Why do we need so many chemical similarity search methods?

Robert P. Sheridan and Simon K. Kearsley

Computational tools to search chemical structure databases are essential to finding leads early in a drug discovery project. Similarity methods are among the most diverse and most useful. We will present some lessons we have gathered over many years experience with in-house methods on several therapeutic problems. The effectiveness of any similarity method can vary greatly from one biological activity to another in a way that is difficult to predict. Also, any two methods tend to select different subsets of actives from a database, so it is advisable to use several search methods where possible.

Robert P. Sheridan*
Simon K. Kearsley
Dept of Molecular Systems
RY50SW-100 Merck Research Laboratories
Rahway, NJ 07065, USA
*tel: +1 732 594 3859
fax: +1 732 594 4224
e-mail: sheridan@merck.com
e-mail: skk@merck.com

▼ Finding leads, drug-like compounds that are worthy of further synthetic or biological study, is a primary goal early in a drug discovery project. Computational search methods have been very useful in this endeavor. The chemist typically starts with a 'query' and uses appropriate software to search a database of chemical structures. The query can be a molecule with some kind of interesting biological activity (e.g. a competitor's compound or an in-house compound from screening), or else a hypothesis about the structural requirements for that activity, for example, a pharmacophore or a quantitative SAR (QSAR) model. The chemist is usually trying to identify other molecules in the database that can then be tested in the appropriate assay. Often the ultimate goal is to find a compound that is different enough (at least from a patent point of view) from the query compound(s), or other previously known compounds, that it could be considered a new class of therapeutic agents.

Why use search methods?

Recently, chemical search methodology has been called 'virtual screening' [1–3] by analogy with HTS, the idea being that these methods are testing large numbers of compounds by computer instead of by experiment. The term originally referred to the screening of virtual combinatorial libraries, but can equally apply to corporate or commercial collections of compounds because the computational methods are identical.

At present, new methods are often justified on the grounds that they support HTS; the entirety of corporate-sized databases (often exceeding 1 million entries) can not be screened at a reasonable cost even by current technology, and computer-based methods can be used to suggest subsets of compounds that are most likely to be active, thus making targeted screening more efficient than a 'test everything' approach. However, all of the chemical searching methods discussed here were in full development long before HTS or combinatorial technology existed, when the expense of experimentally testing a compound relative to scoring it computationally was even greater than it is now.

Classification of virtual screening methods

One possible classification of current virtual screening methods, and the paradigms behind them, is listed in Box 1. This includes the question each method asks of molecules in the database, and what assumption the user is making about the activity. Many methods have been implemented by commercial software vendors, some are licensable from academic groups, but many pharmaceutical companies, including Merck (http://www.merck.com), develop and maintain their own proprietary versions.

Virtual screening methods differ in what information they require before a meaningful query can be formed. There is also a substantial difference in computational expense; the most notable difference is that 3D versions require the generation of reasonable conformers of the molecules in the database. Given the size of corporate databases, a practical search method cannot take more than a few seconds elapsed time per compound.

1359-6446/02/$ – see front matter PII: S1359-6446(02)02411-X

Box 1. Virtual Screening Methods

I. Substructure search
Question: Which molecules in a database contain the specified substructure?
Model: Compounds that contain this substructure are likely to be active.
Query requires: 2D or 3D substructure common to actives ('pharmacophore' in 3D).

II. Similarity
Question: Which molecules in a database are 'similar' to the query molecule(s)?
Model: Compounds globally similar in structure to the query are likely to have a similar activity.
Query requires: One or more active molecules.

III. Docking
Question: Which molecules in a database can fit into the binding site of a known enzyme or receptor?
Model: Compounds that have a high interaction score are likely to bind to the enzyme.
Query requires: Atomic-level structure of receptor (or model built from homolog).

IV. QSAR
Question: Which molecules have the highest predicted activity?
Model: Compounds with high predicted activity are likely to be active.
Query requires: Activity data on enough compounds to form QSAR.

Methods might use a 2D or 3D representation of molecules. 2D or 'topological' refers to information derived from the connection table of a molecule (as represented by the 2D structural drawing). 3D refers to information derived from 3D coordinates. Docking is the only method that works only in 3D.

Why similarity methods are especially useful

The similarity methods [4–8], which came into wide use in the 1980s, have proven extremely useful in the pharmaceutical setting. A few reasons for this are:

(1) Little information is needed to formulate a reasonable query. In particular, no assumption need be made about which part(s) of the query molecule confers activity. In the 2D versions, nothing needs to be known about the active conformation of the query molecule(s). Thus, similarity methods can be used at the beginning of a drug discovery project when there is little information about the target and only one or two known actives. (Later, of course, methods that require more information can be brought to bear.)

(2) Many implementations of similarity methods are computationally inexpensive, so searching large databases can be routinely performed.

(3) The assumption that molecules that are globally similar in structure should exhibit similar biological activity is generally valid [9–11]. The best current phrase for this is that molecules exhibit 'neighborhood behavior' [9].

Superposition- and histogram-based similarity methods

Similarity methods have been under development for decades, and a bewildering variety of approaches have been tried. Obviously, it would be impossible to even briefly discuss them all, so we refer the reader to more comprehensive reviews [4–8]. There is clearly a lot of 'art' involved in defining similarity, and different definitions are useful for different purposes. Any method conceptually consists of two independent aspects: how to represent molecules, and how to calculate the similarity between them. A broad class of superposition methods [12–22] tries to map one molecule onto another. In 2D, this means treating the molecules A and B as graphs and finding the correspondence between atoms in A and atoms in B [12,13]. In 3D, this means finding the best superposition of the molecules as 3D objects (with or without taking flexibility into account). The superposition can be based on atoms [14] or some kind of 'field' (including shape) surrounding the atoms [15–21]. The similarity would take into account the overlap between the fields of A and B. A second broad class of methods transforms 2D or 3D structures into one or more 'spectra' or histograms and then calculates the overlap of the histograms [23,24].

Descriptor-based similarity methods

The third, and most popular, class of methods represents a molecule as a set of numbers or 'descriptors,' such that a molecule can be considered a point in a multidimensional 'descriptor space'. The disadvantage of this, relative to superposition methods, is that the equivalence of parts between one molecule and another is lost, but the advantage is that computation becomes much simpler. A subclass of methods uses a set of numbers where each number is a property of the whole molecule (e.g. molecular weight, log D, number of hydrogen bond donors, Kier index, dipole moment, BCUT parameters [25–27], and so on; reviewed in [28]). The properties can be derived from the topology or 3D structure of the molecule. Similarity between two molecules is some inverse function of the distance between them in descriptor space. Another subclass represents molecules as a set

of user-defined 2D or 3D substructures and their frequencies. The substructures are used as the descriptors. In 'fingerprint' methods, only the presence or absence of a descriptor is noted. Similarity is defined by how many descriptors two compounds have in common normalized by the number of descriptors in each. This subclass of methods is efficient because it is computationally inexpensive to compare lists of pre-computed descriptors, especially in the case of a fingerprints. The most popular commercial similarity methods are of this type [29–31].

Of course, one still has a large number of possible substructures that one could use as descriptors. For 2D representations one could use substructures that are familiar to chemists (hydroxyl, indole, pyridine, and so on) or more abstract substructures consisting of a few atoms and their bonded environments. In 3D, a typical descriptor would be a 3- or 4-point pharmacophore [32,33]. Our group has proposed a quantity called 'fuzziness' [34], which measures how likely they are to occur in two arbitrary molecules. The more fuzzy the descriptor, the more similar two arbitrary molecules will appear. Ideally, one would like a descriptor that is fuzzy enough to enable the detection of similarity between most molecules with the same activity, but not so much that too many false associations are made.

Further complications to the descriptor-based representation can be introduced by 'reducing the dimensionality', for example, considering only a subset of the most relevant descriptors [35–37], or finding a small set of orthogonal latent descriptors [25,38–40]. Cell-based methods introduce a further modification by partitioning the lower-dimensional descriptor space into multi-dimensional rectangular regions [25–27]. Because some reduced-dimensionality methods require activity data on many compounds to help select the best subset of descriptors, they are somewhat like QSAR methods.

One promising area that has not been much explored [35,40,41] has been the combination of more than one molecule in a single query. Another is the combination of scores from different descriptors [34] or different similarity methods [24], the idea being that deficiencies in one method would be compensated for by others. This has been called 'data fusion' [24], and is analogous to 'consensus scoring' [42] in the docking field.

How good is a similarity method as a virtual screening tool?

The reader should note that all the similarity methods can be used for several applications, including:

(1) Clustering: grouping similar compounds together [10].
(2) Diversity: selecting a subset of disparate molecules from a larger set [43].
(3) Virtual screening.

Each of the above is a whole field onto itself, but it is the last we are concerned with here.

Can we identify a consistently better descriptor, or a class of methods that is better than the others for this purpose? First, we need to objectively define 'better'. Two useful criteria are: (1) How good are the methods at selecting the actives from a database? (2) How likely are we to find novel actives?

Any evaluation of goodness in any specific case will depend on the query, the method and the contents of the database, so any comparison between methods must use the same query and database.

The first criterion can be addressed in a retrospective analysis. Assume n_{act} active molecules in a database of N molecules. If molecules are 'tested' in order of decreasing score on some virtual screen, we can plot the total number of actives found versus the total number of molecules tested as a 'cumulative recall curve' [34,44]. If similarity to the query, for example, were a perfect discriminator of activity we would find all the actives at the front of the list, so the curve would have a slope of 1 until n_{act} molecules were tested; thereafter the slope would be zero. If similarity were of no use at all, the curve would be a line with slope n_{act}/N. In practice, the curves are usually hyperbolic, indicating the front of the list is enriched in actives (see Fig. 1). The cumulative recall curve is a useful way of comparing methods because it depends only on the rank of the molecules (the best-scoring molecule is rank 1, the next rank 2, and so on) and not absolute score, which is generally not comparable between methods. One useful resource for retrospectively testing is the MDL Drug Data Report (MDDR) database [45], a licensable database compiled from the patent literature By Molecular Design Ltd (http://www.mdli.com). It contains many diverse drug-like molecules to which most have been assigned a therapeutic category. For our purposes, the most important limit of patent-based databases is 'false inactives': only one or two activities are reported for any given compound but that compound might actually be active in another area, if only it were tested.

For the second criterion, the idea of 'novelty' relative to the query depends on a definition of similarity to the query, and that is the thing we are comparing among methods. Our resolution is to define a separate similarity standard that uses a set of specific descriptors with which only close analogs would be considered 'similar'. By that definition, if the actives have low similarity to the query they can be considered novel.

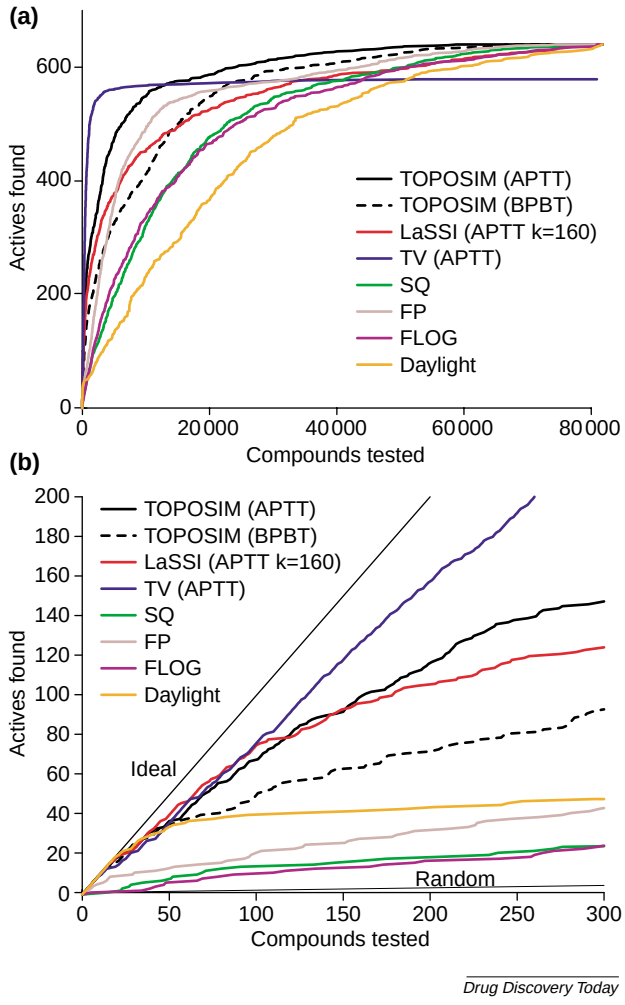

Figure 1. (a) The cumulative recall curves for in-house methods applied to searching the MDL Drug Data Report database (http://www.mdli.com) for HIV-1 protease inhibitors. Better methods find more actives at a given number of molecules tested. APTT indicates that the methods uses a combination of AP and TT descriptors, similarly for BTBT. The TV curve does not reach the same number of actives as the others because some actives were used to create the training set for that method. **(b)** A close-up of **(a)**. On the graph are shown the idealized situations where all the actives are at the front of the list ('ideal') or are randomly distributed in the list ('random'). TOPOSIM, LaSSI, TV and Daylight are 2D methods; SQ, FP and FLOG are 3D methods.

Comparisons of in-house methods for an HIV protease inhibitor query

Here we will present some general observations we have gathered over many person-years experience at Merck with our in-house similarity methods on many therapeutic problems. By no means are we claiming that these methods are special. They are representative of extant methodologies and our conclusions should be broadly applicable. We will consider inhibition of HIV-1 protease as the biological activity. It has current clinical relevance, the crystal structures of many inhibitor-protease complexes are known, and many diverse chemical classes of inhibitors are in the MDDR under the therapeutic category 'HIV-1 protease inhibitor'.

For the descriptor-based topological methods [TOPOSIM (topological similarity), LaSSI (latent semantic structural indexing) and TV (trend vector)] we will use a combination of AP (regular atom pair) [46] and TT (topological torsion) [47]. These descriptors take into account the element, hybridization and number of bonded neighbors of each atom. Alternatively, atoms can be classified into seven 'physiochemical types' (cation, hydrogen bond donor, hydrophobe, and so on) [48] and these types can be used to generate analogous descriptors BP (binding pair) and BT (binding torsion) [34].

For 3D methods, the conformational flexibility of the database molecules must be taken into account by generating low-energy conformers. We have preferred an approach where these conformers are pre-calculated and stored in a 'flexibase' [49,50], and our 3D methods (SQ, FP, FLOG (flexible ligands oriented on grid)] work with the individual conformers. The score of a molecule is taken as that of its highest-scoring conformer. For the 3D methods, atoms are classified into the same physiochemical types as for the BP and BT descriptors.

(1) TOPOSIM [34] uses the Dice normalization [6] to score the similarity of the query versus each database entry. The query consists of the AP and TT descriptors parsed from the connection table of the HIV inhibitor indinavir. In Figs 1 and 2, APTT will indicate the combination of AP and TT.

(2) LaSSI [38–40] uses the same topological descriptors, but these are projected into a lower-dimensional space as k orthogonal latent descriptors. k is an adjustable parameter and here we use the value of k (k = 160) at which the number of returned actives is highest for indinavir as a query.

(3) TV analysis [51] is partial-least squares regression applied to AP and TT descriptors. Being a QSAR method, TV must start with a training set, which was constructed by randomly selecting 10% of the HIV protease inhibitors in the MDDR and 1000 non-inhibitors. The TV summarizes the difference in AP and TT descriptors between inhibitors and other drug-like compounds in the training set. It can then be used to predict the activity of the remainder of the MDDR. Although it is not a similarity method, we include TV because it provides one kind of upper limit for how well we can do with the AP and TT descriptors.

(4) SQ [18] is a 3D superposition similarity method for finding the best rigid overlap of a flexibase conformer with a query molecule. The SQ score takes into account the physiochemical types assigned to the atoms and the distance between the database and query atoms. The query is the conformer of indinavir co-crystallized with HIV protease (Protein Data Bank [52] structure 1HSH), thus it is by definition in the 'correct' conformation.

(5) FP (B. Feuston, pers. commun.) is a descriptor-based similarity method using 3D substructure descriptors. Each descriptor is a triplet of atoms and the distances between them. These are similar to the 3-point pharmacophores described by Pickett *et al.* [32]. We measure Dice similarity between the co-crystalized conformer of indinavir and each conformer in the flexibase.

(6) FLOG [53] is our docking method. Here candidate conformers are docked into a protein site of known structure, in this case 1HSH. FLOG is not a similarity method, but we include it because it uses a model of the receptor and is not dependent on an arbitrary choice of query molecule.

(7) Daylight, a commercial 2D fingerprint method licensed by Daylight Chemical Information Systems (http://www.daylight.com) [29] is one *de facto* standard for topological similarity.

Recall curves for the seven methods are shown in Fig. 1. This example is typical in that we usually observe the following about the ability of the methods to select actives:

(1) TV performs the best: this is not surprising because its training set contains information about many classes of actives, not just the query indinavir.

(2) The other topological methods, TOPOSIM and LaSSI, are equivalent and are good at selecting actives.

(3) Our 3D methods are consistently poorer than the 2D methods. FLOG is usually the worst.

(4) For the indinavir–HIV protease problem, and approximately half the problems we have examined, the effectiveness of Daylight is confined to the beginning of the list.

Note that all the methods select actives from MDDR far better than a random selection would.

2D versus 3D

There has always been an expectation that 3D methods would be better than topological methods in selecting actives. After all, we think it is the 3D presentation of chemical groups, including absolute stereochemistry, that makes a drug bind to a receptor. However, the opposite trend has been seen many times in retrospective studies in the literature using different methods from ours [11,54,55] (although

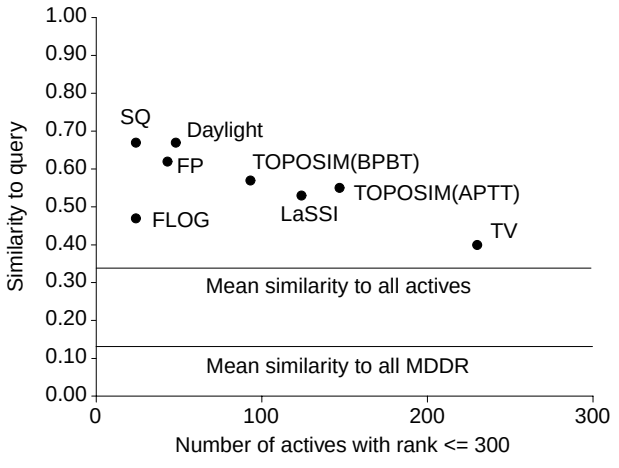

Figure 2. A plot of the mean similarity to the query (using the TT descriptor) of the actives in the first 300 molecules tested (from the methods in Fig. 1). The topological torsion (TT) descriptor is used here because it is specific [34]; only close analogs would have a high similarity to the query. For the TV method, the query is the descriptor average of all the actives in the training set. For all other methods, the query is the topological structure of indinavir. This graph is meant to show whether the methods find novel actives relative to the query; lower similarity would indicate more novelty. A horizontal line marks the average similarity to indinavir of all actives (HIV protease inhibitors) in the MDL Drug Data Report (MDDR; http://www.mdli.com). The fact that all the methods are above this line indicate that no method is sampling the full diversity of the entire set of actives by 300 compounds tested. TOPOSIM, LaSSI, TV and Daylight are 2D methods; SQ, FP and FLOG are 3D methods.

counterclaims should be noted [56,57]), so our observations are probably not a result of problems with our particular implementations. Two reasons have been suggested [55] to explain this discrepancy: (1) Because there are many topologically similar drugs in the patent literature, topological methods will artificially appear to be pulling out more actives; and (2) We do not sample conformational space finely enough in our calculations.

We are not convinced that either of these two explanations is complete. First, if we repeat our searches over a database that includes only topologically diverse compounds from the MDDR (~20,000 out of ~82,000), the curves do not appreciably shift relative to each other. Second, although conformational sampling is probably an important issue, when we search over fourfold more conformers, the curves for the 3D methods do not improve significantly. A third possible explanation is that the description of atoms typically used by 3D methods is not specific enough to capture the subtleties of the requirements for activity (for instance by not distinguishing between aromatic and aliphatic hydrophobes). This idea could have merit because, in our case, using the more

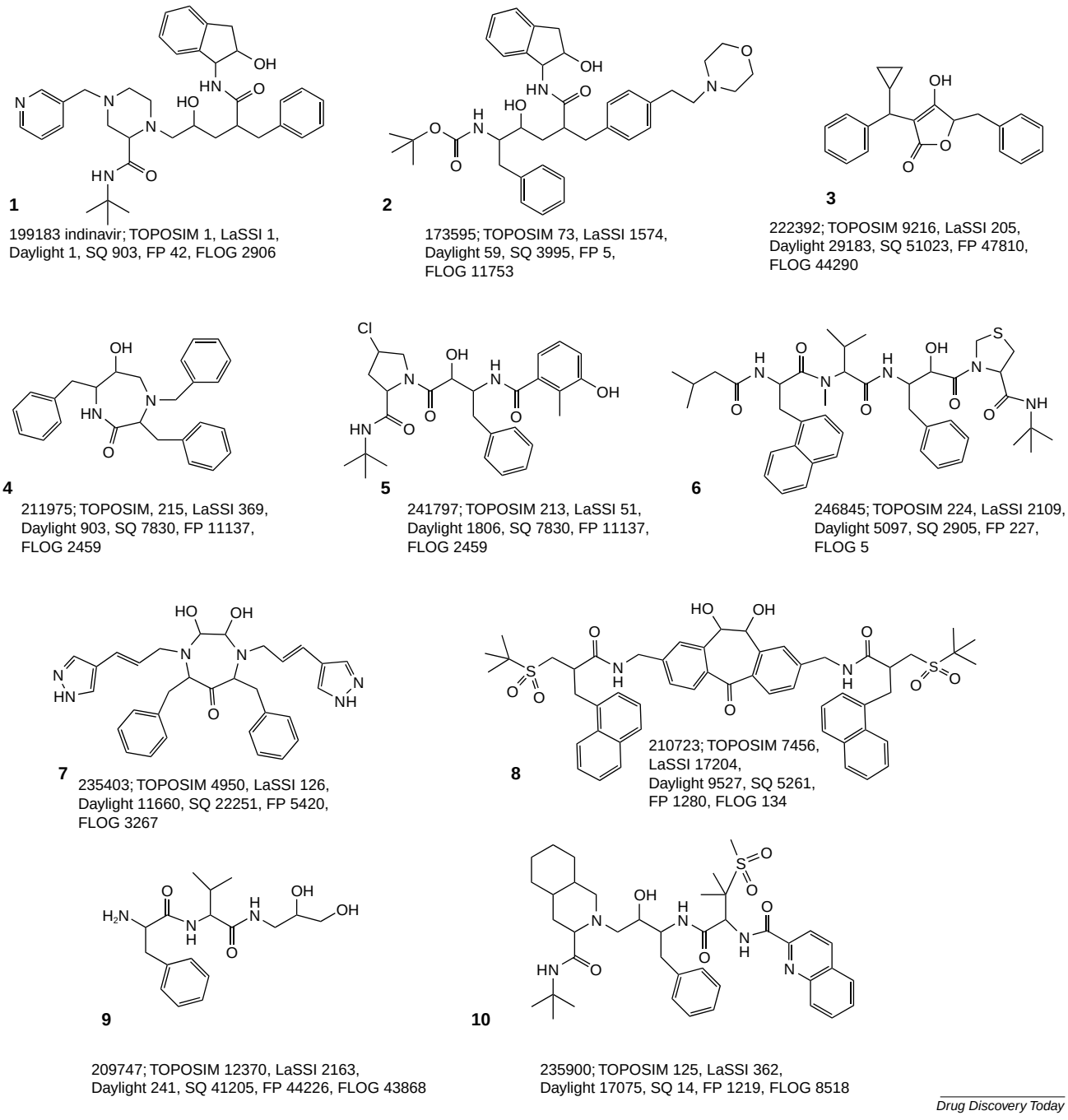

Figure 3. Selected active compounds that are ranked highly by one method and not very highly by most of the others. These were selected by comparing the minimum rank of the compound for all the methods to the median rank or the next highest rank. The caption under each compound lists the MDL Drug Data Report (http://www.mdli.com) external registration number and then the rank of that compound by six methods. It should be noted that because of the limits of representing conformational space by a few tens of explicit conformers, there might be no conformer of indinavir in the database close to the receptor-bound conformer of indinavir, which forms the query. Thus, the similarity rank of indinavir can be >> 1 for the 3D methods. TOPOSIM, LaSSI, TV and Daylight are 2D methods; SQ, FP and FLOG are 3D methods.

generic topological descriptors BP and BT, instead of the specific AP and TT descriptors, usually results in a degradation of the performance of TOPOSIM. A fourth, broader but harder to test, explanation might be that the connection table of a molecule encodes so much implicit information about the 3D structure of a molecule that using actual 3D coordinates adds little more information, but adds noise because of insufficient conformational sampling.

How many actives a method selects is only one criterion. We would also be pleased if a method were able to find actives that are not obvious analogs of the query. Figure 2 shows the mean similarity (TT descriptor) to the query for the actives in the highest-scoring 300 molecules versus the number of such actives for the methods. The smaller the mean similarity, the more 'novel' the actives. Generally, it is felt that 3D methods, which ignore bond topology in favor of atom positions, are better able to 'hop' across chemical classes. Therefore, we might expect that 3D methods offer more novelty in compensation for selecting fewer actives; however, in our hands there is no systematic difference between 2D and 3D methods, at least for drug-like queries. It might seem that for peptide queries, wherein a larger molecule can be folded into a compact shape, the relationship between through-bond and through-space distance would break down and 3D methods would always be better, but it should be noted that 2D methods can be modified easily to find non-peptide actives [58] given a peptide query.

Different methods find different actives

We find little evidence that 3D methods are better than 2D methods, at least by our two criteria. Certainly 2D methods are much more cost-effective. So should we ignore 3D methods, or other methods that do not appear to be doing as well? Generally, the answer is No. Although some methods do better than others on average, it is hard to predict how any method will do on a particular combination of query and activity. Some of our earlier work on descriptors [34] demonstrated, for example, that BP and BT descriptors were on average poorer than AP and TT, but did much better on some specific cases. Moreover, different methods generally rank active compounds differently and, therefore, tend to select different subsets of actives. This is a general phenomenon that can be seen between two topological methods (e.g. TOPOSIM versus LaSSI [39]), or even between two versions of the same method that differ only in the descriptors (e.g. TOPOSIM using APTT versus BPBT descriptors [34]). One way to highlight this is to show molecules that are preferentially selected by one of our methods. These are shown in Figure 3. (TV is omitted because it does too well relative to the others.) Despite the fact that some methods select fewer actives overall, clearly each method can find some actives that all other methods would miss. Some of the newer HIV protease inhibitors displace the conserved water molecule bound between the inhibitor and the 'flaps' of the protease because they have a hydrogen-bond acceptor opposite the required hydroxyl (e.g. 222392, 211975, 238403, 210723). We might naïvely expect that these would be found only by FLOG or TV, because the other methods use only indinavir as the query, and indinavir does not contain an acceptor in that position. However, these newer inhibitors are as likely to be found by TOPOSIM or LaSSI as by FLOG.

Why multiple methods are needed

We have come to regard looking for 'the best' way of searching chemical databases as a futile exercise. In both retrospective and prospective studies, different methods select different subsets of actives for the same biological activity and the same method might work better on some activities than others in a way that is difficult to predict beforehand. In retrospect, this makes sense because receptors are diverse, and chemical groups that appear equivalent to one descriptor might not be equivalent to another. For example, consider two receptors A and B. Receptor A will accept a carboxylate or a tetrazole at a particular site, receptor B accepts only the carboxylate. If we start with a query containing a carboxylate and a database containing carboxylates and tetrazoles, descriptors that distinguished carboxylate from tetrazole, perhaps on the basis of including element type, would appear to work better on activity B because it would score only the carboxylates highly. Descriptors that were more generic, perhaps ones that had the concept 'anion,' would appear to do better on activity A because it would score both carboxylates and tetrazoles highly.

This unpredictability from case to case has consequences for the validation of new similarity methods and the use of existing ones. There is an unfortunate tendency in the literature for investigators to invent a new method, compare it to some standard method (usually Daylight fingerprints) for one or two activities, and then claim global superiority if the new method does better at selecting actives. In particular, it is often implied that some 3D method is superior to 2D methods if it can beat Daylight [19,56]. We feel such claims are unjustified; one can always find one or two activities that will make a method look better (or worse) than the standard. Also, although Daylight is a good method, it too has its weak points [59] and can not be taken as representative of all 2D methods. We would suggest that investigators present at least 10 examples with diverse activities and that there be more than one standard of comparison. Investigators can take comfort in the idea that their new method might be useful, even if not better at finding actives than some standard, if it finds different actives than other methods. By contrast, because nearly every method identifies at least some unique actives, they cannot claim that their new method is special in that respect.

Any search method asks a certain question and makes a set of assumptions. Clearly, biological activity is more diverse and complicated than can be addressed by a single method or set of methods. It is typical practice at Merck, therefore, to use any available search method for which we have sufficient information to formulate a query, and to try several variations of the same method where possible. A large effort of combining the search results and filtering them through our own chemical intuition is necessary as well. It is as if we have a set of imperfect windows through which to view Nature. As computational scientists, we get nearer to the truth by looking through as many different windows as possible.

Aknowledgements

We thank the many developers that have created and validated in-house virtual screening methods at Merck: Brad Feuston, Eugene Fluder, Richard Hull, Michael Miller, Robert Nachbar, Susan Sallamack. Members of the Molecular Systems Applications Group, past and present, applied the methods in practical problems of drug-design. J. Chris Culberson helped us with the in-house implementation of Daylight.

References

1 Walters, W.P. *et al.* (1998) Virtual screening – an overview. *Drug Discov. Today* 3, 160–178
2 Waszkowycz, B. *et al.* (2001) Large-scale virtual screening for discovery leads in the postgenomic era. *IBM Systems J.* 40, 360–376
3 Miller, M.A. (2002) Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.* 1, 220–227
4 Johnson, M.A. and Maggiora, G.M. (1990) *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, New York
5 Downs, G.M. and Willett, P. (1995) Similarity searching in databases of chemical structures. *Rev. Comput. Chem.* 7, 1–66
6 Willett, P. *et al.* (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 38, 983–996
7 Bajorath, J. (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis and virtual screening. *J. Chem. Inf. Comput. Sci.* 41, 233–245
8 Cheng, A. *et al.* (2002) Computation of the physio-chemical properties and data mining of large molecular collections. *J. Comput. Chem.* 23, 172–183
9 Patterson, D.E. *et al.* (1996) Neighborhood behavior: a useful concept for validation of 'molecular diversity' descriptors. *J. Med. Chem.* 39, 3049–3059
10 Brown, R.D. and Martin, Y.C. (1996) Use of structure-activity data to compare clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* 36, 572–584
11 Brown, R.D. and Martin, Y.C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* 37, 1–9
12 Hagadone, T.R. (1992) Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.* 32, 515–521
13 Rarey, M. and Dixon, J.S. (1998) Feature trees: a new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Design* 12, 471–490
14 Wild, D.J. and Willett, P. (1996) Similarity searching in files of three-dimensional chemical structures. Alignment of molecular electrostatic potential fields with a genetic algorithm. *J. Chem. Inf. Comput. Sci.* 36, 159–167
15 Mestres, J. *et al.* (1997) MIMIC: a molecular field matching program. Exploiting applicability of molecular similarity approaches. *J. Comput. Chem.* 18, 934–954
16 Kearsley, S.K. and Smith, G.M. (1990) An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. *Tetrahedron Comput. Method.* 3, 615–633
17 Perry, N.C. and Geerestein, V.J. (1992) Database searching on the basis of three-dimensional molecular similarity using the SPERM program. *J. Chem. Inf. Comput. Sci.* 32, 607–616
18 Miller, M.D. *et al.* (1999) SQ: a program for rapidly producing pharmacologically relevant molecular superpositions. *J. Med. Chem.* 42, 1505–1514
19 Mount, J. *et al.* (1999) Icepick: a flexible surface-based system for molecular diversity. *J. Med. Chem.* 42, 60–66
20 Lemmen, C. and Lengauer, T. (2000) Computational methods for structural alignment of molecules. *J. Comput.-Aided Mol. Design* 14, 215–232
21 Jain, A.N. (2000) Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. *J. Comput.-Aided Mol. Design* 14, 199–213
22 Dixon, S.L. and Merz, K.M. (2001) One-dimensional molecular representations and similarity calculations: methodology and validation. *J. Med. Chem.* 44, 3795–3809
23 Schurr, J.H. *et al.* (1996) The coding of three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* 36, 334–344
24 Ginn, C.M.R. *et al.* (2000) Combination of molecular similarity using data fusion. *Perspect. Drug Discov. Des.* 20, 1–16
25 Menard, P.R. *et al.* (1998) Chemistry space metrics in diversity analysis, library design, and compound selection. *J. Chem. Inf. Sci.* 38, 1204–1213
26 Pearlman, R.S. and Smith, K.M. (1999) Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* 39, 28–35
27 Schnur, D. (1999) Design and diversity analysis of large combinatorial libraries using cell-based methods. *J. Chem. Inf. Comput. Sci.* 39, 36–45
28 Livingstone, D.J. (2000) The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* 40, 195–209
29 Daylight Chemical Information Systems INC/2740 Los altos, Suite #360, Mission Viejo, CA 92691, USA (http://www.daylight.com)
30 Heritage, T.W. and Lowis, D.R. (1999) Molecular hologram QSAR. In *Rational Drug Design: Novel Methodology and Practical Applications*, American Chemical Society Symposium Series 719, (Parrill, A.L. and Reddy, M.R., eds), pp. 212–225
31 McGregor, M.J. and Pallai, P.V. (1997) Clustering large databases of compounds: using the MDL 'keys' as structural descriptors. *J. Chem. Inf. Comput. Sci.* 37, 443–448
32 Pickett, S.D. *et al.* (1996) Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* 36, 1214–1223
33 Mason, J.S. *et al.* (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged structures. *J. Med. Chem.* 42, 3251–3264
34 Kearsley, S.K. *et al.* (1996) Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* 36, 118–127
35 Shemetulskis, N.E. *et al.* (1996) STIGMATA: an algorithm to determine structural commonalities in diverse subsets. *J. Chem. Inf. Comput. Sci.* 36, 862–871
36 Xue, L. and Bajorath, J. (2000) Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* 40, 801–809

37 Xue, L. *et al.* (2001) Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* 41, 746–753

38 Hull, R.D. *et al.* (2001) Latent semantic indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.* 44, 1177–1184

39 Hull, R.D. *et al.* (2001) Chemical similarity searches using latent semantic structural indexing (LaSSI) and comparison to TOPOSIM. *J. Med. Chem.* 44, 1185–1191

40 Singh, S.B. *et al.* (2001) Mining the chemical quarry with joint chemical probes: An application of latent semantic structure indexing (LaSSI) and TOPOSIM (Dice) to chemical database mining. *J. Med. Chem.* 44, 1564–1575

41 Sheridan, R.P. (2000) The centroid approximation for mixtures: calculating similarity and deriving structure-activity relationships. *J. Chem. Inf. Comput. Sci.* 40, 1456–1469

42 Charifson, P.S. *et al.* (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* 42, 5100–5109

43 Lewis, R.A. *et al.* (2000) Computer-aided molecular diversity analysis and combinatorial library design. In *Reviews in Computational Chemistry* 16, (Lipkowitz, K.B. and Boyd, D.B. eds), pp. 1–51

44 Edgar, S.J. *et al.* (2000) Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *J. Mol. Graph. Model.* 18, 343–357

45 MDL Drug Data report licensed by Molecular Design, San Leandro, CA, USA (http://www.mdli.com)

46 Carhart, R.E. *et al.* (1985) Atom pairs as molecular features in structure-activity studies: definition and application. *J. Chem. Inf. Comput. Sci.* 25, 64–73

47 Nilakantan, R. *et al.* (1987) Topological torsions: a new molecular descriptor for SAR applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* 27, 82–85

48 Bush, B.L. and Sheridan, R.P. (1993) PATTY: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* 33, 756–762

49 Kearsley, S.K. *et al.* (1994) Flexibases: a way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Design* 8, 565–582

50 Feuston, B.P. *et al.* (2001) Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. *J. Chem. Inf. Comput. Sci.* 41, 754–763

51 Sheridan, R.P. *et al.* (1994) Extending the trend vector: the trend matrix and sample-based partial least squares. *J. Comput.-Aided Mol. Design* 8, 323–340

52 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242

53 Miller, M.D. *et al.* (1994) FLOG: a system to select quasi-flexible ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Design* 8, 153–174

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

55 Schuffenhauer, A. *et al.* (2000) Similarity searching in files of three-dimensional chemical structues: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* 40, 295–307

56 Makara, G.M. (2001) Measuring molecular similarity and diversity: total pharmacophore diversity. *J. Med. Chem.* 44, 3563–3571

57 Andrews, K.M. and Cramer, R.D. (2000) Toward general methods of targeted library design: topomer shape similarity scoring with diverse structures as queries. *J. Med. Chem.* 43, 1723–1740

58 Sheridan, R.P. *et al.* (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.* 41, 1395–1406

59 Flower, D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* 38, 379–386

Contributions to *Monitor*

We welcome recommendations of papers for review within *Monitor*, in the fields of combinatorial chemistry, pharmacogenomics, pharmacoproteomics, bioinformatics, new therapeutic targets, high throughput screening, new drug delivery technologies and other promising lines of research.

Details of recent papers or those *in press* should be directed to Dr Debbie Tranter, Editor, *Drug Discovery Today*, Elsevier Science London, 84 Theobald's Road, London, UK WC1X 8RR. tel: +44 207 611 4132, fax: +44 207 611 4485, e-mail: deborah.tranter@elsevier.com

Contributions to *Profiles*

We welcome contributions for the *Profiles* series, which gives a commentary on promising lines of research, new technologies and progress in therapeutic areas. Articles should provide an accurate summary of the essential facts together with an expert commentary to provide a perspective. Brief outlines of proposed articles should be directed to the *Monitor* Editor (see below). Articles for publication in *Monitor* are subject to peer review and occasionally may be rejected or, as is more often the case, authors may be asked to revise their contribution. The *Monitor* Editor also reserves the right to edit articles after acceptance.

All suggestions or queries relating to *Monitor* should be addressed to Dr Debbie Tranter, Editor, *Drug Discovery Today*, Elsevier Science London, 84 Theobald's Road, London, UK WC1X 8RR. tel: +44 207 611 4132, fax: +44 207 611 4485, e-mail: deborah.tranter@elsevier.com