Mining the Chemical Quarry with Joint Chemical Probes: An Application of Latent Semantic Structure Indexing (LaSSI) and TOPOSIM (Dice) to Chemical Database Mining

Suresh B. Singh,* Robert P. Sheridan, Eugene M. Fluder, and Richard D. Hull†

Molecular Systems, Merck Research Laboratories, P.O. Box 2000, Rahway, New Jersey 07065

Received September 11, 2000

In this study we use a novel similarity search technique called latent semantic structure indexing (LaSSI) with joint chemical probes as queries to mine the MDL drug data report database. LaSSI is based on latent semantic indexing developed for searching textual databases. We use atom pair and topological torsion descriptors in our calculations. The results obtained with LaSSI are compared with another in-house similarity search technique TOPOSIM. The results from the similarity searches using joint chemical probes are significantly better than searches using single chemical probes for both LaSSI and TOPOSIM. The selected molecules are closely related in activity to their queries and are ranked among the top 300 scoring molecules of the 82 860 entries in the database. Our implementation of LaSSI is very fast and efficient in finding active compounds. The results also show that LaSSI consistently retrieves more diverse chemical structures representative of the joint chemical probes in comparison to TOPOSIM. The use of multimolecule topological probes to identify compounds complements the use of searching databases with 3D pharmacophore hypotheses.

Introduction

Chemical database mining is an essential process in the identification of biologically active molecules. Such mining involves searching a database for compounds that have chemical properties similar to those of a probe molecule(s) in order to identify a set of compounds that potentially have similar biological properties and, perhaps, diverse chemical scaffolds. Several similarity search techniques are available to retrieve structures from chemical databases.[1] Compounds that have a common mechanism of action are traditionally analyzed via 3D molecular modeling tools to generate pharmacophore hypotheses. These pharmacophore hypotheses are used to search chemical databases to identify novel structural classes that act through the same mechanism of action. Our assertion is that chemical similarity searches with multicompound probes using topological descriptors would accentuate common chemical features present in the members of the probe. Hence, joint chemical probes would serve as a 2D equivalent of a pharmacophore hypothesis. There are a few published reports of 2D or topological approaches to search chemical databases with the simultaneous use of multiple compounds[2−4] or with the use of a 2D fingerprint technique.[5]

We describe here an application of a novel 2D similarity search technique called latent semantic structure indexing (LaSSI) to probe a chemical database using a chemical probe constructed from multiple compounds. LaSSI is inspired by latent semantic indexing (LSI)[6,7] developed at Bellcore Laboratories[8] for searching textual databases. LSI can be adapted to search both 2D and 3D databases, but we focus our efforts in this paper on the application of this technique to search a 2D database. LaSSI involves the following steps: (1) a singular value decomposition of a matrix of molecules and its descriptors and (2) generation of a low-dimensional representation of the original chemical descriptor space. This has to be done only once to create the database. The next step involves calculating the similarity of molecules in this database to a given probe and ranking them according to decreasing similarity. The details of the theory and implementation of this technique are presented elsewhere.[6]

We present here an iterative approach for both LaSSI and TOPOSIM that uses the results from a similarity search with a single molecule probe to select structurally diverse active compounds, which we use to construct a joint chemical probe to search the chemical database again. Information about known active compounds can be used to increase the chances of retrieving a greater number of active compounds. We employed here an approach analogous to relevance feedback used in the field of natural language understanding[9] to influence the outcome of similarity searches with LaSSI. The list of active compounds retrieved by single molecule probe searches was used to identify parameters that lead to optimal database retrieval performance by LaSSI using joint chemical probes.

Hull et al.[7] compared the results obtained from similarity searches with single molecule probes across 16 therapeutic categories using LaSSI and TOPOSIM[10] to search MDL's drug data report (MDDR) database. In this study it was shown that the combination of atom pair[11] and topological torsion[12] descriptors on average retrieved the most number of active compounds across the 16 therapeutic categories. These results also showed

* Corresponding author: Suresh B. Singh, RY50SW 100, P.O. Box 2000, Merck Research Laboratories, Rahway, NJ 07065. E-mail: suresh_singh@merck.com. Tel: 732-594-4954, Fax: 732-594-4224.

† Current address: Elagent Corporation, 7011 N. Atlantic Ave., Suite 200, Cape Canaveral, FL 32920.

10.1021/jm000398+ CCC: $20.00
Published on Web 05/03/2001

that LaSSI on average performed as well as TOPOSIM and better when the information regarding the active compounds was used to find the optimal number of singular values.

Methods

We chose five therapeutic categories out of the 16 therapeutic categories from our single molecule probe similarity search results.[7] In two cases LaSSI performed better (dopamine D2 agonists and ACE inhibitors), in two cases TOPOSIM performed better (thrombin inhibitors and leukotriene antagonists), and in one case LaSSI and TOPOSIM were comparable (5HT re-uptake inhibitors). In each therapeutic category, the top 300 compounds from the single probe similarity search results were examined to identify known active compounds. From the list of active compounds, a representative set of compounds were selected to construct the joint chemical probe. These compounds were chosen so as to represent the structural diversity of the active compounds.

Descriptors. We used atom pairs[11] and topological torsion[12] descriptors in all our calculations.

Atom pairs are substructure descriptors of a molecule and are defined as $AT_i - AT_j - r_{ij}$. The distance, r_{ij}, is the distance in bonds along the shortest path between an atom type AT_i and an atom type AT_j. The atom type encodes the element type, the number of non-hydrogen atom neighbors, and the number of π electrons. Topological torsions are substructure descriptors of a molecule and are defined as $AT_i - AT_j - AT_k - AT_l$, where i, j, k, and l are consecutively bonded atoms.

LaSSI. The theory, implementation, and methodology of LaSSI are presented elsewhere.[6] We present here only a brief description of the LaSSI methodology.

LaSSI uses the singular value decomposition (SVD) of the chemical descriptor matrix **X**. The SVD of **X** in $R^{m\times n}$ leads to a left singular matrix **P** ($m \times r$), right singular matrix **Q** ($n \times r$), and a diagonal matrix **Σ** ($r \times r$) and is defined by $\mathbf{X} = \mathbf{P\Sigma Q}^T$. The columns of the **P** matrix are the eigenvectors of $\mathbf{XX}^T$ corresponding to nonzero eigenvalues. The columns of the **Q** matrix are the eigenvectors of $\mathbf{X}^T\mathbf{X}$ corresponding to nonzero eigenvalues. The nonzero elements of the diagonal matrix $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_r)$ are called the singular values. The singular values are square roots of the eigenvalues and possess the property that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$. The kth rank approximation of **X**, $\mathbf{X}_k$, for $k \ll r$, $\sigma_{k+1} ... \sigma_r$ set to 0 can be efficiently computed using variants of the Lanczos algorithm.[13] $\mathbf{X}_k$ is the matrix of rank k that is closest to **X** in the least squares sense; it is called a partial SVD of **X**, and is defined as $\mathbf{X}_k = \mathbf{P}_k\mathbf{\Sigma}_k\mathbf{Q}^T_k$. The rows of $\mathbf{X}_k$ are orthogonal descriptors that are linear combinations of the original descriptors, and the columns are the projection of the molecules into the space of those descriptors. The similarity of two descriptors i and j is calculated by computing the dot product between the ith and jth rows of the matrix $\mathbf{P}_k\mathbf{\Sigma}_k$. The similarity between molecules, represented by vectors **i** and **j** can be calculated by computing the dot product between the ith and jth rows of the matrix $\mathbf{Q}_k\mathbf{\Sigma}_k$. The similarity of a descriptor i to a molecule j can be calculated by computing the dot product between the ith row of the matrix $\mathbf{P}_k\mathbf{\Sigma}_k$ and the jth row of the matrix $\mathbf{Q}_k\mathbf{\Sigma}_k$. Finally, the similarity of a probe to the descriptors and molecules in the database can be calculated by first projecting the probe into the k-dimensional space of the partial SVD and then treating the projection as a molecule for probe−descriptor and probe−molecule comparisons. The projection of a probe vector, v, is defined as $y = v^T\mathbf{P}_k\mathbf{\Sigma}_k^{-1}$. y is treated as a row of $\mathbf{Q}_k$ for the purposes of calculating similarity.

LaSSI does not use the singular values to scale the singular vectors when calculating similarities, however, as is the case for LSI. Instead, the identity matrix **I** is used in place of $\mathbf{\Sigma}_k$. Ignoring the scaling component $\mathbf{\Sigma}_k$ improves the system's ability to select similar molecules regardless of whether the probe's descriptors are well represented in the database.

The connection table of a probe molecule(s) is converted into the descriptor set of the LaSSI database to create a feature vector for the probe. The probe is then projected into a reduced dimensional space as described by Hull et al.[6] The normalized dot products of each molecule in the database with the transformed probe are calculated and the resulting values are sorted in descending order, maintaining the index of the molecule responsible for that value. Then LaSSI generates a list of top-ranking molecules at a chosen cutoff, e.g., usually the highest ranked 300, 500, or 1000 compounds. The calculation of LaSSI similarity between the probe molecule v and the database molecule i is given by

$$\mathrm{Sim}_{vi} = \sum_{x=1}^{k} \frac{v_x q_{ix}}{|v||q_i|}$$

where q_{ix} are elements of $\mathbf{Q}_k$ with x ranging from 1 to k (total number of singular values). Sim_{vi} ranges from −1 (least similar) to 1.0 (most similar). The process of carrying out SVD converts the original descriptors into a new set of descriptors that are a linear combination of the original descriptors. Thus the coefficients on the new descriptors are floating point numbers and no longer represent frequencies. For this reason we feel that the Dice similarity measure defined below is not a viable approach to calculate similarity between the descriptors from the LaSSI database.

By varying the number of singular values (k) we can control the level of fuzziness of the search: larger singular values produce better approximations of the original descriptor space than smaller values. In the extreme case of $k = r$, r being the rank of the matrix **X**, $\mathbf{X}_k = \mathbf{X}$ and hence the descriptors are represented in their original form. Alternatively, if k is very small, much of the distinctive character of the descriptors will be lost. An investigation of the effect of singular values on chemical similarity has been presented by Hull et al.[7] The optimal singular values used in single probe similarity searches with LaSSI across the 16 therapeutic areas ranged from 50 to 800. The assessment of the retrieval of actives at all these singular values leads to an average singular value ∼300. These results are presented by Hull et al.[7] On the basis of these results we selected the compounds from LaSSI single probe similarity rankings at 300 singular values for constructing joint probes.

TOPOSIM. The Dice similarity measure was used in TOPOSIM[10] calculations, and it is defined as follows

$$\mathrm{Sim}_{vi} = \frac{\sum_j \min(d_{jv}, d_{ji})}{0.5[\sum_j d_{jv} + \sum_j d_{ji}]}$$

where d_{jv} is the count of descriptor j in probe molecule v, and d_{ji} is the count of descriptor j in the molecule i. The index j goes over the union of unique descriptors in v and i. Sim_{vi} ranges from 0.0 (nothing in common) to 1.0 (identical). We use the Dice similarity measure for TOPOSIM because empirically it works better than cosine similarity in retrieving actives.[7]

The joint probes for TOPOSIM were constructed from TOPOSIM's single probe similarity searches. The procedure used to select a representative set of compounds for constructing joint probes is shown schematically in Figure 1. For example, we show how LaSSI and TOPOSIM joint probes were generated for 5-HT re-uptake inhibitors (Figure 1). In similarity searches using LaSSI and TOPOSIM with the single chemical probe 170534, there were 11 and 6 active compounds, respectively, in the top-ranking 300 compounds. We selected three structurally diverse representative compounds from the respective set of active compounds to construct each joint chemical probe. The list of MDDR registration numbers comprising the joint chemical probes for each therapeutic category for LaSSI and TOPOSIM is given in Table 1 and the corresponding structures are shown in Figures 2−6.

Joint Chemical Probe. For TOPOSIM, the joint chemical probe is a sum of the frequencies of the descriptors of the

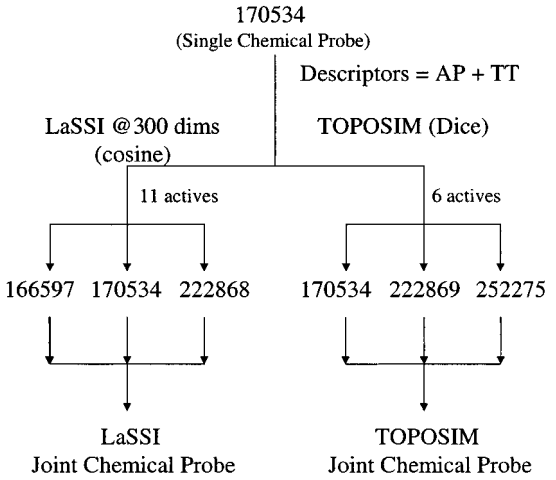

Figure 1. Chemical similarity search paths taken for LaSSI and TOPOSIM.

members of the joint probe divided by the total number of molecules.

$$\text{joint chemical probe} = \sum_{i=1}^{N} \frac{d_{ji}}{N}$$

In the case of LaSSI, the joint chemical probe is generated from summing the frequencies of the descriptors of its constituent molecules.

$$\text{joint chemical probe} = \sum_{i=1}^{N} d_{ji}$$

where d_{ji} is the count of descriptor j of the ith molecule, and N is the total number of molecules in the joint chemical probe. Since the calculation of similarity with LaSSI involves calculation of the angle (cosine) between the probe and the column vectors of the LaSSI database, the normalization of the descriptors does not have an impact on the similarity: the angle between two vectors does not change when the length of one of the vectors changes.

The joint chemical probes from each of the five therapeutic categories were used to rank 82 860 compounds in the MDDR database by the cosine similarity measure for LaSSI and by the Dice similarity measure for TOPOSIM. In the case of LaSSI, we ranked compounds in the MDDR database by including a series of singular values ranging from 10 to 430 at increments of 10. The maximum number of singular values retained for the MDDR database version 98.1 in the current study is 430. These scored lists were used to generate the ranks for the members of the joint chemical probes. We analyzed these ranks to select the singular value at which the rank of the last retrieved joint probe member is the smallest. If there were similar rankings for the last joint probe member at two or more singular values, we chose the results of the smallest k.

Initial Enhancements. This measure of effectiveness is computed by taking a ratio of the number of actives for a particular therapeutic category retrieved in the top-scoring 300 compounds (actives@300) and the number of actives that are expected by pure chance.

$$\text{initial enhancement} = \frac{\text{actives@300}}{\text{nactives} \times 300/82860}$$

where "actives@300" is the number of actives found by LaSSI or TOPOSIM in the top-ranking 300 compounds for a particular therapeutic category, and "nactives" is the total number of actives belonging to the corresponding therapeutic category.

Measure of Diversity. We used the following approaches to assess the diversity of active compounds retrieved by LaSSI and TOPOSIM.

(a) Similarity to the Centroid. For each therapeutic category we generated the list of active molecules retrieved by the joint probe. For instance, in the case of 5-HT re-uptake inhibitors we have 18 and 31 actives retrieved by LaSSI and TOPOSIM, respectively, with the use of joint probes. These actives were used to construct the centroid for each set

$$\text{centroid} = \sum_{i=1}^{N} \frac{d_{ji}}{N}$$

where d_{ji} is the count of descriptor j in the ith molecule, and N is the total number of molecules in the actives set.

Then, for each member of the active set we computed the similarity between it and the centroid using topological torsion descriptors and the Dice similarity measure

$$\text{Sim}_{ci} = \frac{\sum_j \min(d_{jc}, d_{ji})}{0.5[\sum_j d_{jc} + \sum_j d_{ji}]}$$

where d_{jc} is the count of descriptor j in centroid molecule c, and d_{ji} is the count of descriptor j in the molecule i. The index k goes over the union of unique descriptors in c and i.

The mean of the similarities Sim_{ci} is computed as follows

$$\text{mean} = \sum_{i=1}^{N} \frac{\text{Sim}_{ci}}{N}$$

where N is the total number of molecules in the set of actives retrieved by LaSSI or TOPOSIM for a given therapeutic category.

(b) Computing Similarity Matrixes. We used the following methodology to enumerate the number of structural classes retrieved by LaSSI and TOPOSIM in order to assess and display the diversity of retrieved compounds. Self-similarity and cross-similarity matrixes were computed between an individual molecule and all other molecules in that set using topological torsion descriptors and the Dice similarity measure. We clustered compounds together when similarity between any two compounds in a given class was greater than or equal to 0.65.

Results

Joint Chemical Probe Similarity Searches with LaSSI. For each therapeutic category we qualitatively selected a structurally representative set of compounds from the actives retrieved by the single molecule probe similarity search. The descriptors from these compounds were summed to form the joint probe. LaSSI similarity searches were carried out at various singular values ranging from 10 to 430 at increments of 10 to rank the 82 860 molecules in the MDDR database. We analyzed the top 300 compounds from the LaSSI similarity searches at each singular value to rank the members of the joint chemical probe. For each analysis we determined the number of singular values that gave the best ranking of the joint chemical probes (Table 2).

The results of the database searches with joint chemical probes using LaSSI are given in Table 2 (columns 6−10). These results are compared with those obtained with single molecule probes (columns 3−5).[7] The therapeutic categories are indicated in column 1. The total number of actives shown in column 2 for each therapeutic category is based on the count of the

Table 1. Joint Chemical Probe Members

therapeutic category	total actives	MDDR registration numbers	
		LaSSI	TOPOSIM
5HT re-uptake inhibitors	219	166597, 170534, 222868	170534, 222869, 252275
ACE inhibitors	499	090439, 152038, 158135, 158137, 167770, 180880, 204703, 211892	090388, 206856, 220578, 152038, 155108, 165334, 185771
dopamine agonists	127	139393, 143986, 161853, 169745, 174007, 177101, 179378, 224232	143986, 224232, 174007, 161853
leukotriene antagonists	811	146603, 148762, 154326, 162215, 206343	146603, 205152, 154326
thrombin inhibitors	493	090744, 159159, 177193, 184521, 214229, 220363, 248991, 251848	256114, 159160, 256052, 238882, 201822, 177193

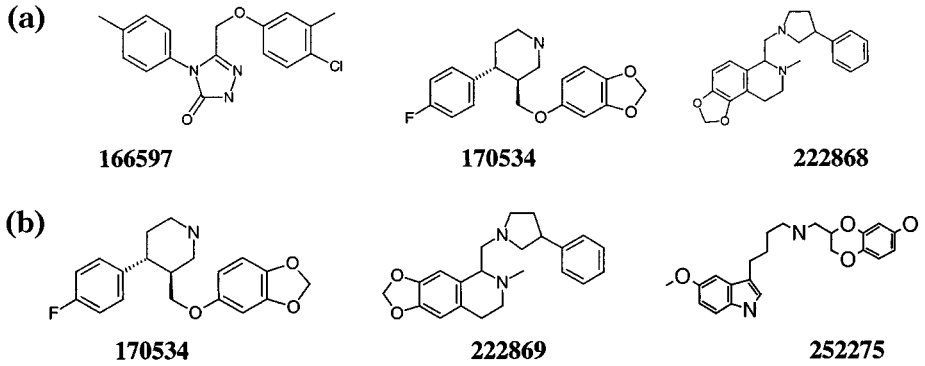

Figure 2. 5HT re-uptake inhibitors joint chemical probe members for (a) LaSSI and (b) TOPOSIM.

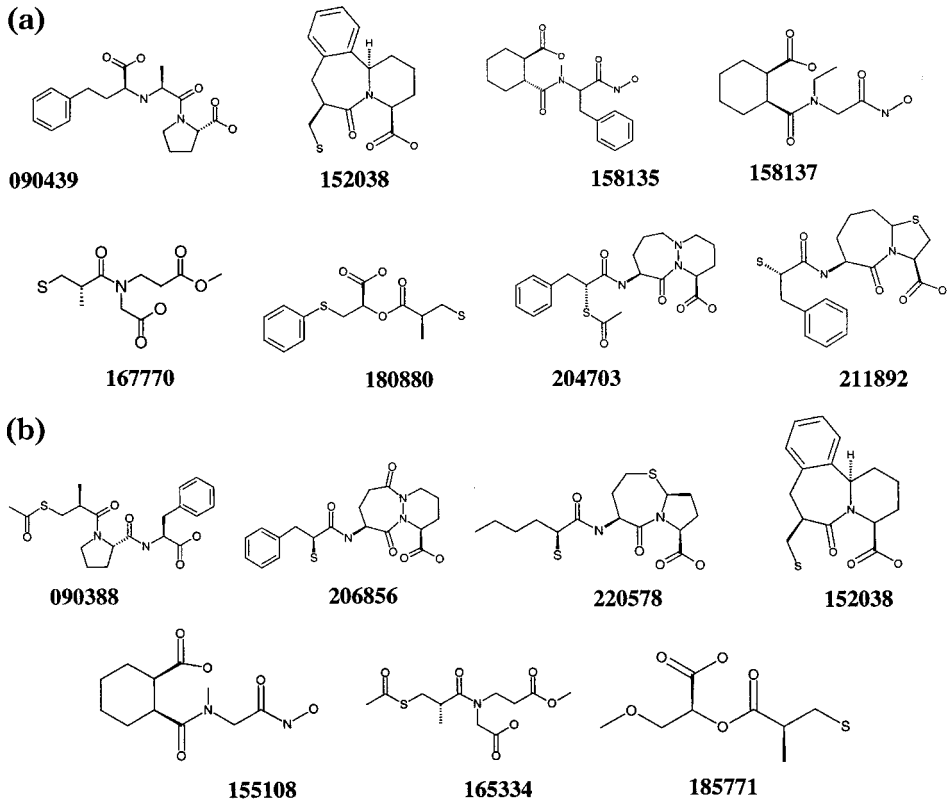

Figure 3. ACE inhibitors joint chemical probe members for (a) LaSSI and (b) TOPOSIM.

compounds labeled as such in the MDDR database. The MDDR registration numbers of chemical probes used for each therapeutic category are given in column 3.

There are 219 compounds labeled as 5-HT re-uptake inhibitors in the MDDR database. The compound with the MDDR registration number 170534 was used as a single probe for similarity searches. A LaSSI search using $k = 300$ retrieved 11 5-HT re-uptake inhibitors in the top-ranking 300 compounds (Table 2). The maximum number of actives identified by LaSSI in the top-scoring 300 compounds are given in column 5, and the corresponding number of singular values (k_{best}) is given in parentheses in column 5. The results of the searches carried out with joint chemical probes are given in columns 6−10. Column 6 gives the MDDR registration numbers of the structurally diverse representatives from single molecule searches. In column 7 are the number of actives from the top 300 ranked compounds

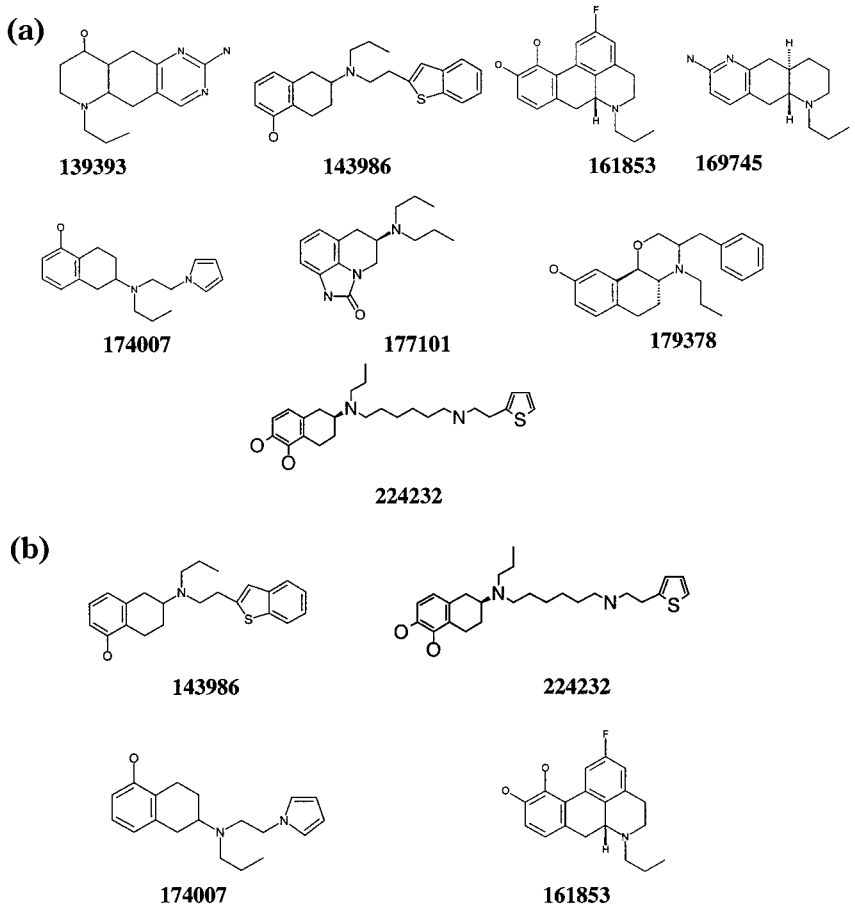

Figure 4. Dopamine agonists joint chemical probe members for (a) LaSSI and (b) TOPOSIM.

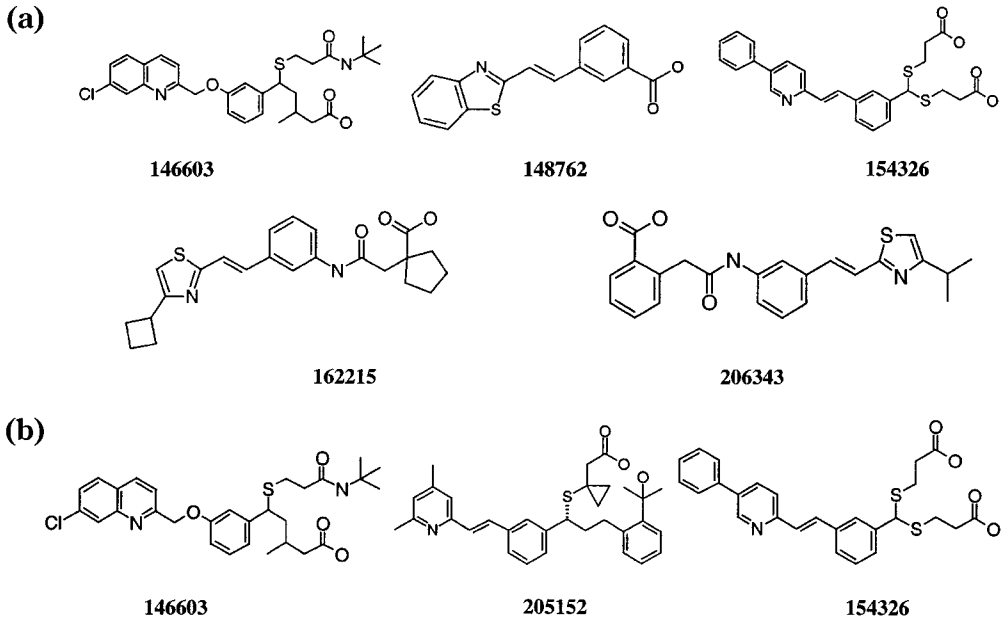

Figure 5. Leukotriene antagonists joint chemical probe members for (a) LaSSI and (b) TOPOSIM.

at a given *k* (column 8). The ranks of the probes in column 9 are those from the results at the value of *k* given in column 8. The value of *k* at which most actives (column 10) are retrieved is given in parentheses in column 10. The last column gives the improvement in the number of actives retrieved by similarity searches with the joint chemical probe over the number of actives retrieved by the similarity searches with the single chemical probe. If the rank of the last joint probe member retrieved was the same at two or more *k* values, we present the results of the smallest *k* value.

For example, in the case of 5-HT re-uptake inhibitors we selected the following three structurally representative compounds from the 11 actives from the single chemical probe similarity search: 166597, 170534, and 222868 (Figure 2a and Table 2). The best ranking of

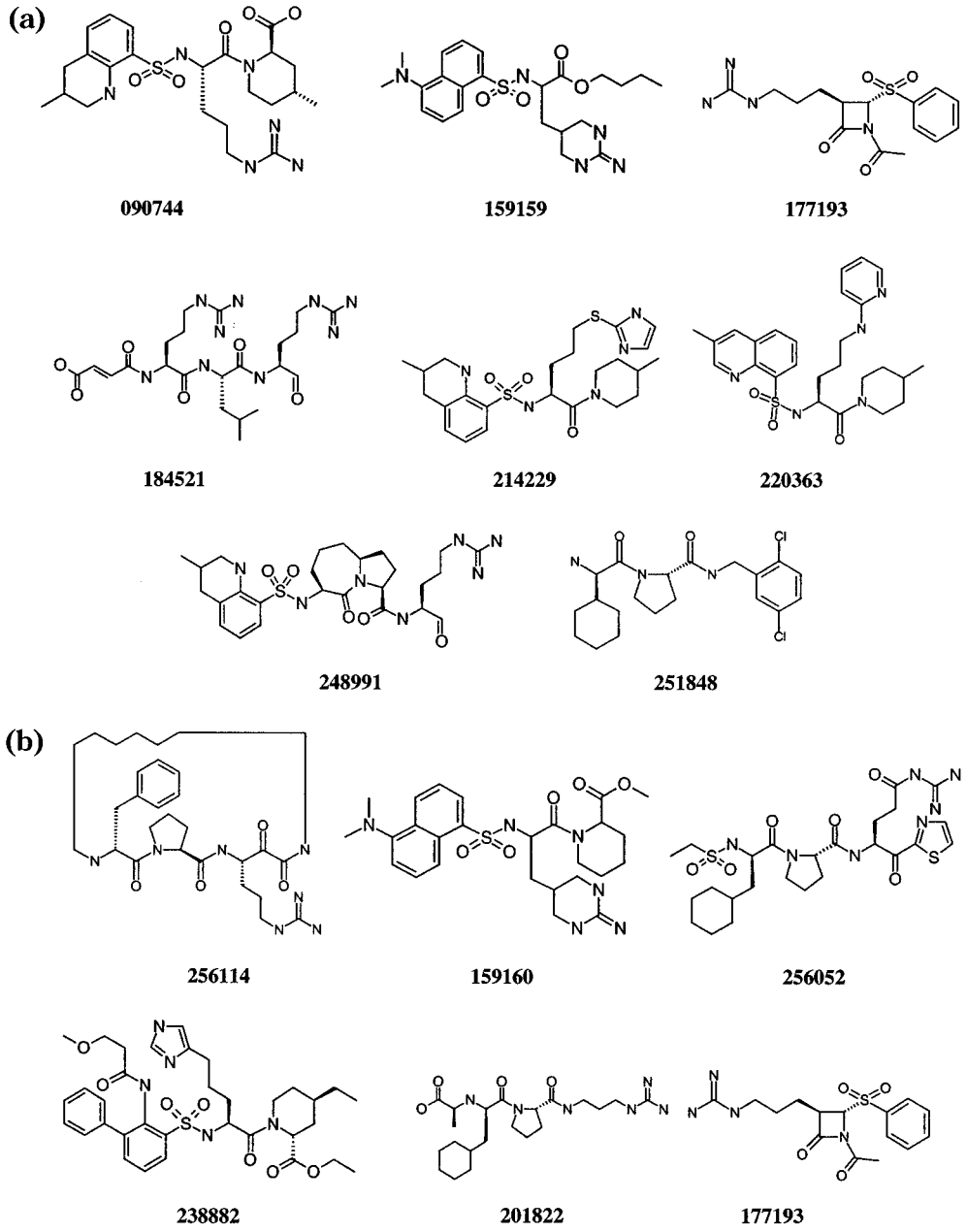

Figure 6. Thrombin inhibitors joint chemical probe members for (a) LaSSI and (b) TOPOSIM.

the joint chemical probes yielded 15 actives (column 7) with the use of 110 singular values (column 8). The best ranking of the joint chemical probe members of 5HT re-uptake inhibitors (column 9) in the scored list is as follows: 170534 ranked 1st, 222868 ranked 4th, 166597 ranked 11th. The highest number of actives were retrieved with the use of 370 singular values yielding 19 actives (column 10), which is better than the results at 110 singular values with the best rankings of the joint chemical probes. There is about 37% improvement (column 11) in the retrieval of the number of actives (15) by the joint chemical probe over the number of actives (11) retrieved by the single chemical probe similarity searches. Percent improvement = [(joint chemical probe actives − single chemical probe actives at $k = 300$)/single chemical probe actives at $k = 300$] × 100.

$$\% \text{ improvement} = (15 - 11)/11 \times 100 = 37\%$$

Joint Chemical Probe Similarity Searches With TOPOSIM

The results of the database searches with joint chemical probes using TOPOSIM are given in Table 3 (columns 5−7). These results are compared with those obtained with single molecule probes (columns 3−4).[7]

For example, in the case of 5-HT re-uptake inhibitors we selected the following three structurally diverse compounds from the six actives (column 4) from the single chemical probe similarity search: 170534, 222869, and 252275 (Figure 2b and Table 3). The rankings of the joint chemical probe members (column 6) of 5-HT re-uptake inhibitors in scored list are as follows: 170534 ranked 1st, 222869 ranked 2nd, 252275 ranked 3rd. There is a 367% improvement (column 8) in the retrieval of the number of actives (28) by the joint chemical probe over the number of actives (6) retrieved by the single chemical probe similarity searches.

Table 2. Results of Joint Chemical Probes Searches with LaSSI

		single chemical probe[a]			joint chemical probe[b]					
therapeutic category	total actives	probe	actives[d] $k = 300$	actives (best k)[e]	JCP[f] members	calibration[c] actives	calibration[c] k	rankings of probes	actives (best k)[e]	improvement[g] (%)
5HT re-uptake inhibitors	219	170534	11	17 (150)	166597	15	110	1:170534	19 (370)	37
					170534			4:222868		
					222868			11:166597		
ACE inhibitors	499	115230	73	78 (410)	090439	152	430	5:204703	152 (430)	108
					152038			9:211892		
					158135			15:152038		
					158137			30:158135		
					167770			38:158137		
					180880			65:090439		
					204703			74:180880		
					211892			87:167770		
dopamine agonists	127	161853	22	26 (320)	139393	44	230	6:174007	48 (290)	100
					143986			9:169745		
					161853			26:143986		
					169745			31:179378		
					174007			53:177101		
					177101			70:161853		
					179378			80:224232		
					224232			99:139393		
leukotriene antagonists	811	205402	106	106 (300)	146603	207	310	1:154326	213 (330)	95
					148762			10:146603		
					154326			23:162215		
					162215			55:206343		
					206343			58:148762		
thrombin inhibitors	493	090744	104	104 (300)	090744	146	370	1:090744	171 (250)	40
					159159			5:214229		
					177193			14:248991		
					184521			23:159159		
					214229			25:220363		
					220363			50:177193		
					248991			55:184521		
					251848			88:251848		

[a] Results presented by Hull et al.[7] [b] This study. [c] Calibration: The scored lists were used to generate ranking for the members of the joint chemical probes. We analyzed these rankings to pick out the singular value at which the members of the joint chemical probe are ranked the best (the rank of the last compound retrieved is the lowest). If there was a tie in the rankings for two or more singular values, we chose the results of the smallest singular value. [d] The number of actives retrieved by LaSSI in the top-scoring 300 compounds out of 82 000 compounds in MDDR database. The compounds used in the probes are not counted among the actives. [e] Best k: The results from the best performing singular value at which most actives are retrieved. [f] Joint chemical probe members. [g] Percent improvement = [(joint chemical probe titration actives − single chemical probe actives at $k = 300$)/single chemical probe actives at $k = 300$] × 100.

$$\% \text{ improvement} = (28 - 6)/6 \times 100 = 367\%$$

Comparison of LaSSI and TOPOSIM Results

The comparison of the retrieval rates and the initial enhancements of the searches using joint chemical probes with LaSSI and TOPOSIM is given in Table 4. The results from the LaSSI searches are given in columns 3−5. The results of the TOPOSIM searches are given in columns 6−8. In column 4, initial enhancements for LaSSI searches are presented. As described in the Methods section, initial enhancement is given by the ratio of the number of actives found in the top 300 compounds and the number of active compounds that can retrieved by chance. In column 7, initial enhancement for TOPOSIM searches are given. The number of actives in the top 300 compounds are given in parentheses in each case. In column 9, the difference in initial enhancements and the number of actives retrieved are given as the percentage change from LaSSI to TOPOSIM. The positive numbers indicate that LaSSI's performance is better than TOPOSIM, the negative numbers indicated that TOPOSIM's performance is better than LaSSI, and the numbers close to zero indicate that both perform equally.

The initial enhancements in all cases show that the use of the joint chemical probes significantly enhances the chances of retrieving active compounds over pure chance. The best initial enhancement for LaSSI is achieved in the case of dopamine agonists (113). However, in the case of TOPOSIM, the best initial enhancement is achieved for thrombin inhibitors (113). The differences in initial enhancements between LaSSI and TOPOSIM show that LaSSI does better in two cases, ACE inhibitors (27%) and dopamine agonists (23%). However, TOPOSIM has better initial enhancements than LaSSI in the cases of 5HT re-uptake inhibitors (−95%) and thrombin inhibitors (−33%). LaSSI and TOPOSIM give similar initial enhancements in retrieving leukotriene antagonists, with both retrieving actives at a rate much better than pure chance.

Discussion

We show that the use of joint chemical probes, in addition to the use of single probes, can significantly enhance retrieval of the active compounds from a database (Table 5). We wanted to verify that the use of the joint chemical probes provide significant advantage in the retrieval of actives over similarity searches with constituents of the joint probe. Therefore, we compared the similarity search results of leukotriene antagonists with the joint chemical probes used by LaSSI and TOPOSIM versus the similarity search results with the constituents of the joint probes. The individual similarity searches with each constituent of the joint probes

Table 3. Results of Joint Chemical Probes Searches with TOPOSIM

therapeutic category	total actives	single chemical probe[a]		joint chemical probe			improvement (%)
		probe	actives[b]	JCP members[c]	rankings of probes	actives[b]	
5HT re-uptake inhibitors	219	170534	6	170534	1:170534	28	367
				222869	2:222869		
				252275	3:252275		
ACE inhibitors	499	115230	62	090388	2:090388	111	79
				152038	4:206856		
				155108	18:220578		
				165334	23:152038		
				185771	144:155108		
				206856	675:165334		
				220578	4156:185771		
dopamine agonists	127	161853	16	143986	1:143986	39	144
				161853	4:224232		
				174007	6:174007		
				224232	42:161853		
leukotriene antagonists	811	205402	186	146603	2:146603	194	4
				154326	8:205152		
				205152	27:154326		
thrombin inhibitors	493	090744	195	159160	4:256114	198	1
				177193	5:159160		
				201822	13:256052		
				238882	22:238882		
				256052	37:201822		
				256114	2175:177193		

[a] Results presented by Hull et al.[7] [b] The number of actives retrieved by TOPOSIM in the top 300 compounds out of 82 000 compounds in MDDR. The compounds used in the probes are not counted among the actives. [c] Joint chemical probe members.

Table 4. Retrieval Rates and Initial Enhancements of LaSSI and TOPOSIM

therapeutic category	total actives	joint chemical probe						
		LaSSI			TOPOSIM			
		JCP members	IE[a] (actives)	ranking of probes	JCP members	IE[a] (actives)	ranking of probes	% diff[b]
5HT re-uptake inhibitors	219	166597	19. (15)	1:170534	170534	37. (28)	1:170534	−95.%
		170534		5:222868	222869		2:222869	(−87)
		222868		11:166597	252275		3:252275	
ACE inhibitors	499	090439	88. (152)	5:204703	090388	64. (111)	2:090388	27.%
		152038		9:211892	152038		4:206856	(27)
		158135		15:152038	155108		18:220578	
		158137		30:158135	165334		23:152038	
		167770		38:158137	185771		144:155108	
		180880		65:090439	206856		675:165334	
		204703		74:180880	220578		4156:185771	
		211892		87:167770				
dopamine agonists	127	139393	113. (44)	6:174007	143986	87. (39)	1:143986	23.%
		143986		9:169745	161853		4:224232	(11)
		161853		26:143986	174007		6:174007	
		169745		31:179378	224232		42:161853	
		174007		53:177101				
		177101		70:161853				
		179378		80:224232				
		224232		99:139393				
leukotriene antagonists	811	146603	49. (207)	1:154326	146603	46. (194)	2:146603	6.%
		148762		10:146603	154326		8:205152	(6.3)
		154326		23:162215	205152		27:154326	
		162215		55:206343				
		206343		58:148762				
thrombin inhibitors	493	090744	85. (146)	1:090744	159160	113. (198)	4:256114	−33.%
		159159		5:214229	177193		5:159160	(−36)
		177193		14:248991	210822		13:256052	
		184521		23:159159	238882		22:238882	
		214229		25:220363	256052		37:201822	
		220363		50:177193	256114		2175:177193	
		248991		55:184521				
		251848		88:251848				

[a] Initial enhancement: ratio of the number of actives found in top 300 ranked compounds and the number active compounds that can be retrieved by chance. [b] % diff = [initial enhancement (actives) of LaSSI − initial enhancement (actives) of TOPOSIM]/[initial enhancement (actives) of LaSSI].

using LaSSI and TOPOSIM retrieved fewer actives than the similarity searches with the joint probe (data not shown).

All the LaSSI searches with the joint probes afforded at least 37% enhancement over searches with the single probes, whereas TOPOSIM searches with joint probes

Table 5. Number of Common Actives Retrieved by LaSSI and TOPOSIM Using Single Chemical (SCP) and Joint Chemical Probes (JCP)

	LaSSI			TOPOSIM			LaSSI and TOPOSIM JCP	
therapeutic category	SCP[a] actives	JCP actives	SCP/JCP common	SCP actives	JCP[b] actives	SCP/JCP common	common	total unique actives
5HT re-uptake inhibitors	11	15	8	6	28 (24)	3 (4)	14 (15)	29 (24)
ACE inhibitors	73	152	55	62	111 (122)	34 (26)	97 (106)	166 (168)
dopamine agonists	22	44	14	16	39 (42)	12 (13)	36 (39)	47 (47)
leukotriene antagonists	106	207	95	186	194 (187)	149 (120)	156 (155)	245 (239)
thrombin inhibitors	104	146	88	195	198 (207)	127 (165)	84 (120)	260 (233)

[a] Similarity search with LaSSI at $k = 300$ using single molecule probe. [b] The values in parentheses are the results from the TOPOSIM similarity searches with LaSSI joint chemical probes.

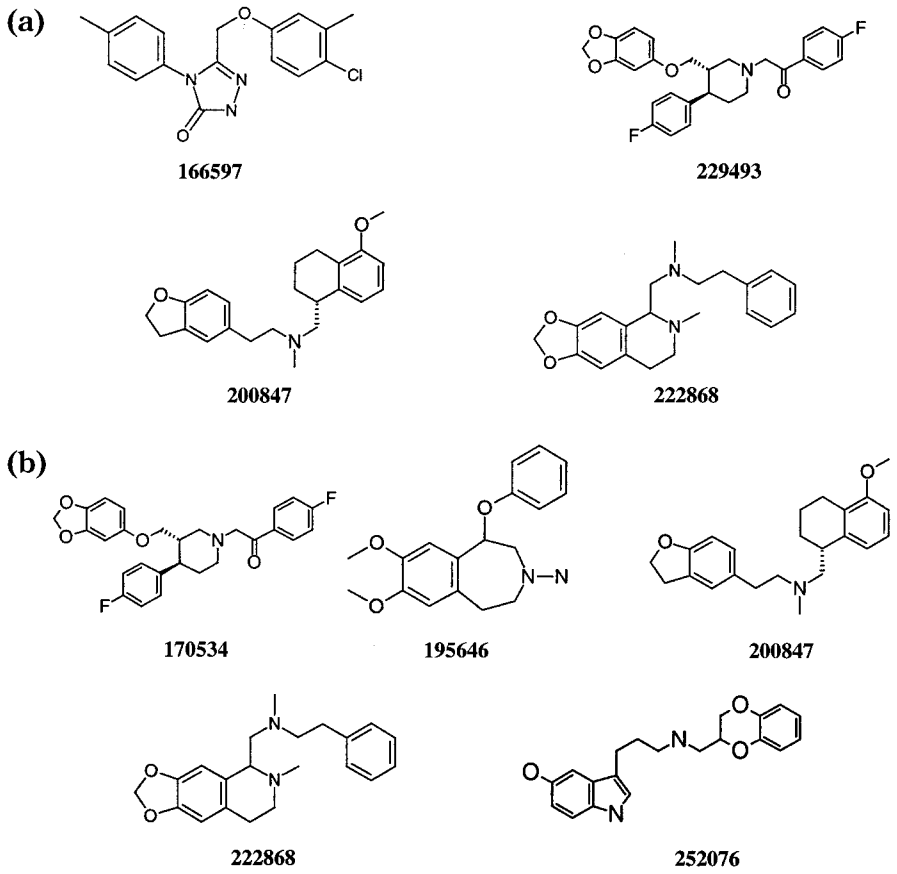

Figure 7. 5HT re-uptake inhibitor structural classes retrieved by (a) LaSSI and (b) TOPOSIM.

afforded significant enhancement in only three out of five cases (see Tables 2 and 3). The method of calibration used here ensures that we can identify the k value a priori at which LaSSI's retrieval rate is close to the best (Table 2). This is true based on the performance in three out of five cases. In two cases where the best k value was not located, the number of actives retrieved were off by less than 26% (5-HT re-uptake inhibitors and thrombin inhibitors, Table 2).

Since LaSSI inherently retrieves diverse chemical structures, it cannot be expected to perform optimally when the actives belonging to a particular therapeutic category are close analogues of each other. LaSSI appears to use the descriptors from the molecules that comprise the joint chemical probe to retrieve compounds that are similar to each of the constituent molecules and similar to those that are hybrids of two or more of these molecules. For example, in the case of ACE inhibitors, the substructure features found in the joint probe members 152038, 158135, and 167770 (Figure 3a) are present in the retrieved compound 219699 (Figure 8a). In another case, the thrombin inhibitor 174301 (Figure 9a) retrieved by LaSSI has constituent elements of 177193 and 214229 (Figure 6a).

The performance of TOPOSIM with the joint chemical probes is significantly enhanced over single probes in three out of five cases. In the case of 5-HT re-uptake inhibitors, there is 367% enhancement in the retrieval of the actives over the single probe searches. There is little enhancement in the retrieval of the actives by use of the joint chemical probes in the searches for leukotriene antagonists and thrombin inhibitors (4% and 1%, respectively, Table 3).

We used two approaches to assess the diversity of compounds in the top-scoring 300 compounds by the two search techniques. In the first approach, we computed the mean similarity of the entire set of actives belonging to a corresponding therapeutic category to their cen-

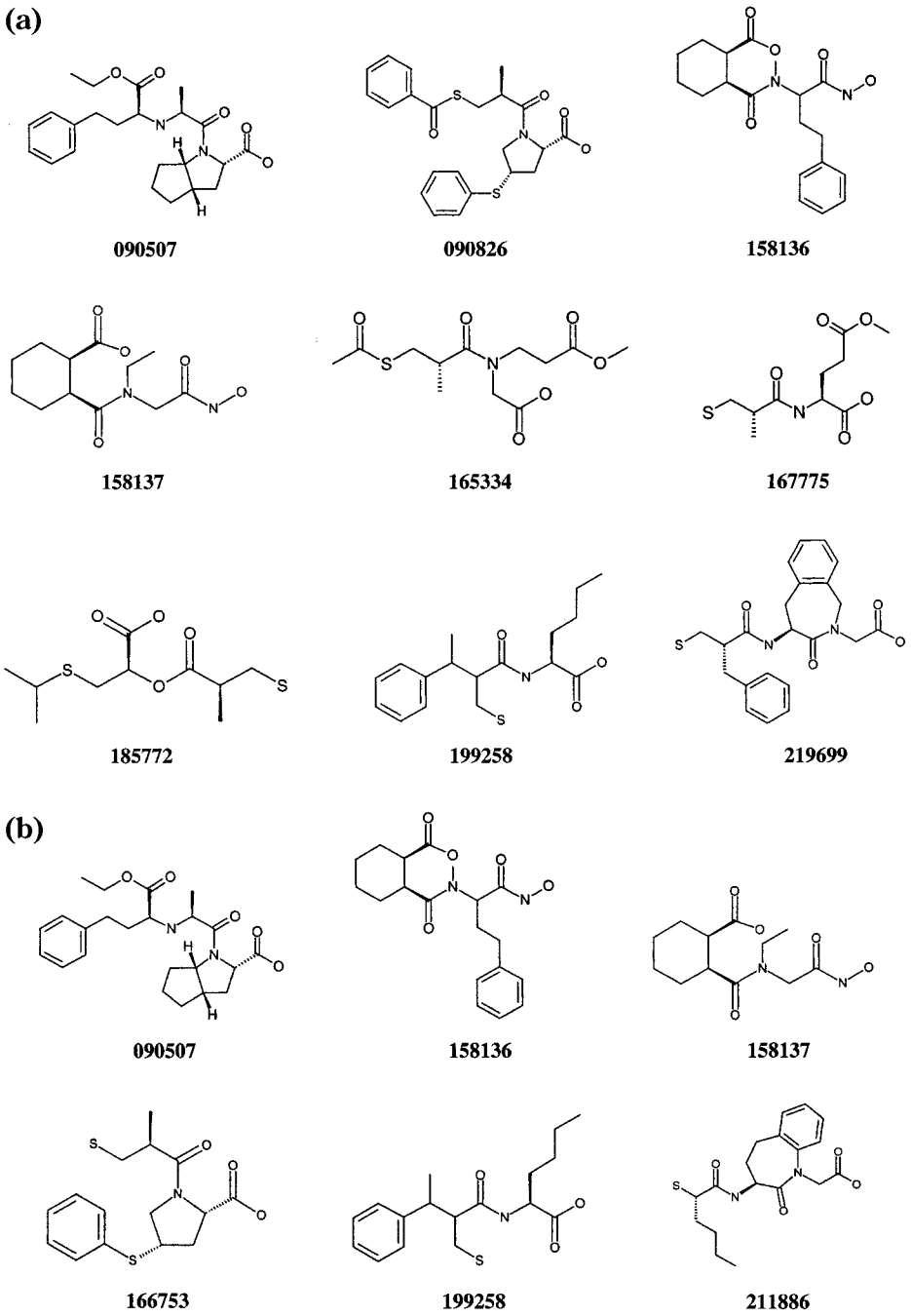

Figure 8. ACE inhibitor structural classes retrieved by (a) LaSSI and (b) TOPOSIM.

troid. The mean, standard deviation, and the range (lowest to highest) of the similarities are given in Table 6. These data show that in all categories ,except the 5-HT re-uptake inhibitors and Leukotriene antagonists, the mean similarity of the retrieved compounds to their centroids is lower for LaSSI than it is for TOPOSIM. These values are a measure of how similar the compounds are to each other and therefore a measure of diversity among the active compounds. In the case of thrombin inhibitors, LaSSI retrieves the most diverse set of compounds (mean similarity of 0.39 for LaSSI actives vs 0.47 for TOPOSIM actives). It is interesting to note that the mean values appear to the mask the wide range of similarity values in each of the cases. In the case of ACE inhibitors, it is interesting to note that the mean similarity values are very similar for LaSSI and TOPOSIM; however, the range of similarity values is wider for LaSSI than for TOPOSIM. Thus, as measured by mean similarity values and the range of average similarity values of the retrieved actives to their centroids, LaSSI retrieves structurally more diverse compounds than TOPOSIM does. Therefore, it is informative to examine the range of similarity values to better understand the diversity of compounds retrieved by each of these techniques.

To assess the issue of diversity from the structural perspective we examined the structural classes retrieved by the two techniques. This was done by generating self-similarity and cross-similarity matrixes between all the actives found by both techniques. We also generated self- and cross-similarity matrixes for the actives that are not in common between the two sets of actives. These similarity matrixes were then used to generate structural classes by grouping compounds together

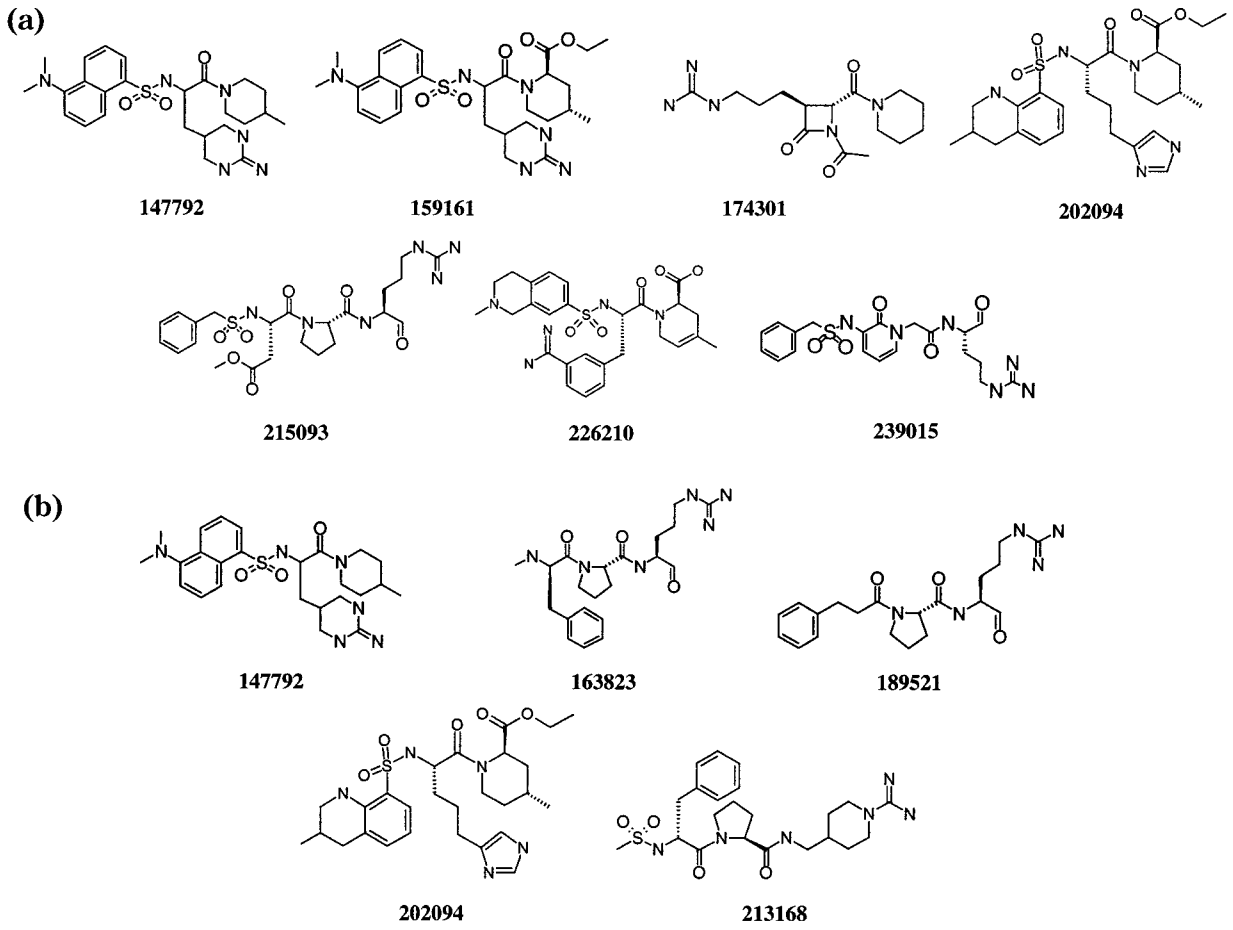

Figure 9. Thrombin inhibitor structural classes retrieved by (a) LaSSI and (b) TOPOSIM.

Table 6. Measure of Diversity of LaSSI and TOPOSIM Similarity Search Results

therapeutic category	LaSSI		TOPOSIM	
	average similarity[a]		average similarity	
	mean ± SD	range	mean ± SD	range
5HT re-uptake inhibitors	0.58 ± 0.13	0.25−0.70	0.53 ± 0.07	0.28−0.63
ACE inhibitors	0.46 ± 0.11	0.10−0.61	0.48 ± 0.09	0.23−0.62
dopamine agonists	0.55 ± 0.12	0.17−0.70	0.59 ± 0.10	0.41−0.74
leukotriene antagonists	0.55 ± 0.12	0.25−0.63	0.53 ± 0.07	0.36−0.66
thrombin inhibitors	0.39 ± 0.07	0.19−0.53	0.47 ± 0.06	0.27−0.60

[a] Average similarity = similarity of each member of the set of retrieved actives to their centroid. The mean of the similarities of each member of the active set to their centroid, standard deviation, and the range (lowest−highest) of similarity values are presented above.

when similarity between any two compounds in a given class was greater than or equal to 0.65. We present the analysis of the results from three therapeutic categories.

In the case of 5-HT re-uptake inhibitors, the self-similarity matrixes revealed that LaSSI retrieved four structurally distinct classes of compounds and TOPOSIM retrieved five structural classes. The cross-similarity matrix computed between LaSSI actives and TOPOSIM actives yielded four structural classes, of which three classes are common between the two methods (see Figure 7). LaSSI found one structural class that TOPOSIM did not and vice versa. The distinct structural classes found by LaSSI and TOPOSIM are similar to the structures used in their joint probes. In the case of TOPOSIM there is one unique structure occurring as a singleton which is not found by LaSSI (195646, Figure 7b).

In the case of ACE inhibitors, LaSSI retrieved nine structurally distinct classes with two or more representative members, and five singletons. On the other hand, TOPOSIM retrieved six structurally distinct classes with two or more representative members, and four singletons. The representative structures from these classes retrieved by each of these two methods are shown in Figure 8. The four structural classes retrieved by the two methods are represented by the following four compounds: 090507, 158136, 158137, and 199258 (Figure 8a). LaSSI has five structural classes distinct from the compounds retrieved by TOPOSIM. These are represented by the following three compounds: 090826, 165334, and 185772 (Figure 8a). TOPOSIM, on the other hand, has no structural class that is distinctly different from that of LaSSI's. In the search for ACE inhibitors using the joint probe with TOPOSIM, we see that TOPOSIM with the Dice similarity measure did not rank two members of the joint chemical probe in the top 300 compounds (165334 and 185771, Table 4). Thus it is not surprising that the searches did not retrieve compounds belonging to the classes represented by those two compounds. This case shows that when there is a wide variety of structural classes represented in the database belonging to a particular therapeutic category, the use of LaSSI with a joint chemical probe yields compounds with greater structural diversity than

does TOPOSIM. This is due to the fact that LaSSI, unlike TOPOSIM, retrieves compounds corresponding to the constituents of the joint probe, and when queried with the probe, it retrieves compounds that have probe descriptors and those descriptors that correlate with them.

In the case of thrombin inhibitors, TOPOSIM found more actives than LaSSI; however, LaSSI found seven structurally distinct classes of compounds whereas TOPOSIM found five (see Figure 9). It appears that TOPOSIM, once it locks into a structural class, retrieves as many compounds as it can corresponding to that class and hence results in fewer structural classes but a greater number of compounds. However, LaSSI has fewer members belonging to the structural classes it retrieves and hence overall retrieves fewer actives than does TOPOSIM. The results presented in Table 4 show that TOPOSIM retrieved only six of its seven members of the joint chemical probe in the top-scoring 300 compounds and hence could not retrieve the corresponding class of compounds belonging to the missing compound (177193, compare Figures 5b and 9b).

Conclusions

We have demonstrated here that the use of joint chemical probes to mine chemical databases significantly enhances the rate of retrieval of active compounds compared with the single molecule probes. The compounds retrieved by both LaSSI and TOPOSIM closely resemble the members of the joint chemical probes; with LaSSI, the compounds retrieved may be hybrid molecules that contain one or more fragments of the compounds comprising the joint probe. In similarity searches with joint chemical probes, LaSSI enhances the retrieval of actives by at least 37% over the searches with single chemical probes. On the other hand, TOPOSIM shows significant enhancement in only three out of five cases. We have presented a calibration method to fine-tune similarity searches with LaSSI to identify the number of singular values at which the greatest number of actives are retrieved from a chemical database. LaSSI inherently retrieves diverse chemical structures through correlated descriptors whereas TOPOSIM retrieves only compounds that share descriptors with the probe. The analysis of our results here show that more than 30% of the actives retrieved from the database are found by both LaSSI and TOPOSIM, and in almost all cases LaSSI retrieves all the structural classes that TOPOSIM retrieves and more. This suggests that an initial search with LaSSI would lead us to identify all possible structural classes that are related to the joint chemical probe and that a subsequent similarity search with TOPOSIM with each individual compound or a combination of compounds will identify most of the actives in a given class.

Acknowledgment. We thank Chris Culberson, Laurie Castonguay, and Bruce Bush for their comments on this manuscript.

References

(1) Willet, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(2) Grethe, G.; Hounshell, W. D. Similarity searching in the development of new bioactive compounds. An application. In *Chemical Structures 2*; Warr, W. A., Ed.; Springer-Verlag: Heidelberg, 1993; pp 399−407.
(3) Sheridan, R. P.; Kearsley, S. K. Using a genetic algorithm to suggest combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.
(4) Nachbar, R. N. Molecular evolution: automated manipulation of hierarchical chemical topology and its application to average molecular structures. *Gen. Prog. Evolv. Mach.* **2000**, *1*, 57−94.
(5) Shemetulskis, Weininger, D.; Blankey, C. J.; Yang, J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.
(6) Hull, R. H.; Singh, S. B.; Nachbar, R. N.; Sheridan, R. P.; Kearsley, S. K.; Fluder, E. M. Latent Semantic Structure Indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.* **2001**, *44*, 1177−1184.
(7) Hull, R. H.; Fluder, E. M.; Singh, S. B.; Kearsley, S. K.; Nachbar, R. N.; Sheridan, R. P. Chemical Similarity Searches Using Latent Semantic Structure Indexing (LaSSI) and Comparison to TOPOSIM. *J. Med. Chem.* **2001**, *44*, 1185−1191.
(8) Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landuaer, T. K.; Harshman R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41* (6), 391−407.
(9) Salton, G.; Buckley, C. Improving retrieval performance by relevance feedback. *J. Am. Soc. Inf. Sci.* **1990**, *41* (4), 288−297.
(10) 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*, 118−127.
(11) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(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) Berry, M.; Do, T.; O'Brien, G.; Krishna, V.; Varadhan, S. SVDPACKC (version 1.0) User's guide (UTK technical report C_93-194, revised March 1996). University of Tennessee, Knoxville, Department of Computer Science.

JM000398+