STUDIA UNIVERSITATIS BABEŞ-BOLYAI, CHEMIA, L, 2, 2005

3D MOLECULAR SIMILARITY; METHOD AND ALGORITHMS

OLEG URSU*, MIRCEA V. DIUDEA*

* *Faculty of Chemistry and Chemical Engineering*
Babes-Bolyai University, 400028 Cluj, Romania

ABSTRACT. This study presents a method and algorithms for calculation of 3D similarity between pairs of chemical structures represented as 3D molecular graphs. Similarity searching in chemical databases is widely used for virtual screening, lead discovery and optimization, and most recently protein amino-acid sequences studies to discover and determine the functionality of a new isolated protein. This method has obvious advantages over other known methods due to the following: (*i*) the superposition method does not depend on the preliminary alignments of the chemical structures; (*ii*) entire conformational space is searched without generation of each conformer; (*iii*) excellent discrimination between geometrical isomers. Although it is a computationally demanding method, recent implementation of maximum clique algorithm and bound smoothing algorithm made possible the optimization of this method and application to similarity searching in chemical databases of non trivial size.

INTRODUCTION

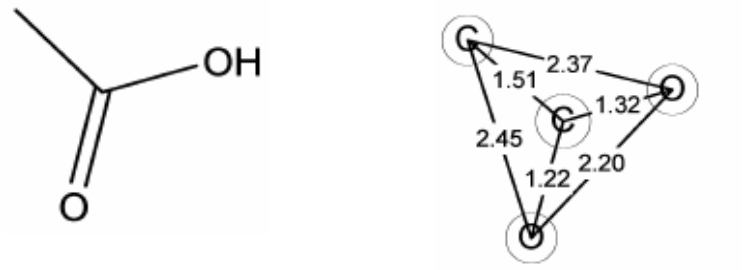
The investigation of molecular structure involves research on its constitution – the number and chemical identity of atoms and bonds joining them along with the configuration in 3D space. Molecular similarity has been studied from two different major view points: (i) topological similarity defined in connectivity and constitutional terms and (ii) geometrical similarity, when geometrical aspects of the molecular structure are taken into account.

Similarity searching in databases of 2D chemical structures is widely used for virtual screening and lead discovery. A similarity measure, that quantifies the degree of structural resemblance between the target structure and each of the database structure, is based on fingerprint or molecular descriptor encoding of the molecular structure with similarity between pairs of such representations being computed using the *Tanimoto* coefficient. Another topological similarity measure of increasing interest (although more computationally demanding) is detection of 2D maximum common subgraphs (MCS).[1,2] The binding affinity of the ligand to the receptor site, which usually express the biological activity, is related to a single geometrical configuration of the ligand.

The procedures and algorithms used in detection of the MCS can be extended to 3D similarity searching, with several modifications, the most important being conformational flexibility in the matching algorithm. This study presents a complete implementation of such an algorithm. The effectiveness of the proposed method in classifying chemical structures with respect to a given bioactive leader is evaluated.

SIMILARITY METHODOLOGY

Chemical graphs. All molecular structures can be represented as simple, undirected graphs. In a 3D chemical graph, the vertices denote atoms but edges here can indicate the geometric distance or a range of distances between pair of atoms (vertices). The main difference between these two representations is that 3D chemical graph is usually weighted by geometrical distance (see figure 1).



(a) 2D chemical graph (b) 3D chemical graph

Figure 1. Acetic acid chemical graphs representations

Distance geometry. In ligand-receptor interaction mechanism, ligand usually exhibits some degree of flexibility and thus the distances between atoms are not fixed.

One approach to cope with this drawback is to generate several conformations of low energy for each structure under consideration and then to compare all possible pairs of the resulting conformations. This approach is, however, computationally demanding, if exhaustive sampling of the molecules' conformational space is to be achieved and still cannot guarantee that the optimal similarity has been identified.

Distance geometry, herein considered, encodes the molecules' conformational flexibility within a single graphical representation.[3] Specifically, each edge of a 3D molecular graph is represented by a range of distances spanning the maximum and minimum allowable distance between two atoms. Distance ranges are imposed by some constrains, e.g., distance and chirality. The distance constrains are simply the lower and upper bounds of the interatomic distances; the chirality includes the handedness of the asymmetric centers in the molecule.

Covalent distance constrains. The local covalent structure of a molecule is easily defined by distance constrains. Unless one is dealing with a highly strained ring system, it is sufficiently to use the exact distance constrains in which the lower and upper bounds are equal. For example, the distances among covalently bonded pairs of atoms, are determined with high precision by the bond order and the types of atoms connected. Similarly, the bond angles can usually be determined from the covalent structure, while for fixed bond lengths there is a one-to-one relation between the bond angle and the geminal distance, so that these distances can also be determined. The relation between the geminal distances and the bond angle θ is given explicitly by the law of cosines:

$$\begin{aligned} d_{13}^2 &= d_{12}^2 + d_{23}^2 - 2d_{12}d_{23}\cos(\theta) \\ d_{13}^2 &= l_{13}^2 + (u_{13}^2 + l_{13}^2)\sin^2\left(\frac{\theta}{2}\right) \end{aligned} \qquad (1)$$

where, $l_{13} = |d_{12} - d_{23}|$ and $u_{13} = d_{12} + d_{23}$ are called the *lower* and *upper* triangle inequality limits, respectively.

Vicinal distance constrains. Similarly, when the incident and geminal distances are held fixed, there is a one-to-one relation between the *absolute value* of the torsion angle φ and the vicinal distance, given by:

$$d_{14}^2 = l_{14}^2 + \left(u_{14}^2 - l_{14}^2\right)\sin^2\left(\frac{\varphi}{2}\right) \tag{2}$$

where l_{14} and u_{14} are *cis* and *trans* limits on the *1*,*4* distance.

Chirality constrains. The chirality χ_{1234} of an ordered quadruple of points numbered *1,2,3,4* is given in terms of their Cartesian coordinates by the sign of the following determinant:

$$\chi_{1234} = \mathrm{sgn}\left(\det\begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{bmatrix}\right) \tag{3}$$

Torsion angle constrains. As shown above, the absolute value of a torsion angle can be constrained to any range of values by means of suitable 1,4 distance constraints, including its *cis* and *trans* limits. Moreover, since the chirality χ_{1234} of a chain of four bonded atoms A_1-A_2-A_3-A_4 is equal to the sign of the torsion angle $\mathrm{sgn}(\varphi) = 0,\pm 1$ about the 2,3 bond, by a suitable combination of distance and chirality constraints we can obtain any range of values with a given sign. This is sufficient to specify the rotameric state (*gauche*$^+$, *gauche*$^-$ or *anti*) about the single bonds.

Steric distance constrains. Since two atoms cannot be in nearly the same place at the same time, in order to obtain reasonable conformations it is necessary to impose lower bound constraints on the distances between all pairs of atoms, separated by more than three bonds. For the sake of simplicity, these lower bounds are generally set to the sum of suitable *hard sphere radii* (van der Waals radii):

$$l_{ij} = r_i + r_j \tag{4}$$

After applying the preceding distance and chirality constrains, we ensure that the structures which satisfy them are not grossly unreasonably on energetic grounds. In order to get the correct conformation, it is necessary to impose constrains on interatomic distances for atoms that are separated by four or more bonds. Such constrains are determined by *bound smoothing* procedures: the *triangle bound smoothing* and *tetrangle bound smoothing*.

Triangle bound smoothing. Triangle inequality bound smoothing is based upon the well-known triangle inequality among the distances:

$$d_{ij} \le d_{ik} + d_{jk} \tag{5}$$

for all triples of atoms *i*, *j*, *k*. It follows that if $d_{ik} \le u_{ik}$ and $d_{jk} \le u_{jk}$ then:

$$d_{ij} \le d_{ik} + d_{jk} \le u_{ik} + u_{jk} \tag{6}$$

So, if $u_{ij} > u_{ik} + u_{jk}$, then $u_{ij} > d_{ij}$ and hence u_{ij} can be replaced by the upper limit $u_{ik}+u_{jk}$ on d_{ij} without eliminating any conformation that satisfy the constrains $d_{ik} \le u_{ik}$ and $d_{jk} \le u_{jk}$ (see figure 2).

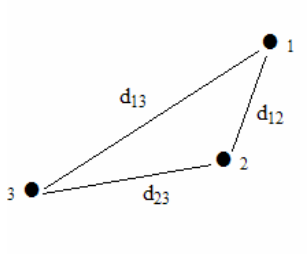


Figure 2. Triangle inequality

Tetrangle inequality bound smoothing. Unfortunately, the triangle inequality limits represent a rather poor approximation to the actual Euclidean limits, so that the triangle inequality bound smoothing is not a very effective approach to locating errors in the bounds. A somewhat more effective (although much more time-consuming) approach looks at four atoms at a time, rather than three. In this case, the algebraic form of the relations among the distances is far more complicated, so that the *tetrangle inequality limits* are best described pictorially as in Figure 3.

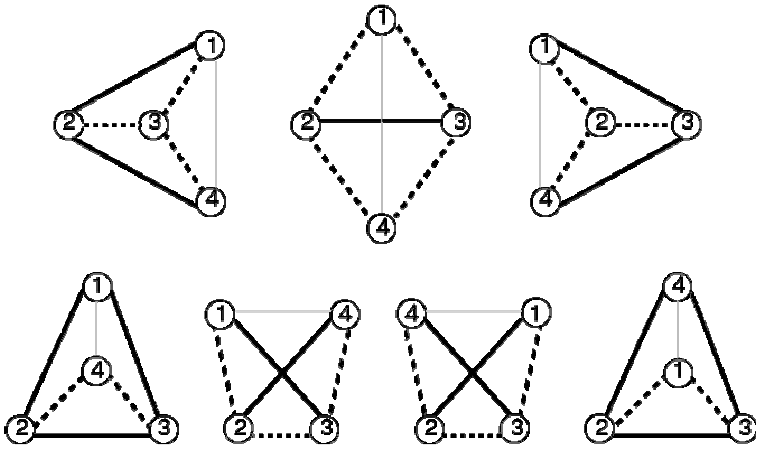


Figure 3. Tetrangle inequality limits

The mathematical form of this inequality can be expressed in terms of *Cayley-Menger* determinants:[4]

$$0 \le CM(d_{12},\ldots,d_{34}) = \det\begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & d_{14}^2 \\ 1 & d_{12}^2 & 0 & d_{23}^2 & d_{24}^2 \\ 1 & d_{13}^2 & d_{23}^2 & 0 & d_{34}^2 \\ 1 & d_{14}^2 & d_{24}^2 & d_{34}^2 & 0 \end{pmatrix} \tag{7}$$

Thus the corresponding mathematical forms of the above tetrangle inequalities are:

$$\begin{aligned} & 0 \le CM(l_{12},u_{13},u_{14},u_{23},u_{24},u_{34}), \\ \textit{or}\quad & 0 \le CM(u_{12},l_{13},l_{14},u_{23},u_{24},u_{34}), \\ \textit{or}\quad & 0 \le CM(u_{12},u_{13},u_{14},l_{23},l_{24},u_{34}), \end{aligned} \tag{8}$$

together with

$$\begin{aligned} & 0 \le CM(u_{12},u_{13},l_{14},l_{23},u_{24},l_{34}), \\ \textit{or}\quad & 0 \le CM(u_{12},l_{13},u_{14},u_{23},l_{24},l_{34}), \\ \textit{or}\quad & 0 \le CM(l_{12},l_{13},u_{14},l_{23},u_{24},l_{34}), \\ \textit{or}\quad & 0 \le CM(l_{12},u_{13},l_{14},u_{23},l_{24},l_{34}), \end{aligned} \tag{9}$$

Figure 4 illustrates a pair of structures and their corresponding 3D MCS; the initial coordinates are generated by *Hyperchem* molecular modeling package.

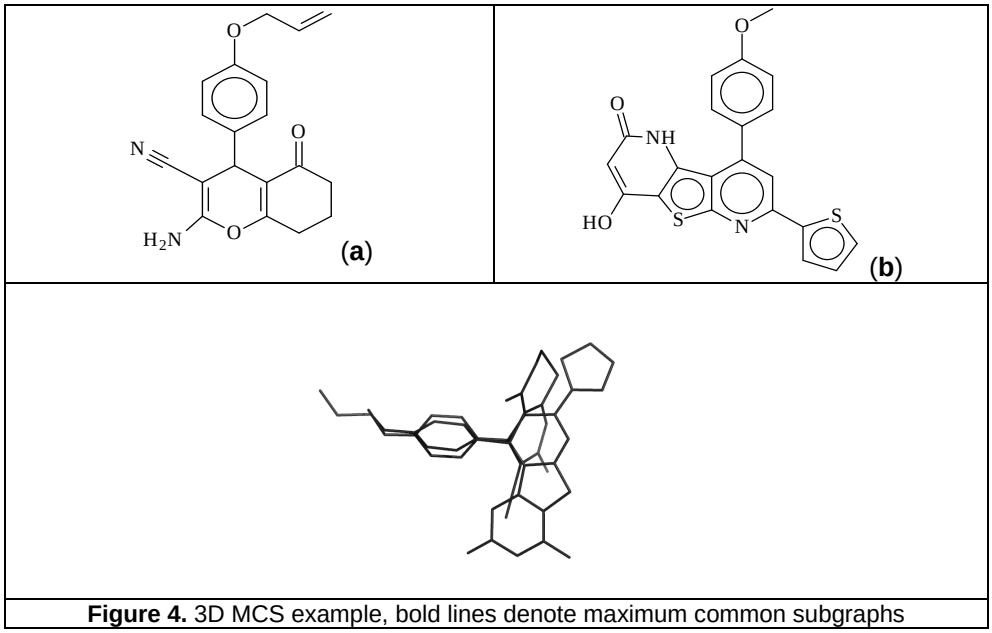
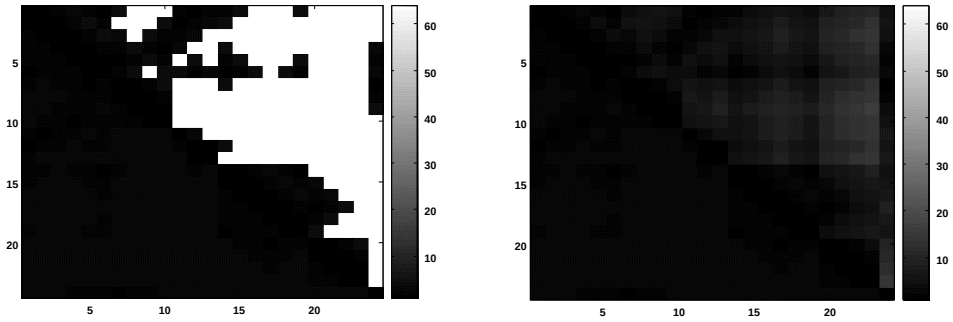


Figure 4. 3D MCS example, bold lines denote maximum common subgraphs

Upper and lower distance matrix for structure (**a**) Figure 4, before bound smoothing and after bound smoothing procedure is illustrated in Figure 5.



(**a**) Before bound smoothing (**b**) After bound smoothing

Figure 5. Distance bound smoothing example

A significant improvement in upper and lower bound is observed after applying distance geometry bound smoothing procedure (see Figure 5).

VALIDATION STUDIES

Validation studies were carried out on a set of 19 dopamine receptor antagonists. Dopamine receptors in the brain are important in modulating motor, endocrine, and emotional functions.[5,6] The antagonist affinity was measured at recombinant receptors selectively expressed in cloned cells (see Figure 6).

The test set used here was published by Brusniak;[7] it contains experimental data $\log(1/K_d)$, where K_d is the dissociation constant of the receptor-antagonist complex (see Table 1). For this simulation, the most active compound ((R) SKF82526) was used as a query; 18 pairwise comparisons were performed, with distance tolerance value ε = 0.1Å. Initial coordinates are obtained by *Hyperchem* program, followed by bound smoothing procedure. The 3D similarity threshold was set to 0.2 to prevent unnecessary pairwise comparisons. On a PC with 2.8 GHz Intel processor, 512 RAM, Windows XP, the computations for overall pairs comparisons took less then 2 s. The results are summarized in Table 1. It is noticeable the fact that receptor can discriminate between stereo-isomers (R)-SKF-82526 and (S)-SKF82526, although the difference is only one carbon atom configuration; the procedure is discriminative, giving appropriate similarity index.

The similarity index values were calculated using a weighing atom scheme. Thus the atoms that ensemble the active scaffold necessary for a structure to be active have a higher rank than the irrelevant atoms. All the active structures contain this scaffold, so that, after the overlapping procedure, it is easy to identify these atoms.

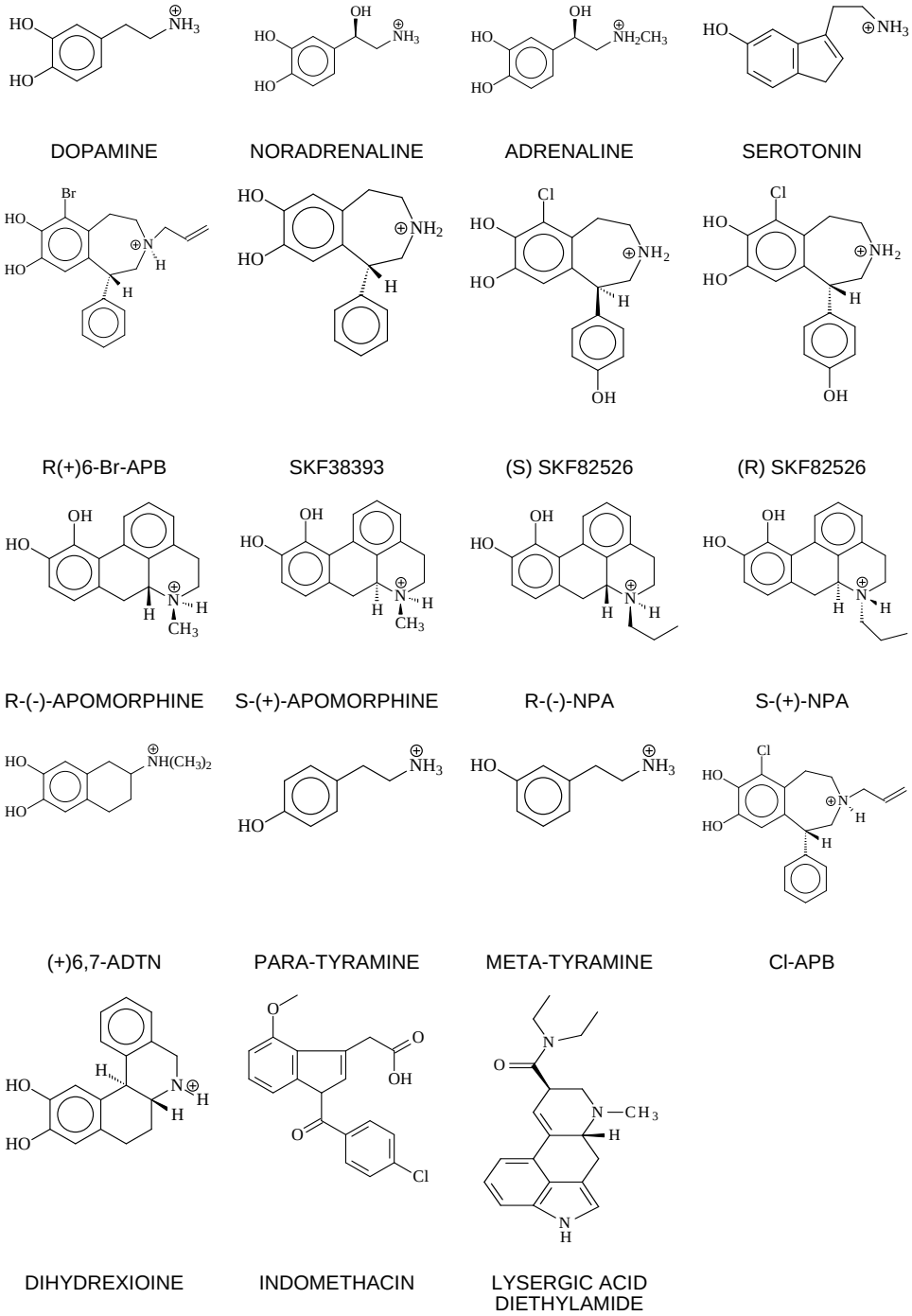


Figure 6. Structures of dopamine receptor antagonists

Table 1.

Drug affinities and similarity calculation results

molecule	$\log(1/K_d)$	Similarity
(+)-6,7-ADTN	-3.66	0.7500
adrenaline	-4.74	0.4444
Cl-APB	-1.92	0.9907
DHX	-3.08	0.8102
dopamine	-3.39	0.6759
m-tyramine	-4.68	0.5833
noradrenaline	-4.69	0.6759
p-tyramine	-5.59	0.5880
(R)-apomorfine	-2.83	0.7685
(R)-(-)-NPA	-3.26	0.7546
(R)-(+)-6-Br-APB	-2.58	0.9861
(S)-(+)-apomorfine	-3.08	0.7639
S-(+)-NPA	-3.72	0.7639
(S)-SKF82526	-3.26	0.7778
serotonin	-3.99	0.6806
SKF38393	-2.18	0.9861
(R)-SKF-82526*	-1.45	1.0000
INDOMETHACIN**	-3.53	0.7479
LSD**	-4.02	0.6818

*query structure
**estimate values for BA

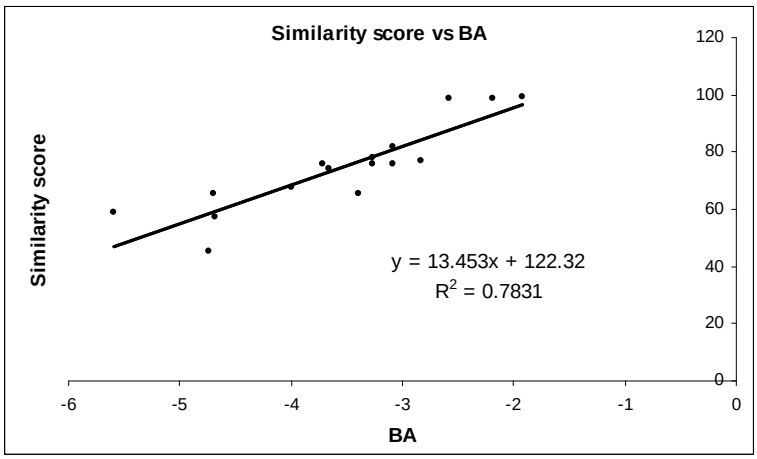


Figure 7. Linear dependence between Similarity scores and BA

Linear regression analysis showed good correlations between similarity score and BA. Thus an attempt to give estimative values of BA for two compounds of interest (INDOMETHACIN and LSD) was made. The predicted values (by using regression eq. in Figure 7) showed mild activity of these known antagonists.

Considering the simplicity of the used model, we can draw the conclusion that the similarity scores can classify correctly the unknowns, which is the most desirable feature (see Figure 7).

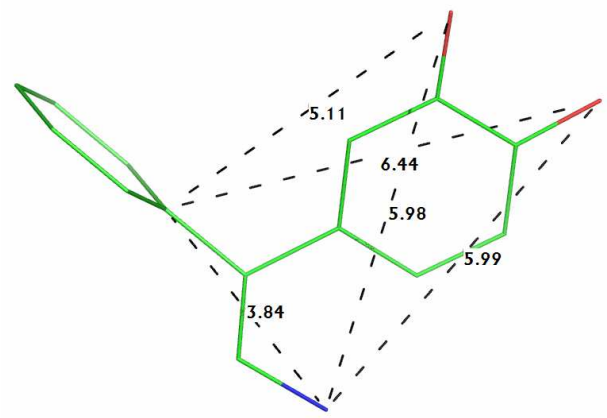


Figure 8. Pharmacophore map for dopamine receptor antagonists

The obtained results reveal the important structural features, *i.e*., the *pharmacophore*, of antagonists of the dopamine receptor (see Figure 8). In the high affinity compounds, the distance between the cationic nitrogen and the m-hydroxyl oxygen ranged from 6 to 6.45 Å with highest activity compound (R)-SKF-82526 having a distance of 6 Å. The distance between the cationic nitrogen and the first carbon in the second benzene cycle ranged from 3.78 to 3.81 Å; in the lowest activity compounds, this pharmacophore is missing.

CONCLUSIONS

In this paper we described an advanced method for the calculation of intermolecular structural similarity, useful in mining databases of 3D structures. This method takes a full account of the conformational flexibility, being in the mean time sufficiently rapid to allow search in databases of nontrivial size. Validation studies demonstrated that the method is more accurate than fingerprint screening, allowing discrimination even between stereo-chemical isomers. It would be also possible to use fingerprints screening procedure prior to graph matching in view of improving the overall efficiency. The method provides an effective extension to current approaches in virtual screening and lead optimization procedures.

REFERENCES

1. Raymond, J.; Gardiner, E.; Willett, P.; RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs, Comput. J., 2002, **45**, 631-644.
2. Raymond, J.; Gardiner, E.; Willett, P.; Heuristics for Rapid Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm, J. Chem. Inf. Comput. Sci. 2002, **42**, 305-316.

3. Crippen, G.; Havel, T.; Distance Geometry and Molecular Conformation, Research Studies Press: **1988**.
4. Easthope, P.; Havel, T. F.; Computational Experience with an Algorithm for Tetrangle Inequality Bound Smoothing, Bull. Math. Biol., 1989, **51**, 173-194.
5. Strange, P. G.; Brain biochemistry and brain disorders; Oxford University Press: New York, **1993**.
6. Waddington, J.; D1:D2 Dopamine receptor interactions; Academic Press: New York, **1993**.