Paper:

Various Defuzzification Methods on DNA Similarity Matching Using Fuzzy Inference System

M. Rahmat Widyanto*, Nurtami Soedarsono, Norihiro Katayama***, and Mitsuyuki Nakao*****

*Faculty of Computer Science, University of Indonesia
Depok Campus, Depok 16424, West Java, Indonesia
E-mail: widyanto@cs.ui.ac.id
**Faculty of Dentistry, University of Indonesia
Salemba Campus, Jakarta, Indonesia
***Biomodeling Laboratory, Department of Applied Information Sciences,
Graduate School of Information Sciences, Tohoku University
6-3-09 Aoba, Aramaki-aza, Aoba-ku, Sendai 980-8579, Japan
[Received December 8, 2009; accepted February 9, 2010]

A DNA similarity matching using fuzzy inference system is proposed to measure a similarity between human STR (Short Tandem Repeat) based DNA (Deoxyribonucleic Acid) profiles. Moreover, various defuzzification methods are also tested to observe their behavior on different DNA data characteristics. Experiment on real human STR based DNA profile data shows that the proposed DNA similarity matching produces more realistic similarity values compared to those of the conventional one. Experiment on various defuzzification methods on DNA similarity matching shows that Sugeno defuzzification method is more suitable than those of other defuzzification methods.

Keywords: DNA similarity matching, fuzzy inference system, defuzzification method, short tandem repeat

1. Introduction

DNA (Deoxyribonucleic Acid) is a unique genetic material that inherits to the descendants, therefore DNA can be used for human identification [1–3]. There are many aims of human identification using DNA, e.g., family relation proof, criminal action evidence, and disaster victim identification. As the development of technology, the DNA analysis method has also been developing. Currently there are various DNA analysis methods [4, 5], e.g., Restriction Fragment Length Polymorphism (RFLP) analysis, Fragment Length Polymorphism (FLP) analysis, mitochondrial DNA analysis, and Short Tandem Repeat (STR) analysis. The Federal Bureau Investigation (FBI) has chosen 15 locus and amelogenin based on STR analysis [6–8] to be the standard for DNA human identification. The standard, called CODIS (Combined DNA Index System) [9], enables a uniform representation of DNA profile and exchange between DNA forensics laboratories. This CODIS has also been adopted by Interpol for DNA data exchange between countries.

National Institute of Standards and Technology (NIST) Human Identity Project Team [10] has developed a method (called STR_MatchSamples) to measure the similarity between two CODIS data. However, DNA is uncertainty data due to different circumstances in sampling, contaminated with other micro cell, noise caused by inhibitor material, etc. Since the STR_MatchSamples developed based on crisp logic, it cannot deal with DNA data having uncertainty and noise.

To deal with the uncertainty and noise in DNA data, a DNA similarity matching using fuzzy inference system is proposed. The proposed method uses 15 locus and amelogenin based on STR analysis as input. The fuzzy rules evaluate the similarity value between two DNA alleles which is the height of the intersection between of the fuzzy triangular numbers. The intersection value is then inferred to the consequent part by 9 fuzzy inference rules where Mamdani and Sugeno Fuzzy Inference System [11] are used. Moreover, various defuzzification methods are also tested to observe their behavior on different DNA data characteristics. Experiment on real human STR based DNA profile data shows that the proposed DNA similarity matching produces more realistic similarity values compared to those of the conventional one. Experiment on various defuzzification methods on DNA similarity matching shows that Sugeno defuzzification method [11] is more suitable than those of other defuzzification methods.

In Section 2, DNA analysis including DNA sample processing, DNA STR analysis and DNA similarity matching using STR_MatchSamples method are discussed. The DNA similarity matching using fuzzy inference system is proposed in Section 3. Experimental results on various data characteristics are given in Section 4. Conclusions are discussed in Section 5.

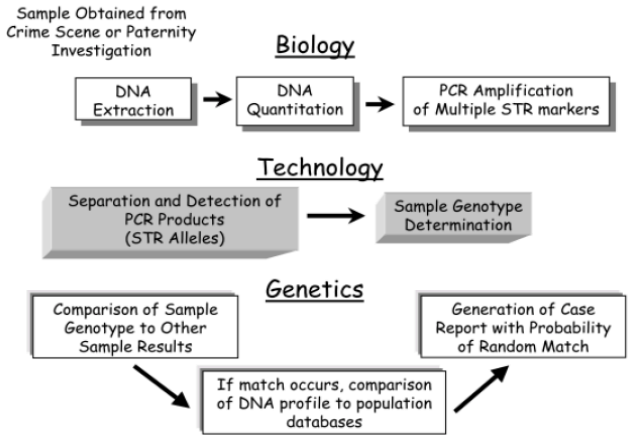

Fig. 1. Steps in DNA sample processing [12].

2. DNA Analysis

DNA (Deoxyribonucleic Acid) is a unique genetic material that inherits to the descendants. Therefore DNA of a descendant has a pair of chromosomes from his father or mother. Due to its unique characteristics, DNA analysis is often carried out for family relation proof, criminal action evidence, and disaster victim identification. This section describes DNA analysis method which consists of DNA sample processing, DNA STR analysis, and the conventional method to measure DNA similarity matching called DNA STR_MatchSamples.

2.1. DNA Sample Processing

Steps in DNA Sample Processing consist of three steps [12], i.e., biology, technology, and genetics steps. **Fig. 1** shows the steps in DNA sample processing.

1. Biology Step

   In this step, DNA is extracted from the corresponding biological material where the DNA is then quantified to calculate the number of DNA those can be recovered [13]. After separating the DNA from their cells, the certain region is copied using Polymerase Chain Reaction (PCR) method. PCR will produce millions copies of DNA therefore there will be enough data to be examined.

2. Technology Step

   The results from PCR are then separated and detected so that can be classified to be the region that will be examined. The region is identified by Short Tandem Repeat (STR) which is a sequence of protein having certain pattern. The separation of STR uses slab gel and capillary electrophoresis, while for the detection the florescence method is usually used. After STR allele detection, the number of repeated sequence can be calculated where this process is called sample genotyping.

3. Genetics Step

   In this step, the results of STR combination from previous step are compared to the other DNA profile to identify the individual identity or relation. If the comparison has a high similarity it can be concluded that the individual having relation with the other sample, otherwise it doesn't.

2.2. DNA STR Analysis

As the development of technology, the DNA analysis method has also been developing. Currently there are various DNA analysis methods [4, 5], e.g., Restriction Fragment Length Polymorphism (RFLP) analysis, Fragment Length Polymorphism (FLP) analysis, mitochondrial DNA analysis, and Short Tandem Repeat (STR) analysis. STR is a repeated pattern from two or more nucleotides without intervention from other nucleotides. For example there are CA and GT which are repeated, called dinucleotide repeat, as follow,

5′—CTAGCTACTG**CACACACACACACACA**CGTGCCGATGC—3′

3′—GATCGATGAC**GTGTGTGTGTGTGTGT**GCACGGCTACG—5′,

then the value of STR for those markers is eight. The combination of several important markers of STR can be used for human identification.

The Federal Bureau Investigation (FBI) has chosen 15 markers (called locus) and amelogenin based on STR analysis to be the standard for DNA human identification. The standard, called CODIS (Combined DNA Index System) [9], enables a uniform representation of DNA profile and exchange between DNA forensics laboratories. This CODIS has also been adopted by Interpol for DNA data exchange between countries. The 15 locus are CSF1PO, FGA, TH01, TPOX, vWA, D21S11, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D2S1338, plus amelogenin. Certain locus has its own characteristics shown by certain repeated pattern [6–8, 10].

2.3. DNA Similarity Matching Using STR_MatchSamples

STR_MathSamples [10] is a Microsoft® Excel based tool used for DNA matching developed by National Institute of Standards and Technology (NIST). NIST is an agency of the U.S. Department of Commerce, was founded in 1901 as the nation's first federal physical science research laboratory. The tool is developed as macro in Microsoft® Excel, therefore the user can input the value of STR to compare, and the tool will calculate the similarity automatically.

The STR_MatchSamples calculates the percentage of allele similarity which show how much resemble a DNA profile to the other. The calculation is as follows;

$$k(\%) = \frac{\sum_{i=1}^{m} sa_i}{\sum_{j=1}^{n} al_i}, \quad \ldots\ldots\ldots\ldots\ldots (1)$$