DOI 10.1007/s10559-021-00347-9
Cybernetics and Systems Analysis, Vol. 57, No. 2, March, 2021

DETERMINATION OF RISK GROUPS FOR THE COVID-19 UNDERLYING DESEASES

A. A. Vagis,[1†] **A. M. Gupal,**[1‡] **and I. V. Sergienko**[1††] UDC 519.217.2

Abstract. *For every disease, there is a certain set of genes whose mutations increase the risk of illness development. DNA sequencing of sick and healthy individuals results in the determination of genes related to certain diseases. Efficient procedures are described in order to determine point mutations in gene sequences of the examined patients. The optimal Bayesian procedure is used to determine risk groups for certain diseases, including the ones that underlie COVID-19.*

Keywords: *DNA sequencing, point mutations, Bayesian recognition procedure.*

INTRODUCTION

The possibility of quick decoding of individual human genome has allowed us to amass vast data arrays of diseases, as well as associated human DNA mutations. It is well-known that DNA mutations cause thousands of genetic diseases and influence the human immune system. Coronaviruses are enveloped RNA viruses that cause respiratory illnesses of different severity levels, from a common cold to a pneumonia with lethal outcome. The COVID-19 virus that has been recorded at the end of 2019 in Wuhan (China) for the first time is aggressively spreading all over the world. Researchers are still studying how easily this virus can be transmitted from one person to another or how steady its circulation will become in a population.

The symptoms of a person who has contracted COVID-19 can be mild or nonexistent at all. However, in the case of some patients, a severe course of the disease with an unfavorable prognosis is observed. The symptoms of COVID-19 include fever, cough, and labored breathing. Patients suffering from a severe form of this disease can exhibit lymphocytopenia and changes characteristic for pneumonia during diagnostic visual testing. The exact COVID-19 latent period is unknown, it is thought to fluctuate between 1 and 14 days. Patients of older age groups have a higher chance to develop a severe disease form. The diagnosis is performed by the means of PCR tests of secretions from upper and lower respiratory tracts, as well as blood serum.

The risk group for patients with COVID-19 include individuals with chronic cardiovascular, respiratory, and endocrine diseases, as well as oncologic pathologies, immunodeficiency, and other types of deficiency.

DESEASES CAUSED BY GENE MUTATIONS

In the present day, decoding (sequencing) of genome of a large amount of people is performed in the developed countries. The obtained information is used for the early diagnosis of different diseases, oncological ones in the first place. The main task of this area is to determine genetic (or innate) predisposition to complex diseases of systems such as cardiovascular diseases, cancer, diabetes, and schizophrenia. For each disease, there exists a certain gene set, mutations in which raise the risk of disease development. Mass DNA sequencing of ill and healthy individuals has led to the determination of genes associated with certain diseases, including the ones that appear with COVID-19.

[1]V. M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, [†]*alexdep@gmail.com*; [‡]*gupalanatol@gmail.com*; [††]*incyb@incyb.kiev.ua*. Translated from Kibernetyka ta Systemnyi Analiz, No. 2, March–April, 2021, pp. 62–68. Original article submitted November 9, 2020.

TABLE 1. Mutation Estimation for Cardiovascular Diseases

Gene	Mutation Indicator	Codon	Codon Mutation	Standard Code
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	+

Annotation: $+^*$ — retained polarity, $-^*$ — violated polarity, and c^* — retained aminoacid with a mutation in the third nucleotide.

The most widespread mutation type that leads to diseases is point mutations, as a result of which a single gene nucleotide is replaced by another nucleotide. Point mutations can arise as a result of spontaneous mutations taking place during DNA replication, as well as the result of mutagen influence, as, for example, the impact of ultraviolet light or X-ray radiation, of high temperatures or chemical substances.

Internet resource data, where diseases were associated with DNA mutations related to them, was used in [1, 2], i.e., pairs of initial and mutated nucleotide triplets and the aminoacids encoded by them, respectively, have been obtained. Mutations induced by autoimmune, oncological, cardiovascular, genetic, and neurodegenerative diseases, as well as psychical disorders and addictions have been studied.

Applying genetic algorithms, optimal genetic codes have been obtained, whose noise immunity is 8.5% higher than that of the standard code. Using genetic disease databases, approximately 400 mutations associated with different disease types have been checked by the standard code, and almost half of them has led to polarity violation or to mutations of the third nucleotide (in this case, the aminoacid does not change; however, the process of intron cutting or splicing stops) [3]. Optimal codes correct polarity violations caused by mutations of the first and the second nucleotides in the codon; however, it is impossible to get rid of mutations in the third nucleotide. Table 1 presents mutation estimates for cardiovascular diseases, which have been obtained by using the standard genetic code (similar tables can be presented for the above-mentioned diseases).

BAYESIAN RECOGNITION PROCEDURE

As it is shown in [4, 5], the Bayesian recognition procedure is optimal. To justify this result, the upper-end error estimates of the Bayesian recognition procedure had to be found and the lower-end problem class complexity had to be obtained. For the sake of simplicity, let us consider the following problem with Boolean variables.

Let there be a finite set X of objects b. Every object $x \in X$ is associated with a Boolean vector $(x_1, x_2, \ldots, x_n, f)$, where n is a natural number. Let us assume that a probability distribution P is determined over the set X and that it is unknown. A training sample V is formed from the set X. Let a certain object be obtained from the set X irrespective of the sample V according to the distribution P, where only the values of indicators $x_1, x_2, \ldots, x_n$ are known. We have to determine the value of an objective indicator f (the state of an object x) in accordance with these values and the training sample V.

We will assume that the recognition of the objective indicator f of an object in accordance with the known indicators x is performed by using a function $A(x)$ by the formula $f = A(x)$. The training sample $V = (V_0, V_1, V_2)$ has the following form:

- V_0 is an $m_0 \times n$ Boolean matrix, where m_0 is the number of rows with each of them being the vector $x = (x_1, x_2, \dots, x_n, f)$ chosen in accordance with the distribution P under the condition $f = 0$;
- V_1 is an $m_1 \times n$ Boolean matrix, where m_1 is the number of rows with each of them being the vector x chosen in accordance with the distribution P under the condition $f = 1$;
- V_2 is a Boolean vector of dimension m_2, whose each component is an observable state of f chosen in accordance with the distribution P. We can assume that $m_2 = m_0 + m_1$.

Inductive Step. Such an inductive proof procedure has to be constructed that it will determine the state f of an object based on measures $x_1, x_2, \dots, x_n$ of any following object and a random sample $V = (V_0, V_1, V_2)$.

Let $d = (d_1, d_2, \dots, d_n)$ be a Boolean vector. We will consider that the distributions P in the case of each d satisfy the following condition:

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

which proves the independence of the indicators x_j for each object class; here, $P(x = d \mid f = i)$ are probability conditions. Let us consider the following random variables $\xi(d, i)$ that depend on parameters d and i:

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

here $k(d_j, i)$ is a number of values equal to d_j and the jth indicator in the jth column of a matrix V_i; $k(i)$ is the number of values of the objective indicator, which are equal to i in the vector V_2. Then, the recognition function is determined by the formula

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

Let us denote the training procedure determined by (1) and (2) by Q_B. Note that $\xi(d, i) / (\xi(d,0) + \xi(d,1))$ are approximate values of the probabilities $P(f = i \mid x_1 = d_1, x_2 = d_2, \dots, x_n = d_n)$ calculated by the Bayes formula; therefore, the recognition procedures Q_B are called Bayesian procedures. As it is shown in [3], for the upper-end estimate of the error Q_B the inequality

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

is fulfilled, where a is an absolute constant. The lower-end problem class complexity differs from (3) by the absolute constant, therefore, in this context, the Bayesian procedure Q_B is optimal.

RISK GROUP DETERMINATION IN THE CASE OF COVID-19 INFECTION

Simplified Variant without Introns. Having analyzed Table 1, we can conclude that patients suffering from a cardiovascular disease and having been infected with COVID-19 have a high probability of point mutation occurrence in certain genes. The data on these patients can be introduced into the "ill" training sample V_1 that is to be divided into age groups, and the data on the patients with negative PCR test results, where their age is also accounted for, can be introduced into the "healthy" testing sample V_0.

We assume that the genes in the first column of Table 1 are indicators for the Bayesian procedure. In order to exclude trivial cases, we assume that for each gene in Table 1 in the sample V_0 there exist representatives with mutations in this gene. Similarly, we assume that there exists data on patients with no mutations in this gene in the sample V_1.

Let us choose the first gene in Table 1 and consider the sample V_0. When comparing sequences of the first gene for certain representatives of the sample V_0 to its sequence for the patient under study, we can obtain the following results:

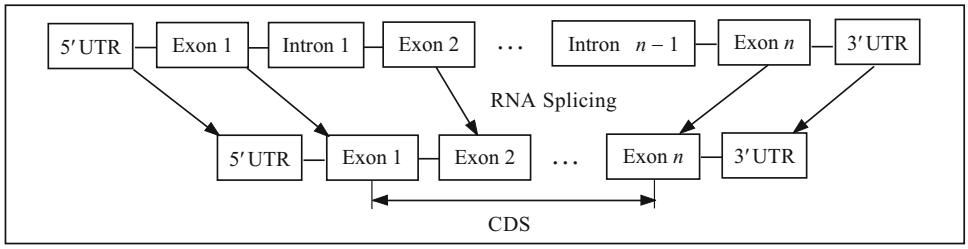

Fig. 1. Gene structure in DNA and matrix RNA.

- 0 — there exist no changes or mutations;
- 1 — there exists a single mutation;
- 2 — there exist two mutations.

Since mutations arise in an arbitrary fashion in the gene sequence, there exists a low probability of mutations appearing in one and the same gene sequence region of two different individuals. (Note that the length of one single human DNA gene can exceed tens of thousands of nucleotides.) The appearance of number 2 during comparison means that there exist mutations in the first gene of the patient. Therefore, in (1), $k(d_1, 0)$ is equal to the number of 2s obtained during comparison. Similarly, in the sample V_1, during the comparison with the patient, $k(d_1, 1)$ is also equal to the number of 2s.

If no number 2 appears in the sample V_0 during the comparison with the examined patient, it signifies the absence of mutations and that $k(d_1, 0)$ is equal to the number of zeroes during calculation in this sample. In that case, number 2 will also not appear in the sample V_1 during the comparison with the patient and that $k(d_1, 1)$ is equal to the number of zeroes during calculation in this sample.

We will apply the above-described calculation scheme to all the genes present in Table 1 and determine the value $\xi(d, i)$ for the samples V_0 and V_1. We will obtain the Bayesian procedure results for the patient in accordance with (2).

General Case. Genes are DNA regions with the length of up to a couple of tens of thousands of bases. The nucleotide sequence in DNA regions determines the structure of a certain protein. The gene structure has become more and more complicated in the evolution process; therefore, the DNA regions that encode genetic information for eukaryotes (organisms whose cells contain a nucleus, namely, plants and animals), have a complex form (Fig. 1). The following main components of eukaryote genes are distinguished in accordance with the function being fulfilled during the protein synthesis:

- the initial and the finite untranslated regions denoted by 5′UTR and 3′ UTR, respectively, which do not take part in the process of protein encoding but influence it indirectly;
- exons that are DNA components directly encoding aminoacid sequences that build a protein using the standard genetic code;
- introns that are DNA regions situated between exons that do not take part in protein synthesis. (In the present time, their purpose in unknown; it may be that introns are protection mechanisms against mutations.)

One human gene has approximately seven exons. Intron length exceeds exon length by more than 10 times. Cases have been described in [3], where mutations took place in introns or at exon-intron boundaries and have stopped the process of cutting (splicing) of introns, as well as caused various diseases. Introns GU–AG and AU–AC can be found in eukaryote genes that encode protein. In the most RNA introns 5′–GU–3′ are the first two intron sequence nucleotides and 5′–AG–3′ are the last two nucleotides. For that reason, they are denoted by GU–AG introns, and all the members of this class are cut in the same way. This feature has been revealed after the discovery of introns and it has been assumed that they will be important for the splicing process.

For example, the mutation of G or T in a DNA copy in the 5′ site of GU–AG intron cutting or the mutation of A or G in the 3′ site of the cut will stop the splicing process, as the correct exon-intron boundary will not be identified. Methods of identification of gene region fragments based on the Markov models with hidden variables are proposed in [6, 7].

By comparing the gene sequences of two representatives of the sample V_0 (V_1), we will determine the number and region of the detected mutations on a computer, assuming that the coincidence probability of point mutations in one place is too low. By comparing the gene sequence of the third representative with the distinguished sequence, we will determine its number of mutations and their location. Similarly, we will find the number of mutations and their location for all the representatives of the sample V_0 (V_1), as well as for the examined patient.

It is possible to determine the number of mutations for three representatives and then do it for all the other participants. By comparing the data on the first and the second representatives from the sample, we obtain the following equation: $M_1 + M_2 = S_1$, where S_1 is the mutation sum; the equation $M_1 + M_3 = S_2$ is the mutation sum of the first and the third representatives and the equation $M_2 + M_3 = S_3$ is the mutation sum of the second and the third representatives. Solving this equation system comprising three equations, we obtain $M_1 = S_1 + S_2 - S_3$, $M_2 = S_1 + S_3 - S_2$, and $M_3 = S_2 + S_3 - S_1$.

Note that during the calculation process based on the Bayesian procedure, it is necessary to take into consideration mutations for the representatives of the samples V_0 (or V_1) and for the examined patient, which have occurred in exons or at exon-intron boundaries, and to not consider intron mutations that do not influence the appearance of diseases.

Thus, knowing the number of mutations of the examined patient, we will determine the values $\xi(d, i)$ for the first gene based on the information from the samples V_0 and V_1. We will apply the above-described scheme for all the genes presented in Table 1 and determine the values $\xi(d, i)$ for the states $i = 0, 1$. We will obtain the results of the Bayesian procedure for the examined patient by (2).

CONCLUSIONS

There exists a certain gene set for each disease, mutations in which increases the risk of disease development. Mass DNA sequencing of ill and healthy individuals has allowed us to determine genes associated with certain diseases, including the ones underlying COVID-19. The individuals with determined diagnoses and those who have recovered from COVID-19 have a high probability degree of having developed point mutations in certain genes.

The proposed procedures for determining mutations and their location in gene sequences allow us to solve the following important problems: to conduct a detailed statistical analysis (including for age groups of patients) in relation to the number of mutations in encoding gene regions (exons) and in introns, as well as to confirm a hypothesis about protecting mechanisms in introns.

Since the Bayesian procedure is widely applied in medical prognosis and in bioinformatics [8, 9], we propose to use it to determine risk groups for diseases underlying COVID-19. The above-described method can be used to determine patient risk groups for different diseases not related to COVID-19.

REFERENCES

1. I. V. Sergienko, A. M. Gupal, and A. V. Ostrovskii, “Noise immunity of genetic codes to point mutations,” Cybern. Syst. Analysis, Vol. 50, No. 5, 663–669 (2014). https://doi.org/10.1007/s10559-014-9656-y.
2. I. V. Sergienko, B. A. Biletskyy, A. M. Gupal, and M. A. Gupal, “Optimal noise-immune genetic codes,” Cybern. Syst. Analysis, Vol. 55, No. 1, 34–39 (2019). https://doi.org/10.1007/s10559-019-00110-1.
3. T. A. Brown, Genomes 3, Garland Sci. (2006).
4. A. M. Gupal, S. V. Pashko, and I. V. Sergienko, “Efficiency of Bayesian classification procedure,” Cybern. Syst. Analysis, Vol. 31, No. 4, 543–554 (1995). https://doi.org/10.1007/BF02366409.
5. I. V. Sergienko, A. M. Gupal, and S. V. Pashko, “Complexity of classification problems,” Cybern. Syst. Analysis, Vol. 32, No. 4, 519–533 (1996). https://doi.org/10.1007/BF02366774.
6. I. V. Sergienko, A. M. Gupal, and A. V. Ostrovsky, “Recognition of DNA gene fragments using hidden Markov models,” Cybern. Syst. Analysis, Vol. 48, No. 3, 369–377 (2012). https://doi.org/10.1007/s10559-012-9416-9.
7. A. M. Gupal and A. V. Ostrovsky, “Using compositions of Markov models to determine functional gene fragments,” Cybern. Syst. Analysis, Vol. 49, No. 5, 692–698 (2013). https://doi.org/10.1007/s10559-013-9556-6.
8. A. M. Gupal, M. A. Gupal, and A. L. Tarasov, “Bayesian procedures of hematologic disease recognition,” Cybern. Syst. Analysis, Vol. 53, No. 6, 925–930 (2017). https://doi.org/10.1007/s10559-017-9994-7.
9. N. Ya. Gridina, A. M. Gupal, A. L. Tarasov, and Yu. V. Ushenin, “Analysis of neurosurgical pathologies using Bayesian recognition procedures for indicators of surface plasmon resonance in the aggregation of blood cells,” Cybern. Syst. Analysis, Vol. 56, No. 4, 550–558 (2020). https://doi.org/10.1007/s10559-020-00271-4.