ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.021

Р.С. Сергеев[1], А.В. Тузиков[2], В.Ф. Еремин[3]

АЛГОРИТМЫ АНАЛИЗА МУТАЦИЙ В ПЕРВИЧНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ БЕЛКОВ ВИЧ-1 СУБТИПА *A*

Одной из проблем терапии вируса иммунодефицита человека (ВИЧ) является его высокая мутагенность. Появление мутаций в определенных участках генома вируса может стать причиной выработки его устойчивости к антиретровирусным препаратам, применяемым в ходе лечения. Поэтому при выборе схемы терапии необходимо знать, какие из мутаций ведут к резистентности и как они связаны друг с другом. Предлагаемый подход к анализу первичных последовательностей белков ВИЧ позволяет оценить, насколько выявленные мутации связаны с применением конкретных лекарственных препаратов, а также определить взаимозависимые мутации. Одновременно с этим излагается опыт его практической реализации и результаты применения к анализу последовательностей от ВИЧ-инфицированных пациентов из Беларуси.

Введение

Современные способы лечения ВИЧ-инфекции (высокоактивная антиретровирусная терапия) замедляют прогрессирование ВИЧ-инфекции и ее переход в стадию СПИДа (синдрома приобретенного иммунодефицита). Высокая степень изменчивости ВИЧ и, следовательно, появление мутаций в геноме вируса, вызывающих изменения в структуре кодируемых белков, которые отвечают за поддержание процессов жизнедеятельности ВИЧ, являются одними из основных факторов, способных существенно снизить эффективность терапии. Несмотря на то что в настоящее время разработаны и применяются разнообразные схемы лечения, сопротивляемость вируса к каждому из используемых в них препаратов (обычно это ингибиторы протеазы, ингибиторы обратной транскриптазы, ингибиторы слияния, а в последнее время и ингибиторы интегразы) часто ведет к развитию кросс-резистентности [1, 2]. Тем не менее практика применения ингибиторов вирусных белков показывает, что лишь некоторые из мутаций способны вызывать появление устойчивости вируса к лекарственным препаратам.

Существует несколько подходов к определению мутаций резистентности, которые используются при анализе образцов от инфицированных пациентов. Однако системы, позволяющие провести тест на наличие мутаций резистентности, как правило, анализируют последовательности, используя заранее определенные правила, полученные на основании результатов других исследований [3–5]. Стоит отметить, что большинство опубликованных данных касается субтипа *B*, в то время как в Беларуси и некоторых других странах Восточной Европы наибольшее число случаев инфицирования связано с распространением ВИЧ-1 субтипа *A*. Поэтому возникает вопрос, как проанализировать эти данные и насколько правомерно это делать, используя системы, обученные на результатах анализа мутаций вируса субтипа *B*. Как ни странно, несмотря на наличие больших публичных баз данных по анализу ВИЧ-1, не существует доступного программного обеспечения, при помощи которого можно было бы полностью решить эту задачу.

В настоящей работе достигнуты следующие результаты:

1. Предложен подход, основанный на методах математической статистики, который может быть использован для решения задачи по выявлению взаимосвязанных мутаций в геноме ВИЧ-1, а также определению мутаций, ассоциированных с приемом ингибиторов.

2. Разработан вариант программной реализации предложенного подхода. Описаны результаты его применения для анализа первичных последовательностей белков протеазы и обратной транскриптазы ВИЧ-1 субтипа *A*, которые были получены в ходе исследования пациентов из Беларуси.

1. Постановка задачи

Объектом проводимого исследования являются первичные последовательности вирусных белков, полученные в результате трансляции соответствующих участков вирусной ДНК. Существует ген полимеразы pol ВИЧ-1, который кодирует два важных для жизнедеятельности вируса белка-фермента: протеазу и обратную транскриптазу. Существуют лекарства, ингибирующие тот или иной белок, – это ингибиторы протеазы, нуклеозидные и ненуклеозидные ингибиторы обратной транскриптазы. Есть и ингибиторы других ферментов, которые не рассматриваются в данной работе.

При лечении ВИЧ-инфицированных пациентов в обоих ферментах могут появляться мутации к конкретному ингибитору. Это приводит к тому, что такой мутировавший экземпляр вируса становится нечувствительным (резистентным) к данному ингибитору. Эти мутации хорошо описаны для ВИЧ-1 субтипа *B* и их можно найти в публичных базах данных [5, 6]. Например, для препарата азидотимидина (AZT, зидовудин) конкретные мутации будут в позициях 41, 67, 70, 210, 215 и 215 фермента обратной транскриптазы.

Однако вместе с тем могут происходить и вторичные, в том числе компенсаторные, мутации, которые изучены мало. В связи с этим представляет интерес задача разработки подхода и его реализации в виде алгоритмического и программного обеспечения для определения корреляции конкретной мутации в конкретной позиции с конкретными мутациями во всех других позициях.

Например, пусть имеется фрагмент выравнивания аминокислотных последовательностей (рис. 1), тире «–» обозначает ту же аминокислоту, что и в первой последовательности. Очевидно, что в приведенном случае замена *I* на *K* в первой позиции коррелирует с заменой *N* на *L* в десятой позиции.

```
1   5    10    16

IRIIRIKSENVVIQIK
------E-----L-K-
-E---R---M------
-------R--------
KI---R--KL----S-
KE-------E---E--
K----L---L------
K--------L----RR
```

Рис. 1. Пример фрагмента выравнивания последовательностей белков с коррелированными заменами

Если вернуться к рассматриваемой задаче, то резистентные мутации, если они присутствуют одновременно, будут, как правило, коррелировать между собой. Так, мутация в позиции 41 фермента обратной транскриптазы будет коррелировать с заменами в позициях 67, 70, 210 и 215 этого же фермента. Однако корреляции с заменами в других позициях, если они существуют, еще не описаны.

Кроме того, поскольку большинство исследований мутаций резистентности относится к ВИЧ-1 субтипа *B* [7], а в пределах Беларуси большинство случаев инфицирования связано с ВИЧ-1 субтипа *A*, актуальна еще и задача поиска резистентных мутаций непосредственно в последовательностях ВИЧ-1 субтипа *A*. Представляет интерес исследование связи таких мутаций с приемом конкретных ингибиторов, а также сравнение этих результатов с известными данными для других субтипов вируса.

2. Алгоритмы и методы

Чтобы решить поставленную задачу, разобьем ее на части, для каждой из которых будет использован свой набор инструментов. В результате нужно получить ответы на следующие вопросы:

– какие мутации вызваны применением лекарственных препаратов?

– какие мутации являются взаимосвязанными между собой?

– какие мутации привели к появлению устойчивости к применяемым препаратам?

Для решения первой и последней подзадачи воспользуемся методом анализа таблиц сопряженности, который позволяет адекватно оценить интенсивности связей признаков. Идея подхода к выявлению взаимосвязанных мутаций заключается в том, что по множественному выравниванию последовательностей строится филогенетическое дерево с корнем. Тогда листьями этого дерева будут последовательности из исходного выравнивания, а вид последовательностей во внутренних узлах дерева можно предсказать, смоделировав ход эволюции. На основании полученной при этом информации можно оценить коррелированность замен в выбранных позициях последовательности, называемых сайтами.

Предполагается также, что все анализируемые последовательности являются выравненными (т. е. позиции в разных последовательностях сопоставлены друг другу), а к каждой последовательности прилагаются подробное описание места и времени ее получения и наименования препаратов, применявшихся в процессе лечения, если оно проводилось. Точечной мутацией считается любое отличие рассматриваемой последовательности от эталонной последовательности исследуемого субтипа вируса. Учитываются мутации выпадения, замены или вставки нуклеотидов.

2.1. Взаимосвязь с антиретровирусной терапией

Для анализа возможной связи появления выявленных мутаций с антиретровирусной терапией воспользуемся методом анализа таблиц сопряженности.

Чтобы определить, вызвана ли мутация в *i*-й позиции последовательности применением некоторого препарата, построим таблицу сопряженности признаков (рис. 2). Таблица строится для каждой мутации в рассматриваемой позиции отдельно по каждому препарату (если есть необходимость определить, каким именно препаратом вызвана мутация). В качестве признаков выбираются наименование препарата (1 – применялся / 0 – не применялся для терапии) и мутация (1 – присутствие / 0 – отсутствие) в рассматриваемом сайте последовательности. Обозначим через n_{kl} число последовательностей, для которых первый признак имеет значение $k \in \{0,1\}$, а второй признак – значение $l \in \{0,1\}$.

X\Y	**Yes**	**No**	Σ
Yes	n_{00}	n_{01}	n_{0*}
No	n_{10}	n_{11}	n_{1*}
Σ	n_{*0}	n_{*1}	n

Рис. 2. Сопряженности признаков X и Y

Имеющаяся ситуация в простейшем случае может быть описана вероятностной моделью, где вся совокупность случайных элементов таблицы сопряженности $(n_{00},...,n_{11})$ – случайная выборка из полиномиального распределения с вероятностями $(q_{00},...,q_{11})$ и фиксированным числом наблюдений *n*, равным числу рассматриваемых последовательностей.

Положив в качестве основной гипотезу о независимости признаков, выполняем процедуру проверки статистических гипотез. Определим нулевую гипотезу независимости признаков

$$H^0 : q_{ij} = q_{i*}q_{*j},$$

где $q_{i*} = \sum_{j=0}^{1} q_{ij}$, $q_{*j} = \sum_{i=0}^{1} q_{ij}$.

Тогда по данным таблицы строим оценку

$$\chi^2 = n \cdot \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{\left(n_{ij} - n_{i*}n_{*j}\right)^2}{n_{i*}n_{*j}} \sim \chi_1^2$$

при $n \to \infty$.

Тест проверки статистических гипотез имеет вид

$$\begin{cases} H^0 : P\left(\chi_1^2\right) \geq \alpha; \\ \overline{H}^0 : P\left(\chi_1^2\right) < \alpha, \end{cases}$$

где $\alpha \in (0,1)$ – задаваемый уровень значимости.

Тогда *p*-значение вычисляется по формуле $P\left(\chi_1^2\right) = 1 - F\left(\chi_1^2\right)$, где $F\left(\chi_1^2\right)$ – функция распределения χ -квадрат с одной степенью свободы.

Приведенный выше критерий для анализа связи и независимости признаков можно использовать, когда число наблюдений достаточно велико и ожидаемые частоты не слишком малы – $\frac{n_{i*} n_{*j}}{n} > 5$ [8]. В противном случае применяется точный тест Фишера [9] с последующей процедурой проверки статистических гипотез.

При анализе генетических последовательностей тесты проверки статистических гипотез, как правило, приходится повторять многократно. Это относится и к рассматриваемому случаю, когда описанная выше процедура должна выполняться для каждого сайта исследуемого выравнивания, который содержит мутации. Чтобы контролировать количество ошибочных выводов, важную информацию для определения порога принятия решения в отдельном тесте может дать знание общего числа проведенных тестов, а также количество принятых нулевых и альтернативных гипотез. Таким образом, например, можно заключить, что устанавливаемый для каждого теста статистической проверки гипотез уровень значимости α приведет к ошибке, равной αm, где m – число отдельных тестов (соответственно, если $\alpha = 0,05$ и необходимо проверить 100 мутаций на наличие ассоциированности с терапией, получим пять ложных срабатываний). Применение процедуры множественной проверки гипотез позволяет с учетом этой информации скорректировать порог принятия решений для индивидуальных гипотез таким образом, чтобы свести число ошибок для всей совокупности проверяемых сайтов до приемлемого уровня, не потеряв действительно существующие зависимости. Для этого предлагается использовать подход, позволяющий контролировать среднюю долю ложных отклонений нулевых гипотез (*FDR*, false discovery rate). Пусть $H_{(1)}^0, \ldots, H_{(m)}^0$ – множество нулевых гипотез; $H_{(i)}^1 = \overline{H}_{(i)}^0$ – множество альтернативных гипотез. Упорядочим соответствующие им *p*-значения по неубыванию: $p_{(1)} \leq \ldots \leq p_{(m)}$. Тогда $FDR(t) = E\left[\frac{F(t)}{S(t)}\right]$, где $S(t)$ – количество $p_{(i)} \leq t$; $F(t)$ – число истинных нулевых гипотез среди принятых альтернативных гипотез [10], а величина $q_{(k)} = \min\limits_{t \geq p_{(k)}} FDR(t)$ демонстрирует долю ошибок, допущенных среди принятых альтернативных гипотез, если бы уровень значимости для индивидуальных тестов был принят равным $p_{(k)}$.

2.2. Выявление коррелированных мутаций

Выявление коррелированных мутаций в сайтах последовательностей белков интересно по нескольким причинам. Во-первых, совместно эволюционирующие сайты могут оказаться близко расположенными друг от друга, если говорить о пространственной структуре белковой молекулы. Во-вторых, именно связанные друг с другом мутации, действуя совместно, часто вызывают устойчивость вируса к применяемому препарату.

Предположим, что выбрано *N* выравненных последовательностей, принадлежащих к одному субтипу. Выясним, насколько коррелированы сайты *i* и *j* в данных последовательностях. Для этого по имеющемуся множественному выравниванию восстанавливается филогенетическое дерево [11]. Здесь каждой последовательности будет соответствовать лист построенного дерева.

Каждому *i*-му сайту (позиции в последовательности) поставим в соответствие вектор $\mathrm{V}_i=\left(\mathrm{V}_{i1},\mathrm{V}_{i2},...,\mathrm{V}_{iN}\right)$, где V_{ik} – количество мутаций на пути от корня филогенетического дерева к листу $m=\overline{1,N}$.

Выберем из выравнивания пару сайтов (i, j) для анализа (рис. 3). При этом позиции *i* будет соответствовать вектор $\mathrm{V}_i=\left(\mathrm{V}_{i1},\mathrm{V}_{i2},...,\mathrm{V}_{iN}\right)$, а позиции *j* – вектор $\mathrm{V}_j=\left(\mathrm{V}_{j1},\mathrm{V}_{j2},...,\mathrm{V}_{jN}\right)$. Тогда для каждой пары сайтов можно вычислить меру тесноты связи. В простейшем случае такой мерой может служить выборочный коэффициент корреляции $\bar{\rho}_{ij}=\dfrac{\mathrm{cov}\left(\mathrm{V}_i,\mathrm{V}_j\right)}{\sigma_{\mathrm{V}_i}\cdot\sigma_{\mathrm{V}_j}}$, где $\mathrm{cov}\left(\mathrm{V}_i,\mathrm{V}_j\right)$ – ковариация; σ_{V_i} и σ_{V_j} – дисперсии элементов соответствующих векторов. Таким образом, чем ближе по модулю величина $\bar{\rho}_{ij}$ к 1, тем увереннее прослеживается тенденция к синхронности изменений, происходящих в рассматриваемых сайтах в ходе эволюции.

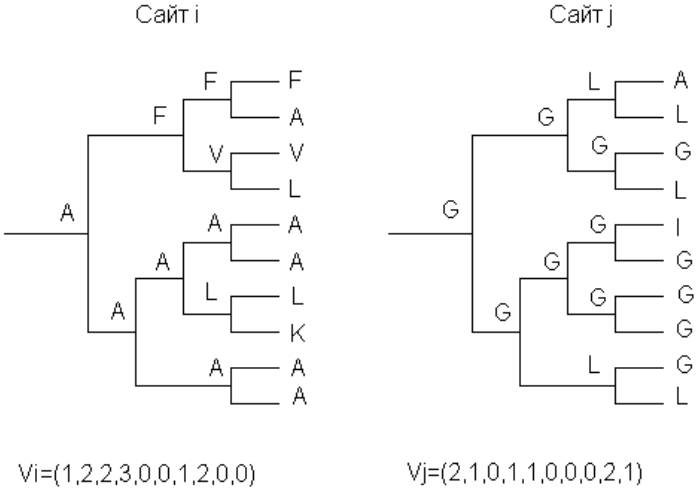

Рис. 3. Определение интенсивностей мутаций в сайтах *i* и *j* по восстановленному филогенетическому дереву

Если сайт *i* коррелирует с сайтом *j*, а *j* – с *r*, то имеем кластер взаимосвязанных мутаций в сайтах *i*, *j*, *r*. Силу их связности можно оценить величиной $\rho=\min\left\{\bar{\rho}_{ij},\bar{\rho}_{jr},\bar{\rho}_{ir}\right\}$.

На практике часто возникают ситуации, когда информации, на основании которой делается вывод о наличии или отсутствии корреляций, недостаточно. Например, величина коэффициента корреляции, рассчитанного указанным выше способом для пары сайтов, для которых векторы V_i и V_j содержат только по одной компоненте, отличной от нуля, будет высока и даже равна единице, если эти векторы совпадают. Однако вызвано это может быть не столько тем, что эти сайты эволюционировали совместно, сколько погрешностями при реконструкции последовательностях в узлах филогенетического дерева или эффектом основателя [12]. Поэтому анализировать имеет смысл только достаточно вариабельные сайты.

Для того чтобы оценить применимость метода для анализа выбранных сайтов, введем коэффициент вариабельности, который будет рассчитываться по формуле $SN_i=10\log_{10}\left(\sum_{k=1}^{N}V_{ik}^2\Big/N\right)$ для каждого *i*-го сайта (логарифм взят для удобства сравнения получаемых значений). Чем больше значение этого коэффициента, тем выше достоверность результатов анализа с участием рассматриваемого сайта. При таком подходе наиболее консервативные сайты автоматически исключаются из рассмотрения.

Для дополнительной проверки полезно применить еще один метод анализа сайтов выравнивания, основанный на подходе к определению взаимной информации сайтов. Ниже приведем методику выполняемых при этом расчетов.

Формально мерой изменчивости *i*-го сайта в выравнивании можно считать энтропию по Шеннону H_i, определенную через вероятности символов $P(x_i)$, которые могут встречаться в аминокислотных последовательностях в позиции номер *i*, где $x_i \in S$. В свою очередь, $S = S_{AA} \cup \{«–», «X»\}$, где S_{AA} – множество всех 20 аминокислот, «–» обозначает мутацию типа выпадение и «*X*» – любой неизвестный символ. Тогда энтропия для всего *i*-го сайта определяется как $H_i = -\sum_{x_i \in S} P(x_i) \log P(x_i)$.

Взаимная энтропия определяется через совместное распределение вероятностей $P(x_i, y_j)$ нахождения символа x_i в позиции *i* и символа y_j в позиции *j* рассматриваемого выравнивания и вычисляется по формуле $H_{ij} = -\sum_{x_i, y_j \in S} P(x_i, y_j) \log P(x_i, y_j)$. Тогда взаимная информация двух сайтов *i* и *j* может быть рассчитана как $MI_{ij} = H_i + H_j - H_{ij}$. Полученная величина всегда неотрицательна и принимает максимальное значение в случае наличия полной согласованности между сайтами. Минимальное значение, равное нулю, достигается, когда сайты эволюционируют полностью независимо либо не имеют мутаций.

Существуют и другие методы, основанные на несколько отличающихся подходах, которые могут быть применены для выявления коррелированных мутаций [13, 14].

3. Детали реализации

Для выполнения расчетов было использовано как существующее свободно распространяемое программное обеспечение, так и собственные разработки. Входящие в состав тестовой системы сторонние компоненты были предварительно установлены из исходных кодов и сконфигурированы под мультипроцессорную платформу. К ним относятся программа ClustalW для выполнения множественного выравнивания, RAxML для филогенетического анализа и пакет GASP для реконструкции последовательностей во внутренних узлах дерева. Собственные разработки представляют собой утилиты, написанные с использованием библиотеки Qt, и скрипты на языке R, которые применяются совместно с перечисленными приложениями и позволяют оценить коррелированность замен в сайтах, выполнить проверку мутаций на ассоциированность с терапией и выделить мутации, которые могут приводить к резистентности.

Предполагается, что исходные последовательности представлены как набор нуклеотидов ДНК – в таком виде они поступают с секвенатора после первичной обработки. Последовательности загружаются для выравнивания в систему в формате FASTA. Множественное выравнивание выполняется программой ClustalW в параллельном режиме [15].

В ходе анализа последовательностей выявляются мутации и проводится их статистический анализ. Обособленно реализован алгоритм анализа последовательностей на предмет выявления коррелированных мутаций. Для этого в виде отдельной разработки реализован метод, который использует эволюционную информацию, полученную из восстановленного филогенетического дерева. Филогенетическое дерево строится методом максимального правдоподобия по набору данных множественного выравнивания средствами пакета RAxML. Для увеличения достоверности результатов применяется процедура бутстрепа. Следует отметить, что построение филогенетического дерева и применение бутстрепа достаточно трудоемкие процедуры, которые выполняются в параллельном режиме [16].

Вид последовательностей-предков во внутренних узлах дерева восстанавливается при помощи программы GASP [17], после чего становится возможным выполнить поиск коррелированных мутаций.

4. Результаты

Для тестирования предложенного подхода на больших выборках данных и проверки достоверности результатов предложенного подхода были использованы последовательности из международных банков данных Стэнфорда и Лос-Аламоса [5, 6]. Из этих источников для целей генотипирования были также получены и включены в выравнивание эталонные последовательности гена pol ВИЧ-1 различных субтипов.

В качестве реальных данных для апробации предложенного подхода был представлен набор последовательностей гена pol, полученных в результате секвенирования образцов от ВИЧ-инфицированных пациентов из Беларуси лабораторией диагностики ВИЧ и сопутствующих инфекций РНПЦ эпидемиологии и микробиологии Республики Беларусь.

Во всех случаях сайты выравнивания, содержавшие большое количество пропусков, по возможности исключались из рассмотрения.

4.1. Анализ взаимосвязи с антиретновирусной терапией

Для анализа взаимосвязи с антиретновирусной терапией было использовано выравнивание, состоящее из 79 последовательностей гена pol ВИЧ-1 субтипа *A*, полученных в результате секвенирования пациентов из Беларуси. Среди них только 18 пациентов проходило курс терапии ингибиторами протеазы и обратной транскриптазы. Все сравнения выполнялись с последовательностями 98UA0116 (Ukraine) и AY500393_03RU20-06-13 (Russia), которые были выбраны в качестве референтных для варианта вируса, распространенного в Беларуси.

В связи с малым объемом данных, доступных для анализа, был использован тест Фишера с последующей процедурой множественной проверки гипотез для выбора порога принятия решений. В качестве порога для принятия альтернативной гипотезы были установлены $p < 0{,}0465$ для мутаций протеазы и $p < 0{,}0023$ для мутаций обратной транскриптазы, что обеспечивало не более одной ошибки среди принятых альтернативных гипотез. Результаты анализа с соответствующими *p*-значениями приведены в табл. 1, где отображены мутации протеазы и обратной транскриптазы ВИЧ-1 субтипа *A*, которые были признаны ассоциированными с приемом соответствующих ингибиторов.

Таблица 1

Мутации ВИЧ-1 субтипа *A*, ассоциированные с терапией ингибиторами протеазы и обратной транскриптазы, по результатам тестирования

Мутации	Позиция	Аминокислота	Пациенты, проходившие терапию, %	Пациенты, не проходившие терапию, %	*P*-значение
Ассоциированные с ингибиторами протеазы	L10	I	6,7	4	0,0092697
	I54	V	2,6	0	0,0464798
Ассоциированные с ингибиторами обратной транскриптазы	D67	N	6,4	1,3	0,0020737
	K103	N	5,1	0	0,0021452
	M184	V	16,4	0	3,246357e-11
	G190	S	5,2	0	0,0022612
	T215	F	5,1	0	0,0021452

4.2. Выявление совместно эволюционирующих сайтов

Для проверки достоверности предложенного подхода к поиску коррелированных замен в сайтах выравнивания было выполнено его тестирование на наборе данных, состоящем из последовательностей протеазы ВИЧ-1 субтипа *B*, а результаты сопоставлены с приведенными в исследовании [18] для этого же набора данных. Сравнительный список из 20 наиболее коррелированных позиций выравнивания представлен в табл. 2. Таким образом, можно видеть, что зависимости, выявляемые при помощи предложенного метода, с точностью до отношений транзитивности обнаруживаются и другими методами.

Таблица 2

Список наиболее коррелированных сайтов в одном и том же выравнивании последовательностей протеазы ВИЧ-1 субтипа *B* по результатам разных исследований

Поз. 1	Поз. 2	Коэф. связи сайтов по результатам исследования [18]			Коэф. $\overline{\rho}_{ij}$	Взаимная информация	Коэф. SN
		Поз.1	Поз.2	Коэф. Phi			
54	82	54	82	0,63	0,751319	0,246775	2,75617
54	90	54	71	0,34	0,676675	0,030537	0,668446
		71	90	0,28			
20	36	20	36	0,41	0,662566	0,100348	-7,26284
35	36	35	36	0,45	0,624214	0,078273	-4,86524
62	90	62	73	0,2	0,595151	0,031533	0,668446
		73	90	0,47			
82	90	82	71	0,26	0,56598	0,017478	0,668446
		71	90	0,38			
32	47	32	47	0,51	0,536928	0,046459	-16,071
10	54	10	54	0,41	0,534811	0,157287	0,872152
10	90	10	90	0,35	0,534618	0,110409	0,668446
62	82	62	73	0,2	0,522542	0,009281	1,03302
		73	71	0,21			
		71	82	0,26			
20	35	20	36	0,41	0,518268	0,043409	-7,26284
		35	35	0,45			
54	62	54	71	0,34	0,517607	0,023783	1,03302
		71	73	0,21			
		73	62	0,2			
10	77	10	93	0,22	0,48648	0,01851	-4,04791
		93	77	0,31			
20	46	20	90	0,22	0,475654	0,036565	-7,26284
		90	10	0,35			
		10	46	0,35			
73	93	73	90	0,47	0,445986	0,009786	-2,78039
		90	93	0,22			
48	62	48	54	0,29	0,440401	0,000261	-12,2689
		54	71	0,34			
		71	73	0,21			
		73	62	0,2			
46	84	46	10	0,35	0,439516	0,026041	-8,64111
		10	84	0,3			
71	90	71	90	0,38	0,437241	0,126232	0,668446
77	90	77	93	0,31	0,426077	0,012273	-4,04791
		93	90	0,22			
10	71	10	71	0,37	0,417702	0,150926	0,705095

Аналогичным образом были проанализированы последовательности участка гена pol вируса ВИЧ-1 субтипа *A* из набора данных, полученных в результате секвенирования образцов от ВИЧ-инфицированных пациентов из Беларуси (табл. 3). Общее выравнивание включало 79 последовательностей.

Таблица 3

Наиболее коррелированные позиции в белках протеазы ВИЧ-1 субтипа А по результатам анализа данных от пациентов из Беларуси

Поз. 1	Поз. 2	Коэф. $\overline{\rho}_{ij}$	Взаимная информация	Коэф. SN
1	2	3	4	5
15	77	0,5306	0,0224	-7,8368
46	89	0,5223	0,1123	-11,1948
63	64	0,4504	0,2163	-7,5150
10	89	0,4056	0,1010	-8,5623
13	15	0,3986	0,0167	-7,8368
20	37	0,3958	0,0928	-8,9763
77	93	0,3542	0,0272	-5,1742

Окончание таблицы 3

1	2	3	4	5
64	89	0,3524	0,1980	-8,5623
10	46	0,3392	0,0514	-11,1948
20	35	0,3375	0,0887	-8,9763
35	63	0,3232	0,0963	-8,1845
10	35	0,3019	0,0519	-8,1845
35	37	0,2913	0,0816	-8,1845
35	89	0,2817	0,0749	-8,5623
63	89	0,2739	0,0969	-8,5623
20	89	0,2605	0,1104	-8,9763
35	46	0,2542	0,0408	-11,1948
35	64	0,2426	0,1212	-8,1845
13	93	0,2066	0,0242	-4,6626
10	20	0,2050	0,0598	-8,9763

Заключение

Одной из проблем подавления ВИЧ является его высокая мутагенность, т. е. способность варьировать свою ДНК и таким образом вырабатывать жизнеспособные мутации даже в неблагоприятных условиях.

В 2000 г. были выработаны рекомендации, призывающие включить в практику лечения больных ВИЧ тестирование на резистентность, результаты которого должны интерпретироваться экспертами.

В настоящей работе предложен метод, позволяющий автоматизировать процесс исследования первичных последовательностей вирусных белков, полученных от ВИЧ-инфицированных пациентов. Особенностью предложенного подхода является применение методов математической статистики для анализа данных и выявления взаимозависимостей, что требует наличия достаточного количества последовательностей. Программно реализованы основные элементы предложенного метода. Для выполнения стандартных процедур использованы хорошо зарекомендовавшие себя пакеты анализа биоинформатических данных с открытым исходным кодом.

Список литературы

1. HIV-1 Protease and Reverse Transcriptase Mutations: Correlations with Antiretroviral Therapy in Subtype B Isolates and Implications for Drug-Resistance Surveillance / S.-Y. Rhee, W.J. Fessel, A.R. Zolopa [et al.] // The Journal of Infectious Diseases. – 2005. – № 192. – P. 456–465.

2. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance / R.W. Shafer [et al.] // AIDS – 2007. – № 21 (2). – P. 215–223.

3. Antiretroviral Drug Resistance Testing in Adult HIV-1 Infection: 2008 Recommendations of an International AIDS Society–USA Panel / M.S. Hirsch, H.F. Gunthard, J.M. Schapiro [et al.] // HIV/AIDS CID – 2008. – № 47. – P. 266-285.

4. Celera ViroSeq HIV-1 Genotyping System [Electronic resource]. – 1998–2011. – Mode of access : https://www.celera.com/cdx/ViroSeq. – Date of access :

5. Stanford University HIV Drug Resistance Database [Electronic resource]. – 1998–2011. – Mode of access : http://hivdb.stanford.edu/. – Date of access :

6. Los Alamos HIV Database [Electronic resource]. – 2005–2011. – Mode of access : http://www.hiv.lanl.gov/content/. – Date of access :

7. International AIDS Society [Electronic resource]. – 2001–2011. – Mode of access : http://www.iasociety.org/. – Date of access :

8. Аптон, Г. Анализ таблиц сопряженности / Г. Аптон. – М. : Финансы и статистика, 1982. – 143 с.

9. Agresti, A. A Survey of Exact Inference for Contingency Tables / A.Agresti // Statistical Science. – 1992. – № 7 (1). – P. 131–153.

10. Storey, D.J. Statistical significance for genomewide studies / D.J. Storey, R.Tibshirani // PNAS 100. – 2003. – № 16. – P. 9440–9445.

11. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach / J.Felsenstein // Mol. Evol. – 1981. – № 17. – P. 368–376.

12. Лукашов, В.В. Молекулярная эволюция и филогенетический анализ / В. В. Лукашов. – М. : Бином, 2009. – 256 с.

13. Correlated Mutations and Residue Contacts in Proteins / U. Gobel [et al] // Proteins: Structure, Functions and Genetics. – 1994. – № 18. – P. 309–317.

14. Pollock, D.D. Coevolving Protein Residues: Maximum Likelihood Identification and Relationship to Structure / D.D. Pollock, W.R. Taylor, N. Goldman // J. Mol. Biol. – 1999. – № 287. – P. 187–198.

15. Parallel CLUSTAL-W for PC Clusters / J. Cheetham [et al.] // LNCS. – 2003. – № 2668. – P. 300–309.

16. Stamatakis, A. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models / A. Stamatakis // Bioinformatics. – 2006. – № 22 (21). – P. 2688–2690.

17. Edwards, R.J. Gapped Ancestral Sequence Prediction for proteins. [Electronic resource]. – 2004. – Mode of access : http://www.biomedcentral.com/1471-2105/5/123. – Date of access : 02.04.2011.

18. Mutation Patterns and Structural Correlates in HIV-1 Protease Following Varying Degrees of Protease Inhibitor Treatment. Journal of Virology / T.D. Wu [et al]. – 2003. – № 77 (8). – P. 4836–4847.

Поступила 08.06.11

[1]*Белорусский государственный университет,*
Минск, пр. Независимости, 4,
e-mail: roma.sergeev@gmail.com

[2]*Объединенный институт проблем информатики НАН Беларуси,*
Минск, ул. Сурганова, 6
e-mail: tuzikov@newman.bas-net.by

[3]*РНПЦ эпидемиологии и микробиологии МЗ Беларуси,*
Минск, ул. Филимонова, 23,
e-mail: veremin@mail.ru

R.S. Sergeev, A.V. Tuzikov, V.F. Eremin

ALGORITMS FOR MUTATION ANALYSIS OF HIV-1 SUBTYPE A PRIMARY PROTEIN SEQUENCES

High variability of human immunodeficiency virus type 1 (HIV-1) is a major obstacle in its therapy. Mutations in specific parts of the virus genome may cause therapy failure because of drug resistance. Therefore, when choosing a therapy modes one needs to know which mutations lead to the drug resistance and how they relate to each other. The proposed approach to analysis of HIV-1 primary protein sequences allows to predict whether the identified mutations are associated with specific drugs and to determine their interdependence. The results of implementation and the preliminary analysis of sequences from HIV-infected patients from Belarus are also presented.