Кібернетика та комп'ютерні технології

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

О.А. ВАГІС, А.М. ГУПАЛ, М.А. ГУПАЛ

ВИЗНАЧЕННЯ ГРУП РИЗИКІВ ПРИ ЗАХВОРЮВАННЯХ, ЩО ВИКЛИКАНІ COVID-19

Для кожного захворювання існує конкретний набір генів, мутації в яких збільшують ризик розвитку хвороби. Показано, що наявність точкових мутацій у декількох генах ДНК людини спричиняє певне захворювання. На основі байєсівської процедури розпізнавання можна ефективно визначати групи ризиків захворювань, які викликані COVID-19.

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

Вступ. Можливість швидко розшифровувати індивідуальні геноми людей дозволила накопичувати великі масиви даних щодо захворювань і пов'язані з ними мутації у генах ДНК людини. Відомо, що мутації у ДНК викликають тисячі генетичних захворювань і також впливають на роботу імунної системи людини. Коронавіруси – це покриті оболонкою РНК-віруси, які викликають респіраторні захворювання різного ступеня тяжкості від звичайної простуди до смертельної пневмонії. Численні коронавіруси, що вперше були виявлені у домашніх птахів у 1930-х роках, викликають у тварин респіраторні, кишкові, печінкові та неврологічні захворювання. Тільки сім коронавірусів викликають захворювання у людини. Три з семи коронавірусів, що викликають значно більш тяжкі, ніж інші коронавіруси, а іноді й летальні респіраторні інфекції у людей, стали причиною значних спалахів смертельної пневмонії у 21-у столітті.

COVID-19 уперше зареєстрований у кінці 2019 року в місті Ухань (Китай), і з того часу активно розповсюджується по всьому світу. Ранні випадки COVID-19 пов’язували з пташиним ринком у місті Ухань, припускаючи, що первинне зараження людей вірусом відбулося від тварин. Передача вірусу від людини до людини відбувається при контакті з інфікованим секретом, що виділяється з дихальних шляхів великими краплями, проте можливо також зараження і при контакті з забрудненою респіраторними виділеннями поверхнею. Дослідники вивчають, наскільки легко цей вірус передається від людини до людини або наскільки стійкою буде його циркуляція в популяції; втім, можливо, що COVID-19 більш заразливий, чим SARS і, мабуть, його розповсюдження більш схоже на розповсюдження грипу.

У людей з COVID-19 симптоми можуть бути незначними або навіть повністю відсутніми, хоча деякі з них важко захворюють і вмирають. Симптоми можуть включати лихоманку, кашель і задишку. У пацієнтів з важчою формою захворювання можуть реєструватися лімфопенія і характерні для пневмонії зміни при діагностичних візуалізуючих дослідженнях. Точний інкубаційний період не відомий; імовірно він варіюється від 1 до 14 днів. З віком збільшується ризик захворіти важкою формою захворювання і смерті від COVID-19. Діагностика проводиться за допомогою ПЦР виділень з верхніх і нижніх дихальних шляхів і сироватки крові. Разом з лабораторіями системи охорони здоров'я, діагностичне тестування на COVID-19 стає все більш доступним у комерційних і лікарняних лабораторіях. ПЦР-аналіз у ліжку хворого також доступний.

У групі ризику з COVID-19 знаходяться особи з такими хронічними захворюваннями: серцево-судинна система; дихальна система; ендокринна система; онкологічні захворювання; імунодефіцитні стани; хворі з нирковою недостатністю.

Мутації у генах викликають різні захворювання. В даний час у провідних країнах світу проводиться розшифрування (секвенування) геномів великої кількості людей. Отримана інформація використовуватиметься для ранньої діагностики різних захворювань, у першу чергу онкологічних. Основним завданням у цій області є визначення генетичних (або природжених) схильностей до складних системних захворювань, таких як хвороби серцево-судинної системи, рак, діабет, шизофренія. Для кожного захворювання існує свій конкретний набір генів, мутації в яких збільшують ризик розвитку хвороби. Масове секвенування ДНК хворих і здорових людей привело до визначення генів, які пов’язані з конкретними захворюваннями у тому числі й із захворюваннями, що виникають при COVID-19.

Можливість швидко розшифровувати індивідуальні геноми людей дозволила накопичити великі масиви даних про захворювання і пов'язані з ними мутації в ДНК. Найбільш поширеним типом мутацій, які приводять до захворювань, є точкові мутації, у результаті яких одиничний нуклеотид гена міняється на інший нуклеотид. Були досліджені мутації, обумовлені такими захворюваннями: аутоімунними, онкологічними, серцево-судинними, генетичними, нейродегенеративними, психологічними розладами, згубними звичками. У роботах [1, 2] використовувалися дані інтернет-ресурсів, де захворюванням ставилися у відповідність пов’язані з ними мутації в ДНК, тобто були отримані пари початкових і мутованих триплетів нуклеотидів, і відповідно кодованих ними амінокислот.

За допомогою генетичних алгоритмів отримано оптимальні генетичні коди, завадостійкість яких на 8,5 % вище, ніж у стандартного коду. На основі баз даних генетичних захворювань стандартним кодом перевірено приблизно чотириста мутацій для різних типів захворювань. Приблизно половина з них привела до порушення полярності або до мутацій третього нуклеотида (амінокислота при цьому не міняється, проте уривається процес вирізування або сплайсинга інтронів). Оптимальні коди виправляють порушення полярності при мутаціях першого і другого нуклеотидів у кодоні, проте позбавитися мутацій у третьому нуклеотиді не можна. У таблиці приведені оцінки мутацій для серцево-судинних захворювань, виконані за допомогою стандартного генетичного коду. Аналогічні таблиці можна представити для вищеперерахованих захворювань.

ТАБЛИЦЯ. Серцево-судинні захворювання

Ген	Ідентифікатор мутації	Кодон	Мутація кодону	Стандартний 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	+

Примітка. $+^*$ – збереження полярності, $-^*$ – порушення полярності, c^* – збереження амінокислоти при мутації третього нуклеотида.

Байєсівські процедури розпізнавання. У роботах [3, 4] показано, що байєсівська процедура розпізнавання – оптимальна. Для обґрунтування цього результату необхідно було вивести верхню оцінку похибки байєсівської процедури розпізнавання і отримати нижню оцінку складності класу задач. Розглянемо наступну задачу з булевими змінними.

Нехай є скінчена множина X об'єктів b. Кожен об'єкт $x \in X$ ототожнюється з булевим вектором $(x_1, x_2, ..., x_n, f)$, де n – натуральне число. Припустимо, на множині X задано розподіл імовірностей P, який нам невідомий. З множини X отримано навчаючу вибірку V. Нехай деякий об'єкт отриманий з множини X незалежно від вибірки V відповідно до розподілу P, причому відомі значення тільки ознак $x_1, x_2, ..., x_n$. Потрібно за цими значеннями та на основі навчаючої вибірки V визначити значення цільової ознаки f (стан об'єкта x).

Вважаємо, що процес розпізнавання цільової ознаки f об'єкта за відомими ознаками x здійснюється за допомогою функції $A(x)$ по формулі $f = A(x)$. Навчаюча вибірка $V = (V_0, V_1, V_2)$ має наступний вигляд: перша частина V_0 – булева матриця розмірністю $m_0 \times n$, де m_0 – число

рядків. Кожен рядок – вектор $x=(x_1,x_2,...,x_n,f),$ який вибраний відповідно до розподілу P за умови $f=0.$ Друга частина V_1 – булева матриця розмірністю $m_1\times n,$ де m_1 – число рядків. Кожен рядок матриці V_1 – вектор $x,$ який вибраний на основі розподілу P за умови $f=1.$ Остання частина V_2 – булевий вектор розмірності $m_2.$ Кожна компонента цього вектора – спостережуване значення стану $f,$ який вибирається відповідно до розподілу $P.$ Можна вважати, що $m_2 = m_0+m_1.$

Індуктивний крок. Потрібно побудувати таку процедуру індуктивного висновку, яка по вимірюваннях $x_1,x_2,...,x_n$ будь-якого наступного об'єкта і навчаючій вибірці $V=(V_0,V_1,V_2)$ визначає стан f об'єкта.

Нехай $d=(d_1,d_2,...,d_n)$ – булевий вектор. Вважаємо, що розподіли P при кожному d задовольняють умові

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

що означає незалежність ознак x_j для кожного класу об'єктів; тут $P(x=d\,|\,f=i)$ позначає умовну імовірність. Розглянемо випадкові величини $\xi(d,i),$ які залежать від d і i як від параметрів:

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

тут $k(d_j,i)$ – кількість значень, рівних $d_j,$ j-ої ознаки в j-му стовпці матриці j; $k(i)$ – кількість значень цільової ознаки, рівних $i,$ у векторі V_2. Тоді функція розпізнавання визначається формулою

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

Процедуру навчання, що визначається співвідношеннями (1), (2), позначимо Q_B. Зазначимо, що величини $\xi(d,i)/(\xi(d,0)+\xi(d,1))$ є наближеними значеннями імовірностей $P(f=i|x_1=d_1,x_2=d_2,...,x_n=d_n),$ обчислених за формулою Байєса, тому процедуру розпізнавання Q_B називаємо байєсівською. У роботі [3] показано, що для верхньої оцінки похибки Q_B виконується нерівність

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

де a – абсолютна константа. Нижня оцінка складності класу задач відрізняється від (3) на абсолютну константу, тому в цьому сенсі байєсівська процедура Q_B – оптимальна.

Визначення груп ризиків при захворюваннях COVID-19. Звертаючись до таблиці можна зробити висновок про те, що у осіб, що перехворіли COVID-19 з діагнозом серцево-судинне захворювання, з високою часткою імовірності мали місце точкові мутації у певних генах.

Цих осіб можна умовно внести в навчаючу вибірку V_1 «хворі», причому її можна розбити на вікові групи. У клас V_0 «здорові» вносяться особи з негативним результатом ПЦР, теж з урахуванням їх віку.

Вважаємо, що гени в лівому стовпчику таблиці – ознаки для байєсівської процедури. Для виключення тривіальних випадків вважаємо, що у вибірці V_0 для кожного гена у таблиці V_0 є представники з мутаціями в цьому гені. Аналогічно вважаємо, що у вибірці V_1 присутні особи без мутацій у цьому гені.

Виберемо перший ген у таблиці і розглянемо вибірку V_0. При порівнянні послідовностей гена 1 окремих представників вибірки V_0 з послідовністю гена 1 досліджуваної особи можна отримати наступні результати: відсутність змін або мутацій (позначаємо цю ситуацію 0); наявність мутацій, їх може бути одна (позначаємо 1) або дві (позначаємо 2). Оскільки мутації з'являються випадковим чином у послідовності гена, імовірність появи мутацій в одному й тому ж місці послідовності гена у двох різних людей має малу імовірність. Зазначимо, що довжина окремого гена в ДНК людини може перевершувати десятки тисяч нуклеотидів. Наявність числа 2 при порівнянні означає, що у пацієнта є мутації в гені 1. Тому у формулі (1) $k(d_1,0)$ дорівнює кількості двійок, отриманих при порівнянні. Аналогічно у вибірці V_1 при порівнянні з пацієнтом $k(d_1,1)$ теж дорівнює кількості двійок.

Якщо ж у вибірці V_0 при порівнянні з пацієнтом двійки не з'являються, то це означає відсутність мутації у пацієнта, і $k(d_1,0)$ дорівнює кількості нулів у таблиці V_0. Тоді для вибірки V_1 при порівнянні з пацієнтом двійки теж не з'являються і $k(d_1,1)$ дорівнює кількості нулів у таблиці V_1.

Описану схему обчислень проводимо для всіх генів у вищепредставленій таблиці й визначаємо значення $\xi(d,i)$ для вибірок V_0 і V_1. Результат байєсівської процедури для пацієнта отримуємо за формулою (2).

Висновки. Для кожного захворювання існує конкретний набір генів, мутації в яких збільшують ризик розвитку хвороби. 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 декількох генах ДНК людини спричиняє певне захворювання. На основі байєсівської процедури розпізнавання можна ефективно визначати групи ризиків захворювань, що викликані COVID-19.

Список літератури

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.

Одержано 14.10.2020

Вагіс Олександра Анатоліївна,
доктор фізико-математичних наук,
провідний науковий співробітник
Інституту кібернетики
імені В.М. Глушкова НАН України, Київ,

Гупал Анатолій Михайлович,
доктор фізико-математичних наук, завідувач відділу,
член-кореспондент НАН України, Інституту кібернетики
імені В.М. Глушкова НАН України, Київ,
gupalanatol@gmail.com

Гупал Микита Анатолієвич,
кандидат фізико-математичних наук,
науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ.

УДК 519.272.2

А.А. Вагис, А.М. Гупал, Н.А Гупал

Определение групп рисков при заболеваниях, вызванных COVID-19

Институт кибернетики имени В.М. Глушкова НАН Украины, Киев
Переписка: gupalanatol@gmail.com

Введение. 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.

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. У людей, переболевших COVID-19 с определенным заболеванием, с высокой долей вероятности имели место точечные мутации в определенных генах. Этих людей можно условно внести в обучающую выборку «больные», в класс «здоровые» вносятся люди с отрицательным результатом ПЦР.

Цель работы. На основе обучающих выборок разработать эффективные методы определения групп рисков заболеваний, которые сопутствуют COVID-19.

Результаты. Считаем, что гены в левом столбце таблицы являются признаками для байесовской процедуры. Работа процедуры выполняется на основе подсчета количества мутаций или их отсутствия в обучающих выборках классов «больные» и «здоровые». Исследуемое лицо соотносим в тот класс «больные» и «здоровые», для которых результат процедуры выше.

Выводы. 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. Показано, что наличие точечных мутаций в генах ДНК человека приводит к определенному заболеванию. На основе байесовской процедуры распознавания можно эффективно определять группы рисков заболеваний, которые сопутствуют COVID-19.

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

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.