Defining Privileged Reagents Using Subsimilarity Comparison

Brett A. Tounge* and Charles H. Reynolds

Johnson & Johnson Pharmaceutical Research and Development, L.L.C., P.O. Box 776,
Welsh and McKean Roads, Spring House, Pennsylvania 19477-0776

Received April 30, 2004

We have developed a new method for assigning a drug-like score to reagents. This algorithm uses topological torsion (TT) 2D descriptors to compute the subsimilarity of any given reagent to a substructural element of any compound in the CMC. The utility of this approach is demonstrated by scoring a test set of reagents derived from the "Comprehensive Survey of Combinatorial Library Synthesis: 2000" (*J. Comb. Chem.*). R-groups were extracted from the most-active compounds found in each of the reviewed libraries, and the distribution of the subsimilarity scores for these monomers were compared to the ACD. This comparison showed a dramatic shift in the distribution of the JCC R-group subset toward higher subsimilarity scores in comparison to the entire ACD database. The ACD was also used to examine the relationship between molecular weight and various subsimilarity scoring algorithms. This analysis was used to derive a subsimilarity score that is less biased by molecular weight.

INTRODUCTION

There has been considerable interest in approaches to constrain compound or reagent selections to structural classes that are considered privileged, because they contain functional elements likely to confer biological activity or because they are otherwise "drug-like" in nature.[1−4] Of course there is no simple or universal definition for the term drug-like. Indeed, whether a compound is drug-like may depend on many factors completely external to the molecule, such as the target of interest. Nevertheless, this concept of drug-like molecules has gained wide acceptance in the pharmaceutical industry as a tool for restricting screening lists, reagents, or virtual libraries to chemistry that might be more likely to produce viable drug candidates. Possibly the best known and most widely used approach for assessing drug-likeness is the Lipinski rule of five which defines drug-like with respect to a range of physical properties.[5,6] Many other variations of this rule have been proposed which use an expanded set of properties or different cutoffs for properties ranges. Common examples of properties used to define drug-like molecules are molecular weight, logP, number of rotatable bonds, and the number of hydrogen bond donors or acceptors.[7,8] In principle, almost any structural property might be used in this approach. Indeed, we have described a scaling procedure that can be used quite generally with any molecular descriptor.[9] While the rule of five is used broadly to limit reagent inputs, or the resulting products, it must be kept in mind that the rule of five was originally derived as property limits for oral bioavailability.

A fundamentally different approach is to ignore any physical/chemical properties or descriptors and simply focus on the structures themselves. The idea is to identify structural motifs that convey the favorable property of drug-likeness without regard to the physical or mechanistic cause behind this effect. To pursue this approach one must have a reference state for drug-like and nondrug-like molecules. Since databases such as the Comprehensive Medicinal Chemistry (CMC) or World Drug Index (WDI) contain late stage drug candidates, they are often seen as drug-like sets.[10,11] For this reason, much work has been done to try to extract information from these databases to guide the synthesis of new compounds.[12−15] Several approaches have been pursued that attempt to identify key substructures that are most prevalent in pharmaceuticals. For example, the RECAP procedure published by Lewell et al. uses "chemical knowledge" to break molecules from the WDI into their building blocks.[16] Bemis and co-workers have taken a shaped based approach to analyzing the CMC and MACCS-II Drug Data Report (MDDR)[17] databases to identify common core and side-chain shapes.[18−20] One key feature of this type of approach is that it is not concerned solely with oral bioavailability but also encodes the concept of privileged structures. This approach will identify structural moieties that are more commonly found in whatever representative drug database is used as a reference, but it makes no distinction between the structural moieties role in potency, bioavailability, or any other drug property.

The goal of our work is also to provide guidance in selecting reagents that have privileged or drug-like structural elements. However, our method does not rely on breaking molecules down into their building blocks but instead uses a similarity measure that tests whether a given reagent is similar to any portion (subsimilarity) of known drug molecules. This subsimilarity measure determines the drug-like score of a given reagent. The method is similar in some respects to the approach published by Lewell and Smith.[21] In their work, reagents were first clustered, and then the cluster centroids were compared to the Standard Derwent File (SDF) using several different algorithms.[22] While they do use subsimilarity as one of the comparison tools, we use a different descriptor and a different formula for calculating

* Corresponding author phone: (215)628-5230; fax: (215)628-4985; e-mail: btounge@prdus.jnj.com.

10.1021/ci049854j CCC: $27.50 © 2004 American Chemical Society
Published on Web 07/20/2004

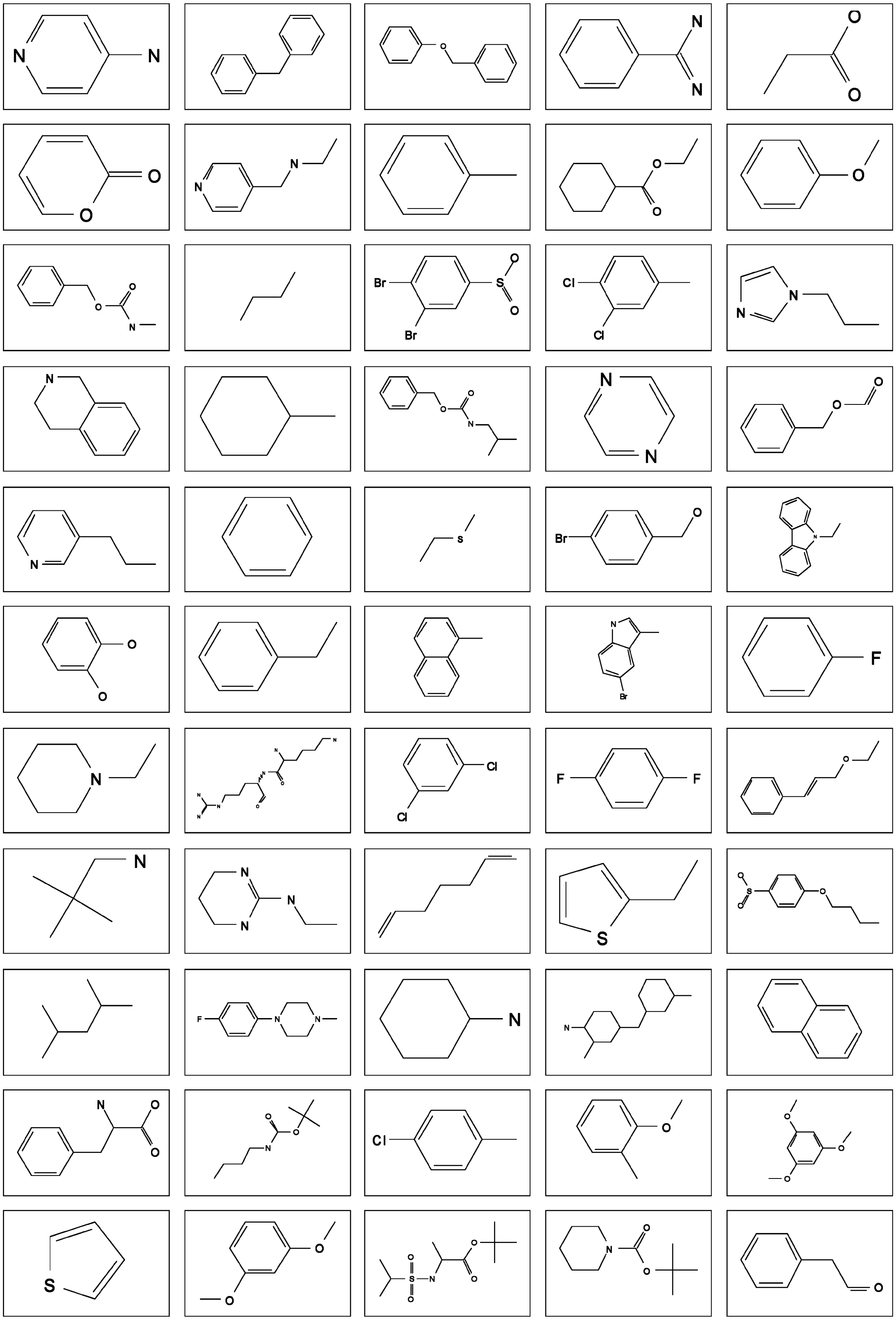

Figure 1. R-groups extracted from the “Comprehensive Survey of Combinatorial Library Synthesis: 2000”.

the subsimilarity score. In addition, we have focused on a method for applying a privileged structure score to individual reagents as opposed to clusters.

We have elected to take a subset of the CMC as the representative set of drug molecules. However, any set of compounds could be chosen for the reference set. The Available Chemicals Directory (ACD)[23] is taken as a source for reagents to test our methodology. The first step in developing this procedure was to examine how well our chosen descriptor (topological torsion) represented the differences in the reagent and drug libraries. This was done using several subsimilarity scoring algorithms. One of our key concerns was to adopt a scoring function that was relatively insensitive to molecular size. After settling on a descriptor and scoring algorithm, we then studied a set of reagents derived from the “Comprehensive Survey of Combinatorial Library Synthesis: 2000”[24] in order to assess the ability of our subsimilarity approach to differentiate the

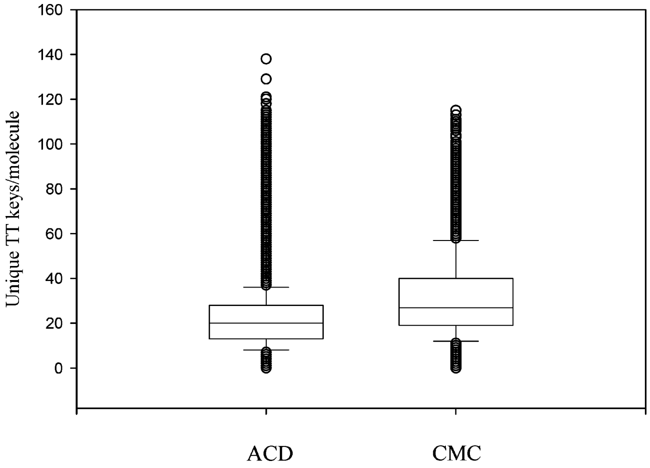

Figure 2. Box-plot of the unique TT keys per molecule for the CMC and ACD. The line in the box is the median of the data set, the box encompasses the 25th to 75th percentiles, and the whiskers (capped lines) represent the 90th and 10th percentiles.

reagents that produced active compounds in the reviewed libraries from the ACD as a whole.

METHODS

Descriptor. The molecules were represented using the topological torsion (TT) descriptor.[25] This is a 2D descriptor which encodes consecutively bonded non-hydrogen atom (sets of 4 in this case) sequences. For each atom, the atom type, number of π electron pairs, and the number of non-hydrogen attachments are recorded. For a given molecule all the unique TT codes were kept along with a count of the frequency of each TT. The generation of these descriptors was done using C++ code developed in-house by Xin Chen.[26] It is likely that other descriptors could be used, but it makes sense to use some type of connected graph descriptor given our intention of examining substructural elements of the CMC.

Databases. Entire ACD: The Available Chemicals Directory (ACD, release 2000.2)[23] was used as the source for the test reagents. The only filtering performed on this set was to remove compounds without structures.

Filtered ACD set: For generating this set, the first step was to extract compounds from the "Entire ACD" set that had any of the following functional groups: acetylene, acid chloride, acid, alcohol, aldehyde, aniline, aryl bromide, aryl chloride, aryl iodide, betaketoester, BOC amino acid, boronic acid, bromoketone, chloroketone, cyanate, FMOC amino acid, isocyanate, isothiocyanate, nitrile, primary amine, primary aromatic amine, secondary aliphatic amine, secondary aromatic amine, stannane, sulfonyl chloride, tertiary aliphatic amine, tertiary aromatic amine, or thioamide. The resulting set of 105 251 compounds ("Reagent ACD") was then filtered to remove outlier compounds. Compounds with AlogP ≥ 5.82, molecular weight ≥ 455, hydrogen bond donor count ≥ 3, and hydrogen bond acceptor count ≥ 9 were removed. This filter was derived from our previous analysis of the ACD. This "Filtered ACD" set contained 83 209 compounds.

CMC: The drug-like compounds were taken from the Comprehensive Medicinal Chemistry (CMC, release 2000.1) database.[11] All compounds in this database have been assigned United States Approved Names (USAN) indicating that they have likely entered at least Phase II testing. To ensure the compounds fit the drug-like criterion, all compounds without activity information and with the following activity classes were removed: adsorption promoters, aerosols, alcohol denaturants, anesthetics, antidotes, antifoams, antioxidants, antiperspirants, astringents, blood substitutes, biochemical reducing agents, buffering agents, bulking agents, chelating agents, contraceptives, cosmetics, dental, diagnostic agents/aids, dietary supplements, disinfectants, dyes, emollients, emulsifiers, food additives, insect repellents, insecticides, herbicides, imaging agents, metal complexes,

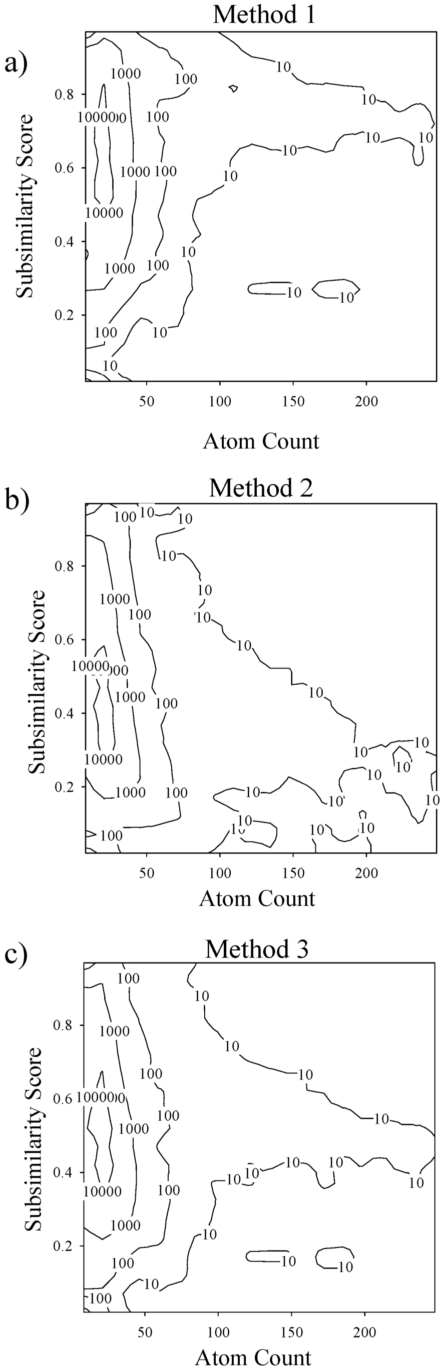

Figure 3. Histogram of the subsimilarity scores as a function of the atom count. The contour lines (*z*-axis) represent the number of molecules with a given atom count and subsimilarity score. Method 3 shows the least correlation between subsimilarity score and molecule size since the average score does not change as the atom count gets higher.

minerals, pharmaceutical aids/tools, pregnancy tests, photosensitizers, plant growth regulators, prosthetic aids, propellants, radio pharmaceuticals, resins, surfactants, sunscreens, surgical aids, sweeteners, topicals, ultraviolet light absorbers, vaccines, and vitamins. After filtering, 5813 of the original 7937 entries were kept.

Reagent test set: A test set of reagents that were known to result in active compounds was taken from the "Comprehensive Survey of Combinatorial Library Synthesis: 2000".[24] For each library that was reviewed, R-groups were taken from the most active compounds (Figure 1). The resulting 55 structures were used as probe molecules to extract reagents from the "Reagent ACD" (see above for description of this set). We collected at most 10 reagents per probe and required that the similarity was ≥ 0.66 (Tanimoto Similarity based on MDL public keys). This step was carried out to ensure that the reagent inputs for the test set were representative of the structures in the ACD. For example, when ranking the reagents from the ACD we are looking at the full monomer without cleaving any reacting groups. Using the cutoff of 0.66 for the similarity measure gives us a set of relatively close analogues of the initial R-groups. The final "active" reagent test set contained 379 compounds.

Subsimilarity Algorithm. For each probe molecule we calculate a subsimilarity to each compound in our drug-like database (CMC) and kept the highest subsimilarity score. The subsimilarity (SS) was computed using the following general formula

$$\mathrm{SS} = \frac{\text{total number TT in target that match TT in probe}}{\mathrm{TT_p}} \qquad (1)$$

where $\mathrm{TT_p}$ is the total number of unique TT codes in the probe molecule (i.e. the reagent). Three different criteria were considered for determining a match. In one case the simple existence of the same TT descriptor in both molecules was considered a match (method 1). This is akin to traditional binary fingerprints. In the second approach the TT descriptor and the count had to be the same to constitute a match (method 2). The final scoring algorithm was a hybrid of these two methods. In this case the existence of the same TT code resulted in a half-match (scored as 0.5). If the counts were also equal another 0.5 was added for this pair. So, a full match (score 1) resulted only from the TT code and count for that TT code being equal (method 3).

RESULTS

Descriptor Space. The first question one must address is whether the CMC truly represents a unique subset of space as compared to the ACD. This can be addressed through a closer investigation of the TT properties of the databases. Figure 2 contains a box-plot of the unique TT keys per molecule for the CMC and ACD. It is evident from this plot, not surprisingly, that the molecules in the CMC are on average more complex: i.e., they have more unique TT codes per molecule. However, if one compares the actual number of unique descriptors for the databases as a whole it can be seen that the space covered by the ACD is in total much larger. The ACD contains 22 373 unique TT codes, whereas the CMC contains only 4299. In addition, the CMC contains only 172 TT codes that are not found in the ACD. Taken together, this indicates that the CMC is effectively a subset of the ACD in terms of the TT descriptor space employed here.

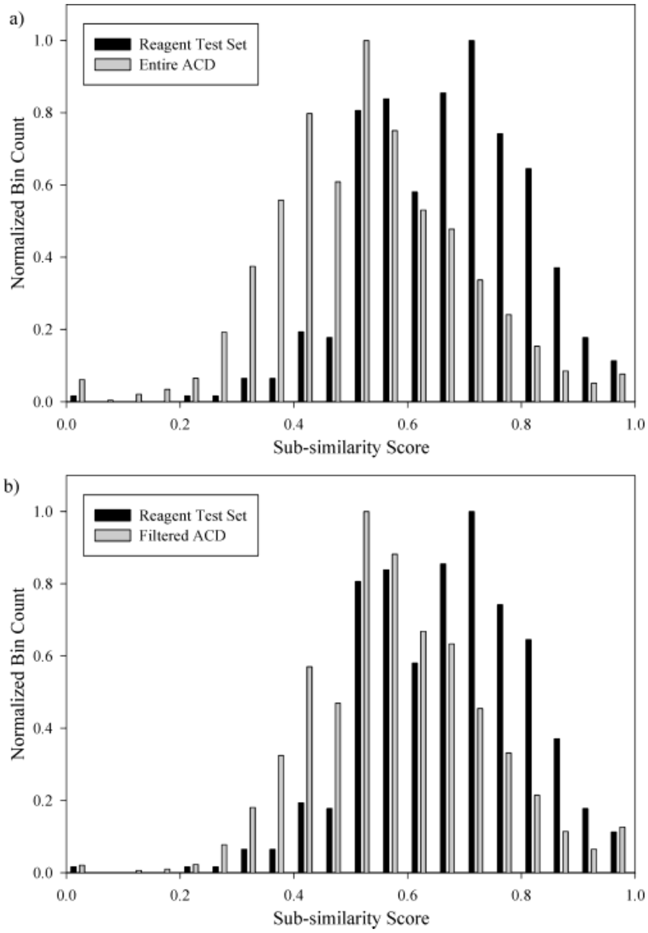

Figure 4. (a) Histograms of the subsimilarity score for the entire ACD (gray) compared to the reagent test set (black). (b) Histograms of the subsimilarity score for the filtered ACD (gray) compared to the reagent test set (black). The test set of active reagents is shifted to higher scores.

Subsimilarity Measure. As was outlined in the Methods section, three different variants of the subsimilarity score were investigated with the goal of probing the size bias. For each algorithm, the entire ACD was scored. Each of these methods results in a different size bias that is illustrated in Figure 3 where we have plotted a series of 3D histograms. The *Z* axis (contour lines) represents a count of the number of reagents that that have a given atom count (*X* axis) and subsimilarity score (*Y*-axis). For method 1 (Figure 3a) there is a shift to higher subsimilarity scores as the atom count goes up. In method 2 (Figure 3b) there is an opposite atom count dependence: the subsimilarity scores are shifted lower as the atom count increases. Given these two results, the intermediate scoring function was evaluated. This method (method 3) leads to a scoring function that is relatively insensitive to the atom count as can be seen in Figure 3c.

Reagent Test Sets. To test the ability of this scoring function to correctly rank drug-like reagents, we applied the method to a set of reagents derived from previously synthesized combinatorial libraries (see Methods section). Figure 4a shows histograms representing the distribution of subsimilarity scores for the ACD as a whole versus the subset

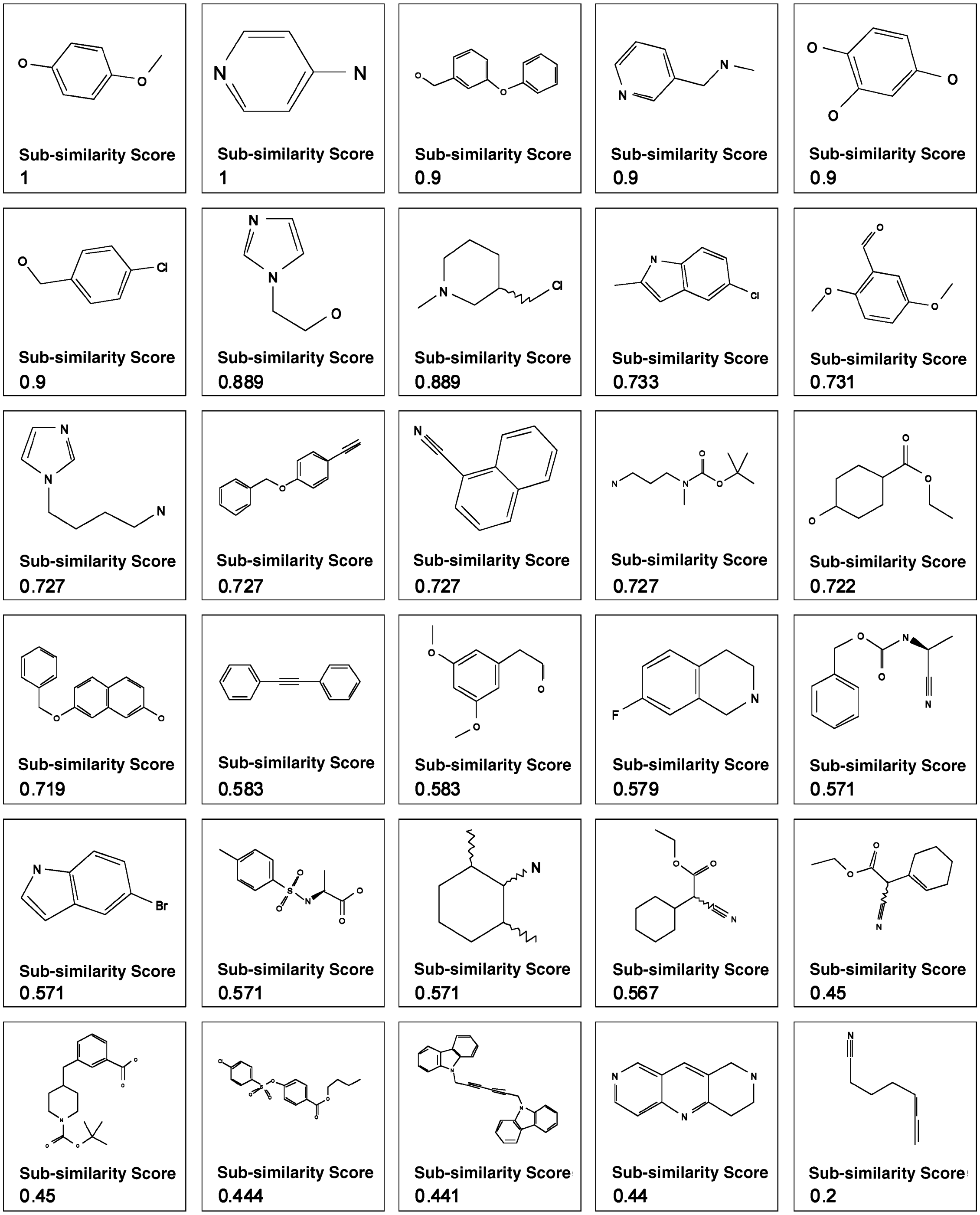
Sub-similarity Score 1
Sub-similarity Score 1
Sub-similarity Score 0.9
Sub-similarity Score 0.9
Sub-similarity Score 0.9
Sub-similarity Score 0.9
Sub-similarity Score 0.889
Sub-similarity Score 0.889
Sub-similarity Score 0.733
Sub-similarity Score 0.731
Sub-similarity Score 0.727
Sub-similarity Score 0.727
Sub-similarity Score 0.727
Sub-similarity Score 0.727
Sub-similarity Score 0.722
Sub-similarity Score 0.719
Sub-similarity Score 0.583
Sub-similarity Score 0.583
Sub-similarity Score 0.579
Sub-similarity Score 0.571
Sub-similarity Score 0.571
Sub-similarity Score 0.571
Sub-similarity Score 0.571
Sub-similarity Score 0.567
Sub-similarity Score 0.45
Sub-similarity Score 0.45
Sub-similarity Score 0.444
Sub-similarity Score 0.441
Sub-similarity Score 0.44
Sub-similarity Score 0.2

Figure 5. Examples of subsimilarity scores for monomers from the reagent test set.

of reagents from the active compounds. While the ACD as a whole has a distribution of scores centered on 0.5, the active reagent set is shifted to a much higher subsimilarity scores (i.e. ~0.8). Since the ACD contains many outliers compounds that would not be considered viable monomers, we also looked at a comparison of the filtered ACD verses the reagent test set (Figure 4b). Even in this filtered set, the reagent test set is well separated. Some examples of the subsimilarity scores for monomers in the reagent test set are shown in Figure 5.

When applied to our test set of reagents, this score differentiates monomers that lead to active compounds from the rest of the ACD. This represents a powerful, and very general, approach for selecting reagent inputs that are privileged for biological activity. In addition, this approach can be applied to any property of interest where a reasonably large reference set of compounds that represent that property can be identified. It makes no assumptions about the underlying physical properties responsible for that property and does not require any a priori definition of structural elements.

CONCLUSIONS

Using the CMC as our drug-like reference set, we have developed an algorithm that uses a subsimilarity metric to identify structural moieties that represent privileged (drug-like) structures. This approach provides a fast and intuitive algorithm for ranking reagents as inputs to medicinal chemistry, either automated (e.g. parallel synthesis) or traditional. It can even be employed in situations where no structural, QSAR, or pharmacophore models are available to aid in library design. Further, this approach is completely general and might be applied to any chemical property where a representative reference database of compounds can be defined.

As a test of the algorithm, we ranked a set of reagents that were known to produce biologically active products. For this set, we were able to demonstrate that the subsimilarity algorithm does in fact differentiate between the active reagents and the ACD as a whole.

ACKNOWLEDGMENT

The authors would like to thank Xin Chen for his work on developing the programs for generating the TT keys as well as for useful discussions and Steven Coats and Dennis Hlasta for their valuable input. We also acknowledge the support of the Johnson & Johnson Corporate Office of Science and Technology.

Supporting Information Available: A comma delimited file containing the reagent test set SMILES strings along with the subsimilarity scores is provided. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical substructures in drug discovery. *Drug Discov. Today* **2003**, *8*, 594−602.

(2) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.

(3) Walters, W. P.; Murcko, M. A. Prediction of "drug-likeness". *Adv. Drug Delivery Rev.* **2002**, *54*, 255−271.

(4) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science (Washington, D. C.)* **2004**, *303*, 1813−1818.

(5) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(6) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2001**, *44*, 235−249.

(7) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.

(8) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355−366.

(9) Tounge, B. A.; Pfahler, L. B.; Reynolds, C. H. Chemical Information Based Scaling of Molecular Descriptors: A Universal Chemical Scale for Library Design and Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 879−884.

(10) *World Drug Index*; Derwent Information; London.

(11) *Comprehensive Medicinal Chemistry*; MDL Information Systems, Inc.; San Leandro, CA.

(12) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H. et al. Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs. *J. Med. Chem.* **2004**, *47*, 224−232.

(13) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103−108.

(14) Schneider, G.; Lee, M.-L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487−494.

(15) Matter, H.; Baringhaus, K.-H.; Naumann, T.; Klabunde, T.; Pirard, B. Computational approaches towards the rational design of drug-like compound libraries. *Comb. Chem. High Throughput Screening* **2001**, *4*, 453−475.

(16) Lewell, X. Q.; Judd, D.; Watson, S.; Hann, M. RECAP−Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

(17) *MACCS-II Drug Data Report*; MDL Information Systems, Inc.: San Leandro, CA.

(18) Ajay; Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.

(19) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(20) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.

(21) Lewell, X. Q.; Smith, R. Drug-motif-based diverse monomer selection: Method and application in combinatorial chemistry. *J. Mol. Graph. Model.* **1997**, *15*, 43−48.

(22) *Standard Derwent File*; Derwent Information: 14 Great Queen Street, London.

(23) *Available Chemicals Directory*; MDL Information Systems, Inc.: San Leandro, CA.

(24) Dolle, R. E. Comprehensive Survey of Combinatorial Library Synthesis: 2000. *J. Comb. Chem.* **2001**, *3*, 477−517.

(25) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(26) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−1414.

CI049854J