M. *The Logic of Chemical Synthesis*; John Wiley and Sons, Inc.: New York, 1989.

(32) Kvasnička, V.; Kratochvíl, M.; Koča, J. *Mathematical Chemistry and Computer Aided Synthesis Planing*; Academia: Prague, 1987 (in Czech).

(33) Lindsay, R.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry*; McGraw-Hill: New York, 1980.

(34) Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. *J. Org. Chem.* **1981**, *46*, 1708.

(35) Munk, M. E.; Lind, R. J.; Clay, M. E. *Anal. Chim. Acta* **1986**, *184*, 1.

(36) The same notation is used in all papers on the synthon model in order to facilitate the reader's task.

Fast Drug-Receptor Mapping by Site-Directed Distances: A Novel Method of Predicting New Pharmacological Leads

ANDREW S. SMELLIE,[†] G. M. CRIPPEN,[*,‡] and W. G. RICHARDS[§]

BioCAD Corporation, 1091 North Shoreline Boulevard, Mountain View, California 94043, College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065, and Physical Chemistry Laboratory, South Parks Road, University of Oxford, Oxford, OX1 4AR England

Received December 3, 1990

The searching and characterization of large chemical databases has recently provoked much interest, particularly with respect to the question of whether any of the compounds in the database could serve as new leads to a compound of pharmacological interest. This paper introduces a fast and novel method of determining whether any of a given series of compounds are able, on geometrical grounds, to interact with an active site of interest. The C program written to implement the method is able to make a qualitative prediction for a given compound in about 1 s per structure (for drug-sized molecules), while still permitting the compound complete conformational freedom. However, the algorithm is sufficiently flexible to permit distance constraints to be placed on the molecules while docking. The test system studied was a family of Baker's triazines docking into the active site of dihydrofolate reductase (DHFR), as defined by a methotrexate/NADPH complex.

INTRODUCTION

(a) Overview. There exist in the literature numerous algorithms for performing docking of small molecules (usually drugs) into a larger binding site (usually a protein). This is the drug-receptor mapping problem, which breaks down into various classes. However for all classes of problem, the ultimate goal is always the same—can the binding conformer(s) of molecules be predicted for a given receptor site? The binding conformers are referred to as *binding modes*.

The following sections describe each class of problem. "Known" is taken to mean that the structure or conformer of the receptor and/or drug is known. "Unknown" means that the topology of the molecule is known, but the conformation is not.

Known Receptor (Site), Known Drug (Molecule). Here a conformation is assumed for the molecule, and predictions are made for the binding of this conformer only. Hopfinger proposed molecular shape analysis as a good criterion for mapping site and molecule[1] but this is not suitable for introducing conformational flexibility in the molecule.

Known Receptor (Site), Unknown Drug (Molecule). The scope of this paper falls into this category. Given a known receptor site and an unknown molecule, an algorithm is presented that will predict possible binding modes for the molecule. Complete conformational flexibility of the molecule is permitted in the site.

The drug-receptor mapping problem is represented as a bipartite graph in which the atoms of the site form one set of nodes (x) and the atoms of the molecule form a second set of nodes (o). By definition the edges in this bipartite graph link nodes from set x to nodes from set o. It can be seen from Figure 1 that the bipartite graph representation is analogous to the docking problem, where graph nodes are site or molecule atoms and graph edges are site–molecule interactions.

Binding modes can be extracted very rapidly from the bipartite graph by the docking graph approach,[2] which is described later in this paper.

Unknown Receptor (Site), Known Drug (Molecule). An example of this problem would be where binding data of several molecules are available, and it is known that they are binding at the same site. However, the structure of the site is not known.

The Voronoi binding site model of Crippen et al.[3,4] proposes a mathematical model of the binding site that is able to reproduce and predict binding data (in the form of equilibrium constants K_i, and hence free energy of binding from $\Delta G_{obs} = RT \ln K_i$). However, the abstract model site can bear little resemblance to the actual protein.

The clique graph algorithms of Kuhl[2] first introduced the docking graph concept used in this paper but without the fast filtering used here to greatly increase computational efficiency.

(b) Definitions. A few graph theory definitions will serve to clarify some of the points to follow, but for a more complete description of graph theory see ref 5.

An undirected *graph*, G, is a set, $u_g = \{u_i\}$, of *nodes* or *vertices* and a set $e_g = \{(u_i, v_i) | u_i, v_i \in u_g\}$ of unordered pairs of nodes interpreted as *edges*. A molecule can be thought of as a special type of graph where the atoms are the nodes and the bonds are the edges of the graph.

A *subgraph*, S, of G is composed of a subset of vertices and edges of G. Thus $u_s \subset u_g$ and $e_s \subset \{(u, v) | (u, v) \in e_g, u, v \in u_s\}$. A graph is *completely connected* if there is an edge between all pairs of vertices. For a completely connected subgraph of G, S_T, with N_T vertices, there are $N_T(N_T - 1)/2$ edges in set e_T where $e_T = \{(u, v)\ \forall\ u, v \in S_T\}$. The graph G* is a *clique* of G if it is a *maximal* completely connected subgraph of G.

A *bipartite* graph, B, is a graph where the nodes have been partitioned into two sets; $u = \{u_1, ..., u_N\}$, $v = \{v_1, ..., v_M\}$ and each edge involves exactly one vertex from set u and one vertex

[†] BioCAD Corporation.
[‡] University of Michigan.
[§] University of Oxford.

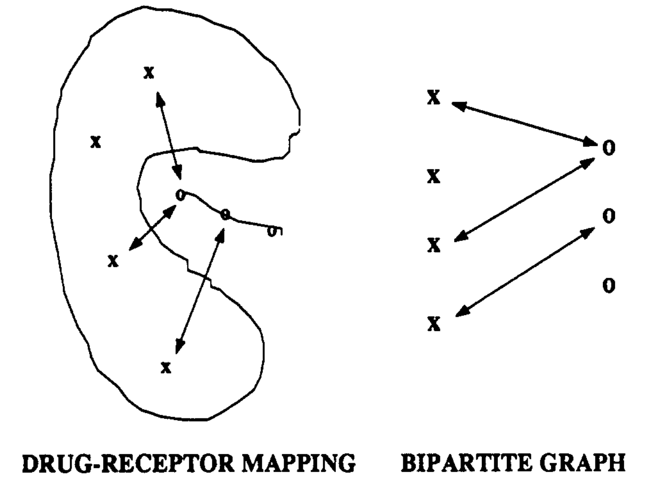

Figure 1. Equivalence of drug-receptor mapping and bipartite graphs.

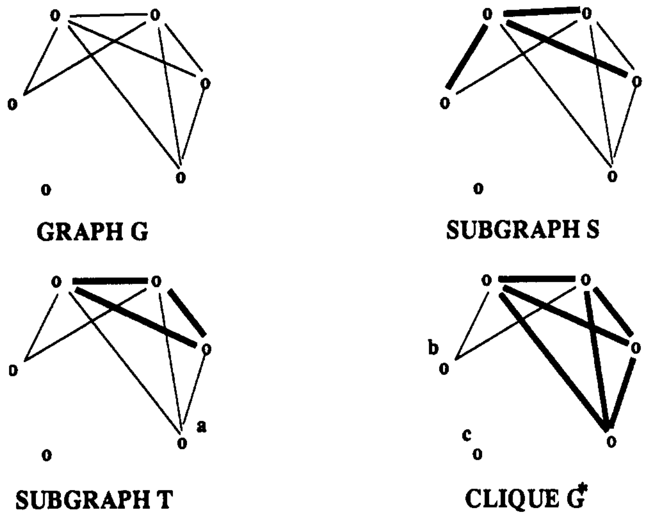

Figure 2. Graphs, subgraphs, and cliques.

from set v: $e_b = \{(u_i, v_j)|u_i \in u, v_j \in v\}$.

Figure 2 illustrates the relationship between graphs, subgraphs, and cliques. Graph S is a subgraph of G because it is composed entirely of nodes and edges that are contained in G. Graph T is a completely connected subgraph of G because all nodes in T are mutually connected. T is not a clique because it not a maximal completely connected subgraph—node **a** could be added to form a clique. G* is a clique of G because (i) it is a subgraph of G, (ii) it is completely connected, and (iii) adding nodes **b** or **c** would cause the subgraph to no longer be completely connected—hence it is maximal.

METHODS

The flow diagram in Figure 3 outlines the prediction of binding modes for flexible molecules into receptor sites. Stages 1 and 2 are virtually interactive for drug-sized molecules. Stage 3 is ultimately the rate-determining step and depends upon the number of site and molecule atoms identified at stage 1. Stage 4 is interactive and produces the binding modes for the molecule. The modes take the form of an ordered list of pairs of atoms, one from the site and one from the molecule, that interact *simultaneously*.

Stage 5 is not actually part of the prediction of binding modes, but is necessary to validate preferred binding modes. We chose to use the EMBED[6,7] algorithm to generate docked structures because it matched our requirements very closely, i.e., to generate structures subject to distance data arising from the requirement that atom pairs interact simultaneously.

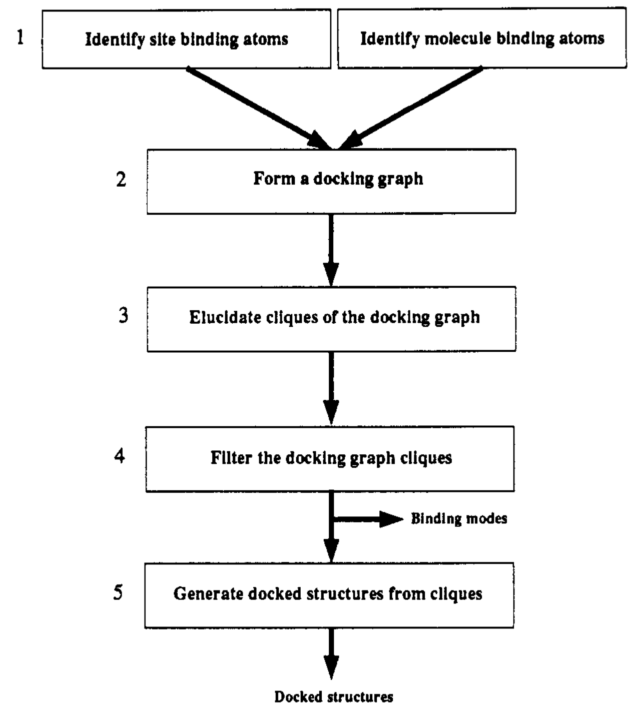
1
Identify site binding atoms
Identify molecule binding atoms
2
Form a docking graph
3
Elucidate cliques of the docking graph
4
Filter the docking graph cliques
Binding modes
5
Generate docked structures from cliques
Docked structures

Figure 3. Prediction of binding modes from bipartite graphs–algorithm.

Table I. Truncated DHFR Sites Derived with Respect to Bound Methotrexate

cutoff (Å)	total no. of atoms	total no. of H-bond atoms[a]
3.5	20	16
4.0	51	27
4.5	72	31
5.0	96	42
5.5	128	56
6.0	150	67

[a] Heavy atoms capable of participating in H-bonding.

The model system used in this study was the active site of *E. coli* dihydrofolate reductase (DHFR) with bound methotrexate/NADPH complex.[8] The test drug molecules were a family of Baker's triazines whose structure–activity properties have been extensively studied experimentally[9,10] and theoretically.[1]

(a) Identification of Molecule and Site Binding Atoms. The number of nodes in the docking graph is directly proportional to the number of molecule and site atoms considered important in the docking process. Obviously the fewer atoms considered, the more rapid the algorithm. For this reason only those atoms capable of H-bonding were considered. For the DHFR system this is not a damaging assumption because it is generally thought that the methotrexate/NADPH complex is predominantly stabilized by electrostatic interactions with the site.

The total number of site atoms was drastically reduced by "skinning" the binding site pocket of the DHFR/methotrexate complex. By this we mean including all atoms of DHFR that lie within a certain cutoff distance of methotrexate in the crystal structure of the complex. Hence this includes only those atoms of DHFR that interact directly with methotrexate. The assumption is made that the triazine binds to the same site as methotrexate and that the site is rigid during docking. Thus the problem has been reduced to mapping the atoms of the molecule onto relatively few fixed points in space. These points are the site points that form the "skin" of the site pocket. Table I summarizes the truncated DHFR sites obtained for various cutoffs to bound methotrexate.

A cutoff of 3.5 Å was selected. This minimized the number of site points to be considered but still provided a mapping site rich in potential H-bond donor/acceptor atoms.

The molecules in this study were representative of general drug-sized molecules and have approximately 5–8 atoms ca-

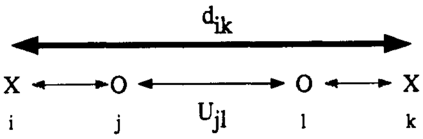

Figure 4. Geometrical limiting case for molecule upper bounds, U_{jl}.

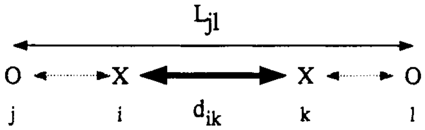

Figure 5. Geometrical limiting case for molecule lower bounds, L_{jl}.

pable of acting as a hydrogen-bond donor or acceptor. No special treatment of the molecules was required to reduce the number of points considered in the mapping study.

(b) Forming the Docking Graph. The lists of site and molecule atoms obtained after stage (a) are comprised only of atoms capable of acting as a hydrogen-bond donor or acceptor. These atoms form a bipartite graph in which the site atoms form one set of vertices (x) and the molecule atoms form another set of vertices (o). Edges between these sets represent interactions between the site and the molecule.

A *docking graph* is a convenient method of reducing this bipartite graph to a normal graph from which the binding modes can be extracted. There is a node in the docking graph for every ordered pair of points from sets x and o. For N_x vertices in set x and N_o vertices in set o, there are N_xN_o vertices in the docking graph. Let g_{ij} represent a node in the docking graph formed from the *i*th node of set x (the site) and the *j*th node of set o (the molecule).

For any two nodes of the docking graph, g_{ij} and g_{kl}, there is an edge between these nodes when

$$d_{ik} \leq U_{jl} + \delta \tag{1}$$

$$d_{ik} \geq L_{jl} - \delta \tag{2}$$

Equations 1 and 2 are the necessary and sufficient conditions for two *simultaneous* interactions between the molecule and the site. Node g_{ij} represents the interaction of the *i*th node of set x and the *j*th node of set o, and node g_{kl} represents the interaction of the *k*th node of set x and the *l*th node of set o.

Equations 1 and 2 also describe the limiting geometrical cases for the two simultaneous interactions as shown in Figures 4 and 5. The site is assumed rigid, and site distances are described by d_{ik}. The molecule is permitted flexibility, and molecule distances are described by an upper bound U_{jl} and a lower bound L_{jl}. Note that complete conformational flexibility can be represented by generous upper and lower bounds on the molecule. On the other hand, distance constraints can be imposed on the molecule and rigid docking can be simulated by setting $U_{jl} = L_{jl}$.

In this study, complete conformational flexibility was permitted by setting lower molecule bounds equal to the sum of van der Waals radii and upper molecule bounds to an arbitrarily large number. These bounds were then triangle (TRNGL) smoothed[6] to deduce geometrically consistent distance bounds.

The parameter δ is a flexibility parameter for the site. A completely rigid site has $\delta = 0.0$. For this study, a value of 6.0 Å was used. The geometrical limiting case of Figures 1 and 2, when applied to H-bonding situations, imply that each edge in the docking graph represents two simultaneous H-bonds. Each H-bond has an approximate length of 3.0 Å. Thus two H-bonds in the limiting geometrical cases require $\delta = 6.0$ Å.

For the special case in this study, where only electrostatic interactions are considered, there is another condition for placing a node, and hence an edge, in the docking graph. The site and molecule atoms involved in each node must be capable of participating in H-bonding. All atoms of the molecule and site were divided into one of four classes: (a) H-bond donor/acceptor type (e.g., hydroxyl oxygen), (b) donor type (e.g., protonated hydroxyl oxygen), (c) acceptor type (e.g., carbonyl oxygen), and (d) neither donor nor acceptor. Each node in the docking graph represents a site/molecule–atom interaction. Thus each node must be formed from two atoms of classes (a)-(a), (a)-(b), (a)-(c), or (b)-(c). Any other classification of atoms is disallowed, and this node is not permitted in the docking graph.

Note that class (b) above is not observed in the current systems of interest, but is included for completeness.

(c) Elucidate Cliques of the Docking Graph. Cliques, G*, of the docking graph are maximal completely connected subgraphs of the docking graph. Hence all pairs of nodes in G* are mutually connected, and the addition or deletion of any node will cause the subgraph to be no longer completely connected.

Consider the graphs illustrated in Figure 2, and assume that they are docking graphs. Each node represents a mapping of one atom of the site and one atom of the molecule, and each edge represents a simultaneous interaction of two pairs of site/molecule atoms. Since cliques are a collection of completely connected nodes, then the site/molecule interactions expressed by each node must, by definition, be able to occur *simultaneously*. These are the binding modes—a collection of simultaneous molecule and site interactions.

There are many clique-finding algorithms in the literature. This study chose the algorithm of Bron and Kerbosch[11]—with some minor modifications.

A single isolated node, mathematically speaking, is a clique. This is clearly not a binding mode because it only represents a single interaction between the site and the molecule. Thus the algorithm in ref 11 was modified to generate cliques with at least five nodes. This means that the binding modes generated consist of at least five simultaneous interactions.

(d) Filtering the Docking Graph Cliques. Clique finding is an NP-complete problem, and hence the complexity and run-time of extracting docking graph cliques is indeterminate but extremely large.[5] Several very rapid filters were applied to each clique as it was generated to drastically reduce the total number of cliques produced.

The first has already been described—no clique with less than five nodes was considered. This cutoff may seem unusually large because it forces each molecule to make five or more simultaneous interactions with the site. As can be seen from the results later, this still returned all the important binding modes of the test triazines. Also this parameter is under user control and can be adjusted to find binding modes with fewer contacts at the expense of generating more cliques.

The second filter used information about the nature of the interaction being studied. H-bonding is generally a one-to-one mapping, though bifurcated H-bonds are fairly common. However since the clique-finding process is a geometrical rather than a chemical problem, many cliques are generated with multiple edges to a single node as shown in Figure 6. Here the cliques are shown expanded into bipartite graphs with site and molecule atoms represented by 'X' and 'O', respectively. Each edge shown represents a single node in the docking graph.

A clique is rejected if its *edge average* >2.0. The edge average is defined as

$$\text{edge average} = \max\left(\frac{N_{c_i}}{N_{x_i}}, \frac{N_{c_i}}{N_{o_i}}\right) \tag{3}$$

where N_{c_i} = total number of edges in clique *i*, N_{x_i} = total number of site points in clique *i*, and N_{o_i} = total number of

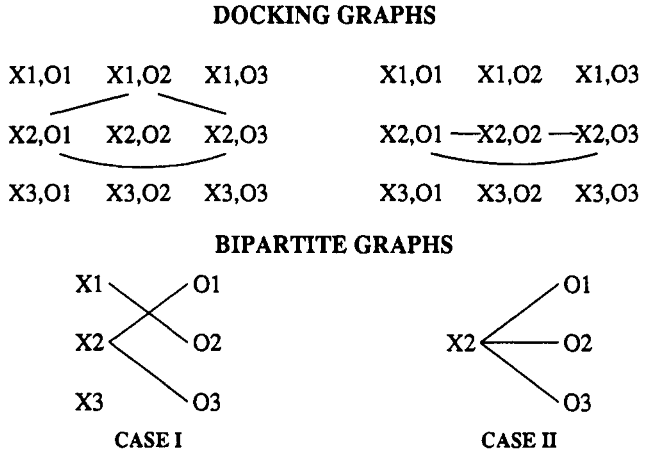

Figure 6. Clique filtering by multiple edges.

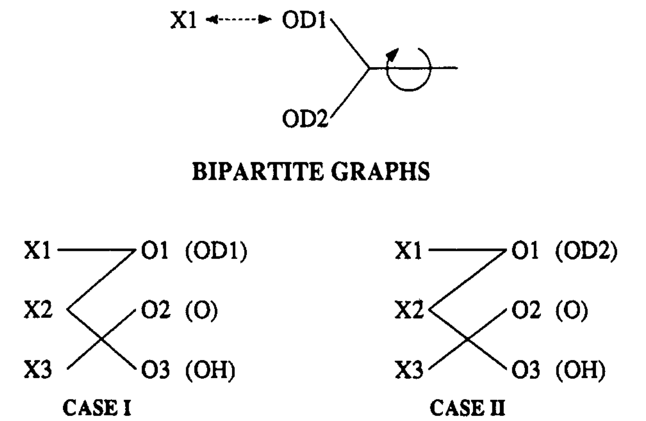

Figure 7. Clique filtering by group degeneracy.

molecule points in clique *i*. This discriminates against cliques with a large number of incident edges per node. A value of 2.0 was used to allow for bifurcated H-bonds. Once again this edge-average criteria is under user control and can be used to reduce the total number of cliques generated. From the bipartite graphs in Figure 6: case I has an edge average of 1.5 and is retained but case II has an edge average of 3.0 and is rejected.

The third and final filter again uses information about the chemistry of each clique generated. Often one clique is degenerate with respect to another. This arises from site or molecule atoms that are degenerate with respect to some internal motion. Consider the bipartite graphs of Figure 7 (with cliques omitted for clarity). Although these are different cliques, they refer to the same interaction because a simple rotation of the carboxylic bond interconverts between the two cliques by switching the carboxylic oxygens. Thus, all such degenerate interactions are pruned out of the final clique lists.

All cliques that pass the filters are ranked in order of weight. A clique is a collection of nodes each of which are in turn a pairing of site/molecule atoms. Oxygen atoms are a given an arbitrary weighting of 3, nitrogen atoms a weighting of 2, and sulfur atoms a weighting of 1. Each clique, therefore, has an overall weight where, for site atoms x and molecule atoms o:

$$G_w^* = \sum_i^{\text{nodes}} (w_i^x + w_i^o) \qquad (4)$$

This ensures that the highest ranked cliques have a larger number of nodes (and hence a large number of simultaneous site/molecule interactions) and/or contain atoms that form particularly strong H-bonds.

(e) Generate Docked Structures from Cliques. The fully filtered cliques obtained after stage 4 above represent the purely geometrical binding modes of a given flexible molecule into an active site. Actual docked structures are not necessary to make qualitative statements as to the binding of the molecules, but are important for validation purposes.

A given clique is a list of nodes in the docking graph, which are in turn a list of simultaneous interactions between the molecule and the site. Hence to generate a docked structure the following information is available at this stage:

A full description of the site either in terms of atomic coordinates or interatomic distances because the site is assumed to be rigid.

A list of upper and lower bounds on the molecules, assuming complete flexibility, obtained from distance geometry triangle smoothing.

A list of site–molecule interactions. The focus of this study has been H-bonding so suitable lower and upper bounds on these interactions are 2.80 and 3.30 Å, respectively.

All the above information was submitted to the EMBED[6] algorithm using the CONSTRICTOR[7] program.

The generation of binding modes used an active site truncated to within 3.5 Å of bound methotrexate. For embedding purposes, a larger site was used to include implicit VDW constraints from the surrounding environment. Thus, the active site for embedding was truncated to within 7.0 Å of bound methotrexate.

It was found that the distance geometry structures were sufficient to fully describe the docked structures while satisfying the imposed distance constraints from the cliques.

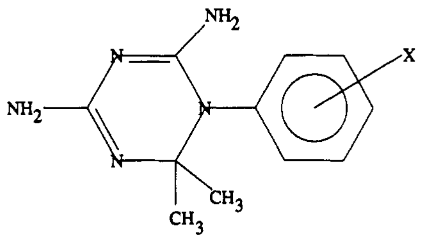

Figure 8. General structural formula of the Baker's triazines.

Table II. Summary of Results for the Baker's Triazines

mol	X	no. of nodes in docking graph	no. of edges in docking graph	total no. of cliques	obs. activity
1	-H	34	17	6	6.92
2	3-Cl,4-OCH$_2$Ph	40	20	37	7.52
3	3-COCH$_3$	40	20	37	6.79
4	4-C(OH)Cl(Ph)	40	20	36	6.45
5	3-COClCH$_2$	40	20	37	6.21
6	3-OCH$_3$	40	20	19	6.17
7	2-OCH$_3$	40	20	10	3.68
8	4-CH$_2$Ph	34	17	6	8.05
			mean time	CPU (s)	7.60

RESULTS

(a) Baker's Triazines—A Docking Study with DHFR. Figure 8 shows a family of eight Baker's triazines studied, with supplemental information shown concerning the generation of cliques. Even with heavy pruning from the filters, several cliques are produced per molecule. It is not feasible here to examine each clique in detail. Instead only the top few cliques are discussed. Recall that the cliques are ranked in order of weight with the highest ranked cliques containing many nodes and/or particularly strong interactions. There are several significant points to be made concerning the cliques found for these molecules:

The principal interaction in this study has been H-bonding. Molecules **1** and **8** from Table II have side

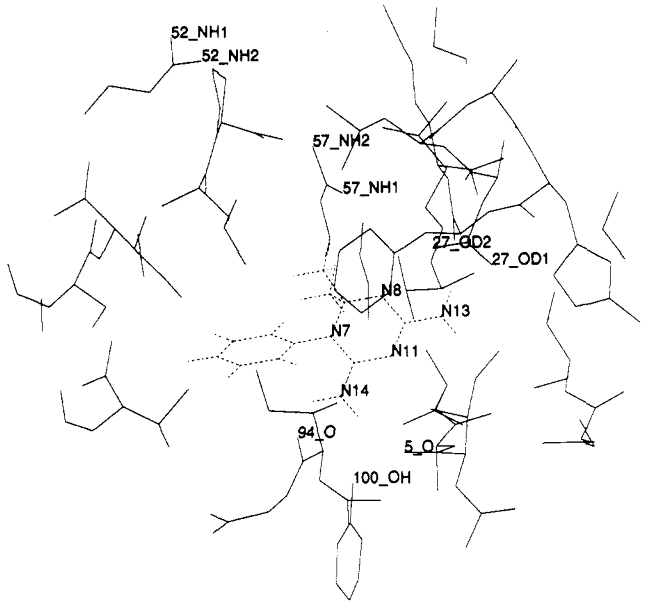

Figure 9. Sample docked structure of highest ranked clique for X = H.

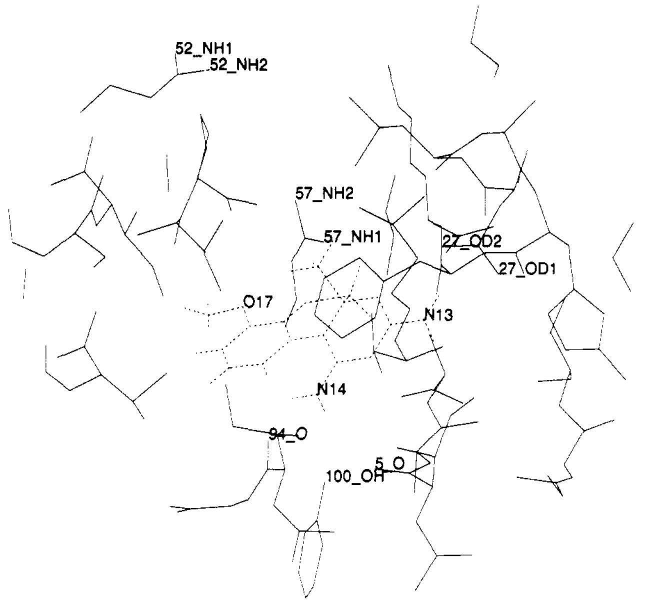

Figure 11. Sample docked structure of highest ranked clique for X = 3-OCH_3.

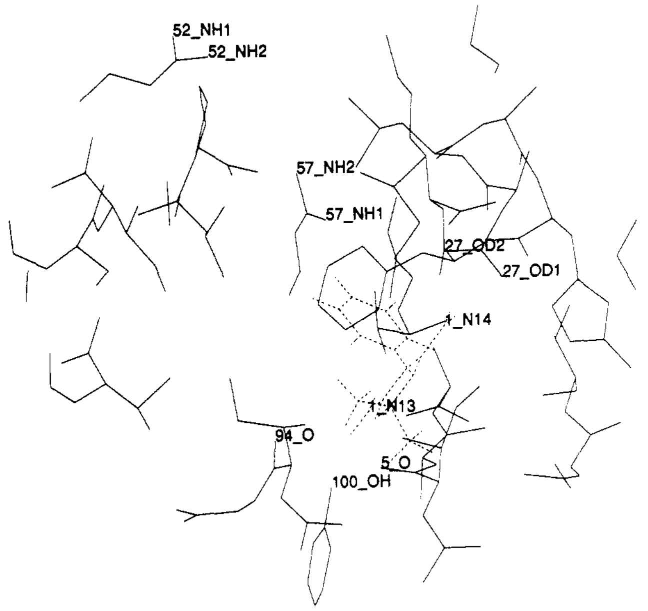

Figure 10. Sample docked structure of 3rd highest ranked clique for X = H.

chains that are incapable of such H-bonding. These side chains cannot enhance binding, and thus these molecules have the same number of cliques and in general fewer cliques than have the other molecules (whose side chains do contain polar atoms). The cliques for molecules **1** and **8** arise from the interaction of the amide groups of the 2,4-diaminotriazine ring system with polar atoms of the active site.

Concentrating on molecule **1**, X = H: To verify the six binding modes predicted for this molecule, the distance geometry EMBED algorithm was used to generate sample structures for each of the binding modes. Distance constraints were read from a file and used to constrain the corresponding atoms from the site and molecule to lie between 2.80 and 3.30 Å (i.e., an idealized H-bond length). The results for two of these binding modes are shown in Figures 9 and 10. Figure 9 shows the clique with highest ranking. This is a legitimate binding mode because EMBED reported no VDW clashes and complete satisfaction of the distance constraints (to within 0.5 Å). Note the *potential* for simultaneous H-bonding involving site/molecule atom pairs 94_O/N14, 100_OH/N14, 27_OD2/N13, and 27_OD1/N13. Figure 10, however, shows the clique with the 3rd highest ranking. It shows the same potential for H-bonding because it is simply the above clique with atoms N13 and N14 interchanged. However the inclusion of a fuller site during embedding clearly shows that this is not a legitimate binding mode—the phenyl ring is bent out of position because it cannot be accommodated in the site. This raises an important point concerning clique finding—binding modes are predicted on the basis of *geometrically allowed* mappings on the molecule onto the site. Regions of the site that intrude into the space of the molecule in any binding are neglected. To obtain true docked structures from these binding modes requires that each mode be embedded to produce or reject docked structures. This is not a major drawback because the cliques are ranked in an order that typically favors true binding modes over false ones (see below).

Molecules **6** (X = 3-OCH_3) and **7** (X = 2-OCH_3) differ only in the position of substitution by the methoxy group in the phenyl ring. There is, however, a large difference in the total *number* of cliques found for each molecule. Since cliques are expressions of geometrically allowed binding modes there must be simple *geometrical* reasons for the differences in clique number. To test this, the highest ranked clique for each molecule was subjected to the EMBED algorithm to produce docked structures. Figure 11 shows a sample docked structure of X = 3-OCH_3, and Figure 12 shows a sample docked structure of X = 2-OCH_3. Note in Figure 11 that substitution at the 3-position of the phenyl ring by the methoxy group permits a conformation that allows the *simultaneous* interaction of atoms 57_NH1/O17, 57_NH2/O17, 94_0/N14, 27_OD1/N13, etc. However substitution at the 2-position does not allow all the above interactions simultaneously. This is reflected in fewer cliques found for X = 2-OCH_3 than for X = 3-OCH_3. The *total number* of cliques found for any molecule can be used as a qualitative indication of the number of binding modes available to a molecule in a known active site.

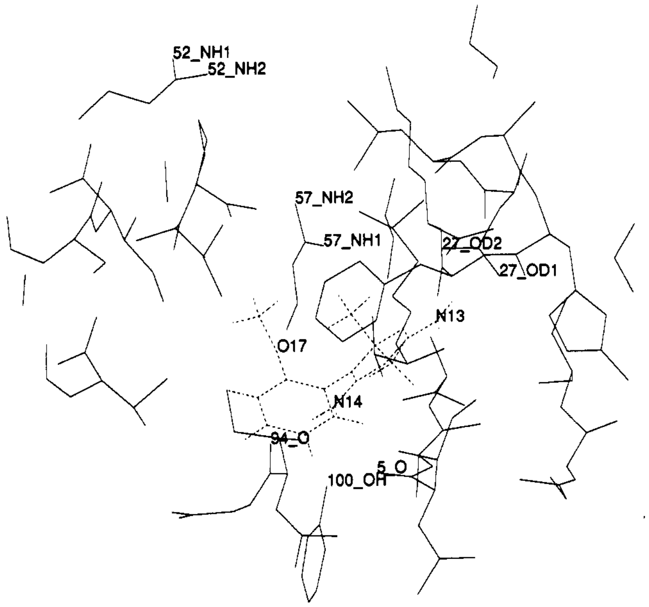

Figure 12. Sample docked structure of highest ranked clique for X = 2-OCH_3.

The following stages of the algorithm are the most time consuming: (a) the derivation of upper and lower bounds on interatomic distances for each molecule, (b) the elucidation of all cliques of the docking graph, and (c) the filtering of each of the cliques. For these eight drug-sized molecules in the truncated active site, the rate-determining step was the derivation of upper and lower bounds on interatomic distances. In any case, no molecule took more than about 1.5 s on a SUN SPARCstation 1+, with an average for each molecule of about 0.95 s. It is hoped that with such speed, especially on larger computers, that algorithm could be used as a fast first pass on 3-D database searches.

In the current implementation, the overriding time is spent in the embedding stage. Every generated clique could, in principle, be a binding mode. Currently this is verified by EMBED, and in the current test set approximately 20% of the passed cliques successfully embedded. It is anticipated in a real database search that faster screening methods are required, and this could form the basis of further exploration.

CONCLUSIONS

An algorithm has been prsented that elucidates all geometrically allowed binding modes of an arbitrary molecule into a well-characterized site. Complete conformational flexibility is permitted for each molecule, with an average run-time for each molecule of less than 1 s on a SUN SPARCstation 1+. The expected run-time is independent of conformational flexibility, since such flexibility is incorporated into the distance bounds.

Important factors in the rapid run-time were the pruning of the active site by 'skinning' and that the only interactions considered were electrostatic in nature. This resulted in few nodes and edges in the molecular docking graph. It may be possible to extend this method to nonbonded interactions by including extra nodes in the docking graph. For example, a phenyl ring or methyl group could be represented by a single dummy atom and edges could be permitted in the docking graph for certain types of nonbonded interactions.

Although site rigidity is implied in the derivation of the docking graph (see eqs 1 and 2), the flexibility parameter, δ, can be interpreted as flexibility in the active site. In this study a value of δ = 6.0 Å was used to allow for the geometrical limiting cases of the formation of two collinear H-bonds (see Figures 1 and 2).

It should be noted that as the flexibility parameter increases, the number of cliques generated will increase. It was found in the current test set that setting $d = 0.0$ resulted in no cliques found due to the inability of simultaneous H-bonds being formed.

The weighting of docking graph edges could be used to rank binding modes. Currently a scheme is used where, for each atom in an interaction, oxygen atoms are given a weight of three, nitrogen atoms a weight of two, and sulfur atoms a weight of one. This simple scheme is sufficient to differentiate electrostatic interactions by ranking each binding mode in order of weight: The higher the weighting, the greater the proportion of energetically favorable O-O, O-N, etc. interactions. For the inclusion of nonbonded interactions a more complex scheme would be required that takes into account nonelectrostatic-type interactions.

Although not pursued in this study, the technique permits a full or constrained search of the conformational space. The docking of each molecule is constructed by using the upper and lower bounds on interatomic distances. These bounds may be set to permit complete flexibility, or constraints may be imposed to limit the search. In particular, work is on-going to apply this algorithm to pharmacophore mapping, where the proposed pharmacophore is used to impose distance constraints on the molecules of interest.

It is hoped that this algorithm may be used as a fast first-pass in a 3-D database search in two important types of problem:

> For a known active site, the database may be scanned for molecules that can interact with the site. Here the number and quality of binding modes (cliques) are the most important factors in predicting likely binding of the molecule to the site.
>
> For a proposed pharmacophore, the database can be scanned for molecules that could interact with the pharmacophore. Here the pharmacophore imposes distance constraints upon the molecules, from which the number and quality of binding modes are again the most important factors in predicting binding.

In either case, complete conformational flexibility is permitted for each molecule or distance constraints can be imposed. With sufficient pruning and filtering of the binding modes (cliques) found per molecule, very rapid speeds could be obtained for any given database.

ACKNOWLEDGMENT

This work was supported through a CASE award with ICI Pharmaceuticals, Alderly Park, Alderly Edge, Cheshire, England, by grants from the National Institutes of Health (GM-37123) and the National Science Foundation (DMB-8705006), and by the University of Michigan Program in Protein Structure and Design. The CONSTRICTOR program is the property of Oxford Molecular Ltd, Terrapin House, South Parks Road, Oxford, England.

Registry No. **1**, 4022-58-6; **2**, 50574-67-9; **3**, 70579-34-9; **4**, 134311-41-4; **5**, 10161-71-4; **6**, 17711-73-8; **7**, 20285-47-6; **8**, 36076-92-3; DHFR, 9002-03-3.

REFERENCES AND NOTES

(1) Hopfinger, A. *J. Am. Chem. Soc.* **1980**, *102*, 7196–7206.
(2) Kuhl, F.; Crippen, G. M.; Friesen, D. *J. Comput. Chem.* **1984**, *5*, 24–34.
(3) Boulu, L.; Crippen, G. M. *J. Comput. Chem.* **1989**, *10* (5), 673–682.
(4) Boulu, L.; Crippen, G. M. Voronoi Receptor Site Models. In *Computer-Assisted Modelling of Receptor-Ligand Interactions: Theoretical Aspects and Applications to Drug Design*; Alan R. Liss Inc.: New York, 1989; pp 267–277.
(5) Gibbons, A. *Algorithmic Graph Theory*. Cambridge University Press: Cambridge, 1988.

(6) Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformations. In *Chemometrics Research Studies Series*; Bawden, D., Ed.; Research Studies Press (Wiley): New York, 1988.

(7) Smellie, A. *CONSTRICTOR*. Oxford Molecular Ltd.: Terrapin House, University Science Area, South Parks Road, Oxford, England, 1989.

(8) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. *J. Biol. Chem.* **1982**, *257*, 13650.

(9) Silipo, C.; Hansch, C. *J. Am. Chem. Soc.* **1975**, *97*, 6849.

(10) Dietrich, S.; Smith, R. N.; Fukunaga, M.; Olney, M.; Hansch, C. *Arch. Biochem. Biophys.* **1979**, *194*, 600.

(11) Bron, C.; Kerbosch, J. *Commun. ACM* **1973**, *16* (9), 575-577.

Computer-Assisted Infrared Identification of Vapor-Phase Mixture Components

BARRY WYTHOFF,[*,†,‡] XIAO HONG-KUI,[§] STEVEN P. LEVINE,[§] and STERLING A. TOMELLINI[†]

Department of Chemistry, University of New Hampshire, Durham, New Hampshire 03824, and School of Public Health, The University of Michigan, Ann Arbor, Michigan 48109

Received January 3, 1991

The IRBASE/MIXIR system was originally tested on interpretation of infrared spectra of condensed-phase mixtures. The system has now been adapted to allow interpretation of vapor-phase mixture spectra. The dynamic interpretation capabilities of the system have been expanded to allow runtime manipulation of complete peak lists, allowing generation of the optimum spectral description for the interpretation at hand. The modifications to the system are described, along with the results of testing on actual mixtures of varying complexity.

INTRODUCTION

Infrared analysis of organic vapors has many potential applications, including on-site measurement of toxic compounds at hazardous waste sites and in the workplace, and analysis of unresolved effluents from GC–FTIR experiments. Efforts at computer-assisted identification of components of mixtures have been largely directed at condensed-phase systems. The analysis may be performed in either of two ways: through a hyphenated technique involving a separation prior to identification, e.g., GC–FTIR, or by analysis of the intact mixture.

Using a separation method prior to identification has the advantage that the data analysis is much simpler and ideally involves a spectral library search using some similarity metric for each of the separated components. Important characteristics include robustness in the presence of noise and experimental variation, and analysis times, particularly when real-time analysis of capillary GC–FTIR spectra is required. One-at-a-time library comparison methods using linear neural networks,[1] boolean logic,[2,3] cluster analysis,[4] parallel processing,[5] and information theory[6] have been demonstrated. Principal components analysis has been used to select library subsets for more detailed analysis,[7] and orthonormalized spectral libraries have been evaluated.[8] Interferometric search has been performed using only 100 data points from the interferograms.[9] The use of a second, coupled spectrometer subsequent to GC separation (GC–FTIR MS) has been demonstrated on a 30-component mixture.[10] The combination of independent sources of spectral information was found to aid the analysis.

A drawback of most library search methods is that they perform poorly on mixtures. There is no guarantee that separation will be complete when using a hyphenated separation–identification scheme. In addition, many compounds can undergo thermal degradation during gas chromatographic analysis or may adsorb irreversibly to liquid chromatographic column packings. Finally, the cost and complexity of instrumentation obviously increase with hyphenated techniques.

Infrared spectra of intact mixtures have been studied mathematically by principal components analysis to determine the number of components[11] and, in the Fourier domain by factor analysis, to quantitate mixtures where the component identities were known.[12] A comparison of four multivariate methods was performed on quantitative analysis of mixtures with known component identities.[13]

Quantitative analysis of vapor-phase mixtures where the component identities are not known has been explored by using least-squares fitting (LSF) techniques.[14,15] Qualitative identification of vapor-phase mixture components has been recently reported by using iterative least-squares fitting techniques (ILSF).[16] An intriguing possibility is the use of a knowledge-based system to reduce the number of components fed into an LSF quantitation program. This should greatly reduce the workload required for the LSF calculations and provide more accurate quantitative results as well.

The IRBASE/MIXIR system is a knowledge-based system developed to identify the likely components of mixtures from infrared spectral data. The original work concerned the development of a compound-specific automated rule generator[17] and a knowledge-based system to manipulate these rules.[18] The experimental test data were condensed-phase mixtures; however, most of the interpretation algorithms and logic are applicable to vapor-phase samples as well. A previous attempt at adapting a condensed-phase expert system for vapor-phase analysis has been made.[19] It was concluded in that research that a peak-based expert system was "not appropriate" to vapor-phase analysis. It was recognized from the outset, therefore, that these data would present a difficult challenge. The goal of this research was to attempt to define the limits of knowledge-based systems for interpreting peak-based information from infrared spectra of mixtures.

While adapting these programs for vapor-phase analysis, many improvements have been made that can also be used for condensed-phase analysis. This paper describes these modifications and enhancements, which represent another phase in the continuing evolution of the MIXIR/IRBASE system.

EXPERIMENTAL SECTION

The vapor-phase infrared spectra used for this study were acquired at a nominal 0.3-cm^{-1} resolution and were transformed to a 2-cm^{-1} resolution representation for this work. The

[*] Corresponding author.

[†] University of New Hampshire.

[‡] Present address: Division of Inorganic Analysis, Center for Analytical Chemistry, National Institute of Standards and Technology, Gaithersburg, MD 20899.

[§] The University of Michigan.

0095-2338/91/1631-0392$02.50/0