# Biometrics System based Human Identification using STR DNA Marker

Saja Dheyaa Khudhur
Computer Engineering Department,
University of Technology,
Baghdad

Muayad Sadik Croock
Computer Engineering Department, University of Technology,
Baghdad

## ABSTRACT
Human identification plays an important role in numerous fields in the world, such as forensic, government institution, medical application, ..etc. Biometric is a metric to measure the biological characteristics that has then been used for identification and verification functions. The verification process is performed by compare the biometric characteristic such as, fingerprint, iris, DNA, *etc*., with a pre-stored record, while the identification process done by find the best match between the biometric characteristic and all records saved in a database. In this paper, we propose a biometrics system based on the DNA as a biometric technology for human identification using sixteen Short Tandem Repeats (STRs) DNA marker. A database of 139 records has been considered for Iraqi Diyala Province Population as a real data. In addition, a million DNA profiles has been generated randomly to test the performance of the proposed system. A database has been built using SQL SERVER software environment that provides a high efficiency in human identification. The proposed identification system introduces different search and matching methods with distinct matching level ratios that ease the utilization by users. The outcome results of the proposal system show a flexibility in term of inserting, searching, updating and a high ratio of matching.

## General Terms
Bio-metrics System, DNA fingerprint.

## Keywords
Biometrics, Identification system, DNA-profile, STR loci, Database, SQL server 2012.

## 1. INTRODUCTION
Biometrics are the one of technologies that deal with physiological and behavioral features of the human in order to verify or identify individual [1]. Human identification plays an important role in many directions. So that, the selection of the identification technology depends on the security issues from the demandable fields. The biometrics technologies provide the security level which is not able to be cheated [2].

The famous biometric techniques as shown in different research are: fingerprint, iris, hand geometry, signature, voice, etc. which had have many applications in our life. In [3], the author presented a comprehensive survey of a biometric research which is based on gender detection upon using the fingerprint as a biometric tool. He combined various strategies and methods that had been elaborated in the paper to handle more accurate results.

Many researchers implicated the uniqueness of a movement of the human eye and using it as a biometric technology to verify and identify individuals. In [4], the author introduced a novel approach of biometric system termed as Eye-tracking system which was based on the eye movement to identifying persons by using the Eyewriter hardware and software.

The behavioral biometrics have many fields in direct on identify or verify functions, a handwriting was considered as a significant method in the identity science when considered as a biometric tool in term of "handwriting biometrics" [5]. In addition, the human hand at all can be utilized as a behavioral biometrics in terms of eigenfinger and eigenpalm [6]. In [7], the author proposed a new and operative method with two levels for identification and verification writer automatically by using his handwriting image with no consideration about what he wrote.

As much as important of these applications, but still have a limitation in the security level that does not provide the requirements of some governmental or private institution [8]. Using a DNA as a metric for the human physiological characteristic can provide a high degree of accuracy. The human genome contain a repetitive hypervariable regions with variable length which repeats in tandem. These regions in 1980s considered as markers used for mapping the human genome, several year later, Alec Jeffreys (1985) concluded that this repeated regions could be utilized for human identification. At the beginning of using DNA for identification purpose, Variable Number Tandem Repeat (VNTR) loci were considered firstly as a marker because of it has high level of heterozygosity. In many cases such as criminal cases, the investigators sometime yields a degraded samples with nanogram or picogram amounts of DNA that is uncomfortable for VNTR analyzing. From that reason, the scientist utilized the Polymerase Chain Reaction (PCR) which is a technology to amplify shorter hypervariable regions (STR) to generating thousands to millions of copies of a particular DNA sequence [9].

In 1990, STR loci considered as a reliable tool for identification purpose [10],and a golden standard in forensic cases and paternity testing [11,12]. In [8], the author created a high accuracy biometric system which used the DNA as a biometric tool to identify individual by comparing his STR marker with the pre-stored STRs in the huge database. DNA profiling in general is a reliable and robust system to gauge the human genetic characteristics which can be used as a biometric data for the identification process [13].

In this paper, we present a new reliable and powerful biometrics system which uses the STRs DNA marker as a biometric tool. This is to satisfy a high degree of the security and accuracy for identification function achieved by building a hug database with capacity of up to ten millions records. We utilize fifteen autosomal STR loci, which are (D3S1358, VWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, TH01, TPOX, CSF1PO, D16S539, D19S433, D2S1338) plus Amelogenin (AMEL) to determine sex. Where the first thirteen loci are the CODIS core loci..

## 2. DNA-PROFILE DATABASE SYSTEM

The database row data was collected from the pre-stored data that was recorded in 2015, which is related to the population of Iraq/Diyala Province [10]. The system provides the ability of adding, updating, searching and matching of the DNA profile across usage the SQL Server 2012 with the utilizing the Graphical User Interfaces (GUI) provided from the Visual Studio. This GUI allows the ordinary user to transact with the DNA database without the need to pre-knowledge about the mechanism of stocking and arranging data.

### 2.1 Database Building

The Relational DataBase Management System (RDBMS) was appeared in 1980s as a standard for all database type where the data marshaled in these system in a relational model. Several years later, almost in mid-1990s a new data platform called Microsoft SQL server (MSSQL) entered the RDBMS market. MSSQL server played a vital rule among all the databases in the world by what it offers of attributes which make it fully good to any application [14].

In this paper, we builds a SQL database by utilizing SQL Server Management Studio (SSMS) which is a software application that provides an integrated environment for managing all SQL Server components. SSMS plays as a graphical management tool which is easy to using and dealing with it [15],[16] and [17]. The underlying database, called " DNAprofileDB", has been built in two tables structure, one for the DNA profiles called "IndivDNAprofile" and one for the personal information for individuals called "IndividualInfo". The "IndivDNAprofile" includes 33 columns that are [ID] and the 16 loci, which are previously introduced, where each loci composed 2 alleles as shown in Figure (1). This figure shows a sample of the entire table due to the size limit.

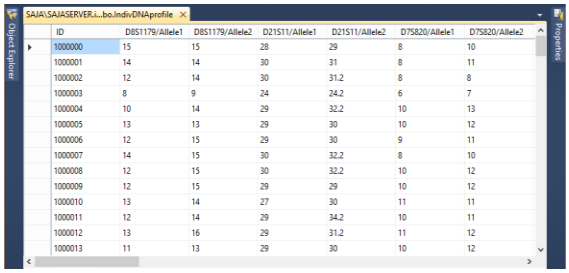



**Figure 1. IndivDNAprofile table.**

The "IndividualInfo" involve 11 columns that are: [ID]; which is a shared columns in both tables to ensure the connection between them, Full Name, Mother Name, Birthday, Career, Work Place, Section Address, Street Address, Home Address and Photograph, as shown in Figure (2). This figure cannot cover the whole table due to the size limit. It is important to note that the ID in both tables is unique to each individual, to hold the identity of each individual.

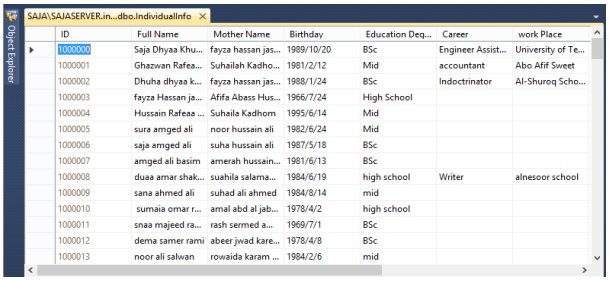



**Figure 2. IndividualInfo table**.

## 3. DNA-PROFILE PROCESSING

The introduced real DNA database involves 139 row, where each row represents a one records. Each record dedicated to one person from Diyala province in Iraq. As mentioned before, the ID column was sheared between the considered two tables. Furthermore, the dealing with the presented database is easy and flexible in terms of adding new record, updating or searching for an existence profile, and matching with all records to satisfy the identity.

### 3.1 Proposed Algorithm

The proposed algorithm composes of four main processes which were explained in the Figure (3) as a flowchart. These processes are: inserting, updating, searching and matching where each one are clearly explained below.

i. Inserting: This process is used to insert a new record to the DNAprofileDB tables. As mentioned, the database has two tables. The IndivDNAprofile table holds the DNA characteristics which is saved as a DNA profile for an individual. In addition, the IndividualInfo table holds the personal information of that individual. Some points must be taken into consideration, which are:

- Field which are marked with (*) must be filled.
- The DNA information must compose all 32 alleles, regardless of the input method, whether manually or uploading a DNA file. The format of the DNA profile file must be at a csv extension.

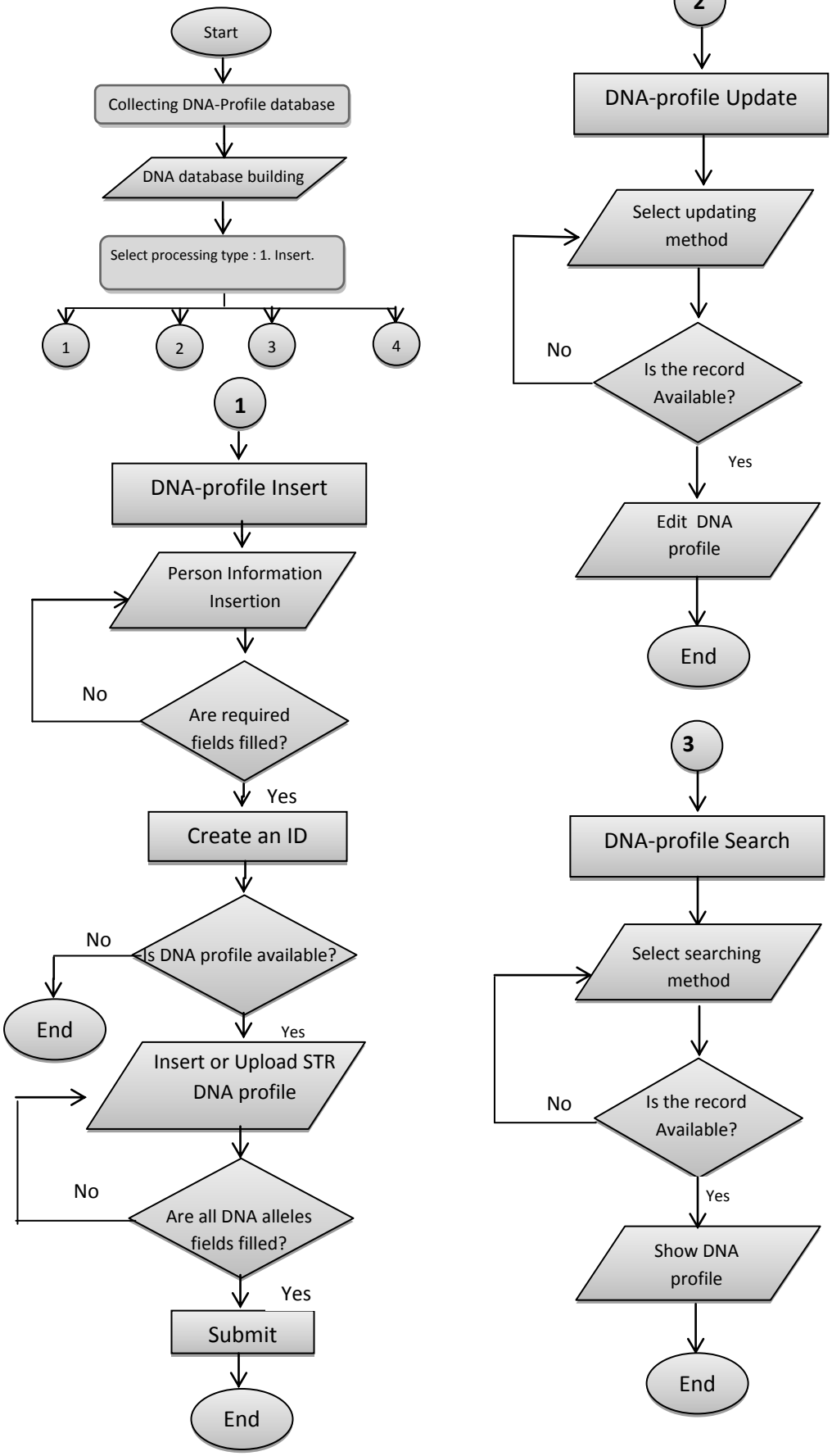
ii. Updating: This process is done on the records that were already included in the database. To find the required record, any of the ID, Full Name or Mother Name of the required record can be used. This process can easily update any field of the whole profile.

iii. Searching: Here, inspection of a desired record can be done by the searching process. As the updating process, ID, Full Name or Mother Name of the required record can be used to search for that record.

iv. Matching: This process is done by insertion the DNA information to get the record that has a matching proportion equal or greater to the matching rate which may be 100%, 70%, or over 50%, with all records stored in the database. the Internal procedure for this process done on two stages:

- Filtering: This done by searching all DNA profile records that have a similarity of at least one allele of the D2S1338 locus because this locus has a higher degree of heterozygosity than the other locus. The outcomes of this process are, reducing the time consuming and utilizing the system resources in an efficient manner.

- Finding: Now, after the filtering process, the finding is done to get the records that have a matching proportion greater or equal to the specified rate.



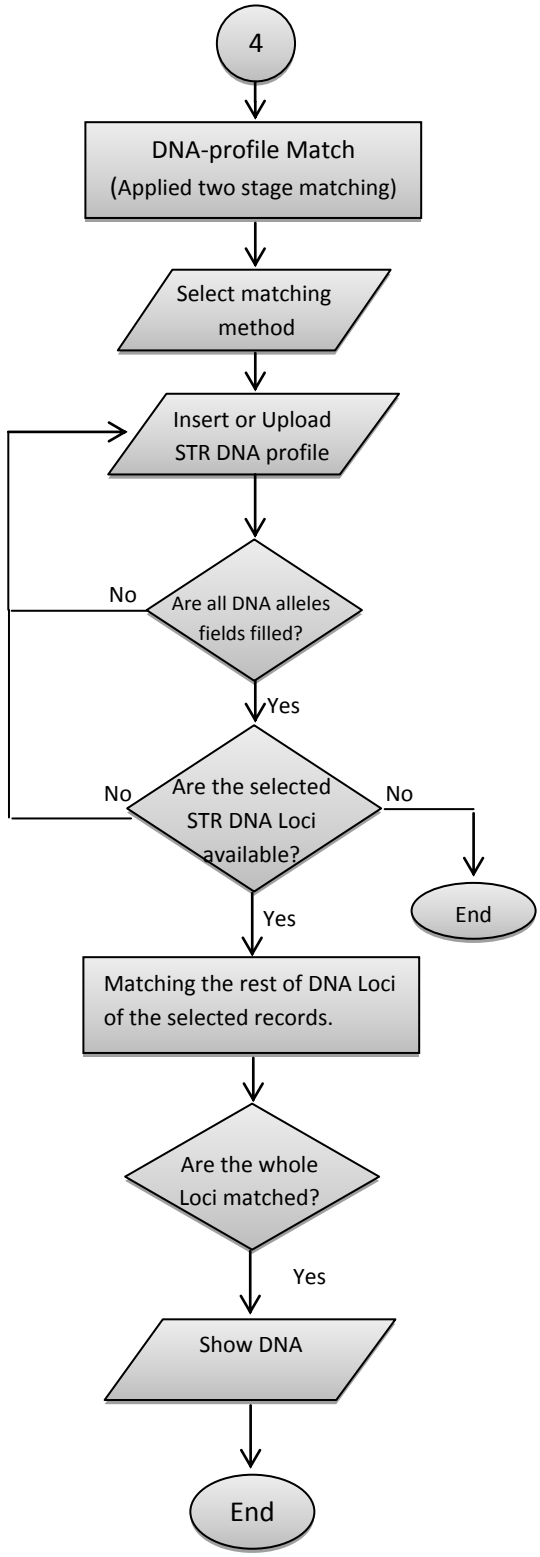



**Figure 3. Flowchart of the proposed algorithm.**

## 3.2 GUI Design

Visual Studio (VS) environment is utilized to design and implement the GUI of the proposed system. Using the VS as a designer application we possible guarantee that the GUI will have easy dealing design and will not require skilled users for dealing with it. Figure (4) shown the system home page which involves four main buttons that are concerning to the DNA profile processing. These are: DNA profile Insert, DNA profile Update, DNA profile Search and DNA profile Match.

- When the user click on the "DNA-profile Insert" button, the personal information fields will appear. Here, the user at least must fill all the fields with * in order to generate an ID to the individual after clicking on "Create an ID" button. These information was considered as record in the IndividualInfo table as shown in Figure (5). Thereafter generating an ID to the new profile, the DNA profile can be inserted in the corresponding fields manually or automatically. In manually case, the user must click on the (Insert) button and then fills all the allele fields. In another case, automatically, the user must click on the (Upload) button and a csv file format which compose the DNA profile must be selected. Figure (6) shows the DNA profile insertion by automatically way. The structure of DNA profile file (CSV file) must be as shown in Figure (7). The CSV file must include 32 column where each two column are specified to one locus. Keep in mind, that the sequence of the loci is a very important and restricted thing. Afterword, the user now must be click on the (Submit) button which is a final step in the DNA profile insertion, to complete the whole profile of the individual in the database.


**Figure 4. DNA Profile Database Form**.


**Figure 5. New Record Insertion.**

- The update process is done when the user click on the "DNA-profile Update" button. The user can utilize the ID, Full name or Mother name to obtain the required DNA profile that wants to be updated.

The outcome of this process is the possibility of the user to update any field of the earned record in an easily series. Figure (8) show the chosen of a profile based on its ID to updating it.


**Figure 6. File with extension of csv Uploading.**

- At the same manner, searching process was done by clicking the user on the "DNA-profile Search" button as shown in Figure (9). The obtained profile can be printed or saved as file with PDF extension and can get screen capture of it.

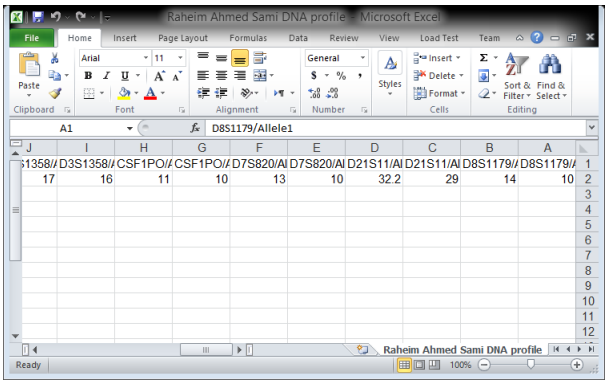



**Figure 7. Structure of file with extension of csv.**

- Now, we are starting to talk about an important process in the proposed biometrics system which is the matching operation. This process begins with an attempt to match the inserted or uploaded DNA profile to all profiles in the database by two stages.


**Figure 8. DNA-Profile Update.**

These stages are: filtering and finding. At the filtering stage, all profiles are searched in order to candidate profiles which have at least one allele of the locus D2S1338 is similar to the corresponding allele in the DNA profile that wants to identify its owner. As a result, this stage can reduce the time consuming and limits the usage of the visual resources. In addition, it can increase system scalability to handle large numbers of records in a very short time.


**Figure 9. DNA-Profile Search.**

The another stage, finding, selects one or more records from the filtered profiles which have matching rate equal or greater than rate that are selected by the user before starting matching algorithm. The matching rate was computed based on the number of matched alleles from all 32 allele. Figure (10) show the matching operation based on matching rate over 50%. Moreover, "About" button gives to the user the highlighted information about the system, whereas "Home" button returns back the user to the home of the system.


**Figure 10. DNA-profile match.**

## 4. RESULTS

The proposed system was tested in terms of insertion function by entering the records of 139 DNA profiles for Diyala Province. As shown in Figure (5), we use the DNA-Profile Insert process and as a result, the insertion is done with high flexibility and without any errors.

In order to test the updating process, we select the method of ID searching, for example the profile of (ID=1000056) as shown in Figure (8). Figure (11) show the output of the update process where each field is ready to update regardless of whether DNA information or personal information.

In term of searching process, as shown in Figure (9), we select the profile with the name (Athraa Amar Hassan). Figure (12) shows the obtained profile that can be printed, a screen capture taken or saved as a pdf file format.

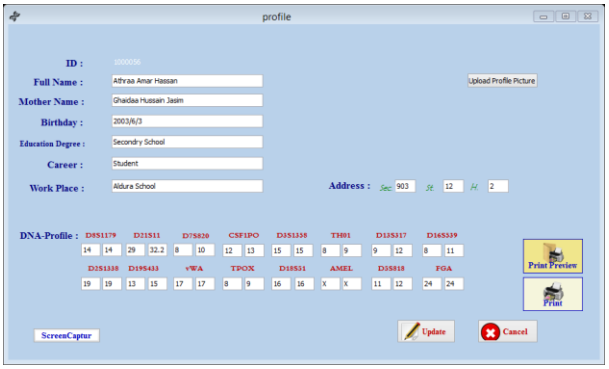



**Figure 11. Output of Update Method.**


**Figure 12. Output of searching method.**

For the matching process, Figure (13) shows a table which contains the profiles that have matching rate equal or greater than 50% as shown in the Figure (10). In addition, by clicking on the any item in this table, the user can view the whole profile that corresponding to it.

In order to test the system for identification, we generate a random DNA profile based on the dedicated rang of each locus in the allelic ladder fetched from the GeneMapper1 v. 3.2 software (Applied Biosystems, Foster City, CA, USA) that was used to analyze the amplification products. We generate 1000 and a million records and the time consumed is shown in Table (1).

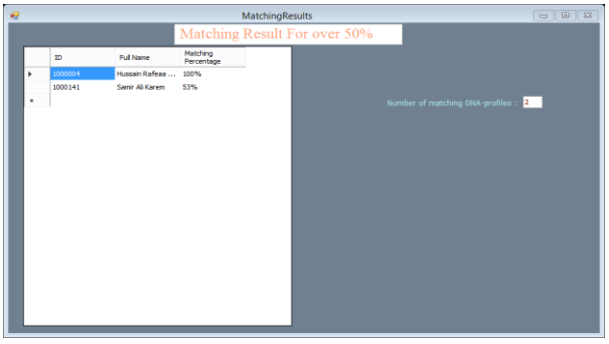



**Figure 13. Output of matching method.**

From the obtained results, all DNA profile processing operations confirm that the proposed biometrics system is efficient and capable to handle a huge amount of records with high degree of flexibility and efficiently. Furthermore, the proposed system provides a high degree of security where the username and password was presented within the connection string of the SQL server.

**Table 1. Processing time consuming table.**

| Type of Process | Number of record | Time consuming | | |
|---|---|---|---|---|
| | | Not found | Average | Random |
| Matching Process | *139* | 12 ms | 530 ms | 527 ms |
| | *1000* | 144 ms | 740 ms | 647 ms |
| | *1000000* | 516 ms | 3043 ms | 5000 ms |
| Searching Process | *139* | 514 ms | 568 ms | 510 ms |
| | *1000* | 503 ms | 576 ms | 562 ms |
| | *1000000* | 512 ms | 640 ms | 664 ms |

*** ms = millisecond**

## 5. CONCLUSION

A biometrics system has been presented. The database was built based on the fifteen autosomal STR markers (the 13 CODIS core loci and D19S433 and D2S1338) pulse Amelogenin for 139 persons from Diyala Province. SQL Server 2012 software was utilized to build that database. It is important to note that the above STR markers were considered in human identification. After testing all DNA profile processing operation including: insert, update, search and match, we conclude that all these processes were performed in efficient way. This system is presented for identification purpose and shows great and satisfactory results in all aspects in terms of time, efficiency, capacity and accuracy. Moreover, Visual Studio environment was utilized to build the GUI forms which allows the ordinary user to transact with the DNA database without the need to pre-knowledge about the mechanism of stocking and arranging data.

## 6. ACKNOWLEDGEMENTS

We are thankful to our colleagues ant to the Forensic DNA Center Research & Training at Al-Nahrain University/ Iraq.

## 7. REFERENCES

[1] Zhang, David D. Automated Biometrics: Technologies and Systems, Springer US, 2000; pp. 1-21.

[2] Faisal Rehman, M. Usman Akram, Naveed Riaz, F. Kunwar, A Sharp Comparison of Biometric Techniques and Dental Biometrics. Proceedings of the 2nd International Conference on Engineering & Emerging Technologies (ICEET), Superior University, Lahore, PK, 26-27 March, 2015.

[3] Gornale, SS. Fingerprint Based Gender Classification for Biometric Security: A StateOf-The-Art Technique, AIJRSTEM 2015; 9(1): 39-49.

[4] Tait, BL. Behavioural Biometrics: Utilizing Eye-Tracking to Generate a Behavioural Pin Using the Eyewriter. In: Jahankhani H, Carlile A, Akhgar B, Taal A,Hessami AG, Hosseinian-Far A. Global Security, Safety and Sustainability: Tomorrow's Challenges of Cyber Security. London: Springer International Publishing 2015; pp. 348-59.

[5] Schomaker, L. Advances in Writer Identification and Verification. Proceedings of the 9th International Conference on ICDAR; 2007 Sept 23-26; Parana: Document Analysis and Recognition 2007.

[6] Ribaric S, Fratric I. A biometric identification system based on eigenpalm and eigenfinger features. Pattern Analysis and Machine Intelligence, IEEE Transactions on 2005; 27(11): 1698-1709.

[7] Bulacu M, Schomaker L. Text-Independent Writer Identification and Verification Using Textural and Allographic Features. Pattern Analysis and Machine Intelligence, IEEE Transactions on 2007; 29(4): 701-17.

[8] Tripathi VA. The Effects of Forensic DNA Typing on the FRR and FAR of a Biometric System. SCJAS April 10, 2015.

[9] Butler J M. Advanced Topics in Forensic DNA Typing: Methodology. London: Academic Press 2012; pp. 69-140.

[10] AL-Zubaidi, MM, Al-Awadi, SJ, Namaa, DS, Saleh, TY, Shehab, MJ, Hameed, SN, and Abd- Alatief, A. Genetic Variation of 15 Autosomal Short Tandem Repeat (STR) Loci in The Diyala-Iraqi Population. International Journal of Biological & Pharmaceutical Research 2014; 5(3): 131-35.

[11] Butler, JM. Review: Genetics and Genomics of Core STR Loci Used in Human Identity Testing. J. Forensic Sci, in press, Mar 2006; 51(2):253-65.

[12] Kashyap A, Kumar, A, Awadhanam S. Investigating contributors of the mixed DNA samples by forensic Bioinformatics; Uncertainty to certainty for Crime laboratories. Proceeding of the 15th International Conference on ICACT; 2013 Sept. 21-22; Rajampet: Advanced Computing Technologies 2013.

[13] Rudin N, Inman K, Stolovitzky G, Rigoutsos I. DNA Based Identification. In: Jain AK, Bolle R, Pankanti S, EDs. Biometrics. Springer US 1996; pp. 287-309.

[14] LEE J., "Oracle vs. MySQL vs. SQL Server: A Comparison of Popular RDBMS", https://blog.udemy.com/.

[15] "SQL_Serv_Man_Studio", http://www.boosla.com. (Accessed December 15, 2015).

[16] Karim AA. Improved Approach to Iris Normalization for iris Recognition System. Eng. & Tech. Journal 2015; 33(2): 213-21.

[17] Alwan N. Developing a Database System for the Laboratory Tests. Eng. & Tech. Journal 2013; 31(18): 52-67.