Hindawi
Advances in Bioinformatics
Volume 2018, Article ID 8602513, 13 pages
https://doi.org/10.1155/2018/8602513

Research Article

Gaussian Fuzzy Number for STR-DNA Similarity Calculation Involving Familial and Tribal Relationships

Maria Susan Anggreainy,[1] **M. Rahmat Widyanto**,[1]
Belawati H. Widjaja,[1] **and Nurtami Soedarsono**[2]

[1] *Faculty of Computer Science, Universitas Indonesia, Depok Campus, West Java 16424, Indonesia*
[2] *Faculty of Dentistry, Universitas Indonesia, Salemba Campus, Jakarta 10430, Indonesia*

Correspondence should be addressed to Maria Susan Anggreainy; maria.susan61@ui.ac.id

Received 18 March 2018; Accepted 9 July 2018; Published 29 July 2018

Academic Editor: Florentino Fdez-Riverola

We performed locus similarity calculation by measuring fuzzy intersection between individual locus and reference locus and then performed CODIS STR-DNA similarity calculation. The fuzzy intersection calculation enables a more robust CODIS STR-DNA similarity calculation due to imprecision caused by noise produced by PCR machine. We also proposed shifted convoluted Gaussian fuzzy number (SCGFN) and Gaussian fuzzy number (GFN) to represent each locus value as improvement of triangular fuzzy number (TFN) as used in previous research. Compared to triangular fuzzy number (TFN), GFN is more realistic to represent uncertainty of locus information because the distribution is assumed to be Gaussian. Then, the original Gaussian fuzzy number (GFN) is convoluted with distribution of certain ethnic locus information to produce the new SCGFN which more represents ethnic information compared to original GFN. Experiments were done for the following cases: people with family relationships, people of the same tribe, and certain tribal populations. The statistical test with analysis of variance (ANOVA) shows the difference in similarity between SCGFN, GFN, and TFN with a significant level of 95%. The Tukey method in ANOVA shows that SCGFN yields a higher similarity which means being better than the GFN and TFN methods. The proposed method enables CODIS STR-DNA similarity calculation which is more robust to noise and performed better CODIS similarity calculation involving familial and tribal relationships.

1. Introduction

Genetics is the study of genes, genetic variation, and heredity in living organisms. Population genetics is a part of evolutionary biology and is a subfield genetic that deals with genetic differences within and between populations [1]. Variations in traits among human populations represent genetic differences that can be inherited from generation to generation. Population genetics is learning about genetic variation in the population, involving the examination and modeling of changes in the frequency of genes and alleles in populations over time and space [2]. Population genetics gives us the opportunity to step back and observe patterns of genetic change over time. Comparing populations to one another can lead to capturing how external factors trigger the evolution of a trait, as well as mapping variants associated with various traits within the population. Population genetics is another way of looking at DNA that can generate insight into the potential to benefit everyone. Many of the genes found in a population will be polymorphic, that is, will occur in a number of different alleles. Mathematical models are used to investigate and predict the occurrence of specific alleles or combinations of alleles in the population; the focus is by comparing data groups or populations or species, not individuals.

A population is a group of individuals with the same characteristics (species) that live in the same place and have the ability to reproduce among each other; evolution also works through populations [3]. Geneticists, on the other hand, view the population as a means or container for the exchange of alleles owned by the individuals of its members. The dynamic frequency of alleles in a population is of major concern in the study of population genetics

DNA regions with short repeat units (usually 2-6 base pairs in length) are called Short Tandem Repeats (STR). STRs are found surrounding the chromosomal centromere. STRs have proven to have several benefits that make them especially suitable for human identification [4]. STRs have become popular DNA markers because they are easily amplified by polymerase chain reaction (PCR) without the problem of differential amplification; that is, the PCR products for STRs are generally similar in amount, making analysis easier. An individual inherits one copy of an STR from each parent, which may or may not have similar repeat sizes. The number of repeats in STR markers can be highly variable among individuals, which make these STRs effective for human identification purposes [5].

Beginning in 1996, the FBI Laboratory launched a nationwide forensic science effort to establish core STR loci for inclusion within the national database known as CODIS (Combined DNA Index System). The 13 CODIS loci are CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11. These loci are nationally and internationally recognized as the standard for human identification. DNA STR markers used in this research were 15 CODIS loci with two additional loci, i.e., D19S433, and D2S1338 has additional loci for an extensive and powerful STR testing battery if required [6, 7].

A person's DNA profile can match DNA profile data similarity to another person. DNA profile plays an important role in solving problems related to the family's father and other family members [8, 9]. This method is a way that is legally used for solving to prove the validity of kinship or family ties of the person, identifying unknown body of war or natural disaster victims, and studying human population [10, 11].

In previous research, it has been noted that although M. R. Widyanto et al. [12–14] are quite sufficient in setting with triangular fuzzy number similarity of the size between the two alleles, the statistical information on the ethnicity of the two profiles' information is lost. To overcome the problem, this research employs novel methods to measure similarity between tribes that gives a better result than previous method. We proposed shifted convoluted Gaussian fuzzy number (SCGFN) and Gaussian fuzzy number (GFN) to represent each locus value as improvement of triangular fuzzy number (TFN) as used in previous research.

Research method was proposed in Section 2. Experimental results on three methods are shown in Section 3. Analyses of statistical and comparison tests are summarized in Section 4.

2. Proposed Research Method

2.1. Fuzzy Sets. Fuzzy sets are held as a basis for the theory of possibility. A fuzzy set A in x is formally defined as follows [15]:

$$\mathbf{A} = \{(\mathbf{x}, \boldsymbol{\mu}_{\mathbf{A}}(\mathbf{x})) \mid \mathbf{x}\, \boldsymbol{\varepsilon} \mathbf{X}\} \quad (1)$$

where x is the universe of discourse and is the membership degree of the x in A. When fuzzy set theory was presented, researches considered decision-making as one of the most attractive application fields of that theory [16].

2.2. Measurement of Similarity Values of Two-Individual STR-DNA. The calculation to find the STR-DNA similarity of two individuals (as shown by Figure 1) is the STR-DNA value of allele 1 of the individual in comparison with the allele value 1 STR-DNA of the reference and the STR-DNA value of allele 2 of the individual with the STR-DNA value of allele 2 of each reference locus. Then we find the intersection point value of the two alleles for each locus and then calculate the average value of the similarity of each locus.

2.3. Two-Individual Matching: Evidence versus Reference with TFN Similarity. A triangular fuzzy number (TFN) α can be defined by a triplet ($a1$, $a2$, and $a3$). The triangular fuzzy number is used to represent uncertainty resulting from imprecision of polymerase chain reaction (PCR) machine. The membership function $\mu a(x)$ is [17]

$$\mu_a(x) = \begin{cases} 0, & x < a_1, a_3 < x \\ \dfrac{x - a_1}{a_2 - a_1}, & a_1 \le x \le a_2 \\ \dfrac{x - a_3}{a_2 - a_3}, & a_2 \le x \le a_3 \end{cases} \quad (2)$$

where $0 \le a1 \le a2 \le a3 \le 1$, $a1$ and $a3$ stand for the lower and upper values of the support of α, respectively, and $a2$ stands for the modal values. Value of every DNA loci is represented by fuzzy triangular number where the fuzziness value is set to be 0.4 through experiments and the center of the fuzziness is the value of the corresponding loci. The similarity value between an allele of DNA profile evidence and DNA profile reference is given by

$$t = \frac{a3 - a1}{2(a3 - a2)} \quad (3)$$

where the value of the first allele < value of the second allele; t is intersection of the two alleles,

a2 is STR-DNA value of the first allele, a3 is a2 + 0.2, and a1 is STR-DNA value of the second allele -0.2.

If t is a result of a negative value calculation, then t is considered zero because it means there is no intersection on both STR values; therefore, t values stay at interval [0, 1].

The following example shows the geometric calculation of the individual intersection points and the reference (as shown by Figure 2) with the D8s1179 locus where the allele values are individual STR-DNA = 13 and the allele value of one in reference STR-DNA = 13.3 and the allele value of two on the individual STR-DNA = 14 and the value of the two alleles in the reference STR-DNA = 14.1.

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

$$t_i = \frac{\sum_{j=1}^{N} \mu(x_i, y_j)}{N} \quad (4)$$

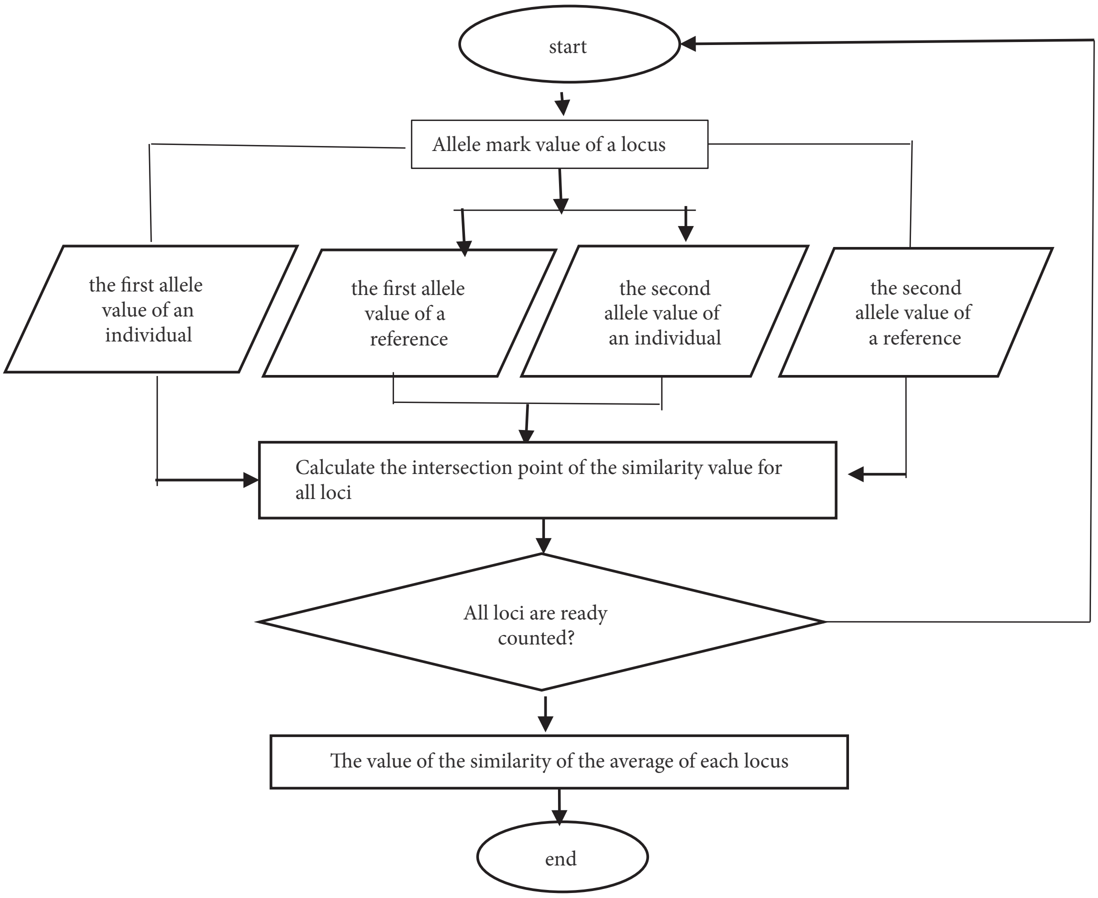

start
Allele mark value of a locus
the first allele value of an individual
the first allele value of a reference
the second allele value of an individual
the second allele value of a reference
Calculate the intersection point of the similarity value for all loci
All loci are ready counted?
The value of the similarity of the average of each locus
end

Figure 1: Calculation flow of similar two individuals.

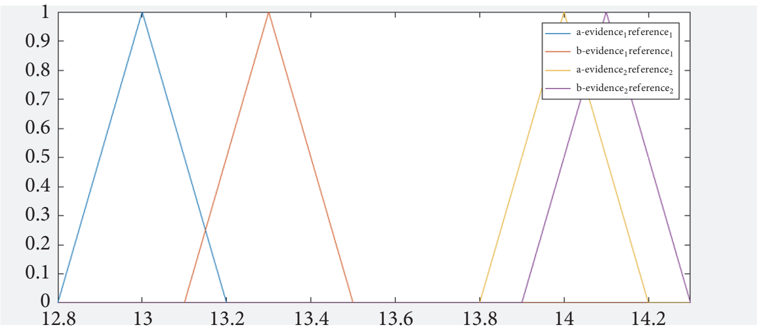


Figure 2: The fuzzy triangular number.

where t_i is the value of similarity between DNA profile individual and DNA profile reference of the ith individual, x_i is a vector of DNA profile individual, and y_j is a vector of DNA profile reference. The vectors x_i and y_j are the N ($\in N$)-dimensional vector consisting of the value of 15 loci without amelogenin as has been used by Federal Bureau of Investigation (FBI).

2.4. Two-Individual Matching: Evidence versus Reference with GFN Similarity. To improve the capability of locus matching, we propose GFN (Gaussian fuzzy number) similarity. Compared to traditional triangular fuzzy number (TFN), GFN is more realistic to represent uncertainty of locus information because the distribution is assumed to be Gaussian. The Gaussian fuzzy function transforms the original values into a normal distribution. The midpoint of the normal distribution defines the ideal definition for the set, assigned a 1, with the remaining input values decreasing in membership as they move away from the midpoint in both the positive and negative directions. The input values decrease in membership from the midpoint until they reach a point where the values move too far from the ideal definition and definitely not in

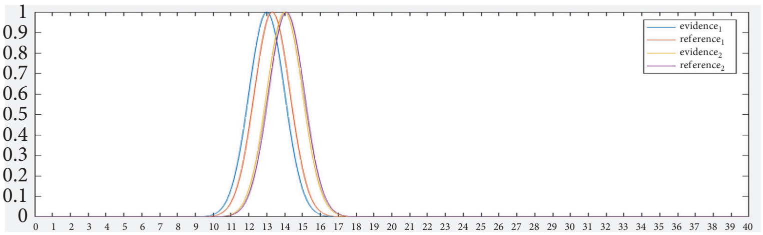


Figure 3: Fuzzy Gaussian similarity.

the set and are therefore assigned zeros. The fuzzy Gaussian function is given below [18]:

$$f(x) = ae^{-(x-\mu_f)^2/2\sigma_f^2} \tag{5}$$

A Gaussian Membership function is specified by two parameters: a Gaussian membership function is determined complete by μ and σ; μ represents the membership ship center (the peak of the curve) and σ determines the membership function width.

Intersection from two Gaussian functions is as follows [19]:

$$v = \begin{cases} 1 & if\ \mu 2 \leq \mu 1 \\ \exp\left[-\left[\left(\dfrac{\mu 1 - \mu 2}{\sigma 1 + \sigma 2}\right)\right]^2\right] & if\ \mu 2 < \mu 1 \end{cases} \tag{6}$$

where $\mu 1$ is value of STR-DNA from individual, $\mu 2$ is value of STR-DNA from reference, $\sigma 1$ is the sigma value of the individual, and $\sigma 2$ is the sigma value of the reference.

The following example shows Gaussian of the individual intersection points and the reference (as shown by Figure 3) with the D8s1179 locus where the allele values are individual STR-DNA = 13 and the allele value of one in reference STR-DNA = 13.3 and the allele value of two on the individual STR-DNA = 14 and the value of the two alleles in the reference STR-DNA = 14.1.

2.5. Two-Individual Matching: Evidence versus Reference with SCGFN Similarity. To improve the capability of locus matching in which ethnic information is involved, we propose SCGFN (shifted convoluted Gaussian fuzzy number) similarity. The original Gaussian fuzzy number (GFN) is convoluted with distribution of certain ethnic locus information. Therefore, the new SCGFN more represents ethnic information compared to original GFN.

The convolution function is a multiplication of the individual fuzzy Gaussian locus function and the fuzzy Gaussian approximation of the population locus. The fuzzy Gaussian function of the population locus approximation is obtained by extracting the mean value at which the mean value of the fuzzy Gaussian population locus is the STR-DNA value of the most population density and deviation of the particular population. Fuzzy Gaussian individual locus obtained, where the mean is the STR-DNA value of an individual locus, with the standard deviation value is the mean value minus 2.

The convolution is a mathematical operation on two functions (f and g) to produce a third function, that is, typically viewed as a modified version of one of the original functions, giving the integral of the pointwise multiplication of the two functions as a function of the amount that one of the original functions is translated [20]. The function f is obtained from the individual and the function g is derived from the reference

$$f(x) = ae^{-(x-\mu_f)^2/2\sigma_f^2}; \tag{7}$$

and

$$g(x) = ae^{-(x-\mu_g)^2/2\sigma_g^2} \tag{8}$$

convolution operator is [21]

$$\begin{aligned} P_{f\otimes g}(x) &= F^{-1}\left[F(f(x))\,F(g(x))\right] \\ &= ae^{-(x-(\mu_f+\mu_g))^2/2(\sigma_f^2+\sigma_g^2)} \end{aligned} \tag{9}$$

where a is the height of fuzzy = 1. For counting means,

$$\mu_{f\otimes g}(x) = \mu_f + \mu_g \tag{10}$$

and standard deviation is as follows:

$$\sigma_{f\otimes g}(x) = \sqrt{\sigma_f^2 + \sigma_g^2} \tag{11}$$

The convoluted fuzzy number will replace the fuzzy number as individual locus representation value. Therefore, to get a stronger tribal relationship individual similarity calculation value is involved then the convoluted Gaussian fuzzy number is shifted approaching to the tribal population reference fuzzy number. The new shifted convoluted fuzzy number is called shifted convoluted Gaussian fuzzy number (SCGFN). This SCGFN is a new representation value of individual locus. Obtaining the mean value of SCGFN is to compare the mean value of individual fuzzy Gaussian (μ_i) with the mean value of fuzzy Gaussian approximation of the locus population (μ_{ip}). And then, the convoluted fuzzy number is shifted approaching to the fuzzy Gaussian approximation of locus

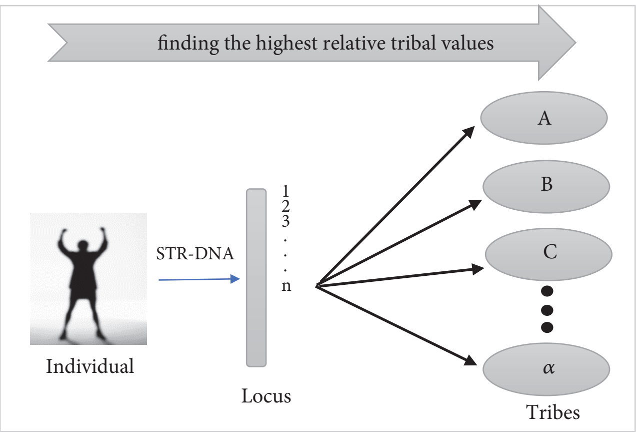


Figure 4: Tribal inference system architecture design.

population of certain tribe. The algorithm for shifting SCGFN is given below:

```
if $(\mu_{ip} < \mu_i)$
    $\mu_{scgfn} = \mu_i - 0.02 * |\mu_{ip} - \mu_i|$;
        elseif $(\mu_{ip} > \mu_i)$
    $\mu_{scgfn} = \mu_i + 0.02 * |\mu_{ip} - \mu_i|$;        (12)
else
    $\mu_{scgfn} = \mu_i$;
end;
```

The standard deviation value of SCGFN is the sum of the standard deviation value of the individual fuzzy Gaussian number (σi) with the standard deviation value of the fuzzy Gaussian approximation of the population locus of certain tribe (σip). The following formula is for the SCGFN standard deviation:

$$\sigma_{scgfn} = \sigma_{ip} + \sigma_i \tag{13}$$

2.6. Measuring Tribal Relative Value from a DNA Profile. In general, the work flow of the tribal inference system is to find the value of the tribal similarity done by calculating the average value of the point of intersection of the value of the individual similarity to the value of the tribal population in the database of 15 loci. The tribe having the highest similarity value to the individual profile will be selected as the ethnic estimation of the profile. Workflow process can be seen in Figure 4.

Tribal matching with triangular fuzzy number is obtained from the intersection of fuzzy triangular individual and triangular fuzzy population approximation. From each tribal population the mean intersection value of the individual fuzzy triangular and fuzzy triangular population approximation for each locus are calculated, and to determine the ethnic population of the individual the greatest value of the mean value of each locus of a tribal population triangular fuzzy individuals is obtained by using formula (3). Fuzzy triangular population approximation is also obtained by using formula (3) where a2 is the STR-DNA value of the largest population of density, a3 is a2 + 0.2, and a1 is a2 - 0.2.

Table 1: The results of the output on population A.

Locus: D3S1358	
Allele :1	
STR-DNA	Number of Individuals
[13]	[3]
[14]	[10]
[15]	[27]
[16]	[31]
[17]	[8]
[18]	[1]

Example is shown in Table 1.

The output for population A is at locus D3S1358 and on allele 1. The value of 13th, 14th, 15th, 16th, 17th, and 18th STR-DNA are 3, 10, 27, 31, 8, and 1 individuals, respectively, as the 16th STR-DNA shows the most number of individuals at 31, it can be concluded that the value of a1 = 15.8, a2 = 16, and a3 = 16.2.

2.7. Tribal Matching with GFN. Tribal matching with Gaussian Fuzzy Similarity was obtained from individual fuzzy Gaussians with fuzzy Gaussian population approximation. Gaussian fuzzy individuals are obtained by using equation of formula (7), where μ_f is individual STR-DNA value and σ_f = 1. Gaussian fuzzy population is obtained by using equation of formula (7), where μ_f means the distribution of the number of individuals and σ_f is the standard deviation from the distribution of the number of individuals. Calculation of standard deviation is as follows:

$$s = \frac{\sqrt{n \sum x_i^2 - \sum (x_i)^2}}{n(n-1)} \tag{14}$$

where n is number of STR-DNA values and x_i is the number of individuals of a locus population.

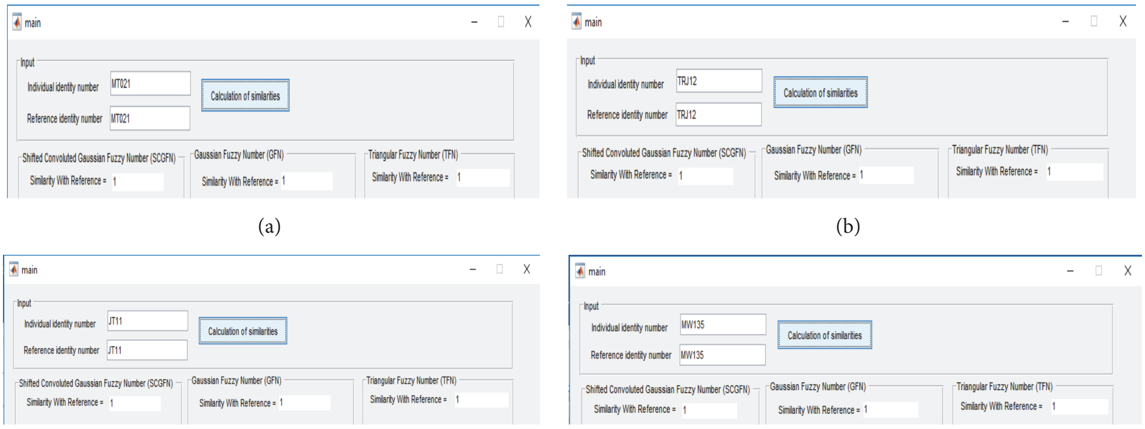
2.8. Tribal Matching with SCGFN. Shifted convoluted Gaussian fuzzy number (SCGFN) will replace the individual fuzzy number. SCGFN is a new individual locus with increasingly strong ethnicity. Tribal matching with SCGFN is calculating the value of similarity intersection between the corresponding SCGFN and population's fuzzy number as in tribal matching in GFN. Then look for the maximum value for all tribes. The maximum tribal value means the individual's tribal value.

3. Experimental Results

The DNA profile data to be entered into the database system is a PCR based DNA identification profile consisting of 15 loci, excluding amelogenin, each consisting of two alleles for each locus. The DNA profile used as an input to a tribal inference system is an Indonesian DNA with a total of 240 DNA data comprising Java (A), Malay (B), Mentawai (C), and Toraja

Table 2: Example results of individual similarity with family reference.

No	Individual1 Id	Relationship	Individual2 Id	Relationship	SCGFN	TFN	GFN
1	0800103	mother	0800101	child	0.938856	0.525	0.798565
2	08006002	mother	0800603	child	0.822547	0.383333	0.684361
3	0800401	mother	0800403	child	0.923285	0.433333	0.747206
4	0800903	mother	0800902	child	0.931517	0.5	0.862006
5	08002F	mother	08002C	child	0.856002	0.333333	0.614684
6	08002M	father	08002C	child	0.838193	0.533333	0.738193
7	0800901	father	0800902	child	0.872816	0.566667	0.736653
8	0800102	father	0800101	child	0.850089	0.383333	0.703589
9	0800402	father	0800403	child	0.848397	0.533333	0.781071
10	08006001	father	0800603	child	0.818216	0.516667	0.739065


main
Input
Calculation of similarities
Shifted Convoluted Gaussian Fuzzy Number (SCGFN)
Gaussian Fuzzy Number (GFN)
Triangular Fuzzy Number (TFN)

Individual identity number MT021
Reference identity number MT021
Similarity With Reference = 1 (SCGFN)
Similarity With Reference = 1 (GFN)
Similarity With Reference = 1 (TFN)

(a)

Individual identity number TRJ12
Reference identity number TRJ12
Similarity With Reference = 1
Similarity With Reference = 1
Similarity With Reference = 1

(b)

Individual identity number JT11
Reference identity number JT11
Similarity With Reference = 1
Similarity With Reference = 1
Similarity With Reference = 1

(c)

Individual identity number MW135
Reference identity number MW135
Similarity With Reference = 1
Similarity With Reference = 1
Similarity With Reference = 1

(d)

Figure 5: Examples from the same person similarity calculation.

(D) tribes. The experiments have been done using the Matlab R2016b. The experiment was conducted with four cases.

3.1. Same Person. From 240 pieces of data experiments were conducted with the same people, with SCGFN, GFN, and TFN; the similarity values are equal 1. Figure 5 is an example of whether the individual identity and reference entered are the same person and the result of the similarity obtained is 1:

(a) Input individual id MT021 and reference id number MT021

(b) Input individual id TRJ12 and reference id number TRJ12

(c) Input individual id JT11 and reference id number JT11

(d) Input individual id MW135 and reference id number

3.2. People Who Have Family Relationships. DNA profiles tested in both biological parents are father and mother. From the experiment, the average individual similarity with family reference is obtained: SCGFN 87%, TFN 39%, and GFN 74%. Table 2 shows ten instances of the result of the similarity of an individual with a reference being a mother or father.

3.3. People Who Belong to the Same Tribe. From the experiment, the average similarity of two individuals who have the same tribe is obtained: SCGFN 89.6%, TFN 21%, and GFN 65.14%. Table 3 shows ten instances of the result of the similarity of two individuals from the same tribe.

3.4. Certain Tribal Populations. Population data consists of four tribes where the number of people in tribe A is 80, the number of people in tribe B is 100, the number of people in tribe C is 20, and the number of people in tribe D is 40. Figure 6 shows an example of the tribal population similarity calculation of the individual identity JT19. From Figure 6 it can be seen that SCGFN, GFN, and TFN displaying the tribe of JT11 are tribe A.

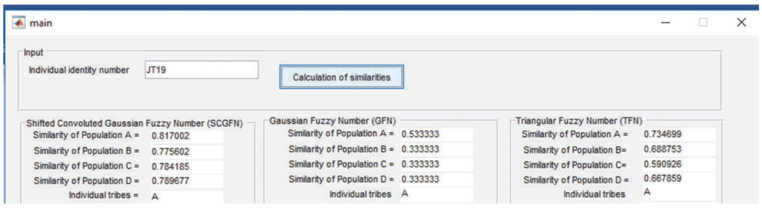
Table 4 shows the result of similarity values with a certain tribal population with SCGFN, TFN, and GFN. From 80 experiments on the A tribe, the average tribe population was found to be 79% with fuzzy convolution, 45% with fuzzy triangular, and 71% with fuzzy Gaussian. The average B population of 100 experiments were 81% with fuzzy convolution, 46% with fuzzy triangular, and 76% with fuzzy Gaussian. The average C population of 40 experiments were 80% with fuzzy convolution, 45% with fuzzy triangular, and 73% with fuzzy

Table 3: Example results of two individuals in the same tribe.

No	Individu1 Id	Tribe	Individu 2 Id	Tribe	SCGFN	TFN	GFN
1	MT097	C	MT104	C	0.944695	0.85	0.911584
2	TRJ19	D	TRJ22	D	0.934938	0.6	0.847523
3	JT11	A	JT13	A	0.903104	0.6	0.746172
4	MW132	B	MW135	B	0.869267	0.216667	0.694363
5	JT11	A	JT12	A	0.856067	0.333333	0.669071
6	JT11	A	JT17	A	0.890797	0.233333	0.609616
7	MT021	A	MT028	A	0.924174	0.333333	0.689898
8	MT019	B	MT036	B	0.873376	0.283333	0.607031
9	MW128	B	MW123	B	0.886405	0.366667	0.671862
10	MT017	C	MT018	C	0.878135	0.333333	0.66973

Table 4: The result of similarity values with a certain tribal population.

Tribe	SCGFN	TFN	GFN
A	0,79	0,45	0,714639375
B	0,81	0,46	0,7615544
C	0,8071442	0,44999925	0,73734225
D	0,846345667	0,64	0,794088667



Input
Individual identity number: JT19

Shifted Convoluted Gaussian Fuzzy Number (SCGFN)
Similarity of Population A = 0.817002
Similarity of Population B = 0.775602
Similarity of Population C = 0.784185
Similarity of Population D = 0.789677
Individual tribes = A

Gaussian Fuzzy Number (GFN)
Similarity of Population A = 0.533333
Similarity of Population B = 0.333333
Similarity of Population C = 0.333333
Similarity of Population D = 0.333333
Individual tribes A

Triangular Fuzzy Number (TFN)
Similarity of Population A = 0.734699
Similarity of Population B= 0.688753
Similarity of Population C= 0.590926
Similarity of Population D = 0.667859
Individual tribes A

Figure 6: Examples of tribal population certain.

Gaussian. The average D population of 20 experiments were 84% with fuzzy convolution, 64% with fuzzy triangular, and 79% with fuzzy Gaussian.

4. Analysis of Statistical and Comparison Tests

To perform analysis of the test results that have been done, statistical tests were used. To know the difference of average value of STR-DNA similarity value of the three methods used in this study, ANOVA and comparative test were used. Interpretation of ANOVA test is that if the test results show that H0 failed to be rejected (no difference), then post hoc test is not done. Conversely, if the test results indicate H0 is rejected (there is a difference), then a post hoc advanced test should be performed. To give a clear explanation why SCGFN is better than GFN and TFN, analysis of statistical and comparison test with Tukey method in ANOVA is provided. Statistical tests were performed using Minitab 16. Statistical tests were performed for 3 cases.

4.1. Individuals Who Have Family Relationships. To find out the different methods used in this method, we used ANOVA and comparative tests. In the one-way ANOVA test, there is only one independent variable for this case as independent variables are individuals who have family relationships. The summary of variance analysis in Algorithm 1 was obtained; P value ≤ 0.001. Associated with the level of significance $(\alpha) = 0.05$, obtained $p < \alpha$ means H0 is rejected so it can be concluded that there is a difference between the three methods. To determine which method is better than the other method, it is further tested by the Tukey method.

In Algorithm 2 it can be seen that

(i) TFN < SCGFN, because it does not contain zero and center negative;

(ii) GFN < SCGFN, because it does not contain zero and center negative;

(iii) GFN > TFN, because it does not load zero and center positive.

Then it can be concluded that TFN < GFN < SCGFN. From the result of similarity test with three methods and boxplot obtained in Figure 7, it can be seen that SCGFN method used in this research has higher similarity value in comparison with GFN and TFN method.

4.2. Individuals Who Belong to the Same Tribe. The summary of variance analysis in Algorithm 3 was obtained; P value ≤

```
Source    DF       SS        MS      F        P
Factor     2  0,90267   0,45133  99,19   0,0005
Error     27  0,12286   0,00455
Total     29  1,02553

S = 0,06746   R-Sq = 88,02%   R-Sq(adj) = 87,13%
```

Algorithm 1: ANOVA individual test results of those who have family relationship.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 98,04%

SCGFN subtracted from:

        Lower     Center     Upper     +---------+---------+---------+---------
TFN  -0,49007   -0,41520  -0,34033     (-- * --)
GFN  -0,20433   -0,12945  -0,05458                (-- * --)
                                       +---------+---------+---------+---------
                                    -0,50     -0,25      0,00      0,25

TFN subtracted from:

       Lower     Center     Upper     +---------+---------+---------+---------
GFN  0,21087   0,28575   0,36062                               (-- * --)
                                      +---------+---------+---------+---------
                                   -0,50     -0,25      0,00      0,25
```

Algorithm 2: Comparison test with Tukey method.

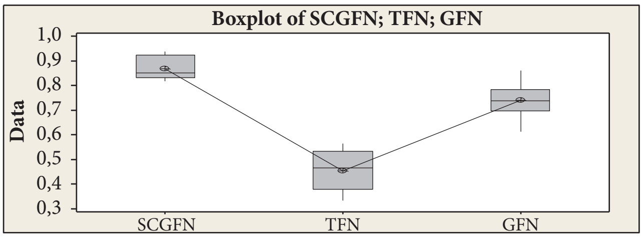


Figure 7: Boxplot people who have family relationships.

0.001. Associated with the level of significance (α) = 0.05 or confidence level 95%, obtained p < α means H0 is rejected so it can be concluded that there is a difference between the three methods. To determine which method is better than the other method, it is further tested by the Tukey method.

In Algorithm 4 it can be seen that

(i) TFN < SCGFN, because it does not contain zero and center negative;

(ii) GFN < SCGFN, because it does not contain zero and center negative;

(iii) GFN > TFN, because it does not load zero and center positive.

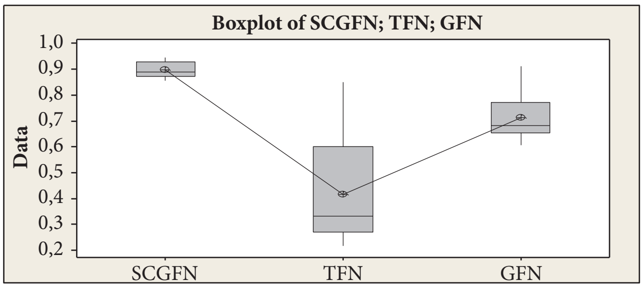


Figure 8: Boxplot Individuals who belong to the same tribe. Comparison test with Tukey method.

Then it can be concluded that TFN < GFN < SCGFN. From the result of similarity test with three methods and boxplot obtained in Figure 8, it can be seen that SCGFN method used in this research has higher similarity value in comparison with GFN and TFN method.

4.3. Certain Tribal Populations. Testing is done with 720 data, that is, 240 data with 3 methods. The summary of variance analysis for A, B, C, and D in Algorithm 5 was obtained; P value ≤ 0.001. Associated with the level of significance

```
Source  DF      SS      MS      F       P
Factor   2  1,1783  0,5891  34,24  0,0005
Error   27  0,4646  0,0172
Total   29  1,6429

S = 0,1312    R-Sq = 71,72%    R-Sq(adj) = 69,63%
```

Algorithm 3: Individual test results with the same tribe using ANOVA.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 98,04%

SCGFN subtracted from:

       Lower    Center    Upper   -+---------+---------+---------+--------
TFN  -0,6267  -0,4811  -0,3355   (---- * ----)
GFN  -0,3300  -0,1844  -0,0388              (---- * ----)
                                  -+---------+---------+---------+--------
                                -0,60     -0,30      0,00      0,30

TFN subtracted from:

       Lower   Center   Upper   -+---------+---------+---------+--------
GFN   0,1511   0,2967  0,4423                            (---- * ----)
                                -+---------+---------+---------+--------
                              -0,60     -0,30      0,00      0,30
```

Algorithm 4: Comparison test with Tukey method.

(α) = 0.05, obtained p < α means H0 is rejected so it can be concluded that there is a difference between the three methods.

The experiment was conducted with 3 methods, where the method of SCGFN is method 1, TFN method is method 2, and GFN is method 3. To determine which method is better than other method then each tribe is tested further. In Algorithm 6 the comparison of three methods in ANOVA test results in tribal population A can be explained:

(1) TFN < SCGFN because it does not load zero and center negative.

(2) GFN < SCGFN because it does not load zero and center negative.

(3) GFN > TFN because it does not contain zero and positive center.

So it can be concluded that TFN < GFN < SCGFN which means SCGFN is better than GFN and GFN is better than TFN.

In Algorithm 7 the comparison of three methods in the tribal population B can be explained:

(1) TFN < SCGFN because it does not load zero and center negative.

(2) GFN = SCGFN because it loads zero.

(3) GFN > TFN because it does not contain zero and positive center.

So it can be concluded that TFN < (GFN = SCGFN), which means that SCGFN and GFN methods are equal and better than TFN.

In Algorithm 8 the comparison of 3 methods in tribal population C can be explained:

(1) TFN < SCGFN because it does not load zero and center negative.

(2) GFN < SCGFN because it does not load zero and center negative.

(3) GFN > TFN because it does not contain zero and positive center.

So it can be concluded that TFN < GFN < SCGFN which means SCGFN is better than GFN and GFN is better than TFN.

In Algorithm 9 the comparison of three methods in the tribal population D can be explained:

(1) TFN < SCGFN because it does not load zero and center negative.

(2) GFN = SCGFN because it loads zero.

```
Analysis of Variance for A , using Adjusted SS for Tests

Source   DF   Seq SS   Adj SS   Adj MS       F        P
Metode    2   7,6061   7,6061   3,8030   438,56   0,0005
Error   251   2,1766   2,1766   0,0087
Total   253   9,7827

S = 0,0931218   R-Sq = 77,75%   R-Sq(adj) = 77,57%

Analysis of Variance for B , using Adjusted SS for Tests

Source   DF   Seq SS   Adj SS   Adj MS       F        P
Metode    2   7,9691   7,9691   3,9845   357,37   0,0005
Error   251   2,7986   2,7986   0,0111
Total   253  10,7677

S = 0,105592   R-Sq = 74,01%   R-Sq(adj) = 73,80%

Analysis of Variance for C , using Adjusted SS for Tests

Source   DF   Seq SS   Adj SS   Adj MS       F        P
Metode    2   8,7934   8,7934   4,3967   444,01   0,0005
Error   251   2,4855   2,4855   0,0099
Total   253  11,2788

S = 0,0995100   R-Sq = 77,96%   R-Sq(adj) = 77,79%

Analysis of Variance for D , using Adjusted SS for Tests

Source   DF   Seq SS   Adj SS   Adj MS       F        P
Metode    2   7,6350   7,6350   3,8175   328,52   0,0005
Error   251   2,9167   2,9167   0,0116
Total   253  10,5517

S = 0,107798   R-Sq = 72,36%   R-Sq(adj) = 72,14%
```

Algorithm 5: ANOVA test results for a particular tribe.

```
Tukey 95,0% Simultaneous Confidence Intervals
Response Variable A
All Pairwise Comparisons among Levels of Metode
Metode = 1 subtracted from:

Metode    Lower   Center    Upper  -------+---------+---------+---------
2       -0,4276  -0,3940  -0,3605 (*-)
3       -0,0957  -0,0622  -0,0286             (-*)
                                   -------+---------+---------+---------
                                      -0,25      0,00      0,25

Metode = 2  subtracted from:

Metode  Lower   Center   Upper   -------+---------+---------+---------
3       0,2984  0,3319   0,3653                              (*-)
                                 -------+---------+---------+---------
                                    -0,25      0,00      0,25
```

Algorithm 6: ANOVA test results on tribal population A.

```
Tukey 95,0% Simultaneous Confidence Intervals
Response Variable B
All Pairwise Comparisons among Levels of Metode
Method = 1   subtracted from:

Method    Lower     Center     Upper   -------+---------+---------+---------
2        -0,4301    -0,3921    -0,3541 (*-)
3        -0,0738    -0,0358    0,0023               (-*)
                                       -------+---------+---------+---------
                                            -0,25     0,00       0,25


Method = 2   subtracted from:

Method   Lower     Center    Upper   -------+---------+---------+---------
3        0,3184   0,3563    0,3942                                (*-)
                                     -------+---------+---------+---------
                                          -0,25     0,00      0,25
```

Algorithm 7: ANOVA test results on tribal population B.

```
Tukey 95,0% Simultaneous Confidence Intervals
Response Variable C
All Pairwise Comparisons among Levels of Metode
Method = 1   subtracted from:

Method    Lower     Center     Upper    --------+---------+---------+--------
2        -0,4482   -0,4123   -0,3765     (-*)
3        -0,0745   -0,0386   -0,0028                    (*-)
                                        --------+---------+---------+--------
                                              -0,25      0,00      0,25


Method = 2 subtracted from:

Method    Lower   Center   Upper   --------+---------+---------+--------
3        0,3380   0,3737   0,4094                                (*)
                                   --------+---------+---------+--------
                                         -0,25     0,00       0,25
```

Algorithm 8: ANOVA test results on tribal population C.

(3) GFN > TFN because it does not contain zero and positive center.

So it can be concluded that TFN < (GFN = SCGFN), which means that SCGFN and GFN methods are equal and better than TFN.

5. Conclusions

In this research, the experiments were conducted to find a better method to obtain higher individual similarity values and to find stronger tribal properties. To improve the capability of locus matching, SCGFN and GFN have been proposed. It performed fuzzy number similarity of the size between the two alleles. Experiments were done for the following cases: people with family relationships, people of the same tribe, and certain tribal populations. In these three cases, ANOVA shows the difference in similarity between SCGFN, GFN, and TFN with a significant level of 95%. In the case of people with family relationship and the case of people of the same tribe with Tukey method in ANOVA shows that SCGFN yields a higher similarity which means better than the GFN and TFN methods. While in the case of certain tribal population with Tukey method in ANOVA shows in population A and population C, SCGFN better than GFN and TFN, whereas, in population B and population D, SCGFN is equal to GFN and better than TFN. The proposed method enables CODIS STR-DNA similarity calculation which is more robust to noise and performed better CODIS similarity calculation involving familial and tribal relationships.

```
Tukey 95,0% Simultaneous Confidence Intervals
Response Variable D
All Pairwise Comparisons among Levels of Metode
Metode = 1   subtracted from:

Metode    Lower     Center     Upper   -------+---------+---------+---------
2       -0,4176   -0,3787   -0,3399    (-*)
3       -0,0624   -0,0236   0,0152                    (*-)
                                       -------+---------+---------+---------
                                          -0,25      0,00       0,25

Metode = 2   subtracted from:

Metode      Lower    Center   Upper    -------+---------+---------+---------
3          0,3164   0,3551   0,3938                                 (*-)
                                       -------+---------+---------+---------
                                          -0,25      0,00      0,25
```

Algorithm 9: ANOVA test results on tribal population D.

Data Availability

The STR-DNA data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper is sponsored by the Indexed Scientific Publication for Student Doctor's Final Project (Hibah Doktor) 2018 of Universitas Indonesia [0588/SK/R/UI/2018].

References

[1] S. K. Sheppard, D. S. Guttman, and J. R. Fitzgerald, "Population genomics of bacterial host adaptation," *Nature Reviews Genetics*.

[2] *Emery and Rimoin's Principles and Practice of Medical Genetics*, Elsevier, 2013.

[3] A. Neil and Campbell., *Campbell Reece*, Pearson Benjamin Cummings, 8th edition, 2008.

[4] M. S. H. Abu Halima, L. P. Bernal, and F. A. Sharif, "Genetic variation of 15 autosomal short tandem repeat (STR) loci in the Palestinian population of Gaza Strip," *Legal Medicine*, vol. 11, no. 4, pp. 203-204, 2009.

[5] M. G. Patel Jignal, Shaikh, and D. Marjadi, "Forensic Conception: DNA Typing of FTA Spotted Samples," *Journal of Applied Biology Biotechnology Vol*, vol. 2, no. 04, pp. 021–029, 2014.

[6] S. J. Venables, R. Daniel, S. D. Sarre et al., "Allele frequency data for 15 autosomal STR loci in eight Indonesian subpopulations," *Forensic Science International: Genetics*, vol. 20, pp. 45–52, 2016.

[7] *Khaleda Parven Forensic use of DNA information : human rights, privacy and other challenges.. University of Wollongong Thesis Collections*, 7 Khaleda Parven Forensic use of DNA information, human rights, 2012.

[8] B. Derbyshire, *Sharing DNA Profiles and Finger Prints Across the EU requires further safeguards*, GeneWatch, UK, 2015.

[9] A. Holobinko, "Theoretical and Methodological Approaches to Understanding Human Migration Patterns and their Utility in Forensic Human Identification Cases," *Societies*, vol. 2, no. 2, pp. 42–62, 2012.

[10] D. Ricke, A. Shcherbina, N. Chiu et al., "Sherlock's Toolkit: A forensic DNA analysis system," in *Proceedings of the IEEE International Symposium on Technologies for Homeland Security, HST 2015*, USA, April 2015.

[11] A. L. Lowe, A. Urquhart, L. A. Foreman, and I. W. Evett, "Inferring ethnic origin by means of an STR profile," *Forensic Science International*, vol. 119, no. 1, pp. 17–22, 2001.

[12] M. Rahmat Widyanto, "Various defuzzification methods on DNA similarity matching suing fuzzy inference system," *Journal of Advanced Computational Intelligence Intelligent Informatics*, vol. 14, no. 3, 2010.

[13] R. N. Hartono, M. R. Widyanto, and N. Soedarsono, "Fuzzy logic system for DNA profile matching with embedded ethnic inference," in *Proceedings of the 2010 2nd International Conference on Advances in Computing, Control and Telecommunication Technologies, ACT 2010*, pp. 69–73, Indonesia, December 2010.

[14] M. R. Widyanto, R. N. Hartono, and N. Soedarsono, "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, no. 6, p. 3103, 2016.

[15] L. A. Zadeh, "Fuzzy sets," *Information and Computation*, vol. 8, pp. 338–353, 1965.

[16] S. Iwamoto, K. Tsurusaki, and T. Fujita, "Conditional decision-making in fuzzy environment," *Journal of the Operations Research Society of Japan*, vol. 42, no. 2, pp. 198–218, 1999.

[17] Liyuan Zhang, Xuanhua Xu, and Li Tao, "Some Similarity Measures for Triangular Fuzzy Number and Their Applications in Multiple Criteria Group Decision-Making," *Journal of Applied Mathematics*, vol. 2013, pp. 1–7, 2013.

[18] K. Kundu, "Image Denoising using Patch based Processing with Fuzzy Gaussian Membership Function," *International Journal of Computer Applications*, vol. 118, no. 12, pp. 0975–8887, 2015,

Department of Computer Science & Engineering Govt. College of Engineering & Textile Technology, Serampore, Hooghl.

[19] H. A. Hefny, H. M. Elsayed, and H. F. Aly, "Fuzzy multi-criteria decision making model for different scenarios of electrical power generation in Egypt," *Egyptian Informatics Journal*, vol. 14, no. 2, pp. 125–133, 2013.

[20] B. Akshay, B. Akash, and P. Avril, "Convolution and application of convolution," *International Journal of Innovative Research In Technology*, vol. 6, pp. 2349–6002, 2014.

[21] P. A. Bromiley, *Products and Convolutions of Gaussian Probability Density Function*, vol. 14, Imaging Sciences Research Group, Institute of Population Health, School of Medicine, University of Manchester, Manchester, 2014.