A Method for Visualizing Recurrent Topological Substructures in Sets of Active Molecules

Robert P. Sheridan*,[†] and Michael D. Miller[‡]

Department of Molecular Design and Diversity, RY50S-100, Merck Research Laboratories,
Rahway, New Jersey 07065, and Department of Molecular Design and Diversity,
WP42-3, Merck Research Laboratories, West Point, Pennsylvania 19486

Received April 15, 1998

We present a method for detecting meaningful common topological substructures among sets of active compounds. A clique-based subgraph detection method is used to find the highest scoring common substructure (HSCS) for each pair of molecules. Only those HSCSs are kept that have a much larger score than would be expected by chance for a randomly selected pair of molecules of the same size. Information on these significant HSCSs is visualized in two complementary ways. Individual HSCSs can be displayed as pairs of molecules with the atoms in the HSCS highlighted. The HSCSs are presented in order of decreasing statistical significance. Alternatively, individual molecules can be displayed such that each atom is labeled with the number of significant HSCSs in which it has appeared. These are presented in order of decreasing maximum number per molecule. Browsing these data gives an impression of what parts of molecules are conserved among actives. The molecules can be simplified by deleting parts that are not conserved. We show three examples taken from the MDDR database. One example highlights common substructures among diverse CCK antagonists. Two examples point out substructures common between actives on different receptors: benzodiazepine agonists vs CCK antagonists and antidepressants vs antihistamines.

INTRODUCTION

Given a set of molecules active on a particular receptor, one often wants to ask “What parts of these molecules are in common?”. The presumption is that common parts of active molecules may be important for binding to a particular receptor (as a 2D “pharmacophore”) or act as a scaffold to hold interacting groups in a particular orientation. In 1988 the term “privileged structure” was used by Evans et al.[1] to mean a substructure that conferred activity to two or more different receptors. The implication was that the privileged structure provides the scaffold and that the substitutions on it provide the specificity for a particular receptor.

In setting up experiments to look for meaningful common substructures it is important to arrange things such that it would be statistically surprising to find commonality. If one looks at a set of very close analogues active on a single receptor, it would be inevitable any two of them would have common substructures that are very large compared to the size of the individual molecules. Such substructures would not be very biologically informative. On the other hand, if one selects a diverse subset of actives, it would be much less likely that the actives in the subset would have large substructures in common and much more likely that the common substructures would be informative. Similarly, if one is comparing sets of molecules active on two different receptors, one must be sure that the receptors do not share many identical molecules as ligands.

In this paper we present a method for detecting and presenting common substructures for pairs of compounds either with the same or different biological activity. We will deal here only with topological substructures and not 3D substructures for two reasons. First, most chemists think in the language of chemical drawings. Second, formulating the problem in 3D would require knowledge of receptor-bound conformations of many actives—knowledge we generally do not have.

We will begin by estimating the probability that two unrelated compounds will have common substructures of a given size; this provides crucial information as to whether a common substructure is meaningful. Given this information, we will show one example in which we find common substructures among a set of molecules with a single activity and two examples in which we find common substructures among sets of compounds with different activities.

METHODS

Common Substructures. Common substructures are the sets of atoms that two molecules A and B have in common. For the purposes of this work, they will be expressed as cliques of non-hydrogen atoms. These are lists of atom correspondences with the following properties:

1. Corresponding atoms in A and B must have the same “atom type”.

2. The topological distances between the substructure atoms in A must be the same as the distances between the corresponding atoms in B.

Topological distance between atoms is the distance in bonds along the shortest path connecting the atoms. “Atom type” may be defined in a number of ways, as will be discussed below.

* To whom correspondence should be addressed.

[†] Merck Research Laboratories, Rahway, NJ.

[‡] Merck Research Laboratories, West Point, PA.

S0095-2338(98)00044-4 CCC: $15.00
Published on Web 08/22/1998

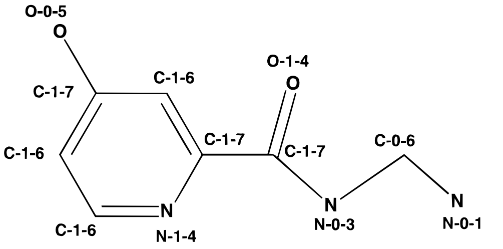

Figure 1. A sample molecule with atoms typed. The type includes the element, the number of π electrons, and the physiochemical type (a number from 1 to 7). The physiochemical types are 1. cation, 2. anion, 3. neutral H-bond donor, 4. neutral H-bond acceptor, 5. polar, 6. hydrophobic, 7. other.

At the core of this work is a new method for detecting common substructures. It is a modification of the ones developed for finding 3D cliques.[2,3] See the Appendix for details. A useful feature of clique-based algorithms is that the common substructures may be discontinuous.

Atom Types. The definition of atom type is critical in any common substructure method. After some experimentation we found a definition of atom type that is pharmacologically relevant and that is a reasonable compromise between generality and specificity. We use this atom typing throughout, but the methods below can be used with any other definition.

The type is a string containing element, number of π electrons, and physiochemical type at near-neutral pH. One exception is that we equivalence all halogens to the "element" Hal. The physiochemical type of an atom is one of seven: 1 = cation; 2 = anion; 3 = neutral H-bond donor; 4 = neutral H-bond acceptor; 5 = polar atom (atoms which are both donors and acceptors, e.g., hydroxy oxygen or either donor or acceptor via tautomerization, e.g., the nitrogens of imidazole); 6 = hydrophobe; 7 = other (nonpolar atoms in a polar environment or polar atoms that cannot accept or donate H-bonds). The automated method to assign these types has been described.[4]

Figure 1 shows an example of a typed molecule. Note that since the number of π electrons is included in the atom type, the bond orders in a substructure can be ignored.

Scores for Substructures. Normally one is interested in the "best" substructure per pair of molecules. Most common substructure algorithms find the "maximum common substructure" (MCS), a continuous bonded substructure with the largest number of atoms that is found in both molecules. When discontinuous substructures are allowed, evaluating a substructure becomes more complicated than just counting the number of atoms. We use

$$\text{Score} = \text{Size} - p^*(\text{Nfrag} - 1)$$

where Size is the number of atoms in the substructure, Nfrag is the number of discontinuous pieces into which the substructure is divided, and p is a "discontinuity penalty". If there is no penalty, one can get completely spurious correspondences with quite high scores. A value of 1.0–2.0 seems useful, and we will use $p = 1.0$ throughout. An example of a substructure is shown in Figure 2. Although our algorithm can produce many substructures for every pair of molecules, for our purposes we will consistently use only one, the highest scoring common substructure (HSCS).

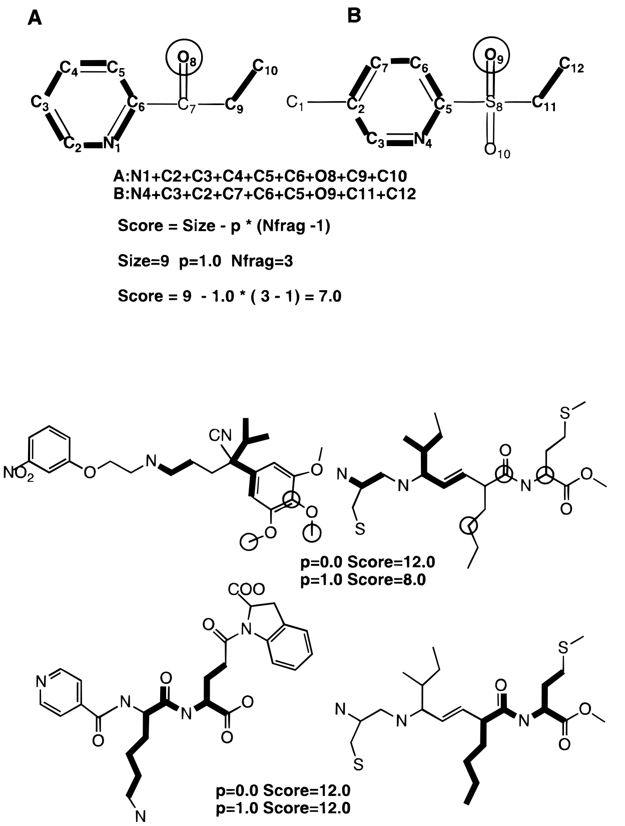
A
B
C_4 C_5 O_8 C_{10}
C_3 C_6 C_7 C_9
C_2 N_1
O_9 C_7 C_6 C_{12}
C_1 C_2 C_5 S_8 C_{11}
C_3 N_4 O_{10}
A:N1+C2+C3+C4+C5+C6+O8+C9+C10
B:N4+C3+C2+C7+C6+C5+O9+C11+C12
Score = Size - p * (Nfrag -1)
Size=9 p=1.0 Nfrag=3
Score = 9 - 1.0 * (3 - 1) = 7.0
NO_2 O N CN O O O N N S S O N O O
p=0.0 Score=12.0
p=1.0 Score=8.0
COO N O N N O O O O N N S O N S O O N
p=0.0 Score=12.0
p=1.0 Score=12.0

Figure 2. (Above). Two molecules A and B with the atoms in the highest-scoring common substructure (HSCS) displayed. The convention used in this paper for chemical drawings is that atoms in the HSCS are joined by bonds rendered in bold; isolated atoms are circled. The score for this substructure is worked out as a function of size, number of fragments, and a "discontinuity penalty" p (see text). (Below) Examples of HSCSs that would have the same score if $p = 0.0$. This demonstrates the usefulness of $p > 0$ in lowering the score of HSCSs with largely spurious correspondences.

Z-Score. Since we are looking at HSCSs to determine what is in common between molecules with particular activities, it is important to ask what type of HSCSs would appear between completely unrelated molecules. For this we need to look at a large sample of pairs of randomly selected molecules. When we do this we find that the mean expected score is a linear function of the size of the smaller molecule in the pair

$$\text{Mean score of HSCS } (n_A,n_B) = M_{\text{mean}}^*\min(n_A,n_B) + B_{\text{mean}}$$

where n_A is the number of atoms in molecule A. M_{mean} and B_{mean} are constants describing the line. A similar relationship exists for the standard deviation in scores:

$$\text{Stdev of HSCS } (n_A,n_B) = M_{\text{stdev}}^*\min(n_A,n_B) + B_{\text{stdev}}$$

The values of M_{mean}, B_{mean}, etc. depend on the atom types, the value of p, and the type of database from which the random molecules are selected. For instance, random compounds from a database of drugs might have more in common than random compounds from a database of general organic compounds.

One can define a "Z-score" for a HSCS between molecules A and B by asking how much larger the score is than expected for a pair of random molecules of the same size:

$$\text{Z-score} = (\text{Score} - \text{Mean}(n_A,n_B))/\text{Stdev}(n_A,n_B)$$

We use the Z-score as a way of selecting HSCSs which are very unlikely to be accidental. Usually we say that a HSCS with a Z-score ≥ 4.0 is "significant". Justification for this cutoff is provided in the Results section.

Visualization of Substructure Information. One could in principle try to construct a representation of a few substructures common to many compounds. Rather than present a summary of that type, we prefer to browse individual molecules with HSCS information projected onto them. While there are many more entities to inspect than there would be for a summary, the molecular context of the substructure is always present. We use two complementary ways to visualize HSCS information. The first, "individual HSCS", is to look at the pairs of molecules sorted by decreasing Z-score of the HSCS. The molecules are displayed side by side and the atoms in the HSCS are highlighted in some user-defined way. For this paper, the convention will be that bonds between atoms in the HSCS are bold; where there is a single isolated atom, it is circled. On a molecular graphics terminal, the atoms in the HSCS can be made "ball-and-stick", while the remaining parts of the molecule remain "wire".

For the second type of visualization, "scored atom", each atom in each molecule receives one "point" for every significant HSCS that contains that atom. Molecules are then sorted by the maximum number of points in the molecule. Atoms in each molecule are labeled as to the number of points, i.e., the number of HSCSs in which it has appeared. For this paper we will simply write the number next to the atom. On a graphics terminal, the numbers can be displayed as color or size of the atom.

Displays of a molecule with scored atoms can be somewhat difficult to interpret. All the atoms of the molecule are displayed whether the atoms are in any HSCS or not, and this distracts the viewer from the information in the numbers. Therefore, we employ a simplified display of the molecules called "conserved atom" such that atoms in a molecule that are not conserved, i.e., not in the majority of the HSCSs of which that molecule is a part, are deleted. An example is shown in Figure 3.

Source of Structures. We will take our molecules from the MDDR (MACCS Drug Data Report) Version 97.2,[5] a licensed database containing data from the patent literature on ~80 000 chemical structures. Each molecule can have one or more "activity" key words. We will be taking five sets of compounds from the MDDR. The first is set of 200 selected randomly from the entire MDDR. We take this set (RANDOM) to represent a sample of nearly all unrelated compounds, from which we can calibrate the significance of HSCSs. The others are selected based on activity label. We chose cholecystokinin (CCK) antagonists and benzodiazepine agonists (BENZ) to see if our method can detect the privileged structure that Evans et al.[1] proposed was common between those activities. There are no molecules in the MDDR with both activities. Antidepressant (ANTIDEP) and antihistaminic (ANTIHIST) molecules were chosen as examples of drugs that work on different receptors. Antidepressants are targeted to various receptors in the central nervous system and antihistamines are targeted to peripheral histamine receptors. Often antihistamines have side-effects on CNS receptors, but there are only three very close analogues listed in the MDDR with both activities.

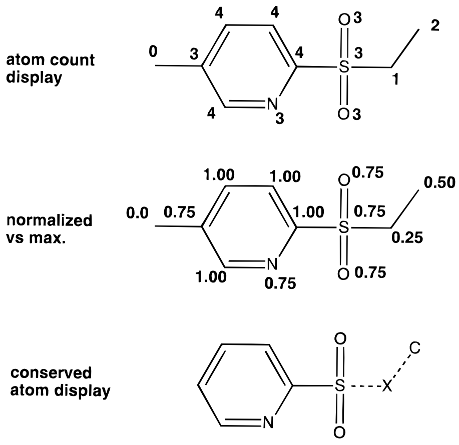

Figure 3. An example of a molecule processed to produce a "conserved substructure" display. We start with each atom of the molecule labeled as to how many significant HSCSs the atom occurs in. Next, each atomic value is normalized by the maximum value in the molecule (here 4). Atoms that have a normalized value below some user-defined threshold are considered "variable" and are marked with X and bonds to them are drawn as dotted; the rest of the atoms are "conserved". Variable atoms not on a path containing two conserved atoms are deleted. Throughout we use the threshold 0.5.

Table 1

molecule set	keywords	no. in original set	no. in diverse set
RANDOM	none	200	200
CCK	CCK antagonist	457	35
BENZ	benzodiazepine agonist	187	39
ANTIDEP	antidepressant	2791	483
ANTIHIST	antihistaminic	590	85

As with most sets of compounds possessing a particular activity, these sets contain many very close analogues. Because of the important concern mentioned in the Introduction, a diverse subset needed to be chosen from each activity so that within each subset no two compounds were very similar in a global sense. For this purpose it was most efficient to employ a method based on topological descriptors. Here we used the regular atom pair descriptor[6] with the Dice similarity index and selected a subset such that no two compounds had a similarity more than 0.6. Experience has shown that this cutoff is very effective at removing close analogues. The subset selection greatly pared down the sets. The final sets are summarized in Table 1.

RESULTS

Sets of HSCSs Calculated. Each molecule in the RANDOM set was paired with every other in that set and the HSCS extracted for each pair, to produce the RANDOM-RANDOM set of substructures. The diverse CCK set was handled similarly to produce the CCK-CCK substructures. A BENZ-CCK substructure was produced by pairing every molecule in the diverse BENZ set with every molecule in the CCK set. ANTIDEP-ANTIHIST substructures were produced similarly. The CPU time for a single pairwise comparison on a drug-like molecule is typically 0.1−0.5 s on an Indigo2 with R10000 processor.

Calibration of Z-Scores. A graph of HSCS score as a function of min(n_A,n_B) for the RANDOM-RANDOM set

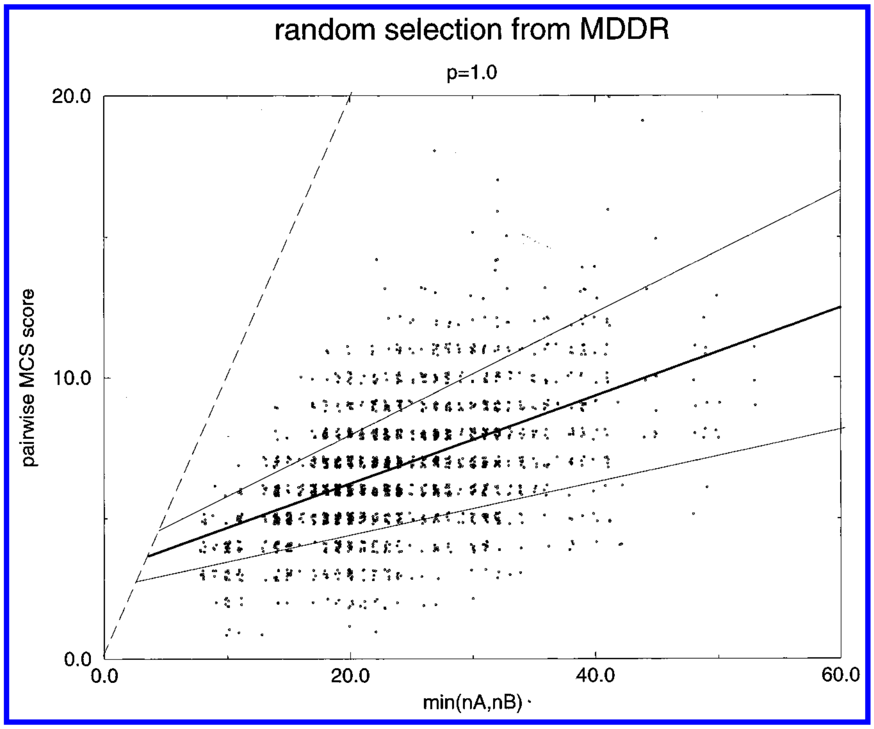

Figure 4. Scores for randomly selected HSCSs for the RANDOM-RANDOM case as a function of the size of the smaller molecule for the scoring function discussed in Methods. Individual points are randomly displaced slightly from their integer values so the distribution of points can be better seen. The thick line indicates the best fit line through the data for the mean score. The thin lines indicate ±1.0 standard deviation from the mean. The dashed line represents the highest possible score.

appears in Figure 4. The best fit line through the data indicates that

$$\text{Mean score of HSCS}(n_A,n_B) = 0.156*\min(n_A,n_B) + 3.08$$

$$\text{Stdev of HSCS}(n_A,n_B) = 0.063*\min(n_A,n_B) + 0.54$$

One consequence of this calibration is that the line for 4.0 standard deviations above the mean and the line for maximum possible score intersect at min(n_A,n_B) ~9, indicating nine non-hydrogen atoms as the minimum size for a molecule to contain a significant substructure. For most drug-size molecules, with 20−30 non-hydrogen atoms, the minimum size for a significant substructure is 13−17 atoms. Since the most common group in drug molecules, the phenyl ring, has six atoms it is very reasonable that the minimum number should be much greater than six.

The distribution of Z-scores using this calibration is shown in Figure 5. It appears nearly normal, with the frequency of Z-scores becoming very small below 4.0, suggesting this as a reasonable cutoff. Figure 6 shows selected HSCSs where min(n_A,n_B) is between 20 and 30 with the Z-scores roughly every 1 unit. After inspecting many examples, we feel that the equivalences between pairs of molecules appears less "meaningful" when the Z-score falls below 4.0, consistent with that number as a useful cutoff.

The total number of HSCSs and the number of significant HSCSs for all the diverse sets is shown in Table 2. Note that the total number of HSCSs may be smaller than the total number of pairs examined because we do not count HSCSs of less than two atoms.

BENZ-CCK. Figure 7 shows all the significant HSCSs for this set. Since there are so few, it is easy to illustrate the use of the two visualization methods. Consistent with the observations of Evans et al.[1], 5-phenylbenzodiazepin is a substructure common between the two activities. However, the substructure drawn by Evans et al. contains the 2-carbonyl, which we see here is not in common. Unexpectedly we found the HSCS 157236-150861, containing a β-carboline ring, which has a higher Z-score than the other two. In the MDDR as a whole, β-carbolines occur in ~80 compounds, of which 52 are marked as CCK antagonists or anxiolytics (probably working via the benzodiazepine receptor). This implies that β-carboline is not a very common group in drug-like molecules but fairly specific for those two activities. As will be seen in the next example, β-carboline appears as a common substructure in the CCK set.

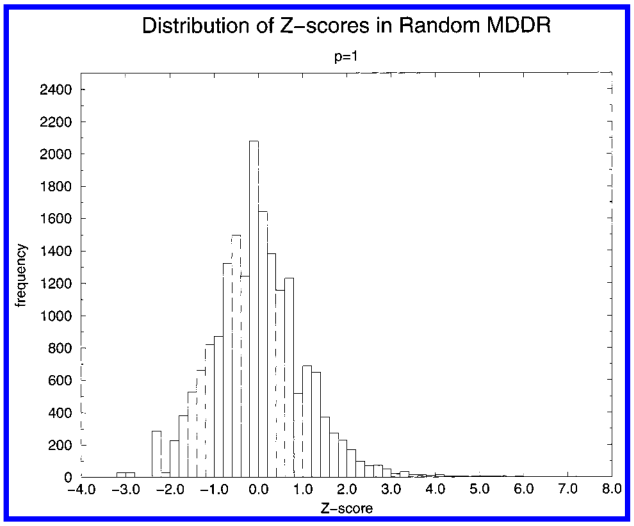

Figure 5. The distribution of Z-scores for the RANDOM-RANDOM set.

Figure 8A shows all the molecules in this set that contain at least one significant HSCS. The molecule 204290 is first

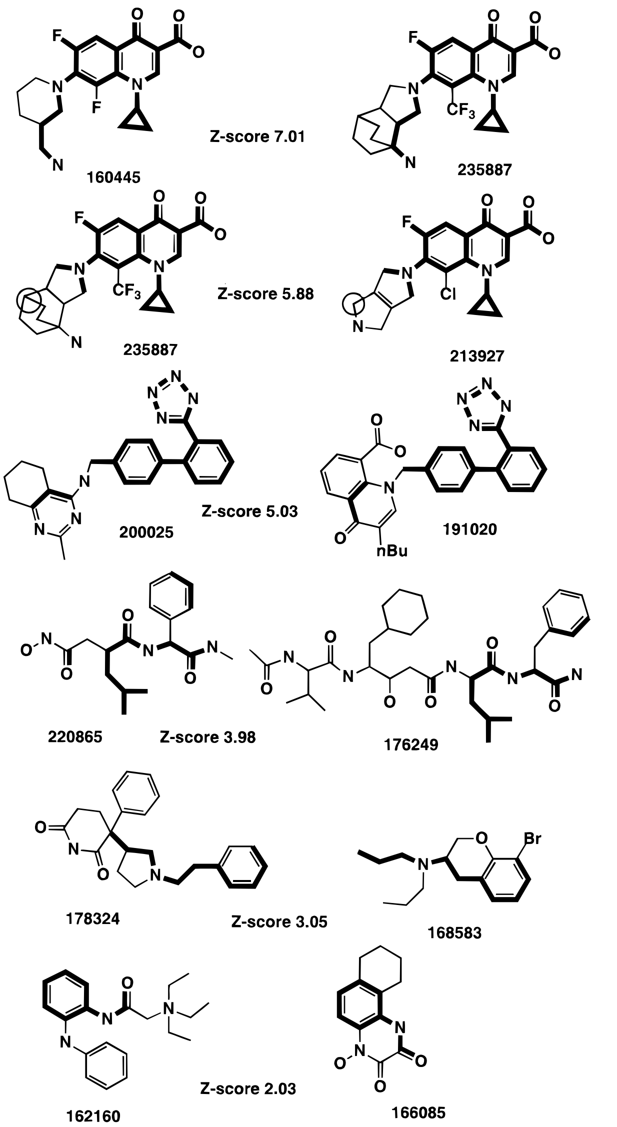

Figure 6. Selected HSCSs from the RANDOM-RANDOM set where min(n_A,n_B) is between 20 and 30, i.e., typical size for a drug-like molecule. The HSCSs were taken roughly 1.0 Z-score unit apart. Molecules are labeled by their external registry number in the MDDR database.

Table 2

set	total no. of HSCSs	no. of Z-scores ≥ 4.0	highest Z-score
RANDOM-RANDOM	18781	57	8.53
CCK-CCK	528	38	7.35
BENZ-CCK	1221	3	6.37
ANTIDEP-ANTIHIST	40973	211	8.03

because the atoms in the 5-phenylbenzodiazepine ring appear in two significant HSCSs: 204290-196309 and 204290-193319. Because no other atoms appear in a significant HSCS, they have a count of zero and are not labeled. The next two 5-phenylbenzodiazepin-containing molecules appear in only one significant substructure each, the ones mentioned above. The next two molecules are from the same HSCS: 157236-150861.

It is clear from this example, that looking at the HSCSs directly emphasizes the most significant HSCSs, while looking at labeled molecules emphasizes the most numerous HSCSs.

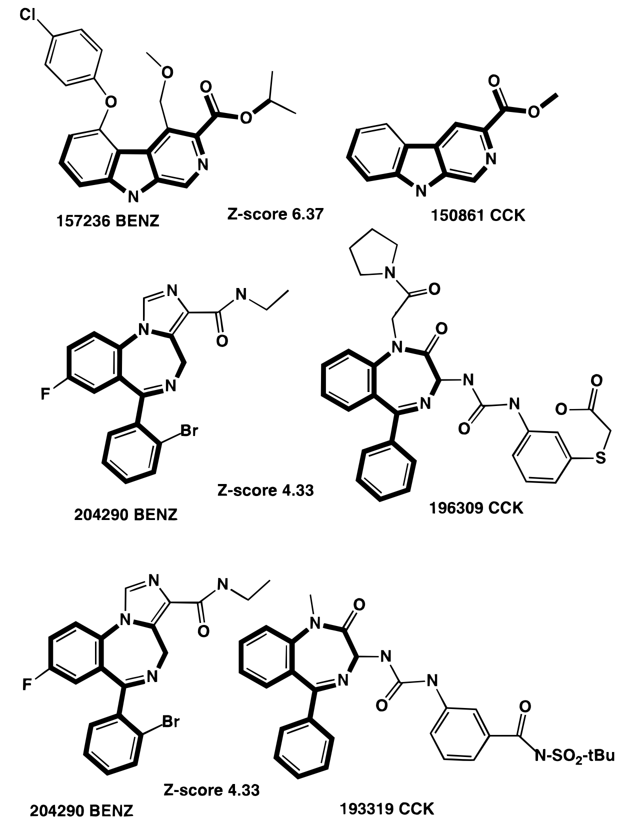

Figure 7. All significant HSCSs for the BENZ-CCK set. The activity of each molecule is indicated.

The conserved atom displayed in Figure 8B shows that there are only two unique conserved substructures among the five molecules in Figure 8A.

CCK-CCK. Figure 9 shows the six most significant HSCSs in this set. There are several types of recurrent substructures present, including a phenyl urea-linked to glycine-*N*-phenyl glycine and a phenyl urea-linked to 5-phenylbenzodiazepines. The 167592-196309 pair shows the commonality between those two types, 167592 being an open-chain version of 196309. (Note that only part of one ring appears in this HSCS because the cyclization changes the shortest-path distance between some atoms in the ring.) There is also a substructure (150861-205698) containing a β-carboline and a substructure (139383-165314) containing a 2-indolyl-glutamate.

Figure 10A shows the six molecules that contain the largest number of significant HSCSs, and Figure 10B shows the conserved parts of these and a few other molecules. Looking at 196309 as an example of benzodiazepine-containing molecules, we see that the benzodiazepine ring, the urea, and the phenyl marked “A” have the highest numbers. The 2-carbonyl of benzodiazepine also seems to be conserved, in contrast to the previous example. Lowe et al.[7] have suggested that the 5-phenyl-3-ureidobenzapepin-2-one is conserved for CCK. However, these results show that this is only partly correct. The 5-phenyl group on the benzodiazepine is not conserved. It is often missing, replaced by 5-cyclohexyl, or a heterocycle with more polar character, e.g., in 190904 and 193712. Substitutions on the 1-position of the benzodiazepine and on the ring A are also not conserved.

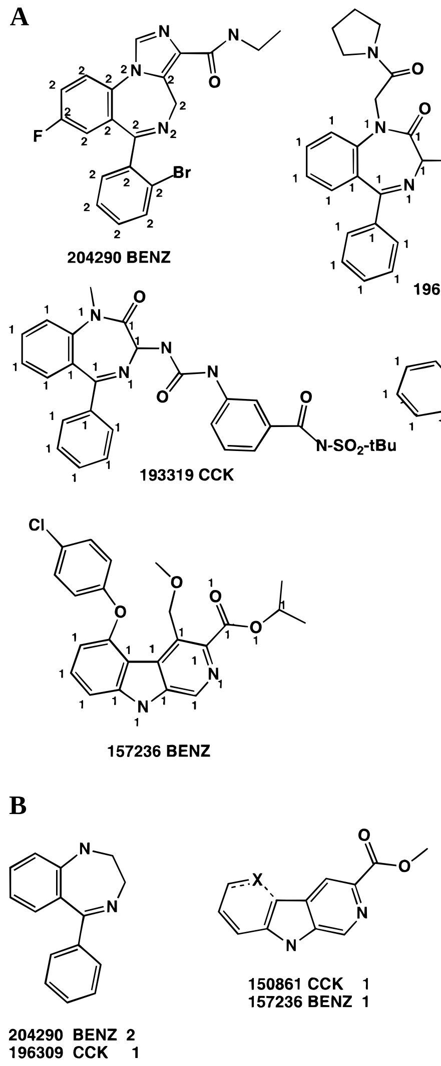

Figure 8. All molecules in the BENZ-CCK set that contain in at least one significant HSCS. A. In "atom count" display. The atoms are labeled with the number of significant HSCSs in which that atom appears, where that number is not zero. B. The unique "conserved atom" representations of the molecules in Figure 8. The labels indicate the molecules from which the representation is taken, and the maximum atom count in that molecule.

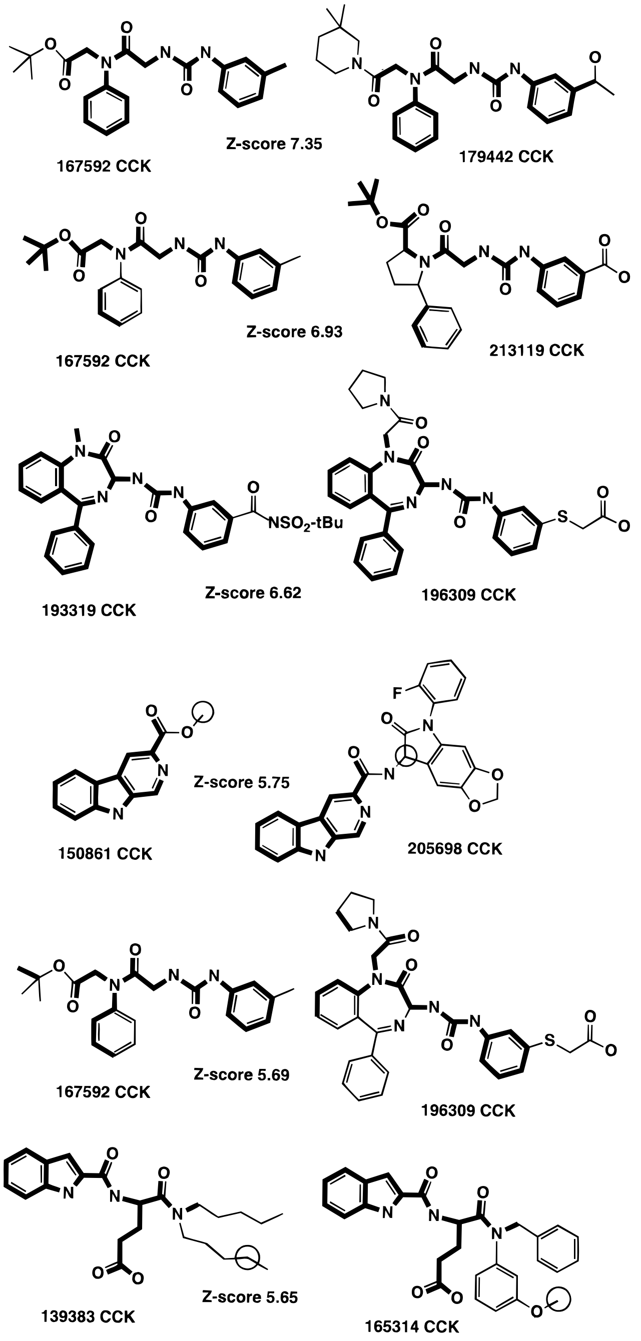

Figure 9. The most significant HSCSs for the CCK−CCK set.

Looking at 167592, which is an open chain variant, we see that the phenyl marked A, part of the phenyl marked B, and the bonds linking them are conserved.

ANTIDEP-ANTIHIST. Figure 11 shows some of most significant HSCSs. The HSCS with the highest Z-score refers to the same compound, 147970. Although there are no other specific compounds which have both activities, antidepressents and antihistamines share several large common substructures as might be expected. Most of the HSCSs are of the form gem-diphenyl-X-(CH2)2-amine, although at least a few other chemical classes are present. Most of the structures have a cation and two aromatic rings.

Figure 12A shows the molecules that are contained within the largest number of significant HSCSs. Simplified versions of the molecules are shown in Figure 12B. In 143217 gem-diphenyl, plus a carbon two bonds away, is the only conserved portion. This is the feature contained in the most HSCSs. (Some of the gem-diphenyls are part of a tricyclic ring system as in 147988.) While almost all ANTIDEP and ANTIHIST compounds contain a cation, the reason that the cation does not appear in the most HSCSs is that the distance to the gem-diphenyl is variable from three to five bonds, with four bonds being the most common (see next paragraph). Since the environment surrounding the cation is so small and so variable, the cation by itself almost never occurs in a significant HSCS.

The majority of the compounds in Figure 12B contain a more extensive conserved substructure: gem-diphenyl-X-(CH2)2-amine, where X is the most variable. Para-halogens appear in some of these.

Another recurrent substructure is aryl-piperazine-(CH2)2-aryl, like that contained in 194163.

DISCUSSION

We have demonstrated a method wherein meaningful common substructures shared by active molecules can be

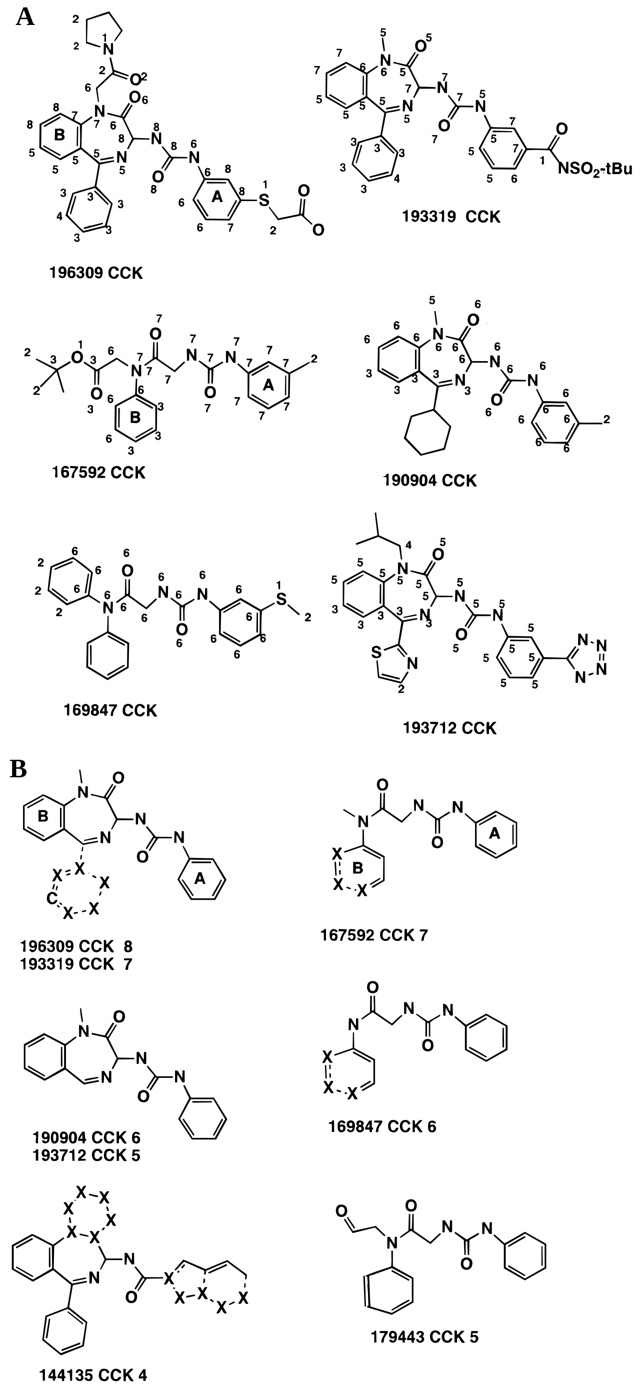

Figure 10. Molecules in the CCK–CCK that contain in the largest number of significant HSCSs. A. In atom count display. B. Conserved atom display.

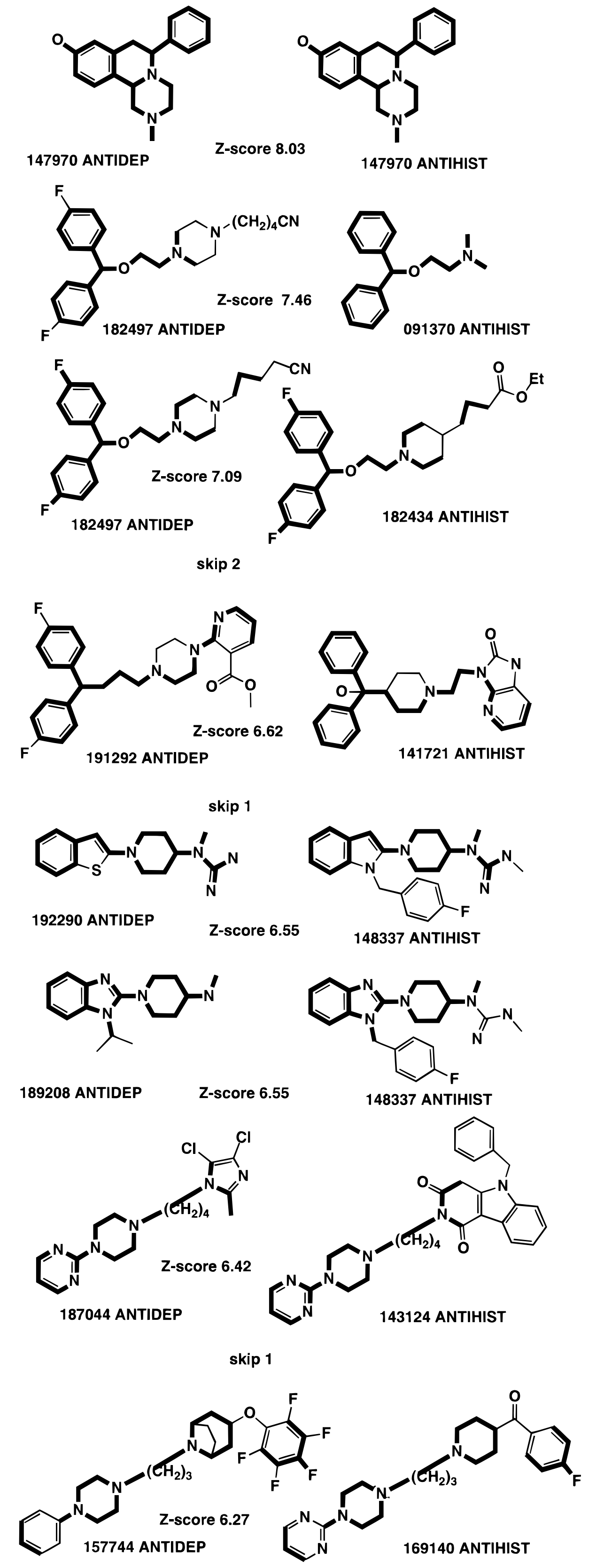

Figure 11. Selected most significant HSCSs for the ANTIDEP-ANTIHIST set.

detected and displayed. The method can be used to find candidate 2D pharmacophores from sets of molecules active on the same receptor or to find privileged structures on sets of molecules active on different receptors. While we have shown examples involving up to two activities, the method is trivially extended to three or more activities. One simply looks at all pairs of molecules where the molecules in each pair have different activities.

Our philosophy for displaying HSCS information is to do so in the context of individual molecules rather than to summarize all the information into a few small substructures. More entities have to be inspected, but more detailed information is retained. Further discussion will focus on three aspects: the interpretation of substructures, extensions to the current method, and comparison to previous work.

How can we interpret the substructures indicated in this or any related method? Most investigators realize that just because a substructure is conserved in many diverse actives

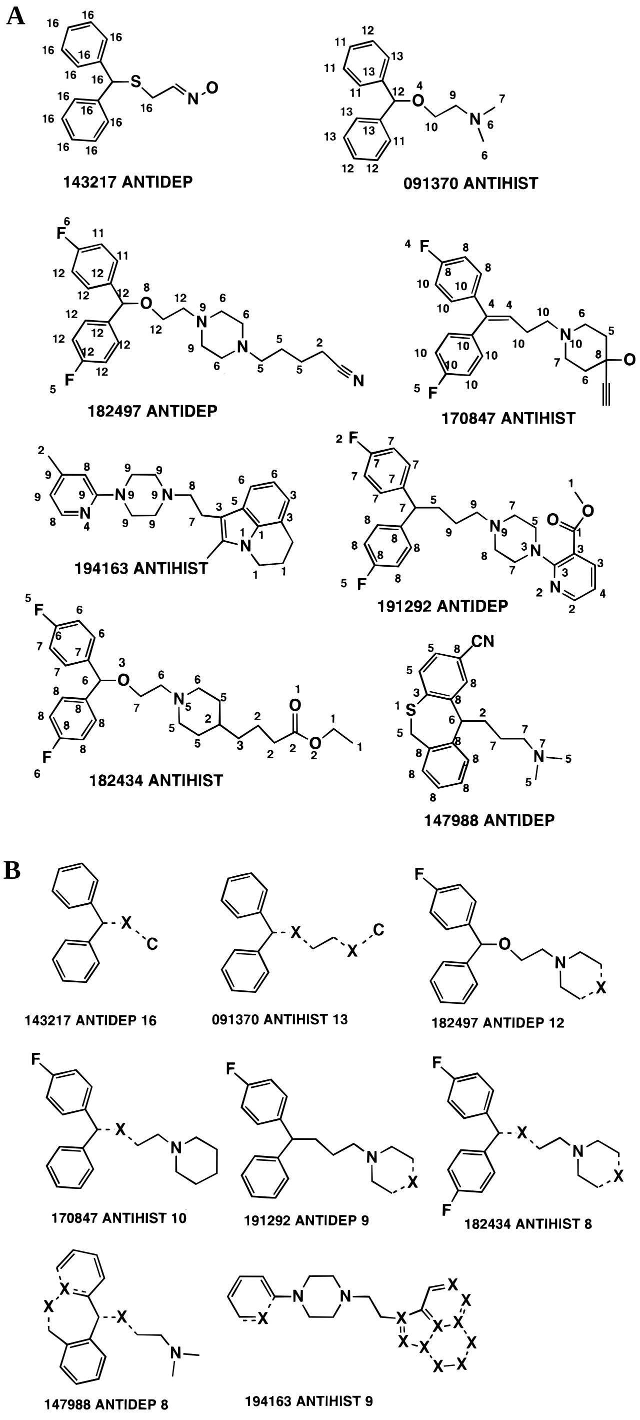

Figure 12. Molecules in the ANTIDEP-ANTIHIST set that contain the largest number of significant HSCSs. A. In atom count display. B. Conserved atom display.

we cannot conclude it is either necessary or sufficient for activity. In practice we have finite sets of actives wherein not every synthetic change has been attempted. Thus what we see may more closely reflect the history of synthetic work in a given therapeutic area than what the receptor requires. The effect of this is that the substructures appear more conserved and larger than they really are. A second problem is that since we are looking at only active molecules, we have no information about whether the substructure is actually correlated with activity. Thus conserved substructures need to be validated by other methods wherein the frequency of the substructure in active vs inactive molecules is statistically compared.

While our method works well as it stands, there are two obvious extensions. At present, the matching and scoring algorithms are very discrete. In particular, two atoms match only if their types are identical and the contribution of matched atoms to the score is 1.0. One can imagine a more continuous scheme where the similarity of atom types is made on a scale of 0 to 1. Atoms would match if their similarities were above a certain threshold, and their contribution to the score would be their similarity. Selection of the highest scoring common substructure would be the same, except that the substructures would now have nonintegral scores.

In the current "scored atom" visualization, only those HSCSs with Z-scores ≥ 4.0 are counted and the rest are ignored. Additional fuzziness could also be added by a weight on each HSCS as a ramp function of Z-score. For instance

$$\text{Z-score} \geq 4.5 \text{ weight} = 1.0$$

$$3.5 \leq \text{Z-score} < 4.5 \text{ weight} = (\text{Z-score} - 3.5)/(4.5-3.5)$$

$$\text{Z-score} < 3.5 \text{ weight} = 0$$

The increment to an atom from a particular HSCS would be weight*contribution, where contribution is the score contribution discussed above.

It is important that users be able to type atoms arbitrarily so they can tune the types for specific purposes. Most common substructure methods in the literature consider only the element and the hybridization of the atom. Our types include more information about local chemical environment, so that, for instance, one distinguishes a −N− in a tetrazole (anion at physiological pH) from a −N− in a basic amine (cation). This is relevant to considering how drugs interact with their receptors.

Many methods for finding maximum common substructures in 2D and 3D have been proposed, for example, refs 8−10. The algorithm we use here could be replaced by any 2D method that is rapid, allows for discontinuous substructures, and allows arbitrary typing of the atoms. There are methods to generate a common substructure for more than two molecules at a time (for instance ref 8). We chose to handle only pairwise comparisons of molecules for three reasons. First, comparing sets of molecules with different activities A and B can only be done sensibly by forming pairs: one molecule from A and one from B. Second, it is much more computationally tractable to examine all possible pairs of molecules from a large set than all triplets, etc. Third, as the number of molecules we compare at one time goes up, the size of the substructure they all have in common will go down. Information could be lost.

The work on STIGMATA[11] has some analogies to our work in that STIGMATA includes a method of displaying molecules so that conserved parts are highlighted, the conservation being expressed in terms of atom path descriptors instead of maximum common substructures. In STIGMATA, one considers the Daylight fingerprints, represented as bit strings, for a set of molecules. The bits occurring in at least a specified percentage of the molecules, say 50%, are collected into a "modal" fingerprint. Given an arbitrary molecule, each atom in the molecule can be scored by what fraction of the bits from the bond paths emanating from that atom are included in the modal fingerprint. The score, from 0 to 1 where 1 is the most conserved, can be used to color the atoms; this is analogous to our atom count display. There is an important difference in approach however. The examples in the original STIGMATA paper[11] show a "top-down" approach in that one is searching for commonality in an entire set of molecules in one step. This is not really applicable when the set contains many different chemical classes or the common substructures are rare in the set; the common features will be lost in the noise. Also, the top-down approach does not directly allow comparison of molecules with two different activities. In contrast, our approach can be considered "bottom-up" in that pairwise commonalities are considered first, and conserved substructures are built from the pairwise commonalities.

There has been at least some work on finding the statistical significance of the descriptor-based similarity for a pair of molecules (e.g., ref 12 for Daylight fingerprints). However, we are aware of no other work that analyzes the significance of the maximum common substructure shared by two molecules. In retrospect this is a serious omission. Drug-like molecules tend to be made of different arrangements of similar parts (the most common being the phenyl ring), and most pairs of randomly selected drug molecules will have something in common. We believe it is indispensable for future work in this field to know how to separate the signal of a meaningful substructure from the noise of spurious small matches.

ACKNOWLEDGMENT

The methods used in this paper were written using tools from the MIX modeling system developed at Merck Research Laboratories. The authors thank the other members of the MIX team for their tireless work.

APPENDIX

Our common substructure algorithm is modified from the "minimum-residual" clique-finding algorithm in Miller et al.[2]

Each substructure C is a clique, which is a set of paired atoms from two molecules A and B. An example of a substructure where the number of pairs, npair, equals 3 is shown below:

$$\begin{aligned} C_A(1) &= 3 \quad & C_B(1) &= 4 \\ C_A(2) &= 5 \quad & C_B(2) &= 8 \\ C_A(3) &= 7 \quad & C_B(3) &= 5 \end{aligned}$$

Atoms 3, 5, and 7 in A correspond to atoms 4, 8, and 5 in B. These atoms are a clique in the sense that atom types of 3, 5, and 7 in A match atoms 4, 8, and 5 in B and the topological distances 3−5, 3−7, and 5−7 in A are the same as the distances 4−8, 4−5, and 8−5 in B.

The algorithm starts at each possible pairing of atom *i* in A and *j* in B. If *i* and *j* are of the same type, they form the first pair in the clique. The algorithm then looks for another pair of atoms that can form a clique with the first pair. It repeats the search, looking for another pair that can form a clique with all the previous pairs, until there are no more suitable pairs left. We use the heuristic that at each step in the clique-building, the through-bond distance of the new pair from the old ones should be a minimum. There is no back-tracking. We present the pseudocode for the algorithm below. Matrix **M** records whether atoms in the two molecules match: $M(i,j) = 1$ means that atom *i* in A and atom *j* in B have the same atom type. $M(i,j) = 0$ means they do not. During clique-building, data mask V_A records whether an atom *i* in A is available for matching ($V_A(i) = 1$) or has already been matched ($V_A(i) = 0$); similarly for V_B. $V_A(*) = 0$ means all elements of V_A are zeroed.

```
For all pairs of i in A and j in B where M(i,j)=1:


     Initialize $V_A(*)=1$, $V_B(*)=1$. Set $V_A(i)=0$, $V_B(j)=0$
     npair=1; $C_A$(npair)=i; $C_B$(npair)=j


     next_pair:
     Smin=100000
     For all pairs of k in A and m in B where the following is
     true: 1) M(k,m)=1
           2) $V_A(k)=1$ and $V_B(m)=1$
           3) The topological distances from k to all the
              atoms $C_A(*)$ matches the distances from m to all
              the atoms $C_B(*)$.
       Calculate the sum of distances S.
       if(S < Smin) Smin=S, k'=k , m'=m
    End loop over k-m pairs


    If there is no k-m pair that meets the criteria, exit.
    Otherwise:
       npair=npair+1; $V_A(k')=0$; $V_B(m')=0$ ; $C_A$(npair)=k';
       $C_B$(npair)=m'


    If npair $\geq$ min($n_A$,$n_B$) , exit
         Go to next_pair


End loop over i-j pairs
```

Note that for molecules A and B there is a potential clique formed for each initial *i*−*j* pair. Some of these cliques have identical correspondences in different orders. For instance

$$\begin{aligned} C_A(1) &= 3 \quad & C_B(1) &= 4 \\ C_A(2) &= 5 \quad & C_B(2) &= 8 \\ C_A(3) &= 7 \quad & C_B(3) &= 5 \end{aligned}$$

is the same as

$$\begin{aligned} C_A(1) &= 5 \quad & C_B(1) &= 8 \\ C_A(2) &= 3 \quad & C_B(2) &= 4 \\ C_A(3) &= 7 \quad & C_B(3) &= 5 \end{aligned}$$

Cliques are canonicalized by reordering the pairs in the clique by ascending atom order in A. (The clique on the left above is already in that order.) Once canonicalization is done, identical cliques can be recognized and eliminated.

REFERENCES AND NOTES

(1) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. L.; Lotti, V. J.; Cerino, D. J.; Chen, T. B.; Kling, P. J.; Kunkel, K. A.; Springer, J. P.; Hirshfield, J. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235−2246.

(2) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select quasi-flexible ligands complementary to a receptor of known three-dimensional structure. *J. Comput-Aided. Mol. Design.* **1994**, *8*, 153−174.

(3) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*, 723−732.

(4) Bush, B. L.; Sheridan, R. P. PATTY: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.

(5) MACCS−II Drug Data Report (V 97.2) is distributed by Molecular Design Ltd., San Leandro, CA, 1997.

(6) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: definition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(7) Lowe, J. A., III; Hageman, D. L.; Drozda, S. E.; McLean, S.; Bryce, D. K.; Crawford, R. T.; Zorn, S.; Morrone, J.; Bordner, J. 5-phenyl-3-ureidobenzazepin-2-ones as cholecystokinin-B receptor antagonists. *J. Med. Chem.* **1994**, *37*, 3789−3811.

(8) Bayada, D. M.; Simpson, R. W.; Johnson, A. P. An algorithm for the multiple common subgraph problem. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 680−685.

(9) Brint A. T.; Willet P. Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152−158.

(10) Wang, T.; Zou, Z. EMCSS: A new method for maximal common substructure search. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 828−834.

(11) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. STIGMATA: an algorithm to determine structural commonalities in diverse subsets *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.

(12) Bradshaw, J.; Sayle, R. A. Some thoughts on significant similarity and sufficient diversity. Presented at the 1997 EuroMUG meeting, 7−8 October, 1997 in Verona Italy. The text is available at the URL http://www.daylight.com/meetings/emug97/Bradshaw/Significant_Similarity.html.

CI980044F