ELSEVIER

Journal of Molecular Structure: THEOCHEM 727 (2005) 81–87

THEO CHEM

www.elsevier.com/locate/theochem

Assessment of the performance of cluster analysis grouping using pharmacophores as molecular descriptors

Alex Rodriguez[a], M. Santos Tomas[b], Juan J. Perez[c,1], Jaime Rubio-Martinez[a,*]

[a]*Dept. de Química Física, Universitat de Barcelona (UB), Martí i Franqués 1, E-08028 Barcelona, Spain*
[b]*Sec. Matemàtiques ETSAB, Universitat Politècnica de Catalunya (UPC), ETS d'Arquitectura, Av. Diagonal 649, E-08028 Barcelona, Spain*
[c]*Department d'Enginyeria Química, Universitat Politècnica de Catalunya (UPC), ETS d'Enginyeria Industrial, Av. Diagonal 647, E-08028 Barcelona, Spain*

Received 27 October 2004; accepted 10 December 2004
Available online 28 June 2005

Abstract

The present work describes the results of a comparative study designed to assess the performance of using four-point pharmacophores as molecular descriptors coupled with cluster analysis as grouping technique for molecular diversity analysis. For this purpose 31 globulin binding steroids were considered as test set. Each of the molecules was investigated for the number of pharmacophores capable to fulfill, considering or not molecular flexibility, respectively. The cosine coefficient was used as similarity measure and cluster analysis as grouping technique. Specifically, two hierarchical clustering methods were used: on the one hand, the group average method and on the other, the Ward's method. The results obtained were compared with the activity of the molecules, as well as with the performance of previous diversity studies published in the literature using the same data set.

Keywords: Cluster analysis; Docking; Pharmacophores; Diversity analysis

1. Introduction

The goal of molecular diversity methods is to group molecules according to similarity criteria, providing a tool to select a subset of molecules based on their neighborhood in a chemical space [1–4]. The information provided by these studies can be used to select either the smallest subset of molecules that cover all the features of the set or alternatively, a subset of neighboring molecules that can be used to characterize fine differences among them to allow selection of the most fitted one for a specific purpose. These procedures have proven to be very useful in the process of discovery molecules with selected properties, being their success based on the structure-activity paradigm according to which, molecules exhibiting similar geometrical features have similar physicochemical properties as well as biological activities [5]. Specifically, these methods have been shown to be very useful to obtain new leads from chemical libraries, with no need to know the 3D structure of the target receptor. To undertake this type of studies, it is necessary to use molecular descriptors to define a chemical space where molecules are represented by points, together with procedures to assess their distribution throughout the entire space, by computing a similarity index between each pair of molecules to finally, group them according to their proximity in the space.

The selection of molecular descriptors is a very critical process, since they should be capable of discriminating between similar and dissimilar molecules. If descriptors have little discrimination power this leads to a poor performance of the grouping process. Descriptors definition is an area of intensive research and can be chosen with different profiles. They range from 1D descriptors that correspond to bulk properties of molecules, to the more sophisticated 2D and 3D descriptors, all of them being used successfully in different applications. Two-dimensional descriptors basically include information regarding the molecular graph or topology of a chemical structure, whereas 3D descriptors are related to the three-dimensional distribution of the molecular structure, including interaction

* Corresponding author. Tel.: +34 93 4021222; fax: +34 93 4021231.
E-mail addresses: juan.jesus.perez@upc.es (J.J. Perez), jaime.rubio@ub.edu (J. Rubio-Martinez).
[1] Tel.: +34-93-4016111; fax: +34-93-4016210.

0166-1280/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.theochem.2005.02.030

maps, electron density or molecular pharmacophores. In the present work, pharmacophores are used as descriptors due to their proven performance and because it is easy to study the effect of including the conformational flexibility of molecules on their grouping performance.

A pharmacophore can be considered a sketch of the requirements of a molecule to grant its recognition to a specific molecular target or in other words, the minimum set of molecular features that represents the interaction between a ligand and its receptor [6]. Accordingly, a pharmacophore is a set of chemical groups and distances or angles between them. Although there are no restrictions in the number of points selected to define a pharmacophore, previous experience suggest that three or four points suffice to provide a reasonable description of the binding site. Different applications using pharmacophores as molecular descriptors have been reported in the literature. These studies are basically concerned with the use of three- or four-point representations [7,8] and even more, some authors have used pharmacophores together with cluster analysis as grouping technique in diversity analysis [9,10]. However, no systematic study has addressed the performance of this combined use. Accordingly, the goal of the present study is to assess the performance of the combined used of four-point pharmacophores as molecular descriptors with two different hierarchical cluster analysis methods: the group average method (GA) [11] and the Ward's clustering method (WA) [12]. Additionally, the effect on the diversity analysis of including or not molecular flexibility is also analyzed.

The study has been performed using a set of 31 globulin binding steroids whose structure and identification numbers are shown in Fig. 1. This data base has been previously used in other classification procedures [13–17] and a good overview of previous results, including comparison with other QSAR studies, can be found in Ref. [17].

2. Methods

2.1. Molecular indicators

All possible four-point pharmacophores were constructed and codified, following the procedure described by Mason et al. [7,8]. This was carried out using the six distances between the different pharmacophoric points, plus an additional parameter that controls their chirality. This was carried out by discretizing the distance between two pharmacophoric groups as follows: first, only pharmacophoric distances between 1.5 and 16.5 Å were considered and discretized using a step of 1.5 Å. This gives rise to a total of 10 binary positions. In the study, six different pharmacophoric features were used to define pharmacophores, including hydrogen bond acceptors (HA), hydrogen bond donors (HD), positive charges (P+), negative charges (N−), hydrophobic moieties (HI) and aromatic rings (AR), plus due to the nature of the database analyzed, a special pharmacophoric point (DB) (carbon–carbon double bond) to account for different electronic density in this type of bonds with respect hydrophobic carbon–carbon bonds. Also, to better describe extension of adjacent hydrophobic regions, usually described by only one HI pharmacophoric feature, we include all $–CH_2–$ moieties as HI points. Following this procedure, a 2,590,532 vector long was assigned to each of the molecules, defined in such a way that each of the components represents one of the different four-point pharmacophores possible. For each molecule a '1' is assigned to a component when the corresponding pharmacophore can be attained by the molecule and a '0', if not.

2.2. Conformational analysis

A library of conformations was generated for each of the molecules studied as follows. The routable bonds of each molecule were systematically rotated in angle increments according to the bond type: for sp^3–sp^3 bonds 120°; for sp^2–sp^3 60° and for sp^2–sp^2 180°. Rotations were performed using the quaternions formalism. This binary representation allows us to use fast and efficient algorithms for clustering.

2.3. Similarity coefficients and distance

The difference to one of the cosine coefficients was used as similarity index between molecules. The cosine coefficient is essentially equivalent to the Carbó index and in contrast to the Tanimoto coefficient, it does not consider a common absence of attributes accountable for the similarity of two molecules. Contrary to what it had been shown for the Tanimoto coefficient when using binary descriptors, this index does not exhibit the properties of a distance [18]. This fact is relevant when pharmacophores are used as descriptors, since we are interested in selecting molecules that fulfill specific pharmacophores aimed at comparing similar affinity profiles. Accordingly, not fulfilling a common subset of pharmacophores it does not seem relevant in assessing the diverse affinity behavior of the molecules.

2.4. Cluster analysis

In the present work we have selected two different hierarchical clustering methods for our analysis: on the first hand, the *group average method* (GA) [11] and on the other, Ward's hierarchical-agglomerative method (WA) [12]. The former has been demonstrated to be a simple and efficient algorithm. It takes as a distance between two clusters the mean of all possible distances between their corresponding elements. This method was also used by Bultinck et al. in a previous work [13]. The latter maximizes the inter-cluster variance whilst minimizing the intra-cluster variance, and has showed a good performance in comparative studies [19,20].