Application of Fuzzy Clustering Analysis to Compound Datasets for Drug Lead Identification

[1]Sinarwati Mohamad Suhaili,[2]Mohamad Nazim Jambli and Abdul Rahman Mat
[1]Centre for Pre-University Studies, [2]Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan
Sarawak, Malaysia
Email: mssinarwati@preuni.unimas.my, jmnazim@fit.unimas.my, marahman@fit.unimas.my

Abstract—**Recently, the increasing number of chemical compound datasets to be screened has been growing rapidly due to the fast developments of high-throughput screening in drug discovery. These compound datasets requires compound selection methods which have become one of the main technique in drug discovery especially in drug lead identification process. Thus, finding the best method in compound selection is needed to the pharmaceutical industry to ensure the accurate results of this process. One of most used compound selection method is cluster-based compound selection, which involves subdividing a set of compounds into clusters and choosing one compound or a small number of compounds from each cluster. In this cluster-based compound selection, non-overlapping methods such as Ward's, Group Average, Jarvis Patrick's and K-means are preferred methods to cluster the diverse set of compounds. However, there are little study on overlapping method such as fuzzy c-mean (FCM) and fuzzy c-varieties (FCV) clustering algorithms. Therefore, these two clustering algorithms are applied and their performance is compared based on the effectiveness of the clustering results in terms of separation between actives and inactives (Pa) into different clusters and mean intercluster molecular dissimilarity (MIMDS). The analysis shows FCM gives the best results compare to FCV in terms of Pa indicating that FCM has a promising use in compound selection algorithms. But, FCV is perform better than the FCM in term of MIMDS when a higher number of compounds and higher fuzziness index value are concerned.**

I. INTRODUCTION

In the past decade, many chemoinformatics approaches such as the compound selection have been applied in the drug discovery process. The drug discovery process is a process of identifying compounds that may become useful new drugs in pharmaceutical area[1]. This process can be divided into an early and a late phase. The early phase is mainly represented by target and lead discovery, whereas the later deals mainly with clinical evaluation and development[2]. In the drug lead discovery process, thousand of molecule structures need to be screened before lead optimization begins. In this process, initial leads for drug development will originate from high-throughput screening (HTS), for example fifty thousands to five millions of compounds are screened and tested in the hope of discovering biologically active compounds. This dataset is likely to be very large with millions of compounds. 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 for pharmaceutical industry. Because of these constraints, there is a need of demand for powerful and reliable techniques to identify high-quality lead drug candidates in order to save time and money. Therefore, the research and development focuses on this interest in order to develop faster and more effective way to produce chemical compounds that become a useful drug.

One of the impending ways to reduce the processing time and cost of drug discovery is using the cluster-based compound selection method in order to select the potential active compound in lead identification process. It involves subdividing a set of compounds into clusters and choosing one compound or a small number of compounds from each cluster. It is also a technique to groups the data into classes or clusters so that the objects within the cluster have high similarity in comparison to one another, but are very dissimilar to those data objects in other clusters[3]. Indirectly, by using this method, it has helped the researches of finding lead compounds faster and more effectively by showing which compound belongs to certain cluster that is similar to known compound rather than screening all the dataset in compound libraries.

The most widely techniques used to cluster the compound is non-overlapping clustering methods [3] which occurs when each compound is member for only one cluster. However, there are fewer study conducted to overlapping clustering method in term of compound selection such as fuzzy clustering. The fuzzy clustering method has been chosen from the overlapping clustering method because the fuzzy concepts obviously provide the way for tackling the problem of conventional clustering methods, where an object can only belong or not belong to a particular cluster. This concept can represent membership degree to which an object belongs to that cluster. Thus if cluster is a group whose members share common properties, then the membership degree of an object indicates the degree to which that object displays these properties with similar objects having high membership of the same cluster(s) [4]. In [5] also claim that fuzzy is expected to perform better, in cases where there are a significant number of outliers, such as molecular dynamics simulations and molecule alignments. This is also supported by [4] who prove that fuzzy

978-1-4673-1938-6/12/$31.00 ©2012 IEEE