AIMS Mathematics, 8(8): 19275–19296.
DOI: 10.3934/math.2023983
Received: 14 March 2023
Revised: 16 May 2023
Accepted: 19 May 2023
Published: 07 June 2023

http://www.aimspress.com/journal/Math

Research article

Mutation of DNA and RNA sequences through the application of topological spaces

A. A. El-Atik[1], Y. Tashkandy[2], S. Jafari[3], A. A. Nasef[4], W. Emam[2] and M. Badr[5,*]

[1] Department of Mathematics, Faculty of Science, Tanta University, Tanta, Egypt

[2] Department of Statistics and Operations Research, Faculty of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia

[3] College of Vestsjaelland South, Herrestraede 11,4200 Slagelse, Denmark

[4] Department of Physics and Engineering Mathematics, Faculty of Engineering, Kafrelsheikh University, Kafrelsheikh 33516, Egypt

[5] Department of Mathematics, Faculty of Science, New Valley University, Egypt

* **Correspondence:** Email: m.shaban@sci.nvu.edu.eg.

Abstract: Topology is branch of modern mathematics that plays an important role in applications of biology. The aim of this paper is to study DNA sequence mutations using multisets, relations, metric functions, topology and association indices. Moreover, we use association indices to study the similarity between DNA sequences. These different ways of identifying a mutation help biologists to make a decision. A decision of mutation that depends on metrics between two sequences of genes and the topological structure produced by their relationship is presented.

Keywords: multiset; topology; mutation; similarity; metric space
Mathematics Subject Classification: 54A05, 54C10, 54D10

1. Introduction

The strings of DNA sequences are shaped from nucleotides which are bonded together. DNA has four nucleotides called guanine G, cytosine C, adenine A and thymine T (or uracil U). G (resp. A) is paired with C (resp. T or U). The interaction between them is folding. The chain of nucleotides may be folding and bonding, but these interactions only occur under specific energy conditions. Here, we utilize nucleotide chains in conformity with the topological model; see [1, 10]. A change or metamorphosis is called a mutation. Chromosome and gene alterations, known as mutations in biology, frequently manifest physically. The consequences of a mutation depend on the region where

the genetic material's sequence has changed. On the other hand, insertion or deletion mutations might result in the production of gene products that are not functional. Large-scale mutations can also occur, resulting in the inversion, insertion, duplication, deletion, transposition, or translocation of lengthy strands of DNA. A mutation's outcome could be negative, positive, neutral, or even barely noticeable. A mutation may result in the removal or addition of a particular function, altered levels of expression, or even mortality in the developing embryo. A lot of scientists have worked hard to find and fix gene mutations. They have employed a few techniques for investigation, including single-stranded DNA oligonucleotide analysis, single-strand conformation polymorphism analysis, two-dimensional gene scanning, protein truncation testing, and denaturing high performance liquid chromatography [5, 35]. The use of computer technology for the management of biological data is known as bioinformatics. To collect, store, analyze, and combine biological and family data for use in the discovery and development of gene-based drugs, information processing systems are utilized. The increase of publicly accessible genetic data as a result of the Human Genome Project has sparked the need for bioinformatics capabilities. A virus may cause mutations or the host may edit them, and sequencing mistakes can further complicate matters.

Multiset theory was introduced by Gostelow [23]. The concept of a multiset (or bag) is the generalization of a set. A member of a multiset has more than one membership (see [7, 8, 29, 36, 37]); the use of multisets in mathematics predates the name multiset by nearly 90 years.

Topology is a branch of geometry with the name of rubber sheet geometry. It has many real-life applications and solves some problems that are directly or indirectly related to continuity. Its study does not depend on the dimension, i.e., increasing or decreasing can occur without cutting [26, 31, 38]. Using the neighborhood system, graphs have been represented topologically, as in [14, 32], and some topologies have been represented by neighborhoods and graphs, as in [28]. Recently, both graphs and rough sets have been used to represent structures such as self-similar fractals [11, 15], the human heart [12, 13, 33] and DNA [17–19], making them useful in physics, medicine and biology [2–4], respectively.

Graph theory is a mathematical tool to solve some real-life problems. Graphs can be used to model many types of relations and processes in physical, biological [30], social and information systems. Many practical problems can be represented by graphs. Many previous studies have investigated the similarity of genetic sequences [20, 21, 27, 34].

Our aim with this paper is to examine the existence of gene mutations based on relations, and through the use of metric space, topological structures and graph-based models. We generate a code that depends on the multiset, the relation and the metric space between the sequences of genes (DNA sequences) to determine the presence of the mutation, its locations, if any, and the amino acids. Graph theory is used to determine mutations of genes, and the similarity between DNA sequences will be studied. Finally, we combine the concept of a multiset and association indices to study the similarity between mutations for genes.

2. Preliminaries

In what follows, a short survey of multisets and the corresponding theories will be given based on the work of Yager in [39]. Also, we introduce a short survey on genetic mutations, as described in [5, 37], and some methods for mutation analysis and mutation detection are mentioned.

Definition 2.1. [9, 24, 39] A multiset M assigned from a nonempty set X and presented by a function $C_M(x) : X \to N$, where N denotes the natural numbers. $C_M(x)$ represents the number of element x which occurs in M. In other words, M from $X = \{x_1, x_2, \cdots, x_n\}$ to N is written as $M = \{\frac{m_1}{x_1}, \frac{m_2}{x_2}, \cdots, \frac{m_n}{x_n}\}$, where m_i is the number of x_i, for $i = 1, 2, \cdots, n$ that can occur in M.

Proposition 2.2. *Let M and N be two multisets assigned on X. Then, the following holds*

(i) $M = N$ if $C_M(x) = C_N(x)\ \forall\ x \in X$.

(ii) $M \subseteq N$ if $C_M(x) \leq C_N(x)\ \forall\ x \in X$.

(iii) $W = M \cup N$ if $C_W(x) = Max\{C_M(x), C_N(x)\}\ \forall\ x \in X$.

(v) $W = M \ominus N$ if $C_W(x) = Max\{C_M(x) - C_N(x), 0\}\ \forall\ x \in X$.

(vi) $W = M \oplus N$ if $C_W(x) = C_M(x) + C_N(x)\ \forall\ x \in X$.

(vii) $W = M \cap N$ if $C_W(x) = Min\{C_M(x), C_N(x)\}\ \forall\ x \in X$.

Here, $\ominus$ and $\oplus$ denote a multiset subtraction and a multiset addition, respectively. It is noted that any set is a special case of multiset.

Definition 2.3. [22] Let K_1 and K_2 be two multisets assigned on X, and have C_{K_1} and C_{K_2}, respectively. The Cartesian product of K_1 and K_2 is defined by $K_1 \times K_2 = \{\frac{(\frac{m}{x}, \frac{n}{y})}{mn} : x \in^m K_1, y \in^n K_2\}$.

Definition 2.4. [22] A submultiset $\mathcal{R}$ of $M \times M$ is said to be a multiset relation on M, if for each $(\frac{m}{x}, \frac{n}{y})$ of $\mathcal{R}$, there is a product of $C_1(x, y)$ and $C_2(x, y)$ which can be counted. The relationship between $\frac{m}{x}$ and $\frac{n}{y}$ can be formulated as $\frac{m}{x}\mathcal{R}\frac{n}{y}$.

In graph theory [10], the set of vertices will be denoted by V of a finite set. The set of edges have the form $E(V) = \{\{u, v\}$ s. t. $u, v \in V\ u \neq v\}$. In other words, u, v are called adjacent vertices. In this paper, the graph will be denoted by $G = (V, E)$, V_G is the vertex of G and E_G is the set of edges. The graph $G = (V_G, E_G)$ is a directed graph if each edge has a direction. The Łkaszyk-Karmowski distance function [43] is a function defining a distance between two random variables or two random vectors. The axioms of this function are as follows:

- $d(x, y) > 0$,
- $d(x, y) = d(y, x)$,
- $d(x, z) \leq d(x, y) + d(y, z)$.

3. Mutations from the viewpoint of multisets, relations and metric spaces

In this section, the concepts of multisets, metric spaces and multiset relations are applied in MSC code building. The MSC code determines the existence of a mutation locations and number of mutations; it also identifies amino acids. The MSC code specifies the number of (A, T, G, C) elements, the relation between them, the places of their difference their numbers, and the different amino acids. Also, the code can be used to study the similarity between the DNA sequences (numbers of elements, matches and mismatches).

Definition 3.1. Let $M_{5'}^{3'}$ be a sense strand of DNA and $M_{3'}^{5'}$ be an antisense strand of DNA. Define a multiset of DNA sequence as $M = \{\frac{m_i}{x} : x \in \{A, T, G, C\}$, where m_i is the time of occurring for $x\}$

Remark 3.2. In Definition 3.1, there are two DNA multisets M_1 and M_2 for $M_{5'}^{3'}$ and $M_{3'}^{5'}$, respectively.

Definition 3.3. Let M_1 and M_2 be DNA multisets. Define the DNA Cartesian product of M_1 and M_2 by $M_1 \times M_2 = \{\frac{(\frac{m}{x},\frac{n}{y})}{mn} : x \in^m M_1, y \in^n M_2\}$.

Definition 3.4. Let M_1 and M_2 be DNA multisets. Define multibinary relation $R \subseteq M_1 \times M_2 = \{(\frac{m}{x}, \frac{n}{y}) : x \in^m M_1, y \in^n M_2\}$.

Definition 3.5. Let M_1 and M_2 be DNA multisets. Define the correlation coefficient between M_1 and M_2 as $C_{M_1,M_2} = \frac{1}{|M_1||M_2|} \sum_{x_i \in M_1, y_i \in M_2} C(x_i)C(y_i)$, where $C(x_i)$ and $C(y_i)$ are time of occurring for x_i and y_i in M_1 and M_2, respectively.

Corollary 3.6. $C_{M_1,M_1} = \frac{1}{|M_1|^2} \sum_{x_i \in M_1} (C(x_i))^2$.

Corollary 3.7. $0 < C_{M_1,M_2} \leq 1$.

Proof. Since $M_1 \neq \phi \rightarrow |M_1| \neq 0$, $M_2 \neq \phi \rightarrow |M_2| \neq 0$ and $|M_1| \geq C_{M_1}(x_i)$,$|M_2| \geq C_{M_1}(y_i)$, then $|M_1||M_2| \geq \sum C_{M_1}(x_i)C_{M_2}(y_i)$. Therefore, $0 < C_{M_1,M_2} \leq 1$. □

Remark 3.8. $n(M_1)$ is the number of elements existing in DNA multiset M_1, and $n(M_1) = 4$ at most.

Definition 3.9. Let M_1 and M_2 be DNA multisets, $n(M_1) = n(M_1)$. Define the distance function between M_1 and M_2 as $d_{DNA}(M_1, M_2) = \frac{C_{M_1,M_2}}{\sqrt{C_{M_1,M_1} \times C_{M_2,M_2}}}$.

Remark 3.10. From Definition 3.9, $d_{DNA}(M_1, M_2) = \frac{\sum_{x_i,y_i} C(x_i)C(y_i)}{\sqrt{\sum_{x_i}(C_{M_1}(x_i))^2 \times \sum_{x_i}(C_{M_2}(y_i))^2}}$.

Theorem 3.11. *$d_{DNA}(M_1, M_2)$ is associated with the following axioms:*

(i) $d_{DNA} > 0$,

(ii) $d_{DNA}(M_1, M_1) = 1$,

(ii1) $d_{DNA}(M_1, M_2) = d_{DNA}(M_2, M_1)$,

(iv) $d_{DNA}(M_1, M_3) + d_{DNA}(M_3, M_2) \geq d_{DNA}(M_1, M_2)$.

Proof. (i) By Corollaries 3.6 and 3.7, $0 < C_{M_1,M_2} \leq 1$, $C_{M_1,M_1} > 0$ and $C_{M_2,M_2} > 0$. Then, $d_{DNA} > 0$.

(ii) By Corollary 3.6, $d_{DNA}(M_1, M_1) = 1$.

(iii) By Definition 3.5, $d_{DNA}(M_1, M_2) = \frac{C_{M_1,M_2}}{\sqrt{C_{M_1,M_1} \times C_{M_2,M_2}}} = \frac{C_{M_2,M_1}}{\sqrt{C_{M_2,M_2} \times C_{M_1,M_1}}} = d_{DNA}(M_2, M_1)$.

(iv) Since $|M_1| \geq C_{M_1}(x_i)$, $|M_2| \geq C_{M_2}(y_i)$ and $|M_3| \geq C_{M_3}(z_i)$, then, by Definition 3.5, $C_{M_1,M_2} = \frac{1}{|M_1||M_2|} \sum_{x_i \in M_1, y_i \in M_2} C(x_i)C(y_i)$, $C_{M_2,M_3} = \frac{1}{|M_2||M_3|} \sum_{z_i \in M_3, y_i \in M_2} C(z_i)C(y_i)$ and $C_{M_1,M_3} = \frac{1}{|M_1||M_3|} \sum_{x_i \in M_1, z_i \in M_3} C(x_i)C(z_i)$. To prove that $d_{DNA}(M_1, M_3) + d_{DNA}(M_3, M_2) \geq d_{DNA}(M_1, M_2)$, it is sufficient to prove that $\frac{C_{M_1,M_2}}{\sqrt{C_{M_1,M_1} \times C_{M_2,M_2}}} \leq \frac{C_{M_1,M_3}}{\sqrt{C_{M_1,M_1} \times C_{M_3,M_3}}} + \frac{C_{M_3,M_2}}{\sqrt{C_{M_3,M_3} \times C_{M_2,M_2}}}$. Since $C_{M_1,M_3} + C_{M_3,M_2} = \frac{1}{|M_1||M_3|} \sum_{x_i \in M_1, z_i \in M_3} C(x_i)C(z_i) + \frac{1}{|M_3||M_2|} \sum_{z_i \in M_3, y_i \in M_2} C(z_i)C(y_i) = \frac{1}{|M_2||M_1||M_3|} |M_2| \sum_{x_i \in M_1, z_i \in M_3} C(x_i)C(z_i) + \frac{1}{|M_1||M_3||M_2|} |M_1| \sum_{z_i \in M_3, y_i \in M_2} C(z_i)C(y_i) \geq \frac{1}{|M_2||M_1||M_3|} |M_2| \sum_{x_i \in M_1, z_i \in M_3} C(x_i)C(z_i) \geq \frac{1}{|M_2||M_1||M_3|} \sum_{y_i \in M_2} C(y_i) \sum_{x_i \in M_1, z_i \in M_3} C(x_i)C(z_i) \geq \frac{1}{|M_1||M_2|} \sum_{x_i \in M_1, y_i \in M_2} C(x_i)C(y_i) = C_{M_1,M_2}$, where $|M_2||M_1| \leq \sum (C(x_i))^2 \times \sum (C(y_i))^2$, $|M_2||M_3| \leq \sum (C(y_i))^2 \times$

$\sum (C(z_i))^2$, $|M_3||M_1| \leq \sum (C(z_i))^2 \times \sum (C(x_i))^2$,$|M_2||M_1| \geq \sqrt{\sum (C(x_i))^2 \times \sum (C(y_i))^2}$, $|M_2||M_3| \geq \sqrt{\sum (C(y_i))^2 \times \sum (C(z_i))^2}$ and $|M_3||M_1| \geq \sqrt{\sum (C(z_i))^2 \times \sum (C(x_i))^2}$. Therefore, $|M_2||M_1| \geq \sqrt{|M_2||M_1|} \geq \sqrt{C_{M_1,M_1} \times C_{M_2,M_2}}$. □

Remark 3.12. We can call the function d_{DNA} a DNA metric space.

Remark 3.13. Theorem 3.11 satisfies the condition of Łkaszyk-Karmowski distance.

Proposition 3.14. *The distance function* $1 - d_{DNA}$ *is a metric space.*

Proof. Refer to Theorem 3.11. □

Theorem 3.15. *If* $d_{DNA}(M_1, M_2)$ *=1, then there is no mutation.*

Proof. Let $M_1 = \{n_1/G, n_2/A, n_3/T, n_4/C\}$, $M_2 = \{m_1/C, m_2/T, m_3/A, m_4/G\}$ and $d_{DNA}(M_1, M_2)$ =1. Then, by Theorem 3.11, we get that $d_{DNA}(M_1, M_2) = 1 = \frac{C_{M_1,M_2}}{\sqrt{C_{M_1,M_1} \times C_{M_2,M_2}}}$. Then, $C_{M_1,M_2} = \sqrt{C_{M_1,M_1}}$ implies that $C^2_{M_1,M_2} = C_{M_1,M_1} \times C_{M_2,M_2}$. Using Definition 3.5, $(n_1m_1 + n_2m_2 + n_3m_3 + n_4m_4)^2 = (n_1^2 + n_2^2 + n_3^2 + n_4^2) \cdot (m_1^2 + m_2^2 + m_3^2 + m_4^2)$. Then, $n_1 = \lambda m_1$, $n_2 = \lambda m_2$, $n_3 = \lambda m_3$ and $n_4 = \lambda m_4$. So, if $\lambda = 1$, then $n_1 = m_1$, $n_2 = m_2$, $n_3 = m_3$ and $n_4 = m_4$. This means that there is no mutation. □

Corollary 3.16. *From Theorem 3.15, we have that* $d_{DNA}(M_1, M_2) \neq 1$*; then, there is a mutation.*

We present the MSC code in the Appendix as an algorithm which is used to generate multisets, relations and a metric space between M_1 and M_2. Some examples are given to illustrate the proposed results and MSC code algorithm.

Example 3.17. ***Arabidopsis thaliana gamma-glutamylcysteine synthetase gene (abbr. CAD2) [44]***

Tair Accession: *1005028114.*
GenBank Accession: *AF068299.*
Sequence Length *5277.*

$5'$ $ATCGATATGTAACACAAT \cdots TGTATGTTTTT$ $3'$*;*

$3'$ $TAGCTATACATTGTGTTA \cdots ACATACAAAAA$ $5'$*. Using the MSC code algorithm, we have* $M_1 = \{\frac{1019}{G}, \frac{1543}{A}, \frac{1859}{T}, \frac{856}{C}\}$, $|M_1| = 5277$*;*

$M_2 = \{\frac{1019}{C}, \frac{1543}{T}, \frac{1859}{A}, \frac{856}{G}\}$, $|M_2| = 5277$.

The distance between M_1 *and* M_2 *equals* 1 *(no mutation). The relation between* M_1 *and* M_2 *according to their MSC code, is described in Table 1. The relation between* M_1 *and* M_2 *is*

$\mathcal{R} = \{(\frac{1859}{T}, \frac{1859}{A}), (\frac{1543}{A}, \frac{1543}{T}), (\frac{1019}{G}, \frac{1019}{C}), (\frac{856}{C}, \frac{856}{G})\}$*. This relation indicates no mutation.*

Table 1. Bonding between nucleotides.

	A	T	C	G
A	0	1859	0	0
T	1543	0	0	0
C	0	0	0	1019
G	0	0	856	0

Example 3.18. *If we do a mutation in CAD2 [44] in Example 3.1. Using the MSC code algorithm, we have* $M_1 = \{\frac{1014}{G}, \frac{1539}{A}, \frac{1859}{T}, \frac{860}{C}\}$, $|M_1| = 5272$*;* $M_2 = \{\frac{1014}{C}, \frac{1543}{T}, \frac{1859}{A}, \frac{856}{G}\}$, $|M_2| = 5272$.

The distance between M_1 and M_2 equals 0.9999979580282978 *according to the MSC code; the result was a mutation and this corresponds to data reported by the National Center for Biotechnology Information (NCBI) [44]. The position of the mutation is*

$[2568, 2578, 2595, 2609, 2639, 5076]$;

$[C, \quad T, \quad , \quad C, \quad C, \quad , \quad G, \quad C]$;

$[T, \quad T, \quad C, \quad T, \quad A, \quad T]$.

The amino acid resulting from the mutation is presented in Table 2. The relation between M_1 and M_2 is outlined in Table 2. The relation between M_1 and M_2 is $\mathcal{R} = \{(\frac{1}{G}, \frac{1}{A}), (\frac{1}{T}, \frac{1}{T}), (\frac{3}{C}, \frac{3}{T}), (\frac{1}{C}, \frac{1}{C}), (\frac{1539}{A}, \frac{1539}{T}), (\frac{1858}{T}, \frac{1858}{A}), (\frac{1013}{G}, \frac{1013}{C}), (\frac{856}{C}, \frac{856}{G})\}$ according to their MSC code as described in Table 3. This relation indicates a mutation.

Table 2. The amino acid formula.

	$5' \cdots 3' = M_1$	$3' \cdots 5' = M_2$	Amino acid $5' \cdots 3' = M_1$	Amino acid $3' \cdots 5' = M_2$	Position
0	T	C	ATT	TAC	2568
1	T	T	TTA	TAT	2578
2	C	C	TGC	ACC	2595
3	T	C	TTT	ACA	2609
4	A	G	AAA	TGT	2639
5	T	C	ATT	TAC	5076

Table 3. Bonding between nucleotides.

	A	T	C	G
A	0	1858	0	1
T	1539	1	3	0
C	0	0	1	1013
G	0	0	856	0

Corollary 3.19. *From Theorem 3.15, the relation will be $R = \{(n_1/G, n_1/C)/n_1n_1, (n_2/A, n_2/T)/n_2n_2, (n_3/T, n_3/A)/n_3n_3, (n_4/C, n_4/G)/n_4n_4\}$.*

Proposition 3.20. *Let R be a relation between DNA multisets M_1 and M_2. Then, if R is either reflexive or transitive, then there is a mutation.*

Proof. Suppose that $M_1 = \{\frac{n_1}{x} : x \in \{A, T, G, C\}$, where n_1 is the number occurrences of $\{x\} = \{n_1/G, n_2/A, n_3/T, n_4/C\}$, $M_2 = \frac{m_1}{y} : y \in \{A, T, G, C\}$, where m_1 is the number occurrences of $\{y\} = \{m_1/C, m_2/T, m_3/A, m_4/G\}$ and $R = \{(\frac{m}{x}, \frac{n}{y})\} : x \in^m M_1, y \in^n M_2\}$. Then, every C is linked with G. On the other side, A is linked with $T \equiv U$. Otherwise, a mutation will be occurred. □

4. Topological structures of DNA mutations

In the study of the congruence and determination of the presence of mutations between DNA sequences, we have four bases $\{A, T, G, C\}$; thus, we have the 12 mutation rates $A \to C$, $A \to G$, $\cdots$,

$T \to G$ at a particular site. In the study of the similarity between the DNA sequences, we have 12 difference rates $A \to T$, $A \to G$, $\cdots$, $T \to A$ at a particular site. We can use these rates to study SARS-CoV-2 through the mutation of its genes. So, our study can yield a model of the pattern of mutations in SARS-CoV-2 and the alternative model for the mutations that occur in SARS- CoV-2 can be developed. If the length $n_{i,j} > 0$, then we suggest $S_i = \{A, T, C, G\}$, which has more than an average likelihood of linking to S_j. For each $i = 1, \cdots, 4$, define $\mathcal{R}_i = S_i \cup S_j$, $n_{i,j} > 0$. The collection $\mathcal{R}_0 = \{R_i\}_i^4$ is not itself a topology, but we extend it to one, defining 0 to be a minimal topological structure on genotypes containing $\mathcal{R}_0$. The topological space τ_0 will be generated by a basis induced by a finite intersection of the sets in $\mathcal{R}_0$. The topological structure is referred as a mutation space and is called a mutation topological structure.

In this section, we use the proposed MSC code algorithm, the following definition is given.

Definition 4.1. Let X be the set of nucleotides of a DNA sequence such that $X = \{A, T, G, C\}$, and let there exist a bonding between $x_i, MSCy_j$ in X, referred to as $n(x_i, y_j) \neq 0$. Otherwise, $n(x_i, y_j) = 0$.

Definition 4.2. Let X be the set of nucleotides of a DNA sequence. Define a relation $R^* = \{(x, y) : n(x, y) \neq 0, x, y \in X\}$.

We state some properties for the cases of mutations and no mutation.
(i) If there is no mutations, then

- R^* is not reflexive,
- R^* is symmetric,
- R^* is transitive.

(ii) If there is a mutation, then R^* may be reflexive, symmetric and transitive.

Example 4.3. *NM 000518.4 Homo sapiens hemoglobin subunit beta (abbr. HBB), mRNA [44] sequence: HBB gene range:* 1 *to* 626*;*
$5^{'}$ $ACATTTGCTT \cdots CATTGC$ $3^{'}$*;*
$3^{'}$ $TGTAAACGAA \cdots GTAACG$ $5^{'}$*;*
$M_1 = \{\frac{157}{G}, \frac{167}{A}, \frac{137}{T}, \frac{165}{C}\}$, $|M_1| = 626$*;*
$M_2 = \{\frac{157}{C}, \frac{167}{T}, \frac{137}{A}, \frac{165}{G}\}$, $|M_2| = 626$.
The relation between M_1 and M_2 according to the MSC code is described in Tables 4 and 5. $\mathcal{R}^* = \{(T, A), (A, T), (G, C), (C, G)\}$. This relation indicates no mutation and is consistent with the report by the NCBI [44].

Table 4. Bonding between nucleotides.

	A	T	C	G
A	0	137	0	0
T	167	0	0	0
C	0	0	0	157
G	0	0	165	0

Table 5. Relation between nucleotides for $n_{i,j} \neq 0$.

	A	T	C	G
A	-	√	-	-
T	√	-	-	-
C	-	-	-	√
G	-	-	√	-

Example 4.4. *If we do a mutation in CAD2, then* $M_1 = \{\frac{1012}{G}, \frac{1541}{A}, \frac{1856}{T}, \frac{856}{C}\}$, $|M_1| = 5265$, *and* $M_2 = \{\frac{1012}{C}, \frac{1540}{T}, \frac{1858}{A}, \frac{855}{G}\}$, $|M_2| = 5265$. *The position of the mutation* $[17, 686, 5073]$ *is* $[G, \ C, \ A]$ *and* $[A, \ A, \ C]$. *The amino acid which results from the mutation is presented in Table 6. The distance between* M_1 *and* M_2 *equals* 0.999999362175112, *according to the MSC code. The relation between* M_1 *and* M_2 *according to the MSC code, is described in Tables 7 and 8.* $\mathcal{R}^* = \{(T, A), (C, A), (G, A), (A, T), (A, C), (G, C), (C, G)\}$. *So, there is a mutation.*

Table 6. The amino acid formula.

	$5' \cdots 3'$	$3' \cdots 5'$	Amino acid $5' \cdots 3'$	Amino acid $3' \cdots 5'$	Position
0	G	A	AGA	TAT	17
1	C	A	TCT	AAA	686
2	A	C	ACA	TGC	5073

Table 7. Relation between nucleotides.

	A	T	C	G
A	0	1856	1	1
T	1540	0	0	0
C	1	0	0	1011
G	0	0	855	0

Table 8. Relation between nucleotides for $n_{i,j} \neq 0$.

	A	T	C	G
A	-	√	√	√
T	√	-	-	-
C	√	-	-	√
G	-	-	√	-

Next, a topological structure will be defined in terms of R^*.

Definition 4.5. Let R^* be a relation on X. Define a subbase $S = \{xR^* : x \in X\}$ for some topology τ_{DNA} on X.

Example 4.6. *(continued from Example 4.3)* $S = \{AR^*, TR^*, CR^*, GR^*\} = S = \{\{A\}, \{T\}, \{C\}, \{G\}\}$. *Then, the base* β *will be* $\{X, \{A\}, \{T\}, \{C\}, \{G\}\}$ *and* $\tau_{DNA} = \{X, \phi, \{A\}, \{T\}, \{C\}, \{G\}, \{A, T\}, \{T, C\}, \{C, G\}, \{G, A\}, \{A, C\}, \{T, G\}, \{A, T, C\}, \{A, T, G\}, \{T, C, G\}, \{A, G, C\}\}$. *Therefore, this space is discrete. This means that, if every subset of* $X = \{A, T, C, G\}$ *is open and closed, then there is no mutation.*

Example 4.7. *(continued from Example 4.4)* $S = \{\{T, C\}, \{A\}, \{A, G\}, \{A, C\}\}$. *Therefore, a base* $\beta = \{X, \{T, C\}, \{A\}, \{A, G\}, \{A, C\}, \{C\}\}$ *and* $\tau_{DNA} = \{X, \phi, \{T, C\}, \{A\}, \{A, G\}, \{A, C\}, \{C\}, \{A, T, C\}, \{A, C, G\}\} \equiv$ *general topology.*

Now, the existence of a mutation will be determined based on the type of topological structure.

Proposition 4.8. *If the DNA sequence has a mutation, then the generated topology is a general topological structure.*

Proof. Let τ_{DNA} be a class of sets on X generated by the mutation of DNA sequences. Consider that $\tau_{DNA} = \{G : G = \bigcup_i (\bigcap_j^n A_{ij}), A_{ij} \in S\}$ is a class of sets of X. Now, it is sufficient to prove that τ_{DNA} is a topological structure.

(i) $\cap_{j\in\phi}^n A_{ij} = X \in \tau_{DNA}$ and $\bigcup_{i\in\phi}(\bigcap_j^n A_{ij}) = \phi \in \tau_{DNA}$.

(ii) $G_1, G_2, \cdots G_n \in \tau_{DNA}$; then, $G_1 = \bigcup_{i_1}(\bigcap_{j_1}^n A_{i_1 j_1})$, $G_2 = \bigcup_{i_2}(\bigcap_{j_2}^n A_{i_2 j_2}), \cdots, G_n = \bigcup_{i_n}(\bigcap_{j_n}^n A_{i_n j_n})$. $G_1 \cap G_2 \cap \cdots \cap G_n = \bigcup_{i_1, i_2 \cdots i_n}(\bigcap_{j_1}^n A_{i_1 j_1}) \cap \bigcap_{j_2}^n A_{i_2 j_2} \cap \cdots \cap \bigcup_{i_n} \bigcap_{j_n}^n A_{i_n j_n}) = \bigcup_{i_1, i_2, \cdot, i_n}(\bigcap_{k_1}^n B_{k_1 j_k})$, where $B_{kj_k} = A_{i_1 j_1} \cap A_{i_2 j_2} \cap \cdots \cap A_{i_n j_n}$. Since each $A_{i_1 j_1}, A_{i_2 j_2} \cdots A_{i_n j_n} \in S$, $B_{kj_k} \in \beta$; therefore, $G_1 \cap G_2 \cap \cdots \cap G_n \in \tau_{DNA}$.

(ii) $G_1, G_2, \cdots G_n \cdots \in \tau_{DNA}$; then, $G_1 \cup G_2 \cup \cdots \cup G_n \cup \cdots = \bigcup_{i_1, i_2 \cdots i_n}(\bigcap_{j_1}^n A_{i_1 j_1} \cup \bigcap_{j_2}^n A_{i_2 j_2} \cup \cdots \cup \bigcup_{i_n} \bigcap_{j_n}^n A_{i_n j_n} \cup \cdots)$. Hence, $G_1 \cup G_2 \cup \cdots \cup G_n \cup \cdots \in \tau_{DNA}$. □

Proposition 4.9. *If there is no mutation in DNA, then* τ_{DNA} *is a discrete topology.*

Proof. Suppose that the DNA sequence has no mutation. Then, $n(x, x) = 0$ and $n(x, y) \neq 0 \; \forall \; x, y \in X$. Hence, $xR^* = \{\{y\} : \forall \; y \in X\}$. This means that $A \to T$, $C \to G$, $T \to A$, $G \to C$. So, $S = \{\{y\} : y \in X\}$. Therefore, τ_{DNA} is discrete. □

The converse of Proposition 4.9 may not be true, in general.

Example 4.10. $S = \{\{T\}, \{A\}, \{C, G\}, \{C\}\}$ *is a subbase for a discrete topological structure. But, there is a mutation because G is bonded with C.*

Example 4.11. *Let a DNA sequence be* $5'$ *ACGT* $3'$ *and* $3'$ *GATC* $5'$. *Then,* $S = \{\{G\}, \{A\}, \{T\}, \{C\}\}$ *and* τ_{DNA} *is a discrete topology. But, there is a mutation, as no gene consists of only four nucleotides (length 4 only or a little more); the length of a gene is measured in kilobytes.*

Corollary 4.12. *If the topological structure generated from the DNA sequences has*

(1)- a general topological structure, then there is a mutation;

(2)- a discrete topology, then there may or may not be a mutation. We use Theorem 3.15, Corollary 3.16 and Proposition 3.20 to figure out if there is a mutation.

5. Mutations by similarity of genes based on the association indices and multisets

Bass et al. [6] provided an overview of commonly used association indices, including the Jaccard index and the Pearson correlation coefficient, and compared their performance on different types of analysis for a biological network. An association index is a measure that quantifies interaction profile similarity. They discussed the differences and similarities between association indices. There exist many association indices:

(i) The Jaccard index: $J_{AB} = \frac{|N_A \cap N_B|}{|N_A \cup N_B|}$;

(ii) The Simpson index: $S_{AB} = \frac{|N_A \cap N_B|}{Min\{|N_A|,|N_B|\}}$;

(iii) The geometric index: $G_{AB} = \frac{|N_A \cap N_B|^2}{|N_A|.|N_B|}$;

(iv) The cosine index: $C_{AB} = \frac{|N_A \cap N_B|}{\sqrt{|N_A \cdot N_B|}}$.

We apply association indices to determine whether there is a mutation by calculating the association indices for each pair of nucleotides $\{A, T, G, C\}$ of the gene. If the associations are zero, there is no mutation; otherwise, there is a mutation.

Example 5.1. *(continued from Example 4.3)*

$e_1 = 167 = e_5$, $e_2 = 157 = e_7$, $e_3 = 165 = e_6$, $e_4 = 137 = e_8$.

Let X-type $= \{A\}$, *Y-type* $= \{T\}$.

Then, we have the following:

Jaccard$= \frac{0}{2} = 0$, *Simpson* $= \frac{0}{1} = 0$, *Geometric* $= \frac{0}{1} = 0$, *Cosine* $= \frac{0}{1} = 0$. *Then, Jaccard= Simpson= Geometric= Cosine= 0. This is for all nucleotides* $\{A, T, G, C\}$. *Hence, there is no mutation. This is shown in Figures 1–3.*

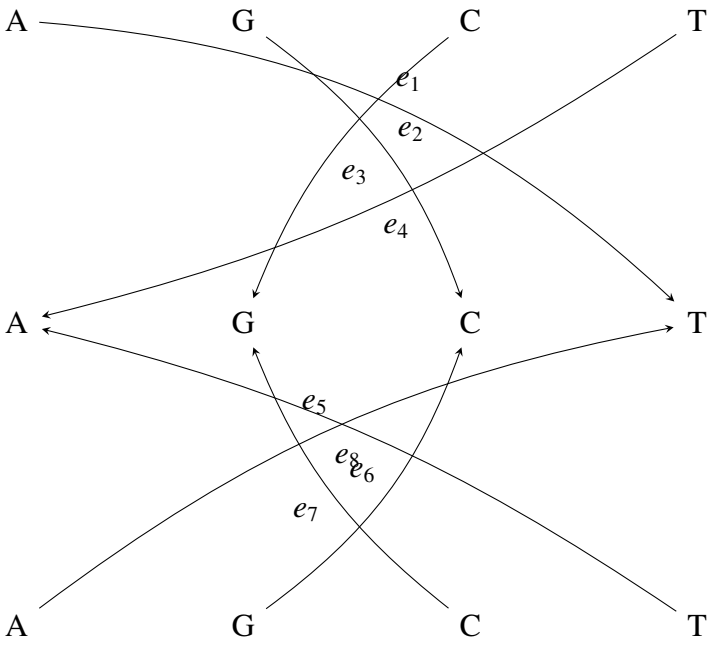

Figure 1. Relations between the nucleotides of HBB gene.

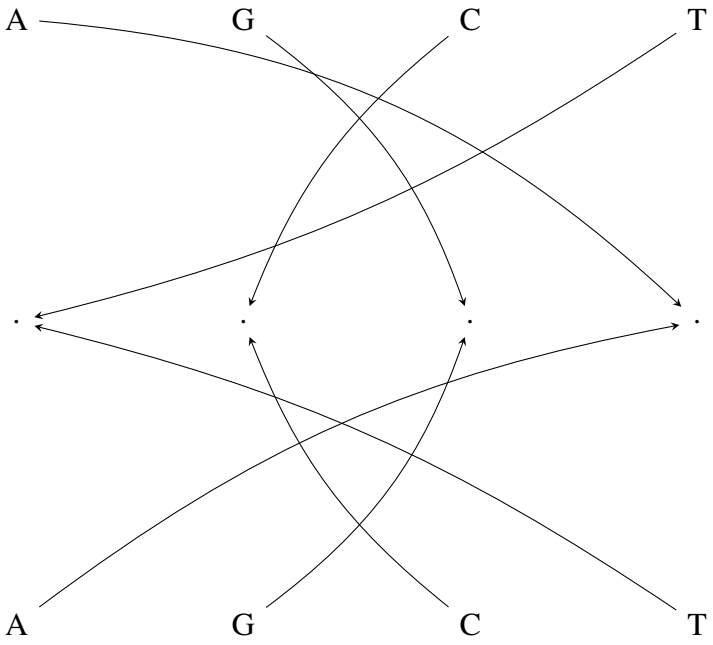

Figure 2. The graph depicts the associations within a nucleotide of the HBB gene.

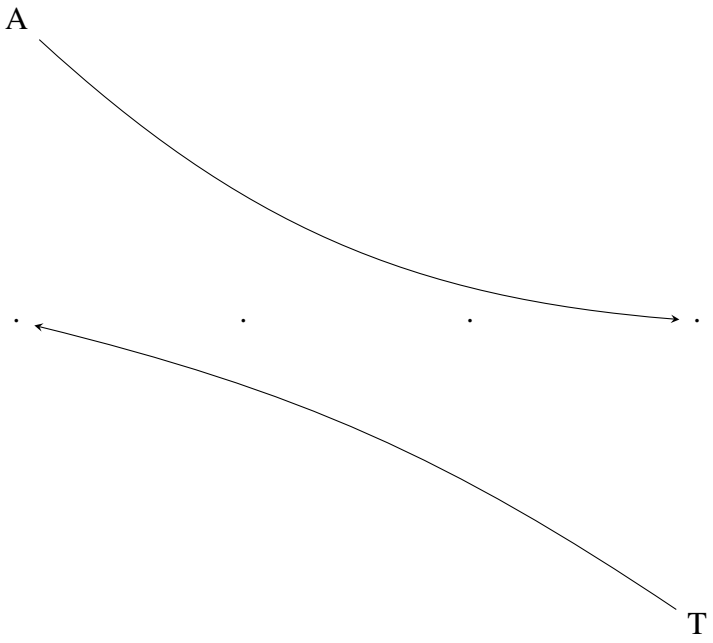

Figure 3. Relation between nucleotides of X-type and Y-type of HBB gene.

Example 5.2. *(continued from Example 4.4)*
Let X-type = $\{A\}$, *Y-type*= $\{G\}$.

Since $e_1 = 1540 = e_5$, $e_2 = 1011 = e_7$, $e_3 = 855 = e_6$ *and* $e_4 = 1856 = e_8$, $e_9 = 1 = e_{12}$, $e_{10} = 1 = e_{13}$ *and* $e_{11} = 1 = e_{14}$, *we have the following:*

Jaccard= $\frac{1}{2}$,

Simpson= $\frac{1}{1}$,
Geometric= $\frac{1}{2}$,
Cosine= $\frac{1}{\sqrt{2}}$.
Then, Jaccard, Simpson, Geometric and Cosine $\neq 0$. *This is for all nucleotides* A, T, G, C. *Hence, there is a mutation. This is shown in Figures 4–7.*

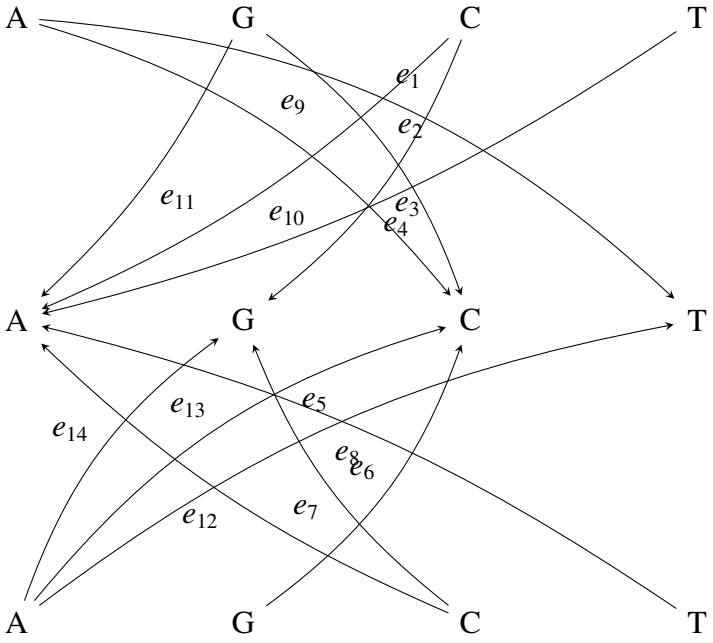

Figure 4. Relations between the nucleotides of a CAD2 gene.

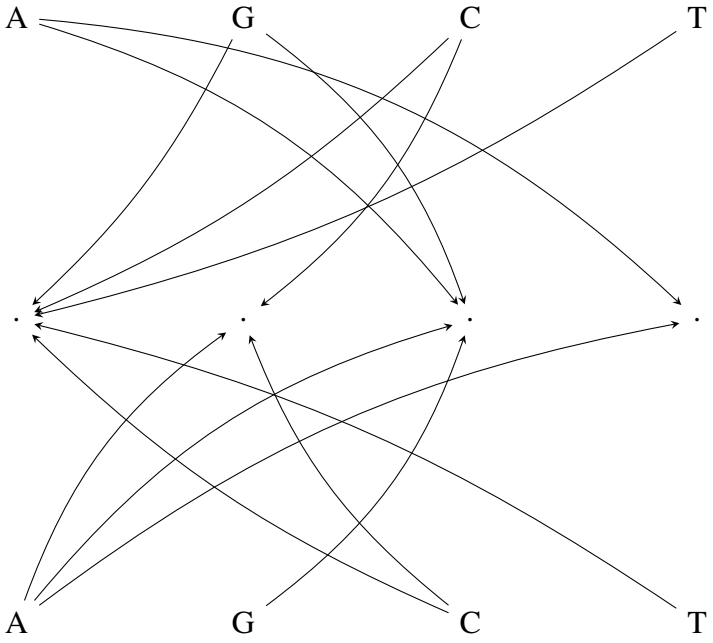

Figure 5. The associations within a nucleotide of a CAD2 gene.

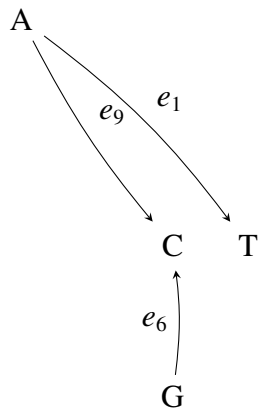

Figure 6. Relations between the nucleotides A,G in a CAD2 gene.

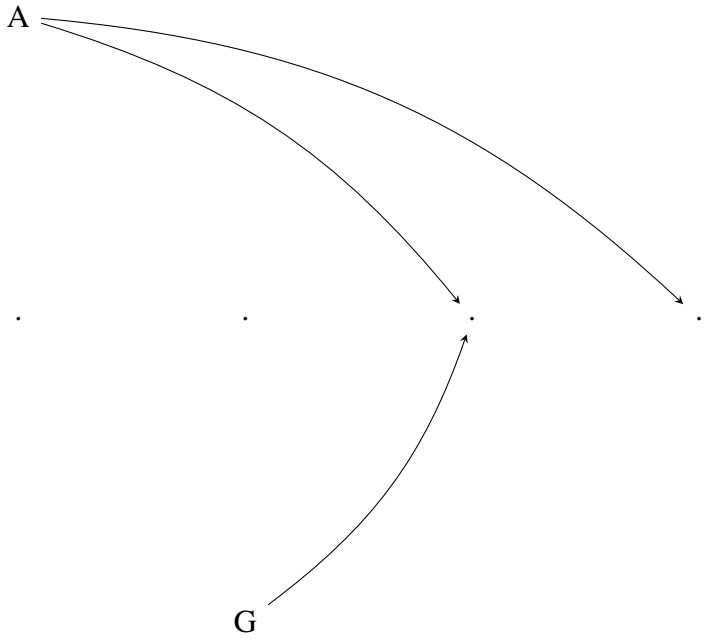

Figure 7. Relation between the nucleotides of X-type and Y-type in CAD2 gene.

Note that the degree of a node A, say, $|N_A|$, is defined as the number of nodes with which it interacts and $|N_A \cap N_B|$ is the shared partners. Biological processes are implemented through complex interaction networks. Metrics known as association indices can be used to quantify the similarity between genes through the use of a multiset. So, $|N_A|$ is the cardinality of a multiset M. Then, the similarity association indices become

$MJ_{AB} = \frac{|N_A \cap N_B|}{|N_A \cup N_B|}$, $MS_{AB} = \frac{|N_A \cap N_B|}{Min\{|N_A|,|N_B|\}}$, $MG_{AB} = \frac{|N_A \cap N_B|^2}{|N_A| \cdot |N_B}$ and $MC_{AB} = \frac{|N_A \cap N_B|}{\sqrt{|N_A| \cdot |N_B|}}$.

The dissimilarity association indices are as follows

$M * J_{AB} = 1 - \frac{|N_A \cap N_B|}{|N_A \cup N_B|}$, $M * S_{AB} = 1 - \frac{|N_A \cap N_B|}{Min\{|N_A|,|N_B|\}}$, $M * G_{AB} = 1 - \frac{|N_A \cap N_B|^2}{|N_A| \cdot |N_B}$ and $M * G_{AB} = 1 - \frac{|N_A \cap N_B|}{\sqrt{|N_A| \cdot |N_B|}}$.

Remark 5.3. (i) $M_{AB} \geq M_{AB}$.

(ii) If $|N_A| = |N_B|$, then $M_{AB} = M_{AB}$.

Theorem 5.4. *The similarity association indices are DNA metric spaces.*

Theorem 5.5. *The dissimilarity association indices are metric spaces.*

Example 5.6. *(continued from Example 5.1)*
Let X-type $= \{A\}$*, Y-type* $= \{T\}$*; also,* e_1*,* e_8 *and* $e_8 = 137$.
$N_A = \{\frac{e_1}{T}, \}$, $N_B = \{\frac{e_8}{A}\}$.
Jaccard$= \frac{0}{304} = 0$,
Simpson $= \frac{0}{137} = 0$,
Geometric $= \frac{0}{22879} = 0$,
Cosine $= \frac{0}{151.26} = 0$,
Then, Jaccard= Simpson= Geometric= Cosine$= 0$*. This is for all nucleotides* $\{A, T, G, C\}$ *since* $e_1 = 1540 = e_5$*,* $e_2 = 1011 = e_7$*,* $e_3 = 855 = e_6$ *and* $e_4 = 1856 = e_8$*,* $e_9 = 1 = e_{12}$*,* $e_{10} = 1 = e_{13}$*;* $e_{11} = 1 = e_{14}$*. Hence, there is no mutation. This is shown in Figure 8.*

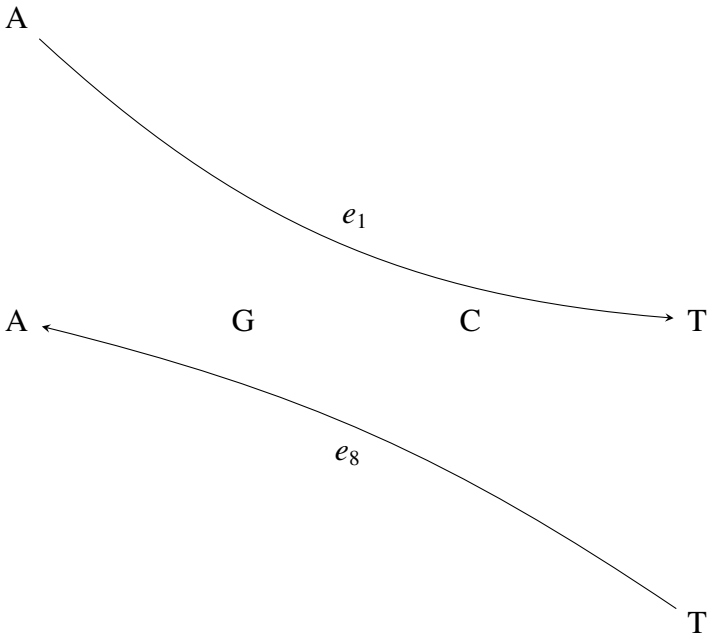

Figure 8. Relation between the nucleotides A,T in a CAD2 gene.

Example 5.7. *(continued from Example 5.2)*
Let X-type$= \{A\}$*, Y-type*$= \{G\}$*,* $e_1 = 1540$*,* $e_6 = 855$ *and* $e_9 = 1$*.* $N_A = \{\frac{e_1}{T}, \frac{e_9}{C}\}$, $N_B = \{\frac{e_6}{C}\}$.
$MJ_{AB} = \frac{|N_A \cap N_B|}{|N_A \cup N_B|} = \frac{|\{\frac{e_1}{T}, \frac{e_9}{C}\} \cap \{\frac{e_6}{C}\}|}{|N_A \cup N_B|} = \frac{|min\{e_9, e_6\}|}{|max\{\{e_6, e_9\}, e_1\}} = \frac{|\{e_9\}|}{|\{e_6, e_1\}} = \frac{1}{2395} \neq 0$,
$MS_{AB} = \frac{|N_A \cap N_B|}{Min\{|N_A|, |N_B|\}} \neq 0$,
$MG_{AB} = \frac{|N_A \cap N_B|^2}{|N_A| \cdot |N_B} \neq 0$,
$MC_{AB} = \frac{|N_A \cap N_B|}{\sqrt{|N_A| \cdot |N_B|}} \neq 0$*. Additionally,* $e_1 = 1540 = e_5$*,* $e_2 = 1011 = e_7$*,* $e_3 = 855 = e_6$*;* $e_4 = 1856 = e_8$*,* $e_9 = 1 = e_{12}$*,* $e_{10} = 1 = e_{13}$ *and* $e_{11} = 1 = e_{14}$.
Then, Jaccard= Simpson= Geometric= Cosine $\neq 0$*. This is for all nucleotides* $\{A, T, G, C\}$*. Hence, there is a mutation. This is shown in Figure 9.*

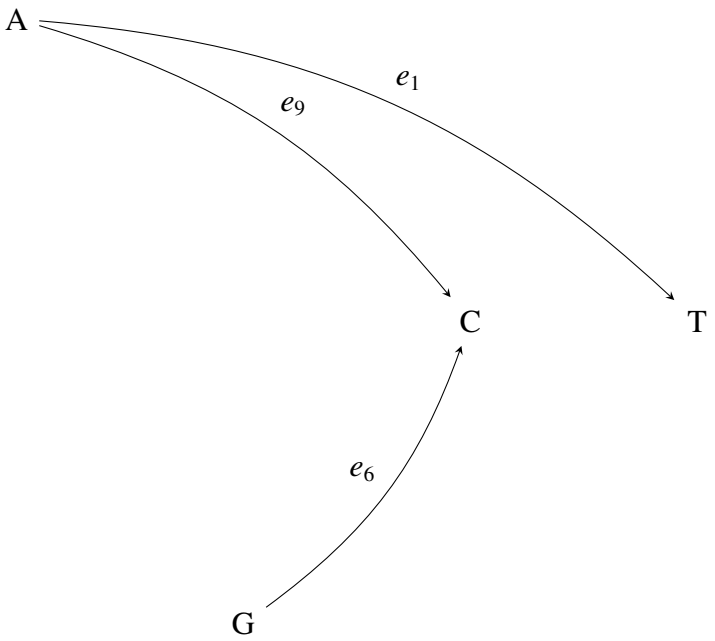

Figure 9. Relation between the nucleotides of X-type = A, Y-type = G in a CAD2 gene.

Example 5.8. *(A similarity and dissimilarity between the sequences of DNA)*
Let $GATACCCCCCGG$, $GATACGACCCGG$, $GATACGCCCCGG$, $CATACGACTCGG$ *and* $GATAGACTCGG$ *be five sequences for DNA. Then,* $A = \{\frac{3}{G}, \frac{2}{A}, \frac{1}{T}, \frac{6}{C}\}$, $|A| = 12$, $B = \{\frac{4}{C}, \frac{1}{T}, \frac{3}{A}, \frac{4}{G}\}$, $|B| = 12$, $C = \{\frac{5}{C}, \frac{1}{T}, \frac{2}{A}, \frac{4}{G}\}$, $|C| = 12$, $D = \{\frac{4}{C}, \frac{2}{T}, \frac{3}{A}, \frac{3}{G}\}$, $|D| = 12$ *and* $E = \{\frac{3}{C}, \frac{2}{T}, \frac{3}{A}, \frac{4}{G}\}$, $|E| = 12$. *Hence,* $M^*J(AB) = 0.286$, $M^*G(AB) = 0.306$, $M^*C(AB) = 0.17$, $M^*S(AB) = 0.17$, $M^*J(AC) = 0.154$, $M^*G(AC) = 0.16$, $M^*C(AC) = 0.084$, $M^*S(AC) = 0.084$, $M^*J(AD) = 0.286$, $M^*G(AD) = 0.306$, $M^*C(AD) = 0.167$, $M^*S(AD) = 0.17$, $M^*J(AE) = 0.4$, $M^*G(AE) = 0.438$, $M^*C(AE) = 0.25$ *and* $M^*S(AE) = 0.25$. *But, the balance of dissimilarity is* $(A, A) = 0$, $(A, B) = 0.17$,$(A, C) = 0.08$, $(A, D) = 0.33$ *and* $(A, E) = 0.25$, *according to the NCBI [44]. By matching the results of the association indices with the reports from the NCBI, it was found that the association indices* M^*C *and* M^*S *are the best.*

6. Conclusions and discussion

The complicated DNA research has become easier by using topology. Recently, many topologists found new methods to examine the mutations of DNA by using a combination of multiset topology and graph theory. Moreover, our presented results for repairing compatibility between the mathematical methods and biological solutions. In addition, we give a decision of mutation that is dependent on the metrics between two sequences of a gene and the topological structure derived from the relations. In the future, we can benefit from mutations by applying them end epidemics and in the fields of industry and agriculture. We have studied and identified mutations and showed how to make new ones, including how to fix mutations and apply Mathematica to construct models. Consequently, they are very significant in decision-making [25, 40–42]. The introduced techniques are very useful in application because they pave the way for more topological applications for real-life problems. We also

have an interesting application of our approaches to DNA sequences. The study of similarity between DNA sequences will be used to solve problems related to diseases and viruses, such as COVID-19 [16], which is an important example of mutations nowadays.

Use of AI tools declaration

The authors declare they have not used artificial intelligence tools in the creation of this article.

Acknowledgment

This work was supported by Researchers Supporting Project number (RSP2023R488), King Saud University, Riyadh, Saudi Arabia.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. C. C. Adamsand, D. F. Robert, *Introduction to Topology: Pure and Applied*, Homewood: Dorsey Press, 2008.
2. T. M. Al-Shami, Soft somewhat open sets: soft separation axioms and medical application to nutrition, *Comp. Appl. Math.*, **41** (2011), 216. https://doi.org/10.1007/s40314-022-01919-x
3. T. M. Al-Shami, Maximal rough neighborhoods with a medical application, *J. Ambient Intell. Human. Comput.*, 2022. https://doi.org/10.1007/s12652-022-03858-1
4. T. M. Al-Shami, On soft separation axioms and their applications on decision-making problem, *Math. Probl. Eng.*, **2021** (2021), 8876978. https://doi.org/10.1155/2021/8876978
5. I. L. Andrulis, H. Anton-Culver, J. Beck, B. Bove, J. Boyd, S. Buys, et al., Comparison of DNA- and RNA-based methods for detection of truncating BRCA1 mutations, *Hum. Mutat.*, **20** (2002), 65–73. https://doi.org/10.1002/humu.10097
6. J. I. F. Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, A. J. M. Walhout, Using networks to measure similarity between genes, association index selection, *Nat. Methods*, **10** (2013), 1169–1176. https://doi.org/10.1038/nmeth.2728
7. W. D. Blizard, Multiset theory, *Notre Dame J. Form. L.*, **30** (1989), 36–66.
8. K. Chakrabarty, R. Biswas, S. Nanda, Fuzzy shadows, *Fuzzy Set. Syst.*, **101** (1999), 413–421. https://doi.org/10.1016/S0165-0114(97)00109-7
9. K. Chakrabarty, R. Biswas, S. Nanda, On Yager's theory of bags and fuzzy bags, *Comput. Informa.*, **18** (2012), 1–17.
10. R. Diestel, *Graph Theory*, New York: Springer, 2005.
11. A. A. El Atik, A. A. Nasef, Some topological structures of fractals and their related graphs, *Filomat*, **34** (2020), 153–165.

12. A. A. El Atik, H. Hassan, Some nano topological structures via ideals and graphs, *J. Egypt. Math. Soc.*, **28** (2020), 41. https://doi.org/10.1186/s42787-020-00093-5

13. A. A. El Atik, A. S. Wahba, Topological approaches of graphs and their applications by neighborhood systems and rough sets, *J. Intell. Fuzzy Syst.*, **39** (2020), 6979–6992.

14. A. A. El Atik, A. S. Wahba, M. Atef, Rough approximation models via graphs based on neighborhood systems, *Granul. Comput.*, **6** (2021), 1025–1035. https://doi.org/10.1007/s41066-020-00245-z

15. A. A. El Atik, A. W. Aboutahoun, A. Elsaid, Correct proof of the main result in "The number of spanning trees of a class of self-similar fractal models" by Ma and Yao, *Inform. Process. Lett.*, **170** (2021), 106117. https://doi.org/10.1016/j.ipl.2021.106117

16. M. K. El-Bably, A. A. El Atik, Soft β-rough sets and their application to determine COVID-19, *Turk. J. Math.*, **45** (2021), 1133–1148. https://doi.org/10.3906/mat-2008-93

17. M. M. El-Sharkasy, M. S. Badr, Modeling DNA and RNA mutation using mset and topology, *Int. J. Biomath.*, **11** (2018), 18500584. https://doi.org/10.1142/S1793524518500584

18. M. M. El-Sharkasy, M. Shokry, Separation axioms under crossover operator and its generalized, *Int. J. Biomath.*, **9** (2016), 16500595. https://doi.org/10.1142/S1793524516500595

19. M. M. El-Sharkasy, W. M. Fouda, M. S. Badr, Multiset topology via DNA and RNA mutation, *Math. Method. Appl. Sci.*, **41** (2018), 5820–5832. https://doi.org/10.1002/mma.4764

20. M. M. El-Sharkasy, Topological model for recombination of DNA and RNA, *Int. J. Biomath.*, **11** (2018), 1850097. https://doi.org/10.1142/S1793524518500973

21. D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, A. Torres, A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets, *J. Theor. Biol.*, **267** (2010), 95–105. https://doi.org/10.1016/j.jtbi.2010.08.010

22. K. P. Girish, S. J. John, Relations and functions in multiset context, *Inform. Sci.*, **179** (2009), 758–768. https://doi.org/10.1016/j.ins.2008.11.002

23. K. Gostelow, Proper termination of flow-of-control in programs involving concurrent processes, *Proc. ACM Annu. Conf.*, **1** (1972), 742–754.

24. S. P. Jena, S. K. Ghosh, B. K. Tripathy, On the theory of bags and lists, *Inform. Sci.*, **132** (2001), 241–254. https://doi.org/10.1016/S0020-0255(01)00066-4

25. H. Jiang, J. Zhan, D. Chen, Covering based variable precision $(\mathcal{I},\mathcal{T})$-fuzzy rough sets with applications to multi-attribute decision-making, *IEEE T. Fuzzy Syst.*, **27** (2018), 1558–1572. https://doi.org/10.1109/TFUZZ.2018.2883023

26. J. L. Kelley, *General Topology*, New York: Courier Dover Publications, 2017.

27. A. Khastan, L. Hooshyar, A computational method to analyze the similarity of biological sequences under uncertainty, *Iran. J. Fuzzy Syst.*, **16** (2019), 33–41.

28. A. M. Kozae, A. El-Atik, A. Elrokh, M. Atef, New types of graphs induced by topological spaces, *J. Intell. Fuzzy Syst.*, **36** (2019), 5125–5134. https://doi.org/10.3233/JIFS-171561

29. D. E. Knuth, Son of seminumerical algorithms, *ACM SIGSAM Bull.*, **9** (1981), 10–11. https://doi.org/10.1145/1088322.1088323

30. A. R. Mashaghi, A. Ramezanpour, V. Karimipour, Investigation of a protein complex network, *T Eur. Phys. J. B*, **41** (2004), 113–121. https://doi.org/10.1140/epjb/e2004-00301-0

31. S. A. Morris, *Topology without Tears*, Biddeford: University of New England, 1989.

32. S. I. Nada, A. A. El-Atik, M. Atef, New types of topological structures via graphs, *Math. Method. Appl. Sci.*, **41** (2018), 5801–5810. https://doi.org/10.1002/mma.4726

33. A. S. Nawar, A. El-Atik, A model of a human heart via graph nano topological spaces, *Int. J. Biomath.*, **12** (2019), 19500062. https://doi.org/10.1142/S1793524519500062

34. J. J. Nieto, A. Torres, D. N. Georgiou, T. E. Karakasidis, Fuzzy polynucleotide spaces and metrics.*Bull. Math. BioL.*, **68** (2006), 703–725. https://doi.org/10.1007/s11538-005-9020-5

35. T. N. Rivera, K. Banas, P. Bialk, K. M. Bloh, E. B. Kmiec, Insertional mutagenesis by CRISPR/Cas9 ribonucleoprotein gene editing in cells targeted for point mutation repair directed by short single-stranded DNA oligonucleotides, *PloS One*, **12** (2017). https://doi.org/10.1371/journal.pone.0169350

36. J. J. Shu, A new integrated symmetrical table for genetic codes, *Biosystems*, **151** (2017), 21–26. https://doi.org/10.1016/j.biosystems.2016.11.004

37. A. Syropoulos, Mathematics of multisets, In: *Workshop on Membrane Computing, WMC 2000. Lecture Notes in Computer Science*, **2235** (2000), 347–358.

38. S. Willard S, *General Topology*, New York: Dover Publications, 2004.

39. R. R. Yager, On the theory of bags, *Int. J. Gen. Syst.*, **13** (1986), 23–37.

40. J. Zhan, B. Sun, J. C. R. Alcantud, Covering based multigranulation $(\mathcal{I},\mathcal{T})$-fuzzy rough set models and applications in multi attribute group decision-making, *Inform. Sci.*, **476** (2029), 290–318.

41. K. Zhang, J. Zhan, W. Wu, J. C. R. Alcantud, Fuzzy β-covering based $(\mathcal{I},\mathcal{T})$-fuzzy rough set models and applications to multi-attribute decision-making, *Comput. Ind. Eng.*, **128** (2019), 605–621.

42. L. Zhang, J. Zhan, Z. Xu, Covering-based generalized IF rough sets with applications to multi-attribute decision-making, *Inform. Sci.*, **478** (2019), 275–302. https://doi.org/10.1016/j.ins.2018.11.033

43. S. Łukaszyk, A new concept of probability metric and its applications in approximation of scattered data sets, *Comput. Mech.*, **33** (2004), 299–304. https://doi.org/10.1007/s00466-003-0532-2

44. https://www.ncbi.nlm.nih.gov .

Appendix: MSC code

```
[caption=Read the data from files and print the length of each DNA sequence,
label ={Read}, language=python]
import pandas as pd

m1 = pd.read_csv('M1.txt', header = None)
M2 = pd.read_csv('M2.txt', header = None)
```

```
print (len(M1.values[0][0]))
print (len(M2.values[0][0]))
output
```

```
[caption=Mh_dna is a function to count A,T,G,C and distance function
to calculate distance, label ={Mh_dna} , language=python] import
math def Mh_dna(x):
count_a=0
count_t=0
count_g=0
count_c=0
for i in x:
if i =='A':
count_a=count_a+1
if i =='T':
count_t=count_t+1
if i =='G':
count_g=count_g+1
if i =='C':
count_c=count_c+1
return count_a,count_t,count_g,count_c

def distance(M1,M2):
#n2=no(A), n3=no(T),n1=no(G),n4=no(C)
#m3=no(A), m2=no(T),m4=no(G),m1=no(C)
n2,n3,n1,n4= Mh_dna(M1)
m3,m2,m4,m1= Mh_dna(M2)
C_aa=n2**2+n3**2+n1**2+n4**2
C_bb=m2**2+m3**2+m1**2+m4**2
C_ab=n1*m1+n2*m2+n3*m3+n4*m4
dist= C_ab/math.sqrt(C_aa* C_bb)
return dist
```

```
def sequence_IDENTICAL(seq_a, seq_b):
len1 = len(seq_a)
len2 = len(seq_b)
mismatches = []
for pos in range (0, min(len1, len2)) :
if seq_a[pos] != seq_b[pos]:
mismatches.append('|')
else:
mismatches.append(' ')
```

```
print (seq_a)
print ("".join(mismatches))
print (seq_b)
```

```
def seq_count_pair(seq_a, seq_b):
len1 = len(seq_a)
len2 = len(seq_b)
columns=['A','T','C','G']
index =['A','T','C','G']
df = pd.DataFrame(0,columns=columns,index=index)
for pos in range (0, min(len1, len2)):
for i,j in enumerate(columns):
if (seq_a[pos] == columns[i] and seq_b[pos] == columns[0]):
k=columns[0]
df[j][k]=df[j][k]+1
elif(seq_a[pos] == columns[i] and seq_b[pos] == columns[1]):
k=columns[1]
df[j][k]=df[j][k]+1
elif(seq_a[pos] == columns[i] and seq_b[pos] == columns[2]):
k=columns[2]
df[j][k]=df[j][k]+1
elif(seq_a[pos] == columns[i] and seq_b[pos] == columns[3]):
k=columns[3]
df[j][k]=df[j][k]+1
return df
```

```
def sequence_complement(seq_a, seq_b):
len1 = len(seq_a)
len2 = len(seq_b)
i=[]
mismatches = []
for pos in range (0, min(len1, len2)) :
if ((seq_a[pos] == 'A' and seq_b[pos] == 'T')
or(seq_a[pos] == 'T' and seq_b[pos] == 'A')
or(seq_a[pos] == 'G' and seq_b[pos] == 'C')
or(seq_a[pos] == 'C' and seq_b[pos] == 'G')):
mismatches.append('')
else:
i.append(pos)
if(len(i)>0):
x=[seq_a[j] for j in i]
y=[seq_b[j] for j in i]
```

```
print(i)
print (x)
print (y)
kg=[]
kg2=[]
for k1 in i:
if k1%3==0:
k2=str(seq_a[k1-2]+seq_a[k1-1]+seq_a[k1])
k3=str(seq_b[k1-2]+seq_b[k1-1]+seq_b[k1])
kg.append(k2)
kg2.append(k3)
elif k1%3==1:
k2=str(seq_a[k1]+seq_a[k1+1]+seq_a[k1+2])
k3=str(seq_b[k1]+seq_b[k1+1]+seq_b[k1+2])
kg.append(k2)
kg2.append(k3)
elif k1%3==2:
k2=str(seq_a[k1-1]+seq_a[k1]+seq_a[k1+1])
k3=str(seq_b[k1-1]+seq_b[k1]+seq_b[k1+1])
kg.append(k2)
kg2.append(k3)
return i,x,y,kg,kg2
```

```
import pandas as pd
M1 = pd.read_csv('M1.txt', header = None)

M2 = pd.read_csv('M2.txt', header = None)

print (len(M1.values[0][0]))

print (len(M2.values[0][0]))

M11=M1.values[0][0] n2,n3,n1,n4= Mh_dna(M11)

print("M1: count A=",n2,"count T=",n3,"count G=",n1,"count C=",n4)

M22=M2.values[0][0] m3,m2,m4,m1= Mh_dna(M22)

print("M2: count A=",m3,"count T=",m2,"count
G=",m4,"count C=",m1)

t=distance(M11,M22) print('distance= ',t)
```

```
df_p=seq_count_pair(M11,M22)
i,x,y,kg,kg2=sequence_complement(M11,M22)

df2 = pd.DataFrame({'Position':i,'M1':x,'M2':y,
   'Amino acid M1':kg,'Amino acid M2':kg2})

columns=['A','T','C','G']

index =['A','T','C','G']

df = pd.DataFrame(0,columns=columns,index=index)
```