INTEGRATED BIOMETRIC DNA IDENTIFICATION SYSTEM

Mahmood K. Ibrahem[1], Alaa S. Sadiq[2]
[1,2]Department of information and Communication Engineering,
College of Information Engineering, Al-Nahrain University, Iraq
mahmood_khalel@yahoo.com[1], alaa_angel_87@yahoo.com[2]
(Received: 16/01/2018; Accepted: 25/02/2018)

***Abstract*-In this paper an integrated biometric identification system based on STR/DNA profiles database has been proposed, each profile composed of 20 autosomal loci and Amelogenin. The objective of this research paper is to design and implement a search engine in such a way that enhance the speediness of searching for matches for Target Profile (TP) regardless, the profile is partial or optimum. This is accomplished by utilizing the alleles distribution nature for the population of interest. In this research more than 15 million profiles were simulated to test the system reliability, accuracy, and speed. The result shows that the proposed system is improving the speed of matching about 72% and all tested profiles were identified successfully.**

Keywords: **Database search, DNA-profile, Filtering, Power of Discrimination, STR loci.**

I. INTRODUCTION

In recent years, DNA databases have significantly affected the system of criminal justice. Serial crimes have successfully been solved and connected. Cases without incipient suspects have been brought to resolution. The innocence of wrongfully incarcerated individuals has been checked when post-conviction evidence has matched another offender [1].

The reason for the efficiency of databases is that the dominant part of the crimes is perpetrated by repeated offenders. Studies have demonstrated that over than 60% of those people is imprisoned for violent crimes who thereafter released are re-captured for an analogous crime in lesser than three years. Significant serial crimes can be linked and their criminals can be ceased via matching biological evidence between offenders and crime scenes. Thus, DNA databases effectiveness increases as their size get larger [1].

Traditionally, DNA analysis requires an entire day to be completed, which makes it challenging to quickly create shortlists of suspects behind crimes. However, the Japanese company Nippon Electric Company (NEC) has introduced new portable DNA analyzer can complete DNA analysis at the crime scene within only 25 minutes [2].

In this research paper we are presenting the design and the implementation of the integrated biometric identification system based on STR/DNA database. This system has designed with a search engine that increases the speediness of matching process of any profile by utilizing the alleles distribution nature for the population of interest. The system can detect degraded profiles such as profiles which encountered drop out or partial profiles. In this system crime scene profile could be searched against suspects profile and crime scene profile could be searched against other crime scene profiles. There are many studies were presented in this field. For instance, [3] and [4], have presented methods that speed the search of STR profiles in database. In [3], the author developed a method to form and preserve a tree-structured index to search multidimensional data utilizing naturally occurring patterns and clusters within the data. Thus, the search and retrieval strategies for a profile in a database implemented effectively. The author generated, in his study, a set of 10000 DNA/STR profiles based on US Caucasians allele frequencies for the 16-loci. Also, Multivariate Statistical analysis is employed to analyze the resulted allele distribution, specifically; the Principal

Component Analysis (PCA) technique is used to reveal clustering patterns amongst the profiles. The analysis demonstrates that by selecting of some loci-pairs, such as D16S539 and D13S17, distinct and good clusters were attainable. In [4], authors were proposed a DNA biometrics system using a database of 16 STR loci which has been constructed using SQL SERVER DBMS. They implanted their study of matching by firstly, filtering DNA profile records based on a match of no less than one allele of the D2S1338 locus which has a higher heterozygosity compared to the other locus. The results showed the efficient use of the system resources besides to the reduction of the time consumption. Secondly, matching records inside the resulting list.

These studies does not take partial profiles into consideration. They are based on specific loci and if one of these loci is partial then the intended profile may not be detected.

In 2013 [5], proposed a DNA biometrics recognition method by converting the letters of DNA sequence into its ASCII code and then training the set through back propagation neural network and save the output network. The matching process done by input a sample in the same network and compare the output with the output of previously enrolled dataset. The author pointed out that the recognition rate has reached 81%.

Other studies [6], [7] and [8] were presented in utilizing the DNA biometrics for security authentication purposes, [6] proposed a secure system that boost the ATM security. In this scenario, the data of DNA samples is digitized and turned into barcode through barcode generator which attached on the user's ATM card. This system can be used to identify the right ATM card owner via capturing fingerprint by using fingerprint scanner and scanning the DNA barcode. Moreover, [7] a study has proposed DNA-based Saliva Security System (D-SSS) for individual authentication which is a new approach to extract DNA pattern and pattern matching. The D-SSS requires very little amount of saliva to efficiently preserve the system security. The authors demonstrate that the proposed DNA acquired form saliva can be utilized in real life system security. Moreover, DNA extracted from Saliva provided user-friendliness in addition to high-level of authentication [7].

Furthermore [8], proposed a method that utilize the DNA profile to improve document security through attaching it to a birth record thus binding the individual to its Birth Certificate (BC). The genotypes are randomly arranged and concatenated with extra context information to create a Profile Data Unit (PDU) which is cryptographically hashed afterwards digitally signed. The author stated that the BC issuer will combine the protected PDU with its digital signature inside an issued BC. Also, Public-key certificates will be combined with BC content to enable the verifier to authenticate the BC's contents.

II. EXPERIMENTAL WORK

A. Software Used

Database was constructed using Oracle 11g. All comparisons, calculations, and rankings were performed using PL/SQL programing language. User's interfaces were built using Microsoft Visual Studio.net 2017 framework, C# programming language.

B. Database

The database of the proposed system composed of two main parts:

First part contains profiles of suspects, offenders, and known victims. Those profiles were simulated based on frequencies of population database for Iraqi Arab population [9]. This process was accomplished as follow:
Firstly, genotype probability is calculated for each possible genotype in Iraqi Arab population. Secondly for each certain genotype generate a number of it so that the number of profiles contain that genotype to the total number of simulated profiles is proportion to that genotype probability. Thirdly shuffle the generated genotypes to increase randomness. This part also includes other information such as case ID, sample ID, full name, email, phone number, address, sample type, sample collecting date, and city. The number of simulated profiles 14000000 profiles.

Second part contains profiles which represent unidentified profiles like unknown victims and stains found at the crime scene. As a simulated data about 1000000 profiles are randomly selected and copied from the first part. Some of them left as they are to represent full profiles. Other profiles are treated to encounter drop out by removing one or more allele from various loci. Partial profiles are also simulated via removing both alleles from one or more loci to test the system capability on identifying partial profiles. Also, this part contains other information such as sample type, sample or body storage place, sample collecting date and sample ID.

C. Formula Used

DNA profile frequency estimation or Random Match Probability (RMP) is calculated by firstly, calculating the genotype probability for all loci. Secondly, utilizing the product rule for multiplying the frequencies together [10]. Nevertheless, Likelihood Ratio (LR) could be obtained from RMP using the expression 1/RMP. In our system, for genotype probability calculation of heterozygotes locus we use National Research Council (NRC) II Recommendation 4.10b [10], since it includes population substructure adjustments for heterozygotes as follow:

$$P_i = \frac{2[p_{i1}(1-\theta)+\theta][p_{i2}(1-\theta)+\theta]}{(1+\theta)(1+2\theta)} \qquad A_{i1} \neq A_{i2} \tag{1}$$

Where θ is the correction factor, A_{i1} and A_{i2} alleles at locus i, P_i genotype probability of it, and p_{i1}, p_{i2} represent frequencies of A_{i1}, A_{i2} respectively.

For homozygotes loci there is may be a probability of drop out included so we use Buckleton, & Triggs [11] formula to calculate LR since it considering probability of dropout Pr(D) and subpopulation correction as shown in table I) [11].

TABLE I

LR FORMULAS INCLUDE DROP OUT PROBABILITY AND SUBPOPULATION CORRECTION

Stain	Suspect	formula
A_iA_i	A_iA_i	$\mathrm{LR} \approx \frac{(1+\theta)(1+2\theta)}{\{2\theta+(1-\theta)p_i\}[3\theta+(1-\theta)[p_i+2\Pr(D)(1-p_i)]]}$ (2)
	A_iA_j	$\mathrm{LR} \approx \frac{(1+\theta)(1+2\theta)\Pr(D)}{\{\theta+(1-\theta)p_i\}[2\theta+2\Pr(D)+(1-2\Pr(D))(1-\theta)p_i]}$ (3)

Eq. 2 and 3 are requiring estimation value for Pr(D) and the assessment of this probability may be difficult. Therefore, some laboratories may prefer the simpler formulations as illustrated in table II) [11]. So, Eq. 4 and 5 are used in the proposed system when Pr(D) is not available.

TABLE II

LR ADJUSTED FOR THE SUBPOPULATION EFFECT USING THE SUBPOPULATION CORRECTION

Stain	Suspect	formula
A_iA_i	A_iA_i	$\mathrm{LR}\approx \frac{(1+\theta)(1+2\theta)}{\{2\theta+(1-\theta)p_i\}[3\theta+(1-\theta)[2-p_i]]}$ (4)
	A_iA_j	$\mathrm{LR}\approx \frac{(1+\theta)(1+2\theta)\Pr(D)}{\{\theta+(1-\theta)p_i\}[4\theta+(1-\theta)(2-p_i)]}$ (5)

III. PROPOSED SYSTEM COMPONENTS

The proposed system is composed, as shown in Fig. 1, of the following components:

a. Population Data Management: in this section user can submit statistical parameters for each locus such as Power of Discrimination (PD) and add its alleles and edit their associated frequencies values. The submitted information should have been approved from specialized institutions and have been published. The importance of this section is that the other components of the system is highly depending on the information submitted here. Also, system has flexibility to accept population data of any population.

b. Cases Management: this section composed of three parts, the first part involves adding, editing, and deleting details of each case. Second part manages profiles of offenders, suspects and identified victims. Third part manages TPs including evidences samples, human remains and unidentified victims. Population Data mentioned in (a) should be set before submitting profiles in this section.

c. Matching: which involves a comparison of unknown samples to a database of offenders; suspects and identified victims also involves comparison with other evidences from other cases to link crimes.

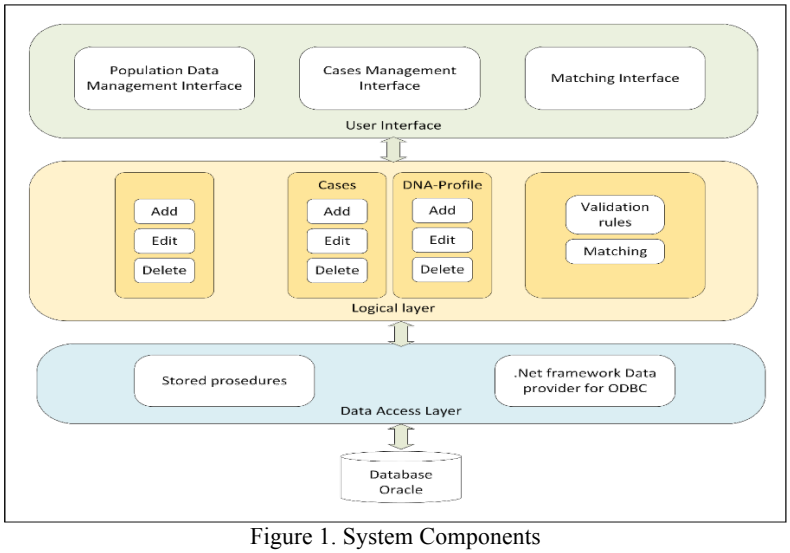


Figure 1. System Components

IV. PROPOSED MATCHING ALGORITHM

Instead of searching all records in database, the proposed system uses some parameters to filter it. Hence, the searching time is reduced, these parameters are:

1) Searching Level (SL): user can search the database at national level or specific city level.

2) Minimum Shared Alleles (MSA): this parameter is set by the user, where Profiles that share less than MSA with TP will be excluded. There are four options for this parameter (10, 20, 30 and 40), when MSA is set to 40 then the returning is the full match profile.

3) Power of Discrimination (PD): this parameter affects the process of selecting the locus that should be used in filtering process. The locus with high PD has higher priority than others. Fig. 2 show the PD for each locus [9].

4) Filtering Locus (FL): affects on the speed of search since records filtering based on the locus with higher PD in the concerned population data.

Fig. 3 illustrates the flow chart of the proposed matching algorithm.

LR becomes 0 if both alleles for the specific locus are heterozygous and no or just one allele is shared at that locus between TP and Reference Profile (RP). Since there is no possibility that TP and RP could be originated from the same source, even if there is dropping out possibility. Otherwise equations (1-5) are used based on the availability or absence of Pr(D), the number of matched alleles for the intended locus between TP and RP, and whether the intended locus is homozygous or heterozygous. Fig. 4 illustrates the flow chart of LR calculation.

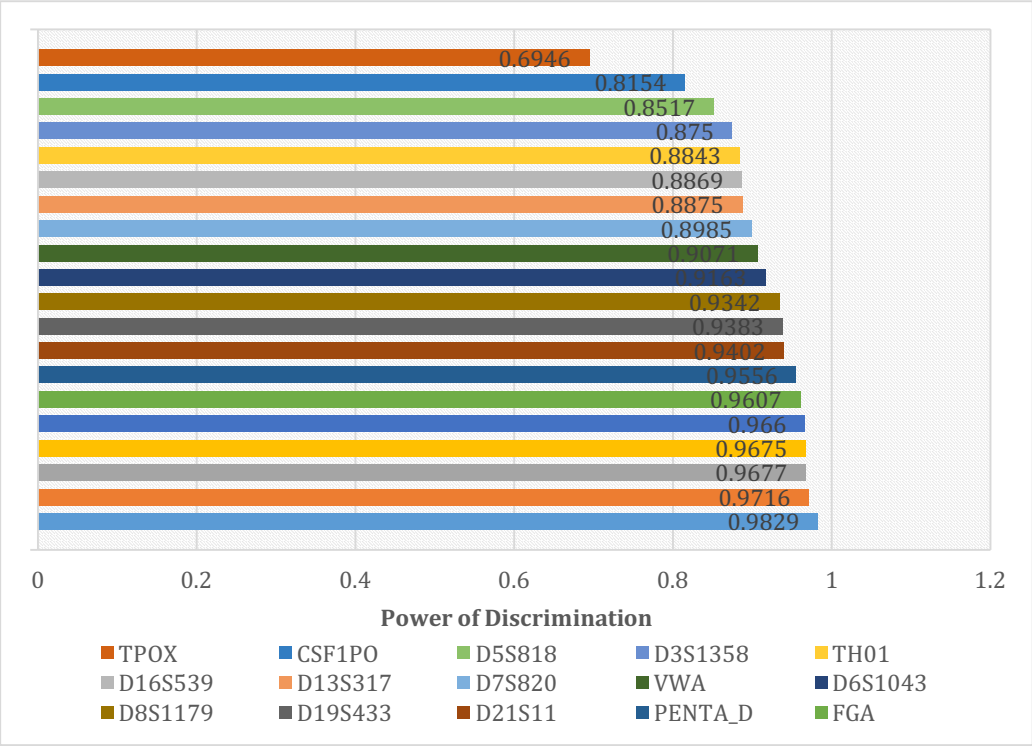


Figure 2. PD at Each Locus in Iraqi Population

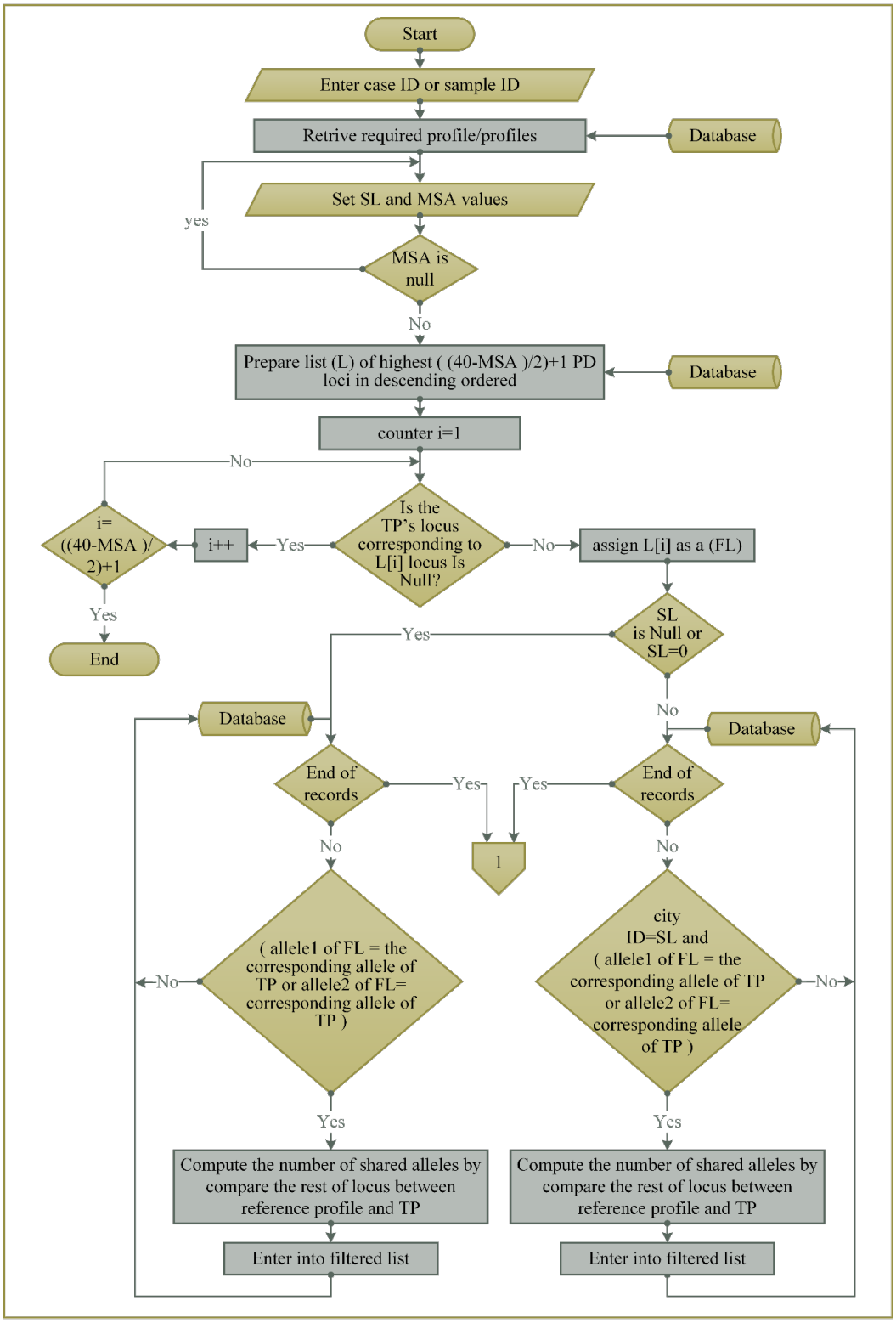


Figure 3. Flowchart of matching process

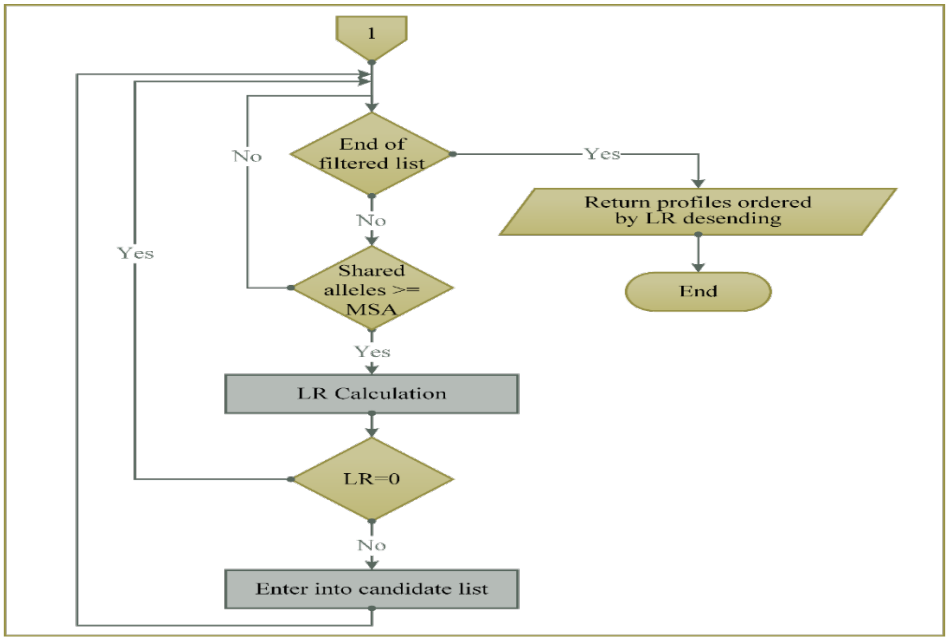


Figure 3. Continued

The maximum value of i is end with the length of the list of possible FLs.

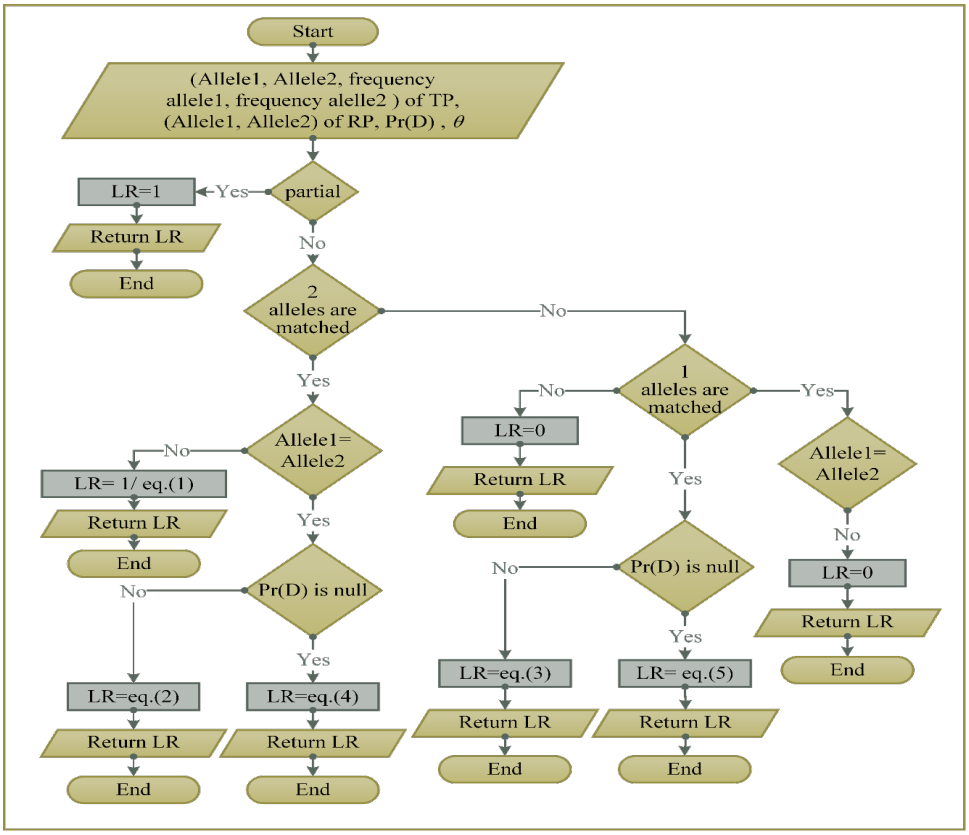


Figure 4. Flow chart of LR calculation

V. RESULTS AND DISCUSSION

The system functionality was tested using both real and simulated data. The insertion, editing and identification functions were evaluated once using real data. This process was accomplished inside the Institute of Forensic Medicine in Baghdad due to the privacy issues. All tested samples have accurately identified with positive results. The other test where number of records was simulated to test reliability of the system and prove that the principle of filtering is based on locus with highest PD. In this research, thirty random samples were selected for testing every time a parameter is changing, or when a database size is changing. The results showed that all the tested samples were identified successfully.

Since there is no previous study used 20 loci profiles in term of increase the speed of search, thus we made a comparison between the proposed system and conventional search method that involves searching all records in the searching pool.

Fig. 5 shows the execution time in case of using the proposed algorithm denoted as "with filter". the other case involves the search of all records that denoted as "without using filter". This figure is also demonstrating how the proposed algorithm is faster using different MSA parameter values. This comparison is performed using simulated database of 15 million records and the FL is PENTA E since all profiles are optimum.

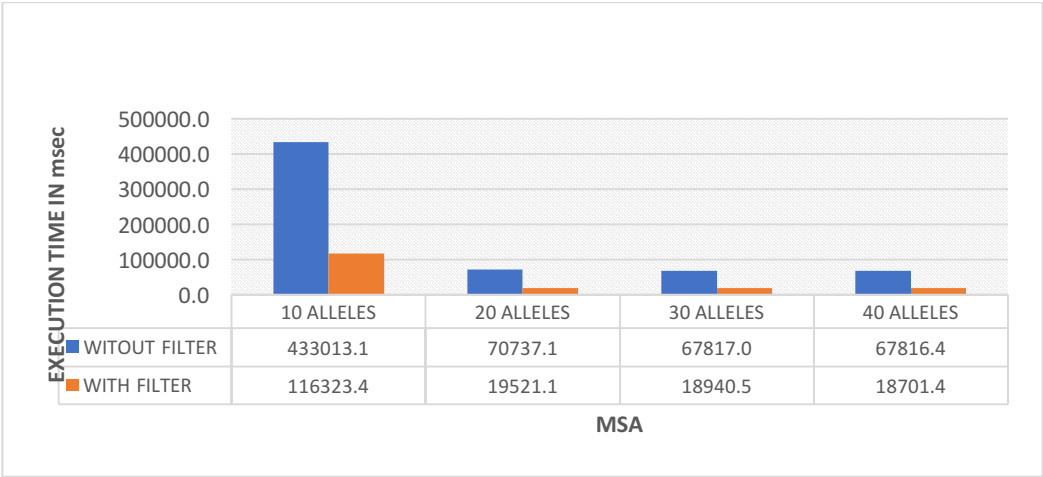
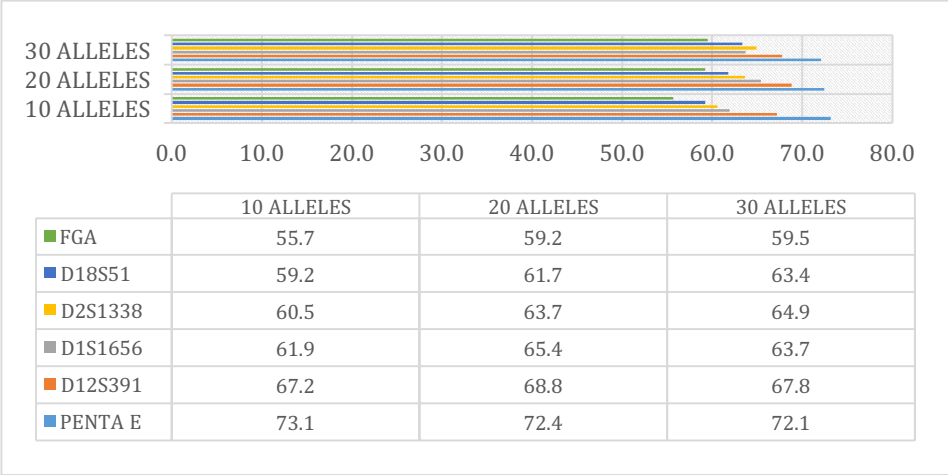


Figure 5. Execution time using database size 15 million record

Fig. 6 shows the average improvement of processing time using six different FL. Each FL tested using three different values of MSA. Firstly, to make the PENTA E as a FL, then PENTA E locus in the tested profiles should contain data, secondly to make D12S391 as FL, then we need tested profiles to be partial in PENTA E and D12S391 contains data, thirdly to make D1S1656 as FL, then we need tested profile to be partial in both PENTA E and D12S391, besides D1S1656 contains data, and so on.

The results proved that the best locus to be selected as FL is proportional to PD of that population, see Fig. 6. Fig. 7 shows the average improvement for each locus at different sizes of database. This test accomplished using all possible MSA values, this figure indicates that our simulated genotypes is randomly distributed though database. All

matching tests was performed in national SL, θ set to zero, Pr(D) is null. System was tested in both local and network environments using windows 10 operating system, Core M-5Y71 CPU 1.20GHz, and 8GB RAM.



	10 ALLELES	20 ALLELES	30 ALLELES
FGA	55.7	59.2	59.5
D18S51	59.2	61.7	63.4
D2S1338	60.5	63.7	64.9
D1S1656	61.9	65.4	63.7
D12S391	67.2	68.8	67.8
PENTA E	73.1	72.4	72.1

Figure 6. Average improvement

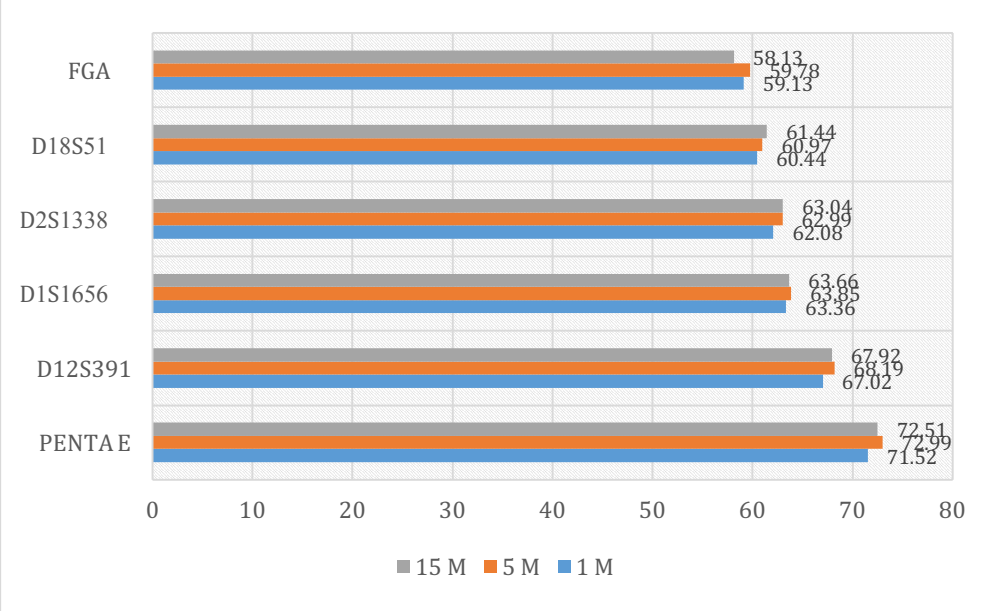


Figure 7. Average improvement of matching process at different loci and database sizes

VI. CONCLUSIONS

To sum up, the process of matching DNA profiles within large databases could be improved via the utilizing of the nature distribution of alleles in the desired population. In addition, the PD assists in selecting the FL process that eliminates the searching time. If the profile is partial at the locus with higher PD then there will be an alternative FL and a profile can be detected. Even when we change parameters such as MSA and database size the improvement percentage still stable. The system is suitable for any population data that entered by a user since the FL is selected at run time.

REFERENCES

[1] J. M. Butler, *Advanced topics in forensic DNA typing: methodology*: Academic Press, 2012.
[2] NEC. *Portable DNA Analyzer*. (December 22, 2017). http://in.nec.com/en_IN/products/public-safety-security/product/portable-dna-analyzer.html
[3] P. Yadav, "Increasing the Speed and Efficiency of Search in FBI/CODIS DNA Database Through Multivariate Statistical Clustering Approach and Development of a Similarity Ranking Scheme," Master's Thesis, University of Tennessee, 2001.
[4] S. D. Khudhur and M. S. Croock, "Biometrics System based Human Identification using STR DNA Marker," *Biometrics,* vol. 138, 2016.
[5] I. A. Saleh, "Software Performance Evaluation Biometric Personal Identification Based on Hybrid Intelligence Technique," PhD., College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq, 2013.
[6] B. Prasanthi, U. P. Jyothi, B. Sridevi, and T. V. Krishna, "Security Enhancement of ATM System with Fingerprint and DNA Data," *International Journal of Advanced Research in Computer Science and Software Engineering,* 2014.
[7] R. Radha, A. J. Blesswin, and G. S. Mary, "A Simple Innovative Approach DNA-based Saliva Security System for User Authentication," *Indian Journal of Science and Technology,* vol. 9, 2016.
[8] M. Tannian, C. Schweikert, and Y. Liu, "Securing Birth Certificate Documents with DNA Profiles," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.
[9] M. M. Farhan, S. Hadi, A. Iyengar, and W. Goodwin, "Population genetic data for 20 autosomal STR loci in an Iraqi Arab population: Application to the identification of human remains," *Forensic Science International: Genetics,* vol. 25, pp. e10-e11, 2016.
[10] J. M. Butler, *Advanced topics in forensic DNA typing: interpretation*. USA: Academic Press, 2014.
[11] J. Buckleton and C. Triggs, "Is the 2p rule always conservative?," *Forensic science international,* vol. 159, pp. 206-209, 2006.