

## 3D Molecular Similarity: Method and Algorithms

Oleg URSU<sup>a\*</sup>, Mircea V. DIUDEA<sup>a</sup> and Shin-ichi NAKAYAMA<sup>b</sup>

<sup>a</sup>Faculty of Chemistry and Chemical Engineering, 11, Babes-Bolyai University  
Arany Janos Str., 400028 Cluj-Napoca, ROMANIA

<sup>b</sup>University of Tsukuba, Research Center for Knowledge Communities  
1-2 Kasuga, Tsukuba, Ibaraki, 305-8550, JAPAN

\*e-mail: [uo510@chem.ubbcluj.ro](mailto:uo510@chem.ubbcluj.ro)

(Received: May 31, 2005; Accepted for publication: October 17, 2005; Published on Web: March 2, 2006)

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.

**Keywords:** Molecular similarity, Maximum common subgraph, Distance geometry, Similarity coefficient

### 1 Introduction

The investigation of a 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) proposed by Raymond *et. al* [1, 2]. The binding affinity of the ligand to the receptor site, which usually expresses 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 one 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.

### 2 Methods

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

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

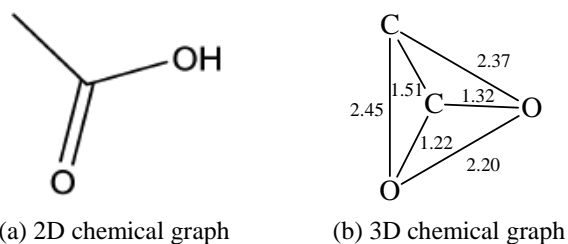


Figure 1. Acetic acid chemical graph representations

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 by Crippen [3]. Specifically, each edge of a 3D molecular graph is represented by a range of distances spanning the maximum and minimum allowable distances 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 sufficient 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  $\theta$  is given explicitly by the law of cosines:

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

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  $\varphi$  and the vicinal distance, given by:

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

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

**Chirality constrains.** The chirality  $\chi_{1234}$  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:

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

**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 constrains, including its *cis* and *trans* limits. Moreover, since the chirality  $\chi_{1234}$  of a chain of four bonded atoms  $A_1-A_2-A_3-A_4$  is equal to the sign of the torsion angle  $\text{sgn}(\varphi) = 0, \pm 1$  about the 2,3 bond, by a suitable combination of distance and chirality constrains we can obtain any range of values with a given sign. This is sufficient to specify the rotameric state (*gauche*<sup>+</sup>, *gauche*<sup>-</sup> 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 constrains 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):

$$l_{ij} = r_i + r_j \quad (4)$$

After applying the preceding distance and chirality constrains, we ensure that the structures which satisfy them are not grossly unreasonable 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 *tetrahedron bound smoothing*.

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

$$d_{ij} \leq d_{ik} + d_{jk} \quad (5)$$

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

$$d_{ij} \leq d_{ik} + d_{jk} \leq u_{ik} + u_{jk} \quad (6)$$

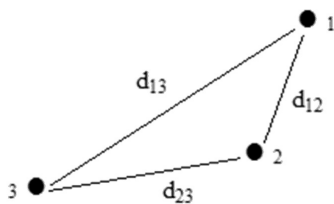


Figure 2. Triangle inequality

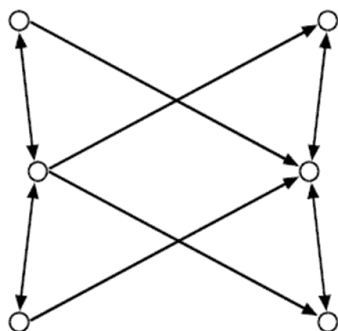


Figure 3. The digraph whose shortest path determines the triangle inequality limits

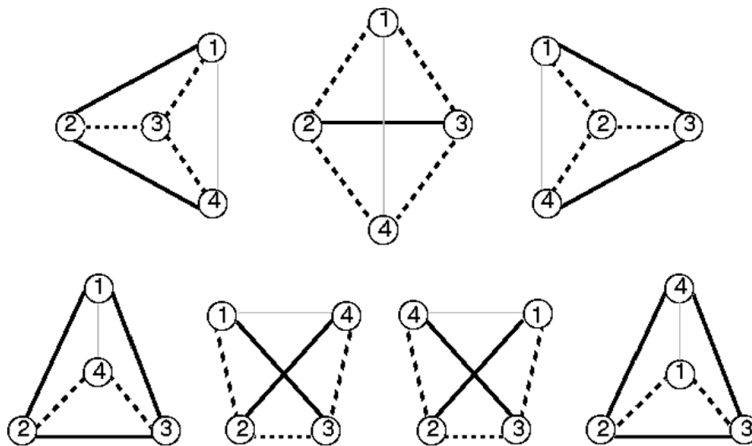


Figure 4. Tetrangle inequality limits

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 conformations that satisfy the constraints  $d_{ik} \leq u_{ik}$  and  $d_{jk} \leq u_{jk}$  (see Figure 2).

In order to compute the triangle inequality limits efficiently, we can reduce their calculations to the all pairs, shortest path problem in a certain digraph. A path in such a digraph is a sequence of nodes such that any consecutive nodes in the sequence are connected by an arc in digraph (Figure 3), whose arrow points from the first to the second. It is easily seen that the upper triangle limits are equal to the lengths of the shortest path in an undirected digraph, whose arc lengths are equal to the given upper bounds. It can be shown that all the lower triangle limits are of the form:

$$\bar{l}_{ij} = l_{km} - \bar{u}_{ik} - \bar{u}_{jm} \quad (7)$$

where  $i, j, k, m$  are not necessarily distinct atom indices,  $l_{ij}$  and  $u_{ij}$  which respectively denote lower and upper bounds on the corresponding indexed atoms, and the overbars indicate the corresponding triangle inequality limits (eq. 5). This shows that the upper limits can be computed independently of the lower and also that the greatest lower limit cannot exceed the greatest lower bound.

Current implementation of the method uses Floyd's [7] shortest path algorithm. This algorithm takes each node  $k$  of the digraph in turn, and then makes a pass through all ordered pairs of other nodes  $(i, j)$ . If the length of the path  $i \rightarrow k \rightarrow j$  is shorter than the length of the direct path  $i \rightarrow j$ , the latter is set to the former. This ensures that after each pass all the path lengths are at least as short as any path that goes through node  $k$ , and hence iterating on this procedure for  $k = \overline{1, N}$  produces the desired matrix of shortest paths.

**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 4.

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

$$0 \leq CM(d_{12}, \dots, d_{34})$$

$$= \det \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & d_{14}^2 \\ 1 & d_{12}^2 & 0 & d_{23}^2 & d_{24}^2 \\ 1 & d_{13}^2 & d_{23}^2 & 0 & d_{34}^2 \\ 1 & d_{14}^2 & d_{24}^2 & d_{34}^2 & 0 \end{pmatrix} \quad (8)$$

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

$$0 \leq CM(l_{12}, u_{13}, u_{14}, u_{23}, u_{24}, u_{34}),$$

$$\text{or } 0 \leq CM(u_{12}, l_{13}, l_{14}, u_{23}, u_{24}, u_{34}), \quad (9)$$

$$\text{or } 0 \leq CM(u_{12}, u_{13}, u_{14}, l_{23}, l_{24}, u_{34}),$$

together with

$$0 \leq CM(u_{12}, u_{13}, l_{14}, l_{23}, u_{24}, l_{34}),$$

$$\text{or } 0 \leq CM(u_{12}, l_{13}, u_{14}, u_{23}, l_{24}, l_{34}),$$

$$\text{or } 0 \leq CM(l_{12}, l_{13}, u_{14}, l_{23}, u_{24}, l_{34}), \quad (10)$$

$$\text{or } 0 \leq CM(l_{12}, u_{13}, l_{14}, u_{23}, l_{24}, l_{34}),$$

The following pseudocode implements the procedures for triangle and tetrangle bound smoothing as described above.

```

procedure Floyd(Natom, Lower, Upper)
1: for k from 1 to Natom do
2:   for i from 1 to Natom-1 do
3:     for j from i+1 to Natom do
4:       /* Path lengths in left-hand network */
5:       if Upper[i, j] > Upper[i, k] + Upper[k, j] then
6:         Upper[i, j] := Upper[i, k] + Upper[k, j];
7:       end if
8:       /* Path lengths in right-hand network */
9:       if Lower[i, j] < Lower[i, k] - Upper[k, j] then
10:        Lower[i, j] := Lower[i, k] - Upper[k, j];
11:      else
12:        if Lower[i, j] < Lower[j, k] - Upper[k, i] then
13:          Lower[i, j] := Lower[j, k] - Upper[k, i];
14:        end if
15:        /* Check for triangle inequality violations */
16:        if Lower[i, j] > Upper[i, j] then
17:          exit ("Bad bounds ");
18:        end if
19:      end for
20:    end for
21:  end for
22: end procedure

```

```

procedure Easthope(Natom, Lower, Upper)
1: for i from 1 to Natom do
2:   for j from i + 1 to Natom do
3:     for k from 1 to Natom do
4:       for m from k + 1 to Natom do
5:         /* See if k and m can be made collinear with
6:         i or j */
7:         if not Collinear(i,k,m,Lower,Upper)
8:           and not Collinear(j,k,m,Lower,Upper) then
9:           /* Tighten k, m upper limit by tetrangle
10:          limit */
11:          test := UpTetLim(i,j,k,m,Lower,Upper);
12:          if test < Upper[k, m] then
13:            Upper[k, m] := test;
14:          end if
15:          /* Tighten k, m lower limit by tetrangle

```

```

16:          limit */
17:          test := LoTetLim(i,j,k,m,Lower,Upper);
18:          if test > Lower[k, m] then
19:            Lower[k, m] := test;
20:          end if
21:        end if
22:        /* Check for tetrangle inequality
23:        violations*/
24:        if Lower[k, m] > Upper[k, m] then
25:          exit("Bad bounds ");
26:        end if
27:      end for
28:    end for
29:  end for
30: end for
31: /* Test for convergence */
32: if any changes were made to the limits then
  return TRUE;
else
  return FALSE;
end if
end procedure

```

Figure 5 illustrates a pair of structures and their corresponding 3D MCS; the initial coordinates are generated by *Hyperchem* molecular modeling package.

Upper and lower distance matrices for structure (a) Figure 5, before bound smoothing and after bound smoothing procedures are illustrated in Figure 6.

A significant improvement in upper and lower bound is observed after applying distance geometry bound smoothing procedure (Figure 6).

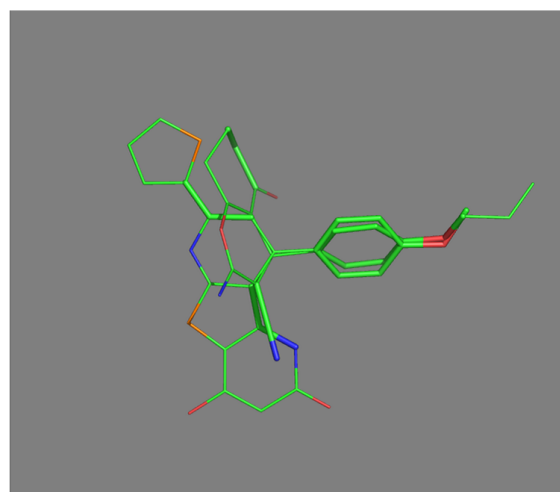
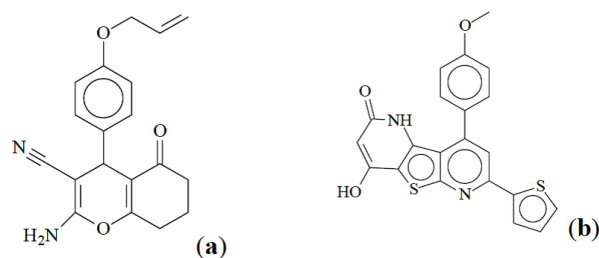
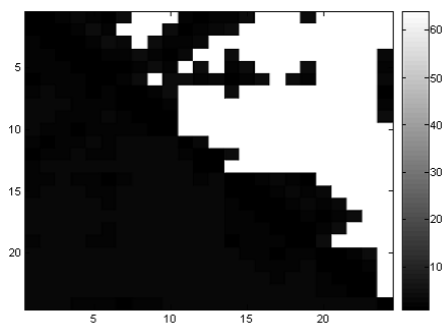
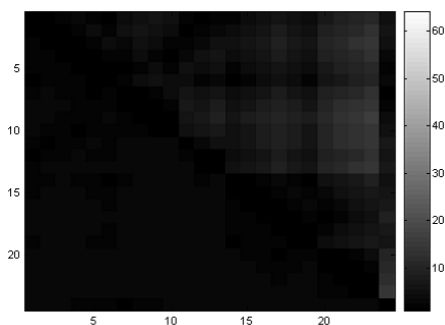


Figure 5. 3D MCS example, bold lines denote maximum common subgraphs



(a) Before bound smoothing



(b) After bound smoothing

Figure 6. Upper and lower bound matrices, the upper (lower) diagonal values indicate upper (lower) bound distances, respectively

Once the set of structures to be queried has been pre-processed and stored using the upper and lower bound values calculated using the previously described distance bounding techniques, similarity searching is performed by using an MCS graph matching algorithm. The algorithm is a clique based method which computes the MCS by determining the maximum clique in the correspondence graph, this process is described in detail by Raymond [1, 2]. In the current approach the maximum clique algorithm as described in Östergard [8] is used. The following pseudocode implements the maximum clique detection algorithm.

```

/* Maximum clique detection algorithm */
function clique(U, size)
1:   if |U| = 0 then
2:     if size > max then
3:       max := size
4:       /* Found new clique save it */
5:       found := true
6:     end if
7:   end if
8:   while U ≠ 0 do
9:     if size + |U| ≤ max
10:      return

```

```

11:     end if
12:     i := min{j|vj ∈ U}
13:     if size + c[i] ≤ max then
14:       return
15:     end if
16:     U := U \ {vi}
17:     clique(U ∩ N(vi), size + 1)
18:     if found = true then
19:       return
20:     end if
21:   end while
22: return

```

```

24: function new()
25:   max := 0
26:   for i := n downto 1 do
27:     found := false
28:     clique(Si ∩ N(vi), 1)
29:     c[i] := max
30:   end for
return

```

### 3 Experimental Part, Results and Discussion

All algorithms were implemented in C# programming language and require an ECMA compliant implementation of .NET framework. The program can run on any platform which supports such an implementation. Binaries are available electronically from authors free of charge. Input file format for program is Hyperchem HIN.

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 by Strange [5] and Waddington [6]. The antagonist affinity was measured at recombinant receptors selectively expressed in cloned cells (Figure 7).

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  $\epsilon = 0.1\text{\AA}$ . Initial coordinates are obtained by *Hyperchem* program, followed by a 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 than 2 s. The results are summarized in Table 1. It is a noticeable fact that the 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.

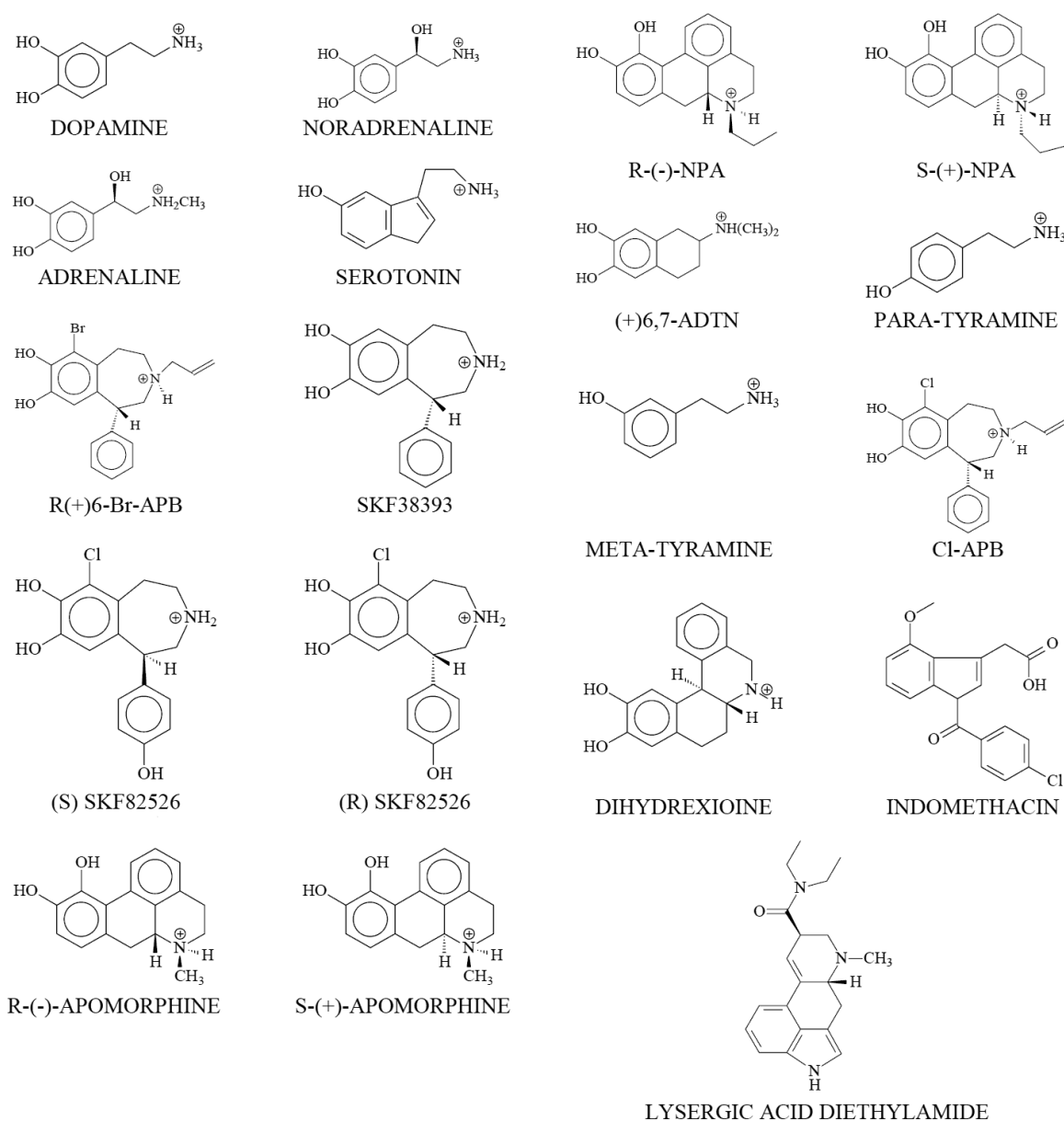


Figure 7. Structures of dopamine receptor antagonists

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.

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 8) 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 8).

The obtained results reveal the important structural features, *i.e.*, the *pharmacophore*, of antagonists of the dopamine receptor (see Figure 9). 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.

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 <sup>a</sup>	-1.45	1.0000
INDOMETHACIN <sup>b</sup>	-3.53	0.7479
LSD <sup>**</sup>	-4.02	0.6818

<sup>a</sup>query structure

<sup>b</sup>estimate values for BA

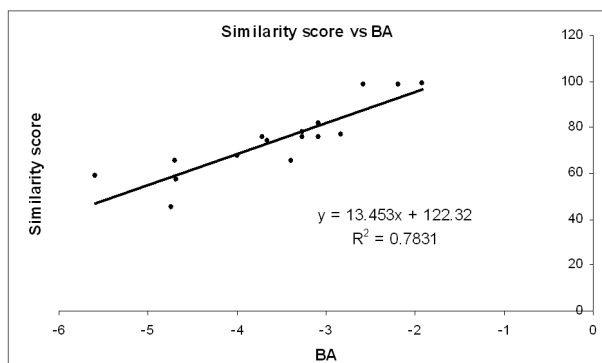


Figure 8. Linear dependence between Similarity scores and BA

## 4 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 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 stereochemical isomers. It would be also possible to use the fingerprint screening procedure prior to graph matching

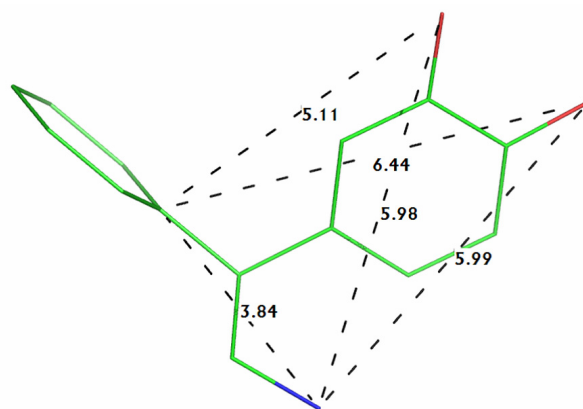


Figure 9. Pharmacophore map for dopamine receptor antagonists

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] J. Raymond, E. Gardiner, P. Willett, *Comput. J.*, **45**, 631-644 (2002).
- [2] J. Raymond, E. Gardiner, P. Willett, *J. Chem. Inf. Comput. Sci.*, **42**, 305-316 (2002).
- [3] G. Crippen, T. Havel, *Distance Geometry and Molecular Conformation*, Research Studies Press (1988).
- [4] P. Easthope, T. F. Havel, *Bull. Math. Biol.*, **51**, 173-194 (1989).
- [5] P. G. Strange, *Brain biochemistry and brain disorders*, Oxford University Press, New York (1993).
- [6] J. Waddington, *D1:D2 Dopamine receptor interactions*, Academic Press, New York (1993).
- [7] R. Floyd, *Communications of the ACM*, **7**, 345 (1962).
- [8] R. J. Östergard, *Discrete Applied Math.*, **120**, 197-207 (2002).

