DNA Profiling Methods and Tools: A Review

Emad Alamoudi[1(✉)], Rashid Mehmood[2], Aiiad Albeshri[1], and Takashi Gojobori[3]

[1] Department of Computer Science, Faculty of Computing and Information Technology (FCIT), King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
ealamoodi0003@stu.kau.edu.sa, aaalbeshri@kau.edu.sa
[2] High-Performance Computing Center, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
RMehmood@kau.edu.sa
[3] Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
Takashi.Gojobori@kaust.edu.sa

Abstract. DNA typing or profiling is a widely used practice in various forensic laboratories, used, for example, in sexual assault cases when the source of DNA mixture can combine different individuals such as the victim, the criminal, and the victim's partner. DNA typing is considered one of the hardest problem in the forensic science domain, and it is an active area of research. The computational complexity of DNA typing increases significantly with the number of unknowns in the mixture. Different methods have been developed and implemented to address this problem. However, its computational complexity has been the major deterring factor holding its advancements and applications. In this paper, we review DNA profiling methods and tools with a particular focus on their computational performance and accuracy. Faster interpretations of DNA mixtures with a large number of unknowns and higher accuracies are expected to open up new frontiers for this area.

Keywords: DNA profiling · Bioinformatics · Forensic science
Likelihood computations · High-performance computing

1 Introduction

According to The American Heritage Medical Dictionary, DNA profiling is "the identification and documentation of the structure of certain regions of a given DNA molecule, used to determine the source of a DNA sample, to determine a child's paternity, to diagnose genetic disorders, or to incriminate or exonerate suspects of a crime [1]." DNA profiling (also named DNA typing, DNA fingerprinting, or DNA testing) which was first introduced in 1985 by Alec Jeffreys has changed the area of forensic science significantly [2]. Dr. Jeffreys has found that there are several regions in the human DNA that contain repeated DNA sequence. He found that these DNA sequence areas may differ from one person to another. Dr. Jeffreys was able to measure

R. Mehmood et al. (Eds.): SCITA 2017, LNICST 224, pp. 216–231, 2018.
https://doi.org/10.1007/978-3-319-94180-6_22

the variation in these DNA sequences by developing a unique identity test called Restriction Fragment Length Polymorphism (RFLP). The repeated DNA areas are called Variable Number of Tandem Repeats (VNTRs).

Today, DNA profiling is helping in many cases to identify an innocent from guilty. Human Identity test can also be used in contexts such as missing people investigation, parentage test, ancestry test, and disaster victim identification.

The DNA typing is considered today to be the most useful tool in the hand of law enforcement. Moreover, computer databases which contain DNA information of criminals which was taken from crime scenes had helped to associate a crime to an offender. Due to having a specific set of Short Tandem Repeat (STR) loci in these massive databases, it is unlikely to see a new set of DNA markers to be introduced shortly [2].

In order for DNA sample to be processed, several steps should be involved [2]. First, obtaining the DNA from a biological source. Second, assessing the amount of DNA recovered. Third, isolating the DNA from its cells and using Polymerase Chain Reaction (PCR), which is a technique for copying specific DNA areas. Finally, the STR alleles which have been generated from the previous step will be examined.

However, many difficulties may occur during the procedure of producing a DNA profile that affects the analysis of a sample. One of these problems is the stochastic effects, which arise during DNA extraction. Other challenges are allele dropout, PCR process, allele sharing, and PCR amplification artifacts. Such difficulties hardened the accurate interpretation of the DNA profile [3].

The result of the DNA sample processing will be compared to other sample or databases to check the similarity. If there is a match or 'inclusion,' this indicates that both samples were taken from the same source. On the other hand, if there is no match, the result would consider as 'exclusion,' which means there is no biological relation between the two samples [2]. A case report will be made by a forensic specialist explaining the result and containing random match probability answering the similarity question.

The Scientific Working Group on DNA Analysis Methods (SWGDAM) advise forensic report to contain a prediction of the number of contributors to the mixture that is under examination [3]. Usually, the number of contributors of a sample that taken from a crime scene is unknown. Therefore, an analyst should estimate it according to the electropherogram obtained. This assumption affects the final weight of DNA evidence [3].

In this paper, we review DNA profiling methods and tools with a particular focus on their computational performance and accuracy. To the best of our knowledge, this is the first review paper on DNA profiling tools. Faster interpretations of DNA mixtures with a large number of unknowns and higher accuracies are expected to open up new frontiers for DNA profiling. In the coming years, the complete genome sequencing technologies in a single or only a few cells will be easily available. These technologies may change the situation of DNA profiling completely. In this case, it is obvious to prepare appropriate statistical methods for that. It will be, therefore, important to prepare the mathematical and statistical algorithms for complete-genome-sequencing-based DNA profile. Emerging computational and big data developments [4], along with

Internet of Things (IoT) [5] and smart society environments [6], will provide opportunities for new services related to DNA profiling.

The rest of the paper is organized as follows. Section 2 describes background concepts related to this paper. Section 3 provides information regarding DNA profiling methods and technologies that are being used to obtain a DNA sample. Section 4 describes a number of approaches that rely upon the calculation of Likelihood ratio to interpret DNA profile. We further discuss the importance of the Number of Contributors (NoC) in profiling a DNA mixture in Sect. 5. Some implementations that estimate the NoC was mentioned in the same section. Section 6 then illustrate notable DNA profiling applications. We conclude and give an outlook for the future of DNA profiling in Sect. 7.

2 Background Material

We now give a brief background of the various concepts and methods related to DNA profiling.

2.1 Forensic Science

Forensic DNA tests had a major influence on the evolution of the criminal justice system. Yet, the advancement of new technologies is enabling forensic labs to expand its capabilities and improved the sensitivity of the DNA interpretation.

Butler [7] thinks that this area would develop in the future in three main areas; DNA technologies will become faster, the sensitivity of extracting relative information will increase, and higher volume of data will be expected due to that sensitive nature. He argued that STR will remain the dominant genetic marker.

According to Butler [7], key challenges in the forensic science field are the subjectivity, inconsistency of the complex DNA mixture interpretations between different laboratories and analysts, and the need for training forensic analyst to enhance interpretation of DNA profiles.

2.2 DNA Mixture

A sample is called a DNA mixture when two or more individuals contribute to it. Under some circumstances, the interpretation of a mixture could be more challenging. Allele sharing is one of the factors that increase the difficulty of interpreting a profile [2].

If we have a two-person mixture, then we expected to observe only four alleles per locus. However, this rule may change if we have alleles overlapping or if we have heterozygous individuals. If we have more than four alleles per locus, then we might deal more than two people mixture [8].

DNA mixtures interpretation is a very demanding task [9]. Perez et al. define the DNA mixtures as when two or more people contribute to the same sample. They added that contributors include victims, perpetrators, or other people who interact with the crime scene. Yet, the mixture can be complex when it became a subject of allele drop-in or/and allele drop-out [10]. A detailed introduction to the DNA analysis on the

forensic science domain was given by [2, 11]. Butler gives a historical overview that explaining the evolution of the area. He also explains the structure of the DNA and its fundamental component. Moreover, how this structure can be different among species which enable us to use it in the identity test. DNA profiling can use in identity tests such as parentage analysis, and disaster victim identification [2].

2.3 Technologies for DNA Profiling

The topic of DNA profiling was improved by the new advances in the technology. Weedn and Foran [11] gave a general overview of the latest updates and challenges in the forensic science domain related to DNA profiling. STR followed by PCR amplification is one of the most used methods that regularly used in forensic labs [11]. Other markers such as Single Nucleotide Polymorphisms (SNP), Y chromosome STRs, and mitochondrial DNA are also considered. Weedn and Foran argued that the forensic DNA typing is the most dominant method in the forensic science laboratory. They mentioned that the forensic test usually performed with taking into consideration the court challenges. Therefore, the forensic science only uses a well-validated procedure, and all the laboratory process should be documented. The protocols should be ready to defend against legal attack.

New technologies had not only increases the quality of profiling the DNA mixture, but also amplified artifacts such as stutter, variabilities, and baseline noise. Monich et al. [12] had introduced a quantitative signal model which forms the variability in a stutter, baseline noise, and allele peak height. They had also applied the chi-squared and Kolmogorov-Smirnov (KS) tests on the true peak heights and noise to test the fitness of various probability distribution classes. They argued that the interpretation of signal measured from a DNA sample used to be accomplished by using thresholding. Nonetheless, using thresholds during DNA analysis yield problem of losing valuable information. For that reason, new methods that don't rely on threshold were developed.

2.4 Factors Increasing the Complexity of DNA Profiles

Different phenomena affect the complexity of interpreting a DNA profile. These factors include: number of contributors, peak heights, stutter, a major peak masking, a stutter peak masking, population, drop-out probability, drop-in probability, analytical threshold. No software had yet considered all these factors in its calculation [13]. Therefore, it is part of the challenges that face people who develop DNA profiler to select which factor to model in their implementation.

2.5 Likelihood Estimator

Likelihood ratio (LR) is the probability comparison between evidence under two propositions [2]. One is called the prosecution hypothesis, which assumes that the DNA collected from a crime scene goes to the suspect, whereas the other is the defendant hypothesis, which assumes that the matches between the suspect and the questioned sample happened coincidentally. The two considered propositions are mutually exclusive.

The Likelihood ratio is calculated by putting the prosecution hypothesis as a numerator while putting the defendant hypothesis as a denominator [2]. The LR equation is:

$$LR = Hp/Hd \tag{1}$$

If we assume that the suspect commits the crime (100% probability), which is the prosecution hypothesis, then $Hp = 1$. Additionally, if the STR typing result is heterozygous, the probability of the defendant hypothesis would be $Hd = 2pq$, where p and q are the occurrences of the allele one and two for a locus in a relevant population [2]. If we have a homozygous STR typing, then the probability of the defendant hypothesis would be $Hd = p^2$. Therefore, the equation would become:

$$LR = Hp/Hd = 1/2pq \tag{2}$$

Butler [2] said that if the final result was greater than one, then this result would support the prosecution side. While if it is less than one, then the defendant theory would be in favor.

Typically, the LR will have higher ratio if the STR genotype is rear because of the reciprocal relationship. LR is the inverse of the locus estimated frequency [2]. Note that the likelihood ratio can be more complex depending on the mixture of the evidence.

The strength of the result of the likelihood ratio in terms of the prosecution's case can be interpreted numerically as presented in Table 1. Column 1 represents the LR value, while Column 2 is showing the corresponding strength of evidence.

Table 1. The strength of evidence according to LR result [2]

Likelihood ratio	Corresponding evidence
1 to 10	Limited support
10 to 100	Moderated support
100 to 1000	Moderated strong support
1000 10,000	Strong support
10,000 or greater	Very strong support

3 DNA Profiling: General Methods

Several methods had been proposed to statistically evaluate a DNA mixture. Likelihood ratio, the combined probability of inclusion/exclusion (CPI/CPE), and a modified random match probability (mRMP) are some examples of these methods [14]. In February 2000, the FBI's DNA Advisory Boar had strongly recommended the first two methods to be used [2]. Moreover, in 2006, the International Society of Forensic Genetics (ISFG) had emphasis the value of likelihood ratio [14]. There are six steps to interpreting a DNA mixture which was first described by Tim Clayton in 1998 [2]. First, we need to identify the existence of a mixture. Second, the Allele peaks should be selected. Third, we need to determine the possible number of contributors. Fourth, an

approximation of the ratio of the people who contribute to the sample. Fifth, we need to calculate all potential genotype combinations. Finally, a reference sample comparison should be made.

In CPI approach, an equal weight is given to all possible genotype combinations. Therefore, a lot of information is being wasted when using this approach which makes it inefficient when working with distinct genotypes [14]. This approach does not require prior knowledge of the number of contributors because it is evaluating all genotypes' combination based on the evidence profile [14].

The Random Match probability (RMP) is usually used with single-source samples; therefore, a modified random match probability (mRMP) is used to refer to the method when it is used with more than single-source sample [14]. Unlike CPI, this approach requires a prior knowledge of the number of contributors in the mixture and will not work well with low-level profiles. An example of two- and three-person mixtures calculations using mRMP was described in [15].

According to Bille et al. LR is the most dominant method of evaluating a DNA mixture. However, both mRMP and LR make use of the available information in the sample where CPI does not tend to do so.

More detailed analysis of the three methods and their advantages and weaknesses can be seen in butler's book "Advanced Topics in Forensic DNA Typing: Interpretation" [14].

4 DNA Profiling Using Likelihood Ratio

LR is considered as the most appropriate and powerful approach for calculating the weight of DNA evidence. There are three methods using LR that are widely described in the literature. The first model is the Binary model, which is the simplest yet it cannot handle complex mixture [16]. Second, the semi-continuous, which is the most used by scientists since it is easy to understand and explain, but it still neglects relevant information [17]. This result in losing information that could be precious. Finally, the continuous which overcomes most of the previous models' shortcomings. It utilizes most of the available information provided by the sample, yet it is harder to be accepted and explained in a courtroom [16]. These models may involve a human or computerized process, depend on of the complexity of the approach. Kelly et al. [17] had made a comparison between these three approaches which are suggested by the DNA Commission of the ISFG.

Many frameworks that interpret complex DNA profiles rely on the likelihood ratios approach such as [10]. Gill and Haned had mentioned set of guidelines which can help to evaluate any complex mixture. In addition, they provide some features for any model that might be dealing with complex interpretation such as the ability to incorporate several contributors. They emphasize the fact that the calculation must be provided in a fast manner.

Most of the likelihood ratio based analysis require the number of contributors be given before to analysis Number of contributors. For instance [18–23] rely on the number of contributors on their analysis.

However, others had tried to avoid using it in their interpretation, such as [24, 25]. Russell et al. had developed a semi-continuous method that can calculate the likelihood ratios without previous knowledge about the contributor's number. Their simple model has the abilities to calculate the statistical weight to inclusions. They had also provided a limit test which will guarantee the absence of any false inclusion by chance. To test the proposed unconstructed likelihood ratio (UCLR) model, researchers had collected a set of DNA mixtures with known contributors in different ratios. The result shows good performance on three people mixture. However, the performance becomes worse as the number of contributors increased.

5 Estimating Number of Contributors for DNA Profiling

Today, most applications that interpreted the DNA profile require the number of contributors to available as an input [24]. Different methods have been developed to conclude the number of contributors in a DNA mixture. One of these methods called Maximum Allele Count (MAC). This approach calculates the minimum number of contributors who might contribute to a sample by counting the observed alleles at each locus. Nevertheless, this method may not be valid to work in a complex mixture because of the complexity of allele-sharing [26]. New methods were proposed that do not only rely on the number of observed alleles, but also on the frequencies of observing the allele in the population. Biedermann et al. [27] had developed a probabilistic method that a perform a Bayesian network to conclude the number of contributors to DNA mixture. The new approach performs better than MAC with degraded DNA sample and a higher number of contributors. Maximum Likelihood Estimator (MLE) is another method used to estimate the number of contributors. It tries to maximize the likelihood value of the DNA profile [28].

Haned et al. [29] had compared MAC and MLE. The efficiency of both methods had been analyzed and compared for identifying two to five-person mixtures. Three different situations used to test both methods. First, when all contributors belong to the same population and when allele occurrences are known. Second, when allele occurrences are not known, which may occur in population subdivision. Finally, a condition of partial profiles and how it could affect the estimation accuracy. MAC method is used to set the lower bound that can clarify the number of alleles in a mixture. Haned et al. believe that MAC is unreliable since there is a chance for allele sharing between people which called the masking effect. The result of the comparison supports the use of MLE when a mixture contains more than three contributors. However, when three or two people contribute to a mixture, MAC would perform better.

However, as the number of contributors increased the risk would increase. Haned et al. [30] Had analyzed the risk of dealing with three-, four-, and five-person mixture. They have done that by comparing the gold standard LR to the casework LR. The gold standard LR is when the number of contributors and genotypes are known which mean the availability of all required information to compute LR per contributor. Authors showed the result and the implied thoughts of analyzing high order mixture in the forensic domain. Haned et al. argued that the low template DNA mixture of three-, four-, and five-person are common in forensic casework, yet it is hard to interpret.

Many methods are used today to evaluate the number of contributors in a sample such as [3, 8, 9, 31]. Perez et al. had created a strategy that could find out the number of contributors from two to four-person mixtures for both low template and high template DNA amounts. The proposed strategy helped to provide a useful tool to differentiate between high and low template two-, three-, and four-person mixtures. The four-person mixtures show some difficulties due to the allele sharing phenomena.

Egeland et al. focus on calculating the number of contributors in a mixture by maximizing the likelihood. The proposed approach is based on single SNP. The method tried to answer two questions: Is it a mixture? And if yes, then how many markers are required and how they should be selected. One of the recommendations that driven from the result was regarding the number of markers needed to calculate the number of contributors which is 100 markers.

A typical algorithm for finding the best allele pair in a locus to interpret a mixture is presented in Algorithm 1. Such a process is essential when calculating the number of contributors in a DNA profile. Moreover, it is a performance bottleneck.

6 Software Tools for DNA Profiling

A number of tools are available that implement various DNA profiling methods. These include DNA MIX [32], Euroformix [18], LRmix [20], LRmix studio [33, 16], TrueAllele [19], LikeLTD [22], LabRetriever [13], CeesIt [21], NOCIt [3], DNAMixture [34], Forensim [35], MixtureCalc, Mixture Analysis [36], FamLink kinship [37], DNA Mixture Separator [38], and STRmix [39]. We will review the most notable of these in this section. We shall describe them here and explain their differences.

6.1 DNAMIX

There were three versions of this software, and all of them are open sources. The third version is the most notable and powerful one among the three, and was based on [32]. This version was written in Java and are appropriate for complex mixtures as well as single-contributor stains. The software will ask for the database, stains, genotype, and hypothesis to be inputted as external files. A simple GUI has been developed in this version.

6.2 LRmix Studio

LRmix Studio is a software designed to interpret complex DNA profiles. It was built on its previous version, which called LRmix; however, LRmix Studio is much faster and more flexible. It can measure the probative value of any (autosomal STR-based) DNA profile [33]. This software is following the semi-continuous model of interpreting DNA profiles. Moreover, it was written in Java, and it is open-source.

Algorithm 1 calculate locus's best allele pair that give best interpretation of the sample

```
1: procedure GeneProbCalc(stepSize, noc, lname, revLoci, forLoci, DNAmass, AlleleAtLoci, LDO)
   //noc=number of contributors, lname=locus name, LDO= Locus Drop Out
2:     locAlleles = AlleleAtLoci[locusname]
3:     MeanAndStd = Meanstd(locusname) //find mean and stddev
4:     for i=0 to stepSize do
5:         g=random array between 0 and 1 with size noc
6:         for j=0 to noc do
7:             for k=0 to 2 do
8:                 r=Generate random number that does not exceed the interval of the locus
9:                 allele = AlleleRange[r] //get the allele in the selected interval for a specific locus
10:                Add allele to Peakscumulative
11:                contMass=g[i-1]*DNAmass
12:                if Rand() < ExpVal(locusName, LDO, contMass) then
13:                    ValidAlleles.add(allele)
14:                    weight[allele] = weight[allele] + contMass
15:                end if
16:            end for
17:        end for
18:        for aName=ValidAlleles.start to ValidAlleles.end do //aName=Allele Name
19:            if locAlleles contains allele then
20:                (mean, variance) =Meanstd(weight[allele]) //find the mean and stddev
21:                if revLoci[lName] && Rand()<ExpVal(lname, RevStutDropO, weight) then
22:                    rMu=ExpVal2(lName, mean, weight[allele]) * allele.height
23:                    rSigma=ExpVal2(lName, Stddev, weight[allele]) * allele.height
24:                    revAlleleStut = aName - 10 // get the reverse
25:                    fowStutPeak = Peakscumulative[allele]
26:                end if
27:                means[revAlleleStut] = means[revAlleleStut] + rMu
28:                variances[revAlleleStut] = variances[revAlleleStut] + rSigma * r
29:                if forLoci[lName] && Rand()>ExpVal(lName, forStutDropO, weight) then
30:                    fMu=ExpVal2(lName, Mean, allele.weight) * allele.height
31:                    fSigma=ExpVal2(lName, Stddev, allele.weight) * allele.height
32:                    fowAlleleStut = aName + 10 // get forward
33:                    fowStutPeak = Peakscumulative[allele]
34:                end if
35:                means[fowAlleleStut] = means[fowAlleleStut] + rMu
36:                variances[fowAlleleStut] = variances[fowAlleleStut] + rSigma * rSigma
37:            end if
38:        end for
39:        for temp=Peakscumulative.start to Peakscumulative.end do
40:            mean.add(temp.allele, MeanAndStd[0])
41:            variances.add(temp.allele, MeanAndStd[1] * MeanAndStd[1])
42:        end for
43:        locusProb=calcLocusPeakHeightsProb(Peakscumulative, means, variances)
44:        Summation+ = locusProb
45:        if locusProb > currMax then
46:            currMax = locusProb
47:            for alleleName=selectedValidAlleles.start to selectedValidAlleles.end do
48:                currMaxAlls.add(alleleName)
49:            end for
50:        end if
51:    end for
52:    result.add(Summation, currMax, currMaxAlls)
53:    Return result
54: end procedure
```

Algorithm 1: A typical algorithm for calculating locus's best allele pair that gives the best interpretation which helps in finding the number of unknowns in a DNA mixture (algorithm inspired by NOCIt tool [3]).

6.3 TrueAllele

It is a software that computes DNA interpretation automatically. It can infer genetic profiles from all sorts of DNA samples. The software applies the continuous model; however, no open source version of the code is available. It was written in Matlab. Analysis followed by a comparison of TrueAllele is presented on [19] using a real information that has been taken from actual cases.

6.4 LabRetriever

LabRetriever is a free software developed to estimate the likelihood ratios that combine a probability of drop-out. It was built on another software called LikeLTD which was written in R language. Authors rewrote the code using C++ to acquire more speed. The software uses the semi-continuous model. It computes likelihood ratios for up to four unknown contributors to a DNA sample.

6.5 CeesIt

CeesIt is a method that integrates two features of the continuous approach to calculate the LR and its distribution which are conditioned on the defense hypothesis and the linked p-value. It combines stutter, dropout, and noise in its calculation. It uses a single source sample with known genotypes. It calculates the LR for a selected POI on a questioned sample, together with the p-value and LR distribution. The software was Witten in Java and is available in a (.jar) format. A deep analysis of the software was presented on [21].

6.6 LikeLTD

LikeLTD is a software that used to computing likelihoods for DNA profile evidence, including complex mixtures. It has been written in R. However, since the fifth version, the computation-intensive areas in code have been rewritten in C to be executed in parallel. This software applies the continuous model of calculating the Likelihood ratio. These areas include the computation of genotype combinations for unknown contributors, computing allele doses for each genotype combination, dose adjustments for relatedness, heterozygosity, dropout, and power.

The runtime of the Peak height model is much slower than the runtime of the discrete model, yet it yields a higher evidence weight (see Table 2). The time complexity of the peak height model scales up with the number of unknown contributors, the number of observed peaks, and the number of replicates in the profile. Other parameters that increase the runtime are the modeling double-stutter or over-stutter. A parallelism was achieved on the C++ code by using shared memory parallelism (OpenMP).

Table 2. The runtime of calculating the Weight of Evidence (WoE) using both the two different models for the laboratory case [18]

Hypothesis	Model	WOE	Runtime (minutes)
Q/X + K1 + U1	Discrete	2.3	14
	Peak height	8.2	23
Q/X + U1 + U2	Discrete	0.5	38
	Peak height	7.8	200

The runtime of the algorithms was recorded using node with eight Intel Core I7 processors (3.1 Hgz per core) and with 15 Gb of RAM. The result is presented in Table 2. The first column describes the hypothesis that was applied. Two hypotheses were used. Q is a contributor to the crime scene profile under the Hp while X is the unknown individual under Hd that assumes to contribute to the profile instead of Q. The hypotheses may specify the number of K is representing the known contributors whereas U is the unknown contributors. The second column indicates the used model whether it uses discrete or peak height. The last two columns are showing the weight of evidence and the corresponding running time.

6.7 DNAMixture

DNAMixture is a statistical model that calculates and analyze DNA sample for one or more contributors [34]. This software has been written in R and follows the "fully continuous" statistical model. Its author claims to develop all methodology within his framework for consistent analysis and transparency. However, the application does not have a graphical user interface, which requires a basic experience with R. The parallelism was applied onto the Bayesian network package that used by this software which called "Hugin." In Hugin package, they used Pthread in the C code.

6.8 EuroForMix

EuroForMix is a software based on the fully continuous approach to estimate STR DNA profiles from a complex DNA sample of contributors with artifacts. It is available as an open source. EuroForMix was written in R language. Nonetheless, the likelihood function was written in C++. The software added a parallel processing, since the 0.5.0 version, using snow R-package. The parallel implementation will only be considered when a number of unknowns are at least 3 (not performed yet for database searching or non-contributor simulation). A number of processes will be similar to the number of random start points required in the optimization.

Euroformix requires a significant amount of computational time when the number of unknown contributors is four or more. Table 3 gives an approximation time complexity for each number of unknown contributors. From the table, it was clear that the time consumed when we have four unknown contributors was too much. A good idea would be to parallelize the code over distributed memory system to reduce that time. The runtime is given in Table 3. Column 1 is showing the number of contributors while Column 2 gives the corresponding time taken.

Table 3. An approximate overview of the time taken to calculate the LR depend on the number of unknown contributors [40]

Number of unknown contributors	Runtime
1	~1 s
2	~1 min
3	~30 min
4	~24 h

6.9 NOCIt

NOCIt [3] analyzes the DNA sample to calculate the number of contributors in a mixture. Java programming language was used to write the software. It determines the number of contributors (from 1 to 5). NOCIt can only interpret an autosomal STRs data which are independent of each other. Moreover, the software is not developed to deal with a stutter.

The execution time of [3] depends on the maximum number of contributors, the number of loci/alleles considered and the processing speed of the computer. It is also dependent on whether multiple runs of NOCIt are occurring at the same time, i.e., two NOCIt interfaces are open at once and running two separate samples. Table 4 provides the runtime of NOCIt. The first column gives the number of contributors, whereas the second column describes the range of time taken to analysis that number. The result was collected from a dual-core laptop with Intel® CoreTM i5-3380 CPU @ 2.9 GHz.

Table 4. The runtime using a different maximum number of contributors [25]

Number of contributors	Range of time (mode)
1	<1 min (0.2 min)
2	15 min–30 min (17 min)
3	30 min–1.5 h (1 h)
4	1 h–5 h (4 h)
5	5 h–20 h (14 h)

6.10 STRmix

STRmix is a probabilistic genotyping application which performs the continuous model of interpreting the DNA profile. It was built to interpret single and mixed DNA profiles. Moreover, it follows the SWGDAM recommendations. It utilizes information that extracts from a DNA sample, such as peak height, to calculate the probability of a DNA profile for all possible genotype combinations. The software considers aspects such as allele drop-in, allele dropout, and stutter. The software has been written in Java, and it's only available for purchase.

6.11 A Comparison of the DNA Profiling Tools

Table 5 compares different software that we had reviewed in this section. The first column gives the names of the software. Columns 2–8 provide information about various features of the software. Column 2 gives information whether the software has a GUI or not. Columns 3 and 4 are illustrating if the selected software considers the phenomena of drop-in and stutter on its interpretation or not. Column 5 gives the model that used to the calculation of LR. The sixth column describes the programming language that used to build the selected software. Column 7 indicates the availability of source code. The last column describes the used parallel framework. The table is missing some information due to either the lack of resource for some software or because of the inability to access the software's source code.

Table 5. A general comparison between the review softwares

	GUI	Drop-in	Stutter	Calculation model	Language	Source code	Parallelism
LRmix studio [13, 16, 33]	Yes	Yes	–	Semi continuous	Java	Yes	Thread pool
TrueAllele [18, 19]	Yes	Yes	Yes	Continuous	Matlab	No	–
DNAMIX V.3 [32]	Yes	–	–	–	Java	Yes	No
Euroformix [18]	Yes	Yes	Yes	Continuous	R, C ++	Yes	Snow package
CeesIt [21]	Yes	Yes	Yes	Continuous	Java	No	Thread pool
NOCIt [3]	Yes	Yes	Yes	Continuous	Java	No	Thread pool
DNAMixtures [34]	No	Yes	Yes	Continuous	R	Yes	No
LikeLTD [22]	No	Yes	Yes	Continuous	R, C	Yes	OpenMP
LabRetriever [13]	Yes	Yes	–	Semi continuous	C ++	Yes	No
STRmix [23, 39]	Yes	Yes	Yes	Continuous	Java	No	–

7 Conclusion

Interpreting DNA mixture is a common practice in forensic science domain. It is a complicated process that requires an extended period of time. We gave an overview of the field of DNA profiling. A historical background, along with its application was mentioned. We, then, discuss the needed steps to sample a DNA mixture and what are the required technologies. After that, we reviewed the literature based on their classification into describing DNA profiling in general. We focus later on approaches that follow the Likelihood Ratio model. We also reviewed the various tools and compared their performance and accuracy

In the end, we would suggest the use of Euroformix and LikeLTD for DNA profiling since they are already performing parallelism. They both utilize most of the available information in the DNA sample because they follow the continuous model for calculating the LR value. The source code for the two software is available for assessment and modification. However, Euroformix provides a GUI which gives it a slight advantage over LikeLTD for users who have no technological expertise.

A frequent necessity to apply these tests might raise the need to speed up the run time of such analysis. The computational complexity has been the major deterring factor holding the area advancements and applications. An improvement would give a chance to interpret mixtures with a larger number of unknowns and within a shorter timeframe. The investigation of the relevant literature reveals that the current approaches for parallelization of DNA profiling rely on shared memory parallelization. A distributed implementation is needed to speed up the computations allowing for the use of a large number of cores/processors. This is our ongoing research, which will be reported in the near future. Faster interpretations of DNA mixtures with a large number

of unknowns and higher accuracies are expected to open up new frontiers for DNA profiling.

In the coming years, the complete genome sequencing technologies in a single or only a few cells will be easily available. These technologies may change the situation of DNA profiling completely. In this case, it is obvious to prepare appropriate statistical methods for that. It will be, therefore, important to prepare the mathematical and statistical algorithms for complete-genome-sequencing-based DNA profile. High performance computing (HPC) will play a key role in speeding up DNA profiling methods, particularly those HPC techniques which exploit domain specific data and algorithmic patterns [41], system heterogeneity (e.g. disks for space, and accelerators for speed) for its advantage [42], and virtual organization models (similar to grids [43]) for information sharing across organizational boundaries. Hierarchical system structures will be needed to localize and optimize data and computations [44]. Internet of Things (IoT) would be integrated in smart city systems to create innovative services [6] and deal with big data-related challenges [5]. Mobile, fog and cloud computing [4, 45, 46] will enable dynamic system environments, seamlessly connecting users and systems.

Acknowledgments. The work carried out in this paper is supported by the HPC Center at the King Abdulaziz University.

References

1. The American Heritage Medical Dictionary. Houghton Mifflin Co., Boston (2007)
2. Butler, J.M.: Fundamentals of Forensic DNA Typing. Academic Press/Elsevier (2010)
3. Swaminathan, H., Grgicak, C.M., Medard, M., Lun, D.S.: NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. Forensic Sci. Int. Genet. **16**, 172–180 (2015)
4. Arfat, Y., Aqib, M., Mehmood, R., Albeshri, A., Katib, I., Albogami, N., Alzahrani, A.: Enabling smarter societies through mobile big data fogs and clouds. Procedia Comput. Sci. **109**, 1128–1133 (2017)
5. Alam, F., Mehmood, R., Katib, I., Albogami, N.N., Albeshri, A.: Data fusion and IoT for smart ubiquitous environments: a survey. IEEE Access. **5**, 9533–9554 (2017)
6. Mehmood, R., Alam, F., Albogami, N.N., Katib, I., Albeshri, A., Altowaijri, S.M.: UTiLearn: a personalised ubiquitous teaching and learning system for smart societies. IEEE Access. **5**, 2615–2635 (2017)
7. Butler, J.M.: The future of forensic DNA analysis. Philos. Trans. R. Soc. Lond. B Biol. Sci. **370**, 577–579 (2015)
8. Paoletti, D.R., Krane, D.E., Raymer, M.L., Doom, T.E.: Inferring the number of contributors to mixed DNA profiles. IEEE/ACM Trans. Comput. Biol. Bioinforma. **9**, 113–122 (2012)
9. Perez, J., Mitchell, A.A., Ducasse, N., Tamariz, J., Caragine, T.: Estimating the number of contributors to two-, three-, and four-person mixtures containing DNA in high template and low template amounts. Croat. Med. J. **52**, 314–326 (2011)
10. Gill, P., Haned, H.: A new methodological framework to interpret complex DNA profiles using likelihood ratios. Forensic Sci. Int. Genet. **7**, 251–263 (2013)

11. Weedn, V.W., Foran, D.R.: Forensic DNA typing. In: Leonard, D.G.B. (ed.) Molecular Pathology in Clinical Practice, pp. 793–810. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-19674-9_54
12. Monich, U.J., Grgicak, C., Cadambe, V., Wu, J.Y., Wellner, G., Duffy, K., Medard, M.: A signal model for forensic DNA mixtures. In: 2014 48th Asilomar Conference on Signals, Systems and Computers, pp. 429–433. IEEE (2014)
13. Inman, K., Rudin, N., Cheng, K., Robinson, C., Kirschner, A., Inman-Semerau, L., Lohmueller, K.E.: Lab retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles. BMC Bioinf. **16**, 298 (2015)
14. Butler, J.M.: Advanced Topics in Forensic DNA Typing: Interpretation. Academic Press (2014)
15. Bille, T., Bright, J.-A., Buckleton, J.: Application of random match probability calculations to mixed STR profiles. J. Forensic Sci. **58**, 474–485 (2013)
16. Garofano, P., Caneparo, D., D'Amico, G., Vincenti, M., Alladio, E.: An alternative application of the consensus method to DNA typing interpretation for Low Template-DNA mixtures. Forensic Sci. Int. Genet. Suppl. Ser. **5**, e422–e424 (2015)
17. Kelly, H., Bright, J.-A., Buckleton, J.S., Curran, J.M.: A comparison of statistical models for the analysis of complex forensic DNA profiles. Sci. Justice **54**, 66–70 (2014)
18. Bleka, Ø., Storvik, G., Gill, P.: EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. Forensic Sci. Int. Genet. **21**, 35–44 (2016)
19. Perlin, M.W., Dormer, K., Hornyak, J., Schiermeier-Wood, L., Greenspoon, S.: TrueAllele casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases. PLoS ONE **9**, e92837 (2014)
20. Gill, P., Haned, H., Eduardoff, M., Santos, C., Phillips, C., Parson, W.: The open-source software LRmix can be used to analyse SNP mixtures. Forensic Sci. Int. Genet. Suppl., Ser (2015)
21. Swaminathan, H., Garg, A., Grgicak, C.M., Medard, M., Lun, D.S.: CEESIt: A computational tool for the interpretation of STR mixtures. Forensic Sci. Int. Genet. **22**, 149–160 (2016)
22. Balding, D.J., Steele, C.: The likeLTD software: an illustrative analysis, explanation of the model, results of performance tests and version history. UCL Genet. Inst. **1**, 1–49 (2014)
23. Moretti, T.R., Just, R.S., Kehl, S.C., Willis, L.E., Buckleton, J.S., Bright, J.-A., Taylor, D. A., Onorato, A.J.: Internal validation of STRmix^{TM} for the interpretation of single source and mixed DNA profiles. Forensic Sci. Int. Genet. **29**, 126–144 (2017)
24. Taylor, D., Bright, J.-A., Buckleton, J.: Interpreting forensic DNA profiling evidence without specifying the number of contributors. Forensic Sci. Int. Genet. **13**, 269–280 (2014)
25. Russell, D., Christensen, W., Lindsey, T.: A simple unconstrained semi-continuous model for calculating likelihood ratios for complex DNA mixtures. Forensic Sci. Int. Genet. Suppl. Ser. **5**, e37–e38 (2015)
26. Paoletti, D.R., Doom, T.E., Krane, C.M., Raymer, M.L., Krane, D.E.: Empirical analysis of the STR profiles resulting from conceptual mixtures. J. Forensic Sci. **50**, JFS2004475-6 (2005)
27. Biedermann, A., Bozza, S., Konis, K., Taroni, F.: Inference about the number of contributors to a DNA mixture: comparative analyses of a Bayesian network approach and the maximum allele count method. Forensic Sci. Int. Genet. **6**, 689–696 (2012)
28. Haned, H., Pène, L., Sauvage, F., Pontier, D.: The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture. Forensic Sci. Int. Genet. **5**, 281–284 (2011)

29. Haned, H., Pène, L., Lobry, J.R., Dufour, A.B., Pontier, D.: Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. **56**, 23–28 (2011)
30. Haned, H., Benschop, C.C.G., Gill, P.D., Sijen, T.: Complex DNA mixture analysis in a forensic context: evaluating the probative value using a likelihood ratio model. Forensic Sci. Int. Genet. **16**, 17–25 (2015)
31. Egeland, T., Dalen, I., Mostad, P.F.: Estimating the number of contributors to a DNA profile. Int. J. Legal Med. **117**, 271–275 (2003)
32. Curran, J.M., Triggs, C.M., Buckleton, J., Weir, B.S.: Interpreting DNA mixtures in structured populations. J. Forensic Sci. **44**, 987–995 (1999)
33. Haned, H., De Jong, J.: LRmix Studio 2.1 user manual (2016)
34. Lauritzen, S.L.: Statistical and computational methodology for the analysis of forensic DNA mixtures with artefacts (2014)
35. Haned, H.: Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics. Forensic Sci. Int. Genet. **5**, 265–268 (2011)
36. Gill, P., Sparkes, R., Pinchin, R., Clayton, T., Whitaker, J., Buckleton, J.: Interpreting simple STR mixtures using allele peak areas. Forensic Sci. Int. **91**, 41–53 (1998)
37. Kling, D., Egeland, T., Tillmar, A.O.: FamLink – A user friendly software for linkage calculations in family genetics. Forensic Sci. Int. Genet. **6**, 616–620 (2012)
38. Tvedebrink, T., Eriksen, P.S., Mogensen, H.S., Morling, N.: Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures. J. R. Stat. Soc. Ser. C (Appl. Stat.) **59**, 855–874 (2010)
39. Developmental validation of STRmixTM: Expert software for the interpretation of forensic DNA profiles. Forensic Sci. Int. Genet. **23**, 226–239 (2016)
40. Bleka, Ø.: An introduction to EuroForMix (v1.8) **2016**, 1–59 (2016)
41. Mehmood, R., Crowcroft, J.: Parallel Iterative solution method for Large Sparse Linear Equation Systems, vol. 22 (2005)
42. Mehmood, R.: Serial disk-based analysis of large stochastic models. In: Baier, C., Haverkort, Boudewijn R., Hermanns, H., Katoen, J.-P., Siegle, M. (eds.) Validation of Stochastic Systems. LNCS, vol. 2925, pp. 230–255. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24611-4_7
43. Altowaijri, S., Mehmood, R., Williams, J.: A quantitative model of grid systems performance in healthcare organisations. In: 2010 International Conference on Intelligent Systems, Modelling and Simulation, pp. 431–436. IEEE (2010)
44. Mehmood, R., Crowcroft, J., Hand, S., Smith, S.: Grid-level computing needs pervasive debugging. In: The 6th IEEE/ACM International Workshop on Grid Computing, 2005, p. 8. IEEE (2005)
45. Tawalbeh, L.A., Mehmood, R., Benkhlifa, E., Song, H.: Mobile cloud computing model and big data analysis for healthcare applications. IEEE Access. **4**, 6171–6180 (2016)
46. Tawalbeh, L.A., Bakhader, W., Mehmood, R., Song, H.: Cloudlet-based mobile cloud computing for healthcare applications. In: 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1–6. IEEE (2016)