Chemical Similarity Using Geometric Atom Pair Descriptors†

Robert P. Sheridan,*,‡ Michael D. Miller,§ Dennis J. Underwood,§ and Simon K. Kearsley†

Department of Molecular Design and Diversity, Merck Research Laboratories, P.O. Box 2000, Rahway, New Jersey 07065, and Sumneytown Pike, West Point, Pennsylvania 19486

Received August 10, 1995⊗

Similarity searches using topological descriptors have proved extremely useful in aiding large-scale screening. In this paper we describe the geometric atom pair, the 3D analog of the topological atom pair descriptor (Carhart et al. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73). We show the results of geometric similarity searches using the CONCORD-build structures of typical small druglike molecules as probes. The database to be searched is a 3D version of the Derwent Standard Drug File that contains an average of 10 explicit conformations per compound. Using objective criteria for determining how good a descriptor is in selecting active compounds from large databases, we compare the results using the geometric versus the topological atom pair. We find that geometric and topological atom pairs are about equally effective in selecting active compounds from large databases. How the two types of descriptors rank active compounds is generally about the same as well, but occasionally active compounds will be seen as very similar to a probe in geometric descriptors, but as fairly dissimilar in topological descriptors. These are of two types: (1) compounds where equivalent groups are in the same spatial arrangement as in the probe but are connected by very different bond paths and (2) compounds that can superimpose onto the probe when they are in a folded conformation.

INTRODUCTION

Similarity searches are now a standard tool for drug discovery.[1,2] A user typically uses an interesting molecule as a probe and searches a large database for other compounds that are "similar". The expectation of such a search is that compounds which are similar in structure will have similar biological activity. A useful feature of similarity searches, as opposed to substructure searches, is that the user need not specify what part(s) of the probe are important.

Most similarity methods use topological (2D) features, i.e., atoms and bonds, since most chemical databases consist only of connection tables. Similarity searches using topological features are able to find active compounds that bear some resemblence to the probe but would not be considered analogs. With the advent of methods[3−5] to convert connection tables to three-dimensional (3D) coordinates, effort has been directed to calculate 3D similarity. The hope is that, since bonds are deemphasized in 3D representations, more active compounds can be found that are very diverse in terms of 2D structure.

There are two approaches to similarity whether 2D or 3D. In direct-comparison methods[6−10] the similarity between two molecules is calculated after finding the best superposition of one molecule on another. In 3D one can compare the "fields" from the atoms or find maximum common substructures of atoms in space. These methods tend to be compute-intensive and, with some exceptions,[6,10] are usually not used in database searches. In descriptor-based similarity,[11−16] sets of coordinates are parsed into chemical descriptors, usually sets of atoms and their geometric relationships, and these descriptors are stored in a separate descriptor database. At search time, the probe is similarly parsed, and the similarity between the probe and each database entry is calculated by comparing the list of descriptors in the probe with the list for the entry. Descriptor-based methods cannot be used to obtain a superposition of the probe with the database entry, but they make up for that by being very fast and thus more suitable for searches of large databases. So far, most descriptor-based methods for addressing 3D similarity address molecular shape,[6,13,15,16] without regard to physiochemical properties.

One set of useful topological descriptors with which we have had experience are those developed at Lederle Laboratories, the atom pair[11] and topological torsion.[12] In our companion paper[17] we described extensions of these descriptors using physiochemical atom types. In this paper we describe the extension of atom pair for use in 3D similarity. We compare the results of similarity searching with the topological atom pair versus the geometric atom pair over a database of "quasi-flexible" molecules. We do an analogous comparison for a search over a database of "rigid" molecules.

METHODS

Definitions of Descriptors. Our descriptors are extensions of the atom pair described by Carhart et al.[11] Atom pairs are substructures of the form

$$\text{atom type i} -(\text{distance}) - \text{atom type j}$$

where (distance) is the distance between atom i and atom j. A molecule with n non-hydrogen atoms will have $n(n-1)/2$ atom pairs, although generally not all of the atom pairs will be unique. In the original atom pair (*ap*), "atom type" includes element, number of neighbors, and number of π electrons, and distance is measured in bonds along the shortest path. In our companion paper,[17] we defined new atom pair descriptors using physiochemical atom types. One

* To whom correspondence should be addressed.

† Keywords: similarity searching, property descriptors, atom pairs, 3D descriptors, Derwent Standard Drug File.

‡ Rahway, NJ.

§ West Point, PA.

⊗ Abstract published in *Advance ACS Abstracts,* January 1, 1996.

0095-2338/96/1636-0128$12.00/0

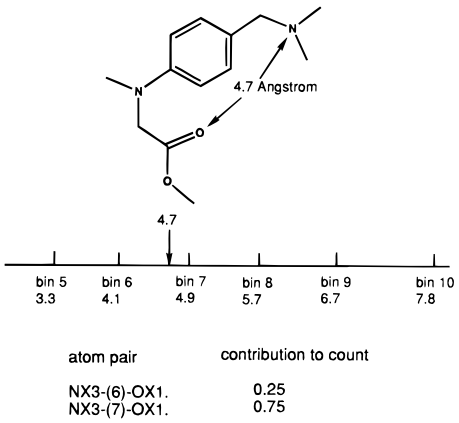

Figure 1. Example of how the through-space distance between atoms in a 3D structure is handled in terms of geometric atom pair descriptors. Distance is divided into 30 bins starting at 1 Å and ending at 75.3. The interval between the first and second bin starts at 0.5 and thereafter the interval increases. Thus the *centers* of the bins are approximately <1.0, 1.5, 2.1, 2.7, 3.3, 4.1, 4.9, 5.7, 6.7, 7.8, 9.0, 10.3, 11.7, 13.3, 15.0, etc. Distances are "fuzzy" in that a particular pair of atoms contributes to two adjacent bins according to how close it is to the bin centers. In the example, the distance 4.7 is $^3/_4$ of the way from bin 6 to bin 7 and therefore contributes $1 - {}^3/_4 = 0.25$ to the total count of the bin NX3-(6)-OX1. and 0.75 to the total count of NX3-(7)-OX1.

descriptor which we will use again here is the binding property pair (*bp*) where atom type is one of seven binding classes[18] (1 = cation, 2 = anion, 3 = H-bond donor, 4 = H-bond acceptor, 5 = polar, 6 = hydrophobic, 7 = other). Figure 1 of our companion paper shows an example of a molecule parsed into *ap*'s and *bp*'s.

Topological atom pairs are easily extended to geometric atom pairs: for a particular conformation of a molecule the distance parameter in the atom pair definition reflects the through-space distance instead of the through-bond distance. Since through-space distance is continuous, we partition distance into discreet "bins". We would like the bins to overlap so that, for instance, a distance of 4.7 Å would be perceived as similar to a distance of 5.0 Å, even though the distances might fall in different bins. This is done by allowing a particular pair of atoms to occupy more than one bin in proportion to where its distance falls between the bin centers. This is demonstrated in Figure 1. The number of times a given descriptor occurs in a molecule, i.e., its count, may not be an integer, but the sum of the counts over all descriptors is still $n(n-1)/2$.

Here we refer to the geometric version of the atom pair as *ag* (i.e., **a**tom pair **g**eometric) and the geometric binding property pair as *bg*.

Construction of Descriptor Databases from Connection Tables. In our implementation, geometric descriptors are derived from *flexibases*, databases that contain an average of ~10 and a maximum of 25 explicit conformations for a given chemical structure. Flexibases are constructed from connection tables by an automatic procedure involving the application of distance geometry and molecular mechanics, followed by steric and geometric filters so that the final conformations are uniformly dissimilar, the minimum root-mean-squared distances between any two conformations being at least 1.2 Å. Flexibases are a way of addressing conformational flexibility without the expense of run-time conformational generation. The details of the construction are given elsewhere.[19]

Conformations in a flexibase can be parsed into geometric descriptors, which are stored in a *geobase*. Each entry in the geobase contains a molecular identifier, a conformation identifier, and the number of each type of the four descriptors (two geometric and two topological). For each of the types are listed the unique descriptors present in the molecule and their counts. As described in our companion paper for topological descriptors, the unique descriptors are identified by two-byte integers, and, as before, the descriptors which use the original atom types (*ag* and *ap*) must be arbitrarily assigned an integer as they are encountered since there are too many possible types.

In our implementation geobases are stored as randomly accessible files. To take advantage of our computer resources, we divide geobases into small fragments, or "stripes", that can be searched concurrently.

Calculation of Similarity. Throughout we will use the index of similarity used by Carhart et al.[11] The similarity of molecules A and B is

$$\mathrm{Sim}_{AB} = \frac{\sum_k \min(f_{Ak}, f_{Bk})}{0.5[\sum_k f_{Ak} + \sum_k f_{Bk}]}$$

where f_{Ak} is the count of descriptor k in molecule A. This is identical to the topological similarity except that the counts need not be integers.

Measures of Merit for Similarity Probes. In our companion paper we proposed a set of measures to determine whether one set of descriptors is better than another based on a simulated screening experiment where N compounds in a database are tested in order of their decreasing similarity to a probe. (1) How many compounds must be tested until half the actives are found. We call this number A50. A50 can be alternatively expressed as a *global enhancement*, the ratio $N/2$ over the A50. (2) How many actives are found after testing an arbitrary small fraction of the total database. For instance the number of actives at 300 compounds tested could be called A@300. A@300 can be expressed as an *initial enhancement*: how many times more actives there are in the set of 300 than expected by chance.

DATABASE USED IN THIS STUDY

In order to measure the merit of the descriptors we need to have a database of molecules for which we know the biological activities. As before we will use Version 6.0 of the Derwent Standard Drug File (SDF),[20] and the same caveats apply.

We generated the SDF flexibase from the licensed version in MAACS format.[21] The flexibase contains ~310 000 conformations of ~30 000 structures. This number of structures is smaller than the number in the topobase used in our companion paper because adequate 3D structures could not be generated from a significant fraction of the connection tables. We generated the SDF geobase from the flexibase. The geobase is divided into 17 stripes.

Databases where there is only one conformation per structure (for example from crystal structures or from CONCORD[22]) are very popular, and so it is important to compare searches over our "flexible" geobase with searches over a "rigid" geobase. We created a rigid geobase by

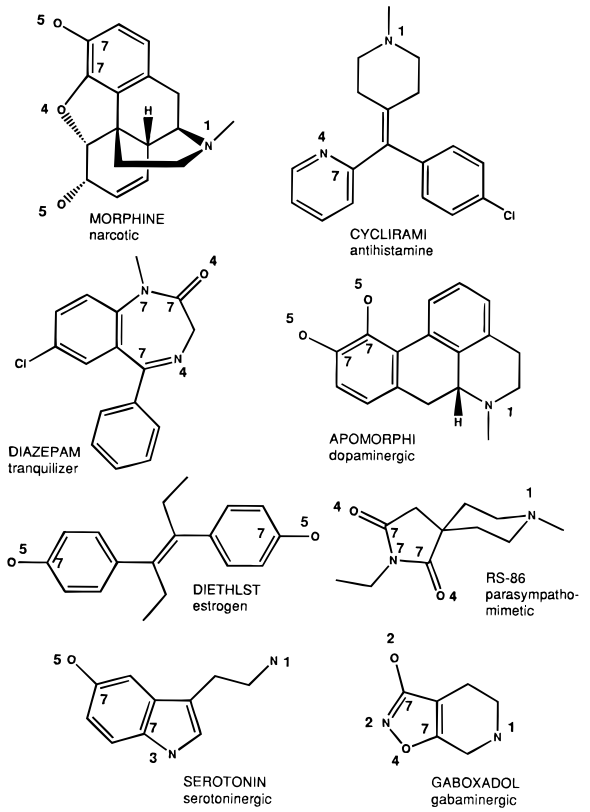

Figure 2. Chemical structures of probes used in this study. The binding property type for each atom is shown where it is not "6".

including only the first entry of each compound from the original flexible geobase. This entry corresponds to the initial 3D structure derived from a 2D drawing and is usually fairly extended.[19]

SIMILARITY PROBES

Choice of Example Probes. Chemical structures of the probes used in this study (named by the SDF external registry number) with the corresponding activity are shown in Figure 2. Table 1 shows how the activities were constructed from keywords in SDF. In the companion paper probes were arbitrarily selected under two constraints. (1) The majority of the actives in the therapeutic area of the probe should work by the same mechanism as the probe. (2) There should be >50 actives. For the purposes of this study, we add an additional constraint that the structure must be conformationally constrained; this ensures that any 3D structure built from the 2D structure of the probe will approximate the active conformation. Here we used the CONCORD conformation. (In practical searches the probe structure would generally not be conformationally constrained, but an active conformation of the probe would have to be specified using additional information.)

How the Probes Were Run. We ran similarity probes with our in-house system GEOSIM. During a search of a geobase, GEOSIM calculates for each database entry the similarity for each of the four descriptors and calculates a score as a user-defined linear combination of the individual similarities. For instance the scores for single descriptors might be

$$ag \text{ score} = ag \text{ similarity}$$

$$bg \text{ score} = bg \text{ similarity}$$

etc.

We can define scores for combination descriptors

$$ag + bg \text{ score} = \text{mean of } ag \text{ and } bg \text{ similarities}$$

$$ap + bp \text{ score} = \text{mean of } ap \text{ and } bp \text{ similarities}$$

Where we do analogous searches over the "rigid" geobase, the geometric descriptors will be called *ar* (i.e., **a**tom pair **r**igid geometric), *br*, and *ar* + *br*.

GEOSIM calculates the similarity of ~120 conformations per second on our cluster of four RS/6000 model 580 workstations running Version 3.2 of the AIX operating system. Four stripes are searched concurrently, and results from each stripe in the geobase are written separately. Once all the stripes are completed, the results are combined and sorted from high to low score. Only the best-scoring conformation is kept per compound. As before we use only the rank of the compounds for comparison.

Measures of Merit for the Examples. Figure 3 shows a graph of the accumulation of actives versus rank for the CYCLIRAMI example. Table 2 lists the measures of merit for all probes. We can get an idea of the overall utility of the descriptors, at least for this set of probes and this database, by taking the mean global enhancement and initial enhancement for over all the probes:

descriptor	mean global enhancement	mean initial enhancement
ag	4.8	18.5
bg	4.2	19.7
ag + *bg*	5.6	21.5
ar	5.2	18.2
br	5.4	19.1
ar + *br*	6.4	20.9
ap	5.3	17.8
bp	5.1	20.4
ap + *bp*	6.7	22.5

All descriptors give enhancements ≫1, indicating general usefulness. For a given atom type description, the three families of searches seem to give roughly equal results on the average (e.g., *ag* ~ *ar* ~ *ap*, *bg* ~ *br* ~ *bp*, etc.). The combination descriptors seem to do better than either

Table 1. Activities from SDF Used in This Study

activity	key words	comments	no. actives
narcotic	("NARCOTIC" and "ANALGESIC") or "OPIOID"	opiate agonists and antagonists	465
antihistamine	"ANTIHISTAMINES-H1"	histamine-H1 antagonists	376
tranquilizer	"TRANQUILIZER" or "BENZODIAZEPINE-AGONIST"	mostly benzodiazepine agonists	359
dopaminergic	"DOPAMINERGICS"	dopamine agonists	240
estrogen	"ESTROGENS"	estrogen agonists	189
parasympathomimetic	"PARASYMPATHOMIMETICS"	muscarinic and nicotinic acetylcholine agonists	129
serotoninergic	"SEROTONINERGICS"	serotonin agonists	66
gabaminergic	"GABAMINERGICS"	GABA agonists	54

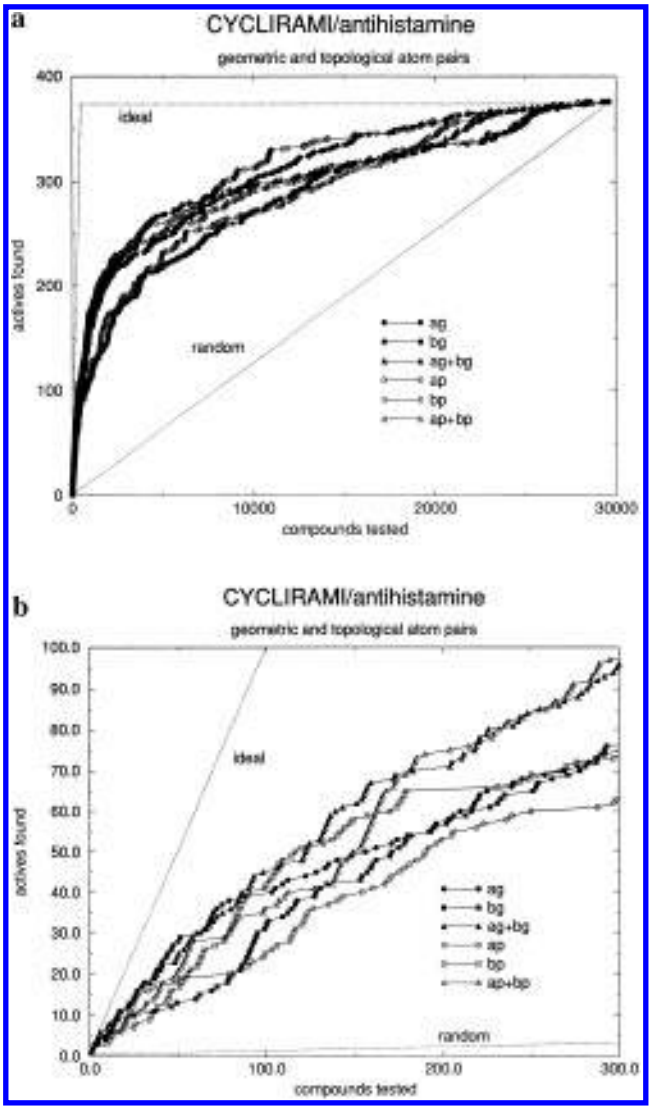

Figure 3. Curves for the accumulation of actives versus rank for the CYCLIRAMI example. Two limiting cases are also shown: "ideal" where all the actives would be at the front of the list, and "random" where all the actives would be evenly distributed throughout the list: a. the curve over the entire database and b. the curve for the first 300 molecules tested.

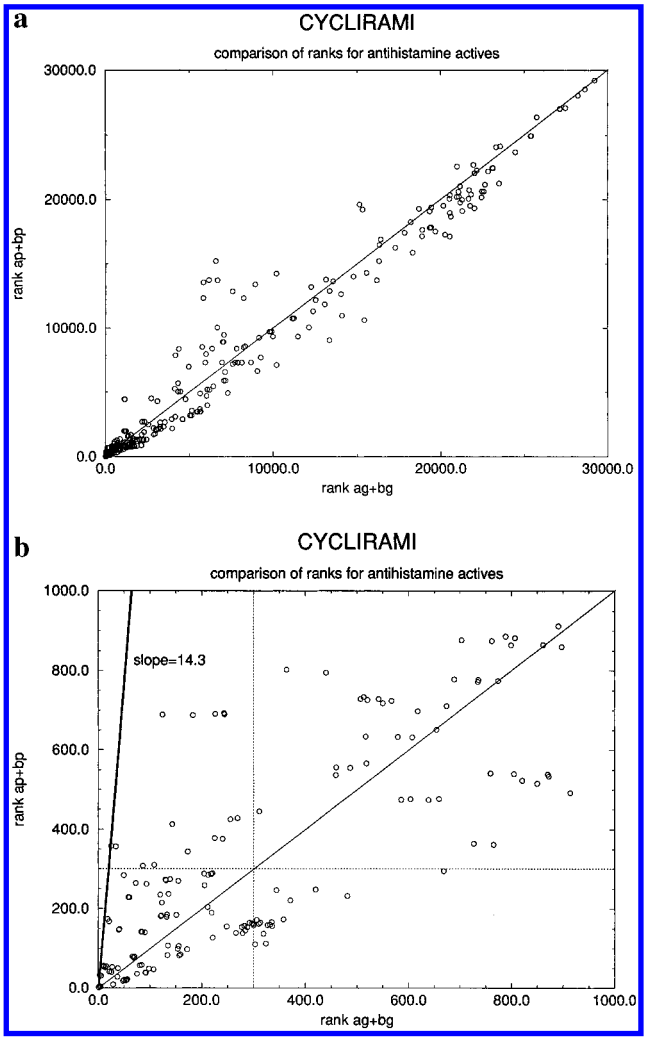

Figure 4. The correlation of rank for the geometric and topological atom pair descriptors ($ap + bp$ versus $ag + bg$) for CYCLIRAMI: a. The scatterplot over the entire database. Each circle represents a compound in SDF with antihistamine activity. b. Closeup of the origin of a. The rank cutoff at 300 is indicated. The line from the origin through ETOLOXAMI (slope = 14.3) is shown as an example of a compound that falls far from the diagonal (slope = 1).

component descriptors in all cases (e.g., $ag + bg > ag$ or bg, $ap + bp > ap$ or bp, etc.). These observations seem to apply to individual cases in Table 2 as well as to the average.

Correlation of Ranks between Descriptors. Besides looking at how well descriptors do at selecting active compounds, one can also look at how descriptors rank the active compounds. Table 3 summarizes the correlation of ranks for different descriptors for all the probes. Consistent with the observations in our companion paper, correlations between descriptors of different atom types do not necessarily correlate very highly (e.g., ag with bg, ap with bp, etc.). However, the correlations among descriptors with the same atom type definition (indicated in bold in Table 3) are always close to 1.0 (e.g., ag with ar and ap, bg with br and bp, etc.). The rankings are highly correlated even if one tries to suppress "covalent information" by ignoring geometric atom pairs containing directly bonded atoms or atoms forming a bond angle (data not shown).

It is easiest to compare the geometric and topological results using the combination descriptors. As an example, consider the graph in Figure 4a. This plots rank by $ap + bp$ score versus $ag + bg$ score for the actives in the CYCLIRAMI/antihistamine example, and Figure 4b shows a closeup of the origin. Generally the ranks cluster along the diagonal. However, there are examples of actives that fall far from the diagonal when one looks at the beginning of the list. To quantitate the deviation, we measure the log of the slope of the line passing through these compounds. Compounds with log(slope) = 0 fall on the diagonal, compounds with log(slope) > 0 score better on $ag + bg$ than on $ap + bp$, and compounds with log(slope) < 0 score better on $ap + bp$ than $ag + bg$.

Table 4A lists those compounds that are especially striking in the comparison of $ag + bg$ with $ap + bp$. They have |log(slope)| > 0.5, and at least one of the ranks <300. Figure 5 shows the structures of selected compounds from Table 4A. Inspection of Table 4A indicates that for three out of eight probes (CYCLIRAMI, APOMORPHI, and RS-86) at least one new class of compounds seen with $ag + bg$ that would be missed by $ap + bp$. In these examples we see compounds where similar chemical groups are held via a different topological connectivity. For instance, in ETOLOXAMI the cation is bonded to one aromatic ring

Table 2. Measures of Merit for Various Descriptors

descriptor	A50	global enhancement	A@300	initial enhancement	descriptor	A50	global enhancement	A@300	initial enhancement	descriptor	A50	global enhancement	A@300	initial enhancement
MORPHINE/narcotic														
ag	2242	6.6	135	28.7	*ar*	1780	8.3	142	30.1	*ap*	1749	8.5	138	29.4
bg	4583	3.2	106	22.6	*br*	3295	4.5	116	21.5	*bp*	3687	4.0	113	24.0
ag + *bg*	2550	5.8	135	28.7	*ar* + *br*	2228	6.6	137	27.3	*ap* + *bp*	2187	6.8	135	28.7
CYCLIRAMI/antihistamine														
ag	1412	10.5	73	19.1	*ar*	1295	11.4	79	20.7	*ap*	1286	11.5	63	16.5
bg	3065	4.8	76	19.9	*br*	2516	5.9	82	21.5	*bp*	2760	5.4	73	19.1
ag + *bg*	1484	10.0	96	25.1	*ar* + *br*	1173	12.6	104	27.3	*ap* + *bp*	1122	13.2	97	25.4
DIAZEPAM/tranquilizer														
ag	2530	5.8	82	22.5	*ar*	2323	6.4	86	23.6	*ap*	2346	6.3	93	25.5
bg	3664	4.0	52	14.3	*br*	3512	4.2	58	15.9	*bp*	3617	4.1	59	16.2
ag + *bg*	2566	5.8	83	22.8	*ar* + *br*	2439	6.1	86	23.6	*ap* + *bp*	2393	6.2	95	26.1
APOMORPHI/dopaminergic														
ag	2575	5.7	50	20.5	*ar*	2751	5.4	40	16.5	*ap*	2827	5.2	38	15.6
bg	3985	3.7	54	22.1	*br*	3395	4.4	44	18.1	*bp*	3155	4.7	43	17.6
ag + *bg*	2313	6.4	58	23.8	*ar* + *br*	2181	6.8	49	20.2	*ap* + *bp*	2228	6.6	43	17.6
DIETHLST/estrogen														
ag	11496	1.3	28	14.9	*ar*	9296	1.6	31	16.1	*ap*	9357	1.6	32	17.0
bg	1776	8.4	56	29.8	*br*	1226	12.1	62	32.3	*bp*	1186	12.5	66	35.1
ag + *bg*	3161	4.7	32	17.0	*ar* + *br*	2587	5.7	33	17.3	*ap* + *bp*	2122	7.0	33	17.6
RS-86/parasympathomimetic														
ag	6421	2.3	11	8.4	*ar*	5419	2.7	7	5.3	*ap*	6024	2.5	6	4.6
bg	4091	3.6	16	12.2	*br*	3732	4.0	9	6.9	*bp*	3445	4.3	12	9.2
ag + *bg*	3644	4.1	17	13.0	*ar* + *br*	3141	4.7	15	16.3	*ap* + *bp*	3573	4.1	21	16.0
SEROTONIN/serotoninergic														
ag	3333	4.4	14	20.9	*ar*	3415	4.3	14	20.9	*ap*	2950	5.0	14	20.9
bg	5129	2.9	10	14.9	*br*	3916	3.8	9	13.4	*bp*	5115	2.9	10	14.9
ag + *bg*	3858	3.8	12	17.9	*ar* + *br*	3664	4.0	12	17.9	*ap* + *bp*	3282	4.5	12	17.9
GABOXADOL/gabaminergic														
ag	8003	1.8	7	12.7	*ar*	9137	1.6	7	12.7	*ap*	7476	2.0	7	12.7
bg	5527	2.7	12	21.8	*br*	5311	2.8	12	21.8	*bp*	5487	2.7	15	27.3
ag + *bg*	3504	4.2	13	23.6	*ar* + *br*	3562	4.2	11	20.0	*ap* + *bp*	2893	5.1	17	30.9

instead of to the atom between the aromatic rings. In LY-206243, N-0436, and N-0435, the conformations most similar to APOMORPHI are bent into a hairpin with the distal aromatic groups taking the place of the aromatic group nearer the cation in APOMORPHI. In U-77053 and OXOTREMOR, the oxygen and cationic amine are in the same relative position as in RS-86 but are connected differently.

There are five molecules where log(slope) < -0.5. Three examples are due to limitations in the 3D building process. AZIDOMORP, DEOXEPAM, and CYTISINE were incompletely optimized and had poor 3D structures. Two other cases are due to limits in the coverage of conformational space. All the conformations of DIENEDIME are puckered around the central bond and thus cannot attain good spatial overlap with DIETHYLST. There is only one conformation of GABA in the flexibase, an extended one, and it does not overlap with GABOXADOL very well.

Table 4B contains the analogous comparisons of *ar* + *br* with *ap* + *bp*. There are three probes in Table 4A which show new compounds. Of these, only CYCLIRAMI shows the same types of compounds in Table 4B. This is not surprising since the compounds (e.g., ETOLOXAMI, MIANSERIN) can fit onto their respective probe while in an extended conformation. In contrast, most of the compounds in Table 4A for APOMORPHI are missing from Table 4B. This is not surprising since the extended forms of these compounds (e.g., LY-206243, N-0436) will not fit on the probe. A similar situation holds for RS-86. The most extended form of OXOTREMOR, for instance, would not superimpose with RS-86. Clearly, "flexible" databases are more complete than "rigid" databases.

DISCUSSION

While there are many papers on 3D similarity methods, very few of them compare results from 2D and 3D searches. We have extended the atom pair descriptor to a geometric form and compared searches with the geometric atom pair against analogous searches with the topological atom pair. It is the direct 2D to 3D comparison that is the major point of this paper. We have not compared geometric atom pairs with other 3D similarity methods. Discussion will cover the following points. (1) how the geometric atom pair descriptor relates to previous types of 3D descriptors, (2) comments on the general usefulness of 3D similarity searches compared to topological searches, and (3) specific behavior of the geometric atom pair versus topological atom pairs.

The requirements for finding the best superposition between small numbers of (perhaps flexible) molecules and searching large databases are very different, and so 3D similarity methods are usually tuned to do one or the other. For instance, descriptor-based approaches are most useful for rapid searching, although a modified maximum common

Table 3. Correlations between Ranks of Active Compounds

	ag	*bg*	*ag* + *bg*	*ar*	*br*	*ar* + *br*	*ap*	*bp*	*ap* + *bp*
				MORPHINE/narcotic					
ag	1.00	0.93	0.98	**0.99**	0.94	0.97	**0.98**	0.94	0.97
bg		1.00	0.99	0.93	**0.99**	0.98	0.93	**0.99**	0.97
ag + *bg*			1.00	0.97	0.99	**0.99**	0.97	0.99	**0.99**
ag				1.00	0.95	0.98	**1.00**	0.95	0.98
bg					1.00	0.99	0.95	**1.00**	0.99
ar + *br*						1.00	0.98	0.99	**1.00**
ap							1.00	0.95	0.98
bp								1.00	0.99
ap + *bp*									1.00
				CYCLIRAMI/antihistamine					
ag	1.00	0.82	0.94	**0.98**	0.80	0.91	**0.97**	0.80	0.92
bg		1.00	0.96	0.81	**0.99**	0.96	0.76	**0.99**	0.95
ag + *bg*			1.00	0.92	0.95	**0.98**	0.88	0.95	**0.98**
ar				1.00	0.81	0.92	**0.99**	0.81	0.93
br					1.00	0.97	0.76	**1.00**	0.96
ar + *br*						1.00	0.89	0.96	**1.00**
ap							1.00	0.77	0.90
bp								1.00	0.96
ap + *bp*									1.00
				DIAZEPAM/tranquilizer					
ag	1.00	0.65	0.90	**0.98**	0.67	0.92	**0.97**	0.68	0.91
bg		1.00	0.90	0.60	**0.98**	0.85	0.61	**0.98**	0.86
ag + *bg*			1.00	0.87	0.91	**0.98**	0.87	0.91	**0.98**
ar				1.00	0.67	0.92	**0.99**	0.65	0.91
br					1.00	0.90	0.68	**0.99**	0.90
ar + *br*						1.00	0.92	0.89	**0.99**
ap							1.00	0.67	0.91
bp								1.00	0.90
ap + *bp*									1.00
				APOMORPH/dopaminergic					
ap	1.00	0.84	0.94	**0.98**	0.81	0.91	**0.96**	0.82	0.91
bg		1.00	0.97	0.84	**0.98**	0.96	0.85	**0.99**	0.96
ag + *bg*			1.00	0.93	0.95	**0.98**	0.93	0.96	**0.98**
ar				1.00	0.84	0.94	**0.99**	0.84	0.93
br					1.00	0.97	0.85	**0.99**	0.97
ar + *br*						1.00	0.95	0.97	**1.00**
ap							1.00	0.85	0.95
bp								1.00	0.97
ap + *bp*									1.00
				DIETHYLST/estrogen					
ag	1.00	0.26	0.71	**0.99**	0.26	0.68	**0.98**	0.26	0.66
bg		1.00	0.82	0.32	**0.99**	0.83	0.32	**0.98**	0.83
ag + *bg*			1.00	0.76	0.82	**0.99**	0.75	0.81	**0.98**
ar				1.00	0.32	0.74	**0.99**	0.32	0.72
br					1.00	0.84	0.32	**0.99**	0.84
ar + *br*							0.73	0.83	**0.99**
ap							1.00	0.32	0.73
bp								1.00	0.84
ap + *bp*									1.00
				RS-86/parasympathomimetic					
ag	1.00	0.33	0.69	**0.99**	0.31	0.64	**0.97**	0.30	0.63
bg		1.00	0.89	0.36	**1.00**	0.91	0.36	**0.98**	0.90
ag + *bg*			1.00	0.72	0.88	**0.99**	0.70	0.85	**0.97**
ar				1.00	0.35	0.68	**0.98**	0.33	0.66
br					1.00	0.91	0.35	**0.98**	0.90
ar + *br*						1.00	0.67	0.88	**0.98**
ap							1.00	0.35	0.68
bp								1.00	0.91
ap + *bp*									1.00
				SEROTONIN/serotoninergic					
ag	1.00	0.87	0.96	**0.99**	0.85	0.95	**0.99**	0.86	0.95
bg		1.00	0.97	0.85	**0.97**	0.95	0.86	**0.99**	0.97
ag + *bg*			1.00	0.94	0.95	**0.99**	0.94	0.96	**0.99**
ar				1.00	0.82	0.93	**0.99**	0.84	0.94
br					1.00	0.97	0.83	**0.96**	0.94
ar + *br*						1.00	0.93	0.95	**0.98**
ap							1.00	0.85	0.95
bp								1.00	0.97
ap + *bp*									1.00

Table 3 (Continued)

	ag	*bg*	*ag* + *bg*	*ar*	*br*	*ar* + *br*	*ap*	*bp*	*ap* + *bp*
GABOXADOL/gabaminergic									
ag	1.00	0.87	0.96	**1.00**	0.71	0.77	**0.99**	0.86	0.95
bg		1.00	0.97	0.71	**1.00**	0.99	0.86	**0.99**	0.97
ag + *bg*			1.00	0.78	0.99	**1.00**	0.94	0.96	**0.99**
ar					0.72	0.78	**0.99**	0.72	0.78
br						0.99	0.69	**1.00**	0.99
ar + *br*							0.76	0.99	**1.00**
ap							1.00	0.85	0.95
bp								1.00	0.97
ap + *bp*									1.00

substructure algorithm[9,10,16] has come close. Geometric atom pairs seem to be a useful descriptor for 3D searches. There have been a few cases[14−16] where an "atom triplet" (three atoms and the three distances between them) has been used as a descriptor. A triplet is closer to a minimal 3D pharmacophore, and it takes some account of correlations between distances. One limit to the use of triplets, however, is that the time to parse a molecule into triplets varies as the cube of the number of atoms, as opposed to the square of the number of atoms for atom pairs. Also, the number of unique triplets per molecule can get very large unless one uses very wide distance bins and/or very few atom types,

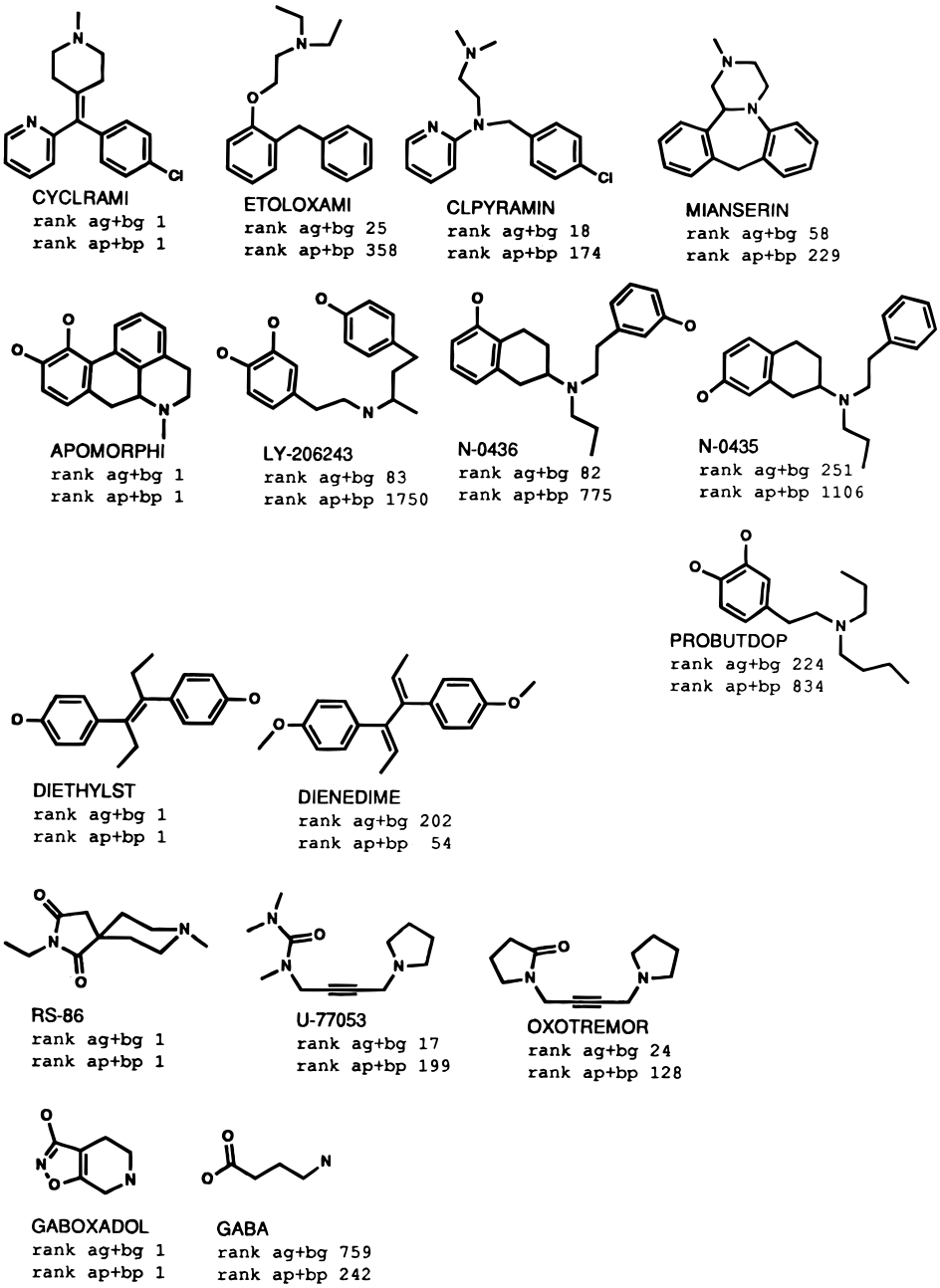

Figure 5. Selected compounds from Table 4A that rank very differently on geometric versus topological descriptors (*ag* + *bg* versus *ap* + *bp*).

Table 4

A. Compounds Very Different in Rank Among "Flexible" Geometric and Topological Descriptors ($ag + bg$ versus $ap + bp$)

probe	name	rank $ag + bg$	rank $ap + bp$	−log (slope)
MORPHINE	AZIDOMORP	269	64	−0.62
CYCLIRAMI	ETOLOXAMI	25	358	1.12
	ETOLOXHCL	34	357	1.02
	CLPYRAMIN	18	174	0.99
	CHLPYRHCL	21	168	0.90
	WV-761	49	284	0.76
	CLEMIZOLE	124	689	0.75
	MIANSERIN	58	229	0.60
	MIANSEHCL	60	228	0.58
	CLEMIZSUL	183	688	0.58
	PHENINDAM	40	147	0.58
	PHETOLCIT	72	264	0.56
	PHENINTRT	41	149	0.56
	BROMTRIPE	86	308	0.55
DIAZEPAM	DEMOXEPAM	239	51	−0.67
APOMORPHI	LY-206243	83	1750	1.32
	RU-24926	96	1624	1.23
	N-0436	82	775	0.98
	N-0936	69	377	0.74
	N-0438	163	798	0.69
	N-0923	184	841	0.66
	N-0434	242	1105	0.66
	N-0924	185	839	0.66
	N-0435	251	1106	0.64
	N-0734	191	803	0.62
	N-0437	208	838	0.61
	PROBUTDOP	224	834	0.57
	JMB-249	114	408	0.55
DIETHYST	DIENEDIME	202	54	−0.57
RS-86	U-77053	17	199	1.07
	BM-5	39	209	0.73
	OXOTREMOR	24	128	0.73
	BM-130	118	609	0.71
	BM-130-A	94	456	0.69
	EK-17	28	108	0.59
	CYTISINE	183	54	−0.53
SEROTONIN	none			
GABOXADOL	GABA	759	242	−0.50

B. Compounds Very Different in Rank Among "Rigid" Geometric and Topological Descriptors ($ar + br$ versus $ap + bp$)

probe	name	rank $ar + br$	rank $ap + bp$	−log (slope)
MORPHINE	AZIDOMORP	232	64	−0.56
CYCLIRAMI	ETOLOXHCL	38	357	0.97
	MIANSERIN	39	229	0.76
	CLPHENTAN	10	56	0.75
	MIANSEHCL	41	228	0.75
	PHENINDAM	29	147	0.71
	PHENINTRT	30	149	0.70
	CLPHENAM	11	54	0.69
	ETOLOXAMI	85	358	0.62
	PHENYLTOL	69	262	0.58
	PHETOLCIT	71	264	0.57
	CARBINOXA	77	22	−0.54
	ROTOXAMIN	68	19	−0.55
DIAZEPAM	DEMOXEPAM	176	51	−0.54
APOMORPHI	PROBUTDOP	165	834	0.70
DIETHYST	none			
RS-86	none			
SEROTONIN	none			
GABOXADOL	HAGABA	358	112	−0.51

and this is the case for the applications that use triplets, for example, shape similarity.[15] We could not use triplets with the original definition of atom types, for which there are >40 possible types.

It should be noted that distance-based descriptors, of which geometric atom pairs and geometric triplets are examples, do not distinguish between enantiomers. While this would be a problem for superposition methods, it is not necessarily bad for database searches.

Whether descriptor-based or not, 3D similarity searches are more expensive than 2D searches because of the complications of molecular flexibility. First, the user must supply the active conformation of the probe, derived from experimental data or deduced by the active analog approach. Second, if one allows for the possibility of flexibility in the database, one must either generate and store conformations beforehand (as we have done here) or allow for on-the-fly flexing of the database entries. Either can be quite time-consuming. The extra expense is justified if 3D similarity searches find more actives than 2D searches or if the 3D searches turn up unexpected classes of actives.

It is clear that geometric atom pairs do not find more actives than topological atom pairs. Also, we were surprised to see how similar the overall ranking by geometric atom pairs was to the ranking by topological atom pairs. Indeed, the ranks are very much more sensitive to changing the definition of the atom type, as was seen in our companion paper, than changing the definition of distance from through-bond to through-space. In retrospect, this is not surprising because through-space distance is roughly proportional to through-bond distances for most pairs of atoms in most druglike molecules, even when the conformations of the molecule are not particularly extended. However, there are exceptions. Through-space descriptors will find different compounds than through-bond descriptors under the following conditions: 1. the active conformation of the probe is compact rather than extended, 2. the database entries are compact rather than extended, and 3. the probe and database entry, although they are both in an extended conformation, have similar groups connected by different bond paths.

Condition 1 is likely to occur only with fairly large probe molecules wherein compact structures are stable, e.g., a hexapeptide bent into a hairpin. Small drug molecules are not likely to have a compact active conformation. Many studies in the literature use databases with only a single conformation per structure. Since these conformations are usually more or less extended, circumstance 2 is not expected. Condition 2 is much more likely to occur if the database contains many conformations per structure, as does the geobase used here. We saw examples of this with the APOMORPHI and RS-86 probes. Condition 3 can occur in databases with both single and multiple conformations, and we found some examples with our CYCLIRAMI probes.

The fact that geometric descriptors occasionally select compounds that are not selected by topological methods argues for the usefulness of these descriptors. Moreover the fact that more such compounds are found in databases that contain multiple conformations argues for the inclusion of "flexibility" in the database.

REFERENCES AND NOTES

(1) Willet, P. *Similarity and Clustering in Chemical Information Systems.* Research Studies Press, Ltd., John Wiley & Sons: New York, 1987.

(2) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.

(3) Ricketts, E. M.; Bradshaw, J.; Hann, M.; Hayes, F.; Tanna, N.; Ricketts, D. M. Comparison of conformations of small molecule structures from the protein data bank with those generated by CONCORD, COBRA, ChemDBS-3D, and Converter and those

extracted from the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 905−925.
(4) Hendrickson, M. A.; Nicklaus, M. C.; Milne, W. A.; Zaharevitz, D. CONCORD and CAMBRIDGE: comparison of computer-generated chemical structures with X-ray crystallographic data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 155−163.
(5) Nicklaus, M. C.; Milne, W. A.; Zaharevitz, D. Chem-X and CAMBRIDGE: comparison of computer generated chemical structures with X-ray crystallographic data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 639−646.
(6) Perry, N. C.; van Geerestein, V. J. Database searching on the basis of three-dimensional molecular similarity using the SPERM program. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607−616.
(7) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Similarity screening of molecular data sets. *J. Comput.-Aided Molec. Design* **1992**, *6*, 513−520.
(8) Burt, C.; Richards, W. G. Molecular similarity: the introduction of flexible fitting. *J. Comput.-Aided Molec. Design* **1990**, *4*, 231−238.
(9) Pepperrell, C. A.; Willet, P. Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. *J. Comput.-Aided Molec. Design* **1991**, *5*, 455−474.
(10) Artymiuk, P. J.; Bath, P. A.; Grindley, H. M.; Pepperrell, C. A.; Poirette, A. R.; Rice, D. W.; Thorner, D. A.; Wild, D. J.; Willet, P.; Allen, F. H.; Taylor, R. Similarity searching in databases of three-dimensional molecules and macromolecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 617−630.
(11) Carhart, R. E.; Smith, D. H.; Ventkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(12) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsions: a new molecular descriptor for SAR applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(13) Bemis, G. W.; Kuntz, I. D. A fast and efficient method for 2D and 3D molecular shape description. *J. Comput.-Aided Molec. Design* **1992**, *6*, 607−628.
(14) Fisanick, W.; Cross, K. P.; Rusinko, III, A. Similarity searching on CAS registry substances. 1. global molecular property and generic atom triangle geometric searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.
(15) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New method for rapid characterization of molecular shapes: applications in drug design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.
(16) Bath, P. A.; Poirrette, A. R.; Willet, P. Similarity searching in files of three-dimensional chemical structures: comparison of fragment-based measures of shape similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.
(17) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, xxxx.
(18) 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.
(19) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: a way to enhance the use of molecular docking methods. *J. Comput.-Aided Molec. Design* **1994**, *8*, 565−582.
(20) Standard Drug File Version 6.0; developed and distributed by Derwent Publications Ltd.: London, England, 1991.
(21) Molecular Access System; Molecular Design, Ltd.: San Leandro, CA.
(22) Rusinko, A.; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. CONCORD: a program for the rapid generation of high quality approximate 3D structures. University of Texas at Austin: CONCORD distributed by Tripos Associates Inc.: St. Louis, MO 63144.

CI950275B