DNA Indexing & Searching Human Authentication using STR Locus based on PCR Process

Milap J. Bhuva
Department of Information Technology
G.K.B.I.E. Rajkot, India

Hiren Kamani
Assistant Professor of CE/IT Department
G.K.B.I.E. Rajkot, India

***Abstract*—DNA Indexing & Search Human Authentication System works on bases of DNA profiling. It contains a DNA Data Bank which has DNA Profiles of civilization. DNA Profile stored with DNA STR Locus or DNA sample and necessity person's information. These samples are stored in database by performing algorithmic procedure, it is the major task of the system which is not stored a DNA sample in physically but it stored with digitize unique DNA signature. It also manages the DNA Data Bank and is responsible for developing, providing, and supporting this Program to federal, state, and local crime laboratories in our country and selected international law enforcement crime laboratories to foster the exchange and comparison of forensic DNA evidence from violent crime investigations. The DNA Data Bank Unit also provides administrative management and support to the police or crime investigation departments for various advisory boards, Forensic Science, Department of Justice (DOJ) grant programs, and legislation regarding DNA.**

⇨ Salient features of this system to resolve…

- ✓ **Rape case.**
- ✓ **Murder case.**
- ✓ **Disputed Paternity & Maternity case.**
- ✓ **Child Swapping & Kidnapping case.**
- ✓ **Unknown Dead bodies Identification case.**

(Like Suicide, Dead body exists at any place, Mass disaster-example: Uttarakhand floods disaster, an Artificial accident-example: Bomb blast.)

***Index Terms*— forensic science, DNA Profiling, polymerase chain reaction, mass fatality incidents, mass disaster, software, D3S1358, vWA, FGA,D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, TH01, TPOX, CSF1PO, D16S539**

I. INTRODUCTION

The DNA Indexing & Search Human Authentication system which contains DNA Data Bank– "It is an effective tool for fighting violent crimes" which has contains DNA samples of each and every people. Some people feel that these DNA databases are a great help to society, giving new hope to forensic cases that would otherwise have no leads. Others feel they are an invasion of privacy, and that these databases should consist only of samples collected from violent crime offenders.

DNA Indexing & Search Human Authentication system works with 24 STR locus in which 13 STR locus commonly used in CODIS[1]. In this paper there are 24 STR loci name suggest as CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, D2S1338, D19S433, Penta D, Penta E, D1S1656, D2S441, D10S1248, D12S391, D22S1045, D6S1043, SE33.

Information on 24 commonly used STR Loci present in commercial kits

STR Marker	Position (MB)	Repeat Motif	Allele Range
D1S1656	230.905	TAGA	8-20.3
TPOX	**1.493**	**AATG**	**4-16**
D2S441	68.239	TCTA/TCAA	8-17
D2S1338	218.879	TGCC/TTCC	10-31
D3S1358	**45.582**	**TCTA/TCTG**	**6-26**
FGA	**155.509**	**CTTT/TTCC**	**12.2-51.2**
D5S818	**123.111**	**AGAT**	**4-29**
CSF1PO	**149.455**	**AGAT**	**5-17**
SE33	88.987	AGAT/AGAC	3-49
D6S1043	92.450	GATA	8-25
D7S820	**83.789**	**TCTA/TCTG**	**5-16**
D8S1179	**125.907**	**GGAA**	**6-20**
D10S1248	131.093	TCAT	7-19
TH01	**2.192**	**TCAT**	**3-14**
vWA	**6.093**	**TCTA/TCTG**	**10-25**
D12S391	12.450	AGAT/AGAC	13.27.2
D13S317	**82.692**	**TATC**	**5-17**
Penta E	97.374	AAAGA	5-32
D16S539	**86.386**	**GATA**	**4-17**
D18S51	**60.949**	**AGAA**	**5.3-40**
D19S433	30.416	AAGG/TAGG	5.2-20
D21S11	20.554	TCTA/TCTG	12-43.2
Penta D	45.056	AAAGA	1.1-19
D22S1045	37.536	ATT	7-20

These alleleic 24 STR Locus presents or says the unique address for the person identity which is very useful to identify the person.

II. Y-CHROMOSOME DNA TESTING

The Y-chromosome is found only in males and therefore genetic markers along the Y-chromosome can be specific to the male portion of a male-femaleDNA mixture such as is common in sexual assault cases. Y-chromosome markers can also be useful in missing persons investigations, some paternity testing scenarios, historical investigations, and genetic genealogy due to the fact that most of the Y-chromosome (barring mutation) is passed from father to son without changes.

A core set of Y-chromosome STR (Y-STR) loci is widely used in laboratories worldwidefor human identity testing and genetic genealogy. The minimal haplotype loci(MHL) were selected in the late 1990s from a meager set of available Y-STRs.The MHL include DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393,and the polymorphic, multi-copy marker DYS385[2]. In 2003, the Y-chromosomesubcommittee of the Scientific Working Group on DNA Analysis Methods (SWGDAM)recommended two additional Y-STRs named DYS438 and DYS439[3].

Characteristics of Commonly Used Y-chromosome STR Loci present in commercial kits.

STR Marker	Position (MB)	Repeat Motif	Allele Range
DYS393	3.19	AGAT	8-17
DYS456	4.33	AGAT	13-18
DYS458	7.93	GAAA	14-20
DYS19	10.13	TAGA	10-19
DYS391	12.61	TCTA	6-14
DYS635	12.89	TSTA	17-27
DYS437	12.98	TCTR	13-17
DYS439	13.03	AGAT	8-15
DYS389	13.12	TCTR	9-17/24-34
DYS438	13.38	TTTTC	6-14
DYS390	15.78	TCTR	17-28
GATA-H4	17.25	TAGA	8-13
DYS385	19.26	GAAA	7-28
DYS392	21.04	TAT	6-20
DYS448	22.78	AGAGAT	17-24

III. TRADITIONAL USE OF AVAILABLE METHODS

There are three types of DNA fingerprints: RFLPs, VNTRs, and STRs. **Restriction fragment length polymorphisms**, or **RFLPs** as they are commonly known, were the first type of DNA fingerprinting which came onto the scene in the mid-1980's. RFLP's focus on the size differences of certain genetic locations. "This technique is the DNA equivalent of screening sand through a progressively finer mesh screens to determine particle sizes"[5].

Variable number tandem repeats, or **VNTRs** represent specific locations on a chromosome in which tandem repeats of 9-80 or more bases repeat a different number of times between individuals. These regions of DNA are readily analyzed using the RFLP approach and a probe specific to a VNTR locus[6]. The fragments are a little shorter than RFLPs (about 1-2 kilo base pairs), but are created through the exact same process.

Some of the advantages of these types of DNAfingerprints are that they are the most stable and reproducible, which is a valuable trait tohave when you are trying to determine an exact match of a person's DNA, which mustexclude billions of other people's DNA with a certain degree of confidence. Some of the disadvantages of RFLPs and VNTRs include they are very time consuming (especially the probe hybridization step), relatively large amounts of DNA must be used to obtain an adequate sample, too many polymorphisms may be present for a short probe, and the cost is very high due to labor and time requirements[7]. And also these types of sequence analysis or DNA profiling require fresh blood sample or evidence.

Currently, the most popular method of DNA fingerprinting is **Short tandem repeats**, or **STRs** for short. Unlike VNTRs which analyze minisatellites that have repeat sequences of 9-80 base pairs, STRs use microsatellites which have repeat sequences ofonly 2-5 base pairs, introducing the "less is more" philosophy to the world of DNA fingerprinting. This was a big step forward in forensic science since the length of DNA fragment being analyzed is short enough to be amplified by polymerase chain reaction (PCR)[8], so now we are able to analyze a very small sample of DNA that is quicker and easier than any previously known method and match it to a person's identity. PCR was developed in the mid 1980's and used the same principles that cells use to replicate DNA to amplify the specified region, which is usually between 150-3,000 base pairs in length. In this type of DNA profiling there is no need to require fresh blood sample. It is applying all the available evidence and also for degraded mix blood samples.

Sources of information regarding DNA Samples:

Evidence	Possible Location of DNA on the Evidence	Source of DNA
Baseball bat or similar Weapon	Handle, end	Sweat, skin, blood, tissue
Hat, bandanna, or mask	Inside	Sweat, hair, dandruff
Eyeglasses	Nose or ear pieces, lens	Sweat, skin
Facial tissue, cotton swab	Surface area	Mucus, blood, sweat, semen, ear wax
Dirty laundry	Surface area	Blood, sweat, semen
Toothpick	Tips	Saliva
Used cigarette	Cigarette butt	Saliva
Stamp or envelope	Licked area	Saliva
Tape or Ligature	Inside/outside surface	Skin, Sweat
Bottle, can, or glass	Sides, mouthpiece	Saliva, Sweat
Used Condom	Inside/outside surface	Semen, vaginal or rectal cells
Blanket, pillow, sheet	Surface area	Sweat, hair, semen, urine, Saliva
"Through and through"Bullet	Outside surface	Blood, tissue
Bite mark	Person's skin or clothing	Saliva
Fingernail, partial fingernail	Scrapings	Blood, sweat, tissue

STRs are currently the most popular type of DNA fingerprint, since the whole PCR process takes only a few hours, compared to RFLP/VNTR probe hybridization and film exposure which can take several days. STRs can use much smaller samples of DNA than RFLPs/VNTRs, and can even use partially degraded DNA to create a fingerprint.

Thus, the integrity and quality of the DNA sample is not as great a factor with STRs than with the traditional methods of DNA fingerprinting. Thecurrent standard forensic protocol analyses 13 core STR loci which have been carefully chosen for their uniqueness[9]. The only disadvantage of the STR approach is it is sensitive to contaminating DNA, so usually the STR approach is used first, followed by a VNTR analysis if contamination is suspected, and enough DNA is available.

There are various STR markers kits available: Applied Biosystems(Identifiler, MiniFiler, NGM, NGM SElect), Promega (PowerPlex 16, PowerPlex ESI 17, PowerPlex ESX 17, PowerPlex 18D), and Qiagen (ESSplex, IDplex) kits.[4]

IV. PROTOTYPE

There are various algorithms are analysed for this kind of purpose. By using dynamic programming approach especially in DNA sequencing Needleman-Wunsch[10] algorithm and Smith-waterman algorithms [12] are more complex in finding exact patternmatching algorithm. By this method the worst case complexity is*O(mn)*. The major advantage of this method is flexibility inadapting to different edit distance functions. The Raita algorithm [11] utilizes the same approach as Horspoolalgorithm[13] toobtaining the shift value after an attempt. Instead of comparingeach character in the pattern with the sliding window from right toleft, the order of comparison in Raita algorithm [11] is carried outby first comparing the rightmost and leftmost characters of thepattern with the sliding window. If they both match, the remainingcharacters are compared from the right to the left. Intuitively, theinitial resemblance can be established by comparing the last andthe first characters of the pattern and the sliding window.Therefore, it is anticipated to further decrease the unnecessarycomparisons. The Aho-Corasick[14] algorithm consists ofconstructing a finite state pattern matching machine from thekeyword and then using the machine to process the text in a singlepass. It can find an occurrence of several patterns in the order of*O(n)* time, where *n* is the length of the text, with pre-processing ofthe patterns in linear time.

In this paper construct anew algorithm which checks the input DNA profile to find all the occurrence of a allelic pattern within this profile, based on skip (put '_'to match the patterns)[15] can be described as follows-

1. Fix the DNA Profile index in a cretin position.
2. Use this position as a starting point of matching.
3. Compare the DNA Profile contents from the defined point with pattern contents.
4. Find the skip value depending on the match number (rangesfrom 1 to y-1)
5. Perform the above sequence while the DNA Profile position does not reaches x-y.

Description ofthe Algorithm:
This algorithm assumes that there is input DNA Profile (T) that has size (x) and there is an Allelic Pattern (P) with size (y). Sothe algorithm proceeds as follows –
1. Input DNA Profile (T) of size (x) and Allelic Pattern (P) of size (y).
2. Output starting index of all substring occurrence of (T) that is equal to (P) and output (-1) if no such substring exists.
3. Initialization step: skip = 1, index i of T = 1. Number of occurrence = 0.
4. Check index. If index <= x-y, then go to step 5. Else go to step 12.
5. Now set index j of T = 1, and save i if (k = i).
6. Check j. If j<=y go to step 7, else go to step 8.
7. Compare P(j) and T(k). If they are equal then increase the value of j and k by 1 and go to step 6.
8. Skip if j=skip.
9. Increase number of occurrence.
10. Now do (I +skip).
11. Go to step 4.
12. Return number of occurrence.

V. IMPLEMENTATION OF THIS ALGORITHM

This algorithm was implemented using C and tested usingdifferent DNA sequences and DNA Profile with different DNA Profile sizes. However this algorithm will be compared with other algorithms innear future.

VI. RESULT ANALYSIS

Due to the lack of resources, mainly the theoretical resultsare considered here. In future experiment based results will becompared for a better understand of this algorithm. This algorithm is very simple to describe and hasfollowing main features –
1. Good time complexity.
2. Unlimited sized of DNAProfile can be handled.
3. Unlimited sized of Allelic Pattern can be handled.
4. Skip technique is used.
5. Can be used different ranges of applications.

VII. FUTURE WORK

Here on this particular paper, mainly emphasizing ontheoretical results. In future I will emphasize more onpractical results. This algorithm will be comparedwith other algorithms that has common features like –
1. Multiple DNA ProfileMatching and Searching
2. Maintaining different types of DNA Profiles (Particular Case wise)
3. No processing time

In near future this algorithm will be compared withother algorithms considering other factors.

VIII. CONCLUSION

A new algorithm is developed for indexing and searching mechanism for multiple DNA Profiling and Allelic Patterns.This algorithm proves performance which is bettercompared to other algorithms. This is a very light weightalgorithm. The space complexity and the time complexity ofthisalgorithm is less compared to other andalgorithms. This algorithm can work with a very long text and pattern which is a great advantage over other algorithms.

ACKNOWLEDGMENT

I would like to thank my Professor Mr.HirenKamaniand Dr.NikunjBrahmbhattfor his time. I am grateful to him for providing me support. Finally, I acknowledge my Professor Mr.HirenKamanifor his kind suggestions on every step of thiswork.

REFERENCES

[1] J. M. Butler*, C. R. HillNational Institute of Standards and Technology, Applied Genetics Group, Gaithersburg, Maryland, United States of America

[2] Butler JM. Forensic DNA typing: biology, technology, and genetics of STR markers. 2nd ed. Elsevier: New York, 2005.

[3] Gill P. Role of short tandem repeat DNA in forensic casework in the UK--past, present, and future perspectives. Bio Techniques 2002;32:366-72.

[4] Budowle, B., et al. (1998). CODIS and PCR-based short tandem repeat loci: DNA Profiling tools. *Proceedings of the Second European Symposium on Human Identification*, pp. 73-88. Madison, Wisconsin: Promega Corporation.

[5] People v. Miles (1991) 577 N.E. Rptr. 2d Series 477 (Ill. App. 4 Dist. 1991), RFLPs (2004)

[6] People v. Castro (1989) 144 Misc. 2d Series 956, 545 N.Y. Supp. 2d Series 985 (Sup.Ct.1989), VNTRs.

[7] Texas Tech University."Molecular Genetic Markers."Genetics@TTU. Fall 2005, 11thNov. 2005.

[8] United States v. Downing (1985) United States Court of Appeals, 3rd Circuit, Vol. 753Federal Series, 2d, pp. 1224. 609 F. Supp. 784 (E.D. Pa. 1985).753 F. Rptr. 2d Series 1227-28 (3rd. Cir. 1985).

[9] Hares, D.R. (2011). Expanding the CODIS core loci in the United States.*Forensic Sci. Int. Genet.*(in press), doi:10.1016/j.fsigen.2011.04.012

[10] Needleman, S.B Wunsch, C.D(1970). "A general method applicable to the search for similarities in the amino acidsequence of two proteins." J.Mol.Biol.48,443-453.

[11] RajuBhukya, DVLN Somayajulu,"An Index Based KPartitionMultiple Pattern Matching Algorithm", Proc.Of International Conference on Advances in Computer Science2010 pp 83-87.

[12] Smith,T.F and waterman, M (1981). Identification of commonmolecular subsequences T.mol.Biol.147,195-197.

[13] Horspool, R.N., 1980. Practical fast searching in strings. Software practice experience, 10:501-506.

[14] Aho, A. V., and M. J. Corasick, "Efficient string matching:an aid to bibliographic Search," Communications of theACM **18** (June 1975), pp. 333 340.

[15] Ziad A.A Alqadi, MusbahAqel&Ibrahiem M.M.EI Emary,Multiple Skip Multiple Pattern Matching algorithms.IAENGInternational.Vol 34(2),2007.