Chemical Similarity Searches Using Latent Semantic Structural Indexing (LaSSI) and Comparison to TOPOSIM

Richard D. Hull,[†] Eugene M. Fluder,* Suresh B. Singh, Robert B. Nachbar, Simon K. Kearsley, and Robert P. Sheridan

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

Received September 6, 2000

Similarity searches based on chemical descriptors have proven extremely useful in aiding large-scale drug screening. Here we present results of similarity searching using **La**tent **S**emantic **S**tructure **I**ndexing (LaSSI). LaSSI uses a singular value decomposition on chemical descriptors to project molecules into a *k*-dimensional descriptor space, where *k* is the number of retained singular values. The effect of the projection is that certain descriptors are emphasized over others and some descriptors may count as partially equivalent to others. We compare LaSSI searches to searches done with TOPOSIM, our standard in-house method, which uses the Dice similarity definition. Standard descriptor-based methods such as TOPOSIM count all descriptors equally and treat all descriptors as independent. For this work we use atom pairs and topological torsions as examples of chemical descriptors. Using objective criteria to determine how effective one similarity method is versus another in selecting active compounds from a large database, we find for a series of 16 drug-like probes that LaSSI is as good as or better than TOPOSIM in selecting active compounds from the MDDR database, if the user is allowed to treat *k* as an adjustable parameter. Typically, LaSSI selects very different sets of actives than does TOPOSIM, so it can find classes of actives that TOPOSIM would miss.

Introduction

Similarity searches are now a standard tool for drug discovery.[1,2] The idea behind such searches is that, given a compound with an interesting biological activity, compounds that are "similar" to it in structure are likely to have a similar activity. In practice, an investigator provides a chemical structure as a "probe", searches over a database of sample-available compounds, and finds those that are most similar, which are then submitted for testing. Similarity searching can be done on the basis of 2D or 3D structure. 2D similarity searches, especially those based on comparing lists of precomputed substructure descriptors, are computationally very inexpensive.

At Merck & Co. Inc. we have set up a system (TOPOSIM) with which a user can specify a probe and search over descriptor databases. One set of useful substructure descriptors with which we have experience are those developed at Lederle Laboratories: the atom pair (AP)[3] and topological torsion (TT).[4] These descriptors are typically able to discover active compounds in different chemical classes from the probe. However, the original AP and TT descriptors are very specific: they distinguish atom types on the basis of element, number of non-hydrogen neighbors, and number of π electrons. This does not allow for the perception of physiochemical equivalence (e.g., carboxylate with tetrazole). It is desirable to add enough "fuzziness" to the searches so that compounds significantly different from the probe will be discovered. One way of doing this is to keep the method of calculating similarity the same but to modify the descriptors. We previously[5] experimented with descriptors of the same form as APs and TTs but with alternative "physiochemical atom types". Another way is to keep the original descriptors but modify the method for calculating similarity, hence the use of LaSSI.

LaSSI (**La**tent **S**emantic **S**tructure **I**ndexing),[6] discussed in detail in a companion paper, provides a radical departure from our usual methods of similarity calculation. Given a large database of molecules, matrix **X** is formed by elements d_{ji} which are the frequency of descriptor *j* in molecule *i*. **X** is expressed as the product of three matrices by singular value decomposition such that:

$$\mathbf{X} = \mathbf{P}\Sigma\mathbf{Q}^T$$

where **P** is the matrix of eigenvectors of $\mathbf{X}\mathbf{X}^T$, **Q** is the matrix of eigenvectors of $\mathbf{X}^T\mathbf{X}$, and Σ is the diagonal matrix of singular values (the square roots of the nonzero eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$). Keeping the *k* largest eigenvalues (also called singular values) gives the best-rank *k* approximation to **X**:

$$\mathbf{X}_k = \mathbf{P}_k\Sigma_k\mathbf{Q}_k{}^T$$

The rows of $\mathbf{P}_k$ are the projected coordinates of the descriptors from the database in a *k*-dimensional space. These "latent descriptors" are orthogonal to each other but linear combinations of the original descriptors. The rows of $\mathbf{Q}_k$ are the projected coordinates of the molecules in that same space. If *k* is small relative to the number of unique descriptors in the database, there are two effects: some descriptors may become less important in calculating similarity, and the descriptors become less

* To whom correspondence should be addressed. Tel: 732-594-5074. Fax: 732-594-4224. E-mail: fluder@merck.com.

[†] Present address: Elagent Corp., 7011 N. Atlantic Ave., Suite 200, Cape Canaveral, FL 32920.

10.1021/jm000392k CCC: $20.00 © 2001 American Chemical Society
Published on Web 03/20/2001

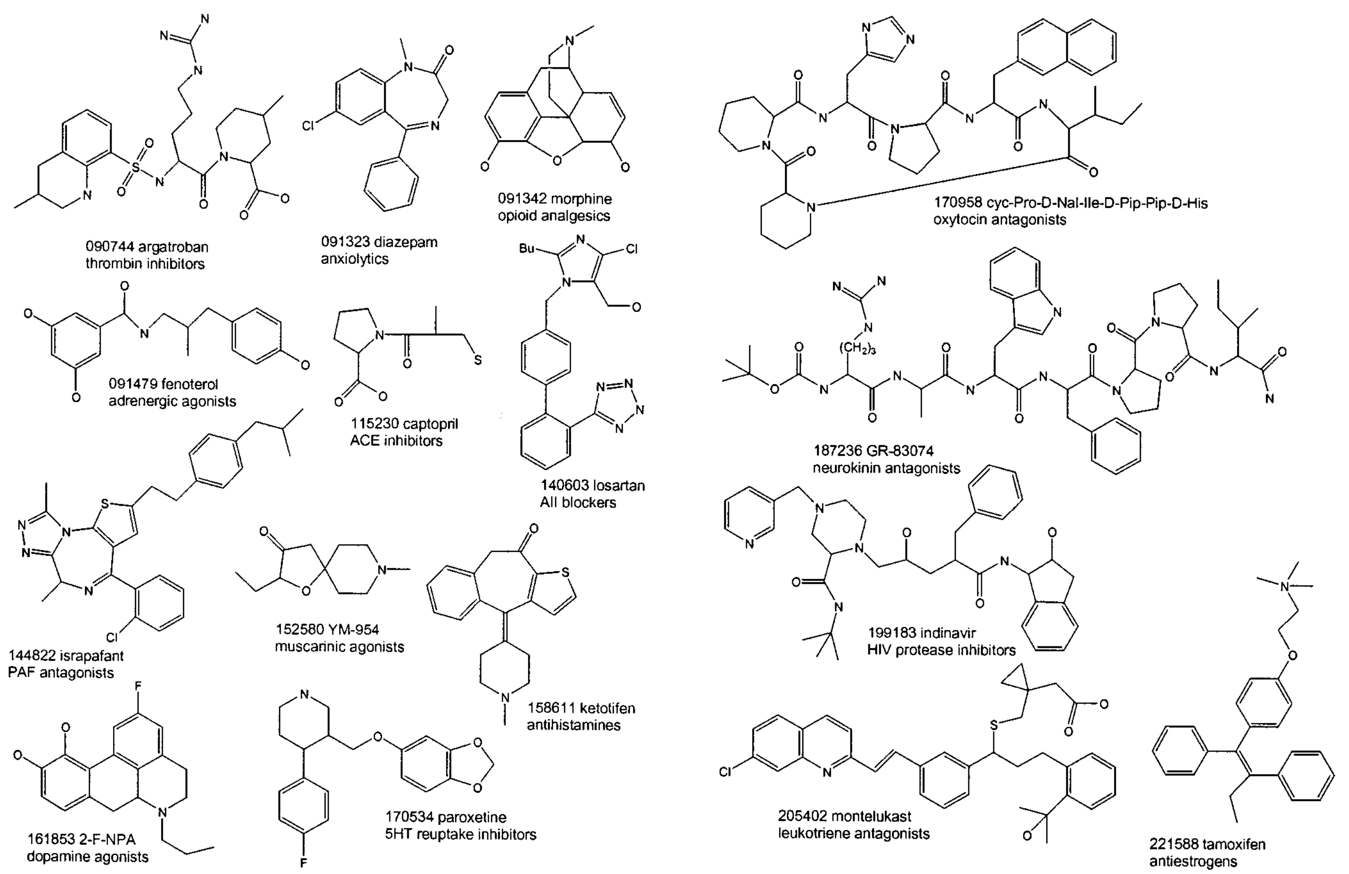

Figure 1. Drug-like probes used in this study. Each is labeled by the MDDR external registry, its name, and associated activity.

Table 1. Probes and Activity Keywords Used in This Study

MDDR registry no. of probe	probe name	activity keywords from MDDR	no. of actives
090744	argatroban	thrombin inhibitor	493
091323	diazepam	anxiolytic benzodiazepine benzodiazepine agonist	3820
091342	morphine	analgesic, opioid opioid agonist κ agonist δ agonist μ agonist	869
091479	fenoterol	adrenergic (β) agonist	161
115230	captopril	ACE inhibitor	490
140603	losartan	angiotensin II blocker	2229
144822	israpafant	PAF antagonist	1240
152580	YM-954	muscarinic (M1) agonist	858
158611	ketotifen	antihistaminic	616
161853	2-F-NPA	dopamine (D2) agonist	127
170534	paroxetine	5-HT reuptake inhibitor	219
170958	cyc-Pro-D-Nal-Ile-D-Pip-Pip-D-His	oxytocin antagonist	176
187236	GR-83074	neurokinin antagonist	150
199183	indinavir	HIV-1 protease inhibitor	641
205402	montelukast	leukotriene antagonist	1165
221588	tamoxifen	antiestrogen	233

independent. Thus, by changing the value of *k* in the LaSSI calculation, the user can adjust the amount of "fuzziness" in the method.

In this paper we demonstrate the utility of LaSSI for searching large databases of chemical structures. Using 16 drug-like probe molecules, we show that LaSSI is about as good as or better than TOPOSIM for selecting active compounds from a large database of drug-like molecules, but LaSSI consistently selects very different sets of actives.

Methods

Definitions of Descriptors. Here we will use the atom pair described by Carhart et al.[3] and the topological torsion described by Nilakantan et al.[4] Atom pairs are substructures of the form:

$$\mathrm{AT}_i - \mathrm{AT}_j - (\mathrm{distance})$$

where AT_i is the atom type of *i* and (distance) is the distance in bonds between atom *i* and atom *j* along the shortest path.

The topological torsion is of the form:

$$\mathrm{AT}_i - \mathrm{AT}_j - \mathrm{AT}_k - \mathrm{AT}_l$$

where *i*, *j*, *k*, and *l* are consecutively bonded atoms. A previous publication[5] gives examples of a molecule partitioned into these descriptors.

Definitions of Similarity for TOPOSIM. The default similarity definition for TOPOSIM is the Dice index. The similarity of molecules *A* and *B* is:

$$\mathrm{Sim}_{AB} = \frac{\sum_j \min(d_{jA}, d_{jB})}{0.5[\sum_j d_{jA} + \sum_j d_{jB}]}$$

where d_{jA} is the count of descriptor *j* in molecule *A*. The index *j* goes over the union of unique descriptors in *A* and *B*. Sim_{AB} ranges from 0.0 (nothing in common) to 1.0 (identity). Two other popular definitions were tried, cosine and Tanimoto (reviewed in ref 2).

TOPOSIM Searches. We run similarity searches with our in-house system TOPOSIM. During a search of a descriptor database, TOPOSIM calculates for each database entry the similarity against the probe for the AP or TT descriptor. In the simplest case, the score of a molecule is its similarity. A combination score APTT is produced by taking the mean of the AP and TT similarities. Previous work[5] shows that combination descriptors are sometimes better at selecting actives than either descriptor alone.

LASSI Scores. Each database entry is already projected into the *k*-dimensional space and resides in the $\mathbf{Q}_k$ matrix. Some probes may also be in the original database. A probe $\mathbf{z}$ not already in the database must be projected into the same space by $\mathbf{y} = \mathbf{z}^T\mathbf{P}_k\Sigma_k^{-1}$. The similarity of the probe and database molecules *B* in LaSSI is the cosine similarity (from −1 to 1):

$$\text{LaSSI similarity} = \sum_{x=1}^{k} y_x q_{Bx}/|\mathbf{y}||\mathbf{q}_B|$$

q_{Bx} is an element of $\mathbf{Q}_k$ and $\mathbf{q}_B$ is a row of $\mathbf{Q}_k$ corresponding to molecule *B*.

Details are presented in our companion paper.[6] It should be noted that similarity in LSI differs from LaSSI; in LSI Σ_k is used in calculating the similarity.

LaSSI Searches. For LaSSI three separate versions of $\mathbf{X}$ and their corresponding singular value decompositions were formed from AP descriptors alone, TT descriptors alone, or both together. For any given probe and combination of descriptors, we calculated the LaSSI similarity of the probe to each of the candidates, treating the number of singular values *k* as an adjustable parameter. We ran each LaSSI search with $k = 2$, 10, 20, ... up to 1000. Note that 1000 singular values is still very small compared to the number of unique descriptors in most databases. In one part of the study we used the value of *k* that gave the highest initial enhancement (see below).

Sorting of Scores. Once all the scores are calculated for a database of molecules, whether from TOPOSIM or LaSSI, they are sorted from high to low score. Ranks are then assigned: the molecule with the highest score is rank 1, the next highest rank 2, etc. We use only the ranks of the compounds in this study, since the distribution of absolute scores varies from one descriptor to another and from TOPOSIM to LaSSI.

Measures of Merit for Similarity Searches. In a previous paper[5] we proposed two measures of goodness for similarity methods. The measures are based on a retrospective screening experiment. Imagine a database of *N* candidates. The candidates are tested in order of decreasing similarity score, and the cumulative number of actives found is monitored as a function of candidates tested. The measures are: (1) How many compounds must be tested until one-half of the actives are found? We called this number A_{50}. A_{50} can be more usefully expressed as a *global enhancement*, the ratio of the A_{50} expected for the random case (*N*/2) over the actual A_{50}. (2) How many actives are found after testing an arbitrary small fraction of the total database? For instance, the number of actives at 300 compounds tested could be called *A*@300. *A*@300 is better expressed as an *initial enhancement*, how many more actives in the top 300 than expected by chance.

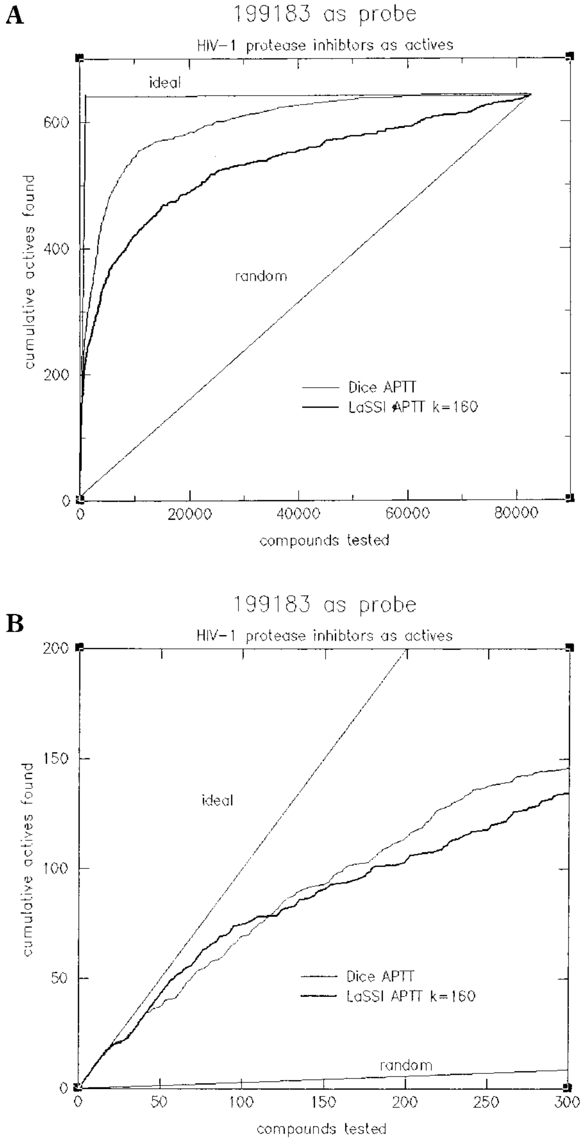

Figure 2. Curve for the accumulation of actives vs rank for the 199183 example. Two limiting cases are also shown: “ideal”, where all the actives would be at the front of the list, and “random”, where all the actives would be randomly distributed throughout the list. The closer the curve approximates the ideal line, the better the method: (A) curve over the entire database; (B) closeup on the origin of A.

Diversity. Our expectation is that LaSSI will find a more diverse set of actives than TOPOSIM, in the sense that we want to see more actives that are not obvious analogues of the probe, especially at ranks ≤ 300. We need a way to measure diversity to confirm this. There is an unavoidable circularity in comparing similarity methods by a diversity measure since diversity itself depends on a particular definition of similarity. Our resolution is to settle on the Dice similarity with the TT descriptor as a standard. In earlier work,[5] the TT was the least fuzzy descriptor, and it has been our experience that only close analogues are recognized as very similar. One simple diversity measure, which we will call the MSP300, is equal to the mean Dice TT similarity of the probe with all the molecules in the top 300 (not including the probe

Table 2. Measures of Merit for Dice and LaSSI Where the Number of Singular Values *k* Is Optimized

	AP			TT			APTT		
probe/activity	Dice	LaSSI	best *k*	Dice	LaSSI	best *k*	Dice	LaSSI	best *k*
				Global Enhancement					
090744/thrombin inhibitors	55.7	35.8	160	33.7	19.0	290	71.6	53.2	170
091323/anxiolytics	1.3	1.1	320	1.5	1.1	20	1.5	1.1	220
091342/opioid analgesics	2.2	1.6	800	1.1	3.3	40	1.7	1.7	470
091479/adrenergic agonists	1.5	28.7	330	27.3	77.3	220	9.4	14.6	170
115230/ACE inhibitors	18.7	14.2	1000	18.1	17.2	650	18.7	17.8	950
140603/AII blockers	36.7	36.0	100	36.6	35.7	110	36.9	36.1	100
144822/PAF antagonists	2.5	1.7	970	1.4	1.3	260	2.0	1.9	850
152580/muscarinic agonists	12.8	16.1	100	6.3	4.7	20	13.5	14.4	70
158611/antihistamines	2.1	2.3	430	1.4	2.0	260	1.6	2.0	430
161853/dopamine agonists	4.5	7.1	760	4.6	27.5	80	5.9	6.6	800
170534/5-HT reuptake inhibitors	3.2	2.0	300	1.6	0.9	170	2.5	2.5	150
170958/oxytocin antagonists	2.8	2.2	100	1.8	3.0	260	2.5	1.7	510
187236/neurokinin antagonist	4.3	1.8	90	3.7	2.3	5	4.6	7.1	100
199183/HIV protease inhibitors	22.1	20.4	60	17.2	6.5	260	21.5	10.9	160
205402/leukotriene antagonists	8.7	7.2	50	6.1	3.2	220	9.2	3.1	420
221588/antiestrogens	2.9	4.1	300	2.9	3.1	270	3.7	5.2	650
mean	11.4	11.4		10.3	13.0		12.9	11.2	
				Initial Enhancement (@300)					
090744/thrombin inhibitors	90.2	79.0	160	89.1	75.1	290	109.2	83.5	170
091323/anxiolytics	4.7	6.2	320	4.4	4.3	20	5.7	6.9	220
091342/opioid analgesics	17.5	23.2	800	30.8	26.1	40	30.2	30.2	470
091479/adrenergic agonists	32.6	34.3	330	44.6	72.1	220	37.7	42.9	170
115230/ACE inhibitors	34.9	76.1	1000	29.3	47.9	650	34.9	71.6	950
140603/AII blockers	37.2	37.2	100	37.2	37.2	110	37.2	37.3	100
144822/PAF antagonists	23.2	29.6	970	32.1	34.1	260	31.2	32.7	850
152580/muscarinic agonists	46.0	49.9	100	29.9	36.7	20	45.1	51.2	70
158611/antihistamines	30.0	44.8	430	51.6	59.2	260	44.8	50.7	430
161853/dopamine agonists	17.4	84.8	760	50.0	60.9	80	34.8	78.3	800
170534/5-HT reuptake inhibitors	18.9	18.9	300	5.0	7.6	170	7.6	22.7	150
170958/oxytocin antagonists	20.4	23.54	100	21.9	18.8	260	20.4	23.5	510
187236/neurokinin antagonists	11.0	16.7	90	12.9	14.7	5	12.9	27.6	100
199183/HIV protease inhibitors	55.6	56.0	60	60.3	69.8	260	62.9	58.2	160
205402/leukotriene antagonists	37.2	37.9	50	42.9	33.0	220	44.1	35.8	420
221588/antiestrogens	54.5	51.0	300	53.3	47.4	270	66.4	65.2	650
mean	33.2	41.8	366 ± 321	37.2	40.3	195 ± 154	39.1	44.9	388 ± 284

Table 3. Enhancements for best *k* vs *k* = 400

	global enhancement			initial enhancement			
probe/activity	Dice APTT	LaSSI APTT best *k*	LaSSI APTT *k* = 400	Dice APTT	LaSSI APTT best *k*	LaSSI APTT *k* = 400	best *k*
090744/thrombin inhibitors	71.6	53.2	6.4	109.2	83.5	57.1	170
091323/anxiolytics	1.5	1.1	1.1	5.7	6.9	5.6	220
091342/opioid analgesics	1.7	1.7	1.3	30.2	30.2	28.0	470
091479/adrenergic agonists	9.4	14.6	34.9	37.7	42.9	27.4	170
115230/ACE inhibitors	18.7	17.8	15.1	34.9	71.6	45.1	950
140603/AII blockers	36.9	36.1	30.0	37.2	37.3	37.2	100
144822/PAF antagonists	2.0	1.9	1.6	31.2	32.7	29.4	850
152580/muscarinic agonists	13.5	14.4	3.0	45.1	51.2	33.2	70
158611/antihistamines	1.6	2.0	1.9	44.8	50.7	50.2	430
161853/dopamine agonists	5.9	6.6	11.6	34.8	78.3	54.4	800
170534/5-HT reuptake inhibitors	2.5	2.5	1.7	7.6	22.7	8.8	150
170958/oxytocin antagonists	2.5	1.7	2.1	20.4	23.5	22.0	510
187236/neurokinin antagonist	4.6	7.1	7.8	12.9	27.6	20.3	100
199183/HIV protease inhibitors	21.5	10.9	4.8	62.9	58.2	43.1	160
205402/leukotriene antagonists	9.2	3.1	3.1	44.1	35.8	35.6	420
221588/antiestrogens	3.7	5.2	3.0	66.4	65.2	51.0	650
mean	12.9	11.2	8.1	39.1	44.9	34.3	

itself if present in the list). One could do the same with only the actives in the top 300, but that would not be as useful because there are many situations where the number of such actives is very small.

Database Used in This Study. To measure the merit of the descriptors we need to have a database of molecules for which we know the biological activities. For this purpose, we use the MDDR (MDL Drug Data Report),[7] which is a licensed database of drug-like molecules compiled from the patent literature. We constructed a database of ~82 000 molecules from version 98.1. There are ~10 200 unique APs and ~5 900 unique TTs in this set. Most structures have one or more key words in the "therapeutic category" field. We will assume that a molecule is active as an HIV protease inhibitor, for instance, if it contains the key word "HIV-1 protease inhibitor" in this field. There are some unavoidable limitations to using patent databases such as MDDR. Since not every compound has been tested in every area, one cannot assume that a compound without a particular key word is inactive. Thus there are probably a number of false inactives. The opposite problem is that for some key words, not all actives work by the same mechanism as the probe (for instance by binding to the same

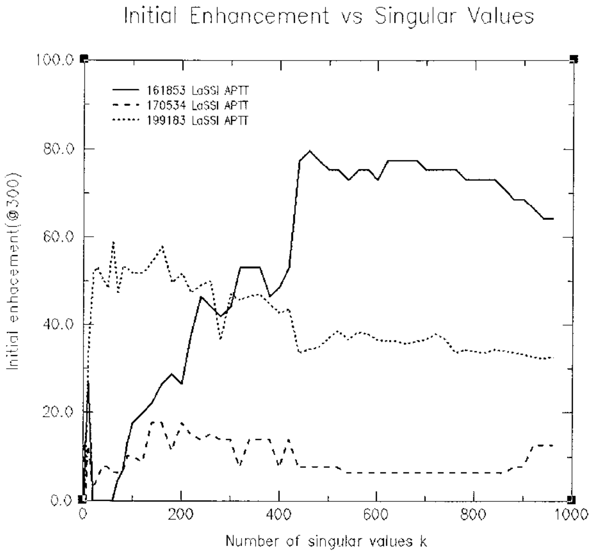

Figure 3. Initial enhancement for LaSSI APTT vs the number of singular values *k* for three examples.

receptor site) and we should not necessarily expect all actives to resemble the probe. Thus there are also some false actives. Another quirk of patent-based databases is that some entries are actually Markush representations. Despite these complications, comparisons between similarity methods should be valid, because for any given probe the number of false actives and false inactives is the same for all methods.

Choice of Example Probes for Similarity Searches. The probes are shown in Figure 1. Table 1 shows how the activities were constructed from key words in MDDR. The probes and the corresponding therapeutic category were selected such that the following were true: (1) the probe itself was typical of a drug-like molecule or at least could be considered a plausible medicinal chemistry lead, (2) compounds in the same therapeutic category as the probe were fairly numerous and several chemical classes were present, (3) the therapeutic category was fairly specific, so that most of the molecules probably work by the same mechanism.

Results

Measures of Merit for Standard Similarity Searches. Figure 2 shows as an example of the graph for the accumulation of actives versus rank for the probe 199183. Table 2 lists measures of merit for Dice vs LaSSI similarities with the optimized value of *k*. The last row of each table shows the enhancement averaged over all the probes. This number can be taken as a qualitative measure of goodness of the method.

For the sake of space, we do not show the results of cosine definition for TOPOSIM, but it is clear that cosine gives measures of merit that are consistently lower than for Dice. Also, for any given probe, the Tanimoto definition gives ranks identical to Dice. Thus we keep Dice as the similarity definition of choice for TOPOSIM. The LSI method for calculating similarity[6] produces much poorer results than the LaSSI similarity.

Table 2 shows that LaSSI and Dice are roughly equivalent in global enhancement and there is no clear advantage to using APTT vs AP and TT individually. However, for initial enhancement there is a clear advantage of LaSSI over Dice. This is not surprising since *k* was adjusted to maximize the initial enhancement. There is also clear advantage in the initial enhancement in using APTT vs AP or TT for both Dice and LaSSI. The optimum *k* for LaSSI varies from 5 to 1000 singular values for AP and TT descriptors and from 70 to 950 for APTT. It is interesting that the mean optimum *k* is much smaller for TT than AP. In our previous work[5] TT was shown to be much less fuzzy than AP, and it is probable that using fewer singular values in LaSSI is adding back some needed fuzziness. For most practical searches, where the number of compounds to be selected is much smaller than the size of the database, the initial enhancement is the more important measure, so henceforth we will emphasize the initial enhancement over the global enhancement. Also, when comparing Dice against LaSSI, we will henceforth consider only the APTT combination since it appears better than either descriptor alone for both methods.

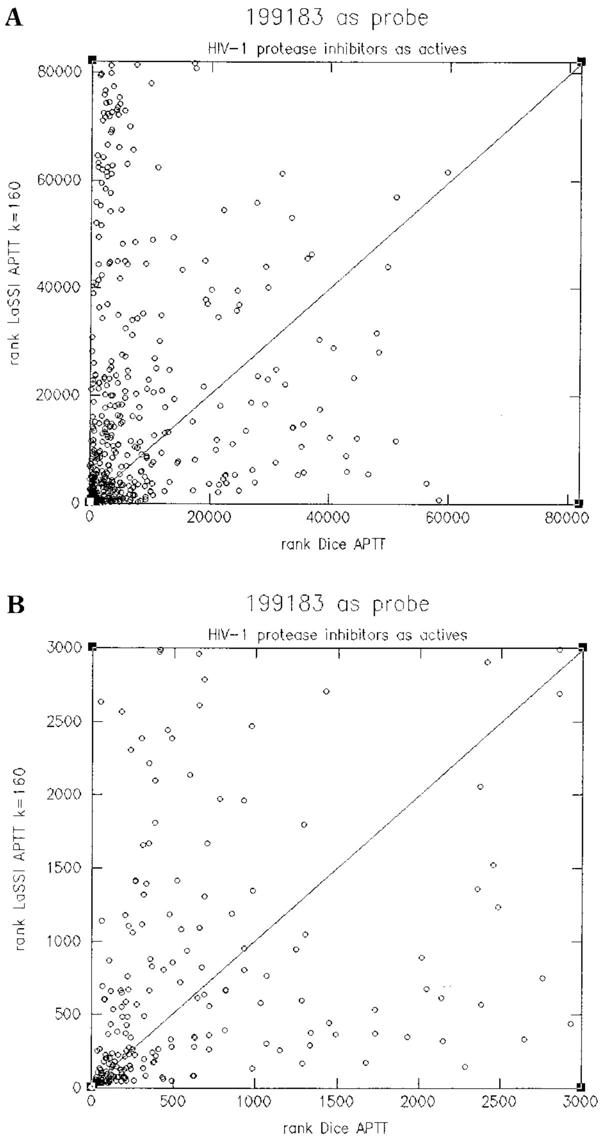

Figure 4. Correlation of rank for Dice APTT and LaSSI APTT. The example is 199183 using 170 singular values. Each circle represents a HIV protease inhibitor: (A) scatterplot over the entire database; (B) closeup of the origin of A.

In a real situation a user would not know the actives in advance. It is therefore critical to know how sensitive the measures of merit are to *k*. Figure 3 shows the initial enhancement as a function of number of *k* for three examples. Clearly the results can be somewhat sensitive to *k* ,and different examples show different sensitivities. If one is to choose a value of *k* to start with, one might

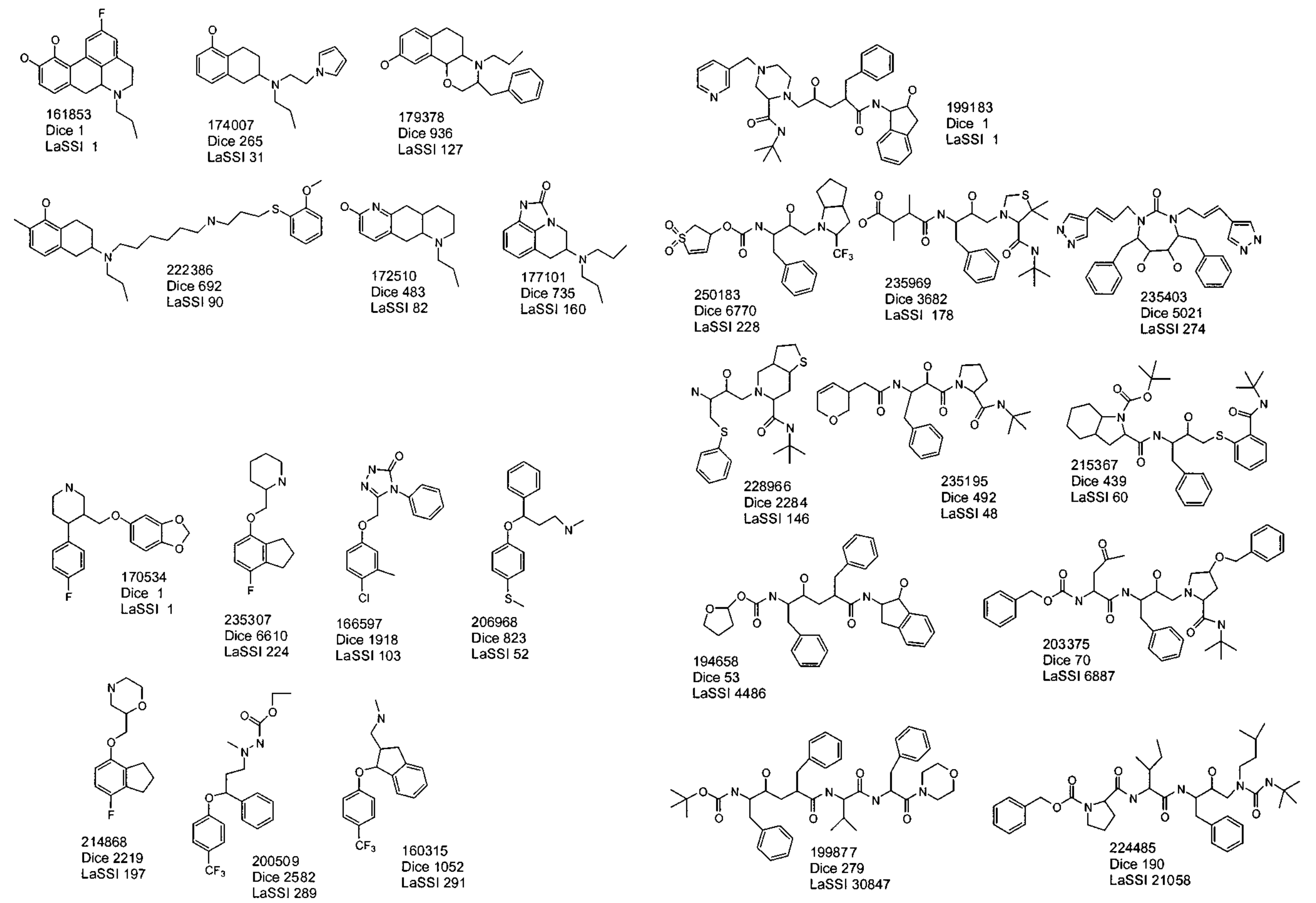

Figure 5. Selected active compounds that have extremely different ranks in Dice APTT vs LaSSI APTT. The examples are 161853 $k = 800$ (dopamine agonists), 170534 $k = 150$ (5-HT reuptake inhibitors), and 199183 $k = 160$ (HIV protease inhibitors), where k is the number of singular values. The ranks in two types of search are indicated.

choose $k = 400$, a number near 388, the mean optimum k over the examples. Table 3 compares the measures of merit for the optimized k vs $k = 400$. For about one-third of the probes there is a significant degradation of the initial enhancement at $k = 400$. These are not necessarily the ones where the optimum k differs the most from 400, however. The degradation at $k = 400$ is never so bad that LaSSI is rendered useless, just somewhat worse than Dice on the average.

Correlation of Ranks between Methods. When we compare the ranks of actives by LaSSI and Dice, we see that there is little to no correlation for any of the probes. An example is shown in Figure 4. The actives are scattered and do not fall near the diagonal. LaSSI is clearly selecting very different actives than Dice. We can select molecules with strikingly different ranks by calculating disparity = log(rank Dice/rank LaSSI). Figure 5 shows examples from three probes where abs(disparity) ≥ 0.5 (the ranks differ by a factor of more than ~3) and one of the ranks > 300 and the other ≤ 300.

Diversity of Actives. Figure 6 shows the MSP300 as a function of k for three probes. For any given probe, the MSP300 for LaSSI is somewhat lower than MSP300 for Dice, indicating an extra bit of “fuzziness” provided by LaSSI. We have found the MSP300 for LaSSI is fairly constant for most probes until k goes below ~20. In other words, for most values of k LaSSI finds different actives than Dice in the top 300, but the diversity of those compounds is not very much larger. On the other hand, for very low k, there is much more fuzziness relative to Dice.

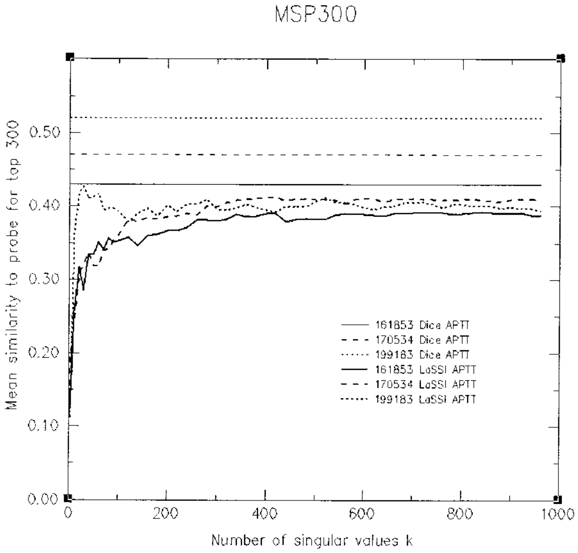

Figure 6. Mean similarity of the probe to each molecule in the top scoring 300 compounds (MSP300) for three examples. The MSP300 for Dice searches are shown as a horizontal lines. For comparison, the MSP300 for random sets of 300 compounds from MDDR would be 0.12−0.14.

Discussion

Similarity searches are the most useful early in a drug-discovery project when few actives are known and little is known about which features of these molecules confer activity. It has been our experience that it is always useful to use different methods of calculating similarity, since each has a potentially different view of chemistry. Rankings of compounds can be strongly affected by descriptor[5] and by definition of similarity.[8] LaSSI certainly selects different actives than does Dice and is thus a useful complement to TOPOSIM. LaSSI and Dice give very different rankings for two reasons: (1) In Dice, all descriptors are treated as distinct entities. In LaSSI, some descriptors may be partly synonymous with others. (2) In Dice, all descriptors are given equal weight. In LaSSI, some descriptors, mostly the very common or rare ones, are strongly down-weighted.

LaSSI adds some useful features, but also adds some complications relative to Dice. For Dice, the absolute similarity of a particular database entry with the probe is independent of other compounds in the database. In contrast, the singular vectors for LaSSI depend on the composition of the database as a whole, and as the database is updated and a new SVD calculated, the LaSSI similarity of a database entry with the probe may change. Also, the fact that LaSSI has the number of singular values k as an adjustable parameter adds flexibility but also requires the user to select an initial k. The goodness of the results can be sensitive to this parameter and the optimum k varies unpredictably from problem to problem. Fortunately, since LaSSI is so fast to run (less than 20 s on an IBM SP2 workstation), it is a trivial matter to run several searches at several different values of k. One useful bootstrap procedure starts with one or two known actives from the database to be searched. One finds the value of k at which the mean ranks of these molecules is a minimum and then tests the high-scoring molecules at that value of k. As more actives are found, one can further adjust k so that all known actives have a minimum mean rank, etc. This will be the subject of another paper in the series.[9]

Acknowledgment. The authors thank J. Chris Culberson, Kate Holloway, and Ralph Mosley for providing in-house examples for testing LaSSI. Phil Harris worked on theoretical comparisons of SVD with PCA. Mike Berry, at the University of Tennessee, Knoxville, provided the implementation of SVD that we currently use.

References

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

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

(3) 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.

(4) 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.

(5) 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.

(6) Hull, R. D.; Singh, S. B.; Nachbar, R. B.; Sheridan, R. P.; Kearsley, S. K.; Fluder, E. M. Latent Semantic Indexing (LaSSI) for Defining Chemical Similarity. *J. Med. Chem.* **2001**, *44*, 1177−1184.

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

(8) Cheng, C.; Maggiora, G.; Lajiness, M.; Johnson, M. Four associated coefficients for relating molecular similarity measures. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 909−915.

(9) Singh, S. B.; Sheridan, R. P.; Fluder, E. M.; Hull, R. D. Mining the chemical quarry with joint chemical probes: an application of LaSSI and TOPOSIM to chemical database mining. *J. Med. Chem.*, in press.

JM000392K