Kütahya Dumlupınar University
Institute of Graduate Studies

Journal of Scientific
Reports-A
E-ISSN: 2687-6167
Number 54, September 2023

RESEARCH ARTICLE

DETECTION OF EXON AND INTRON REGIONS IN DNA SEQUENCES BY THE PROPOSED HASHING FUNCTION

Fatma AKALIN[1,*], Nejat YUMUŞAK[2]

[1]Sakarya University, Faculty of Computer and Information Sciences, Department of Information Systems Engineering, Sakarya, Turkey fatmaakalin@sakarya.edu.tr, ORCID: 0000-0001-6670-915X
[2]Sakarya University, Faculty of Computer and Information Sciences, Department of Information Systems Engineering, Sakarya, Turkey, nyumusak@sakarya.edu.tr, ORCID: 0000-0001-5005-8604

Receive Date: 11.04.2023 *Accepted Date: 16.05.2023*

ABSTRACT

Chromosomes, which are formed by the combination of DNA and special proteins, are structures that can show some changes with the effect of genetic or environmental factors. The DNA molecule in these structures carries vital information in elucidating critical information about life. DNA, which is formed by the combination of sugar, phosphate and organic bases, has exon and intron regions separation. Information about the processes in the life cycle of cells, the changes experienced by stem cells, the regulations in the growth and development stage, the development status of cancer, mutation occurrences and protein synthesis are stored in exon regions. Distinguishing exon regions that form 3% of a cell's DNA is challenging. However, detecting diseases on genetically based facts offers more precise outputs. For this reason, analyses were made on the BCR-ABL gene and BRCA-1 mutation carrier genes to analyse leukemia and breast cancer, which are genetically based diseases. First, these genes obtained from the NCBI gene bank were digitized by integer mapping technique. The digitized sequences were given as input to the hash function. This proposed hash function consists of the steps of finding the logarithmic equivalent of the total number of digitized organic bases, summing all logarithmic equivalents, rounding to the nearest integer, expressing it in binary and placing it in the hash table. These outputs, which define the exon and intron regions, were shown as clusters to find the new input region easily. The collision cluster is the binary representation of key values representing both exon and intron regions for the same region. The main goal is to have a small number of elements in this cluster. With the proposed hierarchy in this study, only one collision occurred for BCR-ABL and BRCA-1 genes. Accuracy rates of the proposed approach based on a mathematical basis and independent of nucleotide length were obtained 93.33%, and 96%, respectively.

Keywords: *DNA sequences, Exon and intron regions, Integer mapping technique, Hashing technique*

JSR A
Journal of Scientific Reports

1. INTRODUCTION

The cell is the basic unit of life and has structural and functional properties. It provides the repair of injured tissues and regeneration of dead cells [1]. However, with the effect of genetic or environmental factors, some changes may occur in the genetic components of the cells [1,2]. These changes are expressed with chromosomes. Chromosomes are formed as a result of the fusion of DNA and special proteins [3]. The DNA molecule, which contains vital functionalities, is a structure formed by the combination of sugar, phosphate and organic bases [3,4]. This structure has a separation of exon and intron gene regions for eukaryotic cells [5]. Analysis of these regions is a source of information for elucidating critical information about life. Because, information about the processes in the life cycle of cells, the changes experienced by stem cells, the regulations in the growth and development stage, the development status of cancer, mutation occurrences and protein synthesis are stored in exon regions. In addition, the development status of cancer and mutation formation can also be evaluated by exon regions [3,6].

Cancer is a malignancy that requires early diagnosis for survival. The medical world uses different methods in the diagnosis of cancer. However, some of these methods produce unclear results. For example, manual assessments on imaging or pathological outputs depend on different parameters such as the person's knowledge, experience, mental intensity and physical fatigue [7,8]. However, the ratios in the blood elements can create a similar curve of change for different diseases [9]. Therefore, analyzing the disease using genetic data builds a successful decision process. However, due to the recent developments in genome technology, manual evaluations for the increasing amount of data [4,10] slow the analysis. On the other hand, since 3% of the DNA of a eukaryotic cell consists of exon regions [11], the investigations of these regions are complex. In addition, the detection process of exon and intron regions has been evaluated as a challenging problem in [4,11-13] studies.

In this study, leukemia malignancy, which is among the most common cancer types, was examined [4]. Leukemia is evaluated in two different ways according to the type of disease in the body. This distinction is characterized as acute and chronic. In acute leukemia, the spread of the disease in the body is rapid. In chronic leukemia, the spread of the disease in the body is slow [1,15]. Therefore, early diagnosis is necessary for the continuation of vitality, especially in acute leukemias. In this study, exon and intron regions in BCR-ABL genes presenting as an important indicator in the diagnosis of ALL and CML malignancies being the main types of leukemia, were analyzed [1,16]. Thus, a genetic-based evaluation was provided.

There is an active research area in the literature for the analysis of exon and intron gene regions from past to present. Until the 2000s, different statistical methods were used to evaluate these regions. In this direction, the detection of exons was provided in the [17] study, in which an estimation algorithm using a quadratic discriminant function for multivariate statistical pattern recognition was presented. Genetic evaluations were made with the information obtained in the [18] study, which provided a comprehensive analysis of various statistical features for human exon regions. In the [19] study, in which the program called GeneParser was developed, the fraction of exons was estimated by means of statistical results obtained from intron and exon regions. On the other hand, analytical and computational studies have been carried out for the interpretation of genomic data since the 2000s [2].

At the same time, the developed digital signal processing approach is preferred for data interpretation [2]. In this framework, the conversion process of the nucleotide sequence to the amino acid sequence was analyzed in the [20] study. Exon regions were distinguished from intron regions by means of entropic measures calculated from amino acid sequences. The [21] study proposed a method based on the Gabor wavelet transform. Thus, the detection of exon regions was provided. However, digital signal processing, which is a strong scientific field on large molecules formed by the polymerization of monomers, has been widely used since 2010 [2]. In this direction, digital signal processing was used in the [22] study in order to detect exon and intron regions and reveal anomalies in these regions. FIR and IIR filters were used to provide a successful estimation of exon regions. In order to detect exon regions in eukaryotic cells in the [23] study, a numerical mapping technique based on Walsh codes has been proposed. In the [24] study, the estimation of exon regions from eukaryotic DNA sequences was provided with the developed bidirectional LSTM and RNN-based deep learning models. In the [25] study, is introduced a convolutional neural network model for the classification of human exon and intron regions. In the [26] study, Frequency Chaos Game Representation and CNN structure were used together. Thus, human exon and intron regions were analyzed.

A study area is available to evaluate exon and intron regions. Especially the developments in the field of digital signal processing [3,4] have emerged as a preferred strong field for exploring the relationships, patterns and periodicity states in the data [10]. However, the [27] and [28] studies using the digital signal processing approach stated that digital signal processing approaches not only serve a successful decision process but also provide an inference based on DNA length. Therefore, it was planned to construct a structure independent of DNA length. In this study, a hierarchy based on the structural and mathematical basis was constructed in order to produce results independent of DNA length. First of all, DNA sequences, which have a symbolic structure in this hierarchy, were digitized according to the rules of the integer mapping technique. Secondly, the digitized sequences were given as input to the hash function. The logarithmic equivalents of the total number of each organic base in the sequences digitized within the scope of the hash function were obtained and rounded to the nearest integer. All integer values generated for exon and intron regions were expressed in binary system. In this study, in which open hashing is preferred, the values that define the exon and intron regions and are expressed in a binary system were placed in the hash table. Finally, it was compared with the results of analysis realized on the BRCA-1 gene [29] seen in breast cancer patients carried high-risk. Accuracy rates of this study for BCR-ABL gene and BRCA-1 gene were obtained as 93.33%, and 96%, respectively.

With this study, the analysis of exon regions that have vital functions related to life has been done. An alternative method for detecting these regions has been proposed compared to statistical methods, digital signal processing approaches, deep learning models, and analytical or computational studies. In this direction, the exon regions, constituting 3% of the eukaryotic cell DNA and having a complex analysis process, were evaluated with the proposed hashing function [4,11-13]. Thus, despite the increasing genome data and unclear manual evaluation methods, genetic-based inferences have been produced that are independent of DNA length, and have a simple hierarchy and a mathematical basis.

JSR A
Journal of Scientific Reports

2. MATERIALS AND METHODS

This study aimed to distinguish exon regions containing vital life information from intron regions using the NCBI dataset supplied https://www.ncbi.nlm.nih.gov/ site. In this direction, the BCR-ABL gene, characterized as an important indicator in the diagnosis of ALL and CML malignancies being the main types of leukemia was used. First, the BCR-ABL gene, which occurs from symbolic letters, was digitized by the integer mapping technique. Then the digitized DNA sequence was given to the hashing function. The values expressed in the binary system as function output were placed in the hash table. In this hierarchy where the open hashing approach is preferred, the outputs corresponding to the same hash value are kept in a list structure. Then, these key values, which define the exon and intron regions, were shown with clusters in order to easily find the region of the new input. At the same time, producing cases of the same output is examined for both exon and intron regions. The intensity of this situation, which is described as collision, was evaluated for the entire clusters. However, in order to generalize with the proposed hierarchy, the same methods were also performed for the BRCA-1 mutation carrier gene [29] seen in high-risk breast cancer patients. Finally, the analyzes performed for two separate DNA sequences were evaluated and comparisons were made. The flow chart of the proposed hierarchy was given in Figure 1.

JSR A
Journal of Scientific Reports

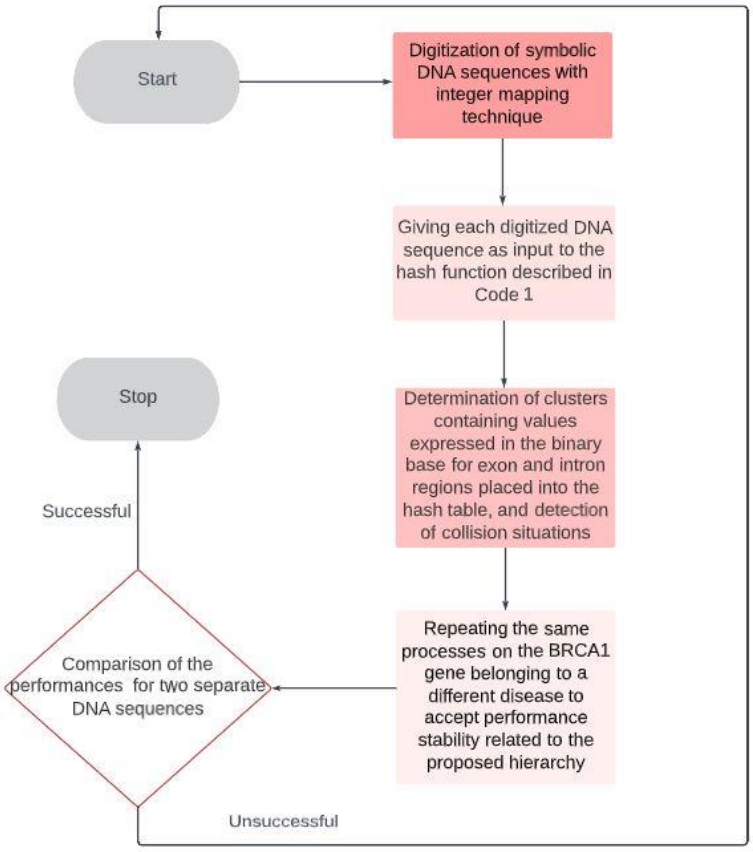

Figure 1. The flow diagram of the hierarchy proposed in this study.

2.1. Dataset

Leukemia, one of the most common types of cancer [14], affects the body's blood production mechanism [1]. Manual evaluations required for the diagnosis of this disease include different parameters. This complicates to produce a clear output. For this reason, the main aim of the study is to produce clearer results in diagnosing the disease using genetically based facts. In this direction, the molecular diagnosis was provided on 4 different BCR-ABL genes used as an indicator in the diagnosis of ALL and CML malignancies, which are the main types of leukemia disease. Nucleotide length information about the exon and intron regions of these genes is given in Table 1.

JSR A
Journal of Scientific Reports

Table 1. Information about the BCR-ABL gene regions obtained from the NCBI dataset.

AM400881.1	**Regions**	**Names Called of Regions**
1-14	Exon Region	(1-Exon)
15-71	Intron Region	(2-Intron)
72-527	Intron Region	(3-Intron)
AM600680.1		
1-29	Exon Region	(4-Exon)
30-578	Intron Region	(5-Intron)
579-1114	Intron Region	(6-Intron)
1115-1180	Exon Region	(7-Exon)
AM886138.1		
1-31	Exon Region	(8-Exon)
32-280	Intron Region	(9-Intron)
281-790	Intron Region	(10-Intron)
791-853	Exon Region	(11-Exon)
EU447303.1		
1-70	Exon Region	(12-Exon)
71-145	Exon Region	(13-Exon)
274-448	Exon Region	(14-Exon)
449-488	Exon Region	(15-Exon)

In this study, the stability of the performance of the proposed framework was also investigated for a different gene type. In this direction, the same hierarchy was applied again for the BRCA-1 mutation carrier gene in breast cancer patients with high risk. The nucleotide length information of the exon and intron regions for the BRCA-1 gene is given in Table 2.

Table 2. Information about the BRCA-1 gene regions obtained from the NCBI dataset.

LC312442.1	**Regions**	**Names Called of Regions**
20-207	Intron Region	(1-Intron)
Y08757.1		
1-746	Intron Region	(2-Intron)
747-834	Exon Region	(3-Exon)
835-1482	Intron Region	(4-Intron)
LC312441.1		
1-203	Intron Region	(5-Intron)
NM_001407963.1		

JSR A
Journal of Scientific Reports

1-94	Exon Region	(6-Exon)
95-183	Exon Region	(7-Exon)
184-323	Exon Region	(8-Exon)
324-429	Exon Region	(9-Exon)
430-475	Exon Region	(10-Exon)
476-552	Exon Region	(11-Exon)
553-3978	Exon Region	(12-Exon)
3979-4067	Exon Region	(13-Exon)
4068-4236	Exon Region	(14-Exon)
4237-4360	Exon Region	(15-Exon)
4361-4551	Exon Region	(16-Exon)
4552-4862	Exon Region	(17-Exon)
4863-4950	Exon Region	(18-Exon)
4951-5028	Exon Region	(19-Exon)
5029-5069	Exon Region	(20-Exon)
5070-5153	Exon Region	(21-Exon)
5154-5208	Exon Region	(22-Exon)
5209-5282	Exon Region	(23-Exon)
5283-5343	Exon Region	(24-Exon)
5344-6851	Exon Region	(25-Exon)

9 exons and 6 intron regions were selected for the BCR-ABL gene. In addition, 4 intron regions and 21 exon regions were created as unbalanced in the preferred BRCA-1 gene. In the last step, the performance of the inferences reached for the BRCA-1 gene was compared with the performance of the inferences reached for the BCR-ABL gene. The title "Names Called to Regions" in Table 1 and Table 2 has definitions related to each example. These definitions are used to illustrate the collision situations in Figure 2, Figure 3, and Figure 4.

2.2. Integer Mapping Technique

DNA is the part of the cell that has vital information for the maintenance of life functions and biological processes [4]. It has a symbolic structure. This complicates the analysis of DNA structure. Digitizing the sequences is a necessary step to provide successful inference. In this study, the integer mapping technique, which is one of the fixed mapping techniques, was chosen [6].

The integer mapping technique is a 1-dimensional mapping technique [28]. In order to digitize the DNA structure with this technique, firstly, the number of organic bases in the sequences is examined and then the assignment is done. The first rule for this mapping technique is that the total number of purine bases (A and G) is greater than the total number of pyrimidine bases (C and T). According to this rule, the 4 bases are assigned as T=0, C=1, A=2, and G=3 respectively. Another rule is that the total number of T organic base is greater than the total number of A organic base and the total number of G organic base is greater than the total number of C organic base. According to this rule, the 4 bases are assigned as A=0, C=1, T=2 and G=3 respectively [6]. In this study, it was also encountered

that the total number of T organic bases is larger than the total number of A organic bases and the total number of G organic bases is smaller than the total number of C organic bases. In such a case, A=0, T=2, C=3 and G=1 assignments were made for 4 bases, respectively. Thus, BCR-ABL genes and BRCA1 genes were digitized by integer mapping technique.

2.3. Hashing Technique

The process of creating a fixed-size output from inputs of different lengths with the help of the address function is called hashing. The outputs produced as a result of the hashing process are placed into a hash table. The data in the hash table, which consists of fixed-size outputs, is accessed with the key. Keys that produce an index are always unique and represent only one value. Each index defined with a key in the hash table is determined by the hash function. The hash function takes the data and places it in memory with the output it produces. This function should be simple to calculate, and produce results without any collision [30].

In this study, a hash function suitable for the structure of the DNA sequence was created. The pseudo code of this function is given below.

Code 1. The pseudo code of the proposed hash function

1-Start
2-Find the total number of organic bases A, T, G, and C for each digitized DNA sequence.
3-Find the logarithmic equivalents of the total number of each organic base in the sequences digitized.
4-Add these 4 separate logarithmic equivalents and round to the nearest integer.
5-Express in the binary system these values obtained in the fourth step.
6-Finish

After the key of each DNA sequence was calculated with the help of the hash function described in Code 1, it was placed in the hash table. By this hash function created suitable to DNA structure, values expressed in the binary system specific to exon and intron regions are obtained. Then the values reached for the exon and intron regions were shown in two separate clusters. The main purpose here is that succeed to avoid a collision situation. In this study, open hashing was preferred within the framework of hashing approach.

Open hashing is a method that provides a list structure solution for the case of hashing the elements calculated with the hash function to the same value. With this method, in the case being of different data corresponding to the same index, elements with the same properties are added to a list [30]. The hash tables of BCR-ABL genes expressed with 4 bits and BRCA-1 genes expressed with 5 bits are given in Figure 2 and Figure 3.

JSR A
Journal of Scientific Reports

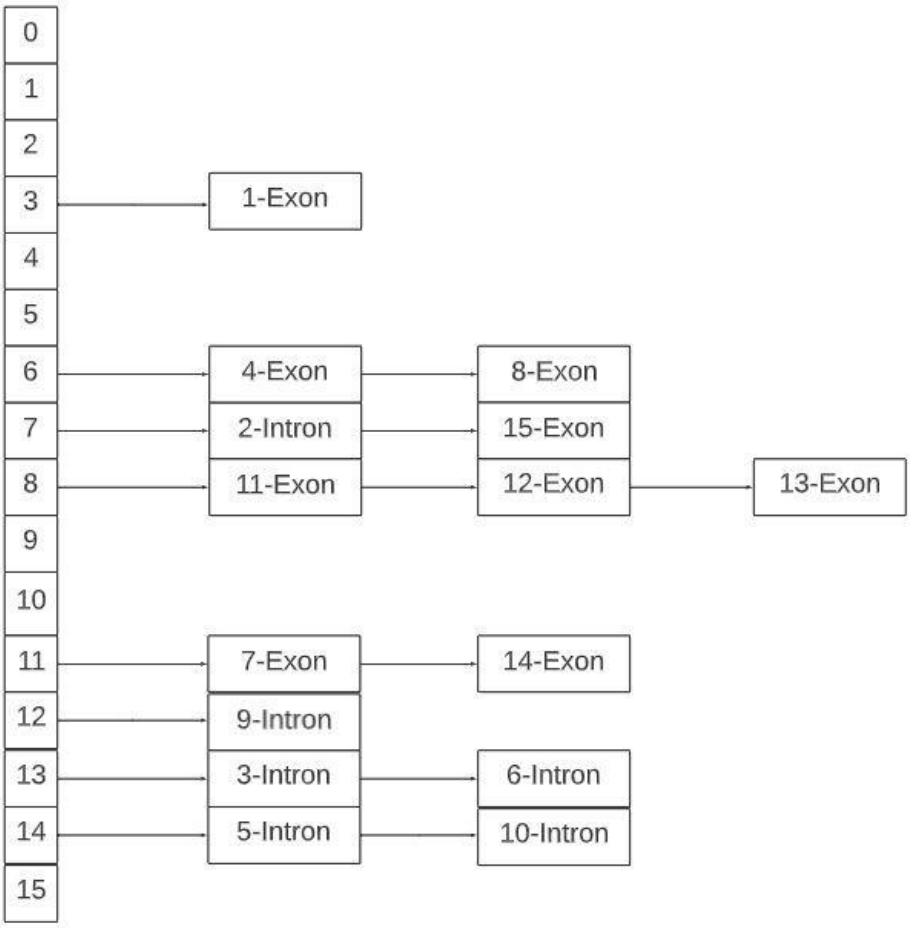

Figure 2. Hash table created fort he BCR-ABL gene.

JSR A
Journal of Scientific Reports

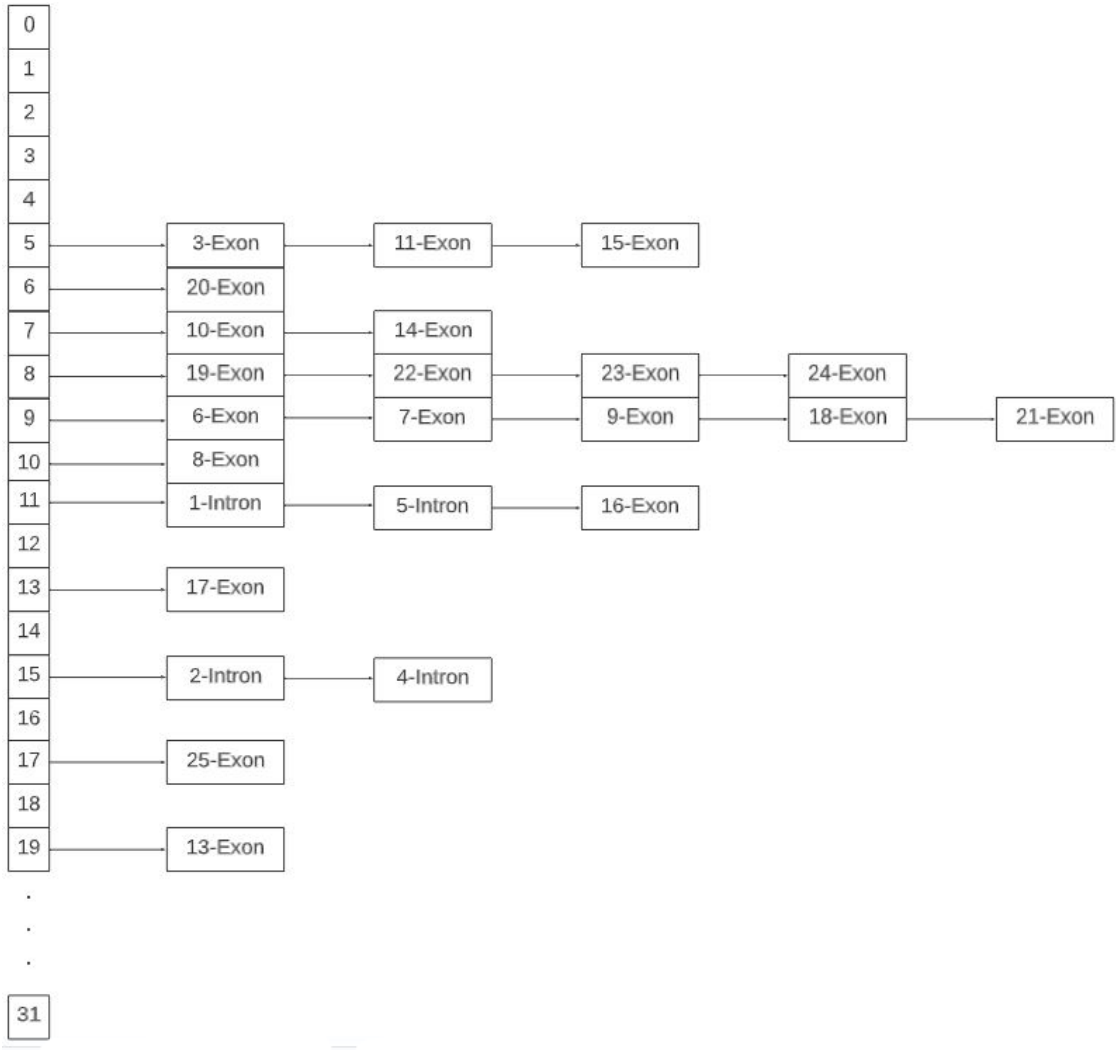

Figure 3. Hash table created for the BRCA-1 gene.

When the hash tables in Figure 2 and Figure 3 are examined, it is seen that there is only one collision for two different gene structures. In this direction, 14 correct detections were made for BCR-ABL genes containing 15 distinct gene regions. At the same time, 24 correct detections were made for BRCA1 genes containing 25 distinct gene regions. As a result of these detections, 93.33% and 96% success rates were produced, respectively. The clusters of numerical values in the binary system that define the exon and intron regions are given in Figure 4.

JSR A
Journal of Scientific Reports

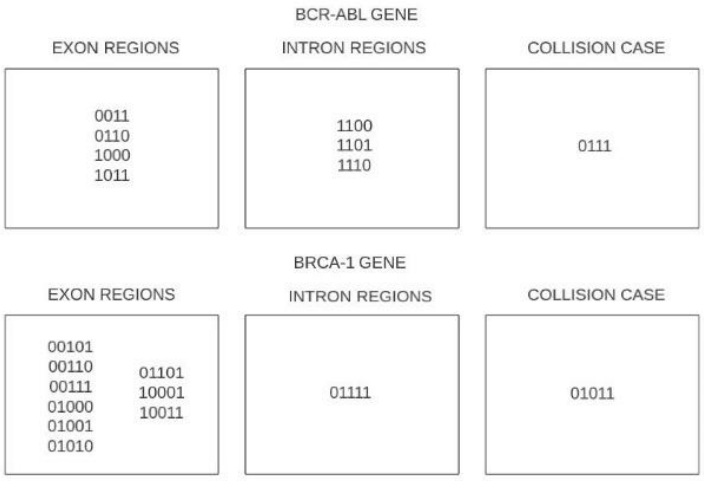

Figure 4. Clusters of numerical representations in the binary system that define exon and intron regions for BCR-ABL and BRCA-1 genes.

In this study, a successful result was produced at the end of the proposed region recognition process for the BCR-ABL gene. In order to examine the stability of this success in a different disease context, the BRCA-1 gene, which consists of an unbalanced number of exons and intron regions with more gene regions, was used and the success of the applied method was tested. At the end of the test process, 1 collision occurred. The proposed hierarchy for the BCR-ABL method was also successful on the dataset with different parameters. This shows the stability and generalizability of the proposed method.

3. RESULTS AND DISCUSSION

In this study, a hierarchy was proposed for the detection of exon and intron gene regions with different nucleotide lengths on the BCR-ABL gene that is an indicator in the detection of ALL and KML malignancies. Thus, the molecular diagnosis was provided on the BCR-ABL gene, which contains 9 exon regions and 6 intron regions. In this diagnostic process, firstly, DNA sequences with a symbolic structure were digitized by integer mapping technique. Then the digitized DNA sequences were given to the hash function. The outputs produced by the Hash function, which consists of 4 basic steps, are the values that define the exon and intron regions. It is expected that there will be any collisions in these values placed in the hash table.

In this study, the open hashing approach was used in the scope of hashing. By this approach, numerical representations defining the same region were placed in a list structure. In the hash table kept as a list structure, there was only 1 collision for the BCR-ABL gene and a success rate of 93.3% was achieved. Then, it was planned to test the performance stability so that the proposed hierarchy can be generalized. Therefore, an analysis was performed on the BRCA-1 mutation carrier gene seen in

high-risk breast cancer patients. In order to clarify the success and stability of the method, the number of selected gene regions was increased and an unbalanced dataset was created. However, the proposed hierarchy produced only 1 collision on a different gene and achieved a success rate of 96%. A successful detection process has been achieved with the stable outputs of this structure, which produces only one wrong prediction for both gene molecules with different parameters.

In addition, this work has two originalities. Its first originality is the generation of a hash function suitable for the structure of DNA. For example, the number of times a paper must be folded to create 4 separate squares is solved by log2(4). Therefore, the organic base density in DNA sequences containing 4 different organic bases should be expressed in logarithm base 4. Thus, a mathematical basis suitable for the structure of DNA was created.

The second feature is to perform a study independent of nucleotide length. In this direction, it has been stated that inferences based on nucleotide length are produced in the [27] study, which uses the signal processing approach that has been prominent, especially since the 2000s. On the other hand, a study independent of nucleotide length was done in the [28] study, in which classification was provided with the structural and statistical features extracted from the sequences. The obtained results were more successful than the outputs produced using the signal-processing approach. On the other hand, a new mapping technique was proposed for digitizing DNA sequences in the [31] study. In this mapping technique, an independent study of nucleotide length was performed depending on the codon distributions. At the same time, it has been stated that more successful results were produced compared to other mapping techniques.

There are studies [3,5,11,13,22,23,32-34] in which statistical analyzes are made, inferences are produced within the scope of signal processing approach or artificial intelligence-based detections are performed. However, in this study, a new perspective was created using the field of data structures in the analysis of DNA molecules.

4. CONCLUSION

Cancer is a malignancy that occurs with the uncontrolled proliferation and spread of cells in a certain tissue or organ. The methods applied to the patient during the diagnosis of this malignancy produce unclear outcomes in some cases. At the same time, some important indicators used in the diagnosis of the disease form a similar curve of change within the scope of different diseases. For this reason, it is advantageous to diagnose the disease with inferences made on genetically based cases.

In this study, ALL and CML malignancies, which are the main types of leukemia, were analyzed with genetically based cases. In this direction, the BCR-ABL gene, which is an important indicator in the diagnosis of ALL and CML malignancies, was analyzed. In this analysis process, firstly, the BCR-ABL gene with symbolic structure was digitized by integer mapping technique. Then the digitized DNA sequences are given as input to the hash function. The proposed hash function is based on a mathematical basis created in suitable for the DNA structure. The outputs for this function are values that define the exon and intron regions. These values can be expressed with clusters in order to easily find the region of the new input. In this study, in which open hashing was used, 1 collision occurred

for the BCR-ABL gene. In addition, the current performance of the proposed hierarchy for the BCR-ABL gene for the gene structure affecting a different disease was also tested. For this, the BRCA-1 gene, which contains more gene regions and is created in an unbalanced way, was used. The proposed hierarchy created 1 collision in the BRCA-1 gene structure. This shows the stability of the proposed hierarchy.

In this study, a structure depended on a mathematical basis and independent of nucleotide length was created. In future studies, different bases, different mathematical calculations or different bit numbers can be used within the preferred data structure hierarchy for the detection of exon and intron regions. Thus, it is planned to reduce the possibility of a collision.

ACKNOWLEDGEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Kocabıyık, V.B. (2011). ALL ve KML'li hastalarda BCR ve ABL genlerindeki mutasyonların incelenmesi. Yüksek Lisans Tezi, Selçuk Üniversitesi Sağlık Bilimleri Enstitüsü, Konya.

[2] Khodaei, A., Feizi-Derakhshi, M.R., and Mozaffari-Tazehkand, B. (2020). A pattern recognition model to distinguish cancerous DNA sequences via signal processing methods. Soft Computing, 24(21), 16315–16334.

[3] Das B., and Türkoglu, I. (2016). Classification of DNA sequences using numerical mapping techniques and Fourier transformation. Journal of the Faculty of Engineering and Architecture of Gazi University, 31(4), 921–932, 2016.

[4] Barman, S., Saha, S., Mandal, A., and Roy M. (2012). Prediction of protein coding regions of a DNA sequence through spectral analysis. 2012 International Conference on Informatics, Electronics and Vision, ICIEV 2012.

[5] Hota, M. K., and Srivastava, V. K. (2010). Performance analysis of different DNA to numerical mapping techniques for identification of protein coding regions using tapered window based short-time discrete Fourier transform. ICPCES 2010 - International Conference on Power, Control and Embedded Systems 2010, 0–3.

[6] Daş, B. (2018). DNA dizilimlerinden hastalık tanılanması için işaret işleme temelli yeni yaklaşımların geliştirilmesi. Doktora Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ, 83s.

[7] Al-jaboriy, S.S., Sjarif, N.N.A., Chuprat, S., and Abduallah, W.M. (2019). Acute lymphoblastic leukemia segmentation using local pixel information. Pattern Recognition Letters, 125, 85–90.

[8] Scotti F. (2005). Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. CIMSA 2005-IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, 20–22.

[9] Kutlu, H., Avci, E., and Özyurt, F. (2020). White blood cells detection and classification based on regional convolutional neural networks. Medical Hypotheses, 135.

[10] Chakraborty, S., and Gupta, V. (2016). DWT based cancer identification using EIIP. Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, CICT 2016, 718–723.

[11] Das, L., Das J.K., and Nanda, S. (2020). Detection of exon location in eukaryotic DNA using a fuzzy adaptive Gabor wavelet transform. Genomics, 112, 4406–4416.

[12] Das, L., Nanda, S., and Das, J.K. (2019). An integrated approach for identification of exon locations using recursive gauss newton tuned adaptive kaiser window. Genomics, 111, 284–296.

[13] Gupta, R., Mittal, A., Singh, K., Bajpai, P., and Prakash, S. (2007). A time series approach for identification of exons and introns. 10th International Conference on Information Technology (ICIT 2007), 91–93.

[14] Hsu, C.H., Chen, X., Lin, W., Jiang, C., Zhang, Y., Hao, Z., and Chung, Y.C. (2021). Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning. Measurement, 175.

[15] Aydın, G. (2017). Quercetinin KML kök hücreleri üzerine sitotoksik etkilerinin moleküler düzeyde incelenmesi. Erciyes Üniversitesi, Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, Kayseri.

[16] Arslan, S. (2014). KML ve ALL Tanılı Hastalarda BCR/ABL füzyon geni mutasyonlarının taranması. Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 76s.

[17] Audic S., and Claverie, J. M. (1998). Self-identification of protein-coding regions in microbial genomes. Proceedings of the National Academy of Sciences of the United States of America, 95(17), 10026–10031.

[18] Zhang, M.Q. (1998). Statistical features of human exons and their flanking regions. Human Molecular Genetics, 7(5), 919–932, 1998.

[19] Snyder, E.E., and Stormo, G.D. (1995). Identification of protein coding regions in genomic DNA. Journal of Molecular Biology, 248(1), 1–18.

[20] Mereuta, S., and Munteanu, V. (2007). A new information theoretic approach to exon - intron classification. ISSCS 2007 - International Symposium on Signals, Circuits and Systems, Proceedings 2007, 2, 497–500.

[21] Mena-Chalco, J., Carrer, H., Zana, Y., and Cesar, R. M. (2008). Identification of protein coding regions using the modified gabor-wavelet transform. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5(2), 198–206.

[22] Kar, S., and Ganguly, M. (2022). Study of effectiveness of FIR and IIR filters in exon identification: a comparative approach. Materials Today: Proceedings, 58, 437–444.

[23] M, R. K., and Vaegae, N. K. (2020). Walsh code based numerical mapping method for the identification of protein coding regions in eukaryotes. Biomedical Signal Processing and Control, 58.

[24] Singh, N., Nath, R., and Singh, D.B. (2022). Splice-site identification for exon prediction using bidirectional LSTM-RNN approach. Biochemistry and Biophysics Reports, 30.

[25] Ben Nasr, F., and Oueslati, A.E. (2021). CNN for human exons and introns classification. 18th International Multi-Conference on Systems. Signals & Devices SSD'21 2021, 249–254.

[26] Ben Nasrand, F., Oueslati, A.E. (2022). A new automatic method for human coding and non-coding zones characterization and classification based on FCGR coding and CNN classifier. International Conference on Advanced Technologies for Signal and Image Processing, ATSIP, 8–9.

[27] Akalın, F., and Yumuşak, N. (2022). Classification of exon and intron regions obtained using digital signal processing techniques on the DNA genome sequencing with EfficientNetB7 architecture. Journal of the Faculty of Engineering and Architecture of Gazi University, 37(3), 1355–1371.

[28] Akalın, F., and Yumuşak, N. (2023). Classification of ALL and CML malignancies being among the main types of leukaemia with graph neural networks and fuzzy logic algorithm. Journal of the Faculty of Engineering and Architecture of Gazi University, 38(2), 707–719, 2023.

[29] Yetim, E. (2018). Meme manyetik rezonans görüntülemede BI-RADS kategori 3 lezyonlar; takip sonuçları. Akdeniz Üniversitesi Tıp Fakültesi Radyoloji Anabilim Dalı, Uzmanlık Tezi.

[30] Yumuşak, N., and Adak. M.F. (2016). C/C++ ile veri yapıları.

[31] Das, B., and Turkoglu, I. (2018). A novel numerical mapping method based on entropy for digitizing DNA sequences. Neural Computing and Applications, 29(8), 207–215.

[32] Marhon, S. A., and Kremer, S. C. (2011). Protein coding region prediction based on the adaptive representation method. Canadian Conference on Electrical and Computer Engineering, 000415–000418.

[33] Li, J., Zhang, L., Li, H., Ping, Y., Xu, Q., Wang, R., Tan, R., Zhen, W., Liu, B., and Wang, Y. (2019). Integrated entropy-based approach for analyzing exons and introns in DNA sequences. BMC Bioinformatics, 20.

[34] Hota, M. K., and Srivastava, V. K. (2012). Identification of protein coding regions using antinotch filters. Digital Signal Processing: A Review Journal, 22(6), 869–877.