

Clustering and its application in multi-target prediction

William Liu¹ & Dale E Johnson^{1,2}

Address

¹University of California at Berkeley, Department of Nutritional Science & Toxicology
119 Morgan Hall, University of California, Berkeley, CA 94720-3104, USA
Email: will.c.liu@gmail.com

²Emiliem Inc, Christie Avenue,
Emeryville, CA 94608, USA
Email: daleejohnson@sbcglobal.net

Correspondence can be addressed to either author

Drug discovery teams are beginning to apply non-screening techniques to make early associations between chemical structure and various biological and druggability characteristics of compound series. The increasing availability of multiple data sets and models of target potency, ADME characteristics, and toxicity allows a researcher from any discipline to draw quick associations with multiple endpoints on sets of compounds or chemical scaffolds. Cluster analysis, for instance, can be used to correlate screening potency with data predicted from both freely available and commercially available models. In the future researchers will be able to draw chemical-biological associations 'on-the-fly' using various clustering or similarity techniques to determine whether the proposed toxicity of a drug is related to its chemical structure or its proposed efficacy mechanism. In this review associations are illustrated with target potency data gleaned from the literature associated with CYP450 substrate predictions from GeneGo's MetaDrug.

Keywords Chemical databases, cluster analysis, computational toxicology, CYP450, QSAR, similarity analysis

Introduction

Classical quantitative structure activity relationship (QSAR) analysis studies are based on a training set consisting of a large, diverse range of compounds. Ideally, these compounds span a large chemical space, such that new test compounds can be reliably evaluated and covered by the chemical space of the training set.

Frequently, however, the researcher is limited to a set of compounds which is neither large nor diverse. In many cases, restricted budgets dictate limited SAR analysis, and therefore a training set consisting of a limited number of chemical scaffolds. This can decrease model quality, and obscure the true SARs in the dataset. Furthermore, limited budgets can imply limited time scales; with the increased cost of drug development, quick and efficient methods of separating data into relevant groups become invaluable, as potential groupings of scaffolds may be non-obvious and difficult to separate without computational methods. Clustering analysis compensates for these limitations by efficiently and effectively separating compounds into distinct clusters, thus improving the predictive power (while limiting the scope) of each individual cluster. This leads to more relevant results, as smaller datasets consisting of just one, or a few, scaffolds controls for the similarities in structure between each compound. Consequently, the differences in structure between compounds are emphasized. A non-diverse set of compounds, consisting of just a few scaffolds, can now be divided into distinct datasets, increasing predictive power and allowing for a

clearer interpretation of structural characteristics which affect activity in each cluster.

In cases where datasets are large and contain a diverse range of compounds, cluster analysis can again play a role in drug development. While with smaller datasets, visual and expert driven clustering is a possible method of grouping compounds, when datasets become extremely large and consist of many scaffolds, it becomes almost impossible to manually sort compounds by structure. Again, computational methods are preferred in this case.

Clustering methodology

Clustering is not a recent invention, nor is its relevance to computational toxicology a recent application. Its theory, however, is often lost within black box treatments used by QSAR programs.

Clustering, in the general sense, is the grouping of objects together based on their similarity, while excluding objects which are dissimilar. One of the first applications of cluster analysis to drug discovery was by Harrison [1], who asserted that locales exist within the chemical space which favor biological activity. Consequently, these localities form clusters of structurally similar compounds. This idea that structure confers activity is also the fundamental premise of all QSAR analyses.

The basic framework for compound clustering consists of three main steps: the computation of structural features,

the selection of a difference metric, and the application of the clustering algorithm.

Molecular descriptors

Before compounds can be clustered, they must first be expressed in a way which is computationally interpretable. For traditional clustering, molecular descriptors are simply numerical, preferably continuous, representations of the structural characteristics of a compound, and can include measures of atom electrotopological state (E-state), connectivity indices, graph measures, and other molecular and physicochemical properties. These descriptors can then be represented as entries in a matrix and, consequently, can easily be used in statistical analyses as well as cluster analysis. The main drawback of traditional molecular descriptors is the possible presence of erroneous descriptors; these may not necessarily offer any strong differentiating measure between compounds, and could result in the grouping of compounds where similarity does not exist.

Other methods of clustering avoid the use of traditional molecular descriptors, instead focusing purely on structural keys and molecular fingerprints [2]. Structural keys can be thought of as descriptors that are assigned a binary value which represents the presence or absence of certain structural patterns [3]. Compounds are then represented as a bitmap, where each bit represents the absence or presence of a particular structural feature. The main advantage to using structural keys is the relative speed of analysis, as computers are quite adept at performing Boolean operations. In terms of clustering, because the composition of a structural key is determined by the researcher, the main drawback becomes a lack of generality. Compounds are difficult to fully represent with a small number of specific patterns expressed as binary variables. Instead, many structural patterns must be used, and not all are guaranteed to be of any significance. This can result in decreased efficiency, as erroneous variables can be selected in the structural key, leading to either poor results or wasted computation.

Molecular fingerprinting, as characterized by the Daylight methodology [3], is an alternate method of describing compounds, where structures are again expressed as bitmaps. Unlike structural keys, molecular fingerprinting is not based on a preset number of structural patterns. Described best by Butina [4], the molecular fingerprint of each compound is generated algorithmically, and is composed of the following: a pattern for each atom, a pattern for each atom and its neighbors, and a set of patterns which describes groups of atoms and their bonds, ranging from paths of 2 to 7 bond lengths. The set of patterns generated is quite large and unique. Because patterns are generated uniquely, and are not predefined, it becomes impossible to express each pattern as a single bit. Instead, each pattern is hashed, and an output of a small set of bits, typically 4 to 5, is used to represent the pattern. This method of describing molecules using non-predefined patterns offers a more complete measure

of the composition of a compound, compared to both traditional molecular descriptors and structural keys.

The fingerprinting and structural keys are composed of descriptors whose values are represented either as bits or as bit-strings. Meanwhile, traditional descriptors are typically not encoded as bit-strings, but rather as numerical values. This implies that matrices made up of traditional descriptors are more easily interpretable, but are not composed of as many elements as fingerprints or structural keys.

Similarity measures

After a suitable method of describing the compounds in the dataset, the next step in clustering is the selection of a similarity measure. In order for compounds to be grouped, there must be some metric defining the differences between compounds. Some of the most popular similarity measures include: Euclidean [5], city-block [6], Minkowski [6], Pearson's correlation coefficient [7], and cosine [8]. A brief summary of each can be seen in Table 1. It should be borne in mind that Pearson's coefficient is not necessarily a measure of similarity as in the other metrics. Rather, it is a measure of the correlation of vectors (association), as opposed to a linear measure of distance. While this is certainly a useful measure, this is not necessarily a good means of determining differences between chemical structures.

The most common method of determining similarity is the Euclidean distance metric [9]. In the context of QSAR analysis, the most-used distance measure is an algorithm which determines the differences in value of molecular descriptors between all compounds. Euclidean distance is typically the optimal metric, given its simplicity and ease of interpretation. It is important to note, however, that for datasets composed of molecular descriptors that have large differences in magnitude, variables must first be normalized, to prevent descriptors characterized by large values from overshadowing descriptors whose values are levels of magnitude smaller. Furthermore, while Euclidean distance can be used to measure datasets composed of continuous or ordinal categorical variables (where magnitude matters), in cases where variables are either nominal (discrete categories without order) or a mix of nominal and continuous, different measures of distance are required [10].

The heterogeneous value difference metric (HVDM) [11], which is shown in Table 2, is an example of a distance metric that can use discrete variables that are either continuous or nominal.

HVDM is the successor of the heterogeneous Euclidean overlap metric (HEOM), which, in comparing observations, separates out variables which are continuous and noncontinuous. Differences between continuous variables are processed in a normalized city-block manner, while nominal variables are processed using an overlap function, where a difference in a nominal variable between

Table 1. Summary of popular similarity measures used for the clustering of compounds.

Distance measure	Description
$d(x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$	Euclidean distance [5].
$d(x,y) = \sum_{i=1}^n x_i - y_i $	City-block distance [6].
$d(x,y) = \left(\sum_{i=1}^n x_i - y_i ^p \right)^{1/p}$	Minkowski distance, for $p = 1$, this measure becomes city-block. For $p = 2$, this measure becomes Euclidean [6].
$d(x,y) = \frac{\sum_{i=1}^n (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_i)^2 \sum_{i=1}^n (y_i - \bar{y}_i)^2}}$	Pearson's Correlation, \bar{x}_i , \bar{y}_i are the average values for attribute i in observation x,y [7].
$d(x,y) = \frac{\sum_{i=1}^n (x_i)(y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2} \sqrt{\sum_{i=1}^n (y_i)^2}}$	Cosine Correlation [8].

observations is assigned a '1', and a '0' is assigned to all other cases. A more robust method of measuring differences in discrete, nominal values is the value difference metric (VDM). The VDM measure of distance between two observations is determined by comparing the class conditional probability distribution for the values that each observation takes on for each variable [12]. The HVDM, similar to HEOM, works in a conditional manner; if a variable is continuous, then a normalized city-block distance metric is used. If a variable is nominal, then the VDM method is used. The square root of the sum of these squared distances is the distance between two observations. Although this method has yet to reach prominence in the context of QSAR cluster analysis, a recent study by Guo *et al* [13] applied this distance metric to a feature-selection analysis of phenols. Because traditional methods of clustering examine descriptors which are either continuous or nominal, and not both simultaneously, this method holds promise in deriving distances between compounds with a large number of continuous and categorical values.

Another popular method of measuring differences between compounds is the Tanimoto coefficient [14], which is particularly effective in cases using molecular fingerprints and structural keys. Because the values used by fingerprints and keys are binary, the Tanimoto similarity can be expressed in the following manner: $C/(A+B-C)$, where C is the number of values in the fingerprint/key for both observations which has a value of 1, or is 'present/positive'. A , B are the number of values which are 'positive/present' or 'positive' in A or B , respectively [15].

Table 2. The heterogeneous value difference metric (HVDM), an example of a distance metric that can use discrete variables that are either continuous or nominal.

Steps	Equations
1.	$HVDM(x,y) = \sqrt{\sum_{i=1}^n d_i^2(x_i,y_i)}$
2.	$d_i(x,y) = \begin{cases} 1 & \text{if } x,y \text{ unknown, else} \\ \text{diff}_i(x,y) & \text{if } i \text{ is linear} \\ \text{vdm}_i(x,y) & \text{if } i \text{ is nominal} \end{cases}$
3a.	$\text{diff}_i(x,y) = \frac{ x-y }{4\sigma_i}$
3b.	$\text{vdm}_i(x,y) = \sqrt{\sum_{c=1}^c \left \frac{N_{i,x,c}}{N_{i,x}} - \frac{N_{i,y,c}}{N_{i,y}} \right ^2}$

HVDM heterogeneous value difference metric.

Clustering algorithms

Clustering algorithms can be divided into two broad realms: hierarchical and nonhierarchical partitioning [16]. Hierarchical clustering can be further divided into two groups: agglomerative and divisive. The general idea behind agglomerative clustering is that clusters are formed

from the smallest clusters, which eventually form to become part of the largest. Each individual compound is considered a cluster, and as the algorithm proceeds, each compound is absorbed into larger and larger clusters, until the dataset is expressed as a single cluster composed of all compounds. Conversely, divisive clustering works in the opposite direction, whereby the initial dataset is considered one large cluster, and is slowly divided into smaller and smaller clusters [17]. In general, agglomerative clustering is the faster type of hierarchical clustering, and as a consequence, is the most popular. In the realm of agglomerative clustering, Ward's method is among the most popular [18]. Methods of non-hierarchical clustering include K-means and Jarvis-Patrick clustering [19,20,21]. A summary of popular clustering methods is presented in Figure 1.

Hierarchical clustering is a method of representing a dataset as a collection of hierarchically-arranged subsets. These subsets are determined by some measure of proximity [17]. In classical QSAR analysis, this measure of proximity is determined by measuring the differences in values of molecular descriptors between different compounds.

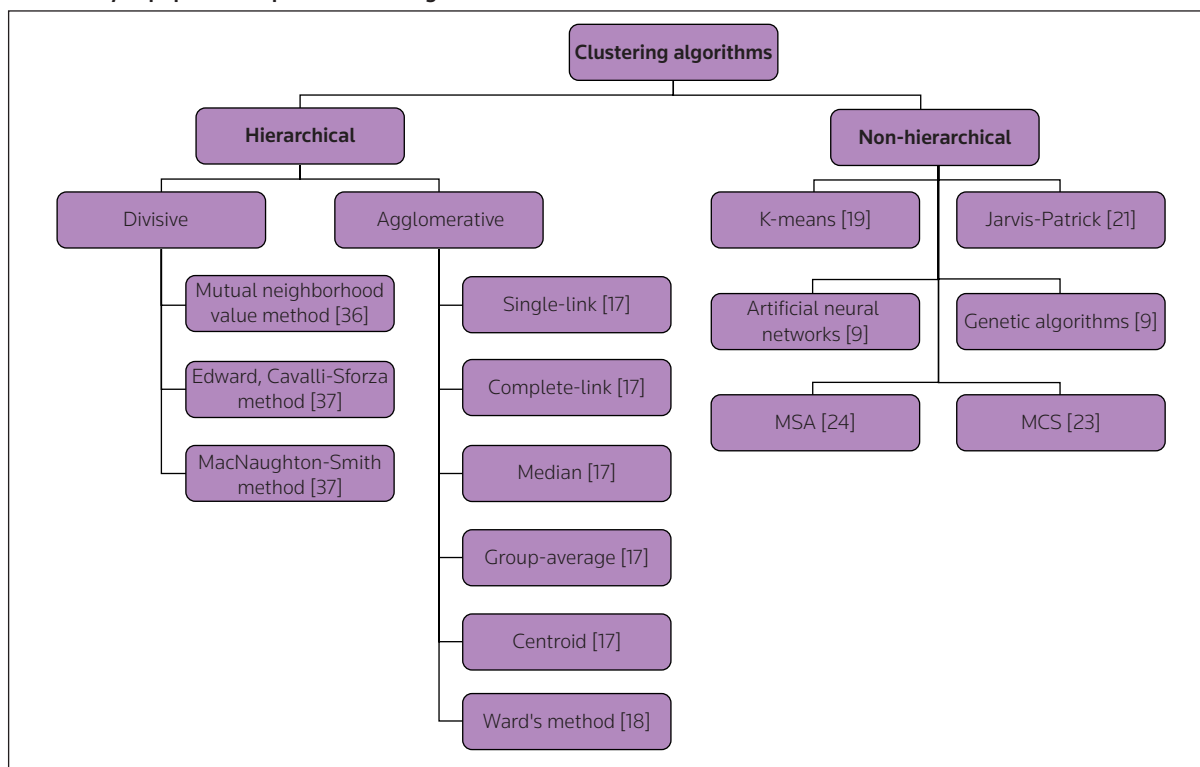
The generalized procedure for agglomerative hierarchical clustering is as follows: after descriptor calculation, a distance metric must be selected. As mentioned previously, this depends on the dataset; binary values are better described using a Tanimoto metric, while continuous

values are better described using a Euclidean metric. After a suitable metric is selected, a distance matrix is created, which, using a particular metric, is composed of rows which express the differences in molecular descriptor values between all of the observations. After this matrix has been created, it is then searched for the two observations which are the closest. These two observations are then combined, and a new matrix is created, with distances between this combination of observations and the remaining observations. Consequently, this results in one less row in the distance matrix. This procedure of combining observations into groups is repeated until all of the observations are integrated into one group consisting of all observations [18].

Ward's method of clustering follows the previously discussed procedure, but includes a measure of error as a metric of combining clusters. The 'error sum of squares' (ESS) is defined as the sum of the squared difference in a certain molecular descriptor for a particular observation in a given cluster and the average value of that molecular descriptor in the cluster. Clusters are combined if the increase in ESS between stages is minimized. It is important to emphasize that Ward's method does not create an optimal set of clusters [17], and it remains up to the computational chemist to determine a logical clustering-stopping point.

In the realm of nonhierarchical clustering, the K-means algorithm is among the oldest and most popular [19].

Figure 1. Summary of popular compound clustering methods.



Nonhierarchical clustering differs from hierarchical clustering in that the former attempts to divide the data into a predefined number of nonrelated clusters, while the latter nests compounds into larger and larger clusters. The K-means algorithm, in particular, follows an iterative approach to nonhierarchical clustering [20]. Given a group of compounds, 'n' reference points are chosen, depending on the desired 'n' number of clusters. The distance of each compound to both reference points is computed. Cluster membership is determined by the distance of each compound to the closest reference point. The K-means algorithm initially follows this approach. After clusters are first computed, the reference point for each cluster is moved into the centroid of the cluster. The distance of this new reference point to each compound is recomputed. If a compound is closer to the reference point of another cluster, then the compound is reassigned to the other cluster. This process is repeated until the clusters are stable and successive trials yield no changes in cluster membership.

An alternate nonhierarchical clustering algorithm is the Jarvis-Patrick method [21]. This algorithm clusters compounds using a nearest-neighbor approach, where the distances between all compounds are computed, using the desired distance metric. Compounds are clustered together if they are suitably close to each other and have, to an extent, matching near-neighbors lists consisting of N near neighbors, where N is some user-defined number. Intuitively, the smaller the value of N, the more inclusive the clusters, while the larger the value of N, the more exclusive the clusters.

Recent developments

The previously described clustering methods have become widely adopted in the field of QSAR analysis. Although the methods remain effective, increased computing power since 2003 has resulted in the introduction of a host of new techniques in cluster analysis.

Maximum common structure (MCS) and macrostructure assembly (MSA) are methods which hold promise in cluster analysis. MCS is a method of searching for the largest substructure in a collection of graphs [22]. The application of MCS to cluster analysis has been characterized by Stahl and Mauser [23]. The procedure begins by creating molecular fingerprints of a dataset which are then clustered according to an exclusion-sphere algorithm. The MCS is then determined for each cluster. For each MCS, a neighbor list is created, where the scaffold for each cluster is compared to scaffolds of other clusters. The singleton clusters created in the first step can now be compared to the MCS of other clusters. Along with these singletons, new clusters can also be formed from clusters with closely related MCSs.

MSA is a related method of structural searching, but also has an inherent application to compound clustering. This method, by Cross *et al* [24], begins with an initial set of predetermined molecular building blocks, which are

then reassembled into larger structures. This process is repeated until an appropriate set of MSAs is reached. These MSAs are defined as a substructural signature, which can then be used to discriminate for membership in a cluster.

Application

Suppose an investigator is given a moderately large set of unique compounds along with their corresponding binding data. Some compounds could be considered active, while others could be considered inactive. By grouping compounds into clusters with similar structures, the researcher could add a greater degree of specificity to the QSAR models. Without any knowledge of the source of the compounds, the investigator would, traditionally, have to separate compounds based on a visual inspection. This could yield erroneous results, as superficial similarities may not indicate accurate groupings. Even if the investigator were to have prior knowledge about the source of the compounds, and could distinguish between the unique scaffolds in the dataset, this does not preclude the use of a clustering algorithm. Scaffolds with a small set of compounds would not make for great models, and could contain structural characteristics that correlate well with structural features of other scaffolds.

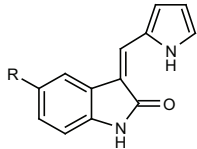
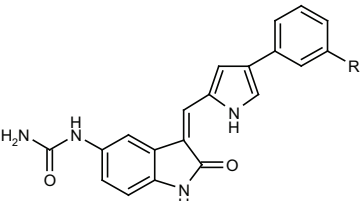
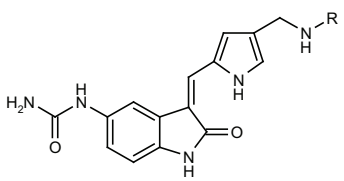
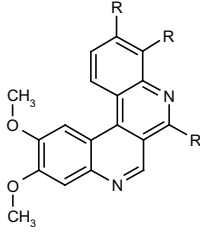
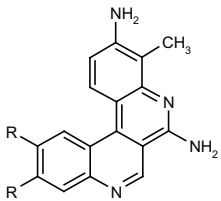
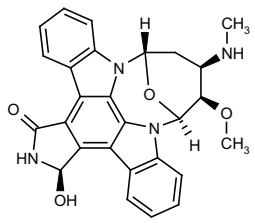
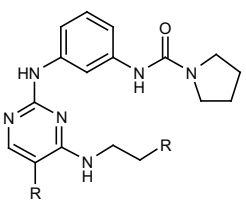
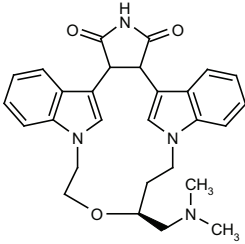
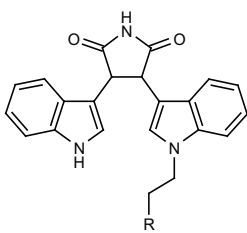
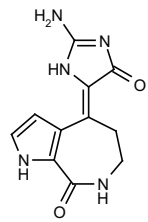
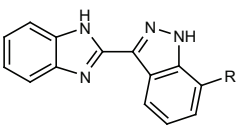
Furthermore, clustering could provide hints about a particular dataset's ability to bind other proteins. For example, in drug development, a researcher could quickly gain important knowledge about the binding ability of clustered lead compounds to off-target enzymes by simply examining the structures prevalent in clusters with increased off-target binding.

One application can be demonstrated as follows. A set of compounds designed to bind pyruvate dehydrogenase kinase, isozyme 1 (PDK1) was compiled, consisting of compounds with distinct scaffolds. PDK1 is a particularly attractive target of inhibition because of its position as an upstream regulator in the phosphoinositide 3-kinase/protein kinase B/mammalian target of rapamycin (PI3K/AKT/mTOR) signaling pathway [25]. Because these compounds are potential cancer therapeutics, it is important to also characterize the metabolic effect these compounds may have, in order to determine effective and ineffective dose levels as compounds are moved forward. These compounds were first manually clustered based on source, then computationally clustered. Models, based on PDK1 binding data, were created using each of the clusters. Using the computationally-derived clusters, associated cytochrome P(CYP)450 Michaelis constant (K_m) values were obtained. The clustering of compounds resulted in significant differences in K_m values between clusters, implying that clustering analysis could be used to determine the scaffolds with enhanced CYP450 substrate specificity.

Method

A dataset modeling PDK1 inhibition, consisting of 98 compounds derived from seven literature sources [26-32] was compiled using bindingdb.org [33,34].

Table 3. Computational clustering of a set of pyruvate dehydrogenase kinase, isozyme 1 (PDK1)-binding compounds using Ward's algorithm.

 <p>Manual Cluster: 2 Ward Cluster: 1,3 Source: 28</p>	 <p>Manual Cluster: 1 Ward Cluster: 1,3 Source: 29</p>	 <p>Manual Cluster: 1 Ward Cluster: 1,3 Source: 29</p>
 <p>Manual Cluster: 3 Ward Cluster: 2 Source: 26</p>	 <p>Manual Cluster: 3 Ward Cluster: 1,2 Source: 31</p>	 <p>Manual Cluster: 4 Ward Cluster: 2 Source: 27</p>
 <p>Manual Cluster: 4 Ward Cluster: 1 Source: 32</p>	 <p>Manual Cluster: 4 Ward Cluster: 1,2 Source: 31</p>	 <p>Manual Cluster: 4 Ward Cluster: 1 Source: 27</p>
 <p>Manual Cluster: 4 Ward Cluster: 3 Source: 30</p>	 <p>Manual Cluster: 4 Ward Cluster: 3 Source: 30</p>	-

These 98 compounds, along with their associated IC_{50} values, were then negative log transformed. Using MDL Information Systems' MDL-QSAR technology, 147 molecular descriptors were generated, and included measures of E-states, molecular connectivity, kappa shape indices, and other molecular properties. Cluster analysis was then performed on this dataset using Ward's algorithm, with standardized values. Manually-derived clusters were grouped according to source; three sources with distinct scaffolds were seen. Because compounds

from the remaining four sources were too few to group into individual clusters, they were grouped into a fourth cluster. Compounds shared between sources 28 and 29 were grouped into cluster 2.

Significant descriptors were found using a genetic algorithm, run with 1000 generations. The pIC_{50} values were then regressed on the molecular descriptor values, and coefficient of determination (R^2), significance, and cross-validation statistics were computed.

Km values for CYPs 1A2, 2B6, 2C9, 2C19, 2D6, and 3A4 were calculated for all 98 compounds in the dataset using GeneGo's MetaDrug CYP450 enzyme substrate models. Because GeneGo's MetaDrug uses a Tanimoto similarity measure to determine the similarity of the test compound with the compounds in MetaDrug's metabolic enzyme models, some attention was paid to this value. However, for the purpose of broadly illustrating the potential application of clustering in directing drug design, even compounds with a Tanimoto similarity percentage of less than 50% were included in this study. Finally, ANOVA testing was performed in the statistical package, R.

Results and discussion

Running cluster analysis on a dataset consisting of obviously differential groupings of compounds allows for a sanity check of the clustering algorithm. Ward's method of clustering satisfactorily separated the compounds extracted from different literature sources. Computational clustering using Ward's algorithm resulted in three groups, as shown in Table 3.

Clusters 1 and 3 consisted of, for the most part, indolinone compounds. Meanwhile, cluster 2 consisted mostly of various forms of dibenzo[c,f][2,7]naphthyridines. The main difference in structure between the indolinone compounds in clusters 1 and 3 was the size of substituents to the main indolinone scaffold. Substituents for the indolinone compounds in cluster 1 were large, consisting of ring structures, and attached to the pyrrole, as opposed to the indolinone compounds in cluster 3, which were smaller and bound to an alkene connected to the main pyrrolidine structure in the scaffold. Compounds that were too few to cluster were found to be scattered within the three computationally-derived clusters. The results presented in Table 4 show that, overall, regression statistics for the computationally-derived clusters were comparable to the regression statistics for the manually-created clusters.

In general, besides cluster 4, which featured abnormally high R^2 and cross-validation statistic (Q^2) values, the computationally-clustered groups fared better than the

manually-clustered groups. This was especially evident in terms of their overall Q^2 values, which is of particular importance because it offers a superior measure of the predictive power of the models, rather than R^2 . It is important to note, however, that, because the clusters contained different numbers of compounds, this comparison of model statistics is simply a non-exact illustration of the effectiveness of automated clustering. In many cases, computational clustering results in clusters which can be highly predictive, and can often be used in lieu of manual clustering, especially in cases of large and diverse datasets.

In cases where the dataset is large, computationally clustering compounds is the preferred method of grouping compounds by structural similarity. This result becomes useful in drug design; models created based on a cluster of compounds with a certain scaffold controls for that particular scaffold, resulting in models which can point the researcher toward relevant substituents that strongly affect activity.

It can be seen from Table 5 that, for CYPs 1A2, 2B6 and 3A4, significant differences exist between clusters 1, 2, and 3 in terms of CYP Km values. This implies that the clustering of compounds results in a strong degree of variation in average Km values.

Data presented in Figure 2 demonstrate that compounds from cluster 2 are markedly different in terms of CYP1A2 Km values. These results correlate well with the cluster analysis; clusters 1 and 3 consist of the indolinone compounds while cluster 2 consists of dibenzo[c,f][2,7]naphthyridines.

Interestingly, in examining CYP2B6, there appears to be a strong distinction in terms of mean Km between cluster 1 and the other two clusters. This suggests that strong structural differences exist between clusters 1 and 3, which also help to dictate Km levels.

While CYPs C9, C19, and 2D6 do not have significant differences in Km between clusters, CYP3A4 exhibits

Table 4. Comparison of regression statistics for the manually-created and computationally-derived clusters of a set of pyruvate dehydrogenase kinase, isozyme 1 (PDK1)-binding compounds.

Cluster	R^2	RMSE	Q^2	F-stat	Number of observations
Manual1	0.6022	0.2582	0.4498	10.22	32
Manual2	0.7222	0.5478	0.6354	20.8	38
Manual3	0.6322	0.5690	0.5024	12.03	15
Manual4	0.9407	0.3783	0.9018	47.58	13
C1	0.7889	0.3609	0.666	20.55	55
C2	0.8778	0.3472	0.7655	28.72	16
C3	0.7543	0.5323	0.659	52.18	27

R^2 coefficient of determination, **RMSE** root mean squared error, Q^2 cross-validation statistic, **F-stat** F-statistic, **Manual** manually-created cluster, **C** computationally-derived cluster.

Table 5. Statistical comparison of clusters 1, 2 and 3 in terms of cytochrome P Michaelis constant (Km) values.

Enzyme	F-stat	Prob > F
CYP1A2	<u>31.44</u>	0.00000
CYP2B6	<u>28.13</u>	0.00000
CYP2C9	1.58	0.2110
CYP2C19	1.05	0.3551
CYP2D6	1.13	0.3267
CYP3A4	<u>8.30</u>	0.0005

F-stat F-statistic, **Prob > F**, probability that the result (F-stat) was obtained, given there is no difference in Km values between clusters, **CYP** cytochrome P450, underlined values imply statistical significance.

a strong difference between cluster 3 and the other 2 clusters. This variation in clusters with marked differences in Km values implies that there are structural features unique to each cluster that determine Km value. These results indicate that inhibitors based on the dibenzo[c,f][2,7]naphthyridine scaffold have a strong effect on CYP1A2, while inhibitors based on the indolinone scaffold have differing effects on CYP2B6 and CYP3A4, depending on the nature and position of substituents.

For the drug developer, if the dibenzo[c,f][2,7]naphthyridines prove to be a particularly strong inhibitor of PDK1, it becomes important to recognize that because it has, on average, a higher Km value, it may be a weaker binder to CYP1A2 compared with the other scaffolds.

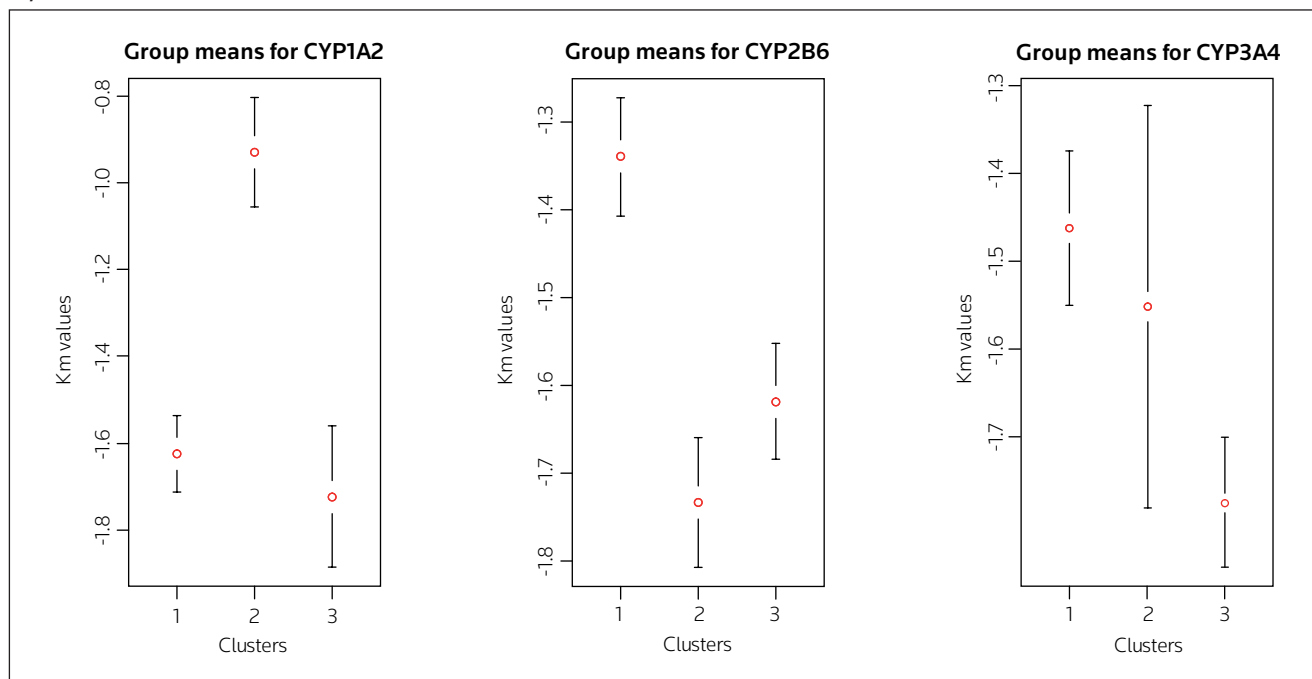
Metabolic characteristics of a resulting lead can therefore be used to further optimize prior to, or in place of, actual screening.

A possible next step in analysis would be to either use a regression method to extract the important descriptors that influence CYP1A2 binding, or use a classification algorithm, which could separate out substructures within cluster 2 that are highly correlated with activity. These methods could further enhance the specificity of drug design.

Conclusions

Statistically significant differences in CYP Km values between clusters suggest that performing cluster analysis on large sets of data can yield important clues to designing drugs to target specific metabolic enzymes.

The use of clustering-directed studies to determine scaffolds with significant metabolic consequences becomes particularly useful, given the concept of ethnic variation in drug metabolism [35]. As the field of pharmacogenetics develops, and as more knowledge is gained in the differing expression of P450 enzymes between ethnicities, it becomes increasingly important to quickly identify groups which could have significant or non-significant differences in drug metabolism. For example, if a particular population were to have a large number of individuals with low or altered expression of CYP1A2, a drug targeting PDK1 may have unintended consequences relating to toxicity or efficacy in these groups. In keeping with the PDK1 example, because it has been characterized

Figure 2. Graphical comparison of mean Michaelis constant (Km) values for clusters 1, 2 and 3, in terms of cytochrome Ps (CYPs) 1A2, 2B6 and 3A4.

CYP cytochrome P, **Km** Michaelis constant. Error bars represent standard errors above and below the mean.

as a possible therapeutic in treating breast cancer, it may be possible to tailor a compound to be more effective in populations with known ethnic variations in drug metabolism.

Certainly, this type of analysis is not limited to metabolic enzymes; clustering-directed studies could include values for *hERG* inhibition or other toxicology endpoints. Potentially, this strategy can be applied to carcinogenicity and mutagenicity data. The wide availability of these data is increasing both as freely-available datasets/models and in commercially-available models, such as Genego's MetaDrug. Clustering in the opposite direction, ie, from a toxicity endpoint into various chemical scaffolds, has application in drug research as well as in environmental chemistry. Such an approach would require that the toxicity models are open-source to allow clustering based on knowledge of structures used to create models.

This brief study is an example of how cluster analysis can be applied to drug discovery, and how it could form a valuable tool in a researcher's toolbox. As computational power increases, chemical databases invariably increase as well. Clustering these databases into manageable and biologically interesting groups increases the effectiveness and efficiency of lead drug development.

References

- of outstanding interest
- of special interest

1. Harrison PJ: **A method of cluster analysis and some applications.** *J Appl Stat* (1968) **17**(3):226-236.
 - This publication was among the first to cite the usefulness of clustering to chemical analysis. This paper is consistently cited in the literature.
2. Raymond JW, Blankley CJ, Willett P: **Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures.** *J Mol Graph Model* (2003) **21**(5):421-433.
3. **Daylight theory - fingerprints:** Daylight, Aliso Viejo, CA, USA (2008). www.daylight.com/dayhtml/doc/theory/theory.finger.html
 - An overview of the Daylight method of fingerprinting, a popular method of representing chemical compounds in the context of creating computational models.
4. Butina D: **Unsupervised database clustering based on Daylight's Fingerprint and Tanimoto Similarity: A fast and automated way to cluster small and large data sets.** *J Chem Inform Comput Sci* (1999) **39**(4):747-750.
5. Black PE: **Euclidean distance.** In: *Dictionary of Algorithms and Data Structures*. Black PE (Ed), US National Institute of Standards and Technology, Gaithersburg, MD, USA (2004). www.nist.gov/dads/HTML/euclidndstnc.html
6. Schneider G: **Drug design.** *Eurekah Bioscience Collection*. Landes Bioscience, Austin, TX, USA (2000-2005).
7. Chen CC, Chu HT: **Similarity measurement between images.** *29th Annual International Computer Software and Applications Conference* (2005).
8. Korenius T, Laurikkala J, Juhola M: **On principal component analysis, cosine and euclidean measures in information retrieval.** *Inform Sci* (2007) **177**(22):4893-4905.
9. Jain AK, Murty MN, Flynn PJ: **Data clustering: a review.** *ACM Comput Surv* (1999) **31**(3):264-323.
 - A comprehensive review of 'classical' clustering analysis, with a brief exploration of evolutionary, genetic algorithm, and simulated annealing-based methods of clustering.
10. Finch H: **Comparison of distance measures in cluster analysis with dichotomous data.** *J Data Sci* (2005) **3**(1):85-100.
11. Wilson R, Martinez T: **Improved heterogeneous distance functions.** *J Artif Intell Res* (1997) **6**:1-34.
 - This paper explores situations where datasets are composed of continuous and nominal data – datasets which are of particular interest in the realm of chemical clustering.
12. Payne T, Edwards P: **Implicit feature selection with the Value Difference Metric.** *Proceedings of the 13th European Conference on Artificial Intelligence* (1998):450-454.
13. Guo G, Neagu D, Cronin MTD: **Using kNN model for automatic feature selection.** *Lect Notes Comput Sci* (2005) **3686**:410-419.
14. Tanimoto TT: **An elementary mathematical theory of classification and prediction.** *IBM Internal Report* (1957) 17th Nov.
 - This article is a commonly cited article in the realm of similarity analysis. This measure of similarity is considered an extension of the Jaccard similarity coefficient.
15. Willett P, Barnard JM, Downs GM: **Chemical similarity searching.** *J Chem Inf Model* (1998) **38**(6):983-996.
16. Willett P, Winterman V, Bawden D: **Implementation of nonhierarchical cluster analysis methods in chemical information structure search.** *J Chem Inform Comput Sci* (1986) **26**(3):109-118.
 - This paper explores the effectiveness of various methods of clustering, and concludes that in certain cases, Jarvis-Patrick is a particularly effective method of clustering.
17. Hubert LJ: **Hierarchical cluster analysis.** In: *Encyclopedia of Statistical Sciences*. John Wiley & Sons Inc, Hoboken, NJ, USA (2006).
18. Mojena R: **Ward's clustering algorithm.** In: *Encyclopedia of Statistical Sciences*. John Wiley & Sons Inc, Hoboken, NJ, USA (2006).
19. Hartigan JA, Wong MA: **Algorithm AS 136: A K-Means clustering algorithm.** *J Appl Stat* (1979) **28**(1):100-108.
 - This paper, while not the initial paper in describing the K-means method of clustering, is instead an efficient and often-used method of K-means clustering.
20. Faber V: **Clustering and the continuous k-means algorithm.** *Los Alamos Science* (1994) **22**: 138-144.
21. Jarvis RA, Patrick EA: **Clustering using a similarity measure based on shared near neighbors.** *IEEE Trans Comput* (1973) **22**(11):1025-1034.
22. Raymond JW, Willett P: **Maximum common subgraph isomorphism algorithms for the matching of chemical structures.** *J Comput Aided Mol Des* (2002) **16**(7):521-533.
23. Stahl M, Mauser H: **Database clustering with a combination of fingerprint and Maximum Common Substructure methods.** *J Chem Inform Model* (2005) **45**(3):542-548.
24. Cross KP, Myatt G, Yang C, Fligner MA, Verducci JS, Blower PE Jr: **Finding discriminating structural features by reassembling common building blocks.** *J Med Chem* (2003) **46**(22):4770-4775.
 - This paper describes a method of clustering currently employed by the Leadscape Predictive Data Miner, a popular application under use by regulatory and industry groups.
25. Liang K, Zeng X, Glazer RI, Jin W, Mills GB, Fan Z: **PDK1 as a preferred target for sensitizing breast cancer cells to gemcitabine.** *Proc Am Assoc Cancer Res* (2004) **45**:Abs 3857.
26. Gopalsamy A, Shi M, Boschelli DH, Williamson R, Olland A, Hu Y, Krishnamurthy G, Han X, Arndt K, Guo B: **Discovery of dibenzo[c,f][2,7]naphthyridines as potent and selective 3-phosphoinositide-dependent kinase-1 inhibitors.** *J Med Chem* (2007) **50**(23):5547-5549.

27. Komander D, Kular GS, Schüttelkopf AW, Deak M, Prakash K, Bain J, Elliot M, Garrido-Franco M, Kozikowski AP, Alessi DR, Van Aalten DM: **Interactions of LY333531 and other bisindolyl maleimide inhibitors with PDK1.** *Structure* (2004) **12**(2): 215-226.
28. Islam I, Bryant J, Chou YL, Kochanny MJ, Lee W, Phillips GB, Yu H, Adler M, Whitlow M, Ho E, Lentz D *et al*: **Indolinone based phosphoinositide-dependent kinase-1 (PDK1) inhibitors. Part 1: Design, synthesis and biological activity.** *Bioorg Med Chem Lett* (2007) **17**(14):3814-3818.
29. Islam I, Brown G, Bryant J, Hrvatin P, Kochanny MJ, Phillips GB, Yuan S, Adler M, Whitlow M, Lentz D, Polokoff MA *et al*: **Indolinone based phosphoinositide-dependent kinase-1 (PDK1) inhibitors. Part 2: Optimization of BX-517.** *Bioorg Med Chem Lett* (2007) **17**(14):3819-3825.
30. Foloppe N, Fisher LM, Francis G, Howes R, Kierstan P, Potter A: **Identification of a buried pocket for potent and selective inhibition of Chk1: Prediction and verification.** *Bioorg Med Chem* (2006) **14**(6):1792-1804.
31. Komander D, Kular GS, Bain J, Elliott M, Alessi DR, Van Aalten DM: **Structural basis for UCN-01 (7-hydroxystaurosporine) specificity and PDK1 (3-phosphoinositide-dependent protein kinase-1) inhibition.** *Biochemistry* (2003) **375**(Pt 2):255-262.
32. Feldman RI, Wu JM, Polokoff MA, Kochanny MJ, Dinter H, Zhu D, Biroc SL, Alicke B, Bryant J, Yuan S, Buckman BO *et al*: **Novel small molecule inhibitors of 3-phosphoinositide-dependent kinase-1.** *J Biol Chem* (2005) **280**(2):19867-19874.
33. Chen X, Lin Y, Liu M, Gilson MK: **The Binding Database: Data management and interface design.** *Bioinformatics* (2002) **18**(1):130-139.
 - *The Binding Database is an openly accessible repository of binding data for a large and diverse set of compounds for a variety of targets.*
34. Liu T, Lin Y, Wen X, Jorrisen RN, Gilson MK: **BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities.** *Nucleic Acids Res* (2007) **35**:D198-D201.
35. Johnson DE, Park K, Smith DA: **Editorial overview - Ethnic variation in drug response: Implications for the development and regulation of drugs.** *Curr Opin Drug Discov Dev* (2008) **11**(1):29-31.
36. Prakash S, Kumar SR, Nagabhushan P, Gowda KC: **Modified divisive clustering useful for quantitative analysis of remotely sensed data.** *Geoscience and Remote Sensing Symposium* (1996) **3**:1858-1860.
37. Chavent M: **A monothetic clustering method.** *Pattern Recogn Lett* (1998) **19**(11):989-996.