The Use of Cluster Analysis in Molecular Design

C. Frank Shen¹, Kimberly A. Mason², and Alan H. Katz³,
Wyeth-Ayerst Research
Princeton, NJ 08543

ABSTRACT

In drug design, it is widely accepted that two molecules have the same biological activity if they interact with the same receptor in the same way. To identify a potential drug, for example, a chemist would fit different conformations of a candidate to the target molecule manually using the Cartesian coordinates and could obtain further information by calculating non-bonded potential energies which resulted from probing the surface around each molecule with hydrogen bond donor, hydrogen bond acceptor, and lipid probes. A statistical procedure combining hierarchical cluster analysis with principal components is developed to identify clusters that best account for similarity among the target structures using the same data mentioned above. This procedure makes it possible to automate the fitting process by comparing orientation and energy of probes simultaneously. Analysis was conducted using statistical procedures in SAS® software. In most cases, the fits obtained by the cluster analysis are better than those obtained manually.

Key words: drug similarity; hierarchical cluster analysis.

1. INTRODUCTION

In drug design, it is widely accepted that two molecules have the same biological activity if they interact with the same receptor in the same way. In an effort to match two molecular structures which have the same interactions, we defined the drug similarity by seven parameters. They are non-bonded potential energies (E) which are the sums of electrostatic energy, hydrogen-bond energy and Lennard-Jones potential energy, three Cartesian coordinates (X, Y, Z) and three rotational angles (θx, θy, θz) around the axis. Ideally, the matching pair of molecules should have comparable orientations with similarly low energies. Recently we presented a new approach, GRASP (1), of comparing molecular structures with similar biological activity based on the results of GRID® (2) and SYBYL® (3) calculations in a UNIX environment on a Silicon Graphics® workstation. We use GRID® to handle the potential energy computations and SYBYL® to conduct the least squares fit for two molecules based on the result from GRID®. The 3-D maps for the fitted structures are also plotted. Nyburg’s BMFIT algorithm (4) was adopted by SYBYL® for the least squares fit. The GRASP method is carried out by probing the surface around each molecule with a hydrogen bond accepting carbonyl oxygen (CO), a hydrogen bond donating amide hydrogen (NH) and a lipophilic defining aromatic hydrogen (AR). The resultant Cartesian coordinates and associated energies are reported by GRID®. Three pairs of data points (one pair per probe) which have comparable positions and favorably low energies, i.e., strong attractive forces, are selected as the input to SYBYL® for fitting the molecules. However, it takes time for an experienced modeler to select an appropriate set of points in order to determine the similarity. It often turns the fitting into a trial-and-error and subjective process. The goal of this study is to develop a more efficient and unbiased method. A computer-based cluster analysis proposed in the next section will fulfill this purpose.

2. METHODOLOGY

Instead of having the modeler manually select a set of points, an algorithm DECLUS shown in Figure 1 uses all available points from both molecules of interest. It contains the methodology which can be carried out by SAS® procedures. Three key components for DECLUS are described as follows:

¹ Biometrics Department
² Chemistry Division
Data Transformation

Before starting the cluster analysis, the raw data extracted from GRID is prepared through a two-step transformation process. The first step is the standardization with respect to each molecule which results in zero mean and unit variance for every variable among the data points. The purpose of this standardization is to bring data to a common origin and remove the possible influence of the gross size and variation of the data from different objects which may introduce artificial importance to a group of objects during the cluster analysis. The second step is the principal component analysis. Although we start with four variables, the resultant eigen system suggests that the dimensionality can be reduced to three due to collinearity between coordinates and energy. Let X be the standardized data matrix with p columns corresponding to the number of variables and n rows corresponding to the number of data sets. The variance-covariance matrix, \( \Sigma \), is equivalent to the correlation coefficient matrix and is given by

\[
\Sigma = \frac{X'X}{(n-1)},
\]

where \( X' \) is the transpose of the matrix X. The first three principal components of the matrix \( \Sigma \) can be computed by solving the equation

\[
(\Sigma - \lambda I) V = 0.
\]

The solution of the above equation will result in a set of non zero eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_p \) with their corresponding eigenvectors \( v_1, v_2, \ldots, v_p \) which are mutually orthogonal to each other. A large proportion of the variation (say, 90%) caused by p variables over n data points can be accounted for by k dimensions \( (k \leq p) \) which gives the directions of a set of k orthogonal axes. The coordinates of these axes are linear combinations of the original p variables and summarize the major dimensions of the variation. By choosing k=3, the matrix of the transformed data, T, is given by

\[
T_{nx3} = X_{nxp} V_{px3}
\]

and is used as the input to the subsequent cluster analysis.

Cluster Analysis

The term "cluster analysis" normally represents a wide variety of procedures that can be used to create a classification. More specifically, a clustering method is a multivariate statistical procedure that places objects into groups of highly similar entities suggested by data. Several types of clusters are possible based on the chosen cluster method. The cluster procedures considered here are methods that create hierarchical clusters in which one cluster may be entirely contained within another cluster, but no other kind of overlap between clusters is allowed. Consequently for n data points, a hierarchical clustering method results in a tree configuration for the total number of clusters from one (includes all the data) to the number of entities n (each cluster has only one member). The objects that are members of a common taxon at a lower level can be members of different taxa at a higher level. The remaining task is to select a criterion which will cut the tree at a proper level. For more information about the hierarchical clustering method, see Sneath and Sokal (5).

The hierarchical clustering method we used for this study is the \( \beta \)-flexible method proposed by Lance and Williams (6). Let \( d_{ij} \) denote the inter-group similarity coefficient between i and j of which the Euclidean distance is a special example. While \( d_{ij} \) is the smallest measure remaining so that (i) and (j) fuse to form a new group (k) with \( n_k \) elements, consider fuse group (k) with a third group (h). Before the fusion, the values of \( d_{ih}, d_{jk}, d_{jk} \) and \( n_i \) are all known and \( d_{kh} \) can be expressed as a linear combination of these five values. Lance and Williams suggested a generalized combinatorial formula for the hierarchical clustering methods:

\[
d_{hk} = \alpha_i d_{ih} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{ih} - d_{hj}|,
\]

where the parameters \( \alpha_i, \alpha_j, \beta \) and \( \gamma \) determine the nature of the clustering. For example, the oldest of the clustering methods, nearest neighbour or single linkage, can be derived from the above equation by setting \( \alpha_i=\alpha_j=1/2, \beta=0 \) and \( \gamma=1/2 \). For the \( \beta \)-flexible method, the following constraints are set:

\[
\alpha_i + \alpha_j + \beta = 1; \alpha_i = \alpha_j = \beta < 1; \gamma = 0.
\]

It is combinatorial by definition. Its flexibility lies in its space-distorting properties. As \( \beta \) approaches unity the system becomes increasingly space contracting and chains completely. As \( \beta \) falls to zero and then becomes negative, the system becomes space dilating, and the elements becomes more intensely grouped. By a suitable choice of \( \beta \) the system can be made space conserving without the problem of chaining. We use \( \beta = 0.25 \) as suggested by Lance and Williams.

The algorithm for cutting the tree is based on the \( R^2 \) criterion (7). The value of \( R^2 \) is equal to unity when there are n clusters containing exactly one data point. As the data are emerged to form tighter clusters, \( R^2 \)
decreases from unity. In the limit, when all the data entities are emerged together to form one big cluster, $R^2$ is equal to zero and the information about grouping is lost. For our purpose, we choose $R^2 = 0.9$, i.e., the number of clusters that accounted for 90% of the variation.

**Centroid Computation**

After the number of clusters is determined, we use PROC TREE to redistribute data points into clusters. The cluster which has the lowest mean standardized energy was considered as the most favorable cluster and the weighted centroid of the cluster was calculated as

$$\left( \frac{\sum_1^n w_i x_i}{\sum_1^n w_i}, \frac{\sum_1^n w_i y_i}{\sum_1^n w_i}, \frac{\sum_1^n w_i z_i}{\sum_1^n w_i} \right)$$

using $w_i = 1/n$. One centroid per probe/molecule was computed and used as the input to SYBYL®.

### 3. CASE STUDIES

**Case I. Comparisons Among Molecules with Known Structures.**

Manaut, et al. (8) tested electrostatic similarity between some neuro-transmitters such as benzylamine (BZA), phenylethamine (PHEEA) and phenylpropylamine (PHEPA) using the Spearman rank coefficient as the measure for the similarity. We test the same molecules using our approach. Three molecules were processed by GRID® using three types of probes mentioned earlier. A cartesian coordinate and energy level was calculated for each favorable interaction between the target molecule and the probe. For each pair of molecules of interest, data were input to the cluster analysis which returned the centroids for the most favorable clusters. These centroids were then used by the FIT function within SYBYL® to perform a least squares fit for two molecules. The results of fit are shown in Figure 2 for BZA versus PHEEA, and Figure 3 for BZA versus PHEPA. Despite the fact that PHEEA is structurally closer to BZA than PHEPA, the resultant fit indicate that BZA is actually more like PHEPA than PHEEA. This observation agrees with general knowledge about chemical reactivity of which the positions of the terminal amine are the same for BZA and PHEPA but not for BZA and PHEEA.

Although our result reaffirm Manaut's conclusion, the method we proposed is quite different from theirs. In their method, an exhaustively iterative procedure was used to search through six variables, three coordinates plus three rotational angles, and find the best Spearman rank coefficient between electrostatic energies from both molecules. In contrast, our methods searches similarity through principal components of three coordinates plus non-bonded energies simultaneously by using the cluster analysis. The resultant centroids are then provided as an initial position for the SYBYL® to continue searching through three rotational angles. Our method is more efficient in terms of searching for similarities by using the cluster analysis. The resultant fit is also more insightful since the non-bonded energies that we used are more informative than the use of electrostatic energy alone.

**Case II. Conformational Analysis for An Angiotensin II (A-II) Antagonist.**

**Background.** Angiotensin II (A-II) is an octapeptide natural ligand and the final product of the renin cascade which causes an increase in blood pressure. A new therapeutic area has emerged recently involving the design of A-II antagonists as alternatives to ACE (angiotensin converting enzyme) inhibitors. Due to the fact that the octapeptide is too small to be constrained by a well formed secondary structure and too large to be rigid and predictable, the modeler has problems validating the structure of the active conformation of A-II by the currently existing conformational searching methods. Consequently, design of antagonists based on any computational model of A-II would be inconclusive. Additionally, the structures of the receptors for the A-II ligand have not yet been determined by crystallographic methods with the result that the intermolecular requirements of the receptor-inhibitor complex are still unknown. Therefore any search process for an A-II antagonist must be confined to the information to be gained from known A-II antagonists, based on an assumption that all antagonists of a specific receptor subtype must be interacting with the receptor in the same way. Furthermore, an active antagonist can be of the form as a combination of conformers (or simply one of its conformers) due to the flexibility of the chemical structure. Study on these individual conformations helps a modeler understand some intrinsic properties of a molecule with respect to the receptor's energy requirement.

**Method.** The case illustrated here involves using the proposed cluster analysis for characterizing a known A-II antagonist, DUP-3174. A conformational search was performed on DUP-3174 using Macromodel and the MM2 force-field in a water solvent continuum. The ten lowest energy conformations from the search were identified and processed by GRID® using three types of probes.
mentioned before. A cartesian coordinate and energy level was calculated for each favorable interaction between the target molecule and the probe. In order to determine the similarity among different conformers, a conformer was randomly chosen as a reference (e.g., conformer 2). Data from both test and reference conformers were used as input to the cluster analysis which returned the centroids for the most favorable clusters. These centroids were then used by the FIT function within SYBYL™ to perform a least squares fit for two conformers.

**Result.** Figure 4 through 6 illustrates that the ten lowest energy conformations of DUP-3174 form two families based on two distinct orientations. A common characteristic of both families is the formation of a lipophilic pocket defined by the large number of aromatic hydrogen and carbonyl oxygen points found in the area between the middle phenyl ring of the biphenyl acid and the imidazole of the headpiece, indicating a possible binding site. Conformers 1, 4 and 6 represent a family of conformations in which the carboxylic acid functional group attached to the biphenyl is turned inward toward the hydrocarbon chain of the headpiece, resulting in a poor fit with conformer 2. In each of the other conformations which fit well with conformer 2, this acid is turned away from the hydrocarbon. A possible cause for the difference in orientation is that some cluster points for the amide probe in the first family are overlapped with the lipophilic pocket while in the second family, the amide cluster points are ~7 Å away from the pocket.

Note that same procedure can be applied to a test molecule and the outcome will be compared with the known A-II antagonist. This comparison is still in progress and will be presented in the future.

**4. CONCLUSION**

A multivariate statistical procedure containing a principal component analysis and a cluster analysis was proposed for assessing the drug similarity. It is more efficient and robust than several existing methods. Above all, this procedure makes the automation of GRASP possible, which may play an important role for further analysis using methods such as ComFA (Comparative Molecular Field Analysis) which requires a good overlap of structures as input (9).

**REFERENCES**

2. GRID: Molecular Discovery Ltd., Oxford, UK.
3. SYBYL: Tripos Associates, St. Louis MO.
Figure 1. DECLUS Algorithm

Read Data;

INPUT: Variables X, Y, Z and ENERGY of atoms were generated by program GRID®. Read 4 variables for both molecules.

Standardization;

Use PROC STANDARD to standardize input variables to (0,1) variables w.r.t. molecule; Merge standardized data from both molecules without differentiation; Output data set "std_data";

Principal Component (PC);

Use PROC PRINCOMP to transform std_data; Output PC1, PC2, PC3 to data set "pc_std";

Cluster Analysis;

PROC CLUSTER groups pc_std using β-flexible method; Output cluster data to "tree";

** Determine the number of clusters should be formed using $R^2$ criterion; PROC TREE on tree to draw a tree diagram and group the data as requested; Output clustered data to "separate"; PROC PRINT and PROC MEANS to separate;

Calculate Centroids;

** Identify the cluster with the lowest standardized energy as the target cluster; Restore molecule identity to the target cluster; PROC MEANS on target cluster by molecule;

Output Centroids;

OUTPUT: Feed centroids to SYBYL® for a least squares fit.

NOTE: (a) ** - Decision point.
(b) Words in bold capitals indicate a statistical procedure from SAS® software was used.
(c) Analysis was done for one type of probes (e.g. NH) at a time.
Figure 3. Case I: The Overlap Between BZA and PHEPA

Figure 4. Case II: The Overlay Between Conformer 1, 3, 4 of DUP-3174 and The Reference Conformer
Figure 5. Case II: The Overlap Between Conformer 5, 6, 7 of DUP-3174 and The Reference Conformer

Figure 6. Case II: The Overlap Between Conformer 8, 9, 10 of DUP-3174 and The Reference Conformer