Journal of Basrah Researches (Sciences) **49**(1), 13 (2023)
DOI: https://doi.org/10.56714/bjrs.49.1.2

Using Genetic Algorithm for DNA Profile Matching

Nawal S. Jabir[1], Zainab A. Kahlaf [2*]

[1] Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah, Iraq.
[2] Department of Computer Science, College of Sciences, University of Basrah, Basrah, Iraq.

ARTICLE INFO

Received 16 November 2022
Accepted 08 January 2023
Published 30 June 2023

Keywords:

Genetic Algorithm, DNA profiling, Bioinformatics, DNA forensic.

Citation: N.S. Jabir, Z. A. Kahlaf, J. Basrah Res. (Sci.) **49**(1), 13(2023). DOI:https://doi.org/10.56714/bjrs.49.1.2

ABSTRACT

The DNA is used in forensic investigations to identify suspects and victims at crime scenes. However, manual matching of DNA profiles is difficult and error-prone, especially in large databases. In Iraq, technology for DNA matching is limited, making manual matching the only option. Regenerate. In this work, we propose a Genetic Algorithm (GA) for DNA dataset matching to provide simple and user-friendly software to be used by law enforcement agencies in Iraq. The genetic algorithm is a type of heuristic search method used in computing science and artificial intelligence. It is based on the theory of natural selection and evolutionary biology and is used to find the best solutions to search problems. Genetic algorithm is robust for searching through big, complicated datasets. Thus, in this paper, the GA is the algorithm of choice to achieve the goal of DNA matching search. The used dataset is actual data that have been collected from the Ministry of Interior at the Basra Investigation Center. Finally, the python simulation results show 100% accuracy where the proposed method managed to find the DNAs under consideration precisely.

1. Introduction

We Despite the fact that biologists have amassed a massive size of DNA sequence data, the specifics of how these sequences data are mostly still understood [1]. Biologists can now convert huge amounts of biological data into usable data, thanks to more improved techniques developed in recent years. Genomics functional is the creation and implementation of global experimental techniques to estimate gene function using structural genomics' knowledge and reagents. High-throughput or large-scale experimental approaches distinguish it, as well as statistical and computer analysis of the outcomes [2][3].It's used to track the expression of a large number of genes at the same time. Gene expression refers to the process of translating a gene's DNA sequence into RNA, which serves as a template for the production of proteins, and gene expression level refers to how active a gene is in a particular tissue at a certain time or in a particular experimental circumstance [4]. This technology entails a number of steps. The complementary DNA (cDNA) molecules, also known as oligos, are initially printed on slides as dots. Following that, sample and control dye-labeled samples are hybridized. The term comparable here refers to the fact that any difference in a gene's measured expression value between two trials should reflect the genuine expression levels of that gene [5].

The majority of gene expression assessments are carried out manually using scant experimental data. There is presently a great need for automatic analysis of the overall connection underlying

***Corresponding author email :** zainab.khalaf@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq

multiple genes from their expression. An optimization algorithm is the study of algorithms that can forecast using past data and learn from it [6]. The optimization algorithm's theoretical parts are founded in informatics and statistics, but computational concerns are also necessary [7].Optimization algorithms could play a crucial part in the analysis process due to the complicated nature of the biological data [8]. One of the challenges in the forensic field is matching the extracted DNA and match it with the DNA database. Even though there are advanced devised to do this task yet, these devices are expensive and hard to maintain. In Iraq, access to thesis devices is limited. Consequently, most law enforcement agencies use manual fashion for DNA matching. In this paper, a Genetic Algorithm (GA) is used as data mining to search efficiently in the DNA database for fast matching for the DNA under consideration. The algorithm has the ability to find the degree of relationship between the introduced DNA and the DNAs in the database (e.g. the person himself, brother, sister, nephew, etc.).The contributions of the current study are to identify typical gene expression analysis problems by using an optimization technique that is used to find the best common gene expression for known (e.g. criminal or suspect) or unknown person (e.g. Victims of war or terrorism) through the match the gene expression within the database. The designed system can be used to assist investigators in revealing the identity of accused or deceased persons and other cases. The main contribution of this work is providing Iraqi law enforcement agencies with simple, free and user-friendly software for forensic investigations and DNA matching for the purposes mentioned above.

The main idea is to focus on applying data mining techniques and optimization techniques in order to match the gene expression in order to reduce the number of falsely matching. The experiment results show 100% accuracy and fast convergence for the GA in finding the DNA match [9]. The rest of the paper is organized as follows: in section 2, related works have been reviewed and discussed. Next, section 3 sheds some light on the Gene expression in cells with some necessary details. Section 4 reviewed the details of the proposed GA and its concept. The proposed model and the data collection subsection are introduced in Section 5. Section 6, illustrated the evaluation method and finally, the results section and the conclusion are presented in sections 6 and 7, respectively [2].

2. Problem Statement

DNA matching is an important part of many parentals, criminal, dead bodies, and individual identification cases. Currently, in Iraq, law enforcement agencies have limited access to advanced equipment that can analyze and match DNA for everyday cases. Manual methods are used, and this method is error prone no matter how long it takes to get results. These challenges have put much stress on these agencies and people alike. The speed of determining the genetic match in criminal cases has an effective and decisive role in the speed of decision-making and access to the perpetrators, as well as the accuracy of the match helps the investigators in identifying the perpetrator and deporting the suspects. Thus, a simple, cheap, and rapid method is needed to mitigate or solve this problem.

3. Related Work

The employment of intelligent computing models has resulted in numerous advancements in the field of DNA categorization. Some of them are briefly described below: In [1] authors presented a strategy for dealing with HDLSS DNA methylation datasets based on the usage of AEs and survival analysis. This approach was utilized to obtain useful information about breast cancer recurrence based on key genes in particular. The discovery of multiple enriched words and related linkages between the genes using functional annotation enrichment analysis was another outcome that validated the methodology. A BIGBIOCL algorithm to derive alternative and equivalent classification models by iteratively deleting chosen characteristics is proposed in [2]. The authors perform classifications that extract numerous methylation sites and their associated genes with high accuracy (>97%). Moreover, in [3], a new intraclass helitron family classification system in the Celegans genome using Frequency Chaos Game Representation (FCGR pictures) as characteristics and optimization algorithm approaches. The pre-trained deep neural network (PTDNN) classifier uses the acquired FCGR images as input, which correlate to helitron sequences as features. These frequency matrices were used as normalized statistical features (NKmers) to categorize helitron families using two classifiers: the Support Vector Machine (SVM) and the Random Forest (RF).For quantifying DNA damage using comet assay images, researchers in [4] compared Convolutional Neural Network (CNN) to other approaches in the literature.

An expert performed the techniques, who tagged the 796 single comet grayscale pictures into four classes with around 200 samples each: G0 (healthy), G1 (poorly defective), G2 (defective), and G3 (healthy) (extremely defective). a method that uses a number of molecules to in vitro learn generic classes of biological operations on DNA instances is presented in [5]. The system trains a binary classifier in this vector space by computing the inner product between embedded vectors in a similar vector space, which is understood as hybridization between DNA molecules. A new deep learning-based strategy for discriminating between N6-methyladenosine sites is presented in [6].The input sequence is encoded using one-hot representation in this approach, which allows for subsequent convolution layers. Then they combine CNN features with the tri-nucleotide composition (TNC) feature extraction approach. Machine learning techniques have an important contribution in the field of DNA profile matching as shown in [7]. The purpose work in [16], was to determine the identities of individuals based on the analysis of DNA profiles, which may be individual or a mixture of multiple profiles. DNA profiling is used to find the number of contributors in a DNA mixture, a major task with challenges caused by allele dropout, stutter, blobs, and noise. The performance of six machine learning algorithms, including Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Stochastic Gradient Descent (SGD), and Gaussian Naïve-Bayes (GNB). The investigation was done using a publicly dataset called PROVED. It is containing DNA mixtures with up to five contributors. Machine learning had received recently attention in this field, but with limited success. The research aimed to advance the state-of-the-art in this field by evaluating the algorithmic performance using confusion matrices and four performance metrics: accuracy, F1-Score, Recall, and Precision. The results showed that Logistic Regression (LR) provided the highest Accuracy of 95% for mixtures with five contributors.DNA matching technique is also used for age prediction as illustrated in [8], in this work, age prediction models for both healthy and diseased samples are proposed using DNA methylation data and machine learning techniques. Four machine learning techniques are used, and the model designed using Random Forest Regression shows the best performance (96 %) The model has a MAD of 2.51 years for training data and 4.85 for testing data in the case of healthy samples, and a MAD of 3.83 years for training and 9.53 years for testing in the case of diseased samples.

Finally, in [9], researchers present a method for automatically learning high-performance deep networks termed DNA computing-inspired networks design (DNAND). They also describe the killing approach, in which "poor" models are stopped from being trained if they fail to meet a certain accuracy level on the validation set, reducing computational costs and speeding up the learning process.

In summary, none of the aforementioned works clearly discuss the challenge of DNA matching, especially in a large database, unlike our work, where the main goal is to find DNA matching by searching a large database accurately with minimum time. Moreover, the accuracy of our work is 100% whereas the previous works had never achieved such accuracy.

4. Gene Expression in Cells

Any known living organism's complete hereditary information is stored in the cell as a deoxyribonucleic acid (DNA) molecule. Each of the smaller units that make up the DNA molecule, known as nucleotides, contains one of four distinct biological components and can represent by a four-symbol alphabet. However, the used database in this paper has represented as numbers.The DNA molecule is usually made up of two of these sequences (strands) and structured as a double helix. The sequences are mutually exclusive; a symbol on one strand dictates the symbol on the other. This arrangement enables successful information replication during gene multiplication or copying into ribonucleic acid (RNA), which differs from DNA by one nucleobase (using uracil instead of thymine) and (ii) normally remains single-stranded [10].

DNA polymerase is responsible for replication, which happens during the transmission of hereditary information to the progeny of any cell or creature. The RNA polymerase facilitates transcription, which is the process by which information encoded in a DNA sequence, called a gene, is transferred to an mRNA sequence [11].

5. Genetic Algorithms

The initial population, selection, crossover, and mutation are all parts of the GA process. To retain population diversity, avoid premature convergence, and speed convergence, the traditional genetic algorithm is paired with a niche strategy and an elitist approach [21]. The initial populace is produced at random. As taken into length, the number of stages affects how long an encoding chromosome is. One of the nodes in the appropriate stage is randomly selected to represent each gene on this chromosome. A genetic algorithm can choose people to be replicated into the next generation using a variety of different selection procedures, Fitness-proportionate selection, Roulette-wheel selection, fitness Scaling selection, tournament selection. An operator called crossover allows two parents to produce new offspring with traits from both parents. Many crossover operators nowadays rely on the application. The crossover and mutation processes will be applied to the chromosomes chosen for the following generation. Due to its capacity to get closer to the optimal solution, it is clear that the modified partially matched crossover with a random two-point mutation worked the best [12]. Fig.1 shows an example of a crossover.

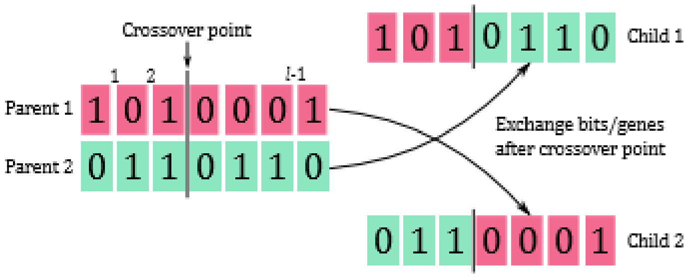

Fig.1. Examples of crossover [12].

The process of mutation involves changing the sites of certain alleles on random chromosomes at different places and probabilities, for as from 0 to 1 or 1 to 0. (Fig.2 and Fig.3). The small mutation is a chance event that might benefit or harm the chromosome, which would then either decline or thrive during the following selection. Finding the best option is the objective, thus a single terrible choice will momentarily harm the population. Additionally, coming up with a workable solution will be quite beneficial [13].

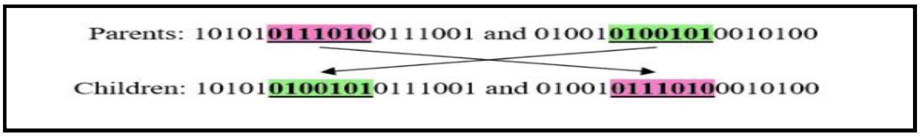

Fig.2. Two-point crossover [13].

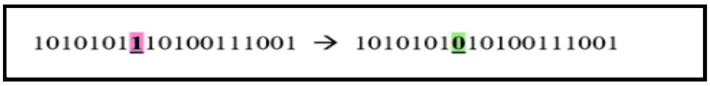

Fig.3.Mutation[13].

6. The Proposed Model

The adopted identification system consists of four main phases, as shown in Fig.4. Data collection, pre-processing data phase, identification phase, and evaluation phase. The pre-processing data phase is an important stage in preparing the DNA data for the next phase. Then, the DNA sequence will be passed to the matching algorithm (GA) to find the best matching results. Finally, the evaluation is applied to compute the system performance.

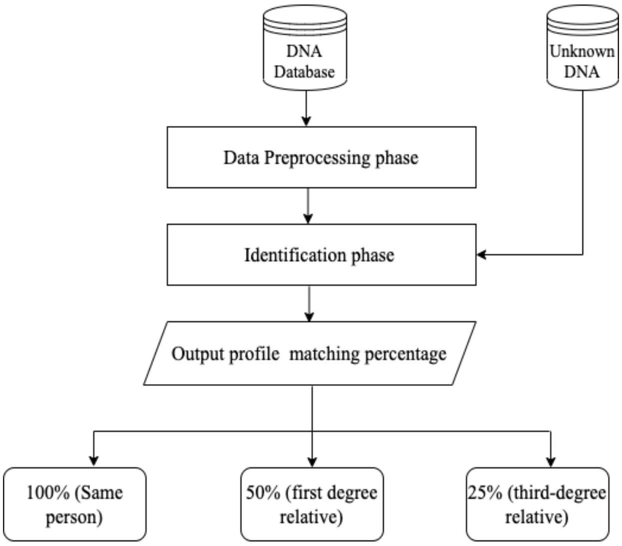

Fig.4. The proposed model phases

5.1.Data Collection

1500 samples, comprising seventeen Y-chromosomal STRs and mtDNA noncoding region sequencing, were gathered for this study from the Iraqi Ministry of Interior/Basra Investigation Center. This study's goals are to ascertain the genetic composition of the Basrah region and assess the significance of these STR loci for forensic genetic applications [14]. Table I shows the dataset details where the table shows the sources of the obtained DNA profiles(markers) dataset that were derived as classes. While Fig.5 illustrates the following attributes (for each person's DNA) that are considered for the process of marching: FGA, D5S818 D18S51, TPOX, VWA, D19S433, D2S1338, D16S539, D13S317, THO1, D3S1358, CSF1PO, D7S820, D21S11, D8S1179, and labels. These genetic markers are frequently employed in forensic DNA analysis to build DNA profiles. A DNA profile is a collection of genetic data that may be used to identify information. The markers previously mentioned are short tandem repeat (STR) markers, which are unique regions of the genome where short DNA sequences are repeated repeatedly. Individual DNA profiles may be constructed using changes in the amount of times the sequence is repeated depending on the individual. The "Core Forensics STR Collection," a set of markers used universally in forensic DNA analysis, is a DNA profiling technique that uses the markers mentioned above.

Table I: The dataset sources of the obtained DNA profiles

Class	**No. of DNA profiles**
Victims of mass graves	311
Unknown bodies	374
Civilians and law enforcement victim	402
suspect individuals	413
Total	1500

Fig.5: The attribute and their value for the DNA profiles.

FGA	D5S818	D18S51	TPOX	VWA	D19S433	D2S1338	D16S539	D13S317	THO1	D3S1358	CSF1PO	D7S820	D21S11	D8S1179	0
237254	153161	289295	227227	168176	120122	321349	269274	235239	172172	126126	322326	261273	203217	129142	1
238250	157157	281381	227235	172184	126134	325329	274274	231239	173185	126130	309322	257261	215215	142160	2
250254	139161	296320	227227	172180	114124	317345	269281	231231	173173	126126	322326	261261	207209	137147	3
230246	153157	292302	227227	164168	114118	345345	277281	227239	173177	118126	322330	261269	198198	125142	4
230246	157161	281286	231243	188188	118122	321325	274277	231235	177185	126134	322334	273281	205223	147151	5
230238	157157	281302	231239	168172	114118	321341	254266	227235	173185	130130	330330	273273	203207	137147	6
258267	157161	298311	231231	172180	114122	313317	274281	219235	173173	126134	330330	265277	207207	137147	7
242250	157157	296307	239243	164180	118118	317329	266274	235239	173185	134134	326326	273281	202205	137151	8
238242	149157	292292	227227	172184	118118	317345	274277	235235	173185	126134	326330	261273	203217	133133	9
230246	153157	286316	227239	164184	110118	325341	266274	235242	173177	118122	322330	269277	198205	142151	10
234242	149157	294298	231243	168188	122122	317317	269277	235235	173185	138138	322326	269269	205209	142155	11
234234	149157	294298	243243	188188	122122	313321	285285	235235	173173	138138	326326	277277	205209	142155	12

5.2.Data Pre-processing Phase

In this study, the removal of DNA sequences (or some regions of the DNA sequence) containing missing data from the DNA profile was used to ensure reliable results.

5.3.GA Phase

GA operates, generally, on binary chain structures, similar to biological organisms (genomes). Structures evolve by the survival of the fittest using a random and orderly information-sharing system. As a result, each generation produces a new set of binary chains based on the fittest members' parts from the previous group. Binary encoding of the parameter space is processed and operated on by a Genetic Algorithm. The information is encoded in binary chains as a result of this coding (which is an important aspect of the design process of GA).

The coding scheme, in our case, includes parameters representing the data used to diagnose the condition. Each parameter can be a symbol in a binary chain. This binary strand is the evolving genome of GA. Each binary sequence creates a gene of the Genetic Algorithm genome. The origins of these genomes constitute the genetic algorithm assembly. In this work, a genetic algorithm is used to match the database as it selects the more correlated DNA in the database to obtain the highest accuracy and the least time in defining the data category as described in the algorithm (1). **Fig.6** shows the processing of GA matching with a neutral example.

Algorithm (1): Genetic algorithm matching DNA

Input: DNA database

Output: the DNA matching

Step1: Read the DNA database

Step2: A cell holds the index DNA in the dataset utilized, and no comparable numbers are repeated in a single chromosome in a random population generation with n individuals (chromosomes) and 4 cells per chromosome.

Step3: Apply fitness function(alignment)

: This function determines each chromosome's effectiveness by comparing it to the new input, DNA.

Step 4: Operation of crossover: This stage separates the two chromosomal exchanges and results in sons' development by replacing the second half of the first chromosome with the second half of the second chromosome utilizing a single point.

Step 5: Mutation operation: by flapping between the third index in the cell of child one and child two

Step 6: Update the population

Step 7: Repeat previous operations (from step 4, until reaching the desired goal or no more chromosomes.

Step 8: Calculate the degree of similarity between the introduced DNA and DNAs in the database.

End

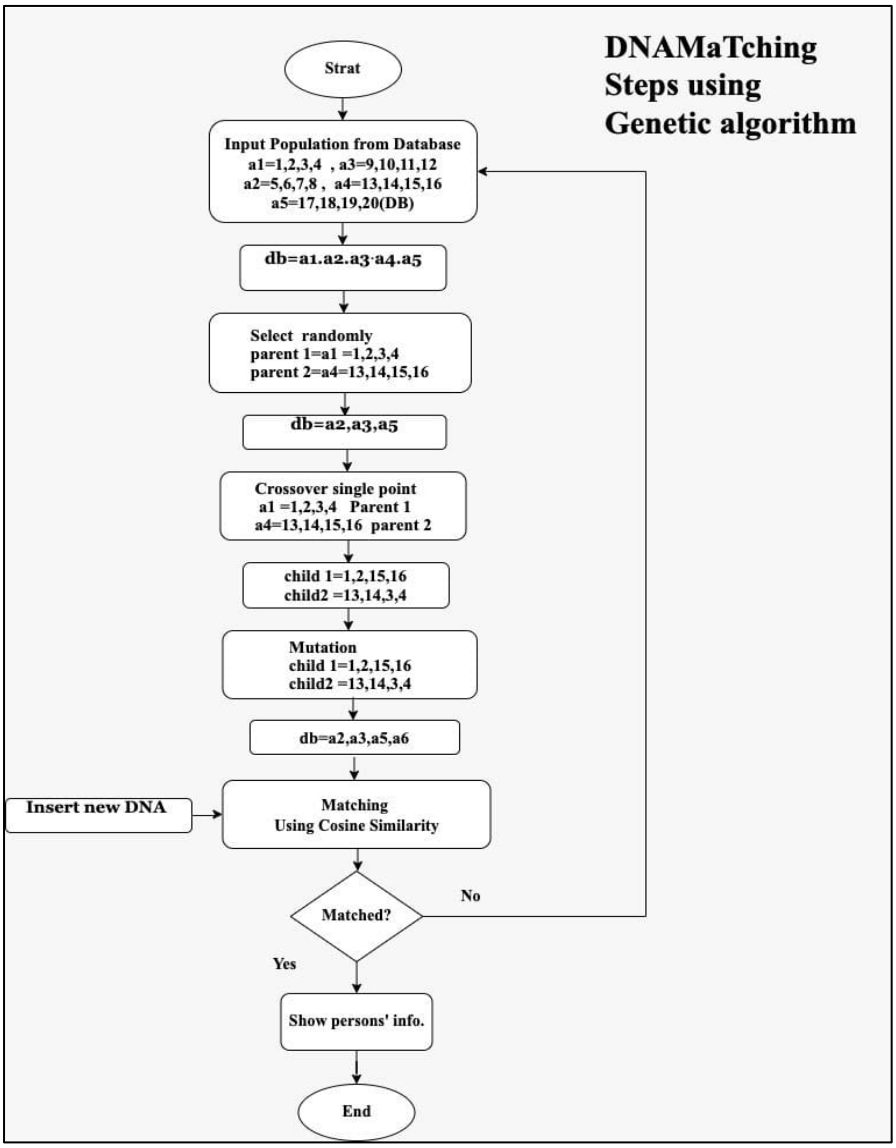

Fig.6. Processing of GA matching

6. DNA Matching Evaluation

The Match confidence scores between two DNA profiles depend on the inheritance of common traits from a common ancestor, and this score decreases with each successive generation. In this paper, the normalized dot product of the two attributes is found using the cosine similarity metric[15]. We will try to calculate the cosine of the angle between the two attributes by calculating the cosine similarity (in our case the DNA attributes). 0° has a cosine of 1, while all other angles have a cosine that is less than 1. Let's suppose *New_unknown*, and *Data_base* represent the database of unknown persons and the predefined DNA sequences, respectively. The length of *New_unknown*, and *Data_base* act as the length of DNA sequence. Also, let k be the length of the DNA sequence in *Newp*, and *DBp* that equals 16 as shown in eq.(1) that represents the equations of cosine similarity equation:

$$(\text{New\_unknown},\text{Data\_base})=\cos(\theta)=\frac{xy}{\|x\|\|yy\|}=\frac{\sum_{i=1}^{n}\text{New\_unknown}_i\,\text{Data\_base}_i}{\sqrt{\sum_{i=1}^{n}\text{New\_unknown}_i^2}\sqrt{\sum_{i=1}^{n}\text{Data\_base}_i^2}} \quad (1)$$

The results of 100% mean $New\_unknown = Data\_base_i$ and the DNA sequence belongs to the same person, while 50% means the first-degree relative (parents and siblings). Whereas 25% denotes second-degree relative (grandparents, half-siblings, aunts/uncles, or double first cousins), and 0% means there is no matching at all. In eq. (1) calculated the fitness value for unknown DNA profiling New_uknown_i of GA. $Data\_base_i$ which has the best matching value is considered the optimum solution. The fitness function is the best matching score calculated for each unknown DNA profiling.

7. Results and Discussion

In this study, DNA sequences containing missing data from the DNA profile were removed to ensure reliable results. The total collected data was roughly 1500 profile, and the numbers of eliminated profiles were 325. In this work, the genetic algorithm was used to match DNA. According to the explanation of the genetic algorithm's mechanism for DNA matching, the primary communities, or chromosomes, are first created. Each chromosome contains four cells, or four numbers, which are used to represent each chromosome (16). Each cell corresponds to a portion of the DNA address.The representation of the data in a genetic algorithm where consisting of a number of chromosomes and each cell contains a DNA site and does not repeat any site number, and this enables us to compare with all cells in a faster way than the traditional method. The selection operation of two chromosomes randomly represents parents in the genetic algorithm after the selection process of two chromosomes, we carry out the process of crossover and produce a new two chromosomes representing the child by using single point cut, in another word, the child1 take the first part from parent1 and the second meddle take from the parent 2, so as child2. Next, the process of mutation is performed, which enables the production of a change in the child, which helps in the process of reaching the goal faster, knowing that the mutation occurs if the probability is greater than 0.5. After that, the algorithm will continue the process of comparison with the first child if it matches by 100%,50%, or 25%, the algorithm will continue to find all the related DNAs. After checking all the databases, the algorithm will show the list of all matching DNAs with their information.The results of 100% matching mean that the DNA sequence under consideration belongs to the same person, while 50% of matching means the first-degree relative (parents and siblings). Whereas 25% denotes second-degree relative (grandparents, half-siblings, aunts/uncles, or double first cousins), and 0% means there is no matching at all. Table I shows a sample of the obtained results. The Table I shows the similarity score (degree of matching), the sequence of the DNA in the database, and the degree of kinship. The results suggested that the introduced DNA matched three profiles in the database with different degrees of kinship.

Table I: A results of the DNA matching examination

Similarity score%	Matching with profile No.	Kinship
50	65	First-degree relative (parents and siblings).
100	77	The same person
25	82	second-degree relative (grandparents, half-siblings, aunts/uncles, or double first cousins)

8. Conclusion

DNA is a valuable tool in forensic science because it can be used to either clear the name of someone who has been wrongly accused or to help identify and prosecute the actual perpetrator of a crime. This is achieved through the analysis of biological evidence collected from a crime scene, which can reveal the genetic material present in DNA. These organic elements can be found in many different forms, including bodily fluids and human remains. Therefore, DNA technology is crucial for identifying people. In this work, the GA algorithm is used for DNA matching for issues related to law enforcement and other issues related to DNA tests. The used database is real data collected from the Iraqi interior ministry.

The results using GA algorithm show 100% accuracy. Furthermore, besides the simulation results, the proposed work has been tested in a real-world case scenario in the Iraqi ministry of interior at the Basra Investigation Center, in which the robustness and accuracy have been confirmed. Finally, the proposed method can replace the manual method that is currently used there (Basra Investigation Center).In the foreseen future, our method will be occupied with a user-friendly interface to be handled by law enforcement that needs such technology.

References

[1] L.M. Macías-García, M. Martínez-Ballesteros, J.M. Luna-Romera, J.M. García-Heredia, J. García-Gutiérrez, J.C. Riquelme-Santos, Artif Intell Med **110**, 101976 (2020).

[2] F. Celli, F. Cumbo, E. Weitschek, Big Data Research **13**, 21 (2018).

[3] R. Touati, I. Messaoudi, A. E. Oueslati, Z. Lachiri, M. Kharrat, IRBM **42**(3), 154 (2021).

[4] Ü. Atila, Y. Y. Baydilli, E. Sehirli, M. K. Turan, Comput Methods Programs Biomed **186**, 105192 (2020).

[5] Y. Wang, M. Alangari, J. Hihath, A. K. Das, M. P. Anantram, BMC Genomics **22**(1), 1(2021).

[6] M. Tahir, M. Hayat, K. T. Chong, Neural Networks **129**, 385 (2020).

[7] H. Alotaibi, F. Alsolami, R. Mehmood, International Journal of Advanced Computer Science and Applications **12**(11), 130 (2021).

[8] A. Zaguia, D. Pandey, S. Painuly, S. K. Pal, V. K. Garg, N. Goel, Comput Intell Neurosci, **2022**, (2022).

[9] G. Zhong, T. Li, W. Jiao, L.-N. Wang, J. Dong, C.-L. Liu, Neurocomputing, **382**, 140 (2020).

[10] M. Inutsuka, "Set-level gene expression data analysis with machine learning," PhD thesis, Czech Technical University in Prague, Prague, Czech, (2014).

[11] T. Ching , J R Soc Interface **15**(141), (2018).

[12] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, V. B. S. Prasath, Information (Switzerland) **10**(12), 390 (2019).

[13] D. Wu, X. Zhu, L. Tan, H. Zhang, L. Sha, X. Fan, Y. Wang, H. Kang, J. Lu, Y. Zhou, Cytogenet Genome Res, **161**(4), 213 (2021).

[14] B. M. O. Medan, "Introduction Chapter," PhD Thesis, University of Basrah, Basrah, Iraq, (2019).

[15] K. Zhou, K. Ethayarajh, D. Card, D. Jurafsky, in Proceedings ofthe 60th Annual Meeting ofthe Association for Computational Linguistics **2**, 401 (2022).

Journal of Basrah Researches (Sciences) **49**(1), 13 (2023)
DOI: https://doi.org/10.56714/bjrs.49.1.2

استخدام الخوارزميات الجينية لمطابقة ملف الحمض النووي

نوال شهاب جابر[1] ، زينب علي خلف*[2]

[1] قسم علوم الحاسوب، كلية التربية للعلوم الصرفة، جامعة البصرة، البصرة، العراق.

[2] قسم علوم الحاسوب، كلية العلوم، جامعة البصرة، البصرة، العراق.

معلومات البحث

الاستلام 16 تشرين الثاني 2022
القبول 08 كانون الأول 2023
النشر 30 حزيران 2023

الكلمات المفتاحية

الأدلة الجنائية، مطابقة البصمة الوراثية، الخوارزمية الوراثية.

Citation: N.S. Jabir, Z.A. Kahlaf., J. Basrah Res. (Sci.) **49**(1), 13(2023). DOI:https://doi.org/10.56714/bjrs.49.1.2

الملخص

الملخص. تحتوي معظم الخلايا في جسم الإنسان على الحمض النووي، وهو فريد لكل شخص ويترك أثرًا أينما ذهب. الطب الشرعي اقسام الأدلة الجنائية في الأجهزة الأمنية يمكن ان تستفيد من هذه الخاصية للتوصل إلى استنتاجات حول هوية المشتبه بهم والضحايا في مسرح الجريمة أوضحاها الإرهاب والمقابر الجماعية و كذلك فحوص تأكيد الابوة. ومع ذلك، فإن مطابقة ملفات تعريف الحمض النووي يدويًا أمر غير عملي وعرضة للخطأ، خاصة في قواعد البيانات الكبيرة. ولهذا فأننا في هذا العمل قمنا بأتمتة هذه العملية من خلال استخدام الخوارزمية الجينة لمطابقة الحمض النووي لمجموعة من الأشخاص مع قاعدة بيانات. في هذا العمل تم استخدام بيانات حقيقة تم جمعها من وزارة الداخلية - مركز تحقيقات البصرة. اثبتت نتائج المحاكاة دقة تبلغ ١٠٠٪لكل النماذج التي تم اختبارها وبسرعة عالية، بينما أظهرت التجربة العملية التي أجريت في مركز تحقيقات البصرة دقة وصلت الى ١٠٠٪ايضا. وفي نهاية التجربة، طلب مركز تحقيقات البصرة نسخة من المحاكاة لاستخدامها في مهمة مطابقة الحمض النووي الخاصة بهم لتحل محل الطريقة اليدوية التي يستخدمونها حاليًا.

***Corresponding author email :** zainab.khalaf@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq