

**FEASIBILITY STUDY OF FUZZY CLUSTERING TECHNIQUES IN
CHEMICAL DATABASE FOR COMPOUND CLASSIFICATION**

VOT 75107

**JABATAN SISTEM MAKLUMAT
FAKULTI SAINS KOMPUTER DAN SISTEM MAKLUMAT**

PENYELIDIK:

ROZILAWATI BT DOLLAH @ MD. ZAIN (Ketua)

ARYATI BT BAKRI

MAHADI BIN BAHARI

PM DR NAOMIE BT SALIM

**FEASIBILITY STUDY OF FUZZY CLUSTERING TECHNIQUES
IN CHEMICAL DATABASE FOR COMPOUND
CLASSIFICATION**

RESEARCHER :

ROZILAWATI BINTI DOLLAH @ MD. ZAIN (HEAD)

ARYATI BINTI BAKRI

MAHADI BIN BAHARI

PM DR. NAOMIE BINTI SALIM

UNIVERSITI TEKNOLOGI MALAYSIA

2006

ACKNOWLEDGEMENT

Alhamdullillah...

First and foremost we would like to express our true and sincere thanks and gratitude to PM Dr Naomie Salim, for the guidance, support and encouragements during the course of this work. Also, we would like to express our thanks to the member of our department for all the help and support during the time spent in the lab. We also take this opportunity to express our appreciation, especially to the staffs and members of the Faculty of Computer Science and Information Systems (FSKSM) for supporting our research.

ABSTRACT

Compound selection methods are important in drug discovery especially in lead identification process. Finding the best method in compound selection has become a need to the pharmaceutical industry because of the increasing number of chemical compound to be screened. One of the best and widely used methods in compound selection is cluster-based selection where the compound datasets are grouped into clusters and representative compounds are selected from each cluster. Non-overlapping methods, such as Ward's clustering method, have been widely used and it was agreed as the most efficient clustering method in compound selection. However, little focus has been given to overlapping method in compound selection or even in lead identification process. The research focused on the fuzzy c-means clustering where the effectiveness of the clusters produced with regard to compound selection is analyzed and compared with other conventional cluster-based compound selection method. Fuzzy c-means have been chosen because it produces clusters by identifying the cluster centroid and their corresponding degree of membership, therefore the compounds may belong to more than one cluster. The results from fuzzy c-means method are compared to Ward's clustering method and also to the results from the fuzzification of Ward's cluster. The analysis shows that fuzzy c-means clustering gives the best result in intermolecular dissimilarity; however it shows poor results of separation of active/inactive structure.

Key Researchers :

Rozilawati bt Dollah @ Md. Zain (Head)

Aryati bt Bakri

Mahadi b Bahari

PM Dr Naomie bt Salim

Email : rozilawati@utm.my

Tel : 07-5532425

Vot : 75107

ABSTRAK

Kaedah pemilihan sebatian merupakan kaedah yang penting di dalam penemuan ubat, terutamanya bagi proses pengenalpastian molekul yang berpotensi menjadi ubat. Penyelidikan untuk mencari kaedah yang terbaik bagi pemilihan sebatian telah menjadi satu kepentingan kepada industri farmasi kerana peningkatan pada jumlah sebatian yang perlu ditapis. Kaedah yang terbaik dan sering digunakan di dalam pemilihan sebatian adalah kaedah pengkelompokan; di mana set-set data sebatian dikumpulkan dalam kelompok masing-masing dan wakil daripada setiap kelompok akan dipilih. Kaedah tidak bertindih seperti kaedah pengkelompokan Ward's merupakan satu kaedah yang paling berkesan di dalam pengkelompokan sebatian dan digunakan dengan meluas di dalam pemilihan sebatian. Namun begitu, kaedah pengkelompokan bertindih tidak diberikan perhatian yang khusus di dalam pemilihan sebatian mahupun di dalam proses pengenalpastian molekul. Fokus kajian ini adalah kepada kaedah *fuzzy c-means* dan keberkesanan kelompok yang dihasilkan oleh kaedah ini dianalisa dan dibandingkan dengan kaedah konvensional pengkelompokan yang lain. Kaedah *fuzzy c-means* ini telah dipilih kerana ia akan menghasilkan kelompok yang baik dengan mengenalpasti titik tengah kelompok dan darjah keahlian bagi setiap ahli di dalam kelompok. Oleh itu, satu sebatian mungkin berada di dalam lebih daripada satu kelompok berdasarkan kepada darjah keahliannya. Hasil daripada eksperimen ini dibandingkan dengan keputusan daripada kaedah pengkelompokan Ward's. Analisa yang diperolehi menunjukkan bahawa pengkelompokan *fuzzy c-means* memberikan keputusan yang terbaik bagi ketidak-samaan molekul bagi pusat kelompok yang terhasil, tetapi ia tidak melakukan pemisahan struktur aktif/tidak aktif dengan baik di dalam kelompok yang berkenaan.

Penyelidik :

Rozilawati bt Dollah @ Md. Zain (Ketua)

Aryati bt Bakri

Mahadi b Bahari

PM Dr Naomie bt Salim

Email : rozilawati@utm.my

Tel : 07-5532425

Vot : 75107

TABLE OF CONTENT

CHAPTER	TITLE	PAGE
	ABSTRACT	iii
	ABSTRAK	iv
	TABLE OF CONTENT	v
	LIST OF TABLES	viii
	LIST OF FIGURES	ix
	LIST OF SYMBOLS	xi
	LIST OF ABBREVIATION	xii
	LIST OF TERMINOLOGY	xiii
1	INTRODUCTION	1
	1.1 Background of Problems	3
	1.2 Problem Statement	5
	1.3 Objectives	5
	1.4 Scope	6
	1.5 Project Plan	6
	1.6 Contribution	7
	1.7 Organization of Report	8
2	LITERATURE REVIEW	9
	2.0 Introduction	10
	2.1 Cluster-based Compound Selection	11
	2.1.1 Overlapping Method	13

	2.1.2 Non-Overlapping Method	14
	2.1.2.1 Non-Hierarchical Clustering	15
	2.1.2.2 Hierarchical Clustering	19
	2.1.2.3 Ward's Technique	21
	2.2 Clustering in Chemical Application	23
	2.3 Fuzzy Clustering Method	24
	2.3.1 Application of Fuzzy Clustering	27
	2.4 Evaluation of Compound Clustering for Compound Selection Purpose	29
	2.4.1 Separation of Active/Inactive Structure	29
	2.4.2 Diversity Analysis	30
	2.4.3 Grouping Similar Compound	31
	2.5 Descriptors for Chemical Databases	31
	2.5.1 Bit String	33
	2.6 Discussion	36
	2.7 Summary	38
3	EXPERIMENTAL DESIGN	39
	3.0 Introduction	40
	3.1 Dataset	41
	3.2 Generation of Descriptors	42
	3.3 Selection of Similarity Measures	43
	3.4 Implementation of Fuzzy Clustering	44
	3.4.1 Ward's Algorithm	47
	3.4.2 Fuzzification of Ward's Cluster using Fuzzy c Means Clustering Method	49
	3.5 Analysis of Results	50
	3.6 Discussion	52
	3.7 Summary	54
4	EXPERIMENTAL RESULT	56
	4.0 Introduction	57
	4.1 Analysis of Fuzzy c-Means Clustering	57

4.1.1	Analysis of Active/Inactive Separation	58
4.1.2	Analysis of Mean Intermolecular Dissimilarity (MIMD)	60
4.2	Comparison of Fuzzy c-Means and Ward's Clustering Method	63
4.2.1	Analysis of Active/Inactive Separation	64
4.2.2	Analysis of Mean Intermolecular Dissimilarity (MIMD)	65
4.3	Fuzzification of Ward's Cluster using Fuzzy c Means	67
4.3.1	Analysis of Active/Inactive Separation	68
4.3.2	Analysis of Mean Intermolecular Dissimilarity (MIMD)	69
4.4	Discussion	70
4.5	Summary	73
5	CONCLUSION	75
5.0	The Analysis of Contribution	77
5.1	Suggestion for Future Work	78
5.2	Summary	79
	LIST OF REFERENCES	81

LIST OF TABLES

TABLE NO	TITLE	PAGE
2.1	Types of compound selection method	11
2.2	Different types of non-hierarchical clustering methods	17
2.3	Different choices of hierarchical-agglomerative clustering	20
2.4	Different types of fuzzy clustering algorithm	26
4.1	Results for proportion of actives from different fuzziness index	58
4.2	Results for intermolecular dissimilarity from different fuzziness index	61
5.1	Comparison of results from other studies	76

LIST OF FIGURES

FIGURE NO	TITLE	PAGE
1.1	Drug discovery process	2
2.1	Examples of overlapping and non-overlapping clusters (Wild, 2003)	13
2.2	Overlapping clustering (MacCuish and MacCuish, 2003)	14
2.3	A broad classification of the most common clustering methods (Barnard and Downs, 2002)	15
2.4	Example of hierarchical-agglomerative and hierarchical-divisive method (Wild, 2003)	19
2.5	A simple representation of bit string (MacCuish and MacCuish, 2003)	33
2.6	Encoding chemical structure as a bit string (Flower, 1997)	34
2.7	A detailed encoding of a bit string (Flower, 1997)	34
3.1	Flowchart for fuzzy clustering algorithm	41
3.2	Fuzzy c-means algorithm	46
3.3	Algorithm for Ward's clustering	48
3.4	Algorithm for Ward's clustering using RNN	48
3.5	Algorithm for fuzzy c-means clustering in Ward's method	50
3.6	Research methodology framework	54
4.1	Result of Proportion of actives (Pa) for Cluster 10	59
4.2	Result of Proportion of actives (Pa) for all clusters (using $q = 1.1, 1.5$ and 2.0)	60
4.3	Result of MIMD for Cluster 10	62
4.4	Result of MIMD for all clusters (using $q = 1.1, 1.5$ and 2.0)	62
4.5	Results from fuzzy c-means and Ward's clustering based on their proportion of actives	64

4.6	Results from fuzzy c-means and Ward's clustering based on their MIMD	66
4.7	Results based on their proportion of actives	68
4.8	Results based on their intermolecular dissimilarity	70

LIST OF SYMBOLS

q	-	fuzziness index
k	-	number of cluster
u_{ij}	-	degree of membership
X_j	-	the data point of the j th compound
M	-	number of data point
U	-	a fuzzy K-partition of the data set
C	-	a set o K prototypes (cluster center)
C_i	-	the centroid of the i th cluster
x_{ik}	-	the attribute value of molecule i in cluster k
n	-	size of cluster
P_a	-	proportion of active structure
$d_2(X_j, C_i)$	-	any inner product metric or the distance measure
$S_{A,B}$	-	similarity measure between compound A and B
a	-	the number of unique fragments in compound A
b	-	the number of unique fragments in compound B
c	-	the number of unique fragments shared by compounds A and B

LIST OF ABBREVIATION

AFC	-	Adaptive Fuzzy Clustering
BCI	-	Barnard Chemical Information
CA	-	Confirmed Active
CI	-	Confirmed Inactive
CM	-	Confirmed Moderately Active
DNA	-	Deoxyribonucleic Acid
E.COLI	-	Escherichia coli
EM	-	Expectation Maximization
ESS	-	Error Sum of Squared
FCM	-	Fuzzy c-Means
FCV	-	Fuzzy c-Varieties
GG	-	Gath-Geva
GK	-	Gustafson-Kessel
HTS	-	High Throughput Screening
MDDR	-	MDL Drug Data Report
MDL	-	Molecular Design Limited
MIMD	-	Mean Intermolecular Dissimilarity
NCI	-	National Cancer Institute
PPP	-	Potential-Pharmacophore-Point
QSAR	-	Quantitative Structure-Activity Relationship
R&D	-	Research and Development
RNN	-	Reciprocal Nearest Neighbor
USD	-	United States Dollar

LIST OF TERMINOLOGY

Alignment	-	Concerned with the relationships between biological sequences
Analyte	-	A sample mixture that is passed through some form of material that will provide resistance by virtue of chemical interactions between the components of the sample and the material
Atom	-	The smallest irreducible constituent of a chemical system
Benign	-	A tumor that is not dangerous to one's health
Bond	-	The force which holds atoms together in molecules
Compound	-	A substance formed from two or more elements, with a fixed ratio determining the composition
E.Coli	-	One of the main species of bacteria that live in the lower intestines of warm-blooded animals
Gene	-	A sequence of DNA that represents a fundamental unit of heredity
Gene Expression	-	Refers to the multi-step process that begins with protein biosynthesis and is followed by folding, post translational modification and targeting.
Lead	-	A molecule that have the potential to become new drug
Malignant	-	A tumor that was formed when the body's own cell divide in an uncontrolled manner
Molecule	-	The smallest indivisible portion of a pure compound that retains a set of unique chemical and physical properties
Organic	-	A branch of chemistry dealing with carbon-based compounds
Outlier	-	An unusual data objects
Sedatives	-	A drug that depresses the central nervous system
Tranquilizer	-	A drug that is used to relax the body system

EDIE

I S \$

Ø¿

øê Døç ù{ ,

àa

* , %

E D I E

J a b

CHAPTER 1

INTRODUCTION

Chemoinformatics is the collection, representation and organisation of chemical data to create chemical information, to which it can be applied to create chemical knowledge. In pharmaceutical and agrochemical industry, chemoinformatics has been used for identification of novel compounds with useful, and commercially valuable, biological properties (Brown and Martin, 1996; Warr, 1997; Tropsha and Zheng, 2001). The drug discovery process is very complex, and it is a multi-disciplinary task with many stages to be performed in a long time, as shown in Figure 1.1. However, the drug discovery process is a very risky business because most of the newly found compounds do not result in a drug. The molecule that has the potential to become drugs may cause unexpected long-term side effects. The drug discovery process can take about 12 years and the costs may reach USD \$350 millions per drug.

The high costs to bring a drug to market have increase the pressure to the pharmaceutical industries. Therefore attention is given to the research and development to develop faster and more effective way to produce chemical compounds that can react to the disease and furthermore, can produce antibodies towards the disease. This has encouraged the study of chemoinformatics and drug discovery as one of a new area in

Malaysia's research and development (R&D) (Law, 2003). Malaysian government's commitment to participate actively in biotechnology industry is proven by the development of Bio Valley Malaysia, in the south of Cyberjaya. Bio Valley Malaysia will conduct a wide spectrum of biotechnology-related activities, especially in drug research.

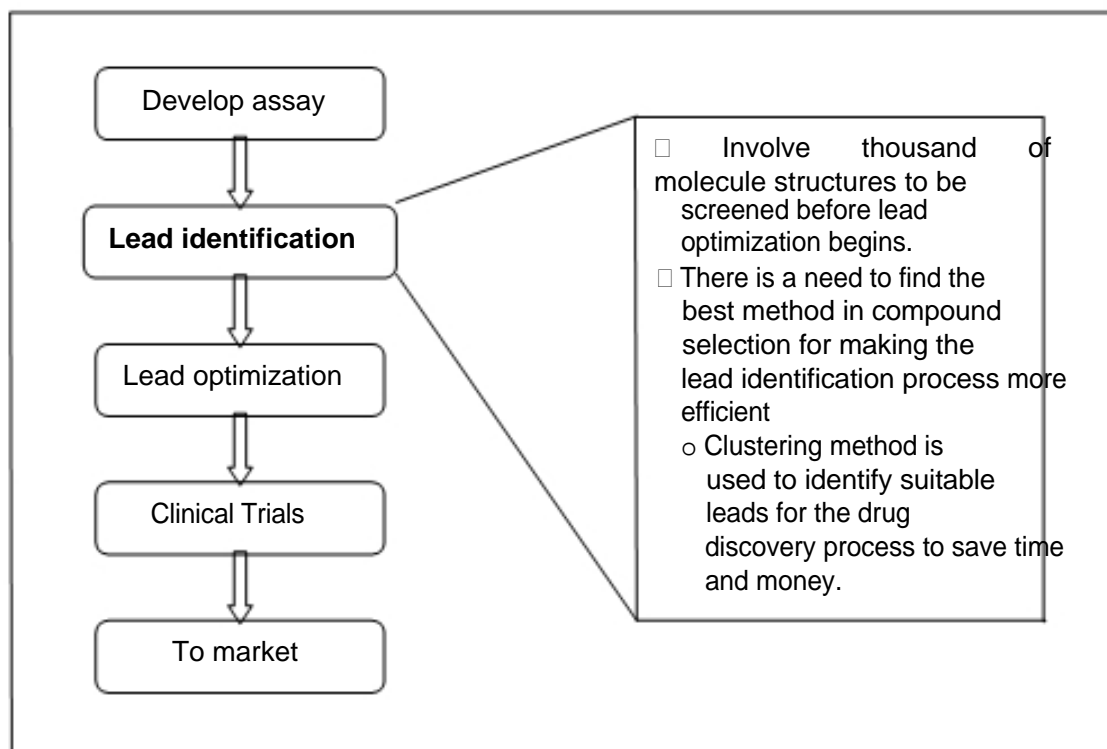


Figure 1.1: Drug discovery process

This project concentrates in the lead identification process. Here, initial leads for drug development will originate from high-throughput screening (HTS), where hundreds of thousands of compounds are tested for biological activity. This slow process of identifying the leads has created constrictions in the drug discovery process, which are time constraint and the huge amount of cost in developing drugs. Because of these constraints, there is a need for rational selection of a subset in the combinatorial chemical library. Here, the maximum amount of information can be obtained just by synthesizing and testing minimum numbers of compounds.

There are many approaches for compound selection such as cluster-based compound selection, dissimilarity-based compound selection, partition-based compound selection and optimization-based compound selection (Salim, 2003). Among these different approaches, cluster-based or clustering has become the most commonly used in compound selection. Clustering is an unsupervised learning problem, where only inputs are available and no target outputs are predefined by the users. Thus, it deals with finding structure in a collection of unlabeled data. It is used to measure the similarity of items in multi-dimensional space. Below are the three main uses of clustering in chemoinformatics for compound selection:

- i) Grouping compounds into chemical series, which is particularly helpful in analyzing large datasets (Wild, 2003).
- ii) Grouping structures which are likely to have similar biological activity (Wild, 2003).
- iii) Choosing small sets of representative compounds from large datasets. The small number of compounds is selected from each cluster to represent the entire dataset to be used as candidates for chemical and biological compound screening (Wild, 2003; Massart and Kaufman, 1993 and Takashi et al., 1980).

By using cluster analysis method, it has helped the researches of finding lead compounds faster and more effectively. Thus, cluster-based is one of the most important unsupervised learning problems in chemoinformatics.

1.1 BACKGROUND OF PROBLEM

One effective way to summarize the content of a chemical database is by using the compound clustering method. This method is a technique to separate the datasets into different groups or clusters where items in one group are similar to each other. According to Downs (2001), clustering is a technique that is being used to understand,

simplify and interpret large amounts of multidimensional data. It has been widely used for researches in biological science. It is likely needed for datasets of chemical structures, since the datasets are likely to be very large and have millions of compounds.

Clusters can be overlapping or non-overlapping. The non-overlapped clustering method is where each compound is a member of exactly one cluster. It has two major categories, hierarchical and non-hierarchical clustering. Most of the clustering methods used in chemical datasets are from non-overlapping method, because the development and analysis of this clustering method is simpler compared to overlapping methods. Willett (1987) has proven that non-overlapping methods are most effective methods for compound selection in his study of comparing Ward's (hierarchical) and Jarvis-Patrick (non-hierarchical).

Most of compound clustering method for compound selection is from the non-overlapping method. The effectiveness of the non-overlapping methods has always been analyzed and compared, in order to find the best clustering method for compound selection. Most pharmaceutical industries are using these methods in the process of drug discovery for lead identification.

There are little focus has been given to overlapping clustering method in term of compound selection. Clusters are overlapping if a compound can be found in more than one cluster, where each compound is a member of all clusters to a certain degree. One example of overlapping clusters is the fuzzy clustering and its effectiveness and efficiency in clustering the compounds should be experimented more and compared to other clustering method in the process of lead identification for discovering new drugs.

1.2 PROBLEM STATEMENT

This study on fuzzy clustering in compound selection is experimented based on their intermolecular dissimilarity of their centroids and their ability to separate active/inactive structures. The different values of fuzziness index and the number of clusters are also experimented to see the effect of these different values to the clusters produced by fuzzy clustering. The combination of fuzzy c-means and Ward's clustering method is analyzed to evaluate the effectiveness and efficiency of both methods.

1.3 OBJECTIVES

Objectives of the project are as follows:

- i) To investigate the fuzzy clustering techniques in chemical database.
- ii) To test the effectiveness of the clusters produced from fuzzy method based on their ability to separate actives/inactives compounds and their intermolecular similarity of their centroids.
- iii) To test the different fuzziness index and their effect on the clusters produced.
- iv) To analyze and compare the result from the fuzzy clustering method with Ward's method.
- v) To combine the fuzzy clustering method and Ward's method to improve the results of both methods.

1.4 SCOPE

The scopes of this project are as follows:

- i) The analysis is done to the National Cancer Institute's (NCI) AIDS.
- ii) String (binary descriptor) is used as representation of the chemical compounds.
- iii) Distance measures between the descriptors are by using the Euclidean distance measures and for intermolecular similarity, the Tanimoto distance measures are used.
- iv) Fuzziness index in the range of 1.1 to 2.0 is tested.
- v) For comparison with overlapping method, the fuzzy clustering method results will compared to the Ward's clustering method.
- vi) For the combination of both methods, the clusters produced by Ward's clustering method will be used as the initial clusters in fuzzy clustering.

1.5 PROJECT PLAN

The project will be carried out in two parts. The first part of the project will focused in understanding of the literature review and the methodology to be used. Here, most time is spent on searching and gathering information from articles in journals. The understanding of cluster-based method in compound selection is important in order to know different methods of clustering. The fuzzy clustering implementation in chemical compounds is also being researched. The first part aim is to have better understanding of compound selection using fuzzy clustering before implementing the project.

For the second part of the project, the implementation of the fuzzy clustering method will be started. The implementation of part two will start with generating descriptors and calculating the distance measure for each descriptor. Then the fuzzy cluster and Ward's programming codes will be written and the results will be analyzed to compare both cluster methods. The combination of both methods will also be done in part two. After comparing the results to see whether fuzzy clustering can produce the better cluster for compound selection, the second part of the report will be written. This will include the experimental result, analysis of results and its comparison with other method.

1.6 CONTRIBUTION

The developed clustering method based on the fuzzy c-means clustering algorithm, was tested and analyzed based on their ability to separate active/inactive structure and the difference of their centroid based on their intermolecular dissimilarity. Thus, the result from the clusters produced gives the information of whether the compounds are suitable to become new drugs. The effectiveness of the clusters produced is also being analyzed based on their different fuzziness index and number of clusters. The comparison of fuzzy clustering method and Ward's clustering method gives better views on the effectiveness of both methods. This has also initiates the analysis on combining both methods where it shows improvement on the effectiveness of clusters produced compared to both methods. The results from all experiments in the project give more diversity in compound selections clustering by using fuzzy c-means clustering method.

1.7 ORGANIZATION OF REPORT

Chapter 1 of this report will give the introduction and the background of problem on why is study is being conducted. Thus, from the background of problem, a problem statement is derived and this will become the aim for this project. It will also give the objectives and scope of study for the project.

Chapter 2 will discuss mainly on the compound clustering and details on the clustering method, its application and types of techniques from the overlapping and non-overlapping method produced in clustering. It will also discuss in detail the fuzzy clustering method as the method to be compared to Jarvis-Patrick or Wards and the application produced by fuzzy clustering. The last part of the chapter will discuss the diversity analysis as the technique to compare all the method in order to find the best method in clustering-based approach in compound selection.

Chapter 3 discussed the methodology used in the project. It will explain in details the basic clustering processes by using fuzzy clustering. It also will discuss the algorithm used to produce the fuzzy clustering. In chapter 4, the result from the fuzzy clustering method will be discussed and analyzed. Here, two types of analyses will be discussed to see whether fuzzy clustering is a suitable method for compound selection. Chapter 5 is the conclusion of the project based from all the previous chapters discussed. It will also discuss the future work and summary from the results of the experiments.

CHAPTER 2

LITERATURE REVIEW

This chapter will discuss the compound selection technique that is chosen for the project, the cluster-based method. In Introduction, we will see the importance of compound selection in chemical studies. Many different approaches of compound selection are introduced and the reason of selecting cluster-based method is mentioned.

The first part of the literature will focus on the clustering method, mainly on the non-overlapping method produced in clustering and the reason why methods from non-overlapping are widely used in compound selection for the last decade. The discussion will also focus on Ward's method because the results from Ward's will be used to be compared and fuzzified with the results from fuzzy clustering technique. Ward's method is chosen because from studies by Brown and Martin (1997), they have found that Ward's method is the best method that is able to separate active and inactive structure. The studies were conducted in identifying the most suitable descriptor and clustering method for the use of compound selection. In their paper, Ward's method was compared to Jarvis-Patrick, group-average and Guénoche method. This was also agreed by Van Geerestein et.al (1997), where they found that Ward's clustering could separate actives/inactives compound in a dataset.

The second part will discuss fuzzy clustering method in details and the studies that were using fuzzy clustering in their experiment of chemical compound. Studies by Feher and Schmidt (2003), Barkó et.al (1999), Guthke et.al (2002) and Rodgers et.al (2004) will be referred as they have successfully used fuzzy clustering in their experiment of chemical, organics, gene clustering and chemical structures, respectively. The last part of the chapter will discussed the diversity analysis and the separation of active/inactive structures as a technique for measuring the performance between fuzzy and Ward's clustering method, and the combination of both methods.

2.0 INTRODUCTION

The main purpose of compound selection in lead identification is because of the existence of millions of compounds. This has made it extremely hard to synthesize all of the library compounds in a short period. It could take a chemist 27 million weeks or 0.5 million years to synthesize 1,000 compounds per week (Tropsha & Zheng, 2002). The similarity of the compound structures will create redundancies in the chemical information contained in the library. This has made the compound selection an important study in chemoinformatics because there is a need to speed up the search in the library compound.

There are many approaches for compound selection such as cluster-based compound selection, dissimilarity-based compound selection, partition-based compound selection and optimization-based compound selection. Table 2.1 summarizes the different types of compound selection:

Table 2.1: Types of compound selection method

Types of Compound Selection Method	Description
Cluster-Based Compound Selection	Techniques used to separate a dataset into groups and clusters, so that the members of one group differ from one another according to a chosen criterion.
Dissimilarity-Based Compound Selection	Techniques are used to identify the compounds that are dissimilar from the selected ones by using some quantitative measure of dissimilarity.
Partition-Based Compound Selection	Compounds that are in the same section in the descriptor space are combined and partitioned.
Optimization-Based Compound Selection	Techniques based on the optimization of a diversity index to quantify the degree of structural heterogeneity in a subset

2.1 CLUSTER-BASED COMPOUND SELECTION

Cluster-based compound selection involves subdividing a set of compounds into clusters and choosing one compound or a small number of compounds from each cluster (Salim, 2003). Clustering was first studied in biological science and it is now being applied to many other areas including chemoinformatics. The items to be clustered in chemoinformatics are the compounds in a chemical database, described by a set of molecular descriptors. It is used to select representative compounds for sample or overview, biological screening and homogeneous subset for StructuralActivity Analysis (Downs, 2001).

Among all of the compound selection approaches, cluster-based compound selection is a useful subset selection based on experiments by Bayada et.al (1999), Brown and Martin (1996), Matter (1997), Taylor (1995) and Van Geerestein et.al (1997). Bayada et.al (1999) experiment using Ward's clustering method, have found that clustering is the best choice for compound selection because it can extract a diverse set of activities from compound file. Ward's method was also agreed to give a better result in experiment by Brown and Martin (1996). They compared Ward's method to group-average clustering, Guënoche clustering from hierarchical-divisive and Jarvis-Patrick from non-hierarchical clustering. While experiment by Matter (1997) shows that hierarchical clustering gives only small difference between median, average and single linkage, when used with 2D descriptors.

Another experiment by Taylor (1995) used cluster sampling based on analysis of nearest neighbors where the molecules that have the highest occurrence in the nearest neighbor lists of other molecules were chosen. After the molecules have been chosen, their own nearest neighbors were excluded from the following selection. This technique tends to get molecules sampled from natural clusters, in the order of the largest clusters down to singletons (Salim, 2003).

Clustering methods can produce overlapping clusters or non-overlapping clusters. Overlapping clusters occur when each compound can exist in more than one cluster, whilst in non-overlapping clusters; each compound belongs to only one cluster. In Figure 2.1, the small blue circles represent items plotted in the 2D space, and the red circles represent "clusters" formed by a clustering algorithm. The three small clusters represent non-overlapping clusters, whilst the two larger clusters show examples of overlapped clusters.

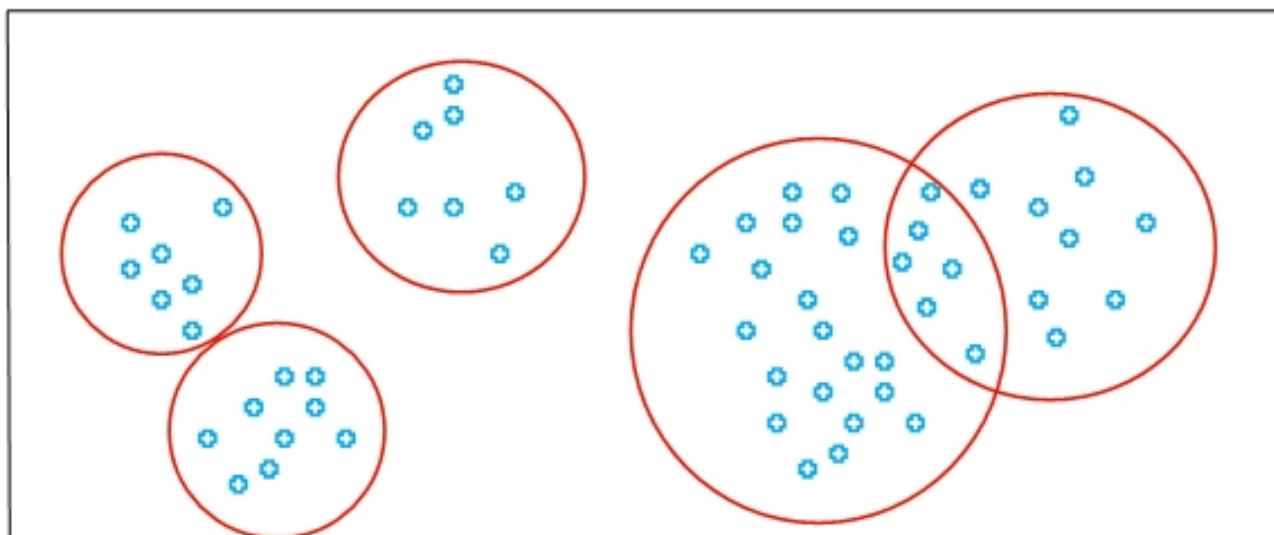


Figure 2.1: Examples of overlapping and non-overlapping clusters (Wild, 2003)

2.1.1 OVERLAPPING METHOD

Clusters can be overlapping or non-overlapping. If a compound occurs in more than one cluster, the clusters are overlapping. At one extreme, each compound is a member of all clusters to a certain degree (Barnard and Downs, 2002). This is an example of fuzzy clustering, in which the degree of membership of an individual compound is in the range 0 to 1. The total membership summed across all clusters is normally required to be 1. Overlapping clusters can be very useful; however, more overlapping will produce more ambiguity, and will be more difficult to interpret (MacCuish and MacCuish, 2003). The diagram is shown in Figure 2.2.

For the past few years, researchers have started to use fuzzy clustering for clustering chemical compounds because it is more realistic than crisp clustering. The comparison between overlapping and non-overlapping methods is important because the effectiveness and efficiency of fuzzy clustering may produce better result in compound selection based on its ability to give membership degree to each compound.

Detailed discussion on Fuzzy clustering method is focused in section 2.4.

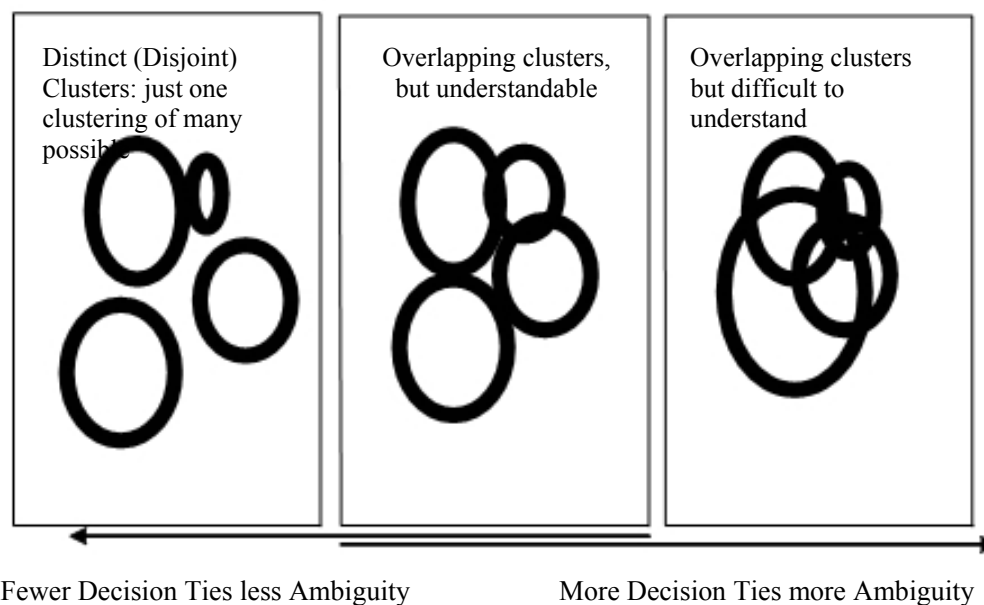


Figure 2.2: Overlapping clustering (MacCuish and MacCuish, 2003)

2.1.2 NON-OVERLAPPING METHOD

Non-overlapping clustering techniques are the most widely used for compound selection (Downs and Willett, 1995). In non-overlapping clustering, the two main non-overlapping clustering methods used in compound clustering are hierarchical methods and non-hierarchical methods. Hierarchical clustering produced method in hierarchical agglomerative method and hierarchical divisive method. The most effective methods in non-hierarchical method are Jarvis-Patrick and K-means. There are more methods produced by the non-hierarchical clustering than hierarchical clustering as in Figure 2.3:

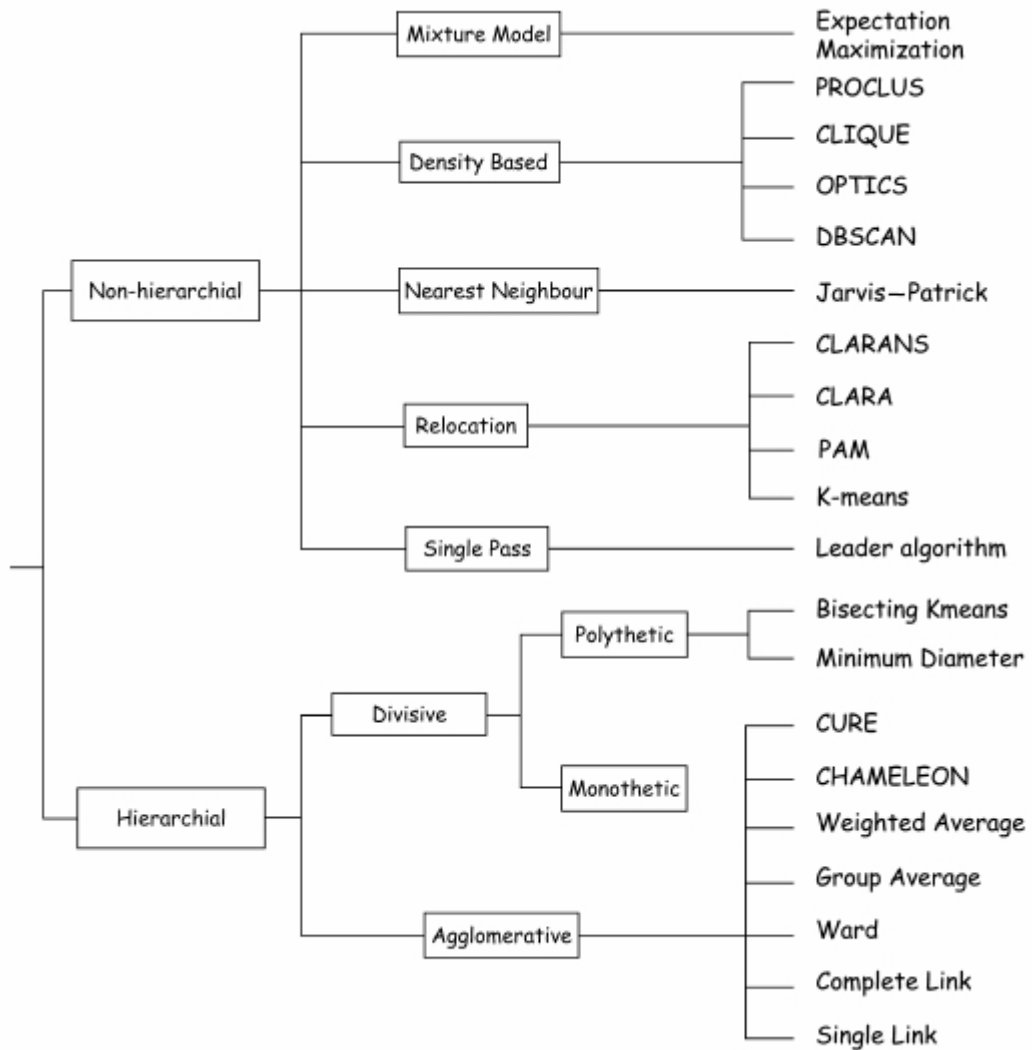


Figure 2.3: A broad classification of the most common clustering methods (Barnard and Downs, 2002)

2.1.2.1 NON-HIERARCHICAL CLUSTERING

Non-hierarchical clustering occurs if the data set is analyzed to produce a single partition of the compounds resulting in a set of clusters. The methods divide a dataset into a number of subsets. These will result in a set of groups with similar

objects in the same cluster is being separated from non-similar objects placed in different cluster. Thus, the clusters have no hierarchical relationships between them.

Non-hierarchical methods cover a wide range of different techniques to build the clusters. The first method is the single-pass method where the partition is created by a single pass through the data set. However, if the partition is randomly accessed, each compound is examined only once to decide which cluster it should be assigned. An example of the single-pass method is the Leader Algorithm, where the cluster is represented by its centroid. The first compound selected becomes the first cluster; a single sequential scan of the dataset and cluster centroids are updated as each compound is assigned to a particular cluster (Barnard and Downs, 2002). According to Barnard and Downs (2002), this method is simple to implement and very fast. However, the disadvantages of the method is that it is order dependent; if the compounds are rearranged and scanned in a different order, then the resulting clusters can be different.

The second method is the relocation method. In this method, compounds are reassigned from one cluster to another in order to improve on the initial estimation of the clusters. Typically, it is accomplished based on improving a cost function (Barnard and Downs, 2002). This is done by guessing where the centers of the clusters are located and then the centers are iteratively refined by shifting the compounds between the clusters until stability is achieved. The best-known relocation method is the K-Means method (Barnard and Downs, 2002), where there exist many variants and different algorithms for its implementation. The K-Means algorithm minimizes the sum of the squared Euclidean distances between each item in a cluster and the cluster centroid.

The nearest neighbor approach is more compound-centered than other nonhierarchical methods (Barnard and Downs, 2002). In nearest neighbor, the environment around each compound is examined in terms of its most similar

neighbor compounds. The criterion used for cluster formation is the commonality between nearest neighbors. Although several nearest-neighbor methods have been developed, the Jarvis-Patrick method is the mostly used for chemical applications.

The next non-hierarchical clustering is the mixture model, where the data are assumed to exist as a mixture of densities. The densities of the data are not known in advance, however usually it was assumed as Gaussian (normal) distributions. The most widely used and most effective general technique for estimating the mixture model parameters is the Expectation Maximization (EM) algorithm (Barnard and Downs, 2002). It finds values of the parameters, which associated with the mixture model, by using an iterative refinement approach. This is almost similar to the K Means relocation method, but the mixture model has not been widely use in the chemical application.

A density-based, or mode-seeking, method is based on the distribution of descriptors across the dataset as generated patterns of high and low density. When these patterns are identified, they can be used to separate the compounds into clusters. Other nonhierarchical methods include topographic and probabilistic methods (Barnard and Downs, 2002). Table 2.2 summarizes different types of nonhierarchical clustering method:

Table 2.2: Different types of non-hierarchical clustering methods

Types	Description	Most used
Single Pass	Partition created by a single pass through the data set.	Leader Algorithm
Relocation	Compounds are reassigned from one cluster to another in order to improve on the initial estimation of the clusters	K-Means method
Nearest Neighbor	The environment around each compound is examined in terms of its	Jarvis-Patrick

	most similar neighbor compounds	
Mixture Model	Data are assumed to exist as a mixture of densities.	Expectation Maximization (EM) algorithm
Density Based	Based on the distribution of descriptors across the dataset as generated patterns of high and low density - used to separate the compounds into clusters	-none-
Topographic	Apply a variable cost function with added restriction that topographic relationships are preserved- the neighboring clusters are close in descriptor space	Kohonen maps
Probabilistic	Generates non-overlapping clusters where the compound is assigned a probability that it belongs to the chosen clusters.	Bayesian

2.1.2.2 HIERARCHICAL CLUSTERING

If a data set is analyzed in an iterative way, where in each step a pair of clusters is merged or a single cluster is divided, the result is hierarchical clustering (Barnard and Downs, 2002). The first method starts with all compounds as single object, known as a singleton, and then merged iteratively until all compounds are in a single cluster. This method is called hierarchical-agglomerative clustering. If the hierarchical method starts with all compounds in a single cluster and iteratively splits one cluster into two until all compounds are singletons, the method is called hierarchical-divisive method.

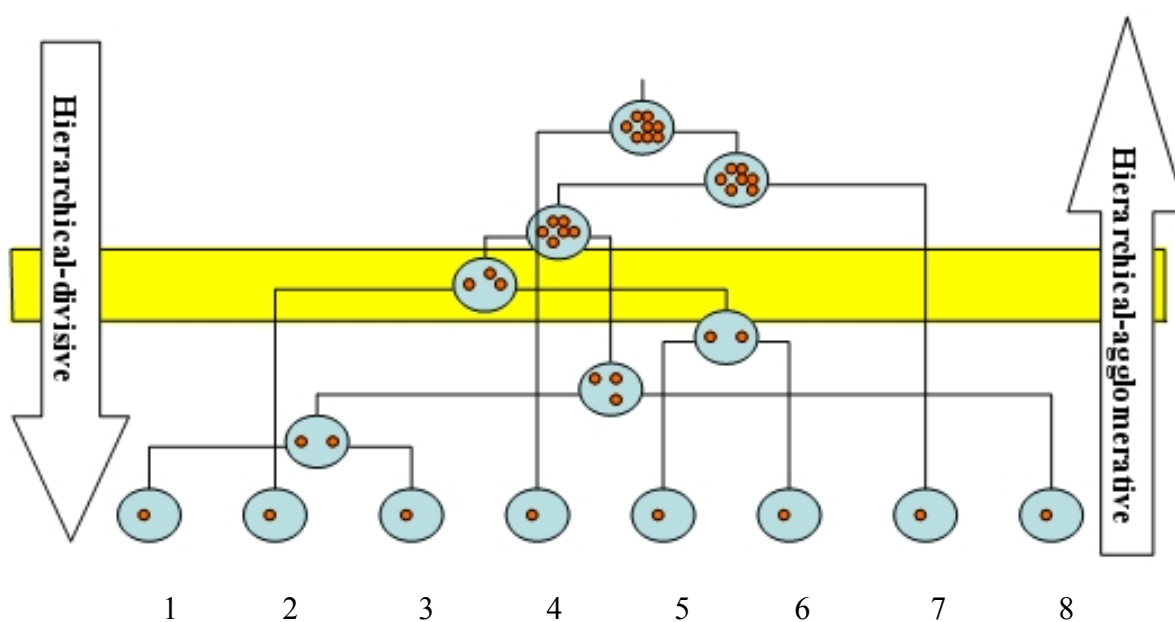


Figure 2.4: Example of hierarchical-agglomerative and hierarchical-divisive method (Wild, 2003)

In the hierarchical-agglomerative method, it starts with each compound in its own cluster (Wild, 2003). The two most similar clusters are merged to form a new cluster and this process will be repeated until all items are merged into one cluster. In Figure 2.4, the two most similar initial clusters are 1 and 3, which are then merged to form a new cluster. After the first merge to form a new cluster, the most similar clusters (8), are then merged with the new cluster. Next, clusters 5 and 6 are merged, this process will continue until we have the cluster at the top of the tree, which contains all of the items.

Table 2.3: Different choices of hierarchical-agglomerative clustering

	Description
Single-Linkage Clustering	<ul style="list-style-type: none"> <input type="checkbox"/> also known as minimum method <input type="checkbox"/> the distance between a pair of cluster to is equal to the shortest distance from any member of the cluster
Complete-Link Clustering	<ul style="list-style-type: none"> <input type="checkbox"/> also known as maximum method <input type="checkbox"/> the distance between one cluster to another is considered to be equal to the maximum distance of any member of the clusters
Average-Link Clustering	<ul style="list-style-type: none"> <input type="checkbox"/> the distance between a pair of cluster is considered equal to the average distance from any member of the cluster <input type="checkbox"/> a midpoint between single-link cluster and complete link cluster
Ward's Method	<ul style="list-style-type: none"> <input type="checkbox"/> most effective clustering technique from hierarchical method <input type="checkbox"/> uses Euclidean distance to find the most similar items

There are different choices of hierarchical-agglomerative clustering, such as simplelinkage, average-linkage, complete-linkage, and Ward's method as in Table 2.3:

The average-link clustering was most commonly used in chemoinformatics and it has been used by Sneath (1966) to classify amino acids based on their physicochemical properties and structural features. Takashi (1980) has also used it to cluster 29 antibiotics based on antibacterial data. The disadvantage of the method is that it produces small clusters of outliers that are unlike every other cluster. All these methods are classified as graph-theoretic or linkage method because it determine the inter-cluster distance by using graph of points in the two clusters.

The second method in the hierarchical clustering is the divisive clustering. Divisive method occurs when the hierarchical method starts with all the compounds in a single cluster, and it iteratively splits the cluster into two until all compounds are singletons. One way to choose a set of clusters is normally by using stopping rules. The stopping rules are used to select the clusters that best represent the populations (Mojena, 1977) and this is represented by the slice across the hierarchy. Any slice should provide a set of non-overlapping clusters that cover all of the items. In Figure 2.4, the slice highlighted would produce two clusters of three items, two clusters of two items, and two singletons. From the divisive method, it will depend on the presence and absence of some chosen features. This means that, after the division of the cluster, only that attribute is present in the clusters, but it will be absent from the attribute in the other new cluster that has been formed.

In hierarchical-divisive clustering, there are two methods on how the cluster splits; namely monothetic and polythetic. Monothetic happens when only one descriptor is used to determine how the cluster is split. Polythetic uses more than one descriptor, but most polythetic methods are slow. According to Barnard and Downs (2002), even though monothetic methods are generally much faster compared to the hierarchical-agglomerative method, it often gives poor performance because they only based on just one attribute. Chu (1974) has made a comparison of a range of classification procedures for 66 structurally diverse molecules as either tranquilizers or sedatives and his research has included monothetic divisive method. For polythetic method, Kaufman et al (1983), Massart and Kaufman (1983) have used the method to cluster coals based on their elemental and mineral compositions.

2.1.2.3 WARD'S TECHNIQUE

Ward's technique links a pair of groups that produce the smallest variance in the merged group. Therefore, for each pair of groups, they are linked and the centroid is determined. The average squared distance to the centroid is calculated and the pair that

produced the smallest variance in the merged group is linked. Thus, Ward's method is classified as geometric or cluster-center method, together with centroid and median method.

Reducibility property concept was introduced by Murtagh (1983), and it was applicable to geometric method. It states that for the merger of two clusters a and b , to form a cluster c , there cannot be another cluster d that is closer to c than to a and b (Murtagh, 1983). Ward's method, implemented using the Euclidean distance, is one of a few geometric methods that satisfy the reducibility property (Barnard and Downs, 2002).

The importance of the reducibility property is that it enables the stored-matrix algorithm to be replaced by the more efficient reciprocal nearest-neighbor (RNN) algorithm that requires only $O(N^2)$ time and $O(N)$ space (Barnard and Downs, 2002). According to Barnard and Downs (2002), RNN algorithm works by tracing paths through proximity space from one point to its nearest neighbor. This is repeated until a point is reached where the nearest neighbor was the previous point in the path. This pair of points is called the reciprocal nearest neighbors.

In 1994, Downs et.al used RNN implementations of the Ward and groupaverage methods to compare methods for clustering compounds based on property data. These two agglomerative methods have been used successfully in comparative studies covering a wide range of non-chemical applications. From this comparison, Ward's showed a consistently reasonable result. Willett (1997) has found that the best result among the hierarchical method was produced by Ward's method. However, this method was not well suited to process large datasets due to the requirement for random access to fingerprints. Brown and Martin's study (1996) confirmed Ward's method superiority for the use in compound selection. Ward's algorithm will be discussed in chapter 3.

2.2 CLUSTERING IN CHEMICAL APPLICATION

Most of the clustering in chemical compounds emphasize on pharmaceutical applications because these applications tend to process very large and high dimensional data sets (Barnard and Downs, 2002). The most widely used of clustering techniques are the hierarchical-agglomerative technique, especially the Ward's technique. This is because Ward's gives the best result in separating actives and inactives structures. The finding was based on studies by Brown and Martin (1996) in their experiment of comparing different cluster method and descriptors for use in compound selection. In the experiment, the performance of group-average clustering and Guënoche method were almost similar and only slightly worse than Ward's. Whereas, Jarvis-Patrick performed the poorest due to the most uneven cluster size and was prone to produce many singletons as well as large diverse clusters (Brown and Martin, 1996).

Van Geerestein et.al (1997) showed that cluster representatives from Ward's clustering provide a significantly better sampling of activity space than random selection. Their research showed that clustering could separate actives from inactives in a dataset. Thus, a cluster containing at least one active compound will likely have more than an average number of other active compounds in the cluster. An example that used the Ward's technique was the CerBeruS, a system that incorporated Ward's clustering and level selection. The system is used for analysis of Johnson and Johnson Company's compound database. The clustering was used to produce smaller, more homogeneous subsets from which one representative compound was selected as a screening candidate. The level selection was used to determine the optimal clustering level (Barnard and Downs, 2002).

Jarvis-Patrick clustering uses a nearest neighbor approach to cluster objects. Since studies by Willett (1987) have found that Jarvis-Patrick to be the best method in clustering, it has become the method used by Daylight Systems to cluster a

chemical database. The Daylight Clustering Package is a set of programs that provide general-purpose clustering of molecules based on their structural connectivity. Jarvis-Patrick clustering has also been used to support QSAR analysis in a system developed at the European Communities Joint Research Center (Barnard and Downs, 1992). The database contains more than 100,000 compounds and it has been clustered using 2D structural descriptors. According to Barnard and Downs (2002), Jarvis-Patrick clustering was used to extract clusters containing sufficient compounds with measured data. This can be used to estimate the properties of the compounds in the clusters that have lack of data.

2.3 FUZZY CLUSTERING METHOD

The goal of traditional clustering is to assign each data point to only one cluster. In contrast, fuzzy clustering assigns different degrees of membership to each point where the membership of a point is shared among various clusters (Fung, 2001). Fuzzy clustering method has been chosen from the overlapping clustering method to be compared to non-overlapped clustering method. This is because fuzzy was expected to perform better, in cases where there are a significant number of outliers, such as molecular dynamics simulations and molecule alignments (Feher and Schmidt, 2003). This is a similar case of compound selection where finding unusual data objects or outliers from the inactive set produced by the clusters can be a result of determining grouping in a set of unlabelled data.

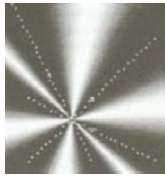
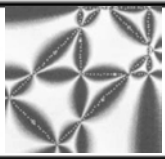
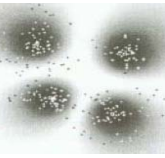
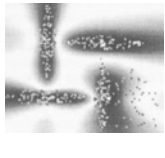
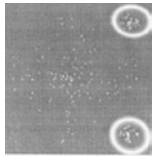
There are few types of fuzzy clustering, such as fuzzy c-varieties (FCV) algorithm, adaptive fuzzy clustering (AFC) algorithm, fuzzy c-means (FCM) algorithm, Gustafson-Kessel (GK) algorithm and Gath-Geva (GG) algorithm. In fuzzy c-varieties algorithm, uses linear subspace of the clustering space as prototypes (Bezdek, 1981). This is useful for detecting lines and other linear structure in the data.

Since the shape of the clusters is often not known, the AFC algorithm is more suitable (Kaymak and Setnes, 2000), where the shape of the clusters is change from point-shaped clusters to straight lines, via elliptic shape.

The third type of the fuzzy clustering is fuzzy c-means clustering algorithm. The shape of clusters produces by fuzzy c-means clustering is determined by the distance measure that is used. Usually, it uses Euclidean distance measure and it is suitable for clusters with spherical shape (Dunn, 1973). Fuzzy c-means algorithm have been used extensively for different tasks such as pattern recognition, data mining, image processing and fuzzy modeling (Kaymak and Setnes, 2000).

Gustafson-Kessel (GK) algorithm is an extension of fuzzy c-means clustering and it used the covariance matrix to capture ellipsoidal properties of clusters. Gustafson and Kessel (1979) extended the fuzzy c-means algorithm for an innerproduct metric norm, where a positive definite matrix is adapted according to the actual shapes of the individual clusters, described approximately by the cluster covariance matrices. The Gath-Geva (GG) algorithm is derived from a combination of the fuzzy c-means algorithm and the fuzzy maximum likelihood estimation (FMLE) (Gath and Geva, 1989). It ignores the objective function and simply replaces it by posterior probability of class given the observation. However it does not give optimal partition in cases of variable cluster shapes and densities. The advantages and disadvantages of different fuzzy clustering are described in Table 2.4.

Table 2.4: Different types of fuzzy clustering algorithm

	Advantages	Disadvantages	Suitable For	Shapes of clusters
Fuzzy C Varieties (FCV) Algorithm	Each cluster represents an r dimensional variety in the dimension of the data space	<ul style="list-style-type: none"> □ The areas of high membership exceed beyond the line segments. □ A higher number of clusters increases the number of local minima 	Lines, planes, and hyper planes.	
Adaptive Fuzzy Clustering (AFC)	Able to recognized elliptic or circular clusters	The eigenvalues have to be computed to update the prototypes, any changes are hardly visible.	Line segments	
Fuzzy C Means (FCM) Algorithm	Few iteration steps already provide good approximation to the final solution	FCM tends to locate centroid in the neighborhood of the larger cluster and misses the small, well-separated cluster.	Spherical shape	
Gustafson Kessel (GK) Algorithm	Faster than AFC. In order to adapt to different structures in data, GK used the covariance matrix to capture ellipsoidal properties of clusters.	The clusters are narrower and the areas with higher memberships are thinner.	Line segments	
Gath-Geva (GG) Algorithm	Unlike FCM and GK algorithm, it is not based on objective function. It is a fuzzification of statistical estimators.	Because the occurrence of the exponential function within the distance, the distance divided into two range, close and remote.	Line segments	

As shown in Table 2.4, fuzzy c means clustering method produces spherical clusters and it does not let a cluster change its shape dependent on the data. Whilst, Gustafson-Kessel proposed a method so a cluster could adapt to hyper-ellipsoidal shapes (Martin, 2003). Experiment by Martin (2003) shows that fuzzy c-means perform better when more number of clusters was used and Gustafson-Kessel algorithm performed best with relatively few clusters. The poorer results obtained by Martin (2003) by using Gustafson-Kessel was assumed to be caused by the clusters being limited to hyper-ellipsoidal shapes. Furthermore, this showed that clusters produced by Gustafson-Kessel did not supplement each other as very well as fuzzy cmeans clustering method. Hence, in among all fuzzy clustering method, fuzzy cmeans using Euclidean distance measures is better used in compound selection.

2.3.1 APPLICATION OF FUZZY CLUSTERING

The study by Barkó et.al (1999) used fuzzy c-means clustering for discrimination of organic compounds using piezoelectric chemical sensor array data of 14 analytes. They applied fuzzy c-means and fuzzy c-lines algorithm for classification and quantitative determination of different volatile organic compounds. Here, the fuzzy clustering algorithm is used to handle the frequency signals of the piezoelectric quartz sensors and all of the sensing materials gave a different response to each analytes. The aim of using fuzzy clustering is to recognize the signals of 14 analytes, where the points related to the analytes are grouped according to their place in n dimensional place. The results were compared using the Principal Component Analysis (PCA) and all of the analytes was successfully identified by PCA algorithm. The fuzzy c-means algorithm has proved to be better in the discrimination of analytes with similar structure, like benzene and toluene (Barkó et.al, 1999). All 14 organic compounds can be distinguished by fuzzy clustering and the similar alcohols, aromatic hydrocarbons and open chain hydrocarbons were easily discriminated.

In another study, Feher and Shmidt (2003) focused on developing and testing algorithm that will provide an applicable clustering approach to deal with a collection of conformers and molecular alignments. Fuzzy clustering was chosen for the study because of the need to select representative conformations or alignments in cases where there are no clear groupings in the data. In this study, Feher and Shmidt (2003) was using five examples, the first three examples shows the application of fuzzy clustering to conformation and the other two examples show the use of fuzzy clustering in flexible alignments. From the result of the study, the alignments process can be optionally incorporated with the quality of alignments using the weighted fuzzy c-means. They also proved that conformers or alignments might belong to more than one cluster with varying degrees of membership. The advantage of fuzzy clustering in this study is that a visual and undistorted representation of the relative differences among conformers is available and visual outliers have little impact on the success rate of the clustering process.

A study by Guthke et.al (2002) used fuzzy c-means algorithm to study the gene expression data and gene functions of the microorganism *Escherichia coli* or *E.coli*. The experiment were conducted by comparing fuzzy c-means algorithm and Gustafson-Kessel algorithm with K-means clustering, Kohonen's self-organizing maps (SOM), Eisen's hierarchical clustering and Quinlan's C4.5 decision tree induction algorithm. Their experiment was using 265 genes that belong to three functional groups.

Among the methods used in the experiments, the highest prediction accuracy was from fuzzy c-means and Gustafson-Kessel with 66.0% and 70.6%, respectively. The accuracy of gene function prediction can be higher using fuzzy technology (Guthke et.al, 2002). Thus it should be favored due to the limited accuracy of gene expression measurements by DNA arrays as well as due to the fact that one gene may be related to more than one physiological function.

The most recent research of fuzzy c-means in chemoinformatics is by Rodgers et.al (2004), where they evaluates the use of fuzzy c-means clustering method for the clustering of files of 2D chemical structures. In their experiment, they used two datasets, Sygenta and Starlist, and compared their findings to K-means and Ward's clustering method.

In their experiments, the results from fuzzy c-means were compared to kmeans clustering and Ward's clustering. The comparison involves simulated property prediction, in which the groupings resulting from a cluster analysis were used to predict the properties of compound within each other (Rodgers et.al, 2004). Their results shows that fuzzy c-means clustering gives the best results as the prediction coefficient values reached up to 0.74, with difference of 0.05 compared with Ward's clustering method 0.03 compared to K-means clustering method.

2.4 EVALUATION OF COMPOUND CLUSTERING FOR COMPOUND SELECTION PURPOSE

The need to ensure coverage of the largest possible expanse of chemical space to search for bioactive molecules means that the selection techniques must aim to maximize the diversity of the library (Bayley et.al, 1999). The evaluation of the effectiveness of the clusters produced by many compound clustering for compound selection can be experimented in many ways. The most used analyses are the ability to separate active/inactive structure, diversity analysis and the ability to group similar compound together.

2.4.1 SEPARATION OF ACTIVE/INACTIVE STRUCTURE

According to Brown and Martin (1996), to select the most suitable clustering method, the clusters produced must cluster together biologically similar structures and separate actives and inactives into different set of cluster. Selecting a representative from each cluster should allow the range of diversity among active compounds to be sampled and number of actives to be missed should be minimized. The degree of separation of actives and inactives in a set of clusters is indicated by the difference of the proportion of structures in the dataset that are actives. Brown and Martin (1996) have given the parameter in calculating the proportion of actives (P_a) in a cluster and the active cluster subset. An active cluster is defined when there are at least one member of a cluster is active. A subset of the active dataset is considered as a set of structures in active cluster (Brown and Martin, 1996).

An increase in the proportion of active (P_a) in the active cluster subset can arise in two ways (Brown and Martin, 1996). First is the result of the clustering itself,

where the active may distribute at no more than one per cluster and thus, inactive clusters will still be formed. Another reason for the increasing of the P_a value is when if there will occur is any greater similarity between pairs of actives than active-inactive pairs. This depends on both the presence of such pairs in a dataset and the ability of a given descriptor to characterize the similarity.

2.4.2 DIVERSITY ANALYSIS

Another evaluation on the clusters for compound selection is by using the diversity analysis, where diversity refers to the degree of structural variation that is present within the set of molecules from a combinatorial synthesis. One of the applications of diversity techniques is subset selection (Wild, 2003). This application requires a small representative set of compounds for a large dataset. It is based on the assumption that the conclusions drawn from the small set is the representative a larger set. Thus, the representative sets represent the variety of the dataset. Diversity measures can be used to compare the diversity of one dataset to another. The measures can be used to see the different of two sets of compounds

Diversity analysis provides a single-number quantification of the degree of structural, property or activity within a dataset. This will give the result of the average inter-molecular similarity where the set that has the minimum average distance to the nearest compound. The representative compound for each cluster is either selected at random or selected as being the closest to the cluster centroid (Bayley et.al, 1999). There are two principal components to a similarity measure (Holliday et.al, 2002). Firstly, the representation that is used to characterize the molecules that are to be compared and this often being a set of descriptors such as 2D fragment substructures or sets of calculated physicochemical properties and secondly the similarity coefficient that is used to

quantify the degree of resemblance between two such representations (Holliday et.al, 2002).

2.4.3 GROUPING SIMILAR COMPOUND

In the ability to group similar compound, the techniques were derived from the similarity searching techniques in chemical databases. Similarity searching involves comparing the set of structural descriptors that characterize a molecule. The molecule will exhibit activity corresponding to the sets of descriptors for each of the database structures. The result of the comparison enables the calculation of a measure of inter-molecular structural similarity and it will determine the structures that are most similar to the target structure or the nearest neighbors.

2.5 DESCRIPTORS FOR CHEMICAL DATABASES

Descriptors are referred to a standardized representation of a molecular feature. A descriptor is a function or an algorithm that accepts a representation of a molecule or an atom as input and outputs some data such as real numbers, bit strings and vectors (Hollas, 2002). It is important to select structural descriptors that are most appropriate for an application and a good descriptor must be able to distinguish between biologically different molecules (Salim, 2003). The descriptors' ability to predict the property or activity of a compound from other compounds will gives the best result in predicting the property values that are very close to the actual values. This will be based on the compounds that are similar to it in term of descriptor similarity.

Descriptors can be classified into 1D descriptors, 2D descriptors and 3D descriptors. Examples of 1D descriptor are physicochemical properties. 2D screens (such as bit strings) and topological indices are examples of 2D descriptors. 3D descriptors consist of 3D screens, potential-pharmacophore-point descriptors or PPP and affinity fingerprints. 3D-descriptor usually changes its values if the molecule shifts to a different spatial conformation.

The use of molecular descriptor is based on the notion that similar molecules generally produce similar biological effects (Gillet, 1999). According to Gillet (1999), the factors affecting the choice of structural descriptors for library design are as follows:

- i) If the descriptor is a good indication of biological activity, the good coverage of biological space can be achieved by covering as diverse range of structural types as possible.
- ii) The speed with which they can be calculated should be fast enough to allow the analysis of the huge numbers of compounds.

For this research, the focus will be on 2D descriptors, that is Barnard Chemical Industries (BCI) dictionary bit string for binary descriptor. 2D descriptors have been chosen because it performs remarkably well in numbers of application (Bajorath, 2001). They are also capable of producing meaningful results in virtual screening. This is also agreed by Brown and Martin (1996) where they compare 2D descriptors available from MACCS, Unity and Daylight with 3D descriptors of Unity and PPP, by using different clustering method. They suggested that all 2D descriptors were able to distinguish actives and inactives better than 3D descriptors. Thus, the results showed that 2D descriptors could be effectively used in similarity calculation to distinguish biological activity. This proves an earlier study by Matter (1997) that showed 2D descriptors are most effective in selecting representative subsets of bioactive compounds.

2.5.1 BIT-STRING

Another example from the 2D description is the 2D screen, where it is categorized as dictionary bit strings and hashed fingerprint. The process converting the molecules into bit string involves splitting a molecule up into fragments and if a particular fragment is present, then a corresponding bit is set in the bit string (Flowers, 1997) as shown in Figure 2.5. There are several methodologies exist for chemical binary representations such as Daylight Chemical Information Systems (Daylight), Molecular Design Limited (MDL) and Barnard Chemical Information Systems (BCI).



Figure 2.5: A simple representation of bit string (MacCuish and MacCuish, 2003)

Dictionary bit string is the identity of a fragment that determines if a bit is set and it is based on a predefined dictionary of fragments and the presence and absence of the fragment in a structure. This means that a fragment can be present once or 100 times, but it would still only set one bit because the bit string does not determine its quantity. It is the number of different types of fragment that determines the number of bits set in a fingerprint (Flowers, 1997). The similarity between two structures is determined from the number of fragments they have in common (Gillet, 1999). The fragments tend to be either specific functional groups or substructures such as carboxylic acids, or different linear atom paths through the corresponding molecular graph such as in Figure 2.6:

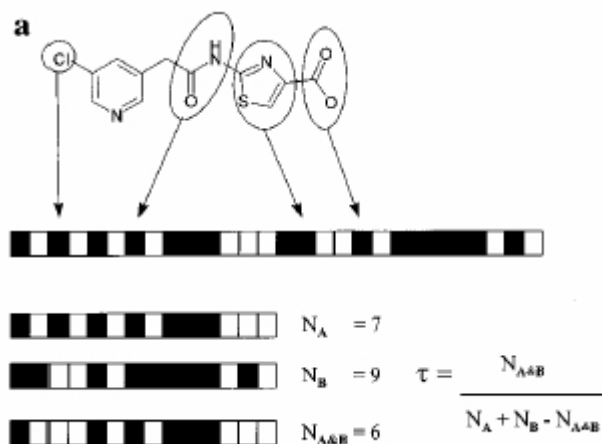


Figure 2.6: Encoding chemical structure as a bit string (Flower, 1997)

A more detailed representation of how bits are set is shown in Figure 2.7. The molecule is decomposed into a set of atom paths of all possible lengths and each of these paths is then mapped to a bit set in a corresponding binary string.

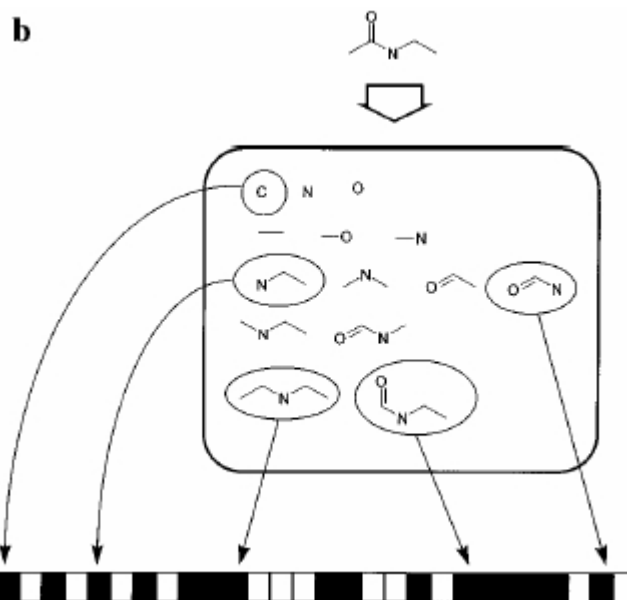


Figure 2.7: A detailed encoding of a bit string (Flower, 1997)

An example of the use of dictionary bit string is in the Barnard Chemical Information Systems (BCI), which combines a bit of both of the Daylight and MDL

approaches. BCI bit string is a 1052-bit structural key-based bit string generated based on presence and absence of fragments in the BCI's standard 1052 fragment dictionary, which encodes augmented atoms, atom sequences, atom pairs, ring components and ring fusions descriptors (Dittmar *et.al*, 1983). BCI dictionary could generate thousands of keys, resulting in molecular fingerprint bit lengths approximately 5,000 bits (MacCuish and MacCuish, 2003).

Another type of the 2D screen descriptors is the hashed fingerprint. It was designed to remove the disadvantage associated with structural keys (Gillet, 1999). This will allow for more generalization where the unique fragments that exists in a molecule is hashed using some hashing function to fit into the length of the bit string. The hashed fragments encode all unique linear, branched, and cyclic fragments, including overlapping fragments (Flower, 1997). Then, each fragment is mapped to an integer randomly in the range 0 to $(2^{31}-1)$ and the integer generated is unique and reproducible for each unique structure.

Daylight algorithm is an example of a hashed fingerprint. A molecular fingerprint is generated from a hash of all the unique connection paths or a subgraph up to a maximum size (typically 8) into a fixed length bit string and the fingerprint learned from the structures themselves (MacCuish and MacCuish, 2003). Typical sizes for Daylight fingerprints are 512 or 1024 bits in length. Molecular Design Limited (MDL) created a key-based fingerprint. This fingerprint uses a pre-defined set of definitions and creates fingerprints based on pattern matching of the structure to the defined key set (MacCuish and MacCuish, 2003). MDL fingerprints could take on a maximum bit length of 966.

2.6 DISCUSSION

In compound selection, the cluster-based or clustering method, especially the non-overlapping clustering, is widely used. Many studies have focused on nonoverlapping clustering method in chemoinformatics since the 1980s; hence, most of the clustering methods that have been widely used are from this approach. Methods that have become particularly popular for clustering chemical structures include Ward's clustering, a hierarchical method and Jarvis-Patrick clustering, a nonhierarchical method (Bajorath, 2001). Hierarchical and non-hierarchical clustering are the two major categories from the non-overlapping clustering.

In comparing the clustering method, the method's efficiency in terms of computational complexity is considered one factor of choosing the best method. Other than the computational efficiency, there are other factor have to be taken into account in choosing the best clustering method for compound selection. These factors are as follows:

- i) The clustering ability to recover the natural clusters that exist in the dataset.
- ii) Their effectiveness in gaining the desired results from their intended applications.

Among all the clustering techniques the most effective are Jarvis-Patrick and Wards techniques, and they have become the choice for large datasets (Downs, 2001). Jarvis-Patrick is a method from the non-hierarchical clustering and it is was popular in the early days of chemical information clustering as it is very fast but the cluster produced are not considered to be of as good quality as Wards and K-means (Wild, 2003). Nevertheless, Jarvis-Patrick method is very computational efficient because the calculations of intermolecular and inter-cluster similarity are simpler. It is used for the generation of nearest neighbor list that can be broken into large

number of small clusters since they are independent of one another. New compounds can also be added without having to re-cluster the whole dataset.

Ward's method is a hierarchical clustering method, where the cluster size distributions are more even than Jarvis-Patrick, but the clusters tend to be spherical (Downs, 2001). In Ward's method, the Euclidean distance is used to determine distances between points. The cluster centroid can be taken as the point in the middle of the distance between the two initial points. Studies also indicate that Ward's methods are better than Jarvis-Patrick for property prediction (Downs et al., 1994) and better than other hierarchical clustering for active/inactive separation (Brown and Martin, 1996).

Brown and Martin's experiment (1996) has compared various clustering methods (including Jarvis-Patrick and Ward's method) for compound selection using various 2D and 3D fingerprints. Their assessment was based on the degree to which clustering separates active from inactive compounds and they have found that Jarvis-Patrick performed the poorest but fastest, which means that, even though Jarvis-Patrick has been widely used for structure-based clustering for compound selection, it may not be the best method for compound selection. There are possibilities that an inactive compound might be selected as representative of one or more actives. The results show that Ward's method gave the best and most consistent results.

In the last few years, fuzzy clustering from overlapping clustering has been used in chemoinformatics. This is because fuzzy clustering represents the real world situation where a compound may belong to several clusters simultaneously with different degrees of membership (Barkó et al., 1999). In many real situations, fuzzy clustering is more natural than other non-overlapping clustering, as objects on the boundaries between several classes are not forced to belong only in one class.

The use of fuzzy clustering has been implemented in the study of separation of malignant and benign tumors and the study of selecting representative conformers (Luke, 2000) and molecular alignments (Feher and Schmidt, 2003). These two studies have shown that fuzzy clustering gives better results in term of having little impact on the success rate. The latest study in fuzzy c-means have also proved that based on simulated property prediction, fuzzy c-means method was at least as effective as traditional, crisp clustering based on 2D fingerprint (Rodgers et.al, 2004). Thus, fuzzy clustering should be tested in term of efficiencies and process time, in the compound selection environment more seriously because the result may gives us other choices in searching the compound libraries.

2.7 SUMMARY

In this chapter, we can see that most clustering used in chemoinformatics focused to the non-overlapping method, where most methods have been tested for efficiency and compared to find the best and fastest method for compound clustering. These include the study of Ward's and Jarvis-Patrick, where both methods have been proven that they are the best method for compound clustering (Brown and Martin, 1996; Downs et.al, 1994; Willett, 1987; Willett et.al, 1986).

However, the efficiency of fuzzy clustering has never been tested in the compound selection environment or compared to the other clustering methods. The need to this comparison is that the results from the comparison will give a variety of compound selection method, especially in cluster-based selection.

CHAPTER 3

EXPERIMENTAL DESIGN

In this chapter, the methodology to implement the fuzzy clustering and the analysis that is done to the results from the clusters produced is discussed. The first two part of the chapter discussed the descriptor chosen and the similarity measures used to calculate the distance matrix. Here, the result of the distance matrix will be fed as input to the fuzzy clustering. Then the fuzzy clustering algorithm is explained in detail and the Ward's algorithm is discussed as the method to be compared to the fuzzy clustering method. The analysis of the active and inactive compounds in each clusters and their inter-cluster dissimilarity between centroids produced by each clustering methods is compared to find the best clustering method. The same analysis is to the results of the fuzzification of Ward's clustering method to evaluate their effectiveness in clustering

3.0 INTRODUCTION

The current main use of clustering for chemical datasets is to find representative subsets from high throughput screening (HTS) and combinatorial chemistry. Another use is to increase the diversity of in-house datasets through selection of additional compounds from other datasets (Downs and Barnard, 2002). Overall, the process of clustering involves four basic steps:

- i) Generate appropriate descriptors for each compound in the dataset
- ii) Select an appropriate similarity measure
- iii) Use an appropriate clustering method to cluster the dataset
- iv) Analyze the result. Repeat the clustering process or select only the best clusters.

These four steps will be the methodology used in this project. However, for step (iii), fuzzy clustering has been chosen for this project. The fuzzy clustering method will be based from the fuzzy c-means algorithm. Fuzzy c-means clustering is an extension of classic K-means using the concepts of fuzzy logic and it is the most prominent fuzzy clustering algorithm. It uses Euclidean distance measures to produce spherical clusters and it does not let a cluster change its shape dependent on the data used in the experiments (Kaymak and Setnes, 2000).

Fuzzy c-means is currently being used in clustering, referring to the work from Feher and Shmidt (2003), Barkó et.al (1999) and Guthke et.al (2002). The fuzzy c-means algorithm has been chosen because of its ability to produce the best clusters by identifying the cluster centroid and their corresponding degree of membership until the threshold is minimize. Figure 3.1 shows the flowchart for the algorithm. To improve the results produced by fuzzy c-means and Ward's clustering method, both methods were combined where the clusters produced by Ward's clustering is given membership degree. This is to see if both methods combined can produced better cluster based on

their separation of active/inactive structure and the inter-cluster dissimilarity of the centroids.

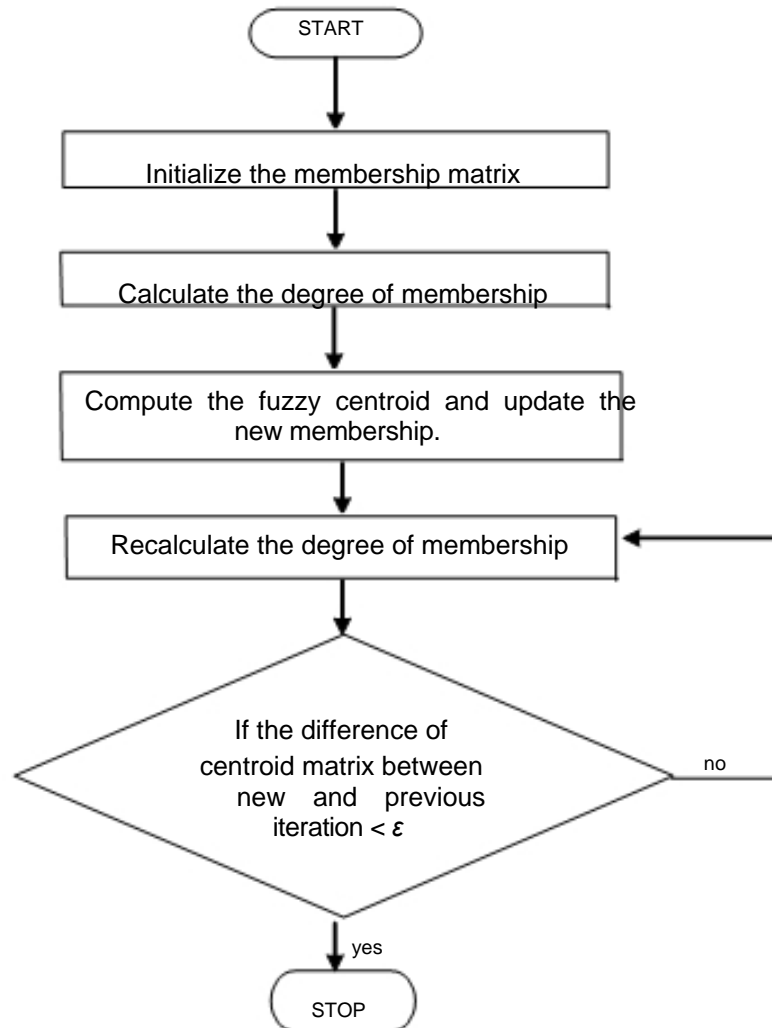


Figure 3.1: Flowchart for fuzzy clustering algorithm

3.1 DATASET

The datasets used in to test the clustering methods is the AIDS dataset obtained from the National Cancer Institute (NCI) in which data are cell-based assay measuring protection from HIV-1 infection. The effectiveness of the clusters produced will be

tested based on the clusters ability to separate actives and inactives compound into different set of clusters. It will also be tested for its effectiveness based on their inter-cluster dissimilarity by using the Tanimoto measures.

In the AIDS dataset, there are 5772 molecules and they are categorized as confirmed active (CA), confirmed moderately active (CM) or confirmed inactive (CI) in each group and molecules. There 247 molecules of CA, 802 molecules that are CM and 4723 CI molecules. However, for this project, only 1000 molecules are tested and analyzed, where it consist of 247 of actives molecules (CA) and 753 inactive molecules (CI).

3.2 GENERATION OF DESCRIPTORS

Descriptors are used in computational chemistry for tasks such as similarity analysis, clustering, and quantitative structure-activity relationship (QSAR) studies. This is because a good descriptor must be able to distinguish between biologically different molecules. Thus, the choice of descriptors plays a crucial role in the analysis of chemical screening data (Root et.al, 2002). Descriptors may include property values, biological properties, topological indices and structural fragments (Downs and Barnard, 2002). Only one type of descriptor is chosen for this project that is bit string (BCI) for binary descriptor. These descriptors are from the 2D descriptors, as mentioned in Chapter 2. BCI descriptors are chosen for binary descriptor because it combines the best strategies from Daylight fingerprint and MDL (MacCuish and MacCuish, 2003).

3.3 SELECTION OF SIMILARITY MEASURE

For similarity measures, we will calculate the distance matrix that choose a subset of the compound space which consist only compounds which have sufficient number of close neighbors. This is obtained based on the descriptor chosen in the earlier step. The similarity measures often used in calculation of similarity between chemical compounds are Euclidean measures, Tanimoto measures and Cosine measures. The similarity measure chosen is the Euclidean distance, which is based on the triangle inequality. Euclidean measure is chosen because it shows that it was best used in fuzzy clustering based on studies from Feher and Shmidt (2003).

Euclidean distances are usually computed from raw data and the advantage of this method is that the distance between any two object is not affected if we add new objects (such as outliers) into the analysis. The similarity measures using Euclidean distance is measured based on inter-point distance $d(x_1, x_2)$ and the equations for binary descriptor is as follows:

$$d(x, x_2) = 1 - \left(\frac{\sqrt{a + b - 2c}}{n} \right) \dots\dots\dots(3.1)$$

where

a: the number of unique fragments in compound A

b: the number of unique fragments in compound B

c: the number of unique fragments shared by compounds A and B

n: the number of fragments in the compounds

After we measures the distance of the similarity matrix, the result gained will be the input for the calculation of the cluster method chosen.

3.4 THE IMPLEMENTATION OF FUZZY CLUSTERING

The third step of clustering using fuzzy clustering is implementing the fuzzy clustering itself. The fuzzy algorithm that will be used is the fuzzy c-means. The first stage of fuzzy c-means is initialing the centroid for each cluster. The primary centroid \hat{C}_i is chosen randomly, depending on the number of clusters defined in each clustering process. The centroid is then used to compute the degree of membership depend on the definition or input of the distance measure. The calculation of the degree of membership (u_{ij}) from centroid i to compound j in the clusters is derived from the equation (Fung, 2001):

$$u_{ij} = \frac{\frac{1}{d^2(X_j C_i)^{\frac{1}{(q-1)}}}}{\frac{1}{d^2(X_j C_k)^{\frac{1}{(q-1)}}}} \dots\dots\dots (3.2)$$

where

$d_2(X_j, C_i)$: any inner product metric or the distance measure q :

fuzziness index

k : number of cluster

The value of fuzziness index (q) is used mostly in the range of 1.0 to 2.0, however there are other studies that used fuzziness index values up to 5.0. The q value has a large impact on the cluster because when a q value is too low, it will not effectively handle noise in the data and if the value is too high, it will produced very poorly separated clusters (Rodgers, 2004). For this experiment, the q value in range of 1.1 to 2.0 is tested.

From the membership matrix, we will derive the fuzzy centroid (\hat{C}_i) from the equation (Fung, 2001):

$$\hat{C}_j = \frac{\sum_{j=1}^M (u_{ij})^q X_j}{\sum_{j=1}^M (u_{ij})^q} \dots\dots\dots (3.3)$$

where

- u_{ij} : degree of membership q :
- fuzziness index
- X_j : the data point of the j th compound M :
- number of data point

Then, the difference between the centroid matrixes will be calculated based on the distance between the centroids as in 3.4. This process is repeated until the difference reached the predetermined value (ϵ). This termination criterion (ϵ) is usually set to 0.01 (Rodgers, 2004).

$$\sum_i (\hat{C}_{jcurrent} - \hat{C}_{jprevious}) < \epsilon \dots\dots\dots (3.4)$$

This will resulted in minimizing the cost-function. The cost-function (J) equation is as follows (Fung, 2001):

$$J_q(U, C) = \sum_{j=1}^M \sum_{i=1}^K (u_{ij})^q d^2(X_j, C_i); K \leq M \dots\dots\dots (3.5)$$

where

- U : a fuzzy K -partition of the data set C : a
- set o K prototypes (cluster center) M :
- number of data point
- k : number of cluster
- u_{ij} : degree of membership q :
- fuzziness index

X_j : the data point of the j th component

C_i : the centroid of the i th cluster

$d_2(X_j, C_i)$: any inner product metric or the distance measure

The parameter q is the weighting exponent for u_{ij} and it controls the fuzziness of the resulting data. It is always any number greater than 1 and from studies by Feher and Shmidt (2003) and Barkó et.al, (2003), the fuzziness index is always set to 2. However, for this project, the fuzziness index will be experimented in the range of 1.1 to 2.0. The algorithm for fuzzy c-means is as in Figure 3.2.

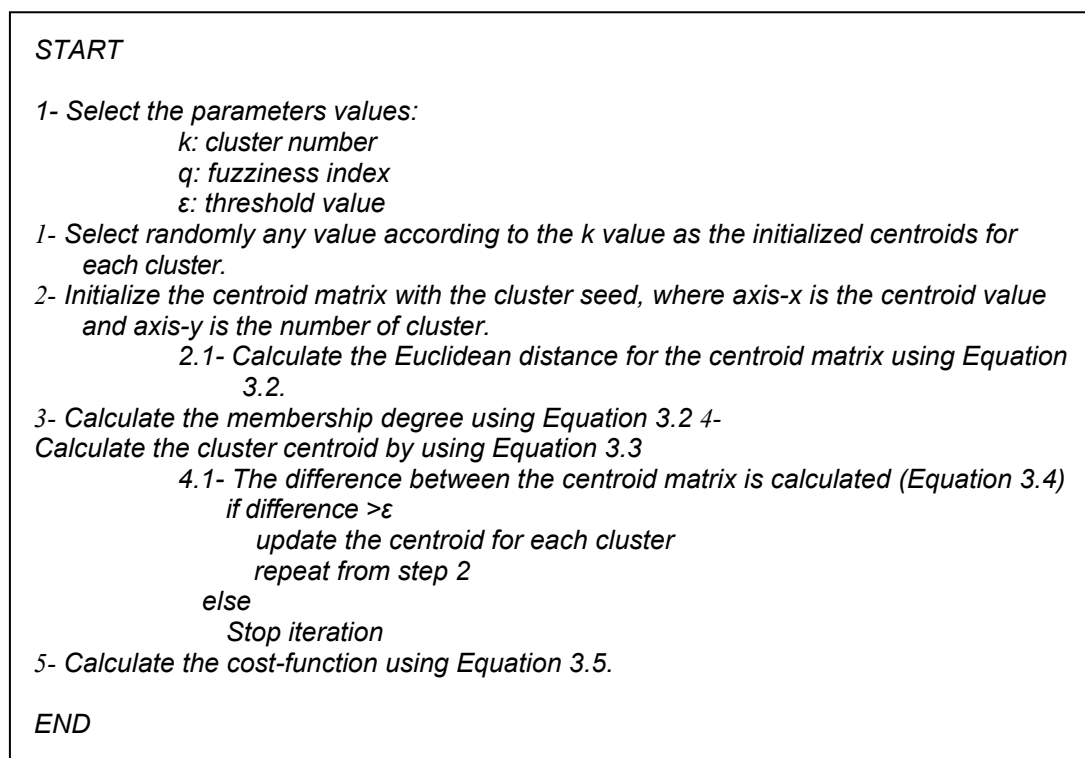


Figure 3.2: Fuzzy c-means algorithm

3.4.1 WARD'S ALGORITHM

Ward's clustering method is implemented by reducing the number of clusters one at a time starting from one cluster per compound and ending which one cluster comprises all the compounds. At each cluster reduction, the method merges the two clusters and this will give the result of the smallest increase in the total sum of squares of the distances of each point to its cluster centroid. Thus, the Ward's algorithm forms clusters by selecting a cluster that minimizes the within cluster sum of squares or the error sum of the squares (ESS).

$$ESS_k = \sum_{i=1}^n x_{ik}^2 - \frac{1}{n} \left(\sum_{i=1}^n x_{ik} \right)^2 \quad \dots\dots\dots (3.6)$$

where:

x_{ik} : the attribute value of molecule i in cluster k

n : size of cluster

The ESS values will be summed together as in:

$$E = \sum_{k=1}^K ESS_k \quad \dots\dots\dots (3.7)$$

where:

K : the number of cluster

The algorithm to perform cluster analysis using the Ward's clustering method is in Figure 3.3:

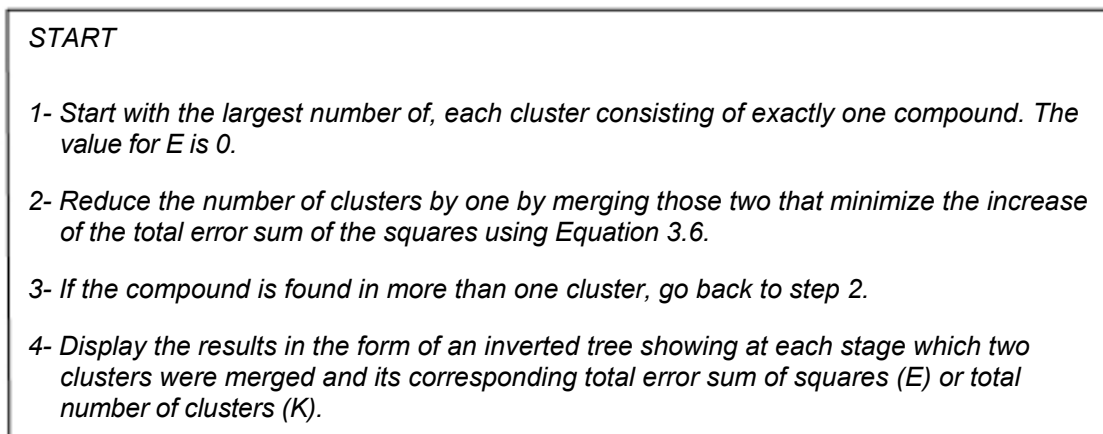


Figure 3.3: Algorithm for Ward's clustering

Ward's clustering method minimizes the variance of groups but it usually took very large of computational resources with time of $O(N^3)$ and memory complexity of $O(N^2)$, where N is the number of objects to be clustered (Borosy *et.al*, 2000). By applying the reciprocal nearest neighbor (RNN) by Murtagh (1983), the memory and time demand can be decreased to $O(N^2)$ and $O(N)$, respectively. Below is the algorithm for Ward's clustering method using RNN.

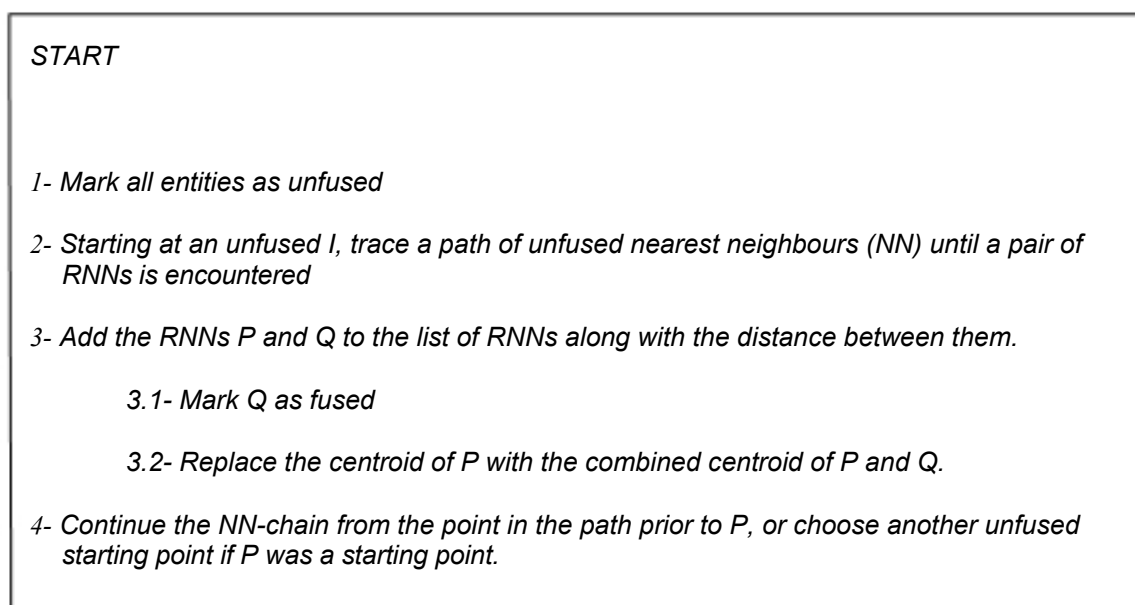


Figure 3.4: Algorithm for Ward's clustering using RNN

The RNN method is applicable to geometric clustering methods where the most similar pair at each stage is defined by a distance measure and the Euclidean distance is used to determine distances between points. The cluster centroid can then be taken as the point in the middle of the distance between the two initial points.

3.4.2 FUZZIFICATION OF WARD'S CLUSTERS USING FUZZY C-MEANS CLUSTERING METHOD

The clusters produced from Ward's clustering method were given membership degree and the centroid for the clusters was based on the RNN algorithm. In RNN, at each stage the pair of clusters is joined with the most similar centroids and these centroids is taken for initializing the fuzzy c-means algorithm.

The clusters' member has also been determined from the results from Ward's clustering. These cluster members are then is given the membership values according to their centroid. The algorithm for fuzzy c-means based on clusters produced by Ward's clustering is as follows:

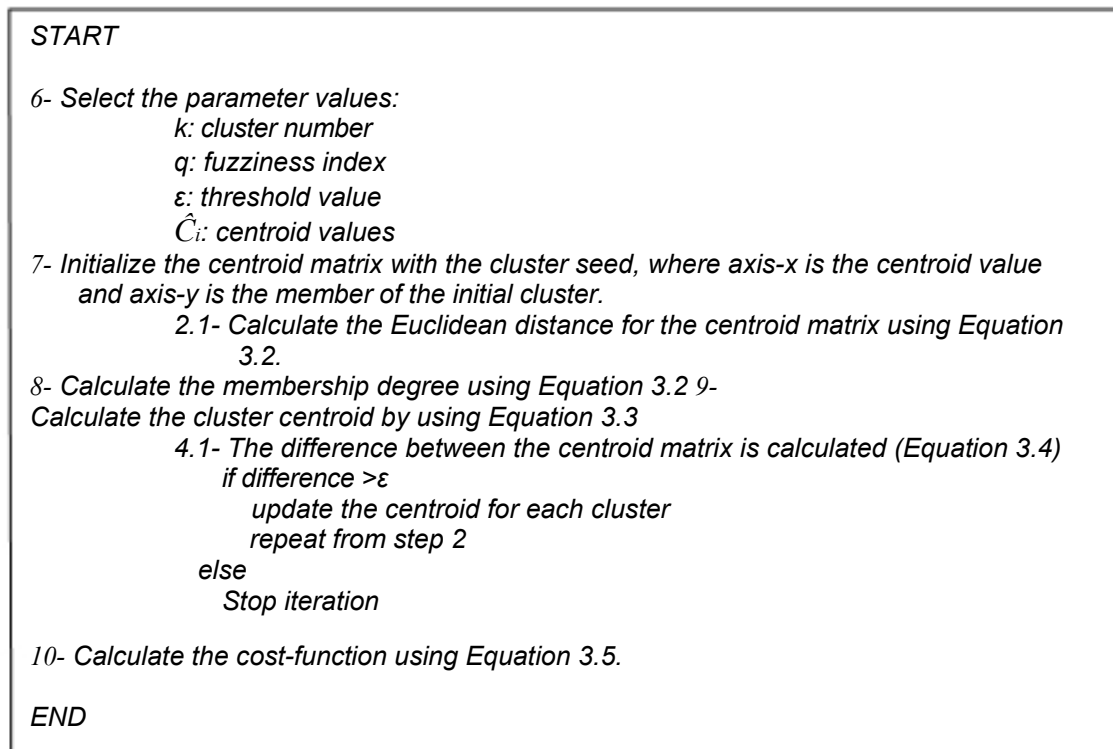


Figure 3.5: Algorithm for fuzzy c-means clustering in Ward's method

The difference on the algorithm is that the initialized centroid values (\hat{C}_i) and the members in each cluster are already determined. By giving membership degree to the member of the clusters, the clusters which are not overlapped earlier, is now overlapped because the members exist in more than one cluster based on their membership degree.

3.5 ANALYZE EXPERIMENTAL RESULTS

After the result from the fuzzy clustering has been obtained, it will be analyzed and compared to other results from the clustering method. The result will be compared based on the effectiveness of the method and the effect of the fuzziness index on the cluster produced. Ward's clustering method has been chosen for the comparison

[]

This will give a measure of relative diversity, of how different the molecules are to each other, and this will be applied to the centroid of the clusters. Thus, the higher the value of MIMD, the centroid of the cluster will become more different.

Another analysis that will be done is based on study by Brown and Martin (1996), the evaluation of the clustering method are based on their ability to cluster together similar structures and their ability to separate actives and inactive into different set of clusters. The active cluster subset must contain compounds from all clusters that have at least one active compound. This will minimize the possibility of having an inactive compound being selected as the representative of a cluster that contain actives compound. The proportion of active structure (P_a) is calculated as:

$$P_a = \frac{\text{no. of actives in dataset}}{\text{no. of structures in active cluster subset}} \dots\dots\dots (3.10)$$

This is the maximum possible proportion of actives for a given size of active cluster subset and this would be obtained only if all active structure were in multimember clusters and none in singletons.

3.6 DISCUSSION

Based from the Fuzzy C-means (FCM) and the algorithm, we can see that no initial clusters are needed since the algorithm calculates the initial fuzzy partition matrix from the distance measure. The computation of the degree of membership u_{ij} depends on the definition of the distance measure $d_2(X_j, C_i)$. The distance measure is obtained from the similarity measure based on the descriptor chosen earlier in the project that is BCI bit string. From studies by Feher and Shmidt (2003), they showed that fuzzy clustering

that fuzzy clustering method can represent a visual and distorted representation of the relative differences among conformers is available (Feher and Schmidt, 2003). While studies by Barkó *et.al*, (1999) showed that, fuzzy clustering method is a reliable of identifying similar compound. Studies from Guthke *et.al* (2002) also showed that fuzzy c-means algorithm was successfully applied to the functional classification of *E.coli* genes (Guthke *et.al*, 2002). The latest study in fuzzy c-means clustering method demonstrates the ability of this method to highlight multicluster membership (Rodgers, 2004). This motivates the project in trying to get the best result of fuzzy clustering for compound selection.

The fuzzy c-means algorithm will be repeated many times until the difference between the centroid matrix from the current and previous iteration is less than a predefined value ϵ . The result from the algorithm will be analyzed to see if the fuzzy clustering can give better result and this result will also be compared to other widely used clustering method such as Ward's clustering method. The comparison will be based on the separation of active and inactive compound in the cluster produced. The greater number of structures in the datasets will give better separation of active compound from the inactive compound in the cluster. The framework of the methodology is as Figure 3.1:

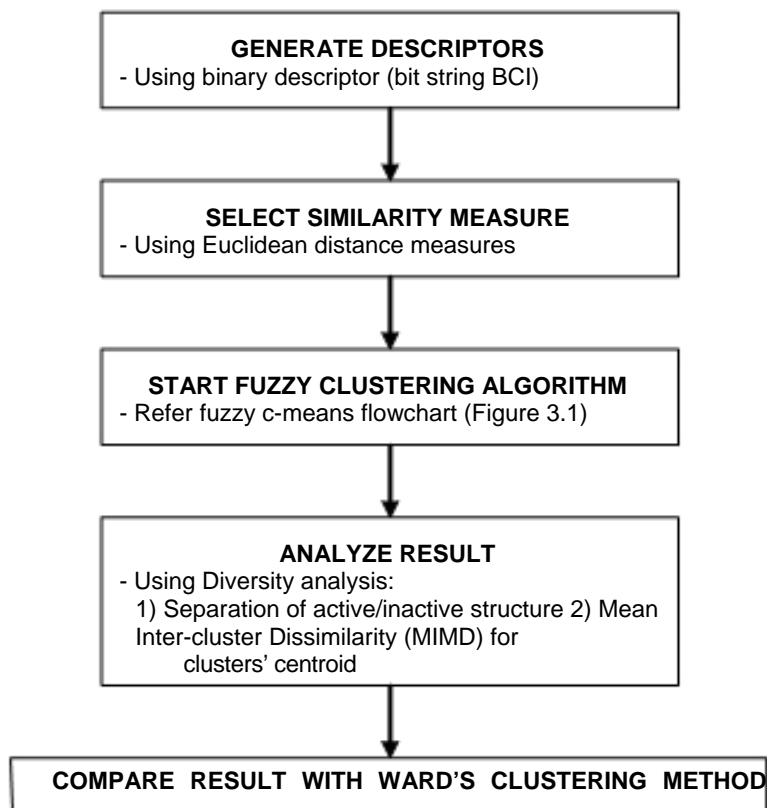


Figure 3.6: Research methodology framework

3.7 SUMMARY

In this chapter, the methodology for the project is defined, where the clustering process is initiated by choosing the descriptor and the similarity measure is calculated. The Euclidean distance measures was selected to measure the distance of the bit string because from studies from Feher and Shmidt (2003) and Guthke *et.al* (2002) where they used Euclidean distance to measure the distance of two molecules.

After we have the distance measures, we initiated the fuzzy c-means algorithm by choosing the appropriate cluster center and calculating the degree of the membership. From the degree of membership, the cluster center is again calculated 55 using the new cluster center and this process is minimized the cost-function. The optimal criterion is determined by using the fuzzy cluster validity measure. The results from fuzzy clustering will be compared from the results from Ward's clustering method. Here, we will compare the clusters produced based on the diversity analysis by measuring the actives and inactives compound in each cluster. The next chapter will discussed the results obtained from fuzzy c-means and Ward's clustering.

CHAPTER 4

EXPERIMENTAL RESULT

Chapter 4 discusses the result from experiments described in Chapter 3. The experiments were conducted using 1000 AIDS data where it consists of 247 active structures and 753 inactive structures. There are two types of analysis been done to the data; analysis based on the ability to separate active/inactive structure in the clusters and the intermolecular dissimilarity between the centroid from clusters produced by fuzzy c-means.

The first part of the chapter discussed mainly on the result produced from fuzzy c-means. Here, the effect of fuzziness index and the number of clusters were analyzed to see the effectiveness of the clusters produced from fuzzy c-means. In the second part of the chapter, results from fuzzy c-means clustering are compared to Ward's clustering from the overlapping method. These also include analysis on the fuzzification of Ward's clustering method using fuzzy c-means.

4.0 INTRODUCTION

The result from the fuzzy c-means clustering will be evaluated based on their diversity analysis. For both of the analyses, different fuzziness index (in range 1.1 to 2.0) and number of clusters (10 to 50 clusters) are used. This is to see the effect of different fuzziness index (q) and number of clusters (k) to the clusters produced. Result from both analyses will be discussed below.

4.1 RESULT OF FUZZY C-MEANS CLUSTERING

The analyses done to results from the fuzzy c-means program are analyzed in two aspects: their ability to separate active/inactive structures and their intermolecular dissimilarity between the clusters. The result was obtained by running fuzzy c-means program by using 1000 molecules (247 actives and 753 inactive) from AIDS dataset. The fuzziness index used is in the range of 1.1 to 2.0 and experimented from 10 to 50 clusters.

The final step of fuzzy c-means clustering involves selecting the molecules to be included in the clusters. The process starts by selecting a minimal membership (μ_{\min}) as the minimum membership function to all the clusters. The results depend on the value that is chosen for the threshold membership function, which was selected in the range of $0.5 \leq \mu_{\min} \leq 0.95$ (Rodgers, 2004). Based on experiment by Rodgers (2004), the highest value of membership function for the clusters peaked at $\mu_{\min} = 0.80$ for all of fuzziness index (q) values, therefore this value was taken for the experiments in this project. For all the structures in the clusters, their memberships are put into descending order. These memberships degree are sum until μ_{\min} are reached; these are the clusters that will be used for both analyses.

4.1.1 ANALYSIS OF ACTIVES/INACTIVE SEPARATION

In the first analysis, this criterion will allow sampling of the range of activities in the datasets and minimize the chances that any activity is missed when an inactive is selected as the representative of a cluster containing actives (Salim, 2003). This means that the more active structure is in a cluster, the higher possibility that an active structure is selected as a representative for further analysis.

Table 4.1: Results for proportion of actives (P_a) from different fuzziness index

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
10	0.2118	0.2081	0.2043	0.2119	0.2087	0.2176	0.2121	0.2118	0.2139	0.2143
20	0.2127	0.2053	0.2074	0.2148	0.2062	0.2122	0.2103	0.2071	0.2130	0.2136
30	0.2104	0.2057	0.2082	0.2106	0.2062	0.2110	0.2084	0.2082	0.2135	0.2114
40	0.2096	0.2067	0.2084	0.2104	0.2070	0.2103	0.2120	0.2086	0.2110	0.2115
50	0.2088	0.2078	0.2082	0.2101	0.2075	0.2095	0.2123	0.2084	0.2123	0.2109

The result for different fuzziness index used is shown in Table 4.1 and it shows that the proportion of actives structure (P_a) becomes higher as the fuzziness index increase. This applied to all clusters, from Cluster 10 to Cluster 50 and the result is shown as in Figure 4.1.

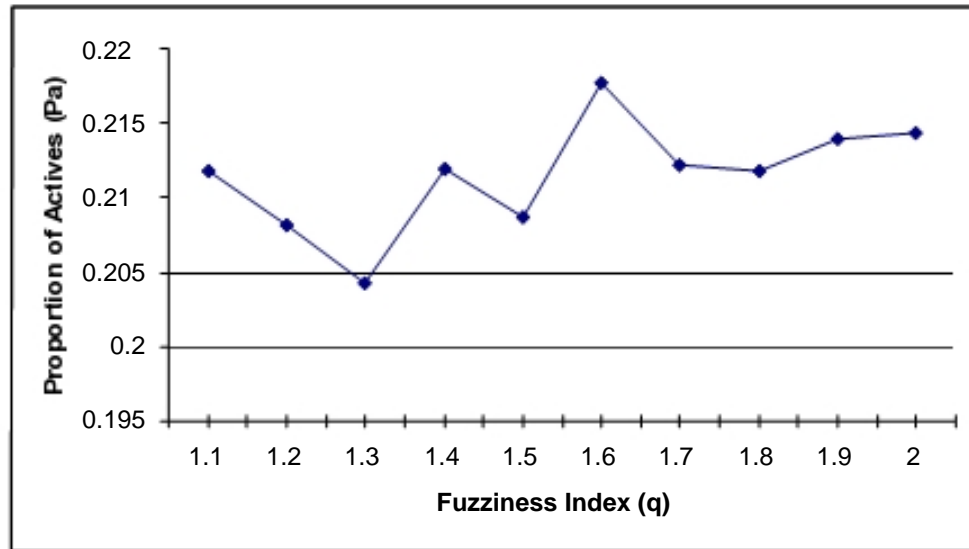


Figure 4.1: Result of Proportion of Actives (P_a) for Cluster 10

The graph shows the increasing of P_a in the fuzziness index for cluster 10. It shows that as the fuzziness index (q) increase, the clusters became fuzzier and the structures become more overlapped. Therefore, the number of active cluster subset also increased. Even though higher fuzziness index (q) gives better result, the result becomes harder to interpret because the clusters are not well separated and becomes too overlapped.

Figure 4.2 shows the result from cluster 10 to cluster 50 using fuzziness indexes (q) 1.1, 1.5 and 2.0. As shown in the graph, the proportions of active structure decreasing as the number of cluster become larger and the highest P_a value is from $q = 2.0$.

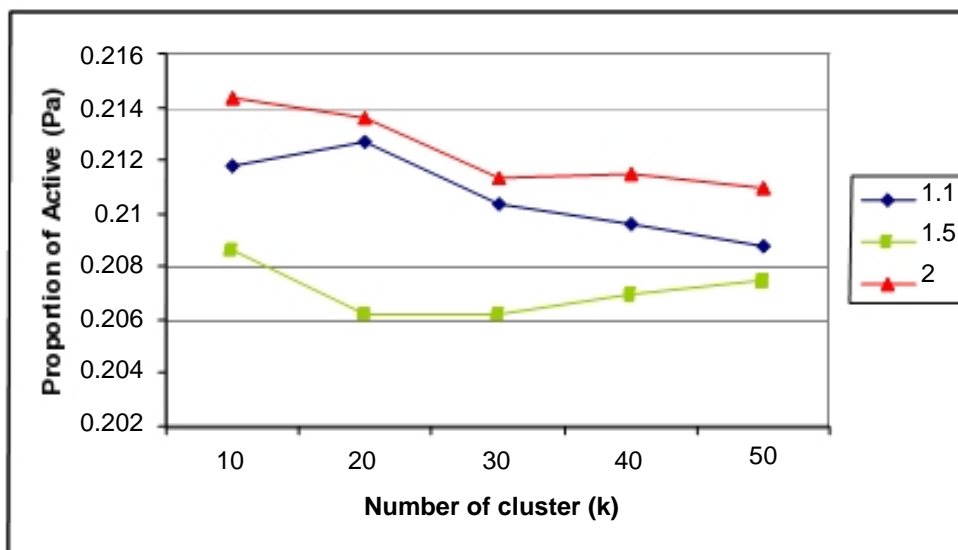


Figure 4.2: Result of Proportion of Actives (P_a) for all clusters using $q = 1.1, 1.5$ and 2.0

As discussed in Chapter 3, the active cluster subset contains compounds from all clusters that contain at least one active compound. This has caused the active cluster subset to increase as the number of clusters becomes larger. All clusters in each experiment from Cluster 10 to Cluster 50 were considered as actives because they contain at least one active structure. Thus, for this analysis, smaller numbers of cluster (k) have better proportion of active structure.

4.1.2 ANALYSIS OF MEAN INTERMOLECULAR DISSIMILARITY (MIMD)

The second analysis is done to test the intermolecular dissimilarity between the centroid of the clusters. This gives a measure of relative diversity of the molecules, to see how different the centroids are to each other. The higher value of the intermolecular dissimilarity shows that the centroids of the clusters are very different between each

other. The dissimilarity measure used in the experiment is Tanimoto distance measure.

The results for fuzzy c-means algorithm are shown as in Table 4.2, where the results are based on different fuzziness index (q). At the lower level of q , the intermolecular dissimilarity values improve as the clusters become better separated. This shows a different impact of the q value as discussed in section 4.1.1.

Table 4.2: Results for intermolecular dissimilarity from different fuzziness index

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
10	0.7422	0.7458	0.6920	0.7354	0.6919	0.7540	0.6941	0.7103	0.7468	0.6907
20	0.7422	0.7394	0.7342	0.7307	0.7181	0.7541	0.7326	0.6927	0.7294	0.7215
30	0.7520	0.7371	0.7399	0.7134	0.7196	0.7530	0.7361	0.6832	0.7277	0.7239
40	0.7497	0.7303	0.7425	0.7182	0.7210	0.7473	0.7090	0.6980	0.7295	0.7285
50	0.7587	0.7260	0.7495	0.7218	0.7274	0.7471	0.7057	0.7087	0.7171	0.7339

This can be seen well in Figure 4.3, where it shows the graph for the mean intermolecular dissimilarity (MIMD) for Cluster 10. Fuzziness index (q) 2.0 gives the lowest value of MIMD, where the centroids of the clusters are more similar to each other and this may not gives better result for further analysis.

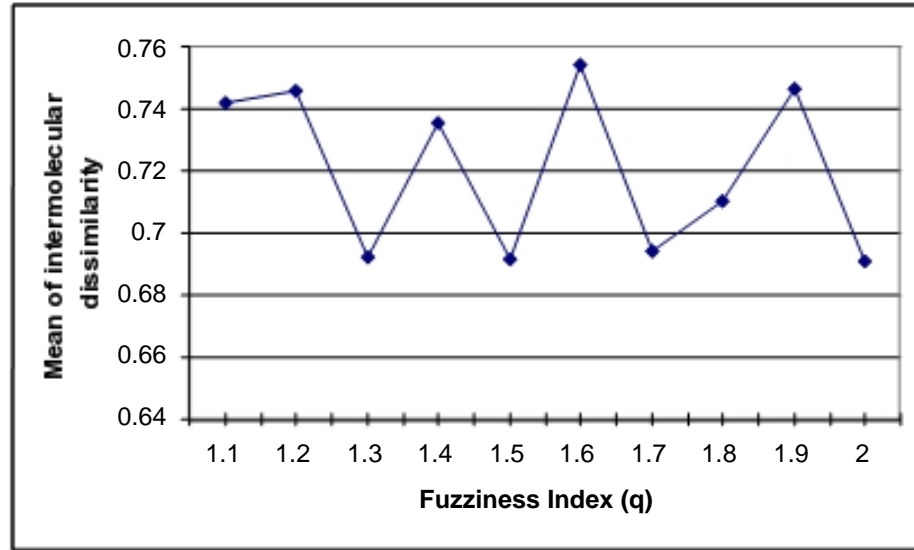


Figure 4.3: Result of mean of intermolecular dissimilarity for Cluster 10

Figure 4.4 shows the result for Cluster 10 to Cluster 50 by using $q = 1.1, 1.5$ and 2.0 . Again, it shows the best result was obtained from the lowest value of q . The graph shows that the MIMD rises rapidly for $q=2.0$, from 0.7422 (Cluster 10) to 0.7587 (Cluster 50). Based from this, the best result is obtained from the largest number of cluster. This is because Cluster 50 has the smallest and tightest cluster of all and thus, the centroid of the clusters are less similar between each other.

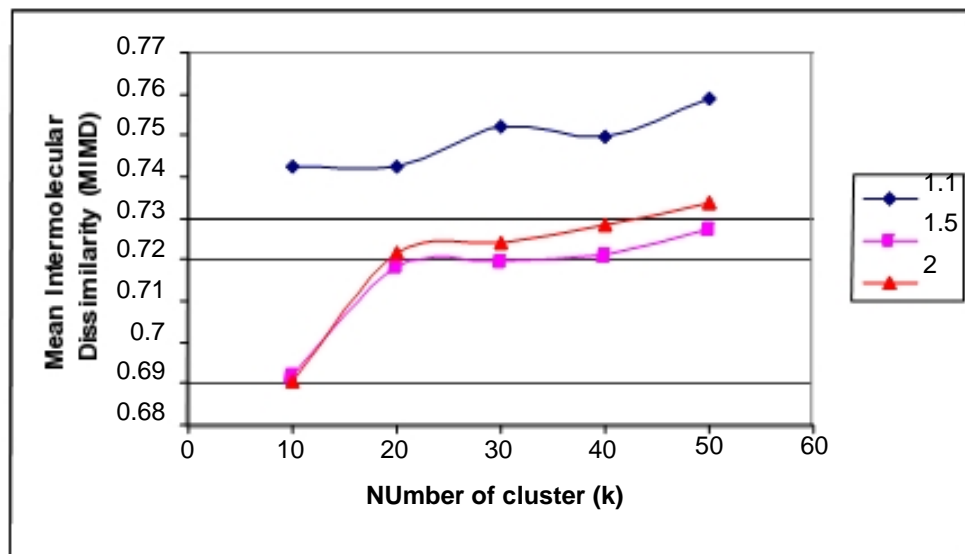


Figure 4.4: Result of MIMD for all clusters (using $q = 1.1, 1.5$ and 2.0)

This shows that the effect of the number of clusters to the result of the intermolecular dissimilarity, where the larger the number of cluster (k), the higher the dissimilarity between the cluster centroid.

4.2 COMPARISON OF FUZZY C-MEANS AND WARDS CLUSTERING METHOD

One of the objectives in this project is to compare fuzzy c-means and Ward's clustering method. The purpose to do the comparison is to test whether fuzzy c-means from the overlapped clustering, can produce better clusters than nonoverlapped clustering. The performances for the clusters produced from both methods are evaluated based on their ability to separate active/inactive structure and their intermolecular dissimilarity.

For both results, the proportion of actives will be used to evaluate their ability to separate the active/inactive structure. For the second analysis, the mean of intermolecular dissimilarity is use to see the difference of the centroids in the clusters produced from both methods. Again, the dataset used for both experiments are from the NCI's AIDS dataset, where 1000 molecules was randomly selected; from which 247 of them are active structures and 753 molecules are inactive structures. For both methods, the experiments were carried out with the number of clusters from 10 clusters to 50 clusters.

4.2.1 ANALYSIS OF ACTIVES/INACTIVE SEPARATION

The results from the fuzzy c-means clustering that were taken for the analysis was from fuzziness index (q) 2.0. The q value was chosen because it produced the highest proportion of active structure compared to other q values. Figure 4.5 shows the graph for the results from Ward's clustering (blue line) and fuzzy c-means (red line) for separation of active/inactive structures. Based from the figure, Ward's gives better result better separation based on the proportion of actives in structure than fuzzy c-means.

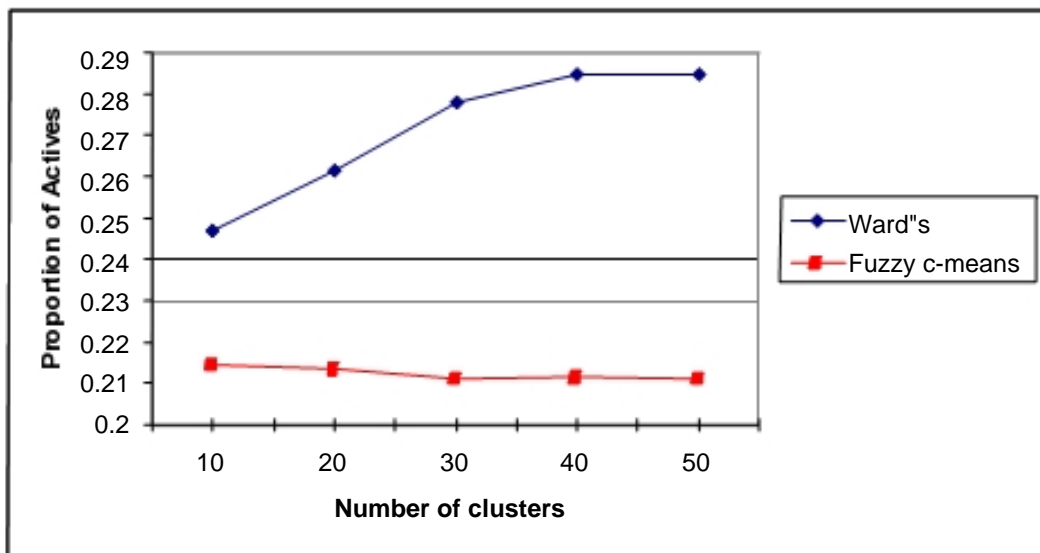


Figure 4.5: Results from fuzzy c-means and Ward's clustering based on their proportion of actives (P_a)

Based on the graph shown in Figure 4.5, the proportion of actives (P_a) from Ward's clustering has increasing value from 0.2470 (Cluster 10) to 0.2849 (Cluster 50). On the other hand, the fuzzy c-means clustering result the P_a value decreased from Cluster 10 ($P_a = 0.2143$) to Cluster 50 ($P_a = 0.2109$). This is because in Ward's clustering, the numbers of active cluster subset decrease as the number of cluster (k) become larger. There may exists clusters that do not have any active structure, therefore the number of structure in the active cluster subset decreased.

The scenario is different for fuzzy c-means clustering; as the number of cluster (k) becomes larger, the number of active cluster subset also increased. This is because, at least one active structure presents in all of the clusters. This has made all of the clusters are considered as active cluster subset, thus the separation for active/inactive structure in larger number of clusters becomes poorer and less effective.

This shows that Ward's clustering gives better separation of active/inactive structure than fuzzy c-means clustering. This may caused by the numbers of structure that exist in fuzzy c-means clusters are in large numbers compared to clusters produced by Ward's clustering. This may caused the clusters produced by fuzzy c-means are harder to interpret because of the large number of structure in overlapped clusters.

4.2.2 ANALYSIS OF MEAN INTERMOLECULAR DISSIMILARITY (MIMD)

In the second analysis of the comparison between fuzzy c-means and Ward's clustering, the intermolecular dissimilarity is used to measure the relative diversity of the molecules. For this analysis the fuzzy c-means results were taken from fuzziness index (q) 1.1, where clusters were better separated and less overlapped. The cluster produced from $q = 1.1$ is considered the best cluster from the fuzzy c-means clustering as discussed in 4.1.2.

For clusters from Ward's clustering, the centroid for each clusters were produced based on the reciprocal nearest neighbors (RNN) algorithm, where it traces a path through the similarity space until a pair of points is reached that are both more

similar to one another than they are to any other points (Salim, 2003). The points are then being combined to form a single new point and it will continue until all points have been combined. In Ward's clustering method, the Euclidean distance is used to determine distances between points. The cluster centroid can then be taken as the point in the middle of the distance between the two initial points.

Figure 4.6 shows the result for mean of intermolecular dissimilarity for both methods. As shown in the graph, the clusters produced from fuzzy c-means gives higher value of mean of intermolecular dissimilarity (MIMD) compared to the Ward's clustering method. For all clusters (except for Cluster 20), the mean of intermolecular dissimilarity value for fuzzy c-means are much higher than Ward's clustering method. The highest difference was 0.0179 for Cluster 50, where the MIMD value for Ward's clustering is 0.7408; compared to fuzzy c-means where it reached the highest peak at 0.7587.

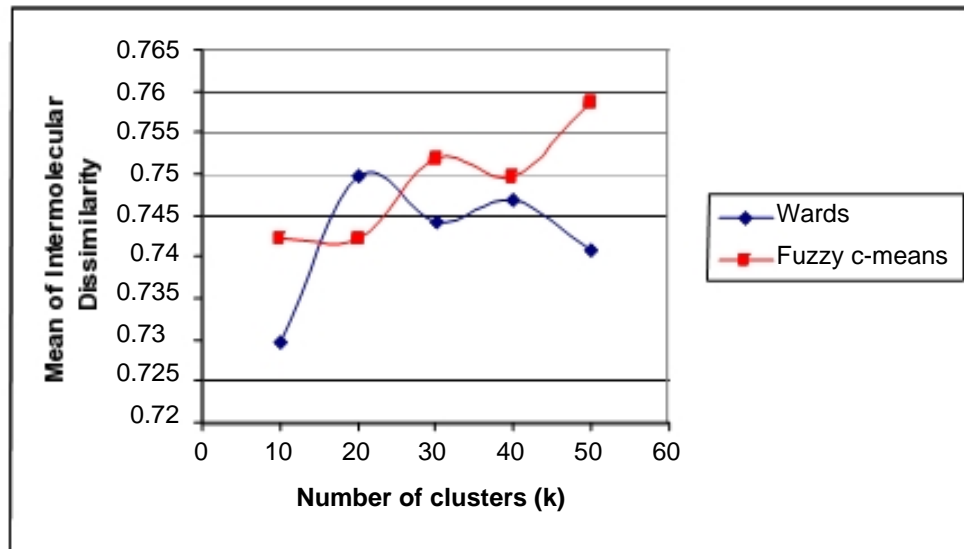


Figure 4.6: Results from fuzzy c-means and Ward's clustering based on their MIMD

This shows that centroid from clusters in fuzzy c-means clustering are far more different between each other than centroid from clusters in Ward's clustering. This may happen because of the centroid from fuzzy c-means clustering that were produced

from equation 3.3 (in Chapter 3) are repeated until the cost function was minimized. The centroids for each cluster were shifted from one molecule to the other and membership degrees for the clusters were recalculated until the iteration stops. Thus, this gives better centroids which were more dissimilar between each other compared to Ward's clustering method.

4.3 FUZZIFICATION OF WARD'S CLUSTERS USING FUZZY C-MEANS

Based on the comparison of both methods, we can see Ward's clustering produced better separation of active/inactive separation based on their proportion of active structures. Whilst, the second analysis shows that fuzzy c-means produced better centroid based on their intermolecular dissimilarity. This has encouraged analyses to be done to clusters produced by combining both methods.

The experimental design for combining both methods is by applying fuzzy cmeans clustering on the clusters produced by Ward's clustering method; in other word, the clusters produced by Ward's clustering method is fuzzified using fuzzy cmeans clustering technique. By doing this, each structure in the clusters produced by Ward's clustering method is given a membership degree. The calculations as in the fuzzy c-means algorithm were applied to each cluster and the results were compared with fuzzy c-means and Ward's clustering methods. Again, the analyses done to the clusters were based on their ability to separate active/inactive structure and their intermolecular dissimilarity. The results for the analyses are discussed below.

4.3.1 ANALYSIS ON ACTIVES/INACTIVE STRUCTURE

For the first analysis, the ability to separate active/inactive structures was compared between the three methods. As discussed earlier, Ward's clustering gives better result compared to fuzzy c-means clusters for this analysis. The low values of proportion of active (P_a) structures in fuzzy c-means clusters were caused by the number of structures in each cluster that were in large numbers. This has also caused for the decreasing value of P_a as the number of clusters become larger, from 10 clusters to 50 clusters.

Otherwise for Ward's clustering, the P_a value increased as the numbers of clusters become larger. The increasing of P_a value also occurred to results from the combination of fuzzy c-means and Ward's clustering method. However, the P_a value is much higher in the combination, as it increased from 0.2940 for Cluster 10 to 0.3789, the highest peak in the graph, for Cluster 50. The result can be seen in Figure 4.6.

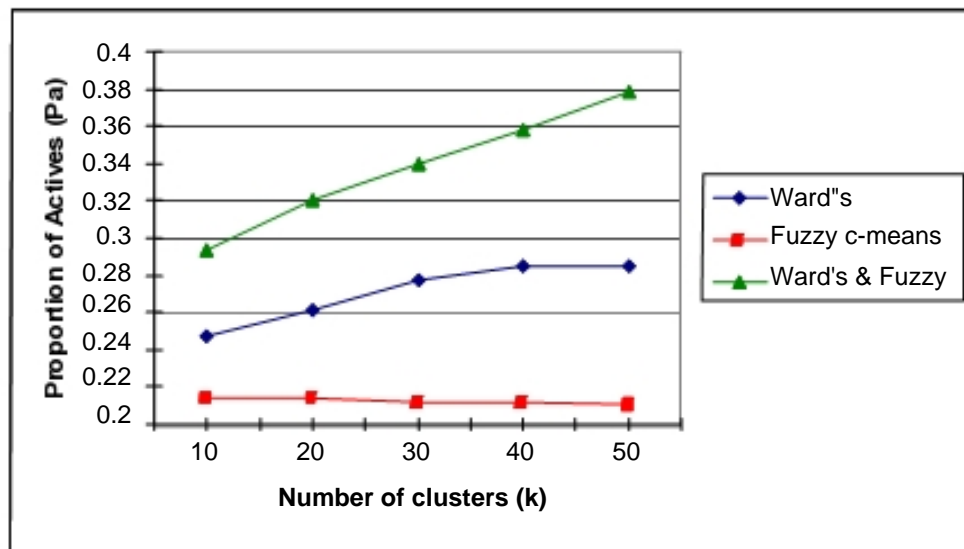


Figure 4.7: Results based on the proportion of actives

This shows that the fuzzy c-means and Ward's clustering results can be improved by combining both of the method. By combining both of the method, their ability to separate active/inactive structures in the clusters improved. This is shown in the graph where the proportion of actives (P_a) gives much higher value than both methods.

4.3.2 ANALYSIS OF MEAN INTERMOLECULAR DISSIMILARITY (MIMD)

Another analysis that is been done to the combining method is the mean of intermolecular dissimilarity (MIMD) for the clusters' centroids. From previous comparison of fuzzy c-means and Ward's clustering, the results shows that fuzzy c means clustering method gives higher value of intermolecular dissimilarity for its centroids. Fuzzy c-means still shows the highest mean of intermolecular dissimilarity compared the other two methods, as in Figure 4.8. The highest peak in the graph is from fuzzy c-means clustering (0.7422) from Cluster 50, with the difference of 0.0008 from the second highest peak from Cluster 20 of the combination method.

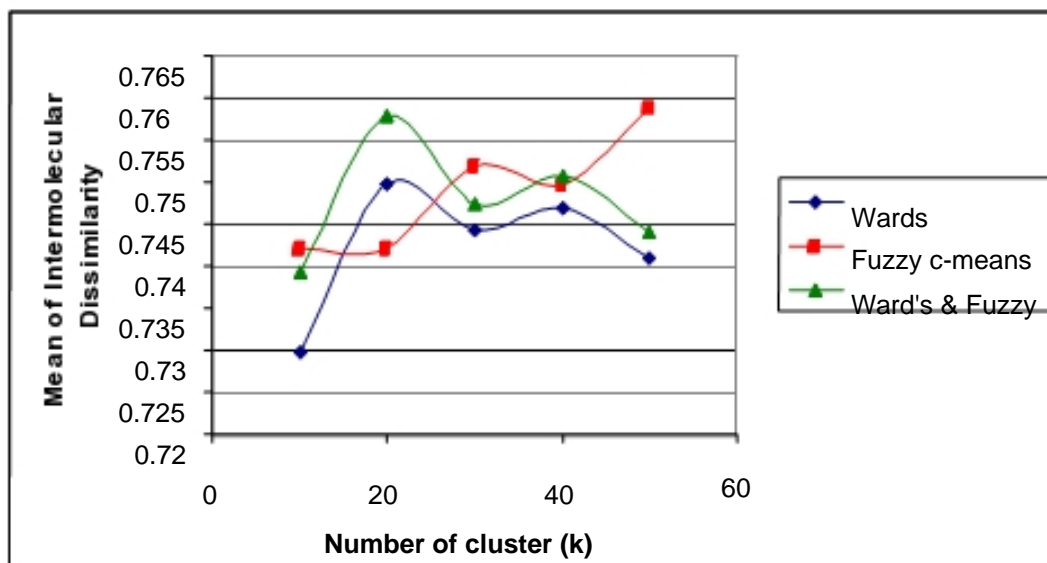


Figure 4.8: Results based on their intermolecular dissimilarity

However, the combination of both clustering method gives better result than Ward's clustering, with the highest difference of 0.010 for Cluster 10. The graph in Figure 4.8 shows that the line for the combination (green-triangle node) and the Ward's clustering (blue-diamond node) has almost the same pattern. This is because, even though fuzzy c-means algorithm was applied to clusters produced from Ward's clustering, most of the centroid from Ward's clustering does not relocate. Only few of the original centroids in the clusters were relocated to a new centroid. As a result, the mean of the intermolecular dissimilarity (MIMD) improved higher than Ward's clustering; however it has the same pattern as the Ward's clustering method.

4.4 DISCUSSION

The experiments were carried out to analyze the clusters produced by fuzzy c-means and Ward's clustering method; to test their effectiveness based on their ability to separate active/inactive structure and their intermolecular dissimilarity.

Based on the first analysis, Ward's clustering method shows better separation of active/inactive structure compared to clusters from fuzzy c-means clustering method. However, fuzzy c-means shows better intermolecular dissimilarity than Ward's clustering method.

For the first analysis, fuzzy c-means clustering shows poor separation of active/inactive structure based on low proportion of active (P_a) value because of the existence of active structures in all clusters. This has caused the number of active cluster subset becomes larger; therefore, the proportion of actives (P_a) becomes lower. The results decrease from $P_a = 0.2118$ to $P_a = 0.2088$, as the number of cluster (k) becomes larger. This was also caused by the larger number of structure in the clusters as it become more overlapped as the fuzziness index (q) increased. Thus, lower value of q gives better result since the clusters were less overlapped. For this reason, only the result from $q = 1.1$ was used for all clusters, from Cluster 10 to Cluster 50 for the comparison Ward's clustering method and for further analysis of the combination of both methods.

Clusters produced from Ward's clustering method gives higher value of proportion of actives (P_a) in their clusters because of the structures are not overlapped in the clusters. As the larger number of cluster (k) being applied to the experiments, the value of P_a becomes higher because lower number of active cluster subset exist in the clusters. There may exist clusters that have none active structure in their clusters and thus the number of active cluster subset decreased. These will gives higher value of P_a as the value k increased.

However, results produced by fuzzy c-means give higher value of mean intermolecular dissimilarity compared to Ward's clustering. This is because of the centroids produced by fuzzy c-means are better than centroid chosen based on RNN for Ward's clustering. This shows that the clusters were more dissimilar between each other in fuzzy c-means clustering rather than Ward's clustering even though the active/

inactive structure were less separated in fuzzy c-means clusters. The value of mean intermolecular dissimilarity (MIMD) also increases as the number of cluster (k) becomes larger.

To have more diversity in the analysis, a combination of fuzzy c-means and Ward's clustering method was evaluated. The combination was done by giving membership degree to the clusters that were produced from Ward's clustering. The membership degree and centroid were obtained from fuzzy c-means algorithm and some of the original centroids were changed. Fuzziness index (q) value of 1.1 was used for the combination method. The results from the combination give better separation of active/inactive structure, and this can be seen as the value of proportion of active (P_a) from the combining method are higher than fuzzy c-means and Ward's clustering.

However, the results based on the second analysis for the combining method gives low value of mean intermolecular dissimilarity than fuzzy c-means and slightly better than Ward's clustering method. The reason for this is the centroids for the combination of both methods still have the same centroid from the Ward's clustering. This can be seen as both methods produced same pattern of the intermolecular dissimilarity. The summary of the results is shown in Table 4.3.

Table 4.3: Summary of results of analyses

		Fuzzy C-Means clustering method	Ward's clustering method	Fuzzy c-means and Ward's clustering method
Separation of active/inactive (P_a)	Fuzziness Index (q)	$P_a \square$ when $q \square$ - Separation of active/inactive becomes better. - Higher value of q gives the best result	-not applied-	-not tested-
	Number of Clusters (k)	$P_a \square$ when $k \square$. - Separation of active/inactive decreased - Lower value of k gives the best result	$P_a \square$ as $k \square$. - Separation of active/inactive becomes better. - Higher value of k gives the best result	
Mean Inter-molecular Dissimilarity (MIMD)	Fuzziness Index (q)	$MIMD \square$ when $q \square$ -Centroids of clusters more similar - Lower value of q gives the best result	-not applied-	-not tested-
	Number of Clusters (k)	$MIMD \square$ when $k \square$ -Centroids of clusters more dissimilar - Higher value of k gives the best result	$MIMD \square$ when $k \square$. -Centroids of clusters more similar - Lower value of k gives the best result	

Based from all of these analyses, we can see that clusters produced from fuzzy c-means method gives better results based on intermolecular dissimilarity but the clusters have lower separation of active/inactive structure. The centroid produced from fuzzy c-means gives better value of intermolecular dissimilarity. However, the existence of active structure in all clusters gives larger number of active cluster subset in the overall clusters. This has caused the clusters less separated as the clusters become more overlapped and fuzzier.

4.5 SUMMARY

Based on the experiments carried out for the analyses, we can conclude that fuzzy c-means gives the best result of clusters produced when compared to Ward's clustering based on their intermolecular dissimilarity. However, the clusters

produced have less value of proportion of active (P_a) compared to Ward's clustering caused by the large number of active cluster subset that exist in fuzzy c-means clusters. For results that combined fuzzy c-means and Ward's clustering method, the results shows improvement in separation of active/inactive structures compared to both method, but for the intermolecular dissimilarity, the result shows slightly improvement than Ward's clustering method. However, the results from fuzzy cmeans clustering still give higher intermolecular dissimilarity.

CHAPTER 5

CONCLUSION

Based on the discussion in literature review, we can see that cluster-based method has been widely used in compound selection. In compound selection, there are four main approaches namely, cluster-based compound selection, dissimilarity-based compound selection, partition-based compound selection and optimization-based compound selection. Cluster-based compound selection is the process of subdividing chemical databases into groups or clusters. The members of one group will differ from one another according to a chosen criterion. As stated by Bayada et.al (1999), Brown and Martin (1996), Matter (1997), Taylor (1995) and Van Geerestein et.al (1997), cluster-based compound selection is the most useful subset selection, thus this has encourage the studies and researches of cluster-based method for compound selection.

One feature of clustering is that the process is unsupervised and clusters can be overlapping and non-overlapping. The clusters are said to overlap when each compound can exist in more than one cluster and it is non-overlapping if each compound belongs to only one cluster. Non-overlapping clustering methods are widely used in compound selection and there are two types of non-overlapping cluster methods, which are hierarchical and non-hierarchical clustering.

In hierarchical clustering, clusters can be agglomerative or divisive. Nonhierarchical clusters produced a single partition of the compounds. According to Willet (1987), Jarvis-Patrick produced the best result compared to the other nonhierarchical method and it is a preferred method in term of computational efficiency. However, comparison among the clustering methods shows that Ward's method from the hierarchical agglomerative clustering, is consistently the best able in term of the separation of actives and inactives (Brown and Martin, 1996).

As far the overlapping clustering methods, the fuzzy clustering is now used in chemical clustering but the use of fuzzy clustering in compound selection has not yet been done. In fuzzy clustering, the degree of membership of a compound is in the range 0 to 1. The different degrees of membership to each compound are shared among various clusters. The results from fuzzy clustering of chemical compound show that the method can produce better clusters (Feher and Schmidt, 2003; Castellano et.al, 2003; Barkó et.al, 1999; Guthke et.al, 2002 and Rodgers, 2004).

Table 5.1: Comparison of results from other studies

Studies by	Dataset Used	Description of Analysis	Results
Barkó, Abonyi, and Hlavay (1999)	Different volatile organic compounds	Comparing fuzzy c-means algorithm with fuzzy c-lines algorithm	<input type="checkbox"/> proved to be better in the discrimination of analytes with similar structure, like benzene and toluene <input type="checkbox"/> All 14 organic compounds can be distinguished
Guthke, Schmidt Heck, Hahn and Pfaff (2002)	265 genes of the microorganism <i>E.coli</i> that belongs to 3 functional group	Comparing fuzzy c-means algorithm with Gustafson Kessel algorithm	<input type="checkbox"/> highest prediction accuracy was from fuzzy c-means (66.0%) and Gustafson Kessel (70.6%)
Feher and Shmidt (2003)	Conformations of: 1) Roseotoxin-B 2) Pentane Flexible Alignments of: 1) Thiorpan 2) Retrothiorpan	Fuzzy c-means clustering for the selection of representatives from assemblies of conformations or alignments.	proved that conformers or alignments might belong to more than one cluster with varying degrees of membership

	3) Estradiol 4) Raloxifene		
Rodgers et.al (2004)	1763 molecules from Starlist database	Comparing fuzzy c-means algorithm with k means and Ward's clustering method	Fuzzy c-means clustering gives best result in simulated property prediction.
Author (2004)	1000 molecules from AIDS dataset	Comparing and combining fuzzy c-means algorithm with Ward's clustering method	Fuzzy c-means clustering gives best centroid for the clusters but did not give better separation of active/inactive structure.

In this study, the fuzzy c-means clustering was chosen to cluster the compounds. This method is based on the cluster center and degree of membership and the process is repeated until the cost-function is minimized. Then the clusters produced will be measured based on the similarity measures and their ability to separate actives and inactives compounds. The result is be compared to the Ward's clustering method and analyzed based on its ability to separate active/inactive structure and the intermolecular dissimilarity between centroids of the clusters.

5.0 THE ANALYSIS OF CONTRIBUTION

As discussed in Chapter 4, the analysis of fuzzy c-means clustering is done by measuring their proportional of actives (P_a) to see their ability to separate active/inactive structure; and also their intermolecular dissimilarity for the centroid in the clusters to see the differences between centroid clusters. For each clusters, fuzziness index (q) of 1.1 was considered as the best q value for all clusters because the clusters produced by fuzzy c-means clustering were less overlapped and also less fuzzy. Thus, the clusters are easier to interpret for further analysis.

However, the results of the analysis shows that fuzzy c-means clustering only gives best result compared to Ward's clustering method based on the intermolecular dissimilarity, as discussed in Chapter 4. The results for separation of active/inactive separation shows less proportion of active for clusters from fuzzy c-means than Ward's clustering. The reason for this was the existence of overlapped active structure in all clusters making the number of active structure subset becomes large. Therefore, fuzzy c-means clustering gives best centroid for the clusters but did not give better separation of active/inactive structure.

Analysis was also done to the clusters that were produced from the combination of the fuzzy c-means and the Ward's clustering methods. By combining both methods, the results give improvement in the proportion of active (Pa) compared to the fuzzy c-means and the Ward's clustering method. However the result only improves compared to Ward's clustering method in term of the intermolecular dissimilarity. Fuzzy c-means clustering shows the best results based on the intermolecular dissimilarity when compared to Ward's clustering and the combination of both methods.

5.1 SUGGESTION FOR FUTURE WORK

For this project, fuzzy c-means clustering were only being experimented for 1000 molecules from the AIDS dataset. The reason for this was the limited system requirement that was used to conduct the experiments. Higher hardware requirements are needed for further analysis of the clusters produced. The number of clusters and data used in the experiment should be increased, to evaluate the effectiveness of the clustering method. This experiment could not be done due to time constraint.

The dataset used for the experiment was represented by dictionary-based bit string from the binary descriptor. From the analyses, the clusters produced by fuzzy c-means clustering show best results for intermolecular dissimilarity, but not for similar property principle. Thus, experiments should also be conducted by using non-binary descriptors such as the topological indices, to see the difference that will be obtained from the two descriptors.

Other fuzzy clustering approaches can also be used in the experimented, such as the Gustafson-Kessel and the Gath-Geva clustering. The combination of fuzzy cmeans and Ward's clustering should also be improved for future work of this project.

5.2 SUMMARY

From the chapters discussed, we can conclude that fuzzy clustering method give a better result compared from Ward's method based on the intermolecular dissimilarity because its ability to give different degree of membership to compounds in different clusters. Thus, each compound does not have to belong in just one cluster. However, in term of the ability to separate active/inactive structure, fuzzy cmeans gives low values of proportion of actives structure as the value of q and k used are higher. This is because the clusters become fuzzier and the actives structures become more overlapped.

The combination of fuzzy c-means and Ward's clustering method, also gives diversity of analysis in compound selection and more research should be done to test the effectiveness of the combination of both methods. By experimenting different methods in the cluster-based approach, more clustering method can be applied in the compound selection for drug discovery. Results from these experiments will gives better results

in compound selection method and therefore will reduce time and money invested in drug discovery process. The number of compounds to be screened before further analysis can also be minimized. Based on the experiments conducted in this project, fuzzy c-means should be used more in compound selection method to produce better and faster result in drug discovery process.

LISTS OF REFERENCES

- Bajorath, J. (2001). Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *Journal of Chemical Information and Computer Science*. 41.233-245.
- Barkó, G., Abonyi, J. and Hlavay, J. (1999). Application of Fuzzy Clustering and Piezoelectric Chemical Sensor Array for Investigation on Organic Compounds. *Analytica Chimica Acta*. 398(2-3). 219-222.
- Barnard, J.M. (2003). Chemical Structure Representation and Search Systems. Barnard Chemical Information Ltd, Sheffield UK.
- Barnard, J. M. and Downs, G.M. (1992). Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *Journal of Chemical Information and Computer Science*. 32. 644-649.
- Barnard, J. M. and Downs, G.M., (2002). Clustering Methods and Their Uses in Computational Chemistry. In: Lipkowitz, K.B. and Donald B. Boyd D.B. (Ed). *Reviews in Computational Chemistry*, 18. 1-40.
- Bayada, D.M., Mamerma, H. and van Geerestein, V.J. (1999). Molecular Diversity and Representativity in Chemical Database. *Journal of Chemical Information and Computer Science*. 39.1-10.
- Bayley, M.J., Gillet, V.J., Willett, P., Bradshaw, J. and Green, D.V.S. (1999). Computational Analysis of Molecular Diversity for Drug Discovery. *Proceeding of the 3rd Annual Conference on research in Computational Molecular Biology*. ACM Press. New York. 321-330.
- Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function*. Plenum Press. New York.

- Borosy, A., Csizmaia, F. and Volford, A. (2000). Structure Based Clustering NCI's Anti-HIV Library. Ivax Drug Research Ltd.
- Brown, R. D. and Martin, Y. C. (1996). Use Of Structure-Activity Data To Compare Structure-Based Clustering Methods And Descriptors For Use In Compound Selection. *Journal of Chemical Information and Computer Sciences*. 36. 572-584
- Brown R. D. and Martin, Y. C (1997). The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *Journal of Chemical Information and Computer Sciences*. 37. 1-9.
- Castellano, G., Fanelli, A.M. and Mencar, C. (2003). A Fuzzy Clustering Approach for Mining Diagnostic Rules. *IEEE*.
- Chen X. and Reynolds C.H. (2002). Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *Journal of Chemical Information and Computer Sciences*. 42. 1407-1414.
- Chu, K. (1974). Applications of Artificial Intelligence to Chemistry. Use of Pattern Recognition and Cluster Analysis To Determine The Pharmacological Activity Of Some Organic Compounds. *Analytical Chemistry*. 46. 1181-1187
- Dittmar, P.G., Farmer, N.A., Fisanick, W., Haines, R.C. and Mockus, J. (1983). General System Design and Selection, Generation and Use of Search Screens. *Journal of Chemical Information and Computer Science*. 23. 93-102.
- Downs, G.M. (2001). Clustering in Chemistry. MathFIT Workshop, Belfast.
- Downs, G.M. and Willett, P. (1995). Clustering of Chemical Structure Databases for Compound Selection. In van de Waterbeemd, H. (Ed.). *Chemometric Methods in Molecular design*. VCH Publishers, New York. 111-130.

- Downs, G.M., Willet, P. and Fisanick, W. (1994). Similarity Searching and Clustering of Chemical Structure Databases Using Molecular Property Data. *Journal of Chemical Information and Computer Science*. 34. 1094-1102.
- Dunn, J. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact, Well-Separated Cluster. *Journal of Cybernetics*. 3(3). 32-57.
- Feher, M. and Schmidt, J.M. (2003). Fuzzy Clustering as a Means of Selecting Representative Conformers and Molecular Alignmen". *Journal of Chemical Information and Computer Science*. 43. 810-818.
- Flowers, D.R. (1997). "On the Properties of Bit String-Based Measures of Chemical Similarity. *Journal of Chemical Information and Computer Science*. 38. 379-386.
- Fung, G. (2001). A Comprehensive Overview of Basic Clustering Algorithms.
- Gath, I. and Geva, A.B. (1989). Unsupervised Optimal Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 11(7). 773-781.
- Gillet, V.J. (1999). Computational Aspects of Combinatorial Chemistry. In: Miertus, S. and Fassina, G. *Combinatorial Chemistry and Technology: Principles, Methods and Applications*. 251-274.
- Gillet, V.J., Wild, D.J., Willett, P. and Bradshaw, J. (1998). "Similarity and Dissimilarity Methods for Processing Chemical Structure Databases". *The Computer Journal*. 41(8). 547-558.
- Gustafson, D. and Kessel, W. (1979). Fuzzy Clustering with a Fuzzy Covariance Matrix. *Proceeding of IEEE CDC*. IEEE. San Diego. 761-766.
- Guthke, R., Schmidt-Heck, W., Hahn, D. and Pfaff, M. (2002). Gene Expression data Mining for Functional Genomics using Fuzzy Technology. *Advanced in Computational Intelligence and Learning Methods and Applications*. Kluwer. 475-487.

- Hollas, B. (2002). An Analysis of the Autocorrelation Descriptor for Molecules. University of Ulm.
- Holliday J.D., Hu C.Y. and Willet P. (2002). Grouping Of Coefficients For The Calculation Of Inter-Molecular Similarity And Dissimilarity Using 2D Fragment Bit-strings. *Combinatorial Chemistry and High Throughput Screening*. 5. 155-166.
- Johnson, S. C. (1967). Hierarchical Clustering Schemes. *Psychometrika*. 2:241-254
- Kaufman, L., Pierreux, A., Rousseeuw, P., Derde, M.P., Detaevernier, M.R., Massart, D.L. and Platbrood, G. (1983). Clustering on a Microcomputer with an Application to the Classification of Coals. *Analytica Chimica Acta*. 153. 257-260.
- Kaymak, U. and Setnes, M. (2000). Extended Fuzzy Clustering Algorithm. *ERIM Report Series Research in Management*. 1-23.
- Lin, T.H., Wang, G.M. and Hsu, Y.H. (2002). Classification of Some Active HIV-1 Protease Inhibitors and Their Inactive Analogues Using Some Uncorrelated Three-Dimensional Molecular Descriptors and A Fuzzy c-Means Algorithm. *Journal of Chemical Information and Computer Science*. 42. 1490-1504.
- MacCuish, J.D. and MacCuish, N.E. (2003). Clustering Ambiguity and Binary Descriptors. MESA Analytics and Computing, New York.
- Martin, E. (2003). Pap-Smear Classification. Technical University of Denmark.
- Massart, D.L. and Kaufman, L. (1983). The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis. Wiley, New York.
- Matter, H. (1997). Selecting Optimally Diverse Compounds from Structural Database: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *Journal of Medicinal Chemistry*. 40. 1219-1229.

- Mojena, R. (1977). Hierarchical Grouping Methods and Stopping Rules: An Evaluation. *Computer Journal*. 20(4). 359-363.
- Murtagh, F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms. *Computer Journal*. 26. 354-359.
- Rodgers S.L., Holliday J.D. and Willet P. (2004). Clustering Files of Chemical Structures Using the Fuzzy k-Means Clustering Method. *Journal of Chemical Information and Computer Science*. 44. 894-902.
- Root, D.E., Kelley, B.P. and Stockwell, B.R. (2002). Global Analysis of Large-Scale Chemical and Biological Experiments. *Current Opinion in Drug Discovery & Development*. 5(3). 355-360.
- Salim, N. (2003). Analysis and Comparison of Molecular Similarity Measures. University of Sheffield. PhD Thesis.
- Sneath, P.H.A. (1966). Relations between Chemical Structure and Biological Activity in Peptides. *Journal of Theoretical Biology*. 12. 157-195
- Takahashi, Y., Miyashita, Y., Abe, H. and Sasaki, S.I. (1980). A Structure Biological Activity Study Based on Cluster Analysis and the Non-Linear Mapping Method of Pattern Recognition. *Analytica Chimica Acta*. 122. 241-247.
- Taylor, R. (1995). Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *Journal of Chemical Information and Computer Science*. 35. 59-67.
- Tropsha, A. & Zheng, W. (2002). Rational Principles of Compound Selection for Combinatorial Library Design. *Combinatorial Chemistry and High Throughput Screening*. 5. 111-123.

- Van Geerestein, V.J., Hamersma and van Helden, S.P. (1997). Exploiting Molecular Diversity: Pharmacophore Searching and Compound Clustering. In: van de Waterbeemd, H., Testa, B. and Folkers, G. (eds.). *Computer-Assisted Lead Finding and Optimization*. Wiley-VCH, Weinheim. 157-178.
- Warr, W. (1997). Combinatorial Chemistry and Molecular Diversity. An Overview. *Journal of Chemical Information and Computer Science*. 37. 134-140.
- Wikipedia Encyclopedia [online]. http://en.wikipedia.org/wiki/Main_Page as retrieved on 19th February 2004.
- Wild, D.J. (2003). Advanced Chemoinformatics Methods. *Chemical Engineering: Introduction to Chemoinformatics Lesson 6*.
- Willett, P. (1987). Similarity and Clustering in Chemical Information System. Research Studies Press. Letchworth.
- Willett, P. (1997). Computational Tools for the Analysis of Molecular Diversity. *Perspectives in drug Discovery and Design*. 7/8. 1-11.
- Y.B. Dato' Seri Law Hieng Ding (2003). Managing Risks and Assessing Opportunities: Biotechnology in a World of Uncertainty. *The National Conference on Biotechnology & Life Science 2003*. Kuala Lumpur.

UNIVERSITI TEKNOLOGI MALAYSIA

**BORANG PENGESAHAN
LAPORAN AKHIR PENYELIDIKAN**

TAJUK PROJEK : FEASIBILITY STUDY OF FUZZY CLUSTERING TECHNIQUES IN
CHEMICAL DATABASE FOR COMPOUND CLASSIFICATION

ROZILAWATI BT DOLLAH @ MD. ZAIN

Saya _____

(HURUF BESAR)

Mengaku membenarkan **Laporan Akhir Penyelidikan** ini disimpan di Perpustakaan Universiti Teknologi Malaysia dengan syarat-syarat kegunaan seperti berikut:

1. Laporan Akhir Penyelidikan ini adalah hakmilik Universiti Teknologi Malaysia.
2. Perpustakaan Universiti Teknologi Malaysia dibenarkan membuat salinan untuk tujuan rujukan sahaja.
3. Perpustakaan dibenarkan membuat penjualan salinan Laporan Akhir Penyelidikan ini bagi kategori TIDAK TERHAD.
4. *Sila tandakan (✓)

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam **(AKTA RAHSIA RASMI 1972)**).

TERHAD

(Mengandungi maklumat **TERHAD** yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan).

TIDAK
TERHAD



TANDATANGAN KETUA PENYELIDIK

ROZILAWATI BT DOLLAH @ MD. ZAIN

KETUA PENYELIDIK VOT 75107

JABATAN SISTEM MAKLUMAT,

FAKULTI SAINS KOMPUTER DAN SISTEM MAKLUMAT, UTM

Nama & Cop Ketua Penyelidik

CATATAN: * Jika Laporan Akhir Penyelidikan ini **SULIT** atau **TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

UNIVERSITI TEKNOLOGI MALAYSIA
Research Management Centre

PRELIMINARY IP SCREENING & TECHNOLOGY ASSESSMENT FORM

(To be completed by Project Leader submission of Final Report to RMC or whenever IP protection arrangement is required)

1. PROJECT TITLE IDENTIFICATION :

FEASIBILITY STUDY OF FUZZY CLUSTERING TECHNIQUES IN CHEMICAL DATABASE FOR
COMPOUND CLASSIFICATION.

Vote No:

75107

2. PROJECT LEADER :

Name :

ROZILAWATI BT DOLLAH @ MD. ZAIN

Address:

DEPARTMENT OF INFORMATION SYSTEMS, FACULTY OF COMPUTER SCIENCE AND

INFORMATION SYSTEMS, UTM, 81310, SKUDAI, JOHOR

Tel : 07-5532425 Fax : 07-5565044 e-mail : zeela@fsksm.utm.my

3. DIRECT OUTPUT OF PROJECT (Please tick where applicable)

Scientific Research	Applied Research	Product/Process Development
<input checked="" type="checkbox"/> Algorithm	<input type="checkbox"/> Method/Technique	<input type="checkbox"/> Product / Component
<input type="checkbox"/> Structure	<input type="checkbox"/> Demonstration / Prototype	<input checked="" type="checkbox"/> Process
<input checked="" type="checkbox"/> Data		<input type="checkbox"/> Software
<input type="checkbox"/> Other, please specify	<input type="checkbox"/> Other, please specify	<input type="checkbox"/> Other, please specify
_____	_____	_____
_____	_____	_____
_____	_____	_____

4. INTELLECTUAL PROPERTY (Please tick where applicable)

- | | |
|--|--|
| <input checked="" type="checkbox"/> Not patentable | <input type="checkbox"/> Technology protected by patents |
| <input type="checkbox"/> Patent search required | <input type="checkbox"/> Patent pending |
| <input type="checkbox"/> Patent search completed and clean | <input type="checkbox"/> Monograph available |
| <input type="checkbox"/> Invention remains confidential | <input type="checkbox"/> Inventor technology champion |
| <input type="checkbox"/> No publications pending | <input type="checkbox"/> Inventor team player |

5. LIST OF EQUIPMENT BOUGHT USING THIS VOT

i) HP NOTEBOOK - 1 UNIT

6. STATEMENT OF ACCOUNT

a)	APPROVED FUNDING	RM : 17,50000.00
b)	TOTAL SPENDING	RM : 17, 714.94
c)	BALANCE	RM : - 214.94

7. TECHNICAL DESCRIPTION AND PERSPECTIVE

Please tick an executive summary of the new technology product, process, etc., describing how it works. Include brief analysis that compares it with competitive technology and signals the one that it may replace. Identify potential technology user group and the strategic means for exploitation.

a) Technology Description

Briefly, the cluster-based method has been widely used in compound selection. In compound selection, there are four main approaches namely, cluster-based compound selection, dissimilarity-based compound selection, partition-based compound selection and optimization-based compound selection. Cluster-based compound selection is the process of subdividing chemical databases into groups or clusters. The members of one group will differ from one another according to a chosen criterion. As stated by Bayada et.al (1999), Brown and Martin (1996), Matter (1997), Taylor (1995) and Van Geerestein et.al (1997), cluster-based compound selection is the most useful subset selection, thus this has encourage the studies and researches of cluster-based method for compound selection.

In this study, the fuzzy c-means clustering was chosen to cluster the compounds. This method is based on the cluster center and degree of membership and the process is repeated until the cost-function is minimized. Then the clusters produced will be measured based on the similarity measures and their ability to separate actives and inactives compounds. The result is be compared to the Ward's clustering method and analyzed based on its ability to separate active/inactive structure and the intermolecular dissimilarity between centroids of the clusters.

b) Market Potential

Result from this study is very useful for Malaysia's research and development (R&D) and agencies related to biotechnology industry or pharmaceutical industries. Bio Valley Malaysia for example, will conduct a wide spectrum of biotechnology-related activities, especially in drug research.

c) Commercialisation Strategies

Extended research is needed to improve the result from this study using higher hardware requirements for further analysis of the clusters produced. The number of clusters and data used in the experiment should be increased, to evaluate the effectiveness of the clustering method. Thus, experiments should also be conducted by using non-binary descriptors such as the topological indices, to see the difference that will be obtained from the two descriptors.

8. RESEARCH PERFORMANCE EVALUATION

a) FACULTY RESEARCH COORDINATOR

Research Status	()	()	()	()	()	()
Spending	()	()	()	()	()	()
Overall Status	()	()	()	()	()	()
	Excellent	Very Good	Good	Satisfactory	Fair	Weak

Comment/Recommendations :

.....
Signature and stamp of
JKPP Chairman

Name :
Date :

b) RMC EVALUATION

Research Status	()	()	()	()	()	()
Spending	()	()	()	()	()	()
Overall Status	()	()	()	()	()	()
	Excellent	Very Good	Good	Satisfactory	Fair	Weak

Comments :-

Recommendations :

- Needs further research
- Patent application recommended
- Market without patent
- No tangible product. Report to be filed as reference

.....
 Signature and Stamp of Dean / Deputy Dean
 Research Management Centre

Name :
 Date :