УДК 519.217.2

А.А. ВАГИС, А.М. ГУПАЛ, И.В. СЕРГИЕНКО

ОПРЕДЕЛЕНИЕ ГРУПП РИСКОВ ПРИ ЗАБОЛЕВАНИЯХ, СОПУТСТВУЮЩИХ COVID-19

Аннотация. Для каждого заболевания существует определенный набор генов, мутации в которых увеличивают риск развития болезни. Массовое секвенирование ДНК больных и здоровых людей привело к определению генов, связанных с конкретными заболеваниями. Описаны эффективные процедуры определения мутаций и их месторасположения в последовательностях генов исследуемых пациентов. Предложено использовать оптимальную байесовскую процедуру определения групп рисков при конкретных заболеваниях, в том числе сопутствующих COVID-19.

Ключевые слова: секвенирование ДНК, точечные мутации, байесовская процедура распознавания.

ВВЕДЕНИЕ

Возможность быстро расшифровать индивидуальный геном человека позволила накапливать большие массивы данных о заболеваниях и связанных с ними мутациях в генах его ДНК. Известно, что мутации в ДНК вызывают тысячи генетических заболеваний, а также влияют на работу иммунной системы человека. Коронавирусы — это покрытые оболочкой РНК-вирусы, которые вызывают респираторные заболевания различной степени тяжести: от обычной простуды до пневмонии с летальным исходом. Вирус COVID-19, впервые зарегистрированный в конце 2019 г. в Ухане (Китай), интенсивно распространяется по всему миру. Исследователи до сих пор изучают, насколько легко этот вирус передается от человека к человеку или насколько устойчивой будет его циркуляция в популяции.

У человека, заразившегося COVID-19, симптомы могут быть слабовыраженными или полностью отсутствовать, хотя у некоторых пациентов наблюдается тяжелое протекание болезни с неблагоприятным прогнозом. Симптомы COVID-19 включают лихорадку, кашель и одышку. У пациентов с более тяжелой формой заболевания могут регистрироваться лимфопения и характерные для пневмонии изменения при диагностических визуализирующих исследованиях. Точный инкубационный период заболевания COVID-19 неизвестен; предположительно он варьируется от одного до 14 дней. У возрастной группы пациентов увеличивается риск иметь тяжелую форму болезни. Диагностика проводится с помощью ПЦР-тестов выделений из верхних и нижних дыхательных путей и сыворотки крови.

В группу риска пациентов, больных COVID-19, входят лица с хроническими заболеваниями сердечно-сосудистой, дыхательной и эндокринной систем, а также с онкологическими патологиями, иммунодефицитными состояниями и почечной недостаточностью.

ЗАБОЛЕВАНИЯ, ВЫЗВАННЫЕ МУТАЦИЯМИ В ГЕНАХ

В настоящее время в развитых станах мира проводится расшифровка (секвенирование) геномов большого количества людей. Полученную информацию используют для ранней диагностики различных заболеваний, в первую очередь, онкологических. Основной задачей в этой области является определение

генетических (или врожденных) предрасположенностей к таким сложным системным заболеваниям, как болезни сердечно-сосудистой системы, рак, диабет и шизофрения. Для каждого заболевания существует конкретный набор генов, мутации в которых увеличивают риск развития болезни. Массовое секвенирование ДНК больных и здоровых людей привело к определению генов, связанных с конкретными заболеваниями, в том числе и с возникающими при COVID-19.

Наиболее распространенным типом мутаций, которые приводят к заболеваниям, являются точечные мутации, в результате которых единичный нуклеотид гена заменяется другим нуклеотидом. Точечные мутации могут возникать в результате спонтанных мутаций, происходящих во время репликации ДНК, а также в результате действия мутагенов, например, влияния ультрафиолетового или рентгеновского излучений, высокой температуры или химических веществ.

В работах [1, 2] использовались данные интернет-ресурсов, где заболеваниям ставились в соответствие связанные с ними мутации в ДНК, т.е. были получены пары исходных и мутированных триплетов нуклеотидов и соответственно кодируемых ими аминокислот. Исследовались мутации, обусловленные аутоиммунными, онкологическими, сердечно-сосудистыми, генетическими, нейродегенеративными заболеваниями, а также психическими расстройствами и пагубными привычками.

С помощью генетических алгоритмов были получены оптимальные генетические коды, помехоустойчивость которых на 8.5 % выше, чем у стандартного кода. С использованием баз данных генетических заболеваний стандартным кодом проверялись приблизительно 400 мутаций для различных типов болезней и почти половина из них привела к нарушению полярности или к мутациям третьего нуклеотида (аминокислота при этом не изменяется, однако прерывается процесс вырезания или сплайсинга интронов) [3]. Оптимальные коды исправляют нарушение полярности при мутациях первого и второго нуклеотидов в кодоне, но избавиться от мутаций в третьем нуклеотиде нельзя. В табл. 1 приведены оценки мутаций для сердечно-сосудистых заболеваний, полученные с помощью стандартного генетического кода. (Аналогичные таблицы можно представить для перечисленных ранее болезней.)

БАЙЕСОВСКИЕ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ

В работах [4, 5] показано, что байесовская процедура распознавания является оптимальной. Для обоснования этого результата необходимо было вывести верхнюю оценку погрешности байесовской процедуры распознавания и получить нижнюю оценку сложности класса задач. Для простоты рассмотрим следующую задачу с булевыми переменными.

Пусть имеется конечное множество X объектов b. Каждый объект $x \in X$ отождествляется с булевым вектором $(x_1, x_2, \ldots, x_n, f)$, где n — натуральное число. Предположим, что на множестве X задано распределение вероятностей P, которое неизвестно. Из множества X образована обучающая выборка V. Пусть некоторый объект получен из множества X независимо от выборки V в соответствии с распределением P, причем известны лишь значения признаков $x_1, x_2, \ldots, x_n$. Требуется по этим значениям и по обучающей выборке V определить значение целевого признака f (состояние объекта x).

Полагаем, что процесс распознавания целевого признака f объекта по известным признакам x осуществляется с помощью функции $A(x)$ по формуле $f = A(x)$. Обучающая выборка $V = (V_0, V_1, V_2)$ имеет следующий вид:

Таблица 1. Оценки мутаций для сердечно-сосудистых заболеваний

Ген	Идентификатор мутации	Кодон	Мутация кодона	Стандартный код
KL	rs 953614	TTT	GTT	$+^*$
KL	rs 9527025	TGC	TCC	$-^*$
ARHGA	rs 2774279	AG**G**	AG**A**	c^*
PCSK9	rs 505151	GGG	GAG	–
APOB	rs 5742904	CGG	CAG	+
APOB	rs 12713559	CGC	TGC	–
LDLR	rs 28940776	GGT	GAT	–
LDLR	rs 28942081	GGC	GAC	–
LDLR	rs 28942082	GGC	GTC	+
TLR4	rs 4986790	GAT	GGT	–
SH2B3	rs 3184504	TGG	CGG	–
BRAP	rs 3782886	AG**A**	AG**G**	c
CHRNA	rs 1051730	TA**C**	TA**T**	c
F5	rs 6025	CGA	CAA	+
GNB3	rs 5443	TC**C**	TC**T**	c
PRKCH	rs 2230500	GTA	ATA	+

Примечание: $+^*$ — сохранение полярности, $-^*$ — нарушение полярности, c^* — сохранение аминокислоты при мутации третьего нуклеотида.

• V_0 — булева матрица размера $m_0 \times n$, где m_0 — количество строк, каждая из которых является вектором $x = (x_1, x_2, \dots, x_n, f)$, выбранным в соответствии с распределением P при условии $f = 0$;

• V_1 — булева матрица размера $m_1 \times n$, где m_1 — количество строк, каждая из которых является вектором x, выбранным на основе распределения P при условии $f = 1$;

• V_2 — булев вектор размерности m_2, каждая компонента которого является наблюдаемым значением состояния f, выбираемым в соответствии с распределением P. Можно считать, что $m_2 = m_0 + m_1$.

Индуктивный шаг. Требуется построить такую процедуру индуктивного вывода, которая по измерениям $x_1, x_2, \dots, x_n$ любого следующего объекта и обучающей выборки $V = (V_0, V_1, V_2)$ определит состояние f объекта.

Пусть $d = (d_1, d_2, \dots, d_n)$ — булев вектор. Полагаем, что распределения P при каждом d удовлетворяют условию

$$P(x_1 = d_1, x_2 = d_2, \dots, x_n = d_n \mid f = i) = \prod_{j=1}^{n} P(x_j = d_j \mid f = i), \ i = 0, 1,$$

что означает независимость признаков x_j для каждого класса объектов; здесь $P(x = d \mid f = i)$ — условная вероятность. Рассмотрим случайные величины $\xi(d, i)$, которые зависят от параметров d и i:

$$\xi(d, i) = \left(\frac{k(i)}{m_2} \right) \prod_{j=1}^{n} \left(\frac{k(d_j, i)}{m_i} \right), \ i = 0, 1, \qquad (1)$$

здесь $k(d_j, i)$ — количество значений, равных d_j, j-го признака в j-м столбце матрицы V_i; $k(i)$ — количество значений целевого признака, равных i, в векторе V_2. Тогда функция распознавания определяется формулой

$$A(d)=\begin{cases}0, & \text{если } \xi(d,0)\ge\xi(d,1),\\ 1, & \text{если } \xi(d,0)<\xi(d,1).\end{cases} \tag{2}$$

Процедуру обучения, определяемую соотношениями (1), (2), обозначим Q_B. Заметим, что величины $\xi(d,i)/(\xi(d,0)+\xi(d,1))$ являются приближенными значениями вероятностей $P(f=i\,|\,x_1=d_1,x_2=d_2,\dots,x_n=d_n)$, вычисленных по формуле Байеса, поэтому процедуру распознавания Q_B называем байесовской. В [3] показано, что для верхней оценки погрешности Q_B выполняется неравенство

$$v(Q_B,C)\le\min\left(1,\ a\sqrt{\frac{n}{m_0}+\frac{1}{m_2}}\right), \tag{3}$$

где a — абсолютная константа. Нижняя оценка сложности класса задач отличается от (3) на абсолютную константу, поэтому в этом смысле байесовская процедура Q_B оптимальна.

ОПРЕДЕЛЕНИЕ ГРУПП РИСКОВ ПРИ ЗАБОЛЕВАНИИ COVID-19

Упрощенный вариант без интронов. Проанализировав табл. 1, можно сделать вывод о том, что у пациентов с диагнозом сердечно-сосудистое заболевание, заразившихся COVID-19, высока вероятность точечных мутаций в определенных генах. Данные об этих пациентах можно внести в обучающую выборку V_1 «больные», причем разбить ее на возрастные группы, а в выборку V_0 «здоровые» внести данные о пациентах с отрицательным результатом ПЦР-тестов и также с учетом их возраста.

Полагаем, что гены в первом столбце табл. 1 являются признаками для байесовской процедуры. Для исключения тривиальных случаев считаем, что в выборке V_0 для каждого гена в табл. 1 есть представители с мутациями в этом гене. Аналогично полагаем, что в выборке V_1 имеются данные о пациентах без мутаций в этом гене.

Выберем первый ген в табл. 1 и рассмотрим выборку V_0. При сравнении последовательностей первого гена у отдельных представителей выборки V_0 с его последовательностью у исследуемого пациента можно получить следующие результаты:

- 0 — отсутствие изменений или мутаций;
- 1 — наличие одной мутации;
- 2 — наличие двух мутаций.

Поскольку мутации появляются случайным образом в последовательности гена, мала вероятность появления мутаций в одном и том же месте последовательности гена у двух разных людей. (Заметим, что длина отдельного гена в ДНК человека может превышать десятки тысяч нуклеотидов.) Наличие числа 2 при сравнении означает, что у пациента есть мутации в первом гене. Поэтому в формуле (1) $k(d_1,0)$ равно количеству двоек, полученных при сравнении. Аналогично в выборке V_1 при сравнении с пациентом $k(d_1,1)$ тоже равно количеству двоек.

Если в выборке V_0 при сравнении с исследуемым пациентом двойки не появляются, то это означает отсутствие мутации у него и $k(d_1,0)$ равно количеству нулей при подсчете в данной выборке. Тогда для выборки V_1 при сравнении с пациентом двойки тоже не появляются и $k(d_1,1)$ равно количеству нулей при подсчете в этой выборке.

Описанную схему вычислений применяем для всех генов, представленных в табл. 1, и определяем значения $\xi(d,i)$ для выборок V_0 и V_1. Результат байесовской процедуры для пациента получаем по формуле (2).

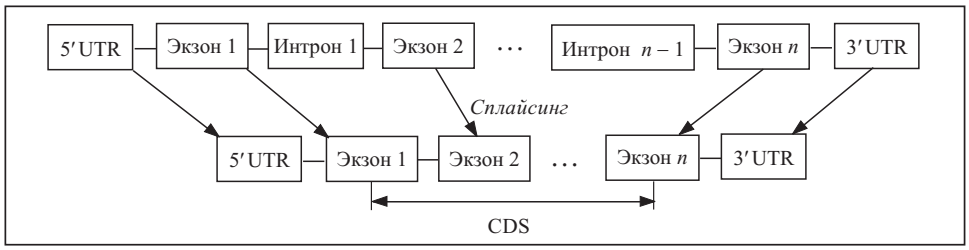

Рис. 1. Структура гена в ДНК и матричной РНК

Общий случай. Гены представляют собой участки ДНК длиной до нескольких десятков тысяч оснований. Последовательность нуклеотидов в участках ДНК определяет структуру конкретного белка. Структура генов усложнялась в процессе эволюции, в результате чего участки ДНК, кодирующие генетическую информацию для эукариотов (организмы, клетки которых содержат ядро, в частности, растения и животные), имеют сложный вид (рис. 1). В соответствии с выполняемой при синтезе белков функции выделяются основные составляющие генов эукариотов:

• начальный и конечный некодирующие участки (untranslated region), обозначенные 5′ UTR и 3′ UTR соответственно, не участвующие в процессе кодирования белков, но косвенно влияющие на него;

• экзоны — участки ДНК, непосредственно кодирующие последовательность аминокислот, из которых состоит белок, с помощью стандартного генетического кода;

• интроны — участки ДНК, расположенные между экзонами, не принимающие участия в синтезе белков. (В настоящее время их предназначение не выяснено, возможно, интроны являются механизмом защиты от мутаций.)

У человека на один ген приходится приблизительно семь экзонов. Длина интронов более чем в 10 раз превышает длину экзонов. В [3] описаны случаи, когда мутации происходят в интронах или на экзон-интронных границах и прерывают процесс вырезания (сплайсинга) интронов, а также вызывают различные заболевания. Интроны GU–AG и AU–AC встречаются в кодирующих белок генах эукариотов. В большинстве интронов РНК 5′–GU–3′ — первые два нуклеотида последовательности интрона, а 5′–AG–3′ — последние два. Поэтому их называют интронами GU–AG и все члены этого класса вырезаются одинаково. Эта особенность была обнаружена после того, как открыли интроны, и предполагалось, что они будут важны в процессе сплайсинга.

Например, мутация G или T в ДНК-копии 5′-сайта вырезания интрона GU–AG либо мутация A или G в 3′-сайте вырезания прервет процесс сплайсинга, потому что правильная экзон-интронная граница не будет опознаваться. В работах [6, 7] предложены методы распознавания фрагментов генов на основе моделей Маркова со скрытыми переменными.

Сравнивая последовательности первых генов двух представителей выборки V_0 (V_1), на компьютере определяем количество и месторасположения обнаруженных мутаций, полагая, что вероятность совпадения точечных мутаций в одном месте мала. Сопоставляя последовательность первого гена третьего представителя с отмеченной последовательностью, определяем для него количество мутаций и их расположение. Аналогично находим количество мутаций и их расположение для всех представителей выборки V_0 (V_1), а также для исследуемого пациента.

Можно определить число мутаций для трех представителей, а затем для всех остальных участников. При сравнении данных первого и второго представителя выборки получаем уравнение $M_1 + M_2 = S_1$, где S_1 — сумма мутаций, данных первого и третьего представителя — уравнение $M_1 + M_3 = S_2$, второго и третьего — уравнение $M_2 + M_3 = S_3$. Решая систему из трех уравнений, находим $M_1 = S_1 + S_2 - S_3$, $M_2 = S_1 + S_3 - S_2$, $M_3 = S_2 + S_3 - S_1$.

Заметим, что в процессе вычисления на основе байесовской процедуры необходимо учитывать мутации для представителей выборок V_0 (или V_1) и у исследуемого пациента, которые имели место в экзонах или на экзон-интронных границах, и не учитывать мутации в интронах, не влияющих на процесс возникновения заболеваний.

Таким образом, зная количество мутаций у исследуемого пациента, определяем значения $\xi(d, i)$ для первого гена на основе информации в выборках V_0 и V_1. Описанную схему вычислений применяем для всех генов, приведенных в табл. 1, и определяем значения $\xi(d, i)$ для состояний $i = 0, 1$. Результат байесовской процедуры для исследуемого пациента получаем по формуле (2).

ЗАКЛЮЧЕНИЕ

Для каждого заболевания существует конкретный набор генов, мутации в которых увеличивают риск развития болезни. Массовое секвенирование ДНК больных и здоровых людей привело к определению генов, связанных с конкретными заболеваниями, в том числе сопутствующими COVID-19. У лиц с установленным диагнозом, переболевших COVID-19, с высокой долей вероятности имеют место точечные мутации в определенных генах.

Предложенные процедуры определения мутаций и их месторасположения в последовательностях генов позволят решать важные проблемы: провести детальный статистический анализ (в том числе для возрастной группы пациентов) относительно количества мутаций в кодирующих участках генов (экзонах) и в интронах, а также подтвердить гипотезу о защитных механизмах в интронах.

Поскольку байесовская процедура широко применяется при прогнозировании в медицине и биоинформатике [8, 9], предлагается использовать ее для определения групп рисков при заболеваниях, сопутствующих COVID-19. Описанную методику можно применять для определения групп рисков пациентов с различными заболеваниями, не связанными с COVID-19.

СПИСОК ЛИТЕРАТУРЫ

1. Сергиенко И.В., Гупал А.М., Островский А.В. Устойчивость генетического кода к точечным мутациям. *Кибернетика и системный анализ*. 2014. Т. 50, № 5. С. 17–24.
2. Сергиенко И.В., Белецкий Б.А., Гупал А.М., Гупал Н.А. Оптимальные помехоустойчивые генетические коды. *Кибернетика и системный анализ.* 2019. Т. 55, № 1. С. 44–50.
3. Браун Т.А. Геномы. Москва-Ижевск: Институт компьютерных исследований, 2011. 924 с.
4. Гупал А.М., Пашко С.В., Сергиенко И.В. Эффективность байесовской процедуры распознавания. *Кибернетика и системный анализ.* 1995. № 4. С. 76–89.
5. Сергиенко И.В., Гупал А.М., Пашко С.В. О сложности задач распознавания образов. *Кибернетика и системный анализ.* 1996. № 4. С. 70–88.
6. Сергиенко И.В., Гупал А.М., Островский А.В. Распознавание фрагментов генов в ДНК с применением моделей Маркова со скрытыми переменными. *Кибернетика и системный анализ*. 2012. Т. 48, № 3. С. 58–67.
7. Гупал А.М., Островский А.В. Использование композиций моделей Маркова для определения функциональных участков генов. *Кибернетика и системный анализ.* 2013. Т. 49, № 5. С. 61–68.
8. Гупал А.М., Гупал Н.А., Тарасов А.Л. Байесовские процедуры распознавания гематологических заболеваний. *Кибернетика и системный анализ*. 2017. Т. 53, №. 6. С. 118–124.

9. Гридина Н.Я., Гупал А.М., Тарасов А.Л., Ушенин Ю.В. Анализ нейрохирургических патологий с применением байесовских процедур распознавания для показателей поверхностного плазмонного резонанса при агрегации клеток крови. *Кибернетика и системный анализ*. 2020. Т. 56, №. 4. С. 35–45.

Надійшла до редакції 09.11.2020

О.А. Вагіс, А.М. Гупал, І.В. Сергієнко
ВИЗНАЧЕННЯ ГРУП РИЗИКІВ ДЛЯ ЗАХВОРЮВАНЬ, СУПУТНИХ COVID-19

Анотація. Для кожного захворювання існує певний набір генів, мутації в яких збільшують ризик розвитку хвороби. Масове секвенування ДНК хворих і здорових людей допомогло визначити гени, пов'язані з конкретними захворюваннями. Описано ефективні процедури визначення мутацій та їхнього розташування в послідовності генів досліджуваних пацієнтів. Запропоновано використовувати оптимальну баєсівську процедуру визначення груп ризиків для конкретних захворювань, зокрема супутних COVID-19.

Ключові слова: секвенування ДНК, точкові мутації, баєсівська процедура розпізнавання.

A.A. Vagis, A.M. Gupal, I.V. Sergienko
DETERMINATION OF GROUPS OF RISKS AT THE DISEASES COVID-19

Abstract. For every disease there is the concrete set of genes the mutations of which multiply the risk of development of illness. Determination of DNA of sick and healthy people resulted in determination of the genes, related to the concrete diseases. The effective procedures are described to determine the point mutations in sequences of the genes. On the basis of Bayesian procedure of recognition it is possible effectively to determine the groups of risks of diseases which COVID-19 accompanies.

Keywords: determination of DNA, the points mutations, Bayesian procedure of recognition.

Вагис Александра Анатольевна,
доктор физ.-мат. наук, ведущий научный сотрудник Института кибернетики им. В.М. Глушкова НАН Украины, Киев, e-mail: valexdep@gmail.com.

Гупал Анатолий Михайлович,
чл.-кор. НАН Украины, доктор физ.-мат. наук, заведующий отделом Института кибернетики им. В.М. Глушкова НАН Украины, Киев, e-mail: gupalanatol@gmail.com.

Сергиенко Иван Васильевич,
академик НАН Украины, доктор физ.-мат. наук, директор Института кибернетики им. В.М. Глушкова НАН Украины, Киев: e-mail: incyb@incyb.kiev.ua.