International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 6, December 2016, pp. 3103~3111
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i6.11810

A Novel Human STR Similarity Method using Cascade Statistical Fuzzy Rules with Tribal Information Inference

M. Rahmat Widyanto, Reggio N. Hartono, Nurtami Soedarsono
Faculty of Computer Science, University of Indonesia, Indonesia

Article Info

Article history:

Received Aug 13, 2016
Revised Oct 20, 2016
Accepted Nov 4, 2016

Keyword:

Cascade fuzzy rules
Short tandem repeat
Statistical distribution
Tribal information

ABSTRACT

A novel human STR (Short Tandem Repeat) similarity method using cascade statistical fuzzy rules with tribal information inference is proposed. The proposed method consists of two cascade Fuzzy Inference Systems (FIS). The first FIS is to discriminate the tribal similarity, and the second FIS is to calculate the STR similarity. By using the allele marker's statistical distribution probability density function as the membership function in the Fuzzy Rules of the first FIS, the new method makes it possible to tell the tribal similarity between two STR profiles. A 727 data acquired from tribal groups of Indonesia is used to examine the method produced promising result, being able to indicate higher tribal similarity score within a tribal group and lower similarity between tribal groups. In the light of Indonesia's diverse tribal groups, these properties are able to be leveraged as a new way to improve the versatility of existing DNA matching algorithm.

Corresponding Author:

M. Rahmat Widyanto,
Faculty of Computer Science,
University of Indonesia,
Depok Campus, Depok-16424, West Java, Indonesia.
E-mail: widyanto@cs.ui.ac.id

1. INTRODUCTION

DNA (Deoxyribonucleic Acid) is a genetic material which contains information that is unique for each individual. For this reason DNA has been used for human identification [1]. Currently there are many DNA analysis methods [2], e.g. Restriction Fragment Polymorphism (RFLP) analysis, Fragment Length Polymorphism (FLP), and Short Tandem Repeat (STR) analysis. The Federal Bureau Investigation (FBI) has chosen 15 locus and amelogenin based on STR analysis to be the standard of human identification [3]. M.R. Widyanto et al. [4] has proposed STR-Based DNA similarity matching using fuzzy inference system for human identification. There are many aims of human identification using DNA, e.g., family relation proof, criminal action evidence, and disaster identification. However, in the case of mass number of DNA suspects, it is important to reduce the numbers of suspects. Information regarding the possible tribe and ethnicity may help reducing DNA suspects for mass screening. Previous research on tribal information on DNA has been investigated. It has been observed that there are statistical and probabilistic properties of DNA in regards of the profile's ethnicity [5]. There are studies suggesting that there are certain Short Tandem Repeat (STR) allele proportions which occur distinctively different across tribal groups [6]. However, the research is conducted on assumption that there is no noise and uncertainty condition during DNA data acquisition.

This paper proposes a novel method to infer the similarity of tribal information using statistical fuzzy rules [7] to deal with uncertainty and impreciseness of STR. This paper is improvement of [4] to include tribal information inference on STR-Based DNA similarity matching using statistical fuzzy rules where the allele marker's statistical distribution probability density function is used as the membership function in the fuzzy rules. The proposed method consists of two cascade Fuzzy Inference Systems (FIS). The first which is based on Takagi-Sugeno FIS [8] is to discriminate the tribal similarity, and the second

which is based on Mamdani FIS [9] is to calculate the STR similarity. In the light of Indonesia's diverse tribal groups, these properties are able to be leveraged as a new way to improve the versatility of existing DNA matching algorithm. The experiment on 727 Indonesian DNA profiles shown that the new method is able to differentiate DNA profiles between tribal groups although the data is highly non-discriminative. In Section 2, the Fuzzy STR similarity method is described. In Section 3, the statistical property of tribal inference is explained. In Section 4, the proposed cascade statistical fuzzy inference system with tribal information inference is discussed. Section 5 discusses the experimental results of the proposed method. Conclusion is described in Section 6.

2. FUZZY STR SIMILARITY

DNA Profile is a set of numbers which reflects a person's DNA makeup, in contrast to full genome sequencing. There are many methods to do this work [2], but STR (Short Tandem Repeats) analysis is preferred [3] as used by the United Kingdom and United States. STR analysis counts the repetition count of patterns of two or more nucleotides which are repeated and the repeated sequences are directly adjacent to each other. The idea is to capture the STR value of a few locations in a person's genetic makeup called locus, as illustrated in Figure 1 (Original picture from National Institute of Standards and Technology (NIST)). Previous research has been undertaken to harness the flexibility of Fuzzy Set to the DNA (Deoxyribonucleic Acid) Profile matching's problem. M. R. Widyanto et al. [4] proposed a method to assign fuzzy similarity measure [10] between two allele markers, in contrast to crisp 0 or 1 similarity measure used in conventional STR based DNA profile matching algorithm [11]. The need to have a fuzzy similarity measure is triggered by the fact that STR profiles often showed real-valued numbers as an allele marker value instead of natural numbers. This is supposedly the effect of noise in the process of analyzing the STR profile. Using fuzzy similarity measure, two alleles with small difference will still get a similarity score instead of a crisp 0, which discard the possibility of the two alleles having similarity value although only differ slightly, which could occur because of the noise during DNA data acquisition.

Figure 1. The Locus in Chromosomal Positions

In DNA profile matching problem, for $M(\in \mathbb{N})$ amount of individuals, the dataset is described as

$$\{\boldsymbol{x_i}, \boldsymbol{y_i}\}_{i=1}^{M}, \tag{1}$$

where $\boldsymbol{x_i}$ is a vector of DNA profile evidence and $\boldsymbol{y_i}$ is a vector of DNA profile reference. The vector $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ are the $N(\in \mathbb{N})$ dimensional vector consisting of the value of 15 loci without amelogenin as has been used by Federal Bureau of Investigation (FBI) [3].

Value of every DNA loci is represented by fuzzy number (as shown by Figure 2) where the fuzziness value is set to be 0.4 through experiments [4] and the center of the fuzziness x_j is the value of the corresponding loci. The similarity value [10] between an allele of DNA profile evidence and DNA profile reference is given by

$$\mu(x_j, y_j) = \frac{\frac{1}{2}\left((x_j+0.2)-(y_j-0.2)\right)}{\left((x_j+0.2)-x_j\right)} \in [0,1], \tag{2}$$

where $x_j \in \mathbb{R}$ is the value of the *j*-th loci of the DNA profile evidence and $y \in \mathbb{R}$ is the value of the *j*-th loci of the DNA profile reference. Further by breaking down the formulation through doing simple multiplication operations, it is obtained

$$\mu(x_j, y_j) = \frac{\frac{1}{2}(x_j - y_j + 0.4)}{0.2} \quad . \tag{3}$$

And by dividing and multiplying the coefficients, finally the simple linear form is begotten

$$\mu(x_j, y_j) = 2.5x_j - 2.5x_j + 1. \tag{4}$$

The similarity between two DNA alleles is thus calculated as the average of the similarity of the entire locus, which in turn is arithmetic mean, which is expressed as

$$t_i = \frac{\Sigma_{j=1}^{N} \mu(x_j, y_j)}{N}, \tag{5}$$

where t_i is the value of similarity between DNA profile evidence and DNA profile reference of the *i*-th individual. The next step to obtain the STR similarity value is to calculate the similarity between t_i and r_i ($\in \mathbb{R}$) which is the family reference value of the *i*-th individual through Mamdani FIS [9] which consist of 9 rules as follow.

1. If t_i is *low* and r_i is *low* then STR Similarity is *low*,
2. If t_i is *low* and r_i is *medium* then STR Similarity is *low*,
3. If t_i *low* is and r_i *high* is then STR Similarity is *medium*,
4. If t_i *medium* is and r_i is *low* then STR Similarity is *medium*,
5. If t_i is *medium* and r_i is *high* then STR Similarity is *medium*,
6. If t_i is *medium* and r_i is *high* then STR Similarity is *medium*,
7. If t_i is *medium* and r_i is *low* then STR Similarity is *medium*,
8. If t_i is *high* and r_i is *medium* then STR Similarity is *medium*,
9. If t_i is *high* and r_i is *high* then STR Similarity is *high*.

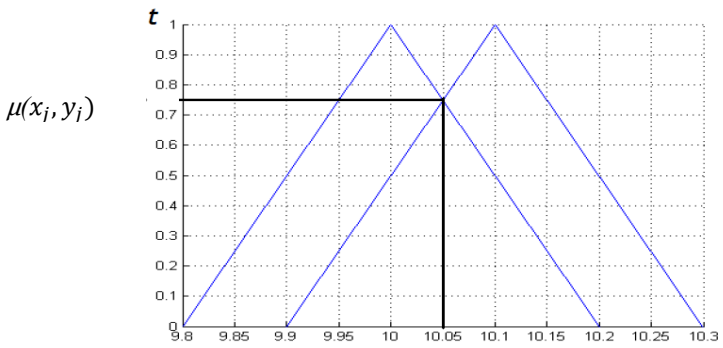

Figure 2. Fuzzy Similarity between Two Locus

For the antecedent the fuzzy membership function for *low*, *medium*, and *high* are given by Figure 3. The *low* membership function value is 1 when t_i is 0 until 0.18, and become 0 when t_i is 0.3. The *medium* membership function has its peak at t_i is 0.35 and having zero value at t_i is 0.2 and 0.5. The *high* membership function has zero value at t_i is 0.4 and 1 at t_i is 0.5 until 1.

For the consequent the fuzzy membership function for *low*, *medium*, *high* are given by Figure 4. The *low* membership function value is 1 when similarity is 0, and becomes 0 when similarity is 0.4. The *medium* membership function has its peak at similarity is 0.5 and has zero value when similarity is 0.3 and 0.8. The *high* membership function has zero value at similarity is 0.8 and 1 at similarity is 1.

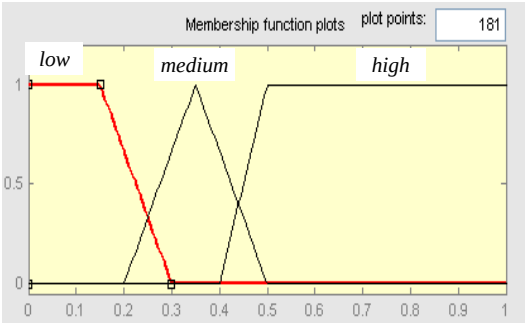

Figure 3. Fuzzy Membership Function for Antecedent

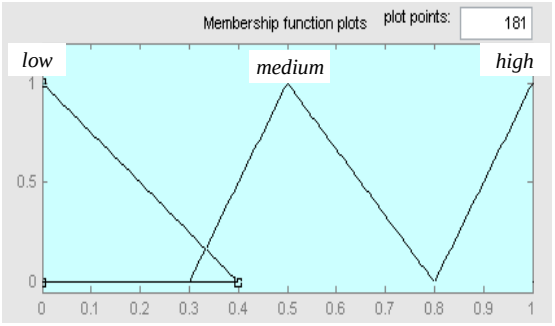

Figure 4. Fuzzy Membership Function for Consequent

3. THE STATISTICS OF TRIBAL INFERENCE

In forensic science, it has been the standard to profile DNA samples with the Short Tandem Repeat (STR) characteristic of certain loci [11-12]. The Federal Bureau Investigation (FBI) has chosen 15 markers (called locus) and amelogenin based on STR analysis to be the standard for human identification. The standard is called CODIS (Combined DNA Index System) [3], enables a uniform representation of DNA profile and exchange DNA between countries. The 15 locus are CSFF1PO, FGA, TH01, TPOX, VWA, D21S11, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D2S1338, plus amelogenin. In the frequent case where no identical STR match is found, any intelligence which could be inferred from the sample is very precious in aiding the investigation, especially in reducing the number of interviews and suspects. Information regarding the probable ethnicity and tribal information of an otherwise totally unknown offender may help investigators to look for targets and setting priorities for mass screening or interviews [3]. One of the information that could be inferred is the tribal information of the person from which the DNA sample is profiled. There have been a lot of studies concerning the variation of the STR profile proportion between tribal populations [13-14]. In one study, Gill et al. [15] have described distributions for three tribal groups employing four different single locus probes. The confirmation that the technique in [15] can provide qualified indications, but not tribal categorical conclusions are given, it has been described by Evett et al. [6]. In contrast, in this paper, the novel method to infer tribal information of a person is proposed.

4. THE PROPOSED TRIBAL INFERENCE

The development of the Fuzzy Inference System (FIS) for the DNA profile matching was focused on how to infer statistically useful information from the DNA profiles. The idea of the proposed method is to incorporate the statistical property of Probability Density Function of the distribution of allele markers of certain tribal information and incorporate the calculation using Takagi-Sugeno FIS [8]. Figure 5 shows the overview of how the algorithm is carried out. Start step is to compare the DNA profile evidence with DNA profile reference from DNA database. If all references have not been checked, pick the next reference profile. If all loci have not been compared then compare the next loci using the proposed FIS. If all locus have been compared, then check whether all references have been checked or not. If all references have been checked then calculate the average tribal similarity. After that, calculate the aggregate tribal similarity value.

The proposed FIS can be configured to use as many profiles as intended, and therefore giving the investigator the flexibility in testing different tribal groups. The following explains in detail the walkthrough of the framework that is embodied in the proposed FIS, which uses the allele marker's distribution as membership function. And then adjust the settings of the Takagi-Sugeno FIS and perform some calculations to obtain the value that reflects the tribal information similarity between two STR-Based DNA profiles. The hypothesis in embedding the tribal information inference in the proposed FIS is to somehow incorporate the statistical property of the tribal characteristic so that the tribal information is not lost in the STR similarity calculation. It has been noted that although M. R. Widyanto et al [4] were adequate in assigning a fuzzy similarity measure between two alleles, the statistical information regarding the two profile's tribal information is lost. This lost information is what the system is trying to regain. Initial statistical analysis of the DNA profiles showed an interesting pattern. It seems that the allele markers' probability distributions within a tribal group form a Gaussian distribution with different standard deviations between tribal groups.

Figure 6 depicts the plot of the distribution of the D19S433 allele markers of a tribal group in Indonesia i.e. Javanese.

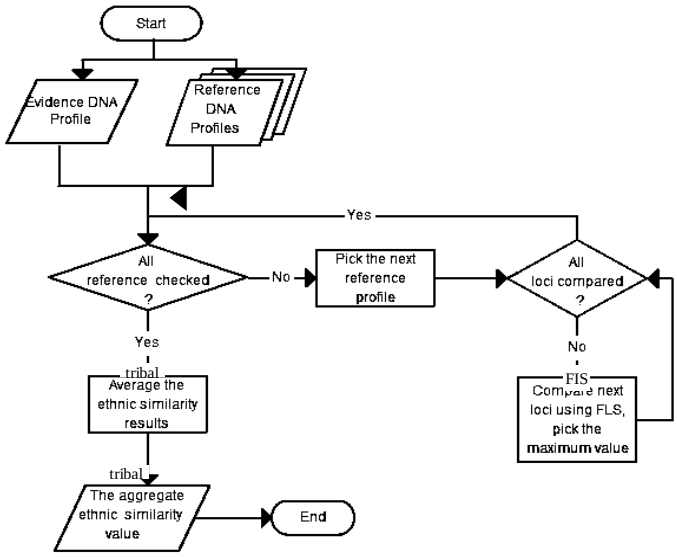

Figure 5. The Overview of the Tribal Information Inference

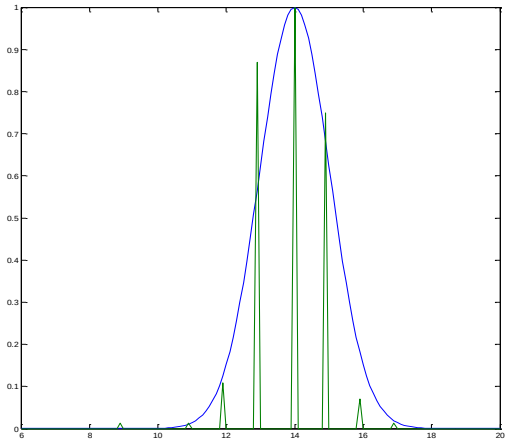

Figure 6. The Gaussian distribution of Allele Marker Value

Although it has been said that each tribal group has their unique allele distribution, there is a caveat that the difference in their distribution is very tight and hardly discriminative. Taking this statistical property, the logical hypothesis is to embed this Gaussian Probability Density Function (PDF) as the membership function in the proposed FIS that is built. This way, the Gaussian function acts similar to a membership function that assign a value to an allele marker, to which tribal group it belongs to. The Gaussian PDF is described as

$$f(x_j) = ae^{-\frac{(x_j-b)^2}{2c^2}} \quad (6)$$

Where in the light of the statistical information, $a = 1$, $b \in \mathbb{R}$ is the mean of the distribution (the peak of the curve), and $c \in \mathbb{R}$ is the standard deviation of the distribution. Figure 6 shows how the Gaussian membership function generated from the average and standard deviation of the data fits the marker distribution of the allele marker (marked by the spiky line fluctuations), thus modeling the probability density

function. Apparently, this approach could assign a membership value, or in this case, the probability that a marker value x_j falling within the Gaussian interval belongs to a particular tribal group. Figure 7 shows the discrete distribution of statistical tribal data: raw data and Figure 8 shows the adjusted discrete distribution of statistical tribal data: after natural adjustment.

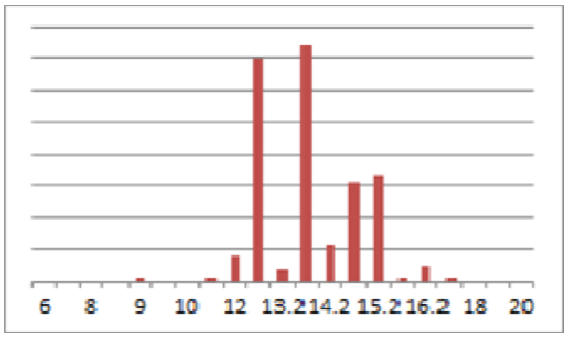

Figure 7. The Discrete Distribution of Statistical Tribal Data: Raw Data

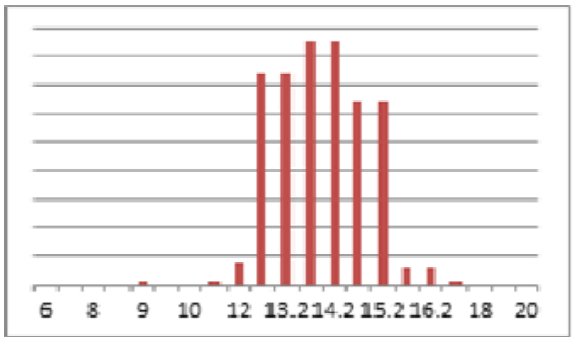

Figure 8. The Adjusted Discrete Distribution of Statistical Tribal Data: After Natural Adjustment

The proposed method consists of two cascade Fuzzy Inference Systems (FIS). The first which is based on Takagi-Sugeno FIS [8] is to calculate the tribal similarity, and the second that is based on Mamdani FIS [9] is to calculate the STR similarity. The Takagi-Sugeno FIS takes two inputs and two outputs, with the first output being the input for the second FIS and the second output is the tribal similarity that indicates how close the control profile and the query profile ethnically. This is in contrast to previous existing DNA matching algorithms which only produce one STR similarity score [4]. The tribal similarity score is only useful when the evidence profile is compared to a group of reference profiles whose ethnicity is known beforehand. Figure 9 shows the arrangement of the FIS.

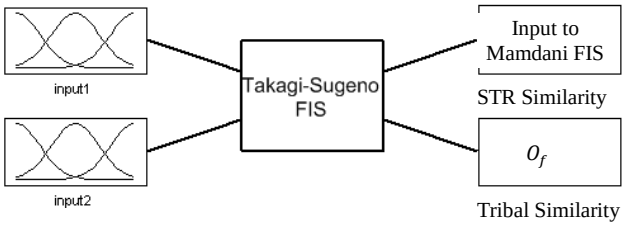

Figure 9. The Configuration of the Takagi-Sugeno FIS

5. EXPERMENTAL RESULTS

To examine the discriminating ability of the proposed fuzzy inference system (FIS), a prototype is implemented and an experiment using a real dataset containing 727 profiles from the people of Indonesia is conducted. The proposed cascade Fuzzy Inference System (FIS) is implemented in Microsoft .NET platform using C# as the programming language and using Visual Studio 2015 IDE. The database is stored in a MySQL DBMS storing 727 profile data. The Fuzzy Logic Library is based on FuzzyLogicLibrary (http://fuzzynet.sourceforge.net/), modified to some extent to support the features used in the FIS design which is previously unavailable in the default package. The database is accessed and mapped using the ORM Vici Coolstorage (http://viciproject.com/).

5.1. STR Similarity Experiment Design

The first experiment is to examine the STR Similarity performance by comparing the proposed Cascade FIS, the previous Fuzzy Short Tandem Repeat (STR) Similarity [4], and the conventional STR match [11]. The data is 727 of Indonesia STR Based DNA profiles. The experiment is design to calculate the similarity between each STR profile with his or her family STR profile i.e. father or mother or sibling available in the database. If the STR similarity is greater than 0.5 then the family relationship is correct otherwise false. The Table 1 below shows the correctness percentage of the three methods compared. The

table shows that the proposed method achieves the accuracy of 100%, meanwhile the previous Fuzzy STR Similarity method [4] also achieves the accuracy of 100%. In contract, the conventional STR match [11] only achieves 76.3% accuracy. This is supposedly the effect of noise in the process of analyzing the STR profile. The false calculation is due to the crisp calculation where two locus having slightly differ value got similarity score 0. Using fuzzy similarity measure, two alleles with small difference will still get a similarity score instead of a crisp 0. The result shows that the proposed Cascade Fuzzy Inference system does not under-grade the performance or previous Fuzzy STR Similarity as has been proposed in [4]. Moreover, the proposed Cascade Fuzzy Inference System is able to calculate the tribal similarity which cannot be found in the previous method. The experiment on how the experiment of the proposed method in calculating tribal similarity is discussed in the next sub-section.

Table 1. Correctness Percentage of Family Relationship Similarity

Compared Method	Correctness Percentage
The proposed Cascade Fuzzy Inference System	100%
The previous Fuzzy STR Similarity [4]	100%
The conventional STR Match [11]	76.3%

5.2. Tribal Similarity Experiment Design

The experiment objective is to see whether the proposed Tribal Similarity is able to indicate correctly that profiles which belong to different tribal groups should get lower tribal similarity score than those which belong to the same tribal group. This was done by first labeling the profiles to two separate groups, those who belong to tribal group *A*, and those which do not belong to group *A*. Then, every profile from each of the tribal groups was compared in different ways as illustrated in the diagrams below and the statistics of tribal similarity score results are observed.A total of 727 profiles were available, with 398 profiles known to be of Javanese tribal group. The rest of the profiles' ethnicity was unknown, so they were tested against the Javanese profiles. There were three modes of testing i.e. every Javanese compared against every Javanese, all non-Javanese against the Javanese, and every profile against themselves. Profiles with known ethnicity were needed as control group because the model assumes known ethnicity of the profile being used as comparison. For all the 3 modes of experimental comparison, various types of membership function discussed in the previous section are used in the experiments to see which one is most suitable for this particular statistical classification problem. The tables in the following show the experimental results. Naturally it would expected that the result from the same tribal group to score higher than those from different tribal group. Each modes and types of testing produce different results and the similarity scores of all the test within a mode is thenaveraged to see how good it is generally.

Table 2 shows tribal similarity average, max, and mean value.

Table 2. Tribal Similarity Average, Max, and Mean Value

Continuous Form				
	Mode 1	Mode 2	Mode 3	Δ
Average	0.89742	0.39634	0.59658	0.50108
Max	0.92329	0.52076	0.6214	0.40253
Min	0.5905	0.19099	0.2909	0.399951
Discrete Form				
	Mode 1	Mode 2	Mode 3	Δ
Average	0.78958	0.38558	0.58674	0.404
Max	0.88685	0.58182	0.68333	0.30503
Min	0.5908	0.28778	0.28864	0.30302
Adjusted Discrete Form				
	Mode 1	Mode 2	Mode 3	Δ
Average	0.92436	0.19038	0.59153	0.73398
Max	0.98764	0.58242	0.68395	0.40522
Min	0.60337	0.10023	0.20112	0.50314

The Δ column shows the difference of the similarity score between Mode 1 and 2. The positive Δ value confirmed the correctness of the proposed Takagi-Sugeno FIS. The benchmark is, the bigger the difference, the better discriminated are the profiles. The result shows that the use of discrete form of membership function to model the distribution of the allele markers resulted in averagely lower

discrimination of the profiles. However when adjusted, the results showed up a significant improvement. The most important thing is, the result shown that the proposed method of applying Takagi-Sugeno FIS with three different membership functions has excellent results.

5. CONCLUSIONS

A novel human STR (Short Tandem Repeat) similarity method using cascade statistical fuzzy rules with tribal information inference is proposed. The proposed method consists of two cascades Fuzzy Inference Systems (FIS). The first FIS is to discriminate the tribal similarity, and the second FIS is to calculate the STR similarity. By using the allele marker's statistical distribution probability density function as the membership function in the Fuzzy Rules of the first FIS, the new method makes it possible to tell the tribal similarity between two STR profiles. A 727 data acquired from tribal groups of Indonesia is used to examine the method produced promising result. In the first experiment the proposed method achieves 100% accuracy in recognizing family relationship which is the same performance compared to the previous Fuzzy STR Similarity method. In contract, the conventional STR match only achieves 76.3% accuracy. This is supposedly the effect of noise in the process of analyzing the STR profile. However the proposed method is able to calculate Tribal Similarity that cannot be performed by the previous Fuzzy STR Similarity. In the light of Indonesia's diverse tribal groups, these properties of the proposed method are able to be leveraged as a novel and new way to improve the versatility of existing DNA matching algorithm.

REFERENCES

[1] H. Kitakami, *et al.*, "Yamato and Asuka: DNA Database Manag.ement System," *Proc. of 28th Annual Hawaii Int. Conf. on System Sciences*, 2005.
[2] R. C. Michaelis, *et al.*, "A Litigator"s Guide to DNA: From the Laboratory to the Courtroom," Elsevier Academic Press, 2008.
[3] C. M. Ruitberg, *et al.*, "STRBase: A Short Tandem Repeat DNA Database for the Human Identity Testing Community," *Nucleid and Acid Research*, vol/issue: 29(1), 2001.
[4] M. R. Widyanto, *et al.*, "Various defuzzification methods on DNA similarity matching suing fuzzy inference system," *Journal of Advanced Computational Intelligence & Intelligent Informatics*, vol/issue: 14(3), 2010.
[5] A. L. Lowe, *et al.*, "Inferring ethnic origin by means of an STR profile," *Forensic Science International*, vol/issue: 119(1), 2001.
[6] I. W. Evett, *et al.*, "An investigation of the feasibility of inferring ethnic origin from DNA profiles," *J. Forensic Sci*, vol/issue: 32(4), 1992.
[7] Z. Liu and H. X. Li, "A probabilistic fuzzy logic system for modeling and control," *IEEE Trans Fuzzy Systems*, vol/issue: 13(6), pp. 848- 859, 2005.
[8] M. Sugeno, "Industrial applications of fuzzy control," Elsevier Science Pub. Co., 1985.
[9] E. H. Mamdani, "Application of fuzzy algorithms for control of simple dynamic plant," *Proc. of the Institution of Electrical Engineers*, vol/issue: 121(12), 1974.
[10] G. Taofik and D. Benslimean, "Fuzzy Similarity Measure," Springer, 2006.
[11] J. M. Butler, "STRBase and Information Resources on Forensic DNA," National Institute of Standards & Technology – U.S. Dept of Commerce, 2012.
[12] T. M. Clayton, *et al.*, "Analysis and interpretation of mixed strains using DNA STR profiling," *Forensic Science International*, vol/issue: 91(1), 1998.
[13] J. M. Butler, "Forensic DNA Typing: Biology and Technology Behind STR Markers," Academic Press, 2001.
[14] X. Fosellaa, *et al.*, "Assigning individuals to ethnic groups based on 13 STR loci," *Proc. of International Congress Series*, pp. 1261, 2004.
[15] M. Graydon, *et al.*, "Inferring ethnicity using 15 autosomal STR loci—Comparisons among populations of similar and distinctly different physical traits," *Forensic Science International: Genetics*, vol. 3, pp. 251–254, 2009.

BIOGRAPHIES OF AUTHORS

M. Rahmat Widyanto received B.Sc. from Faculty of Computer Science University of Indonesia in 1998, Master and Doctor Degree from Department of Computational Intelligence, Tokyo Institute of Technology Japan in 2003 and 2016 respectively. Currently he is a senior lecturer at Faculty of Computer Science University of Indonesia.

Reggio N. Hartono received B.Sc. from Swiss-German University, Master Degree from Faculty of Computer Science Universityof Indonesia. Currenty pursuing his Doctor Degree at Auckland University of Technology Australia.

Nurtami Soedarsono received drg. from Faculty of Dentistry University of Indonesia, and Doctor Degree from Tokyo Medical and Dental University Japan in 2006. Currently he is a senior lecturer at Departement of Oral Biology Faculty of Dentistry University of Indonesia.