Кібернетика та комп'ютерні технології

УДК 519.272.2 DOI:10.34229/2707-451X.20.3.3

А.А. ВАГИС, А.М. ГУПАЛ, Н.А. ГУПАЛ

ОПРЕДЕЛЕНИЕ ГРУПП РИСКОВ ПРИ ЗАБОЛЕВАНИЯХ, ВЫЗВАННЫХ COVID-19

Для каждого заболевания существует конкретный набор генов, мутации в которых увеличивают риск развития болезни. Показано, что наличие точечных мутаций в нескольких генах ДНК человека приводит к определенному заболеванию. На основе байесовской процедуры распознавания можно эффективно определять группы рисков заболеваний, сопутствующих COVID-19.

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

Введение. Возможность быстро расшифровывать индивидуальные геномы людей позволила накапливать большие массивы данных относительно заболеваний и связанные с ними мутации в генах ДНК человека. Известно, что мутации в ДНК вызывают тысячи генетических заболеваний и также влияют на работу иммунной системы человека. Коронавирусы – это покрытые оболочкой РНК-вирусы, вызывающие респираторные заболевания различной степени тяжести от обычной простуды до смертельной пневмонии. Множество коронавирусов, впервые обнаруженных у домашних птиц в 1930-х годах, вызывают у животных респираторные, желудочно-кишечные, печеночные и неврологические заболевания. Только семь коронавирусов вызывают заболевание у человека. Три из семи коронавирусов вызывают гораздо более тяжелые, чем другие коронавирусы, а иногда и летальные респираторные инфекции у людей, они послужили причиной крупных вспышек смертельной пневмонии в 21-м веке.

COVID-19 впервые зарегистрирован в конце 2019 года в городе Ухань (Китай), и с тех пор активно распространялся по всему миру. Ранние случаи COVID-19 связывали с птичьим рынком в городе Ухань, предполагая, что первоначальное заражение людей вирусом произошло от животных. Передача от человека к человеку происходит при контакте с инфицированным секретом, выделяемым из дыхательных путей, главным образом, крупными каплями, однако возможно также заражение и при контакте с загрязнённой респираторными выделениями поверхностью. Исследователи всё ещё изучают, насколько легко этот вирус передается от человека к человеку или насколько устойчивой будет его циркуляция в популяции; впрочем, возможно, что COVID-19 более заразен, чем SARS и, пожалуй, его распространение более похоже на распространение гриппа.

У людей с COVID-19 симптомы могут быть незначительными или даже полностью отсутствовать, хотя некоторые из них тяжело заболевают и умирают. Симптомы могут включать 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 не известен; предположительно он варьируется от 1 до 14 дней. С возрастом увеличивается риск заболеть тяжёлой формой заболевания и смерти от COVID-19. Диагностика проводится с помощью ПЦР выделений из верхних и нижних дыхательных путей и сыворотки крови. Наряду с лабораториями системы здравоохранения, диагностическое тестирование на COVID-19 становится все более доступным в коммерческих и больничных лабораториях. ПЦР-анализ у постели больного также коммерчески доступный.

В группе риска у людей с COVID-19 находятся лица с такими хроническими заболеваниями: сердечно-сосудистая система; дыхательная система; эндокринная система; онкологические заболевания; иммунодефицитные состояния; больные с почечной недостаточностью.

Мутации в генах вызывают различные заболевания. В настоящее время в ведущих станах мира проводится расшифровка (секвенирование) геномов большого количества людей. Полученная информация будет использоваться для ранней диагностики различных заболеваний, в первую очередь онкологических. Основной задачей в этой области является определение генетических (или врожденных) предрасположенностей к сложным системным заболеваниям, таким как болезни сердечно-сосудистой системы, рак, диабет, шизофрения. Для каждого заболевания существует свой 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 заболеваниями, которые возникают при COVID-19.

Возможность быстро расшифровывать индивидуальные геномы людей позволила накопить большие массивы данных о заболеваниях и связанных с ними мутациях в ДНК. Наиболее распространенным типом мутаций, которые приводят к заболеваниям, являются точечные мутации, в результате которых единичный нуклеотид гена меняется на другой нуклеотид. Были исследованы мутации, обусловленные такими заболеваниями: аутоиммунными, онкологическими, сердечно-сосудистыми, генетическими, нейродегенеративными, психологическими расстройствами, пагубными привычками. 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.

С помощью генетических алгоритмов получены оптимальные генетические коды, помехоустойчивость которых на 8,5 % выше, чем у стандартного кода. На основе баз данных генетических заболеваний стандартным кодом было проверено примерно четыреста мутаций для различных типов заболеваний. Примерно половина из них привела к нарушению полярности или к мутациям третьего нуклеотида (аминокислота при этом не меняется, однако прерывается процесс вырезания или сплайсинга интронов). Оптимальные коды исправляют нарушение полярности при мутациях первого и второго нуклеотидов в кодоне, однако избавиться от мутаций в третьем нуклеотиде нельзя. В таблице приведены оценки мутаций для сердечно-сосудистых заболеваний, выполненные с помощью стандартного генетического кода. Аналогичные таблицы можно представить для вышеперечисленных заболеваний.

Байесовские процедуры распознавания. В работах [3, 4] было 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. Рассмотрим следующую задачу с булевскими переменными.

ТАБЛИЦА. Сердечно-сосудистые заболевания

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	+

Примечания. $+^*$ – сохранение полярности, $-^*$ – нарушение полярности, c^* – сохранение аминокислоты при мутации третьего нуклеотида.

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

Считаем, 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: первая часть V_0 – булева матрица размерности $m_0 \times n$, где m_0 – число строк. Каждая строка представляет собой вектор $x = (x_1, x_2, ..., x_n, f)$, который выбран в соответствии с распределением P при условии $f = 0$. Вторая часть V_1 – булева матрица размерности $m_1 \times n$, где m_1 – число строк. Каждая строка матрицы – вектор x, который выбран на основе распределения P при условии $f = 1$. Последняя часть V_2 – булев вектор размерности m_2. Каждая компонента этого вектора – наблюдаемое значение состояния f, которое выбирается 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 $x_1, x_2, ..., x_n$ любого следующего объекта и обучающей выборке $V = (V_0, V_1, V_2)$ определяет состояние f объекта.

Пусть $d = (d_1, d_2, ..., d_n)$ – булев вектор. Считаем, что распределения P при каждом d удовлетворяют условию

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

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

$$\xi(d,i) = \left(\frac{k(i)}{m_2}\right) \prod\nolimits_{j=1}^{n} \left(\frac{k(d_j, i)}{m_i}\right); \quad i = 0,1; \tag{1}$$

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

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

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

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

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

Определение групп рисков при заболеваниях COVID-19. Обращаясь к таблице можно сделать вывод о том, у лиц, переболевших COVID-19 с диагнозом сердечно-сосудистое заболевание, с высокой долей вероятности имели место точечные мутации в определенных генах. Этих людей можно условно внести в обучающую выборку V_1 «больные», причем ее можно разбить на возрастные группы. В класс V_0 «здоровые» вносятся лица с отрицательным результатом ПЦР, тоже с учетом их возраста.

Полагаем, что гены в левом столбце таблицы – это признаки для байесовской процедуры. Для исключения тривиальных случаев считаем, что в выборке V_0 для каждого гена в таблице V_0 есть представители с мутациями в этом гене. Наоборот, в выборке V_1 присутствуют лица без мутаций в этом гене.

Выберем первый ген в таблице и рассмотрим выборку V_0. При сравнении последовательностей гена 1 отдельных представителей выборки V_0 с последовательностью гена 1 исследуемого лица можно получить следующие результаты: отсутствие изменений или мутаций (обозначаем эту ситуацию 0); наличие мутаций, их может быть одна (обозначаем 1) или две (обозначаем 2). Поскольку мутации появляются случайным образом в последовательности гена, вероятность появления мутаций в одном и том же месте последовательности гена у двух разных людей имеет ничтожно малую вероятность. Заметим, что длина отдельного гена в ДНК человека может превосходить десятки тысяч нуклеотидов. Наличие числа 2 при сравнении означает, что у исследуемого лица есть мутации в гене 1. Поэтому в формуле (1) $k(d_1,0)$ равно количеству двоек, полученных при сравнении. Аналогично в выборке V_1 при сравнении с исследуемым лицом $k(d_1,1)$ тоже равно количеству двоек.

Если же в выборке V_0 при сравнении с исследуемым лицом двойки не появляются, то это означает отсутствие мутации у исследуемого лица, и $k(d_1,0)$ равно количеству нулей в таблице V_0. Тогда для выборки V_1 при сравнении с исследуемым лицом двойки тоже не появляются и $k(d_1,1)$ равно количеству нулей в таблице V_1.

Описанную схему вычислений проводим для всех генов в вышепредставленной таблице и определяем значения $\xi(d,i)$ для выборок V_0 и V_1. Результат байесовской процедуры для исследуемого лица выполняется по формуле (2).

Выводы. 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 конкретними заболеваниями, в том числе и с заболеваниями, которые возникают при COVID-19. Показано, что наличие точечных мутаций в нескольких генах ДНК человека приводит к определенному заболеванию. На основе байесовской процедуры распознавания можно эффективно определять группы рисков заболеваний, которые сопутствуют 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. 5 . С. 17–24.
2. Сергиенко И.В., Белецкий Б. А., Гупал А.М., Гупал Н.А. Оптимальные помехоустойчивые коды. *Кибернетика и системный анализ.* 2019. 1 . С. 44–50.
3. Гупал А.М., Пашко С.В., Сергиенко И.В. Эффективность байесовской процедуры распознавания. *Кибернетика и системный анализ.* 1995. 4 . С. 76–89.
4. Сергиенко И.В., Гупал А.М., Пашко С.В. О сложности задач распознавания образов. *Кибернетика и системный анализ.* 1996. 4 . С. 70–88.

Получено 14.10.2020

Вагис Александра Анатольевна,
доктор физико-математических наук,
ведущий научный сотрудник Института кибернетики имени В.М. Глушкова НАН Украины,

Гупал Анатолий Михайлович,
доктор физико-математических наук, член-корреспондент НАН Украины,
заведующий отделом Института кибернетики имени В.М. Глушкова НАН Украины,

Гупал Никита Анатольевич,
кандидат физико-математических наук,
научный сотрудник Института кибернетики имени В.М. Глушкова НАН Украины.

УДК 519.272.2

О.А. Вагіс, А.М. Гупал*, М.А. Гупал

Визначення груп ризиків при захворюваннях, що викликані COVID-19

Інститут кібернетики імені В.М. Глушкова НАН України, Київ
* *Листування: gupalanatol@gmail.com*

Вступ. У групі ризику у людей з COVID-19 знаходяться обличчя з такими хронічними захворюваннями: серцево-судинна система; дихальна система; ендокринна система; онкологічні захворювання; імунодефіцитні стани; хворі з нирковою недостатністю. Для кожного захворювання існує свій конкретний набір генів, мутації в яких збільшують ризик розвитку хвороби. Масове секвенування ДНК хворих й здорових людей привело до визначення генів, пов'язаних з конкретними захворюваннями, у тому числі й із захворюваннями, які виникають при COVID-19. В осіб, що перехворіли COVID-19 з певним захворюванням, з високою часткою ймовірності мали місце точкові мутації в певних генах. Цих людей можна умовно внести в навчаючу вибірку «хворі», у клас «здорові» вносяться персони з негативним результатом ПЦР.

Мета роботи. На основі навчаючих вибірок розробити ефективні методи визначення груп ризиків захворювань, які супроводять COVID-19.

Результати. Вважаємо, що гени в лівому стовпці таблиці є ознаками для байєсівської процедури. Робота процедури виконується на основі підрахунку кількості мутацій або їх відсутності в навчаючих вибірках класів «хворі» і «здорові». Досліджувану особу співвідносимо в той клас, для яких результат процедури вище.

Висновки. Масове секвенування ДНК хворих і здорових людей привело до визначення генів, пов'язаних з конкретними захворюваннями, у тому числі й із захворюваннями, які виникають при COVID-19. Показано, що наявність точкових мутацій у генах ДНК людини приводить до певного захворювання. На основі байєсівської процедури розпізнавання можна ефективно визначати групи ризиків захворювань, які супроводять COVID-19.

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

UDC 519.272.2

A.A. Vagis, A.M. Gupal*, N.A. Gupal

Determination of Groups of Risks at the Diseases COVID-19

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv
* *Correspondence: gupalanatol@gmail.com*

Introduction. In the group of risk at people with COVID-19 there are persons with the such chronic diseases: heart-vessel system; respiratory system; endocrine system; oncologic diseases; immune-deficit states; patients with kidney insufficiency.

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 diseases which arise up at COVID-19. At persons having by had COVID-19 with the certain disease, with the high stake of probability took place points mutations in certain genes. These people can be brought in a teaching sampling «sick», in a class «healthy» persons are brought in with the negative result of PCR.

Purpose of the article. On the basis of teaching selections to develop the effective methods of determination of groups of risks of diseases which COVID-19 accompanies.

Results. We consider that genes in a left table column are signs for Bayesian procedure. Work of procedure is executed on the basis of count of amount of mutations or their absence in the teaching selections of classes «sick» and «healthy». We correlate the explored person in that class «sick» and «healthy», for which result of procedure higher.

Conclusions. Determination of DNA of sick and healthy people resulted in determination of the genes related to the concrete diseases, including with the diseases which arise up at COVID-19. It is shown that the presence of points mutations in the genes of DNA of man results in the certain disease. 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.