Calculating Similarities between Biological Activities in the MDL Drug Data Report Database

Robert P. Sheridan* and Joseph Shpungin

RY50S-100 Merck Research Laboratories, Rahway, New Jersey 07065

Received October 30, 2003

There are a number of licensed databases that assign biological activities to druglike compounds. The MDL Drug Data Report (MDDR), compiled from the patent literature, is a popular example. It contains several hundred distinct activities, some of which are therapeutic areas (e.g., Antihypertensive) and some of which are related to specific enzymes or receptors (e.g., ACE inhibitor). There are several data mining applications where it would be useful to calculate a similarity between any two activities. Two distinct activity labels can have a significant similarity for a number of reasons: two activities can be nearly synonymous (e.g., CCK B antagonist vs Gastrin antagonist), one activity may be a subset of another (e.g., Dopamine (D2) agonist vs Dopamine agonist), or an activity can be the mechanism by which another activity works (e.g., ACE inhibitor vs Antihypertensive), etc. In an ideal world, similarities for two activities could be calculated simply by comparing the compounds they have in common, but in hand-curated databases such as the MDDR the assignment of activities to compounds are inevitably inconsistent and incomplete. We propose a number of methods of calculating activity−activity similarities that hopefully compensate for errors in hand-curation. Two of these, TIMI and trend vector, show promise. Soft clustering of the activities using a union of similarity methods shows a reasonable association of therapeutic areas with their mechanisms.

INTRODUCTION

There are a number of licensed databases that assign biological activities to compounds. One of the most popular is the MDL Drug Data Report (MDDR),[1] which has been compiled by Prous Science (www.prous.com) from the patent literature since 1988, and distributed as an ISIS-readable database by Molecular Design Limited (www.mdli.com). Typically a molecule is assigned anywhere between 1 and 5 activity records. The activity record system is proprietary to Prous (Prous Science, personal communication). An activity record consists of an activity index (e.g., 31000) and a descriptive string (e.g., Antihypertensive). One clear advantage of the MDDR is that there is a limited set of activities to choose from, in contrast to some other databases that indicate activity by unstructured text. Activity records in the MDDR may indicate a general therapeutic area (e.g., Antihypertensive), indicate a specific mechanism of action (e.g., ACE inhibitor), or be merely descriptive of the chemical class (e.g., Carbapenem). The version of MDDR that we use here (2000.2) has ∼119 000 compounds and ∼700 distinct activity indices.

The MDDR is a very valuable resource, but certain limits should be kept in mind. A general issue with patent-centric databases is that each molecule has been tested for only a few activities chosen by the filers of the patents, so there are probably many molecules that lack an activity record that they might have had if only they were tested for that activity. Also, the criteria for a calling a molecule active may be different from source to source, especially in regards to whether the activity is in vivo or in vitro. Finally, one has limited confidence that any particular compound has a particular activity, since there is rarely independent confirmation by another laboratory for the majority of compounds.

Some MDDR activities are clearly related to others. For instance, some may be nearly synonymous (e.g., CCK B antagonist vs Gastrin antagonist), some are clearly subsets of others (e.g., Dopamine (D2) agonist vs Dopamine agonist), and some are related by mechanism (e.g., ACE inhibitor is a mechanism for Antihypertensive). In an ideal world it would be easy to find such relationships by comparing lists of compounds the activities have in common, but casual inspection of the MDDR shows that there are gaps in internal consistency and completeness. For instance, not all compounds that are "Dopamine (D2) Agonists" are also "Dopamine Agonists", not all "ACE inhibitors" are "Antihypertensive," etc. Such inconsistencies are almost inevitable with hand-curated databases compiled over a period of years. Some possible reasons follow.

1. New receptor subtypes are discovered, and receptor nomenclature changes with time.

2. Different filers of patents use different names for the same therapeutic area.

3. The accepted mechanism for a therapeutic effect may change over time, and more steps in the process may be discovered.

4. Some patents emphasize specific receptors without mentioning therapeutic uses, and vice versa. As recombinant technology becomes easier, we can expect more attention to be paid to receptor-based assays.

5. There is a tendency to mention only the most specific activity, e.g. "Dopamine (D2) agonists" without "Dopamine agonists".

* Corresponding author phone: (732) 594-3859; fax: (732) 594-4224; e-mail: sheridan@merck.com.

10.1021/ci034245h CCC: $27.50 © 2004 American Chemical Society
Published on Web 02/27/2004

6. Human compilers, even the most knowledgeable, have finite ability to remember past activity assignments of related compounds, to remember all synonyms for a given receptor, to know which receptors are associated with which therapeutic area, etc.

There are a number of applications for which it would be potentially useful to be able to define a similarity between two activities for the purposes of finding out which activities might be related. The MDDR is commonly used to benchmark clustering or virtual screening methods. In that application, one usually defines a set of compounds with a certain activity as "active" and assumes all other compounds are "inactive". For instance if we take "CCK B antagonists" as the activity, compounds that are labeled "Gastrin antagonists" but not "CCK B antagonists" will be falsely considered inactive. Another application involves deciding, given a list of activities, how many truly different activities there are. This would be useful for some of our earlier work on multiactivity substructures.[2] A third application is to be able to mine the MDDR for the mechanistic bases for certain therapeutic areas. For instance, we could find all the mechanism-specific activity labels associated with "antihypertensive" without having to look it up in Goodman & Gilman's *The Pharmacological Basis of Therapeutics*.[3] Fourth, it may be possible to find previously unnoticed structural similarities between compounds effective on unrelated activities. Finally, any attempt to "clean up" the MDDR for data mining purposes would have to start by finding similarities between the activities not obvious in the original data.

It is possible to group at least some of the MDDR activities by hand. For instance, Schuffenhauer et al.[4] created a hierarchical arrangement of the receptors and enzymes mentioned in the MDDR. However, this type of effort requires a great deal of knowledge and time on the part of the user. It would be very useful to have an automated method for calculating similarities between activities in the MDDR (including therapeutic areas) that does not require that level of knowledge, and to have the method work so as to compensate for inconsistency and incompleteness in the assignment of activities. We examine several such methods in this paper.

METHODS

Activity Labels. MDDR activities are in the form of a 5-digit activity code (e.g., 71000) and a brief description (e.g., Antiviral). The activity code and the description are almost always correlated (but not perfectly, as will be seen later), so the strings were concatenated to produce a single activity label, all punctuation and embedded blanks were changed to "_" and the string was forced to upper-case (71000_ANTIVIRAL).

Chemical Descriptors. Here we will be using the topological torsion (TT) substructure descriptor,[5] which has the form

$$\mathrm{AT_i} - \mathrm{AT_j} - \mathrm{AT_k} - \mathrm{AT_l}$$

where i, j, k, and l are consecutively bonded atoms. The atom type takes into account the element, the number of nonhydrogen neighbors, and the number of π electrons. A previous publication[6] gives examples of a molecule partitioned into these descriptors.

SC (Shared Compound) Similarity. We can think of each activity label i as a list of compounds with that activity, or as a vector $\mathbf{f_i}$ with the value $f_{ik} = 1$ if compound k is active on i and $f_{ik} = 0$ otherwise. The cosine similarity between activities i and j based on "shared compounds" (SC) is

$$\mathrm{SC}_{ij} = \sum_k f_{ik} f_{jk} / |\mathbf{f_i}||\mathbf{f_j}| \quad (1)$$

where k goes over all compounds in the database. If two activity labels have all their molecules in common, the similarity of the activities would be 1.0; if no molecules are in common, the similarity would be 0.0. There are other vector-based similarity definitions (Dice, Tanimoto, etc.), but cosine is the most consistent with the other methods in this paper and does not penalize two vectors for being of very different lengths (having different numbers of compounds for the two activities). Note that the cosine similarity assumes a symmetric similarity $\mathrm{SC}_{ij} = \mathrm{SC}_{ji}$, as do all the methods in this paper. When we list pairs of activities, they will be in alphabetical order of the strings.

TIMI Similarity. LSI (Latent Semantic Indexing),[7] a branch of LSA (Latent Semantic Analysis),[8] is a linguistic method for document searching that uncovers latent relationships between documents, between words, and between documents and words. TIMI[9] (Text Influenced Molecular Indexing) is a chemistry-oriented extension by which one can uncover latent relationships between documents, molecules, and chemical descriptors. TIMI has the same mathematical underpinnings as LSI. Given a large database of objects (documents, molecules, etc., depending on the application) matrix $\mathbf{X}$ is formed by elements d_{ji} which are the frequency of term j in object i. $\mathbf{X}$ is expressed as the product of three matrices by singular value decomposition such that

$$\mathbf{X} = \mathbf{P}\Sigma\mathbf{Q}^{\mathbf{T}} \quad (2)$$

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

$$\mathbf{X_k} = \mathbf{P_k}\Sigma_{\mathbf{k}}\mathbf{Q_k}^{\mathbf{T}} \quad (3)$$

The rows of $\mathbf{P_k}$ are the projected coordinates of the terms from the database in a k-dimensional space. These "latent terms" are orthogonal to each other but are linear combinations of the original terms. The rows of $\mathbf{Q_k}$ are the projected coordinates of the objects in that same space. If k is small relative to the number of unique terms in the database, there are two effects: some terms may become less important in calculating similarity, and the terms become less independent. Thus, by changing the value of k in the TIMI calculation, the user can adjust the amount of "fuzziness" in the method.

```
MOLECULE 090094
SOURCE  wyeth wyeth_ayerst
PATTITLE etodolac for inhibition of joint ankylosis
ACTIVITY 02000_ANTIARTHRITIC
ACTIVITY 02100_ANTIINFLAMMATORY
ACTIVITY 78454_CYCLOOXYGENASE_2_INHIBITOR
DESCRIPTORS C20C40C20C10 C20O20C40C20 C20O20C40C20 C21C21C31C20 C21C31C20C10
C21C31C31C20 C31C20C40C20 C31C21C21C21 C31C21C21C21 C31C31C20C10 C31C31C20C20
C31C31C20C20 C31C31C21C21 C31C31C21C21 C31C31C21C21 C31C31C31C20 C31C31C31C20
C31C31C31C21 C31C31C31C21 C31C31C31C21 C31C31C31C31 C31C31C31C31 C31C31C40C20
C31C31C40C20 C31C40C20C10 C31C40C20C31 C31C40O20C20 C31N20C31C31 C31N20C31C31
C31N20C31C31 C40C31C31C20 C40C31C31C31 C40C31N20C31 C40O20C20C20 N20C31C31C20
N20C31C31C20 N20C31C31C21 N20C31C31C21 N20C31C31C31 N20C31C31C31 N20C31C40C20
N20C31C40C20 O11C31C20C40 O11C31C20C40 O20C20C20C31 O20C40C20C10 O20C40C20C31
O20C40C31C31 O20C40C31N20
ACTION _ known analgesic and antiinflammatory agent now claimed to be of particular
utility in inhibiting joint ankylosis in the treatment of ankylosing spondylitis
proven to dose_dependently inhibit the pathological and immunological
characteristics of ...

MOLECULE 090109
SOURCE  aventis
PATTITLE
ACTIVITY 28000_CARDIOTONIC
ACTIVITY 37240_CAMP_PHOSPHODIESTERASE_INHIBITOR
ACTIVITY 78417_PHOSPHODIESTERASE_III_INHIBITOR
DESCRIPTORS C21C31C21C21 C21C31C21C21 C21C31C21C21 C21C31C21C21 C21C31S20C10
C21C31S20C10 C31C21C21C31 C31C21C21C31 C31C31C21C21 C31C31C21C21 C31C31C31C10
C31C31C31C21 C31C31C31C21 C31C31C31C31 C31N20C31C10 C31N20C31C31 C31N20C31C31
C31N20C31C31 N20C31C31C10 N20C31C31C31 N20C31C31C31 N20C31C31N20 N20C31N20C31
N20C31N20C31 O11C31C31C21 O11C31C31C21 O11C31C31C31 O11C31C31N20 O11C31N20C31
O11C31N20C31 S20C31C21C21 S20C31C21C21
ACTION _ cardiostimulant phosphodiesterase inhibitor indications _ congestive heart
failure presentation _ solution (injection) 100 mg/20 ml...
```

Figure 1. The first two molecules of the MDDR in a recast database for processing by TIMI. Each molecule includes one or more ACTIVITY labels, TT substructure DESCRIPTORS, and words from three MDDR fields: SOURCE, PATTITLE, and ACTION. For TIMI, the words, activity labels, and descriptors are collectively known as "terms". TT descriptors are a concatenation of the atom types of 4 bonded atoms. For instance, C20C40C20C10 is interpreted as C20−C40−C20−C10. C20 means a carbon with two non-hydrogen neighbors and zero π electrons, i.e., $-CH_2-$.

For our TIMI application, the columns will be entries in the MDDR, i.e., molecules. The rows will be terms consisting of the activity labels, TT descriptors for the parent structure of the molecule, plus words derived from three MDDR fields we feel contain additional clues about the activity: the SOURCE, the PATTITLE, and the ACTION fields. (Not all molecules have all fields filled.) The first two molecules are shown in Figure 1. Note that we are using only the TT descriptors whereas our previous efforts with TIMI[9] used both AP[10] (atom pairs) and TT descriptors. There were so many AP descriptors per molecule that they overwhelmed the influence of the relatively few words. This is much less of a problem when we use TT only.

For this exercise we used three flavors of the **X** matrix: TIMI-WC (words and chemistry) has all the terms described above. TIMI-W (words only) has the activity labels and words, with no TT descriptors. TIMI-C (chemistry only) has the activity labels and TT descriptors with no words. Usually there is some preprocessing of terms with TIMI. For example, words are "stemmed"; i.e., suffixes are removed. Also, very frequent and very uncommon words are usually deleted. However, for this application we found it best to keep all terms whatever their frequency. For instance, when we removed terms that occurred in more than 25% of the molecules (standard for TIMI), some interesting relationships were lost. There were 45 297, 35 631, and 10 366 unique terms in TIMI-WC, TIMI-W, and TIMI-C, respectively.

The TIMI similarity between terms i and j is

$$\mathrm{TIMI}_{ij} = \sum_x p_{ix}p_{jx}/|\mathbf{p_i}||\mathbf{p_j}| \qquad (4)$$

where x goes from 1 to k and $\mathbf{p_i}$ and $\mathbf{p_j}$ are rows corresponding to terms i and j in matrix $\mathbf{P_k}$. If the length of either row $\mathbf{p_i}$ or $\mathbf{p_j}$ is zero, the similarity is set to zero.

Here we are interested only in the terms that are activity labels. Note that TIMI similarities may be positive or negative.

Which value of k to use is not obvious a priori. We tried $k = 100$, 1000, 2000.

Trend Vector Similarity. As originally implemented, trend vector analysis[11] is a QSAR method that uses a sample-based PLS approach to summarize the biological "response" for a set of training molecules as a function of the presence or absence of chemical descriptors in those molecules. In the current application, we will construct a QSAR that relates the presence of each activity label with the presence or absence of terms. Generating a QSAR requires constructing a training set for each activity label. We randomly chose up to 1000 molecules with the activity label. These were given a response of "1". We then added 1000 randomly selected compounds without the label. These were given a response of "0". For example, for 11124_DOPAMINE__D1_AGONIST there were 118 compounds with the activity label and 1000 without.

11124_DOPAMINE__D1__AGONIST		11125_DOPAMINE__D2__AGONIST	
dopamine	0.136	dopamine	0.151
O10C31C31O10	0.115	N30C20C20C10	0.119
abbott	0.089	d2	0.091
agonists	0.087	C21C20N20C20	0.083
O10C31C31C21	0.078	N30C20C20C20	0.080
o		o	
o		o	
o		o	
C30C20C20C20	-0.024	C21C31C31C21	-0.370
C31C21C31C21	-0.026	agents	-0.043
agents	-0.027	C31N30C30C20	-0.043
C31C30C20C20	-0.027	N20C31C31C21	-0.046
O11C31C31C31	-0.028	C31C21C31C20	-0.047

TV-WC Similarity=0.427

Figure 2. Examples of trend vectors for two sample activities. For each, the terms are listed in decreasing correlation with the activity. One can clearly see that these two activities share some of the most positive and most negative terms and therefore their similarity (the normalized dot product of the vectors) would be significantly positive.

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 5 PLS components. If the significance of the first component is <3, the trend vector is unreliable. A final trend vector has a coefficient associated with each term in the training set. For instance in Figure 2, the terms "dopamine" and "O10C31C31O10" (part of a catechol) are most correlated with the presence of the activity label 11124_DOPAMINE__D1__AGONIST, and "O11C31C31C31" is the most anticorrelated.

As with TIMI, we constructed three sets of trend vectors. TV-WC includes words and TTs as terms, TV-W includes only words, and TV-C includes only TTs.

The similarity of two trend vectors TV_{ij} is the cosine of the angle θ between them, equivalent to taking the dot product of the normalized vectors. This is analogous to eq 1 except f_{ik} is the floating point value of term k in trend vector i and the sum is over the union of terms in the two vectors. Two trend vectors pointing in the same direction have a similarity of 1, and two trend vectors pointing in the opposite direction have a similarity of -1. The similarity is meaningful only if both trend vectors are statistically significant.

Label Similarity. Some activity labels are obviously related, regardless of the molecules associated with them, simply because their strings are similar, for example 11124_DOPAMINE__D1__AGONIST and 11125_DOPAMINE__D2__AGONIST. One way of describing the strings such that a similarity can be easily calculated would be to break the labels into overlapping letter-triplet descriptors. For instance, the string 11124_DOPAMINE__D1__AGONIST would be broken up into 111, 112, 124, 24_, 4_D, etc. We soon realized that the presence of common words such as "INHIBITOR", "AGONIST", etc. made unrelated activities (e.g., 077001_DOPAMINE__D1__AGONIST vs 07707_ADENOSINE__A1__AGONIST) look very much alike, and also that substrings such as "INHIBITOR" and "BLOCKER" would not be perceived as equivalent even though they are practically synonymous. For that reason, we preprocessed the activity labels by replacing certain words with shorter, more consistent strings, before parsing the labels into letter triplets:

_INHIBITOR → _INH
_BLOCKER → _INH
_ANTAGONIST → _INH
_TREATMENT(FOR) → _INH
_STIMULANT → _AGO
_AGONIST → _AGO
_PROMOTOR → _AGO
_AGENT(_FOR) → _AGF
_DISORDER → _DIS

The similarity between labels would be the cosine similarity, analogous to eq 1 except the sum would be over letter-triplet descriptors instead of molecule identifiers.

Clustering. Given a similarity matrix, a number of methods can be used to cluster the activity labels. For instance, one can produce hierarchical clustering dendrograms. Hierarchical clustering is good at detecting "low similarity" relationships of objects, but the dendrograms are sometimes hard to interpret if the number of objects is large. In contrast, nonhierarchical clustering methods group objects into smaller, more manageable sets. These require an arbitrary cutoff similarity above which the objects can be considered "neighbors." "Crisp" clustering algorithms such as that described by Butina[12] assign objects to one and only one cluster, but that is less desirable here. Therefore, we implemented a version of the soft clustering algorithm of Ibrahimov[13] which is a variant on the Butina algorithm. In these algorithms, objects are first sorted by decreasing number of neighbors they have. The object with the largest number of neighbors is taken as the center of the first cluster. Objects too similar to this center are eliminated as candidate centers. The next object not already eliminated becomes the second cluster center; objects too similar to it are eliminated, etc. After all objects have been assigned as cluster centers or eliminated, each eliminated object is assigned to a cluster center. In the Ibramov algorithm, an object can be assigned to more than one cluster center. Here the objects to be clustered are activity labels. Having compared crisp clusters with soft clusters, we feel the soft clusters are a great improvement in grouping together associated labels, at the expense of some redundancy where there may be clusters sharing many of the activity labels. This is expected since

in pharmacology one mechanism-specific activity may be associated with more than one therapeutic area and vice versa.

How Do We Know When Activities are Related? Ideally one would like to have some objective criterion by which to say whether a similarity method is valid. Some amount of objectivity is achievable in some cases. For instance, in the field of chemical similarity one can compare descriptors and clustering methods by how well they group compounds with the same biological activities.[14−16] Here, however, the biological activities themselves are being compared, and we know of no master list of truly related biological activities that does not suffer from the same incompleteness and inconsistency issues that the MDDR itself suffers from. Thus, we have to rely on our own judgment. Very few people (certainly not the authors) have a priori knowledge of all the therapeutic areas in the MDDR and their mechanisms, and we have had to educate ourselves in the course of this project. We have used the following rules of thumb to confirm that activities are related using information already in the MDDR.

1. If the activity labels mention the same receptor, we can assume they are related.

2. If two activities have many compounds in common (SC similarity is large) and the compounds are from many different patent sources, the activities are probably related.

3. If the structures associated with the two activities are similar, the activities may be related. One can also look at the ACTION field for information about the compounds

Other sources outside the MDDR are also helpful.

1. Many therapeutic areas and at least some of their mechanisms are discussed in the Goodman & Gilman text.

2. Web search engines can be used to find literature citations on drugs that act on specific targets and the relationship of the targets to therapeutic areas. We found that nearly all therapeutic areas we looked at have a support group on the Web that keeps track of the associated medical literature.

RESULTS

There are roughly ∼200 000 activity−activity pairs. There may not be exactly the same number of pairs for each method because the methods require different information, which may be missing in some cases. For instance, it is not possible to calculate TIMI or TV similarities if there are no connection tables associated with one of the activities, but it still possible to calculate Label similarity. The distribution of similarities for all the methods is shown in Figure 3. The similarities center around zero, as expected. All SC similarities are nonnegative. It is obvious from this figure that the TIMI-WC $k = 100$ distribution is much too broad, making it impossible to distinguish "meaningful" and "noise" similarities, and we therefore prefer $k = 1000$. TIMI $k = 2000$ (not shown) gives a slightly narrower distribution. The results for $k = 1000$ and $k = 2000$ are otherwise quite similar, and we will henceforth discuss only TIMI $k = 1000$. For the TV methods, we show the distribution of all activity−activity pairs regardless of the significance of the individual vectors in the pair. The spread of values for those pairs where both TVs are significant is nearly identical. 96% of the pairs for TV-WC have both TVs significant, 86% for TV-W, and 91% for TV−C. The distribution of Label similarity shows an extra peak around 0.1, indicating that we did not completely succeed removing spurious resemblances between strings.

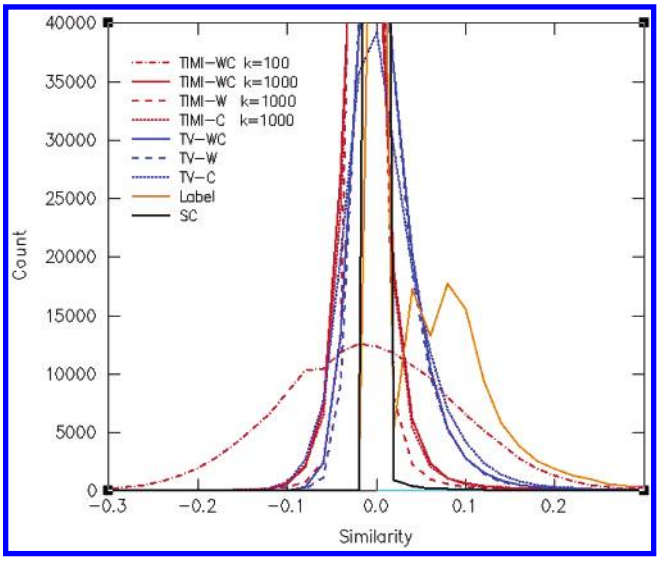

Figure 3. A histogram showing the distribution of activity−activity similarities as calculated by several methods. The similarity is divided into bins of 0.01 unit.

From the distributions in Figure 3 we can get an idea of what activity−activity similarity is in the "noise" for any given method. Another way of finding the noise level is to inspect pairs of activities at a variety of similarity levels and determine at what similarity cutoff most of the pairs are truly related. From such inspections we judge that similarities ≥ 0.25 are probably meaningful for the TV methods. The cutoff 0.25 is well above the bulk of the pairwise similarities for all the TV distributions. Similar rule-of-thumb cutoffs are 0.20 for TIMI $k = 1000$, 0.10 for SC, and 0.30 for Label.

SC Similarity. Despite the inconsistencies and incompleteness, we expect high SC similarity to indicate real activity−activity relationships, and we use it as a baseline of comparison for the other methods, in the sense that we hope the other methods will find all the relationships in SC, plus others. The 20 most similar activity−activity pairs are listed in Table 1. The majority indicate a relationship between a therapeutic area and a mechanism, although there is at least one synonym: 42713_CCK_B_ANTAGONIST versus 42714_GASTRIN_ANTAGONIST.

TIMI-WC versus SC. A graph of TIMI-WC similarity versus SC similarity is shown in Figure 4. Each point in the graph represents a pair of activities. Generally the two types of similarity correlate, with two types of outliers:

(1) pairs for which the SC similarity is near zero, but the TIMI-WC similarity is high;

(2) pairs that fall far below the majority of the pairs.

The most discrepant points are listed in Table 2 in order of the difference between the two types of similarity. (A more sophisticated method of normalizing the similarities by their spread before subtracting does not much change the order and is harder to interpret.) Clearly, TIMI-WC is able to find many relationships that SC has missed because the activities share few or no compounds in common. There is one example where the spelling of the activity description is

Table 1. Highest Activity−Activity Similarities by SC (Shared Compounds)

	SC
43200_ANTIDIABETIC__SYMPTOMATIC 43210_ALDOSE_REDUCTASE_INHIBITOR	0.862
27510_PULMONARY_EMPHYSEMA__AGENT_FOR 78323_ELASTASE_INHIBITOR	0.839
37100_ANTICOAGULANT 3110_THROMBIN_INHIBITOR	0.738
12150_BOTULINUM_TOXIN 18345_EYE_MUSCLE_DISORDERS__AGENT	0.707
35560_PROSTATE_DISORDERS__AGENT_FOR 78335_STEROID__5ALPHA__REDUCTASE_INHIBITOR	0.704
57310_LIVER_FIBROSIS__AGENT_FOR 78337_PROTOCOLLAGEN_PROLYL_ HYDROXYLASE_INHIBITOR	0.667
42713_CCK_B_ANTAGONIST 42714_GASTRIN_ANTAGONIST	0.645
06233_5_HT3_ANTAGONIST 12350_ANTIEMETIC	0.638
36320_G_CSF 36321_GM_CSF	0.630
31341_VASOPRESSIN_V1_ANTAGONIST 31342_VASOPRESSIN_V2_ANTAGONIST	0.619
54110_ANTISECRETORY__GASTRIC 54112_H__K__ATPASE_INHIBITOR	0.617
68000_ANTIBACTERIAL 68210_QUINOLONE	0.575
44200_UTERINE_RELAXANT 44210_OXYTOCIN_ANTAGONIST	0.574
16110_ADRENERGIC__ALPHA2__AGONIST 22000_NASAL_DECONGESTANT	0.574
06246_5_HT1D_AGONIST 12342_ANTIMIGRAINE	0.569
03000_TREATMENT_OF_GOUT 03200_XANTHINE_OXIDASE_INHIBITOR	0.567
27210_LEUKOTRIENE_ANTAGONIST 27212_LEUKOTRIENE_D4_ANTAGONIST	0.550
37200_PLATELET_ANTIAGGREGATORY 37260_GPIIB_IIIA_RECEPTOR_ANTAGONIST	0.544
37100_ANTICOAGULANT 37121_FACTOR_XA_INHIBITOR	0.530
12452_NEURONAL_INJURY_INHIBITOR 12455_NMDA_RECEPTOR_ANTAGONIST	0.528

different, 78403_GLYCINAMIDE_RIBONUCLEOTIDE-_FORMYLTRANSFERASE_INHIBIT versus 78403_GLYCINAMIDE_RIBONUCLEOTIDE_FORMYLTRANSFERASE_INHIBITOR, and an example where the activity description is the same, but the activity index is different 08420_MAO_B_INHIBITOR versus 11130_MAO_B_INHIBITOR. There are examples where TIMI-WC finds relationships between agonists and antagonists on the same receptor: 06280_HISTAMINE_H3_AGONIST versus 09226_HISTAMINE_H3_ANTAGONIST, 06211_BENZODIAZEPINE_AGONIST versus 06212_BENZODIAZEPINE-_AGONIST_ANTAGONIST. It is interesting that TIMI-WC recognized that proteoglycanase inhibitors are a subset of matrix metalloproteinase inhibitors and that calpain inhibitors and cathepsin L inhibitors are examples of thiol protease inhibitors. The relationship between 49200_MINERAL and 75347_VANADIUM_COMPLEX is not obvious until one inspects the molecules; they share structurally very similar metal chelators.

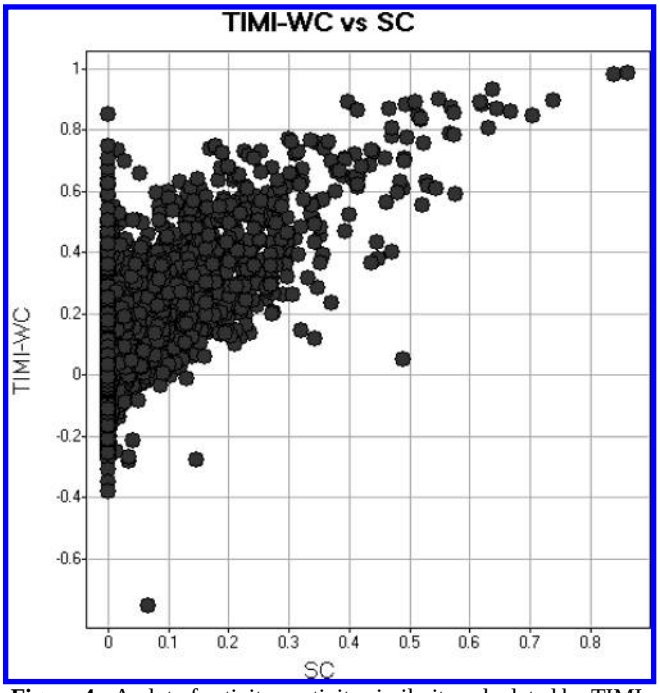

Figure 4. A plot of activity−activity similarity calculated by TIMI-WC vs SC (see text for definitions). Each point represents a pair of activities.

The examples at the bottom of Table 2 are "false negatives" in the sense that the activities are obviously related (one even has a large SC), but TIMI-WC finds a near-zero or negative similarity. Particularly egregious is 31430_ANGIOTENSIN_II_BLOCKER versus 31432_ANGIOTENSIN_II_AT1_ANTAGONIST, which should be an example of a subset relationship. However, it should be noted that there are only a handful of false negatives out of ∼200 000 pairs.

Chemistry versus Words and Chemistry: TIMI-C versus TIMI-WC. The plot of TIMI-WC versus TIMI-C is in Figure 5. The two are clearly correlated without many extreme outliers. This is not particularly surprising in retrospect because the number of words is less than the number of descriptors in most molecules. The most discrepant are in Table 3. Most of the pairs in Table 3 are clearly related. In some cases TIMI-WC gives higher similarities, and in some cases TIMI-C does. There does seem to be a slightly larger number of related pairs with higher TIMI-WC similarity. Thus, we feel there is a slight advantage to having both words and TT descriptors versus descriptors alone.

TV-WC versus SC. This plot is in Figure 6. Only the pairs where both trend vectors in the pair are significant are shown. As with TIMI-WC, TV-WC is generally correlated with SC. Again there are many outliers falling above the majority of the points, but this time there are none falling much below. The most discrepant points are shown in Table 4. The top of Table 4 shares some pairs in common with Table 2, but mostly the pairs are different. There is an additional synonym 37252_PROSTAGLANDIN versus 54121_PROSTAGLANDIN, and different subset relationships, e.g., 02520_BRADYKININ_ANTAGONIST versus 02522_BRADYKININ_BK2_ANTAGONIST. It is significant that the pair 31430_ANGIOTENSIN_II_BLOCKER versus 31432_ANGIOTENSIN_II_AT1_BLOCKER, which is given a very negative similarity in TIMI, is seen as similar by TV. The pairs at the bottom of Table 4 do not have very different similarities between TIMI-WC and SC.

Table 2. Most Discrepant Similarities between TIMI-WC and SC

	SC	TIMI-WC
49200_MINERAL 75347_VANADIUM_COMPLEX	0.000	0.851
78403_GLYCINAMIDE_RIBONUCLEOTIDE_ FORMYLTRANSFERASE_INHIBIT 78403_GLYCINAMIDE_RIBONUCLEOTIDE_ FORMYLTRANSFERASE_INHIBITOR	0.000	0.748
06280_HISTAMINE_H3_AGONIST 09226_HISTAMINE_H3_ANTAGONIST	0.017	0.735
78398_PROTEOGLYCANASE_INHIBITOR 78432_MATRIX_METALLOPROTEINASE_ INHIBITOR	0.000	0.708
50160_CATHEPSIN_L_INHIBITOR 78375_THIOL_PROTEASE_INHIBITOR	0.000	0.674
42110_GROWTH_HORMONE_SECRETION_ INHIBITOR 42111_SOMATOSTATIN_ANALOG	0.027	0.699
34410_ANTIDIURETIC_HORMONE_ ANTAGONIST 44200_UTERINE_RELAXANT	0.000	0.637
78375_THIOL_PROTEASE_INHIBITOR 78439_CALPAIN_INHIBITOR	0.000	0.625
06211_BENZODIAZEPINE_AGONIST 06212_BENZODIAZEPINE_AGONIST_ ANTAGONIST	0.000	0.617
02520_BRADYKININ_ANTAGONIST 02522_BRADYKININ_BK2_ANTAGONIST	0.053	0.656
57100_ANTICHOLELITHOGENIC 57200_CHOLERETIC	0.000	0.590
08420_MAO_B_INHIBITOR 11130_MAO_B_INHIBITOR	0.000	0.582
o o o		
64400_PENEM 64500_CARBAPENEM	0.0000	−0.38
75504_ANTIMITOTIC 75506_TAXANE_DERIVATIVE	0.147	−0.277
75505_MICROTUBULE_INHIBITOR 75506_TAXANE_DERIVATIVE	0.489	0.052
31430_ANGIOTENSIN_II_BLOCKER 31432_ANGIOTENSIN_II_AT1_ANTAGONIST	0.065	−0.754

TV-C versus TV-WC. In Figure 7, we see again the similarities with descriptors only and the similarities with words and descriptors are highly correlated. Most of the outliers are below the majority of points, where pairs appear more similar to TIMI-C than to TIMI-WC. Discrepant pairs are in Table 5. The top of Table 5 has a number of pairs that do not seem related, e.g., 42770_VASOACTIVE_INTESTINAL_PEPTIDE__VIP_ versus 75800_CANCER_IMMUNOTHERAPY and 42734_NEUROKININ_AGONIST versus 75800_CANCER_IMMUNOTHERAPY. On the other hand, those pairs at the bottom of Table 5 are clearly related. Inspection of the structures associated with the activities at the top of Table 5 shows a reason for these false positives. The majority of structures associated with all the activities are large peptides, which naturally appear very similar if one considers TT descriptors alone. On the other hand, the words are not similar, and thus, TV-WC correctly gives a smaller similarity than TV-C.

Label Similarity. It is reasonable to expect the labels to be correlated with SC, but Figure 8 shows that there is no detectable correlation. Label similarity does not correlate with TIMI-WC or TV-C either (not shown). If one looks only at the activity index, the integer part of the label, one gets similarly poor correlations between the integer differences and SC (not shown). Generally, then, the label is unexpectedly uninformative in terms of relationships between activities. However, there are some interesting cases revealed by Label similarity. Many more "spelling synonyms" are revealed. There is one pair where the activities have the same activity index and similar spelling of the description but no compounds in common, e.g., 78440_PROLYL_PEPTIDYL_ISOMERASE_INHIBITORS versus 78440_PROLYL_PEPTIDYL_ISOMERASE_INHIBITOR. There are

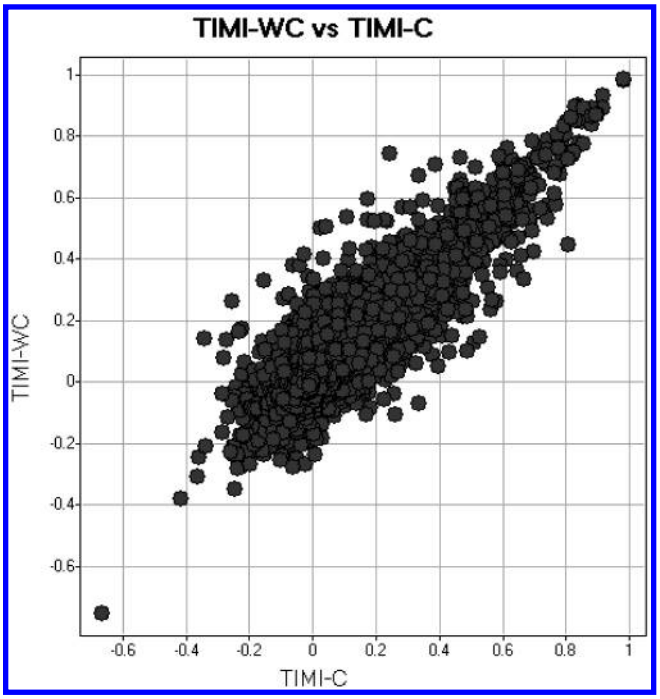

Figure 5. A plot of activity−activity similarity calculated by TIMI-WC (words and chemical descriptors) vs TIMI-C (chemical descriptors alone).

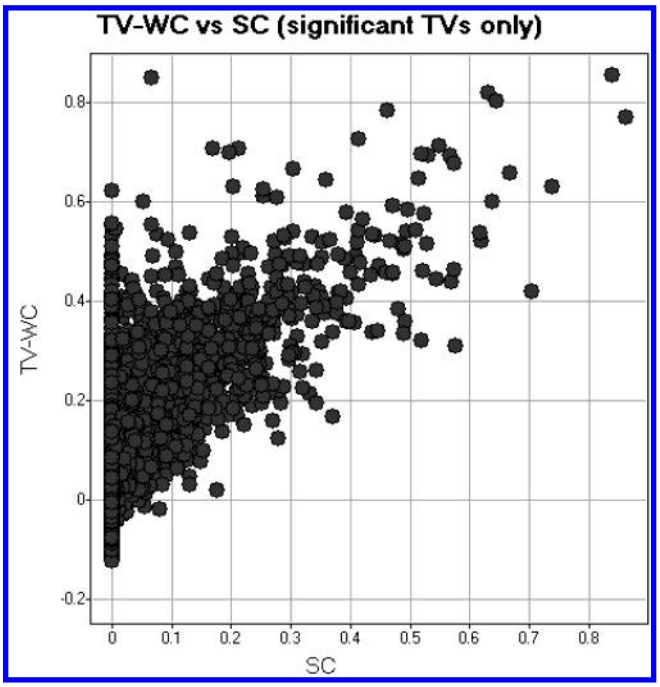

Figure 6. A plot of activity−activity similarity calculated by TV-WC vs SC. We show only the similarities in which the two trend vectors are statistically significant.

Table 3. The Most Discrepant Similarities between TIMI-WC and TIMI-C

	TIMI−WC	TIMI−C
31651_ENDOTHELIN_AGONIST 31656_ENDOTHELIN_ETB_ANTAGONIST	−0.073	0.337
31650_ENDOTHELIN_ANTAGONIST 35500_RENAL_FAILURE__AGENT_FOR	0.101	0.489
31650_ENDOTHELIN_ANTAGONIST 31651_ENDOTHELIN_AGONIST	0.125	0.508
31650_ENDOTHELIN_ANTAGONIST 31655_ENDOTHELIN_ETA_ANTAGONIST	0.145	0.526
31651_ENDOTHELIN_AGONIST 31655_ENDOTHELIN_ETA_ANTAGONIST	−0.109	0.259
06100_SEDATIVE_HYPNOTIC 42650_MELATONIN_AGONIST	0.444	0.806
75505_MICROTUBULE_INHIBITOR 75506_TAXANE_DERIVATIVE	0.052	0.398
06211_BENZODIAZEPINE_AGONIST 42650_MELATONIN_AGONIST	0.097	0.432
31650_ENDOTHELIN_ANTAGONIST 31656_ENDOTHELIN_ETB_ANTAGONIST	0.335	0.669
02448_CYTOKINE_MODULATOR 02455_IL_6_INHIBITOR	0.235	0.561
o o o		
12100_ANTISPASTIC 12200_SKELETAL_MUSCLE_RELAXANT	0.396	0.113
06100_SEDATIVE_HYPNOTIC 06213_BENZODIAZEPINE_ANTAGONIST	0.303	0.018
06100_SEDATIVE_HYPNOTIC 10000_ANTICONVULSANT	0.521	0.236
07707_ADENOSINE__A1__AGONIST 08450_ADENOSINE__A1__ANTAGONIST	0.290	−0.003
08450_ADENOSINE__A1__ANTAGONIST 78353_ADENOSINE_DEAMINASE_INHIBITOR	0.291	−0.008
09229_NICOTINIC_AGONIST 09350_ANTISMOKING	0.525	0.226
52504_LANOSTEROL_14ALPHA_METHYL_ DEMETHYLASE_INHIBITOR 52505_2_3_OXIDOSQUALENE_LANOSTEROL_ CYCLASE_INHIBITOR	0.433	0.119
78398_PROTEOGLYCANASE_INHIBITOR 78432_MATRIX_METALLOPROTEINASE_ INHIBITOR	0.708	0.387
42731_SUBSTANCE_P_ANTAGONIST 42732_NEUROKININ_ANTAGONIST	0.331	0.005

Table 4. The Most Discrepant Similarities between TV-WC and SC

	SC	TV-WC
07707_ADENOSINE__A1__AGONIST 07708_ADENOSINE__A2__AGONIST	0.0658	0.849
78403_GLYCINAMIDE_RIBONUCLEOTIDE_ FORMYLTRANSFERASE_INHIBIT 78403_GLYCINAMIDE_RIBONUCLEOTIDE_ FORMYLTRANSFERASE_INHIBITOR	0.000	0.620
57100_ANTICHOLELITHOGENIC 57200_CHOLERETIC	0.000	0.556
07708_ADENOSINE__A2__AGONIST 78325_S_ADENOSYL_L_HOMOCYSTEINE_ HYDROLASE_INHIBITOR	0.000	0.551
02520_BRADYKININ_ANTAGONIST 02522_BRADYKININ_BK2_ANTAGONIST	0.053	0.598
31340_VASOPRESSIN_ANTAGONIST 31342_VASOPRESSIN_V2_ANTAGONIST	0.007	0.544
60100_VACCINE 60110_BACTERIAL_VACCINE	0.170	0.706
07707_ADENOSINE__A1__AGONIST 78325_S_ADENOSYL_L_HOMOCYSTEINE_ HYDROLASE_INHIBITOR	0.000	0.527
64203_ISOCEPHEM 64220_CARBACEPHEM	0.000	0.509
31650_ENDOTHELIN_ANTAGONIST 31656_ENDOTHELIN_ETB_ANTAGONIST	0.195	0.697
31650_ENDOTHELIN_ANTAGONIST 31655_ENDOTHELIN_ETA_ANTAGONIST	0.212	0.706
37252_PROSTAGLANDIN 54121_PROSTAGLANDIN	0.000	0.493
31430_ANGIOTENSIN_II_BLOCKER 31432_ANGIOTENSIN_II_AT1_ANTAGONIST	0.065	0.554
o o o		
16000_ANTIGLAUCOMA 16200_CARBONIC_ANHYDRASE_INHIBITOR	0.492	0.359
68000_ANTIBACTERIAL 68241_OXAZOLIDINONE	0.342	0.196
75505_MICROTUBULE_INHIBITOR 75506_TAXANE_DERIVATIVE	0.489	0.335
75400_ANTINEOPLASTIC_ANTIBIOTIC 75410_ANTHRACYCLINE	0.280	0.124
43136_ADRENOCEPTOR__BETA3__AGONIST 53000_ANTIOBESITY	0.518	0.321
50000_CALCIUM_REGULATOR 50100_CALCITONIN_ANALOG	0.369	0.166
68000_ANTIBACTERIAL 68210_QUINOLONE	0.575	0.309
35560_PROSTATE_DISORDERS__AGENT_FOR 78335_STEROID__5ALPHA__REDUCTASE_ INHIBITOR	0.704	0.418

more cases where the activity index is different but the description is identical, e.g., 16100_ADRENERGIC__BETA__BLOCKER versus 31250_ADRENERGIC__BETA__BLOCKER, 28500_VASODILATOR versus 31300_VASODILATOR, 02500_IMMUNOMODULATOR versus 62000_IMMUNOMODULATOR. We found 5 pairs where the Label similarity > 0.8, but where there is no significant similarity by SC, TIMI, or TV. Generally, this is because the activities have very different types of compounds associated with them. Five more such pairs could be "rescued" by TIMI or TV.

As might be expected, there are a large number of labels that are completely dissimilar but have very high SC similarity (many compounds in common); usually they are mechanistically related, e.g., 35560_PROSTATE_DISORDERS__AGENT_FOR versus 78335_STEROID__5ALPHA__REDUCTASE_INHIBITOR (SC = 0.704), 27510_PULMONARY_EMPHYSEMA__AGENT_FOR versus 78323_ELASTASE_INHIBITOR (SC = 0.839).

TIMI-WC versus TV-WC. We have two methods TIMI and TV that appear to work well, and in both cases there is at least a slight preference for both words and descriptors. Figure 9 shows a plot of TIMI-WC versus TV-WC. There is some correlation, but the large amount of scatter indicates that the two methods are picking up different types of similarity. The most discrepant pairs are in Table 6. TV correctly assigns some of the false negatives from TIMI, but TV misses some of the related areas found by TIMI. Also, TIMI finds relationships that TV cannot find because one or both of the relevant trend vectors is not significant. One example is the pair with highest TIMI-WC similarity (0.851): 49200_MINERAL versus 75347_VANADIUM_COMPLEX. Other interesting pairs of this type are

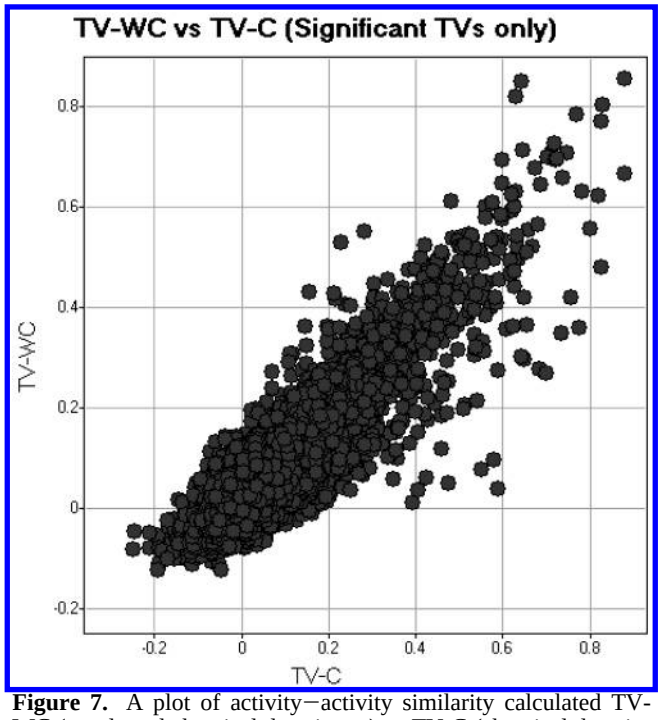

Figure 7. A plot of activity−activity similarity calculated TV-WC (words and chemical descriptors) vs TV-C (chemical descriptors) alone.

09370_TREATMENT_OF_OPIOID_DEPENDENCY versus 111222_DOPAMINE_RELEASING_DRUG (TIMI-WC = 0.426) and 68000_ANTIBACTERIAL versus 78388_ALANINE_RACEMASE_INHIBITOR (TIMI-WC = 0.343). In all cases, the SC similarity is near zero.

Clustering. Individually SC, TIMI-WC, and TV-WC are recognizing similarities that the others miss. Clearly we would like to use some union of the methods so as not to miss any relationships. We will take the following simple approach: Since none of those methods produces false positives, for the purposes of clustering, it seems reasonable to take max(SC, TIMI-WC, TV-WC) as the similarity for each pair and use the Ibrahimov algorithm on that similarity. We will use 0.25 as the clustering cutoff, which is appropriate for the TV-WC, the method for which the histogram in Figure 3 is furthest to the right and needs the highest cutoff. Strictly speaking, this simple approach is viable only if the cutoffs for all three methods are the same, but for our purposes, they are probably close enough. We did try more sophisticated approaches[17] where the activities were clustered within each method using its own cutoff, and then the cluster memberships were combined such that two activities were considered neighbors if both were in the same cluster by any of the three methods. The final clusters were not very different from those produced by the simple approach, except for being slightly larger. (This is expected because the more sophisticated approaches include some borderline activity pairs where all three similarities < 0.25, but the SC similarity> 0.1, or the TIMI similarity > 0.20.) On the other hand, the more sophisticated approaches remove the notion of the similarity of an activity to the cluster center, which is useful in deciding how much credence to put in its membership. Thus, we prefer the simpler approach because of its interpretability.

The simple soft clustering method produces 246 clusters. In Table 7, the largest clusters are listed by the cluster center and the number of activities in the cluster. The majority of the cluster centers are therapeutic areas. That is consistent with our hope that therapeutic areas would be "surrounded" with a number of mechanism-related activities. As expected, a number of activity labels occur in a large number of clusters: 12452_NEURONAL_INJURY_INHIBITOR occurs in 12 clusters, 06200_ANXIOLYTIC in 12 clusters, 07000_ANTIPSYCHOTIC in 11 clusters, 33456_ANTIISCHEMIC in 11 clusters, etc.

The memberships of representative large clusters are in Table 8. In most cases, each center is grouped with related therapeutic areas and perhaps several mechanism-based activities. The members of the cluster are listed in order of decreasing Max similarity to the cluster center. The maximum value can come from SC, TIMI-WC, or TV-WC. Most of the relationships seem reasonable in terms of synonyms or mechanisms. However, there are a few false associations. The cutoff 0.25 strikes a reasonable balance between catching most of the interesting relationships and not having too many false associations.

We will discuss one cluster in detail: 31000_ANTIHYPERTENSIVE. Antihypertensives are especially interesting

Table 5. The Most Discrepant Similarities between TV-WC and TV-C

	TV-WC	TV-C
42720_VASOACTIVE_INTESTINAL_PEPTIDE__VIP_ 75800_CANCER_IMMUNOTHERAPY	0.039	0.588
42734_NEUROKININ_AGONIST 75800_CANCER_IMMUNOTHERAPY	0.096	0.578
42720_VASOACTIVE_INTESTINAL_PEPTIDE__VIP_ 60120_VIRAL_VACCINE	0.077	0.552
60120_VIRAL_VACCINE 75815_CANCER_VACCINE	0.269	0.701
42720_VASOACTIVE_INTESTINAL_PEPTIDE__VIP_ 42734_NEUROKININ_AGONIST	0.049	0.475
60120_VIRAL_VACCINE 80506_DIAGNOSTIC_FOR_AIDS	0.360	0.777
60000_IMMUNIZING_AGENT__ACTIVE_ 60120_VIRAL_VACCINE	0.277	0.683
60120_VIRAL_VACCINE 60121_AIDS_VACCINE	0.349	0.733
42720_VASOACTIVE_INTESTINAL_PEPTIDE__VIP_ 75815_CANCER_VACCINE	0.012	0.392
42720_VASOACTIVE_INTESTINAL_PEPTIDE__VIP_ 80506_DIAGNOSTIC_FOR_AIDS	0.036	0.405
o		
o		
o		
37268_P2T_PURINORECEPTOR_ANTAGONIST 78325_S_ADENOSYL_L_HOMOCYSTEINE_ HYDROLASE_INHIBITOR	0.307	0.111
07708_ADENOSINE__A2__AGONIST 78353_ADENOSINE_DEAMINASE_INHIBITOR	0.415	0.215
34610_ATRIAL_NATRIURETIC_POLYPEPTIDE 43111_INSULIN_DERIVATIVE	0.271	0.071
07707_ADENOSINE__A1__AGONIST 07708_ADENOSINE__A2__AGONIST	0.849	0.644
11124_DOPAMINE__D1__AGONIST 11125_DOPAMINE__D2__AGONIST	0.427	0.214
07708_ADENOSINE__A2__AGONIST 78325_S_ADENOSYL_L_HOMOCYSTEINE_ HYDROLASE_INHIBITOR	0.551	0.282
07707_ADENOSINE__A1__AGONIST 78353_ADENOSINE_DEAMINASE_INHIBITOR	0.430	0.155
07707_ADENOSINE__A1__AGONIST 78325_S_ADENOSYL_L_HOMOCYSTEINE_ HYDROLASE_INHIBITOR	0.527	0.228

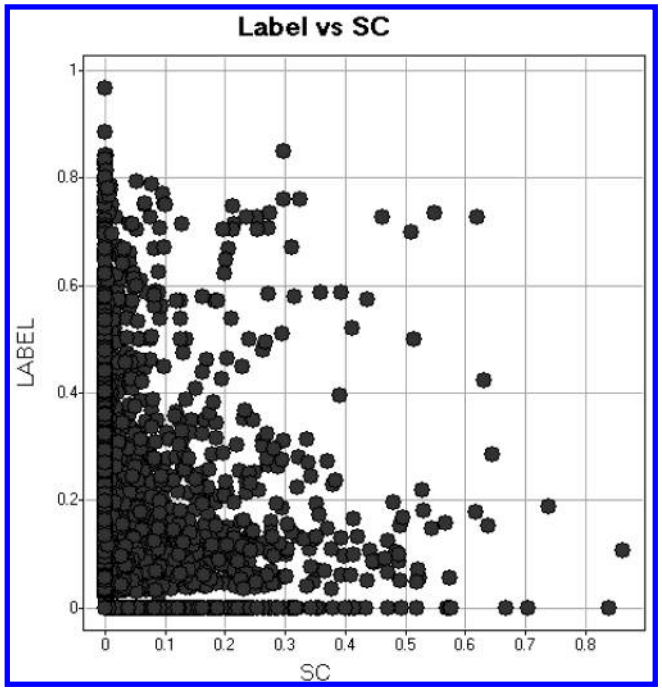

Figure 8. A plot of activity−activity similarity calculated by comparing activity labels vs SC.

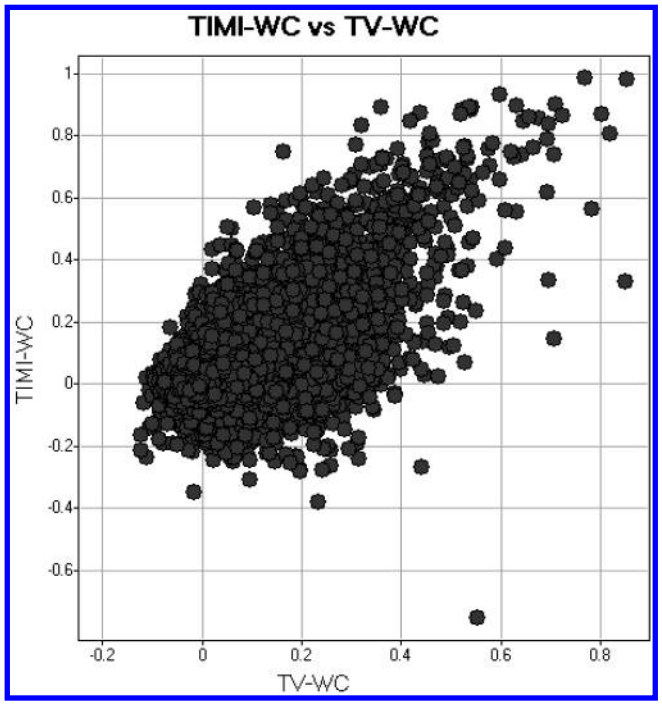

Figure 9. A plot of activity−activity similarity calculated by TIMI-WC vs TV-WC.

because historically a number of mechanisms have been explored. The following mechanisms in that cluster are mentioned in chapter 33 of the Goodman & Gilman text: 34000_DIURETIC, 31300_VASODILATORS, 31260_ADRENERGIC__ALPHA__BLOCKER, 31500_CALCIUM_CHANNEL_BLOCKER, 31410_ACE_INHIBITOR, 31430_ANGIOTENSIN_II_BLOCKER, 31432_ANGIOTENSIN_II_AT1_ANTAGONIST. There are a number of mechanisms that have been explored for antihypertensives, but have not yet made it into Goodman & Gilman because there are not many drugs at present that work by those mechanisms: 31281_DOPAMINE_BETA_HYDROXYLASE_INHIBITOR, 31420_RENIN_INHIBITOR, 31652_ENDOTHELIN_FORMATION_INHIBITOR, 31520_POTASSIUM_CHANNEL_ACTIVATOR. A few activities have a "transitive" relationship to 31000_ANTIHYPERTENSIVE. For example, 34620_NEUTRAL_ENDOPEPTIDASE_INHIBITOR has a high similarity to 31410_ACE_INHIBITOR (all methods), and 31410_ACE_INHIBITOR is a mechanism for antihypertension. Some activities, while not specifically antihypertensive, are cardiovascular, e.g., 28000_CARDIOTONIC, 30000_ANTIANGINAL, 29000_ANTIARRHYTHMIC. The Max similarity is coming from TV-WC. Indeed, the words "cardiovascular", "angiotensin", and "endothelin" are among the more positive components of the 31000_ANTIHYPERTENSIVE trend vector and the trend vectors of these other activities. This is enough to put their similarities over the cutoff of 0.25. Two activities 08000_ANTIDEPRESSANT and 06200_ANXIOLYTIC are clearly false associations and have a similar cause. The nonspecific words "inhibitors", "merck", and "agents" are among the most

Table 6. The Most Discrepant Similarities between TIMI-WC and TV-WC

	TIMI-WC	TV-WC
1430_ANGIOTENSIN_II_BLOCKER 31432_ANGIOTENSIN_II_AT1_ANTAGONIST	−0.754	0.554
64200_CEPHALOSPORIN 64220_CARBACEPHEM	−0.270	0.440
64400_PENEM 64500_CARBAPENEM	−0.380	0.234
31650_ENDOTHELIN_ANTAGONIST 31655_ENDOTHELIN_ETA_ANTAGONIST	0.145	0.706
42120_GRF_ANALOG 60100_VACCINE	−0.243	0.314
37121_FACTOR_XA_INHIBITOR 37260_GPIIB_IIIA_RECEPTOR_ANTAGONIST	−0.264	0.259
07707_ADENOSINE__A1__AGONIST 07708_ADENOSINE__A2__AGONIST	0.329	0.849
75504_ANTIMITOTIC 75506_TAXANE_DERIVATIVE	−0.277	0.243
64500_CARBAPENEM 64530_TRIBACTAM	−0.209	0.296
16250_PROSTAGLANDIN 54121_PROSTAGLANDIN	−0.192	0.312
o o o		
33500_SEPTIC_SHOCK__TREATMENT_FOR 71580_CACHEXIA__TREATMENT_FOR	0.577	0.138
42111_SOMATOSTATIN_ANALOG 42112_SOMATOSTATIN_ANTAGONIST	0.499	0.059
60120_VIRAL_VACCINE 71300_VIRAL_HEPATITIS__AGENT_FOR	0.506	0.050
11100_ANTIPARKINSONIAN 11123_DOPAMINE_AGONIST	0.767	0.310
31330_RENAL_VASODILATOR 37240_CAMP_PHOSPHODIESTERASE_INHIBITOR	0.567	0.104
43136_ADRENOCEPTOR__BETA3__AGONIST 53000_ANTIOBESITY	0.833	0.321
39500_ANTIGLUCOCORTICOID 40231_PROGESTERONE_ANTAGONIST	0.891	0.359
72100_ANTIMALARIAL 72300_ANTIAMEBIC	0.745	0.162

Table 7. The Largest Soft Clusters

cluster center	no. in cluster
08000_ANTIDEPRESSANT	34
06200_ANXIOLYTIC	28
07000_ANTIPSYCHOTIC	25
27200_ANTIALLERGIC_ANTIASTHMATIC	24
02000_ANTIARTHRITIC	23
11100_ANTIPARKINSONIAN	21
31000_ANTIHYPERTENSIVE	19
09200_COGNITION_DISORDERS__AGENT_FOR	17
75000_ANTINEOPLASTIC	17
12452_NEURONAL_INJURY_INHIBITOR	17
28000_CARDIOTONIC	16
33456_ANTIISCHEMIC__CEREBRAL	15
12342_ANTIMIGRAINE	14
34000_DIURETIC	13
52000_HYPOLIPIDEMIC	13
27100_BRONCHODILATOR	12
30000_ANTIANGINAL	12
01200_ANALGESIC__NON_OPIOID	11
42705_CCK_A_AGONIST	11
60000_IMMUNIZING_AGENT__ACTIVE	11

negative components of the trend vectors of 31000_ANTIHYPERTENSIVE and these activities.

DISCUSSION

While calculating similarities between molecules is a staple of molecular modeling, we believe ours is the first attempt to calculate similarities between biological activities. Nothing in this paper should be taken as a criticism of the MDDR. The MDDR is a valuable resource, and the activity labeling is more than adequate for the purpose for which it was intended, as a human-readable database field. To humans familiar with pharmacology, small inconsistencies such as differences in spelling or the use of two different names for the same receptor are inconsequential. However, the amount of information in the MDDR makes it an attractive database to which to apply automated data mining techniques, and for that purpose, inconsistent and incomplete data is a large problem. Here we have proposed methods to calculate similarities between activities in such a way as to at least partly compensate for the imperfections unavoidable in hand-curation. Both the TIMI and TV methods appear promising at finding relationships that are not obvious on the basis of the original information explicit in the database. We are able in many cases to find relationships between activities that are synonymous but differ in the spelling of their labels, find relationships between agonists and antagonists, find mechanistic relationships, etc. In this regard we should again note that what sources say about compounds in their patents (i.e., the words) is at least as important as the chemical structures (i.e., the descriptors) for detecting similarities and making distinctions. On the downside, not all relationships are caught. For instance, there are a handful of activity pairs where the labels are nearly identical, but the compounds and/or words are different enough that no relationship can be detected by SC, TIMI, or TV. Overall, however, the methods look useful, and we expect them to be applicable to other biological activity databases: World Drug Index (www.derwent.com), National Cancer Institute (cactus.nci.nih.gov), Investigational Drugs (www.current-drugs.com), etc.

Given that we can generate similarities, most of the potential applications mentioned in the Introduction are feasible. The one application where our methodology has not proved adequate is finding structurally similar compounds among previously unrelated activities using TIMI-C or TV-C. For instance, one might hope to find the common tricyclic substructure among members of 08000_ANTIDEPRESSANT and 27300_ANTIHISTAMINIC.[18] The main difficulty is that most activities include several chemical classes and a global comparison of the activities is usually not sensitive enough to pick out the similarity of subclasses. A better way of approaching that problem is by comparing compounds pairwise to find the similar compounds, then noting that they have different activities, as in our previous work.[2] A third approach might be to break the activities up into subclasses based on chemical structure and follow the procedures in this paper.

We should also note a limit to the interpretation of the similarities, whatever the method. We expect the similarities to be somewhat sensitive to the version of the database, the type of chemical descriptor, the value of k for TIMI, etc. Also, we treat similarities as symmetric when clearly in some cases they are not. For instance, one might expect most 31410_ACE_INHIBITOR compounds would also be 31000_ANTIHYPERTENSIVE, but only a minority of 31000_ANTIHYPERTENSIVE compounds would be 31410_ACE_INHIBITOR. Thus, the similarity values are at best semiquantitative, so it does not make much sense to discriminate between, say, a similarity of 0.4 and 0.6. For this reason, we prefer to avoid any kind of hierarchical clustering, which is sensitive to differences of this type, and to stick to nonhierarchical methods, which require only a reasonable cutoff that distinguishes "probably real" from "within noise".

There are several fundamental issues with any kind of similarity calculation. First, one must make arbitrary decisions about what one means by "similar" and what features should count toward the similarity. This, in turn, dictates how to represent the objects to be compared and how to relate the representations. For example, in comparing molecules, we would choose a particular type of chemical descriptors (whole-molecule, substructure, etc.), and then choose an appropriate similarity measure (Euclidean distance, Tanimoto index, etc.) to compare the descriptors. In our application, do we want "similar" to mean "involves the same receptor", e.g., include agonists with antagonists, to mean "has very similar compounds in common", or "has the same therapeutic effect"? It is not necessarily clear which is the most useful so we have tried a number of methods with various subsets of words and chemical descriptors.

A corollary of the above issue, in this work as in molecule similarity, is that each method gives slightly different results and no one method finds all the relationships we might hope to see. Obviously, each method we examine here presents its own set of strengths and weaknesses. SC similarity is incapable of generating a falsely high similarity between activities A and B in the sense that if the similarity > 0, it is guaranteed that at least one compound has been claimed to have both activities. On the other hand, because of inconsistencies and incompleteness in hand-curation there are many falsely low similarities. Also, there is no expectation that the similarities are necessarily transitive, i.e., if A is similar to B and B to C, A will not necessarily be similar to C. It is well established that LSI/LSA[8] can extract latent concepts from text and infer a measure of similarity between

Table 8. Selected Large Clusters

	SC	TIMI-WC	TV-WC	Max
08000_ANTIDEPRESSANT 34				
06245_5_HT_REUPTAKE_INHIBITOR	0.303	0.658	0.416	0.658
06200_ANXIOLYTIC	0.471	0.399	0.591	0.591
08410_MAO_A_INHIBITOR	0.141	0.524	0.195	0.524
06247_5_HT1D_ANTAGONIST	0.238	0.517	0.237	0.517
08415_NOREPINEPHRINE_UPTAKE_INHIBITOR	0.129	0.517	0.306	0.517
07000_ANTIPSYCHOTIC	0.219	0.121	0.506	0.506
06235_5_HT1A_AGONIST	0.285	0.420	0.470	0.470
08400_MAO_INHIBITOR	0.096	0.449	0.175	0.449
11100_ANTIPARKINSONIAN	0.163	0.135	0.443	0.443
06240_5_HT1A_ANTAGONIST	0.192	0.422	0.385	0.422
09400_PSYCHOSEXUAL_DYSFUNCTION__AGENT_FOR	0.068	0.383	0.137	0.383
11126_DOPAMINE_REUPTAKE_INHIBITOR	0.135	0.381	0.237	0.381
10000_ANTICONVULSANT	0.066	0.027	0.377	0.377
31262_ADRENERGIC__ALPHA2__BLOCKER	0.133	0.373	0.249	0.373
08010_ADRENOCEPTOR__ALPHA2__ANTAGONIST	0.100	0.359	0.161	0.359
09200_COGNITION_DISORDERS__AGENT_FOR	0.098	0.029	0.343	0.343
06250_5_HT2C_ANTAGONIST	0.137	0.341	0.274	0.341
12342_ANTIMIGRAINE	0.069	−0.007	0.335	0.335
07701_DOPAMINE__D2__ANTAGONIST	0.043	−0.002	0.332	0.332
01200_ANALGESIC__NON_OPIOID	0.054	0.004	0.317	0.317
06248_5_HT2A_ANTAGONIST	0.073	0.120	0.300	0.300
12452_NEURONAL_INJURY_INHIBITOR	0.025	−0.025	0.296	0.296
08420_MAO_B_INHIBITOR	0.079	0.274	0.231	0.274
33456_ANTIISCHEMIC__CEREBRAL	0.014	−0.022	0.270	0.270
06246_5_HT1D_AGONIST	0.036	−0.040	0.269	0.269
07712_SIGMA_ANTAGONIST	0.043	0.026	0.269	0.269
31000_ANTIHYPERTENSIVE	0.052	0.020	0.268	0.268
06249_5_HT2B_ANTAGONIST	0.082	0.268	0.175	0.268
06100_SEDATIVE_HYPNOTIC	0.036	0.021	0.267	0.267
84900_PHARMACOLOGICAL_TOOL	0.010	−0.051	0.262	0.262
78393_COMT_INHIBITOR	0.051	0.258	0.064	0.258
11130_MAO_B_INHIBITOR	0.054	0.257	0.127	0.257
10712_GABA_UPTAKE_INHIBITOR	0.089	0.254	0.101	0.254
27200_ANTIALLERGIC_ANTIASTHMATIC 24				
78351_LIPOXYGENASE_INHIBITOR	0.449	0.380	0.472	0.472
27240_MEDIATOR_RELEASE_INHIBITOR	0.244	0.461	0.340	0.461
02100_ANTIINFLAMMATORY	0.236	0.134	0.428	0.428
27210_LEUKOTRIENE_ANTAGONIST	0.348	0.285	0.420	0.420
27220_LEUKOTRIENE_SYNTHESIS_INHIBITOR	0.275	0.266	0.377	0.377
59300_ANTIPSORIATIC	0.106	0.062	0.356	0.356
27100_BRONCHODILATOR	0.122	0.061	0.353	0.353
02000_ANTIARTHRITIC	0.103	−0.004	0.343	0.343
27212_LEUKOTRIENE_D4_ANTAGONIST	0.264	0.247	0.335	0.335
78444_TRYPTASE_INHIBITOR	0.075	0.314	0.087	0.314
78331_CYCLOOXYGENASE_INHIBITOR	0.132	0.132	0.311	0.311
26105_RHINITIS__AGENT_FOR	0.092	0.308	0.291	0.308
59200_ANTIINFLAMMATORY__TOPICAL	0.037	−0.005	0.292	0.292
33500_SEPTIC_SHOCK__TREATMENT_FOR	0.095	0.055	0.287	0.287
78418_PHOSPHODIESTERASE_IV_INHIBITOR	0.221	0.200	0.284	0.284
27300_ANTIHISTAMINIC	0.148	0.230	0.273	0.273
27261_PAF_ANTAGONIST	0.108	0.070	0.268	0.268
62200_IMMUNOSUPPRESSANT	0.022	−0.021	0.263	0.263
54120_ANTIULCERATIVE	0.058	0.003	0.258	0.258
01200_ANALGESIC__NON_OPIOID	0.061	−0.011	0.257	0.257
27214_LEUKOTRIENE_B4_ANTAGONIST	0.160	0.218	0.256	0.256
42733_NEUROKININ_NK2_ANTAGONIST	0.132	0.254	0.216	0.254
02000_ANTIARTHRITIC 23				
02450_IL_1_INHIBITOR	0.205	0.477	0.298	0.477
02454_TNF_INHIBITOR	0.204	0.228	0.393	0.393
02100_ANTIINFLAMMATORY	0.056	−0.029	0.387	0.387
33500_SEPTIC_SHOCK__TREATMENT_FOR	0.172	0.154	0.371	0.371
78371_COLLAGENASE_INHIBITOR	0.250	0.366	0.299	0.366
59300_ANTIPSORIATIC	0.108	0.080	0.359	0.359
50050_TREATMENT_FOR_OSTEOPOROSIS	0.093	0.058	0.359	0.359
27200_ANTIALLERGIC_ANTIASTHMATIC	0.103	−0.004	0.343	0.343
59200_ANTIINFLAMMATORY__TOPICAL	0.088	0.160	0.341	0.341
62200_IMMUNOSUPPRESSANT	0.098	0.073	0.336	0.336
75000_ANTINEOPLASTIC	0.119	0.025	0.328	0.328
02448_CYTOKINE_MODULATOR	0.054	0.328	0.135	0.328
75751_ANTIANGIOGENIC	0.089	0.113	0.322	0.322
02150_DISEASE_MODIFYING_DRUG	0.041	0.319	0.099	0.319
78432_MATRIX_METALLOPROTEINASE_INHIBITOR	0.228	0.308	0.298	0.308
02455_IL_6_INHIBITOR	0.070	0.283	0.166	0.283
27520_ARDS__AGENT_FOR	0.117	0.238	0.282	0.282
71580_CACHEXIA__TREATMENT_FOR	0.048	0.273	0.130	0.273
78351_LIPOXYGENASE_INHIBITOR	0.039	−0.042	0.268	0.268
12340_MULTIPLE_SCLEROSIS__AGENT_FOR	0.096	0.267	0.261	0.267
02451_IL_1BETA_CONVERTING_ENZYME_INHIBITOR	0.137	0.262	0.181	0.262
11100_ANTIPARKINSONIAN 21				
11123_DOPAMINE_AGONIST	0.301	0.767	0.310	0.767

Table 8 (Continued)

	SC	TIMI-WC	TV-WC	Max
08420_MAO_B_INHIBITOR	0.189	0.674	0.305	0.674
78393_COMT_INHIBITOR	0.132	0.591	0.170	0.591
11126_DOPAMINE_REUPTAKE_INHIBITOR	0.203	0.550	0.283	0.550
11125_DOPAMINE__D2__AGONIST	0.183	0.525	0.274	0.525
42610_PROLACTIN_SECRETION_INHIBITOR	0.145	0.492	0.207	0.492
11130_MAO_B_INHIBITOR	0.100	0.472	0.195	0.472
08000_ANTIDEPRESSANT	0.163	0.135	0.443	0.443
07000_ANTIPSYCHOTIC	0.130	0.070	0.395	0.395
11124_DOPAMINE__D1__AGONIST	0.142	0.388	0.231	0.388
10000_ANTICONVULSANT	0.061	0.051	0.359	0.359
06200_ANXIOLYTIC	0.085	0.048	0.359	0.359
12452_NEURONAL_INJURY_INHIBITOR	0.048	0.011	0.345	0.345
09200_COGNITION_DISORDERS__AGENT_FOR	0.127	0.110	0.344	0.344
31262_ADRENERGIC__ALPHA2__BLOCKER	0.105	0.319	0.220	0.319
78440_PROLYL_PEPTIDYL_ISOMERASE_INHIBITOR	0.071	0.299	0.118	0.299
33456_ANTIISCHEMIC__CEREBRAL	0.023	0.006	0.290	0.290
12340_MULTIPLE_SCLEROSIS__AGENT_FOR	0.065	0.283	0.138	0.283
12455_NMDA_RECEPTOR_ANTAGONIST	0.021	0.003	0.255	0.255
12342_ANTIMIGRAINE	0.012	−0.022	0.251	0.251
31000_ANTIHYPERTENSIVE 19				
28000_CARDIOTONIC	0.229	0.164	0.454	0.454
31281_DOPAMINE_BETA_HYDROXYLASE_INHIBITOR	0.096	0.390	0.064	0.390
30000_ANTIANGINAL	0.185	0.134	0.360	0.360
31410_ACE_INHIBITOR	0.216	0.355	0.280	0.355
31430_ANGIOTENSIN_II_BLOCKER	0.343	0.118	0.259	0.343
31260_ADRENERGIC__ALPHA__BLOCKER	0.095	0.341	0.144	0.341
31420_RENIN_INHIBITOR	0.332	0.313	0.214	0.332
31432_ANGIOTENSIN_II_AT1_ANTAGONIST	0.319	0.142	0.224	0.319
34000_DIURETIC	0.166	0.319	0.299	0.319
31300_VASODILATOR	0.157	0.208	0.307	0.307
34620_NEUTRAL_ENDOPEPTIDASE_INHIBITOR	0.165	0.299	0.240	0.299
31652_ENDOTHELIN_FORMATION_INHIBITOR	0.107	0.199	0.274	0.274
31500_CALCIUM_CHANNEL_BLOCKER	0.246	0.223	0.271	0.271
35500_RENAL_FAILURE__AGENT_FOR	0.106	0.253	0.269	0.269
08000_ANTIDEPRESSANT	0.053	0.020	0.268	0.268
31520_POTASSIUM_CHANNEL_ACTIVATOR	0.193	0.227	0.259	0.259
06200_ANXIOLYTIC	0.036	−0.011	0.253	0.253
29000_ANTIARRHYTHMIC	0.052	0.009	0.251	0.251
75000_ANTINEOPLASTIC 17				
02000_ANTIARTHRITIC	0.119	0.025	0.328	0.328
18310_RETINOPROTECTOR	0.020	0.126	0.007	0.126
75751_ANTIANGIOGENIC	0.178	0.301	0.323	0.323
78370_TYROSINE_SPECIFIC_PROTEIN_KINASE_INHIBITOR	0.229	0.241	0.317	0.317
59300_ANTIPSORIATIC	0.123	0.050	0.310	0.310
75650_INTERCALATING_AGENT	0.126	0.169	0.303	0.303
33451_RESTENOSIS__AGENT_FOR	0.112	0.117	0.301	0.301
50050_TREATMENT_FOR_OSTEOPOROSIS	0.083	0.031	0.295	0.295
75855_SIGNAL_TRANSDUCTION_INHIBITOR	0.073	0.044	0.294	0.294
75721_AROMATASE_INHIBITOR	0.204	0.289	0.184	0.289
75200_ALKYLATING_AGENT	0.123	0.174	0.288	0.288
62200_IMMUNOSUPPRESSANT	0.066	0.009	0.283	0.283
78374_PROTEIN_KINASE_C_INHIBITOR	0.123	0.161	0.277	0.277
62000_IMMUNOMODULATOR	0.051	0.000	0.276	0.276
75711_ANTIESTROGEN	0.129	0.262	0.253	0.262
75100_ANTIMETABOLITE	0.200	0.256	0.260	0.260
71000_ANTIVIRAL	0.088	−0.037	0.255	0.255
50060_BONE_RESORPTION_INHIBITOR 9				
50050_TREATMENT_FOR_OSTEOPOROSIS	0.276	0.558	0.608	0.608
50000_CALCIUM_REGULATOR	0.108	0.534	0.499	0.534
51300_BONE_REGENERATION__AGENT_FOR	0.018	0.528	0.331	0.528
50150_BISPHOSPHONATE	0.278	0.493	0.384	0.493
51310_BONE_FORMING_FACTOR	0.000	0.387	0.053	0.387
28330_ADENYLATE_CYCLASE_ACTIVATOR	0.000	0.310	0.042	0.310
18310_RETINOPROTECTOR	0.000	0.299	0.116	0.299
40210_ESTROGEN	0.066	0.286	0.160	0.286
71521_ANTIVIRAL__AIDS_ 7				
71522_REVERSE_TRANSCRIPTASE_INHIBITOR	0.414	0.619	0.541	0.619
71523_HIV_1_PROTEASE_INHIBITOR	0.480	0.593	0.384	0.593
71000_ANTIVIRAL	0.205	0.127	0.470	0.470
71527_HIV_INTEGRASE_INHIBITOR	0.152	0.438	0.201	0.438
71524_TAT_INHIBITOR	0.083	0.400	0.132	0.400
78330_PROTEASE_INHIBITOR	0.230	0.361	0.248	0.361
75100_ANTIMETABOLITE 7				
78362_THYMIDYLATE_SYNTHETASE_INHIBITOR	0.421	0.679	0.565	0.679
78403_GLYCINAMIDE_RIBONUCLEOTIDE_..._INHIBIT	0.235	0.442	0.303	0.442
78403_GLYCINAMIDE_RIBONUCLEOTIDE_..._INHIBITOR	0.166	0.453	0.303	0.453
78407_S_ADENOSYL_L_METHIONINE_DECA.._INHIBITOR	0.090	0.191	0.323	0.323
78389_DIHYDROFOLATE_REDUCTASE_INHIBITOR	0.221	0.298	0.308	0.308
75000_ANTINEOPLASTIC	0.200	0.256	0.260	0.260

words that agrees well with human perception. That it can do this is remarkable, given that LSI knows only about word co-occurrence, ignores word order, and knows nothing about the meaning of the words. This useful behavior has also been previously demonstrated for words and descriptors in our derived method TIMI.[9] However, such methods suffer from a number of drawbacks. First, results depend on *k*, and it is not obvious a priori what value of *k* is optimum for a given problem. This must be approached empirically, sometimes without an objective measure of goodness. Second, given that the similarity is a normalized dot product over latent terms, each of which in turn is a linear sum of the original terms, it is sometimes very hard to understand why TIMI finds certain relationships and misses others. (This is a recognized issue with LSI/LSA.[8]) In our particular exercise, we have a handful of persistent false negatives for which we have no easy explanation. Trend vectors are more easily interpretable, being a linear combination of the original terms. Given that the similarity is the normalized dot product of two vectors on the original terms, one may easily monitor which terms are contributing the most to the sum. On the other hand, statistically significant trend vectors cannot be calculated for all activities. Also, in TV each term is treated as a distinct string, and there is no partial synonymy as there is in TIMI. Clearly, then, one must use a combination of methods to attack this problem, and no doubt more methods can be explored.

ACKNOWLEDGMENT

The authors thank the following people: Dr. Eugene Fluder implemented the Ibrahimov soft clustering method. Dr. Robert Nachbar implemented the similarity algorithms. Dr. Simon Kearsley wrote the current implementation of trend vectors. Dr. Richard Hull and Dr. Suresh Singh did the early work on TIMI.

Supporting Information Available: Comma-separated table of all activity-activity similarities using five methods. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) MDDR licensed by Molecular Design, Ltd., San Leandro, CA. www.mdli.com.

(2) Sheridan, R. P. Finding multiactivity substructures by mining databases of drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037−1050.

(3) *Goodman and Gilman's Pharmacological Basis of Therapeutics*, 10th ed.; Hardman, J. G., Limbird, L. E., Goodman, A. G., Eds.; McGraw-Hill: New York, 2001.

(4) Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J.; Lecchini, S.; Jacoby, E. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947−955.

(5) 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.

(6) 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.

(7) 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*, 391−407.

(8) Landauer, T. K.; Foltz, P. W.; Laham, D. An introduction to Latent Semantic Analysis. *Discourse Processes* **1998**, *25*, 259−284.

(9) Singh, S. B.; Hull, R. D.; Fluder, E. M. Text influenced molecular indexing (TIMI): a literature database mining approach that handles text and chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 743−752.

(10) 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.

(11) 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.

(12) Butina, D. Unsupervised database clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large datasets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(13) Ibrahimov, O.; Sethi, I.; Dimitrova, N. A novel similarity based clustering algorithm for grouping broadcast news. Proceedings of the SPIE Conference, Orlando, FL, Apr 1−4, 2002; Vol 4730, pp 394−404.

(14) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of 'molecular diversity' descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(15) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(16) Wild, D. J.; Blankley, C. J. Comparison of 2D fingerprint types and hierarchical level selection methods for structural groupings using Ward's clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155−162.

(17) Strehl, A.; Ghosh, J. Cluster ensembles−a knowledge reuse framework for combining multiple partitions. *J. Machine Learning Res.* **2002**, *3*, 583−617.

(18) Sheridan, R. P.; Miller, M. D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915−924.

CI034245H