Development of a Three-Dimensional Substructure Search Program for Organic Molecules

Hiroaki Kato and Yoshimasa Takahashi*

Laboratory for Molecular Information Systems, Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology, Tempaku-cho, Toyohashi 441

(Received September 6, 1996)

A computer program, called SS3D, has been developed for the three-dimensional substructure searching of organic molecular structures. Our approach is based on a clique-finding algorithm that compares chemical graphs containing edges labelled with inter-atomic distances. The program allows one to specify some allowance of the distance for the geometrical matching of 3D-substructures; it also allows one to define some constraints for the query substructure or for the geometrical pattern to be used in the search. The details concerning the program are discussed and several illustrative examples are given.

An understanding of the three-dimensional structural features of drug molecules is necessary for many problems in chemistry. In particular, a substructural analysis or a functional group analysis is essential for structure-activity (or property) studies and rational molecular design based on them. Such a process involving a structural (substructural) feature analysis could be done manually for a small set of molecules with two-dimensional structural information. However, the work is quite tedious and time consuming for a large set of molecules, even if it is handled in a topological or two-dimensional space. It is clear that computerized techniques are very useful in this case.[1—4] One of the most basic and common techniques for a computerized structural feature analysis is substructure searching, which has developed in the area of chemical information systems, including structure databases.[5—8] Until recently, substructure searching has been carried out only on 2D structures. However, it is clear that molecular properties, including biological function, relay not only on atom connectivity or the topological mean but also on the three-dimensional geometrical arrangement of the atoms.

For the last decade the increasing availability of 3D structural information has prompted many researchers to establish 3D structure search and substructure search techniques.[9—13] 3D substructure searching and its analogies were originally developed for the specific purpose of carrying out searches for pharmacophore patterns in drug molecules with similar pharmacological activity;[14—16] they have been integrated with other tools in computer-aided molecular design systems.[17,18] With the same view point, the authors have developed a computer program for finding pharmacophore patterns within sets of molecules, called COMPASS (COMmon geometrical PAttern Search System).[19,20] However, this system was designed for use only in our computer-aided molecular design system.[21] The geometric search of substructural features or 3D substructure searching for more general purposes is still highly desired for various areas of computer-aided chemistry, not only in molecular design. We are currently engaged in a project to develop computational techniques for the 3D structural feature analysis of proteins which can provide a basis for automatic motif finding, and have recently discussed the used of 3D substructure search technique for the efficient implementation of a protein motif search.[22] Here, we report on the development of a computer program for 3D substructure search (SS3D) which can be employed as a basic tool for various lines of research.

Methods

Basic Concept of the Approach: The basic concept of 3D substructure search used in the present work is shown in Fig. 1. We have made the assumption that the 3D structures of the molecules presented here are rigid. Each molecule is treated as a set of points that correspond to its constituent atoms in the 3D-space. The set of points is described by a matrix representation, of which each element involves the inter-atomic distance within the molecule. Thus, since the set of points can be regarded as being an edge-weighted graph, the graph is a complete graph in which every pair of vertices are connected. In other words, we can represent the

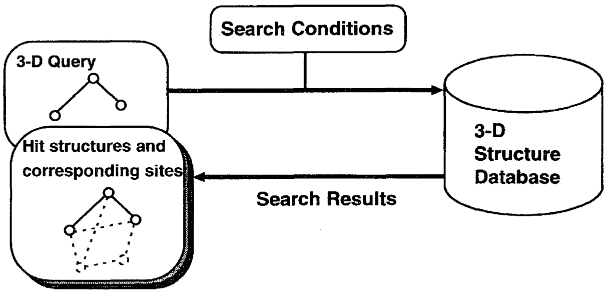


Fig. 1. Basic concept for the 3D substructure search in the present work.

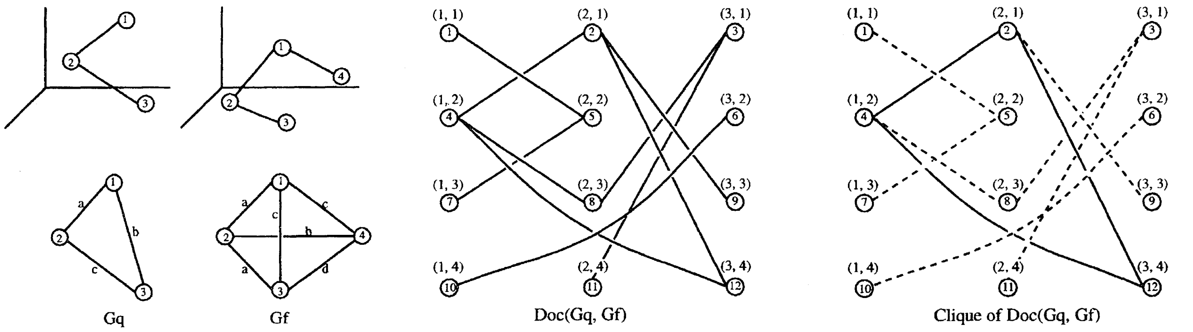


Fig. 2. Docking graph and clique obtained from graphs G_q and G_f.

structural information of a molecule, including its 3D geometry, with a weighted (or labeled) graph of which the nodes and edges correspond to atoms and the inter-atomic distances between them, respectively. On the basis of this, a 3D substructure search can be treated as a subgraph matching problem, which is the basis of an ordinary topological or 2D substructure search.

Algorithms and Implementation: The algorithm presented here is based on that of COMPASS which was reported in our previous paper.[19] First, the algorithm starts with the three-dimensional atomic coordinate data of a 3D-query substructure and that of the molecule to be compared. Assuming that all of the atoms of their structures are equivalent, both of them can be expressed by sets of points in three-dimensional rectangular coordinate space. It then becomes possible to describe the geometry for both the 3D-query and the file molecule by simply using a distance matrix corresponding to the set of points. Such distance matrices are easily created from the three-dimensional atomic coordinate data of the query and file molecule. A distance matrix so produced can be regarded as a representation of an edge-weighted complete graph consisting of the same number of vertices as that of the constituent atoms in the query and the file molecule. Thus, the problem for a three-dimensional substructure search can be handled in topological and graph theoretical ways, such as a subgraph search.

In the present work, subgraph matching is implemented by the use of a docking graph[23] and a clique finding algorithm.[24] The complete graphs derived from the 3D-query and the file molecule are compared with each other according to the criteria of edge-equality based on the edge-weight (i.e. interatomic distance) with a tolerance value specified in advance. On this basis an alternative graph, called a docking graph, is produced. The docking graph (Doc) of G_q for the query and G_f for the file one is defined as follows:

$$\mathrm{Doc}(G_q, G_f) = < \boldsymbol{V}, \boldsymbol{E} >,$$
$$\text{where } \boldsymbol{V} = < (\sigma, \mu) \mid \sigma \in G_q, \mu \in G_f >$$
$$\text{and } \boldsymbol{E} = < [(\sigma_i, \mu_k), (\sigma_j, \mu_l)] \mid\mid w_q(i,j) - w_f(k,l)| \leq \delta > .$$

Here, $w_q(i,j)$ is the weight on the edge between vertices i and j in graph G_q, and $w_f(k,l)$ is the weight on the edge between vertices k and l in graph G_f. Here, $\boldsymbol{V}$ and $\boldsymbol{E}$ represent the sets of vertices and edges in the docking graph, respectively. σ and μ denote the vertices contained in graphs G_q an G_f, respectively. δ is the allowance for the weights at which they are considered to be equivalent. The so-produced docking graph is also an ordinary connected or disconnected graph. A clique is a maximal complete subgraph in which every vertex is connected to every other vertex, and which is not contained in any larger complete subgraph. These are illustrated in Fig. 2. Thus, the present problem of the 3D substructure search is to examine whether the docking graph has one or more cliques or not, in which the size (the number of vertices) is the same as that of the query, and then to identify any clique(s) that it finds. For clique finding, a tree-search method based on a back-tracking procedure was used. Since the basic algorithm for the clique search used here has already been reported elsewhere,[19] the details will not be repeated here.

A three-dimensional substructure search program (SS3D) was developed based on the algorithm described above. A flowchart depicting the overall scheme of SS3D is shown in Fig. 3. All of the programs were written in the FORTRAN 77 and C language, and implemented on a UNIX workstation, a Sun SPARCstation 5.

Results and Discussion

The validity of the program has been demonstrated by using a few selected examples. The results are described in the following section. The first trial was carried out with a small data set. The data set involved only nine molecules (in Fig. 4), which were selected to validate the algorithm and to illustrate the overview of the 3D substructure search by the SS3D. The three-dimensional atomic coordinates

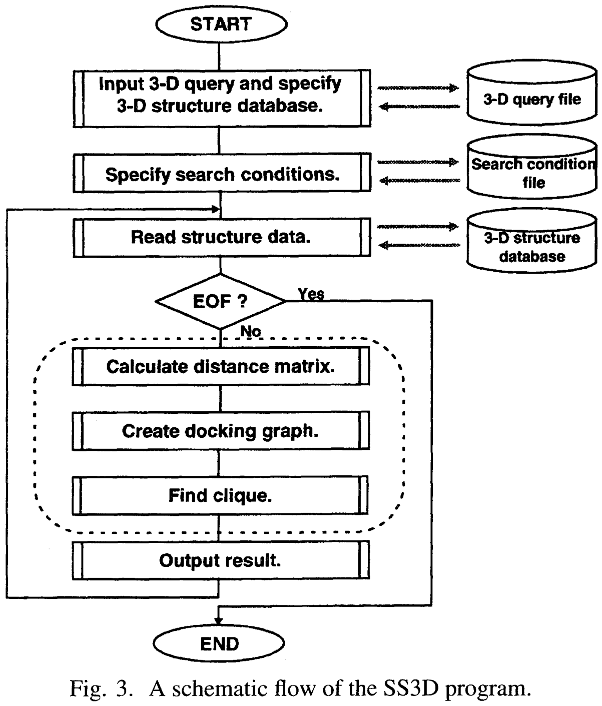


Fig. 3. A schematic flow of the SS3D program.

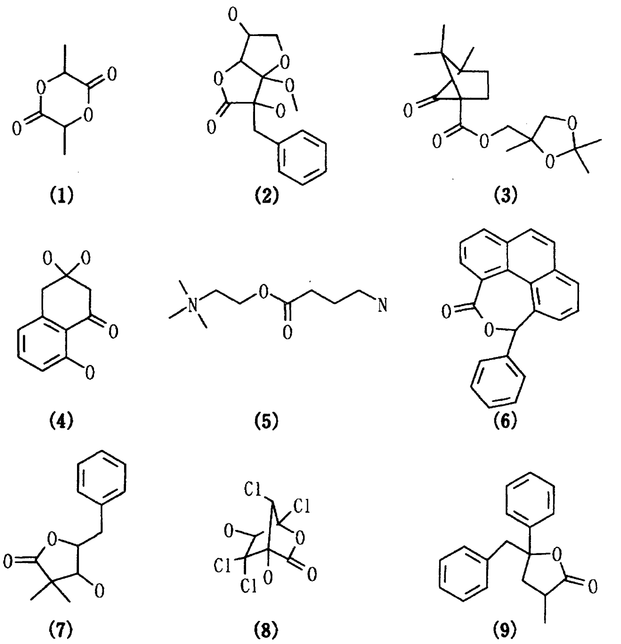
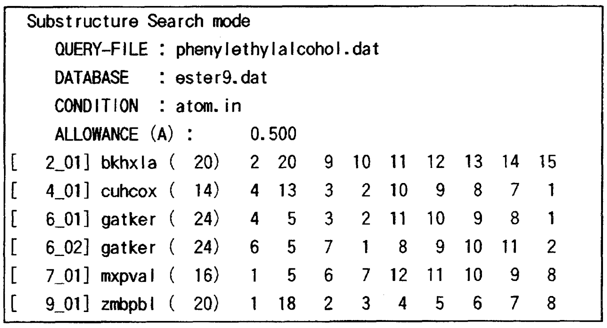


Fig. 4. A data set used for the search trial by the SS3D program.



```
  Substructure Search mode
     QUERY-FILE : phenylethylalcohol.dat
     DATABASE   : ester9.dat
     CONDITION  : atom.in
     ALLOWANCE (A) :         0.500
[   2_01] bkhxla (  20)    2  20   9  10  11  12  13  14  15
[   4_01] cuhcox (  14)    4  13   3   2  10   9   8   7   1
[   6_01] gatker (  24)    4   5   3   2  11  10   9   8   1
[   6_02] gatker (  24)    6   5   7   1   8   9  10  11   2
[   7_01] mxpval (  16)    1   5   6   7  12  11  10   9   8
[   9_01] zmbpbl (  20)    1  18   2   3   4   5   6   7   8
```

Fig. 5. The result of the 3D substructure search trial for the nine molecules presented in Fig. 4. (Part of the output of the SS3D program.) The query used in the current search trial was the phenylethylalcohol moiety of the molecule (2). This output summarizes the search conditions, sample & site identifier, the reference code of each compound in the CSD, the number of atoms (excepting hydrogen atoms in this case) and the detected site that is represented with a set of identifiers of the corresponding atoms.

of the molecules were taken from the Cambridge Structural Database (CSD). It can be noticed that these molecules are very different from each other in terms of the geometrical structure, including their sizes. A molecular fragment, a 2-phenylethyl alcohol moiety, derived from the molecule (2) in Fig. 4, was used as the 3D-query substructure for the present analysis. A search trial was carried out under the condition that the maximum allowed for the distance (δ) is 0.5 Å, and that the different types of elements are distinguished during the matching process. The results are given in Figs. 5 and 6. Figure 5 shows part of the output of our program which summarizes the search condition, the hit molecules

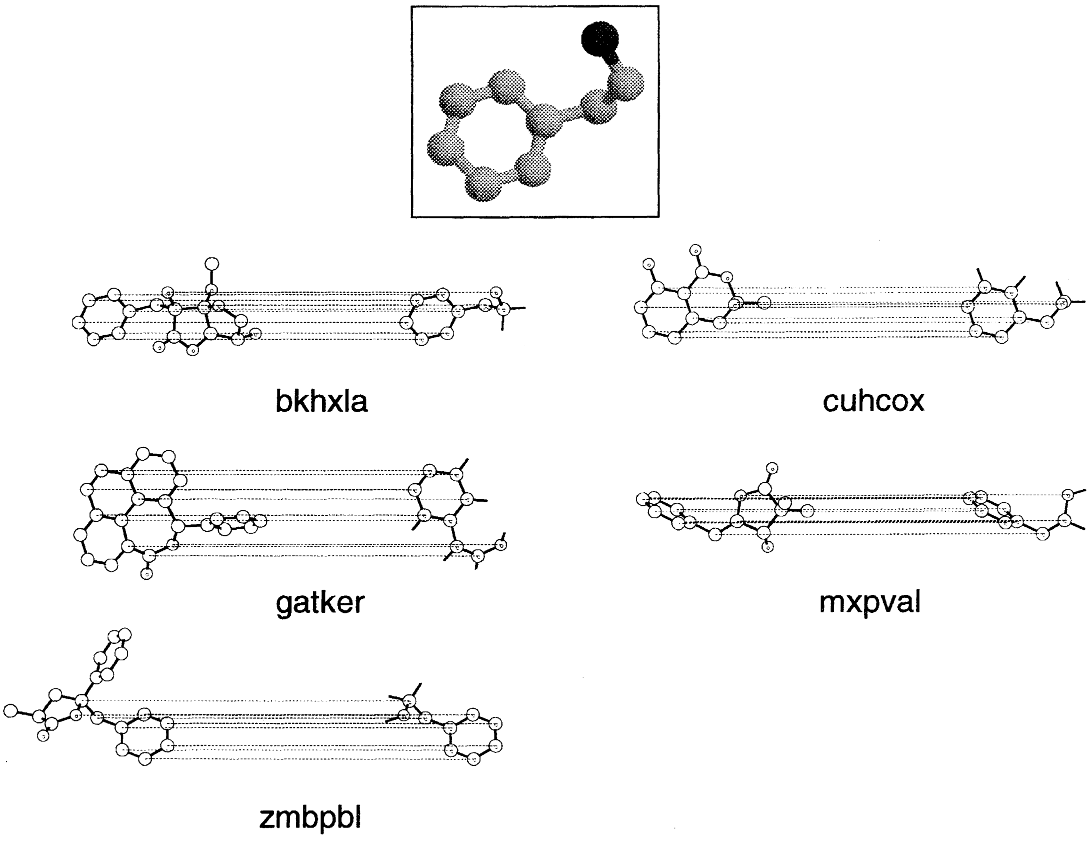
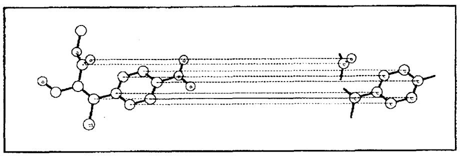


Fig. 6. Three-dimensional graphical views of the hit molecules and the hit sites in Fig. 5.

and the hit sites. It is worth noting that SS3D can identify all possible sites within a molecule that contains two or more corresponding sites within the molecule. Unfortunately, no methods have been established to systematically evaluate the validity of such a substructure search result. Thus, the present search result was validated by manually checking the distance matrices. Figure 6 shows some graphical views of the search result and presents the 3D structures of the five hit molecules and their sites corresponding to the query (the first one only is the case in which two or more corresponding sites exist). This result also affords a visual validation that the current search has been performed correctly.

The second trial was carried out in a more practical sense. In this trial 352 esters extracted from the CSD file were used to prepare the test database. A pair of 3D molecular fragments, a benzyl group and a carbonyl moiety, derived from FALPEN in Fig. 7, were used for the current query. In this case it should be noted that the 3D-query substructure means a disconnected substructure. In the present analysis, the effect of the allowance of the distance was examined, initially. Several search trials were executed with different search conditions, such that the value of the allowance of the distance varied from 0.2 to 1.0 Å. In every case, different kinds of elements were distinguished during atom matching. The results are summarized in Table 1. It is obvious that a larger value of the allowance gives a larger number of hit molecules and hit sites. Additionally, some trials with the different ways of weighting the atoms were also carried out. For these trials, the allowance of the distance was kept with the value of 0.6 Å. The results are summarized in Table 2. While the search with a set of simple points (no weighting) gave twenty seven hit sites for twenty two compounds, the searches with the element type and the hybridization-distinguished mode gave thirteen and five hit compounds, respectively. Obviously, stricter weighting results in a lower number of hits. On the other hand, it also should be noticed that a search trial using partial charges for weighting atoms gives a slightly different result. This weighting is not for an exact matching of atoms, but for similarity matching based on their functionality. This fact is quite important, particularly since 3D substructure searches are used in many structure-activity studies. The current program, SS3D allows us to include such additional information for weighting the atoms.



```
3-D Query (benzyl+carbonyl from FALPEN)
METHYL(E)-3-CHLORO-2-FORMYL-3-(4-NITROPHENYL)CINNAMATE FALPEN
CambridgeDB 3D geometry + Gaussian88 charge data
 9
 1  6  0  0  3   2  4   6 -4   7  1
 2  6  0  0  2   1  4   3 -4
 3  6  0  0  2   2 -4   4  4
 4  6  0  0  2   3  4   5 -4
 5  6  0  0  2   4 -4   6  4
 6  6  0  0  2   1 -4   5  4
 7  6  0  0  1   1  1
 8  6  0  0  1   9  2
 9  8  0  0  1   8  2
 1    0.2187   -0.7659   -0.1347    0.0186
 2    1.1149   -1.5055    0.6171   -0.0467
 3    2.4414   -1.1320    0.6778   -0.0462
 4    2.8476   -0.0407   -0.0546    0.0835
 5    1.9863    0.6879   -0.8300   -0.0508
 6    0.6726    0.3210   -0.8686   -0.0541
 7   -1.2267   -1.1243   -0.1519    0.0547
 8   -1.9473    1.1362    0.4723    0.3438
 9   -1.4416    1.4660    1.4986   -0.2666
```

Fig. 7. The 3D-query substructure for the second search trial for 352 esters. The connection table which contains the atomic coordinates and the atomic charges used in the SS3D program is also displayed. The first three lines are comments, the fourth line shows the number of atoms, the next nine lines are the atom connectivity information, and the last nine lines are for the 3D coordinates and atomic charges. The same specification is also used for the description of each file molecule.

Table 1. The Results of the SS3D Search for 352 Esters Using Different Values of the Distance Allowance

Distance allowance (Å)	No. of hit compounds	No. of hit sites
0.2	1	1
0.4	5	5
0.6	13	13
0.8	26	34
1.0	80	183

Table 2. The Results of the SS3D Search for 352 Esters Using Various Atom Weighting Schemes

Atom weighting	No. of hit compounds	No. of hit sites
None	22	27
Atom	13	13
Hybiridization	5	5
Charge	6	6

Conclusions

A computer program used for 3D substructure searches, SS3D, has been developed. This program can identify all of the query-corresponding sites within each molecule in the database. It is executable for a query using a connected substructure and a set of 3D structural fragments (disconnected query substructures) that possess a specific geometrical arrangement. The atom-weighting scheme presented in this work allows us to make use of various attributes of the atoms in the 3D substructure search. The present approach is not specific to only 3D substructure searching, and SS3D provides an expandable in-house tool for a structural feature analysis of 3D molecules.

We gratefully acknowledge the support of this work by Special Coordination Funds for Promoting Science and Technology, Science and Technology Agency, Japan.

References

1) G. W. Adamson, M. F. Lynch, and W. G. Town, *J. Chem. Soc. C*, **1971**, 3702.

2) W. E. Brugger, A. J. Stuper, and P. Jurs, *J. Chem. Inf. Comput. Sci.*, **16**, 105 (1976).

3) G. Klopman, *J. Am. Chem. Soc.*, **106**, 7315 (1984).

4) C. R. Carhart, D. H. Smith, and R. Venkatarahavan, *J. Chem. Inf. Comput. Sci.*, **25**, 64 (1985).

5) E. H. Sussenguth, *J. Chem. Doc.*, **5**, 36 (1965).

6) J. Figueras, *J. Chem. Doc.*, **12**, 237 (1972).

7) J. R. Ullmann, *J. ACM*, **16**, 31 (1976).

8) P. Willett, *J. Chemometrics*, **1**, 139 (1987).

9) A. T. Brint and P. Willett, *J. Mol. Graphics*, **5**, 49 (1987).

10) R. P. Sheridan, R. Nilakantan, A. Rusinko, N. Bauman, K. S. Haraki, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, **29**, 255 (1989).

11) Y. Martin, M. G. Bures, and P. Willett, "Reviews in Computational Chemistry," ed by K. B. Lipkowitz and D. B. Boyd, VCH, New York (1990), p. 213

12) P. Willett, "Three-Dimensional Chemical Structure Handling," Research Study Press, Taunton (1991).

13) Y. C. Martin, *J. Med. Chem.*, **35**, 2145 (1992).

14) P. Gund, *Prog. Mol. Subcell. Biol.*, **5**, 117 (1977).

15) A. M. Lesk, *Commun. ACM*, **22**, 219 (1979).

16) T. Esaki, *Chem. Pharm. Bull.*, **30**, 3657 (1982).

17) J. H. Drie, D. Weininger, and Y. C. Martin, *J. Comput.-Aid. Mol. Design*, **3**, 225 (1989).

18) T. Hurst, *J. Chem. Inf. Sci.*, **34**, 190 (1994).

19) Y. Takahashi, S. Maeda, and S. Sasaki, *Anal. Chim. Acta*, **200**, 363 (1987).

20) Y. Takahashi, T. Akagi, and S. Sasaki, *Tetrahedron Comput. Method.*, **3**, 27 (1990).

21) Y. Takahashi, K. Hosokawa, F. Yoshida, M. Ozaki, and S. Sasaki, *Anal. Chim. Acta*, **217**, 61 (1989).

22) H. Kato and Y. Takahashi, "Proceedings of Genome Informatics Workshop 1994," Universal Academy Press, Tokyo (1994), p. 162.

23) F. S. Kurl, G. M. Crippen, and D. K. Friesen, *J. Comput. Chem.*, **5**, 24 (1984).

24) C. Bron and J. Kerbosh, *Commun. ACM*, **16**, 575 (1973).