# Fuzzy Methods for Analysis of Microarrays and Networks

**Yanfei Wang**
**Bachelor of Science (Information and Computation Sciences)**
**Master of Science (Applied Mathematics)**

**Thesis submitted for the degree of Doctor of Philosophy in**
**Discipline of Mathematical Sciences**
**Faculty of Science and Technology**
**Queensland University of Technology**

**2011**

**Principal supervisor: Professor Vo Anh**
**Associate supervisor: Professor Zuguo Yu**

# Keywords

DNA microarray; clustering; fuzzy set; type-2 fuzzy set; fuzzy c-means; fuzzy parameters; empirical mode decomposition; noise; uncertainty; disease-related genes; type-2 membership function; protein protein interaction networks, sub-networks, protein complexes; fuzzy relation; communities; graph model.

# Abstract

Bioinformatics involves analyses of biological data such as DNA sequences, microarrays and protein-protein interaction (PPI) networks. Its two main objectives are the identification of genes or proteins and the prediction of their functions. Biological data often contain uncertain and imprecise information. Fuzzy theory provides useful tools to deal with this type of information, hence has played an important role in analyses of biological data. In this thesis, we aim to develop some new fuzzy techniques and apply them on DNA microarrays and PPI networks. We will focus on three problems: (1) clustering of microarrays; (2) identification of disease-associated genes in microarrays; and (3) identification of protein complexes in PPI networks.

The first part of the thesis aims to detect, by the fuzzy C-means (FCM) method, clustering structures in DNA microarrays corrupted by noise. Because of the presence of noise, some clustering structures found in random data may not have any biological significance. In this part, we propose to combine the FCM with the empirical mode decomposition (EMD) for clustering microarray data. The purpose of EMD is to reduce, preferably to remove, the effect of noise, resulting in what is known as denoised data. We call this method the fuzzy C-means method with empirical mode decomposition (FCM-EMD). We applied this method on yeast and serum microarrays, and the silhouette values are used for assessment of the quality of clustering. The results indicate that the clustering structures of denoised data are more reasonable, implying that genes have tighter association with their clusters. Furthermore we found that the estimation of the fuzzy parameter *m*, which is a difficult step, can be avoided to some extent by analysing denoised microarray data.

The second part aims to identify disease-associated genes from DNA microarray data which are generated under different conditions, e.g., patients and normal people. We developed a type-2 fuzzy membership (FM) function for identification of disease-associated genes. This approach is applied to diabetes and lung cancer data, and a comparison with the original FM test was carried out. Among the ten best-ranked

genes of diabetes identified by the type-2 FM test, seven genes have been confirmed as diabetes-associated genes according to gene description information in Gene Bank and the published literature. An additional gene is further identified. Among the ten best-ranked genes identified in lung cancer data, seven are confirmed that they are associated with lung cancer or its treatment. The type-2 FM-*d* values are significantly different, which makes the identifications more convincing than the original FM test.

The third part of the thesis aims to identify protein complexes in large interaction networks. Identification of protein complexes is crucial to understand the principles of cellular organisation and to predict protein functions. In this part, we proposed a novel method which combines the fuzzy clustering method and interaction probability to identify the overlapping and non-overlapping community structures in PPI networks, then to detect protein complexes in these sub-networks. Our method is based on both the fuzzy relation model and the graph model. We applied the method on several PPI networks and compared with a popular protein complex identification method, the clique percolation method. For the same data, we detected more protein complexes. We also applied our method on two social networks. The results showed our method works well for detecting sub-networks and give a reasonable understanding of these communities.

## Declaration of Original Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher educational institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed:

Date:

## Lists of Papers

1. Yan-Fei Wang, Zu-Guo Yu, Vo Anh. Fuzzy C-means method with empirical mode decomposition for clustering microarray data. *International Journal of Data Mining and Bioinformatics* (Accepted 30 April 2011).

2. Yan-Fei Wang, Zu-Guo Yu. A type-2 fuzzy method for identification of disease-related genes on microarrays. *International Journal of Bioscience, Biochemistry and Bioinformatics*, Vol. 1, No. 1, pp. 73-78, 2011.

3. Yan-Fei Wang, Zu-Guo Yu, Vo Anh. Fuzzy c-means method with empirical mode decomposition for clustering microarray data. *IEEE International Conference on Bioinformatics and Biomedicine, 2010, Hongkong, China*.

4. Yan-Fei Wang, Zu-Guo Yu, Vo Anh. Type-2 fuzzy approach for disease-associated gene identification on microarrays. *IACSIT International Conference on Bioscience, Biochemistry and Bioinformatics, 2011, Singapore, Singapore*.

5. Yan-Fei Wang, Zu-Guo Yu, Vo Anh. Identification of protein complexes from PPI networks based on fuzzy relation and graph model (to be submitted).

## Acknowledgements

I would like to sincerely express my deep gratitude to Professor Vo Anh, my principal supervisor. Not only for the thesis writing was suggested by him, but his constant guiding and help were also essential for carrying out my studies. His inspiration, responsibility and his warm personality have won my highest respect and love.

I also would like to appreciate Professor Zu-Guo Yu, my associated supervisor, not only for his valuable suggestion and discussion, but also for the help he affords us generously when I am living in Australia.

I also would like to appreciate Queensland University of Technology and China scholarship Council to offer me a valuable opportunity studying in QUT in Australia.

My appreciation also give to the Discipline of Mathematical Sciences of QUT, for the excellent research condition and support they offer to me in these three years, and the school staffs who always give me heart-warmed help.

Special thanks to Shaoming, Shiqiang, Qianqian, Qiang Yu, Zhengling and my friends in O415 for their support and assistance throughout the time here.

I want to especially appreciate my families for their support and encouragement. This thesis is dedicated to my grandparents, my parents and Danling.

# Content

## List of Figures

## List of Tables

# Chapter 1

# Introduction

## 1.1 The research problems

Bioinformatics is defined as the application of computers, databases and mathematical methods to analyses of biological data, especially genetic sequences, microarrays and protein structures. The two main research fields in bioinformatics are genomic analysis and proteomic analysis (Herrero and Flores 2008). Genomic analysis aims to extract information from large amounts of gene data, while proteomic analysis has an objective to determine protein functions from protein databases (Lee and Lee 2000, Mann and Jensen 2003).

Abundant positive results suggest analysis of DNA microarray is a significant way to discovering meaningful information about DNA structures and their functions. Protein complex (or multi-protein complex) is a group of two or more proteins in protein- protein interaction (PPI) networks. Most proteins seem to function with complicated cellular pathways, interacting with other proteins either in pairs or as components of large complexes. So identification of protein complexes is crucial for understanding the principles of cellular organization and functions.

Although biological experiments can provide a wealth of information on genes and proteins, these experiments are expensive and time-consuming (Sokal and Rohlf 1995). Hence computational prediction methods are needed to provide valuable information for large DNA microarray and protein data whose structures or functions cannot be determined from biological experiments (Zar 1999). As new biological technologies advance, the growth in DNA data available to researchers is unparalleled. For example, Gene Bank, a major public database where DNA data are stored, doubles in size approximately every year. It has become important to improve new theoretical methods to make analysis of these data more efficient and precise.

The data for DNA and protein biological analyses contain plenty of uncertain and imprecise information. Fuzzy set theory has many advantages in dealing with this type of data; therefore, approaches based on fuzzy set theory have been taken into consideration to analyse DNA microarrays and PPI networks. There are several applications of fuzzy set theory in bioinformatics. The results show that fuzzy method is a way to render precise what is imprecise in the world of bioinformatics.

**This thesis aims to study fuzzy methods on the analysis of DNA microarrays and PPI networks in three related aspects: (i) clustering analysis on DNA microarrays; (ii) identification of disease-related genes on microarrays; (iii) identification of protein complexes in PPI networks. The fuzzy c-means clustering method, type-2 fuzzy method, and fuzzy relation clustering method will be used to investigate these three problems.**

(i) Microarray techniques have revolutionized genomic research by making it possible to monitor the expression of thousands of genes in parallel. The enormous quantities of information data generated have led to a great demand for efficient analysis methods. Data clustering analysis is a useful tool and has been extensively applied to extract information from gene expression profiles obtained with DNA microarrays. Existing clustering approaches, mainly developed in computer science, have been adapted to microarray data. Among these approaches, fuzzy c-means (FCM) method is an efficient one. However, a major problem in applying the FCM method for clustering microarray data is the choice of the fuzziness parameter *m*. Commonly, $m = 2$ is used as an empirical value but it is known that $m = 2$ is not appropriate for some data sets and that optimal values for *m* vary widely from one data set to another. On the other hand, microarray data contain noise and the noise would affect clustering results. Some clustering structure can be found from random data without any biological significance. In this part of the thesis, we propose to combine the FCM method with the empirical mode decomposition (EMD) for clustering microarray data in order to reduce the effect of the noise. We call this method fuzzy c-means method with empirical mode decomposition (FCM-EMD).We applied this method on yeast and serum microarray data respectively and the silhouette values are used for assessment of quality of clusters.

(ii) Comparison of gene microarray expression data in patients and those of normal people can identify disease associated genes and enhance our understanding of disease. In order to identify the disease-associated genes, we usually need to determine for each gene whether the two sets of expression values are significantly different from each other. Measuring the divergence of two sets of values of gene expression data is an effective approach.

The word "different" itself is a fuzzy concept and fuzzy theory has many advantages in dealing with data containing uncertainty, therefore fuzzy approaches have been taken into consideration to analyse DNA microarrays. Liang et al. (2006) proposed a fuzzy set theory based approach, namely a fuzzy membership test (FM-test), for disease genes identification and obtained better results by applying their approach on diabetes and lung cancer microarrays. However, some limitations still exist. The most obviously one is when the values of gene microarray data are very similar and lack over-expression, in which case the FM-d values are very close or even equal to each other. That made the FM-test inadequate in distinguishing disease genes. Meanwhile, DNA microarray data contains noise, hence yielding uncertain information in the original data. When deriving the membership function for evaluation, all of these uncertainties translate into uncertainties about fuzzy set membership function. Traditional fuzzy sets are not able to directly model such uncertainties because their membership functions are totally crisp.

To overcome these problems, we introduce type-2 fuzzy set theory into the research of disease-associated gene identification. Type-2 fuzzy sets can control the uncertainty information more effectively than conventional type-1 fuzzy sets because the membership functions of type-2 fuzzy sets are three-dimensional. It is the new third dimension of type-2 fuzzy sets that provides additional degrees of freedom that makes it possible to directly model uncertainties.

(iii) Identification of protein complexes is very important for better understanding the principles of cellular organisation and unveiling their functional and evolutionary mechanisms. It is known that dense sub-networks of protein-protein interactions (PPI) networks represent protein complexes or functional modules. Therefore, the problem of identifying protein complexes is equivalent to that of searching sub-networks in the original networks. Many methods for mining protein complexes have mostly focused on detecting highly connected sub-networks. An extreme example is to identify all fully connected sub-networks. However, it is too restrictive to be useful in real biological networks because there are many protein complexes which are not fully connected sub-networks. In this problem, we propose a novel method which combines the fuzzy clustering method and interaction probability to identify the overlapping and non-overlapping community structures in PPI networks, then to detect protein complexes in these sub-networks. Our method is based on both the fuzzy relation model and the graph model. Fuzzy theory is suitable to describe the uncertainty information between two objects, such as ‘similarity’ and ‘differences’. On the other hand, the original graph model contains clustering information, thus we don’t ignore the original structure of the network, but combine it with the fuzzy relation model. We apply the method on yeast PPI networks and compare the results with those obtained by a standard method, CFinder.

## 1.2 Clustering analysis of microarrays

### 1.2.1 Biological background and literature review

A DNA microarray is a multiplex technology applied in molecular biology. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles of a specific DNA sequence, known as probes or reporters. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. Therefore, microarray has dramatically accelerated many types of investigation. Microarray technology evolved from southern blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment (Maskos and Southern, 1992). The first reported application of this method was the analysis of 378 arrayed bacterial colonies each harbouring a different sequence which were assayed in multiple replicas for expression of genes in multiple normal and tumor tissue (Augenlicht and Kobrin, 1982). This was expanded to analysis of more than 4000 human sequences with computer driven scanning and image processing for quantitative analysis of the sequences in human colonic tumors and normal tissue (Augenlicht et al., 1987) and then to comparison of colonic tissues at different genetic risk (Augenlicht et al., 1991).

Following preparation of an array support with DNA of genes of interest, the basic steps in a microarray experiment are as follows: (1) mRNA isolation from cells; (2) Generation of cDNA by reverse transcription with a fluorescent tag attached; (3) Hybridization of the cDNA mixture with the DNA array; (4) Image generation by scanning of the array with lasers (Duggan et al., 1999; Lbelda and Sheppard, 2000; Bowtell, 2000). We show this process in Figure 1.1.

As we see in Figure 1.1, the raw output of a microarray is presented as the actual image of the colours of the array spots. However, quantification of the intensity of the fluorescence and assignment of the numerical values are needed for analysis of the data. Presentation and analysis of the vast data generated by microarrays are an ongoing challenge, and some standards have recently been adopted (Ball et al., 2002). Active advances in the fields of statistics, computational biology, system biology, and bioinformatics promise to enhance our ability to interpret the large amount of data from microarrays in the future (Jason and Christie, 2005).

Microarray is considered as an important tool for advancing the understanding of the DNA information, molecular mechanism, biology and pathophysiology of critical illness. The expression of thousands of genes can be assessed, complex pathways can be more fully evaluated in a single experiment (Jason and Christie, 2005). Thus, microarrays could lead to discovery of new genes involved in disease processes. Meanwhile, microarrays can potentially be used to predict disease states on the basis of the expression profiles of specific cell populations, such as predicting development of sepsis in at-risk populations (Jason and Christie, 2005; Slonim, 2002). In addition, microarray can be used to monitor the biological response to new drugs in treatment trials (Brachat et al., 2002). Furthermore, different expression value of genes would be useful to classify different “flavors” of a syndrome, such as sepsis, on the basis of a molecular mechanism (Brunskill et al., 2011).

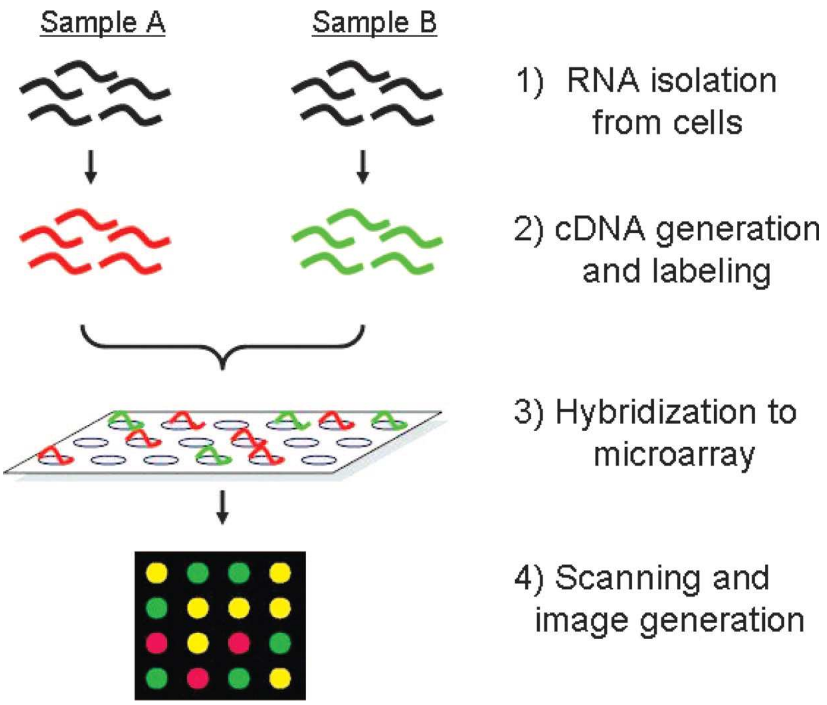

Sample A
Sample B
1) RNA isolation from cells
2) cDNA generation and labeling
3) Hybridization to microarray
4) Scanning and image generation

Figure 1.1 Four basic steps of the microarray experiment. (1) mRNA is isolated from cells; (2) cDNA is generated from the mRNA by reverse transcription, and a fluorescent tag is attaced; (3) the resulting tagged cDNA solution is hybridized to the DNA array; (4) the array is imaged by a laser fluorimeter and the color of each sopt is analysed. This figure is from (Lbelda and Sheppard, 2000). In this figure, a red spot indicates sample A>B; a green spot indicated sample A<B, and the yellow one indicated sample A=B. The illustrated example is for a comparative hybridization experiment. A relative intensity experiment would involve only one sample corrected for background expression or normalized with control genes.

Nowadays, DNA microarrays can be used in many bioscience and bioinformatics fields. These applications include: 1. Gene expression profiling. In an mRNA or gene expression profiling experiment, the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases and developmental stages on gene expression. For instance, gene expression profiling based on microarray can be applied to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected to that in uninfected cells or tissues (Schena et al., 1995; Lashkari et al., 1997). 2 Comparative genomic hybridization. Microarray can assess genome content in different cells or closely related organisms (Pollack et al., 1999; Moran et al., 2004). 3 Chromatin immunoprecipitation. The first chromatin immunoprecipitation assay was developed by Gilmour and Lis (1984, 1985, 1986) as a technique for monitoring the association of RNA polymerase II with transcribed and poised genes in Escherichia coli and Drosophila. 4 GeneID. Small microarrays can be used to check IDs of organisms in food and feed, mycoplasms in cell culture, or pathogens for disease detection (Kulesh et al., 1987). 5 SNP detection. For instance, people can use microarrays to identify single nucleotide polymorphism among alleles within or between populations (Hacia et al., 1999). 6 Fusion genes microarray. A fusion gene microarray can detect fusion transcripts from cancer specimens. The principle behind this is building on the alternative splicing microarrays (Lovf, et al., 2011). 7 Tiling array. Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. It is can be used to empirically detect expression of transcripts or alternatively splice forms which may not have been previously known or predicted (Bertone et al., 2005; Zacher et al., 2010).

In these applications, techniques such as cluster analysis, principal component analysis, and latent class models are most widely used. These methods all aim to group genes with similar expression profiles and then analyse the function and relationship between these grouped genes and disease (Slonim, 2002). Clustering analysis is the most common method for microarray evaluation. There are vast methods for clustering analysis, such as hierarchical or non-hierarchical methods; changing the distance measure that the cluster analysis uses to group genes; “supervising” the clustering with known information about biological relationships of the genes (Qu and Xu, 2004; Xiao et al, 2008) or using “unsupervised” methods to obtain the clustering structure and clustering numbers automatically (Boutros and Okey, 2005). Different methods may suit for different study designs and data. If more than one method is used and the results appear the same, it strengthens the conclusions (King and Sinha, 2001). The output from cluster analysis can be simplified by using boxes of artificial colours to represent changes in genes relative to each other as we do in Chapter 2. In this method, groups of genes with similar expression can be visualized according to the colour boxes, representing differences ore similarities in expression pattern (Chinnaiyan et al., 2001).

There is a vast literature about clustering methods on microarrays. Belacel et al. (2006) gave a general view of clustering techniques used in data analysis of microarray gene expressions. In their work, they provided a survey of various methods available for gene clustering and illustrated the impact of clustering methodologies on the fascinating and challenging area of genomic research. The strengths and weaknesses of each clustering technique are pointed out. Meanwhile, the development of software tools for clustering is also emphasized (Belacel et al, 2006).

As the development of clustering methods on microarrays continues, some outstanding achievements have been obtained in the past decades. Statistical tools are widely used in this field. Medvedovic et al. (2004) developed different variants of Bayesian mixture based clustering procedures for clustering gene expression data with experiment replicates. In this method, a Bayesian mixture model is used to describe the distribution of microarray data. Clusters of co-expressed genes are created from the posterior distribution of clustering, which is estimated by a Gibbs sampler. In their work, they demonstrated that the Bayesian infinite mixture model with “elliptical” variances structure is capable of identifying the underlying structure of the data without knowing the correct number of clusters. This is a useful unsupervised clustering method.

Because of the experiment condition or some other factors, microarray data are sometimes incomplete. Missing values may affect the clustering result. Zhang and

Zhu (2002) proposed a novel clustering approach to overcome data missing and inconsistency of gene expression levels under different conditions in the stage of clustering. It is based on the smooth score, which is defined for measuring the deviation of the expression level of a gene and the average expression level of all the genes involved under a condition. The algorithm was tested intensively on random matrices and yeast data. It was shown to perform well in finding co-regulation patterns in a test with the yeast data.

Many bioinformatics problems can be tackled form a fresh angle offered by the network perspective. Zhu et al. (2005) proposed a gene clustering approach based on the construction of co-expression networks that consist of both significantly linear and non-linear gene associations together with controlled biological and statistical significance. This method is used to group functionally related genes into tight clusters despite the expression dissimilarities (Zhu et al., 2005). According to comparison with some traditional approaches on a yeast galactose metabolism dataset, their method performed well in rediscovering the relatively well known galactose metabolism pathway in yeast and in clustering genes of the photoreceptor differentiation pathway.

Getz et al. (2000) presented a coupled two-way clustering approach to gene microarray data analysis. The main idea is to identify subsets of the genes and samples, such that when one of these is used to cluster the other, stable and significant partitions emerge. This algorithm is based on iterative clustering and especially suitable for gene microarray data. It is applied to two gene microarray datasets, colon cancer and leukaemia respectively. The results showed this method is able to discover partitions and correlations that are masked and hidden when the full dataset was used in the analysis. Some of these partitions have clear biological interpretation (Getz et al., 2000).

How to choose the cluster numbers is a critical problem for supervised clustering analysis. Ma and Huang (2007) proposed a method based on gap statistic to determined the optimal number of clusters. This method is a clustering threshold gradient descent regularization (CTGDR) method, for simultaneous cluster selection and within cluster gene selection. This approach was applied to binary classification and censored survival analysis. Compared with the standard TGDR and other regularization methods, the CTGDR takes into account the cluster structure and carries out feature selection at both the cluster level and within-cluster gene level (Ma and Huang 2007).

Thalamuthu et al. (2006) made a comparison on six clustering methods. They are hierarchical clustering, K-means, PAM, SOM, mixture model-based clustering and tight clustering. Performance of the methods is assessed by a predictive accuracy analysis through verified gene annotations. The results show that tight clustering and model-based clustering consistently outperform other clustering methods both in simulated and real data, while hierarchical clustering and SOM perform poorly. Their analysis provides insight for the complicated gene clustering problem using expression profile and serves as a practical guideline for routine microarray cluster analysis (Thalamuthu et al., 2006).

De Bin and Risso (2011) presented a general framework to deal with the clustering of microarray data based on a three-step procedure: (i) gene filtering; (ii) dimensionality reduction; (iii) clustering of observations in the reduced space. Via a nonparametric model-based clustering approach they obtained promising results.

Gaussian mixture models are also widely used for gene clustering analysis. McNicholas and Murphy (2010) extended a family of eight mixture models which utilize the factor analysis covariance structure to 12 models and applied to gene expression microarray data for clustering analysis. This family of models allows for the modelling of the correlation between gene expression levels even when the number of samples is small. This expanded family of Gaussian mixture models, known as the expanded parsimonious Gaussian mixture model family, was then applied to two well-known gene expression data sets. The performance of this family of models is quantified using the adjusted rank index. Their method performs well relative to existing popular clustering techniques when applied to real gene expression data.

There is also vast literature about the application of other clustering methods to microarray data. Koenig and Youn (2011) proposed a hierarchical signature clustering method for microarray data. In their work, they proposed a new metric instead of Euclidean metric. Subhani, et al. (2010) introduced a pairwise gene expression profile alignment and defined a new distance function that is appropriate for time-series profiles. Extensive experiments on well-known datasets yield encouraging results of at least 80% classification accuracy. Macintyre et al. (2010) developed a novel clustering algorithm, which incorporates functional gene information from the gene ontology into the clustering process, resulting in more biologically meaningful clusters. Romdhane et al. (2010) developed an unsupervised "possibilistic" approach for mining gene microarray data. The optimal number of clusters is evaluated automatically from the data using the information entropy as a validity measure. Experimental results using real-world data sets reveal a good performance and a high prediction accuracy from this model.

### 1.2.2 Methods

*Fuzz set theory and fuzzy c-means*

For the work of clustering analysis of microarrays, we applied fuzzy c-means method which is a widely used clustering method in many fields. Traditional hard clustering methods, such as K-means or SOM which assign each gene exactly to one cluster, are poorly suited to the analysis of microarray data because the clusters of genes frequently overlap in such data (Dembel and Kanstner, 2003). Fuzzy theory has many advantages in dealing with data containing uncertainty, thus it is introduced into analysis of DNA microarrays (Fu and Medico, 2002).

Zadeh (1965), the first publication on fuzzy set theory, shows the intention to generalize the classical notion of a set and a proposition to accommodate fuzziness in human judgment, evaluation, and decisions. Since its appearance, the theory of fuzzy sets has advanced in a variety of ways and in many disciplines. Nowadays, there are more than 30,000 publications about fuzzy theory and methods and their applications (Zimmermann, 2010). Roughly speaking, fuzzy set theory has developed along two lines during the last decades: (1) As a formal theory that, when maturing, becomes more sophisticated and specified and is enlarged by original ideas and concepts as well as by "embracing" classical mathematical areas, such as algebra (Dubois and Prade, 1997; Liu, 1998), graph theory, topology, and by generalizing or "fuzzifying" them. This development is still ongoing. (2) As an application-oriented "fuzzy technology", that is, as a tool for modeling, problem solving, and data mining that has been proven superior to existing methods in many cases and as attractive "add-on" to classical approaches in other cases (Zimmermann, 2010).

Applications of this theory can be found, for example, in artificial intelligence (Freeman, 1994), computer science (Yager and Zadeh, 1992), medicine (Maiers, 1985), control engineering (Tong et al, 2010), decision theory (Liu, 2008), expert systems (Siler and Buckley, 2005), logic (Ross, 2004), management science (Grint, 1997), operations research (Herrera and Verdegay, 1997), pattern recognition (Pedrycz, 1990), and robotics (Fukuda and Kubota, 1999).

With the development of electronic data processing, clustering analysis of these data becomes more and more important. Classical methods for data mining, such as clustering techniques, were available, but sometimes they did not match the needs. Because clustering techniques, for instance, assume that data could be subdivided crisply into clusters, they would not fit the structures that existed in reality. Fuzzy set theory seems to offer good opportunities to improve existing concepts. Bezdek (1978, 1981) was the first one to develop fuzzy clustering method with the goals to search for structure in data to reduce complexity and to provide input for control and decision making. He proposed and developed the most famous fuzzy clustering method: Fuzzy C-means Method (FCM). Nowadays, FCM and its combined methods are applied in various fields.

FCM combined methods are widely used for image segmentation. Ji et al. (2011) proposed a modified possibilistic fuzzy c-means clustering algorithm for fuzzy segmentation of magnetic resonance (MR) images that have been corrupted by intensity inhomogeneities and noise. By combining a novel adaptive method to compute the spatially local weights in the objective function, this method is capable of utilizing local contextual information to impose local spatial continuity, thus allowing the suppression of noise and helping to resolve classification ambiguity. Comparisons with other approaches demonstrate the superior performance of the proposed algorithm and this method is robust to initialization.

Zhang and Chen (2004) proposed a fuzzy segmentation method for magnetic resonance imaging data. This algorithm is realized by modifying the objective function in the conventional fuzzy c-means algorithm using a kernel-induced distance metric and a spatial penalty on the membership functions. In this method, the original Euclidean distance is replaced by a kernel induced distance, and then a spatial penalty is added to the objective function in FCM to compensate for the intensity inhomogeneities of MR image and to allow the labelling of a pixel to be influenced by its neighbours in the image. Experimental results on both synthetic and real MR images show that the proposed algorithm has better performance when noise and other artifacts are present than the standard algorithms (Zhang and Chen, 2004).

Chuang et al. (2006) proposed a fuzzy c-means algorithm that incorporates spatial information into the membership function for clustering. The spatial function is the summation of the membership function in the neighbourhood of each pixel under consideration. The advantages of the new method are the following: (1) it yields regions more homogeneous than those of other method. (2) it reduces the spurious blobs, (3) it removes noisy spots, and (4) it is less sensitive to noise than other techniques. This technique is a powerful method for noisy image segmentation and works for both single and multiple-feature data with spatial information (Chuang et al., 2006).

In bioinformatics, FCM and its combined methods are also efficient tools for clustering analysis. Dembel and Kastner (2003) firstly applied FCM on clustering analysis of microarray data. A major problem in applying the FCM method for clustering microarray data is the choice of the fuzziness parameter *m*. Usually, researchers use $m = 2$, however, it is not appropriate for some data sets. Thus they proposed an empirical method, based on the distribution of distances between genes in a given data set, to determine an adequate value of *m*. By setting threshold levels for the membership values, genes which are tightly associated to a given cluster can be selected. Using a yeast cell cycle data set as an example, it is shown that the selection increases the overall biological significance of the genes within the cluster (Dembel and Kastner, 2003).

Wang et al. (2003) proposed a novel FCM method for tumor classification and target gene prediction. In this method gene expression profiles are firstly summarized by optimally selected self-organizing maps (SOMs), followed by tumor sample classification by fuzzy c-means clustering. Then, the prediction of marker genes is accomplished by either manual feature selection or automatic feature selection. This method is tested on leukemia, colon cancer, brain tumors and NCI cancer cell lines. The method gave class prediction with markedly reduced error rates compared to other class prediction approaches, and the important of feature selection on microarray data analysis was also emphasized (Wang et al., 2003).

Asyali and Alci (2005) discussed reliability analysis of microarray data using FCM and normal mixture modeling based classification methods. A serious limitation in microarray analysis is the unreliability of the data generated from low signal intensities. Such data may produce erroneous gene expression ratios and cause unnecessary validation or post-analysis follow-up tasks. Therefore, the elimination of unreliable signal intensities will enhance reproducibility and reliability of gene expression ratios produced from microarray data. In their work, they applied fuzzy c-means and normal mixture modeling based classification methods to separate microarray data into reliable and unreliable signal intensity populations. According to the comparison between these two methods, fuzzy approach is computationally more efficient.

Seo et al. (2006) identified the effect of data normalization for application of FCM on clustering analysis of microarray. In their work, they used three normalization methods, the two common scale and location transformations and lowest normalization methods, to normalize three microarray datasets and three simulated datasets. They found the optimal fuzzy parameter *m* in the FCM analysis of a microarray dataset depends on the normalization method applied to the dataset during preprocessing. Lowest normalization is more robust for clustering of genes from microarray data, especially when FCM is used in the analysis.

*Empirical mode decomposition*

Data analysis helps to construct models for practical problems and understand phenomena in many research fields. However, the data available often have different characteristics, such as trends, seasonality and non-stationarity (Huang et al. 1998). In these cases, researchers have to try various methods to deal with different features. Spectral analysis has been applied in the study of these problems in the frequency domain. Spectral analysis has many constraints, such as linearity and stationarity (Conte and de Boor 1980). The related spectrogram needs the traditional Fourier transform and slides along the time axis (Huang et al. 1998). This method can work well on piecewise stationary data, but it needs to choose window width (Huang et al. 1998). Evolutionary spectral analysis extends Fourier spectral analysis to generalized basis (Priestley 1965). This method has a family of orthogonal basis indexed by time and frequency, and the signal function was expressed with Stieltjes integration of these orthogonal functions and the amplitude (Priestley 1965). However, a constraint of this method is to define the basis function.

The empirical mode decomposition was proposed to obtain more information from data. It defines a class of functions called intrinsic mode functions that have some specific properties. For example, the number of extrema and the number of zero-crossings are almost the same, and the mean value of the envelope formed by local maxima and the envelope formed by local minima is zero (Huang et al. 1998). These characteristics not only are used as the traditional narrow band for a stationary Gaussian process, but also reduce the unnecessary fluctuations by asymmetric waves. Janušauskas et al. (2006) used EMD and wavelet transform to process ultrasound signals for human cataract detection. They decomposed the signal and enhanced specific features with both methods. In their results, EMD performed better in the detection of signal than the discrete wavelet transform.

Shi et al. (2007) studied the functional similarity of proteins using the EMD method, and they compared the results with those from the pair-wise alignment and PSI-BLAST. However, their work did not cover complete comparisons, and still needs further improvement.

## 1.3 Identification of disease-associated genes

### 1.3.1 Biological background and literature review

Disease-associated gene identification is one of the most important areas of medical research today. Many current methods for disease-associated gene identification are based on protein-protein interaction networks and microarray data. It is known that certain diseases, such as cancer, are reflected in the change of the expression values of certain genes. For instance, due to genetic mutations, normal cells may become cancerous. These changes can affect the expression level of genes. Gene expression is the process of transcribing a gene’s DNA sequence into RNA. A gene’s expression level indicates the approximate number of copies of that gene’s RNA produced in a cell and it is correlated with the amount of the corresponding proteins made (Mohammadi et al., 2011). Analysing gene expression data can indicate the genes which are differentially expressed in the diseased tissues. In the past decades, both kinds of methods have important breakthroughs and progresses.

Shaul et al. (2009) proposed a new algorithm for predicting disease-causing genes (causal genes) based on gene networks established according to gene expression values. The algorithm relies on the assumption that in the disease state, one or more causal genes are disrupted, leading to the expression changes of downstream (disease-related) genes through signaling regulatory pathways in the network. Gene expression data under disease conditions have been used to highlight a set of disease-related genes that are assumed to be in close proximity to the causal genes in the gene network. Then based on this assumption, a greedy heuristic that recovers putative causal genes as those admitting pathways to a maximal number of disease-related genes has been applied.

It is believed that a large number of genes are involved in common human brain diseases. Liu et al. (2006) proposed a novel computational strategy for simultaneously identifying multiple candidate genes for genetic human brain diseases from a brain-specific gene network level perspective. This approach includes two main steps as follows. (1) Construction of the human brain-specific gene network based on the expression value; (2) Identification of the sub-network.

Kohler et al. (2008) have investigated the hypothesis that global network-similarity measures are better suited to capture relationships between disease proteins than are algorithms based on direct interactions or shortest paths between disease genes. In this approach, 110 disease-gene families have been defined and a protein-protein interaction network has been established based on a total of 258314 experimentally verified or predicted protein-protein interactions. This approach adapts a global distance measure based on a random walk with restart (RWR) to define similarity between genes within the protein-protein interaction network and ranks candidates on the basis of this similarity to known disease genes.

Sun et al. (2011) combined four clustering methods to decompose a human PPI network into dense clusters as the candidates of disease-related clusters, and then a log likelihood model that integrates multiple biological evidences was proposed to score these dense clusters. They identified disease-related clusters using these dense clusters if they had higher scores. The efficiency was evaluated by a leave-one-out cross validation procedure. Their method achieved a success rate of 98.59% and recovered the hidden disease-related clusters in 34.04% cases when one known disease gene is removed. They also found that most of the disease-related clusters consist of tissue-specific genes that were highly expressed only in one or several tissues, and a few of those were composed of housekeeping genes (maintenance genes) that were ubiquitously expressed in most of the tissues.

Firneisz et al. (2003) applied a friends-of-friends algorithm to identify significant gene clusters on microarray data. Using a set of cDNA microarray chip experiments in two mouse models of rheumatoid arthritis, they identified more than 200 genes based on their expression in inflamed joints and mapped them into the genome.

Liang et al. (2006) proposed an innovative approach, the fuzzy membership test (FM-test), based on fuzzy set theory to identify disease associated genes from microarray gene expression profiles. They applied this method on diabetes and lung cancer data. Within the 10 significant genes identified in diabetes dataset, 6 of them have been confirmed to be associated with diabetes in the literature. Within the 10 best ranked genes in lung cancer data, eight of them have been confirmed by the literature.

Among numerous existing methods for gene selection, the support vector machine-based recursive feature elimination (SVMRFE) has become one of the leading methods, but its performance can be reduced because of the small sample size, noisy data and the fact that the method does not remove redundant genes. Mohammadi et al. (2011) proposed a novel framework for gene selection which uses the advantageous features of conventional methods and addresses their weaknesses. They have combined the Fisher method and SVMRFE to utilize the advantages of a filtering method as well as an embedded method. Furthermore, a redundancy reduction stage is added to address the weakness of the Fisher method and SVMRFE. The proposed method has been applied to colon, Diffuse Large B-Cell Lymphoma (DLBCL) and prostate cancer datasets. It predicts marker genes for colon, DLBCL and prostate cancer with a high accuracy. The predictions made in this study can serve as a list of candidates for subsequent wet-lab verification and might help in the search for a cure for cancers (Mohammadi et al. 2011).

Yoon et al. (2006) presented a new data mining strategy to better analyze the marginal difference in gene expression between microarray samples. The idea is based on the notion that the consideration of gene's behavior in a wide variety of experiments can improve the statistical reliability on identifying genes with moderate changes between samples. This approach was evaluated via the re-identification of breast cancer-specific gene expression. It successfully prioritized several genes associated with breast tumor, for which the expression difference between normal and breast cancer cells was marginal and thus would have been difficult to recognize using conventional methods. Maximizing the utility of microarray data in the public database, it provides a valuable tool particularly for the identification of previously unrecognized disease-related genes.

Watkinson et al. (2010) presented a computational methodology that jointly analyse two sets of microarray data, one in the presence and one in the absence of a disease, identifying gene pairs whose correlation with disease is due to cooperative, rather than independent, contributions of genes, using the recently developed information theoretic measure of synergy. High levels of synergy in gene pairs indicates possible membership of the two genes in a shared pathway and leads to a graphical representation of inferred gene-gene interactions associated with disease, in the form of a “synergy network”. They applied this technique on a set of publicly available prostate cancer expression data. The results show that synergy networks provide a computational methodology helpful for deriving "disease interactomes" from biological data. When coupled with additional biological knowledge, they can also be helpful for deciphering biological mechanisms responsible for disease.

Li et al. (2010a) proposed a method for prediction of disease-related genes based on hybrid features. In their study, multiple sequence features of known disease-related genes in 62 kinds of disease were extracted, and then the selected features were further optimized and analysed for disease-related genes prediction.

Zhang et al. (2010) adopted the topological similarity in human protein-protein interaction networks to predict disease-related genes. This method is specially designed for predicting disease-related genes of single disease-gene family based on PPI data. The application results show a significant abundance of disease-related genes that are characterized by higher topological similarity than other genes.

### 1.3.2 Methods

We introduced type-2 fuzzy set theory into the research of disease-associated gene identification. Type-2 fuzzy set is an extension of traditional fuzzy set introduced by Zadeh (1975). Of course, employment of type-2 fuzzy sets usually increases the computational complexity in comparison with type-1 fuzzy sets due to the additional dimension of having to compute secondary grades for each primary membership. However, if type-1 fuzzy set does not perform satisfactorily, employment of type-2 fuzzy sets for managing uncertainty may allow us to obtain desirable results (Hwang and Rhee, 2007). Mizumoto and Tanaka (1976) studied the set theoretic operations of type-2 sets and the properties of membership grades of such sets, and examined their operations of algebraic product and algebraic sum (Mizumoto and Tanaka, 1981). Dubois and Prade (1980) discussed the join and meet operations between fuzzy numbers under minimum t-norm. Karnik and Mendel (1998, 2000) provided a general formula for the extended sup-star composition of type-2 relations. Choi and Rhee (2009) did some work on the methods for establishing interval type-2 fuzzy membership function for pattern recognition. Greenfiled et al. (2009) discussed the collapsing method of defuzzification for discretised interval type-2 fuzzy sets. Mendel (2007) introduced some important advances that have been made during the past 5 years for both general and interval type-2 fuzzy sets and systems.

Type-2 fuzzy sets have already been used in a number of applications, including decision making (Chaneau et al., 1987; Yager, 1980), solving fuzzy relation equations (Wagenknecht and Hartmann, 1988), and pre-processing of data (John et al., 1998), neural networks (Rafik et al., 2011), controller design (Kumbasar et al., 2011), genetic algorithms (Wu and Tan, 2006) and so on.

Huarng and Yu (2005) presented a type-2 fuzzy time series model for stock index forecasting and made a comparison with type-1 fuzzy model. Most conventional fuzzy time series models (Type-1 models) utilize only one variable in forecasting. Furthermore, only parts of the observations in relation to that variable are used. To utilize more of that variable's observations in forecasting, this study proposes the use of a Type-2 fuzzy time series model. The Taiwan stock index, the TAIEX, is used as the forecasting target. Their empirical results show that type-2 model outperforms type-1 model.

Jeon et al. (2009) designed a type-2 fuzzy logic filter for improving edge-preserving restoration of interlaced-to-progressive conversion. In their work, they focused on advance fuzzy models and the application of type-2 fuzzy sets in video deinterlacing. The final goal of the proposed deinterlacing algorithm is to exactly determine an unknown pixel value while preserving the edges and details of the image. In order to address these issues, they adopted type-2 fuzzy set concepts to design a weight evaluating approach. In the proposed method, the upper and lower fuzzy membership functions of the type-2 fuzzy logic filters are derived from the type-1 fuzzy membership function. The weights from upper and lower membership functions are considered to be multiplied with the candidate deinterlaced pixels. Experimental results showed that the performance of the proposed method was superior, both objectively and subjectively, to other different conventional deinterlacing methods. Moreover, the proposed method preserved the smoothness of the original image edges and produced a high-quality progressive image (Jeon et al., 2009).

Balaji and Srinivasan (2010) presented a multi-agent system based on type-2 fuzzy decision module for traffic signal control in a complex urban road network. The distributed agent architecture using type-2 fuzzy set based controller was designed for optimizing green time in a traffic signal to reduce the total delay experienced by vehicles. A section of the Central Business District of Singapore simulated using PARAMICS software was used as a test bed for validating the proposed agent architecture for the signal control. The performance of the proposed multi-agent controller was compared with a hybrid neural network based hierarchical multi-agent system (HMS) controller and real-time adaptive traffic controller (GLIDE) currently used in Singapore. The performance metrics used for evaluation were total mean delay experienced by the vehicles to travel from source to destination and the current mean speed of vehicles inside the road network. The proposed multi-agent signal control was found to produce a significant improvement in the traffic conditions of the road network reducing the total travel time experienced by vehicles simulated under dual and multiple peak traffic scenarios (Balaji and Srinivasan, 2010).

Leal-Ramirez et al. (2010) proposed an age-structured population growth model based on a fuzzy cellular structure. An age-structured population growth model enables a better description of population dynamics. In this paper, the dynamics of a particular bird species was considered. The dynamics is governed by the variation of natality, mortality and emigration rates, which in this work are evaluated using an interval type-2 fuzzy logic system. The use of type-2 fuzzy logic enables handling the effects caused by environment heterogeneity on the population. A set of fuzzy rules, about population growth, are derived from the interpretation of the ecological laws and the bird life cycle. The proposed model is formulated using discrete mathematics within the framework of a fuzzy cellular structure. The fuzzy cellular structure allows us to visualize the evolution of the population’s spatial dynamics. The spatial distribution of the population has a deep effect on its dynamics. Moreover, the model enables not only to estimate the percentage of occupation on the cellular space when the species reaches its stable equilibrium level, but also to observe the occupation patterns (Leal-Ramirez et al., 2010).

Fazel Zarandi et al. (2007) presented a new type-2 fuzzy logic system model for desulphurization process of a real steel industry in Canada. In this research, the Gaussian mixture model was used for the creation of second order membership grades. Furthermore, a reduction scheme was implemented which results in type-1 membership grades. In turn, this leads to a reduction of the complexity of the system. The result shows that the proposed type-2 fuzzy logic system is superior in comparison to multiple regression and type-1 fuzzy logic systems in terms of robustness and error reduction.

Fazel Zarandi et al. (2009) also applied type-2 fuzzy set theory to stock price analysis. They developed a type-2 fuzzy rule based expert system on the forecast of stock price. The proposed method applies the technical and fundamental indexes as the input variables. This model is tested on stock price prediction of an automotive manufactory in Asia. Through the intensive experimental tests, the model has successfully forecasted the price variation for stocks from different sectors. The results are very encouraging and can be implemented in a real-time trading system for stock price prediction during the trading period (Fazel Zarandi et al., 2007).

## 1.4 Identification of protein complexes from PPI networks

### 1.4.1 Biological background and literature review

In the "post-genome" era, proteomics (Palzkill, 2002; Waksman, 2005) has become an essential field and drawn much attention. Proteomics is the systematic study of the many and diverse properties of proteins with the aim of providing detailed descriptions of the structure, function, and control of biological systems in health and diseases.

A particular focus of the field of proteomics is the nature and role of interactions between proteins. Protein-protein interactions (Palzkill, 2002; Park et al., 2009; Peink et al., 1998; Pellegrini et al., 1999; Qi et al., 2007; Rao and Srinivas, 2003; Rumelhart et al., 1986) play different roles in biology depending on the composition, affinity, and lifetime of the association. It has been observed that proteins seldom act as single isolated species while performing their functions in vivo. The study of protein interactions is fundamental to understand how proteins function within a cell.

Protein-protein interaction plays a key role in the cellular processes of an organism. An accurate and efficient identification of protein-protein interaction is fundamental for us to understand the physiology, cellular functions, and complexity of an organism. Before the year 2000 most theoretical methods to predict protein-protein interactions are based on available complete genomes such as the phylogenetic profiles, domain fusion or Rosetta stone method, and gene neighbor method, etc.

The knowledge of protein-protein interaction can provide important information on the possible biological function of a protein. Much effort has been done to detect and analyze protein-protein interactions using experimental methods such as the yeast two-hybrid system which is well known. Recently, several algorithms have been developed to identify functional interactions between proteins using computational methods which can provide clues for the experimental methods and could simplify the task of protein interaction mapping. As the prediction task becomes harder the need for methods that can accommodate high levels of missing values and are directly interpretable by biologists increases.

***Phylogenetic profiles***

The phylogenetic profile (Cubellis et al., 2005; Hoskins et al., 2006; Karimpour-Fard et al., 2007), which is also called the co-conservation method, is a computational method which has been used to predict functional interactions between pairs of proteins in a target organism by determining whether both proteins are consistently present or absent across a set of reference genomes. This method was first introduced by Pellegrino et al. (1999) and it has been successfully applied to the prediction of protein function by several groups and proved to be more powerful than sequence similarity alone at predicting protein function.

Hoskins et al. (2006) took E. coli K12 as the target genome and performed three steps:

i. Creating a phylogenetic profile vector where $P_{ij} = 1$ indicating a homolog exists between protein *i* in the target genome and a protein j in a reference genome;

ii. Calculating similarity measurements on the profile vectors for each pair of genes in the target genome;

iii. Defining protein interactions in the target genome based on proteins sharing a profile similarity value greater than a threshold value.

They measured the performance and reliability of their method over previous methods through comparing the number of interacting proteins, the number of predicted unknown proteins and the functional similarity of proteins sharing a protein-protein interaction. They showed that the selection of reference organisms had a substantial effect on the number of predictions involving proteins of previously unknown function, the accuracy of predicted interactions, and the topology of predicted interaction networks. They proved predicted interactions are influenced by the similarity metric that is employed and differences in predicted protein interactions are biologically meaningful.

***PPI prediction with neural networks***

Neural networks (Schalkoff, 1997) are now a subject of interest to professionals in many fields and it is also a tool for many areas of problem solving. Just as human brains can be trained to master some situations, neural networks can be trained to recognize patterns and to do optimization and other tasks. Some researchers have used neural networks to predict protein-protein interaction.

Fariselli et al. (2002) proposed a method to predict PPI sites with neural networks. Their method was a feed-forward neural network (Rao and Srinivas, 2003; Rumelhart et al., 1986) trained with the standard back-propagation algorithm. The network system was trained and tested to predict whether each surface residue was in contact with another protein or not. The network architecture contains an output layer which consists of a single neuron representing contact or non-contact. They tested their predictor using different numbers of hidden neurons and the best performance was obtained with a hidden layer containing four nodes. They analyzed the possibility of predicting the residues forming part of protein-protein interacting surfaces in proteins of known structure. They used two very basic sources of information: evolutionary information as accumulated in sequence profiles derived from family alignments, and surface patches in protein structures identified as sets of neighbor residues exposed to solvent. The result is surprising because their prediction could come up with an average accuracy of about 73%.

### *Mixture-of-feature-experts method*

There are two important difficulties for the PPI prediction task. First, previous classification methods estimate a set of parameters that are used for all input pairs. However, the biological datasets used contain many missing values and highly correlated features. Thus, different samples may benefit from using different feature sets. The second difficulty is that researchers who want to use these methods to select experiments cannot easily determine which of the features contributed to the resulting prediction. Because different researchers may have different opinions regarding the reliability of the various feature sources, it is useful if the method can indicate, for every pair, which feature contributes the most to the classification result. So some researchers proposed a mixture-of-feature-experts (MFE) method (Qi et al., 2007) for protein-protein interaction prediction.

There are many biological data sets that may be directly or indirectly related to PPIs. Qi et al. (2007) have tried to collect as many sets as possible for yeast and human being. For different data sources, each of them has its own representative form. These researchers collected a total of 162 feature attributes from 17 different data sources for yeast and a total of 27 feature attributes from 8 different data sources for human being, and then divided the biological data sources into four feature categories, which are referred to as feature experts in the paper:

Expert P: direct high-throughput experimental PPI data
Expert E: indirect high-throughput data
Expert S: sequence based data sources
Expert E: functional properties of proteins.

After that Qi et al. (2007) used the MFE framework as classifiers to modify the weights of different feature experts. To measure the ability of the MFE method to predict PPIs, they compared it with other popular classifiers that have been suggested in the past for this task and showed that the MFE method improved the classification outcome. This method is useful for overcoming problems in achieving high prediction performance arising due to missing values which are a major issue when analyzing biological data sets.

***Properties of PPI networks***

The simplest representation of PPI networks takes the form of a mathematical graph consisting of nodes and edges (or links). Proteins are represented as nodes and an edge represents a pair of proteins which physically interact. The degree of a node is the number of other nodes with which it is connected. It is the most elementary characteristic of a node.

A protein-protein interaction network has three main properties (Hu and Pan, 2007): scale invariance, disassortativity and small-world effect. Much work has been done to study these properties and to find new ones.

Scale invariance: in scale-free networks, most proteins participate in only a few interactions, while a few participate in dozens of interactions.

Small-world effect means that any two nodes can be connected via a short path of a few links. The small-world phenomenon was first investigated as a concept in sociology and is a feature of a range of networks arising in nature and technology such as the most familiar one: Internet.

Disassortativity: in protein-protein interaction networks the nodes which are highly connected are seldom link directly to each other. This is very different from social networks in which well-connected people tend to have direct connections to each other. All biological and technological networks have the property of disassortativity.

***Protein-protein interaction network and protein complexes***

Protein complex (or multi-protein complex) is a group of two or more proteins. No protein is an island entire of itself or at least very few proteins are. Most proteins seem to function with complicated cellular pathways, interacting with other proteins either in pairs or as components of large complexes. So identification of protein complexes is crucial for understanding the principles of cellular organization and functions. As the size of protein-protein interaction sets increases, a general trend is to represent the interaction as network and to develop effective algorithms to detect significant complexes in such networks. Various methods have been used to detect protein complexes.

Partitional clustering approaches can partition a network into multi separated sub-networks. As a typical example, the Restricted Neighbourhood Search Clustering (RNSC) algorithm (King et al., 2004) developed the best partition of a network by using a cost function. The method starts with randomly partitioning a network, and iteratively moves a vertex from one cluster to another to decrease the total cost of clusters. When some moves have been reached without decreasing the cost function, it ends. This method can obtain the best partition by running multi-times. However, it needs the number of clusters as prior knowledge and its results depend heavily on the quality of initial clustering. Moreover, it cannot get the overlapping protein complexes since it requires each vertex belonging to a specific cluster.

Hierarchical clustering approaches build (agglomerative), or break up (divisive), a hierarchy of clusters. The traditional representation of this hierarchy is a tree (called a dendrogram). Agglomerative algorithms start at the top of the tree and iteratively merge vertices, whereas divisive algorithms begin at the bottom and recursively divide a graph into two or more sub-graphs. For iteratively merging vertices, the similarity or distance between two vertices should be measured. The Super Paramagnetic Clustering (SPC) algorithm (Spirin and Mirny, 2003) is an example of iterative merging. For recursively dividing a graph, the vertices or edges to be removed should be selected properly. The Highly Connected Sub-graph (HCS) algorithm (Hartuv and Shamir, 2000) uses the minimum cut set to remove edges recursively. Girvan and Newman (Girvan and Newman, 2002; Newman, 2004) decomposed a network based on the graph theoretical concept of betweenness centrality. Luo et al. (2007) also used betweenness and developed a new algorithm named MoNet. Hierarchical clustering approaches can display the hierarchical organisation of biological networks. To our knowledge, all methods of predicting PPIs cannot avoid yielding a non-negligible amount of noise (False Positives, FP). As a disadvantage, the hierarchical clustering approaches are sensitive to noisy data (Cho et al., 2007).

Density-based clustering approaches detect densely connected sub-graphs from a network. An extreme example is to identify all fully connected sub-graphs (cliques) of $d$ = 1 (Spirin and Mirny, 2003). However, all methods of protein interaction predictions are known to yield a non-negligible rate of false positives and to miss a fraction of existing interactions. Thus, only mining fully connected sub-graphs is too restrictive to be used in real biological networks. In general, sub-graphs are identified by using a density threshold. A variety of alternative density functions have been proposed to detect dense sub-graphs (Bader and Hogue, 2003; Altaf-Ul-Amin et al., 2006; Pei and Zhang, 2007). The Clique Percolation Method (CPM) (Palla et al., 2005) detects overlapping protein complexes as k-clique percolation clusters. A k-clique is a complete sub-graph of size k. On the basis of CPM, a powerful tool named CFinder (Adamcsek et al., 2006) for finding overlapping protein complexes has been developed.

There are some other methods for protein complex detection. Habibi et al. (2010) proposed a protein complex prediction method which is based on connectivity number on sub-graphs. This method was applied to two benchmark data sets, containing 1142 and 651 known complexes respectively and it performed well. Jung et al. (2010) proposed a protein complex prediction method based on simultaneous protein interaction networks. This concept is introduced to specify mutually exclusive interactions (MEI) as indicated from the overlapping interfaces and to exclude competition from MEIs that arise during the detection of protein complexes. Ozawa et al. (2010) introduced a combinatorial approach for prediction of protein complexes focusing not only on determining member proteins in complexes but also on the PPI organization of the complexes. Cannataro et al. (2010) proposed a new complexes meta-predictor which is capable of predicting protein complexes by integrating the results of different predictors. It is based on a distributed architecture that wraps predictor as web/grid services that is built on top of the grid infrastructure.

### 1.4.2 Methods

We combine fuzzy relation clustering method with the graph model. Let $X_1,\ldots, X_n$ be n universes. An n-ary fuzzy relation $R$ in $X_1\times\ldots\times X_n$ is a fuzzy set on $X_1\times\ldots\times X_n$. An ordinary relation is a particular case of fuzzy relations, whose membership value is just 0 or 1. Since the proposal of fuzzy set theory by Zadeh in 1965, the work on fuzzy relation clustering has been vast (Zadeh, 2005; Baraldi, et al., 1999; Borgelt, 2009).

Dib and Youssef (1991) followed Zadeh's work and gave a new approach to Cartesian product, relations and functions in fuzzy set theory. A concept of fuzzy Cartesian product is introduced using a suitable lattice. A fuzzy relation is then defined as a subset of the fuzzy Cartesian product analogously to the crisp case. For fuzzy equivalence relations, they obtained similar results to those of ordinary equivalence relations. For fuzzy functions, they obtained a generalization of Zadeh's definition in terms of a family of ordinary functions. These introduced concepts and provided new tools to attack many problems in fuzzy mathematics.

Dudziak (2010) studied graded properties (α-properties) of fuzzy relations, which are parameterized versions of properties of a fuzzy relation defined by Zadeh. They took into account fuzzy relations which are α-reflexive, α-irreflexive, α-symmetric, α-antisymmetric, α-asymmetric, α-connected, α-transitive, where $\alpha \in [0,1]$. They studied the composed versions of these basic properties, e.g. an α-equivalence, α-orders as well. They also considered the so-called "weak" properties of fuzzy relations which are the weakest versions of the standard properties of fuzzy relations. They took into account the same types of properties as in the case of the graded ones. Using functions of n variables they considered an aggregated fuzzy relation of given fuzzy relations. They gave conditions for functions to preserve graded and weak properties of fuzzy relations.

Ciric et al. (2008) introduced and studied the concepts of a uniform fuzzy relation and a uniform F-function. They gave various characterizations and constructions of uniform fuzzy relations and uniform F-functions and showed that the usual composition of fuzzy relations is not convenient for *F*-functions; thus they introduced another kind of composition, and established a mutual correspondence between uniform *F*-functions and fuzzy equivalences. They applied the uniform fuzzy relations in some fuzzy control problem and the result shows uniform fuzzy relations are closely related to the defuzzification problem.

Dudziak and Kala (2008) studied bipolar fuzzy relations. This relation turns out to be an equivalence in the family of all bipolar fuzzy relations in a given set. It also has many other properties which seem to be useful in applications. Moreover, they proposed new types of properties for bipolar fuzzy relations which are compatible with standard relations.

A fuzzy relation can effectively describe the uncertain information between two objectives, like the concepts “similar” and “different” (Zadeh, 1965). The clustering methods based on fuzzy relation are widely applied in many fields.

Wang (2010) proposed a clustering method based on fuzzy equivalence relation for customer relationship management. In real world, customers commonly take relevant attributes into consideration for the selection of products and services. Further, the attribute assessment of a product or service is often presented by a linguistic data sequence. To partition these linguistic data sequences of customers’ assessment on a product or service, the fuzzy relation clustering method is applied in Wang’s research. In the clustering method they proposed, the linguistic data sequences are presented by fuzzy data sequences, and a fuzzy compatible relation is first constructed to present the binary relation between two data sequences. Then a fuzzy equivalence relation is derived by max–min transitive closure from the fuzzy compatible relation.

Based on the fuzzy equivalence relation, the linguistic data sequences are easily classified into clusters. The clusters representing the selection preferences of different customers on the product or service will be the base for developing customer relationship management (CRM).

Sun et al. (2009) adopted the fuzzy analytic hierarchy process which is a clustering method based on fuzzy relation to determine the weightings for evaluation dimension among decision makers on industrial cluster problems. From their analysis, the factor condition is the most important driving force for advancing the industrial cluster performance. Moreover, the promotion of international linkage policy and broader framework policies rank the first two priorities for cluster policy.

## 1.5 Contributions of the thesis

Chapter 2 of the thesis addresses the problem of clustering analysis on DNA microarrays. Clustering analysis is an efficient way to find potential information in microarray data. A clear cluster structure is important and necessary for the ensuing analysis on DNA functions and relations.

The fuzzy c-means clustering method (FCM) and the empirical mode decomposition method (EMD) are combined to be applied in this part. It is the first time that these two methods are combined in clustering analysis of DNA microarrays. We combine the FCM with EMD for clustering microarray data in order to reduce the effect of the noise. We call this method fuzzy c-means method with empirical mode decomposition (FCM-EMD). We applied this method on yeast and serum microarray data respectively and the silhouette values are used for assessment of the quality of clustering. Using the FCM-EMD method on gene microarray data, we obtained better results than those using FCM only. The results suggest the clustering structures of denoised data are more reasonable and genes have tighter association with their clusters. The cluster structures are much clearer than before by combining EMD with FCM. Denoised gene data without any biological information contains no cluster structure. We find that we can avoid estimating the fuzzy parameter *m* in some extent by analysing denoised microarray data. This makes clustering more efficient. Using the FCM-EMD method to analyse gene microarray data can save time and obtain more reasonable results.

In Chapter 3, we perform the identification of disease-related genes based on DNA microarray data. We applied type-2 fuzzy set theory which is an extension of traditional fuzzy set theory, and established type-2 fuzzy membership function to describe the differences of the gene expression values generated from normal people's genes and patients' genes.

Type-2 fuzzy sets can control the uncertainty information more effectively than conventional type-1 fuzzy sets because the membership functions of type-2 fuzzy sets are three-dimensional. This is the first time in the literature that type-2 fuzzy set theory is applied to identify disease-related genes. We call our method type-2 fuzzy membership test (type-2 FM-test) and applied it to diabetes and lung cancer data. For the ten best-ranked genes of diabetes identified by the type-2 FM-test, 7 of them have been confirmed as diabetes associated genes according to genes description information in Genebank and the published literature. One more gene than the original approaches is identified. Within the 10 best ranked genes identified in lung cancer data, 7 of them are confirmed by the literature as associated with lung cancer. The type-2 FM-d values are significantly different, which makes the identifications more reasonable and convincing than the original FM-test.

Chapter 4 concentrates on identification of protein complexes from protein-protein interaction networks. We propose a novel method which combines the fuzzy clustering method and interaction probability to identify the overlapping and non-overlapping community structures in PPI networks, then to detect protein complexes in these sub-networks.

Our method is based on both the fuzzy relation model and the graph model. Fuzzy theory is suitable to describe the uncertainty information between two objects, such as 'similarity' and 'differences'. On the other hand, the original graph model contains clustering information, thus we don't ignore the original structure of the network, but combine it with the fuzzy relation model. We apply the method on yeast PPI networks and compare the results with those obtained by a standard method, CFinder. For the same data, although the precision of matched protein complexes is lower than CFinder, we detected more protein complexes. We also apply our method on two social networks, Zachary’s karate club network and American college football team network. The results showed that our method works well for detecting sub-networks and gives a reasonable understanding of these communities.

# Chapter 2

# Fuzzy c-means method with empirical mode decomposition for clustering microarray data

## 2.1 Introduction

### 2.1.1 Microarray clustering analysis

Bioinformatics is defined as the application of computers, databases and mathematical methods to analyses of biological data and especially genetic sequences and protein structures. The objectives of bioinformatics are the identification of genes and the prediction of their function. The scope of bioinformatics covers completely functional genomics, and the study of genomic information has especially influenced biology and related fields. In the past decade or so, there has been an increasing interest in unravelling the mysteries of deoxyribonucleic acids (DNA). How to gain more bioinformation from DNA is a challenging problem. The growth in DNA data available to researchers is unparalleled. Genbank, a major public database where DNA data are stored, doubles in size approximately every year. It has become important to improve new theoretical methods to conduct DNA data analysis.

Microarray techniques have revolutionized genomic research by making it possible to monitor the expression of thousands of genes in parallel. Since the work of Eisen and colleagues (1998), clustering methods have become a key step in microarray data analysis because they can identify groups of genes or samples displaying a similar expression profile. Such partitioning has the main scope of facilitating data visualization and interpretation, and can be exploited to gain insight into the transcriptional regulation networks underlying a biological process of interest. It has been reported that, due to the complex nature of biological systems, microarray datasets tend to have very diverse structures, some even do not have well defined clustering structures. As a result, none of the existing clustering algorithms performs significantly better than the others when tested across multiple data sets.

There are many methods for cluster analysis, such as K-means (Macqueen, 1967; Lioyd, 1982; Hamerly and Elkan, 2002), K-nearest neighbours (Cover and Hart, 1967; Terrell and Scott, 1992; Hall et al., 2008), Fuzzy C-means (Bezdek, 1981; Groll and Jakel, 2005), hierarchical clustering (Ward and Joe, 1963; Szekely and Rizzo, 2005), self-organizing maps (Kaski, 1997; Hosseini and Safabakhsh, 2003), simulated annealing (Kirkpatrick et al., 1983; Cerny, 1985; Granville et al., 1994; De Vicente et al., 2003) and graph theoretic approaches (Augustson and Minker, 1970; Kuznetsov and Obiedkov, 2001). These algorithms are also applied to analysis of microarrays. *K-means clustering* is a method which is used to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. It is simple and and has been applied in many fields. Hu and Weng (2009) proposed a method which is combined K-means and mathematical morphology. They applied it on segmentation of microarray image processing. The result of the experiment shows that the method is accurate, automatic and robust. Kim et al. (2009) proposed MULTI-K algorithm based on K-means. They newly devised the entropy-plot to control the separation of singletons or small clusters. Compared with the original approach, MULTI-K is able to capture clusters with complex and high-dimensional structures accurately. The *K-nearest neighbour* algorithm (K-NN) is a method for classifying objects based on closest training examples in feature space. Liu et al. (2004) combined genetic algorithm and KNN to subtypes of renal cell carcinoma using a set of microarray gene profiles. The result shows this combined method can be efficiently used in identifying a panel of discriminator genes. In statistics, *hierarchical clustering* is a method of cluster analysis which seeks to build a hierarchy of clusters. Qin et al. (2003) describe a generalization of the hierarchical clustering algorithm that efficiently incorporates high-order features by using a kernel function to map the data into a high-dimensional feature space. Chipman and Tibshirani (2006) proposed a hybrid clustering method that combines the strengths of bottom-up hierarchical clustering with that of top-down clustering and they illustrate the technique on simulated and real microarray datasets. A *self-organizing map* (SOM) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples. Hautaniemi et al.

(2003) applied SOM to analysis and visualization of gene expression microarray data in human cancer. Their results show SOM is capable of helping finding certain biologically meaningful clusters. Clustering algorithms could be used for finding a set of potential predictor genes for classification purposes. Comparison and visualization of the effects of different drugs is straightforward with SOM. Torkkola et al. (2001) applied SOM to exploratory analysis of yeast DNA microarray data. They found SOM not only enabled quick selection of the gene families identified in previous work, but also facilitated the identification of additional genes with similar expression patterns. Alon and colleagues (1999) applied *simulated annealing* for identification of tumor genes. Maulik et al. (2010) combine simulated annealing with fuzzy clustering method for analysing microarray data. *Graph theoretic approaches* are also widely used in bioinformatics. Sharan and Shamir (1999) proposed a graph-theoretic method based on computing minimum cut and applied it on analysis of gene expression data. This method is an unsupervised approach which does not make any prior assumptions on the number or the structure of the clusters. Potamias (2004) presents a novel graph-theoretic clustering (GTC) method which relies on a weighted graph arrangement of genes, and the iterative partitioning of the respective minimum spanning tree of the graph. GTC utilizes information about the functional classification of genes to knowledgeably guide the clustering process and achieve more informative clustering results.

DNA microarray data contain uncertainty and imprecise information (Glonek and Solomon, 2004; Brown et al., 2001). Hard clustering methods such as K-means, KNN and self-organizing maps, which assign each gene to a single cluster, sometimes are poorly suited to the analysis of microarray data because the clusters of genes frequently overlap in such data (Dembele and Kanstner, 2003; Sharan and Shamir, 2003). Fuzzy theory has many advantages in dealing with data containing uncertainty, therefore, fuzzy clustering approaches have been taken into consideration to analyse DNA microarrays (Chen et al., 2006; He et al., 2006; Wang et al., 2008; Avogadri and Valentini, 2009). The most widely applied fuzzy clustering method is the fuzzy C-means (FCM) algorithm. Dembele and Kastner (2003) applied FCM to analysis of microarray data and proposed a newly method for estimation of the fuzzy parameter *m*. Wang et al. (2003) applied FCM to tumor classification and marker gene prediction. Kim et al. (2006) discussed the effect of data normalization on FCM clustering of DNA microarray data. Fu and Medico (2007) devised a cluster analysis software (GEDAS) based on FCM and the SOM algorithm.

However, when implementing fuzzy algorithms, it is important to choose appropriate values for parameters such as the fuzziness exponent *m*. Especially, in fuzzy models the minimization criterion for the objective function depends on *m*. In the fuzzy clustering literature, a value of $m = 2$ is commonly used, but this values is not appropriate for gene expression data (Dembel and Kanstner, 2003; Kim et al, 2006). How to estimate the value of fuzziness parameter *m* is a problem in applying the FCM method to DNA microarray data clustering. The optimal values for *m* vary a lot from one dataset to another. Although some researchers have already given some methods for choosing the values of *m*, these methods usually are time-consuming (Dembel and Kanstner, 2003; Yang et al., 2007). In Dembel and Kanstner's work, DNA microarray data contain noise which would affect clustering results (Li and Johnson, 2002; Ma, 2006; Someren et al., 2006; Wang et al., 2006). Research into normalizing and removing noise from datasets has been an important component of previous works on clustering analysis (Kim et al, 2006; Bertoni and Valentini, 2006).

In this chapter, we propose to combine FCM method with empirical mode decomposition (EMD) for clustering microarray data. The EMD method was first proposed by Huang et al. (1998) and then Lin et al. (2009) proposed an alternative EMD. Usually, EMD is used to analyse the intrinsic components of a signal. These components are called intrinsic mode functions (IMFs). Most noisy IMFs are considered as noise in the signal. If we remove most noisy IMFs from the raw data, the trend component can be obtained. Then we use the trend as denoised data to perform clustering analysis. Shi et al. (2007) used EMD to remove noise in protein sequences and studied the functional similarity of these sequences. RecentlyYu et al., (2010) used the EMD method in Lin et al. (2009) to get the trend and simulate geomagnetic field data. Here we propose to remove noise in DNA microarray data by the EMD method in Lin et al. (2009). Comparing with the results obtained by Dembele and Kastner (2003), we can get better clustering structure by using denoised data and choosing $m = 2$ which avoids the estimation of the value of the fuzziness parameter in some extent. We can also get better clustering structure results using denoised data and the estimated value of $m$ according to silhouette measure which has been used to assess the quality of clusters.

### 2.1.2 Fuzzy theory

Most of our traditional tools for modelling, reasoning and computing are crisp, deterministic, and precise in character. By crisp we mean dichotomous, that is, yes-or-no-type rather than more-or-less type. In conventional dual logic, for instance, a statement can be true or false and nothing in between. In classical set theory, an element can either belong to a set or not, and in optimization, a solution is either feasible or not. Precision assumes that the parameters of a model represent exactly either our perception of the phenomenon modelled or the features of the real system that has been modelled. Generally precision also implies that the model is unequivocal, that is, that it contains no ambiguities. This is 0 and 1 logic (Klir and Yuan, 1995; Zimmermann, 2001; Chen et al., 2001).

However, more often than not, the problems encountered in the real physical world are not always yes-or-no type or true-or-false type. Real situations are very often uncertain or vague in a number of ways. Due to lack of information the future state of the model might not be known completely. This type of uncertainty (stochastic character) has long been handled appropriately by probability theory and statistics. This Kolmogorov type probability is essentially frequentist and based on set-theoretic considerations. Koopman’s probability refers to the truth of statements and therefore based on logic. On both types of probabilistic approaches it is assumed, however, that the events or the statements, respectively, are well defined. We shall call this type of uncertainty or vagueness stochastic uncertainty by contrast to the vagueness concerning the description of the semantic meaning of the events, phenomena or statements themselves, which we shall call fuzziness. Fuzziness can be found in many areas of daily life, such as in engineering, medicine, meteorology, manufacturing. It is particularly frequent, however, in all areas in which human judgment, evaluation, and decisions are important. These are the areas of decision making, reasoning, learning, and so on. Some reasons for this have already been mentioned. Others are that most of our daily communication uses “natural languages” and a good part of our thinking is done in it. For instance, instead of describing the weather tody in terms of the exact percentage of cloud cover, we can just say that it is sunny. In order for a term such as sunny to accomplish the desired introduction of vagueness, however, we cannot use it to mean precisely 0% cloud cover. Its meaning is not totally arbitrary, however; a cloud cover of 100% is not sunny, and either, in fact, is a cloud cover of 80%. We can accept certain intermediate states, such as 10% or 20% of cloud cover, as sunny. But where do we draw this line? If, for instance, any cloud cover of 25% or less is considered sunny, does this mean that a cloud cover of 26% is not? This is clearly unacceptable, since 1% of cloud cover hardly seems like a distinguishing characteristic between sunny and not sunny. We could, therefore, add a qualification that any amount of cloud cover 1% greater than a cloud cover already considered to be sunny ( that is, 25% or less) will also be labelled as sunny. We can see, however, that this definition eventually leads us to accept all degrees of cloud cover as sunny, no matter how gloomy the weather looks! In order to resolve this paradox, the term sunny may introduce vagueness by allowing some gradual transition from degrees of cloud cover that are considered to be sunny and those that or not (Klir and Yuan, 1995).

Fuzziness has so far not been defined uniquely semantically, and probably never will. It will mean different things, depending on the application area and the way it is measured. However to solve the problems encountered in the real world, fuzzy theory was proposed and developed. Fuzzy theory was proposed by L. A. Zadeh in 1965. From the inception of the theory, a fuzzy set has been defined as a collection of objects with membership values between 0 (complete exclusion) and 1 (complete membership). The membership values express the degrees to which each object is compatible with the properties or features distinctive to the collection. A fuzzy set can be defined mathematically by assigning to each possible individual in the universe of discourse a value representing its grade of membership in the fuzzy set.

This grade corresponds to the degree to which that individual is similar or compatible with the concept represented by the fuzzy set. Thus, individuals may belong in the fuzzy set to a greater or lesser degree as indicated by a larger or smaller membership grade. As already mentioned, these membership grades are very often represented by real-number values ranging in the closed interval between 0 and 1. Thus, a fuzzy set representing our concept of sunny might assign a degree of membership of 1 to a cloud cover of 0%, 0.8 to a cloud cover of 20%, 0.4 to a cloud cover of 30%, and 0 to a cover of 75%. These grades signify the degree to which each percentage of cloud cover approximates our subjective concept of sunny, and the set itself models the semantic flexibility inherent in such a common linguistic term. Because full membership and full non-membership in the fuzzy set can still be indicated by the values of 1 and 0, respectively, we can consider the concept of a crisp set to be a restricted case of the more general concept of a fuzzy set for which only these two grades of membership are allowed. Research on the theory of fuzzy sets has been growing steadily since the inception of the theory in the mid-1960s. The body of concepts and results pertaining to the theory is now quite impressive. Research on a broad variety of applications has also been very active and has produced results that are perhaps even more impressive (Klir and Yuan, 1995).

In the next section, we introduce the theoretical background needed for a description of the fuzzy c-means method with empirical mode decomposition (FCM-EMD) detailed in Section 2.3. We will apply this method on yeast and serum microarray data respectively in Section 2.4, and the silhouette values are used for assessment of quality of clusters.

## 2.2 Theoretical background

### 2.2.1 Fuzzy sets

Let $X$ be a space of points (objects), called the universe, and $x$ an element of $X$. Membership in a classical subset $A$ of $X$ is often viewed as a characteristic function $\mu_A$ from X to {0, 1} such that a fuzzy set is characterized by a membership function mapping the elements of a universe of discourse $X$ to the unit interval [0, 1]. That is,

$\mu_A(x) \in [0, 1]$. $\mu_A(x)$ is the grade of membership of $A$. Clearly, A is a subset of $X$ that has no sharp boundary.

$A$ is completely characterized by the set of pairs of the elements in $X$ and their membership values,

$$A = \{(x,\ \mu_A(x)),\ x \in X\}. \tag{2.1}$$

Sometimes a sum notation is used. This allows us to enumerate only elements of $X$ with nonzero grades of membership in the fuzzy set. For instance, if $X = \{ x_1,\ x_2, \ldots, x_n \}$, then the fuzzy set $A = \{(a_i / x_i \,|\, x_i \in \mathrm{X}\,)\}$, where $a_i = \mu_A\,(x_i)$, $i = 1, \ldots, n$, may be denoted by

$$A = a_1 / x_1 + a_2 / x_2 + a_3 / x_3 + \ldots + a_n / x_n = \sum_{i=1}^{n} a_i / x_i\ . \tag{2.2}$$

In this notation the sum should not be confused with the standard algebraic summation; the only purpose of the summation symbol in the above expression is to denote the set of the ordered pairs. Also, note that when $A = \{a / x\}$, that is, when there exists only one point $x$ in a universe for which the membership degree is non null, we have a fuzzy singleton. In this sense, we may also interpret the summation symbol as union of singletons. Equivalently, one can summarize $A$ as a vector, meaning that $A = [a_1, a_2, \ldots, a_n]$. When the universe $X$ is continuous, we use, to represent a fuzzy set, the following expression:

$$A = \int_x a / x\,, \tag{2.3}$$

where $a = \mu_A(x)$ and the integral symbol should be interpreted in the same way as the sum given above.

Two fuzzy sets $A$ and $B$ are said to be equal, denoted $A = B$ if and only if (iff)

$$\forall\, x \in X,\ \mu_A(x) = \mu_B(x)\,. \tag{2.4}$$

### 2. 2. 2 Membership function

The value of $\mu_A(x)$ describes a degree of membership of $x$ in $A$ and we define $\mu_A(x)$ as membership function. For instance, consider the concept of high temperature in, say, an environmental context with temperatures distributed in the interval [0, 50] defined in $^{\circ}C$. Clearly $0\,^{\circ}C$ is not understood as a high temperature value, and we may assign a null value to express its degree of compatibility with the high temperature concept. In other words, the membership degree of $0\,^{\circ}C$ in the class of high temperatures is zero. Like wise, $30\,^{\circ}C$ and over are certainly high temperatures, and we may assign a value of 1 to express a full degree of compatibility with the concept. Therefore, temperature values in the range [30, 50] have a membership value of 1 in the class of high temperatures. The partial quantification of belongingness for the remaining temperature values through their membership values can be pursued as exemplified in Figure 2.1, which actually is a membership function $H$: $\mathrm{T} \rightarrow [0,1]$ characterizing the fuzzy set $H$ of high temperatures in the universe T = [0, 50].

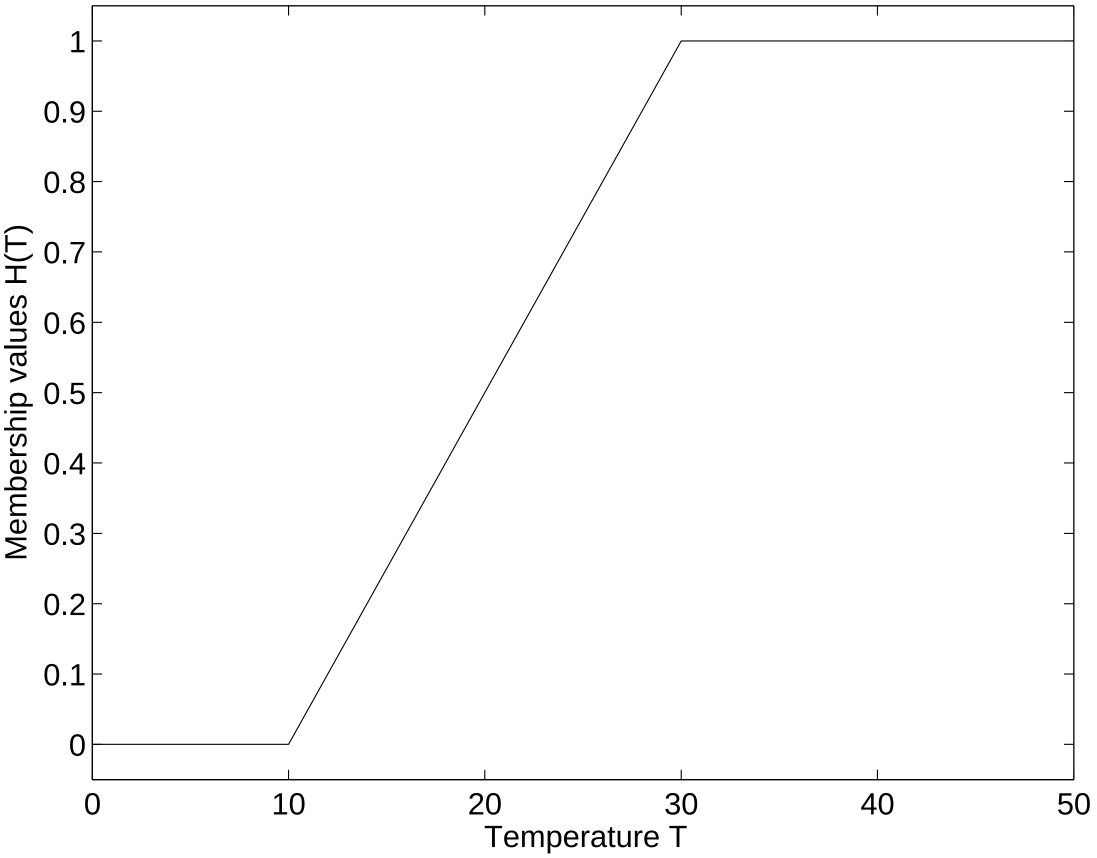


Figure 2.1 Membership function of environment temperature.

In principle any function of the form $A$: $\mathrm{X} \rightarrow [0, 1]$ describes a membership function associated with a fuzzy set A that depends not only on the concept to be represented, but also on the context in which it is used. The graphs of the functions may have very different shapes, and may have some specific properties. Whether a particular shape is suitable can be determined only in the application context. In certain cases, however, the meaning semantics captured by fuzzy sets is not too sensitive to variations in the shape, and simple functions are convenient (Klir and Yuan, 1995; Dubois and Prade, 1980).

Triangular-shaped function, trapezoidal-shaped function, Gaussian-shaped function and S-shaped function are the simplest membership functions. Their equations and plots are as follows:

1. Triangular-shaped membership function:

$$f(x,a,b,c,d) = \begin{cases} 0, & \text{if } x \le a \\ \dfrac{x-a}{b-a}, & \text{if } a \le x \le b \\ \dfrac{c-x}{c-d}, & \text{if } b \le x \le c \\ 0, & \text{if } x \ge c \end{cases}, \tag{2.5}$$

2. Trapezoidal-shaped membership function:

$$f(x,a,b,c,d) = \begin{cases} 0, & \text{if } x \le a \\ \dfrac{x-a}{b-a}, & \text{if } a \le x \le b \\ 1, & \text{if } b \le x \le c \\ \dfrac{d-x}{d-s}, & \text{if } c \le x \le d \\ 0, & \text{if } x \ge d \end{cases}, \tag{2.6}$$

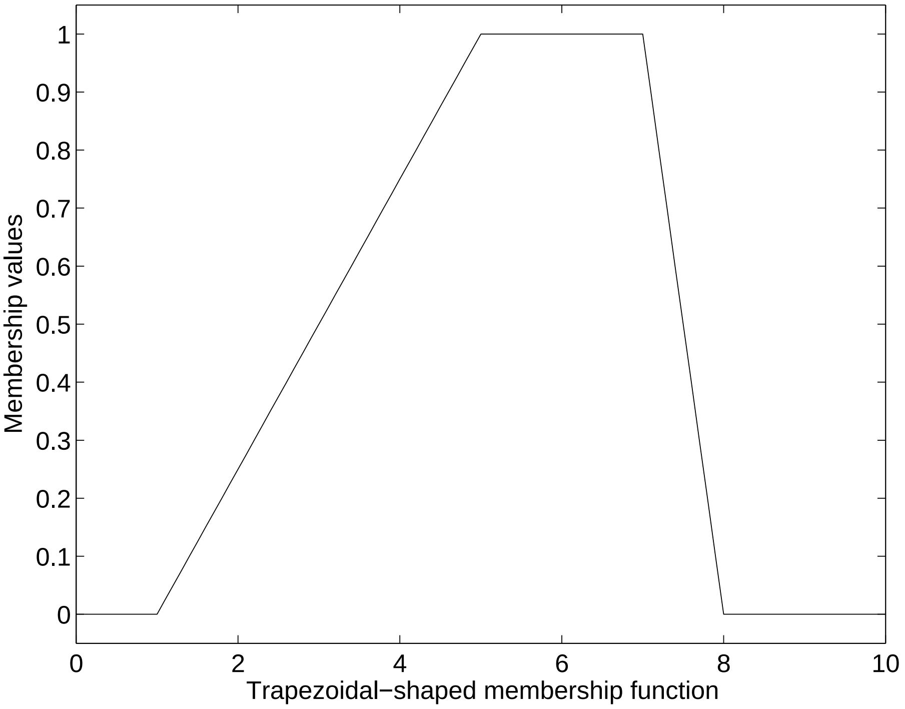
3. Gaussian-shaped membership function:

$$f(x,\sigma,c) = \exp[-\frac{(x-c)^2}{2\sigma^2}], \tag{2.7}$$

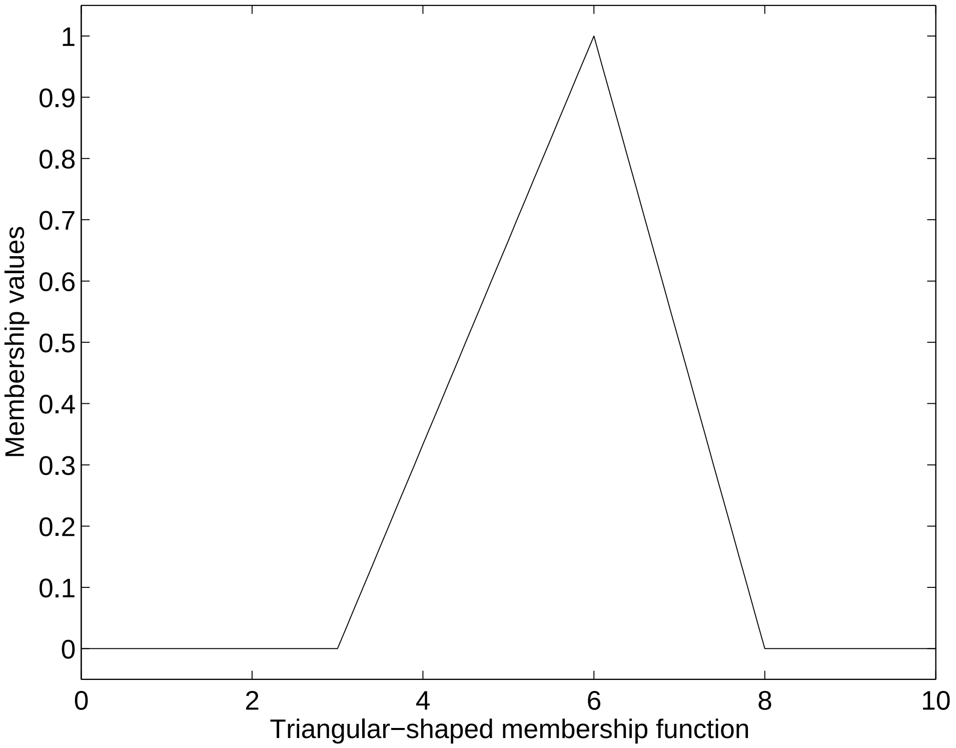
4. S-shaped function

$$S(x;a,b,c) = \begin{cases} 0, & \text{if} \quad x \le a \\ 2\bullet(\frac{x-a}{c-a})^2, & \text{if} \quad a < x \le b \\ 1-2\bullet(\frac{x-a}{c-a})^2, & \text{if} \quad b < x \le c \\ 1, & \text{if} \quad x > c \end{cases}, \tag{2.8}$$

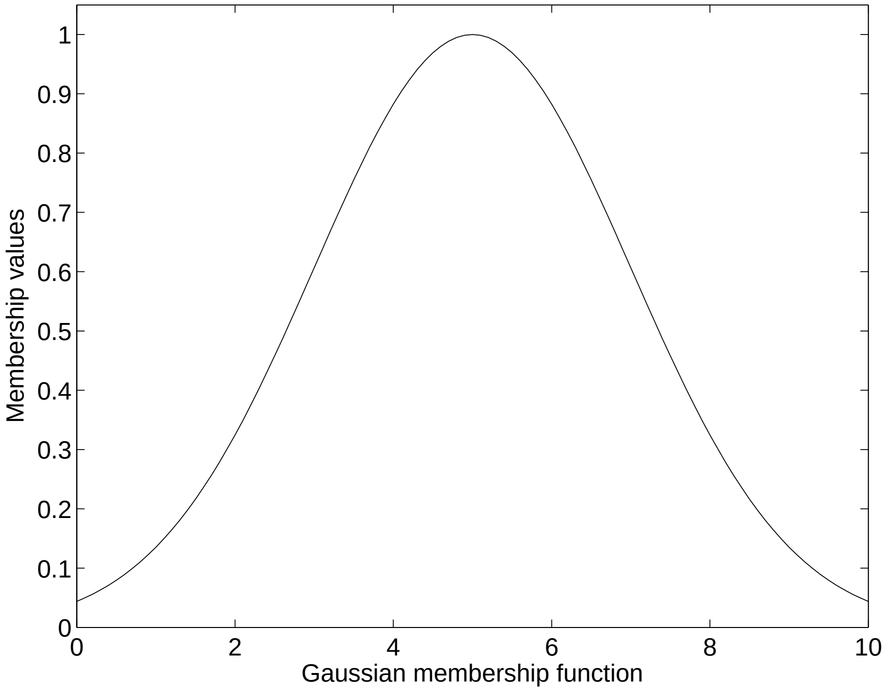
where $b = \frac{a+c}{2}$.



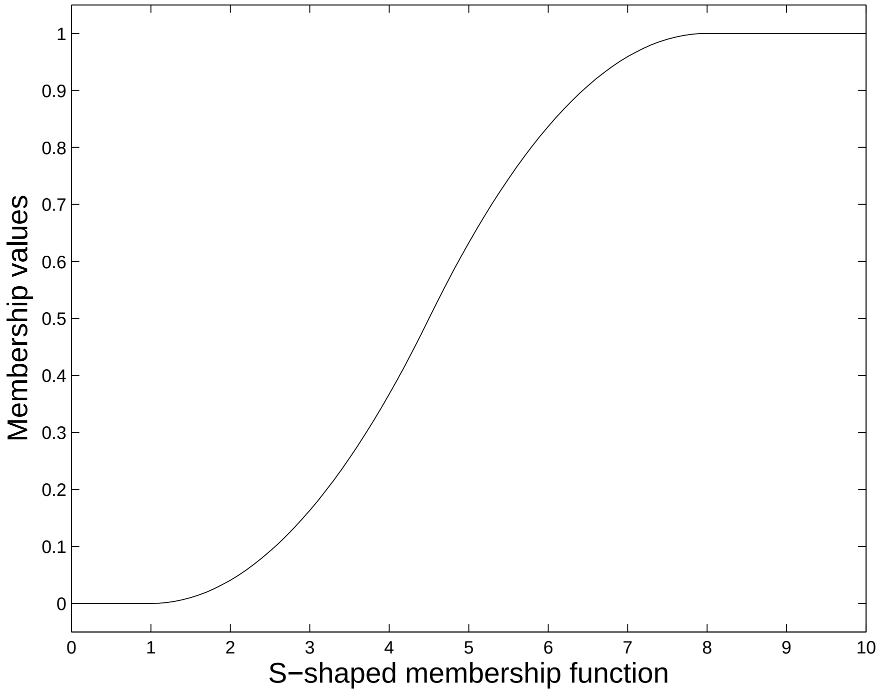
(a)


Membership values
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
2
4
6
8
10
Triangular−shaped membership function

(b)


Membership values
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
2
4
6
8
10
Gaussian membership function

(c)



(d)

Figure 2.2 Four simplest membership functions. (a) Trapezoidal shape, (b) Triangular shape, (c) Gaussian shape, (d) S shape.

As mentioned above, even for similar contexts, fuzzy sets representing the same concept may vary considerably. In this case, however, they also have to be similar in some key features, irrespective of choice of membership function. It is convenient to use a simple shape to describe the “temperature changing” by a trapezoidal-shaped membership function.

## 2. 2. 3 Fuzzy set operations

As a classical set, fuzzy set also has its operations: fuzzy complement, intersection and union. The classical union and intersection of ordinary subsets of X can be extended by the following formulas, proposed by Zadeh (1965)

*Fuzzy complement* $\mu_{\bar{A}} = 1 - \mu_A(x)$;

*Fuzzy intersection* $\mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)]$;

*Fuzzy union* $\mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)]$;

for all $x \in X$ . These operations are called the standard fuzzy operations.

However, we can easily see that the standard fuzzy operations perform precisely as the corresponding operations for crisp sets when the range of membership grades is restricted to the set {0, 1}. That is, the standard fuzzy operations are generalizations of the corresponding classical set operations. It is now well understood, however, that they are not the only possible generalizations. For each of the three operations, there exists a broad class of functions whose members qualify as fuzzy generalizations of the classical operations as well. Functions that qualify as fuzzy intersections and fuzzy unions are usually referred to in the literature as *t-norms* and *t-conorms*, respectively.

Since the fuzzy complement, intersection and union are not unique operations, contrary to their crisp counterparts, different functions may be appropriate to represent these operations in different contexts. That is, not only membership functions of fuzzy sets but also operations functions on fuzzy sets are context-dependent. The capability to determine appropriate membership functions and meaningful fuzzy operations in the context of each particular application is crucial for making fuzzy set theory practically useful.

Among a variety of fuzzy complements, intersections, and unions, the standard fuzzy operations possess certain properties that give them special significance. The standard fuzzy intersection (min operator) produces for any given fuzzy sets the largest fuzzy set from among those produced by all possible fuzzy intersections (*t-norms*). The standard fuzzy union (max operator) produces, on the contrary, the smallest fuzzy set among the fuzzy sets produced by all possible fuzzy unions (*t-conorms*). That is the standard fuzzy operations occupy specific positions in the whole spectrum of fuzzy operations: the standard fuzzy intersection is the weakest fuzzy intersection, while the standard fuzzy union is the strongest fuzzy union.

A desirable feature of the standard fuzzy operations is their inherent prevention for the compounding of errors of the operands. If any error *e* is associated with the membership grades $\mu_A(x)$ and $\mu_B(x)$, then the maximum error associated with the membership grade of $x$ in $\overline{A}$, $A \cap B$ and $A \cup B$ remains $e$. Most of the alternative fuzzy set operations lack this characteristic (Klir and Yuan, 1995; Dubois and Prade, 1980; Zadeh, 1965)..

### 2.2.4 The differences between fuzziness and probability

Fuzziness is often mistaken for probability. Therefore it is necessary to distinguish these concepts. In science the two following types of uncertainty are distinguished (there are also other kinds):

1. Stochastic uncertainty;
2. Lexical uncertainty.

Stochastic uncertainty means uncertainty of occurrence of an event, which is itself precisely defined; lexical uncertainty means uncertainty of the definition of this event. Uncertainty of the definition means its fuzziness. Fuzzy system theory is engaged in methods of creating models employing fuzzy concepts, which are used by people. It should be mentioned that people also employ, apart from lexical fuzzy concepts, intuitive concepts and pictures not connected at all with any vocabulary. There are people who know no language; there are also animals, which create intuitive, non-lexical information about reality enabling them to function and survive in it. The theory of intuitive modelling may probably be the continuation of the theory of fuzzy modelling in the future.

To understand the distinction between fuzziness and randomness, it is helpful to interpret the grade of membership in a fuzzy set as a degree of compatibility (or possibility) rather than probability. As an illustration, consider the proposition “They got out of Roberta’s car” (which is a Pinto). The question is : How many passengers got out of Roberta’s car? (Zadeh 1978) -assuming for simplicity that the individuals involved have the same dimensions.

Let *n* be the number in question. Then, with each *n* we can associate two numbers $\mu_n$ and $p_n$ representing, respectively, the possibility and the probability that *n* passengers got out of the car. For example, we may have for $\mu_n$ and $p_n$:

Table 2.1 The values of $\mu_n$ and $p_n$

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\mu_n$ | 0 | 1 | 1 | 1 | 0.7 | 0.2 | 0 |
| $p_n$ | 0 | 0.6 | 0.3 | 0.1 | 0 | 0 | 0 |

in which $\mu_n$ is interpreted as the degree of ease with which *n* passengers can squeeze into a Pinto. Thus $\mu_5 = 0.7$ means that, by some specified or unspecified criterion, the degree of ease of squeezing 5 passengers into a Pinto is 0.7. On the other hand, the possibility that a Pinto may carry 4 Passengers is 1; by contrast the corresponding probability in the case of Roberta might be 0.1.

This simple example brings out three important points. First, that possibility is not an all or nothing property and may be present to a degree. Second, the degrees of possibility are not the same as probabilities. And third, that possibilistic information is more elementary and less context-dependent than probabilistic information. But, what is most important as a motivation for the theory of fuzzy sets is that much, perhaps most, of human reasoning is based on information that is possibilistic rather than probabilistic in nature.

## 2.3 Methods

### 2.3.1 Fuzzy c-means algorithm

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. The fuzzy clustering algorithm links each gene to all clusters via a real-valued vector of indexes. The values $\mu_{ki}$ of the components of this vector lie between 0 and 1. For a given gene, an index close to 1 indicates a strong association to the cluster. Conversely, indexes close to 0 indicate the absence of a strong association to the corresponding cluster. The vector of indexes defines thus the membership of a gene with respect to the various clusters. Membership vector values $\mu_{ki}$ and cluster centroids $\mathrm{c}_k$ can be obtained after minimization of the total inertia criterion (Bezdek, 1981):

$$J(K,m)=\sum_{k=1}^{K}\sum_{i=1}^{N}(\mu_{ki})^m d^2(\mathrm{x}_i,\mathrm{c}_k)\,, \tag{2.9}$$

$$d^2(\mathrm{x}_i,\mathrm{c}_k)=(\mathrm{x}_i\text{-}\mathrm{c}_k)^T A_k(\mathrm{x}_i\text{-}\mathrm{c}_k)\,, \tag{2.10}$$

$$\text{with }\sum_{k=1}^{K}\mu_{ki}=1;\ 0<\sum_{i=1}^{N}\mu_{ki}<1\ , \tag{2.11}$$

where $1 \le i \le N$, $1 \le k \le K$.

In equation (2.9), $K$ and $N$ are respectively the number of clusters and the number of samples (or genes) in the data, $m$ is a real-valued number which controls the 'fuzziness' of the resulting clusters, $\mu_{ki}$ is the degree of membership of gene $\mathrm{x}_i$ in cluster $k$, and $d^2(\mathrm{x}_i,\mathrm{c}_k)$ is the square of the distance from gene $\mathrm{x}_i$ to centroid $\mathrm{c}_k$. In equation (2.10), $A_k$ is a symmetric and positive definite matrix.

Equation (2.11) indicates that empty clusters are not allowed. The scalar $m$ is any real-valued number greater that 1. When $A_k$ is the identity matrix, then $d^2(\mathrm{x}_i,\mathrm{c}_k)$ corresponds to the square of the Euclidian distance. From equation (2.9), parameters of interest are the cluster centroid vectors $\mathrm{c}_k$ and the components of the membership vectors $\mu_{ki}$ . These unknown parameters can be obtained using the following algorithm (Bezdek, 1981):

(i) Initialization: Fix $K$, $m$ and choose any product norm metric for calculation of $d^2(\mathrm{x}_i,\mathrm{c}_k)$. Select randomly $K$ samples as initial centroids $\mathrm{c}_k^{(0)}$ and then form partitions of all others samples around these centroids to obtain the initial partition matrix $U^{(0)} = [\mu_{ki}]$, k = 1, …, *K* and *i* = 1, …, *N*. At step *l*, *l*=1, 2, …, perform the following steps:

(ii) Computation of centroids $\mathrm{c}_k{}^{(l)}$:

$$\mathrm{c}_k{}^{(l)} = \frac{\sum_{i=1}^{N} (\mu_{ki}^{(l)})^m \mathrm{x}_i}{\sum_{i=1}^{N} (\mu_{ki}^{(l-1)})^m}; \qquad k = 1,2,\ \ldots,\ K, \tag{2.12}$$

(iii) Computation of membership values $\mu_{ki}^{(l)}$:

$$I_i = \{k / 1 \le k \le K; d^2(\mathrm{x}_i, \mathrm{c}_k{}^{(l)}) = 0\},$$

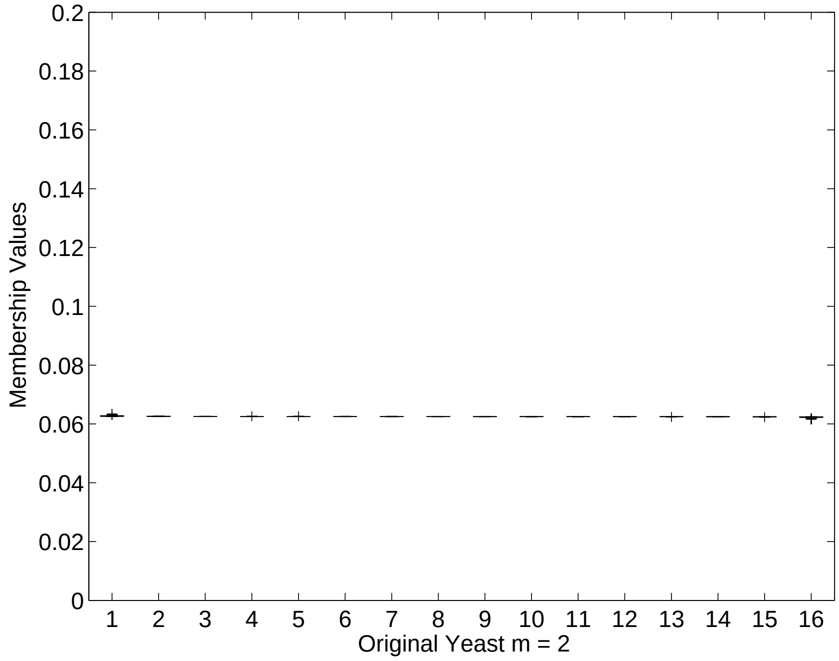
$$\tilde{I}_i = \{1, 2, ..., K\} - I_k,$$

$$\mu_{ki}^{(l)} = \begin{cases} \dfrac{1}{\sum_{s=1}^{K} \left[ \dfrac{d^2(\mathrm{x}_i, \mathrm{c}_k{}^{(l)})}{d^2(\mathrm{x}_i, \mathrm{c}_s{}^{(l)})} \right]^{\frac{1}{(m-1)}}}, & \text{if } I_i = \varnothing \\ 0, & \text{if } I_i \ne \varnothing, \forall i \in \tilde{I}_i \\ \dfrac{1}{|I_i|}, & \text{if } I_i \ne \varnothing, \forall i \in I_i \end{cases}, \tag{2.13}$$

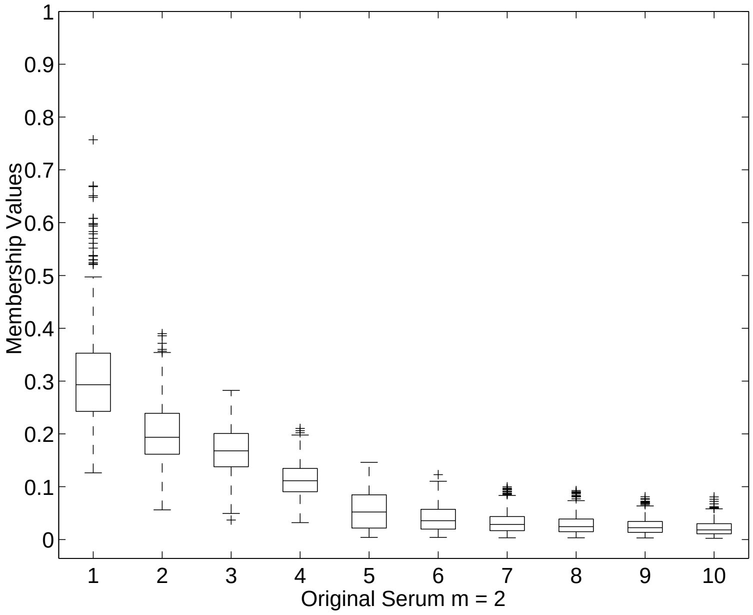
(iv) Repetition of (2.12) and (2.13) until stabilization, i.e. $\left\| U^{(l)} - U^{(l-1)} \right\| \le \varepsilon$, *l* > 1.

After several passes through (2.12) and (2.13), the algorithm will stop, i.e. the error between two consecutive values of the constrained fuzzy partition matrix *U* will be smaller than a priori specified level. Convergence of FCM has been proven by Bezdek (1981).

In most works about FCM, to avoid complicated computation of the membership $\mu_{ki}$, *m* is commonly fixed to 2. However, the value 2 is not appropriate for every data set. For example, in Fig. 2.3 when we used *m* = 2 for the yeast microarray data, we observed that all the membership values were similar. That means FCM failed to extract any clustering structure. On the other hand, for the serum data set, although a clustering structure was found, all memberships have low values. This means that this FCM setting failed to tightly associate any gene to any cluster.



(a)



(b)

Figure. 2.3 The affect of fuzzy parameter *m* on Yeast and Serum data sets. (a) is yeast data, (b) is serum data. The horizontal axis is number of clusters, the vertical axis is membership values.

Based on observing computations on different data sets, Dembele and Kastner (2003) proposed a hypothesis that when *m* varies, there might be a relationship between the FCM membership values and the coefficient of variation of the set of distances between genes. They proposed a method for estimation the fuzziness parameter *m*. The details of this hypothesis and method are as follows:

It was shown that when *m* goes to infinity, the values of $\mu_{ki}$ go to $\frac{1}{K}$. Thus, for a given data set, there is an upper bound value for *m* ($m_{ub}$), above which the membership values resulting form FCM are equal to $\frac{1}{K}$. As a first step towards the evaluation of an appropriate value for *m*, Dembele and Kastner (2003) first attempted to estimate $m_{ub}$. From (10), they note that membership values $\mu_{ki}$ depend on the distances between genes and cluster centroids. For complex data sets, it is reasonable to make the approximation that the cluster centroids will be close to some genes. Thus they made the hypothesis that when *m* varies, there might be a relationship between the FCM membership values and the coefficient of variation (cv) of the set of distances between genes:

$$Y_m = \{[d^2(\mathrm{x}_i, \mathrm{x}_k)]^{\frac{1}{m-1}}; k \neq i = 1, 2, ..., N\}. \tag{2.14}$$

Note that $Y_m$ depend only on the initial data set and *m*, and are thus completely independent of the FCM results. To test the above hypothesis, they used the iris dataset and two generated data sets. For each data set, they varied *m* and determined the *cv* and $Y_m$. They also ran the FCM algorithm to determine the distribution of the membership values. In each case, they observed that the values of *m* which lead to membership values close to $\frac{1}{K}$ gave a cv of $Y_m$ close to 0.03*p*, *p* being the data dimension. However, they offered no theoretical justification for this observation. They proposed to use it to solve the following equation to evaluate $m_{ub}$:

$$cv\{Y_m\} = \frac{\sigma_{Y_m}}{\overline{Y}_m} \approx 0.03p\,, \tag{2.15}$$

where $\sigma_{Y_m}$ and $\overline{Y}_m$ are respectively the standard deviation and the mean of the set $Y_m$.

Dembele and Kastner (2003) solved equation (2.15) numerically by using the dichotomy search strategy. Initially they set $m$ = 2 and computed $cv\{Y_2\}$. This value allowed them to decide the direction of search: in [1, 2] if $cv\{Y_2\} < 0.03p$, in [2, $\infty$] if $cv\{Y_2\} > 0.03p$. If $m_{ub}$ was not equal to 2, they performed successive choices of $m$ in the correct direction and computed $cv\{Y_m\}$ until $cv\{Y_m\} \approx 0.03p$.

The closer $m$ gets to 1, the less fuzzy membership values become. (Bezedk, 1981). Dembele and Kastner (2003) proposed to choose $m$ lower or equal to 2, to get high membership values for genes strongly related to clusters. More precisely, they chose $m$ = 1 + $m_0$ where $m_0$ = 1 if $m_{ub} \geq 10$ and $m_0$ = $\frac{m_{ub}}{10}$ if $m_{ub} \leq 10$. This choice leads to $m$ = 2 when $m_{ub}$ >10 and $m$ < 2 when $m_{ub}$ <10. In this section we also apply this method to estimate the value of the fuzzy parameter $m$.

### 2.3.2 Empirical mode decomposition

The method called empirical mode decomposition is originally designed for non-linear and non-stationary data analysis by Huang et al. (1998), and has been applied to signal processing in various fields since 1998. Lin et al. (2009) briefly described the traditional empirical mode decomposition (EMD) and presented a new approach to EMD. We outline some content of Lin et al. (2009) here. The traditional EMD decomposes a time series into components called intrinsic mode functions to define meaningful frequencies of a signal. An intrinsic mode function (IMF) is defined with two conditions (Huang et al., 1998; A. Janusauskas et al., 2005).

(i) In the whole data set, the number of extrema and the number of zero crossings must be either equal or differ at most by one;

(ii) The mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

By the definition of IMF, the decomposition called shifting process can be followed by using envelopes.

The original EMD is obtained through an algorithm called *shifting process*. Let $X(t)$ be a function representing a signal and $\{t_j\}$ be the local maxima for $X(t)$. We use $X$ denote the values in this signal. The cubic spline $E_U(t)$ connecting the points $\{( t_j, X(t))\}$ is referred as the *upper envelope* of $X(t)$. Similarly, with the local minima $\{s_j\}$ of $X$ we also have the lower envelope $E_L(t)$ of $X(t)$. Then we define the operator $S$ by

$$S(X(t)) = X(t)\text{-}\frac{1}{2}\,(\,E_U(t) + E_L(t))\frac{1}{K}\ , \tag{2.16}$$

In the shifting algorithm, the finest IMF in the EMD is given by

$$I_1(t) = \lim_{n\to\infty} S^n(X(t))\,, \tag{2.17}$$

Subsequent IMFs in the EMD are obtained recursively via

$$I_k(t) = \lim_{n\to\infty} S^n(X(t) - I_1(t) - I_2(t) - \ldots - I_{k-1}(t))\,, \tag{2.18}$$

The process stops when $Y = X - I_1 - I_2 - \ldots - I_k$ has at most one local maximum or local minimum. This function $Y(t)$ denotes the trend of $X(t)$.

Lin et al. (2009) proposed a new algorithm for EMD. Instead of using the envelopes generated by spline, in the new algorithm we use a low pass filter to generate a “moving average” to replace the mean of the envelopes. The essence of the shifting algorithm remains. Let L be an operator that is a low pass filter, for which L($X$)($t$) represent the “moving average” of $X$. Now define

$$T(X) = X - L(X). \qquad (2.19)$$

In this approach, the low pass filter L is dependent on the data $X$. For a given $X(t)$, we choose a low pass filter $L_1$ accordingly and set $T_1 = I - L_1$, where I means the identical operator. The first IMF in the new EMD is given by $\lim_{n\to\infty} T_1^n(X)$, and subsequently the $k$-th IMF $I_k$ is obtained first by selecting a low pass filter $L_k$ according to the data $X - I_1 - I_2 - \ldots - I_k$ and iterations $I_k = \lim_{n\to\infty} T_k^n(X - I_1 - I_2 - \ldots - I_{k-1})$, where $T_k = I - L_k$ . Again the process stops when $Y = X - I_1 - I_2 - \ldots - I_k$ has at most one local maximum or local minimum. Lin et al. (2009) suggested to use the filter $Y$ = L($X$) given by $Y(n) = \sum_{j=-m}^{m} a_j X(n+j)$. We select the mask $a_j = \dfrac{m-|j|+1}{m+1}$, j = -$m, \ldots, m$ in this thesis.

Let $r(t) = X(t) - I_1(t) - \ldots - I_{k-1}(t)$. The original signal can be expressed as

$$X(t) = \sum_{i=1}^{K_1} I_i(t) + r(t), \qquad (2.20)$$

where the number $K_1$ can be chosen according to a standard deviation (SD). In our work the number of components in IMFs is set as 4. The empirical mode decomposition can be considered as an extraction of the different frequency components of the original series.

### 2.3.3 The CLICK algorithm

Before we ran FCM, we have to determine the number of clusters K. In this thesis we use the Cluster Identification via Connectivity Kernels (CLICK) to estimate the number of clusters. The CLICK algorithm was proposed by Sharan and Shamir (2000). It combines graph-theoretic and statistical techniques for automatic identification of clusters in a data set. We firstly turn the microarray marix into a weighted graph, and then perform cluster analysis of this graph. After this work is done, we can obtain the number of clusters K. The method to generate a weighted graph is as follows:

Let $S$ be a pairwise similarity matrix for gene microarray data matrix $X$, where $S_{ij}$ is the inner product of the vectors of genes $i$ and $j$, i.e.,

$$S_{ij} = \sum_{k=1}^{p} x_{ik} x_{jk} \ , \tag{2.21}$$

then we can transform the microarray matrix into weighted similarity graph $G = (V, E)$. In this graph, vertices correspond to elements and edge weights are derived from the similarity values. The weight $w_{ij}$ of an edge $(i, j)$ reflects the probability that $i$ and $j$ are mates, and is set to be

$$w_{ij} = \log \frac{p_{i,j\in\Omega}\sigma_F}{(1-p_{i,j\in\Omega})\sigma_T} + \frac{(S_{ij}-\mu_F)^2}{2\sigma_F^2} - \frac{(S_{ij}-\mu_T)^2}{2\sigma_T^2}, \tag{2.22}$$

here $f(S_{ij}\,|\,i, j\in\Omega) = f(S_{ij}\,|\,\mu_T, \sigma_T)$ is the value of mate probability density function at $S_{ij}$, $\Omega$ is the set of element who are neighbours:

$$f(S_{ij}\,|\,i, j\in\Omega\ ) = \frac{1}{\sqrt{2\pi}\sigma_T} e^{-\frac{(S_{ij}-\mu_T)^2}{2\sigma_T^2}} \ . \tag{2.23}$$

Similarly, $f(S_{ij}\,|\,i, j\in\bar{\Omega}) = f(S_{ij}\,|\,\mu_F, \sigma_F)$ is the value of the non-mate probability density function at $S_{ij}$, $\bar{\Omega}$ is the set of elements who are not neighbours:

$$f(S_{ij}\,|\,i, j\in\bar{\Omega}\ ) = \frac{1}{\sqrt{2\pi}\sigma_F} e^{-\frac{(S_{ij}-\mu_F)^2}{2\sigma_F^2}} \ , \tag{2.24}$$

hence

$$w_{ij} = \log \frac{p_{i,j\in\Omega}\sigma_F}{(1-p_{i,j\in\Omega})\sigma_T} + \frac{(S_{ij}-\mu_F)^2}{2\sigma_F^2} - \frac{(S_{ij}-\mu_T)^2}{2\sigma_T^2} \ . \tag{2.25}$$

The basic CLICK algorithm can be described recursively as follows: The algorithm handles some connected component of the sub-graph induced by the yet-unclustered elements in each step. If the component contains a single vertex, then this vertex is considered as a singleton and is handled separately. Otherwise, a stoping criterion is checked. If the component satisfies the criterion, it is declared a kernel. Otherwise, the component is split according to a minimum weight cut. The algorithm yields a list of kernels which serves as a basis for the eventual clusters. After the algorithm is finished, we can obtain the number of clusters *K* (Sharan and Shamir, 2000).

### 2.3.4 Silhouette method

To assess the quality of clusters, we used the silhouette measure proposed by Rousseeuw (1987) which is based on the comparison of the clusters tightness and separation. To calculate the silhouette value $s(i)$ of a gene $x_i$, firstly we must estimate two scalars $a(\mathrm{x}_\mathrm{i})$ and $b(\mathrm{x}_\mathrm{i})$. Suppose gene $x_i$ belongs to cluster *A*, when cluster *A* contains other genes apart from $x_i$, then we can compute

$$a(x_i) = \textit{average distance of gene i to all other genes of cluster A.} \qquad (2.26)$$

Then we consider any other cluster *C* which is different from *A*, and compute

$$d(i, C) = \textit{average distance of gene i to all objects of cluster C.} \qquad (2.27)$$

After computing $d(i, C)$ for all clusters $C \neq A$ , we select the smallest of those values and denote it by

$$b(x_i) = \min\{d(i, C)\},\ \ C \neq A. \qquad (2.28)$$

Suppose cluster B is the cluster for which this minimum is obtained, that is, $d(i, B) = b(x_i)$, then we call it the neighbour of gene $x_i$. Now $s(x_i)$ can be obtained by combining $a(x_i)$ and $b(x_i)$ as follows:

$$s(x_i) = [b(x_i) - a(x_i)\ ] / \max\{a(x_i), b(x_i)\}. \qquad (2.29)$$

From the above definition we can easily see that $s(x_i)$ is located in [-1, 1]. When $s(x_i)$ is close to 1, it implies that the ‘within’ distance $a(x_i)$ is much smaller than the smallest ‘between’ distance $b(x_i)$. Therefore, we can consider gene $x_i$ is tied with its cluster and it is ‘well-clustered’. Another situation is that $s(x_i)$ is around 0 which means $a(x_i)$ and $b(x_i)$ are almost equal, hence it is not clear whether gene $\mathrm{x}_\mathrm{i}$ should belong to either cluster *A* or *B*. This situation is considered as an ‘intermediate case’.

However, the worst situation is $s(x_i)$ is close to -1. It shows $a(x_i)$ is much larger than $b(x_i)$, thus gene $x_i$ is much closer to *B* than to *A*. Therefore, we consider this as a 'bad cluster' (Rousseeuw, 1987).

## 2.4 Data analysis and discussion

### 2.4.1 Testing

We used two different data sets downloaded from two databases. The first set is the **Serum data**. This data set contains 517 genes which were described and used by Iyer et al. (1999). Each gene contains 13 expression values. It can be downloaded from: http:// www.sciencemag.org/feature/data. The expression of these genes varies in response to serum concentration in human fibroblasts. The second set is the **Yeast data**. The original yeast micorarray data contains 6200 yeast genes which were measured every 10 min during two cell cycles in 17 hybridization experiments (Cho et al., 1998). We used the same 2945 genes selected by Tavazoie et al. (1999). In this selection, the data exclude values at time points 90 and 100 minutes. These data sets have already been normalized in such a way that the average expression values of each gene is zero and the standard deviation of each gene is one. For comparison, we generate random microarray data for different data sets as follows: To the first gene in the list of the data set, we associate an expression value selected randomly from the N values of the experiment *j*. To the second gene in the list, we associate an expression value selected randomly from the remaining (*N*-1) values of experiment j. We repeat this process until we associate the remaining expression values to the last gene in the list.

For different data sets we estimated the optimal values of *m* as in Dembele and Kanstner (2003), which are listed in Table 2.2. We used the same values of *m* for random data. For comparison, we also used $m = 2$ for each data set.

Table 2.2 Parameters and number of clusters used for FCM

| Data name | Number of genes | *m* used | ***Number of clusters*** |
|---|---|---|---|
| Serum (original) | 517 | 1.25 | 10 |
| Serum (denoised) | 517 | 1.58 | 10 |
| Yeast (original) | 2945 | 1.17 | 16 |
| Yeast (denoised) | 2945 | 1.48 | 16 |

Figure 2.4 illustrates the clustering structures of serum and yeast microarray data without noise removal. Using original data, we see both serum and yeast data have no clustering structure when *m* is set to 2. Especially, in yeast data the 16 memberships for each gene to 16 clusters are very similar to each other, suggesting a poor clustering result. To avoid this problem, we estimated the optimal values of *m* for the two data sets. Then we obtain clearer clutsering structure results. However, in the two randomized data sets there are still clear clustering structures. This observation shows that noise in data affects clustering results, and that clustering structures still can be found even in data sets which do not contain any biological significance. In order to remove noise in microarray data, we applied EMD to the original data. We denoised the serum data 4 times and the yeast data 5 times. We showed the noise removing process for serum data in Figure 2.5 as an example. After denoising it 4 times, we obtain a smooth trend which we used as denoised serum data to do clustering analysis.

For the denoised microarray data, firstly we set *m* to 2. We also estimated the optimal values of *m* for the two new data set and generated random data for the two denoised data sets respectively. We show the clustering results for the denoised data in Figure 2.6. It is seen that both denoised serum and yeast data have clear clustering structures when *m* is equal to 2 and the results are similar to the result on the original data when the estimated values of *m* are used. This observation suggests *m* = 2 is suitable for new data. When we used the estimated values of *m*, the results become more extreme. The highest membership values become closed to 1 which shows genes have tight association which cluster they belong to. However, for the random data sets, there is no clustering structure. Because we have removed noise in the original data, now it is reassuring that there is no clustering structure in random data without biological significance.

### 2.4.2 Assessment of quality of clusters

We show scatter plots of original data and denoised data in Figure 2.7. The horizontal axis represents the highest membership values of each gene and the vertical axis represents the second highest membership values. For serum data, we obtained similar results when we used $m$ = 1.25 and 2 for the original and denoised data respectively. When we used the estimated value $m$ = 1.58 for denoised serum data, the sum of the two highest membership values for each gene is closed to 1, which means the behaviour of each gene in denoised serum data can be almost entirely determined by its first and second membership values. However in the original yeast data, when we used the proper value $m$ = 1.17, we obtained a very dispersed distribution of the two highest memberships. After our denoising step, we got a better scatter plot when $m$ is set equal to 2. The sum of the two highest membership values is close to 1 when we used $m$ = 1.48.

Figure 2.8 illustrates the assessment of quality of clusters. The silhouette values lies between -1 and 1. When the value is less than zero, the corresponding gene is poorly classified. For serum data, we see that clustering results of the original data ($m$ = 1.25) and denoised data are similar. To some extent, the result for denoised data is better than that for the original data because the main part of the box plot is higher than 0.4. On the other hand, the result for denoised serum data ($m$ = 1.58) is much better than the above two results. For yeast data, we obtain the same result. However, the assessment of the 14th cluster of denoised yeast data is not satisfactory. The silhouette values of some genes are even lower than 0.4 which suggests poor clustering.

Figure 2.9 gives another way to assess the quality of clusters. This figure is generated by Gene Expression Data Analysis Studio (GEDAS) which is a cluster software designed by Fu (2007). The colours represent the values of each gene at each time point. The lower the value is, the greener the colour is. The higher the value is, the redder the colour is. In this figure, although we use the same method FCM to do cluster analysis on original and denoised yeast data respectively, we see the denoised microarray data show better separated and homogeneous clusters.

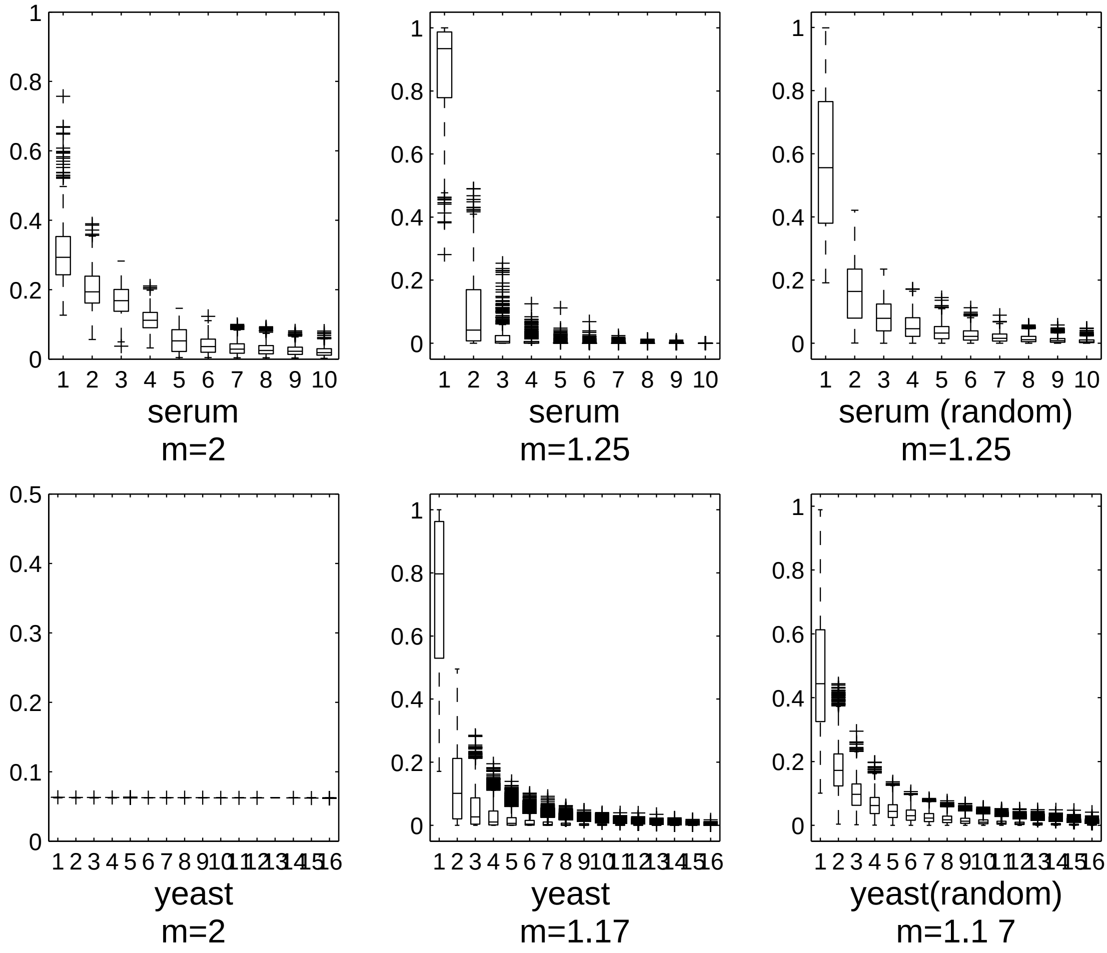
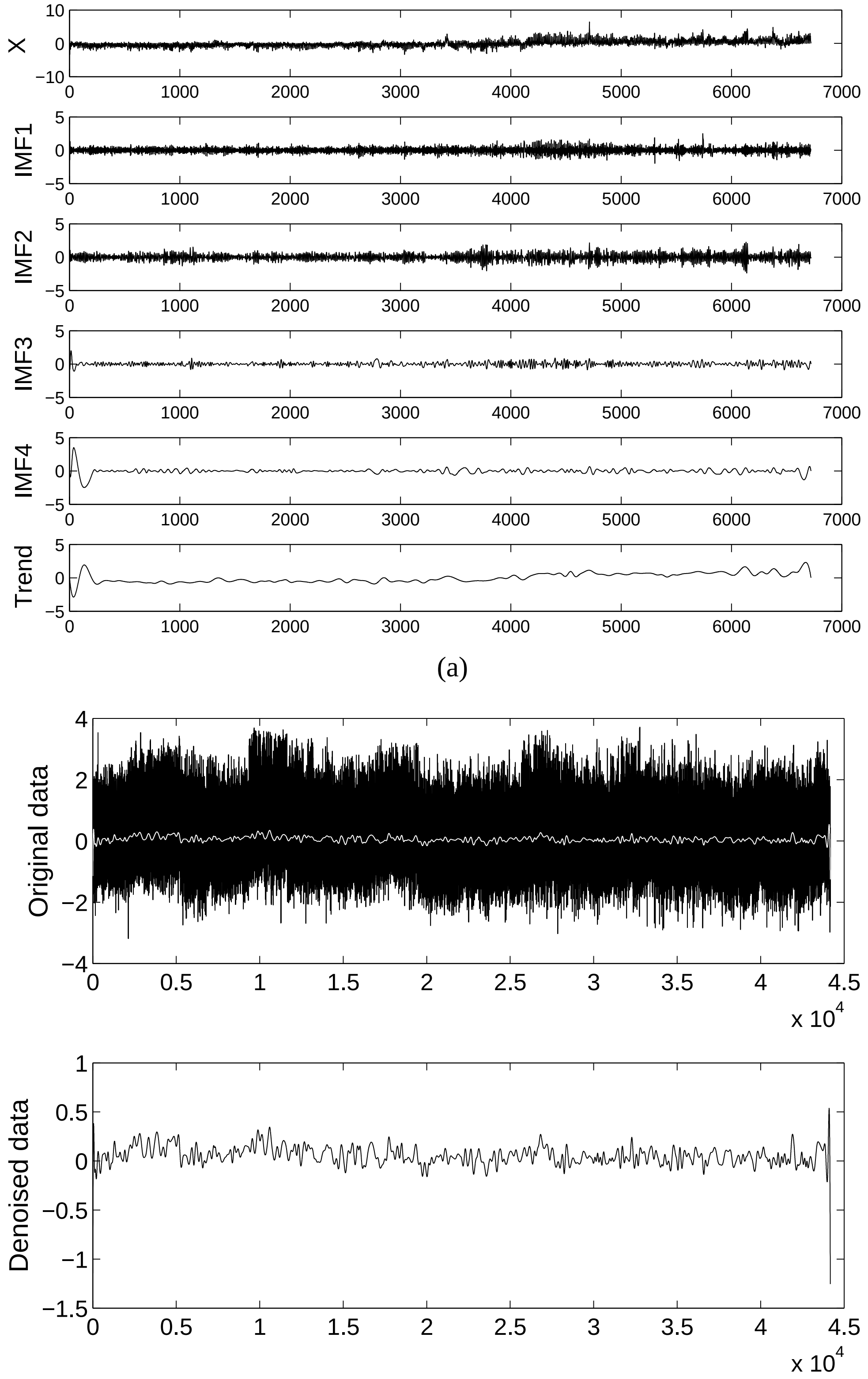


Figure 2.4 Influence of the fuzzy parameter and noise on the distribution of membership values. The horizontal axis represents the sorted number of membership. The vertical axis represents of membership values.



(a)

(b)

Figure 2.5 Noise removing process on the serum microarray data. (a) Noise removing process. (b) Comparison between denoised data and original data.

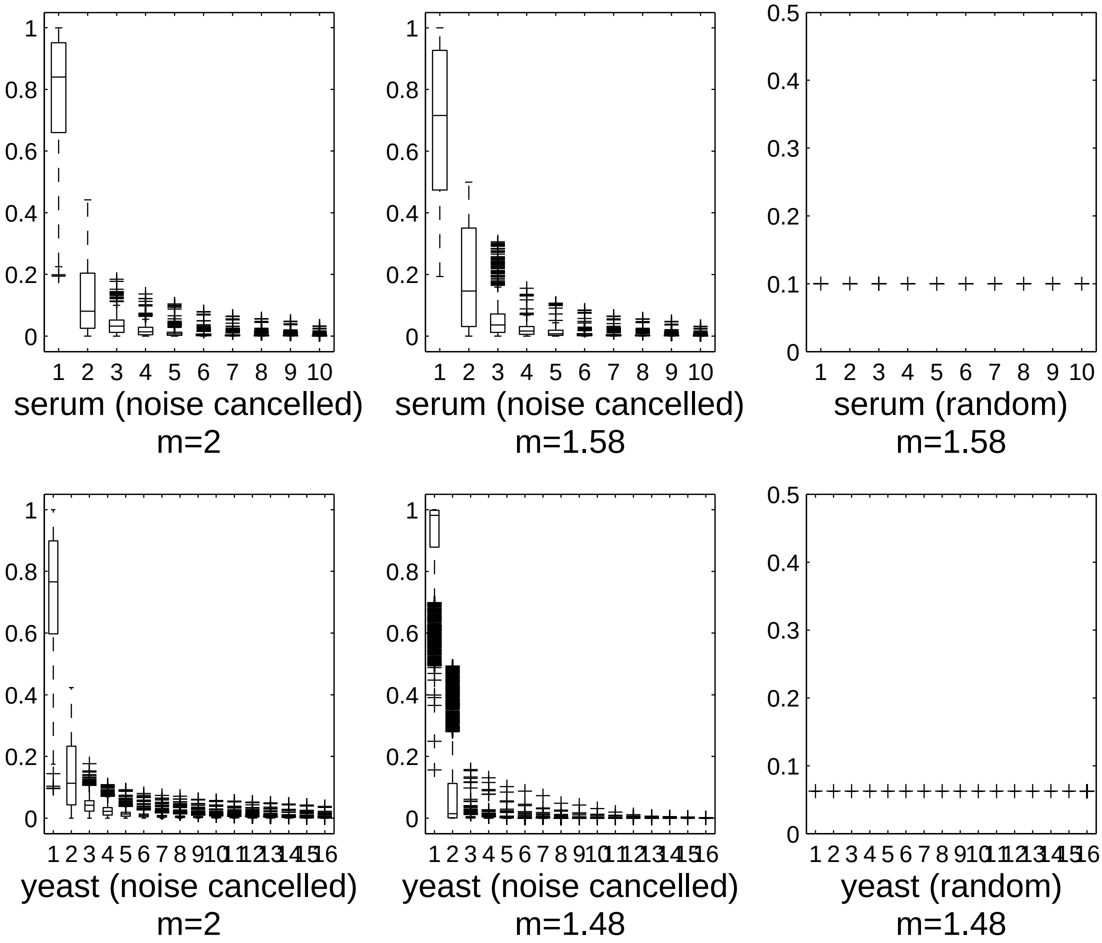


Figure 2.6 Cluster structure of noise cancelled data and random data. The horizontal axis represents the sorted membership. The vertical axis represents membership values. In noise cancelled serum and yeast data, we obtain clear cluster structure, however in the random data which contains none biological significance, there is no cluster structure.

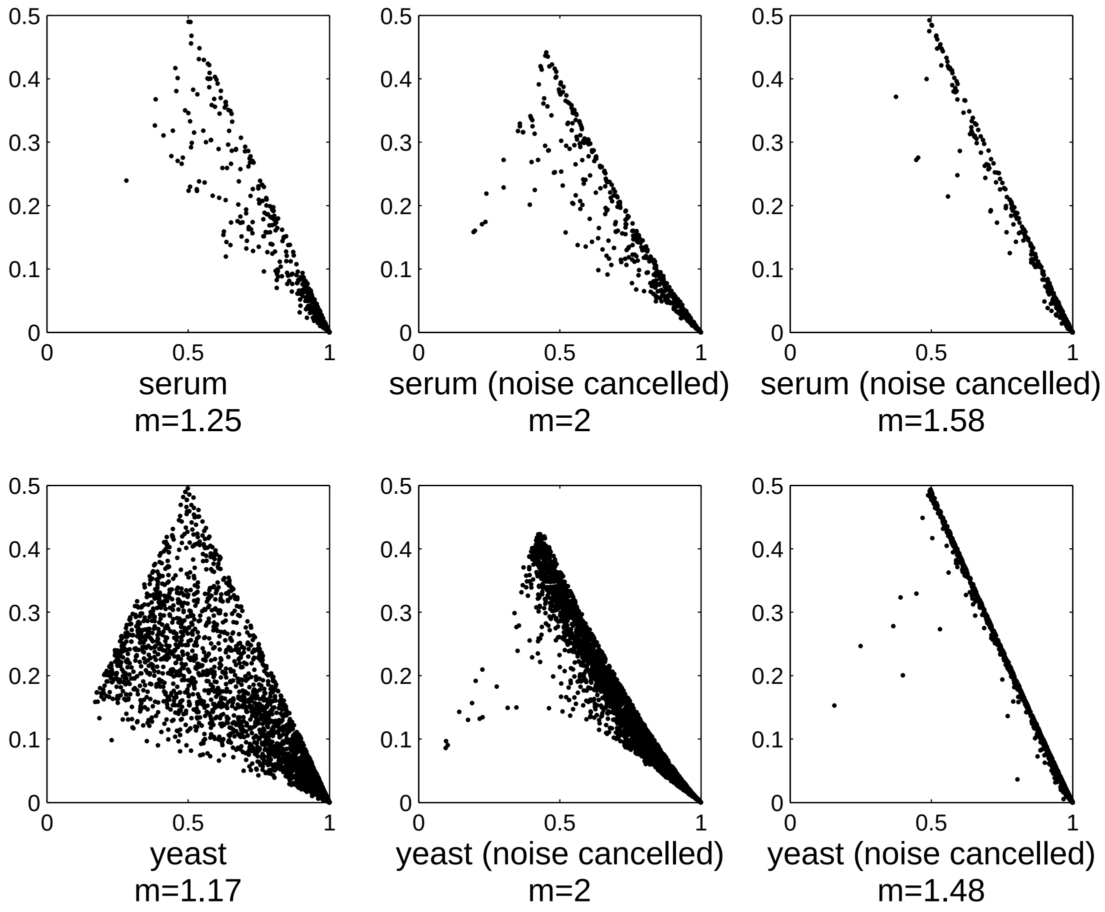


Figure 2.7 Scatter plots of the two highest membersip values of all genes in the serum and yeast data sets. The horizontal axis represents the highest membership values. The vertical axis represents the second highest membership values. After denoising, the sum of the two highest membership values is much close to 1 which suggests we can group genes easily from the two highest membership values.

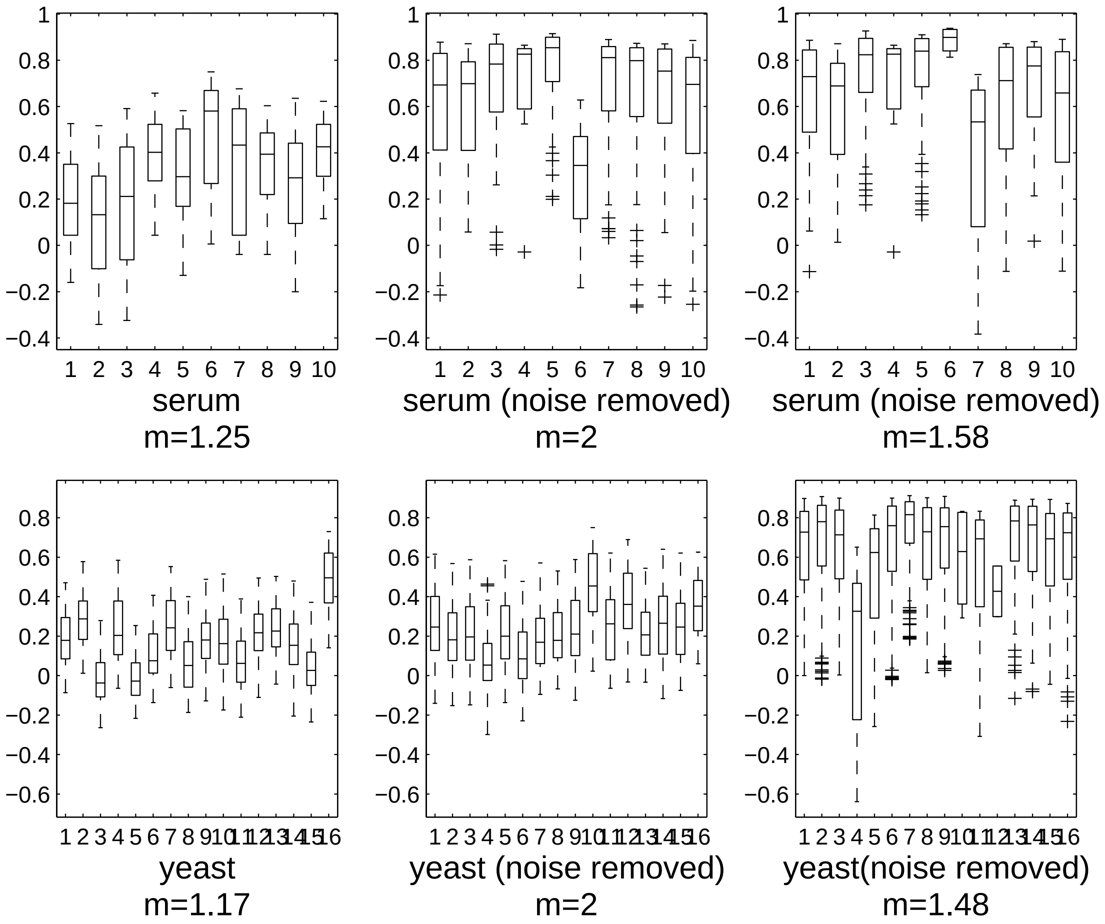


Figure 2.8 Box plots of silhouette values of genes in clusters. The horizontal axis represents number of cluster. The vertical axis is silhouette values.

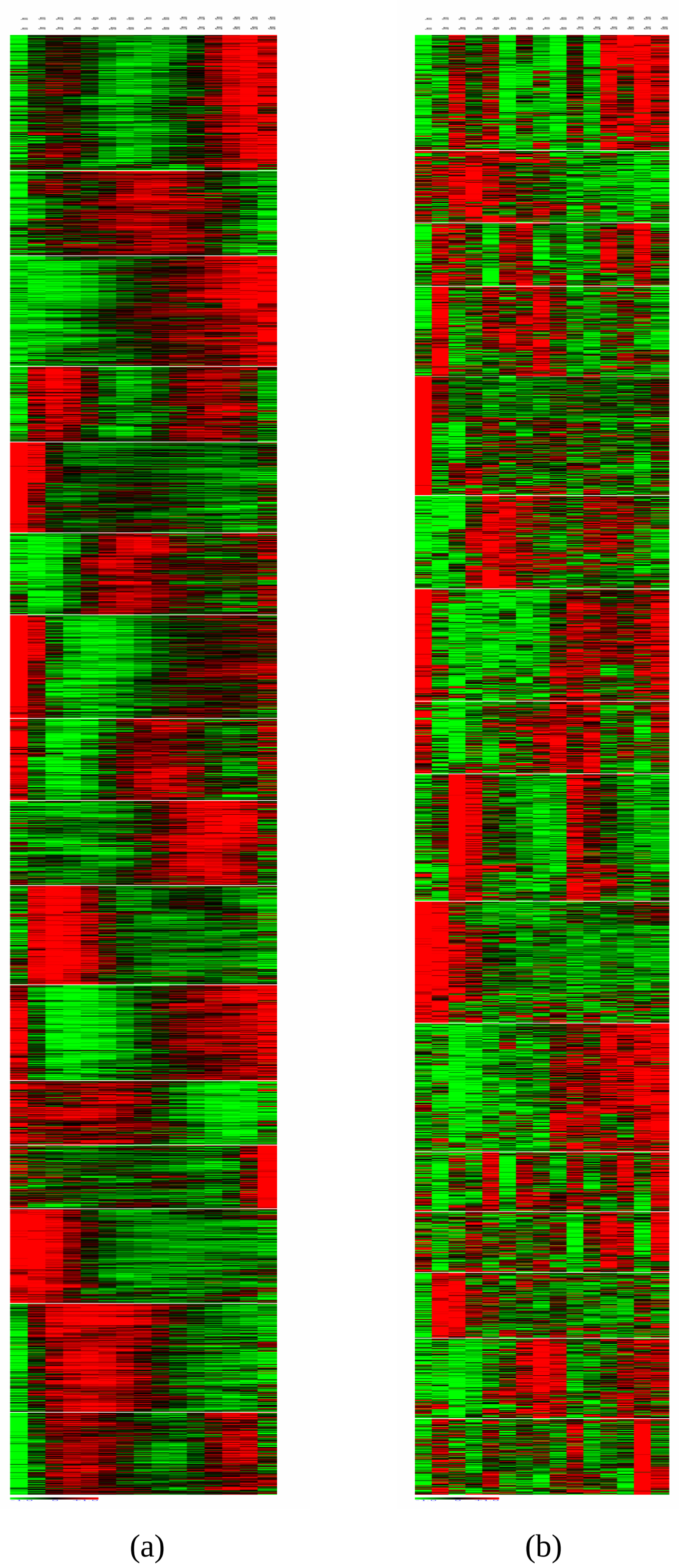


Figure 2.9 Cluster structure plot generated by GEDAS (Fu et al., 2002). (a) Cluster structure of denosied yeast data. (b) Cluster structure of original yeast data

## 2.5 Conclusion

Fuzzy clustering methods have been widely used for analysing gene expression data (Dougherty et al., 2002). However the estimation of the value of the fuzzy parameter *m* is still a problem. Dembele and Kastner (2003) proposed a predetermining method using distances between genes, but this method is based on observation and has no theoretical justification. On the other hand, FCM is sensitive to initialization. To avoid this problem, we have to run the program many more times. The FCM process and estimation of *m* are all time-consuming.

In this chapter, we proposed to combine the FCM method with empirical mode decomposition for clustering microarray data. Based on the analysis of clustering serum and yeast gene microarray data by FCM-EMD, the results suggest noise removing is necessary. For both data sets, we found clearer clustering structures from denoised data than from the original data. Especially, we cannot find any clustering structure in denoised random data which contains no biological significance. It suggests the noise has been almost removed and has little effect on the clustering results. Comparing with the clustering results on original data, we can even avoid estimation the fuzzy parameter *m* for denoised data to some extent. We can just use 2 as the parameter value and obtain better results than original data using estimating values. This makes clustering works more efficient.

We introduced the EMD method here to remove noise in microarray data. However, the number of times for this noise removal is still uncertain. When the signal becomes smooth, we consider noise has been removed, but this may not be sufficiently precise. Another problem is that the more times we denoise the more information we would lose in microarray data. Therefore, determination of the number of times of denoising is a pressing problem to be addressed.

# Chapter 3

## Type-2 fuzzy approach for disease-associated gene identification on microarrays

### 3.1 Introduction

Disease-associated gene identification is one of the most important areas of medical research today. It is known that certain diseases, such as cancer, are reflected in the change of the expression values of certain genes. For instance, due to genetic mutations, normal cells may become cancerous. These changes can affect the expression level of genes. Gene expression is the process of transcribing a gene's DNA sequence into RNA. A gene's expression level indicates the approximate number of copies of that gene's RNA produced in a cell and it is correlated with the amount of the corresponding proteins made (Mohammadi et al., 2011). Analysing gene expression data can indicate the genes which are differentially expressed in the diseased tissues. Several important breakthroughs and progress have been made (Liang et al., 2006).

One effective approach of identifying genes that are associated with a disease is to measure the divergence of two sets of values of gene expression. Usually, they are patients' and normal people's expression data. In order to identify the genes that are associated with disease, one need to determine from each gene whether or not the two sets of expression values are significantly different form each other (Liang et al., 2006). The two most popular methods to measure the divergence of two sets of values are t-test and Wilcoxon rank sum test (Rosner, 2000). According to Liang et al. (2006), both of these two methods have some limitations. The limitation of t-test is that it cannot distinguish two sets with close means even though the two sets are significantly different from each other. Another limitation is that it is very sensitive to extreme values. Although rank sum test overcomes the limitation of t-test in sensitivity to extreme values, it is not sensitive to absolute values. This might be advantageous to some application but not to others. To overcome these disadvantages, Liang et al. (2006) proposed the FM test. However, some limitations still exist. The most obvious one is when the values of gene microarray data are very similar and lack over-expression, in which case the FM-d valves are very close or even equal to each other. That made the FM-test inadequate in distinguishing disease genes.

To overcome these problems, we introduce type-2 fuzzy set theory into the research of disease-associated gene identification. Type-2 fuzzy set is an extension of traditional fuzzy set, introduced by Zadeh (1975). Of course, employment of type-2 fuzzy sets usually increases the computational complexity in comparison with type-1 fuzzy sets due to the additional dimension of having to compute secondary grades for each primary membership. However, if type-1 fuzzy sets would not produce satisfactory results, employment of type-2 fuzzy sets for managing uncertainty may allow us to obtain desirable results (Hwang and Rhee, 2007). Mizumoto and Tanaka (1976) have studied the set theoretic operations of type-2 sets, properties of membership grades of such sets, and have examined the operations of their algebraic product and algebraic sum (Mizumoto and Tanaka, 1981). Dubois and Prade (1980) have discussed the join and meet operations between fuzzy numbers under minimum t-norm. Karnik and Mendel (1998, 2000) have provided a general formula for the extended sup-star composition of type-2 relations. Type-2 fuzzy sets have already been used in a number of applications, including decision making (Chaneau et al., 1987; Yager, 1980), solving fuzzy relation equations (Wagenknecht and Hartmann, 1988), and pre-processing of data (John et al., 1998).

In this chapter we establish the type-2 fuzzy membership function for identification of disease-associated genes on microarray data of patients and normal people. We call it type-2 fuzzy membership test (type-2 FM-test) and apply it to diabetes and lung cancer data. For the ten best-ranked genes of diabetes identified by the type-2 FM-test, 7 of them have been confirmed as diabetes associated genes according to genes description information in Genebank and the published literature. One more gene than original approaches is identified. Within the 10 best ranked genes identified in lung cancer data, 7 of them are confirmed by the literature which is associated with lung cancer treatment. The type-2 FM-d values are significantly different, which makes the identifications more reasonable and convincing than the original FM-test. In the next section, we introduce the theoretical background needed for a description of the type-2 FM-test detailed in Section 3.3, and we will give our results in Section 3.4.

## 3.2 Theoretical background

### 3.2.1 Type-2 fuzzy sets

The concept of a type-2 fuzzy set was introduced by Zadeh (1975) as an extension of the concept of an ordinary fuzzy set (which we can call it type-1 fuzzy set). The transition from ordinary sets to fuzzy sets tells us, when we cannot determine the membership of an element in a set as 0 or 1, we use fuzzy sets of type-1. Similarly, when the circumstances are so fuzzy that we cannot determining the membership grade even as a crisp number in [0, 1], we can use fuzzy sets of type-2. If we continue thinking along this line, we can say that no finite-type fuzzy set (type-∞) can completely represent uncertainty.  However, as we go on to higher types, the complexity of computation increases rapidly. Therefore in this chapter we just deal with type-2 fuzzy sets.

We now give the definition of a type-2 fuzzy set and associated concepts.

**Definition 3.1** A type-2 fuzzy set, denoted as $\tilde{A}$, is characterized by a type-2 membership function $\mu_{\tilde{A}}(x, u)$, where $x \in$ X and u$\in$ J$_x \subseteq [0, 1]$,

$$\tilde{A} = \left\{ \left( (x,u), \mu_{\tilde{A}}(x,u) \right) \middle| \forall x \in X, \forall u \in J_x \subseteq [0,1] \right\}, \tag{3.1}$$

in which $0 \le \mu_{\tilde{A}}(x, u) \le 1$. $\tilde{A}$ can also be expressed as

$$\tilde{A} = \int_{x \in X} \int_{u \in J_x} \mu_{\tilde{A}}(x,u) \big/ (x,u), \quad J_x \subseteq [0,1] \ , \tag{3.2}$$

where ∬ denotes union over all admissible $x$ and $u$.

In Definition 3.1, the restriction that $\forall\, u \in$ J$_x$ is the same with type-1 constraint that $0 \le \mu_A(x) \le 1$. That is, if the blur disappears, then a type-2 membership function must reduce to a type-1 membership function, in which case the variable $u$ equals $\mu_A(x)$

and $0 \leq \mu_A(x) \leq 1$. $0 \leq \mu_{\tilde{A}}(x, u) \leq 1$ is an additional restriction which is consistent with the fact that the amplitudes of a membership function should lie between or be equal to 0 and 1 (Mendel, 2001).

**Definition 3.2**: At each value of *x*, i.e. $x = x'$, the 2D plane whose axes are u and $\mu_{\tilde{A}}(x', u)$ is called a vertical slice of $\mu_{\tilde{A}}(x, u)$. A secondary membership function is a vertical slice of $\mu_{\tilde{A}}(x, u)$. It is $\mu_{\tilde{A}}(x = x', u)$ for $x' \in X$ and $\forall$ u$\in$ $J_{x'} \subseteq [0,1]$,

$$\mu_{\tilde{A}}(x = x', u) \equiv \mu_{\tilde{A}}(x') = \int_{u \in J_{x'}} f_{x'}(u)/u\ ,\ J_{x'} \subseteq [0,1]\,, \tag{3.3}$$

in which $0 \leq f_{x'}(\mathrm{u}) \leq 1$. Because $\forall x' \in X$, we drop the prime notation on $\mu_{\tilde{A}}(x')$ and refer to $\mu_{\tilde{A}}(x')$ as a secondary membership function; it is also a type-1 fuzzy set, which we also refer to as a secondary set (Mendel, 2001).

Based on the concept of secondary sets, we can reinterpret a type-2 fuzzy set as the union of all secondary set,

$$\tilde{A} = \left\{ \left( x, \mu_{\tilde{A}}(x) \right) \middle| \forall x \in X \right\}, \tag{3.4}$$

or, as

$$\tilde{A} = \int_{x \in X} \mu_{\tilde{A}}(x)/x = \int_{x \in X} \left[ \int_{u \in J_x} f_x(u)/u \right] \Big/ x \quad J_x \subseteq [0,1]\,. \tag{3.5}$$

**Definition 3.3**: The domain of a secondary membership function is called the primary membership of *x*. In 3.5, $J_x$ is the primary membership of *x* where $J_x \subseteq [0,1]$ for $\forall x \in X$ (Castillo and Melin, 2008).

**Definition 3.4**: The amplitude of a secondary membership function is called secondary grade. In 3.5, $f_x(u)$ is a secondary grade; in (3.2) $\mu_{\tilde{A}}(x = x', u = u')$ is a secondary grade.

If *X* and $J_x$ are both discrete, no matter by problem formulation or by discretization of continuous universes of discourse, then the type-2 fuzzy set can be expressed as

$$\tilde{A} = \sum_{x \in X} \left[ \sum_{u \in J_x} f_x(u)/u \right] \Big/ = \sum_{i=1}^{N} \left[ \sum_{u \in J_{x_i}} f_{x_i}(u) \Big/ u \right] \Big/ x_i$$

$$=\left[\sum_{k=1}^{M_1} f_{x_1}(u_{1k})\Big/u_{1k}\right]\Big/x_1+\ldots+\left[\sum_{k=1}^{M_N} f_{x_N}(u_{NK})\Big/u_{Nk}\right]\Big/x_N\ . \tag{3.6}$$

We can observe that x has been discretized into N points and at each of these values u has been discretized into $M_i$ values. However, the discretization along each $u_{ik}$ does not have to be the same number. The expressions similar to (3.5) can be written for the mixed cases when *X* is continuous but $J_x$ is discrete, or vice-versa. The most important case for us in the thesis will be equation (3.6), because when a type-2 membership function is programmed it must be discretized, not only over *X* but also over $J_x$.

There are many choices for the secondary membership functions, such as Gaussian, Trapezoidal and Triangular. We associate the type-2 set with the name of its secondary membership functions. If the secondary membership functions are Gaussian, then we can call it a Gaussian type-2 fuzzy set.

Note that when $f_x(u)=1$, $\forall\, u\in J_x\subseteq[0,1]$, then the secondary membership functions are interval sets; we call this kind of type-2 fuzzy sets interval type-2 fuzzy sets. Interval secondary membership functions reflect a uniform uncertainty at the primary membership of *x*. In this chapter we apply interval type-2 fuzzy sets, which can reduce the computational complexity significantly, to identification of disease-related genes (Mendel, 2001).

**Definition 3.5**: Assume that each of the secondary membership functions of a type-2 fuzzy set has only one secondary grade equal to 1. A principal membership function is the union of all such points at which this occurs, i.e.,

$$\mu_{principal}(x)=\int_{x\in X} u/x\text{, where } f_x(u) = 1. \tag{3.7}$$

The principal membership function for the Gaussian type-2 fuzzy set is the solid Gaussian curve in Figure 3.1 (a) (Castillo and Melin, 2008).

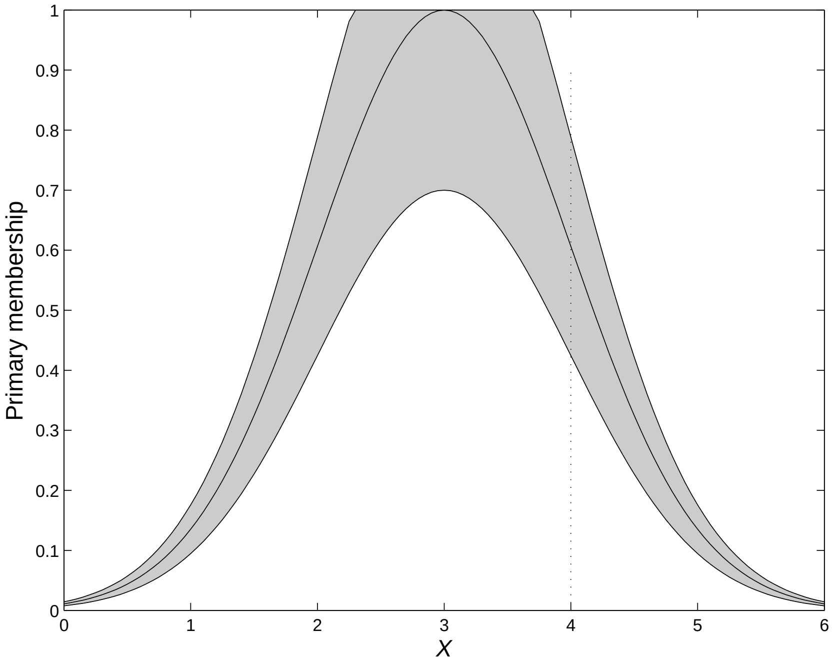
**Definition 3.6**: Uncertainty in the primary memberships of a type-2 fuzzy set, $\tilde{A}$, consists of a bounded region that we call the footprint of uncertainty (FOU). It is the union of all primary memberships, i.e.,

$$FOU(\tilde{A}) = \bigcup_{x\in X} J_x \ . \tag{3.8}$$

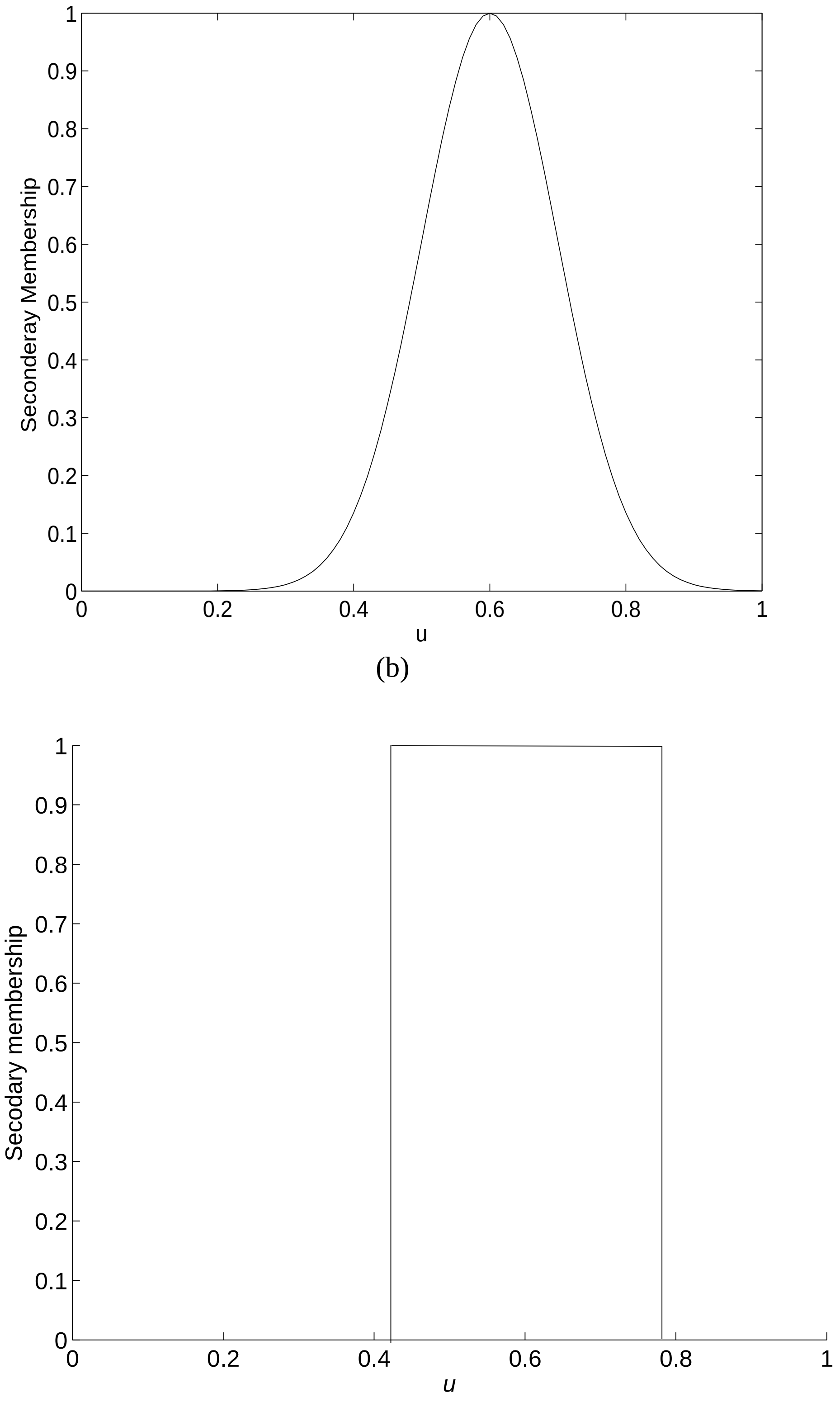
The term FOU is very useful, because it not only focuses our attention on the uncertainties inherent in a specific type-2 membership function, whose shape is a direct consequence of the nature of these uncertainties, but also provides a very convenient verbal description of the entire domain of support for all the ssecondary grades of a type-2 membership function. An example of a FOU is the shaded regions in Figure 3.1 (a). The FOU is shaded uniformly to indicate that it is for a Gaussian type-2 fuzzy set.

**Definition 3.7**: Consider a family of type-1 membership functions $\mu_A\,(x|p_1, p_2, \ldots, p_v)$ where $p_1, p_2, \ldots, p_v$ are parameters, some or all of which vary over some range of values, i.e., $p_i\in P_i$ ($i$ = 1, …, $v$). A primary membership function is any one of these type-1 membership functions, e.g., $\mu_A(x|p_1 = p_{1'}, p_2 = p_{2'}, \ldots, p_v{=}p_{v'})$ .

It is subject to some restrictions on its parameters. The family of all primary membership functions creates a FOU.



(a)



(b)

(c)

Figure 3.1 Gaussian type-2 fuzzy set. (a). FOU of a Gaussian type-2 fuzzy set. (b). Gaussian type-2 secondary membership function. (c). Interval secondary membership function.

### 3.2.2 Type-2 fuzzy set operations

In this part, we will give some introduction of set theoretical operations of type-2 fuzzy sets. We will explain how to compute the union, intersection and complement for type-2 fuzzy sets. Consider two type-2 fuzzy sets $\tilde{A}$ and $\tilde{B}$, i.e.

$$\tilde{A} = \int_{x} \mu_{\tilde{A}}(x)/x = \int_{X} \left[ \int_{J_x^u} f_x(u)/u \right] \Big/ x, \quad J_x^u \subseteq [0,1], \tag{3.9}$$

and

$$\tilde{B} = \int_{x} \mu_{\tilde{B}}(x)/x = \int_{X} \left[ \int_{J_x^w} g_x(u)/u \right] \Big/ x, \; J_x^w \subseteq [0,1]. \tag{3.10}$$

**Union of type-2 fuzzy sets**

The union of $\tilde{A}$ and $\tilde{B}$ is another type-2 fuzzy set, just as the union of type-1 fuzzy sets *A* and *B* is another type-1 fuzzy set,

$$\tilde{A} \cup \tilde{B} \Leftrightarrow \mu_{\tilde{A}\cup\tilde{B}}(x,v) = \int_{x\in X} \mu_{\tilde{A}\cup\tilde{B}}(x)/x = \int_{x\in X} \left[ \int_{v\in J_x^v \subseteq [0,1]} h_x(v)/v \right] \Big/ x, \tag{3.11}$$

where

$$\int_{v\in J_x^v} h_x(u)/v = \varphi\left( \int_{u\in J_x^u} f_x(u)/u, \int_{w\in J_x^w} g_x(w)/w \right) = \varphi\left(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\right), \tag{3.12}$$

here, $\varphi$ plays the role of *f* in (3.9), which is a t-conorm function of the secondary membership functions, $\mu_{\tilde{A}}(x)$ and $\mu_{\tilde{B}}(x)$, which are type-1 fuzzy sets. $\varphi$ is a t-conorm function because the union of two type-1 fuzzy sets is equivalent to the t-conorm of their membership functions. Following the prescription of the right-hand side of (3.9), we see that

$$\varphi\left( \int_{u\in J_x^u} f_x(u)/u, \int_{w\in J_x^w} g_x(w)/w \right) = \int_{u\in J_x^u} \int_{w\in J_x^w} f_x(u) \bullet g_x(w)/\varphi(u,w), \tag{3.13}$$

when we consider $\varphi$ is the maximum operation $\vee$, then (3.11) and (3.13) can be expressed as

$$\mu_{\tilde{A}\cup\tilde{B}}(x) = \int_{v\in J_x^v \subset [0,1]} h_x(v)/v = \int_{u\in J_x^u} \int_{w\in J_x^w} f_x(u) \bullet g_x(w)/(u \vee w), \tag{3.14}$$

where $\bullet$ indicates minimum or product, and $\iint$ indicates union over $J_x^u \times J_x^w$.

Another way to express (3.14) is in terms of the secondary membership functions of $\tilde{A}$ and $\tilde{B}$ which is proposed by Mizumoto and Tanaka (1976):

$$\mu_{\tilde{A}\cup\tilde{B}}(x) = \int_{u\in J_x^u}\int_{w\in J_x^w} f_x(u)\bullet g_x(w)/v \equiv \mu_{\tilde{A}}(x) \coprod \mu_{\tilde{B}}(x), \tag{3.15}$$

where $v \equiv u \vee w$ and $\coprod$ indicates the so-called join operation (Mizumoto and Tanaka, 1976).

Equation (3.15) indicates that to perform the join between two secondary membership functions, $\mu_{\tilde{A}}(x)$ and $\mu_{\tilde{B}}(x)$, $v = u \vee w$ must be performed between every possible pair of primary memberships *u* and *w*, such that u$\in$ $J_x^u$ and w$\in$ $J_x^w$ and that the secondary grade of $\mu_{\tilde{A}\cup\tilde{B}}(x)$ must be computed as the t-norm operation between the corresponding secondary grades of $\mu_{\tilde{A}}(x)$ and $\mu_{\tilde{B}}(x)$, $f_x(u)$ and $g_x$(x), respectively. According to (3.11), this work must be done for any *x* in *X* to obtain $\mu_{\tilde{A}\cup\tilde{B}}(x)$.

**Intersection of type-2 fuzzy sets**

The intersection of $\tilde{A}$ and $\tilde{B}$ is also another type-2 fuzzy set, just as the intersection of type-1 fuzzy sets *A* and *B* is another type-1 fuzzy sets,

$$\tilde{A}\cap\tilde{B} \Leftrightarrow \mu_{\tilde{A}\cap\tilde{B}}(x,v) = \int_{x\in X} \mu_{\tilde{A}\cap\tilde{B}}(x)/x, \tag{3.16}$$

the development of $\mu_{\tilde{A}\cap\tilde{B}}(x)$ is the same as that of $\mu_{\tilde{A}\cup\tilde{B}}(x)$, except that in the present case φ is the minimum or product function $\wedge$,

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \int_{u\in J_x^u}\int_{w\in J_x^w} f_x(u)\bullet g_x(w)/v. \tag{3.17}$$

Another way to express (3.17) is in terms of the secondary membership functions of $\tilde{A}$ and $\tilde{B}$, as

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = \int_{u\in J_x^u}\int_{w\in J_x^w} f_x(u)\bullet g_x(w)/v \equiv \mu_{\tilde{A}}(x)\prod\mu_{\tilde{B}}(x), \tag{3.18}$$

where $v \equiv u \vee w$ and $\prod$ denotes the so-called meet operation (Mizumoto and Tanaka, 1976).

Equation (3.18) indicates that to perform the meet between two secondary membership functions $\mu_{\tilde{A}}(x)$ and $\mu_{\tilde{B}}(x)$, $v = u \wedge w$ must be performed between every possible pair of primary memberships *u* and *w*, such that $u \in J_x^u$ and $w \in J_x^w$, and the secondary grade of $\mu_{\tilde{A}\cap\tilde{B}}(x)$ must be computed as the t-norm operation between the corresponding secondary grades of $\mu_{\tilde{A}}(x)$ and $\mu_{\tilde{B}}(x)$, $f_x(u)$ and $g_x$(x), respectively. According to (3.18), this must be done for any *x* in *X* to obtain $\mu_{\tilde{A}\cap\tilde{B}}(x)$.

**Complement of a type-2 fuzzy set**

The complement of $\tilde{A}$ is another type-2 fuzzy set, just as the complement of type-1 fuzzy set *A* is another type-1 fuzzy sets:

$$\bar{\tilde{A}} \Leftrightarrow \mu_{\bar{\tilde{A}}}(x,v) = \int_{x\in X}\mu_{\bar{\tilde{A}}}(x)/x. \tag{3.19}$$

In this equation $\mu_{\bar{\tilde{A}}}(x)$ indicates a secondary membership function; i.e., at each value of x, $\mu_{\bar{\tilde{A}}}(x)$ is a function:

$$\mu_{\bar{\tilde{A}}}(x) = \int_{u\in J_x^u} f_x(u)/(1-u) \equiv \neg\mu_{\tilde{A}}(x), \tag{3.20}$$

where $\neg$ denotes the so-called negation operation (Mizumoto and Tanaka, 1976). Equation (3.20) indicates that to perform the negation of the secondary membership function $\mu_{\bar{\tilde{A}}}(x)$, 1-*u* must be computed at $\forall\ u \in J_x^u$, and the secondary grade of $\mu_{\bar{\tilde{A}}}(x)$ at 1-u is the corresponding secondary grade of $\mu_{\tilde{A}}(x)$ and $f_x(u)$. According to (3.19), this must be done for any *x* in *X* to obtain $\mu_{\bar{\tilde{A}}}(x)$.

Examples of operations of type-2 fuzzy sets, join, meet, negation, are as follows: Consider two type-2 fuzzy sets :

$$\tilde{A}=\frac{\mu_{\tilde{A}}(x_1)}{x_1}+\frac{\mu_{\tilde{A}}(x_2)}{x_2}+\frac{\mu_{\tilde{A}}(x_3)}{x_3},\ \tilde{B}=\frac{\mu_{\tilde{B}}(x_1)}{x_1}+\frac{\mu_{\tilde{B}}(x_2)}{x_2}+\frac{\mu_{\tilde{B}}(x_3)}{x_3},$$

where

$$\mu_{\tilde{B}}(x_1)=\frac{0.3}{0.4},\ \mu_{\tilde{A}}(x_2)=\frac{0.4}{0.2},\ \mu_{\tilde{B}}(x_2)=\frac{0.1}{0.5}+\frac{0.4}{0.6},\ \mu_{\tilde{A}}(x_3)=\frac{0.5}{0.6}+\frac{0.9}{0.7},\ \mu_{\tilde{B}}(x_3)=\frac{0.9}{0.8}.$$

Following (3.14), we have

$$\mu_{\tilde{A}\cup\tilde{B}}(x_1)=\mu_{\tilde{A}}(x_1)\coprod\mu_{\tilde{B}}(x_1)=\left(\frac{0.3}{0.1}\right)\vee\left(\frac{0.3}{0.4}\right)=\frac{0.3\wedge 0.3}{0.1\vee 0.4}=\frac{0.3}{0.4},$$

$$\mu_{\tilde{A}\cup\tilde{B}}(x_2)=\mu_{\tilde{A}}(x_2)\coprod\mu_{\tilde{B}}(x_2)=\left(\frac{0.4}{0.4}\right)\vee\left(\frac{0.1}{0.5}+\frac{0.4}{0.6}\right)$$

$$=\frac{0.4\wedge 0.1}{0.2\vee 0.5}+\frac{0.4\wedge 0.4}{0.2\vee 0.6}=\frac{0.1}{0.5}+\frac{0.4}{0.6},$$

$$\mu_{\tilde{A}\cup\tilde{B}}(x_3)=\mu_{\tilde{A}}(x_3)\coprod\mu_{\tilde{B}}(x_3)=\left(\frac{0.5}{0.6}+\frac{0.9}{0.7}\right)\vee\frac{0.9}{0.8}$$

$$=\frac{0.5\wedge 0.9}{0.6\vee 0.8}+\frac{0.9\wedge 0.9}{0.7\vee 0.8}=\frac{0.5}{0.8}+\frac{0.9}{0.8}=\frac{0.9}{0.8},$$

then,

$$\tilde{A}\cup\tilde{B}=\frac{0.3/0.4}{x_1}+\frac{0.1/0.5+0.4/0.6}{x_2}+\frac{0.9/0.8}{x_3}.$$

We also can obtain meet and negation of the two type-2 fuzzy sets following (3.17) and (3.20):

$$\tilde{A}\cap\tilde{B}=\frac{0.3/0.1}{x_1}+\frac{0.4/0.2}{x_2}+\frac{0.5/0.6+0.9/0.3}{x_3},$$

$$\bar{\tilde{A}}=\frac{0.3/0.9}{x_1}+\frac{0.4/0.8}{x_2}+\frac{0.5/0.4+0.9/0.7}{x_3}.$$

### 3.2.3 Type-2 fuzzy membership function

In this chapter we apply interval Gaussian type-2 fuzzy sets to identification of disease-related genes, therefore we give some examples of type-2 fuzzy sets with Gaussian primary membership function.

Consider the case of a Gaussian primary membership function having a fixed mean, *m*, and an uncertain standard deviation that takes on values in $[\sigma_1, \sigma_2]$, i.e.,

$$\mu_A(x) = \exp\left[-\tfrac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right],\ \sigma \in [\sigma_1, \sigma_2]. \tag{3.21}$$

Corresponding to each value of $\sigma$ we will get a different membership curve. Here we set $\sigma \in [1, 2]$, *m* = 5; we obtain Figure 3.2

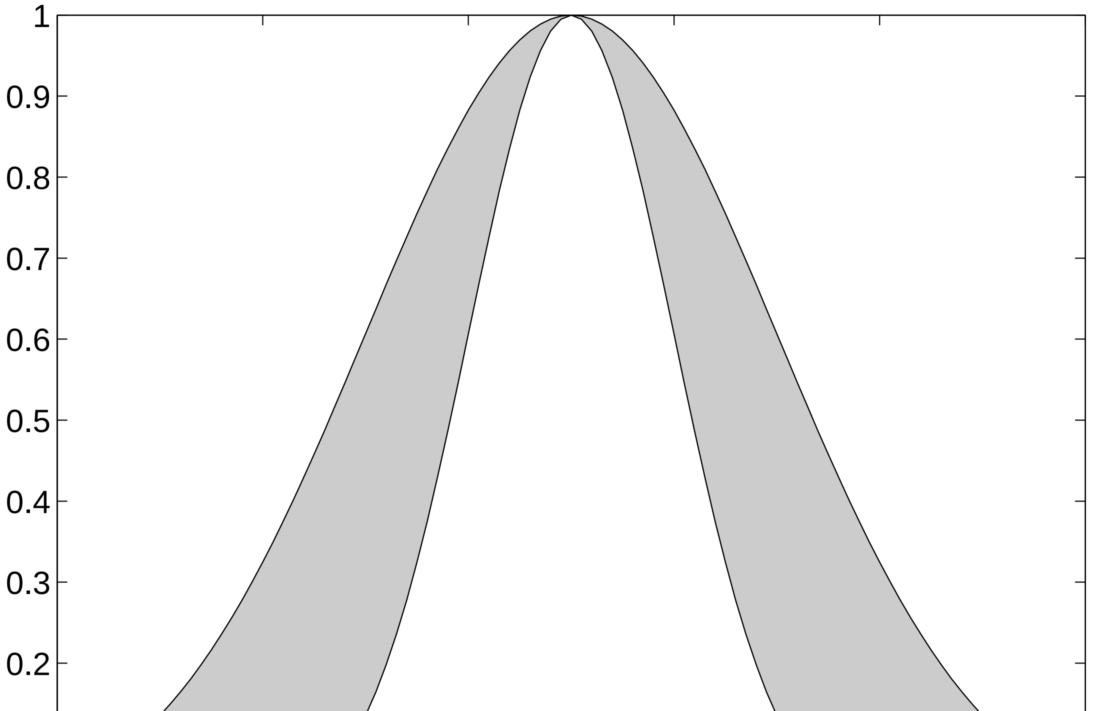


Figure 3.2: FOU for Gaussian primary membership function with uncertain standard deviation.

Consider the case of a Gaussian primary membership function having a fixed standard deviation $\sigma$, and an uncertain mean that takes on values in{m$_1$, m$_2$}, i.e.,

$$\mu_A(x) = \exp\left[-\tfrac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right], \quad m\in \{m_1, m_2\}. \tag{3.22}$$

Corresponding to each value of *m*, we will get a different membership curve. Here we set $\sigma$ = 2, $m_1$ = 4, $m_2$ = 7; we obtain Figure 3.3

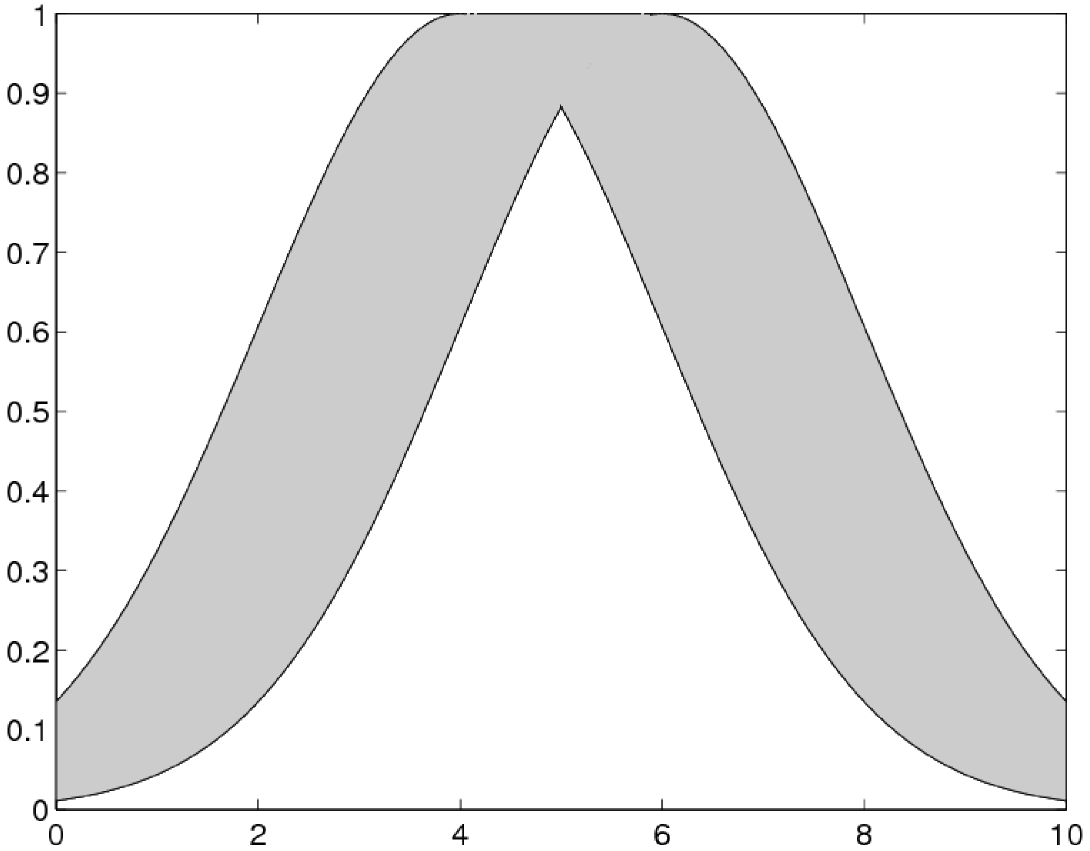


Figure 3.3 FOU for Gaussian primary membership function with mean, m$_1$ and m$_2$.

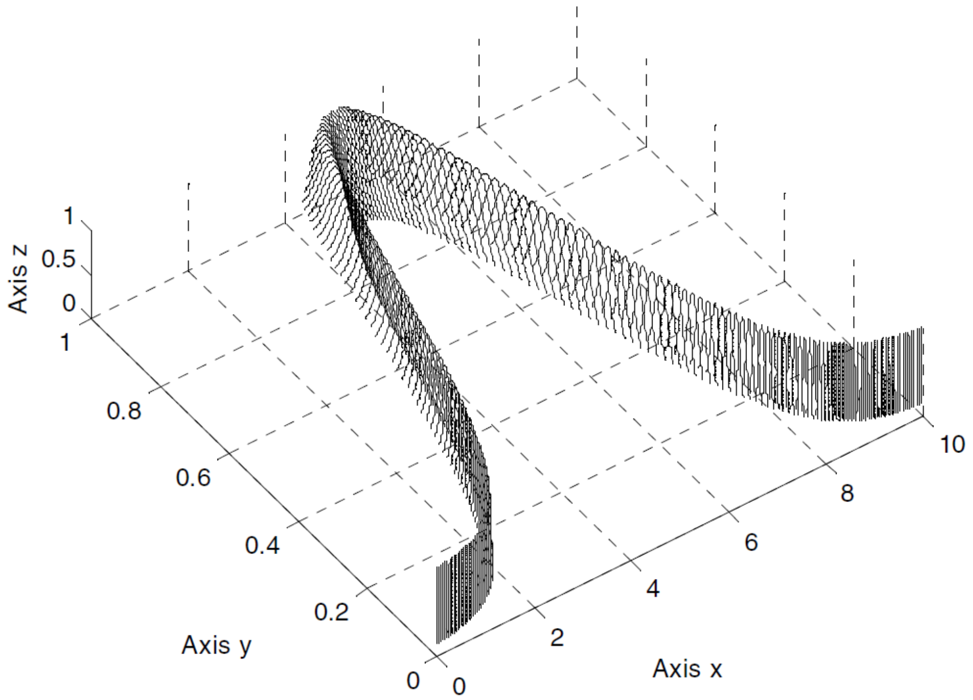


Figure 3.4 Three-dimensional view of a type-2 membership function.

In Figure 3.4 we have a three-dimensional view of a type-2 Gaussian membership function. The structure of primary membership function and secondary membership function are clearly showed in this figure.

The FOU can be described in terms of upper and lower membership functions. In the application we use upper and lower membership functions to establish primary membership functions of diabetes data and lung cancer data.

**Definition 3.8**: An upper membership function and a lower membership function (Mendel and Liang, 1999) are two type-1 membership functions which are bounds for the FOU of a type-2 fuzzy set $\tilde{A}$. The upper membership function is associated with the upper bound of FOU(Ã), and is denoted $\bar{\mu}_{\tilde{A}}(x)$ , $\forall x \in X$ . The lower membership function is associated with the lower bound of FOU(Ã), and is denoted $\underline{\mu}_{\tilde{A}}(x), \forall x \in X$ , i.e.,

$$\bar{\mu}_{\tilde{A}}(x) = \overline{FOU(\tilde{A})}, \ \forall x \in X \ , \tag{3.23}$$

and

$$\underline{\mu}_{\tilde{A}}(x) = \underline{FOU(\tilde{A})} \ \forall x \in X \ . \tag{3.24}$$

Since the domain of a secondary membership function has been constrained in equation (3.2) to be contained in [0, 1], lower and upper membership functions always exist. From (3.10), we see that

$$\overline{FOU(\tilde{A})} = \bigcup_{x \in X} \overline{J_x} \ , \tag{3.25}$$

and

$$\underline{FOU(\tilde{A})} = \bigcup_{x \in X} \underline{J_x} \ , \tag{3.26}$$

where $\overline{J_x}$ and $\underline{J_x}$ denote the upper and lower bounds on $J_x$, respectively; hence, $\bar{\mu}_{\tilde{A}}(x) = \overline{J_x}$ and $\underline{\mu}_{\tilde{A}}(x) = \underline{J_x}$, $\forall x \in X$ .

We can express (3.2) in terms of upper and lower membership functions as

$$\begin{aligned} \tilde{A} = \mu_{\tilde{A}}(x,u) = \int_{x \in X} \mu_{\tilde{A}}(x) \big/ x = \int_{x \in X} \left[ \int_{u \in J_x} f_x(u)/u \right] \Big/ x \\ = \int_{x \in X} \left[ \int_{u \in [\underline{J_x}, \overline{J_x}]} f_x(u)/u \right] \Big/ x \, . \end{aligned} \tag{3.27}$$

We see from this equation that the secondary membership function $\mu_{\tilde{A}}(x)$ can be expressed in terms of upper and lower membership function as

$$\mu_{\tilde{A}}(x) = \int_{\mu \in [\underline{J_x}, \overline{J_x}]} f_x(u)/u \ , \tag{3.28}$$

in the special but important case when the secondary membership functions are interval sets, then (3.27) simplifies to

$$\tilde{A} = \int_{x \in X} \left[ \int_{u \in J_x} 1/u \right] \Big/ x = \int_{x \in X} \left[ \int_{u \in [\underline{J_x}, \overline{J_x}]} 1/u \right] \Big/ x \, . \tag{3.29}$$

We use upper and lower membership functions to compute the differences between genes in this chapter.

For the Gaussian primary membership function with uncertain mean (Figure 3.3), the upper membership function $\bar{\mu}_{\tilde{A}}(x)$ is

$$\bar{\mu}_{\tilde{A}}(x) = \begin{cases} N(m_1,\sigma;x) & x < m_1 \\ 1 & m_1 \le x \le m_2 \\ N(m_1,\sigma;x) & x > m_2 \end{cases}, \tag{3.30}$$

where, for instance, $N(m_1,\sigma;x) \equiv \exp[-\tfrac{1}{2}(\frac{x-\mu}{\sigma})^2]$. The upper thick solid curve in Figure 3.3 denotes the upper membership function. The lower membership function, $\underline{\mu}_{\tilde{A}}(x)$, is

$$\underline{\mu}_{\tilde{A}}(x) = \begin{cases} N(m_2,\sigma;x) & x \le \dfrac{m_1+m_2}{2} \\ N(m_1,\sigma;x) & x > \dfrac{m_1+m_2}{2} \end{cases}. \tag{3.31}$$

The thick lower curve in Figure 3.3 represents the lower membership function.

From this example we see that the upper or lower membership functions cannot be denoted by just one mathematical function over its entire x-domain. It may consist of several branches and each is defined over a different segment of the entire x-domain. When the input x is located in a specific x-domain segment, we call its corresponding membership function branch an active branch (Liang and Mendel, 2000); e.g., in (3.31), when x > ($m_1$+$m_2$) / 2, the active branch for $\underline{\mu}_{\tilde{A}}(x)$ is $N(m_1,\sigma;x)$.

For the Gaussian primary membership function with uncertain standard deviation (Figure 3.2), the upper membership function, $\bar{\mu}_{\tilde{A}}(x)$, is

$$\bar{\mu}_{\tilde{A}}(x) = N(m\ ,\sigma_2;x), \tag{3.32}$$

and the lower membership function, $\underline{\mu}_{\tilde{A}}(x)$, is

$$\underline{\mu}_{\tilde{A}}(x) = N(m\ ,\sigma_1; x)\,. \tag{3.33}$$

The upper thick solid curve in Figure 3.2 denotes the upper membership function, and the lower thick solid curve denotes the lower membership function. We see that the upper and lower membership functions are simpler for this example than for the preceding one.

These two examples illustrate how to define the upper and lower membership functions so that it is clear how to define them for other situations. However, for the problem in this chapter, the upper and lower membership functions we established contain uncertainty both in mean and standard deviation. The plot is close to Figure 3.1 (a).

### 3.2.4 Centroid of type-2 fuzzy sets and type-reduction

Type-reduction methods are "extended" versions of type-1 defuzzification methods. These methods give us a type-1 starting from the type-2 set obtained at the output of the inference engine which is very important for fuzzy logic system and fuzzy clustering methods (such as type-2 fuzzy c-means). Defuzzification is considered as a task of finding the centroid of a fuzzy set. This centroid itself, as an output of a fuzzy logic system, can mostly represent the fuzzy set and describe the fuzzy concept. The centroid of a type-1 set A, whose domain is discretized into N points, is given as

$$C_A = \frac{\sum_{i=1}^{N} x_i \mu_A(x_i)}{\sum_{i=1}^{N} \mu_A(x_i)}\,, \tag{3.34}$$

similarly, the centroid of a type2 fuzz set $\tilde{A}$ whose domain is discretized into N points so that

$$\tilde{A} = \sum_{i=1}^{N} \left[ \int_{u \in J_{x_i}} f_{x_i}(u)/u \right] \Big/ x_i\ , \tag{3.35}$$

can be defined using the Extension Principle as follows (Karnik and Mendel, 1998, 1999)

$$C_{\tilde{A}} = \int_{\theta \in J_{x_1}} \dots \int_{\theta_N \in J_{x_N}} \left[ f_{x_1}(\theta_1) \bullet \dots \bullet f_{x_N}(\theta_N) \right] \Big/ \frac{\sum_{i=1}^{N} x_i \theta_i}{\sum_{i=1}^{N} \theta_i} , \qquad (3.36)$$

where $C_{\tilde{A}}$ is a type-1 fuzzy set.

**Definition 3.9**: For discrete universes of discourse *X* and *U*, an embedded type-2 fuzzy set $\tilde{A}_e$ has N elements, where $\tilde{A}_e$ contains exactly one element from $J_{x_1}, J_{x_2}, \dots, J_{x_N}$, namely $\theta_1, \theta_2, \dots, \theta_N$, each with its associated secondary grade, namely $f_{x_1}(\theta_1)$, $f_{x_2}(\theta_2), \dots,$ $f_{x_N}(\theta_N)$, i.e.,

$$\tilde{A}_e = \sum_{i=1}^{N} \left[ f_{x_i}(\theta_i) / \theta_i \right] \Big/ x_i , \quad \theta_i \in J_{x_i} \subseteq U = [0,1] . \qquad (3.37)$$

**Definition 3.10**: For discrete universes of discourse X and U, an embedded type-1 set A$_e$ has N elements, one each from $J_{x_1}, J_{x_2}, \dots, J_{x_N}$, namely $\theta_1, \theta_2, \dots, \theta_N$, i.e.,

$$A_e = \sum_{i=1}^{N} \theta_i \Big/ x_i , \quad \theta_i \in J_{x_i} \subseteq U = [0,1] . \qquad (3.38)$$

From the above equation we see that the set A$_e$ is actually the union of all primary memberships of the type-2 fuzzy set $\tilde{A}$.

Every combination of $\theta_1, \dots, \theta_N$ and its associated secondary grade $f_{x_1}(\theta_1) \bullet \dots \bullet f_{x_N}(\theta_N)$ forms an embedded type-2 fuzzy set $\tilde{A}_e$. Each element of $C_{\tilde{A}}$ is determined by computing the centroid $\sum_{i=1}^{N} x_i \theta_i \Big/ \sum_{i=1}^{N} \theta_i$ of the embedded type-1 set A$_e$ that is associated with $\tilde{A}_e$ and computing the t-norm of the secondary grades associated with $\theta_1, \dots, \theta_N$, namely $f_{x_1}(\theta_1) \bullet \dots \bullet f_{x_N}(\theta_N)$. The complete centroid $C_{\tilde{A}}$ is determined by doing this for all the embedded type-2 sets in $\tilde{A}_e$.

Let $\theta = [\theta_1, ..., \theta_N]^T$,

$$a(\theta) \equiv \frac{\sum_{i=1}^{N} x_i \theta_i}{\sum_{i=1}^{N} \theta_i} \ , \tag{3.39}$$

and

$$b(\theta) \equiv f_{x_1}(\theta_1) \bullet ... \bullet f_{x_N}(\theta_N) \, , \tag{3.40}$$

then $C_{\tilde{A}}$ can also be expressed as

$$C_{\tilde{A}} = \int_{\theta_1 \in J_{x_1}} ... \int_{\theta_N \in J_{x_N}} b(\theta) / a(\theta) \, , \tag{3.41}$$

in terms of $a(\theta)$ and $b(\theta)$, the computation of $C_{\tilde{A}}$ involves computing the tuple ($a(\theta)$, $b(\theta)$) many times. Suppose, ($a(\theta)$, $b(\theta)$) is computed α times, then, we can consider the computation of $C_{\tilde{A}}$ as the computation of the α tuples ($a_1$, $b_1$), ($a_2$, $b_2$), …, ($a_\alpha$, $b_\alpha$). If two or more combinations of vector $\theta$ give the same point in the centroid set, then we keep the largest value of b(θ).

From (3.31), we see that the domain of $C_{\tilde{A}}$ will be an interval $[a_l(\theta), a_r(\theta)]$ , where

$$a_l(\theta) = \min_\theta a(\theta) \, , \tag{3.42}$$

and

$$a_r(\theta) = \max_\theta a(\theta) \, . \tag{3.43}$$

A practical sequence of computations to obtain $C_{\tilde{A}}$ is summarized as follows:

1. Discretize the *x*-domain into *N* points $x_1, ..., x_N$ .
2. Discretize each $J_{x_j}$ into a suitable number of points, denoted by $M_j$
3. Enumerate all the embedded type-1 fuzzy sets; there will be $\prod_{j=1}^{N} M_j$ of them.
4. Compute the centroid using (3.31), for example, compute the α tuples ($a_k$, $b_k$), k=1,2,…, $\prod_{j=1}^{N} M_j$ , where $a_k$ and $b_k$ are given in (3.29) and (3.30), respectively.

For an interval type-2 fuzz set, (3.26) reduces to

$$C_{\tilde{A}} = \int_{\theta \in J_{x_1}} \cdots \int_{\theta_N \in J_{x_N}} 1 \Bigg/ \frac{\sum_{i=1}^{N} x_i \theta_i}{\sum_{i=1}^{N} \theta_i} \ . \qquad (3.44)$$

In this chapter, we use interval type-2 fuzzy set to establish the similarity membership function between patients and normal people data. In the application, we do not use type-2 fuzzy logic system. The type-reduction step in our problem is aimed to obtain the final membership value of the similarity which is the basis to verify the differences of expression values of genes in the two different data sets.

## 3.3 Methods

In this section, we introduce two methods: fuzzy membership test and type-2 fuzzy membership test, which are applied to identification of disease-associated genes in the next section and we also make some comparison between these two methods.

### 3.3.1 Fuzzy membership test

The fuzzy membership test (FM-test) is proposed by Liang (2006). In this approach, a new concept of fuzzy membership d-value (FM d-value) is defined to quantify the divergence of two sets of values. They applied FM-test to diabetes and lung cancer expression data sets, respectively. The details of this method are as follows.

Let $S_1$ and $S_2$ be two sets of values of a particular feature for two groups of samples under two different conditions. For the problem we plan to solve, the two sets can be patient’s and normal people’s gene expression values. The basic idea of this approach is to consider the two sets of values as samples from two different fuzzy sets. For each fuzzy set, a membership function is established and the membership value of each element is examined with respect to the other fuzzy set. By calculating the average of membership values, the divergence of the original two sets can be measured. In particular, the following steps are performed:

1. Compute the sample mean and standard deviation of $S_1$ and $S_2$ respectively.

2. Characterize $S_1$ and $S_2$ as two fuzzy sets $FS_1$ and $FS_2$ whose fuzzy membership functions, $f_{FS_1}(x)$ and $f_{FS_2}(x)$, are defined with the sample means and standard deviations. The fuzzy membership function $f_{FS_i}(x)$ ($i$ = 1, 2) maps each value $x_j$ to a fuzzy membership value that reflects the degree of $x_j$ belonging to $f_{FS_i}(x)$ ($i$ = 1, 2). For each gene, the value $x_j$ is the expression value of patients or normal people, where $j$ = 1, 2,…, $N$.

3. Quantify the convergence degree of two sets $S_1$ and $S_2$ by the two fuzzy membership functions, $f_{FS_1}(x)$ and $f_{FS_2}(x)$. We will give the definition of the convergence degree below.

4. Define the divergence degree (FM d-value) between the two sets based on the convergence degree.

Liang (2006) applied the Gaussian function as the fuzzy membership function, then the mean and standard deviation are calculated.

The sample mean $\mu_1$ of $S_1$ is calculated as

$$\mu_1 = \frac{1}{N_1} \sum_{x_i \in S_1} x_i \,, \tag{3.45}$$

where $N_1$ is the number of elements in $S_1$, and the sample standard deviation $\sigma_1$ of $S_1$ is calculated as

$$\sigma_1 = \sqrt{\frac{1}{N_1 - 1} \sum_{x_i \in S_1} (x_i - \mu_1)^2} \,, \tag{3.46}$$

then, the fuzzy membership function of set $S_1$ is defined as

$$f_{FS_1}(x) = e^{-(x-\mu_1)^2 / 2\sigma_1^2} \,. \tag{3.47}$$

The function $f_{FS_1}(x)$ maps each value $x$ in $S_1$ to a fuzzy membership value to quantify the degree that $x$ belongs to $FS_1$. A value equal to the mean has a membership value of 1 and belongs to fuzzy set *FS*1 to a full degree; a value that deviates from the mean has a smaller membership value and belongs to *FS*1 to a smaller degree. The further the value deviates from the mean, the smaller the fuzzy membership value is. Similarly, the fuzzy membership function for $S_2$ is defined as

$$f_{FS_2}(x) = e^{-(x-\mu_1)^2/2\sigma_2^2} \quad , \tag{3.48}$$

where $\mu_2$ and $\sigma_2$ are the mean and standard deviation of $S_2$ respectively.

Since the fuzzy membership functions can overlap, one element can belong to more than one fuzzy set with a respective degree for each. For an element in $S_1$, we measure the degree that it belongs to $FS_1$ by applying its value to $f_{FS_1}(x)$. Similarly we can apply its value to $f_{FS_2}(x)$ to measure the degree that it belongs to $FS_2$. The idea of FM-test is to consider the membership value of an element in $S_1$ with respect to $S_2$ as a bond between $S_1$ and $S_2$, and vice versa; then the aggregation of all these bonds reflects the overall bond between these two sets. The weaker this overall bond is, the more divergent these two sets are. The strength of the overall bond between two sets is quantified by their *c*-value, which aggregates the mutual membership values of elements in $S_1$ and $S_2$ and is defined as follows.

**Definition 3.11** (FM c-value): Given two sets $S_1$ and $S_2$, the convergence degree between $S_1$ and $S_2$ in FM-test is defined as

$$c(S_1, S_2) = \frac{\sum_{e \in S_1} f_{FS_2}(e) + \sum_{e \in S_2} f_{FS_1}(f)}{|S_1| + |S_2|} \, . \tag{3.49}$$

**Definition 3.12** (FM d-value): Given two sets $S_1$ and $S_2$, the FM d-value degree between $S_1$ and $S_2$ in FM-test is defined as

$$d(S_1, S_2) = 1 - c(S_1, S_2) = 1 - c(S_1, S_2) = 1 - \frac{\sum_{e \in S_1} f_{FS_2}(e) + \sum_{e \in S_2} f_{FS_1}(f)}{|S_1| + |S_2|} \, . \tag{3.50}$$

### 3.3.2 Type-2 fuzzy membership test

In this section, based on the FM-test of Liang (2006), we propose type-2 fuzzy membership test for disease-associated gene identification. We also consider $S_1$ and $S_2$ as two sets of values of a particular feature for two groups of samples under two different conditions, but this time we will establish type-2 fuzzy membership for the two sets $\tilde{S}_1$ and $\tilde{S}_2$. We choose the Gaussian function as the primary membership function. To avoid computational complexity, we apply the interval secondary membership function for this problem, which means all the secondary membership values are 1. Following the theoretical basis we introduced above, we should establish the upper and lower primary membership functions to describe the uncertainty in the gene expression data. In particular, this method is performed as follows:

1. Use the Gaussian function as the primary membership function to compute the mean ($\mu_1$, $\mu_2$) and standard deviation ($\sigma_1$, $\sigma_2$) of $S_1$ and $S_2$.

2. For each set, we establish the upper and lower primary membership functions. Here, both the mean and the standard deviation will be uncertain. We use two parameters $\alpha$ and $\beta$ which are in [0, 1] to control the uncertainty in mean and standard deviation respectively. Based on the FM-test and the rules of establishing upper and lower primary memberships (3.30-3.33) for $\tilde{S}_1$, we obtain the upper primary membership as

$$\bar{\mu}_{\tilde{S}_1}(x) = \begin{cases} e^{-[x-(1-\alpha)\mu_1]^2/2(1+\beta)\sigma_1^2}, & \text{if } x < (1-\alpha)\mu_1 \\ 1, & \text{if } (1-\alpha)\mu_1 \le x \le (1+\alpha)\mu_1 \\ e^{-[x-(1+\alpha)\mu_1]^2/2(1+\beta)\sigma_1^2}, & \text{if } x > (1+\alpha)\mu_1 \end{cases}, \tag{3.51}$$

and the lower primary membership as

$$\underline{\mu}_{\tilde{S}_1}(x) = \begin{cases} e^{-[x-(1+\alpha)\mu_1]^2/2(1-\beta)\sigma_1^2}, & \text{if } x \le \mu_1 \\ e^{-[x-(1-\alpha)\mu_1]^2/2(1-\beta)\sigma_1^2}, & \text{if } x > \mu_1 \end{cases}, \tag{3.52}$$

We can obtain the upper and lower primary membership functions similarly for $\tilde{S}_2$:

$$\overline{\mu}_{\tilde{S}_2}(x) = \begin{cases} e^{-[x-(1-\alpha)\mu_2]^2/2(1+\beta)\sigma_2^2}, & \text{if } x < (1-\alpha)\mu_2 \\ 1, & \text{if } (1-\alpha)\mu_2 \le x \le (1+\alpha)\mu_2 \\ e^{-[x-(1+\alpha)\mu_2]^2/2(1+\beta)\sigma_2^2}, & \text{if } x > (1+\alpha)\mu_2 \end{cases}, \tag{3.53}$$

$$\underline{\mu}_{\tilde{S}_2}(x) = \begin{cases} e^{-[x-(1+\alpha)\mu_2]^2/2(1-\beta)\sigma_2^2}, & \text{if } x \le \mu_2 \\ e^{-[x-(1-\alpha)\mu_2]^2/2(1-\beta)\sigma_2^2}, & \text{if } x > \mu_2 \end{cases}, \tag{3.54}$$

3. Use the upper and lower primary membership functions $\overline{\mu}_{\tilde{S}_i}(x)$ and $\underline{\mu}_{\tilde{S}_i}(x)$, $i$ = 1,2; and the secondary membership values $f_x(u)$ to quantify the convergence of $S_1$ and $S_2$. Type-reduction work is needed in this step. Here, since we use the interval type-2 fuzzy set, $f_x(u)$ =1, $\forall u \in J_x \subseteq [0,1]$. The secondary memberships are all uniformly weighted for each primary membership of $x$.

4. Calculate the divergence degree between the two sets based on the convergence degree.

Type-reduction is an important step for type-2 fuzzy sets. In our application, $\forall x \in X$ , a primary membership interval [ $\underline{\mu}_{\tilde{S}_i}(x)$ , $\overline{\mu}_{\tilde{S}_i}(x)$ ] can be obtained. We discretize it into $N$ points, where $a_1 = \underline{\mu}_{\tilde{S}_i}(x)$ and $a_N = \overline{\mu}_{\tilde{S}_i}(x)$; then the final membership of $x$ can be obtained as

$$\mu(x) = \frac{\sum_{i=1}^{N} a_i \times f_x(a_i)}{N} \quad . \tag{3.55}$$

This type reduced membership $\mu(x)$ maps each value $x$ in $S_1$ or $S_2$ into a membership value to quantify the degree that x belongs to type-2 fuzzy set $\tilde{S}_1$ or $\tilde{S}_2$. For simplicity, we put $a_i = (a_{i-1} + a_{i+1}) / 2$, $i = 2,\ldots, N-1$, while $f_x(a_i) = 1$, $\forall$ x$\in$ X, $i = 1,\ldots, N$. (3.55) can be expressed as

$$\mu(x) = \frac{\underline{\mu}_{\tilde{S}_i}(x) + \bar{\mu}_{\tilde{S}_i}(x)}{2}, \tag{3.56}$$

to compute the overall bond between $S_1$ and $S_2$, we define type-2 FM c-values based on Liang et al. (2006).

**Definition 3.13** (Type-2 FM c-values): Given two sets $S_1$ and $S_2$, the convergence degree between $S_1$ and $S_2$ in FM-test is defined as

$$c(S_1, S_2) = \frac{\sum_{x\in S_1} \mu_{\tilde{S}_2}(x) + \sum_{y\in S_2} \mu_{\tilde{S}_1}(y)}{|S_1| + |S_2|} . \tag{3.57}$$

We define the divergence value as follows:

**Definition 3.14**: Given two sets $S_1$ and $S_2$, the divergence degree between $S_1$ and $S_2$ in the FM-test is defined as

$$d(S_1, S_2) = 1 - c(S_1, S_2). \tag{3.58}$$

Because the membership function maps the elements of $Si$ into a type-2 fuzzy set $\tilde{S}_j$, $i \neq j$, the aggregation of all membership values in the two sets can be used to quantify the similarity of $S_1$ and $S_2$. It can be considered as an overall bond between these two sets. The weaker this overall bond is, the more divergent these two sets are. In this case, for a given gene, the expression values between patients and normal people can be significantly different. If the elements in both $S_1$ and $S_2$ have high membership values, $S_1$ is very similar to $S_2$. In this case, for a given gene, the expression values do not change a lot between patients and normal people.

## 3.4 Data analysis and discussion

In this section, we apply type-2 FM-test to a diabetes expression dataset and a lung cancer expression dataset, respectively. Meanwhile, we make a comparison with the results of traditional FM-test by Liang et al. (2006).

### 3.4.1 Analysis of diabetes data

The first dataset is a diabetes dataset of microarray gene expression data. It contains 10831 genes and is downloaded from Yang et al. (2002). For each gene in this dataset, there are 10 expression values, five from a group of insulin-sensitive (IS) people and five from a group of insulin-resistant (IR) people. Table 3.1 is an example of the gene expression values under two conditions. To make this data more reliable, only the genes that have null expression values are included in this analysis. Meanwhile, we also require that, for a gene to be included, at least five out of its ten expression values are greater than 100.

Table 3.1: The gene expression values of diabetes data under two conditions.

| Gene | IR | | | | | IS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 123 | 142 | 11 | 406 | 220 | 305 | 398 | 707 | 905 | 688 |
| 2 | 200 | 191 | 220 | 83 | 197 | 49 | 81 | 116 | 111 | 135 |
| 3 | 750 | 559 | 649 | 695 | 639 | 310 | 359 | 135 | 97 | 178 |
| 4 | 246 | 213 | 232 | 134 | 67 | 86 | 79 | 77 | 94 | 61 |
| 5 | 598 | 424 | 695 | 451 | 141 | 342 | 260 | 266 | 229 | 234 |

Ten best-ranked genes of diabetes identified by the type-2 FM-test and the original FM-test are shown in table 3.2. From this table we see that the results of the two methods are not too much different. The bold letters are names of genes which are associated with diabetes.

Table 3.2 Ten best-ranked genes associated with diabetes.

| Type-2 FM-test | | |
|---|---|---|
| Probe Set | Gene Description | T2 d-value |
| **U49573** | Human phosphatidylinositol (4,5) bisphosphate | 0.6733 |
| **X53586** | Human. mRNA for integrin alpha 6 | 0.6131 |
| **M60858** | Human. nucleolin gene | 0.6080 |
| U61734 | Homo sapiens transmembrane emp24-like trafficking protein 10 | 0.5831 |
| D85181 | Homo sapiens mRNA for fungal sterol-C5-desaturase homolog | 0.5808 |
| **Z26491** | Homo sapiens gene for catechol o-methyltrans-fease | 0.5773 |
| **L07648** | Human MXII mRNA | 0.5769 |
| M95610 | Human alpha 2 type IX collagen (COL9A2) mRNA | 0.5760 |
| **L07033** | Human hydroxymethylglutaryl-CoA lyase mRNA | 0.5749 |
| **X81003** | Homo sapiens HCG V mRNA | 0.5525 |
| FM-test | | |
| Probe Set | Gene Description | FM d-value |
| **U45973** | Human phosphatidylinostiol (4,5) bisphosphate | 0.9988 |
| **M60858** | Human nucleolin gene | 0.9351 |
| D85181 | Homo sapiens mRNA for fungal sterol-C5-desaturase homolog | 0.8918 |
| M95610 | Huamn alpha 2 type IX collagen (COL9A2) mRNA | 0.8718 |
| **L07648** | Human MXII mRNA | 0.8575 |
| **L07033** | Human hydroxymethylglutaryl-CoA lyase mRNA | 0.8554 |
| **X53586** | Human mRNA for integrin alpha 6 | 0.8513 |
| **X81003** | Homo sapiens HCG V mRNA | 0.7914 |
| X57959 | Ribosomal protein L7 | 0.7676 |
| U06452 | Melan-A | 0.7566 |

For type-2 FM-test, within the 10 significant genes identified, 7 of them have been confirmed to be associated with diabetes according to genes description information in Genebank and the published literature. One more gene than original approaches is identified. According to the further research in the published literature, we have the following information.

Human phosphatidylinositol (4, 5) bisphosphate 5-phosphatase homolog (gene U45973) was found to be differentially expressed in insulin resistance cases. Over-expression of inositol polyphosphate 5-phosphatase-2 SHIP2 has been shown to inhibit insulin-stimulated phosphoinositide 3-kinase (PI3K) dependent signalling events. Analysis of diabetic human subjects has revealed an association between SHIP2 gene polymorphism and type II diabets mellitus. Aso knockout mouse studies have shown that SHIP2 is a significant therapeutic target for the treatment of type-2 diabetes as well as obesity (Dyson et al., 2005). Schottelndreier et al. (2001) have described a regulatory role of integrin alpha 6 (gene X53586) in Ca2+signalling that is known to have a significant role in insulin resistance (Kulkarni et al., 2004). Csermely et al. (1993) reported that insulin mediates phosphorylation/dephosphorylation of nucleolar protein nucleolin (gene M60858) by simulating casein kinase II, and this may play a role in the simultaneous enhancement in RNA efflux from isolated, intact cell nuclei (Csermely et al., 1993).

For gene Z26491, the Homo sapiens gene for catechol o-methyltrans-fease (COMT) was found to be differently expressed and helpful for treatment in diabetic rat. Wang et al (2002) compared the activity of COMT in the livers of diabetic rats with that in normal rats; the results suggested the activity of COMT is lower in diabetic rats than in normal rats. Lal et al. (2000) examined the effect of nitecapone, an inhibitor of the dopamine-metabolizing enzyme COMT and a potent antioxidant, on functional and cellular determinants of renal function in rats with diabetes. The results suggested that the COMT inhibitory and antioxidant properties of nitecapone provide a protective therapy against the development of diabetic nephropathy. These works proved that gene Z26491 is related with diabetes or treatment. C-myc is an oncogene that codes for transcription factor Myc that along with other binding partners such as MAX plays an important role widely studied in various physiological processes including tumor growth in different cancers. Myc modulates the expression of hepatic genes and counteracts the obesity and insulin resistance induced by a high-fat diet in transgenic mice overexpressing c-myc in liver (Riu, et al., 2003). Max interactor protein, MXI1 (gene L07648) competes for MAX thus negatively regulates MYC function and may play a role in insulin resistance. In the presence of glucose or glucose and insulin, lecucine is utilized more efficiently as a precursor for lipid biosynthesis by adipose tissue. It has been shown that during the differentiation of 3T3-L1 fibroblasts to adipocytes, the rate of lipid biosynthesis from leucine increase at least 30-fold and the specific activity of 3-hydroxy-3-methylglutaryl-CoA lyase (gene L07033), the mitochondrial enzyme catalysing the terminal reaction in the leucine degradation pathway, increases 4-fold during differentiation (Frerman et al., 1983). HCGV gene product (gene X81003) is known to inhibit the activity of protein phosphatise-1, which is involved in diverse signalling pathways including insulin signalling (Zhang et al., 1998).

In summary, from Table 3.2 we see, for the result obtained by the FM-test, genes U49573, M60858, L07648, L07033, X53586 and X81003 are associated-disease genes. For the result obtained by type-2 FM-test, genes U49573, X53586, M60858, Z26491, L07648, L07033 and X81003 are confirmed to be associated with diabetes disease. One more gene than FM-test is identified. Gene X57959, D85181, M95610 and U06452 are recommended by Liang et al. (2006) as candidate genes which are associated with diabetes disease. Here we recommend U61734 as a candidate gene for the future research in this field.

### 3.4.2 Analysis of lung cancer data

The lung cancer dataset contains 22283 genes and is downloaded from Wachi (2005). For each gene, there are 10 expression values. The first five values are from squamous lung cancer biopsy specimens and the others are from paired normal specimens. We also use type-2 FM-test and FM-test on this dataset and then make a comparison.

The results are shown in Table 3.4. From the table we see that the results obtained by the two methods are very different. The bold letters are names of genes which are associated with lung cancer disease.  For the result obtained by type-2 FM-test, 7 genes in ten best ranked are identified. For the result obtained by traditional FM-test, 8 genes are identified. However, when we applied the FM-test on lung cancer data, there are more than 80 genes having the same FM d-values; they are all equal to one, which makes it difficult to rank and distinguish disease associated genes from others. We have to choose the overexpressed genes from these 80 genes for analysis, which made the task more complicated, and it may miss some important genes. The reason is that the gene expression values in lung cancer microarray data are very close to each other, and the original data is noisy. Table 3.3 gives some example genes in this data set. These reasons imply the dataset contain more uncertainty information and the traditional fuzzy set does not seem to be able to deal with these factors suitably.

Table 3.3 The gene expression values of lung cancer data under two conditions

| Gene | Normal | | | | | Squamous lung cancer | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9.185 | 9.618 | 9.369 | 9.61 | 9.372 | 10.529 | 10.343 | 10.484 | 10.934 | 11.332 |
| 2 | 6.282 | 6.389 | 6.402 | 6.395 | 6.34 | 6.803 | 6.717 | 6.616 | 6.6 | 7.067 |
| 3 | 6.508 | 6.48 | 6.587 | 6.658 | 6.799 | 6.514 | 6.427 | 6.557 | 6.486 | 6.436 |
| 4 | 8.945 | 9.004 | 9.145 | 9.032 | 8.719 | 8.898 | 9.017 | 9.017 | 8.791 | 8.725 |
| 5 | 3.974 | 4.142 | 4.296 | 4.043 | 4.043 | 4.007 | 4.157 | 4.294 | 4.068 | 4.082 |

As shown in Table 3.4, 8 genes in ten overexpressed genes are identified as the disease associated genes. Cytokeratines are a polygenic family of insoluble proteins and have been proposed as potentially useful markers of differentiation in various malignancies including lung cancers (Camilo et al., 2006). Dystonin (DST/BPAG1) is a member of plakin protein family of adhesion junction plaque proteins. A recent study showed the expression of BPAG1 in epithelial tumor cells (Schuetz et al., 2006). Maspin (SERPINB5) was has been shown to be involved in both tumor growth and metastasis such as cell invasion, angiogenesis, and more recently apoptosis (Chen and Yates, 2006). Tumor protein p73-like (TP73L/P63) is implicated in the activation of cell survival and antiapoptotic genes (Sbisa et al., 2006) and has been used as a marker for lung cancer. It has been suggested that the p63 genomic amplification has an early role in lung tumorigenesis (Massion et al., 2003). CLCA2 belongs to calcium sensitive chloride conductance protein family and has been used in a multi-gene detection assay for Non Small Cell Lung Cancer (NSCLC) (Hayes et al., 2006). Plakophilins (PKPs) are members of the armadillo multigene family that function in cell adhesion and signal transduction, and also play a central role in tumorigenesis (Schwarz et al., 2006). Desmoplakin (DSP) is a desmosomeprotein that anchors intermediate filaments to desmosomal plaques. Microscopic analysis with fluorescencelabeled antibodies for DSP revealed high expression of membrane DSP in Squamous cell Carcinomas (SCC) (Young et al., 2002). The data analysis also identified cell cycle regulatory proteins such as CDC20 and Cyclin B1. Overexpression of CDC20 has been shown to be associated with premature anaphase promotion, resulting in mitotic abnormalities in oral SCC cell lines (Mondal et al., 2006). Mini chromosome maintenance2 (MCM2) protein is involved in the initiation of DNA replication and is marker for proliferating cells (Chatrath et al., 2003). Here in Liang et al. (2006)'s conclusion, gene NM_023915 and NM_019093 are suggested as potential candidates for biological investigation (Liang et al., 2006).

Table 3.4 Ten best-ranked genes related with lung cancer.

| Type-2 FM-test | | |
|---|---|---|
| Probe Set | Gene Description | T2 d-value |
| **NM_002405** | MFNG: MFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase | 0.7435 |
| **NM_001335** | CTSW: cathepsin W | 0.7285 |
| NM_017761 | PNRC2: praline-rich nuclear receptor coactivator 2 | 0.7266 |
| AV728526 | DTX4: deltex homolog 4 (Drosophila) | 0.7265 |
| **NM_0002694** | ALDH3B1: aldehyde dehydrogenase 3 family, member B1 | 0.7259 |
| **NM_024830** | LPCAT1: lysophosphatidylcholine acyltransferase 1 | 0.7243 |
| BE789881 | RAB31: member RAS oncogene family | 0.7204 |
| **AA888858** | PDE3B: phosphodiesterase 3B, cGMP-inhibited | 0.7194 |
| **NM_006079** | CITED2: cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain,2 | 0.7186 |
| **AF026219** | DLC1:deleted in liver cancer 1 | 0.7145 |
| FM-test (Overexpressed) | | |
| Probe Set | Gene Description | FM d-value |
| **NM_173086** | KRT6E: Keratin 6E | 1 |
| **NM_001723** | DST: Dystonin | 1 |
| **NM_002639** | SERPINB5: Serpin peptidase inhibitor, clade B (ovalbumin), member 5 | 1 |
| **AB010153** | TP73L: Tumor protein p73 like | 1 |
| NM_023915 | GPR87: G protein-coupled receptor 87 | 1 |
| **NM_006536** | CLCA2: Chloride channel, calcium activated, family member 2 | 1 |
| **NM_001005337** | PKPI: Plakophilin 1 ( ectodermal dysplasia/skin fragility syndrome) | 1 |
| **AF043977** | CLCA2: Chloride channel, calcium activated, family member 2 | 1 |
| **NM_004415** | DSP: Desmoplakin | 1 |
| NM_019093 | UGTIA9: UDP glucuronosyltransferase ! family, polypeptide A9 | 1 |

7 genes in ten are identified by type-2 FM-test.  MFNG is a member of the fringe gene family which also includes radical and lunatic fringe genes. They all encode evolutionarily conserved secreted proteins that act in the Notch receptor pathway. The activity of fringe proteins can alter Notch signaling (Gene bank). Activation of the Notch 1 signaling pathway can impair small cell lung cancer viability (Platta et al., 2008). The protein encoded by CTSW is found associated with the membrane inside the endoplasmic reticulum of natural killer (NK) (Gene Bank). NK cells play a major role in the rejection of tumors and cells infected by viruses (Oldham et al., 1983). ALDH3B1 is highly expressed in kidney and lung (Gene Bank). Marchitti et al. (2010) found ALDH3B1 expression was upregulated in a high percentage of human tumors; particularly in lung cancer cell the value is highest. Increasing ALDH3B1 expression in tumor cells may confirm a growth advantage or be the result of an induction mechanism mediated by increasing oxidative stress (Marchitti et al., 2010). LPCAT1 activity is required to achieve the levels of SatPC essential for the transition to air breathing (Bridges et al., 2010) and it is also upregulated in cancerous lung (Mansilla et al., 2009). Gene PDE3B was mentioned in (Lo et al., 2008) as the most significantly amplified gene in the tumors. CITED2 is required for fetal lung maturation (Xu et al., 2008). Researchers found CITED2 was highly expressed in lung cancer but not in normal tissues, which demonstrates that CITED2 plays a key role in lung cancer progression (Chou et al., 2010). Gene DLC1 encodes protein deleted in liver cancer (Liang et al., 2006). This gene is deleted in the primary tumor of hepatocellular carcinoma. It maps to 8p22-p21.3, a region frequently deleted in solid tumors. It is suggested that this gene is a tumor suppressor gene for human liver cancer, as well as for prostate, lung, colorectal and breast cancers (Gene Bank). Our analysis also identified NM_017761, AV_728526, BE789881. Here we suggest these genes as potential candidates in this field.

## 3.5 Conclusion

Fuzzy approaches have been taken into consideration to analyse DNA microarrays. Liang et al. (2006) proposed a fuzzy set theory based approach, namely a fuzzy membership test (FM-test), for disease genes identification and obtained better results by applying their approach on diabetes and lung cancer microarrays. However, some limitations still exist. The most obvious limitation is when the values of gene microarray data are very similar and lack over-expression, in which case the FM-d values are very close or even equal to each other. That makes the FM-test inadequate in distinguishing disease genes.

To overcome these problems, we introduced type-2 fuzzy set theory into the research of disease-associated gene identification. Type-2 fuzzy sets can control the uncertainty information more effectively than conventional type-1 fuzzy sets because the membership functions of type-2 fuzzy sets are three-dimensional. In this chapter we established the type-2 fuzzy membership function for identification of disease-associated genes on microarray data of patients and normal people. We call it type-2 fuzzy membership test (type-2 FM-test) and applied it to diabetes and lung cancer data. For the ten best-ranked genes of diabetes identified by the type-2 FM-test, 7 of them have been confirmed as diabetes associated genes according to genes description information in Genebank and the published literature. One more gene than original approaches is identified. Within the 10 best ranked genes identified in lung cancer data, 7 of them are confirmed by the literature which is associated with lung cancer treatment. The type-2 FM-d values are significantly different, which makes the identifications more reasonable and convincing than the original FM-test.

# Chapter 4

# Identification of protein complexes in PPI networks based on fuzzy relationship and graph model

## 4.1 Introduction

Protein-protein interactions are fundamental to the biological processes within a cell. Beyond individual interactions, there is a lot more systematic information contained in protein interaction networks. Complex formation is one of the typical patterns in the network and many cellular functions are performed by these protein complexes (Qi, 2008). Identification of protein complexes from the PPI network is useful for better understanding the principles of cellular organisation and unveiling their functional and evolutionary mechanisms (Li et al., 2010).

In general, a protein interaction network is represented by an undirected and unweighted network $G(V, E)$, where proteins are vertices and interactions are edges in the network. On the assumption that members in the same protein complex strongly bind to each other, a protein complex can be considered as a connected sub-network with in a protein interaction network. Many sub-network clustering algorithms have been proposed in recent years. Generally, these methods can be categorised into three groups: partitional clustering (King et al., 2004), hierarchical clustering (Girvan and Newman, 2002; Newman, 2004; Cho et al., 2007) and density-based clustering (Sprin and Mirny, 2003; Palla et al., 2005; Adamcsek et al., 2006; Zotenko et al., 2006, Guldener et al., 2005).

Density-based clustering methods are widely used in this field. This approach detects densely connected sub-graphs from a network. For a sub-network with n vertices and m edges, the density is measured with $d = 2m/(n(n\text{-}1))$. An extreme example is to identify all fully connected sub-networks of $d = 1$ (Spirin and Mirny, 2003). The most popular density-based clustering method is the Clique Percolation Method (CPM) proposed by Palla et al. (2005) for detection of overlapping protein complexes as k-clique percolation clusters. A k-clique is a complete sub-network of size k. Based on CPM, a powerful tool named CFinder for identifying overlapping protein complexes has been developed by Adamcsek et al. (2006). In general, less protein complexes can be identified for larger values of k. The authors of CPM suggest using the values of k between 4 and 6 to analyse PPI networks. However, mining fully connected sub-network is too restrictive to be useful in real biological networks. There are many other topological structures that may represent a complex on a PPI network, for example, the star shape, the linear shape, and the hybrid shape. In Figure 4.1 we show some examples of real complexes with different topologies.

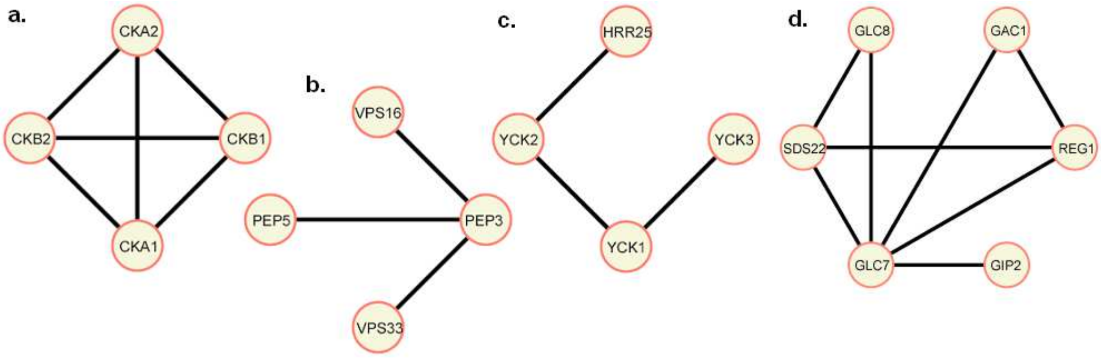


Figure 4.1 Projection of selected yeast MIPS complexes. This figure is taken from Qi (2008). a. Example of a clique. All nodes are connected by edges. b. Example of a star-shape, also referred to as the spoke mode. c. Example of a linear shape. d. Example of a hybrid shape where small cliques are connected by a common node.

Therefore, if we just identify the fully connected sub-networks, we will miss lots of protein complexes with the shape described in Figure 4.1 and the amount of identified protein complexes will decrease. To overcome this problem, we combine the fuzzy relation clustering method with the graph model. Since the fuzzy set theory was proposed by Zadeh in 1965, fuzzy clustering has been applied in many fields (Zadeh, 2005; Baraldi et al., 1999; Borgelt, 2009). Fuzzy relation can effectively describe the uncertainty information between two objectives, like the concepts “similar” and “different” (Zadeh, 1965). Thus we establish a fuzzy relation model between every pair of nodes in the network and use the operations of fuzzy relation to obtain sub-networks. However, we cannot ignore the original structure of the network which contains important information for clustering analysis. That's why we consider the sub-networks obtained from fuzzy relation model as the skeleton and compute the interaction probability of each node to identify the overlapping and non-overlapping sub-networks. In these sub-networks, some protein complexes exist.

We applied the method on yeast PPI networks and compared with the clique percolation method. For the same data, we detected more protein complexes. We also applied our method on two social networks. The results showed our method work well for detecting sub-networks and give a reasonable understanding of these communities.

In the next section, we introduce the theoretical background needed for a description of the fuzzy relation combined graph model method detailed in Section 4.3. We will apply this method on two social networks and yeast PPI networks respectively in Section 4.4. The conclusion will be given in Section 4.5.

## 4.2 Theoretical background

### 4.2.1 Topological properties of PPI networks

The topology of a network concerns the relative connectivity of its nodes. Different topologies affect specific network properties. In bioinformatics, the topological structures have been analysed for the following reasons (Han et al., 2005).

1. The architectural features of molecular interaction networks within a cell are often reflected to a large degree in other complex systems as well, such as the Internet, World Wide Web or organizational networks. The unexpected similarity indicates that similar laws may govern most complex networks in nature. This enables the expertise from large and well-mapped non-biological systems to be utilized for characterizing the complicated inter-relationships that govern cellular functions (Barabasi et al., 2004).

2. Cellular function is a contextual attribute of complex interaction patterns between cellular constituents (Barabasi et al., 2004). The quantifiable tools of networks theory offer possibilities for providing insights into properties of the cell's organization, evolution and stability.

3. The relative positions of proteins within the interaction networks might indicate their functional importance. For instance, a positive correlation between biological essentiality and graphical connectivity has been demonstrated (Han et al., 2005), suggesting a relationship between topological centrality and functional essentiality.

Therefore it is important to describe the topological and dynamic properties of various biological networks in a quantifiable manner. The literature on topological analysis of real networks is vast; therefore in this chapter we just give a briefly discussion on the related concepts and properties. Comprehensive reviews can be found in (Han et al., 2005; Faloutsos et al., 1999; Chakrabarti et al., 2005; Virtanen et al., 2003). Here, we give an example of one part of the yeast PPI network in Figure 4.1 by which we can understand these concepts better.

**Definition 4.1** A graph (or network) is a ordered pair $G = (V, E)$, where
(i) $V = \{v_1, v_2, \ldots, v_n\}$, $V \neq \varnothing$ , is called the vertex or node set of $G$;
(ii) $E = \{e_1, e_2, \ldots, e_m\}$ is the edge set of $G$ in which $e_i = \{v_j, v_t\}$ or $<v_j, v_t>$ is the edge linking two nodes $v_j$ and $v_t$.

**Definition 4.2** If every edge in a graph $G$ is undirected, the graph $G$ is called an undirected graph; if every edge in a graph $G$ is directed, the graph $G$ is called a directed graph.

**Definition 4.3** The two nodes linked by one edge are called adjacent nodes; the edges linking the same node are called adjacent edges.

Networks are naturally represented in matrix form. A graph of *N* nodes is described by an *N*×*N* adjacency matrix *A* whose non-zero elements $a_{ij}$ indicate connections between nodes. For undirected networks, a non-diagonal element $a_{ij}$ of an adjacency matrix is equal to the number of edges between nodes *i* and *j*, and so the matrix is symmetric. In our method, adjacency matrix is used to calculate the similarity between two different nodes.

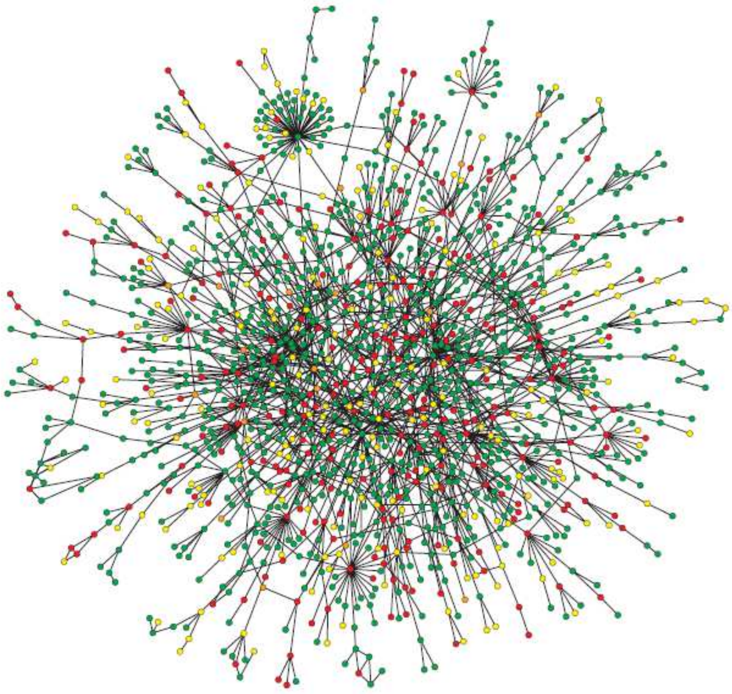


Figure 4.2 An example of protein-protein interactions network in yeast. This figure is obtained from (Han et al, 2005), included as background information only. It is a fully connected network and a few highly connected nodes (hubs) hold the network together.

**Definition 4.4** A simple graph is an undirected graph that has no loops and no more than one edge between any two different nodes. A connected graph is one in which there is at least one path connecting any two different nodes in the graph. A graph is a weighted graph if a weight is assigned to each edge.

**Definition 4.5** In the weighted graph, the shortest path is a path between two vertices such that the sum of weights of its constituent edges is minimized. In the unweighted graph, the shortest path is the minimum number of edges linked two vertices.

The shortest path can be considered as the distance between two vertices. For any positive integer k, the k-distance neighbourhood of *v* contains every vertex with a distance from v that is not greater than *k*. Thus, for *k* = 1, those vertices which are adjacent to *v* can be called the direct neighbors of *v*, denoted by $N_1(v)$. For $k > 1$, we can call these neighbors the indirect neighbors of *v*, denoted by $N_k(v)$ (Mete et al., 2009; Palla et al., 2005).

A real network may be a disconnected graph (the whole network can be divided into some connected sub-networks). If there is no path connecting two given vertices, then conventionally their distance is defined as infinite. The standard algorithms to find shortest paths such as Dijkstra's algorithm, or the breadth-first search method have been proposed in Cormen et al. (2001), Sedgewick (1988) and Ahuja et al. (1993).

**Definition 4.7** The network diameter *D* is defined as the maximum value of the lengths of all shortest paths between any two nodes in the network.

**Definition 4.8** The characteristic path length L is defined as the average of the lengths of all shortest paths in the network *G*, i.e.,

$$L = \frac{\sum_d df(d)}{\sum_d f(d)}, \tag{4.1}$$

where *f*(*d*) is the frequency of shortest paths with length *d*.

The characteristic path length describes the divergence of the nodes in the network, that is, how small the network is. A surprised finding in the study of complex networks is that the characteristic path length of many real complex networks is much smaller than expected. This is the so-called "small-world effect", which was originally observed in the research on social networks and is often characterized as the famous "six degrees of separation" (Chakrabarti, 2005). Figure 4.1 shows that cellular networks are different from social networks in terms of connections between hub nodes. In PPI networks, highly connected nodes avoid linking directly to each other and instead connect to proteins with only a few interactions, whereas in social networks, well-connected people tend to know each other (Han et al., 2005).

**Definition 4.9** The small-world property means that the characteristic path length *L* and the number of nodes *N* have the following relationship:

$$L \sim \log(N)\,. \tag{4.2}$$

**Definition 4.10** The degree $K_v$ of node *v* in a graph *G* is the number of edges that connect to it, i.e.,

$$K_v = |e(u,v)|, u, v \in V\,. \tag{4.3}$$

The degree distribution is the probability distribution of these degrees over the whole graph; it is independent of the size of the graph.

**Definition 4.11** The scale-free property means the degree distribution of a network has a power law (Newman and Watts, 1999)

$$p(k) \approx k^{-r}\,, \tag{4.4}$$

where $\gamma \in [2,3]$ for a common case and it is called power law exponent. The degree distribution appears linear when plotted on the log-log scale (Figure 4.3d). The significance of power law distributions has to do with their being heavy tailed, which means that they decay more slowly than the exponential or Gaussian distribution (referred to as random networks, Figure 4.3c). Thus, a power law degree distribution would be much more likely to have nodes with a very high degree than the other two distributions (Chakrabarti, 2005) (Figure 4.3). Many cellular interaction networks have been shown to be scale-free. Such a distribution indicates that most proteins in the network participate in only a few interactions, while a few proteins participate in many (hubs). Figure 4.2 shows a protein interaction map of the yeast as predicted by previous systematic two-hybrid screens. Most proteins participate in only a few interactions, and only a few participate in dozens: this is typical of scale-free network (Han et al., 2005; Stelzl et al., 2005).

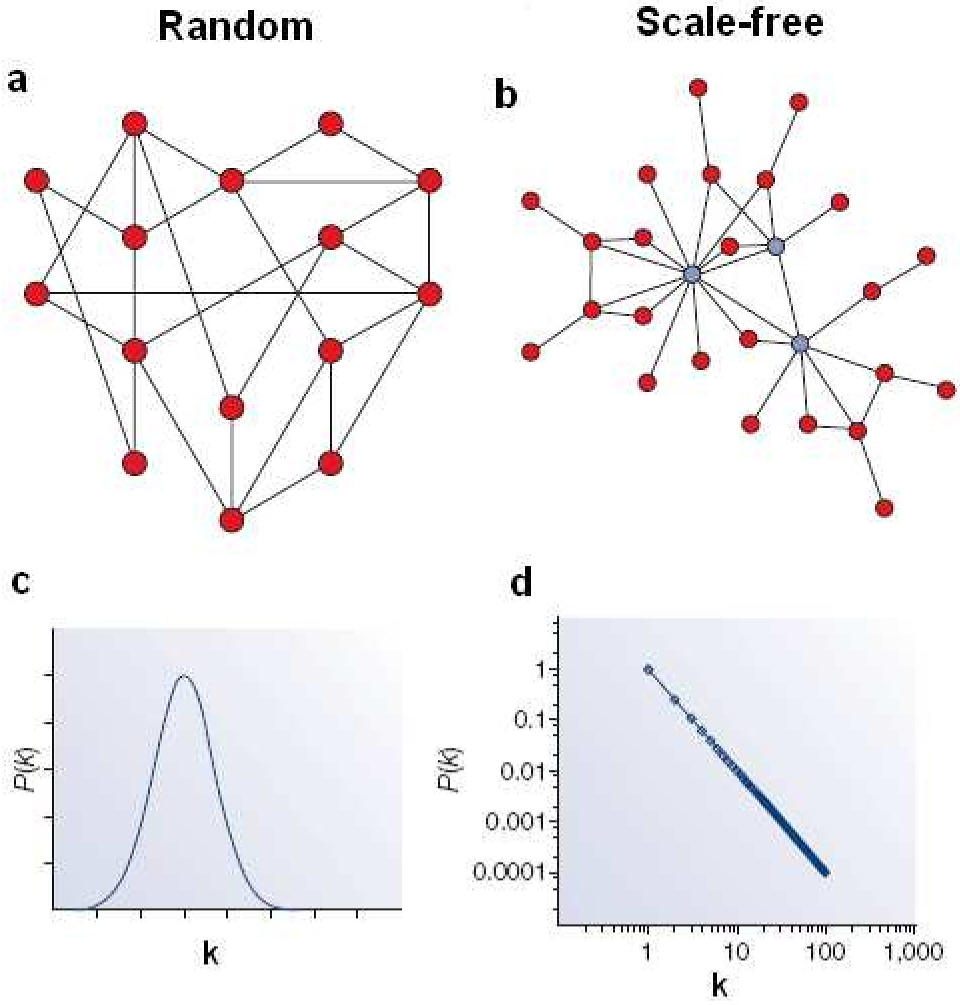


Figure 4.3: Degree distribution of random network versus scale-free network. The Figure is modified from Box 2 of (Han et al., 2005), included for background information only. (a) A schematic representation of a random network; (b) A schematic representation of a scale-free network. (c) The degree distribution of random network obeys a Gaussian distribution, (d) The degree distribution of scale-free network obeys a power-law distribution.

**Definition 4.12** Let the degree of node *v* be $K_v$ and the number of edges present among its $K_v$ adjacent nodes be $E_v$; then the clustering coefficient of *v* is

$$C_v = \frac{2E_v}{K_v(K_v - 1)} = \frac{E_v}{C_{K_v}^2} . \tag{4.5}$$

The clustering coefficient of a node quantifies how close its neighbours are. The clustering coefficient of a network is defined as the mean value of the clustering coefficients of all nodes in the network.

**Definition 4.13** The edge clustering coefficient (Radicchi et al., 2004) is defined as the number of triangles which really include this edge divided by the number of all triangles which possibly include this edge. Let $K_u$ and $K_v$ be the degrees of nodes u and *v* respectively. Then the clustering coefficient of the edge linking *u* and *v* is

$$C_{u,v}^{(3)} = \frac{Z_{u,v}^{(3)}}{\min\{K_u - 1, K_v - 1\}} , \tag{4.6}$$

where $Z_{u,v}^{(3)}$ means the number of triangles built on the edge. However, this definition is not feasible when the network has few triangles. Errors will occur when the number of possible triangles is zero. To avoid this limitation, Sun et al. (2011) modified the definition of edge clustering coefficients by calculating the common neighbours instead of the triangles. Thus a new definition of edge clustering coefficient is given:

$$C_{u,v} = \frac{|N_v \cap N_u| + 1}{\sqrt{N_v \bullet N_u}} , \tag{4.7}$$

where $N_v$ and $N_u$ represent the sets of neighbours of nodes *v* and *u* respectively. $C_{u,v}$ is a local variable; it quantifies how similar the two nodes *v* and *u* are connected by the edge $e_{u,v}$. If there is no edge between node v and u, then we consider $C_{u,v} = 0$. If *v* and $u$ are the same node, then we let $C_{u,v} = 1$. From the definition we can see that the larger the value is, the more similar the two nodes are. In our method we use $C_{u,v}$ to calculate the similarity value between two proteins in PPI networks and transfer the adjacent matrix of PPI networks into a similarity matrix. We then use the fuzzy relation method in clustering analysis to find the sub-networks, which is possibly protein complexes in PPI networks.

**Property**: The range of $C_{u,v}$ is [0,1].

Proof : (1) If there is no path between vertices $v$ and $u$ in the graph $G$, $C_{u,v} = 0$.

(2) If vertices $v$ and $u$ are the same node, then $C_{u,v} = 1$.

(3) If $v$ and $u$ are connected by an edge and $v \neq$ u, let $N_v = m$, $N_u = n, m \geq n \geq 1$.

Then $N(v)$ is the number of direct neighbours of $v$.

There are two extreme situations: $|N(v) \cap N(u)| = 0$, or $|N(v) \cap N(u)| = n - 1$.

(i) If$\in$ $|N(v) \cap N(u)| = 0$, then $\frac{|N_v \cap N_u| + 1}{\sqrt{N_v \bullet N_u}} = \frac{1}{\sqrt{mn}}$.

Since $m \geq n \geq 1$, we have $0 < C_{u,v} \leq 1$.

(ii) If $|N(v) \cap N(u)| = n - 1$, then, $\frac{|N_v \cap N_u| + 1}{\sqrt{N_v \bullet N_u}} = \frac{n}{\sqrt{mn}} = \sqrt{\frac{n}{m}} \leq 1$.

Therefore, $C_{u,v} \in [0, 1]$.

### 4.2.2 Fuzzy relation

Fuzzy relation is also proposed by Zadeh. In this chapter we will give some introduction on fuzzy relation theory. The letter '*R*' can denote not only a fuzzy relation, but also a fuzzy matrix based on the fuzzy relation.

***Concept of fuzzy relation***

**Definition 4.14** Let $U$ and $V$ be nonempty sets. A fuzzy relation $R \in F(U \times V)$ is a fuzzy set of the Cartesian product $U \times V$, $F(U \times V)$ is the set of all the fuzzy relations of $U \times V$ (Klir and Yuan, 1995).

$\forall\,(u, v) \in U \times V$, $R(u, v)$ can be interpreted as the grade of membership of the ordered pair $(u, v)$ in $R$. If $U = V$, then we can say that $R$ is a binary fuzzy relation in $U$. Here, we apply the binary fuzzy relation for identification of protein complexes. We give an example to explain the definition as follows.

**Example 1** Let $X = (-\infty, +\infty)$, the fuzzy relation concept $R$ "less than" can be defined as: $\forall x, y \in X$ ,

$$R(x, y) = \begin{cases} 0 & x \geq y, \\ [1 + \dfrac{100}{(y-x)^2}]^{-1} & x < y. \end{cases}, \tag{4.8}$$

then $R$ is a fuzzy relation on $X$, such as $R(0, 1) = 0.010$, $R(10, 20)=0.5$, $R(100, 400) = 0.990$.

**Example 2** Let $U = \{u_1, u_2, u_3, u_4\}$, $V=\{v_1, v_2, v_3\}$. For every pair $(u_i,v_j)$, if we have a membership value in Table 4.1, then the fuzzy relation $R$ between $U$ and $V$ is also determined.

Table 4.1 The membership values of fuzzy relation $R$ between $U$ and $V$

| | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | 0.7 | 0.5 | 0.3 |
| $x_2$ | 0.2 | 0.9 | 0 |
| $x_3$ | 0.4 | 0.6 | 0.8 |
| $x_4$ | 0 | 0.4 | 0.3 |

From the above table we see that a fuzzy relation $R$ can be expressed as a fuzzy matrix

$$R=(r_{ij})_{n\times m}\,,\ r_{ij}=R(u_i,v_j)\in[0,1]\,. \tag{4.9}$$

If $r_{ij}\in\{0,1\}$, then $R$ is an Boolean matrix and it is a classic relation. Therefore, the membership function maps a fuzzy relation to a fuzzy matrix, and a classic relation would be mapped to a Boolean matrix (Wolkenhauer, 2001).

***Operation of fuzzy relations***

Because a fuzzy relation $R$ is also a fuzzy set, thus it also can be performed as fuzzy set operations such as complement, intersection and union, and these operations can be put in fuzzy matrix form. Let $R=(r_{ij})_{n\times m}$, $S=(S_{ij})_{n\times m}$, $T=(t_{ij})_{n\times m}$ then we have the following operations:

Intersection $R\bigcap S=(r_{ij}\wedge s_{ij})_{n\times m}$;

Union $R\bigcup S=(r_{ij}\vee s_{ij})_{n\times m}$;

Complement $R^c=(1-r_{ij})_{n\times m}$;

$$R\bigcup R=R\,,R\bigcap R=R\,,(R^c)^c=R\,;$$
$$(R\bigcup S)^c=R^c\bigcap S^c\,,(R\bigcap S)^c=R^c\bigcup S^c\,;$$
$$R\bigcup S=S\bigcup R\,,R\bigcap S=S\bigcap R\,;$$
$$(R\bigcup S)\bigcup T=R\bigcup(S\bigcup T)\,;$$
$$(R\bigcap S)\bigcap T=R\bigcap(S\bigcap T)\,;$$
$$(R\bigcup S)\bigcap T=(R\bigcap T)\bigcup(S\bigcap T)\,,(R\bigcap S)\bigcup T=(R\bigcup T)\bigcap(S\bigcup T)\,.$$

**Property 4.1** $\forall R\in F(U\times V)$**,** we have

$$0\subseteq R\subseteq E\,,0\bigcup R=R\,,0\bigcap R=0\,;$$
$$E\bigcup R=E\,,\ E\bigcap R=R\,;$$

where

$$0=\begin{pmatrix} 0 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 0 \end{pmatrix} \text{ and } E=\begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 1 \end{pmatrix}.$$

They are called zero matrix and full matrix respectively. (See Dubois and Prade, 1980)

**Property 4.2** If $R \subseteq S \in F(U \times V)$, then we have

$$R \bigcup S = S \,, R \bigcap S = R \,, R^c \supseteq S^c \,.$$

(see Dubois and Prade, 1980)

**Property 4.3** If $R_1 \subseteq S_1, R_2 \subseteq S_2$, then

$$R_1 \bigcup R_2 \subseteq S_1 \bigcup S_2 \,,\; R_1 \bigcap R_2 \subseteq S_1 \bigcap S_2 \,.$$

Note: $R \bigcup R^c \neq E$, $R \bigcap R^c \neq 0$.

(see Dubois and Prade, 1980)

**Definition 4.15** Let $R = (r_{ij})_{n \times m}$, $\forall \lambda \in [0,1]$, we have $R_\lambda = (r_{ij}(\lambda))_{n \times m}$, where

$$r_{ij}(\lambda) = \begin{cases} 1, & r_{ij} \geq \lambda \\ 0, & r_{ij} < \lambda \end{cases}, \tag{4.10}$$

We call $R_\lambda$ the $\lambda$ cut matrix of *R*, and if

$$r_{ij}(\lambda)=\begin{cases}1, & r_{ij}>\lambda \\ 0, & r_{ij}\le\lambda\end{cases}, \tag{4.11}$$

then we call $R_\lambda$ the strong $\lambda$ cut matrix of *R*. If this chapter, we apply strong cut set to transfer a fuzzy matrix to a Boolean matrix for clustering sub-networks.

***Compositions of fuzzy relations***

**Definition 4.16** Let $R\in F(U\times V)$, $S\in F(V\times W)$. The composition between *R* and *S* is a new fuzzy relation from U to W, which is denoted by $R\circ S$, and its membership function is

$$(R\circ S)(u,w)=\underset{v\in V}{\vee}(R(u,v)\wedge S(v,w)),$$

when $R\in F(U\times U)$, we have $R^2=R\circ R$, $R^n=R^{n-1}\circ R$. We still use Example 1 to explain the definition. If R is a fuzzy relation "*x* is less than *y*", then the composition fuzzy relation $R\circ R$ means "*x* is far less than *y*". We need to obtain the membership function $(R\circ R)(x,u)$:

From definition, $\exists z$ which make $R(x,z)$ and $R(z,y)$ exist, and

$$(R\circ R)(x,y)=\underset{z}{\vee}(R(x,z)\vee R(z,y))=R(x,z_0).$$

Let *R*(*x*,*z*) = *R*(*z*,*y*), then we have $z_0=\frac{x+y}{2}$. The membership function therefore becomes

$$R\circ R(x,y)=\begin{cases}0 & x\le y \\ [1+\frac{100}{(\frac{x+y}{2})^2}]^{-1} & x>y\end{cases}, \tag{4.12}$$

If the domain is finite, then the composition of fuzzy relations can be expressed by product of fuzzy matrices.

**Definition 4.17** Let $Q=(q_{ik})_{m\times t}\in F(U\times V)$ , $R=(r_{ik})_{m\times t}\in F(V\times W)$ , then the composition of $Q$ and $R$ is

$$Q\circ R=S=(s_{ij})\in F(U\times W),$$

where $s_{ij}=\overset{t}{\underset{k=1}{\vee}}(q_{ik}\wedge r_{kj})$, $(i=1,2,...,m,\ j=1,2,...n)$.

From definition we see that the operation of product of fuzzy matrices is very similar that of traditional matrices. Here we give an example to show how to calculate the product of fuzzy matrices.

**Example 2**

$$Q=\begin{pmatrix}0.3 & 0.7 & 0.2\\ 1 & 0 & 0.9\end{pmatrix}\ ,\ R=\begin{pmatrix}0.8 & 0.3\\ 0.1 & 0.8\\ 0.5 & 0.6\end{pmatrix},$$

then $$Q\circ R=\begin{pmatrix}s_{11} & s_{12}\\ s_{21} & s_{22}\end{pmatrix},$$

where $s_{11}=(0.3\wedge 0.8)\vee(0.7\wedge 0.1)\vee(0.2\wedge 0.5)=0.3$;

$s_{12}=(0.3\wedge 0.3)\vee(0.7\wedge 0.8)\vee(0.2\wedge 0.6)=0.7$;

$s_{21}=(1\wedge 0.8)\vee(0\wedge 0.1)\vee(0.9\wedge 0.5)=0.8$;

$s_{22}=(1\wedge 0.3)\vee(0\wedge 0.8)\vee(0.9\wedge 0.6)=0.6$;

then $$Q\circ R=\begin{pmatrix}0.3 & 0.7\\ 0.8 & 0.6\end{pmatrix}.$$

The properties of composition of fuzzy relations are as follows:

**Properties 4.4** (1) $(Q \circ R) \circ S = Q \circ (R \circ S)$;

(2) $R^m \circ R^n = R^{m+n}$;

(3) $0 \circ R = R \circ 0 = 0, I \circ R = R \circ I = I$;

where, 0 is Zero Relation $\Leftrightarrow$ 0(u, v) = 0, I is Identical Relation $\Leftrightarrow$ $I(u,v) = \begin{cases} 1 & u = v \\ 0 & u \neq v \end{cases}$.

(4) $Q \subseteq R \Rightarrow Q \circ S \subseteq R \circ S$, $Q \subseteq R \Rightarrow Q^n \subseteq R^n$;

(5) $S \circ (Q \bigcup R) = (S \circ Q) \bigcup (S \circ R)$, $(Q \bigcup R) \circ S = (Q \circ S) \bigcup (R \circ S)$.

(see Dubois and Prade, 1980).

***Fuzzy equivalence relation***

Before doing clustering analysis based on fuzzy matrix, we have to make sure the fuzzy relation is a fuzzy equivalence relation. Here, we give the definition of fuzzy equivalence relation and fuzzy equivalence matrix.

**Definition 4.18** Let $R \in F(U \times U)$. R is a fuzzy equivalence relation if it satisfies the following conditions:

(1) Reflexivity: $\forall u \in U, R(u,u) = 1$;

(2) Symmetry: $\forall (u_i, u_j) \in U \times U, R(u_i, u_j) = R(u_j, u_i)$;

(3) Transitivity: $R \supseteq R^2$.

If *U* is finite, then the fuzzy relation *R* on *U* can be expressed by fuzzy matrix, which can be called a fuzzy equivalence matrix.

**Definition 4.19** A fuzzy matrix $R(r_{ij})_{n \times m}$ is a fuzzy equivalence matrix if it satisfies the following conditions:

(1) Reflexivity: $r_{ii} = 1$;

(2) Symmetry: $r_{ij} = r_{ji}$;

(3) Transitivity: $r_{ij} \ge \underset{k=1}{\vee} (r_{ik} \wedge r_{kj})$.

Because of reflexivity, $r_{ii} = 1$, then we have

$$\overset{n}{\underset{k=1}{\vee}} (r_{ik} \wedge r_{kj}) \ge r_{ii} \wedge r_{ij} = r_{ij} .$$

Because of transitivity, we have $R^2 = R$. If the fuzzy matrix just has transitivity, then it is called a transitive fuzzy matrix.

From definition we see that a fuzzy equivalence relation is a very stable relation. It won't change by the composition of itself. Therefore, based on this stable relation, we turn the fuzzy matrix to a Boolean matrix by $\lambda$-cut set and then perform clustering analysis. However, in practice, it is hard to obtain a fuzzy equivalence relation. Mostly, we just find fuzzy relations which satisfy reflexivity and symmetry; this kind of fuzzy matrices is called fuzzy similarity matrices. To turn a fuzzy similarity matrix to a fuzzy equivalence matrix, we need to compute its transitive closure.

**Definition 4.20** Let *R* be a fuzzy matrix. The smallest transitive fuzzy matrix of *R* is called the transitive closure of *R*, denoted by *t*(*R*). The transitive closure of *R*, *t*(*R*), should satisfy the following conditions:

(1) $t(R) \circ t(R) \subseteq t(R)$;

(2) $t(S) \supseteq S$;

(3) $S \supseteq R, S \circ S \subseteq S \Rightarrow S \supseteq t(R)$.

**Theorem 4.1** Let *R* be a fuzzy similarity matrix, then there is a smallest nature number *k* ($k \leq n$) such that $t(R) = R^k$. On the other hand, for any *l* greater than *k*, we always have $R^l = R^k$.

(see Klir and Yuan, 1995).

The above theorem suggests *t*(*R*) is a fuzzy equivalence relation and the fuzzy matrix based on it is a fuzzy equivalence matrix. We can transfer a fuzzy similarity matrix to a fuzzy equivalence matrix by computing the transitive closure *t*(*R*). For simplicity, we use the method of squares to compute *t*(*R*):

$$R \to R^2 \to R^4 \to \ldots \to R^{2k} \to \ldots,$$

If $R^i \circ R^i = R^i$, then $R^i$ is the transitive closure *t*(*R*).

Now we know that, to use fuzzy matrix to perform clustering, the fuzzy matrix should be a fuzzy equivalence matrix. In practice, mostly fuzzy matrices established are fuzzy similarity matrices, thus we need to compute its transitive closure by the method of squares. After we obtain its transitive closure, we need to transfer it to a Boolean matrix by computing its $\lambda$-cut matrix. Here we give more details about how to use fuzzy relation matrix to perform clustering analysis.

***Clustering method based on fuzzy equivalence matrix***

In fuzzy clustering analysis, the objects we need to analyse are called samples. To cluster them reasonably, we need to know the observation values of each sample. Suppose there are n samples,

$$X = \{x_1, x_2, \ldots x_n\},$$

each $x_i$ having m observation values, that is,

$$x_i = (x_{i1}, x_{i2}, \ldots x_{im}),$$

then the observation value matrix of samples is

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}.$$

where $x_{ij}$ means the *j*-th observation value of the *i*-th sample.

1. Data normalization.

   Because the order of magnitude in all observations may be different, the effect of observation values with large order of magnitude would be exaggerated, and the effect of observation values with small order of magnitude would be underestimated. This would make the clustering unreasonable. Thus to make the observation values in the same order of magnitude, we need the normalization step, usually we make the mean of observations be zero and variance be one by

$$x_{ij}^{'} = \frac{x_{ij} - \overline{x}_j}{\sigma_j},$$

where

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}.$$

2. Establishing fuzzy similarity matrix.

   After normalization of observation values, we can establish fuzzy similarity matrix via computing the similarity relation between any two samples. For $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jm})$, we compute the similarity value between them, which should satisfy $0 \le r_{ij} \le 1$, i, j=1, 2, …, n. Then we obtain a fuzzy similarity matrix R which shows the similarity between every pair sample:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}.$$

To compute the similarity between sample *i* and sample *j*, we use the following methods:

(i) Dot product

$$r_{ij} = \begin{cases} 1 & i = j \\ \dfrac{1}{M}\sum_{k=1}^{m} x_{ik} \bullet x_{jk} & i \neq j \end{cases}, \text{ where } M \geq \max_{i\cdot j}\Big(\sum_{k=1}^{m} x_{ik} \bullet x_{jk}\Big);$$

(ii) Correlation coefficient

$$r_{ij} = \frac{\sum_{k=1}^{m} \left|x_{ik} - \bar{x}_i\right|\left|x_{jk} - \bar{x}_j\right|}{\sqrt{\sum_{k=1}^{m}(x_{ik} - \bar{x}_i)^2} \bullet \sqrt{\sum_{k=1}^{m}(x_{jk} - \bar{x}_j)^2}}, \text{ where } \bar{x}_i = \frac{1}{m}\sum_{k=1}^{m} x_{ik}, \bar{x}_j = \frac{1}{m}\sum_{k=1}^{m} x_{jk};$$

(iii) Max-Min

$$r_{ij} = \frac{\sum_{k=1}^{n} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{n} \max(x_{ik}, x_{jk})};$$

(iv) Arithmetic mean minimum

$$r_{ij} = \frac{\sum_{k=1}^{n} \min(x_{ik}, x_{jk})}{\frac{1}{2}\sum_{k=1}^{n} \max(x_{ik} + x_{jk})};$$

(v) Geometric mean minimum

$$r_{ij} = \frac{\sum_{k=1}^{n} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{n} \sqrt{(x_{ik} \bullet x_{jk})}};$$

(vi) Absolute value index

$$r_{ij} = e^{-\sum_{k=1}^{n} |x_{ik} - x_{jk}|};$$

(vii) Absolute value subtractor

$$r_{ij} = \begin{cases} 1 & i = j \\ 1 - c\sum_{k=1}^{n} |x_{ik} - x_{jk}| & i \neq j \end{cases}, \text{ where } c \text{ can make } 0 \le r_{ij} \le 1.$$

In practice, we need to choose a proper method to compute the similarity value. We can also define a new method which is suitable for the clustering analysis in the problem. In our problem we apply the clustering coefficient defined by Sun et al. (2011) based on the interaction matrix of a network:

$$r_{ij} = \begin{cases} \dfrac{|N_i \cap N_j| + 1}{\sqrt{N_i \bullet N_j}}, & \text{if } (i, j) \in E, i \neq j \\ 0, & \text{if } (i, j) \notin E \\ 1, & \text{if } i = j \end{cases}, \tag{4.13}$$

where $N_i$ and $N_j$ are sets of neighbours of vertices $i$ and $j$ respectively. $|N_i \cap N_j|$ represents the number of joint neighbours of vertices $i$ and $j$. Note that vertices $i$ and $j$ are also connected by an edge. We proved that the clustering coefficient is in [0, 1] in Section 4.2.1.

3. Computing the transitive closure of the fuzzy similarity matrix via the method of squares.

4. Transforming the transitive closure to a Boolean matrix via computing the $\lambda$-cut matrix. The Boolean matrix is the skeleton of clustering result.

## 4.3 Method

### 4.3.1 The FRIPH method

In this part we introduce our method on identification of protein complexes. We combine fuzzy relation clustering analysis with IP value and hub structure in sub-networks, which we call the FRIPH method.

We can obtain the cluster skeleton of a PPI sub-network via the Boolean matrix transformed from the transitive closure of a fuzzy similarity matrix. Some protein complexes may be in these clusters. However, some protein complexes are overlapping on each other, which means each protein may be involved in multiple complexes. This is particularly true for protein interaction networks for most proteins having more than one biological function. For instance, there are 2750 proteins in the CYGD database (Guldener et al., 2005), however the amount of protein complexes is 8931. Thus, it is very significant to identify overlapping protein complexes. Li et al. (2010) proposed a new concept, Interaction Probability $IP_{vi}$ , to measure how strongly an outside vertex v connects to another sub-network which doesn't contain *v*. Interaction probability $IP_{vi}$ of any vertex v with respect to any sub-network *i* of size$|V_i|$ is defined as

$$IP_{vi} = \frac{|E_{vi}|}{|V_i|}, \tag{4.14}$$

where $|E_{vi}|$is the number of edges between the vertex *v* and the sub-network *i*. As shown in Figure 4.3 below, the $IP_{vi}$ of the vertex *v* to the sub-network *i* is 0.5.

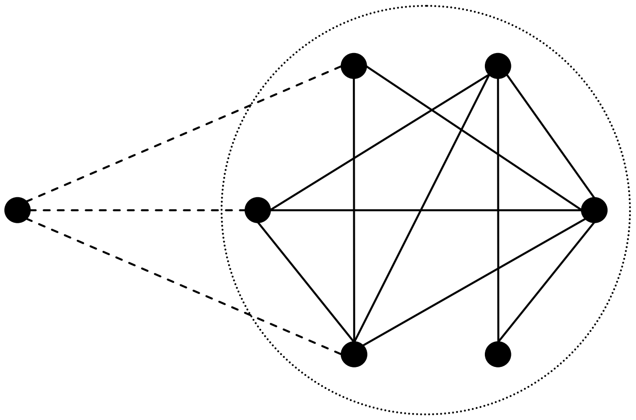


Figure 4.4 The interaction probability $IP_{vi}$ of a vertex *v* with respect to the sub-network *i* is 0.5

For every vertex *v* in the original PPI network, we calculate its $IP_{vi}$ in all sub-networks, *i* =1, 2, 3, …, m. Suppose vertex *v* is in sub-network *j*. If sub-network *i* has the greatest $IP_{vi}$ with vertex *v*, then *v* can be "added" to sub-network *i*, thus sub-network *i* will overlap with sub-network *j*. To summarize,

$$If\ IP_{vi} = \max_k (IP_{vk}), k = 1, 2, ..., m\text{, } then\ v\ is\ also\ in\ sub\text{-}network\ i.$$

However, sometimes vertex *v* has the same greatest *IP* value with several sub-networks. In this situation, we need to compare the nodes connected to vertex *v* in these sub-networks. If vertex *v* is connected with a hub in sub-network *i*, then *v* can be also in sub-network *i*. Figure 4.4 show a hub structure in PPI networks.

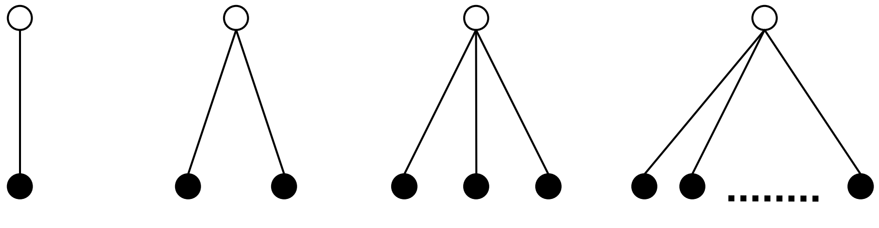


Figure 4.5 The hub structures in PPI networks.

The algorithm FRIPH can be divided into the following steps: 1. Generate an adjacency matrix from PPI data. 2. Choose a suitable method to compute similarity between each node in the network. 3. Compute transitive closure of the fuzzy matrix. 4. Transform the transitive closure to Boolean matrix via $\lambda$-cut matrix. 5. Compute IP values and compare hub structure in the original network to make sub-networks overlap. We give an example to explain our method. Sun et al. (2011) showed this example in their articles; here we use it to illustrate our method.

**Example 3** Consider a small network containing six nodes and 7 edges. Its adjacency matrix *A* is as follows:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

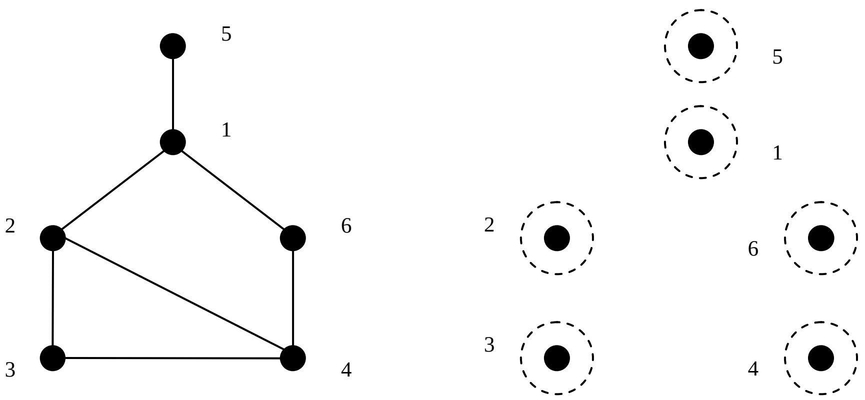
We use equation (4.13) to compute the similarity of each pair of nodes and obtain the fuzzy matrix R:

$$R = \begin{pmatrix} 1 & 0.33 & 0 & 0 & 0.58 & 0.41 \\ 0.33 & 1 & 0.82 & 0.67 & 0 & 0 \\ 0 & 0.82 & 1 & 0.82 & 0 & 0 \\ 0 & 0.67 & 0.82 & 1 & 0 & 0.41 \\ 0.58 & 0 & 0 & 0 & 1 & 0 \\ 0.41 & 0 & 0 & 0.41 & 0 & 1 \end{pmatrix}.$$

After obtaining the fuzzy matrix, we need to compute its transitive closure *t*(*R*):

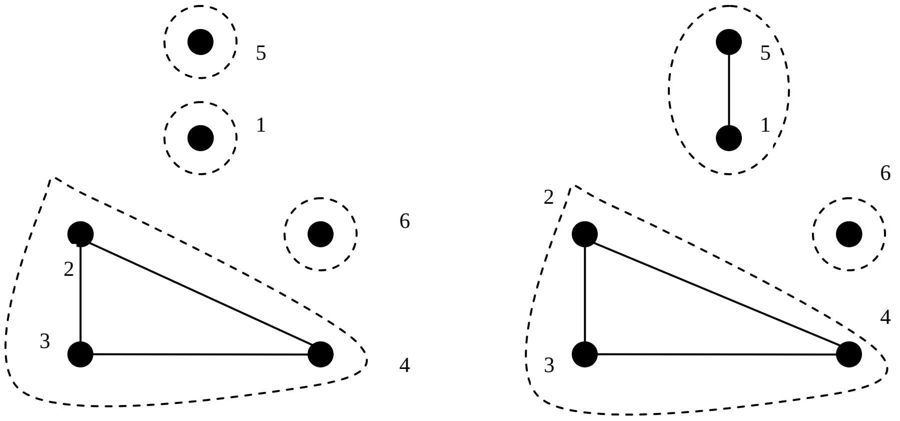
$$t(R) = \begin{pmatrix} 1 & 0.41 & 0.41 & 0.41 & 0.58 & 0.41 \\ 0.41 & 1 & 0.82 & 0.82 & 0.41 & 0.41 \\ 0.41 & 0.82 & 1 & 0.82 & 0.41 & 0.41 \\ 0.41 & 0.82 & 0.82 & 1 & 0.41 & 0.41 \\ 0.58 & 0.41 & 0.41 & 0.41 & 1 & 0.41 \\ 0.41 & 0.41 & 0.41 & 0.41 & 0.41 & 1 \end{pmatrix}.$$

According to the transitive closure matrix, we can choose a proper $\lambda$ cut set for clustering. In the matrix $t(R)$ there are four different values, 0.41, 0.58, 0.82, 1 respectively. For each value we can obtain one $\lambda$ cut set. We need to find out the best $\lambda$ cut set which is the skeleton of the overlapping sub-networks.



(a) (b)

$$A=\begin{pmatrix}0&1&0&0&1&1\\1&0&1&1&0&0\\0&1&0&1&0&0\\0&1&1&0&0&1\\1&0&0&0&0&0\\1&0&0&1&0&0\end{pmatrix}\Rightarrow\ t(R)_{\lambda\in(0.82,1]}=\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix}$$



(c) (d)

$$t(R)_{\lambda>0.58} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad t(R)_{\lambda>0.41} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Figure 4.6 Different $\lambda$ cut sets and clustering structure.

In Figure 4.6, (a) is the original graph of the network, and *A* is its adjacency matrix. (b) When $\lambda$ is in (0.82, 1], each node is clustered as one non-overlapping sub-network. This is an extreme situation. When we set $\lambda$ greater than 0.58, then we have 4 sub-networks. Nodes 2, 3, 4 become one bigger sub-network. If we set $\lambda$ greater than 0.41, then node 5 and node 1 become one sub-network, nodes 2, 3, 4 is one sub-network and node 6 is separated. This clustering structure can be considered as a skeleton of non-overlapping sub-networks.

After we obtain the skeleton of the non-overlapping sub-networks, we need to compute the IP value of each node with the other sub-network and check whether some nodes’ neighbour has hub structure in the original network. We would make these sub-networks overlapped via *IP* values and hub structure. From the original network of graph 4.6 (a) we see that node 6 is connected with node 1 and node 4, its *IP* value with sub-network nodes 5, 1 is 0.5 and the *IP* value with sub-network 2, 3, 4 is 0.33; thus node 6 can belong to sub-network node 5, 1. Then a new sub-network is generated consisting of nodes 5, 1, 6. Node 1 and node 2 are connected. Both of them have hub structure. Therefore, these two sub-networks can overlap on each other. In Figure 4.5 we show the details. We give a graphic diagram of the FRIPH method.

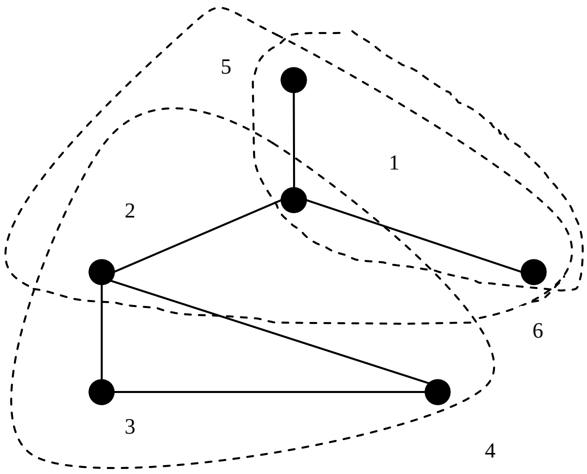


Figure 4.7 Overlapping sub-networks with respect to *IP* values and hub structure. According to the IP values and hub structure, the three separated sub-networks become three overlapping sub-networks.

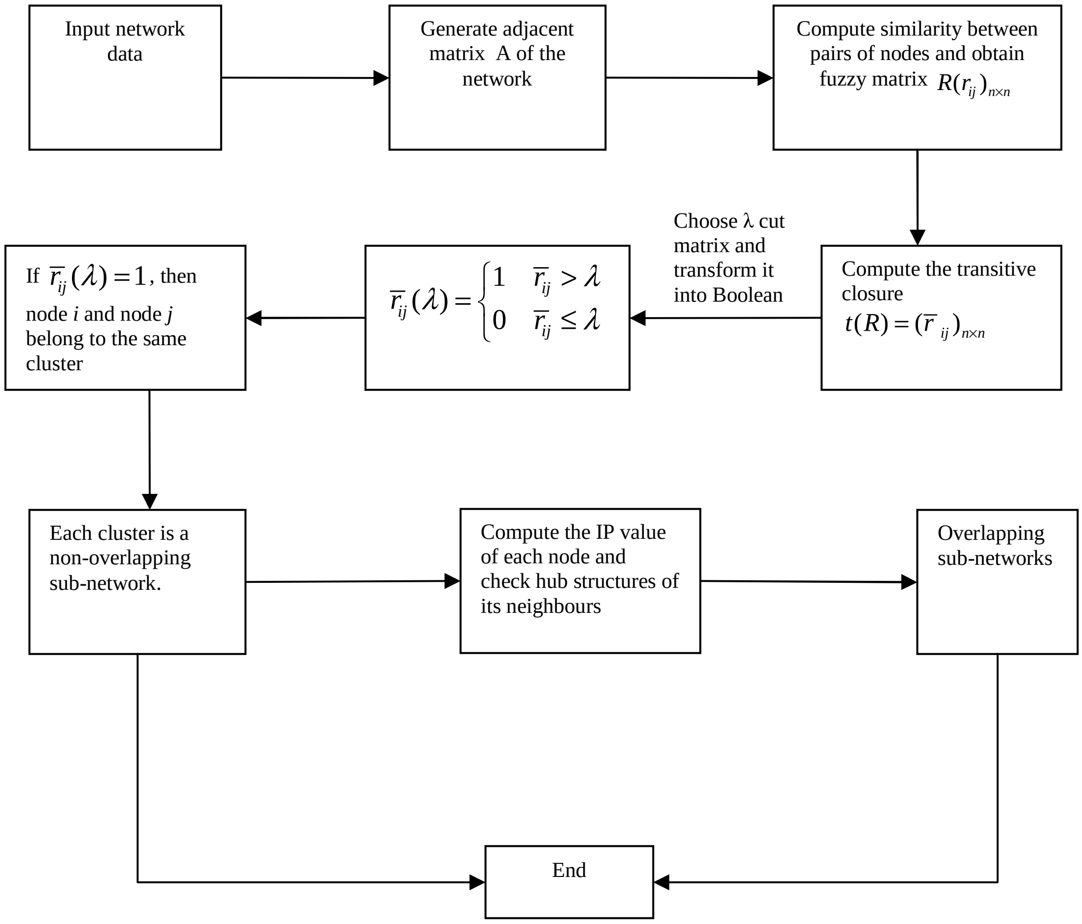

Input network data
Generate adjacent matrix A of the network
Compute similarity between pairs of nodes and obtain fuzzy matrix $R(r_{ij})_{n\times n}$
Compute the transitive closure $t(R) = (\overline{r}_{ij})_{n\times n}$
Choose λ cut matrix and transform it into Boolean
$\overline{r}_{ij}(\lambda) = \begin{cases} 1 & \overline{r}_{ij} > \lambda \\ 0 & \overline{r}_{ij} \le \lambda \end{cases}$
If $\overline{r}_{ij}(\lambda) = 1$, then node *i* and node *j* belong to the same cluster
Each cluster is a non-overlapping sub-network.
Compute the IP value of each node and check hub structures of its neighbours
Overlapping sub-networks
End

Figure 4.8 The graphic diagram of FRIPH

### 4.3.2 CFinder software

CFinder (the community cluster finding program) is one of the most popular software for protein complex identification. It uses the clique percolation method (CPM), which is proposed by Palla et al. (2005), to locate the k-clique communities of unweighted, undirected networks. Complete sub-graphs in a network are called *k*-cliques, where k refers to the number of nodes in the sub-graph, and a k-clique-community is defined as the union of all *k*-cliques that can be reached from each other through a series of adjacent k-cliques. Two k-cliques are said to be adjacent if they share *k*-1 nodes. The outline of the community finding algorithm is as follows:

(1) The k-clique community finding algorithm implemented in CFinder first extracts all such complete sub-graphs of the network that are not included in any larger complete sub-graph. These maximal complete sub-graphs are simply called cliques (the difference between k-cliques and cliques is that k-cliques can be subsets of larger complete sub-graphs).

(2) Once the cliques are located, the clique-clique overlap matrix is prepared. In this symmetric matrix each row (and column) represents a clique and the matrix elements are equal to the number of common nodes between the corresponding two cliques, while each diagonal entry is equal to the size of that clique.

(3) The k-clique-communities for a given value of k are equivalent to such connected clique components in which the neighbouring cliques are linked to each other by at least k-1 common nodes. These components can be found by erasing every off-diagonal entry smaller than k-1 and every diagonal element smaller than k in the matrix, replacing the remaining elements by one, and then carrying out a component analysis of this matrix. The resulting separate components will be equivalent to the different k-clique-communities.

In the next section, we will make a comparison between our method and CFinder.

## 4.4 Results and discussion

### 4.4.1 Application to two social networks

Firstly, we apply our method to two social networks. The first one is Zachary's karate club network. The second one is Network of American college football teams. We aim to identify the non-overlapping sub-networks in the two networks.

***Zachary's karate club network***

This is a widely used data as a test example for methods of identifying sub-networks in complex networks. In this data, there are 34 nodes representing 34 people. Zacahry observed them for more than 2 years. During this study, a disagreement developed between the administrator (node 34) of the club and the club's instructor (node 1), which ultimately resulted in the instructor's leaving and starting a new club, taking about a half of original club members with him. Zachary constructed the network between these members in the original club based on their friendship with each other and using a variety of measures to estimate the strength of ties between individuals. Figure 4.9 shows the graph of the network. There are 78 edges and two non-overlapping sub-networks in the graph, representing two groups of people with the administrator (circle label) and the instructor (square label). We apply our FRIPH to try to identify the two groups.

Following the step of FRIPH described in Figure 4.8, we separated the original networks into two sub-networks and two single nodes when we choose the value of $\lambda$ as 0.75. Figure 4.10 shows the result we obtain.

Comparing Figure 4.10 with the original network in Figure 4.9, the instructor group is perfectly separated from the original network. For the administrator group, node 10 and node 28 are not in the group but as two single points. The remaining nodes are all in administrator's group. Then we calculate the IP values of node 10 and node 28. For node 28 in Figure 4.9, it is connected with nodes 34, 24 and 25, which all belong to administrator's group; only node 3 belongs to instructor's group, thus the

IP value of node 28 in administrator's group is greater than that of node 28 in instructor's group. Node 28 should belong to administrator's group. For node 10, it is just connected with nodes 34 and 3. However, node 34 is the administrator which is the hub of that group. Therefore, node 10 also belongs to administrator's group. From the result of karate club data, the FRIPH method detects the two sub-networks correctly. However, the edges in the sub-networks are totally changed; these new edges have no meaning in the sub-network. But they have no effect on the correctness of groups of sub-networks.

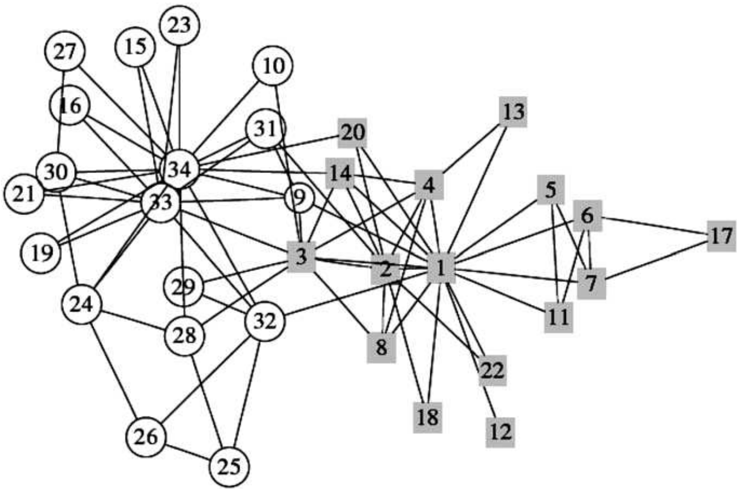


Figure 4.9. Zachary's karate club network. Square nodes and circle nodes represent the instructor's faction and the administrator's faction, respectively. This figure is from Newman and Girvan (2002).

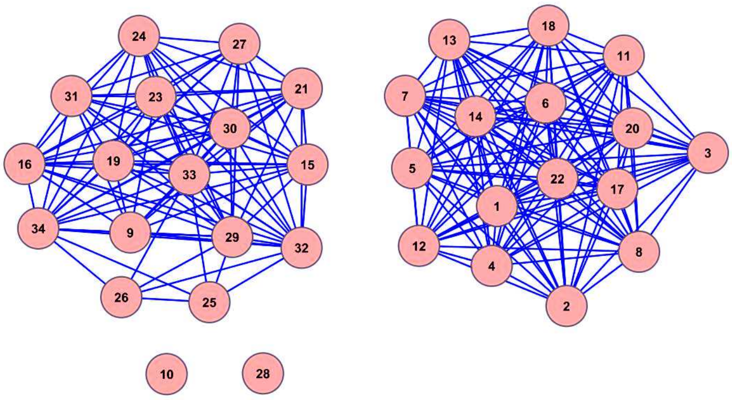


Figure 4.10 Sub-networks of  Zachary's karate club network, obtained by FRIPH

### *American college football teams network*

The second social network we test is the network of American college football teams which represents the game schedule of the 2000 season of Division I of the US college football league. In this data set, there are 115 nodes representing the teams and 613 edges presenting games played in the course of the year. The teams are divided into 12 conferences containing around 8-12 teams each. We apply our method on this data set and obtain the result showed in Figure 4.11. However, the result is not satisfactory.  We make a comparison with the result of Zhang et al. (2007) which is considered as a good one and shown in Figure 4.12. For our result most nodes in the last three sub-networks belong to the Sunbelt conference and should be in the same group of the grey points in Figure 4.12, but they divide into three sub-networks and group with members of the Western Athletic conference. This happens because the Sunbelt teams played nearly as many games against Western Athletic teams as they did against teams in their own conference (Girvana and Newman, 2002). Thus our method fails in this case. Meanwhile, there are 8 points which cannot be grouped in any sub-networks. In Figure 4.12, the same problem exists and these points are shown in red colour. That’s because these nodes generally connect evenly with more than one community, thus our method cannot group them into one specific sub-network correctly. These nodes are the “fuzzy” nodes which cannot be classified correctly by the current edge information. Generally, these points play a “bridge” role in two or more sub-networks of the original network.

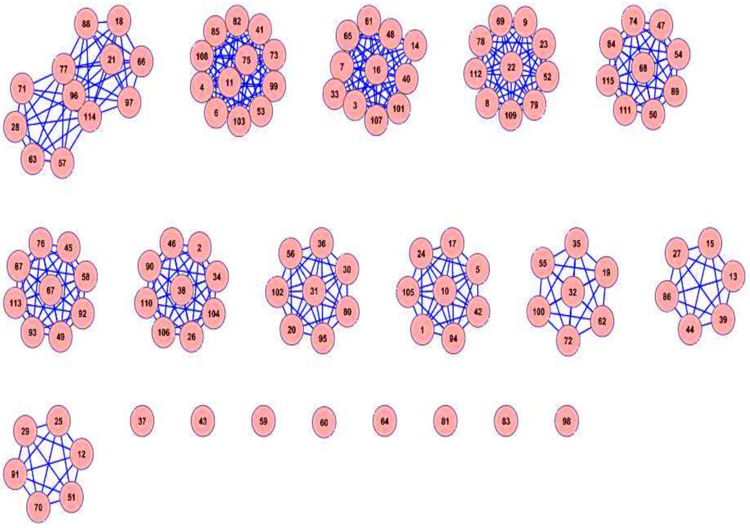


Figure 4.11 Sub-networks of American college football team network, obtained by FRIPH.

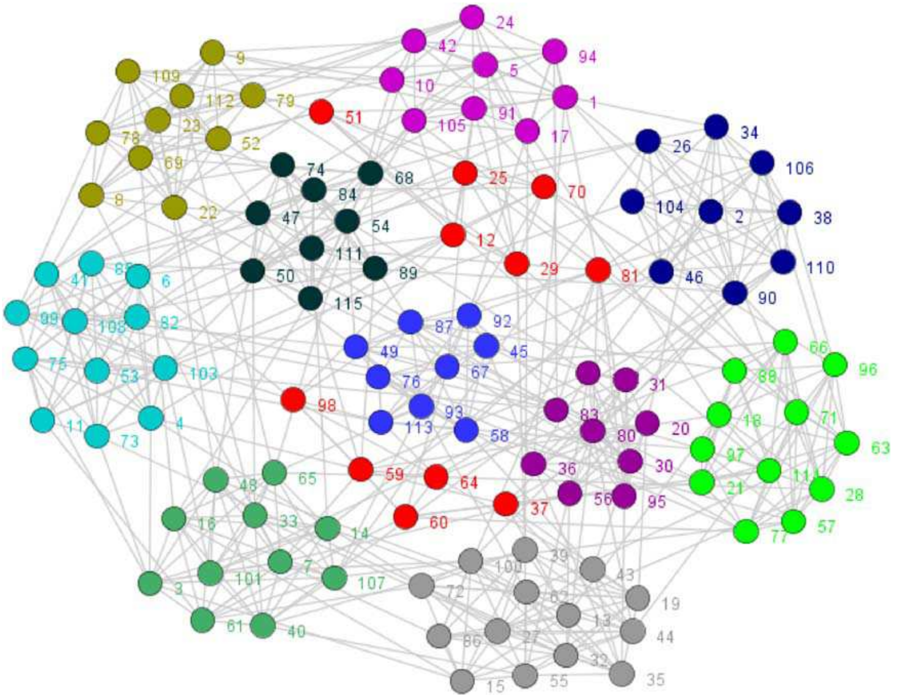


Figure 4.12 Sub-networks of American college football team network. This figure is taken from Zhang et al. (2007).

### 4.4.2 Identification of protein complexes

Identification of protein complexes from PPI network is crucial to understanding principles of cellular organisation and predicting protein functions. Cui et al. (2008) have shown that sub-networks such as cliques and near-cliques indeed represent functional modules or protein complexes. Thus identification of sub-networks from a complex network becomes an important issue. In this section, we apply our method on the protein interaction network of Saccharomyces cerevisiae, which was downloaded form the MIPS database (Mewes, 2006) and make a comparison with the popular software CFinder.

After removing all the self-connecting interactions and repeated interactions, the final network includes 4546 yeast proteins and 12319 interactions. The network diameter is 13 and the average shortest path length is 4.42. According to the annotate in MIPS database for Sacchromyces cerevisiae, there are 216 protein complexes identified by experiment, which consist of two or more proteins. The largest complex contains 81 proteins, the smallest complex just contains 2 proteins and the average size of all the complexes is 6.31.

To evaluate the effectiveness of FRIPH for identifying protein complexes, we compare the predicted clusters with known protein complexes in the MIPS database. There are 216 manually annotated complexes which consist of two or more proteins. We use the scoring scheme which is also applied in King et al. (2004), Altaf-UI-Amin et al. (2006), and Bader and Hogue (2003) to determine how effectively a Predicted Cluster (*Pc*) matches a Known Complex (*Kc*). The overlapping score between a predicted cluster and a known complex is calculated by the following formula:

$$OS(Pc, Kc) = \frac{i^2}{|V_{Pc}| \times |V_{Kc}|}, \tag{4.15}$$

where *i* is the number of nodes which are the intersection set of size of predicted cluster and known complex, $|V_{Pc}|$ is the size of predicted sub-network and $|V_{Kc}|$ is the size of known complex. If a known complex does not have the same protein in a predicted sub-network, then the overlapping score is 0, and if they perfectly match with each other, the overlapping score is 1. A known complex and a predicted cluster are considered as a match if their overlapping score is larger than a specific threshold. The number of matched known complexes with respect to different overlapping score threshold is shown in Figure 4.13 and Table 4.2

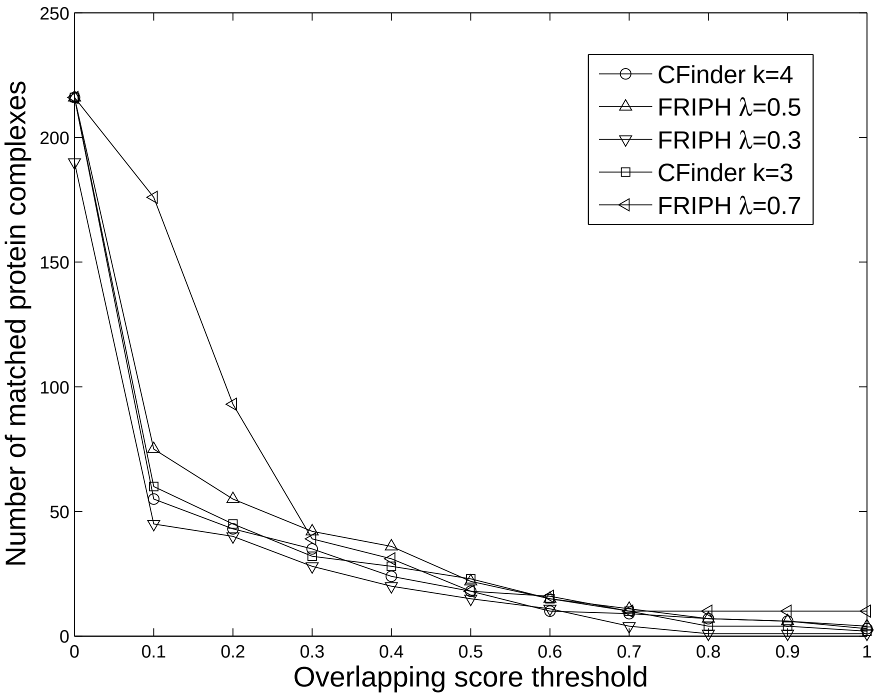


Figure 4.13. The number of known complexes matched by predicted sub-networks of FRIPH and CFinder with respect to different parameters and overlapping score.

As shown in Figure 4.13 and Table 4.2, CFinder obtains best matching when *k* = 3. The number of known complexes matched to the predicted sub-networks detected by CFinder using *k* = 3, 4, 5, 6 are 55, 43, 20 and 11 with respect to *OS*(*Pc*, *Kc*) = 0.2. The number of matched protein complexes decreases as k increases. In the work of Zhang et al. (2006) and Jonsson et al. (2006), this result was also deduced. That's because when k is determined, CPM just identifies the complexes which contain k or more proteins. For the FRIPH method, when λ=0, the PPI network doesn't change and all the nodes are in the same group. As λ increases, the number of matched complexes increases. When *λ* = 0.9, FRIPH obtains the best result and the number of matched complexes whose overlapping score is larger than 0.5 is stable. That is because when *λ* is increasing, the number of single proteins is increasing, thus the protein complexes with 2 or 3 proteins can be found much easier out of the original PPI network.

Table 4.2 The number of known complexes matched by predicted sub-networks of FRIPH and CFinder with respect to different parameters and overlapping score.

| Overlapping Score | CFinder (K=4) | CFinder (K=3) | FRIPH λ=0 | FRIPH λ=0.3 | FRIPH λ=0.5 | FRIPH λ=0.7 | FRIPH λ=0.9 |
|---|---|---|---|---|---|---|---|
| 0 | 216 | 216 | 1 | 190 | 216 | 216 | 216 |
| 0.1 | 55 | 75 | 1 | 45 | 60 | 176 | 216 |
| 0.2 | 43 | 55 | 0 | 40 | 45 | 93 | 104 |
| 0.3 | 35 | 42 | 0 | 28 | 32 | 39 | 40 |
| 0.4 | 24 | 36 | 0 | 20 | 28 | 31 | 28 |
| 0.5 | 18 | 22 | 0 | 15 | 23 | 18 | 20 |
| 0.6 | 10 | 15 | 0 | 11 | 15 | 16 | 20 |
| 0.7 | 9 | 11 | 0 | 4 | 10 | 10 | 20 |
| 0.8 | 7 | 7 | 0 | 1 | 4 | 10 | 20 |
| 0.9 | 6 | 6 | 0 | 1 | 4 | 10 | 20 |
| 1 | 3 | 4 | 0 | 1 | 2 | 10 | 20 |

In Figure 4.14, for a known complex of 10 proteins, the overlapping score obtained by FRIPH is 0.83. CFinder groups another 8 proteins which do not belong to the known complex and the overlapping score is 0.56 when *k* = 3. However, when *k* = 4, CFinder can identify the protein complexes perfectly.

In Figure 4.15, for a known complex of 14 proteins, the overlapping score obtained by FRIPH is 0.7. CFinder groups another 5 proteins which do not belong to the known complex and the overlapping score is 0.61 when *k* = 4. However, when *k* = 6, CFinder can produce a sub-network matching the known complexes with overlapping score 0.875.

As shown in Figure 4.16, for a known complex of 7 proteins, FRIPH identifies the sub-networks which perfectly match the protein complexes while CFinder groups two other proteins and miss one protein. These figures suggest FRIPH is more suitable for identifying sparse sub-networks which do not have too many edges than CFinder.

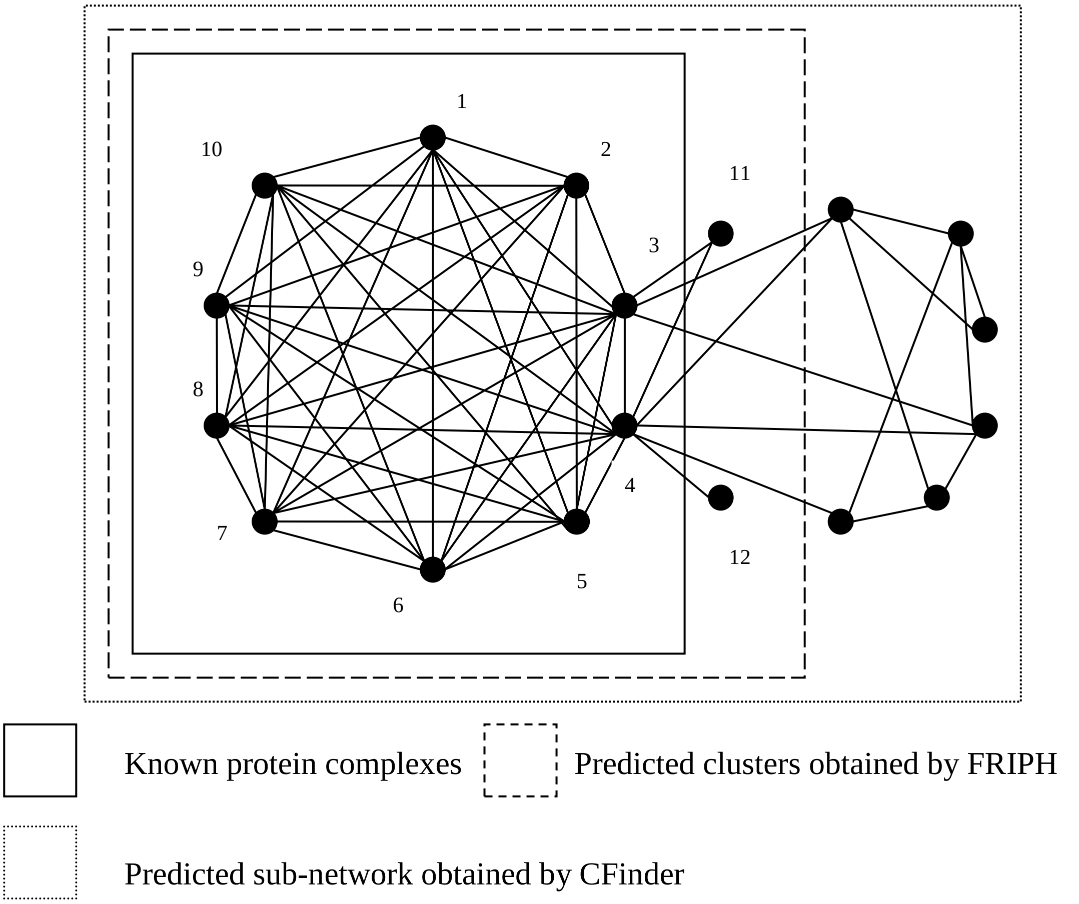
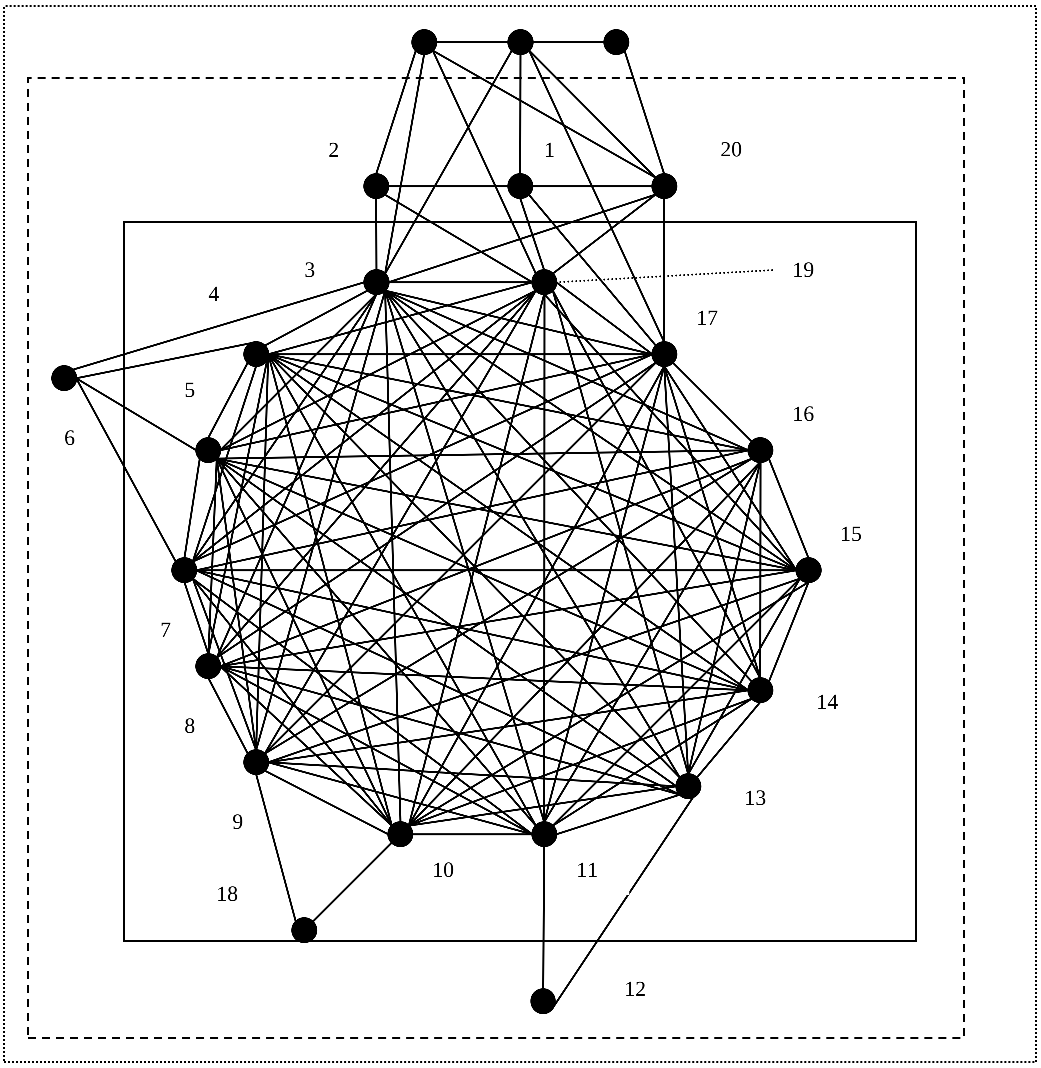

1
10
2
11
3
9
8
4
7
12
5
6
Known protein complexes
Predicted clusters obtained by FRIPH
Predicted sub-network obtained by CFinder

Figure 4.14 A known protein complex of 10 proteins and the matched sub-network generated by FRIPH and CFinder. The overlapping scores obtained by FRIPH and CFinder are 0.83 and 0.56, respectively.

The names of proteins are 1.YDR108w, 2.YOR115c, 3.YKR068c, 4.YML077w, 5.YDR472w, 6.YDR246w, 7.YMR218c, 8.YGR116w, 9.YBR254c, 10.YDR407c, 11. YIL004, 12.YLR078c.



□ Known protein complexes ⬚ Predicted clusters obtained by FRIPH

⬚ Predicted sub-network obtained by CFinder

Figure 4.15 A known protein complex of 14 proteins and the matched sub-network generated by FRIPH and CFinder. The overlapping scores obtained by FRIPH and CFinder are 0.7 and 0.61, respectively.

The names of proteins are 1.YDR290c, 2.YHR041c, 3. YOL148c, 4. YDR392w, 5.YDR448w, 6 YBR448w 7.YDR448w, 8. YDR176w, 9.YDR392w, 10.YOL112c, 11.YBR198c, 12.YGL066w, 13.YPL254w, 14. YDR145w, 15. YMR236w, 16. YHR099w, 17. YBR081c, 18.YCL010c, 19.YGR252c, 20.YNL235w.

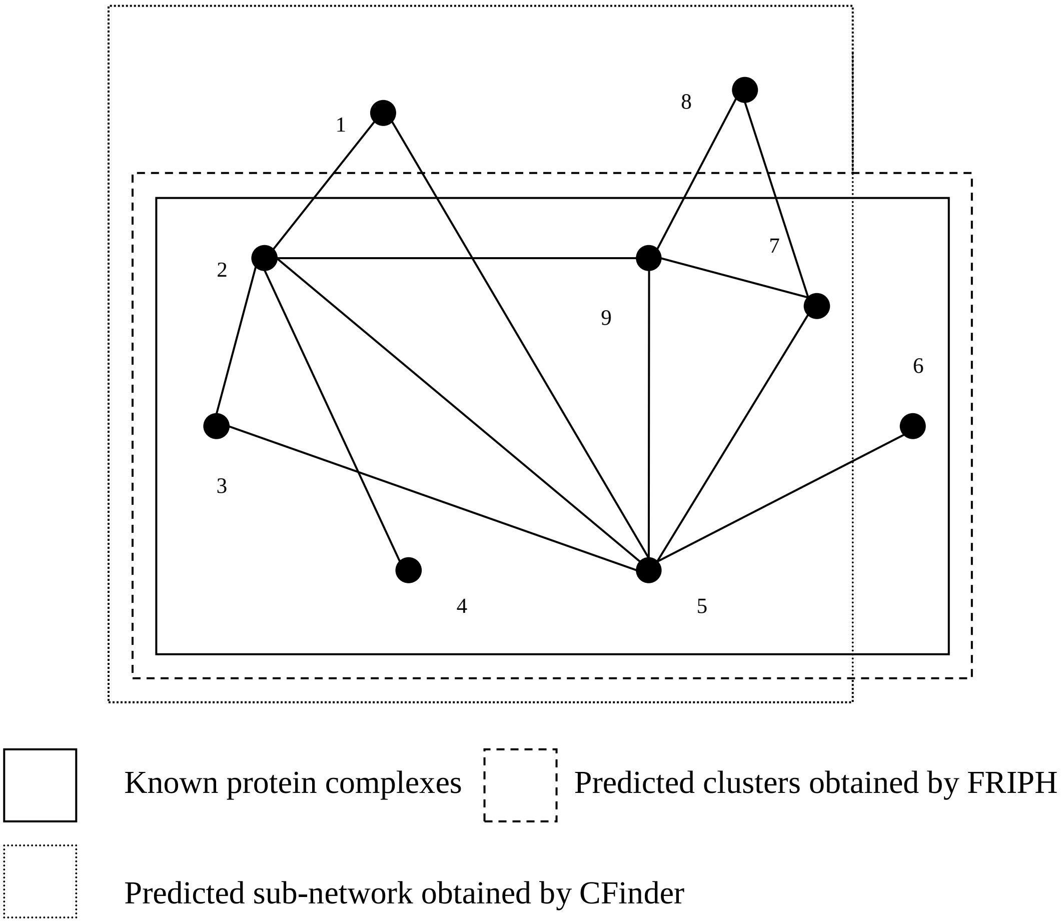

1
8
2
7
9
6
3
4
5
Known protein complexes
Predicted clusters obtained by FRIPH
Predicted sub-network obtained by CFinder

Figure 4.16 A known protein complex of seven proteins and the matched sub-network generated by FRIPH and CFinder. The sub-network generated by FRIPH perfectly matches the protein complexes.

The names of proteins are 1.YPR041w, 2.YMR309c, 3.YBR079c, 4.YNL244c, 5.YOR361c, 6.YNL062c, 7.YDR429c, 8.YPL106c.

***Recall-precision analysis***

Recall and precision are two important methods to estimate the performance of algorithms for identifying protein complexes. Recall is the fraction of the True-Positive (TP) predictions out of all the true predictions. Precision is the fraction of the true-positive prediction out of all the positive predictions. They are defied as follows:

$$recall = \frac{TP}{TP + FN},$$

$$precision = \frac{TP}{TP + FP},$$

where TP is the number of matched sub-networks and FN is the number of not matched known complexes. FP is the number of the remaining identified sub-networks. According to the assumption in Bader and Hogue (2003), a predicted sub-network and a known complex are considered to be matched if the overlapping score is larger than 0.2. Thus we also use 0.2 as the matched overlapping threshold. Table 2 compares recall and precision of the two methods. In Table 4.3, for FRIPH, the recall is increasing when the parameter value is increasing. That’s because when the parameter value increases, more and more nodes are separated from the original network and compose sub-networks from which protein complexes can be identified. The extreme case is when λ = 1, every node is considered as a sub-network. In this case, the complexes which just contain 2 or 3 proteins can be easily identified. Thus the amount of identified protein complexes increases as the parameter increases. However, when the number of sub-networks increases, the numbers of sub-networks which are not protein complexes also increases; that’s why the precision is decreasing as the parameter is increasing for FRIPH. On the contrary, for CFinder, as the parameter is increasing, the recall is decreasing and the precision is increasing. That is because the CPM algorithm aims to find k-cliques in the original network. The larger k is, the less cliques it will find out of the original PPI network. That’s why the authors of CPM suggest using the values of k between 4 and 6 to analyse PPI networks.

Table 4.3 The comparison of FRIPH and CFinder on recall and precision.

| Algorithm | Parameter | Recall | Precision |
|---|---|---|---|
| FRIPH | λ=0.3 | 18.5% | 21.05% |
| | λ =0.5 | 20.8% | 10.1% |
| | λ =0.7 | 43.5% | 9.26% |
| | λ =0.9 | 48.1% | 6.67% |
| CFinder | k=3 | 25.5% | 27.6% |
| | k=4 | 19.9% | 53.3% |
| | k=5 | 9.8% | 75.2% |
| | k=6 | 3.3% | 79.3% |

## 4.5 Conclusion

Identification of protein complexes is very important for better understanding the principles of cellular organisation and unveiling their functional and evolutionary mechanisms. Many methods are proposed for the identification of protein complexes. The Clique Percolation Method (CPM) is one of the most popular one. The CPM is a density-based method which aims to detect densely connected sub-networks (cliques) from a network. However, in real PPI network, it is not enough to just identify cliques because many protein complexes do not just have the clique shape, some have star shape, hybrid shape, or even linear shape. The software CFinder which is developed based on CPM is a powerful tool for identifying protein complexes, but it is very sensitive to the value of *k*.

In this chapter, we proposed a novel method which combines the fuzzy clustering method and interaction probability to identify the overlapping and non-overlapping community structures in PPI networks, then to detect protein complexes in these sub-networks. Our method is based on both the fuzzy relation model and the graph model. Fuzzy theory is suitable to describe the uncertainty information between two objects, such as ‘similarity’ and ‘differences’. On the other hand the original graph model contains significant clustering information, thus we do not ignore the original structure of the network, but combine it with the fuzzy relation model. We applied the method on yeast PPI networks and compared with CFinder. For the same data set, although the precision of matched protein complexes is lower than CFinder, we detected more protein complexes. We also applied our method on two social networks. The results showed that our method works well for detecting sub-networks and gives a reasonable understanding of these communities.

# Chapter 5

# Summary and future research

In this thesis we proposed several new fuzzy approaches to analyze DNA microarray data and protein-protein interaction networks. We focused on three research problems: (i) fuzzy clustering analysis on DNA microarrays, (ii) identification of disease-associated genes in microarrays, and (iii) identification of protein complexes on PPI networks.

## 5.1 Research conclusion

In Chapter 2, we addressed the problem of detecting, by the fuzzy c-means (FCM) method, clustering structures in DNA microarrays corrupted by noise. We introduced a more efficient method for clustering analysis of DNA microarrays which contain noise and uncertainty information.

Because of the presence of noise, some clustering structures found in random data may not have any biological significance. In this part, we combined the FCM with the empirical mode decomposition for clustering microarray data. Applied on yeast and serum microarrays, this combined method detected clearer clustering structures in denoised data, implying that genes have tighter association with their clusters. Furthermore we found that the estimation of the fuzzy parameter *m*, which is a difficult step, can be avoided to some extent by analysing denoised microarray data.

In Chapter 3 we approached the problem of identifying disease-associated genes from DNA microarray data which are generated under different conditions. Making comparison of these gene expression data can enhance our understanding of onset, development and progression of various diseases.

We developed a type-2 fuzzy membership (FM) function for identification of disease-associated genes. This approach was applied to diabetes and lung cancer data, and a comparison with the original FM test was carried out. Among the ten best-ranked genes of diabetes identified by the type-2 FM test, seven genes have been confirmed as diabetes-associated genes according to gene description information in Gene Bank and the published literature. An additional gene is further identified by our method. Among the ten best-ranked genes identified in lung cancer data, seven are confirmed that they are associated with lung cancer or its treatment. The type-2 FM-*d* values are significantly different, which makes the identifications more convincing than the original FM test.

In Chapter 4 we addressed the problem of identifying protein complexes in large interaction networks. Identification of protein complexes is crucial to understand the principles of cellular organisation and to predict protein functions.

In this part, we proposed a novel method which combines the fuzzy clustering method and interaction probability to identify the overlapping and non-overlapping community structures in PPI networks, then to detect protein complexes in these sub-networks. Our method is based on both the fuzzy relation model and the graph model. We applied the method on several PPI networks and compared with a popular protein complex identification method, the clique percolation method. For the same data, we detected more protein complexes. We also applied our method on two social networks. The results showed our method works well for detecting sub-networks and gives a reasonable understanding of these communities.

## 5.2 Possible future work

Fuzzy methods on clustering analysis of DNA microarrays is a worthy research problem. DNA microarray data contain noise and uncertainty information, and fuzzy methods are suitable for dealing with this problem. Many methods have been proposed over the past several decades, and the demand of understanding functions and groups of DNA requires more efficient methods. In our research, although we can reduce the influence of noise on the clustering results, the more times we denoise the microarray data, the more information in them we would miss. Thus, more efficient denoising methods are needed. On the other hand, FCM, which is a simple and efficient fuzzy method for clustering analysis, has been widely used in many fields. However, as a supervised clustering method, FCM still requires to determine the number of clusters firstly. We cannot apply FCM on the data that we don't know the clustering number. Therefore, developing an unsupervised fuzzy method is very significant for analysis of DNA microarrays. Type-2 FCM method has been proposed (Rhee, 2007). An application of this method to the unsupervised fuzzy problem would be promising.

There are many methods for identification of disease-associated genes. In Chapter 3, we proposed type-2 FM test method. However, the computation complexity of type reduction of type-2 fuzzy set is high. We have applied interval type-2 fuzzy set to this problem, but the interval type-2 fuzzy set may not properly describe the differences between expression values under two different conditions. Thus establishing a good membership function to compute the divergence of the two sets is an important step. Meanwhile, most methods are sensitive to different data sets. Thus it is necessary to devise a strategy to combine different methods to obtain the best result.

For the identification of protein complexes, although we identified more complexes than CFiner, the accuracy rate is low. Thus we need to improve the accuracy rate of FRIPH. Meanwhile, the edges in identified sub-networks are not the original edges. We need to connect nodes in sub-networks based on the original network. We also need to define the IP value in a different way, based not only on the relation between the nodes and other sub-networks, but also on the relation between the nodes and their neighbours. We also need to develop type-2 fuzzy relation membership function on the network to describe the similarity between pair of nodes in a network. Type-2 fuzzy relation method would be a useful approach for solving this problem.

# References

Adamcsek, B., Palla, G., Farkas, I., Derenyi, I. and Vicsek, T. (2006) CFinder: locating cliques and overlapping modules in biological networks, *Bioinformatics*, Vol. 22, No.8, pp. 1021-1023.

Aliev, R.A, Pedrycz, W., Guirimov, B.G., Aliev, R.R., IIHan, U., Babagil, M., Mammadli, S. (2011). Type-2 uzzy neural networks with fuzzy clustering and differential evolution optimization, *Information Sciences*, Vol. 181, Issue. 9, pp. 1591-1608.

Alon, U., Barkai, N., Notterman, D.A., Gish, G., Ybarra, S., Mack, D. and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *PNAS*, Vol. 96, pp. 6745-6750.

Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K. and Kanaya, S. (2006), Development and implementation of an algorithm for detection of protein complexes in large interaction networks, *BMC Bioinformatics*, Vol. 7, No. 207, pp.1-13

Asyali, M.H. and Alci, M. (2005). Reliabiltiy analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods, *Bioinformatics,* Vol. 21, Issue. 5, pp. 644-649.

Augenlicht, L.H. and Kobin, D. (1982). Cloning and screening of sequences expressed in a mouse colon tumor, *Cancer Research,* Vol. 42, Issued. 42, 1088-1093.

Augenlicht, L.H., Wahrman, M.Z., Halsey, H., Anderson, L., Taylor, J. and Lipkin, M. (1987). Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro, *Cancer Research*, Vol. 47, pp. 6017-6021.

Augenlicht, L.H., Taylor, J., Anderson, L. and Lipkin, M. (1991). Patterns of gene expression that characterize the colonic mucosa in patients at genetic risk for colonic cancer, *Proc Nati Acad Sci USA*, Vol. 88, Issue. 8, pp. 3286-3289.

Augustson, J.G., Minker, J. (1970). An analysis of some graph theoretical cluster techniques, *Journal of the Association for Computing Machinery*, Vol. 17, No. 4, pp. 571-588.

Avogadri, R. and Valentini, G. (2009). Fuzzy ensemble clustering based on random projections for DNA microarray data analysis, *Artificial Intelligence in Medicine*, Vol. 45, Issue. 2-3, pp. 173-183.

Ball, C.A., Sherlock, G., Parknson, H. (2002), Rocca-Sera, P., Brooksbank, C., Causton, H.C., Cavalieri, D., Gaasterland, T., Hingamp, P., Holstege, F., Ringwald, M., Spellman, P., Stoeckert, C.J.J., Stewart, J.E., Talyor, R, Brazma, A., Quackenbush, J. (2002). Microarray gene expression data: standards for micorarray data. *Science*, Vol, 298 :18.

Balaji, P.G. and Srinivasan, D. (2010). Type-2 fuzzy logic based urban traffic management, *Engineering applications of Artificial Intelligence*, Vol. 24, Issue. 1, pp. 12-22.

Barabasi, A., Oltvai, Z. (2004). Network biology: Understanding the cell’s functional organization, *Nat Rev Genet*, Vol. 5, pp. 101-113.

Baraldi, A., Blonda, P. (1999), A survey of fuzzy clustering algorithms for pattern recongnition, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 29, pp. 778-785.

Belacel, N., Wang, Q. and Cuperlovic-culf, M. (2006). Clustering methods for microarray gene expression data, *A Journal of Integrative Biology*, Vol.10, No.4, pp.507-532.

Bertone, P., Gerstein, M. and Snyder, M. (2005). Applications of DNA tilling arrays to experimental genome annotation and regulatory pathway discovery, *Chromosome Research*, Vol. 13, No. 3, pp. 259-274.

Bertoni, A. and Valentini, G. (2006). Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses, *Artificial intelligence in Medicine*, Vol. 37, Issue. 2, pp. 85-109.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Fucntion Algorithms,* New York: Plenum Press.

Borgelt, C. (2009). Accelerating fuzzy clustering, *Information Sciences*, Vol.179, No.23,pp.3985-3997.

Bowtell, D.D. (1999). Options available-from start to finish-for obtaining expression data by microarray. *Nature genetics,* Vol. 21, Suppl. 1, pp. 25-32.

Boutros, P. and Okey, A. (2005). Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data. *Briefings in Bioinformatics*, Vol. 6, No. 4, pp. 331-343.

Brachat, A., Pierrat, B., Xynos, A., Brecht, K., Simonen, M., Brungger, A. and Heim, J. (2002). A microarray-based, integrated approach to identify novel regulators of cancer drug response and apoptosis, *Oncogene*, Vol. 21, pp. 8341-8371

Bridges, J. P., Ikegami, M., Brilli, L. L., Chen, X., Mason, R. J., Shannon, J. M. (2010). LPCAT1 regulates surfactant phospholipid synthese and is required for trasitioning to air breathing in mice, *J. Clinical Investigation*, Vol. 120, 1736-1748.

Brown, C.S., Goodwin, P.C., Sorger, P.K. (2001). Image metrics in the stratistical analysis of DNA microarray data, *Proceeding of the National Academy of Sciences of the United States of America,* Vol. 98, Issue. 16, pp. 8944-8949.

Brunskill, E.W., Lai, H.L., Jamison, D.C., Potter, S.S and Patterson, L.T. (2011). Micorarrays and RNA-Seq identify molecular mechanisms driving the end of nephron production. *BMC Developmental Biology*, Vol. 11, pp. 15-27.

Cannataro, M., Guzzi, P.H., Veltri, P. (2010). IMPRECO: Distributed prediction of protein complexes, Future Generation Computer Systems-The International Journal of Grid Computing-Theory Methods and Applications, Vol. 26, Issue. 3, pp. 434-440.

Cerny, V. (1985). A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm, *Journal of Optimization Theory and applications*, Vol. 45, pp. 41-51.

Chakrabarti, D. (2005). *Tools for large graph mining. PhD thesis, School of computer Science*, Carnegie Mellon University.

Chaneau, J. L., Gunaratne, M., Altschaeffl, A.G. (1987), An application of type-2 sets to decision making in engineering, in: J.C. Bezdek, Analysis o fFuzzy information-vol.II: *Artificial intelligence and Decision systems*, CRC Press, Boca Raton, FL.

Chen, G.R. and Pham, T.T. (2001*). Introduction to fuzzy sets, fuzzy logic, and fuzzy control systems*, CRC press, Boca Raton, FL.

Chen, S.L., Li, J.G., Wang, X.G. (2005). *Fuzzy set theory and its application, Science press*, Beijing.

Chen, L.C., Yu, P.S., Tseng, V.S. (2011). WF-MSB: A weighted fuzzy-based biclusering method for gene expression data. *Int. J. of Data Mining and Bioinformatics*., Vol. 5, No. 1, pp. 89-109.

Chinnaiyan, A.M., Huber-Lang, M., Kumar-sinha, C., Barrette, T.R., Shankar-Sinha, S., Sarma, V.J., Padgaonkar, V.A., Ward, P.A., (2001). Molecular signatures of sepsis: Multiorgan gene expression profiles of systemic inflammation. *AM. J. Pathol*, Vol. 159, No.4, pp. 1199-1209.

Chipman, H. and Tibshirani, R. (2006). Hybrid hierarchical clustering with applications to micoarray data, *Biostatistics*, Vol. 7, Issue. 2, pp. 286-301.

Cho, R.J., Campbell, M.J., Einzeler, E.A., Steinmets, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998). A genome-wide transcriptional anaysis of the mitotic cell cycle, *Mol. Cell.*, Vol. 2, pp. 65-73.

Cho, Y.R., Hwang, W., Ramanathan, M., and Zhang, A.D. (2007), Semantic integration to identify overlapping functional modules in protein interaction networks, *BMC Bioinformatics*, Vol.8, No. 265, pp.1-13.

Choi, B.I. and Rhee, C.H. (2009). Interval type-2 fuzzy membership function generation methods for pattern recognition, *Information Sciences*, Vol, 179, Issue. 13, pp. 2102-2122.

Chou, Y. T., Hsu, C.F. Kao, Y. R., Wu, C. W. (2010). Cited2, a novel EGFR-induced coactivator, plays a key role in lung cancer progression, *AACR 101st Annual Meeting*, Poster Presentation.

Chuang, K.S., Tzeng, H.L., Chen, S., Wu, J., Chen, T.J. (2006). Fuzzy c-means clustering with spatial information for image segmentation, *Computerized Medical Imaging and Graphics*, Vol. 30, Issue. 1, pp. 9-15.

Ciric, M., Lgnjatovic, J., and Bogdanovic, S. (2009). Uniform fuzzy relations and fuzzy functions, *Fuzzy sets and Systems*, Vol. 160, Issue. 8, pp. 1054-1081.

Cover, T.M. and Hart, P. E (1967). Nearest neighbour pattern classification, *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 21-27.

Cubellis, M.V., Caillez, F., Blundell, T.L., Lovell, S.C. (2005). Properties of polyproline Ⅱ, a secondary structure element implicated in protein-protein interactions. *Proteins*, Vol. 58, pp. 880-892.

De Bin, R. and Risso, D. (2011). A novel approach to the clustering of microarray data via nonparametric density estimation, *BMC Bioninformatics*, Vol. 12, No. 49.

Dembele, D. and Kanstner, P. (2003). Fuzzy C-means method for clustering microarray data, *Bioinformatics*, Vol. 19, pp. 973-980.

De Vicente, J. Lanchaes, J. Hermida, R. (2003). Placement by thermodynamic simulated annealing, *Physics Letters A*, Vol. 317, Issue. 5-6, pp. 415-423.

Dib, K.A., and Youssef, N.L. (1991). Fuzzy Cartesian product, fuzy relations and fuzzy functions, *Fuzzy sets and Systems*, Vol. 41, Issue 3, pp. 299-315.

Dougherty, E.R., Barrera, J., Brun, M., Kim, S., Cesar, R.M., Chen, Y., Bittnet, M. and Trent, J.M.(2002). Inference form clustering with application to gene-expression microarrays, *J. Comput.Biol.*, Vol. 9, pp. 105-126.

Dubois, D. and Prade, H. (1979). Fuzzy real algebra: some results. *Fuzzy set and systems*, Vol. 2, pp. 327-348.

Dubois, D. and Prade, H. (1980). *Fuzzy sets and systems: Theory and applications*, Academic press, INC, Chestnut Hill, MA.

Dudziak, U., Kala, B.P. (2010). Equivalent bipolar fuzzy relations, *Fuzzy Sets and Systems*, Vol. 161, Issue. 2, pp. 1054-1081.

Dudziak, U. (2010). Weak and graded properties of fuzzy relations in the context of aggregation process, *Fuzzy Sets and Systems*, Vol. 161, Issue 2, pp. 216-233.

Duggan, D.J, Bittner, M and Chen, Y. (1999). Expression profiling using cDNA microarrays, *Nature Genetics*, Vol. 21, Suppl.1, pp. 10-14.

Eckenrode, SE., Ruan, QG., Collins, CD., Yang, P. Mclndoe, RA., Muir,A., She, JX. (2004). Molecular pathways altered by insulin b9-23 immunization," *Ann N Y Acad Sci*, Vol. 1037, pp: 175-185.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wid expression patterns, *em Proc. Natl Acad. Sci.USA*, Vol. 95, pp. 14863-14868.

Faloutsos, M., Faloutsos, P., Faloutsos, C. (1999). *On power-law relationships of the internet topology*. ACM Press.

Fariselli P., Pazos F., Valencia A., Casadio R. (2002): Prediction of protein-protein interaction sites in hetero-complexes with neural networks. *Eur. J. Biochem*., Vol. 269, pp. 1356-1361.

Fazel Zarandi, M.H., Turksen, I.B., Torabi Kasbi, O. (2007). Type-2 fuzzy modeling for desulphurization of steel process, *Expert Systems with Application*, Vol. 32, Issue.2, pp. 157-171.

Fazel Zarandi, M.H., Rezaee, B. Turksen, I.B., Neshat, E. (2009). A type-2 fuzzy rule-based expert system model for stock price analysis, *Expert Systems with Applications*, Vol. 36, Issue. 1, pp. 139-154.

Firneisz, G., Zehavi, I. Vermes, C., Hanyecz, A., Frieman, J.A., and Glant, T.T. (2003). Identification and quantification of disease-related gene clusters, *Bioinformatics*, Vol.19, Issue. 14, pp. 1781-1786.

Freeman, J.A. (1994). Artificial Intelligence: Fuzzy systems for control applications: The truck Backer-upper, *The Mathematica Journal,* Vol.4, Issue, 1. pp.64-69.

Fu, L. and Medico, E. (2002). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data, *BMC Bioinformatics*, Vol. 8, pp. 311-325.

Fukuda, T. and Kubota, N. (1999). An intelligent robotic system based on a fuzzy approach, *Proceddings of the IEEE*, Vol. 87, Issue. 9, pp. 1448-1470.

Getz, G, Levine, E and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci USA*. Vol. 97, No. 22, 12079-12084.

Gilmour, D.S. and Lis, J.T. (1984). Detecting protin-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes. *Proc NatI Acad Sci*, Vol. 81, pp. 4275-4279.

Gilmour, D.S. and Lis, J.T. (1985). In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster, *Mol cell Biol*, Vol. 5, pp. 2009-2018.

Gilmour, D.S. and Lis, J.T. (1986). RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Mol Cell Biol*, Vol. 6, pp. 3984-3989.

Girvan, M. and Newman, M. (2002), Community Structure in social and biological networks, *PNAS*, Vol.99, No.12, pp.7821-7826.

Glonek, G.F.V. and Solomon, P.J. (2004). Factorial and time course designs for cDNA microarray experiments, *Biostatistics*, Vol. 5, Issue.1, pp. 89-111.

Granville, V., Krivanek, M., Rasson, J.-P. (1994). Simulated annealing: A proof of convergence, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No.6, pp. 652–656.

Greenfield, S., Chiclana, F., Coupland, S., and John, R. (2009). The collapsing method of defuzzification for discretised interval type-2 fuzzy sets, *Information Sciences*, Vol. 179, Issue. 13, pp. 2055-2069.

Grint, K. (1997). *Fuzzy Management*, Oxford University Press.

Groll, L. Jakel, J. (2005). A new convergence proof of fuzzy c-means. *IEEE Trans. Fuzzy Systems*. Vol. 13, pp. 717-720.

Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., Van Helden, J., Lemer, Cl, Richelles, J., Wodak, S., Garcia-Martinez, J., Perez-Ortin, J., Michael, H., Kaps, A., Talla, E., Dujon, G., Andre, B., Souciet, J., De Montigny, J., Bon, E., Gaillardin, C. and Mewes, H., (2005), CYGD: the comprehensive yeast genome database, *Nucleic Acides Res*., Vol. 33, pp. D364-D368.

Hacia, J.G., Fan, J.B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R.A., Sun, B., Hsie, L., Robbins, C.M., Brody, L.C., Wang, D., Lander, E.S., Lipshutz, R., Fodor, S.P. and Collins, F.S. (1999). Determination of ancestral alleles for human sigle-nucleotide polymorphisms suing high-density oligonucleotide arrays, *Nat Genet*, Vol.22, pp. 164-167.

Hall, P. Park, B.U., Samworth, R.J. (2008). Choice of neighbour order in nearest-neighbor classification, *Annals of Statistics*, Vol. 36, No. 5, pp. 2135-2152.

Hamerly, G. and Elkan, C. (2002). Alternatives to the k-means lgorithm that find better clusterings, *Proceedings of 11th International Conference on Information and Knowledge Management* (*CIKM*).

Han, J.D., Dupuy, D., Bertin, N., Cusick, M.E., Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, Vol 23, No.7, pp. 839-944.

Hautaniemi, S., Yli-Harja, O., Astola, J., Kauraniemi, P., Kallioniemi, A., Wolf, M., Ruiz, J., Mousses, S., Kallioniemi, O.P. (2003). Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps, *Machine learning*, Vol. 52, Issue. 1-2, pp. 45-66.

He, Y.C., Tang, Y.C., Zhang, Y.Q. and Sunderraman, R. (2006). Adaptive fuzzy association rule mining for effective decision support in biomedical applications, *Int. J. of Data Mining and Bioinformatics.*, Vol. 1, No. 1, pp. 3-18

Herrera, F. and Verdegay, J.L. (1997). Fuzzy sets and operations research: perspectives, *Fuzzy Sets and Systems*, Vol. 90, pp. 207-218.

Herrero A. and Flores E. (2008). *The Cyanobacteria: Molecular Biology, Genomics and Evolution (1st ed.)*. Caister Academic Press, Sevilla.

Hoskin,s J., Lovell ,S.C., Blundell, T.L. (2006). An algorithm for predicting protein-protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. *Protein Sc*i., Vol. 15, Issue. 5, pp. 1017-1029.

Http://www.ncbi.nlm.nih.gov/genbank/

Hu X.H., Pan Y. ( 2007). Knowledge *Discovery in Bioinformatics: Techniques, Methods, and Applications*. WILEY.

Hu, Y.J. and Weng G.R. (2009). Segmentation of cDNA microarray spots using K-means clustering algorithm and mathematical morphology, *Wase international conference on information engineering*, Vol. 2, pp.110-113.

Huarng, K.H. and Yu, H.K. (2005). A type-2 fuzzy time series model for stock index forecasting, Physica A: *Statistical Mechanics and its Applications*, Vol. 353, pp. 445-462.

Huang, N.E., Shen, Z., Long, S., Wu, M., Shih, H.H., Zhang, Q., Yen, N., Tung, C.C. and Liu, H.H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proceedings of the Royal Society of London Series A.*, Vol. 454, pp. 903-995.

Hwang, C. and Rhee, F. C.-H. (2007). Uncertain fuzzy clustering :Interval type-2 fuzzy approach to C-Means, *IEEE Transactions on Fuzzy Systems*, Vol. 15, No.1, pp. 107-120.

Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J.J., Bogosk, M.S. (1999). The transcriptional program in the response of human fibroblast to serum. *Science*, Vol. 283, pp. 83-87.

Janusauskas A., Jurkonis R., and Lukosevicius A. (2005). The empitiacl mode decomposition and the discrete wavelet transform for detection of human cataract in ultrasound signals, *Informatica*, 16, 4, 541-556.

Jason, D. and Christie, M. (2005). Microarrays, *Cit Care Med*, Vol. 33, No.12, Suppl. pp. 449-452.

Jeon, G., Anisetti, M., Bellandi, V., Damiani, E., and Jeong, J. (2009). Designing of a type-2 fuzzy logic filter for improving edge-preserving restoration of interlaced-to-progressive conversion, *Information Sciences*, Vol. 179, Issue. 13, pp. 2194-2207.

Ji, Z.X. Sun, Q.S. and Xia, D.S. (2011). A modified possibilistic fuzzy c-means clustering algorithm for bias field estimation and segmentation of brain MR image, *Computerized Medical Imaging and Graphics*, Vol. 35, Issue. 5, pp. 383-397.

John, R.I., Type 2 fuzzy sets: an appraisal of theory and applications, *Int, J. Unvertainty, Fuzziness Knowledge-Based Systems*, Vol.6, No.6, pp. 563-576.

Jung, S.H., Hyun, B., Jang, W.H., Hur, H.Y., Han, D.S. (2010). Protein complex prediction based on simultaneous protein interaction network, Bioinformatics, Vol.26, Issue. 3, pp. 385-391.

Karimpour-Fard A., Hunter L., and Gill R.T. (2007). Investigation of factors affecting prediction of protein-protein interaction networks by phylogenetic profiling. *BMC Genomics*, Vol.8, 393.

Karnik, N.N., Mendel, J.M. (1998), Introduction to type-2 fuzzy logic systems, *IEEE fuzzy conference*, Anchorage, AK, May

Karnik, N.N. and Mendel, J.M. (1999) Type-2 fuzzy logic systems," *IEEE Trans on Fuzzy Systems*, Vol. 7, pp. 643-658.

Karnik, N.N., Mendel, J.M. (2000), Operations on type-2 fuzzy set, *Fuzzy Sets Systems*. Vol. 122, pp. 327-348.

Kaski, S. (1997). Data exploration using self-organizing maps, *Acta Polytechnica scandinavica, Mathematics, Computing and Management in Engineering Series*, No. 82, pp. 57-71.

Kim, S.Y., Lee, J.W. and Bae, J.S. (2006). Effect of data normalization on fuzzy clustering of DNA microarray data, *BMC Bioinformatics*, Vol. 7, pp. 134-147.

Kim, E.Y., Kim, S.Y., Ashlock, D. and Nam, D. (2009). MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering, *BMC Bioinformatics*, Vol. 10, pp. 260-271.

King, H.C, Sina, A.A. (2001). Gene expression profile analysis by DNA microarrays: promise and pitfalls. *JAMA*, Vol. 286, pp. 2280-2288.

King, A.D., Przulj, N. and Jurisica, I. (2004), Protein complex prediction via cost-based clustering, *Bioinformatics*, Vol. 20, No. 17, pp. 3013-3020.

Kirkpatrick, S. Gelatt, C.D, Vecchi, M.P. (1983). Optimization by simulated annealing, *Science, New Series*, Vol. 220, No. 4598, pp. 671-680.

Klir, J.G. and Yuan, B. (1995). *Fuzzy sets and fuzzy logic theory and applications*. Published by Prentice Hall Inc, upper saddle River, NJ, 07458.

Koenig, L. and Youn, E. (2011). Hierarchical signature clustering for time series microarray data, Software Tools and Algorithms for Biological Systems, Vol. 696, pp. 57-65.

Kohler S., Bauer S., Horn D. (2008). Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, Vol. 82, No. 4, pp: 949-958.

Kulesh, D.A., Clive, D.R., Zarlenga, D.S. and Greene, J.J. (1987). Identification of interferon-modulated proliferation-related cDNA sequences, *Proc NatI Acad Sci USA*, Vol. 84, No. 23, pp. 8453-8457.

Kumbasar, T., Eksin, L., Guzelkaya, M., and Yesil, E. (2011). Interval type-2 fuzzy inverse controller design in nonlinear IMC structure, *Enginering Applications of Artificial Intelligence*, Vol. 24, Issue. 6, pp. 996-1005.

Kuznetsov, S. O. and Obiedkov, S. A. (2001), Algorithms for the construction of concept lattices and their diagram graphs, *Principles of Data Mining and Knowledge Discovery, Lecture Notes in Computer Science*. Springer-Verlag, pp. 289-300.

Lal, M.A., Korner, A., Matsuo, Y., Zelenin, S., Cheng, X., Jarmko, G. DiBona, G. Eklof, A. Aperia, A. (2000) Combined Antioxidant and COMT inhibitor treatment reverses renal abnormalities in diabetic rats, *Diabetes*, Vol. 49, pp.1381-1389.

Lashkari, D.A., DeRisi, J.L., McCusker, J.H. and Namath, A.F. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis, *Proc NatI Acad Sci USA*, Vol. 94, No. 24, pp. 1305-13062.

Lbelda, S.M. and Sheppard, D. (2000). Functional genomics and expression profiling: Be there or be square. Am J Respir Cell Molec Biol, Vol. 23, pp. 265-269.

Leal-Ramirez, C., Castillo, O., Melin, P., and Rodriguez-Diaz, A. (2010). Simulation of the bird age-structured population growth based on an interval type-2 fuzzy cellular structure, *Infromation Sciences,* Vol. 181, Issue. 3, pp. 519-535.

Lee P. S. and Lee K. H. (2000). Genomic analysis. *Curr. Opin. Biotechnology,* 11(2), 171-175.

Lee T. I., Rinaldi N. J., Bobert F., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science*, 298 (25), 799-804.

Li, J. and Johnson, J. A. (2002). Time-dependent Changes in ARE-driven gene expression by use of a noise-filtering process for microarray data, *Physiological Genomics*, Vol. 9, Issue. 3, pp. 137-144.

Li, M., Wang, J.X. and Chen, J.E. (2008). A fast agglomerate algorithm for mining functional modules in protein interaction networks, *Proceeding of ICBEI*, pp. 3-8.

Li, M.X., Li, Z.B., Jang, Z.R., Li, D.D. (2010 a). Prediction of disease-related genes based on hybrid features, *Current Proteomics*, Vol. 7, Issue. 2, pp. 82-89.

Li, M., Wang, J.X., Zhao, C., Chen, G. (2010). Identifying the overlapping compleses in protein interaction networks, *Int. J. Data Mining and Bioinformatics,* Vol. 4 No. 1, 2010.

Liang, L.R., Lu, S., Wang, X., Lu, Y., Mandal, V., Patacsil, D. and Kumar, D. (2006). FM-test: a fuzzy-set-theory-based approach to differential gene expression data analysis, *BMC Bioinformatics*, Vol. 7, Suppl No. 4, S7.

Lin, L., Wang, Y. and Zhou, H. (2009). Iterative filtering as an alternative for empiricalmode decomposition, *Advances in Adaptive Data Analysis*, Vol. 1, No. 4, pp. 543-560.

Lioyd., S. P. (1982). Least squares quantization in PCM, *IEEE Transactions on Information Theory*, Vol. 28, No. 2, pp. 129-137.

Liu, X.(1998). The fuzzy theory based on AFS algebras and AFS structures. *Journal of Mathematical Analysis and Applications*, Vol. 217, pp. 479-489.

Liu, D.Q., Shi, T., DiDonat, J.A., Carpten, J.D., Zhu, J.P., Duan, Z.H. (2004). Application of genetic algorithm / K-nearest neighbour method to the classification of renal cell carcinoma, *Proceedings of IEEE Computational Systems Bioinformatics Conference*, pp. 558-559.

Liu B., Jiang T., Ma S., Zhao H., Li J., Jiang X. P., Zhang J. (2006). Exploring candidate genes for human brain diseases from a brain-specific gene network. *Biochem BIophys Res Commun*. Vol. 349, pp. 1308-1314.

Liu, X.D and Liu, W.Q. (2008). The framework of axiomatics fuzzy sets sets based fuzzy classifiers, *Journal of industrial and management optimization*, Vol. 4, Number 3, pp. 581-609.

Lo, K. C., Stein, L. C., Panzarella, J. A., Cowell, J. K., Hawthorn, L. (2008). Identification of genes involved in squamous cell carcinoma of the lung using synchronized data from DNA copy number and transcript expression profiling analysis, *Lung Cancer*, Vol. 59, pp. 315-331.

Lovf, M., Thomassen, G.O., Bakken, A.C., Celestino, R., Fioretos, T., Lind, G.E., Lothe, R.A. and Skotheim, R.I. (2011). Fusion gene microarray reveals cancer type-

2specificity among fusion genes, *Genes Chromosomes Cancer*, Vol. 50, No. 5, pp. 348-357.

Ma, P.C.H., Chan, K.C.C., Yao, X. and Chiu, D.K.Y. (2006), An evolutionary clustering algorithm for gene expression microarray data analysis, *IEEE Transactions on Evolutionary Computation*, Vol. 10, Issue.3, pp. 296-314.

Ma, S. and Huang, J. (2007). Clustering threshold gradient descent regualriation: with applications to microarray studies. *Bioinformatics*, Vol. 23, No. 4, pp.466-472.

Macintyre, G., Bailey, J., Gustafsson, G., Haviv, I. Kovlczyk, A. (2010). Using gene ontology annotations in exploratory microarray clustering to understand cancer etiology, *Pattern Recognition Letters*, Vol. 31, Issue, 14, pp. 2138-2146.

Macqueen, J.B. (1967). Some methods for classification and analysis of multivariate observations, *Proceeding of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 291-297.

Maiers, J.E. (1985). Fuzzy set theory and medicine: the first twenty years and beyond, *Proc Annu Symp Comput Appl Med Care*, pp. 325-329.

Mann M. and Jensen O. N. (2003). Proteomic analysis of post-translational modifications, *Nat. Biotechnol.*, 21(3), 255-261.

Mansillla, F., Costa, K., Wang, S., Kruhoffer, M., Lewin, T. M., Orntoft, T. F., Coleman, R. A., Birkenkamp-Demtroder, K., (2009). Lysophosphatidylcholine acyltransferase 1 (LPCAT1) overexpression in human colorectal cancer, *J. Mol Med*, Vol. 87, pp. 85-97.

Maskos, U. and Southern E.M. (1992). Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides sysnthesised in situ, *Nucleic Acides Res*, Vol. 11, pp. 1679-1684.

Maulik, U. and Mukhopadhyay, A. (2010). Simulated annealing based automatic fuzzy clustering combined with ANN classification for analysing microarray data, *Computers and Operations Research*, Vol. 37, Issue. 8, pp. 1369-1380.

Marchitti, A.S., Orlicky, D.J., Brocker, C., Vasiliou, V. (2010). Aldehyde Dehydrogenase 3B1 (ALD3B1): Immunohistochemical tissue distribution and cellular-specific localization in normal and cancerous human tissues, *J. Histochemistry and Cytochemistry*, Vol. 58, pp. 765-783.

McNicholas P.D. and Murphy, T.B (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models, Bioinformatics, Vol. 26, Issue. 21, pp. 2705-2712.

Medvedovic, M., Yeung, K.Y., Bumgarner, R.E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, Vol. 20, No. 8, pp. 1222-1232.

Mendel, J.M. and Bob John, R.I. (2000). Type-2 fuzzy sets made simple", *IEEE Trans. On Fuzzy Systems*, Vol. 10, No. 2, pp: 117-128.

Mete, M., Tang, F., Xu, X., Yuruk, N. (2008). A structural approach for finding functional modules from large biological networks, *BMC Bioinformatics*, Vol.9, pp.1-14.

Mizumoto, M. and Tanaka, K. (1976), Some properties of fuzzy sets of type-2, *Inform. Control*, Vol. 31, pp. 312-340.

Mizumoto, M. and Tanaka, K. (1981), Fuzzy sets of type-2 under algebraic product and algebraic sum. *Fuzzy sets Systems*, Vol. 5, pp 277-290.

Mohammadi, A., Saraee, M.H, and Salehi, M. (2011). Identification of disease-causing genes using microarray data mining and gene ontology, *BMC Med Genomics*, Vol. 4, pp. 12-23.

Moran, G., Stokes, C., Thewes, S., Hube, B., Coleman, D.C., and Sullivan, D. (2004). Comparative genomics using candida albicans DNA microarrays reveals absence and divergence of virulence-associated genes in candida dubliniensis, *Microbiology*, Vol.150, pp. 3363-3382.

Newman, M. E. J. and Wattes, D.J. (1999), Scaling and percolation in the small-world network model, *Phys. Rev. Lett.*, Vol. 60, pp. 7332-7342.

Newman, M. E. J. (2004), Fast algorithm for detecting community structure in networks, Phys. Rev.E, Vol.69, No.6, pp.66-133.

Oldham, R. (1983). Natural killer cells: Artifact to reality: an odyssey in biology, *Cancer Metastasis Reviews*, Vol. 2, pp. 323-326.

Ozawa, Y., Saito, R., Fujimori, S. Kashima, H., Ishizaka, M., Yanagawa, H., Miyamoto-Sato, E., Tomita, M. (2010). Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions, *BMC Bioinformatics*, Vol. 11, No. 350.

Palla, G., Derenyi, I., Farkas, I., Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, Vol.435, No. 7043, pp. 814-818.

Palzkill, T. (2002), *Proteomics*, Kluwer Academic Publisheds.

Park, S. H., Reyes J. A., Gilbert, D. R., Kim, J. K., Kim, S. (2009). Prediction of protein-protein interaction types using association rule based classification. *BMC Bioinformatics*, Vol. 10, No. 36.

Pederycz, W. Fuzzy-sets in pattern-recognition: Methodology and methods, *Pattern Recognition*, Vol. 23, Issue. 1-2, pp. 121-146.

Peink, J., Alber, M. (1998). Improved multifractal box-counting algorithm, virtual phase transitions, and negative dimensions, *Physical Review E.*, Vol. 57, No. 5, pp. 5489-5493.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic porgies. *Proc Natl Acad Sic USA*, Vol. 96, Issue. 8, pp. 4285-4288

Platta, C.S., Greenblatt, D.Y., Kunnimalaiyaan, M., Chen, H. (2008). Valproic acid induces Notch 1 signaling in small cell lung cancer cells, *J. Surgical Research*, Vol. 148, pp. 31-37.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B. Pergamenschikow, A., Williams, C.F., Jeffrey, S.S., Botstein, D. and Brown, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays, *Nat Genet*, Vol. 23, No. 1, pp. 41-46.

Potamias, G. (2004). Knowledgeable clustering of microarray data, *Biological and Medical Data Analysis*, Vol. 3337, pp. 491-497.

Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z. (2007). A mixture of feature experts approach for protein-protein interaction prediction, *BMC Bioinformatics*, Vol. 8 (Suppl 10): S6.

Qi, Y.J. (2008). *Learing of Protein Interaciton Networks*, PhD Thesis, Carnegie Mellon University.

Qin, J., Lewis, D.P., Noble, W.S. (2003). Kernel hierarchical gene clustering from microarray expression data, *Bioinformatics*, Vol. 19, Issue. 16, pp. 2097-2104.

Qu, Y. and Xu, S. (2004). Supervised cluster analysis for micorarray data based on multivariate Gaussian mixture, *Bioinformatics*, Vol. 20, No. 12, pp, 1905-1913.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. (2004). Defining and identifying communities in networks, *Proc. Natl. Acad. Sci, USA*, Vol. 101, pp. 2658-2663.

Rao M.A., Srinivas J. (2003). *Neural Networks: Algorithms and Applications*, Alpha Science International Lrd.

Rhee, C.H. (2007). Uncertain fuzzy clustering: Insights and recommendations, *IEEE Computational Intelligence Magazine*, Vol. 2, No. 1, pp. 44-56.

Romdhane, L.B., Shili, H., Ayeb, B. (2010). Mining microarray gene expression data with unsupervised possibilistic clustering and proximity graphs, *Applied Intelligence*, Vol. 33, Issue. 2, pp. 220-231.

Rome, S., Clement, K., Rabasa-Lhoret, R., Loizon, E., Poitou, C., Barsh, GS., Riou, JP., Laville, M., idal, H. (2003). Microarray profiling of human skeletal muscle reveals that insulin regulateds approximately 800 genes during a hyperinsulinemic clamp," *J Biol Chem*, Vol. 278, No. 20, pp. 18063-18068.

Rosner, B. (2000). *Fundamentals of Biostatistics*. In Pacific Grove 5th edition, CA: Duxbury Press.

Ross, T. (2004). *Fuzzy logic with engineering applications*, John Wiley & Sons Ltd, The atrium, Southern Gate, Chichester, West Sussex, England.

Rousseeuw, J.P. (1987). Silhouettes: a graphical aid to the interpration and validation of cluster analysis, *J. Comp. Appl. Math.*, Vol. 20, pp. 53-65.

Rumelhart D.E., Hinton G.E., Williams R.J. (1986). Learning representations by back-propagating errors, *Nature*, Vol. 323, pp. 533-536.

Schalkoff R.J. (1997): *Artificial Neural Networks*. MIT.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, Vol. 270, No. 5235, pp. 467-470.

Shah-Hosseini, H. and Safabakhsh, R. (2003). TASOM: A new time adaptive self-organizing map, *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, Vol. 33, No. 2, pp. 271-281.

Sharan, R. and Shamir, R. (2000). CLICK: a Clustering algorithm with application to gene expression analysis, *Proceedings of AAAI-ISMB*, pp. 307-316.

Shaul K., Hermona S., Sharan R. (2009) A network-based method for predicting disease-causing genes. *Journal of Computational Biology*, Vol.16, pp. 181-189

Shi, F., Chen, Q. and Niu, N. (2007). Functional similarity analyzing of protein sequences with empirical mode decomposition, *Proceeding of The fourth international conference on fuzzy systems and knowledge discovery*.

Siler, W. and Buckley, J. (2005). *Fuzzy expert systems and fuzzy reasoning*, John Wiley and Sons, Inc., Hoboke, New Jersey.

Sokal R. R. and Rohlf F. J. (1995). *Biometry: The principles and practice of statistics in biological research*, W. H. Freeman, New York.

Sol A.D., O'Meara P. (2005). Small-world network approach to identify key residues in protein-protein interaction. *Protein*, Vol. 58, pp. 672-682

Someren, E., Wessels, L., Reindersa, M. and Backer, E. (2006). Regularization and noise injection for improving genetic network models, In: Zhang, W. and Shmulevich,I. *Computational and statistical approaches to genomics* (*2nd edition*), (pp. 279-295) Springer Science+Business Media, Inc.

Sonim, D.K. (2002). From patterns to pathways: gene expression data analysis comes of age, *Nature, genetics*, Vol. 32, pp. 502-508.

Spirin, V and Mirny, L.A. (2003). Protein complexes and functional modules in molecular networks, *PNAS*, Vol. 100, No. 21, pp. 12123-12128.

Stelzl, U., Worm, U., Lalowski, M, Wanker, E. (2005). A human protein-protein interaction network: a resource for annotation the proteome. *Cell*, Vol. 122, No.6, pp. 957-968.

Subhani, N., Rueda, L., Ngom, A., Burden, C.J. (2010). Multiple gene expression profile alignment for microarray time-series data clustering, *Bioinformatics*, Vol. 26, Issue. 18, pp. 2281-2288.

Sun, C.C., Lin, T.R., and Tzeng, G.H. (2009). The evaluation of cluster policy by fuzzy MCDM: Empirical evidence from HsinChu Science Park, *Expert Systems with Applications*, Vol. 36, Issue. 9, pp. 11895011906.

Sun, P.G., Gao, L. and Han, S. (2011). Prediction of human disease-related gene clusters by clustering analysis, *Int J Biol Sci*, Vol. 7, pp.61-73.

Sun, P.G., Gao, L. and Han, S.S. (2011). Identificaition of overlapping and non-overlapping community structure by fuzzy clustering in complex networks, *Information Sciences*, Vol. 181, pp. 1060-1071.

Szekely, G.J. and Rizzo, M.L. (2005). Hierarchical clustering via joint between-within distances: extending ward's Minimum variance method, *Jouranl of Classification*, Vol. 22, pp. 151-183.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999). Systematic determination of genetic network architecture, *Nat.Genet*, Vol. 22, pp. 281-285.

Terrell, D.G. and Scott, D.W. (1992). Variable kernel density estimation, *Annals of Statistics*, Vol. 20, No. 3, pp.1236-1265.

Thalamuthu, A, Mukhopadhyay, L, Zheng, X., Tseng, G.C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, Vol.22, No. 19, pp. 2405-2412.

Tong, S.C., Liu, C.L., Li, Y.M. (2010). Robust adaptive fuzzy filters output feedback control of strict-feedback nonlinear systems. *Applied Mathematics and computer science*, Vol. 20, No. 4, pp. 637-653.

Torkkola, K., Gardner, R.M., Kaysser-Kranich, T., Ma, C. (2001) Self-organizing maps in mining gene expression data, *Information Sciences*, Vol. 139, Issue, 1-2, pp. 79-96.

Virtanen, S.E. (2003). Properties of nonuniform random graph meodels. Research Report A77, *Helsinki University of Technology, Laboratory for Theoretical Computer Science*, Espoo, Finland.

Wachi, S., Yoneda, K., Wu, R. (2005) Interactome-transcriptone analysis reveals the high centrality of genes differentially expressed in lung cancer tissues, *Bioinfromatics*, Vol. 21, pp. 4205-4208.

Wagenknecht, M., Hartmann, K. (1988), Application of fuzzy sets of type 2 to the solution of fuzzy equation systems, *Fuzzy Sets Systems*, Vol. 25, pp. 183-190.

Wang, J., Liu, I., Tzeng, T., Cheng, J. (2002). Decrease in catechol-o-methyltransferase activity in the liver of streptozotocin-induced diabetic rats," Clin Exp Pharmacol & Physiol, Vol. 29, pp. 419-422.

Wang, J.B., Bo, T.H., Jonassen, I., Myklebost, O. and Hovig, E. (2003). Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data, *BMC Bioinformatics*, Vol. 4, No.60.

Wang, J., Coombes, K.R., Baggerly, K., Hu, L., Hamilton, S.R.and Wang, W. (2006). Statistical considerations in the assessment of cDNA microarray data obtained using amplification, In: Zhang, W. and Shmulevich,I. *Computational and statistical approaches to genomics* (*2nd edition*), (pp. 21-36), Springer Science+Business Media, Inc.

Wang, Y.J. (2010). A clustering method based on fuzzy equivalence relation for customer relationship management, *Expert Systems with Application*. Vol. 37, Issue. 9, pp. 6421-6428.

Ward and Joe, H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, Vol. 58, No. 301, pp. 236-244.

Watkinson, J., Wang ,X.D., Zheng, T. and Anastassiou, D. (2008). Identification of gene interactions associated with disease from gene expression data using synergy networks, *BMC Systems Biology*, Vol. 2, No. 10.

Waksman, G. (2005). *Protein Reviews*, Springer.

Wolkenhauer, O. (2001). *Data Engineering: Fuzzy mathematics in systems theory and data analysis*. Published by John Wiley &Sons, Inc.

Xiao, J., Wang, X., and Xu, C. (2008). Comparision of supervised clustering methods for the analysis of DNA microarray expression data, *Agricultural Sciences in China*, Vol.7, Iss. 2, pp. 129-139.

Xu, B. Qu, X., Doughman, Y., Watanabe, M., Dunwoodie, S. L., Yang, Y. (2008) Cited2 is required for fetal lung maturation, *Dev Biol*, Vol. 317, pp. 95-105.

Wu, D.R. and Tan, W.W. (2006). A simplified type-2 fuzzy logic controller for real-time control, *ISA Transactions*, Vol. 45, Issue. 4, pp. 503-516.

Yager, R.R. (1980), Fuzzy subsets of type II in decisions, *J. Cybernet*, Vol.10, pp. 137-159.

Yager, R.R., Zadeh, L.A. (1992). *An introduction to fuzzy logic applications in intelligent systems*, Kluwer Academic.

Yang, X., Pratley, R.T., Tokraks, S., Bogardus, C., Permana, P.A. (2002). Microarray profiling of skeletal muscle tissues from eqally obese, non-diabetic insulin-sensitive and insulin-resistant Pima Indians," *Diabetologia*, Vol. 45 pp. 1584-1593.

Yang, W., Rueda, L., Ngom, A. (2007). On finding the best parameters of fuzzy k-means for clustering microarray data, *Journal of Multiple-Valued Logic and Soft Computing*, Vol. 13, Issue. 1-2, pp. 145-177.

Yoon, S., Yang, Y. Choi, J., and Seong, J. (2006). Large scale data mining approach for gene-fpecific standardization of microarray gene expression data, *Bioinformatics*, Vol. 22, Issue. 23, pp. 2898-2904.

Yu, Z. G., Anh, V., Wang, Y., Mao, D. and Wanliss, J. (2010). Modelling and simulationof the horizontal component of the magnetic field by fractional stochastical differential equations in conjunction with empirical mode decomposition, *J. Geophys. Res.*, Vol. 115, Art. No. A10219, doi:10.1029/2009JA015206.

Zacher, B., Kuan, P.F. and Tresch, A. (2010). Starr: Simple tiling array analysis of affymetrix ChIP-chip data, *BMC Bioinformatics*, Vol. 11, pp. 194-200.

Zadeh, L.A. (1965). Fuzzy sets, *Information and Control*, Vol. 8, pp. 338-353.

Zadeh, L.A. (1975). The concept of a linguistic variable and its application to approximate reasoning – 1, *Inform. Sci.*, Vol. 8, pp. 199-249.

Zadeh, L.A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, Vol. 1, pp. 3-28.

Zadeh, L.A. (2005). Towrd a generalized theory of uncertainty, *Information Sciences*, Vol. 172, pp. 1-40

Zar J. H.(1998). *Biostatistical analysis*, Prentice Hall, Upper Saddle River, Prentice Hall, NJ.

Zhang, L.X. and Zhu, S. (2002). A new clustering method for microarray data analysis, *Proceeding of CSB.*

Zhang, D.Q. and Chen, S.C. (2004). A novel kernelized fuzzy c-means algorithm with application in medical image segmentation, *Artificial Intelligence in Medicine*, Vol. 32, Issue. 1, pp. 37-50.

Zhang, S.H., Wang, R.S., Zhang, X.S. (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering, *Physica A*, Vol. 374, pp. 483-490.

Zhang, L., Hu, K. and Tang, Y. (2010). Predicting disease-related genes by topological similarity in human protein-protein interaction network, Central European Journal of Physics, Vol. 8, Issue. 4, pp. 672-682.

Zhu, D., Hero, A.O., Cheng, H., Khanna, R., Swaroop, A. (2005). Network constrained clustering for gene microarray data, *Bioinformatics*, Vol.21, No.21, pp 4014-4020.

Zimmermann, H.-J. (2001). *Fuzzy set theory—and its applications*, Kluwer Academic Publishers, Dordrecht, Boston.

Zimmermann, H.-J. (2010). Fuzzy set theory, *WIREs, Comp*, Stat, Vol. 2, pp. 317-332.

Zotenko, E., Guimaraes, K.S., Jothi, R. and Przytycka, T.M. (2006), Decomposition of overlapping protein complexes: a graph theoretical method for analysing static and dynamic protein associations, *Algorithms for Molecular Biology*, Vol. 1, P.7.