Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical Structure

Yoshimasa Takahashi,* Masayuki Sukekawa, and Shin-ichi Sasaki

Department of Knowledge-Based Information Engineering, Toyohashi University of Technology, Tempaku, Toyohashi 441, Japan

Received July 2, 1992

This paper describes an approach for automatic identification of the similar structural features among molecules. Here, each structure is described with an abstracted chemical graph of which each node describes a functional atomic group and the edge is weighted by the topological path length (the number of bonds) between two functional atomic groups to be considered. Isosterism for functional groups or substructures to be considered can be defined as a kind of knowledge file, the contents of which depend on the problem. A clique-finding algorithm was used for the search of common or similar structural features. Details of the algorithms will be discussed here, and a couple of illustrative examples in structure–activity problems will be shown for the applications.

INTRODUCTION

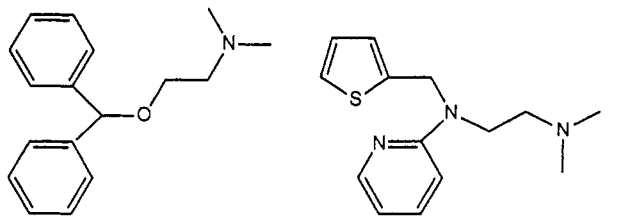
Molecular similarity is quite an important concept in chemistry. But it is difficult to give a universal interpretation to the similarity because there are many aspects of its definition, e.g., structural aspects, molecular properties, local reactivities, biological activities, etc. Thus, the evaluation of the molecular similarity always depends on the problem to be focused on.

There have been many attempts in this direction, see refs 1 and 2. Most of the approaches are based (1) on the parametrization of the chemical structure by means of various specific property values and with the assistance of statistical analysis or pattern recognition methods to establish a numerical model. (2) A second method consists of the examination of structural commonalities or differences directly from the structural information of the chemical compounds themselves. Within the latter class, the approaches based on the topological representation of the structure are very useful because they allow the comparison and examination of similarities between a comparatively large number of compounds at the level of structural formulas. A maximal common substructure problem is also one of the interesting subjects along with this interest.[1,3-5] We developed and reported a system, called MAXFIT,[6] directed to the automated recognition of the maximal common substructure in a set of chemical structures. MAXFIT had as its original objective the search and examination of a maximal common substructure within a set of structures in terms of a connected subgraph.

However, on examination of the problems related to structure–activity relationships, it is often pointed out the relevance of structural features associated with some of the substructures mutually apart from each other with particular distance.[7-9] Consider the two structures in Chart I. If one focuses on the maximal common substructure, one can just get a dimethylaminoethyl moiety. So, it might be said they are not very similar. From the other point of view, however, one may say they are very similar because that both of them have two aromatic rings and a tertiary nitrogen atom, in addition the topological path length between one of the rings and the nitrogen is the same with 5 in both cases. In actuality, both of them have the same biological activity, antihistamine activity.

From this standpoint, it is desirable to have a new approach for the finding of "common" or "similar" structural features

Chart I. These Two Are Similar or Not?



composed of several substructures corresponding to a disconnected subgraph. Taking into account all this, this paper describes an object-oriented approach for automatic identification of molecular similarity based on the knowledge file.

REDUCED REPRESENTATION OF CHEMICAL STRUCTURE

The present approach uses as the sole input the structural information in the connection table for the chemical compound. However this information is not handled directly in that form but is transformed into a reduced structural representation that is based on functional groups. This representation is an abstracted graph formed by the connection of all the nodes after they have been corresponded to the functional atomic groups. Every edge of the graph is weighted by the topological path length between the atomic groups related to the terminal nodes of each edge. Accordingly, the reduced representation of chemical structure simplifies the exclusion of parts of the molecule that are not relevant in a structure–activity relationship analysis, replacing them by means of a 'spacer', and thus increases the effectiveness of the structure handling process.

First, the system reads the connection table and proceeds with the preprocessing of this information, registering the node numbers of the functional groups perceived in the structure (Figure 1). This perception is performed by means of a substructure search technique.[10] A molecular structure, however, possesses frequently many interrelations between atoms, each of them yielding different functional groups. When this happens, the system assumes all the possible combinations and looks for all the functional groups that are in agreement with the desired ones. The functional atomic groups to be

0095-2338/92/1632-0639$03.00/0

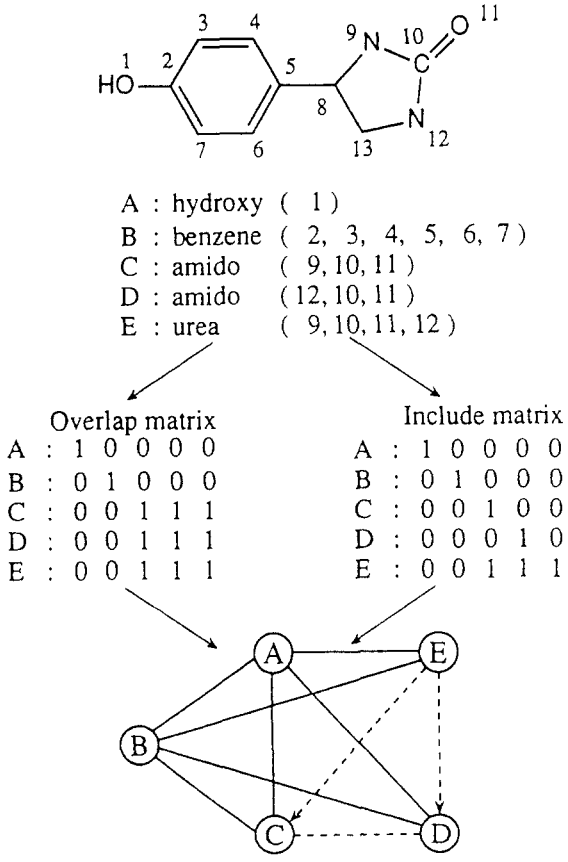


Figure 1. Generation of reduced-graph representation of a chemical structure.

considered are defined in advance and stored in a special file (substructure definition file) by the user.

Once all the functional atomic groups are perceived, the interrelations between them are checked. The relationships evaluated are described in terms of matrix expression, and they are divided into the following two cases:

(i) The case in which two functional groups are partially overlapping.

(ii) The case in which one of the functional atomic groups is completely included by another one.

For the former case, the relationship is described in the overlapping matrix, and the latter is described in the inclusion matrix to avoid the duplication of nodes in the reduced-graph representation. They are later used at the stage of determination of path length between functional groups and at the stage of docking graph construction.

In the next step, the distance between functional atomic groups is determined. This distance is called the topological path length. If the structure has ring(s), there are several possible paths that can be drawn simultaneously between functional groups. In such a case, some of edges are weighted with multiple values (Figure 2). In the present work, the backtrack procedure was used for finding all the paths between two functional atomic groups. At that time, all the bonds contained in the functional groups are neglected and all the meaningless paths such as those that cover the functional group itself are not taken into consideration. Consequently, the chemical structure in Figure 1 is expressed with a node and edge-weighted graph as shown in Figure 3.

In practice, the complete reduced-structure representation is computed by means of a computer program according to this procedure.

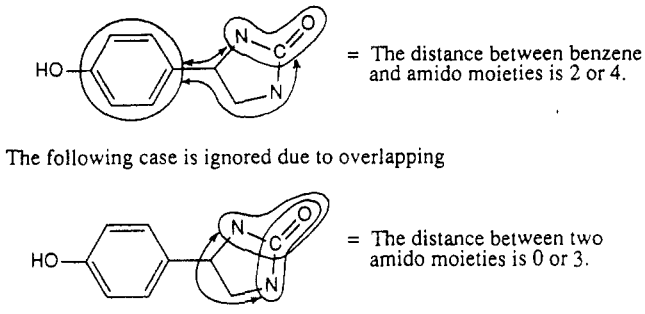


Figure 2. Determination of the distance between functional atomic groups. In the upper case, the distance (3 or 3) between the benzene ring and the alternative amido moiety is taken account, too.

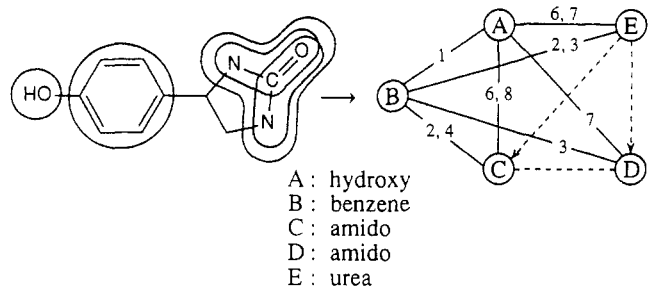


Figure 3. Example of the complete reduced-graph representation of the chemical structure.

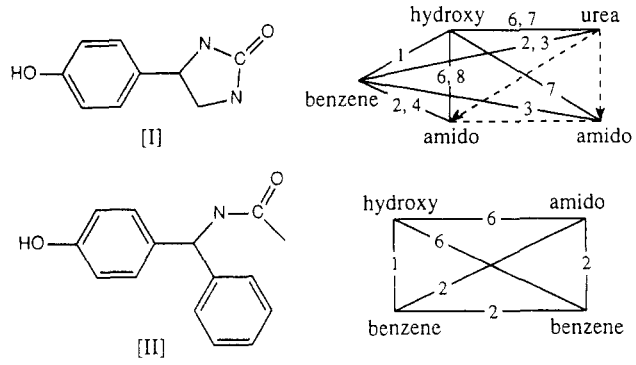


Figure 4. Chemical structures and their reduced-graph representations based on the functional atomic groups.

COMMON STRUCTURAL FEATURE SEARCH

The reduced-graph representation in Figure 3 corresponds to a graph composed of weighted edges and with nodes equal to the number of functional atomic groups perceived. Therefore, the common structural feature search between molecules becomes a problem that can be handled by means of graph-theoretical approaches.

We now examine the common structural features between the structures I and II in Figure 4. The process of the feature search is divided roughly into two steps. They are (1) the generation of the docking graph[11] of the two graphs and (2) the determination of the maximal cliques[12] of the docking graph. The docking graph, here, is an attempt to pairwise map like functional groups in the two original graphs. The mapping works if the distances between any two groups also fit. A clique in the docking graph corresponds to a grouping of functional groups in the original graphs, where all the intragrouping distances are the same in both original graphs.

Docking graph generation: Figure 5 summarizes the generation of a docking graph of the two reduced graphs in Figure 4. The docking graph was created on the basis of the common weight criteria for nodes and edges of the graphs. Here, nodes of the docking graph are all possible combinations of a functional group from graph I with a like functional

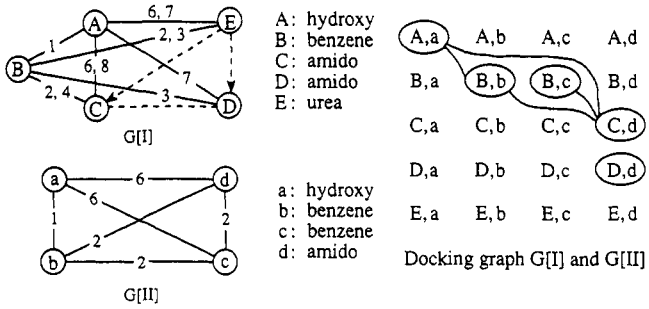


Figure 5. Generation of the docking graph with two reduced graphs in Figure 4.

group from graph II. If the distance between the two functional groups in graph I equals the distance between the corresponding groups in graph II, then the pertinent vertices in the docking graph are linked by an edge. Thus, the docking graph could also be termed "fitting graph". In Figure 5 both "A" and "a" have the same node label, hydroxy group, and "B" and "b" are benzene rings, etc. Consequently, only five nodes should be considered for the following process. On the other hand, focusing on the edges, the edge AB of G(I) and ab of G(II) have the same weight, 1. Thus the nodes (A,a) and (B,b) are linked as shown in Figure 5. In a similar way, (A,a) and (C,d), (B,b) and (C,d), and (B,c) and (C,d) are linked. Thus the docking graph is built using these functional atomic group-based graphs, by means of a common weighing criteria for nodes and edges.

The so-created docking graph is an ordinary connected or disconnected graph possessing the same information on weighted nodes and edges as the original reduced-graph representation of the chemical structures. Consequently, the finding for the clique of the docking graph is equivalent to searching for the common structural features between the original chemical structures.

Clique finding of docking graph: One of aims of the present work is to identify not only the maximal connected common structural features but also other possible common features, including disconnected ones, simultaneously. Along with this aim, the clique-finding process has been designed to search and store not only the real clique but also other complete subgraphs in order to get every size of common structural feature. A clique is a maximal complete subgraph in which every node is connected to every other node and which is not contained in any larger subgraph with this property. Every candidate of the clique obtained by this process, that is, the set of nodes and edges that forms the common subgraph, could be corresponded to a common structural feature between molecules under the analysis.

For the clique finding, a tree-search method based on the back-track procedure was used (Figure 6). As the basic algorithm for clique search used here had already been reported elsewhere,[12,13] the details will not be repeated here. The current example of a common structural feature search is illustrated in Figure 7. It should be noticed that every subgraph of a clique defines common structural features. Further, the edge weight of 1 between (A,a) and (B,b) in Figure 7 means the direct connection between the two functional atomic groups expressed with the symbols (i.e., A-B or a-b shows a phenol moiety in each structure).

SIMILAR STRUCTURAL FEATURE SEARCH USING ISOSTERISM KNOWLEDGE

The preceding discussion defines the fundamentals of the "common" structural feature search in our approach. Now,

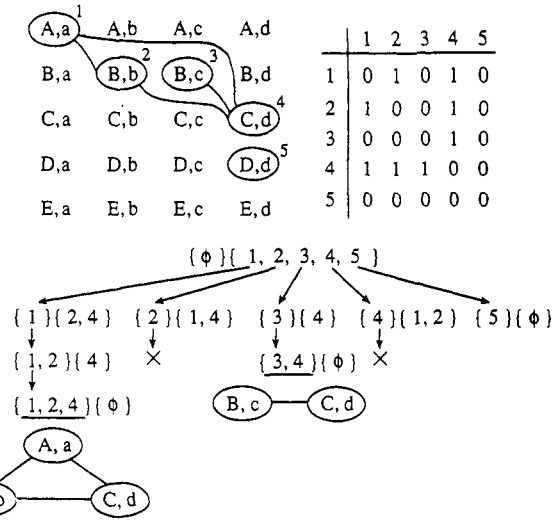


Figure 6. Clique finding of the docking graph.

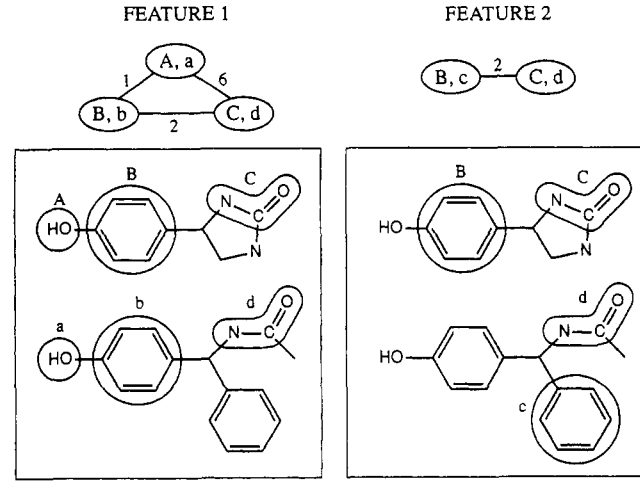


Figure 7. Resulting common structural features for two structures in Figure 4.

returning to the structures in Chart I, one can get just a dimethylaminoethyl moiety as the maximal common substructure between the two, still. We now show how to get the "similar" structural features, like a topological pharmacophoric pattern, as discussed in the beginning of this paper. This problem could be overcome by setting up equivalence groups of edge pairs and/or node pairs in the process of docking graph generation. Allowances for differences in edge weights can be specified when corresponding edges during docking graph formation. On the other hand, a node isosterism definition file, that is a kind of substructure knowledge file, can take into account extended commonality (i.e., similarity) for the nodes. An overview of the process is shown in Figure 8.

The contents of the substructure knowledge file definitely depends on the problem at hand. In this work, knowledge on the functional isosterism has been defined and stored as substructure knowledge from the standpoint of structure–activity relationships. However, functional atomic groups may be corresponding based on a variety of different rules. In this way, the system can refer to the functional groups from several different points of view. For example, the -NH- group may be referred to in several means (a member of amine fragments, H-bonding donor, or electron pair donor). Thus, even if the real fragment does not survive in a "common" structural feature, the alternative feature may be concluded as the "similar" structural feature under the reasoning of the

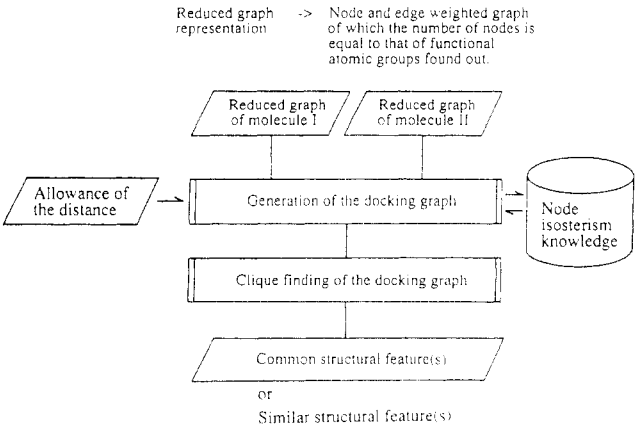


Figure 8. Schematic flow for the similar structural feature search by the use of knowledge file.

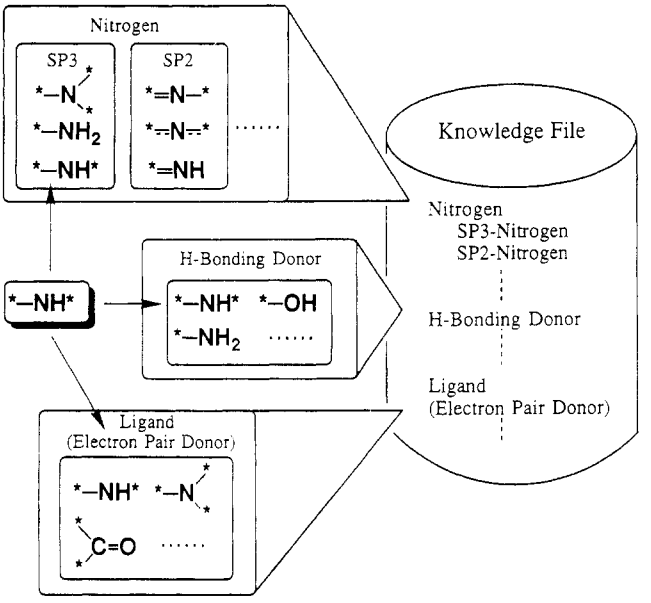


Figure 9. Illustrative example of the substructure knowledge file used in the present system.

substructure knowledge. Figure 9 illustrates the substructure knowledge file used here.

OVERVIEW OF THE SYSTEM

Basic flow of the system is shown in Figure 10. The processes are summarized as follows: (1) the system reads atomic connection tables, (2) functional atomic groups or substructures to be considered are perceived by the substructure search with a substructure definition file, (3) interrelations between them are examined and stored into the related matrices, (4) the distances between the substructures are evaluated, (5) reduced-graph representations of the original chemical graphs (chemical structures) are generated on the basis of the node and edge weighted graph, (6) the docking graph of the two is generated according to the isosterism knowledge for substructures, (7) the common or similar subgraph search is carried out by means of a clique-finding algorithm, and (8) the abstracted topological feature(s) obtained are translated to the corresponding structural features as "common" or "similar" structural features between the chemical structures of interest.

For three or more molecules, the system carries out the common structural feature search for the next molecule and the stepwise found common structural features already found at the preceding step.

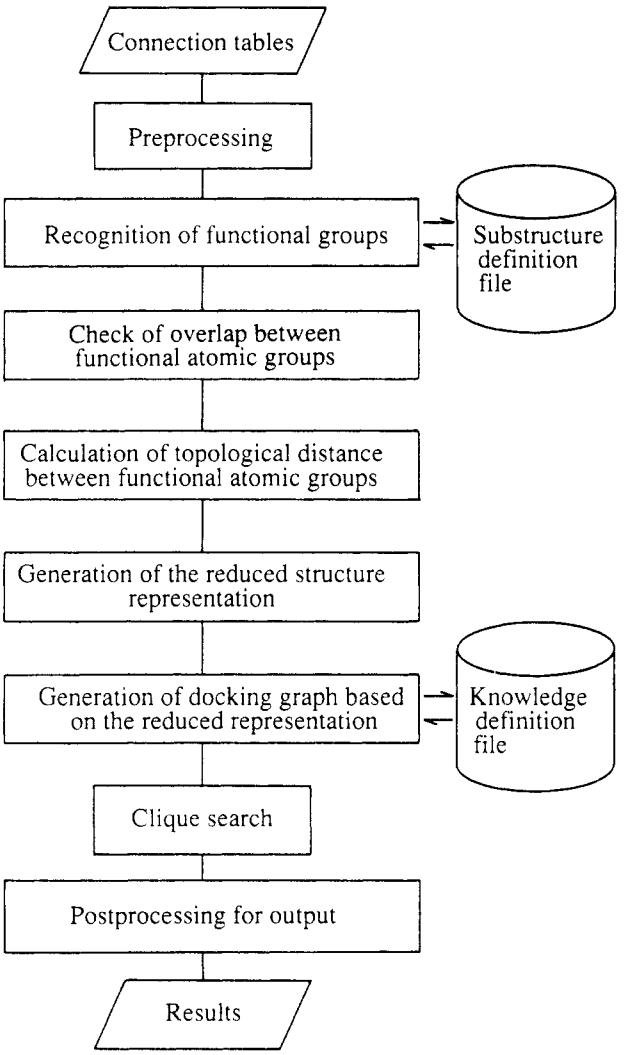


Figure 10. Basic flow of the computerization of the present approach for similar structural feature search.

All of the programs for the present system were written by the FORTRAN77 program on the Data General AV-300 UNIX workstation.

EXAMPLES OF APPLICATION

Using the program implemented on the algorithm so long explained, two brief examples are presented to illustrate structure–activity problems for which the present approach might be useful. In addition, the different results obtained for different specification of similarity among chemical structures.

The first example finds the similar structural features among five structurally diverse antihistamines (diphenhydramine, methapyrilene, cyproheptadine, dimethindene, and promethazine). The results are summarized in Figure 11. In this case, all of the structures possess two aromatic rings and a tertiary nitrogen atom at least. Furthermore, their two aromatic rings and the nitrogen atom are separated from each other by either 5 or 6 bonds, respectively, and the two aromatic rings are separated by either 2 or 3 bonds. In promethazine, two nitrogen atoms are corresponded to the common feature of the nitrogen atomic group. Because of that, all possible paths were examined for the distance determination and edge weighting when generating the reduced-graph representation of the chemical structure. Thus, the system computed the path length between each benzene ring and the nitrogen atom

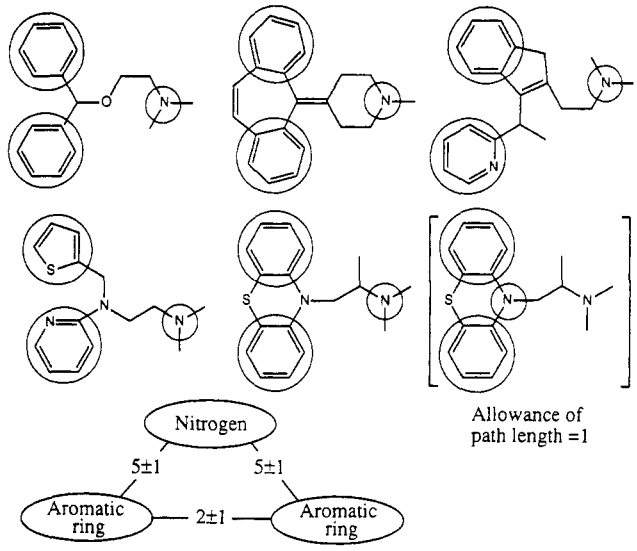


Figure 11. Result of the similar structural feature search for five antihistamines (diphenhydramine, methapyrilene, cyproheptadine, dimethindene, and promethazine).

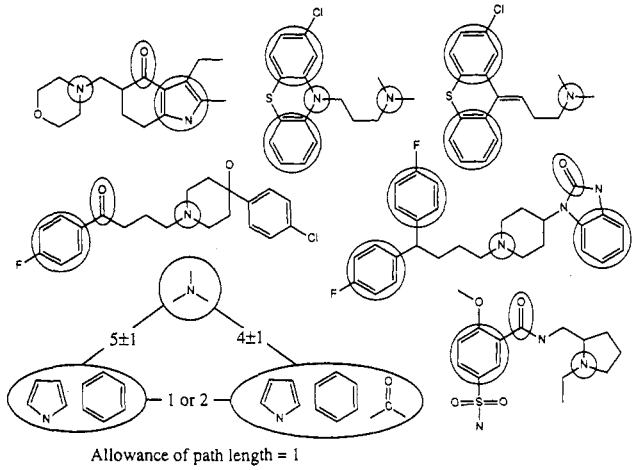


Figure 12. Result of the similar structural feature search for six antipsychotropics (molindone, chlorpromazine, chlorptothixine, haloperidol, pimozide, and sulpiride).

of the phenothiazine ring as well as between each benzene ring and the other nitrogen atom of dimethylamino group. This can be avoided by taking into account just the shortest path. The current system allows such choices in the search condition. The search result for these five antihistamines have the structural similarity that is expressed in the topological triangle based on the two aromatic rings and a tertiary nitrogen atom as shown in lower part of Figure 11.

As another illustration, a structurally more diverse set of six antipsychotropic agents (molindone, chlorpromazine, chlorptothixine, haloperidol, pimozide, and sulpiride) were selected for the search of their molecular similarities. The result is summarized in Figure 12. Olson et al.[14] suggested for the structure–activity relationships of these compounds that an electrostatic interaction and two π-electron interactions play the important roles for their bindings to the receptor site. Additional knowledge on the similarity for aromatic rings and carbonyl group was added into the knowledge file for the execution of this example, as these groups are regarded as equivalent atomic groups in the view of molecular interactions between π-electrons. The result obtained here shows that all of these structures have a tertiary nitrogen atom, an aromatic ring, and another aromatic ring (or a carbonyl group) that might be related to the π-electron interaction with their receptor. The particular topological positional relations with each other are also shown in Figure 12.

CONCLUSION

It follows from the above that the present system facilitates the rational exhaustive search of common or similar structural features for a set of chemical structures. To achieve this, the reduced-graph representation was devised for the description of chemical structure. The reduced representation was expressed as a node and edge-weighted graph in which each node is corresponded to a functional atomic group, and the weight of each edge is corresponded to the number of bonds between the functional atomic groups embedded in the original structure. This approach was computerized, and by the use of the knowledge file for the similarity between substructures one can identify not only common structural features but also similar structural features among chemical structures automatically. As mentioned in the beginning of this paper, evaluation and interpretation of molecular similarity depend on the problem to be handled. The use of knowledge file gives an object-oriented approach for such problems. The present approach provides a new direction of research works for the handling of molecular similarity by the computer.

ACKNOWLEDGMENT

We gratefully acknowledge the financial support of this work by Special Coordination Funds for Promoting Science and Technology, Science and Technology Agency of Japan. We also thank the reviewers for useful comments and suggestions which added to the clarity of presentation.

REFERENCES AND NOTES

(1) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, England, 1987; and references cited therein.
(2) Johnson, M. A.; Maggiora, G. M. *Concepts and Application of Molecular Similarity*; Wiley-Interscience: New York, 1990; and references cited therein.
(3) Armitage, J. E.; Crowe, J. E.; Evans, P. N.; Lynch, M. F. Documentation of Chemical Reactions by Computer Analysis of Structural Changes. *J. Chem. Doc.* **1967**, *7*, 209–215.
(4) Cone, M. M.; Venkataraghavan, R.; McLafferty, F. W. Molecular Structure Comparison Program for the Identification of Maximal Common Substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668–7671.
(5) Varkony, T. H.; Shiloach, Y.; Smith, D. H. Computer-Assisted Examination of Chemical Compounds for Structural Similarities. *J. Chem. Inf. Comput. Sci.* **1977**, *19*, 104–111.
(6) Takahashi, Y.; Satoh, Y.; Suzuki, H.; Sasaki, S. Recognition of Largest Common Structural Fragment among a Variety of Chemical Structures. *Anal. Sci.* **1987**, *3*, 23–28.
(7) Morita, K.; Oka, Y. Synthetic Drugs Containing Nitrogen, with Special Reference to Classification according to Functional Group Pair. *Kagaku, Zokan (Kyoto)* **1979**, *79*, 141–175.
(8) Golender, V. E.; Rozenblit, A. B. *Logical and Combinatorial Algorithms for Drug Design*; Research Studies Press: Letchworth, England, 1983.
(9) Carhart, R.; Smith, D. H.; Venkataraghavan, R. Atom pairs as Molecular Features in Structure–Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(10) (a) Sussenguth, E. H. A Graph-Theoretic Algorithm for Matching Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 36–43. (b) Figueras, J. Substructure Search by Set Reduction. *J. Chem. Doc.* **1972**, *12*, 237–244.
(11) Kuhl, F. S.; Crippen, G. M.; Friesen, D. K. A Combinatorial Algorithm for Calculating Ligand Binding. *J. Comput. Chem.* **1984**, *5*, 24–34.
(12) Bron, C.; Kerbosch, Finding All Cliques of an Undirected Graph. *J. Commun. ACM* **1973**, *16*, 575–577.
(13) Takahashi, Y.; Maeda, S.; Sasaki, S. Automated Recognition of Common Geometrical Patterns among a Variety of Three-Dimensional Molecular Structures. *Anal. Chim. Acta* **1987**, *200*, 363–377.
(14) Olson, G. L.; Chiang, C. E.; Berger, L. In *Dopamine Receptors*; Kaiser, C., Kebabian, J. W., Eds.; ACS Symposium Series 224; ACS: Washington, DC, 1983; Chapter 11, p 251.