OPTIMIZED SUBTRACTIVE CLUSTERING FOR CLUSTER-BASED COMPOUND SELECTION

KUIK SOK PING

UNIVERSITI TEKNOLOGI MALAYSIA

OPTIMIZED SUBTRACTIVE CLUSTERING FOR CLUSTER-BASED COMPOUND SELECTION

KUIK SOK PING

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

APRIL 2006

ACKNOWLEDGEMENT

Praises to God for giving me the patience, strength and will to go through and complete my study. I would like to express my appreciation to my supervisor, Associate Professor Dr. Naomie bte Salim, for her support and guidance during the course of this study and the writing of the thesis. I would also like to extend my thanks to fellow friends : Sok Mun, Yean Hui, Pei Chuen, Mei Serr, Angeline, Karen, Tammy, Siow Chin, Eric, and who have given me the encouragement and support when I needed it.

My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Especially to my boss Mr See Mong Kiong, I cannot finish my study without his toleration. Unfortunately, it is not possible to list all of them in this limited space. Finally, I would like to dedicate this thesis to my family. Without their love and support I would have never come this far.

ABSTRACT

Compound selection method is important in drug discovery especially in lead identification process. Finding the best method in the compound selection has become a need to pharmaceutical chemistry 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. Among all fuzzy clustering method, fuzzy c-means using Euclidean Distance measures is better used in compound selection. Fuzzy c-means clustering gives the best result in intermolecular dissimilarity; however it shows poor results of separation of active/inactive structure. The research focused on the subtractive 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. Subtractive clustering has been chosen because it considers each data point as a potential cluster center and defines a measure of the potential of data point and it also resolves the problem of how many clusters need to be taken for the data. Subtractive clustering will produce the number of cluster automatically together with the value of radii cluster and squash factor. The results from subtractive clustering are compared to fuzzy c-means method and K-means. The analysis shows that subtractive clustering gives the worst result in separation of active/inactive structure among the fuzzy c-means and K-means. K-means produced the highest proportion of active structure in this research. For subtractive clustering, good values of squash factor are between 0.375 and 0.45 and the radii cluster from 0.35 to 0.45 because they always hit the highest proportion of active structures.

ABSTRAK

Kaedah pemilihan sebatian merupakan kaedah yang penting dalam penemuan ubat, terutamanya dalam proses pengenalpastian molekul yang berpotensi untuk dijadikan ubat. Penyelidikan untuk mencari kaedah yang terbaik bagi pemilihan sebatian telah menjadi satu keperluan industri farmasi berikut peningkatan jumlah sebatian yang perlu ditapis. Kaedah yang terbaik dan kerap digunakan di dalam pemilihan sebatian ialah kaedah pengkelompokan; di mana set-set data sebatian dikumpulkan dalam kelompok masing-masing dan wakil daripada setiap kelompok akan dipilih. Kaedah fuzzy c-means 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. Kajian ini menekankan subtractive clustering dan keberkesanan kelompok yang dihasilkan berdasarkan Topological Indixes. Hasil kaedah ini dianalisa dan dibandingkan dengan kaedah pengkelompokan konvensional yang lain. K-means merupakan kaedah yang terbaik untuk mengelompokan sebatian berbanding dengan kaedah subtractive clustering dan fuzzy c-means. Jejari antara 0.35 dan 0.45 serta faktor squash antara 0.375 dan 0.45 merupakan julat yang baik untuk menghasilkan struktur aktif yang tinggi di dalam kelompok berkenaan.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbf{g}$	-	fuzziness index
$\mathbf{c}$	-	number of cluster
$\mathbf{u_{ij}}$	-	degree of membership
$\mathbf{X_j}$	-	the data point of the jth compound
$\mathbf{K}$	-	number of data point
$\mathbf{U}$	-	a fuzzy C-partition of the data set
$\mathbf{C_i}$	-	the centroid of the ith cluster
$\mathbf{x_{ik}}$	-	the attribute value of molecule i in cluster k
$\mathbf{n}$	-	size of cluster
$\mathbf{P_a}$	-	proportion of active structure
d_{ik}	-	any inner product metric or the distance measure
P_i	-	Potential value of data point i
x_{jA}	-	value of *j*th dimension in molecule A
ra	-	Radii or radius defining a neighborhood
rb	-	Squash Factor
$x_1{}^*$	-	first cluster center point
$x_k{}^*$	-	k'th cluster center point
$P_k{}^*$	-	Potential value of $x_k{}^*$

$\bar{\varepsilon}$	-	Accept Ratio
$\underline{\varepsilon}$	-	Reject Ratio
G i	-	partitioned into *c* groups
$\boldsymbol{P}_1{*}$	-	potential value of first cluster center point
exp	-	exponent

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

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 or 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
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

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

The drug design technologies have already produced a tremendous amount of data that requires proper methods of data analyzing. The dramatic increase of resulting compound data has encouraged researchers in the field to look at ways of applying various machine learning techniques and intelligent techniques for data analysis. The main reasons a compound cannot become drugs are inactive, toxicity, flexibility and size molecule. A main issue in analyzing chemical data is preferably to specify different actives and inactive into different clusters.

In the early stages of a drug discovery project, the emphasis is on lead generation, in which an attempt is made to optimize the molecular diversity of the initial library produced. Due to the similar property principle (Johnson and Maggiora, 1990), structurally similar compounds can be expected to exhibit similar properties and biological activities. It is thus undesirable to test a large number of structurally similar compounds for many reasons. Maximizing the diversity of a subset is assumed to enhance the chances of finding active compounds of various structural types in screening experiments. It will reduce the time of the chemist to find out the certain property of the chemical compound. (Everitt, 1993).

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.

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 PROBLEM STATEMENT

The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis. Moreover, most of the chemical data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Nevertheless, finding these groupings or trying to categorize the data is not a simple task for humans unless the data is of low dimensionality. This is why some methods in soft computing have been proposed to solve this kind of problem.

There are various types of clustering methods, the most popular clustering methods is fuzzy clustering such as fuzzy c-means, fuzzy k-mean, Gustafson-Kessel, and the Gath-Fava. In the last few years, fuzzy clustering from the overlapping clustering has been used in chemoinformatics. It represents the real world situation where a compound may belong to several clusters simultaneously with different degrees of membership (Feher, 2003). The evaluation 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 (Sharin and Naomie, 2004). The results of the analysis show that fuzzy c-means clustering only gives best the result compared to Ward's clustering method based on the intermolecular dissimilarity. The results for separation of active/inactive structure show less proportion of active for clusters from fuzzy c-means than Ward's clustering. Many studies have proven that non-overlapping methods are most effective methods for compound selection.

We would like to apply the subtractive clustering method in the clustering compound selection. Subtractive clustering is a method introduced by Chiu (1994) and the efficiency of this method has not been tried in chemoinformatic area. The efficiency of the subtractive clustering method by applying subtractive clustering technique in the clustering compounds selection will be found out in this research, especially in the compound selection application.

1.2 OBJECTIVES

Identifying objectives are very important in defining the goals to be achieved in this project. The followings are the objectives of the project:

i. To apply subtractive clustering technique to chemical compound clustering.
ii. To measure the efficiency of subtractive technique in clustering the chemical compound for compound selection purpose.
iii. To compare the results of subtractive clustering with Fuzzy c-means and K-means method based on ability of the method to separate the data to different partition with certain portion of active and inactive compound.

1.3 SCOPE OF WORKS

The project scope must be identified in order to keep the project running on the right track. The followings are the scopes of the project that have been identified:

i. The dataset used is chemical compound dataset obtained from the MDL Drug Data Report Database.
ii. The algorithm that will be used is the subtractive clustering method.
iii. The descriptors used are Topological Indices only.
iv. *K-fold cross-validation* method will be applied in chemical compound clustering by using *subtractive* clustering algorithm, and observing the proportion of active compounds from the clusters.

1.4 PROJECT PLAN

This project will be carried out in two semesters. The first part of the project is done in the first semester where the understanding of literature review and methodology to be used are focused. With that, most of the time is spent in searching and gathering information from articles in journals such as Journal of Chemical and Computer Science from the American Chemical Society (ACS), Lecturer Note in computer Sciences.

In this project, it is important to understand the chemoinformtic, process of similarity searching, clustering method and subtractive clustering. At the end of Project I, the main goal is to have better understanding of the terms and topics that have been mentioned previously. For the first part of the project, the report includes the Introduction, Literature Review and Methodology of the project. All of these are done during the first semester.

In the second semester, the second part of the project is done that involves the generate descriptors, development and implementation of subtractive algorithm is carried out. The development process of Project II will start with generating descriptors from MDDR database. The research focus on the subtractive technique in clustering chemical compound where the effectiveness of the clusters produced with regard to compound selection is analyzed. Dataset will be divided to training and testing dataset with actives and inactive compound using cross validation technique. The results from subtractive clustering will be compared between the dataset experiments.

The second part of the report will be written after implementation of the project. This part of the report will include the Experimental Result, Analysis of Results and Conclusion of the project.

1.5 ORGANIZATION OF REPORT

Chapter I is the introduction to the project that has been conducted. It contains discussions on the problem background, problem statements, project aim, objectives as well as scopes of project. The significance and knowledge contributions are also stated in this segment.

Chapter II discussed the literature reviews that have been combined in order to make up the whole project. This includes the background knowledge on the terms that are involved in the project mainly on cheminformatic and statistic based clustering method.

Chapter III is about the methodology that is used in this project. In this section, the techniques that are involved are discussed which are subtractive clustering algorithm. The hardware and software requirements for this project are also discussed in this section.

Chapter IV discussed about the results from applying subtractive clustering in this project. It is then analyzed by determining which method produced good result in clustering chemical compound.

Chapter V is the conclusion of the project based on the four previous chapters that has been discussed. There are also discussions and future works that can be done to enhance this project.

LIST OF REFERENCES

Aminzadeh, F. and S. Chatterjee, 1984/85, Applications of clustering in exploration seismology, Geoexploration, v23, p.147-159.

Augen, J. “The evolving role of information technology in the drug discovery process”, *Drug Discov. Today*, 2002, *7*, 315-323.

Bezdek J C 1981 Pattern recognition with fuzzy objective function algorithms; New York: Plenum Press.

Borosy, A., Csizmaia, F. and Volford, A. (2000). Structure Based Clustering NCI’s Anti-HIV Library. Ivax Drug Research Ltd.

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.

Emami, M.R. Turksen, Burhan. Goldenberg A.A., 1998. Development of A Systematic Methodology of Fuzzy Logic Modeling, *IEEE transaction on Fuzzy Systems*, 6(3).

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.

Frank Brown, "Chemoinformatics: What is it and How does it Impact Drug Discovery" Annual Reports in Medicinal Chemistry 33: 375-384, 1998

Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. “Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries”, *J. Med. Chem*., 1994, *37*, 1233-1251.

Gath, I. and Geva, A.B. (1989). Unsupervised Optimal Fuzzy Clsuetring. IEEE Transactions on Pattern Analysis and Machine Intelligence. 11(7). 773-781.

G. M. Downs and P. Willett, in Reviews in Computational Chemistry, K. B. Lipkowitz and D. B. Boyd, Eds., VCH Publishers, New York, 1995, Vol. 7, pp. 1–66. Similarity Searching in Databases of Chemical Structures.

Hall, D. G.; Manku, S.; Wang, F. Solution- and Solid-Phase Strategies for the Design, Synthesis, and Screening of Libraries Based on Natural Product Templates: A Comprehensive Survey, *J. Comb. Chem*., 2001, *3*, 125-150

J. A. Hartigan and M. A. Wong, “A k-means clustering algorithm,” Applied Statistics, 28:100--108, 1979.

Jang, J.S.R. Adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man and Cybernetics 23 (1993) 665–685

Jang, J.S.R., Sun, C.T., Mizutani, E.: Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Prentice Hall (1997)

Jian, A. K. And Dubes, R. C. 1988. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

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.

L. Kaufman and P. J. Rousseeuw, Finding Groups Analysis, Wiley-Interscience, New York, 1990.

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.

M Hann and R Green "Chemoinformatics a new name for an old problem?" Current Opinion in Chemical Biology 3:379- 383, 1999

Mojena, R. (1977). Hierarchical Grouping Methods and Stopping Rules: An Evaluation. Computer Journal. 20(4). 359-363.

P. Willett, J. M. Barnard, and G. M. Downs, J. Chem. Inf. Comput. Sci., 38 (6), 983 (1998). Chemical Similarity Searching.

P. Willett, Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, UK, 1987.

R. D. Brown, in Computational Methods for the Analysis of Molecular Diversity, P. Willett, Ed., Perspectives in Drug Discovery and Design, Vol. 7/8, Kluwer/ESCOM, Dordrecht, The Netherlands, 1997, pp. 31–49. Descriptors for Diversity Analysis.

R. D. Brown and Y. C. Martin, J. Chem. Inf. Comput. Sci., 36 (3), 572 (1996). Use of Structure–Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection.

R. D. Brown and Y. C. Martin, J. Chem. Inf. Comput. Sci., 37 (1), 1 (1997). The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding.

Salim, N. (2003). Analysis and Comparison of Molecular Similarity Measures. University of Sheffield. PhD Thesis.

S.L.Chiu, “Fuzzy model identification based on cluster estimation”, J.Intell.Fuzzy Syst., vo1.2, pp.267-278,1994

The MathWorks, Inc., “Fuzzy Logic Toolbox – For Use With MATLAB,” The MathWorks, Inc., 1999.

Timothy Ritchie. Chemoinformatics; manipulating chemical information to facilitate decision- making in drug discovery. Drug Discovery Today 6(16) : 813- 814, Aug. 2001

Tropsha, A. & Zheng, W. (2002). Rational Prinsiples 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.

Yager, R.R., Filev, D.P., 1994. Approximate clustering via the mountain method. IEEE Trans. Systems, Man, Cybernet. 24 (8), 1279–1284.

Yannis L. Loukas. Adaptive Neuro-Fuzzy Inference System: An Instant and Architecture-Free Predictor for Improved QSAR Studies" *J. Med. Chem.* **2001,** *44*, 2772-2783