DNA-Profile Database Building Using STR DNA Marker For Diyala Province Population

Saja Dheyaa Khudhur, Dr. Muayad Sadik Croock, Mohammed Mahdi AL-Zubaidi

Abstract— **Many systems for human identification based on short tandem repeats (STRs) DNA marker are widely used in forensic such as, crime detection and identification of unknown person of mass disasters. In this paper, a DNA profile database system has been built based on fifteen autosomal STR loci, which are (D3S1358, VWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, TH01, TPOX, CSF1PO, D19S433, D2S1338, D16S539) plus Amelogenin (AMEL) to determine sex. These fifteen STR loci have been chosen from the total of 139 human genomic DNA sample donners. The DNA profile database of 139 donners is implemented using SQL SERVER. This is due to high flexibility in terms of adding or updating records as well as the searching for a specific DNA profile based on the related STR-marker. The test of the implemented system has been done in terms of insertion, updating and searching. The outcome of the presented database system shows a high efficiency and flexibility.**

Index Terms— **Database, DNA-profile, STR loci, SQL server 2012.**

I. Introduction

DNA-Profiling is one of the forensic techniques that has been used to identify individuals (Unknown person) basing on their DNA characteristics. Since 1990 the STR markers were considered as an effective tool for human identification and a golden standard in forensic sciences. The variation of STR loci among individuals leads to consider these loci as an important markers in the science of forensic for different cases such as, identifying of missing people and in the disasters. This is because of the high degree of heterozygosity [1]. In 1996, the FBI Laboratory has begun to consider core STR loci as CODIS (Combined DNA Index System) that is used to identify genetic markers within the available database [2]. These genetic markers has been identified at thirteen STR loci, plus Amelogenin (AMEL) to determine sex. Most European countries, including the United Kingdom are using eight of the thirteen core STR loci with additional markers that is known as the golden standard in forensic cases and paternity testing [3]–[4].

In reference [4], the authors introduced forensic bioinformatics approach to analysis mixed DNA profiling of seventeen suspects, done on fifteen distinct STR loci. They proved that their approach has the capability to handle large amount of data and can detect the identity of person from sample which handle mixed DNA using "MS-Excel database management tool (MixSTR)". In reference [5], new classes of lineage markers (Y-chromosome markers and the mitochondrial DNA (mtDNA)) and polymorphic loci (single nucleotide polymorphisms (SNPs)) has been used to tackle the problem of human identification.

In references [6] and [7], a new concept based on SNPs role in human identification was presented in forensic science termed Forensic DNA Phenotyping (FDP). FDP was simulated by the DNA-profiling limitations where the DNA profile of the putative person or his relatives does not exist in forensic DNA profile database. In references [8] and [9], the authors provided knowledge information regarding the genes related to the appearance of humans limited compared to the genes, related to diseases, such as coloration of head hair, iris and skin.

In reference [10], the first system of eye color prediction based on FDP was published in 2010/2011 termed IrisPlex which developed by Walsh et al. This system has been adopted in the prediction model presented by Liu et al. [11] to predict brown and blue eye color basing on six eye color predictive. In reference [12], Walsh et al. presented a new developed system termed HIrisPlex that include two prediction models, the previously developed, IrisPlex model, and a model for hair color prediction.

In this paper, we presented a solution to the limits that facing the investigative process which based on STR markers by establishing huge database system that includes two main tables. The first one contains the human DNA STR profile. In addition, the second tables consists of the information profile that enables the investigators to identify people by only using his DNA characteristics extracted from samples of biological material. The introduced DNA profile database system can efficiently decrease the required time of searching for a specific DNA profile using the name, generated ID, or mother name.

II. Paper DNA-profile Database System

DNA profile should be collected under using of same sample size of a specific population in order to determine the frequency of each allele in the corresponding STR, which is

Manuscript received March, 2016.

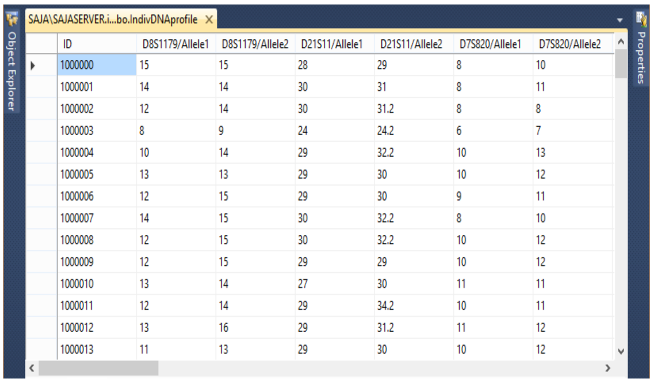
Saja Dheyaa Khudhur*, Computer Engineering Department, University of Technology, Baghdad, Iraq.*
Dr. Muayad Sadik Croock*, Computer Engineering Department, University of Technology, Baghdad, Iraq,*
Mohammed Mahdi AL-Zubaidi*, Forensic DNA Center Research & Training, Al-Nahrain University, Baghdad, Iraq.*

the number of copies of the repeat sequence within each of the STR locus. The distribution of alleles depends on racial or ethnic group and not on population subdivisions [12,13]. This project provides the ability of accessing, searching and updating of the DNA profiles for the population of Diyala Province, Iraq recorded previously in [1]. In addition, the utilizing of SQL Server 2012 across graphical interfaces built by Visual Studio enables the ordinary users to deal with the database without the need to write any SQL code or viewing the mechanism of storing and organizing data to be as a software package.

A. Database Building

In 1980s, the standard of database was appeared under the name of relational database management systems (RDBMS), which became standard for all database type. Data organized in these systems as a groups of tables in a relational model. In the mid-1990s, Microsoft SQL Server (MSSQL) entered the RDBMS market as a serious contestant. MS-SQL Server acts as the key player among the global data platforms around the world. SQL Server 2012 consists of an impressive range of features that make it well to any institution [14]. These amazing features can be summarized as : the characteristic of integrated reports and integration with SharePoint 2010, new Semantic Model feature, the strong performance, and the Surprisingly thing is that SQL Server 2012 can process 57 thousand of process per second. In addition, all these possibilities, the SQL Server 2012 has new improvements on all axes. These axes that are outlined by experts: the performance of critical mission in highest degree of confidence, and the usage of cloud capabilities according to exactly the needs required by business [15].

In this paper, the database is built using SQL Server Management Studio (SSMS), because it provides an integrated environment for accessing, configuring, managing, administering, and developing all components of SQL Server [16], [17] and [18]. A DNA Database has been built to include two tables. The first one composes the STR DNA profile called, "IndivDNAprofile", that consist of 33 columns, which are ID and 16 loci of 2 alleles as shown in Fig. (1). This Fig. shows the included columns and samples of data. It is important to note that the Fig. cannot show whole columns due to the size limit.

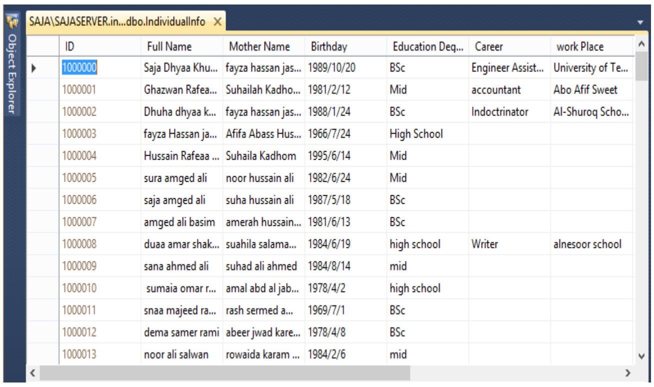


ID	D8S1179/Allele1	D8S1179/Allele2	D21S11/Allele1	D21S11/Allele2	D7S820/Allele1	D7S820/Allele2
1000000	15	15	28	29	8	10
1000001	14	14	30	31	8	11
1000002	12	14	30	31.2	8	8
1000003	8	9	24	24.2	6	7
1000004	10	14	29	32.2	10	13
1000005	13	13	29	30	10	12
1000006	12	15	29	30	9	11
1000007	14	15	30	32.2	8	10
1000008	12	15	30	32.2	10	12
1000009	12	15	29	29	10	12
1000010	13	14	27	30	11	11
1000011	12	14	29	34.2	10	11
1000012	13	16	29	31.2	11	12
1000013	11	13	29	30	10	12

Fig. (1). IndivDNAprofile table.

The other table contains the personal information for individuals, called "IndividualInfo", that consist of 11 column: ID, Full Name, Mother Name, Birthday, Career, Work Place, Section Address, Street Address, Home Address and Photograph, as shown in Fig. (2). This Fig. shows the involved columns of the underlying table with samples of information. It is also known that the Fig. is not able to cover the whole considered columns.

It is important to note that the ID is generated for each individual to be primary and unique to keep the identity of each person. In addition, the ID is a shared columns in both tables to ensure the connection between them.



ID	Full Name	Mother Name	Birthday	Education Deq...	Career	work Place
1000000	Saja Dhyaa Khu...	fayza hassan jas...	1989/10/20	BSc	Engineer Assist...	University of Te...
1000001	Ghazwan Rafea...	Suhailah Kadho...	1981/2/12	Mid	accountant	Abo Afif Sweet
1000002	Dhuha dhyaa k...	fayza hassan jas...	1988/1/24	BSc	Indoctrinator	Al-Shuroq Scho...
1000003	fayza Hassan ja...	Afifa Abass Hus...	1966/7/24	High School		
1000004	Hussain Rafeaa ...	Suhaila Kadhom	1995/6/14	Mid		
1000005	sura amged ali	noor hussain ali	1982/6/24	Mid		
1000006	saja amged ali	suha hussain ali	1987/5/18	BSc		
1000007	amged ali basim	amerah hussain...	1981/6/13	BSc		
1000008	duaa amar shak...	suahila salama...	1984/6/19	high school	Writer	alnesoor school
1000009	sana ahmed ali	suhad ali ahmed	1984/8/14	mid		
1000010	sumaia omar r...	amal abd al jab...	1978/4/2	high school		
1000011	snaa majeed ra...	rash sermed a...	1969/7/1	BSc		
1000012	dema samer rami	abeer jwad kare...	1978/4/8	BSc		
1000013	noor ali salwan	rowaida karam ...	1984/2/6	mid		

Fig. (2). IndividualInfo table.

B. DNA-profile Processing

The considered database includes 139 records for 139 persons in Diyala Province. Each record is represented as a row in the tables with keeping the sheared column ID for both tables. Moreover, the presented system allows the users to deal with IndivDNAprofile database in a high degree of flexibility in terms of adding, updating and searching for a record.

- Proposed Algorithm

  In this sub-Section, the proposed algorithm that explains the operation of underlying system is presented as a flowchart shown in Fig. (3). Throughout the introduced flowchart, the following three processing is proposed:

  1. *Inserting*: This process is used for inserting new records of DNA profile for an individual. The personal information of the entered new record is saved in the IndividualInfo table, while the DNA profile is saved in the IndivDNAprofile table. The IndivDNAprofile table is used later on for searching or updating. The following points have to be considered:
     a. All fields with (*) must be filled.
     b. If the DNA profile file is available, it must be at a csv format.
     c. All 32 allele must be entered.
  2. *Updating*: The updating of the already included DNA profiles is performed here. The finding of a specific record desired to be updated is done using the ID, Full Name or Mother Name searching. If the record is founded the editing can be performed to each field individually with high flexibility.

3. *Searching*: Here, the searching engine is used for reporting the finding records. The search can be done using the following fields ID, Full Name or Mother Name.

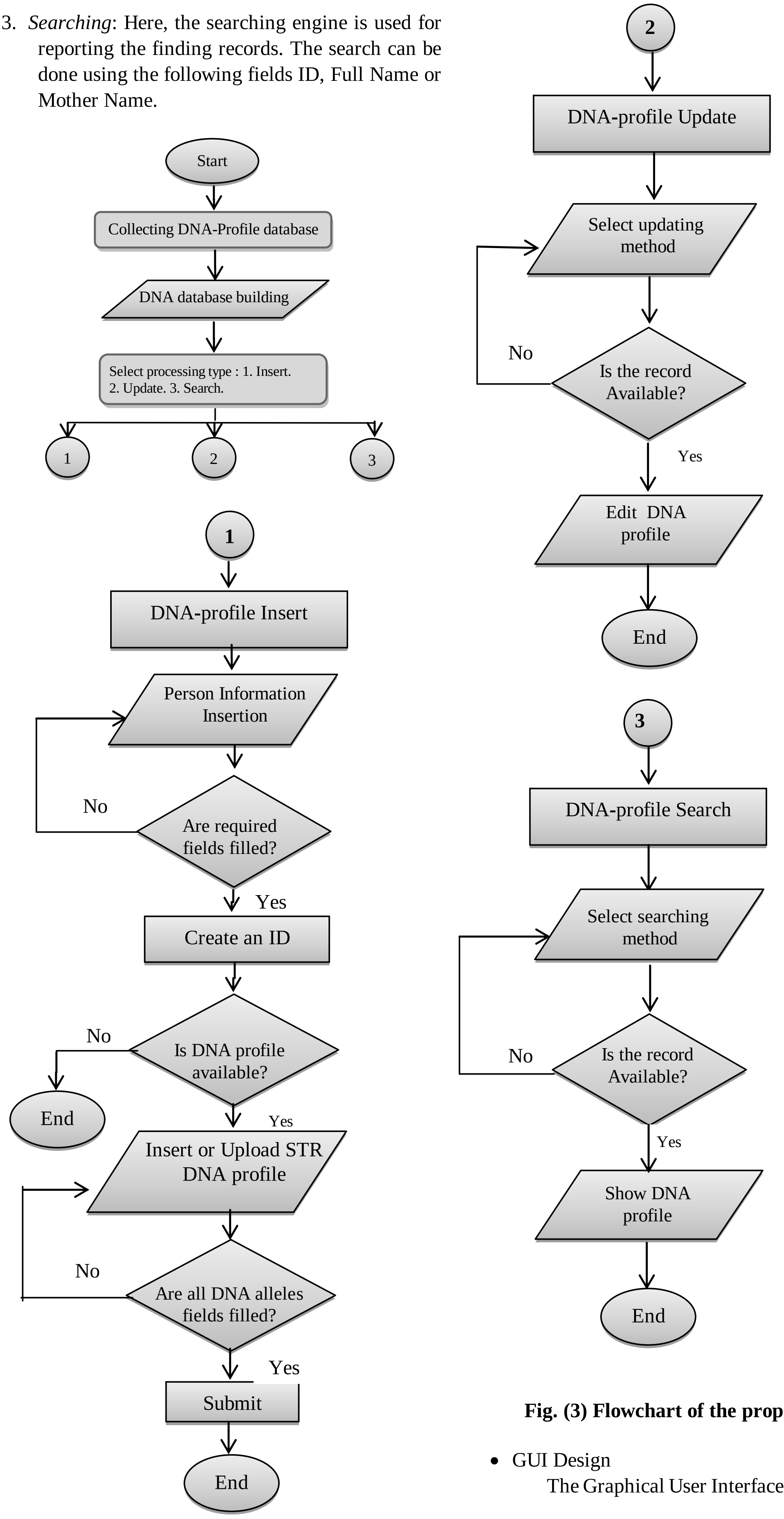


Fig. (3) Flowchart of the proposed algorithm

- GUI Design

The Graphical User Interface (GUI) of the proposed system is designed and implemented in the Visual Studio (VS) environment. VS provides easy dealing with SQL server and doesn't require user to be skilled in SQL server. It also provides a high security level to the database.

The home page form of the presented system composes three main buttons that are related to the DNA profile processing explained above as shown in Fig. (4). These buttons are DNA-Profile Insert, DNA-Profile Searching and DNA-Profile Updating.

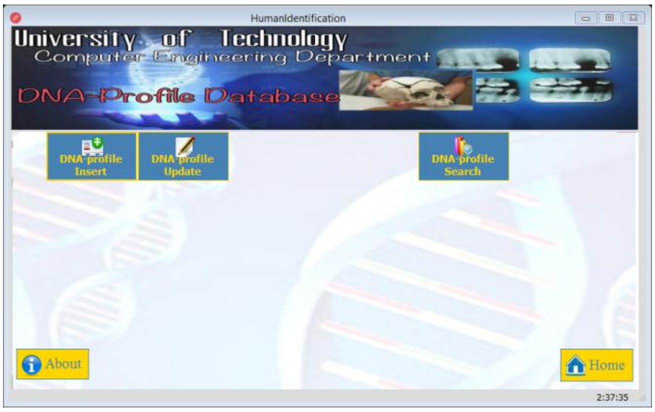
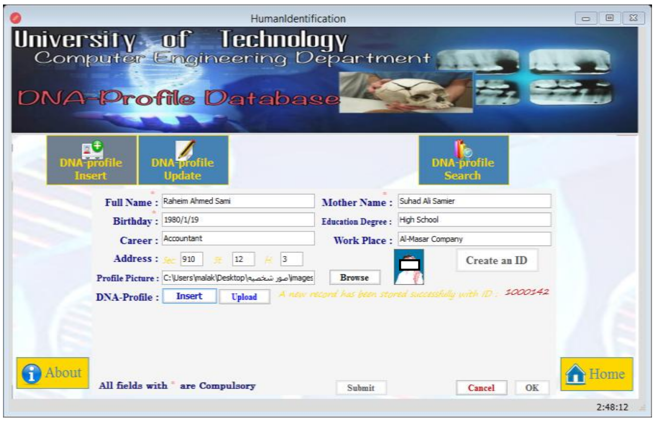


Fig. (4). DNA Profile Database Form.

1. *DNA-Profile Insert:* when the user click on the "DNA-profile Insert", fields for personal information are appeared. Now user can insert the personal information of the individuals that is considered as a new record in the IndividualInfo table, all fields with * must be filled, as shown in Fig. (5). After inserting the information, an ID to the new individual will be created when clicking on (Create an ID) button.

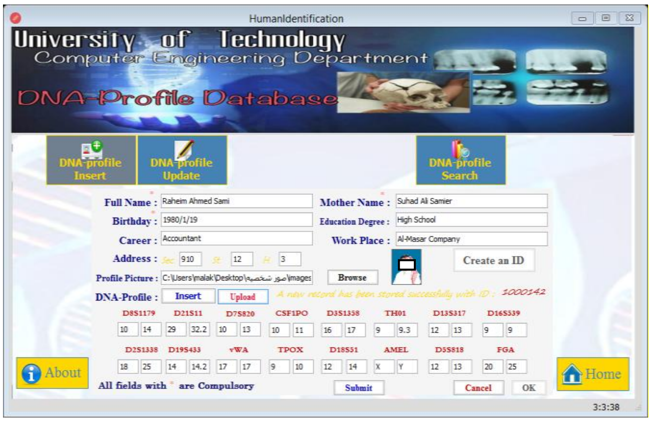
   Now the individual has an ID and then DNA profile can be inserted into locus fields by two ways: Manually, by click on the (Insert) button or automatically, by click on the (Upload) button when the file with csv extension of DNA profile is available as shown in Fig. (6).



Full Name : Raheim Ahmed Sami
Mother Name : Suhad Ali Samier
Birthday : 1980/1/19
Education Degree : High School
Career : Accountant
Work Place : Al-Masar Company
Address : Sec 910 St 12 H 3
Profile Picture : C:\Users\malak\Desktop\صور شخصيه\images
DNA-Profile :
A new record has been stored successfully with ID : 1000142
All fields with * are Compulsory

Fig. (5). New Record Insertion.

The structure of the CSV file must be as the structure in Fig. (7), which must contain 32 column and every two columns are dedicated to one loci. Additionally, the sequence of the loci should be taken on the consideration. The next step is clicking on the (Submit) button to complete the individual's profile.

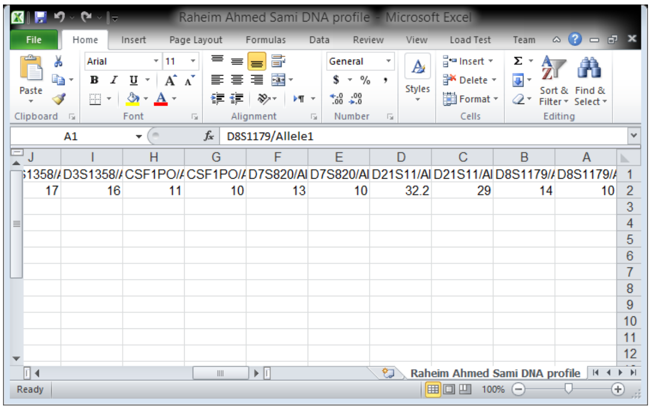


Full Name : Raheim Ahmed Sami
Mother Name : Suhad Ali Samier
Birthday : 1980/1/19
Education Degree : High School
Career : Accountant
Work Place : Al-Masar Company
Address : Sec 910 St 12 H 3
Profile Picture : C:\Users\malak\Desktop\صور شخصيه\images
DNA-Profile :
A new record has been stored successfully with ID : 1000142

D8S1179		D21S11		D7S820		CSF1PO		D3S1358		TH01		D13S317		D16S539	
10	14	29	32.2	10	13	10	11	16	17	9	9.3	12	13	9	9
D2S1338		**D19S433**		**vWA**		**TPOX**		**D18S51**		**AMEL**		**D5S818**		**FGA**	
18	25	14	14.2	17	17	9	10	12	14	X	Y	12	13	20	25

All fields with * are Compulsory

Fig. (6). File with extension of csv Uploading.

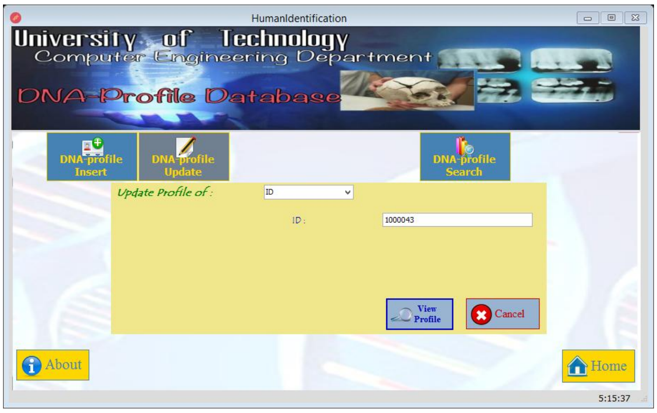


A1 fx D8S1179/Allele1

	J	I	H	G	F	E	D	C	B	A
1	S1358/A	D3S1358/A	CSF1PO/A	CSF1PO/A	D7S820/Al	D7S820/Al	D21S11/Al	D21S11/Al	D8S1179/A	D8S1179/A
2	17	16	11	10	13	10	32.2	29	14	10

Fig. (7). Structure of file with extension of csv

2. *DNA-profile Update*: is used to update the existing records at the database in terms of each field individually. The selecting of the investigated DNA profile for updating is performed by choosing the searching field that can be ID, Full name or Mother name. The obtained record can be modified easily by editing the underlying fields. Fig. (8) show the selection of profile basing on its ID to updating it.



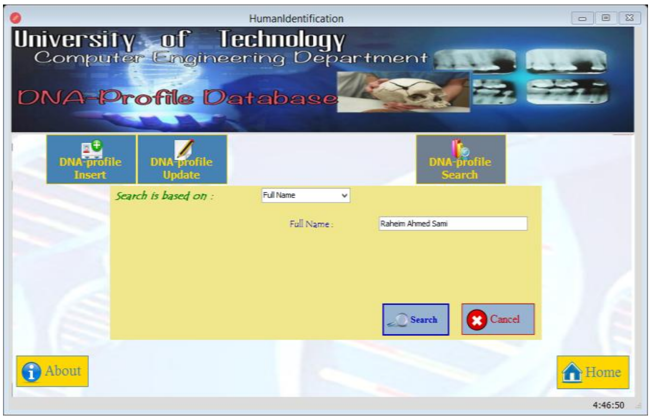
Update Profile of : ID
ID : 1000043

Fig. (8). DNA-Profile Update.

3. At the same manner, searching for any profile can be done by click on "DNA-profile Search" as shown in

ISSN: 2278 – 1323

the Fig. (9). The outcome of the searching is reported as a show at screen or file with pdf extension.



Search is based on : Full Name

Full Name : Raheim Ahmed Sami

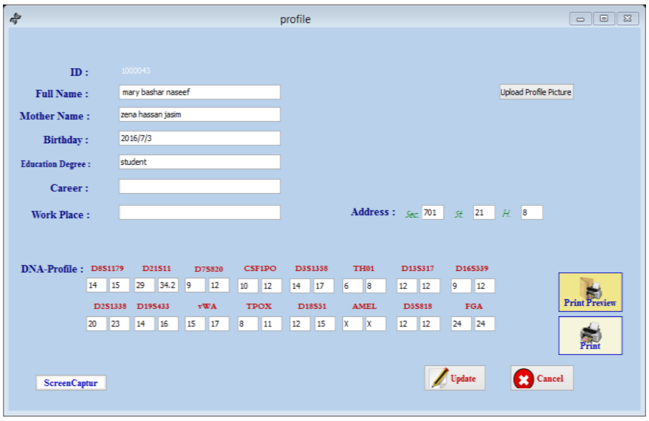
Fig. (9). DNA-Profile Search.

On the other hand, "About" button provides information about the application, whereas "Home" button retunes the user to the home of the application.

III. Results

In order to test the presented DNA profile database system, we entered the 139 DNA profiles for Diyala Province using the DNA-Profile Insert process shown in Fig. (5). The insertion process is done without any errors and with high flexibility.

In terms of updating process, Fig. (10) shown the profile of ID (1000043) that is selected to be updated following the procedure shown in Fig. (8). This DNA profile is now ready for editing any desired field of information regarding the personal and DNA information. In addition, all information shown for each DNA profile can be printed and saved as file with a pdf extension as well as screen capture can be taken.



ID : 1000043

Full Name : mary bashar naseef

Mother Name : zena hassan jasim

Birthday : 2016/7/3

Education Degree : student

Career :

Work Place :

Address : Sec 701 St 21 H 8

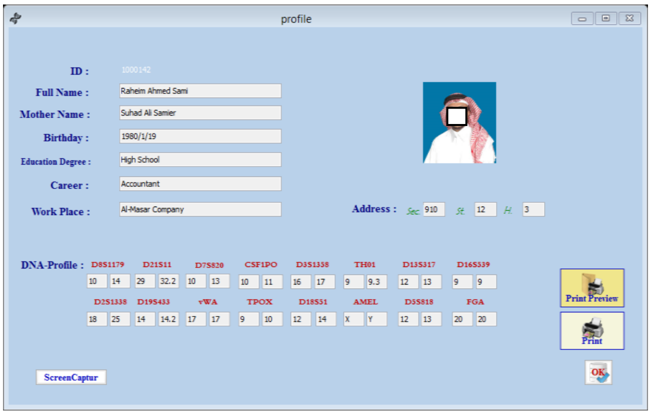
DNA-Profile :

D8S1179		D21S11		D7S820		CSF1PO		D3S1358		TH01		D13S317		D16S539	
14	15	29	34.2	9	12	10	12	14	17	6	8	12	12	9	12
D2S1338		**D19S433**		**vWA**		**TPOX**		**D18S51**		**AMEL**		**D5S818**		**FGA**	
20	23	14	16	15	17	8	11	12	15	X	X	12	12	24	24

Fig. (10). Output of Update Method.

For the searching process, Fig. (11) shows the profile under the name of (Raheim Ahmed Sami, “fake name”), which is required to be shown following the search steps explained in Fig. (9). It is important to note that the searched DNA profile is inserted to the database using the file record uploading process shown in Fig. (6). All information of the individual can be printed, and saved as file with pdf extension in addition to taking screen capture.

The observed results from the insertion, updating and searching methods proved that the proposed database system is able to handle a huge amount of data in an efficient and flexible manner with less time consuming. Moreover, a high degree of security can be set by dedicating the user name and password with the connection string of the SQL server. This is due to the using of SQL-SERVER 2012 software environment.



ID : 1000142

Full Name : Raheim Ahmed Sami

Mother Name : Suhad Ali Samier

Birthday : 1980/1/19

Education Degree : High School

Career : Accountant

Work Place : Al-Masar Company

Address : Sec 910 St 12 H 3

DNA-Profile :

D8S1179		D21S11		D7S820		CSF1PO		D3S1358		TH01		D13S317		D16S539	
10	14	29	32.2	10	13	10	11	16	17	9	9.3	12	13	9	9
D2S1338		**D19S433**		**vWA**		**TPOX**		**D18S51**		**AMEL**		**D5S818**		**FGA**	
18	25	14	14.2	17	17	9	10	12	14	X	Y	12	13	20	20

Fig. (11). Output of searching method.

References

[1] AL-Zubaidi, M.M., Al-Awadi, S.J., Namaa, D.S., Saleh, T.Y., Shehab, M.J., Hameed, S.N, 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, Vol. 5, No. 3, 2014, pp.131-135.

[2] Forensic DNA Testing System/STR, http://www.forensicdnacenter.com.

[3] Butler, J.M., "Review: Genetics and Genomics of Core STR Loci Used in Human Identity Testing", J. Forensic Sci, in press (Mar 2006 issue).

[4] Kashyap A., Kumar, A., and Awadhanam S., "Investigating contributors of the mixed DNA samples by forensic Bioinformatics; Uncertainty to certainty for Crime laboratories", Helix, Vol. 1, No. 2, 2012, pp.101-108.

[5] Chakraborty, R., Stivers, D.N., Su, B., Zhong, Y., and Budowle, B., "The utility of short tandem repeat loci beyond human identification: Implications for development of new DNA typing systems", Electrophoresis, Vol. 20, 1999, pp.1682-1696.

[6] Kayser, M., and Schneider, P.M., "DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations", Forensic Sci. Int. Genet. Vol. 3, No. 3, 2019, pp.154-161.

[7] Kayser, M., and Knijff, P.D., "Improving human forensics through advances in genetics, genomics and molecular biology", NATURE REVIEWS | Genetics, Vol. 12, 2011.

[8] Stranger, B.E., Stahl, E.A., and Raj, T., "Progress and promise of genome-wide association studies for human complex trait genetics", Genetics Vol. 187, No. 2, 2011, pp.367–383.

[9] M. Kayser, "Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes", Forensic Sci. Int. Genet., Vol. 18, 2015,pp.33-48, http://dx.doi.org/10.1016/j.fsigen.2015.02.003.

[10] Walsh, S., Liu, F., Ballantyne, K.N., Oven, M.V., Lao, O., Kayser, M., "IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye color in the absence of ancestry information", Forensic Sci. Int. Genet. Vol. 5, No. 3, 2011, pp.170–180.

[11] Liu F., Duijn, K.V., Vingerling, J.R., Hofman, A., Uitterlinden, A.G., Janssens, A.C., Kayser, M., "Eye color and the prediction of complex phenotypes from genotypes", Current Biology Vol. 19, No. 5, 2009, pp.192–193.

[12] Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., Kayser, M., "The HIrisPlex system for simultaneous prediction of hair and eye color from DNA", Forensic Sci. Int. Genet. Vol. 7, 2013, pp.98–115.

[13] Panneeerchelvam, S., Norazmi, M.N., "Forensic DNA profiling and database", The Malaysian Journal of Medical Sciences, Vol. 10, No. 2, 2003, pp.20-26.
[14] LEE J., "Oracle vs. MySQL vs. SQL Server: A Comparison of Popular RDBMS", https://blog.udemy.com/.
[15] "creating_DB_SQL", http://www.boosla.com.
[16] "SQL_Serv_Man_Studio", http://www.boosla.com.
[17] Dr. Abdulamir A. karim, “Improved Approach to Iris Normalization for iris Recognition System”, Eng. & Tech. Journal, Vol.33,Part (B), No.2, pp 213-221, 2015.
[18] Nahla Alwan, “Developing a Database System for the Laboratory Tests”, Eng. & Tech. Journal, Vol. 31, Part (A), No.18, pp 52-67, 2013.