The Centroid Approximation for Mixtures: Calculating Similarity and Deriving Structure−Activity Relationships

Robert P. Sheridan†

Department of Molecular Systems, RY50S-100 Merck Research Laboratories, P.O. 2000, Rahway, New Jersey 07065

Received May 17, 2000

Compounds are often synthesized and tested as mixtures. We propose the idea that the descriptor representation of a mixture may be approximated as the descriptor average of its individual component molecules. This centroid approximation has several potential advantages: the representation is very compact, calculating similarities and deriving structure−activity relationships (SARs) of mixtures involves very little computation, and existing software can be directly applied to mixtures as if they were single molecules. Here we use the atom pair and topological torsion descriptors. We run several types of simulations using mixtures composed of druglike molecules from the MDL Drug Data Report database. We show that similarity searches using mixtures as queries and/or database entries yield reasonable results, with the caveat that a correction is necessary for mixture−mixture comparisons where at least one of the mixtures contains very diverse molecules. We also show that predictive SARs in the form of trend vectors can be derived from mixtures.

INTRODUCTION

Traditionally compounds have been synthesized and tested as single pure entities. With the advent of mix-and-split combinatorial chemistry methods,[1−3] it is now very common for many compounds to be synthesized simultaneously in a single container. These mixtures may contain up to several thousand distinct chemical entities. Also, combinatorial products are often tested on biological assays as mixtures. Even when compounds are synthesized individually, it is a common practice in high-throughput screening (HTS) to deliberately mix, say 10 compounds per sample, so that the number of samples that must be tested in an assay is reduced by a factor of 10. One drawback of testing mixtures is that, once a mixture is found to be active on an assay, one must then find out which individual molecule(s) is responsible for the activity. The "deconvolution" of combinatorial mixtures or "resolution" of a HTS mixture may be the rate-limiting step in some lead-discovery projects.

Some pharmaceutical companies maintain large databases of combinatorial or HTS mixtures and biological data associated with them. One often needs to ask questions such as, What other mixture in my database contains compounds similar to this mixture that has just shown an interesting activity" or Is there any discernible quantitative structure−activity relationship (QSAR) from this set of mixtures" These are the kinds of questions that have been addressed for single compounds by similarity methods[4,5] and QSAR methods for a long time. It would be very useful to extend those methods to mixtures. However, directly applying these methods to the single compounds within mixtures requires the structures of molecules in these mixtures to be explicitly stored, which is costly in disk space, or generated de novo from generic structures or reaction schemes, which is time-consuming.

In this paper we will propose a very simple approximation to address this issue. Given a substructure descriptor representation, one may approximate the descriptor representation of a mixture as the "centroid" of the molecules contained in it. Representing a set of compounds by a descriptor centroid has been used extensively in the past as an intermediate stage in clustering compounds[6] or in selecting a diverse subset.[7] Here we will treat the centroid of a mixture as a "pseudomolecule", so that mixtures can then be handled by the same computational machinery that handles single molecules.

We use druglike compounds from a large database, partitioned into mixtures by the activity of the compounds or by structural classes, to do two series of simulations. The first series is meant to simulate the situations where either a single molecule or a mixture is used as a query, and similarity searches are done over a database of diverse single molecules or a database of mixtures. We show that the results appear to be reasonable and potentially useful for several examples, with the caveat that some correction is necessary for mixture−mixture comparisons of very diverse mixtures. The second series of simulations tests whether QSAR in the form of trend vectors can be extracted from mixtures and used to predict the activities of single compounds. We look at two situations. The first is where the mixtures are of various sizes and composed of very similar compounds, analogous to experiments where combinatorial mixtures are tested for biological activity. The second situation is where the mixtures are of uniform size and composed of arbitrarily selected compounds, analogous to some HTS experiments. In both situations, we show it is possible to derive a predictive QSAR.

METHODS

Descriptors. We assume a molecule is represented by a list of substructure descriptors and their frequencies. Two

† E-mail: sheridan@merck.com. Telephone: 732-594-3859. FAX: 732-594-4224.

10.1021/ci000045j CCC: $19.00
Published on Web 09/13/2000

descriptors we have found very useful are the atom pair (AP)[8] and topological torsion (TT),[9] and here we use them exclusively. Details on the definition and behavior of these descriptors have been given in a previous paper.[10]

Similarity Definitions. We define the centroid *C* of a mixture as the descriptor average of all the molecules contained therein:

$$f_{Ck} = \sum_{i}^{M} f_{ik}/M \qquad (1)$$

where f_{ik} is the frequency of descriptor *k* in molecule *i*. *M* is the number of compounds in the mixture. The centroid of a single molecule is, of course, the molecule itself. It is assumed here that the molecules are represented in equimolar proportions, but the definition is trivially extendable to unequal mixtures. Note that the notation for the centroid, a list of descriptors and their frequencies, is indistinguishable from that of a single molecule. Later we will make reference to a special centroid *U*, which is taken as the description of the universe of druglike compounds.

One measure we can take of a mixture is its "internal similarity", which we can define as the mean similarity of each individual compound *i* with the centroid of the mixture:

$$\mathrm{IS}_C = \sum_{i}^{M} \mathrm{Sim}_{iC}/M \qquad (2)$$

Later we will show the relationship of the IS of a mixture to its similarity to *U*.

We have not yet defined Sim_{ij}, where *i* and *j* are molecules or centroids. Three popular definitions of descriptor-based similarity are Dice, cosine, and Tanimoto.[4,5] For each, similarity may range from 0 (nothing in common) to 1 (identical). In our laboratory, we generally work with the Dice definition:

$$\text{Dice Sim}_{ij} = 2\sum_{k}\min(f_{ik}, f_{jk})/(\sum_{k} f_{ik} + \sum_{k} f_{jk}) \qquad (3)$$

and the reader should assume Dice similarity unless otherwise stated. We have found that for similarity searches on single molecules, Dice gives better results than cosine.[11] For any given query, Tanimoto is monotonic with Dice.[5]

Similarity Corrections for Very Diverse Mixtures. In examining pairwise similarities of mixtures we often see an artifact that does not occur for single molecules. These are "false hits" where some mixture pairs have Sim's approaching 1, although a chemist would not think there is a real similarity among the individual compounds making them up. At least one of the members of the pair in a false hit is usually a very large or diverse mixture. We believe the false hits are due to the following: In the limit of very high diversity, the centroid of a mixture tends to approximate *U*, the centroid of druglike compounds. As would be true of any three points in similarity space, the more mixtures *i* and *j* resemble *U*, the more they will tend to resemble each other. One heuristic to correct Sim_{ij} so that the false hits are eliminated is to compare the actual value of Sim_{ij} with the expected value based on the similarity of *i* and *j* to *U*. As will be shown in the Results section, there is a linear relationship:

$$\text{expected Sim}_{ij} = c1 \cdot \mathrm{Sim}_{iU}\,\mathrm{Sim}_{jU} + c0 \qquad (4)$$

The scatter of points above and below the line represented by this equation will be summarized as the root-mean-square error σ. A correction we have found useful is

$$\mathrm{cSim}_{ij} = \mathrm{Sim}_{ij} + \text{penalty} \qquad (5)$$

where penalty = 0 when $Z \geq 3.0$ and $C\,(Z - 3.0)$ otherwise.

$$Z = (\mathrm{Sim}_{ij} - \text{expected Sim}_{ij})/\sigma \qquad (6)$$

That is, if the similarity is far above (3 standard deviations or more) the expected "background" value for the two mixtures, leave the score alone, but exact an increasing penalty as it approaches or goes below the expected value. The value $C = 0.1$ was found by trial and error and is used throughout. This correction is sufficient to eliminate the most egregious false hits in similarity searches. It should be noted that negative cSim's are possible, but this is acceptable since in similarity searches we are more interested in the ordering of database entries than their absolute similarities, and we are usually interested only in the most similar entries.

Clustering of MDDR Compounds and Selection of Diverse Compounds. For some experiments we will need to simulate mixtures of compounds closely related by structure and, for other experiments, sets of diverse single compounds. The rapid method we use for partitioning sets of single compounds into clusters and selecting diverse subsets is the following:

1. For all compounds *i*, find neighbors *j* within a similarity cutoff. Our method for doing this rapidly is an implementation (S. K. Kearsley, unpublished) of the method of Holliday et al.,[7] which uses cosine similarity. Here we use the TT descriptor, and a cosine similarity cutoff of 0.85 (roughly equivalent to a Dice similarity of 0.70).

2. Sort the compounds *i* in order of decreasing number of neighbors. Each compound *i* is a candidate "exemplar", or most representative compound for a cluster.

3. For every exemplar *i*, eliminate any remaining candidate examplars *j* further down the list that are more similar to *i* than the cutoff. This gives a smaller set of final exemplars.

4. Although most compounds *j* are eliminated as exemplars, they still remain neighbors of one or more of the final exemplars. Assign each compound *j* to the exemplar *i* to which it is most similar.

The end result of this procedure is a set of mutually exclusive clusters, each with the member with the most neighbors designated as the exemplar. To extract a diverse subset of the original set of compounds, one simply takes the exemplar from each cluster.

Similarity Search Experiments. In a conventional similarity search one selects an interesting molecule as a query or "probe" and calculates the similarity of the probe to every entry in a database. The database entries are then sorted in order of decreasing similarity, the most similar being rank 1, the second rank 2, etc. Generally we are interested in the relative ranking of entries rather than their absolute similarity. One application of similarity searches is to find entries with activity similar to that of the probe, which should be toward the front of the list if similarity is predictive of activity. Traditionally both the probe and database entries are single molecules. With the centroid approximation, one may make

either the probe or the database entries a mixture, and there are three combinations not previously explored: mixture probe/mixture database, single-molecule probe/mixture database, and mixture probe/single-molecule database.

QSAR by Trend Vector. We will use our implementation of trend vector analysis[8,12] as a descriptor-based QSAR method. This method uses a sample-based partial least-squares (PLS) approach[12] to summarize the correlation of descriptors with biological activity for large sets of samples. (Traditionally each sample is a single molecule.) This analysis needs the descriptors for each sample and the measured activity. Only the presence or absence of a descriptor is used in this method; the frequencies are ignored. To avoid overfitting, a randomization method is used to ensure that the correlation is statistically significant during the fit of each PLS component. Our convention is that a PLS component is significant if its length is ≥ 3 standard deviations above the length expected by chance. We generally allow up to five PLS components. The final trend vector has a coefficient associated with each descriptor in the training set. The descriptors with positive coefficients are associated with active molecules, and the descriptors with negative coefficients are associated with inactive molecules. For most applications, the length of the trend vector is not important, only its "direction".

The trend vector can be used to calculate the predicted activity of a molecule outside the training set; the predicted activity is the projection of the new molecule's descriptors onto the trend vector, again using only the presence or absence of descriptors in the new molecule. Before a prediction can be made, we require at least 95% (AP) or 85% (TT) of the unique descriptors in the new molecule must be present in the training set from which the trend vector was derived. For the purposes of this exercise, molecules that do not meet this criterion are given an arbitrarily low score, unless otherwise stated. Details of how trend vectors are calculated and how predictions are made are in ref 12. In this paper the samples may be mixture centroids as well as single molecules, either for training the trend vector or for prediction. For a mixture, we count a descriptor as "present" if it is in any molecule of the mixture.

The vector format makes combining and comparing trend vectors straightforward. Related trend vectors may be averaged by normalizing each vector, and then averaging the individual components. The similarity of two trend vectors may be represented as the cosine of the angle θ between them, equivalent to taking the dot product of the normalized vectors. Two trend vectors pointing in the same direction have $\cos\theta = 1$. Given the high dimensionality of the vectors (typically hundreds to thousands of unique descriptors), any two arbitrary trend vectors will have a $\cos\theta$ very close to zero.

In the QSAR experiments a training set of either single molecules or mixture centroids is defined and trend vector calculation extracts what descriptors are associated with a particular activity. Then trend vector is used to select actives from a database of diverse single compounds.

Measures of Goodness for Searches. The ability of each similarity probe or trend vector to select active compounds is measured as a simulated screening experiment:

1. The similarity or predicted activity of each of the database entries is calculated and the entries are sorted in decreasing order of this value. Compounds are "tested" for activity in this order.

2. The cumulative number of actives is monitored as a function of number of compounds tested. This usually gives a hyperbolic "accumulation curve" (see ref 10). Assume there are D database entries. Let A_{50} be the number of entries tested at which half the actives are found. If the similarity probe or trend vector were useless in predicting activity, i.e., the actives were randomly distributed in the list, A_{50} would be approximately equal to $D/2$. In practice, half the actives are found much earlier. The "global enhancement" is the ratio of $D/2$ divided by A_{50}. The higher the enhancement, the better the prediction. The maximum possible global enhancement is D/n_{actives}. A global enhancement of 1 indicates no selection above chance.

Source of Compounds. For this paper we need a source of nonproprietary druglike compounds that are grouped into sets of various sizes and diversities. For this we use Version 98.1 of the MDL Drug Data Report (MDDR),[13] a licensed database derived from the patent literature. This contains ~90 000 entries, about 94% of which have chemical structures from which we can generate descriptors. Of these ~65 000 have been assigned one or more therapeutic categories. The therapeutic categories are named as a five-digit integer and a title (e.g., 42711 CCK Antagonist). To avoid confusion with molecules, which are named by six-digit integers, we prepend the therapeutic category with a T, e.g., T42711. There are 658 unique therapeutic categories, of which 558 have at least one molecular structure associated with them. Many molecules are in more than one therapeutic category. For the some of the simulations, we will take molecules in a specified therapeutic category as "active" in that area, and all others as "inactive". Some therapeutic categories are problematic in that they include subsets of very diverse molecules that probably work by different mechanisms.

Sets of Mixtures and Single Compounds. For the similarity search experiments we will use the 558 therapeutic categories as sample mixtures. Therefore, we calculated the centroid for each. We will refer to these as the "therapeutic category centroids" (the TC set). We take the centroid of the 65 000 MDDR compounds with a therapeutic category as an approximation of U.

Some experiments require mixtures of closely related compounds and diverse singles. We randomly divided the 65 000-compound MDDR in two (a "training half" and a "search half") and clustered the training half as described above into 14 161 clusters. As with most databases of druglike compounds, there were a few large clusters (maximum $M = 197$) and roughly half of the clusters (8385) were singletons ($M = 1$). We calculated the centroid for each cluster. We will refer to these as the "cluster centroids" (the CC set). Note that the clustering depends only on chemical structure and is independent of therapeutic category. We extracted the exemplars to obtain what we will call the "diverse training singles" (the TS set). A set of 14 116 diverse search molecules (the SS set) was also extracted in the same way from the search half.

RESULTS

Calibration of cSim. A plot of IS_i for centroid i against $\log M$ for the TCs is shown in Figure 1A. The TCs with the

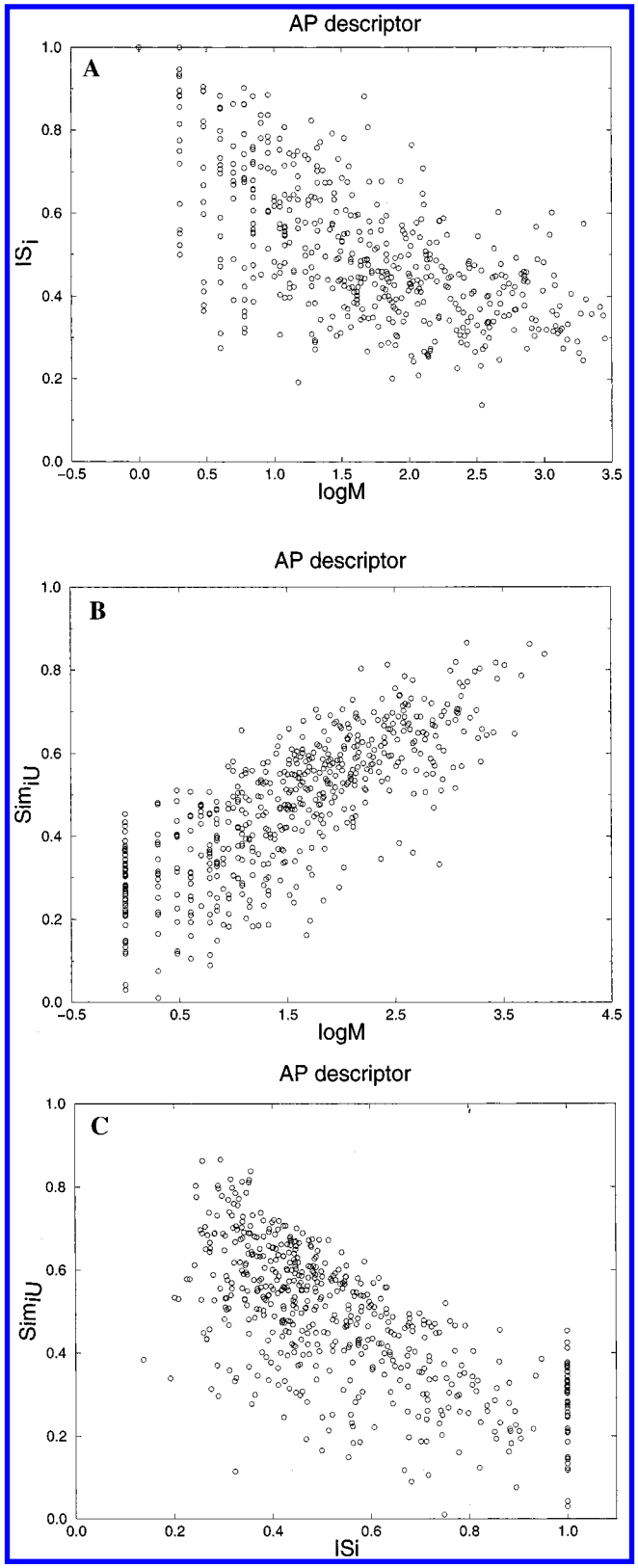

Figure 1. Distribution of similarity properties for the therapeutic category centroids. The AP descriptor is shown. (A) IS_i vs log M. If $M = 1$, IS_i must be 1.0 by definition. (B) Sim_{iU} vs log M for the therapeutic category centroids. U is approximated by the centroid of ~65 000 MDDR compounds. (C) Sim_{iU} vs IS_i.

highest Ms tend to have the lowest ISs; that is, they are more internally diverse. Single compounds (log $M = 0$) have IS $= 1$ by definition. As might be expected, Figure 1B shows that categories with the largest M's also have high values of

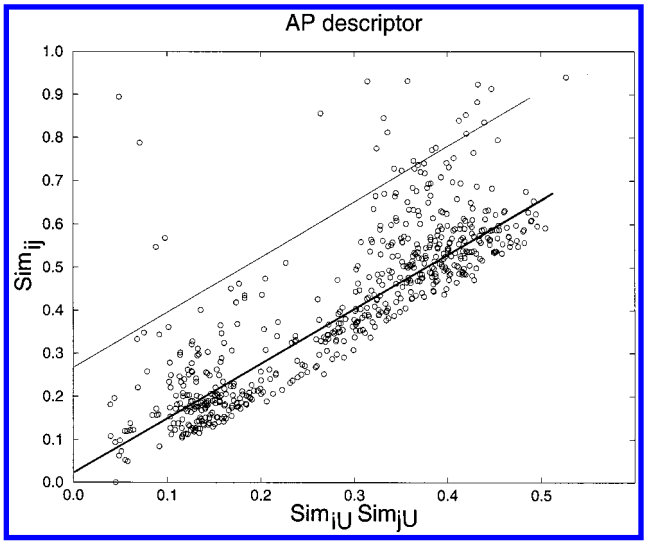

Figure 2. AP Sim for arbitrary pairs of therapeutic category centroids as a function of the product of the similarity of the centroids with the centroid of compounds from the MDDR. The bold line represents the expected Sim. The thin line represents 3 σ above the expected Sim.

Sim_{iU}; that is, they more closely resemble U. Although there is some scatter, single compounds generally have lower values. Figure 1C shows the relationship between Sim_{iU} and IS_i. The TCs with the highest internal diversity most closely resemble U.

Six hundred pairs from the TC set were selected arbitrarily for the calibration of cSim. Sim_{ij} and $Sim_{iU}Sim_{jU}$ were calculated for each pair. The calibration plot for the AP descriptor is shown in Figure 2. Most of the points fall near a line. These represent "background" similarities, and the outliers represent what we would consider interesting similarities discernible from the background. Expected Sim_{ij} was found by standard linear regression once obvious outliers were eliminated. The regression parameters are

$$\text{AP Dice } c1 = 1.31 \quad c0 = 0.025 \quad \sigma = 0.080$$

$$\text{TT Dice } c1 = 1.41 \quad c0 = 0.027 \quad \sigma = 0.057$$

Similarity Search 1: Mixture Probe/Mixture Database. Here we simulate searching for mixtures which are similar to a probe mixture.

Three therapeutic categories T42711 (CCK Antagonist), T31430 (Angiotensin II Blocker), and T75711 (Antiestrogen) are shown as examples. Figure 3 shows representative structures from these therapeutic categories. There is more than one class of compounds in each. For instance, in T42711 there are 3-substituted benzodiazepine analogues and indole-containing molecules, many of which are Trp-containing peptides. Each of the three centroids C-T42711, C-T31430, and C-T75711 was taken as a probe, and we calculated Sim and cSim for each TC against each probe.

The usual method (e.g., ref 10) of assigning database entries as "actives" based on their having the same therapeutic category as the probe cannot work here since all the database entries by definition are distinct therapeutic categories. Instead we took a subjective approach. For each probe TC, we inspected a randomly selected subset of compounds from each therapeutic category and made a judgment whether the majority of the compounds "resemble"

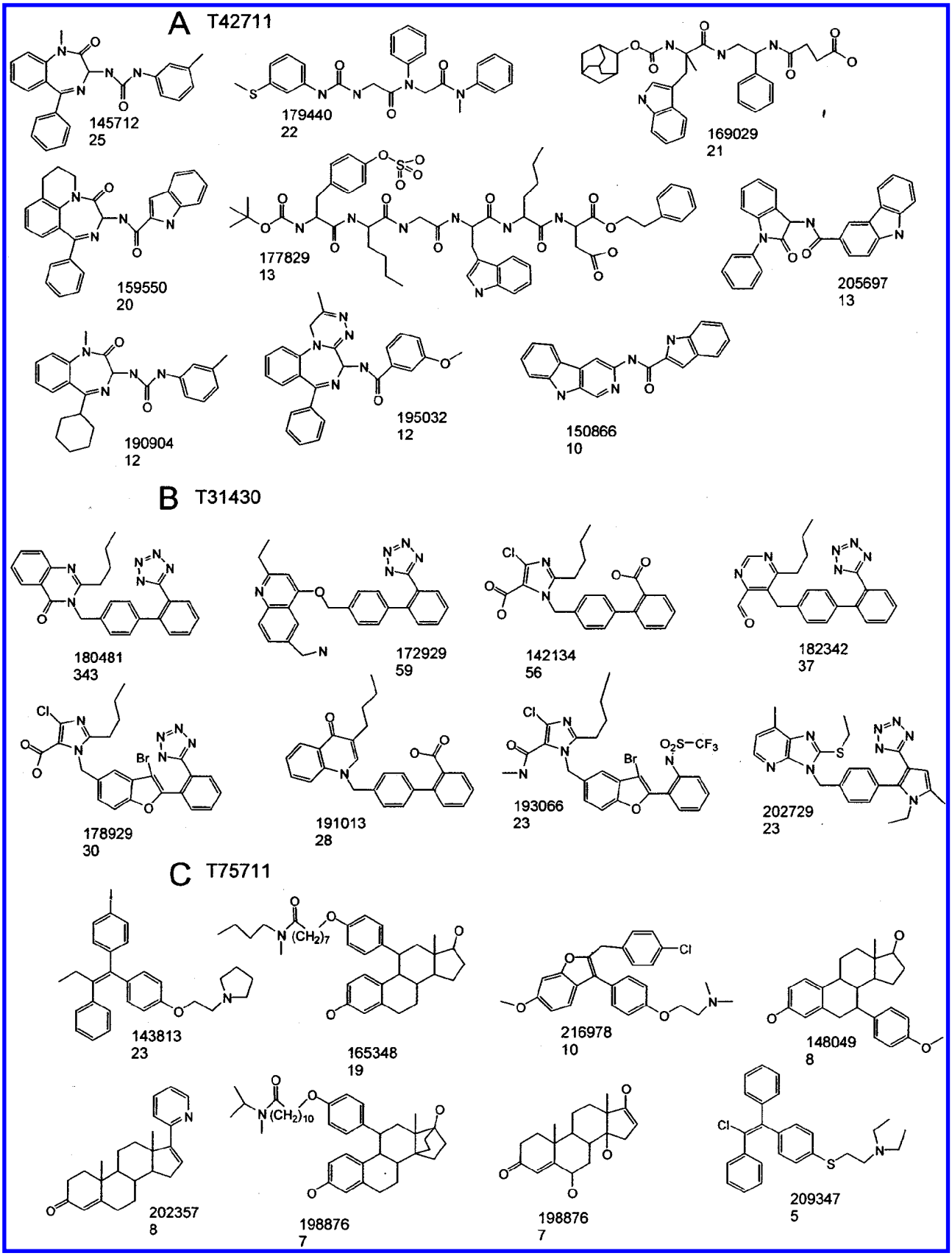

Figure 3. Representative compounds for three therapeutic category probes in Similarity Search 1. Each set of compounds was clustered. The figures show the exemplars from the largest clusters. Listed are the name of the exemplar and the number of compounds in the cluster. (A) T42711. (B) T31430. (C) T75711.

the compounds in the probe. For C-T42711 as the probe we looked for other therapeutic categories that contained 3-substituted benzodiazepines or Trp-containing peptides. For C-T31430 we looked for categories that contained compounds resembling a biphenyl acid with an aromatic "headgroup". For C-T75711 we looked for categories containing

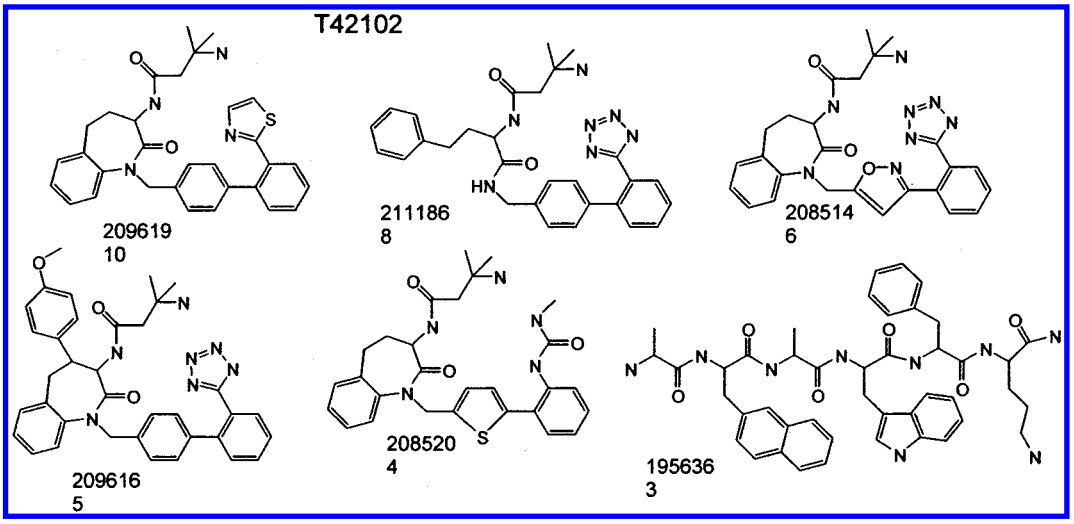

Figure 4. Representative compounds for T42102 (Growth Hormone Release Promoting Agent), which contains molecules that resemble those from T42711 (because of 3-substituted benzodiazepine-like rings) and T31430 (because of the biphenyl rings).

Table 1. Measures of Goodness for Similarity Searches

				global enhance. cSim		global enhance. Sim	
probe	search over	actives	$n_{actives}$	AP	TT	AP	TT
		Similarity Search 1					
C-T42711	TC[a]	benzodiazepine/	11	55	55	46	55
		indole-containing	8	5	8	2	2
C-T31430	TC	biphenyl-containing	9	69	69	55	69
C-T75711	TC	steroid/	22	10	21	5	8
		tamoxifen-containing	8	7	10	2	3
		Similarity Search 2					
193319	TC	benzodiazepine-containing	11	28	46	28	46
180481	TC	biphenyl-containing	9	69	69	69	69
221588	TC	tamoxifen-containing	8	22	68	18	78
		Similarity Search 3					
C-T42711_training	SS[b]	T42711	79	59	71	59	71
C-T31430_training	SS	T31430	254	51	53	51	53
C-T75711_training	SS	T75711	22	35	37	25	23
		Similarity Search 4					
193319	SS	T42711	79	32	23	32	23
180481	SS	T31430	254	45	50	45	50
221588	SS	T75711	22	1	1	1	1
		T75711 Tamox.	4	3529	706	3529	706

[a] TC = therapeutic category centroids. [b] SS = diverse single "search" compounds.

steroids or tamoxifen-like compounds. Thus we were able to build an appropriate list of TC "actives" for the probes and calculate a global enhancement. Sometimes the lists of actives overlap. For instance, the active list for the probes C-T42711 and C-T31430 share the category T42102, which includes molecules with parts resembling 3-substituted benzodiazepines and parts resembling biphenyl acids. Typical members of T42102 are shown in Figure 4.

The global enhancements for the three mixture probes are shown in Table 1. If there was more than one type of possible active, each was treated as a separate actives list. For instance, for C-T42711 there was a list of 11 benzodiazepine-containing categories and a list of 8 indole-containing categories. Clearly, the global enhancements are $\gg 1$, indicating a good selection of actives. Looking at these and other examples, we generally find that cSim gives better search results than Sim, especially for the AP descriptor. Thus we prefer to use cSim where both the probe and database entries are mixtures.

The 10 TCs most similar to each probe based on cSim are shown in Table 2. These contain a large proportion of actives, indicated by a structural notation (e.g., Benz) in the first column. Sometimes the actives are obviously related to the probe (e.g., T42714 and T42712 contain compounds in common with T42711), but some are unexpected. For instance, T42102 (Growth Hormone Release Promoting Agent) and T31341 (Vasopressin V1 Antagonist), like T42711, are G-protein coupled receptor activities, but not otherwise related to T42711.

As examples of false hits that are seen in Sim but are eliminated in cSim, we note that for the probe C-T75711, T52000 (Hypolipidemic), T52700 (ACAT inhibitor), T59300 (Antipsoriatic), and T31000 (Antihypertensive) are ranked very high at 2, 4, 5, and 6 for the AP descriptor (Sim 0.65, 0.63, 0.62, and 0.62). Figure 5 shows representative compounds for T752000. Clearly individual compounds in T52000 do not resemble those in T75711 (Figure 3C) despite T52000 being ranked as the second closest mixture to the

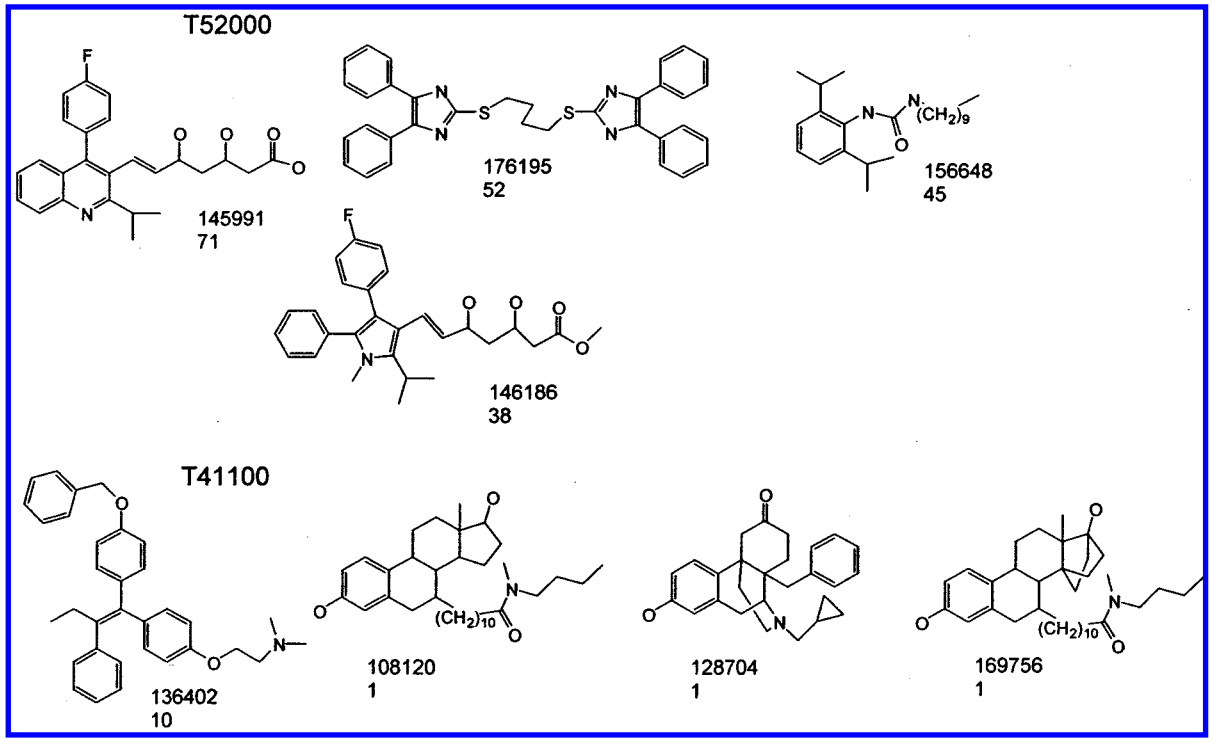

Figure 5. Representative compounds for T52000 (Hypolipidemic) and T41100 (Antiinfertility agent). T52000 is ranked most similar to T57511 (Figure 3A) by the AP Sim calculation. T41100 is ranked most similar to T75711 by the AP cSim calculation. Since individual molecules in T52000 clearly do not resemble those in T75711, it is an example of a false hit that is eliminated by calculating cSim.

probe (T75711 itself is rank 1). These are quite diverse mixtures (Sim_{iU} of 0.85, 0.92, 0.96, and 0.98) and the probe itself is diverse (Sim_{iU} 0.85), so high uncorrected similarities are expected. Correction changes the ranks to 46, 17, 86, and 120 (cSim 0.32, 0.37, 0.27, and 0.23). Once the false hits are eliminated, the list of the 10 most similar mixtures is dominated by steroid-containing compounds (Table 2C). T41100 (Antiinfertility Agent) is then the second closest mixture to the probe, and Figure 5 shows that T41100 is clearly a more reasonable match for T57511 than T52000.

We note from Tables 1, 2A, and 2C that if two structural classes are represented in a probe, the class in the majority is the one selected in the search. We see that C-T42711 selects benzodiazepine-containing compounds much more than Trp-containing peptides, and C-T75711 selects steroids over tamoxifen-like compounds. This is consistent with the observation that, of the molecules making up T42711, the ones most similar to C-T42711 are benzodiazepines. Similarly, of the molecules making up T75711, the most similar to C-T75711 are steroids.

Similarity Search 2: Single-Molecule Probe/Mixture Database. Here we simulate looking for a mixture that is similar to a single molecule. We arbitrarily selected as probes three molecules from the therapeutic categories in the last experiment. These are shown in Figure 6. Since the probes are from the same therapeutic categories and the database entries are TCs, the same list of TC actives is used as in Similarity Search 1.

Global enhancements are shown in Table 1. Looking at these examples and others from the next two experiments, we see that there is no consistent preference of Sim over cSim when either the probe or database entries are single molecules. Usually Sim and cSim give very similar global enhancements. This is not surprising. Single molecules have low values of Sim_{iU}, the product $Sim_{iU}Sim_{jU}$ is therefore low, and thus there would be few spuriously high values of Sim_{ij}. This reasoning also explains why corrections are not needed in similarity calculations where both probe and database entries are single molecules.

The TCs most similar to the probe in terms of Sim are shown in Table 3. In the case of 193319 and 180481, the therapeutic categories from which the probes were taken are among one or two most similar. This is not true for 221588 (tamoxifen), again because C-T75711 more closely resembles steroids than tamoxifen-like compounds.

Similarity Search 3: Mixture Probe/Single-Molecule Database. Here we simulate a screening experiment using a mixture as a probe and searching over a diverse database of single compounds. For the probe we took compounds in the training half of the MDDR belonging to a therapeutic category, say T42711, and constructed a centroid from them called C-T42711_training. For the database we used the SS set.

Here the "actives" in the database are those in the same therapeutic categories as the probe, for instance SS molecules that belong to T42711 would be considered actives for the probe 193319. The global enhancements are shown in Table 1 and the SSs most similar to the probes are shown in Figure 7. The global enhancement for T75711 TT is actually deceptively low. One can see many steroid-like compounds in Figure 7C that do not belong to the T75711 category, but could be considered very similar to T75711.

Similarity Search 4: Single-Molecule Probe/Single-Molecule Database. This experiment is a control for Similarity Search 3. The probes in Figure 6 are used to search the SS, and the set of actives is the same as before. The most important thing from this comparison is that the global enhancements from Similarity Search 4 are not higher than

Table 2. Most Similar Mixtures to C-T42711, C-T31430, and C-T75711 by CSim

therapeutic category	cSim	*M*	title
		(A) C-T42711	
		AP	
T42711 Benz[a]	1.00	427	CCK Antagonist
T42714 Benz	0.85	181	Gastrin Antagonist
T42712 Benz	0.82	87	CCK A Receptor Antagonist
T42713 Benz	0.82	121	CCK B Receptor Antagonist
T42102 Benz	0.59	42	Growth Hormone Release Promoting Agent
T57500 Benz	0.57	107	Pancreas Disorders, Agents for
T31341 Benz	0.55	18	Vasopressin V1 Antagonist
T37121	0.54	33	Factor Xa Inhibitor
T31430	0.53	1970	Angiotensin II Blocker
T42100 Benz	0.53	5	Growth Hormone
		TT	
T42711 Benz	1.00	427	CCK Antagonist
T42714 Benz	0.75	181	Gastrin Antagonist
T42712 Benz	0.74	87	CCK A Receptor Antagonist
T42713 Benz	0.74	121	CCK B receptor antagonist
T57500 Benz	0.60	107	Pancreas Disorders, Agent for
T42102 Benz	0.52	42	Growth Hormone Release Promoting Agent
T42710 Benz	0.52	90	CCK Cholecystokinin Agonist
T31433	0.50	40	Angiotensin II AT2 Antagonist
T31341 Benz	0.50	18	Vasopressin V1 Antagonist
T42100 Benz	0.46	5	Growth Hormone
		(B) C-T31430	
		AP	
T31430 Biphen[b]	1.00	1970	Angiotensin II Blocker
T31432 Biphen	0.71	58	Angiotensin II AT1 Antagonist
T31433 Biphen	0.68	40	Angiotensin II AT2 Antagonist
T09210 Biphen	0.65	32	Nootropic
T31341 Biphen	0.63	18	Vasopressin V1 Antagonist
T07709 Biphen	0.62	17	Neurotensin Receptor Antagonist
T27210	0.56	1183	Leukotriene Antagonist
T42714	0.56	181	Gastrin Antagonist
T27212	0.54	51	Leukotriene D4 Antagonist
T27220	0.53	58	Leukotriene Synthesis Inhibitor
		TT	
T31430 Biphen	1.00	1970	Angiotensin II Blocker
T31432 Biphen	0.69	58	Angiotensin II AT1 Antagonist
T31433 Biphen	0.64	40	Angiotensin II AT2 Antagonist
T09210 Biphen	0.60	32	Nootropic
T07709 Biphen	0.52	17	Neurotensin Receptor Antagonist
T42100 Biphen	0.49	5	Growth Hormone
T42102 Biphen	0.49	42	Growth Hormone Release Promoting Agent
T31341 Biphen	0.48	18	Vasopressin V1 Antagonist
T27210	0.45	1183	Leukotriene Antagonist
T02522	0.43	22	Leukotriene Synthesis Inhibitor
		(C) C-T75711	
		AP	
T75711 Steroid[c] Tamox[d]	1.00	146	Antiestrogen
T41100 Steroid Tamox	0.57	13	Antiinfertility Agent
T40300 Steroid	0.51	74	Contraceptive
T40231 Steroid	0.49	87	Progesterone Antagonist
T39500 Steroid	0.46	35	Adrenocortical Suppressant
T81220 Steroid	0.45	63	Antagonist to Narcotics
T40340 Steroid	0.43	32	Postcoital Contraceptive
T49110	0.42	129	Vitamin D Analog
T40210 Steroid	0.42	69	Estrogen
T33442	0.41	21	Peripheral Vascular Disease, Agent for
		TT	
T75711 Steroid Tamox	1.00	146	Antiestrogen
T40210 Steroid	0.57	69	Estrogen
T75721 Steroid	0.52	474	Aromatase Inhibitor
T41100 Steroid Tamox	0.51	13	Antiinfertility Agent
T59500 Steroid	0.50	865	Antiacne
T35560 Steroid	0.49	709	Prostate Disorders, Agent for
T40300 Steroid	0.49	74	Contraceptive
T40220 Steroid	0.48	58	Progestin
T40120 Steroid	0.48	416	Antiandrogen
T59813 Steroid	0.47	288	Hair Growth Promotor

[a] Contains molecules with substituted benzodiazepines. [b] Contains compounds with biphenyl acids resembling A-II antagonists. [c] Contains compounds resembling steroids. [d] Contains compounds resembling tamoxifen.

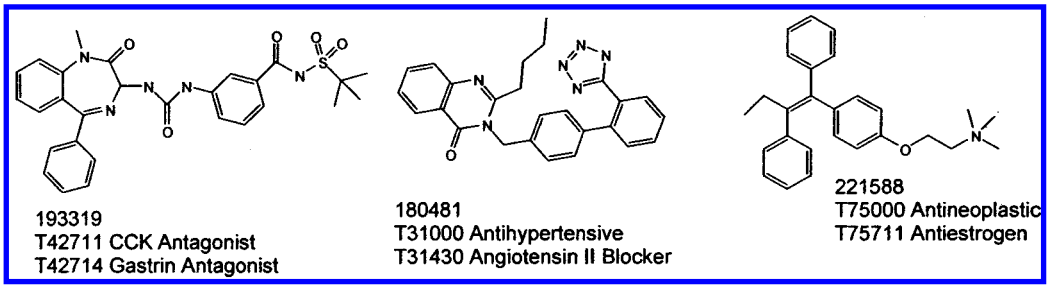

Figure 6. Probe molecules for Similarity Search 2. The probe 221588 is tamoxifen. The therapeutic areas that contain the probes are given.

those from similarity search 3. If that were the case, it might be possible to conclude that there was something detrimental about centroids as probes vs single molecules. In fact, the opposite is true. This is not very surprising in retrospect, since a centroid contains information about more than one class of compound, while a single probe molecule cannot. An especially striking example is with the probe 221588. Since 18 of the 22 T75711 actives in the SS are steroids, which do not especially resemble 221588 by the AP or TT descriptors, the overall global enhancement is at chance level. Again this value is deceptively low due to the fact that tamoxifen-like compounds are a minority in T75711. Inspection shows that the four tamoxifen-like actives in the SS are among the 12 database entries most similar to 221588. If we considered only these as actives, the global enhancement would be extremely high.

QSAR Experiment 1: Trend Vector from Mixtures of Similar Compounds and Single Compounds, Predict Single Compounds. We will simulate an experiment where QSAR is derived for sets of mixtures, each containing closely related compounds. This is hopefully analogous to a situation where mixtures from many different combinatorial syntheses are tested. To train the trend vector, we took the CCs where $M \geq 5$. This is the "Clusters" set of samples ($N = 1283$). The simulated biological activity associated with each sample was assumed to be proportional to the number of "actives" in the cluster, given a selected therapeutic category. For instance, if there were eight molecules from the therapeutic category T31430 in a sample where $M = 100$, the biological activity for a T31430 "assay" would be 0.08. Of course, most samples would be expected to have an activity of zero. For comparison purposes, we also created the "Singles" sample set by taking only the exemplar molecule from each cluster represented by the 1283 CCs. The activities for these samples were 1 or 0 depending on whether the single molecule in each sample was in T31430. The ability of the trend vector to select active compounds from the SS was measured by the global enhancement.

Of the three therapeutic categories T42711, T31430, and T75711, we got statistically significant trend vectors only for the second. This is because for T31430 activity there were many samples which showed activity: 53 samples out of 1283 with activity >0 for the Singles set and 57 for the Clusters set (for which the mean M is 12.4). This is in contrast to ≤ 15 active samples for the other activities, insufficient to ensure statistical significance in the fit. The statistical significance level for both Clusters and Singles trend vectors for T31430 is excellent, all five PLS components being used. The global enhancements for the trend vectors are in Table 4 and the accumulation curves are in Figure 8. Clearly, both Clusters and Singles sample sets give very good predictions, the global enhancements being very close to the maximum. The trend vectors are also very similar. The cos θ between the Singles trend vector and the Clusters trend vector is 0.70 (AP) and 0.86 (TT) i.e., they are very similar. Inspection of the trend vector shows that the descriptors most associated with activity have to do with the tetrazole and the short aliphatic chain, i.e., those features associated with molecules in T31430 to the exclusion of most other druglike molecules.

QSAR Experiment 2: Trend Vectors from Single Compounds, Predict Mixtures. Here we simulate using a QSAR derived from single compounds to search over mixtures. We used the T31430 Singles trend vector from QSAR experiment 1 to sort the TCs by predicted activity. The TCs with the highest predicted activities are shown in Table 5. We realized for this specific exercise to work that we would need to eliminate the requirement that the 95% or 85% of the unique descriptors in the database entry be present in the sample set, since the TCs with large Ms may have a very large number of unique descriptors. The global enhancements for biphenyl-containing actives are 18.4 (AP) and 39.4 (TT), clearly above chance. Relevant therapeutic categories appear in Table 5, but it should be noted that for the AP descriptor, T31430 itself appears near the end of the list, not near the beginning as would be hoped.

QSAR Experiment 3: Trend Vectors from Mixtures of Diverse Compounds, Predict Single Compounds. Here we simulate an HTS experiment where individual compounds are arbitrarily mixed to form samples, the activity of the samples is measured, and a trend vector analysis extracts what descriptors are associated with activity. Then the trend vector is used to select actives from a diverse set of single compounds.

Let n_{training} be the number of actives in the TS set, M be the desired number of compounds per sample, and R be the ratio of inactives to actives. N is the total number of samples we want. Samples were constructed as follows:

1. A pool of compounds was assembled from the actives in TS, plus Rn_{training} inactives from the MDDR. NM compounds were randomly selected from the pool and shuffled. (If there were insufficient actives to complete the pool, i.e., if $Rn_{\text{training}} < NM$, molecules could be chosen more than once with replacement.)

2. The selected compounds were divided into N samples of M compounds each. Descriptors of the M compounds in each sample were generated, and the centroid was calculated. The simulated biological activity of the sample was assigned proportional to the number of actives in the sample, as in the previous experiment.

Table 3. Most Similar Mixtures to Molecules 193319, 180481, and 221588 by Sim

therapeutic category	Sim	*M*	title
		(A) Molecule 193319	
		AP	
T42714 Benz[a]	0.54	181	Gastrin Antagonist[b]
T42711 Benz	0.51	430	CCK Antagonist[b]
T42713 Benz	0.47	121	CCK B Receptor Antagonist
T07709	0.47	17	Neurotensin Receptor Antagonist
T09250	0.47	21	Neurotrophic Factor
T50160	0.46	4	Cathepsin L Inhibitor
T42100 Benz	0.45	28	Growth Hormone
T31433	0.44	40	Angiotensin II AT2 Antagonist
T31432	0.44	58	Angiotensin II AT1 Antagonist
T42102 Benz	0.44	49	Growth Hormone Release Promoting Agent
		TT	
T42714 Benz	0.56	181	Gastrin Antagonist[b]
T42711 Benz	0.48	430	CCK Antagonist[b]
T42713 Benz	0.45	121	CCK B Receptor Antagonist
T42100 Benz	0.41	28	Growth Hormone
T10712	0.40	1	GABA Uptake Inhibitor
T42712 Benz	0.40	87	CCK A Receptor Antagonist
T07709	0.39	17	Neurotensin Receptor Antagonist
T42102 Benz	0.39	49	Growth Hormone Release Promoting Agent
T31432	0.39	58	Angiotensin II AT1 Antagonist
T31341 Benz	0.38	18	Vasopressin V1 Antagonist
		(B) Molecule 180481	
		AP	
T31430 Biphen[c]	0.71	1970	Angiotensin II Blocker[b]
T09210 Biphen	0.60	32	Nootropic
T31432 Biphen	0.58	58	Angiotensin I AT1 Antagonist
T07709 Biphen	0.58	17	Neurotensin Receptor Antagonist
T31341 Biphen	0.56	18	Vasopressin V1 Antagonist
T31433 Biphen	0.55	40	Angiotensin II AT2 Antagonist
T42100 Biphen	0.51	5	Growth Hormone
T42714	0.50	181	Gastrin Antagonist
T69200	0.50	12	Antileprosy
T42711	0.50	427	CCK Antagonist
		TT	
T31430 Biphen	0.67	1970	Angiotensin II Blocker[b]
T31432 Biphen	0.63	58	Angiotensin I AT1 Antagonist
T07709 Biphen	0.59	17	Neurotensin Receptor Antagonist
T42100 Biphen	0.56	5	Growth Hormone
T09210 Biphen	0.53	32	Nootropic
T31433 Biphen	0.53	40	Angiotensin II AT2 Antagonist
T31341 Biphen	0.49	18	Vasopressin V1 Antagonist
T42102 Biphen	0.48	42	Growth Hormone Release Promoting Agent
T38500	0.44	1	Blood Additive
T35520	0.43	1	Antiuremic
		(C) Molecule 221588	
		AP	
T41100 Tamox[d]	0.59	13	Antiinfertility Agent
T75710 Tamox	0.51	1	Estrogen
T09224	0.48	26	Acetylcholine-releasing Agent
T78401	0.43	1	Phosphatidylinositol Kinase Inhibitor
T31341	0.42	18	Vasopressin V1 Antagonist
T33451	0.41	12	Restenosis, Agent for
T33450	0.41	1	Restenosis, Agent for
T09210	0.41	32	Nootropic
T06247	0.40	27	5 HT1D Antagonist
T69200 Tamox	0.40	12	Antileprosy
		TT	
T41100 Tamox	0.61	13	Antiinfertility Agent
T75710 Tamox	0.48	1	Estrogen
T07709 Tamox	0.42	17	Neurotensin Receptor Antagonist
T10712	0.42	1	GABA Uptake Inhibitor
T78401	0.42	1	Phosphatidylinositol Kinase Inhibitor
T09210	0.41	32	Nootropic
T12458	0.39	2	Glutamate Release Inhibitor
T12460	0.39	6	Adenosine Reuptake Inhibitor
T02510	0.37	2	Leumedin
T78420	0.37	2	Glutathione S-Transferase Inhibitor

[a] Contains molecules with substituted benzodiazepines. [b] Therapeutic categories that contain the probe molecule. [c] Contains compounds with biphenyl acids resembling A-II antagonists. [d] Contains compounds resembling tamoxifen.

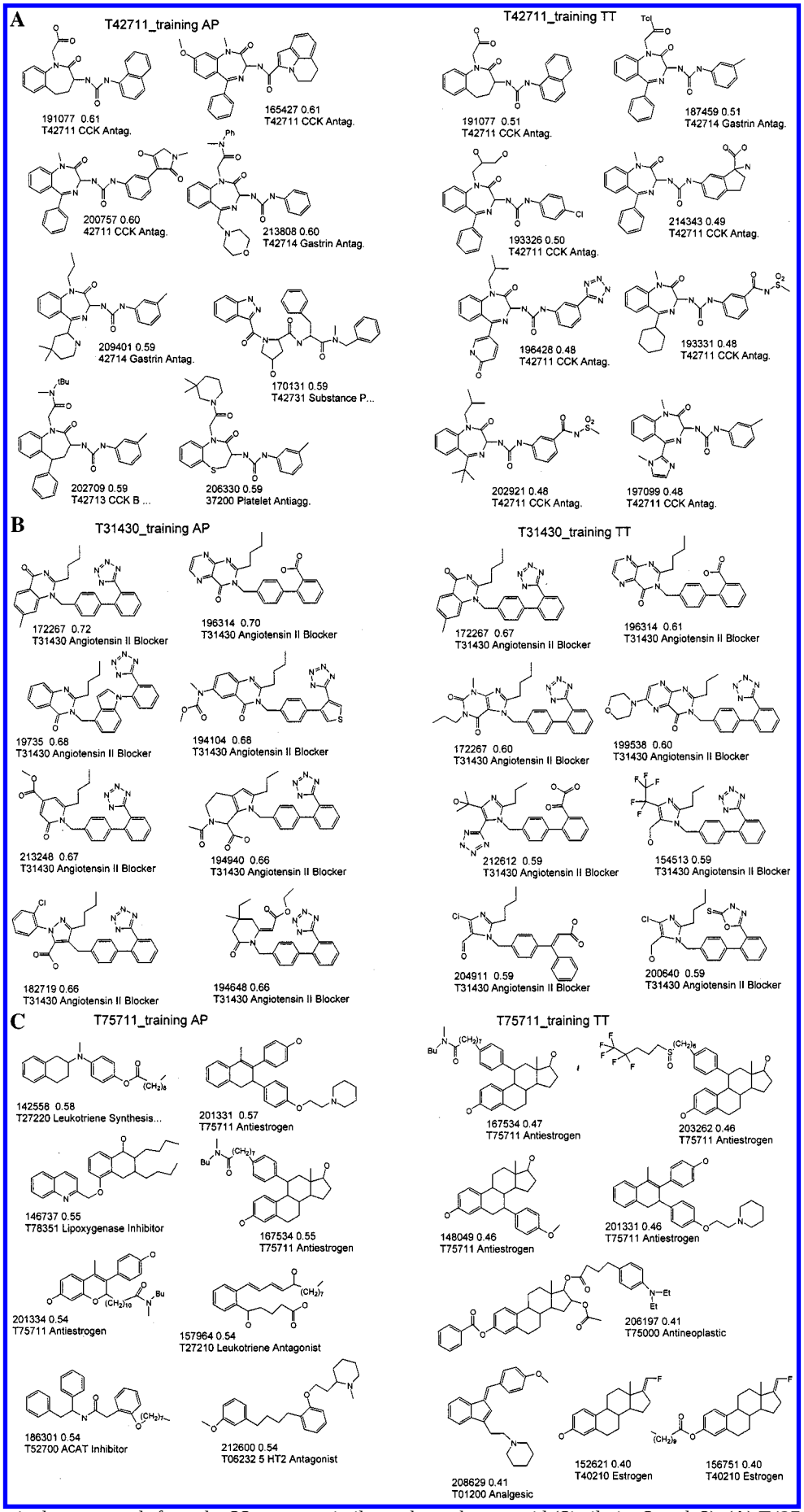

Figure 7. Diverse single compounds from the SS set most similar to the probe centroid (Similarity Search 3). (A) T4271. (B) T31430. (C) T75711.

Table 4. Trend Vector Fit and Predictions for QSAR Experiment 1

training set	descriptor	global[a] enhancement for prediction
Singles	AP	55.1
	TT	55.1
Clusters	AP	54.7
	TT	55.1

[a] Maximum possible global enhancement = 55.6.

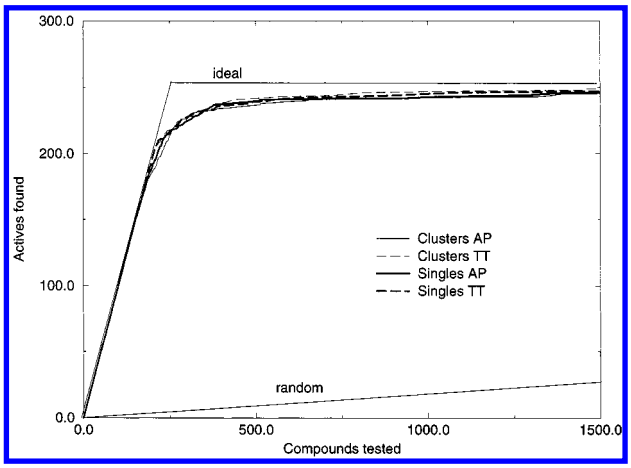

Figure 8. Actives found as a function of the SS compounds tested for QSAR Experiment 1. Curves for two limiting cases are also shown: "random" is the case where the actives are randomly distributed in the database, and "ideal" is the case where all actives are at the beginning of the list. 14 116 diverse compounds were tested, but the *x*-axis is truncated at 1500.

We set $N = 500$ samples throughout. We let $M = 1$, 10, 50, 100, and 500 and $R = 10$, 50, and 100. In most screening experiments cutoffs are set so at most only a few percent of tested compounds are active, so values of $R \geq 50$ are realistic. $M = 1$ is equivalent to testing single compounds. We generated a trend vector for each combination of M and R, and the ability of each trend vector to select active compounds from the SS was measured by the global enhancement. Chance effects are likely to be important in this experiment. Therefore, for every combination of M and R, we repeated the experiment five times to monitor the variation.

Again we show T31430 as the example activity, although we have found qualitatively similar results for T42711 and T75711. It is hard to summarize the statistical significance of the trend vectors in this experiment because it varies quite a bit, even within a given combination of M and R, so we will consider only the accumulation curves and global enhancements, which seem more consistent and are more relevant to actual prediction. Figure 9 shows the accumulation curves for $R = 50$ and various values of M for the TT descriptor. One can see there how much chance-based variation in global enhancement there is among the five attempts. Results of trend vector generation and predictions summarized over the five attempts are in Table 6. There are three clear trends in global enhancement. First, at any given M, global enhancement decreases slowly as R increases. Second, at any given R, global enhancement decreases quickly as M increases Third, for a given M and R, global enhancement for the TT descriptor is generally better than for the AP descriptor where $M > 1$. Even at a realistically large $R \geq 50$, the global enhancement at $M = 10$ is comparable to that for $M = 1$ (testing single compounds). The global enhancement is still fairly good for $M = 50$ for the TT descriptor, indicating that QSAR can be derived from mixtures with a few tens of compounds. Table 6 also shows that, for any given R, $\cos\theta$ falls quickly as M increases, indicating that the vector for, say $M = 100$, has lost meaningful similarity with the vector for $M = 1$. This is consistent with the loss of global enhancement.

One important question is whether these results depend on having quantitative activities being assigned to the samples, because the precision of an actual screening experiment might be limited to "detectable activity" vs "no detectable activity". We repeated QSAR Experiment 3, assigning the activity as "1" if any of the compounds in the sample were active, and "0" if not. The global enhancements were close to those shown in Table 6, with the following caveat: As M gets much larger than R, the probability of a sample having at least one active in it approaches 1. Thus the activity of all samples approaches the same value, and no trend vector can be calculated.

DISCUSSION

Our simulations suggest that, using the centroid approximation with substructure descriptors, similarity methods and QSAR methods that work for single molecules can also give useful results for mixtures. Some of our previous work[14] and that of others[15] refers to "joint probes" wherein a similarity probe is the descriptor centroid of several related compounds, but the number of compounds is always less than 10. Here the concept is extended to much larger values of M. As far as we know, there has been no previous work deriving QSAR from mixtures. Considering that the concept of the centroid is often used in clustering, and considering how natural it is to represent a mixture by a centroid, it is surprising that this idea has not been reported before. Perhaps it was felt that the structural "signal" in a set of compounds would be quickly lost in the noise by averaging over many compounds. Clearly that is not the case with the examples we present here.

That said, it is important to note the limits to the centroid approximation for similarity searches. The two most important are "low resolution" and "noise". We have noticed that if mixtures contain more than one class of compound, the centroid resembles the majority class. For instance, using C-T75711 as a probe selects steroids over tamoxifen-like compounds. Thus Sim measures the resemblance of mixtures as global entities and there is no way to detect that two mixtures might have minority molecules in common. That is, the resolution of searches is limited. There is also an unavoidable source of noise in the approximation. As many diverse molecules are averaged, the centroid approximation ceases to be meaningful because the centroid looks more like U, the centroid of druglike space, than any particular compound class. The therapeutic category centroids we use here are problematic in this respect, particularly in regard to mixture−mixture comparisons, in that some contain very diverse compounds.

The problem of noise can be eliminated in three ways. The first is to restrict ourselves to mixtures with limited diversity. For many useful applications, this is not a problem. Combinatorial mixtures, for instance, are usually rather

Table 5. Highest Predicted Activities of Therapeutic Category Mixtures for QSAR Experiment 2

therapeutic category	pred. activity	*M*	title
		AP	
T09210 Biphen[a]	0.47	32	Nootropic
T43120	0.36	6	Sulfonylurea
T78347	0.29	4	GABA Transaminase Inhibitor
T50160	0.28	4	Cathepsin L Inhibitor
T37430	0.28	1	Vitamin K Deficiency, Agent for
T78400	0.27	11	Lysyl Oxidase Inhibitor
T07709 Biphen	0.27	17	Neurotensin Receptor Antagonist
T53800	0.27	11	Metabolic Disease, Agent for Miscellaneous
T09250	0.27	15	Neurotrophic factor
T31433 Biphen	0.26	40	Angiotensin II AT2 Antagonist
		TT	
T31432 Biphen	1.58	58	Angiotensin II AT1 Antagonist
T31433 Biphen	1.50	40	Angiotensin II AT2 Antagonist
T31520	1.29	606	Potassium Channel Activator
T31430 Biphen	1.27	1970	Angiotensin II Blocker[b]
T09200	1.21	4062	Cognition Disorders, Agent for
T82100	1.21	71	Drug Delivery System
T31000 Biphen	1.21	7751	Antihypertensive
T16000	1.10	744	Antiglaucoma
T28000	1.10	1490	Heart Failure, Agents for
T75000	1.10	5631	Antineoplastic

[a] Contains compounds with biphenyl acids resembling A-II antagonists. [b] Activity from which trend vector was constructed.

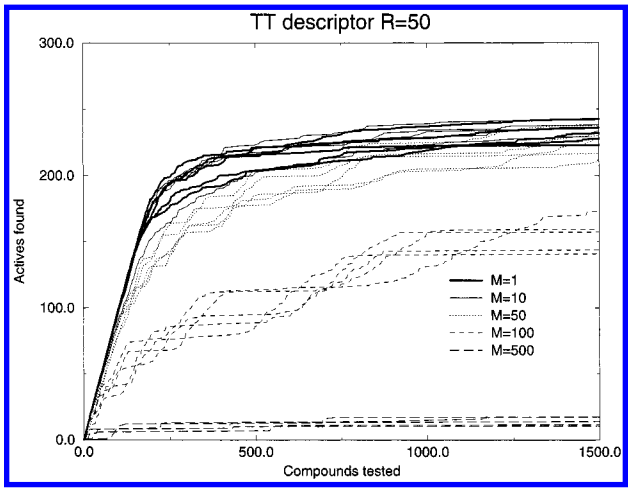

Figure 9. Actives found as a function of the SS compounds tested for QSAR experiment 3. The case shown is $R = 50$ using the TT descriptor with various values of *M*. 14 116 diversse compounds were tested, but the *x*-axis is truncated at 1500.

internally self-similar, and thus their centroids rarely approximate *U*. For instance, when we look at all the combinatorial mixtures synthesized at Merck, the highest Sim_{iU} is 0.6 for both AP and TT, even for $M > 1000$, in contrast to the highest values of over 0.9 for the TCs. Because of their higher homogeneity, then, we would expect low resolution and noise to be less of a problem for combinatorial mixtures. A second approach is to correct the raw similarities using cSim, as we have done here for mixture−mixture comparisons. A third approach (not shown here) is to break diverse mixtures into smaller sets based on structural class. There would be more than one centroid per mixture, but each centroid would represent a more homogeneous set of compounds. This is not necessarily limited to the most diverse mixtures. For instance, the T75711 mixture might be broken into steroid-like compounds and tamoxifen-like compounds.

Next we discuss the suitability of descriptors for the centroid approximation. Our similarity definitions for mixtures can be applied to any type of substructure descriptors for which the frequencies are available, e.g., the AP and TT descriptors. For similarity measures, centroids cannot be usefully represented as a "fingerprints", wherein only the presence or absence of a descriptor is noted (Daylight,[16] MDL,[17] and UNITY[18] fingerprints being popular examples). When all frequencies equal 1, as in a fingerprint, more diverse mixtures will have more unique descriptors and will tend to have higher similarities to any given probe. We have confirmed this undesirable behavior by calculating Sim using only the presence or absence of APs and TTs (data not shown).

In the realm of QSAR, in contrast, many methods including trend vectors (ref 12 and this work) and HQSAR (ref 18), use only the presence or absence of substructure descriptors, so fingerprint descriptors could potentially be

Table 6. Trend Vector Fit and Predictions for QSAR Experiment 3

		global enhancement for prediction[a]		cos θ relative to $M = 1$[b]	
M	*R*	AP	TT	AP	TT
1	10	54.7 ± 0.9	54.6 ± 0.3	1.00	1.00
10	10	49.2 ± 2.9	54.2 ± 0.3	0.43	0.70
50	10	14.5 ± 3.8	35.6 ± 6.1	0.15	0.34
100	10	4.5 ± 3.9	10.0 ± 2.4	0.08	0.21
500	10	0.7 ± 0.1	0.9 ± 0.2	0.01	0.07
1	50	44.1 ± 4.2	53.5 ± 0.7	1.00	1.00
10	50	24.3 ± 9.0	52.0 ± 2.5	0.40	0.69
50	50	9.2 ± 3.9	39.5 ± 3.2	0.12	0.30
100	50	2.3 ± 0.7	8.9 ± 0.9	0.07	0.16
500	50	0.7 ± 0.1	1.0 ± 0.4	0.02	0.06
1	100	42.5 ± 7.4	52.0 ± 2.1	1.00	1.00
10	100	15.1 ± 5.3	48.5 ± 6.9	0.40	0.55
50	100	3.1 ± 2.3	19.6 ± 8.2	0.11	0.23
100	100	1.8 ± 1.3	5.4 ± 2.2	0.06	0.13
500	100	0.7 ± 0.1	0.7 ± 0.1	0.01	0.02

[a] Mean ± 1 std dev over five trials. Maximum possible global enhancement = 55.6. [b] For each *M* and *R* combination, five trend vectors were averaged. Comparisons are made between the averaged trend vectors. cos θ between the five vectors for each *M* and *R* combination is in the range 0.3−0.7.

useful for deriving QSAR from mixtures. Since frequencies are ignored, one would really be using the descriptor "union" of molecules instead of the centroid. The union of descriptors has been used in the past to represent a set of conformations for a single molecule (for instance ref 19 and cited references) and to represent a "modal fingerprint" to summarize a set of active molecules.[20] The union appproximation has its own limits in QSAR. Clearly, as shown in QSAR experiment 3, "dilution" of the descriptors of active compounds by those from inactives becomes important when *M* becomes large.

Similarity and trend vector studies on single compounds have shown that whether AP or TT descriptors give better results is strongly problem-dependent, so that there is no overall preference. However, our present work on mixtures shows consistently better results for the TT descriptor, especially for the QSAR experiments. An obvious explanation for this lies in the fact that the AP descriptor is more "fuzzy" than the TT descriptor, as defined in our previous work.[10] Generally, the more fuzzy a descriptor, the more likely it will occur in two randomly selected molecules. Thus, the problematic artifacts mentioned in the above paragraphs are more pronounced for the AP. This puts a restriction on the type of descriptor that can be used for mixtures. One would need a descriptor comparable in specificity to the TT.

Despite the drawbacks, there are four important advantages to a centroid representation of the descriptors in mixtures relative to a descriptor representation of individual compounds in the mixtures. First, the representation is very compact. The size of a file containing a centroid increases linearly with the number of unique descriptors in the mixture, which varies approximately as log *M*. For instance, in our implementation storing the descriptors for 1000 explicit molecules takes about 3.2 MB, but storing the centroid of those molecules takes only 0.2 MB. As organizations accumulate mixtures in their databases, the space advantage can become important. A second consideration is computational speed. There is no accepted definition of the similarity between two mixtures A and B when the mixtures are represented as single compounds. One might define the similarity between A and B as the mean similarity, or perhaps the maximum similarity of all single-molecule pairs from A and B, or a random sample of pairs thereof. In any case, a large number of similarity calculations would be involved. In contrast, if the mixtures are represented by their centroids, only a single calculation is necessary. Third, mixtures can be handled with existing software that is designed to handle the descriptor representation of single molecules, albeit with some optional correction in the case of mixture−mixture similarity calculations. Finally, in the QSAR realm, structure−activity rules can be derived directly from mixtures without knowing specifically which compound(s) is (are) active.

Demonstration of the use of centroid-based similarity and QSAR on combinatorial mixtures will be published as the next paper in the series.

ACKNOWLEDGMENT

A number of members of the MIX software development team produced software that was useful for this project. Dr. Robert B. Nachbar wrote **simtab**, which supports a large number of pairwise similarity/distance measurements. Dr. Simon K. Kearsley wrote three useful programs: **topogen** generates AP and TT descriptors for large databases; **topodis**, our implementation of the method of Holliday et al.,[7] rapidly calculates neighbor lists; **topopls** is our implementation of trend vectors. Ms. Katherine Dayem and Ms. Souzy Sawiris provided technical support. Dr. Nachbar provided a critical reading of the manuscript.

REFERENCES AND NOTES

(1) Thompson, L. A.; Ellman, J. A. Synthesis and applications of small molecule libraries. *Chem. Rev.* **1996**, *96*, 555−600.

(2) Berk, S. C.; Rohrer, S. P.; Degrado, S. J.; Birzin, E. T.; Mosley, R. T.; Hutchins, S. M.; Pasternak, A.; Schaeffer, J. M.; Underwood, D. J.; Chapman, K. T. A combinatorial approach toward discovery of non-peptide, subtype-selective somatostatin receptor ligands. *J. Comb. Chem.* **1999**, *1*, 388−396.

(3) Berk, S. C.; Chapman, K. T. Spatially arrayed mixture (SPAM) technology: synthesis of a two-dimensionally indexed orthogonal combinatorial libraries. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 837−842.

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

(5) Willet, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

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

(7) Holliday, J. D.; Ranade, S. S.; Willett P. A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.

(8) 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.

(9) 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.

(10) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R.P. Chemical similarity using physicochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(11) Hull, R. D.; Fluder, E. M.; Singh, S. B.; Nachbar, R. B.; Kearsley, S. K.; Sheridan, R. P. Chemical similarity searches using Latent Semantic Structural Indexing (LaSSI). *J. Med. Chem.* submitted for publication.

(12) Sheridan, R. P.; Nachbar, R. B.; Bush, B. L. Extending the trend vector: the trend matrix and sample-based partial least squares. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 323−340.

(13) MDL Drug Data report licensed by Molecular Design Ltd., San Leandro, CA.

(14) Sheridan, R. P.; Kearsley, S. K. Using a genetic algorithm to suggest combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.

(15) Nachbar, R. B. Molecular evolution: automated manipulation of hierarchical chemical topology and its application to average molecular structures. *Genetic Program. Evolvable Hardware* **2000**, *1*, 57−94.

(16) Daylight Chemical Information Systems, Inc., 2740 Los altos, Suite #360, Mission Viejo, CA 92691.

(17) McGregor, M. J.; Pallai, P. V. Clustering large databases of compounds: using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(18) Heritage, T. W.; Lowis, D. R. Molecular hologram QSAR. In *Rational Drug Design: Novel Methodology and Practical Applications*; ACS Symposium Series 719; Parrill, A. L., Reddy, M. M., Eds.; American Chemical Society: Washington, DC, 1999; pp 212−225.

(19) Matter, H.; Potter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(20) 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.

CI000045J