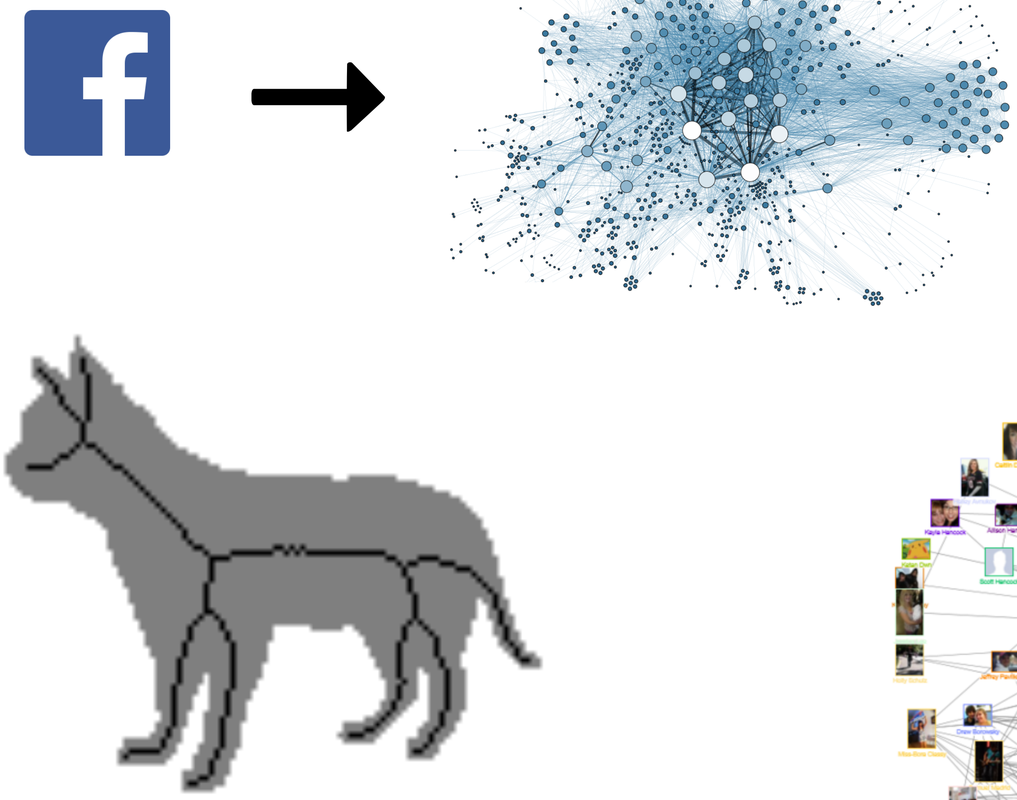
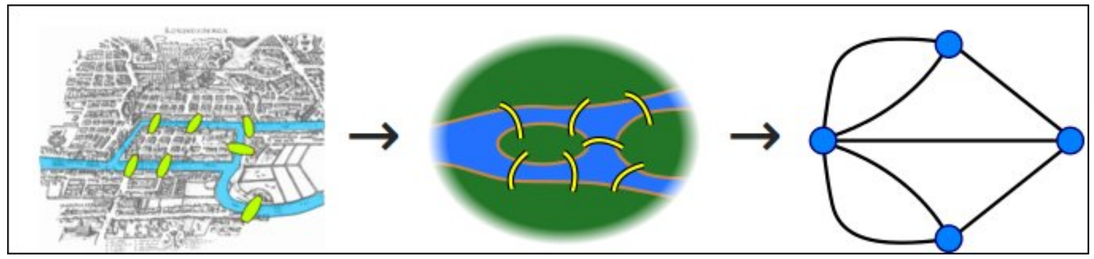
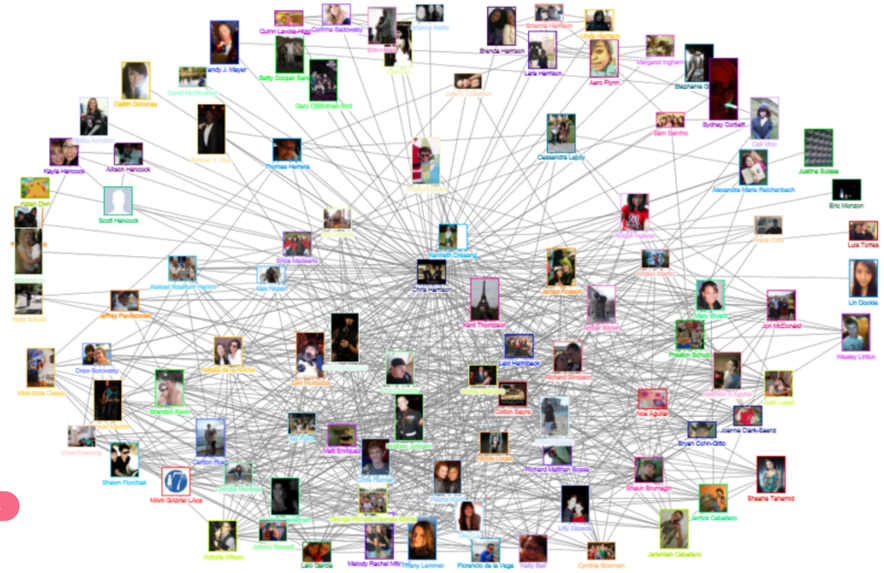
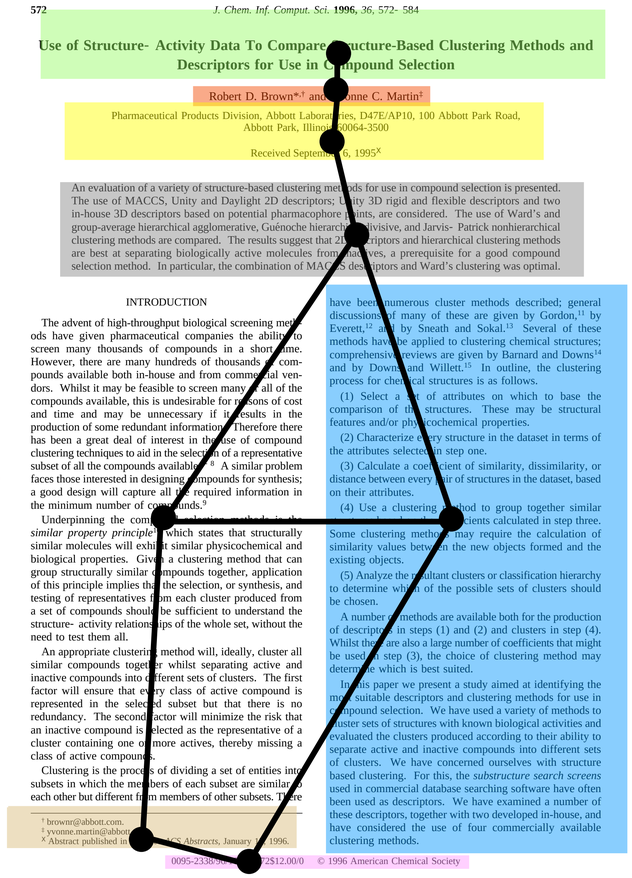








572 *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572- 584

Use of Structure- Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection

Robert D. Brown*,† and Yvonne C. Martin‡

Pharmaceutical Products Division, Abbott Laboratories, D47E/AP10, 100 Abbott Park Road, Abbott Park, Illinois 60064-3500

Received September 6, 1995[X]

An evaluation of a variety of structure-based clustering methods for use in compound selection is presented. The use of MACCS, Unity and Daylight 2D descriptors; Unity 3D rigid and flexible descriptors and two in-house 3D descriptors based on potential pharmacophore points, are considered. The use of Ward's and group-average hierarchical agglomerative, Guénoche hierarchical divisive, and Jarvis- Patrick nonhierarchical clustering methods are compared. The results suggest that 2D descriptors and hierarchical clustering methods are best at separating biologically active molecules from inactives, a prerequisite for a good compound selection method. In particular, the combination of MACCS descriptors and Ward's clustering was optimal.

INTRODUCTION

The advent of high-throughput biological screening methods have given pharmaceutical companies the ability to screen many thousands of compounds in a short time. However, there are many hundreds of thousands of compounds available both in-house and from commercial vendors. Whilst it may be feasible to screen many if all of the compounds available, this is undesirable for reasons of cost and time and may be unnecessary if it results in the production of some redundant information. Therefore there has been a great deal of interest in the use of compound clustering techniques to aid in the selection of a representative subset of all the compounds available.[1-8] A similar problem faces those interested in designing compounds for synthesis; a good design will capture all the required information in the minimum number of compounds.[9]

Underpinning the compound selection methods is the *similar property principle*,[10] which states that structurally similar molecules will exhibit similar physicochemical and biological properties. Given a clustering method that can group structurally similar compounds together, application of this principle implies that the selection, or synthesis, and testing of representatives from each cluster produced from a set of compounds should be sufficient to understand the structure- activity relationships of the whole set, without the need to test them all.

An appropriate clustering method will, ideally, cluster all similar compounds together whilst separating active and inactive compounds into different sets of clusters. The first factor will ensure that every class of active compound is represented in the selected subset but that there is no redundancy. The second factor will minimize the risk that an inactive compound is selected as the representative of a cluster containing one or more actives, thereby missing a class of active compounds.

Clustering is the process of dividing a set of entities into subsets in which the members of each subset are similar to each other but different from members of other subsets. There have been numerous cluster methods described; general discussions of many of these are given by Gordon,[11] by Everett,[12] and by Sneath and Sokal.[13] Several of these methods have been applied to clustering chemical structures; comprehensive reviews are given by Barnard and Downs[14] and by Downs and Willett.[15] In outline, the clustering process for chemical structures is as follows.

(1) Select a set of attributes on which to base the comparison of the structures. These may be structural features and/or physicochemical properties.

(2) Characterize every structure in the dataset in terms of the attributes selected in step one.

(3) Calculate a coefficient of similarity, dissimilarity, or distance between every pair of structures in the dataset, based on their attributes.

(4) Use a clustering method to group together similar structures, using the coefficients calculated in step three. Some clustering methods may require the calculation of similarity values between the new objects formed and the existing objects.

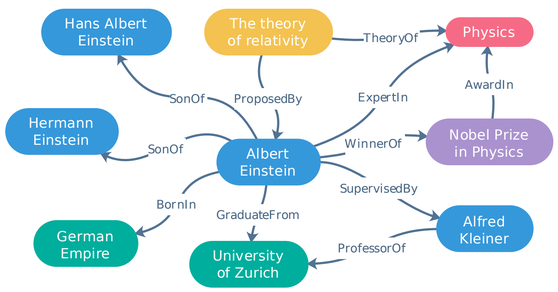
(5) Analyze the resultant clusters or classification hierarchy to determine which of the possible sets of clusters should be chosen.

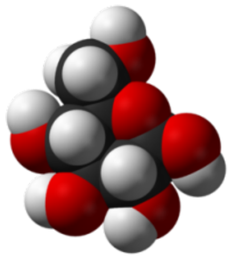
A number of methods are available both for the production of descriptors in steps (1) and (2) and clusters in step (4). Whilst there are also a large number of coefficients that might be used in step (3), the choice of clustering method may determine which is best suited.

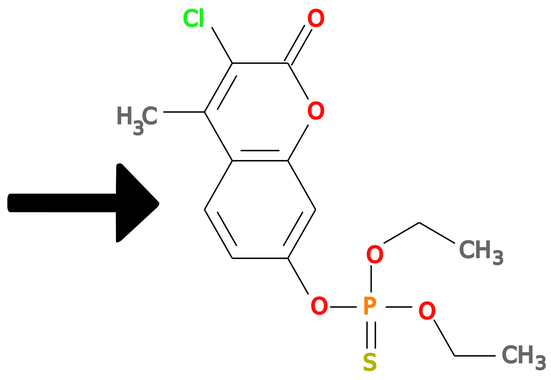
In this paper we present a study aimed at identifying the most suitable descriptors and clustering methods for use in compound selection. We have used a variety of methods to cluster sets of structures with known biological activities and evaluated the clusters produced according to their ability to separate active and inactive compounds into different sets of clusters. We have concerned ourselves with structure based clustering. For this, the *substructure search screens* used in commercial database searching software have often been used as descriptors. We have examined a number of these descriptors, together with two developed in-house, and have considered the use of four commercially available clustering methods.

[†] brownr@abbott.com.
[‡] yvonne.martin@abbott.com.
[X] Abstract published in *Advance ACS Abstracts*, January 15, 1996.

0095-2338/96/1636-0572$12.00/0







Current state \ Input	0	1
S_1	S_2	S_1
S_2	S_1	S_2