Биомедицинская химия, 2017 том 63, вып. 5, с. 457-460.

КОМПЬЮТЕРНЫЙ ПРОГНОЗ РЕЗИСТЕНТНОСТИ ВИРУСА ИММУНОДЕФИЦИТА ЧЕЛОВЕКА К ИНГИБИТОРАМ ОБРАТНОЙ ТРАНСКРИПТАЗЫ ВИЧ

О.А. Тарасова*, Д.А. Филимонов, В.В. Поройков

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; эл. почта: olga.a.tarasova@gmail.com

ВИЧ, являющийся причиной синдрома приобретённого иммунодефицита человека (СПИД), приводит к смерти более миллиона человек ежегодно. Высокоактивная антиретровирусная терапия (ВААРТ) является золотым стандартом терапии против ВИЧ. Нуклеозидные (НИОТ) и ненуклеозидные (ННИОТ) ингибиторы обратной транскриптазы (ОТ) ВИЧ являются одним из основных компонентов ВААРТ. Эффективность терапии во многом определяется устойчивостью штаммов вируса, возникающей в результате точечных мутаций, которые приводят к изменениям в пространственной структуре белков ВИЧ. Целью данной работы являлась разработка компьютерного подхода к прогнозированию устойчивости варианта ВИЧ с известной последовательностью аминокислот к конкретному антиретровирусному препарату из группы ингибиторов ОТ ВИЧ. Ранее нами был разработан метод компьютерного прогноза устойчивости вариантов ВИЧ к препарату на основе позиционно-специфичных дескрипторов, в которых учитывался однобуквенный код аминокислоты и позиция, определяемая в результате множественного выравнивания различных вариантов ВИЧ. В данной работе предлагается использовать в качестве дескрипторов аминокислотной последовательности пентапептидные фрагменты, что позволяет не проводить предварительно выравнивание последовательностей для получения оценки резистентности мутантного штамма. В качестве обучающей выборки использовали последовательности аминокислот более 1900 вариантов ОТ ВИЧ из базы данных HIV Drug Resistance Database, для которых известны результаты тестов на устойчивость в отношении 10 препаратов. Применение методов машинного обучения (метод опорных векторов, Байесовский подход, “случайный лес”, искусственные нейронные сети) позволило достичь максимальной точности прогноза 99%, при средней точности прогноза 89%.

Ключевые слова: ВИЧ/СПИД, ингибиторы, резистентность, обратная транскриптаза

DOI: 10.18097/PBMC20176305457

ВВЕДЕНИЕ

Вирус иммунодефицита человека (ВИЧ) является серьёзной угрозой для человечества: более 1,8 млн вновь инфицированных ежегодно; более миллиона смертей от ВИЧ/СПИД-ассоциированных заболеваний ежегодно; более 36 млн человек, живущих с ВИЧ [1]. Одной из основных проблем, возникающих при подборе терапии пациентам с ВИЧ, является достаточно быстро нарастающая устойчивость к проводимому лечению [2]. Это существенным образом ограничивает возможности элиминации вируса из организма человека и достижение устойчивого снижения вирусной нагрузки. Поэтому, наряду с разработкой новых эффективных и более безопасных препаратов против ВИЧ, является актуальным прогнозирование резистентности конкретного варианта ВИЧ к проводимой терапии (используемым препаратам и/или их комбинациям). Разработаны методы прогноза резистентности на основе аминокислотной последовательности белков ВИЧ (комбинации препаратов, режим дозирования и т.п.) [3, 4]. Ранее нами был разработан метод компьютерного прогноза устойчивости варианта ВИЧ к антиретровирусным препаратам на основе анализа последовательностей обратной транскриптазы (ОТ) ВИЧ [5]. Подход основан на представлении конкретного варианта аминокислотной последовательности ВИЧ в виде множества позиционно-специфичных дескрипторов (комбинации однобуквенного кода аминокислоты и позиции этой аминокислоты в последовательности по результатам множественного выравнивания). В настоящей работе, в отличие от ранее разработанных методов, мы использовали в качестве дескрипторов аминокислотной последовательности набор пентапептидных фрагментов. Применение в качестве дескрипторов фрагментов аминокислотной последовательности небольшой длины даст возможность избежать процедуры выравнивания, и непосредственно осуществить прогноз резистентности, что снизит вероятность ошибок прогноза, обусловленных несовершенством выравнивания аминокислотных последовательностей. Была исследована точность прогноза на основе разработанного нами метода с использованием различных методов машинного обучения: метод опорных векторов, байесовский классификатор, “случайный лес”, искусственные нейронные сети.

МЕТОДИКА

В данной работе в качестве обучающей выборки мы использовали 1900 последовательностей обратной транскриптазы ВИЧ, для которых соответствующие варианты ВИЧ были протестированы на устойчивость к препаратам в тест-системе Phenosense, содержащихся в Стэнфордской базе данных (БД) HIV Drug Resistance

* - адресат для переписки

Database (СтБД) [6]. Выбор данных, полученных с применением именно этой тест-системы, обусловлен более высокой точностью прогноза резистентности, продемонстрированной нами ранее [5], по сравнению с точностью, рассчитанной с применением данных по устойчивости тест-системы Antivirogram (данные с применением тест-системы Antivirogram также доступны в StDB).

В СтБД содержатся данные об устойчивости последовательностей ОТ ВИЧ к десяти зарегистрированным антиретровирусным лекарственным препаратам (в скобках приведены общепринятые сокращения латиницей): ламивудин (3TC), абакавир (ABC), зидовудин (AZT), ставудин (D4T), диданозин (DDI), эфавиренц (EFV), этравирин (ETR), невирапин (NVP), рилпивирин (RPV) и тенофовир (TDF). Каждая последовательность ОТ характеризуется значением FR для каждого антивирусного препарата, обозначающим отношение (IC_{50}/IC_{50_WT}) IC_{50} этого препарата по отношению к ОТ с данной последовательностью к IC_{50_WT} ОТ дикого типа ВИЧ .

Около 38% аминокислотных последовательностей выборки Phenosense содержат несколько (от 2-х до 4-х) символов, обозначающих однобуквенный код аминокислот в одной позиции. Мы предполагали равновероятным появление каждой из этих аминокислот в данной позиции. Для того, чтобы сформировать выборку последовательностей, в которой в каждой позиции находится только одна аминокислота, нами были сгенерированы все возможные варианты последовательностей с включением в каждую позицию по одному символу – однобуквенному коду с присвоением каждой вновь сгенерированной последовательности нового идентификатора и значения FR исходной последовательности, в которой содержались несколько кодов аминокислотных остатков в одной позиции. Таким образом, нами получено 14223 возможных значений аминокислотных последовательностей ОТ ВИЧ-1 с соответствующими значениями FR. При этом для некоторых последовательностей отсутствуют данные о FR в отношении того или иного препарата, поэтому общее количество последовательностей в выборке для конкретного препарата меньше, чем суммарное число последовательностей.

В данной работе мы проводим оценку принадлежности заданной аминокислотной последовательности к классам “устойчивый” или “чувствительный” к терапии конкретным препаратом. Для построения соответствующих моделей необходимо было задать пороговое значение FR, при превышении которого, последовательность попадает в класс “устойчивый”. Величины FR соответствуют значениям, приведённым в ранее опубликованной работе [5], где проведена оценка применимости различных пороговых значений FR для построения наиболее точных моделей. Выбранные пороговые значения FR и количество последовательностей в обучающих выборках для каждого препарата указаны в таблице 1.

Из последовательностей обучающих выборок были сгенерированы пентапептиды. Выбор пентапептидов обусловлен тем, что именно комбинация пентапептидов, позволяет описать последовательность наиболее компактно и при этом учесть вариабельность аминокислот в отдельных позициях согласно предшествующим исследованиям в области анализа аминокислотных последовательностей, [7]. Учитывая, что аминокислотные остатки определены в среднем для первых 260 позиций каждой последовательности, генерацию производили для первых 260 позиций. В результате было сгенерировано 4122 уникальных пентапептида, составивших словарь пентапептидов. На основе словаря пентапетидов были реализованы дескрипторы (фингерпринты): наличие или отсутствие конкретного пентапептида из данной аминокислотной последовательности в словаре пентапептидов обозначали как “1” или “0” соответственно.

В качестве методов машинного обучения были выбраны следующие: метод опорных векторов, Байесовский классификатор, “случайный лес”, вложенные нейронные сети. Обучение классификаторов проводили в среде Weka 3.8.0 [8]. Тестирование метода прогноза проводили десятикратным методом разбиения на обучающую и тестовую выборку в пропорции 19:1.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Результаты прогноза устойчивости ВИЧ с данной последовательностью по отношению к отдельным препаратам приведены в таблице 2. В процессе

Таблица 1. Характеристики выборок аминокислотных последовательностей ВИЧ

Препарат	FR[1]	Общее количество последовательностей[2]	Чувствительные варианты	Устойчивые варианты
3TC	1,5	13934	777	13157
ABC	8,5	5555	1324	2231
AZT	1,5	13968	1745	12223
D4T	5	13959	7587	6372
DDI	1,5	13963	1672	12291
EFV	10	14067	2089	11978
ETR	10	511	430	81
NVP	1,5	14078	1854	12224
RPV	3	178	96	82
TDF	2	13702	7359	6343

Примечание: 1 - указанное значение FR взято из статьи O.A. Tarasova et al., 2017 [5]; 2 - количество последовательностей с точно определённым значением FR.

Таблица 2. Результаты прогноза (сбалансированная точность) принадлежности аминокислотных последовательностей ОТ ВИЧ-1 к классам устойчивых вариантов ВИЧ

Препарат	I[1]	II[2]	III[3]	IV[4]
3TC	0,8	0,84	0,91	0,91
ABC	0,84	0,89	0,95	0,95
AZT	0,87	0,86	0,99	0,96
D4T	0,94	0,92	0,98	0,96
DDI	0,88	0,88	0,96	0,96
EFV	0,86	0,90	0,99	0,96
ETR	0,87	0,85	0,92	0,89
NVP	0,91	0,88	0,94	0,96
RPV	0,74	0,69	0,80	0,80
TDF	0,82	0,82	0,94	0,92
Средняя точность	0,85	0,85	0,94	0,93

Примечание: 1 - метод опорных векторов; 2 - байесовский классификатор; 3 -случайный лес; 4 - вложенные нейронные сети

моделирования были выбраны параметры классификаторов, обеспечивающие наилучшую точность прогноза. Для метода опорных векторов выбраны следующие параметры: функция оптимизации – метод стохастического градиентного спуска, для классификатора “случайный лес”: 1000 деревьев, для вложенных нейронных сетей: 3 слоя по 100 нейронов с двумя вложенными слоями по 5 нейронов.

Как следует из таблицы 2, несмотря на несбалансированность обучающих выборок, точность прогноза превышала 90% (максимально – 97%) для классификаторов “случайный лес” и вложенных нейронных сетей. Для метода опорных векторов и Байесовского классификатора для большинства выборок точность прогноза варьирует от 80% до 99% в зависимости от препарата, устойчивость к которому оценивалась. Полученная точность, в среднем, сопоставима с точностью прогноза устойчивости вариантов ВИЧ на основе аминокислотных последовательностей ранее опубликованных методов [3-5]. Точность прогноза в настоящей работе сопоставима с точностью прогноза, достигнутой в ранее опубликованной нами работе [5], однако несколько ниже для диданозина, этравирина и тенофовира. Необходимо дополнительное исследование применимости пентапептидов в качестве дескрипторов в алгоритме PASS, представленном в ранее опубликованной работе, для более корректного сравнения результатов.

Наименьшие значения точности были получены для препаратов: этравирин, рилпивирин, тенофовир, причём точность прогноза в отношении устойчивости этих препаратов минимальна при всех используемых классификаторах. Мы считаем, что снижение точности прогноза устойчивости к этравирину и рилпивирину обусловлено относительно небольшим количеством последовательностей как устойчивых, так и чувствительных вариантов ВИЧ-1 в обучающих выборках. Вероятно, точность прогноза в значительной степени зависит от полноты и согласованности данных и пропорции последовательностей, содержащих два или более символов однобуквенных кодов аминокислот в конкретной позиции (например, такие причины могут обусловливать относительно низкую точность для тенофовира).

Относительно высокая точность прогноза устойчивости к большинству препаратов, по нашему предположению, может быть обусловлена, во-первых, относительно небольшим числом пентапептидов, сгенерированных из множества последовательностей обучающей выборки. При общем числе последовательностей более 14000 – уникальных пентапептидов было сгенерировано чуть более 4000. Это указывает на то, что вариабельные участки в последовательности ОТ ВИЧ-1, ассоциированные с конкретными позициями в последовательности и заменами в них, вероятнее всего, являются источником наибольшего разнообразия пентапептидов и вносят наибольший вклад в получаемые оценки.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

В настоящем исследовании приведены результаты прогноза устойчивости вариантов ВИЧ к антиретровирусным препаратам. Прогноз устойчивости выполняли на основе аминокислотных последовательностей вариантов ВИЧ, представленных в виде множества пентапептидов. Это позволяет не применять для оценки устойчивости множественное выравнивание аминокислотных последовательностей с последующим выявлением аминокислотных замен в позициях, для которых известно, что изменения в них приводят к устойчивости. Показано, что предлагаемый метод обладает точностью прогноза более 90% для большинства препаратов – ингибиторов обратной транскриптазы.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке гранта РНФ 17-75-10187.

ЛИТЕРАТУРА

1. http://www.unaids.org/ru (31/08/2017)
2. *Garbelli A., Riva V., Crespan E., Maga G.* (2017) Biochem. J., **474**, 1559-1577.
3. *Heider D., Verheyen J., Hoffmann D.* (2010) BMC Bioinform., **11**, 37.
4. *Van Westen G., Hendriks A., Wegner J., Ijzerman A., van Vlijmen H., Bender A.* (2013) PLoS Comput. Biol., **9**, e1002899.
5. *Tarasova O., Filimonov D., Poroikov V.* (2017) J. Bioinform. Comput. Biol., **15**, 1650040.
6. *Rhee S., Gonzales M., Kantor R., Betts B., Ravela J., Shafer R.* (2003) Nucl. Acids Res., **31**, 298-303.
7. *Khrustalev V., Khrustaleva T., Barkovsky E.* (2013). Biochimie, **95**, 1745-1754.
8. *Smith T., Frank E.* (2016) Meth. Mol. Biol., **1418**, 353-378.

Поступила: 31. 08. 2017.
Принята к печати: 19. 09. 2017.

COMPUTATIONAL PREDICTION OF HUMAN IMMUNODEFICIENCY RESISTANCE TO REVERSE TRANSCRIPTASE INHIBITORS

O.A. Tarasova, D.A. Filimonov, V.V. Poroikov

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; e-mail: olga.a.tarasova@gmail.com

Human immunodeficiency virus (HIV) causes acquired immunodeficiency syndrome (AIDS) and leads to over one million of deaths annually. Highly active antiretroviral treatment (HAART) is a gold standard in the HIV/AIDS therapy. Nucleoside and non-nucleoside inhibitors of HIV reverse transcriptase (RT) are important component of HAART, but their effect depends on the HIV susceptibility/resistance. HIV resistance mainly occurs due to mutations leading to conformational changes in the three-dimensional structure of HIV RT. The aim of our work was to develop and test a computational method for prediction of HIV resistance associated with the mutations in HIV RT. Earlier we have developed a method for prediction of HIV type 1 (HIV-1) resistance; it is based on the usage of position-specific descriptors. These descriptors are generated using the particular amino acid residue and its position; the position of certain residue is determined in a multiple alignment. The training set consisted of more than 1900 sequences of HIV RT from the Stanford HIV Drug Resistance database; for these HIV RT variants experimental data on their resistance to ten inhibitors are presented. Balanced accuracy of prediction varies from 80% to 99% depending on the method of classification (support vector machine, Naive Bayes, random forest, convolutional neural networks) and the drug, resistance to which is obtained. Maximal balanced accuracy was obtained for prediction of resistance to zidovudine, stavudine, didanosine and efavirenz by the random forest classifier. Average accuracy of prediction is 89%.

Key words: HIV/AIDS, inhibitors, resistance, reverse transcriptase