# Application of an automatic fuzzy logic based pattern recognition method for DNA Microarray Reader

M.Sc. Wei Wei
M.Sc. Xiaodong Wang
Prof. Dr.-Ing Werner Neddermeyer
Prof. Dr.-Ing Wolfgang Winkler
Prof. Dr.-Ing Michael Schnell
University of Applied Science in Gelsenkirchen
Neidenburger Str. 43, 45897 Gelsenkirchen
Germany

*Abstract:* -The past decade has brought about tremendous advances in genetics and molecular biology. Nowadays DNA technology is widely applied in laboratories in order to diagnose accurately and effectively genetic illnesses of patients. The Solas2 is a DNA microarray reader, which is specifically designed for medical laboratory. It bases on the image processing and pattern detection technology to analyze DNA microarrays, extract the feature of genome, and visualize the early-stage diseases. With the help of the DNA microarray reader users can more easily detect and treat those diseases, which could occur in the future.

Basing on the fuzzy logic algorithms, a special method was developed for the classification of the extracted DNA features. Because of the irregular distribution of pattern space with classic fuzzy algorithms, the result of pattern detection would not be satisfying. This article will introduce a method which is derived and improved from classic fuzzy logic methods FCM and FML. With this method the recognition of DNA features can be proceeded correctly and effectively.

*Key-Words:* - DNA Microarray Reader, Fuzzy Logic, FCM, FML, Pattern Detection, Pattern Space

## 1 Introduction

With the rapid development of DNA technology DNA diagnostics has been widely applied in private and hospital laboratories for the purpose of diagnosing various genetic illnesses, because the approach enables the detection of pathogens that cannot be detected by traditional methods. DNA diagnostics combined with microchip technology is a fast and reliable method for detecting many genetic illnesses especially infectious diseases.

The gene diagnoses to a type of disease do not result from a change in a particular genome. But sequence deviations in genome group are usually analyzed to decide whether patients have been infected by a disease or not. Compared with other methods of analysis, DNA microarray technology has the advantage to make the simultaneous analysis of many genomes.

There are many kinds of DNA miroarrays applied by our DNA microarray reader e.g. osteo/check, thrombo/check, coro/check etc. Each microarray has its own layout and various groups of polymorphisms, which indicate the changes of genomes in chemical reaction. Every polymorphism has three different statuses: wildtyp (++), heterozygot (+/-) and homozygot (-/-) (Fig. 1). The statuses are defined with PM and MM values (Fig. 2), which indicate the result from a chemical reaction. The task of pattern recognition is to classify polymorphisms into different statuses.

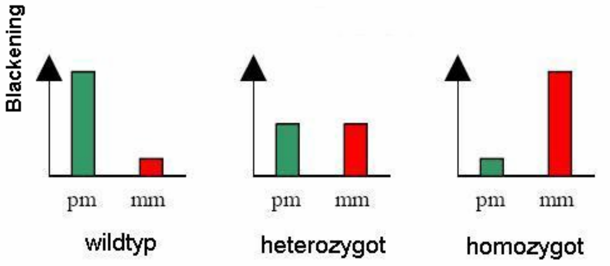

Fig. 1: three statuses of polymorphisms

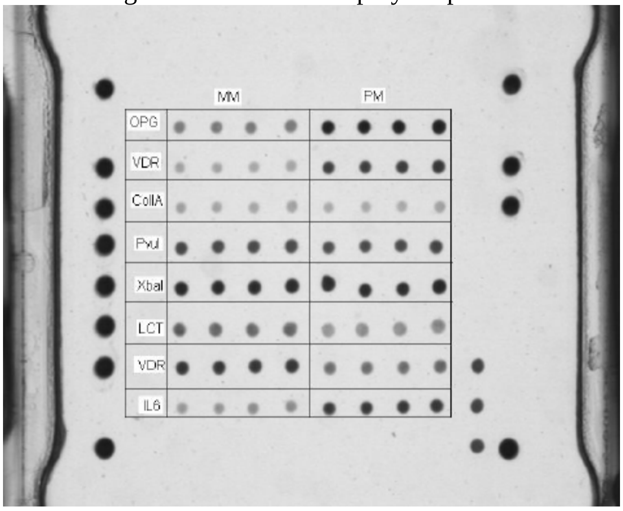

Fig. 2: DNA Microarray Chip

The array tube, as shown in Fig. 3, is a standard micro reaction container. The DNA microarray is located at the bottom of it.

During a measurement, the array tube should be read by the DNA microarray reader. After that the PM and MM values for every polymorphism will be calculated.

An automatic effective method is required in order to classify polymorphisms into standard statuses.

Fig. 3: Array Tube

## 2 Conventional solution of gene analysis

The conventional solution bases on the manually determined classificator to separate the clusters. The disadvantages of this method are described in following points:

1. This method bases on one dimension pattern space. If a large number of pattern elements exist in pattern space, the pattern elements would cover almost the whole pattern space. Thus it is difficult to separate these clusters.
2. Subjective separation is not accurate.
3. The classes are separated with simple thresholds; therefore it could cause false classification.

In order to evaluate the performance of the conventional solution, a test has been carried out. The result, as shown in Table 1, indicates that with manually determined classificator the error rate is up to 3%.

| Genome | Error | Correctness |
|---|---|---|
| CETP | 2% | 98% |
| CYP7A1 | 1% | 99% |
| ITGB | 2% | 98% |
| MTHFR | 3% | 97% |
| PAI-1 | 3% | 97% |
| PON1 | 2% | 98% |

Table 1: Performance of classification by conventional solution

For the purpose of analyzing mutation of gene automatically and drawing a conclusion correctly, a method which bases on fuzzy logic technology has been developed.

## 3 Automatic method for pattern recognition

This automatic method of pattern recognition should be proceeded in three steps:

1. Establishing of pattern space;
2. Applying fuzzy logic algorithms to calculate the classificator;
3. Analyzing polymorphisms for single miroarray.

### 3.1 Establishing of pattern space

According to the conventional pattern recognition method, the pattern space should be established at the first step. This is very important because it strongly influences the performance of pattern recognition. There are two fundamental problems in establishing of pattern space. The first one is concerned with the representation of input data obtained by measurements on objects that are to be recognized: the sensing problem. In general, each object is represented by a vector of measured values of s variables, x = [x1, x2, ..., xs] this vector is usually called a pattern vector. Here the PM and MM values would be chosen as pattern vector. The second problem concerns the extraction of characteristic features from the input data in terms of which dimensionality of pattern vectors xi can be reduced: the feature extraction problem. Valid features must characterize attributes by which the given pattern classes are well discriminated [1]. Different pattern features cause various distributions of clusters in a pattern space. A good pattern space may have following characters:

- The pattern elements in a cluster must be converged as much as possible.
- The gravities between different clusters must be separated as far as possible.

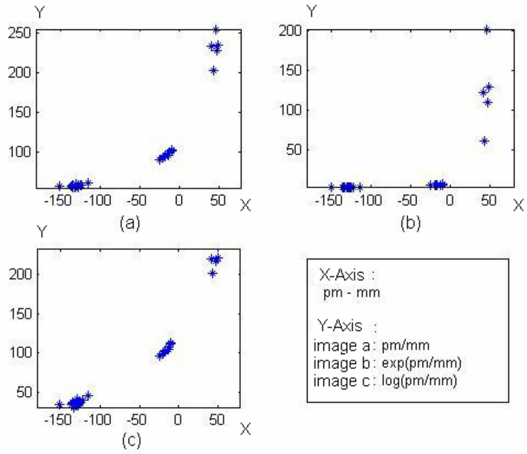

Fig. 4: Comparison of pattern spaces with different pattern features

In terms of comparing the cluster distributions, which are shown in Fig. 4, the best pattern feature is PM-MM in X axis and log (PM/MM) in Y axis.

## 3.2 Method for calculating classificator

### 3.2.1 Fuzzy Cluster Analysis

Fuzzy cluster analysis is a simple and commonly used pattern recognition technology in the field of image processing. Grouping items by their similarities and separating them by their differences are often applied to control those items. The main process of classic fuzzy cluster analysis is:

- Step 1: Setting the initial value of gravity for every cluster, the number of iteration and other parameters.
- Step 2: Calculating the distance between every data and every class.
- Step 3: Calculating the affiliation degree between every data and every class.
- Step 4: Calculating the new gravity for every cluster, then repeat step 2 and step 3 until the difference of the class gravities between two iterations is smaller than the threshold or the number of iteration excesses the predefined limitation.

Usually there is no training phase for fuzzy cluster analysis. All data are directly proceeded during one cluster analysis. After that, the class information will be obtained. But the disadvantage of this method is: this process can not be executed until enough data are available. In the daily diagnoses, the amount of patient’s data can’t be collected in a short period. Therefore the classificator must be calculated before classification. So both training phase and use phase must be included in our method. The calculated gravity of cluster, which is the most important parameter in fuzzy cluster analysis, is defined as classificator. In training phase the classificator is calculated; in use phase the affiliation degrees between every data and every class are calculated with the help of the classificator, and then the class information will be obtained.

### 3.2.2 Fuzzy C Means (FCM) and Fuzzy Maximal Likelihood (FML)

Fuzzy-C-Means is a basic algorithm in fuzzy cluster analysis. Fuzzy-Maximal-Likelihood is derived from Fuzzy-C-Means. It is specially designed to deal with the following problem: the forms and sizes of clusters in a pattern space are usually different. In project Solas2 both algorithms are analyzed (Fig. 5).

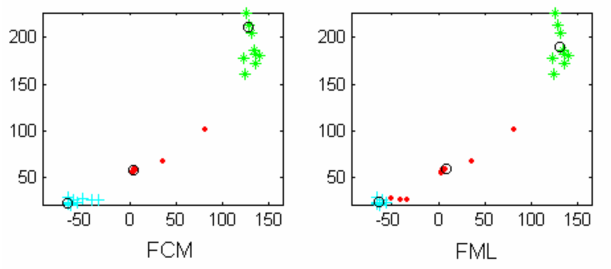

Fig. 5: Comparison between FCM and FML

A comparison between FCM and FML is shown in Table 2.

| | |
|---|---|
| Complexity(computational costs) | FCM < FML |
| Dependence on Initial value of gravity | FCM < FML |
| Accuracy of calculated gravity | FCM < FML |
| Accuracy of affiliation degree | FCM > FML |

Table 2: A comparison of FCM and FML

The advantage of FCM is that its affiliation degree is more accurate than FML’s. On the other side the advantage of FML is that its calculated gravity is more accurate than FCM’s. Thus a new method, which combines the advantages from FCM and from FML, is designed and implemented.

The main process of this new method is (This process is described with 3 classes and n input data as example.):

- Step 1: Setting the initial value of gravity for every cluster, the number of iteration and other parameters. The initial value of gravity is not arbitrarily defined. It is calculated in order that the initial value of gravity is located close to the center of the class.
- Step 2: Executing FCM. The gravity for every cluster S1= {s11, s12, s13} is calculated. The affiliation degree between every data and every class Z1= $\{\{z11, z12, z13\}_1, \{z11, z12, z13\}_{2\ldots}\ \{z11, z12, z13\}_n\}$ is calculated.
- Step 3: Executing FML. The gravity S1 is used here as initial value. Then the new gravity for every cluster S2= {s21, s22, s23} is calculated. The affiliation degree between every data and every class Z2= $\{\{z21, z22, z23\}_1, \{z21, z22, z23\}_{2\ldots}\ \{z21, z22, z23\}_n\}$ is calculated.
- Step 4: Calculating the new affiliation degree between every data and every class Z3= $\{\{z31, z32, z33\}_1, \{z31, z32, z33\}_{2\ldots}\ \{z31, z32, z33\}_n\}$ with the mathematic formula which is used in FCM.

In training phase the gravity for every cluster S2 is calculated through step1 to step3 and defined as classificator. In use phase step 4 is executed in order to calculate affiliation degree Z3. According to this affiliation degree, data will be classified. For example, there are two clusters and after step 4 the affiliation degree between every data and every class is calculated, namely 0.6 and 0.4. Because 0.6>0.4, this data is classified in the first class (affiliation degree=0.6).

Usually there are some critical data in pattern space. Sometimes it is not clear into which class certain critical data should be divided. Hence a new class named uncertain class should be defined. All critical data should be divided into this new class. In order to solve this kind of problem, a validity criterion is often used. For example, in project Solas2, a validity criterion is defined as follows: the affiliation degree between every data and every class Z3= {$\{z31, z32, z33\}_1$, $\{z31, z32, z33\}_{2\ldots}$ $\{z31, z32, z33\}_n$} is calculated. The sum of affiliation degrees between certain data and three classes should equal one: Z31+Z32+Z33=1. If Z31= Max{Z31, Z32, Z33} and Z31 is less than validity criterion, then this data is divided into uncertain class.

In Fig. 6 there is one critical data pointed by an arrow. According to validity criterion this data is divided into uncertain class.

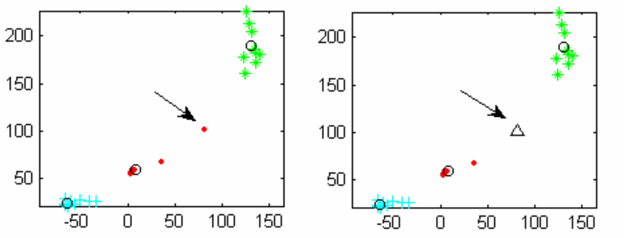

Fig. 6: critical data (uncertain)

### 3.3 Analysis of gene groups for single chip

This process is called use phase. In use phase the image processing should be executed in order to get gene features from polymorphisms. According to the classificator, which has been generated in training phase, the affiliation degrees to each class will be calculated. With the predefined validity criterions, the polymorphisms will be classified.

### 3.4 Testing the method

In order to evaluate the performance of the fuzzy method, a test has been executed. The result shown in Table 3 and Fig. 7 indicates that after classification no error is generated; on the average, only about 4% data is invalid. The total rate of correctness is about 96%.

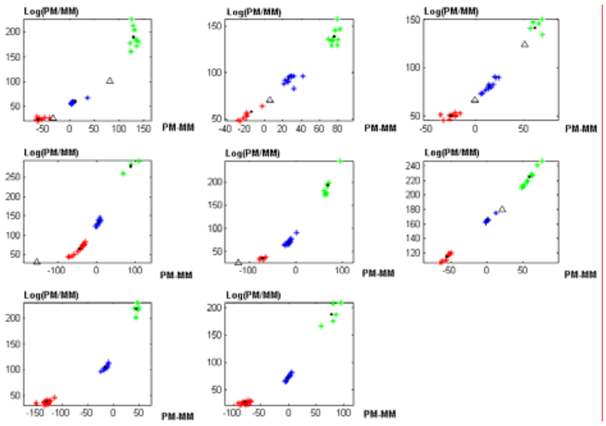

Fig. 7: Classification by fuzzy method

| Genome | Error | Invalid | Correctness |
|---|---|---|---|
| IL6 | 0% | 6% | 94% |
| VDR | 0% | 3% | 97% |
| Colla1 | 0% | 6% | 94% |
| Pvu | 0% | 3% | 97% |
| Xba | 0% | 3% | 97% |
| LCT | 0% | 3% | 97% |
| OPG-209 | 0% | 0% | 100% |
| OPG-245 | 0% | 0% | 100% |

Table 3: Performance of classification by fuzzy method

Comparing to the conventional method, the average correctness by the fuzzy logic method is not obviously increased. But the conventional method causes 3% error, which should be absolutely avoided in clinical gene diagnoses. Because it might cause serious accidents in diagnoses and treatments. For this reason the fuzzy logic method is clearly better than the conventional method.

## 4 Conclusion

Basing on the fuzzy logic method, the data could be automatically and effectively classified. After the classification no error has been generated

The rate of correctness by classification is more than 96%. This performance satisfies the requirement of gene diagnoses and makes them more reliable.

References:

[1] Klir, G. J. and Yuan B., Fuzzy Sets and Fuzzy Logic. Theory and Applications. Prentice Hall PTR, Upper Saddle River, 1995.

[2] Bezdek, J. C., Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York, 1981.

[3] Klir, G. J., St.Clair, U. H. and Yuan B., Fuzzy Set Theory. Foundations and Applications. Prentice Hall PTR, Upper Saddle River, 1997.

[4] Gustafson und Kessel, W.C.: Fuzzy Clustering with a Fuzzy Covariance Matrix, IEEE CDC, San Diego,

Californien, 761-766,1979.

[5] Klawonn, F. und Kruse, R.: Automatic Generation of Fuzzy Controllers by Fuzzy Clustering. Proc. IEEE Int. Conf. on Systems, Man and Cybernetics, 2040-2045, Vancouver, 1995.

[6] Dempster, A.P., Laird, N.M. und Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 39(B), 1-38, 1977.