Family relation and STR-DNA matching using fuzzy inference

Maria Susan Anggreainy[1], M. Rahmat Widyanto[2], Belawati Widjaja[3],
Nurtami Soedarsono[4], Putut Tjahjo Widodo[5]
[1,2,3]Faculty of Computer Science, Universitas Indonesia, Indonesia
[4]Faculty of Dentistry, Universitas Indonesia, Indonesia
[5]DNA Laboratory, Indonesian National Police Center of Health and Medicine, Indonesia

Article Info

ABSTRACT

Deoxyribose Nucleic Acid (DNA) are the basic elements that make up a whole section of an individual. The basic elements store information that is unique to each individual and will be passed down the generations. DNA also helps in identifying the father in paternity testing, locating missing person investigations, identifying victims in mass disasters. Identification of the victims has a problem if the comparison the father and mother no instance the victim's parents have died or are very far away from where the victim. Therefore, it is necessary to try to identify Short Tandem Repeat (STR) Inference of live family such as sibling, grandfather/grandmother, uncle/aunt, cousin and nephew. In this paper, we performed a method to measure the similarity of human DNA profiles using fuzzy similarity. In this fuzzy system, DNA profile data is used as an input that stores human identity along with its DNA profile. The data entered is the result of polymerase chain reaction (PCR) identification which is an electropherogram consisting of 16 loci with two alleles for each locus. Output in this fuzzy system is the value of individual similarity with reference and with similarity levels, namely small, medium and high.

Corresponding Author:

Maria Susan Anggreainy,
Faculty of Computer Science,
Universitas Indonesia,
Depok Campus, Depok-16424, West Java, Indonesia.
Email: maria.susan61@ui.ac.id

1. INTRODUCTION

Deoxyribose Nucleic Acid (DNA) is used to identify criminals, clear suspects and exonerate persons mistakenly accused or convicted of crimes with incredible accuracy when biological evidence exists [1]. DNA Profile is related to the prevention or detection of crime, related to identification of a decreased person, is in the interest of National Security or in a counter-terrorism investigation [2]. If there is damage to all or part of the body of the victim or suspect there is little evidence of a crime, there will be difficulties identifying which causes problems, issues such as the long of the settlement of a case. Therefore, victims or suspects used DNA as the primary means of identification. This is done because the DNA can be found in almost all of the human body, in addition to the unique properties of DNA can be used for identification. Identification of DNA consists of analyzing samples to isolate a unique set of DNA markers. An analyst then compares the DNA profiles to determine whether a person's DNA sample was matched with evidence obtained from the crime scene or of a family relationship.

In a family, a child's DNA profile is a combination of both parents' DNA profiles because the child has a chromosome that one allele handed down by his father and the other allele is derived from the mother. An individual can be declared as a child of a father or mother if it has a similar DNA about 50% of the total DNA because 50% of the DNA is directly passed down by the father or mother [3]. Furthermore, it can be

concluded that the similarity with the biological grandmother or grandfather was about 25%. In previous studies used a comparison to families in the form of the father and the biological mother, and biological grandparents. The comparator is considered to be very good because the resulting match value is very high or can be already perfect. However, behind the perfection of it, there is a drawback that may be difficult to overcome, there is no certainty that such a comparison is there to take the DNA profile of the victim, for example, the parents of the victims have died or parents are incomplete or very far away from where the victim. It is, therefore necessary for case identification DNA wider family relations such as uncles, aunty, cousins, and all the other possibilities surviving family. Based on the research that has been done, the similarity between an individual and his siblings is approximately 45% to 54%, in other words about 50% to note that these siblings have a father and mother together with individuals who are being identified (full sibling). In contrast, if only to have the same father or mother, not both, then the resemblance there is 25% (half-sibling). Another family member, namely uncle or aunt also similar 25% and 25% with a record nephew had the same parental relationship. From there it can be concluded that there is only 12.5% similarity with the cousins [4].

DNA profiles will be compared have 16 loci, each DNA locus is comprised of two alleles, or copies, of the marker-one inherited from the mother and one inherited from the father. Sixteen locus should be compared everything to make a decision whether there is any relationship between the DNA profile evidence the biological by comparison. The sixteen loci are as follows, CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S720, D8S1179, FGA, TH01, TPOX, VWA, and Amelogenin [5].

In previous research algorithms matching DNA-based STR profiles uses 0 or 1 measure of similarity [6]. We propose methods for building fuzzy similarities to identify matching because STR-DNA data is produced by PCR machines containing impresition while fuzzy logic is designed to handle data containing impresition and uncertainty. The need to have fuzzy resemblance steps is triggered by the fact that STR profiles often show real values as allele markers, not natural number, this must be a noise effect in the process of analyzing STR profiles [7]. Using fuzzy inference measure of similarity, two alleles with small differences will still get similarity scores instead of sharp 0, which eliminates the possibility of two alleles that have the same value although only slightly different, which can occur due to noise during DNA data acquisition.

2.   PROPOSED METHOD

To improve the accuracy of DNA analysis, the process of checking to do as much as possible so as to get objective results. This requires the presence of DNA samples in large numbers. On the other hand, the sample DNA is highly susceptible to noise, such as blood mixed with other DNA samples or damage due to temperature and weather. This can increase the habits of the data so that analysis of the process to be less valid identification. In forensic science, known to some kinds of methods of DNA profiles, namely:

1. Restriction fragment length polymorphism (RFLP) analysis:

RFLP is a method first introduced. RFLP is a DNA fingerprinting technique based on the detection of DNA fragments of varying length. With the development of DNA analysis techniques and the newer and more streamlined, RFLP no longer is used because it requires a relatively large sample of DNA. In addition, samples are usually obtained also usually degraded by environmental factors, such as dirt or mildew that can not be used for RFLP [8].

2. Mitochondrial DNA (mtDNA) analysis:

Mitochondrial DNA (mtDNA) is very good to use as a tool for the analysis of DNA because it has three important properties, namely DNA has a high copy number of about 1000-10000 and is in the cells which have no nucleus such as red blood cells or erythrocytes. Mitochondrial DNA can be used for analysis despite the limited number of samples found, easily degraded and in conditions that do not allow for analyzing the DNA core. Second, the human mitochondrial DNA is passed down maternally, so that each individual on the same maternal line have identical mitochondrial DNA types. Characteristics of mitochondrial DNA can be used for the investigation of cases of missing persons or determine a person's identity by comparing the mitochondrial DNA of victims' brother lysed maternal lineage. Thirdly, the mitochondrial DNA polymorphism has a high rate of the rate of evolution is about 5-10 times faster than nuclear DNA. Mitochondrial DNA is a technique that is very expensive and exclusive matrilineal and therefore less informative.

3. Y-Chromosome analysis:

Analysis of Y chromosome is used for investigation on human evolution and for forensic purposes or analysis father [9]. DNA-polymorphisms on the human Y chromosome are valuable tools for understanding human evolution, migration and for tracing relationships among males. Majority of the length

of the human Y chromosome is inherited as a single block in linkage from father to male offspring as a haploid entity.

4. Single nucleotide polymorphism (SNP) typing:

Single Nucleotide Polymorphism (SNP) typing is a DNA sequence variation occurring when a single nucleotide (A, T, C, or G) in the genome sequence is altered. For example, an SNP may change the DNA sequence AAGGCTAA to ATGGCTAA [10]. Excess SNP is useful in some SNP loci are positioned very close together to define their haplotypes and haplotype development tags. Disadvantages SNP is in need of genetic sequence information for a target gene analysis and require procurement of equipment and materials that are costly and needs a large multiplexing test.

5. PCR

PCR is used to make millions of copies of DNA from biological samples. Amplification of DNA using PCR caused DNA analysis on biological samples requires only a small sample and can be obtained from the sample as fine as hair. The ability of PCR to amplify small amounts of DNA makes it possible to analyze samples that are degraded, though. Still, it must be prevented contamination with other biological materials for the identification, collection and preparing samples [11].

6. STR

STR, a popular method used to replace RFLP method. By agency FBI (Federal Bureau of Investigation), this method is proposed as a standard to do a DNA profile. As a result, the profile STR profiling method accepted by many forensic laboratories in the world as a method for profiling. Furthermore, the Agency FBI also proposed to use a number of loci STR profiling results for the purposes of identification of human identity. Some loci are then referred to as the human DNA profiles. A person's DNA profile can be matched DNA profile data resemblance to one another. For the similarity matching process, the agency NIST (National Institute of Standards and Technology) makes a software (software) named auxiliary STR_MatchSamples. Matching results could lead to certain conclusions required by the parties concerned. A problem arises, if the profiling process contamination on biological evidence collected by other chemicals (degraded quality). As a result, the DNA profile obtained will contain the value uncertainty (uncertainty). For this case, the software is not able to handle it STR_MatchSamples aids for software STR_MatchSamples working with crisp logic [12].

2.1. Calculation of locus similarities between two individuals

In previous studies, proposed formulas follow four conditions, namely: the father and the biological mother there, one from the father and the biological mother is not there, there are siblings of individuals who want to be identified, if biological parents are not there, siblings used to represent both parents. The result of fuzzy inference of the proposed method, siblings can be used as a substitute for a parent because the resulting value is quite good and quite close to the value of similarity with parents. The present study also follows the same method, namely by paying attention to external factors such as temperature and weather that could cause a change or shift in the value of STR DNA. Following the assumption of a triangle as in the study, with the STR position as the midpoint of the triangle. Then assumed Similarly, a shift that could happen is a maximum of 0.2 so that the base of the triangle is equal to 0.4 to 0.2 values to the left of the value of STR and 0.2 values to the right value STR. The height of the triangle is assumed to 1 (one) because the value is the greatest similarity value.

The human genome is composed of repetitive DNA strand units in various sizes are patterned. Regio DNA with a short repetition units (roughly along the 2-6 bp) is called Short Tandem Repeats (STR). An individual inherits one copy of each STR parents. The repetition of this unit become STR DNA markers that have high variation in a group of individuals, so it is very effective DNA STR markers are used for identification purposes only human [13].

The smaller the STR alleles that make up the better, because the level of variation given the higher and given that STR found at forensic testing could degradation the result of environmental influences. In addition, the size of the small can make an STR can be separated easily in the DNA to avoid election locus adjacent to avoid disruption of patterns the random distribution of the statistical analysis. The value of STR used in forensic test is the Combined DNA Index System (CODIS), which consists of 16 loci, namely CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51,D21S11, D19S433 and D2S1338 and amelogenin to determine the sex. CODIS issued by the FBI Laboratory as an international standard for identification of individuals.

Alleles of a locus DNA profile human the resulting from the process identification of DNA evidence sometimes worth improper. This could several have been the caused by factors such as the effects of weather and temperature, evidence of contamination by other substances even the possibility of errors due to PCR machine. This will be the main cause misidentification of victims in DNA profile matching performed crisp. For minimize errors identification of DNA profiles then matching DNA profiles is done with using fuzzy.

If a victim has allele short tandem repeat (STR) 20 and reference allele had 20.2 STR, then both alleles have value 0.5 similarity so that both alleles can be said to similar. Fuzzy similarity measurement is performed by measuring the DNA profile a semblance of an allele. Assuming that a triangular shaped allele with short tandem repeat (STR) of an allele show middle value, the second leg is the same distance of 0.4 and higher for the same allele with 1. Then to similarity measure each alleles of a locus compared to use the equation.

$$t = \frac{a2-b1}{2(a2-a1)} \tag{1}$$

where: Individual allele focus 1 <individual allele focus 2 and the value of t, a1, a2 and b1 used double data type.
t = similarity score
a2 = first allele
a2 = a1 + 0.2
b1 = second allele –0.2
replace the symbols a3, b1 and a2 (referring to their original definition) with only the allele value 1 and as the allele value 2, and as a symbol of similarity, it becomes:

$$S = \frac{\frac{1}{2}(x+0.2)-(y-0.2)}{(x+0.2)-x} \tag{2}$$

further breaking down the formulation by doing simple multiplication operations:

$$S = \frac{\frac{1}{2}(X-Y+0.4)}{0.2} \tag{3}$$

finally, divide and multiply the coefficients, and finally the simplest linear form is begotten:

$$S = 2.5X - 2.5Y + 1 \tag{4}$$

As obvious as it seems, the final formulation turns out to be a simple linear equation. This new function fits perfectly into first-order Takagi-Sugeno FIS output, which takes arguments as the variable of the function and building a linear model as an output.

2.2. DNA similarity measurement method

In Figure 1 given a DNA sample allele distribution of both parents to three children. The first child had inherited one allele from her mother, while the second allele from the father. The second child has one of the alleles father and two alleles of the mother. The third child has one of the alleles father and two alleles of the mother. From the example of Figure 1 can be inferred that, if a child has alleles one of the fathers, the two alleles ascertained mother's, and vice versa.

The first allele will be compared with those of the patrilineal, then the second allele will be compared with those of the matrilineal. This causes there will be two times calculation for each comparison. The first is to compare the first with reference allele father and comparing allele both with reference mothers. The second is to compare both with reference alleles and allele father first with reference mothers. The idea of these methods is (eg a comparison that there father and mother do not exist):
1. Take the first allele of a sibling, if the allele is present at one allele from the father, then the second allele certainly belonged to the mother. If no, then it is the first allele that belonged to the mother.
2. Do the same for subsequent siblings

The value taken is the greatest of these two possibilities in accordance Generalized modus ponens (GMP). Since its introduction, in L. A. Zadeh's paper, Generalized Modus Ponens (GMP) has become one of the most powerful tools in approximate reasoning [14]. However, GMP has been used without any assumptions, which if verified, would increase the specificity of the inferred conclusion. One such hypothesis is the gradual relationship between the premise and the consequence. Figure 2 is an example of a family tree that contains the complete family members.
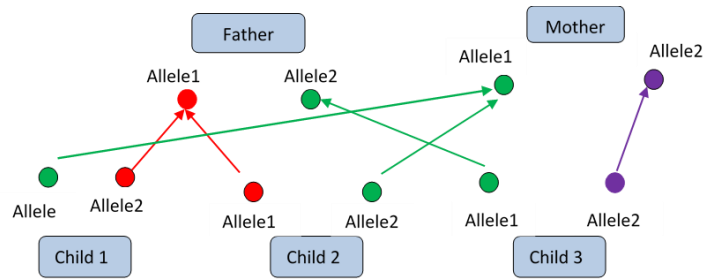
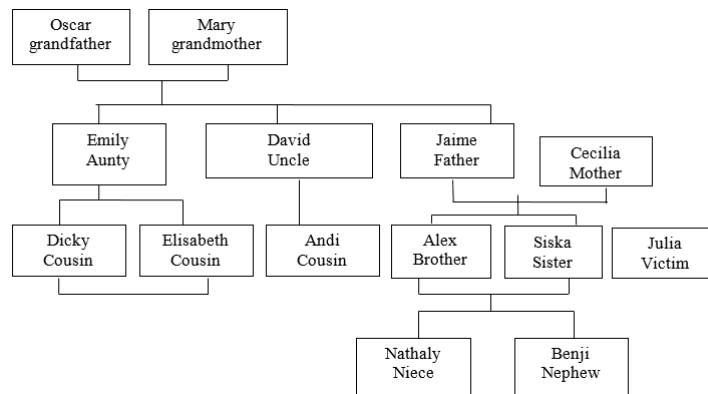Figure 1. DNA sample allele distribution of parent and child [11]

Figure 2. Family tree example

As for the formula proposed for this research are as follows:
1. If there is a comparison that is the biological father and mother, then siblings in the family should not be compared. The equation applies :

Similarity=like (Victim, Father)+like (Victim, Mother)

Variable compatibility is a match value obtained in the process identification, in other words, the final result of the implementation. Predicate similar (A, B) is a function that maps STR values of A and B the real numbers [0; 0,5] where 0 indicates that A and B are not similarities and 0.5 are that A and B have similar perfect [15].
2. If one of the biological father or mothers is not there, but there are biological sibling from the victim who wish to be identified. Siblings will be a substitute for a parent who does not exist because in DNA siblings there are parental DNA. The equation applies:

Simalarity=like (Victim, Father)+PseudoLike1 (Victim, Father, Sibling)

To calculate the similarity with the parent, if the father or mother is not there then one of the functions similar will be worth 0. It certainly would reduce the value of accuracy the calculation to be performed. thus introduced a new function that is PseudoLike (A, B, C) that is, a function which maps STR A, B, and a set of STR-C to the value [0; 0.5] in which the value generated is the value of "hope" resemblance between a parent in addition to B, for example, B is the father, then the value of similarity "hope" that is generated between the victim and his mother. Variable C on function is set of the value of each sibling STR, because it is assumed siblings opted could be more than one.
3. If both parents are not there, then the decision directly taken from biological siblings, siblings will be used to represents both parents. The equation applies:

Similarity=PseudoLike2 (Victim, Sibling)

In the above function use a new function PseudoLike2 (A, B) where the function of mapping the value of the set STR A and B into a value similarity [0,1] with 0 indicates that there is no match between A with family and 1 indicates a perfect match.

4. If both parents are not there, there is no siblings, grandparents are still there. The equation applies:

Similarity=like (Victim, GrandFather)+like (Victim, GrandMother)

Similar predicate (A, B) is a function that maps STR values of A and B to the real numbers [0; 0.25] where 0 indicates that A and B do not have similar and 0.25 states that A and B have a perfect resemblance.

5. If both parents are absent, there is no siblings, one grandparent still exists. The equation applies:

Similarity=like (Victim, GrandFather) or like (Victim, GrandMother)

6. If both parents are absent, there is no siblings, grandparents are absent, aunty and uncle are still there, the equation applies:

Similarity=like (Victim, Uncle)+like (Victim, Aunty)

Similar predicate (A, B) is a function that maps STR values of A and B to the real numbers [0; 0.25] where 0 indicates that A and B do not have similar and 0.25 states that A and B have a perfect resemblance.

7. If both parents are absent, there is no siblings, grandparents are absent, one aunty/uncle still exists, the equation applies:

Similarity=like (Victim, Uncle) or like (Victim, Aunty)

8. If both parents are absent, there is no siblings, grandfather/grandmother is missing, aunty/uncle did not exist, there is still a cousin. The equation applies:

Similarity=PseudoLike3 (Victim, Cousins)

In the above function use a new function PseudoLike3 (A, B) in which the function is mapped value set STR A and B into a similarity value [0,1] with 0 indicates that there is no match between A and his family and one declared a perfect match.

9. If both parents are absent, there is no siblings, grandparents are absent, aunty / uncle is absent, cousins are absent, nephew still exists. The equation applies:

Similarity=PseudoLike4 (Victim, Nephews)

In the above function use a new function PseudoLike4(A, B) in which the function is mapped value set STR A and B into a similarity value [0,1] with 0 indicates that there is no match between A and his family and one declared a perfect match.

2.3. Fuzzy inference of each DNA profile locus

Input for fuzzy inference is the value of the similarity of two alleles on the corresponding locus, and the result is the similarity value of each locus of the DNA profile. There are two methods used in this inference: Sugeno and Mamdani methods. In the implementation of which is different from the two methods, the defuzzification technique is used and the membership set outputs are fuzzy inference systems. The implementation equation lies in the number of input fuzzy inference systems, input membership sets, and inference rules used. There are two fuzzy inference systems, allele1 and allele2. The two alleles have the same membership set. Geometrically, an overview of the membership set can be seen in Figure 3.
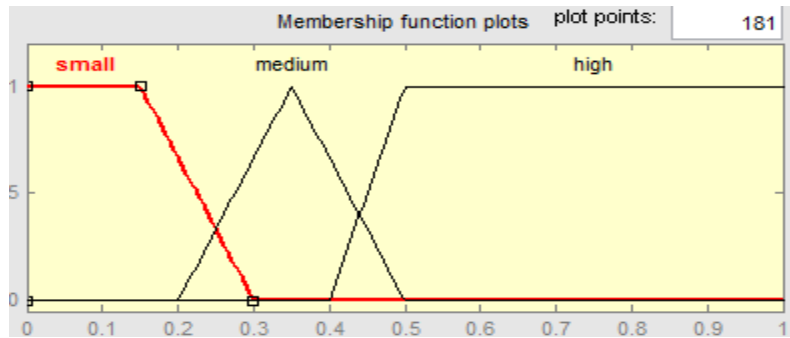
Figure 3. Graph membership functions [15]

The membership degree of the two alleles is determined by the similarity value produced. The fuzzy inference system (FIS) is a system that uses fuzzy set theory to map inputs to outputs using fuzzy logic [16]. FIS methods are often used there are two namely Mamdani and Sugeno methods. The Mamdani fuzzy inference engine takes fuzzy inputs and produces fuzzy output based on the pre-defined rules, on the other hand, Takagi-Sugeno fuzzy inference system takes fuzzy inputs and produces crisp outputs [17]. In this paper, Sugeno's fuzzy method is used to conclusion. Conclusion proposed is the average value for all similarity locus of reference fathers with an average value for all similarity locus of reference mothers. The two statements would be the premise for systems that generate value match individual membership in the family. Corresponding previous studies each similarity value (reference father and mother) to follow the membership function as follows,

$$f_{small}(x) = \begin{cases} 0 & if\ x \geq 0,3 \\ 1 & if\ x \leq 0.15 \\ \frac{0,3-x}{0,15} & if\ 0,15 < x < 0,3 \end{cases} \tag{5}$$

$$f_{medium}(x) = \begin{cases} 0 & if\ x \leq 0,2\ dan\ x \geq 0,5 \\ \frac{x-0,2}{0,15} & if\ 0,2 < x \leq 0,35 \\ \frac{0,5-x}{0,15} & if\ 0,35 < x < 0,5 \end{cases} \tag{6}$$

$$f_{high}(x) = \begin{cases} 0 & if\ x \leq 0,4 \\ 1 & if\ x \geq 0,5 \\ \frac{x-0,4}{0,1} & if\ 0,4 < x < 0,5 \end{cases} \tag{7}$$

Fuzzy inference is a process of obtaining new knowledge through existing knowledge using fuzzy logic [18]-[20]. The fuzzy rules are applied are as follows [21] shown in Figure 4.

1. If (father is small) and (mother is small) then (similarity is small) (1)
2. If (father is small) and (mother is medium) then (similarity is small) (1)
3. If (father is small) and (mother is high) then (similarity is medium) (1)
4. If (father is medium) and (mother is small) then (similarity is small) (1)
5. If (father is medium) and (mother is medium) then (similarity is medium) (1)
6. If (father is medium) and (mother is high) then (similarity is medium) (1)
7. If (father is high) and (mother is small) then (similarity is small) (1)
8. If (father is high) and (mother is medium) then (similarity is medium) (1)
9. If (father is high) and (mother is high) then (similarity is high) (1)

Figure 4. Fuzzy rule

Based on Sugeno method, it must find fuzzification weights for weighted average for each degree of similarity, namely small, medium, and large.

g_Small=0.15 (point change in slope down into a flat line)
g_Medium=0.35 (the midpoint between 0.2 and 0.5)
g_High=0.5 (point change in slope rising to flat line)

The weighted average method chosen for the calculations is quite easy and the number of degree of similarity only slightly. The greatest value that can be generated by these functions also follows that the value is in the diagram fuzzy, thus turning it into a conclusion could very easily because only a mapped value defuzzification results with the values in the diagram fuzzy.

Defuzzification value will be calculated for each resemblance to his father and mother. Furthermore, the value of both is added together to get the total similarity value. Once calculated, the maximum value that can be obtained from the defuzzification function is 0.5. This happens because the value of 0.5 is the highest value that may be generated by the proposed similarity function. In other words, the total value could be generated is 1 which means full similarity.

3. RESULTS AND ANALYSIS

Data used in the experiment is a DNA profile data obtained from the Faculty of Dentistry, University of Indonesia, which consists of 100 DNA profile data consisting of data on 43 men and 57 women of data. That then the data is stored on a database of DNA profiles. For trial similarity measurement reference DNA profile with the biological family data used consists of 10 data including the data contained individuals which has a biological relationship.

The process of this experiment will be conducted in some cases already happened with the results positive that the victims were indeed members of the family. The experiment will be conducted for some conditions, namely with the help of one of the parents and all my siblings without a parent or with the only sibling, or with grandparents, aunts uncles, cousins, and nephews. In each such instance will be calculated for each value matches the number of families that will eventually look how many families were required to produce a very high value.

This system is implemented by using Matlab R2016b and MySQL as database management system. The database consists of two tables: a biodata table and a DNA reference table. In the DNA reference table there are 34 fields/columns namely id (serial number of references), classification (kinship relationship between references with individuals consisting of: father, mother, siblings, grandparents, uncles, aunts, cousins and nieces) And 32 locus values of 16 locus (sixteen loci are CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S720, D8S1179, FGA, TH01, TPOX, VWA and Amelogenin) respectively has two alleles. Allele data entered with double type except for locus amelogenin varchar type with character length 1. DNA profile data of PCR identification result which is electropherogram.

Measurement similarity of DNA profiles using size fuzzy similarity involves assigning values to the similarity each allele are then produced from a value similarity loci. The average of all loci similarity value is the value similarity of DNA profiles. The DNA profile suitable match can be said if the value of similarity >0.5.

In Table 1 shown the greater the number of siblings, the greater the accuracy value. From this result, several things can be analyzed. First, the number of siblings. One person will only produce one of the two senior alleles represented, so the error value or entropy is 50% where an error will occur if the allele victim is not owned by the sibling. Meanwhile, if the number of siblings more than one allele that is not owned by the represented parent can be found, then the entropy value can be 0 if both alleles are found and a maximum of 0.5 if all siblings have the same allele. values. The second thing that can be analyzed is that the number of siblings cannot reach 100%.

Table 1. The Average Value of a Match with Fuzzy

Number of siblings	Number of Cases	Father	Mother	Without Parents
1	20	0.75	0.75	0.70
2	20	0.75	0.77	0.75
3	20	0.80	0.80	0.77

Table 2 shown similarities appear to rise and reach a maximum in the number of siblings 3, but only reach a maximum of 0.3 for the column without parents and 0.41 for the mother column even though the value is monotonous up. It can be seen that the similarity values generated by the t function are lower than the values generated by the proposed method, namely fuzzy.

Table 2. The Average Value of a Match with the Function t

Number of Siblings	Number of Cases	Father	Mother	Without Parents
1	20	0.25	0.36	0.21
2	20	0.30	0.41	0.27
3	20	0.30	0.41	0.33

The absence of one parent causes a lack of one source of alleles that can be used for counting. The calculation of the function t and the fuzzy function is almost the same as with fuzzy similarity values obtained will be divided into two, so that the maximum value is 0.5. This is because only one of the alleles is compared in two individuals, so that it can only be said half of all alleles that are expressed are similar. Whereas the other alleles will be compared to other parents' lineages. Calculations with the function t, the value contained in each locus is very important, therefore if there is a change to just one locus in a DNA reference, there could be a significant change in calculating the match value

Measurements of fuzzy similarity of DNA profiles with reference to DNA profile data were done by measuring fuzzy similarity of each allele of each query loci with each allele at all loci with each record contained in the DNA profile database. For each record that has been compared or measured similarity will result in a similarity score. As output on the system interface from fuzzy profile similarity measurements the DNA with this DNA profile database reference is 16 of the largest similarity values of all records compared.

Table 3 is an example of similarity measurements fuzzy profile of that DNA the system is done against three records stored in the database profile DNA based on query DNA profile and show the value of similarity between the DNA profiles of victims with reference to the biological mother and siblings is 1. Due to the presence of the father as a reference is not available then replaced by a biological father's parents are siblings of victims in order to obtain biological evidence is more accurate because if one allele inherited from the victim's biological father then that allele course there passed on to siblings. When reference is compared only with the biological mother only similarity value between DNA profiles of victims by using fuzzy 0.3. After adding a reference siblings to replace the biological father. Value similarity between DNA profiles of victims rose to 1.

Table 4 shows the value of the similarity between the victim's DNA profile and the number of cousins using fuzzy increases and the function t increases. But with fuzzy functions it looks better than the function t. Analysis at function t is lower because the calculation is done by calculating the average value of the suitability of the cousin's DNA profile and comparing it with the individual STR values. This value is considered suitable for a locus, if one of the two alleles at the locus is the same. Then look for the average values for all loci and then look for the average for all siblings tested. Whereas with fuzzy cousins are considered to replace the father's allele and the mother's allele so that both father and mother alleles are added.

If the presence of the father and the biological mother is not there, then it can be identified families of the victims who are still alive. Figure 5 shows the value of similarity with the functions t and fuzzy between the DNA profiles of victims with references that uncle/aunt, grandfather/grandmother, cousins, nephew. the trial of 10 cases of identification of victims with reference uncle/aunty 56% by similarity function t, whereas with fuzzy 100%, and with reference grandfather/grandmother 52% by similarity function t, whereas with fuzzy 100%. For identification of the victim with reference cousins and nephew do with trial 15 cases, the results of similarity with the reference function t cousins 38% and the similarity with the reference fuzzy nephew 48%. For reference cousins and nephew are done with the number of siblings as much as 3, while the number of siblings 2 results similarity with the reference cousin t function 38% and the similarity with a fuzzy 69%, with the number of siblings 1 results similarity with reference cousin t function 21% and the results of similarity with 48% fuzzy. That the more the number of siblings, the more the number of loci that match.

Table 3. Measurement of Similarity with Mother Biological and Siblings

Loci Name	Queri		Mother		t	Sibling1		t	Sibling2		t
	Alel1	Alel2	Alel1	Alel2		Alel1	Alel2		Alel1	Alel2	
Amel	X	X	X	Y	0	X	Y	1	X	Y	1
D3S1358	17	17	15	16	0.5	16	18	0	15	16	0
D7S820	8	9	11	11	1	11	11	0	11	11	0
D8S1179	13	16	16	17	0.5	10	13	0.5	14	14	0.5
D211S11	31	32.2	29	31.2	0	28	33.2	1	31.2	33.2	1
CSF1P0	10	12	11	11	1	12	12	0	11	12	0.5
TH01	7	7	6	8	0	7	7	9	9	0	0.5
D13S317	10	10	8	11	0	11	11	0.5	10	13	0.5
D16S539	9	10	9	11	0	12	12	0.5	9	9	0.5

Family relation and STR-DNA matching using fuzzy inference (Maria Susan Anggreainy)

Table 3. Measurement of Similarity with Mother Biological and Siblings *(Continue)*

Loci Name	Queri		Mother		t	Sibling1		t	Sibling2		t
	Alel1	Alel2	Alel1	Alel2		Alel1	Alel2		Alel1	Alel2	
D2S1338	19	22	18	19	0	19	22	0.5	23	25	0.5
D19S433	13	14	15	15.2	0	13	15.2	0	13	15.2	0
VWA	17	18	15	17	0	17	17	1	16	19	1
TPOX	9	11	8	9	0	11	11	0.5	8	8	0.5
D18S51	15	19	15	16	0.5	15	18	0.5	8	8	0.5
DS5818	12	13	10	10	1	12	12	0.5	11	11	0.5
FGA	22	22	21	21	1	22	25	0	20	21	0
		Fuzzy			0.3	Fuzzy		1	Fuzzy		1
		Function (t)			0	Function (t)		0.5	Function (t)		0.75

Table 4. Similarity Measurement DNA Profile with Cousins

| Loci Name	Queri		Cousin1		t	Cousin2		t	Cousin3		t
	Alel1	Alel2	Alel1	Alel2		Alel1	Alel2		Alel1	Alel2	
Amel	X	X	X	X	0	X	Y	0.5	X	X	0
D3S1358	16	16	16	17	0.5	16	17	0	8	12	0.5
D7S820	11	12	8	9	0	8	8	0.5	8	12	0.5
D8S1179	12	15	15	16	0.5	12	16	1	14	14	0.5
D211S11	30	30	30	30	1	31	31.2	0.5	29	30	1
CSF1P0	12	14	11	11	0	11	12	0.5	29	30	1
TH01	7	9	6	9	0.5	9	10	0.5	11	11	0.5
D13S317	8	8	11	12	0	8	10	0.5	6	7	0.5
D16S539	10	12	10	11	0.5	9	13	0.5	11	12	1
D2S1338	24	24	19	19	0	17	24	1	19	19	0.5
D19S433	13	14	13	16.2	0.5	13	14.2	0.5	14	15.2	0.5
VWA	16	17	14	18	0	14	16	0.5	14	15.2	0.5
TPOX	8	11	8	8	0.5	8	8	1	8	8	0.5
D18S51	15	16	15	16	1	13	16	1	14	16	1
DS5818	10	13	12	13	0.5	10	12	0	12	13	1
FGA	23	25	22	24	0	19	24	0	22	23	0.5
		Fuzzy			0.3	Fuzzy		0.65	Fuzzy		1
		Function (t)			0	Function (t)		0.25	Function (t)		0.5
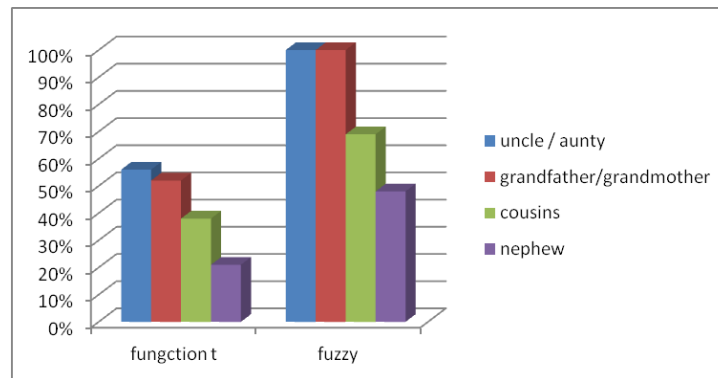
Figure 5. Comparison chart t functions with fuzzy

4. CONCLUSION

In conclusion, the fuzzy inference based on the results of the proposed method, siblings can be used as a substitute for a parent because the value generated enough good and quite close to the value of comparison with parents. Matching DNA profiles of individuals (query) with the DNA profile database Indonesia's country or with the biological family requests made by measuring the similarity of each allele at the locus DNA profile sixteenth using fuzzy similarity. The full biological family is used as a reference, the higher the similarity values measured DNA profile and a larger number of loci that match. If the similarity value is relatively small but alleged that the victim was the biological father and mother of children existing references, the process required a re-examination of biological material evidence of casualties. DNA profile similarity measurement results using fuzzy similarity is very satisfying. Of all the experiments carried out to

deliver results in accordance with the correct data. The proposed method is better than the conventional method that has been used. In addition, the measurement system fuzzy similarity of human DNA profiles is expected to be used to help the police. In future work is expected to get more data needed and can be validated with the root mean squared error (RMSE) which can measure the average error magnitude.

REFERENCES

[1] S. El-Difraway, *et al.*, "A Numerical Optimization Approach for Color Correction in Forensic DNA Genotyping," *Signals, Systems and Computer, Conference Record of the Thirty-Seventh Asimolar Conference*; vol.2, pp. 2088-2092 2003.

[2] GeneWatch UK, "Sharing DNA Profiles and FingerPrints Across the EU Requires Further Safeguard," Buxton, Derbyshire, December 2015.

[3] Lamb, *et al*, "Sibling Relationships: Their Nature and Significance Across the Lifespan," *Routledge*, 1982.

[4] Starr Barry, "Relatedness, " The Tech Museum of Innovation, 2013.

[5] Andrew Rennison, "Guidance: Allele Frequency Databases and Reporting Guidance for the DNA Short Tandem Repeat Profiling. Forensic Science Regulator Overseeing Quality," *Codes of Practice and Conduct*, 2014.

[6] J. M. Butler, "STRBase and Information Resources on Forensic DNA, National Institute of Standards & Technology," U.S. Dept of Commerce, 2012.

[7] M. R. Widyanto, *et al*, "A Novel Human STR Similarity Method using Cascade Statistical Fuzzy Rules with Tribal Information Inference," *International Journal of Electrical and Computer Engineering (IJECE)* vol. 6(6), pp. 3103-3111, December 2016.

[8] Pankaj Datta, *et al*, "DNA Profiling in Forensic Dentistry," *J Indian Acad Forensic Med,* vol. 34(2), April-June 2012.

[9] Budowle B, *et al*, "Interpretation Guidelines for Mitochondrial DNA Sequencing," http://www.promega.com, Available from:http://www.promega.com/geneticidproc/uss ymp10proc/37budowle.pdf , (Cited on 2011 Jul 9) .

[10] Forensic DNA Laboratories, Interpretation Guidelines for Y-Chromosome STR Typing. Scientifc Working Group on DNA Analysis Methods, 2014.

[11] D. Ricke, *et al*, "Sherlock's Toolkit: A forensic DNA Analysis System," Technolgies for Homeland Security (HST), *IEEE International Symposium* 1:10, 2015.

[12] Hartono, Reggio N. "Embedded Ethnic Inference in Fuzzy Logic System for DNA Profile Matching with Discrete Membership Function," Thesis S2. Depok: Fakultas Ilmu Komputer, Universitas Indonesia, 2010.

[13] Darrell O. Ricke, *et al*, "Human CODIS STR Loci Profiling from HTS Data," Bioengineering Systems and Technologies. Massachusetts Institute of Technology, *IEEE*, 2016.

[14] Phuc-Nguyen Vo, *et al*, "Gradual Generalized Modus Ponens," *Laboratoire d'Informatique de Paris VI.Universite Pierre et Marie Curie*, 2013.

[15] Tirmidzi Faizal Aflahi, "Human Identification Based on Profile. Parents and Siblings Use the Fuzzy Inference System-Identifikasi Manusia Berdasarkan Profil DNA Orangtua dan Saudara Kandung Menggunakan Sistem Inferensi Fuzzy," Thesis, Universitas Indonesia, 2012.

[16] Fatih Topaloglu, Huseyin Pehlovan, "Comparison of Mamdani Type and Sugeno Type Fuzzy Inference Systems in Wind Power Plant Installations," in *Proceedings - 2018 6th International Symposium on Digital Forensic and Security (ISDFS).* DOI: 10.1109/ISDFS.2018.8355384.

[17] George, Aby K., Harpreet Singh, "DNA Implementation of Fuzzy Inference Engine: Towards DNA Decision-Making Systems," *IEEE transactions on nanobioscience* (2017).

[18] Schott, D. J., F. Höflinger, R. Zhang, L. M. Reindl, and Hai Yang. "Fuzzy Inference System Assisted Inertial Localization System," In *Engineering, Technology and Innovation (ICE/ITMC)*, 2017 International Conference on, pp. 89-93. IEEE, 2017.

[19] Ahamed, N. U., Benson, L., Clermont, C., Osis, S. T., and Ferber, R., "Fuzzy Inference System-based Recognition of Slow," Medium and Fast Running Conditions using a Triaxial Accelerometer. *Procedia Computer Science*, vol.114, pp. 401-407. 2017.

[20] Kepski, M., Kwolek, B., and Austvoll, I., "Fuzzy Inference-Based Reliable Fall Detection Using Kinect and Accelerometer," In *International Conference on Artificial Intelligence and Soft Computing* pp. 266-273. Springer, Berlin, Heidelberg, April 2012.

[21] M. R. Widyanto, *et al*, "Various Defuzzification Methods on DNA Similarity Matching Suing Fuzzy Inference System," *Journal of Advanced Computational Intelligence & Intelligent Informatics*, vol. 14(3), 2010.