

FULL PAPER

## 3DFS: 3D Flexible Searching System for Lead Discovery – New Version 1.2

Ting Wang and Jiaju Zhou

Laboratory of Computer Chemistry (LCC), Institute of Chemical Metallurgy, Chinese Academy of Sciences, P.O. Box 353, Beijing 100080, China

Received: 18 May 1999/ Accepted: 1 September 1999/ Published: 19 November 1999

**Abstract** 3DFS is a 3D flexible searching system for lead discovery. Version 1.0 of 3DFS was published recently (Wang, T.; Zhou, J. *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 71-77). Here version 1.2 represents a substantial improvement over version 1.0. There are six major changes in version 1.2 compared to version 1.0.

1. A new rule of aromatic ring recognition.
2. The inclusion of multiple-type atoms and chains in queries.
3. The inclusion of more spatial constraints, especially the directions of lone pairs.
4. The improvement of the query file format.
5. The addition of genetic search for flexible search.
6. An output option for generating MOLfiles of hits.

Besides the above, this paper supplies:

1. More query examples.
2. A comparison between genetic search and Powell optimization.
3. More detailed comparison between 3DFS and Chem-X.
4. A preliminary application of 3DFS to K<sup>+</sup> channel opener studies.

**Keywords** Pharmacophore, 3D search, Conformational search, Hydrophobe recognition

### Introduction

The last decade has witnessed a continuing rise in the development and application of 3D database searching techniques [1-4] to drug design and discovery. A number of 3D searching systems have been developed, such as the in-house

programs: 3DSEARCH [5] and ALADDIN [6] and commercial software: MACCS-II/3D [7], ISIS/3D [8], Chem-X(ChemDBS-3D) [9], SYBYL(3DB Unity) [10] and Catalyst [11]. Recently, the 3D searching technique has been successfully applied to the discovery of novel protein kinase C agonists [12], HIV-1 protease inhibitors [13-14], HIV-1 integrase inhibitors [15-17], angiotensin II antagonists [18] and farnesyl protein transferase inhibitors [19].

We have also developed a 3D searching system-3DFS [20] in our laboratory and have been attempting to improve its utility. The new version 1.2 of 3DFS will be described in detail in the following sections.

*Correspondence to:* T. Wang, European Molecular Biology Laboratory, Meyerhofstr.1, 69117 Heidelberg, Germany. Tel: +49(0)6221 387466; Fax: +49(0)6221 387517; E-mail: Ting.Wang@EMBL-Heidelberg.de

## Query definition

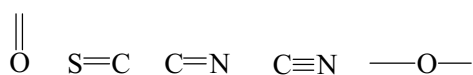
In most 3D searching systems, the elements of a query pharmacophore are defined as actual atoms or functional groups such as a nitrogen atoms or a carbonyl groups. A basic nitrogen atom usually acts as a positive charge center and a carbonyl oxygen usually plays the role of hydrogen bond acceptor in the ligand-receptor interaction. If a positive charge center or hydrogen bond acceptor is used as the query element instead of a nitrogen atom or carbonyl group, 3D searching will retrieve more functionally equivalent but structurally diverse hits because a positive charge center or hydrogen bond acceptor can represent much more groups that serve the same binding function as a nitrogen or carbonyl group. So, it is necessary to introduce the generalized binding function definition into queries.

3DFS considers four important ligand-acceptor interactions: hydrogen bonding, charge interactions, hydrophobic interactions and  $\pi$ - $\pi$  interactions and thus defines six binding sites as the functional elements of a query. They are hydrogen bond acceptors, hydrogen bond donors, positive charge centers, negative charge centers, hydrophobic regions and aromatic ring centers.

Greene [21] has proposed detailed definitions for hydrogen bond acceptors/donors, charge centers, and hydrophobes. On basis of Greene's studies, we propose simpler and more practical definitions for the above functional elements, especially for aromatic rings and hydrophobes.

### Hydrogen bond acceptors and donors

Generally, any nitrogen, oxygen, fluorine or sulfur atom with at least one nondelocalized lone pair can be an acceptor, but a too generalized definition will result in an overload of hits, which may reduce the selectivity of hits. Therefore, 3DFS only considers some nitrogen or oxygen atoms which often appear in bioactive molecules as acceptors. As shown in Scheme 1, they include:



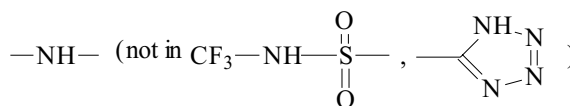
**Scheme 1** Definition of hydrogen bond acceptors

- oxygen with a double bond
- nitrogen attached to carbon with a double or triple bond
- sulfur attached to carbon with a double bond
- ether oxygen

Hydrogen bond donor atoms are primarily oxygen or nitrogen atoms with one or two electropositive hydrogen atoms. As shown in Scheme 2, they include:

- hydroxy groups not attached to a C=O, S=O or P=O group

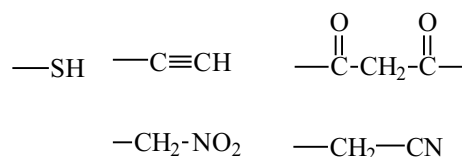
—OH (not attached to C=O, S=O, P=O), —NH<sub>2</sub>,



**Scheme 2** Definition of hydrogen bond donors

- amino groups that are not part of a trifluoromethyl-sulfonamide or tetrazole moiety

In addition, weaker groups shown in Scheme 3 can also be taken into account if necessary.



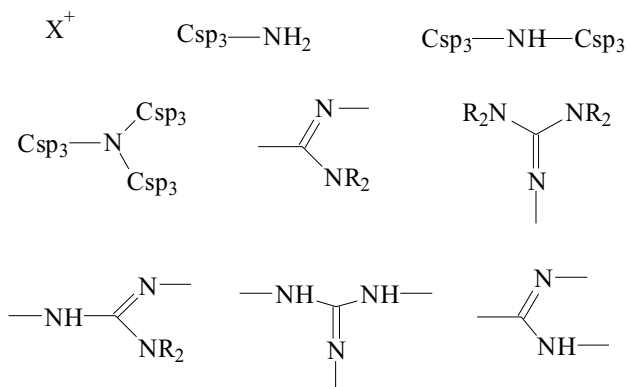
**Scheme 3** Weaker hydrogen bond donors

### Charge center

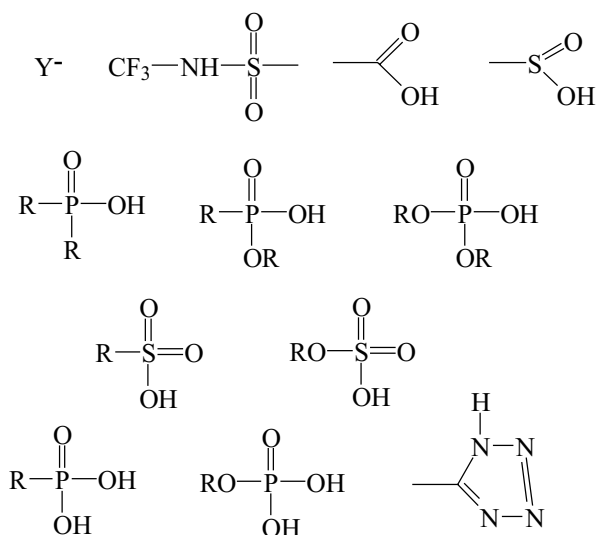
An atom with a formal charge is a charge center, but a neutral moiety might also be considered as a charge center if it would be ionized at physiological pH. For example, an aliphatic amine can be protonated to be a positive charge center whereas a carboxylic acid can be deprotonated to be a negative charge center at physiological pH.  $\pi$ -Delocalized systems such as carboxylate, guanidine and amidine can also make charge centers.

The positive charge centers (Scheme 4) include:

- atoms with a formal positive charge
- nitrogen in aliphatic amines



**Scheme 4** Definition of positive charge centers

**Scheme 5** Definition of negative charge centers

c. imino nitrogen of N,N-disubstituted amidines or N,N,N,N-tetrasubstituted guanidines

d. centroid of nitrogens in guanidines bearing at least one hydrogen on each amino nitrogen or in amidines bearing at least one hydrogen on the amino nitrogen.

The negative charge centers (Scheme 5) include:

- atoms with a formal negative charge
- nitrogen in trifluoromethylsulfonamide
- centroid of the oxo and hydroxyl oxygens in carboxylic, sulfinic, or phosphinic acid
- centroid of the oxo and hydroxyl oxygens in phosphoric diesters or phosphonic ester
- centroid of the two oxo oxygens and the hydroxyl oxygen in sulfuric or sulfonic acid
- centroid of the oxo oxygen and the two hydroxyl oxygens in phosphoric monoester or phosphonic acid
- amino nitrogen in a non-N-substituted tetrazole

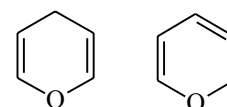
#### Aromatic ring centers

Aromatic ring centers are primarily the centroids of 5- or 6-member aromatic rings such as thiophene and benzene rings. These aromatic rings participate in  $\pi$ - $\pi$  interactions with the  $\pi$ -systems of a protein receptor. 3DFS system can automatically identify the aromatic rings in a database structure by its 2D substructure matching algorithm using the rule that an aromatic ring obeys Hückel's rule of  $4n+2$   $\pi$ -electrons and all carbon atoms in the ring have unsaturated bonds. This recognition rule of aromatic rings can eliminate non-aromatic rings such as a pyran that satisfy Hückel's rule (see Scheme 6).

The sum of  $\pi$ -electrons in a ring is calculated according to the atom type and hybridization type. Table 1 lists the number of  $\pi$ -electrons of C, N, O and S atoms in different hybridization states.

**Table 1**  $\pi$ -electrons of C, N, O and S atoms

atom	$\pi$ - electrons		
	$sp^3$	$sp^2$	$sp$
C	0	1	1
N	2	1	1
O	2		
S	2	2	

**Scheme 6** Structures of  $\alpha$ -pyran(left) and  $\gamma$ -pyran(right)

#### Hydrophobic regions

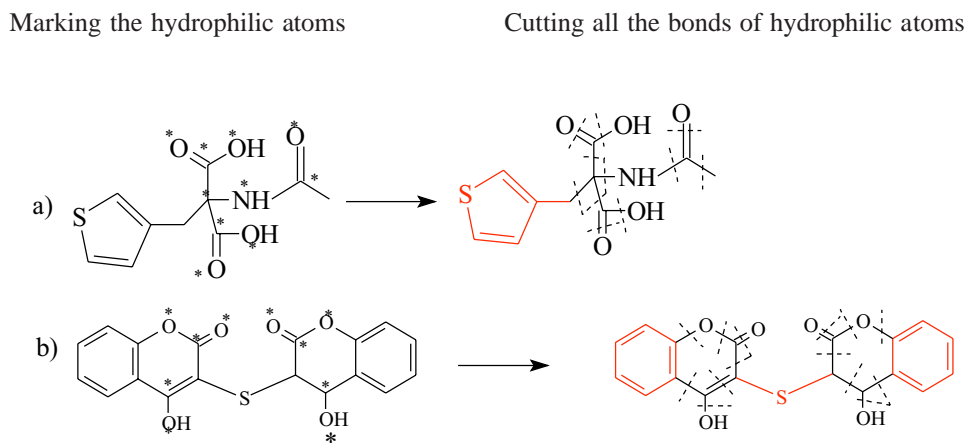
3DFS uses a special algorithm to identify hydrophobic regions in a database structure, since it is almost impossible to list all hydrophobic fragments just as we do for hydrogen bond acceptors/donors and charge centers. We employed the assumption that the hydrophobicity of a hydrophobic region is the sum of atomic contributions [22] and classified hydrophilic atoms into ten types (see Table 2). Our hydrophobe recognition algorithm is as follows:

Firstly, mark all the hydrophilic atoms according to Table 2, and then cut off all the bonds connected to these atoms. If we regard the database structure as a graph, then, at this time the graph is cut into many subgraphs disconnected to each other. Traversing each subgraph to numerate its non-hydrogen atoms, if the number of atoms is not less than three, which ensures that a hydrophobic region has enough surface area, the subgraph is considered as a hydrophobic region. The position of a hydrophobe is represented by the geometric center of the atoms in the hydrophobe.

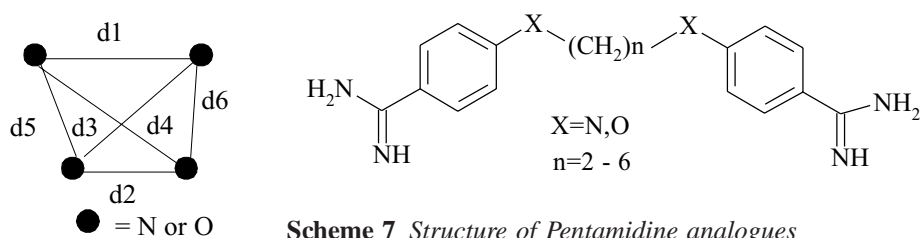
**Table 2** Hydrophilic atom types

Type	Description
1	N,O
2	S in SH
3	S with double bond
4	Charged atom ( $X^+$ ) or its neighbor atom ( $Y-X^+$ )
5	Atom bonding to OH, NH or $NH_2$ ( $Y-OH$ , $Y-NH$ , $Y-NH_2$ )
6	Atom bonding to SH ( $Y-SH$ )
7	Atom bonding to O with double bond ( $Y=O$ )
8	Atom bonding to S with valence $> 2$ (e.g. $Y-SO_2$ )
9	Atom with two or more neighboring O or N
10	two bonds away from two or more O with double bond (e.g. $-CO-Y-CO-$ , $-SO-Y-SO-$ )

**Figure 1** Two Examples of Recognizing Hydrophobic Regions. The atoms marked with "\*" are hydrophilic atoms and the red regions are hydrophobes.



**Figure 2** HIV-1 integrase inhibitor pharmacophore,  $d1=11.5 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d2=8.6 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d3=9.5 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d4=10.5 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d5=2.6 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d6=2.6 \text{ \AA} \pm 0.7 \text{ \AA}$



**Scheme 7** Structure of Pentamidine analogues

This method is simple and rapid. Figure 1 shows two examples of identifying the hydrophobic regions. The atoms with "\*" are hydrophilic atoms according to Table 2. Structure a has one hydrophobic region (red region) and Structure b has three hydrophobic regions (red regions).

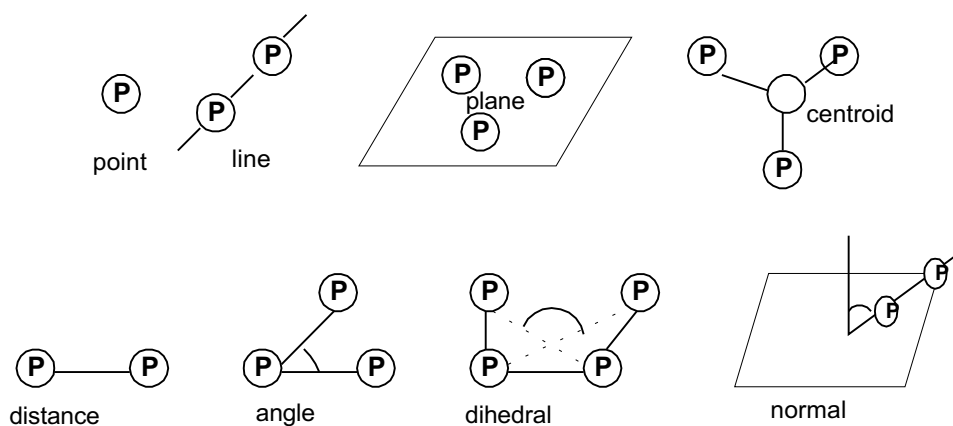
#### Multiple-type atoms and chains

Some queries contain heteroatoms whose atom types are not determined, for example, in HIV-1 integrase inhibitor pharmacophore [17] shown in Figure 2, the four points may be N

or O atoms. In order to be applicable to such queries, 3DFS defines three multiple-type elements Da, Db and Dc, representing N or O atoms, N, O, or S atoms, and O or S atoms, respectively.

In addition, the length of chain (mainly carbon chains) is sometimes important for the bioactivity of a compound, for example, the pentamidine analogues [23] exhibit bioactivity against *Leishmania mexiacna amazonensis* only when the length of the carbon chain is between two and six atoms (see Scheme 7). Considering this situation, 3DFS also defines chain chains as a element with a symbol of Cn in a query.

**Figure 3** 3D objects and constraints allowed in 3DFS



## Spatial constraints in queries

The 3D objects in queries allowed by 3DFS can be a point, a line, a plane, normal, and (or) a centroid. The geometric constraints of these objects can be distances, angles or dihedrals with allowed tolerances (see Figure 3).

Specially, the directionality of lone pairs of Hbond acceptor atoms is an important spatial constraint in some queries. So the directions of lone pairs of the following Hbond acceptor atoms are defined in 3DFS (see Figure 4a-4d, LP represents lone pair):

$sp^2$  Oxygen or Sulfur (see Figure 4a): The lone pairs are in the C-C=O/S plane at  $60^\circ$  angle to C=O line.

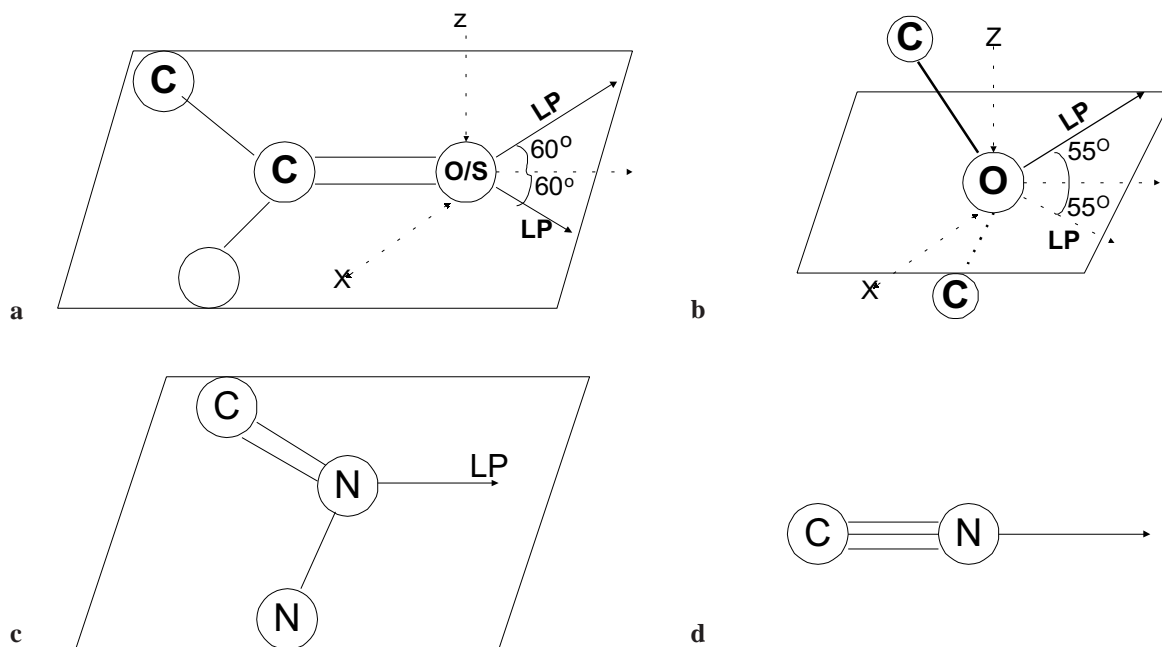
$sp^3$  Oxygen (see Figure 4b): The lone pairs are in the plane which bisects the covalent bonds of the C-O-C fragment at  $55^\circ$  angle to the bisect line.

$sp^2$  Nitrogen (see Figure 4c): The lone pair is in the C=N-C plane and bisects the covalent bonds of the C=N-C fragment.

$sp$  Nitrogen (see Figure 4d): The lone pair is along the C≡N line

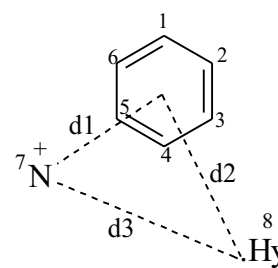
## Query file-BIP format

A pharmacophore usually consists of several disconnected fragments and their spatial relationships, and a fragment may contain a single atom or some atoms connected with each other by bonds. 3DFS allows at most six disconnected fragments in a query pharmacophore.



**Figure 4** Directions of lone pairs of Hbond acceptor atoms defined in 3DFS. a)  $sp^2$  O or S atom, b)  $sp^3$  O atom, c)  $sp^2$  N atom, d)  $sp$  N atom

**Scheme 8** Pharmacophore of  $d$ -selective opioid ligands,  $d1=4.5 \text{ \AA} \pm 1.0 \text{ \AA}$ ,  $d2=6.7 \text{ \AA} \pm 1.0 \text{ \AA}$ ,  $d3=7.6 \text{ \AA} \pm 1.0 \text{ \AA}$



The query file in 3DFS is defined as BIP (BIOphore) format (.bip) and 12 record items are currently available (see Table 3):

A BIP file ends by the ">END" data line. The information provided in the BIP file is read in free format. The details of the BIP format are given in the Appendix. For example, the query pharmacophore of  $\delta$ -selective opioid ligands [24] (see Scheme 8) has such a query file as shown in Figure 5.

## Search strategy

After the query file is input to the system, the search starts. The search consists of four steps: (1) 1D screening, (2) 2D substructure search, (3) rigid 3D search, (4) conformationally flexible search. Each of these steps will be described in detail.

**Figure 5** Query file of  $\delta$ -selective opioid ligands

```

>ATOMS 8
  1 C
  2 C
  3 C
  4 C
  5 C
  6 C
  7 Pc N
  8 Hy 3 50
>CENTROIDS 1
  CR01 1 2 3 4 5 6
>BONDS 6
  1 2 1
  1 6 2
  2 3 2
  3 4 1
  4 5 2
  5 6 1
>DISCONS 3
  1
  7
  8
>DISTANCE CONSTRAINTS 3
  CR01 7 4.5 1.0
  CR01 8 6.7 1.0
  7 8 7.6 1.0
>END

```

### 1D Screening

1D screening is a rapid prescreen to eliminate database structures that cannot possibly satisfy the query. This step compares only types and numbers of atoms in a database structure with those in the query, so called 1D screening. For example, if a query contains two oxygen atoms and two nitrogen atoms, then only the database structures which contain at least two oxygen atoms and two nitrogen atoms can pass the 1D screening. For the function elements in the query, the process is similar:

1. a hydrogen bond acceptor /donor means a oxygen, nitrogen or sulfur atom.
2. a positive charge center means a atom bearing a formal positive charge or a nitrogen atom.
3. a negative charge center means a atom bearing a formal negative charge or an oxygen or nitrogen atom.
4. an hydrophobic region means at least one carbon atom.
5. an aromatic ring center means any atom type.

The 1D screening in 3DFS is much rougher than the key screening used in other 3D searching systems, but we think it is adequate, for two reasons:

1. most pharmacophores contain only a few atoms and simple structures, it is unnecessary to construct a complex key set (For example, ISIS/3D uses 962 keys in key screening), and
2. the importance of the prescreen has decreased because subsequent atom-by-atom substructure searching has been made much faster by the use of effective algorithms and powerful computers.

### 2D Substructure search

The database structures passing the 1D screening need an exact substructure search to check whether atoms interrelate

as defined in the query, i.e. whether the query is a substructure of the database structure.

Most 3D searching systems use the Ullmann algorithm [25] for substructure searching, but 3DFS system uses the GMA algorithm proposed by Xu [26-27].

The GMA algorithm is a partial-ordering-based backtracking substructure search algorithm. It consists of two steps:

Step 1. Reorder the query graph (QG) by a depth-first traversal procedure to get a Partial Order Set (POS) of the QG, i.e. a traversal route on QG.

Step 2. Use POS as a instruction set to walk on the target graph (TG). This is a constrained back-tracking procedure. If the node in TG matches the route node in POS, the walk continues, otherwise it tracks back to the last matched node and selects an alternative walk direction on TG. If all the route nodes in POS find matched nodes in the TG, i.e. the walk is complete, then QG and TG are homomorphic or isomorphic.

Because the match procedure is carried out under the direction of POS, the GMA algorithm is also called the directed match algorithm. The match route in the GMA algorithm is clearer and more perceptual than that in the Ullmann algorithm used in many commercial substructure searching systems. Moreover the GMA's computing complexity is much less than the factorial computing complexity (see ref. [27] for more details). Once a 2D match mapping is found, then the match mapping is submitted to subsequent rigid 3D search.

### Rigid 3D search

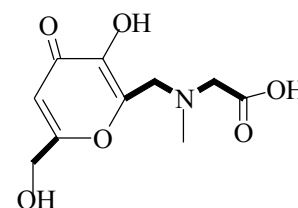
This step checks whether the atoms in the match mapping meets the spatial constraints (distance, angle, dihedral) in the query. The word "rigid" means that only the conformation stored in the database structure is checked.

If all the differentials between constraint values and measured values are within the corresponding tolerances, the structure is declared as a hit, otherwise, the database structure must undergo a further and final conformationally flexible search.

### Conformationally flexible search

In general, the database stores only a single low-energy conformation (or a limited number of such conformations) for each molecule. Accordingly, a rigid search is likely to fail to identify a large number of matching molecules that can adopt

**Scheme 9** This structure has five rotatable bonds (bold black lines)





a conformation containing the query pattern but are represented in the database by a low-energy conformation that does not contain this pattern.

The conformation space is the torsional space of the rotatable bonds in a flexible molecule. We define a rotatable bond as an acyclic single bond except for the terminal single bond. Thus, the structure shown in Scheme 9 has five rotatable bonds (bold black lines). The coordinates of the atoms in the structure will change if the rotatable bonds rotate. Conformationally flexible search means turning rotatable bonds in the database structure to fit the spatial constraints in the query. 3DFS supplies two methods to perform this process: Genetic search and Powell optimization. A user can choose one of them.

#### Genetic search

Genetic algorithms [28-30] are a class of nondeterministic algorithms that provide good, though not necessarily optimal, solutions to combinatorial optimization problems at a low computational cost. In general, a genetic algorithm can process any combinatorial optimization problem once a suitable encoding method and evaluation function for the problem are determined. In the conformationally flexible search, the genetic algorithm encodes the torsion angle of a rotatable bond in a molecule by a binary string of eight bits. One bit corresponds to an angle increment of approximately  $1.4^\circ$  (i.e.,  $360^\circ/255$ ). Each conformation of the molecule is then represented by the concatenation of these strings to form a bit string of length  $8N$ , where  $N$  is the number of rotatable bonds in the molecule. Each conformation (the bit string of length  $8N$ ) will be evaluated using an evaluation function, RMS function defined in Eq.1. Two genetic operators, crossover and mutation, are used.

The first three terms on the right-hand side in Eq.1 describe the distance, angle and dihedral angle differences between current model measured values and constraint values in query, respectively.  $I$ ,  $J$ , and  $K$  are the numbers of distance constraints, angle constraints and dihedral constraints in a query, respectively. The  $P$  term accounts for the additional penalty if there is a plane-side constraint or van der Waals energy constraint in the query.

$$RMS = \sqrt{\frac{\sum_i^I (d - d_i)^2 + \sum_j^J (\theta - \theta_j)^2 + \sum_k^K (\omega - \omega_k)^2}{(I + J + K)}} + P \quad (1)$$

The detailed steps of the genetic search are as follows:

1. An initial population of size of  $SIZE$  of the bit strings are created randomly, and RMS values are calculated.
2. Each two (parents) of  $SIZE \cdot CROSS\_RATE$  randomly-chosen bit strings swap a substring of length of  $CROSS\_BITS$  to form two new bit strings (children). The parameter  $CROSS\_RATE$  is the rate of crossover and  $CROSS\_BITS$  is the length of substrings of crossover. The RMS values of the children are calculated and the children replace the parent if the children have smaller RMS value.

**Table 3** Record items in BIP format (see Appendix for details)

tag	required
>ATOMS	Yes
>CENTROIDS	No
>PLANES	No
>LONE PAIRS	No
>BONDS	Yes
>DISCONS	Yes
>DISTANCE CONSTRAINTS	No
>ANGLE CONSTRAINTS	No
>PLANE_LINE ANGLE CONSTRAINTS	No
>PLANE_PLANE ANGLE CONSTRAINTS	No
>DIHEDRAL ANGLE CONSTRAINTS	No
>PLANE SIDE CONSTRAINTS	No
>END	Yes

3. The  $MUT\_BITS$  bits of each (parent) of  $SIZE \cdot MUT\_RATE$  randomly-chosen bit strings are switched to opposite values to form a new string (child). The parameter  $MUT\_RATE$  is the rate of mutation and  $MUT\_BITS$  is the length of substrings of mutation. The RMS value of the children is calculated and the child replaces the parent if the child has smaller RMS value.

4. Stop if the constraints of the query are fulfilled in the conformation generated in any step above. Otherwise return to step 2 until the maximal generation  $MAX\_GENT$  is reached

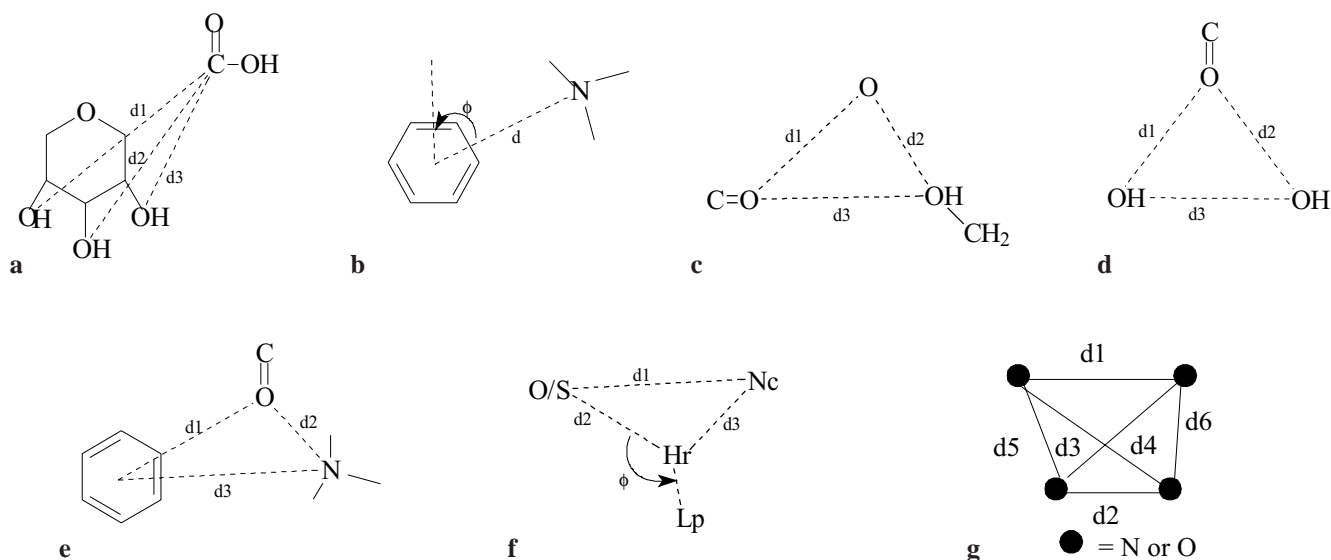
#### Powell optimization

The Powell optimization method [31] is an accelerated conjugated direction method without calculating derivatives. In the application to conformational search, each of the torsion angles of rotatable bonds is a variable and the optimization is to minimize the RMS deviation between constraint values in the query and fitted-model measured values by modifying the torsion angles (variables). The parameters include an interval tolerance  $EPS1$ , function value tolerance  $EPS2$  and a maximal iteration number  $MAX\_ITER$ .

Both the genetic and Powell optimization have the advantage of generality. The RMS function need not be continuously differentiable, and these two methods can be suitable for any type of spatial constraints.

If the conformational optimization fails, the search will return to the step of the 2D substructure search for an alternative match mapping, then repeat rigid 3D search and flexible search until a match mapping satisfying the spatial constraints is found, or no alternative 2D match mapping exists, or maximal match mapping number  $MAX\_MAPPINGS$  is reached. In order to try as many match mappings as possible, all match mappings must be non-redundant.

In the process of conformational optimization, a van der Waals energy calculation can be included optionally in order



**Figure 6** a) Pharmacophore of selectin inhibitor,  $d1$ ,  $d2$ ,  $d3=11.0 \text{ \AA} \pm 1.0 \text{ \AA}$ . b) Pharmacophore of central nervous system (CNS) drug,  $d=5.0 \text{ \AA} \pm 1.0 \text{ \AA}$ ,  $\phi=90^\circ \pm 10^\circ$ ,  $\phi$  is the angle between the normal line of phenyl ring plane and the line of centroid of phenyl ring and N atom. c) Pharmacophore of protein kinase C (PK-C) agonist,  $d1=6.0 \text{ \AA} \pm 0.25 \text{ \AA}$ ,  $d2=6.4 \text{ \AA} \pm 0.6 \text{ \AA}$ ,  $d3=5.7 \text{ \AA} \pm 0.6 \text{ \AA}$ . d) Pharmacophore of HIV-1 protease (HIV-1 PR) inhibitor,  $d1=5.4 \text{ \AA} \pm 1.0 \text{ \AA}$ ,  $d2=5.1 \text{ \AA} \pm 1.0 \text{ \AA}$ ,  $d3=2.8 \text{ \AA} \pm 1.0 \text{ \AA}$ . e) Pharmacophore of 5-HT<sub>3</sub> antagonist,  $d1=3.25 \text{ \AA} \pm 0.25 \text{ \AA}$ ,  $d2=5.25 \text{ \AA} \pm 0.75 \text{ \AA}$ ,

$d3=6.5 \text{ \AA} \pm 1.5 \text{ \AA}$ . f) Pharmacophore of angiotensin-converting enzyme (ACE) inhibitor,  $d1=4.3 \text{ \AA} \pm 0.8 \text{ \AA}$ ,  $d2=7.4 \text{ \AA} \pm 0.8 \text{ \AA}$ ,  $d3=3.8 \text{ \AA} \pm 1.0 \text{ \AA}$ , O/S represents O atom or S atom, Nc and Hr represent negative charge center and hydrogen bond acceptor, respectively, and Lp represents the lone pair direction of Hr; the 3-point angle of O/S-Hr-Lp:  $\phi=90^\circ \pm 10^\circ$ , the dihedral angle of Lp-Nc-Hr-O/S is between  $135^\circ$  and  $180^\circ$ . g) Pharmacophore of HIV-1 integrase (HIV-1 IN) inhibitor,  $d1=11.5 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d2=8.6 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d3=9.5 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d4=10.5 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d5=2.6 \text{ \AA} \pm 0.7 \text{ \AA}$ ,  $d6=2.6 \text{ \AA} \pm 0.7 \text{ \AA}$

to avoid hit conformations with unfavorable steric interactions. The vdW energy is calculated using the Lennard-Jones 6-12 potential and the vdW interactions between atoms separated by one, two, or three bonds are ignored. If the energy difference between the candidate conformation and the reference conformation stored in the database exceeds *CUTOFF*, a pre-specified parameter, then the conformation is penalized by a pre-specified amount. The addition of a vdW energy calculation ensures that hit conformations not only match the spatial constraints but are also of low steric energy, however, the vdW calculation can result in a considerable increase in CPU time for the search. This can be seen in the latter search examples.

## Performance measurement

### 3DFS system

The 3DFS system is programmed in C and runs on an AMD-K62/300 PC. 3DFS can implement three-level structure search: 2D substructure search, rigid 3D search and flexible 3D search. The 3D structure database is one part of the Open

NCI Database, containing a set of 126,705 compounds collected at the National Cancer Institute [32].

The input of 3DFS system is a query file (\*.bip) and the output is a search result file with HIT format (\*.hit). Besides the HIT output file, each hit structure may be written as a MOL file so as to be read in some free 3D visualization softwares such as RASMOL[33] or molecular modeling softwares such as ACD-3D [34] for further analysis.

In a HIT file, the first three lines record the name of the query file, the search level and starting time. From the fifth line, the hit information is recorded including the id number and NSC number of the hit, RMS and the match mapping. For example, the PK-C query (see Figure 6c) gave such a rigid 3D search output file (pkc.hit) as shown in Figure 7. The first hit in pkc.hit has NSC No. of 34318, its atom 24, 5, 2, 17 and 12 match atom 1, 2, 3, 4 and 5 of query, respectively, and its RMS of spatial constraints is 0.321906.

### Query examples

In order to investigate the efficiency and selectivity of 3DFS and its suitability for queries, seven typical pharmacophore examples from the literature were used as search examples (see Figure 6a-6g). They were the pharmacophores of selectin inhibitor [35], central nervous system (CNS) drug [36], HIV-1



```

* Query pkc.bip
* Rigid search
begin time is Tue Dec 29 08:56:49 1998

1 NSC 34318 rms=0.321906
1 2 3 4 5
24 5 2 17 12

2 NSC 36494 rms=0.311331
1 2 3 4 5
23 8 3 17 12

3 NSC 114831 rms=0.317112
1 2 3 4 5
10 12 8 17 14

4 NSC 165158 rms=0.244414
1 2 3 4 5
1 27 22 40 35

5 NSC 207108 rms=0.411611
1 2 3 4 5
1 18 14 19 15

6 NSC 207109 rms=0.411611
1 2 3 4 5
1 18 14 19 15

7 NSC 269220 rms=0.408892
1 2 3 4 5
21 14 7 25 23

end time is Tue Dec 29 09:00:33 1998

7 hits

```

**Figure 7** Rigid 3D search output file (*pkc.hit*) of PK-C query shown in Figure 6c

protease inhibitor [14], HIV-1 integrase inhibitor [17], protein kinase C (PK-C) agonist [12], 5-HT3 antagonist [37] and angiotensin-converting enzyme(ACE) inhibitor [38]. Of the seven queries, the majority were atom-based queries and only one, ACE, was a function-based query. The spatial constraints involved in distances, 3-point angles, dihedral angles, angles of line and plane and lone pair direction.

## Results and discussions

For each of the seven queries (see Figure 6a-6g), 2D search, rigid 3D search and flexible 3D search (Powell and Genetic)

were conducted on the NCI database. The number of hits and search times were listed in Table 4.

The parameter sets of genetic search were as follows: *SIZE*=50, *MAX\_GENT*=10, *CROSS\_RATE*=0.04, *CROSS\_BITS*=5, *MUT\_RATE*=0.1, *MUT\_BITS*=3. Because of the nondeterministic nature of a genetic algorithm, each search was repeated three times, and the mean results were reported.

The parameter sets of Powell method were as follows: *EPS1*=1.0, *EPS2*=1.0, *MAX\_ITER*=5.

According to Table 4, for the six atom-based queries, the 2D search times were less than 3.5 minutes. Considering that it took about two minutes only reading the 126,705 compounds in the database and computing the implicit hydrogens of each atom (since this information is not stored in the 3D database), the 2D search algorithm used in 3DFS were very efficient.

The search performance varied widely with the query. Generally, the more complex (more elements and/or more spatial constraints) the query, the fewer the hits. Of the seven queries, the selectin query was the most complex, so its hit number was lowest. For each query, flexible search identified many more hits than rigid search, but taking more time. If the vdW energy was checked, the number of hits of flexible search decreased while the search time increased. In the PK-C example, the flexible search without the vdW check took 5 times as long as the rigid search, returning 72 times as many hits, but the vdW check resulted in double search time and half as many hits.

### Comparison between genetic search and Powell optimization

The following three situations were seen from Table 4:

1. When the vdW energy check was off, for all the queries except for selectin query, genetic searches retrieved less hits in more search time than Powell optimizations.
2. When the vdW energy check was on, for the CNS, HIV-1 PR and HIV-1 IN queries, genetic searches retrieved more hits in less search time than Powell optimization, and for the other 4 queries, genetic searches retrieved less hits in less search time.
3. From without vdW energy check to with vdW energy check, for almost each queries, the increment of the search time of the genetic search was less than that of the Powell optimization.

As we know, the run time and optimization ability of a genetic algorithm may be influenced by its parameter set. Usually, increasing the population size or maximal generation number can improve the optimization ability, but increase the run time. For the exceptional selectin query in the first situation, we increased *MAX\_GENT* to 50 and repeated the genetic search. The resulting number of hits increased to 17 (equal to that of Powell optimization), but the search time increased to 188 seconds, much more than that of Powell optimization. So, We can conclude that the Powell optimization exhibits higher efficiency ( ratio of hit number to search

**Table 4** Search results of example pharmacophores in Figures 6a-6g

	Selectin		CNS		PKC		HIV-1 PR		5-HT3		HIV-1 IN		ACE	
	hits	time	hits	time	hits	time	hits	time	hits	time	hits	time	hits	time
2D	43	124s	17592	3min	2114	3min	8148	2min	7278	3.5min	67142	2min	26108	3h
Rigid	0	140s	2379	4min	7	3.5min	571	2.5min	24	5min	28	3.5min	164	4.5h
Flex[a]	17	165s	9809	29min	506	18min	2754	49min	974	42min	660	11h	3607	9h
Flex[b]	3	193s	5230	72min	232	34min	1735	121min	343	73min	516	14h	1367	11h
Flex[a,c]	14	167s	9753	33.5min	382	22min	2071	59min	866	45.5min	598	11.5h	3001	10h
Flex[b,c]	27	186s	6754	60min	176	25.5min	1863	80min	326	55.5min	533	12.5h	1356	11h
Rigid			<b>Adding</b>		3	3.5min	379	3.5min						
Flex[a]			<b>hydrophobe</b>		394	17min	2106	50min						
Rigid							<b>Function-based</b>		130	98min				
Flex[a]							<b>Query</b>		2902	4.5h				

[a] flexible search without vdW check; [b] flexible search with vdW check; [c] Genetic search

time) than the genetic search in the case without the vdW energy check.

The second situation tells us that the genetic search has the advantage over the Powell optimization in the case with the vdW energy check. This is mainly due to the fact that the genetic algorithm can easily add the vdW constraint into the evaluation function without affecting the algorithm itself, but an extra constraint apart from the objective function (RMS function) can result in the lower efficiency of the Powell method.

As an accelerated conjugate direction method, the Powell method can give fast convergence but it sometimes does not

find a minimum and misses some hits. However, the genetic algorithm can avoid becoming stuck in local minima by virtue of crossover and mutation and thus finds more hits. On the other hand, the random nature of the genetic algorithm lowers its speed. Hence, if high search efficiency (ratio of hit number to time) is desired, the Powell method is more suitable, and if as many as possible hits are desired, the genetic search should be used, especially in the case with the vdW check. If not specified, the flexible search appearing in the sections below means Powell search.

**Table 5** Structures of two active hits of the PK-C query. Groups with square frames are pharmacophoric groups

Compound	Structure
NSC 54564	
NSC 237672	

**Table 6** Structures of five active hits of the HIV-1 PR query

Compound	Structure
NSC 32180	
NSC 41234	
NSC 251156	
NSC 373937	
NSC 115497	

*Comparison between 3DFS and Chem-X [9]*

In recent years, the National Cancer Institute (NCI) has successfully applied the 3D searching technique to the discovery of novel protein kinase C (PK-C) agonists [12], HIV-1 protease inhibitors [14], and HIV-1 integrase inhibitors [17].

The searching software used was Chem-X (Chem3DB-3D) (July 1994 version) and the 3D database was the open NCI 3D database of 206,976 compounds. The search details were reported in the literature, more importantly, the open NCI 3D database contains the data set of 127,065 compounds used in the 3DFS system. So we think it is necessary and possible to compare 3DFS with Chem-X.

**Table 7** Structures of two hits of the atom-based 5-HT<sub>3</sub> antagonist query

Compound	Structure
NSC 120294	
NSC 91456	

**Protein Kinase C Agonist Query.** 535 hits for this query were found and the search took 40 hours in ref. [12]. However, Table 4 shows that our search for this query took only 18 minutes and retrieved 506 hits. When we increased the *MAX\_MAPPINGS* to 50, the search time increased to 40 minutes and the hit number increased to 554. This illustrates that 3DFS systems has higher efficiency than Chem-X.

In addition, a hydrophobic moiety is requisite for PK-C affinity, and the hits lacking a necessary hydrophobe had to be discarded manually after the search (see in ref. [12]). However, this problem can be resolved conveniently by adding a hydrophobe (Hy) into the query in the 3DFS system. The modified query yielded three rigid search hits and 394 flexible search hits in a search time of 3.5 and 17 minutes (see Table 4, without vdW check), respectively. The addition of a hydrophobe increased the accuracy of the query and led to the exclusion of inactive compounds, thus enhancing the selectivity of the hits. Of the 11 novel active compounds reported in ref. [12], only two (NSC 54564 and NSC 237672, see Table 5, groups with square frames are pharmacophoric groups) are in our database and both were found by the 3DFS search.

**HIV-1 protease inhibitor query.** 2,368 hits for this query were found (search time not given) in ref. [14]. Table 4 shows our search retrieved 2,754 hits (49 minutes). Of the 15 active compounds listed in ref. [14], only five are in our database (NSC 32180, NSC 41234, NSC 251156, 373937 and NSC 115497, among which NSC 32180 is the most active, see

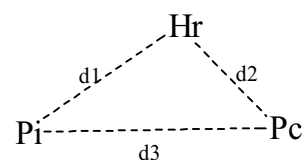
**Figure 8** Function-based query of 5-HT<sub>3</sub> antagonist. Pi: Aromatic Center, Hr: Hydrogen bond acceptor, Pc: Positive charge center,  $d1=3.0-3.5$  Å,  $d2=4.5-6.0$  Å,  $d3=5.0-8.0$  Å

Table 6). 3DFS found all five compounds. Similar to the PK-C agonist, a hydrophobe is also requisite to HIV-1 protease inhibitors. The query modified by adding a hydrophobe (Hy) yielded 2,106 hits also containing the five active compounds (50 minutes).

**HIV-1 integrase inhibitor query.** 179 hits for this query were found (search time not given) in the ref. [17]. Table 4 shows our search retrieved 660 hits (11 hours). Because this query has a very simple 4-point structure and all the molecules (many of the 67,142) contain the four atoms and must undergo a rigid 3D search or even a flexible search, the search time was as long as 11 hours. Unfortunately, of the 39 compounds assayed in the ref. [12], none is in our database, so our 660 hits do not include the 20 novel active compounds reported in ref. [17].

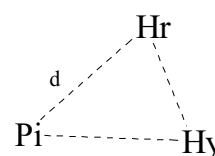
In the searches of the above three examples, 3DFS retrieved more hits within a satisfactory search time from a smaller database compared to Chem-X. This is mainly because of the efficient search algorithms used in the 3DFS system. Moreover, because 3DFS supports function-based queries and uses a rapid hydrophobe recognition algorithm, the selectivities of hits of the PK-C query and HIV-1 PR query were increased while the numbers of hits are decreased.

#### Comparison between atom-based and function-based queries

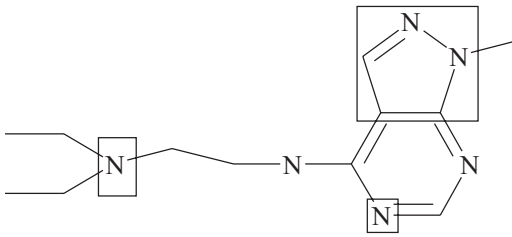
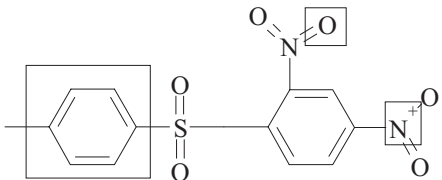
The PK-C and HIV-1 PR examples above have shown the importance of functional elements in query. Here we take the 5-HT<sub>3</sub> query as a example to illustrate further the meaning of a function-based query for the discovery of structurally novel compounds.

Table 4 shows that 5-HT<sub>3</sub> query retrieved 24 rigid search hits and 974 flexible search hits (without vdW check). Two (NSC 120294 and NSC 91456) of the hits are shown in Table 7.

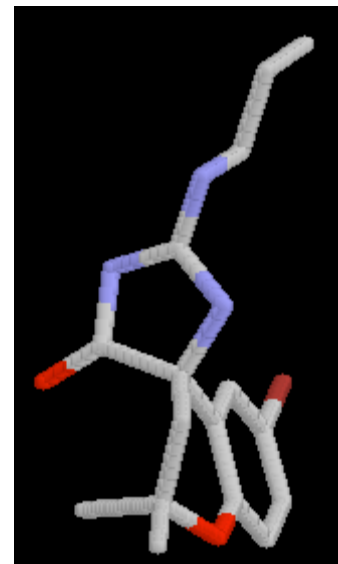
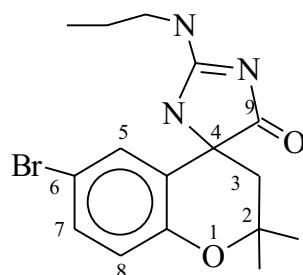
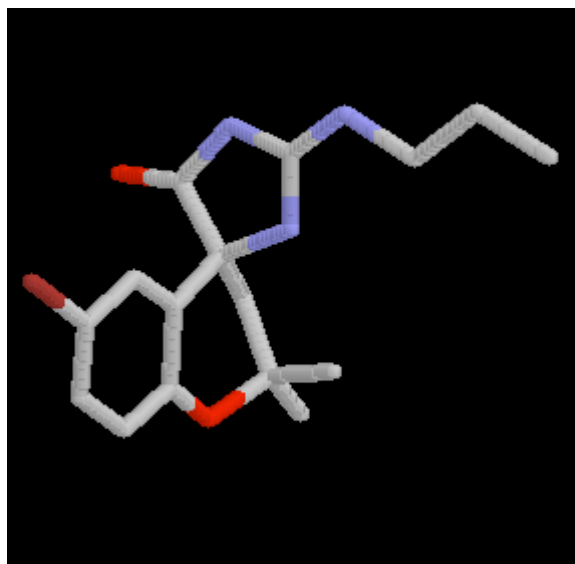
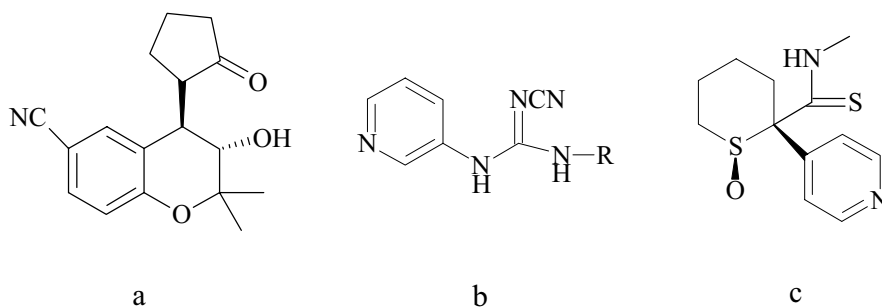
According to the report of Hibert [37], the carbonyl oxygen in the query was serving as a hydrogen bond acceptor, the basic nitrogen was serving as a charge center and the

**Figure 9** Pharmacophore of three types of KCOs shown in Scheme 10. Pi: Aromatic center, Hy: hydrophobic region, Hr: Hydrogen bond acceptor and  $d=3.8-5.2$  Å

**Table 8** Structures of two hits of the function-based 5-HT<sub>3</sub> antagonist query

Compound	Structure
NSC 11633	
NSC 25830	

**Scheme 10** Structures of three types of KCOs of cromacalim (a), pinacidil (b), and aprical (c)



**Scheme 11** Structure of template KCO and its two enantiomers. The left is the active S isomer and the right is the inactive R isomer



phenyl ring was serving as a aromatic ring center. Thus a function-based query (see Figure 8) can be defined by substituting hydrogen bond acceptor, positive charge center and aromatic ring center for the oxygen, nitrogen and phenyl ring, respectively. This function-based query yielded 130 rigid search hits and 2,902 flexible search hits (see Table 4). Two (NSC 11633 and NSC 25830) hits are shown in Table 8.

Apparently, the function-based query retrieved more hits than the atom-based query because of generality of the function elements. Another more important result was that the structural diversity of the hits were also increased. Though with different backbones, the two hits in Table 7 have some common substructures, at least one phenyl ring, one carbonyl group and one amine group. However, the two hits in Table 8 are completely different structures without any common substructures.

### Application -K<sup>+</sup> channel openers

#### K<sup>+</sup> channel opener query

Ion channels are protein gates that permit rapid passage of certain ions. Potassium ion channel openers (KCO) have gained attention as effective smooth muscle relaxants as antihypertensive and bronchodilating drugs [39-41].

Despite the absence of 3D structural data for the K<sup>+</sup> channel, efforts have been made in the past by several groups to study the structure-activity relationship and to identify a common pharmacophore shared by structurally diverse KCOs by molecular modeling approaches. The previous studies [42] from our laboratory has proposed a common pharmacophore (see Figure 9) for the three types of KCOs of cromacalim, pinacidil and aprical (see Scheme 10). The pharmacophore consists of an aromatic ring, a hydrogen bond acceptor and a hydrophobic region with a distance constraint 3.8 - 5.2 Å between the aromatic ring centroid and the hydrogen bond acceptor (see Figure 9).

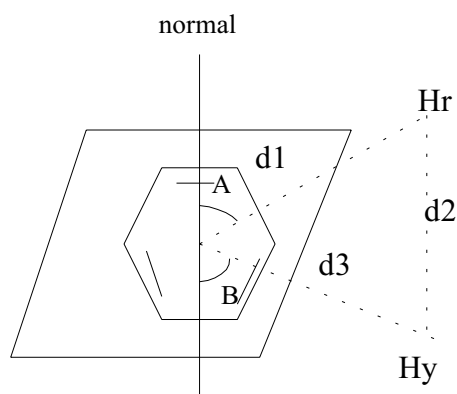


Figure 10 Query pharmacophore of KCOs

In order to obtain a more accurate pharmacophore, we used a highly active KCO molecule with a rigid spiro-ring (see Scheme 11) as a template. First, the geometry of the template molecule was optimized by the MOPAC (AM1) [43] method in the SYBYL6.0 package [44] using the default parameters running on a SGI workstation. The pharmacophore in the low energy conformation involves the phenyl ring, the C<sub>9</sub> carbonyl as the hydrogen bond acceptor (Hr) and the two C<sub>2</sub> methyl groups as a hydrophobic region (Hy). The distance between the oxygen at C<sub>9</sub> and the centroid of the phenyl ring is 4.030 Å, the distance between the oxygen at C<sub>9</sub> and the centroid of the hydrophobic region is 4.969 Å, and the distance between the centroid of the phenyl ring and the centroid of the hydrophobic region is 4.283 Å. The locations of the oxygen at C<sub>9</sub> and the phenyl ring are fixed while the centroid of the hydrophobe may move if the C<sub>3</sub> atom are also regarded as a hydrophobic atom. This suggests that relatively larger tolerances should be used for the two distances involv-

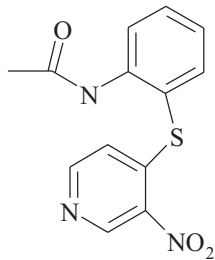
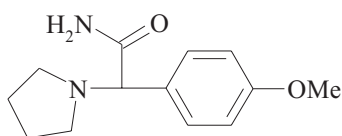
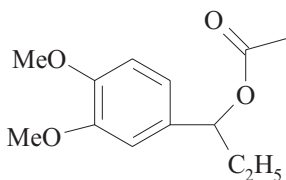
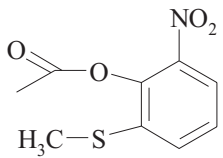
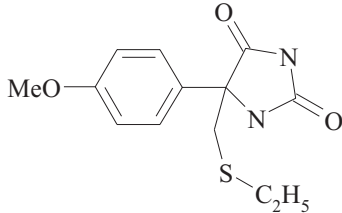
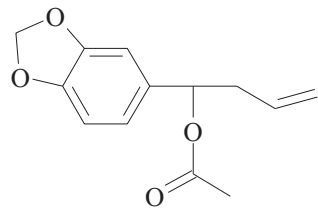
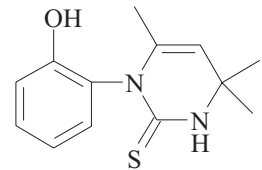
```

>ATOMS 8
 1 C
 2 C
 3 C
 4 C
 5 C
 6 C
 7 Hr *
 8 Hy 3 6
>CENTROIDS 1
CR01 1 2 3 4 5 6
>PLANES 1
PL01 1 2 3
>BONDS 6
 1 2 2
 1 6 1
 2 3 1
 3 4 2
 4 5 1
 5 6 2
>DISCONS 3
 1
 7
 8
>DISTANCE CONSTRAINTS 3
CR01 7 4.0 0.5
CR01 8 4.3 1.0
 7 8 5.0 1.0
>PLANE_LINE ANGLE CONSTRAINTS 2
PL01 CR01 7 52.0 10.0
PL01 CR01 8 75.0 10.0
>PLANE SIDE CONSTRAINTS 1
PL01 7 || 8
>END

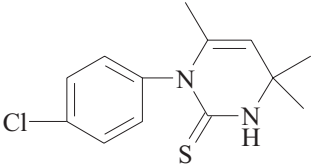
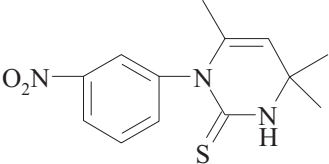
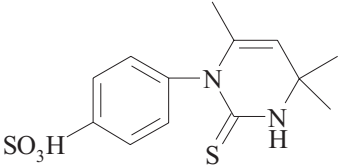
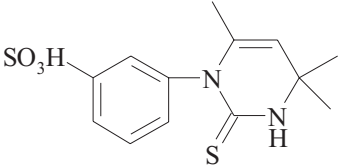
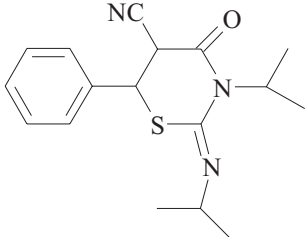
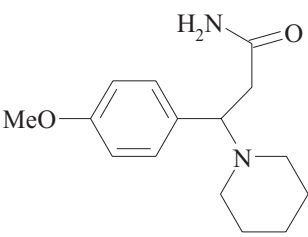
```

Figure 11 Query file (kco.bip) of KCOs pharmacophore

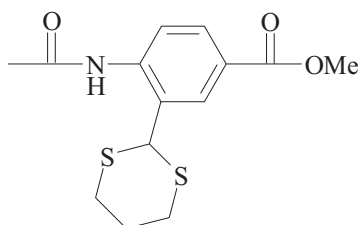
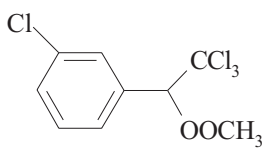
**Table 9a** Structures of some hits found by 3DFS using the query in Figure 10

Compound	RMS	Structure
NSC 155706	0.490	
NSC 163362	0.352	
NSC 15713	0.414	
NSC 298304	0.356	
NSC 59325	0.605	
NSC 6623	0.414	
NSC 49826	0.322	

**Table 9b** Structures of some hits found by 3DFS using the query in Figure 10

Compound	RMS	Structure
NSC 49829	0.323	
NSC 49831	0.323	
NSC 49840	0.323	
NSC 5808	0.323	
NSC 298212	0.499	
NSC 140742	0.371	

**Table 9c** Structures of some hits found by 3DFS using the query in Figure 10

Compound	RMS	Structure
NSC 298310	0.193	
NSC 59894	0.351	

ing the hydrophobic region while a smaller tolerance should be used for the distance between the oxygen at C<sub>9</sub> and the phenyl ring. In addition, the number of atoms in the hydrophobic region is limited to 3 - 6 since the most KCO molecules have small hydrophobic region.

The normal line of the phenyl ring plane has an angle of 51.85° to the line of the centroid of the phenyl ring and the oxygen at C<sub>9</sub>, and an angle of 74.95° to the line of the oxygen at C<sub>9</sub> and the hydrophobe. Both the two angle constraints have the tolerance of 10° in the query.

It should be noted that the 4*S* isomer of the template molecule is highly active, but the 4*R* isomer is almost inactive. The oxygen at C<sub>9</sub> and the hydrophobe locate in the opposite sides of the phenyl ring plane in the 4*S* isomer while in the same side in the 4*R* isomer. So, the plane side constraint was added into the query.

The pharmacophore query which incorporates all these aspects is shown in Figure 10.

### Search results

KCO query file- pkc.bip (see Figure 11) was input to the 3DFS system and searches were conducted on the open NCI database of 126,705 compounds. A total of 147 rigid search hits (54min) and 2,425 flexible search hits with vdW energy check (9h) were retrieved. Table 9 shows 15 hits with chemical novelty and synthetic accessibility. The synthesis and bioassay of some hits are under way.

### Summary

3DFS is a computer program which searches a 3D database for compounds matching a given pharmacophore query. Its characteristics lie in three aspects:

1. using a set of practical binding function definitions for a function-based query, in which, besides the function libraries of Hbond acceptors/donors and charge centers, a new aromatic ring recognition rule and a rapid hydrophobe recognition algorithm were proposed.

2. using the highly effective GMA algorithm for 2D substructure search, unlike other 3D searching systems using Ullmann algorithm.

3. supplying two conformationally flexible search methods: genetic search and Powell optimization, which have advantages in the term of either hit yield or speed.

The searches of the typical pharmacophore queries demonstrated the high utility of the 3DFS system, even advantages over the commercial software Chem-X. Finally, the 3DFS system was applied to the K<sup>+</sup> channel opener studies.

**Supplementary material available** For all compounds in Table 5-9 3D structure pictures (gif format) and cartesian coordinates (MDL mol format), for the stereoisomers in Scheme 11 cartesian coordinates (MDL mol format) are available as supplementary material. The 3DFS program and example files are available from the author upon request.

### References

1. Martin, Y. C. *J. Med. Chem.* **1992**, *35*, 2145-2154.
2. Pearlman, R. S. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; Leiden: ESCOM Science Publishers, 1993; 41-47.
3. Willett, P. *J. Mol. Recogn.* **1995**, *8*, 290-303.
4. Finn Paul, W. *DDT* **1996**, *1*, 363-370.
5. Sheridan, R.; Nilakantan, R.; Rusinko III, A.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255-260.

6. Van Drie, J.; Weininger, D.; Martin, Y. *J. Comp.-Aided Mol. Design* **1989**, *3*, 225-251.
7. MACCS-II/3D is a product of MDL Information Systems, Inc., San Leandro, CA, USA. <http://www.mdli.com>.
8. ISIS/3D is a product of MDL Information Systems, Inc., San Leandro, CA, USA. <http://www.mdli.com>.
9. Chem-X is a product of Chemical Design Ltd., Roundway House, Cromwell Park, Chipping Norton, Oxon OX7 5SR, UK. <http://www.oxmol.com/prods/chem-x>.
10. SYBYL/3DB Unity is a product of TRIPOS Associates, Inc., St. Louis, MO, USA. <http://www.tripos.com>.
11. Catalyst is a product of Molecular Simulation, Inc., San Diego, CA., USA. <http://www.msi.com>.
12. Wang, S.; Zaharevitz, D. W.; Sharma, R.; Zaharevitz, D. W.; Sharma, R.; Marquez, V. E.; Lewin, N. E.; Du, L.; Blumber, P. M.; Milne, G. W. *J. Med. Chem.* **1994**, *37*, 4479-4489.
13. Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, D. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N. *Science* **1994**, *263*, 380-384.
14. Wang, S.; Milne, G. W. A.; Yan, X. ; Posey, I. J.; Nicklaus, M. C.; Graham, L.; Rice, W. G. *J. Med. Chem.* **1996**, *39*, 2047- 2054.
15. Nicklaus, M. C.; Neamati, N.; Hong, H.; Mazumder, A.; Sunder, S.; Chen, J.; Milne, G. W.; Pommier, Y. *J. Med. Chem.* **1997**, *40*, 920-929.
16. Hong, H.; Neamati, N.; Wang, S. *J. Med. Chem.* **1997**, *40*, 930-936.
17. Neamati, N.; Hong, H.; Sunder, S.; Milne, G. W.; Pommier, Y. *Mol. Pharmacol.* **1997**, *52*, 1041-1060.
18. Kiyama, R.; Homma, T.; Hayashi, K.; Ogawa, M.; Hara, M.; Fujimoto, M.; Fujishita, T. *J. Med. Chem.* **1995**, *38*, 2728-2741.
19. Kaminski, J. J.; Rane, D. F.; Snow, M. E.; Weber, L.; Rothofsky, M. L.; Anderson, S. D.; Lin, S. *J. Med. Chem.* **1997**, *40*, 4103-4112.
20. Wang, T.; Zhou, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 71-77.
21. Greene, J.; Kahn, S.; Hamid, S.; Sprague, P.; Teig, S. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297-1308.
22. Ghose, A.; Crippen, G. *J. Comput. Chem.* **1986**, *319*, 199-203.
23. Montanaari, C. A.; Tute, M. S.; Beezer, A. E.; Mitchell, J. C. *J. Comput. Aided Mol. Des.* **1996**, *10*, 67-73.
24. Huang, P.; Kim, S.; Loew, G. *J. Comput. Aided Mol. Des.* **1997**, *11*, 21-28.
25. Ullmann, J. R. *J. of the Association for Computing Machinery* **1976**, *23*, 31-42.
26. Xu, J.; Zhang, M. *Tetrahedron Comput. Methodol.* **1989**, *2*, 75-83.
27. Xu, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 25-34.
28. Holland, J. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
29. Jones, G.; Willett, P.; Glen, R. C. *Genetic Algorithms and Their Application to Problems in Chemical Structure Handling*. Proceedings of the 1994 Chemical Information Conference, Annecy.
30. Cartwright, H. M. *Pesticide Science* **1995**, *45*, 171-178.
31. Powell, M. J. *Computer Journal* **1964**, *7*, 155-162.
32. Open NCI Database is accessible to general public via <ftp://helix.nih.gov/ncidata/3D/nciopn3d.mol.Z> or <http://www2.ccc.uni-erlangen.de/ncidb>.
33. RASMOL v2.6 is a product of BioMolecular Structures Group, Glaxo Research & Development, Greenford, Middlesex, UK.
34. ACD-3D is a product of Advanced Chemistry Development, Inc., 133 Richmond St. West, Suite 605, Toronto, Ontario, M5H 2L3, Canada. <http://www.acdlabs.com>.
35. Rao Narasinga, B.N.; Anderson, M. B.; Musser, J.H.; Gilbert, J.H.; Schaefer, M.E.; Foxall, C.; Brandley, B. K. *J. Biol. Chem.* **1994**, *269*, 19663-19666.
36. Lloyd, E. J.; Andrews, P.R. *J. Med. Chem.* **1986**, *29*, 453-462.
37. Hibert, M. F.; Hoffmann, R.; Miller, R. C.; Carr, A. *J. Med. Chem.* **1990**, *33*, 1594-1599.
38. Mayer, D.; Naylor, C. B.; Motoc, G. R.; Marshall, G. R. *J. Comput. Aided Mol. Des.* **1987**, *1*, 3-16.
39. Edwards, G.; Weston, A. H. *Trends Pharmacol. Sci.* **1990**, *11*, 417-422.
40. Atwal, K. S. *Med. Res. Rev.* **1992**, *12*, 569-591.
41. Longman, S. D.; Hamilton, T. C. *Med. Res. Rev.* **1992**, *12*, 73-148.
42. Chen, H. ; Pang, S.; Zhou, J.; Xu, Z. *Computers and Applied Chemistry* **1997**, *14*, 31-33 (Chinese).
43. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902-3909.
44. The SYBYL package is a product of TRIPOS Associates, Inc., St. Louis, MO, USA. <http://www.tripos.com>.

---

#### Appendix : BIP query file format

##### >ATOMS M

The record item provides the information about atoms contained in query. M is the number of atoms(M≤125). One data line is for one atom.

##### Format:

atom\_id atom\_type(impH)

##### Fields:

atom\_id(integer) = id number of atom.

atom\_type(string) = symbol of actual atom such as C, O, N or Cl; Symbols of such a point as

\*(Any atom type), Cn(Chain), Hr(Hbond acceptor), Hd(Hbond donor), Pc(Positive charge center), Nc(Negative charge center), Hy(Hydrophobe), Pi(Aromatic ring center), Da(N or O), Db(N, O or S) or Dc(O or S).

If atom\_type is Hy, two integers must follow representing the min number and max number of atoms in the hydrophobe, respectively. The default values are 3 and 50.

If atom\_type is Hr or Hd, the type(string) of main atom in Hbond acceptor or Hbond donor must follow. The default is \*(any atom type).



impH(string) = implicit hydrogen atom only if implicit hydrogen of atom must be matched.

Example:

```
>ATOMS      4
 1          N
 2          CH2
 3          Hy      3      6
 4          Hr      N
```

In this example, there are four atoms in the query, the first atom is N, the second one is C with two hydrogens, the third one is a hydrophobe with the number of atoms between 3 and 6, and the fourth one is a Hbond acceptor which main atom is N atom.

>CENTROIDS M

The record item provides the information about centroids in the query, if defined. M is the number of centroids ( $M \leq 10$ ). One data line is for one centroid.

Format:

CR\_id atom\_id atom\_id ... atom\_id

Fields:

CR\_id (string) = id number of centroid

atom\_id (integer) = id number of component atom of centroid

Example:

```
>CENTROIDS  2
CR01        1  2  4  5  7
CR02        3  5  7  6
```

In this example, two centroids are defined in the query, the first one CR01 determined by atom 1, 2, 4, 5 and 7 and the second one CR02 determined by atom 3, 5, 7, and 6.

>PLANES M

The record item provides the information about planes in the query, if defined. M is the number of planes ( $M \leq 5$ ). One data line is for one plane.

Format:

plane\_id atom\_id atom\_id ... atom\_id

Fields:

plane\_id (string) = id number of plane.

atom\_id (integer) = id number of component atom of plane

Example:

```
>PLANES    2
PL01       1  2  7
PL02       3  5  6
```

In this example, two planes are defined in the query, the first one PL01 is determined by atom 1, 2 and 7, the second one PL02 is determined by atom 3, 5 and 6.

>LONE PAIRS M

If the direction of a lone pair of an atom is constrained in the query, the atom is specified in this record item. M is the number of such atoms ( $M \leq 5$ ). One data line is for one atom.

Format:

lone-pair\_id atom\_id

Fields:

lone-pair\_id (string) = id number of lone pair

atom\_id (integer) = id number of atom with constrained lone pair direction

Example:

```
>LONE PAIRS  1
LP01         4
```

In this example, the lone pair direction of one atom is constrained, the id number of the atom is 4. The id number of the lone pair is LP01.

>BONDS M

The record item provides the information about bonds in query. M is the number of bonds ( $M \leq 125$ ). One data line is for one bond.

Format:

origin\_atom\_id target\_atom\_id bond\_type

Fields:

origin\_atom\_id (integer) = id number of the atom on one end of bond

target\_atom\_id (integer) = id number of the atom on the other end of bond

bond\_type (integer) = bond type; 1 for single bond, 2 for double bond, 3 for triple bond.

Example:

```
>BONDS      3
 1  2  1
 1  3  1
 2  4  2
```

In this example, there are three bonds in the query. The first bond is a single bond connecting atom 1 and 2. The second one is a single bond connecting atom 1 and 3. The third one is a double bond connecting atom 2 and 4.

>DISCONS M

The record item provides the information of disconnectivity of a pharmacophore query. M is the number of disconnected fragments in query ( $M \leq 6$ ). One data line is for one disconnected fragment.

Format:

atom\_id

Fields:

atom\_id (integer) = id number of any atom in disconnected fragment

Example:

```
>DISCONS 2
1
4
```

In this example, the query consists of two disconnected fragments, atom 1 is the starting atom of the first fragment and atom 4 is the starting atom of the second fragment.

>DISTANCE CONSTRAINTS M

The record item provides the information about distance constraints in the query. M is the number of distance constraints ( $M \leq 10$ ). One data line is for one constraint.

Format:

```
from_point_id to_point_id dis error_d
```

Fields:

from\_point\_id (integer or string) = id number of atom (or centroid) on one end of the distance

to\_point\_id (integer or string) = id number of atom (or centroid) on the other end of distance

dis (float) = distance value (Å)

error\_d (float) = allowed tolerance (Å)

Example:

```
>DISTANCE CONSTRAINTS 3
1 2 3.4 0.5
2 3 5.4 0.6
CR01 3 6.5 1.0
```

In this example, there are three distance constraints in the query. The distance between atom 1 and atom 2 is 3.4 and the tolerance is 0.5. The distance between atom 2 and atom 3 is 5.4 and the tolerance is 0.6. The distance between centroid CR01 and atom 3 is 6.5 Å and the tolerance is 1.0.

>ANGLE CONSTRAINTS M

The record item provides the information about 3-point angle constraints in the query. M is the number of angle constraints ( $M \leq 10$ ). One data line is for one constraint.

Format:

```
left_point_id cent_point_id right_point_id ang error_ang
```

Fields:

left\_point\_id (integer or string) = id number of left atom(centroid, or lone pair) used to define a 3-point angle

cent\_point\_id (integer or string) = id number of the *vertex* atom(centroid) used to define a 3-point angle

right\_point\_id (integer or string) = id number of right atom(centroid) used to define a 3-point angle

ang (float) = angle value (degree)

error\_ang(float) = allowed tolerance (degree)

**Note:** If the left (or right) point of a 3-point angle is lone pair, the vertex must be the atom which the lone pair belongs to.

Example:

```
>ANGLE CONSTRAINTS 3
1 2 3 60.0 5.0
1 CR01 3 50.0 10.0
LP01 2 3 35.0 5.0
```

In this example, there are three 3-point angle constraints in query. The angle  $\angle 1-2-3$  is  $60^\circ$  and the tolerance is  $5^\circ$ . The angle  $\angle 1-CR01-3$  is  $50^\circ$  and the tolerance is  $10^\circ$ . The angle  $\angle LP01-2-3$  is  $35^\circ$  and the tolerance is  $5^\circ$  and LP01 is the lone pair of atom 2.

>PLANE\_LINE ANGLE CONSTRAINTS M

The record item provides the information about angle constraints between the normal line of a plane and a line defined by two points in the query. M is the number of such angle constraints ( $M \leq 5$ ). One data line is for one constraint.

Format:

```
plane_id from_point_id to_point_id g error_g
```

Fields:

plane\_id (string) = id number of plane

from\_point\_id (integer or string) = id number of one atom(centroid, or lone pair) used to define a line

to\_point\_id (integer or string) = id number of the other atom(centroid) used to define a line

g(float) = angle value (degree)

error\_g (float) = allowed tolerance (degree)

Example:

```
>PLANE_LINE ANGLE CONSTRAINTS 2
PL01 2 3 60.0 5.0
PL01 CR01 3 50.0 10.0
```

In this example, there are two plane\_line angle constraints in the query. The angle between the normal line of plane PL01 and the line of atom 2 and atom 3 is  $60^\circ$  and the tolerance is  $5^\circ$ . The angle between the normal line of plane PL01 and the line of centroid CR01 and atom 3 is  $50^\circ$  and the tolerance is  $10^\circ$ .

>PLANE\_PLANE ANGLE CONSTRAINTS M

The record item provides the information about angle constraints between two planes in the query. M is the number of such angle constraints ( $M \leq 5$ ). One data line is for one constraint.

Format:

```
plane_id plane_id p error_p
```

Fields:

plane\_id (string) = id number of one plane

plane\_id (string) = id number of the other plane

p (float) = angle value (degree)

error\_p(float) = allowed tolerance (degree)

Example:

```
>PLANE_PLANE ANGLE CONSTRAINTS 1
PL01 PL02 60.0 5.0
```

In this example, there is one plane\_plane angle constraint in query. The angle between plane PL01 and plane PL02 is 60° and the tolerance is 5°.

#### >DIHEDRAL ANGLE CONSTRAINTS M

The record item provides the information about dihedral angle constraints in the query. M is the number of dihedral angle constraints (M≤10). One data line is for one constraint.

##### Format:

point1\_id point2\_id point3\_id point3\_id dih error\_dih

##### Fields:

point1\_id (integer or string) = id number of first atom (centroid or lone pair)

point2\_id (integer or string) = id number of second atom (centroid)

point3\_id (integer or string) = id number of third atom (centroid or lone pair)

point4\_id (integer or string) = id number of fourth atom (centroid)

dih (float) = angle value (degree)

error\_dih (float) = allowed tolerance (degree)

##### Example:

```
>DIHEDRAL ANGLE CONSTRAINTS 2
  1      2      3      4      60.0      5.0
CR01    2      4      5      50.0      10.0
```

In this example, there are two dihedral angle constraints in the query. The dihedral angle of atom 1→2—3→4 is

60° and the tolerance is 5°. The dihedral angle of CR01→2—3→4 is 50° and the tolerance is 10°.

#### >PLANE SIDE CONSTRAINTS M

The record item specifies that two points lie on the same side or opposite side of a plane defined in the query. M is the number of the plane side constraints (M≤5). One data line is for one constraint.

##### Format:

plane\_id point1\_id || (&) point2\_id

##### Fields:

plane\_id (string) = id number of plane

point1\_id (integer or string) = id number of one atom (centroid)

point2\_id (integer or string) = id number of the other atom (centroid)

& = same side

|| = opposite side

##### Example:

```
>PLANE SIDE CONSTRAINTS 2
PL01    1      ||      3
PL02    1      &      4
```

In this example, there are two plane side constraints in the query. Atom 1 and atom 2 lie on the opposite side of plane PL01. Atom 1 and atom 4 lie on the same side of plane PL02.