A Review on Y-Chromosomal based DNA Profiling and Bayesian Networks for Crime Evidence Investigations in Forensic Labs

Swinky Arora
Assistant Professor
Chandigarh University,
Gharuan

ABSTRACT
The field of Bioinformatics has extended the scope of its applications in various fields like genetic engineering, phylogenies, protein synthesis, gene expressions and many more. A Promising application of Bioinformatics is in the field of Forensic DNA analysis for crime evidence investigations. DNA Profiling or DNA Typing is used in Forensic Labs for investigating the evidences of crimes like homicide, murder, rape and in mass destruction people identifications based on the DNA samples collected from the crime or disaster scenes. DNA analysis employs extremely sensitive PCR-based techniques to analyze biological material. Criminals and victims can be linked to crime scenes, or one crime scene to another, using DNA evidences collected from very small components like the saliva on a cigarette butt, skin cells on a steering wheel, cheeks swabs or pet hairs on clothing[1]. The aim of this article is to focus on the steps followed to carry out DNA typing and to explain the Y-STR Profiles for DNA sample identification along with Bayesian networks for statistical analysis of evidences. Y-STR Profiles will focus on the Y- Chromosomal structure of DNA and Bayesian networks will provide a probability or likelihood of the evidence collected from victim and the suspect.

General Terms
DNA Analysis, Crime evidences, Homicide, Chromosomes, loci

Keywords
DNA Profiling, Bayesian Networks, Y-STR Analysis, PCR

1. INTRODUCTION
The term Forensic DNA analysis was first introduced in 1985 and actually implemented in 1986 for identifying the suspect in a case of homicide of two young girls [9]. The initial DNA analysis was done using whole genome recognition where the whole genome was sampled, amplified and tested for culprit identification. But with advancements in the technologies, the searches get limited. The analysis of only a small portion of DNA has revolutionized the forensic investigations[11]. It has been proven that 99.5% of DNA is similar in all human beings except in case of monozygotic twins. The difference that uniquely identifies a person lies in the remaining 0.5%. These differences occur in non-coding regions that do not code for any genes. DNA Profiling techniques focus on these 0.5% loci of DNA and identify the person uniquely [6]. Forensic DNA Profiling requires the use of such techniques that detects the genetic variations among humans. The techniques known so far include analysis of Single Nucleotide Polymorphism (SNP), Variable Number Tandem Repeats (VNTR), Short Term Tandem Repeats (STR), haplotypes and Y Chromosomal based STRs.

The brief introduction of DNA Profiling techniques is as follows:

1. **Whole Genome Sequencing:** This technique amplifies DNA in an unbiased manner that results in greater quantities of DNA that can be analyzed. This technique has good discrimination power.
2. **VNTR analysis:** This technique finds a location in genome where a short nucleotide sequence is organized as a tandem repeat that shows variations in the number of repeats among individuals. This technique is carried out using RFLP (Restriction Fragment Length Polymorphism) analysis and gel electrophoresis.
3. **STR analysis:** This technique finds short repeating sequences on DNA loci of 2-5 base pairs long. The number of repeats can range from 3-51. The analysis is performed on nucleus of DNA which is amplified using Polymerase Chain Reaction (PCR) and then resolved by gel electrophoresis. 13 core STR loci have been decided by the US Government on the basis of which an individual's genetic profile can be generated.
4. **SNP analysis:** This technique finds that locus on DNA where only a single nucleotide varies on the DNA strand with which the genome becomes bi-allelic. This method is carried out using PCR technique and gel electrophoresis as that in STR but with a little difference that here only a single nucleotide is considered that exist in polymorphic state [13].
5. **Y-STR analysis:** This technique finds its usefulness in crime cases where mostly men are involved because this analysis targets on Y- chromosome that is present in males. This gives specific information about the male counterpart when a crime involves samples of male- female mixtures as in case of rape. This Profiling is helpful even in the presence of multiple folds of female DNA. The frequencies of these alleles can be determined using Bayesian inferences [12].
6. **Mitochondrial Analysis:** The nucleus of the cell contains only 2-3 copies of DNA and if that portion is found contaminated then the mitochondrial part of cell help the forensic experts. The mitochondrial

part provides energy to the individual and is inherited from the mother [14]. It contains multiple copies of the DNA and thus can be analyzed in case when nucleus of the cell is not helpful.

The following table summarizes the scope, advantages and limitations of the DNA profiling techniques [1]:

Table 1.1 DNA Profiling Techniques

S.No.	Technique	Scope of Analysis	Advantage	Limitation
1	Whole Genome Sequencing	Whole DNA segment	Least chance of DNA similarity	Complex
2	VNTR Analysis	Variable Repeatable sequences in some portion of DNA	Show variations in length among individuals, show pattern of bands unique to each individual	Very long and variable number of base pairs has to be examined.
3	STR Analysis	Repeating sequences of 2-5 base pairs of DNA on a specific loci	High discrimination power.	DNA segments on all 23 pairs of chromosomes are examined.
4	SNP Analysis	A change in single nucleotide in a DNA segment	Usable for very degraded DNA	Due to bi-allelic nature, they provide low discrimination power.
5	Y-STRs Analysis	Repeat sequences on Y-Chromosome only.	Useful in cases where males are the culprits like rapes	Lack high power of discrimination.
6	Mitochondrial Analysis	Focus on mt-DNA present in the cell	Useful in degraded samples where nucleus DNA cannot be retrieved easily	Heteroplasmy, Possible population structure problems

This article will focus on the Y-STR technique for identifying the culprit based on the profiling of the DNA collected from the crime scene. After profiling the DNA, we will focus on the strategic mathematical method to analyze the likelihood of the correctness of the evidence from which the profiling is done using Bayesian Networks (as a strategic mathematical model) to determine the probability of the correctness of the evidence. Bayesian Networks defines a posterior probability equation form which the unknown parameters are observed form the known parameters. In the next sections of this article we will study Y-STR analysis and Bayesian networks in detail.

2. Y-STR ANALYSIS

The human DNA is segmented into 23 pairs of chromosomes that contain DNA segments and genetic information. Out of these 23 pairs, 22 pairs are autosomal chromosomes or homologues chromosomes and one pair is gender chromosome which identifies the gender of the individual. XX identifies a female and XY identifies a male. Thus, the Y-Chromosome is specific to males.

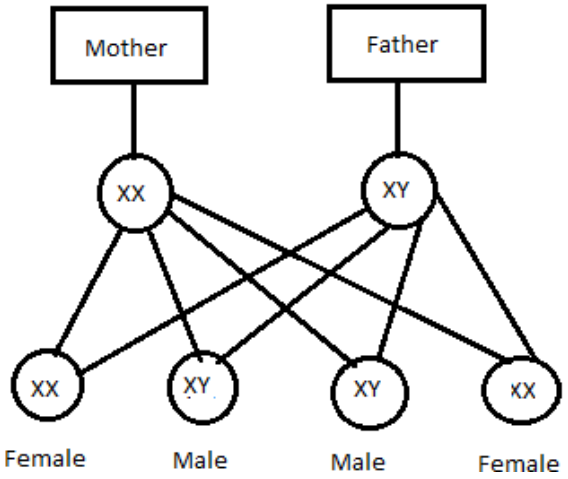

Fig 1.1 Chromosome combinations during meiosis

During DNA typing, only one strand of chromosome is typed because the other strand is known due to complementary base pairing. Consider the following example,

Male 1: GTA***TGGTGGTGGTGGTGG***CTGACGT

Male 2: ATC***TGGTGGTGG***ACTGTACTGCTGAC

Male 1 has TGG repeating STR with frequency of 5 and Male 2 has the same repeating STR but with frequency of 3. These two alleles are analyzed and differences between them are recorded. The STR profiles are called haplotypes.

Apart from STR markers, SNP markers can also be recorded on the Y- chromosome. The main difference is that SNP marker consists of a single nucleotide difference. The SNP profiles are called haplogroups. Consider the following example,

Male 1: GTA***T***GTCATGCTGATCGTAGC

Male 2: GTA***C***GTCATGCTGATCGTAGC

It depends on the forensic expert to either analyze the STR markers or the SNP marker on the Y- Chromosome. This article discusses the Y- Chromosome analysis using STR markers due to their good discrimination power.

For Y-STR analysis, we will consider the DNA sample from different individuals or a single individual and then perform the following steps [4]:

1. **Collection:** A mixture of the sample is taken from the crime scene and then reference samples are taken from the various persons that are considered culprits by the police team. The bottleneck in this step is the problem faced by the experts when the sample collected is found to be contaminated. Suitable measures are thus required to be followed to ensure that the sample collected is free from any contamination.
2. **Extraction and Amplification:** After the sample is collected, the DNA is observed under microscope to find the STR markers. From the observation of DNA it can be found that the sample collected belongs to male or female with the help of Amelogenin markers. Once identifying the gender, the next step involves finding the Y- chromosome in case of male and autosomal chromosomes in case of females. STR markers are identified on the Y Chromosomes and are then amplified using PCR reaction.

   PCR (Polymerase Chain Reaction) is enzyme driven amplification process that confines the DNA search to our area of interest by amplifying only those portions of DNA which are under the interest of our consideration. The process involves around 30 temperature cycles that amplifies the target DNA regions by doubling the copy of DNA in every cycle. Thus the STR markers identified on the Y-chromosome are amplified for next step using PCR.
3. **Separation and Detection using Peak Profiling:** The DNA fragments amplified are detected and separated using electrophoresis where the amplified DNA is injected into a gel and electric current is applied that separates the different alleles according to their frequency or repeat numbers and then a graph is formed that describes the loci and the amount of DNA on each allele. This peak profile information is unique to each individual and thus can be helpful in finding the identity of the person or matching the culprit's identity if the match is found between the reference sample and the collected sample [5].

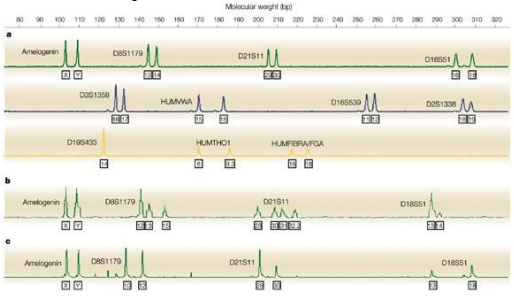

Fig. 2.1 Peak Profile of different individuals

At present, the cutting edge in DNA identification is obtained by matching 13-17 of nuclear STR markers of a victim's sample to family references to identify kinship relationship of the victim and then same matching is performed on the reference individuals that are considered as culprit after police investigation. A system of 13 STRs constitute CODIS (Combined DNA Index System) [15] which is a forensic sample database of STR markers used in USA and Canada. However, various countries can have their own databases and systems. Various commercial kits are available that amplifies the STRs like AMPFistr and SGMPlus etc. [10]

The main **advantage** that this analysis provides to the forensic experts is that only a single chromosome is analyzed and amplified for investigation. This type of analysis is useful in cases where a male or group of males is involved. Even if the sample collected is composed of female extracts but this technique is so powerful that male extracts can be easily identified due to the male specific Y- Chromosome consideration.

The **disadvantage** of Y-STR analysis lies in its advantage only i.e. this technique restricts itself to the identification of the male culprits and also the discrimination power is not as much correct as in case of autosomal STRs. But once it is proved that the culprit is a male then this technique is a promising one to be carried out.

Once the DNA sample is analyzed then it becomes important to verify the decision taken i.e. a statistical method is to be followed so that we can be provided with a likelihood of the correctness of the decision. The prosecutors in the court room can challenge the forensic report if a statistical proof of the report is not provided. So, in the next section of this article, Bayesian networks are discussed that provides a probability distribution based on the likelihood or truthfulness of the evidence proof.

3. BAYESIAN NETWORKS

Bayesian networks or Bayes nets are the probabilistic graphical models that basically exist in the form of Directed Acyclic Graph G= (V, E) that represents the knowledge about the unknown domain. Given knowledge about particular evidence, Bayes nets can determine the unknown attributes based on the known evidence. The nodes of the graph represent random variables and the edges represent probabilistic dependencies among the variables that can be determined using statistical methods.[8]

Mathematically, the nodes in the Bayesian networks represent a set of random variables, $\mathbf{R}=R_1,\ldots Ri\ldots Rn$, from a given domain and the directed arcs connecting the vertices $Ri \rightarrow Rj$ represents the conditional dependency between the variables that are quantified by the conditional probability distributions. We use the following notations to define the Bayesian theory [8]:

a: observed data

b: unknown parameter

p (b): marginal probability of b

p (a): marginal probability of a

p (b, a): joint probability of a and b

p (a| b): conditional probability of a given b, called, sampling distribution

p (b| a): conditional probability of b given a, called, posterior distribution

Using Bayes inference, we compute the posterior distribution as follows:

$$p(b|a) = p(b, a) / p(a)$$

Where, p (b, a) is defined as:

$$p(b, a) = p(a|b).p(b)$$

Thus,

$$p(b|a) = p(a|b).\ p(b)\ /p(a)$$

In Forensic Investigations, this inference is used to test several hypotheses about the samples available and then determine the most likely combination of the contributors whose samples have been collected from the crime scene. [10]

We assumed that evidence or a sample found at crime scene may contain DNA combinations of the culprit, victim or any unknown person if the suspect was not actually present on the crime scene. Thus, we are considering the possibility of the presence of an unknown person and thus determine the posterior probability as mentioned below.

The evidence available can have the following genotype state space:

$$E = \{\text{culprit, victim, unknown person}\}$$

And the hypotheses are:

$$H_0: C\&V,\ H_1: U\&V$$

Where, C denotes culprit, V denotes Victim and U denotes Unknown person contributed to the sample.

Now, the likelihood or probability of the correctness of the sample is calculated as:

$$L = P(E|H_0) / P(E|H_1)$$

Where,

$$P(E|H_0) = P(H_0|E).P(H_0)$$

And

$$P(E|H_1) = P(H_1|E).P(H_1)$$

Once the likelihood of the evidence collected has been calculated then in the next step a network model is developed in several levels. At the top most level, the alleles from all the possible contributors are combined together leading to the sample along with the mixing factors between the contributors. As the networks build level by level, the search and the traversal gets more refined and limited.

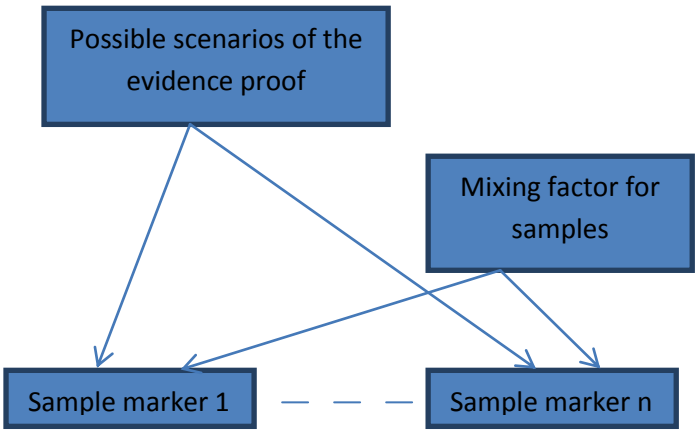

Fig 3.1 Bayesian network for the top most level having contributions from multiple markers

Let us consider a real case, where a body is found on a location and is burnt to so much extent that it is unable to be recognized. The Police supposed that it can be the body of some suspect ***sr*** that they were finding for so long. DNA is available from body, from the known person of the body, let us say, his wife ***wr*** and from two children ***s1*** and ***s2***. The hypotheses node consists of indications that ***sr*** might be identical to the body or he may be treated as unknown person.

Thus for this case, the Bayesian network will be represented as the below figure:

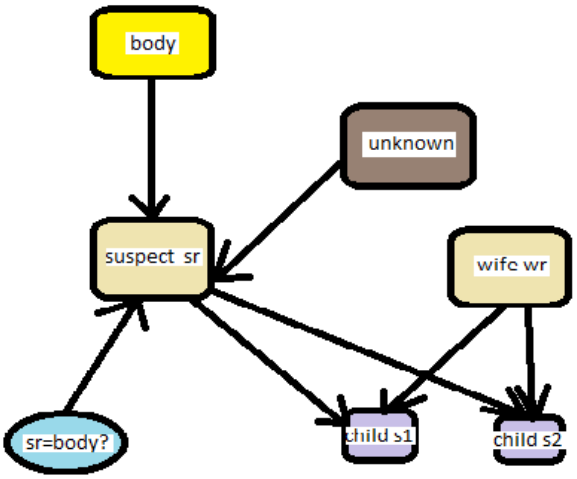

Fig 3.2 Network for suspect identification

Thus, after determining the likelihood of the evidence and the conditional dependency relationship, forensic experts can tell to how much extent the identified person can be considered as a suspect.

Apart from suspect identification, Bayesian inference and networks can also be used to study familial searches based on DNA and kinship relationships between various persons.

The main **advantages** that Bayesian networks provide are that they reduce the confusion as they represent the evidence relationship in a very logical way and generally minimum calculations are required to determine the likelihood of the evidence collected.

The confusion can arise in case of multiple DNA samples but with the hierarchical generation of the networks, the research can be refined in each and every step leading the experts to a more confined solution.

The use of Bayesian networks for forensic evidence calculations is increasing rapidly due to the software packages available in the market like HUGIN, XBAIES, Genie, FINEX etc.

4. PROPOSED SCENARIO FOR FORENSIC LABS

The following points define the steps that can be followed by the forensic experts for crime evidence analysis using Y-STRs and Bayesian networks:

1. Collect DNA sample: The DNA samples are collected from the crime scene either from the blood stains or from the persons who were supposed to be present on the scene at that time. The DNAs collected are first checked for contamination level according to which they are considered perfect for investigation. After the DNA is found to be perfect for analysis then it can be observed under high power microscope.
2. Amplify: The next step is to amplify only that portion of DNA that is to be analyzed. This is done using PCR amplification (already explained in section 2 of this article). The PCR amplification multiplies the DNA to

many folds that make it suitable for correct analysis. Once the DNA is amplified then using Amelogenin markers it is identified whether the DNA belongs to a male or a female.

3. Y-STR analysis: Once the DNA is recognized with a particular gender then the tandem repeats of 2-5 base pair long are checked on the Y-Chromosome in case of male and on the autosomal chromosomes in case of females. If the DNA found is contaminated then we can use mitochondrial part of the cell for better results. The process for Y-STR analysis has already been explained in section 2 of this article.

   Based on this analysis, the different alleles and their frequencies are noted down in the peak height format on the basis of which the individuals are identified uniquely as different individuals have different peak heights and frequencies of the alleles.

   The Forensic labs of various countries have made a database of the various DNA samples like GenBank, CODIS etc. The samples analyzed from the individuals present on the crime scene are matched with those present in the databases to find a match for their community or kinship relations or familial relationships.

4. Posterior distribution using Bayesian Networks: Once the individuals have been identified then the next step is to determine the likelihood of the evidence truthfulness i.e. determine to what extent the match is correct. This is done by first computing the posterior distribution using Bayesian inference and then developing a network that defines the relationship among the various hypotheses we supposed. With a given dimension, the experts can determine the unknown one to see whether the match or the conclusion that they have found using STR analysis is correct up to how much extent. The Bayesian networks can be implemented in MATLAB using Bayesian Network Toolkit (BNT). Software like HUGIN can be very helpful for direct analysis of the probability obtained.[9]

The above mentioned points can be summarized as follows:

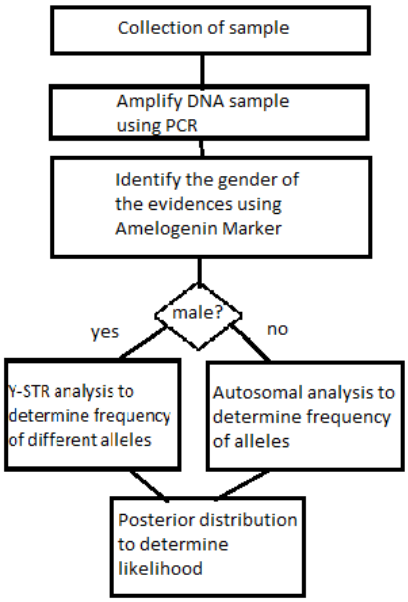

Fig 4.1 Summarized scenario for Evidence Investigation

5. CASE STUDY

Let us now discuss a simple case study to understand what we have presented in the previous sections of this article:

Let us assume that a child is killed by his own father due to some personal issues and the father after being arrested claims that he is not the actual father of the kid and hence is not guilty. The forensic experts will now solve the case to see whether the terms given by the suspected father are true and if not then with how much probability the terms are false.

Suppose the evidence collected is represented by E, the suspect sample by S, M is the sample of child's mother and C is the sample of the victim child.

Prosecutor hypotheses H_p = Suspect is father of child

Defense hypotheses H_d= C is killed by some other man

Evidence index EI is given by:

$$EI = Pr(E|H_p) / Pr(E|H_d)$$

Let gc, gm and gs represents the genotypes of child, mother and suspected father respectively and only one locus of the two alleles is considered for investigation. Thus the above equation now becomes:

$$EI = Pr(gc, gc, gs|H_p) / Pr(gc, gm, gs|H_d)$$

Now, three nodes are required to describe each individual in Bayesian network. First two nodes represent genes of mother and suspect father passed down to the individual. Let the nodes are named as N1 and N2 and the genes of these nodes end with Pg and Mg and the gene of final node ends with Gt. Now, the third node can have three possible values: N1N1, N1N2 and N2N2 with probability of N1 is 0.1 and that of N2 being 0.9.

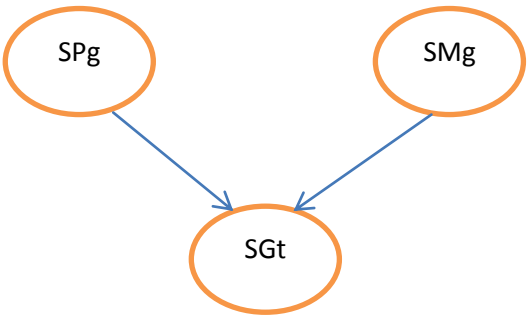

Fig 5.1 Genes of the three individuals

Thus the probability of SGt is calculated as:

Table 5.1 Probability of SGt

SMg	N1		N2	
SPg	N1	N2	N1	N2
N1-N1	1	0	0	0
N1-N2	0	1	1	0
N2-N2	0	0	0	1

The above presented table considers that the suspect is a father of the victim child but it is genuine to consider both the possibilities. So, a hypotheses node is build that returns a true or false result by checking that the suspect is a true father or

not. Thus the previous mentioned Bayesian network in now modified to include the hypotheses node:

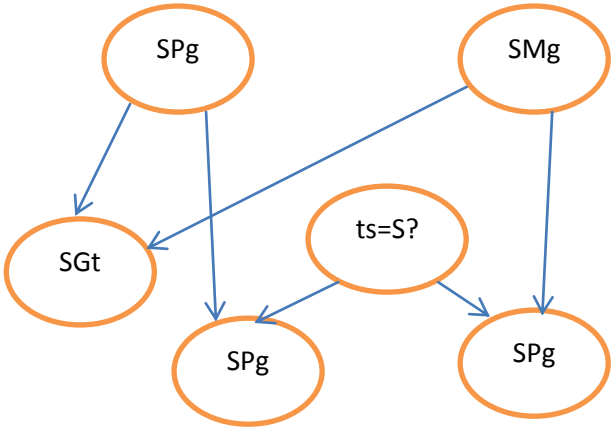

Fig 5.2 Bayesian network considering hypotheses

If the value of the hypothesis is true then the values of tsPg and tsMg are determined by values from SPg and SMg otherwise the probabilities are simply the allele frequencies of respective individuals.

Table 5.2 Probabilities considering hypotheses

Ts=S	Yes		No	
SPg	N1	N2	N1	N2
N1	1	0	0.1	0.1
N2	0	1	0.9	0.9

From the above table, it is clear that the hypotheses considered takes into account the likelihood of the evidence present and thus the suspect can be claimed to be the actual culprit if the allele frequency matches with the hypotheses considered. Otherwise it will be cleared that the suspect is not guilty and is free from any claim.

6. CHALLENGES FACED BY FORENSIC EXPERTS

The scenario presented in this article is quite helpful in forensic evidence investigations but each and every technique has its own limitations that are must to be overcome. The Y-STR analysis and Bayesian networks have to overcome the following mentioned challenges in order to give a perfect evidence investigation:

1. Fake DNA evidences can be implanted on the crime scene i.e. the suspect may leave trace of an innocent person to save himself. So the sample collected must be tested to see that it is an actual sample and the conclusions of the experts should be revised with perfection before reaching a final decision.
2. Artificial DNA is created by some suspects by whom they can save themselves by claiming that they are innocent. Techniques must be designed to effectively distinguish between the original and the artificial DNA sample.
3. Y-STR analysis is difficult in case of contaminated DNA. In that case, purification of DNA is to be carried out. Otherwise, experts can follow mitochondrial DNA analysis that ignores the contaminated nucleus of the DNA.
4. Searching DNA for samples is quite challenging. The process is subjected to errors.
5. Excluding someone on the basis of mt-DNA or Y-STR testing is a different scenario than excluding it on the basis of STR testing of autosomal chromosomes.
6. Y-Chromosome is passed exclusively among males. It may happen that the unrelated individuals may have the same profile.
7. DNA Profiling becomes very difficult in case of monozygotic twins because they have the same profile.
8. Mixtures of DNA can be very complex and difficult to separate and analyze.
9. In case of multiple culprits, the DNA amount collected is so small that it is not possible for the experts to judge even the number of culprits involved.
10. A more sophisticated mathematical model is to be developed that can use a good distribution model for determining the likelihood of the evidences.

7. CONCLUSION AND FUTURE SCOPE

The article started with an introduction to the various DNA Profiling techniques where the scope, advantages and limitations of the various techniques were presented.

From the presented techniques, Y-STR technique was the main focus of this article. Section 2 of the article presented this technique in detail where we saw the steps followed and strengths and weaknesses of the technique.

Section 3 of the article gave an overview of the Bayesian networks where an equation determining the likelihood of the truthfulness of the evidence was defined and based on that equation a network diagram was presented. Bayesian networks were concluded to be a good statistical model used in Forensic analysis.

The proposed scenario from section 4 combined the steps of DNA profiling and Bayesian inferences that can practically be implemented in MATLAB using BNT and PGET toolboxes.

Then the article studied a case study in section 5 as an example of the scenario presented and at last challenges faced by the forensic experts were brought into limelight in section 6.

The article can be concluded with the terms that Y-STR analysis is a useful technique for DNA typing when the culprits are often males and once the alleles are formed from different individuals then the Bayesian networks use allows us to determine the likelihood of the evidence collected.

The author suggests the reader to overcome the challenges that are faced by the forensic experts by improving the techniques to minimize the contamination in DNA sample, to distinguish between actual and fake DNA, to study the Bayes inference for multiple DNA samples and develop a more sophisticated model to get the likelihood of the evidence with more accuracy.

8. ACKNOWLEDGMENT

The author acknowledges Mr. Sanjay Sharma, Professor, Chandigarh University, Gharuan for his valuable suggestions and reviewing of the final manuscript.

9. REFERENCES

[1] MA Jobling, P Gill 2004 Encoded Evidence: DNA in Forensic Analysis Nature Reviews Genetics, nature.com.

[2] E. Giardina, A. Spinella, G. Novelli 2011 Past, Present and Future of Forensic DNA Typing Nanomedicines, futuremedicine.com.

[3] N Hu, B Cong, S Li, C Ma, L Fu, X Zhang 2014 Current Developments in Forensic Interpretation of mixed DNA samples- A Review Biomedical reports, ncbi.nlm.nih.gov.

[4] FR Santos, DR Carvalho-Silva, SDJ Pena 1999 PCR Based profiling of human Y Chromosomes, Springer.

[5] V Keerl 2010 Genotyping of a worldwide panel of raplidly mutating Y-STR, dare.uva.nl.

[6] L Bianchi, Peitro Lio 2007 Forensic DNA and Bioinformatics, oxfordjournals.org.

[7] Bruce Budowle, Angela van Daal 2009 Extracting evidence from forensic DNA analyses: future molecular biology directions, BioTechniques 46:339-350.

[8] Ben Gai I 2007 Bayesian Networks, Wiley and Sons.

[9] Sharif Marya 2012 Statistical issues in modeling the ancestry from Y-Chromosome and Surname data, Phd Thesis, University of Glasgov.

[10] J.J. van Wamelen 2010 Bayesian Networks in Forensic DNA Analysis M.Tech Thesis, University of Laiden.

[11] Gill, P., A. J. Jeffereys and D.J. Werrett 1985 Forensic Application of DNA fingerprints nature 318:577-579.

[12] M. Kayser, M. Nagy 1997 Applications of microsatellite based Y-chromosome haplotyping, Electrophoresis 18:1602-1607.

[13] C.P. Kimpton, P.Gill 1993 Automated DNA Profiling employing multiplex amplification of short tandem repeat loci PCR methods Appl.3:13-22.

[14] M.R. Wilson, J.A. DiZinno 1995 Validation of mitochondrial DNA sequencing for forensic casework analysis Int. J. Legal med. 1-8:68-74.

[15] CODIS-NDIS Statistics. Federal Bureau of investigation (http://www/fbi.gov/hq/lab/codis/clickmap.htm).