

# Methods for Clustering Mass Spectrometry Data in Drug Development

Huiru Zheng<sup>1</sup>, Sarabjot Singh Anand<sup>1</sup>, John G Hughes<sup>1</sup> and Norman D Black<sup>1</sup>

**Abstract.** Isolation and purification of the active principle within natural compounds plays an important role in drug development. MS (mass spectrometry) is used as a detector in HPLC (high performance liquid chromatography) systems to aid the determination of novel compound structures. Clustering techniques provide useful tools for intelligent data analysis within this context. In this paper, we analyse some representative clustering algorithms, describe the complexities of the mass spectrometry data generated by HPLC-MS systems, and provide a new algorithm for clustering, based on the needs of drug development. This new algorithm is based on the definition of a dynamic window within the instance space.

## 1 INTRODUCTION

Chemists make literally thousands of analogues in order to develop drugs from natural sources [4]. Figure 1 shows a simplistic view of the drug development process.

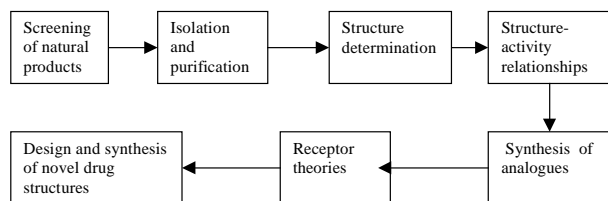


Figure 1. A Simplistic view of the drug development

Screening natural compounds for biological activity continues today in the never-ending quest to find new compounds for discovering drugs. Isolating and purifying the new compounds for the active function plays an important role in the pattern. Next, the structures of the new compounds have to be determined before the structure-activity relationships (SARs) can be analysed. Although the new compounds have useful biological active function, chemists have to use synthetic analogues to reduce any serious side effects, and finally, develop the novel drugs.

Our project is targeted at the isolation, structure elucidation and synthesis of biologically active substances of potential pharmacological significance from the venom of frogs. Several isolation techniques have been developed such as freeze-drying,

filtration, centrifugation, and, in particular chromatography have been developed. MS (mass spectrometry), a technique used to characterize and separate ions by virtue of their mass/charge ( $m/z$ ) ratios [7], can be helpful in structure determination as the fragmentation can give useful clues about the structure.

In our experiment, we use a HPLC (high performance liquid chromatography) system to isolate the natural product. The chromatographic separations are based upon the distribution of components of a mixture between a liquid phase and a stationary phase. The MS (mass spectrometry) is used as the detector. We use a number of different samples of the frog's venom that result in millions of two-dimensional records describing the mass and the separation time of isolated molecules. Clustering is applied to discover the functional molecules for determining structures of the pharmacological compounds.

The aim of this paper is to ascertain whether current clustering algorithms are suitable for the data generated by HPLC-MS. In section 2 we describe the features of the mass spectrometry data generated by our experiments in drug discovery. In section 3 we give an analysis of current algorithms. In section 4, we describe a new algorithm for clustering on the needs of drug development, which is based on the dynamic window concept. In section 5, we will provide the result and discuss it for further work.

## 2 FEATURES OF THE MASS SPECTROMETRY DATA

The venom of frog is collected and injected into the HPLC. The entire isolation time is about 80 minutes. Every 0.03 or 0.04 minutes during these 80 minutes, the system will perform a scan to isolate the molecules from the venom analyte. The mass spectrometry detects the isolated molecules. However, it takes an isolated molecule about 1 minute to travel from one end of the column in the HPLC system to the other end. One molecule could be detected in about 33 (if the scan time interval is 0.03 minute) or 25 (if the scan time interval is 0.04 minute) scans. Therefore, data obtained from the HPLC-MS contains a lot of noise, which needs to be eliminated. For each frog, the amount of the original data is large; for example, frog "Aur" has about 446,900 records. Therefore, clustering technique can be applied to smooth the noise data and get the real data, i.e. identify the true values of mass/charge ratios and isolation time of the molecules.

We aim to search for the same functional molecules between the pharmacological frogs. For each frog, even after eliminating noise, the amount of data is still large, for examples, frog "Aur" contains 24,549 points and frog "Inf" has 50,772. There are different

<sup>1</sup> Faculty of Informatics, University of Ulster at Jordanstown, Newtownabbey, Co. Antrim, N. Ireland  
Email: {H.Zheng, ss.anand, jg.hughes, nd.black}@ulst.ac.uk

species of frogs, each species has different types, and for each type, there are hundreds of different frogs. Therefore, in order to mine for the functional molecules, clustering techniques are used.

Fig 2 shows a sample of data generated by HPLC-MS after eliminating noise. Mass spectrometry data from HPLC-MS has two dimensions: *Time* and *Mass*. *Time* describes the isolated time of the molecule and *Mass* represents the mass/charge ratios. *No.* is the index for each record and *Frog* is the name of the frog from which the data of the venom is collected.

No	Time	Mass	Frog
1	3.51	1114.55	Aur
2	3.64	609.0467	Aur
3	44.75	141.46	Aur
4	16.9382	176.934	Inf
5	37.8	234.14	Caer
6	15.52	617.963	Aur
7	15.1618	617.986	Caer
8	15.43	618.004	Inf
9	60.252	1998.44	Inf
10	41.68	1998.9	Inf

Figure 2. Mass Spectrometry Data

Mass spectrometry data has some unusual features. Only molecules which have similar *Time* and have similar *Mass* will be the same, and therefore in the same cluster. As shown in Figure 2, though molecule No.1 and molecule No.2 have similar *Time*, since their *Mass* values are quite different, they are not the same molecule, and cannot appear in the same cluster. For molecule No. 9 and No. 10, their *Mass* values are similar but their isolated *Time* values are quite different. Therefore they should not be in the same cluster. Molecules No. 6, No. 7 and No. 8 are similar in both *Time* and *Mass* and therefore they are deemed to be the same molecule even though they have been isolated from different frogs.

Another feature of the data is that the ranges of *Time* and *Mass* are large, and the intervals of *Time* and *Mass* between the data of two molecules are small. There is some regularity in the data and it is not necessary to consider the entire data space.

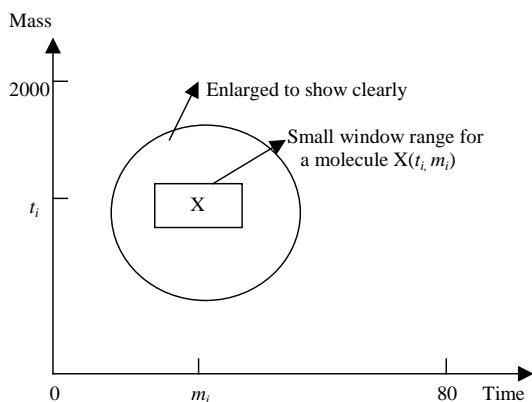


Figure 3. Data space of mass spectrometry data generated by HPLC-MS

Figure 3 shows an example of the data within the instance space.  $X(t_i, m_i)$  is a molecule in the data space. Comparing to the entire data space, the range of the window is very small need to be enlarged. All of the data points within the window represent the

same molecule  $X(t_i, m_i)$ , and all of the data points representing the same molecule  $X(t_i, m_i)$  will fall in the window. Therefore it is unnecessary to search the entire space for the same molecule.

All these features give us a heuristic to select our algorithm for clustering of mass spectrometry data.

### 3 CLUSTERING TECHNIQUE FOR INTELLIGENT DATA ANALYSIS

The various clustering concepts available can be grouped into two broad categories: hierarchical methods and nonhierarchical (or partitioning) methods [6].

Nonhierarchical (or partitioning) methods include those techniques in which a desired number of clusters are assumed at the starting point. Data points are reallocated among clusters so that a particular clustering criterion is optimized. A possible criterion is the minimization of the variability within clusters, as measured by the sum of the variance of each parameter that characterizes a point. Given a set of objects and a clustering criterion, nonhierarchical clustering obtains a partition of the objects into clusters such that the objects in a cluster are more similar to each other than to objects in different clusters. The basic idea of these algorithms is demonstrated by K-means and  $k$ -medoid methods [5]. For K-means, the centre of gravity of the cluster represents each cluster; for K-medoid, each cluster is represented by one of the objects of the cluster located near the center. A well known algorithm is CLARANS (Clustering Large Applications based on RANdomized Search) which uses a randomized and bounded strategy to improve the performance [8], and can be viewed as an extension of these methods for large databases.

The K-means method constructs a partition of a database of  $N$  objects into a set of  $K$  clusters. It requires as input the number of clusters, and starts with an initial partition, uses a minimizing of the distance of each point from the cluster centre as the search criterion and an iterative control strategy to optimize an objective function.

The  $k$ -medoid method is used by PAM (Partitioning Around Medoids) [5] to identify clusters. PAM selects  $K$  items arbitrarily as medoids and swaps with other items until  $K$  items qualify as medoids. Then an item is compared with the entire data space to obtain a medoid. It requires as input a value for input  $K$ . CLARANS is an efficient improvement of this  $k$ -medoid method. However, since it uses randomized search, it cannot be guaranteed to converge when the amount of data is large.

Andrew Moore [3] and Dan Pelleg [1] use the  $kd$ -tree data structure to reduce the large number of nearest-neighbor queries issued by the traditional K-means algorithm, and use the EM (Expectation Maximization) method for finding mixture models. They use a hyper-rectangle  $h$  as an additional parameter to determine the new centroids. The initial value of  $h$  is the hyper-rectangle with all of the input points in it. It updates its counters using the centre of mass and number of points that are stored in the  $kd$ -node corresponding to  $h$ . Otherwise it splits  $h$  by recursively calling itself with the children.

Hierarchical methods include those techniques where the input data are not partitioned into the desired number of classes in a single step. Instead, a series of successive fusions of data are performed until the final number of clusters is obtained. A hierarchical clustering is a nested sequence of partitions. Agglomerative hierarchical clustering starts by placing each object

in its own cluster and then merges these atomic clusters into larger and larger clusters until all objects are in a single cluster. Data partitioning based hierarchical clustering starts the process with all objects in a cluster and subdividing into smaller pieces. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) uses data partitioning according to the expected cluster structure called CF-tree (Cluster Feature Tree) which is a balanced tree for storing the clustering features [9]. STING (Statistical information grid-based method) is based on a quad-tree-like structure, and DBSCAN [2] relies on a density-based notion of cluster and uses an R\*-tree to achieve better performance.

Hierarchical algorithms do not need  $K$  as an input parameter. This is an obvious advantage over the nonhierarchical algorithms, though they require a termination condition to be specified.

BIRCH uses a CF-tree for incrementally and dynamically clustering the incoming data points to produce a condensed representation of the data, and applies a separate cluster algorithm to the leaves of the CF-tree. It uses several heuristics to find the clusters and to distinguish the clusters from noise. It is one of the most efficient algorithms because it condensed data. However, BIRCH is sensitive to the order in which the data is input and so different cluster may result due to simply a change in ordering the data.

DBSCAN [2] defines clusters as density-connected sets. For each point, the neighborhood of a given radius has to contain a minimum number of points - the density in the neighborhood has to exceed some threshold. DBSCAN can separate the noise and discover clusters of arbitrary shape. STING [8] uses a quad-tree-like structure for condensing the data into grid cells. Hierarchical grid clustering algorithms organize the data, sort the block the blocks by their density, and then scan the blocks iteratively and merge blocks. The order of the merges forms a hierarchy. It is crucial to determine a proper criterion to merge grids and to terminate the clustering.

We propose DYWIN (DYnamic WINdow- based) clustering algorithm with full consideration of the complexities of the mass Spectrometry data to overcome the above disadvantages.

## 4 DYNAMIC WINDOW-BASED CLUSTERING

In drug development, it is difficult to predict the number of biologically active compounds. Also, it is difficult for us to predict the number of functional molecules in the samples. In other words, it is difficult to input the value of  $K$  for nonhierarchical clustering.

For density-based or grid-based hierarchical methods, we divide the entire data space into grids, use the density of each grid as a criterion to merge clusters, and obtain the dominant clusters. According to the features of our data, the points in a single cluster are within a range  $W$  [ $\Delta t$ ,  $\Delta m$ ] of the two dimensions of *Time* and *Mass*. When two grids are merged, the size of the new cluster should be within the range  $W$  also. Therefore, it is not easy to select the size of the grids. If the size of the grid is too small, performance of the algorithm will deteriorate, if too large, the merging process will fail.

Furthermore, in our case, we need not search the whole data space. Rather, we can use the features of the data outlined in section 2 to assist in clustering.

The algorithm presented here is based on work by Andrew Moore [3] and Dan Pelleg [1] on grid-based algorithms, but we use dynamic windows defined on the instance space instead of hyper-

rectangles. We use the density of the window, but do not split the entire space into grids. The positions of windows are not fixed, so we call it *dynamic window-based clustering* (DYWIN).

Assume the number of frogs is  $T$ , input data  $P_i = (t_i, m_i)$ ,  $W$  is a window with width  $\Delta t$  in the time dimension and height  $\Delta m$  in the mass dimension.  $D_i$  is the density of  $P_i$  in each  $W_i$  corresponding to cluster  $C_i$ . Figure 4 shows a simplified data space containing from three different frogs.

The algorithm DYWIN developed contains two main steps: to eliminate noise in the data from same frogs to get the real molecule data and to search the common functional molecules between different frogs that have the similar pharmacological significance.

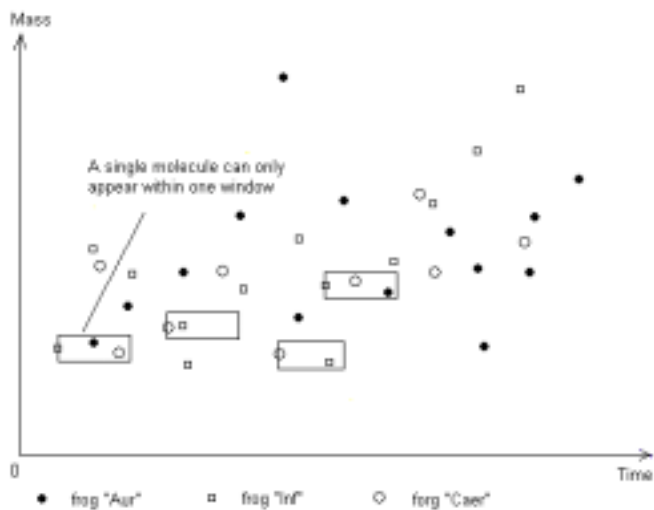


Figure 4. Clustering based on dynamic window

**Step-1** Remove noise in data of same frog, the algorithm is shown below:

```

Cluster1( )
{
  Input  $\Delta t, \Delta m$ 
  Sort Data  $X = \{P_1, P_2, \dots, P_n\}$ 
  Candidates =  $X$ 
  While Candidates exist
  {
    pick next Candidates  $P_i = (t_i, m_i)$ 
    define Window,  $W_j$  by the 4 vertices
       $\{(t_i, m_i + \Delta m/2), (t_i + \Delta t, m_i + \Delta m/2),$ 
       $(t_i, m_i - \Delta m/2), (t_i + \Delta t, m_i - \Delta m/2)\}$ 
     $k = 1, C_j = 0$ 
    while  $i + k$  is contained in  $W_j$ 
    {
       $k++$ 
      Candidates = Candidates -  $\{P_{i+k}\}$ 
    }
  }
}

```

$$C_j = C_j \cup \{P_{i+k}\}$$

Here,  $\Delta t, \Delta m$  are empirical values provided by chemists.

After data of each frog is clustered to remove the noise, we can move to the next step.

**Step-2** Search the function molecules between different frogs

Here, the input data is from different frogs within which the noise has been eliminated. A density threshold  $Td$  is an input value depends on what we expect to find in these different frogs. When we are searching for the common molecules that appear in all the frogs with same functions,  $Td$  is set to the number of the types of the frogs. If we want to query about the molecules which might play a special function appears in two frogs of all the frogs clustered by us, then is equal to 2. If the function is only appeared in one frog, the value of  $Td$  is set to 1.

```

Cluster2( )
{
  Input  $Td$ 
  Input data  $\{P_1, P_2, \dots, P_n\}$  collected from
                                     different frogs
  Cluster1( )
  Calculate density  $D_i$  for each Cluster  $C_i$ 
  If  $D_i \geq Td$  then
    keep  $C_i$ 
    output  $C_i$ 
  end if
}

```

## 5 RESULT AND DISCUSSION

We have applied this algorithm to real mass spectrometry data generated from HPLC-MS and have successfully identified the molecule components in the different frogs.

To date, we have got data from 12 types of frogs. Three of them, "Aur", "Inf" and "Caer" are different type of frogs from the same species, and have some of the same active functions.

After clustering the data to remove the noise in the data, "Aur" contains about 24500 molecules, "Inf" has about 50700 molecules and "Caer" contains about 40200 molecules. When using a value of 3 for  $Td$ , DYWIN output 42 clusters, in other words, 42 common molecules are distinguished in these three frogs. One of them is the molecule that we use to isolate the venom sample. It is definitely right to appear in the result. For  $Td = 2$ , we find 4192 clusters. Since the name of the frog is recorded during clustering, the result clusters show that there are about 1130 common molecules between "Inf" and "Caer" but not in "Aur".

Figure 5 give a simple explanation to the relations of the frogs, functions and molecules.

Frogs	Functions	Molecules
Aur	F1,F3,F4,F5	ABCDEF G
Inf	F1,F2,F5,F6	B D HIS
Caer	F1,F2,F4,F7	AB E H PQ

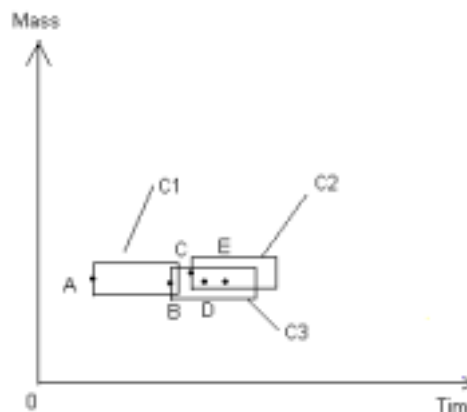
**Figure 5.** Relations of frogs, functions and molecules

When  $Td = 3$ , we get the molecule "B" which exists in all of the three frogs, then, we can infer that molecule "B" play an important role in the function "F1" which is the common function in these three frogs. When  $Td = 2$ , we also get the molecules "D", "E", and "H". According to the frog name from which we collect the data, we can also inform that "D" is important to function "F5" for frog "Aur" and "Inf", "E" may affect "F4" which appears in frog "Aur" and "Caer", and so on. While it is easy to manually decipher the SAR from the data in Fig 5, this is generally difficult to achieve in real data, as the number of molecules is very large. Thus we will, in the future, employ a classification technique for this purpose.

This work is an initial step in the complex process of drug development. We will get more data from the different types in same species with similar function to isolate and identify the functional molecules. Further more, we can establish the database from the results to distinguish the species and the types of different individual frogs by the functional molecule clusters.

The other point we are going to discuss is that, though the benefits of applying data specific clustering techniques are obvious, we still need to do further work on it.

From the experiments, we find the selection of the empirical value of  $\Delta t, \Delta m$  may affect the result. We find a case show as Figure 6.



**Figure 6.** A case need to be discussed

Dots A, B, C, D and E represent five points in the data space. According to DYWIN, molecule A and B are in the same cluster C1, and C, D and E are clustered to C2, though, it seems more reasonable to cluster B, C, D, and E into the same cluster C3. The problem is caused by the size of the window which is determined by  $\Delta t, \Delta m$ . We are considering applying adaptive values of  $\Delta t, \Delta m$  to determine the size of the windows. However this adjustment needs the supports from the chemists.

## ACKNOWLEDGEMENT

We would like to thank Professor Chris Shaw and Dr. Stephen McClean for providing the experimental data and giving helpful suggestions for analyzing the mass spectrometry data.

## REFERENCE

- [1] Dan Pelleg, Andrew Moore (99) Accelerating exact  $k$ -means algorithms with geometric reasoning, *KDD'99*, pp 277- 281
- [2] Ester M, Kriegel H, Sander J, Xu X(96) A density-Based Algorithm for discovering clusters in large spatial databases with noise, *Proceedings of 2<sup>nd</sup> international conference on KDD*
- [3] Andrew Moore (98) Very fast EM-based mixture model clustering using multi-resolution kd – tree, *Neural information processing system conference*,1998
- [4] Graham L. Patrick(1997), *An introduction to medical chemistry*, Oxford, pp 82-89
- [5] Kaufman L, Rousseeuw PJ (1990) *Finding Groups in data: an Introduction to Cluster Analysis*. John Wiley & Sons, Chichester
- [6] Rakesh Agrawal, Johannes Ges Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan (1998) Automatic subspace clustering of high dimensional data for data mining applications, *Proc. of the ACM SIGMOD Int'l Conference on Management of Data*, Seattle, Washington, June 1998
- [7] Stephen McClean(1999), PhD Thesis: *An Investigation of Modern Analytical Techniques for the Identification and Determination of Selected Drugs and Pollutants, their egradation Products and Metabolites*, University of Ulster, U.K
- [8] Wang W, Yang J, Muntz R (1997) STING: A Statistical Information Grid Approach to Spatial Data Ming, *Proceedings of the 23<sup>rd</sup> VLDB conference*, Athens, Greece, pp 186-195
- [9] Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: An Efficient Data Clustering Method for Very Large Database. In *Proceedings of the 1996 ACM SIGMOD International Conference on Mangement of Data*, Montreal, Canada, pp 103 – 114
- [10] Gholamhosein S,Surojit C and Aidong Z(2000) WaveCluster: a wavelet-based clustering approach for spatial data in very large databases, *VLDB Journal*(2000) 8:289-304