Using Molecular Equivalence Numbers To Visually Explore Structural Features that Distinguish Chemical Libraries

Yong-Jin Xu and Mark Johnson*

Discovery Medicinal Chemistry, Pharmacia Corporation, St. Louis, Missouri 63167, and
Pannanugget Consulting, Kalamazoo, Michigan 49006

Received May 10, 2002

A molecular equivalence number (meqnum) classifies a molecule with respect to a class of structural features or topological shapes such as its cyclic system or its set of functional groups. Meqnums can be used to organize molecular structures into nonoverlapping, yet highly relatable classes. We illustrate the construction of some different types of meqnums and present via examples some methods of comparing diverse chemical libraries based on meqnums. In the examples we compare a library which is a random sample from the MDL Drug Data Report (MDDR) with a library which is a random sample from the Available Chemical Directory (ACD). In our analyses, we discover some interesting features of the topological shape of a molecule and its set of functional groups that are strongly linked with compounds occurring in the MDDR but not in the ACD. We also illustrate the utility of molecular equivalence indices in delineating the structural domain over which an SAR conclusion is valid.

INTRODUCTION

The rapid growth of the compound collections relevant to drug discovery has dramatically increased the need for better methods of organizing collections of molecular structures. We frequently need to cluster compounds with similar structural features so that the molecular properties within a compound cluster and across different clusters can be studied. For example, one may want to know what regions in chemical space are poorly represented in a compound collection, how two compound collections compare in that regard, or what regions are well represented by a data set underlying an SAR.

Two genre of methods broadly used for grouping and organizing the compounds in a data set or database are substructure searching and distance-based methods. Substructure searching is the most popular and powerful method in retrieving specific subsets of structures. However, the manual specification of each subset prohibits the automation of such analyses. Distance-based methods, on the other hand, are readily automated, but the high-dimensional spaces associated with these methodologies are difficult to visualize. Consequently, the analyses and the results can become quite unintuitive.

We approach this problem of organizing structures by linearly ordering them with respect to various classes of structural features. Each class of structural features defines a molecular equivalence relation[1] which partitions compounds into disjoint sets called equivalence classes. For example, the molecular equivalence relation may be defined with respect to the cyclic system of a compound so that the equivalence class containing a particular compound consists of all compounds having that same cyclic system. Alternatively, the molecular equivalence relation may be defined with respect to the functional groups of a compound. In this case, the equivalence class containing a particular compound might consist of all compounds having that same set of functional groups.

The equivalence classes that arise in the specification of a molecular equivalence relation differ in two fundamental ways from the clusters of compounds that arise in cluster analysis. First, the clusters from cluster analysis are always defined with respect to the particular set of compounds that are clustered. To illustrate, suppose two compounds A and B fall in the same cluster after clustering the set X of compounds. If compounds are now added to or subtracted from X to form a new set Y and then the compounds in set Y are clustered, one cannot guarantee that compounds A and B will lie in the same cluster. In contrast, the proposed equivalence classes are defined across all compounds with reference to the underlying notion of molecular equivalence. Thus, for example, if compounds A and B fall in the same cyclic-system equivalence class in one context, they will fall in that same cyclic-system equivalence class in all other contexts.

As we shall see, each equivalence class has associated with it a distinct labeled graph (with possibly multiple edges and loops) that constitutes a specification of that class. This gives rise to the second fundamental difference between clusters and equivalence classes. The molecular equivalence classes illustrated here are specifiable categories. For example, when the underlying notion of equivalence is that of two compounds having the same cyclic system, then one completely specifies one of its equivalence classes simply by selecting a member of the class and saying the class consists of those compounds with the same cyclic system as that selected member. As a labeled graph, this cyclic-system specification can be unambiguously and visually displayed. Frequently, as would often be the case for the cyclic system, there exists a well-defined chemical name for an underlying graph of

* Corresponding author phone: (616)349-7599; fax: (616)349-8680; e-mail: mark@pannanugget.com.

10.1021/ci025535l CCC: $22.00
Published on Web 06/08/2002

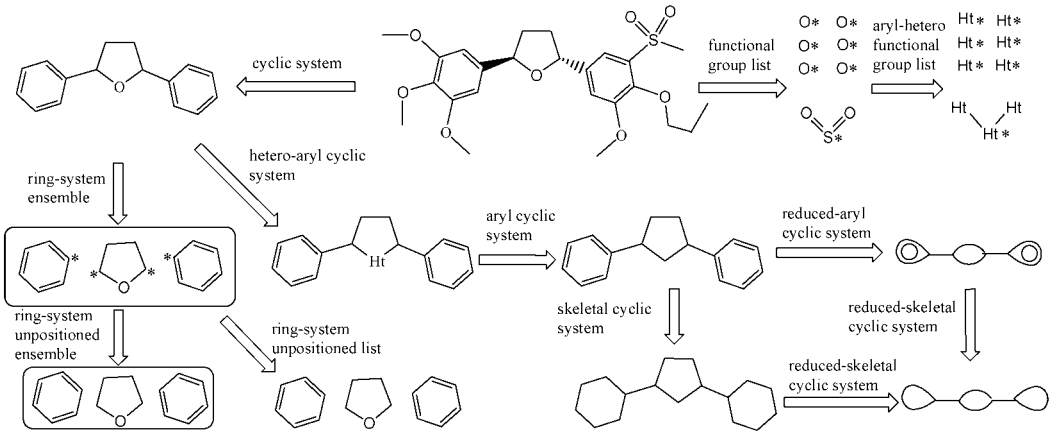

Figure 1. The hierarchical relationships from specific to general of some structural feature classes.

the equivalence class. Such names provide a familiar and facile way of speaking and thinking about the class. In contrast, specifying the commonality of compounds in a particular cluster reflects the chemical descriptors used to represent the structures, the measure used to quantify similarity or distance between two structures, and the method used to group the structures into clusters. Visual inspection may suggest important commonalties of the compounds in the cluster, but an exact specification of a cluster is seldom obvious. Because of these important distinctions, we feel that the notions of molecular equivalence and their associated molecular equivalence indices (MEQIs) presented here constitute a new genre of tools for cheminformatics and QSAR analysis.

There are many notions of molecular equivalence of possible interest in constructing MEQIs. This paper defines and illustrates the use of a wide range of MEQIs in the context of comparing two compound libraries. The libraries were constructed by taking random samples from the MDDR and ACD collections for a number of reasons. First, similar comparisons have interested others in specifying what makes a compound drug-like.[2] Second, both collections are somewhat diverse and thus represent possibly one of the more difficult contexts in which to find structural features that distinguish one collection from another. And third, our goal is to illustrate the use of MEQIs in visually comparing two libraries and to demonstrate that interesting and reliable conclusions can be drawn from that use even when those libraries represent rather small samples for two rather different and diverse libraries. For these purposes, it is helpful to have libraries that differ in a manner (drug-like vs nondrug-like) familiar and accessible to many readers.

That familiarity has a downside in that one could easily confuse our purpose with that of characterizing the drug-like and nondrug-like regions of chemical space. Such a characterization would require a comparison of much larger libraries than our two samples, a careful distillation from the ACD of those compounds that are or have been drugs and from the MDDR those compounds that have not been, and a much more systematic and extensive analysis. Properly addressing any of these three issues quickly goes beyond what is relevant to our purposes of presenting the concept of a MEQI and illustrating its use. However, there is a sense in which some useful information arises from this study that is relevant to the problem of distinguishing drug-like compounds from nondrug-like compounds. A random sample from the MDDR clearly contains a much higher proportion of drugs than does a random sample from the ACD. In that sense, the SAR conclusions that arise out of our analyses should be viewed as potentially relevant to the problem of distinguishing drug-like and nondrug-like regions of chemical space. This is especially true if one focuses on the form of the SAR conclusions, and in what manner these forms may differ from the forms of SAR conclusions drawn using other methods for comparing drug-like and nondrug-like libraries.

CONSTRUCTION OF A MOLECULAR EQUIVALENCE NUMBER

Equivalencing Functions. The construction of a meqnum consists of two steps. The first step is the extraction of the relevant structural feature that defines the equivalence class for the compound in question. This is illustrated for a number of structural features in Figure 1. Before defining the various types of structural features depicted there, a few general remarks are in order. In each case, the extracted structural feature constitutes a specification of a corresponding equivalence class of compounds. We restrict our attention to those extractable structural features that can be represented by a labeled pseudograph where a pseudograph[3] is a graph in which loops and multiple edges are allowed and is labeled if labels, such as atom types and bond types, are assigned to its vertices and edges. The reason for this restriction will become apparent in the next section.

The depiction of the compound in Figure 1 cannot constitute a depiction of an equivalence class satisfying our definition because it contains stereochemical distinctions that are not preserved in the labeled pseudograph naturally abstracted from that drawing. However, representing the atoms of the compound by the vertices of a graph suitably labeled by atom type and representing the bonds of the compound by the edges of that graph suitably labeled by bond type, we obtain the chemical graph of that compound. That chemical graph is a labeled pseudograph that specifies the important equivalence class containing that compound and all of its stereochemical isomers.

Because of valence constraints, we need not, and did not, depict the hydrogen atoms of the compound in Figure 1. Deleting the hydrogens from the chemical graph gives rise to the hydrogen-reduced chemical graph. Consequently, we can speak of the equivalence class of compounds having the same hydrogen-reduced chemical graph. The equivalence

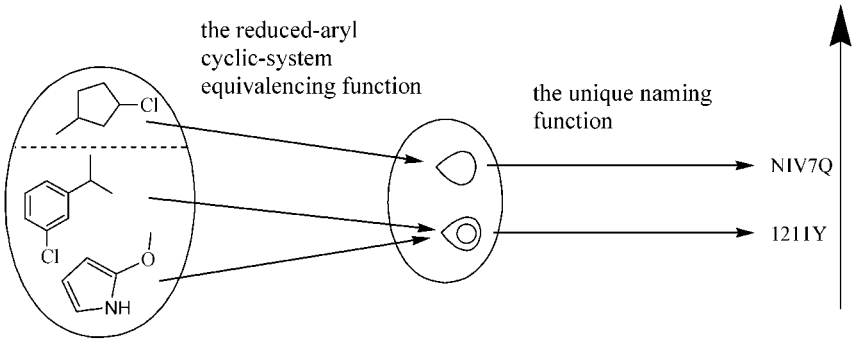

Figure 2. Illustrating the mapping of three compounds to their reduced-aryl cyclic system molecular equivalence numbers. The equivalencing function maps each compound to a labeled pseudograph and, thereby, groups compounds into equivalence classes marked off by the dashed line. The naming function maps each pseudograph to a meqnum on the line of lexicographically ordered strings.

classes associated with the hydrogen-reduced chemical graph are identical to those based on the chemical graph. Consequently, the two corresponding notions of molecular equivalence are interchangeable. (This does not mean that these two representations are equivalent. Substructure searching quickly reveals that they are not. But we will be interested in the equivalence classes themselves, and these are identical.) Because of its greater simplicity, we will restrict our interest to labeled pseudographs extracted from the hydrogen-reduced chemical graph of a compound, and when we refer to the chemical graph of a compound, we shall be referring to its hydrogen-reduced counterpart unless we specifically indicate otherwise.

The equivalence classes created by the chemical-graph relation are too fine for our purposes of finding structural features that distinguish one diverse library from another. Labeled pseudograph specifications of much larger equivalence classes are given in Figure 1 where the arrows indicate which associated structural features are derivable from another. For example, the aryl cyclic system, which ignores atom-type distinctions, can be derived from the hetero-aryl cyclic system, which takes some atom-type distinctions into account. However, the hetero-aryl cyclic system cannot be derived from the aryl cyclic system. Intuitively, it follows that each equivalence class of the hetero-aryl cyclic-system relation is a subset of a corresponding equivalence class of the aryl cyclic-system relation. Formally, we state this by saying that the former equivalence relation is finer than the later, and the latter equivalence relation is coarser than the former. The arrows in Figure 1 depict this hierarchical structuring.

The definition of the abstracted features that define the hetero-aryl cyclic system is almost self-evident from the example in Figure 1. We first abstract the chemical graph of the compound and then iteratively delete all vertices of degree 1, which deletes the side chains, to obtain the cyclic system of the compound. Then we convert all single, double, and triple bond types to single and leave the aromatic bond types as they are. We also convert all atom types to carbon, halogen, or hetero. (The words "single", "aromatic", "hetero", etc. are convenient labels for this discussion. For computation, much shorter bond labels such as "1" and "4" are used as are much shorter atom labels such as "C", "Hl", and "Ht".) In Figure 1, the heteroatoms are labeled Ht, and the aromatic bonds are either indicated by the alternating single and double bond convention or by inserting a circle in a ring as is done for the depiction pointed to by the reduced-aryl cyclic system arrow. The result is the labeled pseudograph following the hetero-aryl cyclic system arrow.

The equivalence class associated with that labeled pseudograph consists of all compounds for which the preceding derivation results in the depicted hetero-aryl cyclic system. If for a second compound that same derivation generates a different labeled pseudograph (i.e. the two labeled pseudographs are nonisomorphic in the graph-theoretic sense), then it lies in a different equivalence class. The variety of ways this derivation can be computed is incidental to the definition of the associated notion of molecular equivalence. However, the computation must be applicable to all compounds and must generate the same labeled pseudograph for all admissible ways of presenting the bonding information to that computation. Thus the derivation can be mathematically viewed as many-to-one mapping from the space of compounds to the space of labeled pseudographs.

As a mapping, this derivation induces an equivalence relation on the space of compounds where an equivalence class of this relation consists of all those compounds mapped to the same labeled pseudograph. This equivalence mapping or function is depicted schematically by the first set of arrows in Figure 2. (The second set of arrows in Figure 2 will be discussed later.) The two compounds in this figure whose cyclic systems consist of a simple aromatic ring are mapped by the reduced-aryl cyclic-system equivalencing function (yet to be defined) to the same labeled pseudograph consisting of a single loop labeled aromatic. The other compound whose cyclic system consists of a simple aliphatic ring is mapped to a different labeled pseudograph consisting of a single loop labeled single.

Returning to Figure 1, the derivation of the labeled pseudograph specification following the aryl cyclic-system arrow is simply obtained from that of the hetero-aryl cyclic-system derivation by changing all atom types to "carbon". If, in addition, we change all bond types to single, we obtain the labeled pseudograph following the skeletal cyclic-system arrow. The equivalence classes associated with the skeletal cyclic system equivalencing function, along with those of a number of other interesting equivalencing functions, have been extensively explored by Bemis and Murcko.[4]

Returning to the derivation of the loops in Figure 2 and the reduced-aryl and reduced-skeletal cyclic systems in Figure 1, Balaban, Filip, and Balaban[5] introduced the concept of a reducible vertex in computing the cycles in the chemical graph. We use a trivially modified version of their idea in which only vertices of degree 2 are reducible. In deriving

the bottom two labeled pseudographs in the right-hand column of Figure 1 from the counterparts at the heads the arrows leading to them, we iteratively delete the reducible vertices as follows: Let G_i denote the graph at step i. Let v be any reducible vertex of G_i and let uv and vw denote the two edges incident to v such that $u \neq v$ and $w \neq v$ (although it may be that u and w denote the same vertex). Now let the next graph G_{i+1} denote the graph obtained from G_i by deleting v, uv, and vw and adding the edge uw. One needs a convention for assigning an edge type to uv when edges uv and vw differ in their type. In our case the edge types are numbers 1, 2, 3, or 4 corresponding to whether the corresponding bond is single, "double", "triple", or aromatic. We simply assign to uv the higher of the edge types assigned to uv and vw. Continue this iterative process until a graph H is obtained that has no reducible vertices. The exact sequence in which the vertices are deleted is unimportant as the final graphs will all be isomorphic to each other.

Note that starting with a simple cycle such as the chemical graph of hexane, the next to the last step in this iterative reduction of vertices creates a graph with two vertices x and y connected by two edges. Both vertices of this graph are reducible. Let v in our reducing scheme denote the vertex with the lower ASCII-valued label. Assume this is vertex x. Then u and v necessarily both denote the vertex y. Deleting x and its two incident edges uv and uw which represent the two copies of the multiedge xy, we find that the added edge vw is simply the loop yy. Thus, our scheme terminates in the graph with a single vertex y having a single loop yy. In this way, starting with the aryl cyclic system of the compound in Figure 1, we obtain its reduced-aryl cyclic system, and starting with the benzene in Figure 2, we obtain the aryl loop in the middle of Figure 2. The reduced-skeletal cyclic-system in the bottom right corner of Figure 1 can be derived from the skeletal cyclic system in the same manner or can be derived from the reduced-aryl cyclic system of that figure by changing all edge types to single.

Salts are examples of structures stored in databases that have more than one component. We typically delete from such multicomponent structures all but the largest component as defined by the number of atoms in the component. Ties, in terms of the number of atoms for the largest component (a rare occurrence), are broken by selecting whichever member of that tie was first encountered. Consequently, all structures entering into our analysis have connected chemical graphs. As noted earlier, by iteratively deleting all single degree vertices, we are left with the cyclic system of a molecule that is necessarily a connected graph. A bridge of that cyclic system is any edge whose deletion results in a disconnected graph. If we delete all bridges along with any vertices that are incident to only bridges, we are left with the ring systems of a molecule which have proven useful in organizing chemical structures.[7] Such a feature group is given in the lower-left-hand corner of Figure 1. We can view it as a single graph with three components or as a list of three distinct graphs. For now we take the first view and refer to it as the ring-system unpositioned ensemble.

We qualify this ensemble as "unpositioned" because the ensemble gives no clue as to how the ring systems are positioned relative to each other in the molecule. By suitably distinguishing which atoms of the ring systems were incident to a bridge, we obtain the feature group called the ring-system ensemble. The distinguished atoms are asterisked in the ring-system ensemble in Figure 1. In our computations, they are distinguished by appending an "*" to the atom label. Among other things, this feature group implies that the originating compound was structured as a linear array of three rings with the furan ring in the middle because of its two asterisks.

It remains to define the feature groups involving the functional groups. There are many options for defining functional groups. We selected one that excluded the possibility that the resulting functional groups would share atoms.[1b] To do this, we partitioned the atoms of the chemical graph into two types: spacers and nonspacers. Keeping in mind that hydrogens have been excluded, all non-carbon atoms were put into the nonspacer group as were all carbon atoms connected by a double or aromatic bond to a heteroatom. The remaining carbon atoms formed the spacer group. The functional groups we are defining are the components of the chemical graph that remain after deleting all of the spacer atoms. We shall refer to them as maximal functional groups. Aspects of the positioning of a maximal functional group within the molecule are incorporated into the labeled pseudograph by appending an asterisk to any atom of that group that was incident to a spacer atom. This positioning information is part of the definition of a maximal functional group. Note that the carbons alpha to the nitrogen in pyridine are nonspacer atoms, while those alpha to the nitrogen in piperidine are spacer atoms. Consequently, the nitrogen functionality is represented in these to cases by C*∼N∼C* and N*, respectively, where the ∼ denotes a bond type of aromatic.

Figure 1 shows six functionalities related to the isolated oxygens and one related to the sulfone. Analogous to our representation of the nitrogen functionality in piperidine, the labeled pseudograph for each ether or hydroxyl group is a graph with a single vertex labeled O* and no edges, while the labeled pseudograph associated with the sulfone has a vertex labeled S* and two edges of type double incident to two other vertices both labeled O.

Our representation of the oxygen functionality obviously does not distinguish the differences between hydroxyl and ether groups. This could be overcome by using a more restrictive version of the hydrogen-reduced chemical graph that includes hydrogens bonded to heteroatoms, but we have not done so in this study.

A Unique Naming Function. We have restricted our attention to equivalence classes that can be specified by simply drawing their associated labeled pseudographs. Such visual specifications reflect the core of chemical nomenclature and understanding. However, for many purposes, simpler (although less informative) specifications suffice and are more efficient. To illustrate, suppose we wished to count how many cyclic systems were represented in a database. We could compute the cyclic-system graph for each compound and then compare that graph to all of the previously encountered cyclic system graphs. If it had not been previously encountered, we would increase our tally by one. But checking to see if two cyclic-system graphs are isomorphic is computationally expensive. These rich and intuitive structures are more useful for communication purposes. It would be much simpler to compute the cyclic system, assign it a short name, and then check if that cyclic-system name had been previously encountered.

Because our equivalence-class specifications share the mathematical structure of a labeled pseudograph, the problem of assigning unique labels to these classes reduces to the problem of assigning unique names to their associated labeled pseudographs. For example, in Figure 2 we show two compounds of the equivalence class of compounds whose cyclic system is a simple aromatic ring. As stated earlier, our labeled pseudograph specification of that equivalence class is the graph consisting of a single vertex and a single loop with a bond type of aromatic. In Figure 2 that labeled pseudograph is assigned the name 1211Y. The other labeled pseudograph having a single vertex and a single loop, but having a bond type of single, is assigned the name NIV7Q.

The two labeled pseudographs in Figure 2 are much more intuitive and informative specifications of their associated equivalence classes than their names. From these two pseudographs, I can quickly deduce that compounds in both equivalence classes have simple rings for cyclic-systems and that in one case these rings are aromatic and in the other they are not. *But I cannot order these labeled pseudographs*, and ordering things is implicit in almost any statistically based analysis operation. For example, plotting objects along an axis implies that the objects can be ordered. However, the names assigned to these two labeled pseudographs, and consequently assigned to their associated equivalence classes, can be lexicographically ordered. This is indicated by positioning those names along an arrow in Figure 2. We refer to these names as molecular equivalence numbers or simply meqnums because their interpretation is always with reference to the equivalence function specifying which compounds are mapped to the label pseudograph assigned that name.

Returning to the problem of counting cyclic systems, the introduction of a naming step in our computations potentially introduces two kinds of error. It may be that two different compounds have isomorphic but differently drawn or represented cyclic system graphs. Unless our naming algorithm is immune to such differences, it may assign different names to compounds in the same cyclic-system class and thereby inflate our tally. That is a type I error: isomorphic graphs assigned different names. These errors usually result from poor programming or rounding errors. Conversely, it may be that two compounds with nonisomorphic cyclic-system graphs have the same name assigned to those graphs and thereby deflates our tally. That is a type II error: nonisomorphic graphs assigned the same name. Reducing the type II error to 0 encounters NP-completeness issues associated with determining if two graphs are isomorphic.[7] We have only encountered difficulties with the type II errors, but have published a unique naming algorithm[8] that will be used in this study in which these difficulties have been resolved.

Besides issues of speed and error, one must consider the types of names that might be assigned. Short names increase computational speed, better utilize storage capacity, and are easier to remember. However, if they are too short, there will not be enough names to go around. We use a base 35 "license-plate" number consisting of the digits 0−9 and the capital letters excluding O (because of its visual similarity to 0). Consequently, we have 35^5 or 52 521 875 possible names of length 5. For practical purposes, the length of the name can vary with each equivalencing function. For example, there would seem to be vastly more equivalence classes when the structural features of interest are chemical graphs than when they are functional groups. Consequently, we use names of length 5 in the former case and of length 4 in the latter.

Construction of a Simple Molecular Equivalence Index. Regardless of the algorithms one uses to derive the labeled pseudograph specifications of the equivalence classes of interest and the algorithm one uses for naming these labeled pseudographs, the construction of a molecular equivalence index is always the same. For example, all essential elements of the construction of reduced-aryl cyclic-system index for a "library" of three compounds is given in Figure 2. The first step computes the relevant labeled pseudograph, in this case the reduced-aryl cyclic-system pseudograph. Although different algorithms can and will be used, they must arrive at isomorphic pseudographs when computed on the same compound. The second step is to compute the name for that pseudograph. Here again different naming algorithms can and will be used. Different algorithms will generally assign different names when computed on the same labeled pseudograph. That will not matter as long as there is a one-to-one correspondence between the assigned names. The names are simply tags for the various equivalence classes that are formed in step 1. As will become evident, any tagging scheme will work as long as it is one-to-one, much as nicknames work for people in place of full names or other personal identification schemes.

This combination of an equivalencing function with the naming function constitutes a simple molecular equivalence index (MEQI). The column of meqnums formed for a compound library can be entered into any graphics or statistical package with the facility for handling text strings in order to construct a histogram of its values. The ordering of the histogram bars will clearly depend on the naming algorithm employed, but the set of classes corresponding to those bars is invariant to the naming algorithm as are the heights of the bars displaying the class counts.

Construction of a Composite Molecular Equivalence Index. Returning to the labeled pseudograph depicted for the ring-system unpositioned ensemble in Figure 1, we see it has three components. When treated as a unitary object, it is assigned a single meqnum that, in our case, has length 5. Alternatively, we can construct a composite meqnum that lists the simple meqnums of the three component graphs. This is illustrated in Figure 3. Row 1 shows the three-component graph of the ring-system unpositioned ensemble in Figure 1 and its simple meqnum. Rows 2 and 3 show the simple meqnums of the component graphs. Rows 4 and 5 show the composite meqnum listing of the simple meqnums of the component graphs for two of the three possible orderings of the component graphs in row 1. Rows 6 shows the change in the simple meqnum when a cyclopropane is added to form a four-component graph, and rows 7 and 8 show the corresponding changes to the composite meqnums under two different orderings of the four components.

In constructing the composite meqnum, two types of errors can arise. Type I errors arise when isomorphic ring-system lists from two different compounds are assigned different composite meqnums. This happens when a canonical scheme for ordering the simple meqnums that comprise the list has not been defined. Such errors are avoided by resolving lexicographically any ties that occur as a result of any scheme

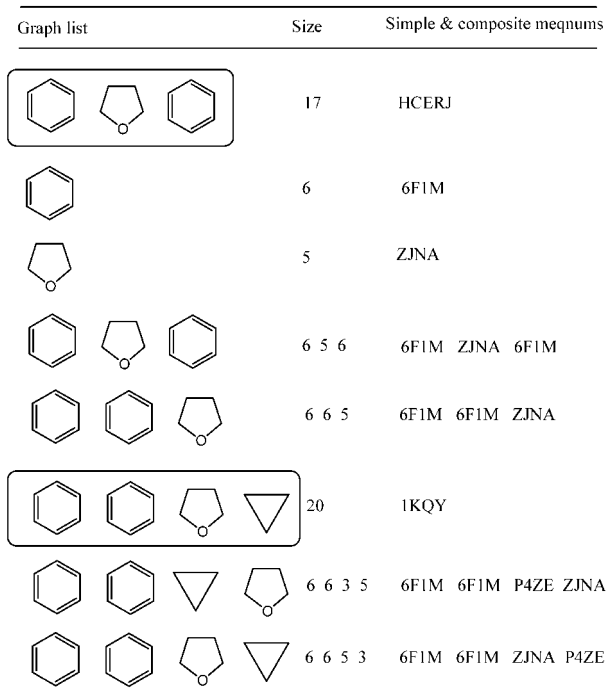

Figure 3. Constructing a composite meqnum. Row 1 shows a single graph of three components and the simple meqnum computed for that graph. Rows 2 and 3 show the simple meqnums of the component graphs. Rows 4 and 5 show the composite meqnums associated with two of the three possible orderings of the components in the first graph. Row 6 shows the simple meqnum when cyclopropane is added to the ring-system unpositioned ensemble. Rows 7 and 8 show the lexicographical and size-based orderings of the composite meqnums.

one may use for ordering the simple meqnums in a composite meqnum. Type II or missed opportunities errors arise when the ordering scheme poorly relates to the analysis issue. To illustrate, consider two cases: case 1 the component meqnums are ordered lexicographically and case 2 the component meqnums are first ordered inversely to the size of their associated graphs and then ties in that ordering are resolved lexicographically. Both ordering schemes will eliminate any type I errors. Now construct three hypothetical histograms H1, H2, and H3 where the *x*-axis for H1 is the simple MEQI, that for H2 is the lexicographically ordered composite MEQI, and that for H2 is the size and then lexicographically resolved composite MEQI. All three histograms reflect the same set of equivalence classes, and, since there are no type I errors, all will be comprised of the same set of histogram bars. *But these bars will be ordered differently*.

Consider the different placement of the bars associated with the three ring-system (TRS) and four-ring-system (FRS) ensembles in Figure 3. In H1, the FRS bar would be far to the left of the TRS bar as the 1 in the 1KQY meqnum precedes the H of the HCERJ meqnum by 16 of the 35 ASCII characters we use in their construction. In H2, the FRS bar would be just to the left of the TRS bar based on an agreement in the first nine characters the two lexicographically ordered composite codes. However, this is an accidental agreement. Had the cyclopropane meqnum preceded the benzene meqnum in the lexicographical ordering, the lexicographical composite meqnums for these two equivalence classes would have differed in their first character. In H3, the FRS bar would be just to the right of the TRS bar based on an agreement in the first 14 characters of their principally size-ordered composite codes. This agreement is not accidental. It reflects an a priori judgment that agreement in the nature of the larger ring systems of two compounds often leads to their having a number of similar properties. This problem of ordering a list of component meqnums has important similarities with notions associated with ordering keyword lists, some of which we have explored in the context of studying compound profiles.[9]

There is an important added analysis benefit in using these composite-meqnum specifications. Composite meqnums lend themselves to substring searching in a manner that complements substructure searching. For example, if we had a column of a table containing the composite meqnums associated with the ring-system unpositioned list we have been discussing, a substring search with "ZJNA" as the query string would return as hits all rows in our table of compounds having one or more furan ring systems.

Construction of a MEQI-Based Structural Browsing Index (SBI). One might ask the following: why one would ever use a simple-meqnum specification of an ensemble structural feature when the corresponding composite specification has these additional capabilities. The obvious answer is that when these additional capabilities are not needed, the simple-meqnum specification is shorter and more efficient. This often occurs in the construction of a structural browsing index (a fuzzy notion corresponding to any index that meaningfully orders a set of structures in an interpretable manner) from a list of MEQIs.

As we have just seen, composite-valued MEQIs can be designed to rationally order structures. However, composite-valued MEQIs have some drawbacks. In particular, one cannot know which types of equivalence classes will head the ordering. That basically reflects the first member of the composite meqnums and predicting that member's value is next to impossible. Starting from that vantage point, one might consider the simple chemical descriptor giving the number of atoms. Then one knows that the smallest compounds will head the ordering. However, we seldom consider only those compounds with a particular number of atoms (although ordering compounds by some version of their molecular formula is commonly done in compound catalogs). On the other hand, we often do consider only those compounds with a particular cyclic system. Consequently, it makes sense to order structures in a way organized around the concept of the cyclic system but somehow using counts to generally locate where one is in that structural ordering.

As a simple example of such an ordering, we might order the compounds first by the number of ring systems in the cyclic system, resolving ties in that count by the meqnums of the ring-system unpositioned ensembles, resolving ties in the ring-system unpositioned-ensemble meqnums by the meqnums of the cyclic-system, resolving ties in the cyclic-system meqnum by the meqnums of the chemical graphs, and finally resolving any remaining ties in the chemical-graph meqnums by the compound identifier. Ignoring the compound identifier, denote the first four variables by RSCnt, RSUEMeqnum, CSMeqnum, and CGMeqnum, respectively. For the compound in Figure 1, the values for these four variables are 3, HCERJ, HYE23, and LR7KS. We now form a new MEQI-based SBI, which we might denote by RSCnt||RSUEMeqnum||CSMeqnum||CGMeqnum by concatenating the values of the preceding four variables suitably

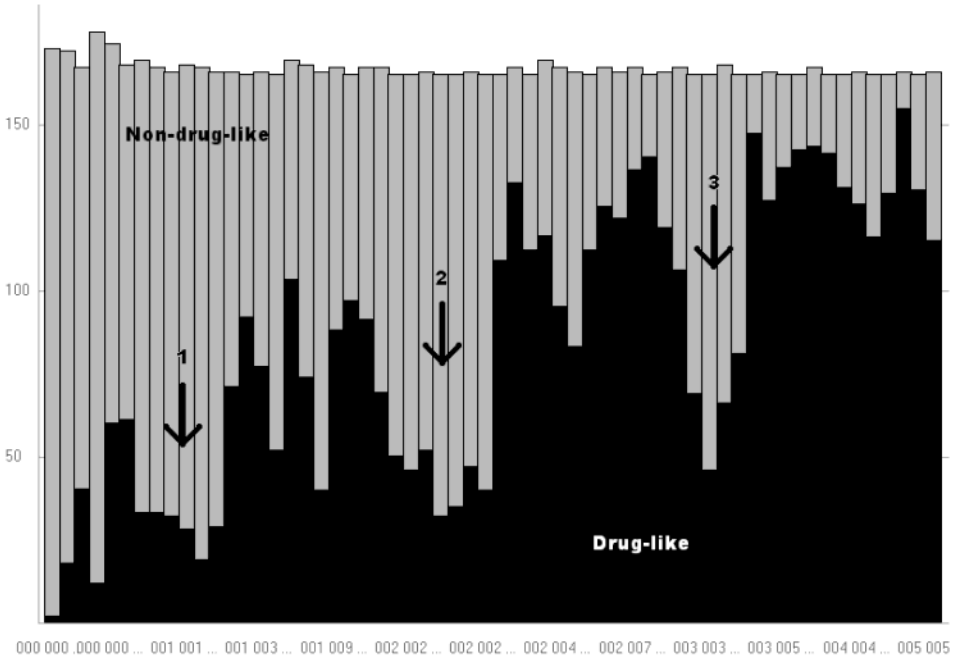

Figure 4. Histogram giving the compound counts for each value of the binned cyclic system fine ordering (CFSO). The binning intervals are adjusted to contain roughly the name number of compounds so that the height of the shaded portion corresponds roughly to the proportion of compounds coming from MDDR library.

separated by spaces. Its value in this case is "3 HCERJ HYE23 LR7KS". When we order the rows of a data table by lexicographical ordering of the values of this SBI, we accomplish the desired structural ordering. Note that replacing the simple meqnum HCERJ by its corresponding composite meqnum "6F1M 6F1M ZJNA" gains us little but greatly lengthens and complicates the value of the concatenated variable.

These MEQI-based SBIs are not new. The Beilstein Lawson Number[10] is a well-known structural ordering that effects such a hierarchical ordering via a concatenated list of structural codes. As the number of MEQIs increases along with counts based on their associated pseudographs, an almost limitless number of structural orderings emerges. We will be illustrating the use of one such SBI that we call the cyclic-system fine ordering (CSFO). It is a subjectively based and rather complicated affair involving the concatenation of 24 variables composed of 10 MEQIs and 14 MEQI-based counts. Most of the logic of this structural ordering has little to do with the scope and purpose of this paper. Its component MEQIs and MEQI-based counts that are necessary to the interpretation of the results and logic of our analysis will be detailed as they are needed.

USING MEQIS AND SBIS TO FIND STRUCTURAL FEATURES THAT DIFFERENTIATE TWO LIBRARIES

To illustrate the utility of MEQIs and SBIs in comparing two libraries, we selected two small compound sets consisting of 5000 compounds randomly picked from the MDL Drug Data Report database with over 90 000 compounds and 5000 compounds randomly picked from the Available Chemical Directory and Chemtrack database with over 250 000 compounds. As noted in the Introduction, we will loosely refer to a region in chemical space represented by a subgroup of compounds with a preponderance of its compounds coming from the MDDR as being drug-like and to a region in chemical space represented by a subgroup of compounds with a preponderance of its compounds coming from the ACD as being nondrug-like.

Using a MEQI-Based Structural Browsing Index. When studying how a categorical variable, such as drug-like or nondrug-like, defined on a set of compounds varies with some other structural classification of those compounds, it is natural to construct a histogram based on that structural classification and proportionally color the heights of the bars by that categorical variable. A structural class strongly associated with one of the values of that categorical variable will then be dominated by the color associated with that value. Such colored-histogram visualizations work well when the structural classes of the histogram contain similar numbers of compounds so that the heights of the bars are comparable. This is not the case for most MEQI-based SBIs as will become evident in the figures to follow.

We address this problem by marking off equal sized intervals along the structural ordering of interest so that the first k compounds in that ordering constitute the first class, the next k compounds constitute the second class, and so forth until there are k or fewer compounds from which to constitute the last class. Figure 4 presents a histogram based on such a binned-structural ordering constructed using the CSFO as the structural ordering with respect to which the equal-sized intervals were formed. Both the binning operation and histogram were constructed using the Spotfire[11] data visualization package version 3.3 which gives approximately, but sometimes not exactly, equal-sized bins. For publication reasons, we used the gray scale coloring capability in Spotfire rather than the full color scale. The proportion of the compounds in a class with the proportion coming from the more drug-like MDDR collection is colored black, and the remaining proportion coming from the less drug-like ACD collection is colored light gray.

Although we know little about MEQI and MEQI-based counts or how they were concatenated to form the CSFO, we can still make effective use of that histogram. Clearly, the proportion of MDDR increases as we move toward the right of the ordering. Moreover, there are clear departures from that trend. By systematically examining structures along the *x*-axis ordering, we would first encounter all acyclic structures, then structures with 1 ring system, followed by structures with 2 ring systems and so forth. By examining the structures in regions 1, 2, and 3, we find the following:

1. Compounds in region 1 have one aromatic six-membered ring.

2. Compounds in region 2 have two linked six-membered aromatic rings.

3. Compounds in region 3 have three five- or six-membered linearly linked aromatic rings, with at least two six-membered rings.

At this point, the CSFO browsing index has served its principal purpose—suggesting something to investigate further. Although the structure of the CSFO had much to do with the departures from the trend that caught our eye in Figure 4, our understanding of the structure of the CSFO played no part in the serendipitous discovery captured by statements 1−3. This is critical, because effective structural orderings can be quite complicated. If their effective use required a through knowledge of their construction, that use would be limited to experts in their formalism. As we have just seen, they can be effectively used as well by experts only in the scientific subject matter.

However, an intuitive sense for the basic rational of a MEQI-based SBI is helpful for interpreting trends and departures from those trends. Obviously, the overall rising trend in the MDDR proportions must reflect the values of the first variable in the concatenation of variables that resulted in the CSFO: the count of the number of ring systems. This count is indicated in the *x*-axis labels which are of the following form: "00x 00y ...". Here the x stands for the number of ring systems. For most of the compounds, this is a single digit number. Had it been a two or three digit number, there would have been one or no, respectively, leading zeros. Using leading spaces, we get a lexicographical ordering to correctly reflect a numerical ordering.

In Figure 4, roughly the first decile of CSFO values begins with 000. They represent acyclic compounds. Roughly, the next two deciles begin with 001 and represent those compounds whose cyclic system consists of a single ring system. Roughly the next four deciles begin with 002 and represent compounds with 2 ring systems. Then come the deciles reflecting 3, 4, 5, or more ring systems. This confirms our earlier observation that the proportion of MDDR compounds tends to increase as the number of ring system increases. What this might mean with regard to drug-likeness is difficult to say. The number of ring systems is a very crude correlate of a compound's size and complexity which may have something to do with drug-likeness, or it may reflect nothing more that the fact that compounds from the ACD are often used as components for making larger compounds.

Again, as we have just seen, departures from a general trend often turn out more interesting than the trend it self. We will be following up on these departures, but for the moment, it is informative to see what they tell us about the rationale underlying the CSFO. Note first that observation 3 is closely akin to a verbal specification of an equivalence class of the reduced-aryl cyclic-system (RACS) equivalence relation indicated in Figure 1. Each RACS class consists of those compounds whose cyclic system is mapped to the same reduced-aryl cyclic-system pseudograph. Just by visual browsing of the structures associated with the different bars of Figure 4, one finds that all those in region 3 have the same reduced-aryl cyclic-system pseudograph and that those outside of this region do not. Similar statements hold for regions 1 and 2.

To see what this implies concerning the CSFO, we shall say equivalence relation R is finer than equivalence relation S if every equivalence class of S is a subset of an equivalence class of R, and, conversely, we say R is coarser than S. Clearly the cyclic-system equivalence relation is finer than the ring-system count equivalence relation because if two compounds have the same cyclic system, they necessarily have the same set of ring systems. However, the converse need not be true. By the same argument, the cyclic-system equivalence is finer than the reduced-aryl cyclic-system (RACS) equivalence relation indicated in Figure 1. Had the cyclic-system MEQI preceded the RACS MEQI in the CSFO, compounds with three six-membered aromatic rings linearly linked would have been in a different region in Figure 4 than those with two six-membered aromatic rings and one five-membered aromatic ring linearly-linked. Thus, the RACS MEQI necessarily precedes the cyclic-system MEQI in our ordering.

We can say more. Suppose the first variable in our CSFO was the count of the number of atoms. This would have sent fragments of the interval of compounds defined by observation 3 all along the axis of the figure. However, the ordering of the compounds implied by observation 3 would be left untouched by any MEQI preceding the RACS MEQI in a cyclic-system ordering if that MEQI's equivalence relation is coarser than that of the RACS MEQI. Our CSFO satisfies this hierarchical constraint as would most other hierarchically constructed cyclic-system orderings.

Using a Simple Molecular-Equivalence Index. As we have seen, the concatenated structure of the CSFO made it a useful tool in suggesting structural features that may be important in differentiating two libraries. The structural ordering revealed both a trend and departures from that trend that suggested interesting subgroups to analyze. However, this concatenated structure hinders our exactly delineating the nature of these subgroups because of the difficulty in deciding where one group of compounds with an interesting defining structural feature ends and another begins. Although a simple MEQI lacks the ordering capabilities needed for a general SBI, it gives an exact demarcation of such information. Consequently, simple MEQIs can be very useful for further delineating the nature of an SAR suggested by a general SBI.

In examining the importance of the types and arrangements of the ring systems in the cyclic system, we turn first to the reduced-skeletal cyclic-system (RSCS) MEQI. Figure 5 presents a variation of a proportionally colored histogram using the RSCS MEQI as the classification variable that splits the colored bars into separate but vertically juxtaposed subbars. Thus in Figure 5 we see the histogram for MDDR compounds juxtaposed above that for the ACD compounds in such a manner that the MDDR bar for a particular RSCS

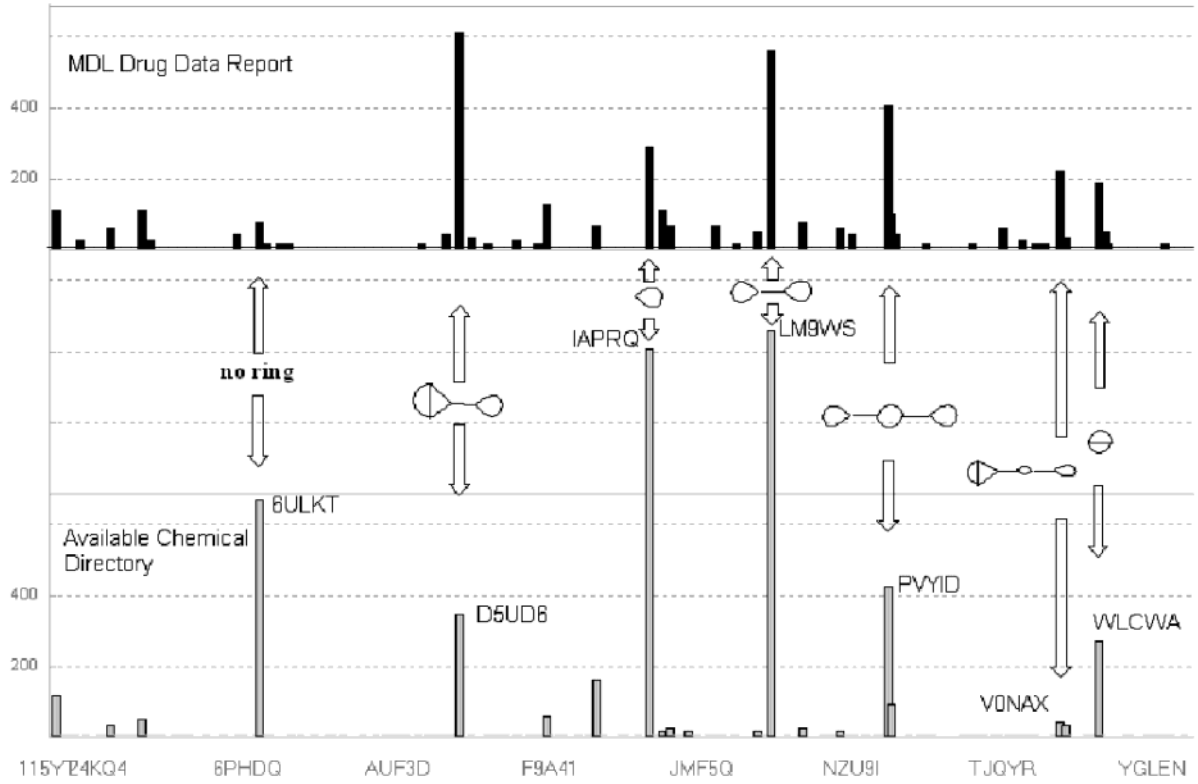

Figure 5. Vertically juxtaposed histograms giving the compound counts for each value of the reduced-skeletal cyclic system MEQI.

equivalence class is directly above the corresponding ACD bar. The meqnum specifications for a few of the RSCS equivalence classes are displayed along the bottom of the *x*-axis. Selected bars are annotated by their associated meqnums.

Not unexpectedly, the heights of the histogram bars for this simple MEQI show extreme variation. In fact, the *x*-axis in the upper MDDR histogram is darkened and thickened because there are so many RSCS equivalence classes represented by only one compound. A similar statement holds for the lower ACD histogram except the histogram bars of height 1 are less dense and more lightly shaded. In contrast, some of the classes contain more than 500 compounds. The acyclic compound class, whose RSCS MEQI value is 6ULKT, is one of these. As was suggested in our analysis of Figure 4, it is associated with a region of nondrug-like chemical space. The labeled pseudograph specifications of a number of other large RSCS classes are also given in the figure. These RSCS pseudographs are not deducible from their RSCS MEQI values but can be easily deduced from knowledge of the structure of any of the member compounds.

Those labeled pseudograph specifications associated with regions 1, 2, and 3 in Figure 4 have RSCS MEQI values of IAPRQ, LM9WS, and PVYID, respectively. But notice that these classes are not as nondrug-like as these three regions in Figure 4 would seem to suggest, especially the class PVYID containing those compounds with three simple rings linearly arranged. This leads one to suspect that compounds so structured but with either one or more nonaromatic rings might be more drug-like than those with all aromatic rings. A natural next step is to further resolve these three classes into their reduced-aryl cyclic system (RACS) subclasses. These comparisons are displayed in Figure 6.

Figure 6 is a juxtaposition of three histograms colored to show the proportion of MDDR compounds. Because of the extreme differences in the heights of the histogram bars, a corresponding pie chart is displayed above to show the MDDR proportions for the shorter bars. The lower histogram contains the two RACS subclasses of the RSCS IAPRQ class, i.e., those compounds whose cyclic system is either a simple aliphatic ring or a simple aromatic ring. The middle histogram shows the three RACS subclasses of the RSCS LM9WS class, and the upper histogram shows the six RACS subclasses of the RSCS PVYID class.

Neither of the lower group of subclasses is drug-like. This reflects the general trend in Figure 4. However, of the two, the aromatic subclass is the least drug-like. This subclass corresponds to region 1 in Figure 4. For the middle group, the nonaromatic subclass tends to be more drug-like than nondrug-like, and, again, the most aromatic subclass is the least drug-like. This latter subclass corresponds to region 2 of Figure 4. This trend continues for the upper group involving three linearly arranged rings. Those compounds with either 2 or 3 nonaromatic rings are decidedly drug-like, those compounds with 2 aromatic rings are less so, and those with 3 aromatic rings are decidedly nondrug-like. This latter RACS subclass forms region 3 of Figure 4. Thus we see a trend, that strengthens with the number of simple ring systems, of drug-likeness being inversely related to aromaticity. We also see the general trend in Figure 4 of the proportion of MDDR compounds increasing with the number of ring systems corroborated within the three subclasses of compounds on the left of Figure 6 having no aromatic rings.

Taken together, Figures 4−6 well illustrate the complementary use of a highly structured SBI and one or more of its component MEQIs. The former can often suggest which of its component MEQIs should be examined in greater detail

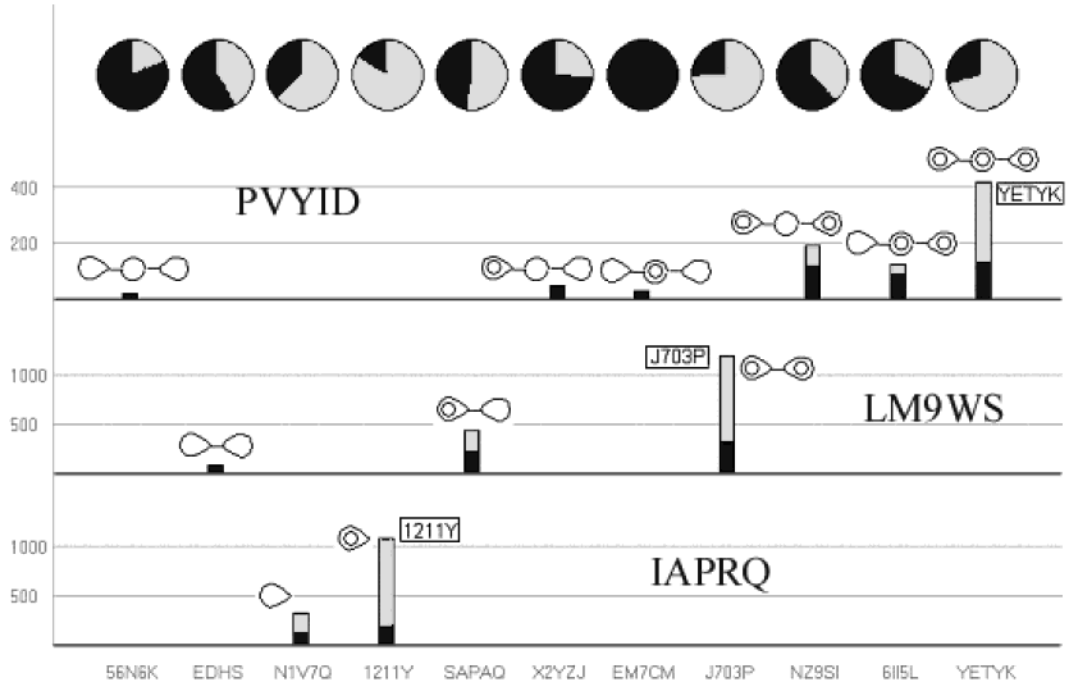

Figure 6. Vertically juxtaposed histograms giving the compound counts for each value of the reduced-aryl cyclic systems with the proportion of MDDR compounds shaded in black. The lower, middle, and upper histograms correspond to the sublibraries having IAPRQ, LM9WS, and PVYID for their respective reduced-skeletal cyclic systems. Position along positions on the *x*-axis are first ordered by a number of aromatic rings and then by the lexicographical ordering of their meqnums. The vertically juxtaposed pie charts indicate the MDDR proportions for the corresponding compound classes.

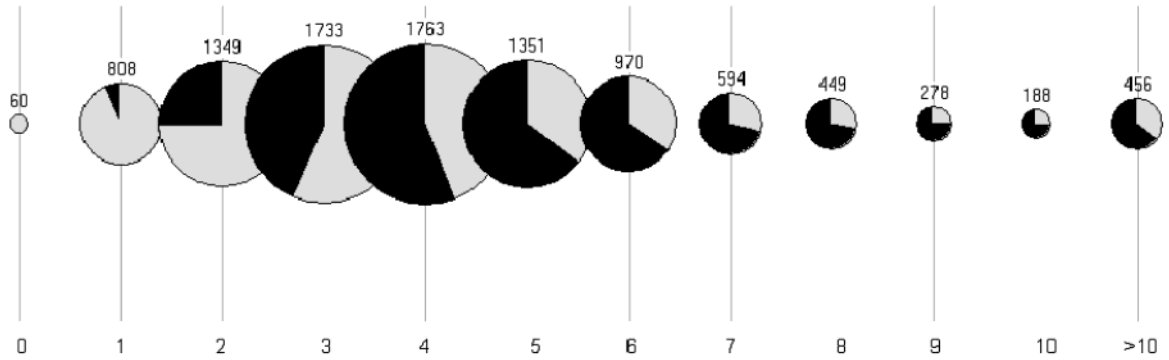

Figure 7. Pie charts broken down by functional-group count and giving the proportion of MDDR compounds. The size and annotation of the pie chart indicates the number of compounds with a particular functional group count.

and the latter can often be used to flesh out the structural features that gave rise to these suggestions.

Using a Composite Molecular-Equivalence Index. Besides the cyclic system, which defines the framework of the molecular shape, the functional groups, which incorporate many of the binding forces, can be expected to play a part in characterizing the structural commonalties and differences in two libraries. For example, the pie-chart array in Figure 7 shows a rising trend between the maximal functional-group count and the proportion of MDDR compounds. But when the functional group count exceeds 10, drug-likeness starts to decrease. This finding is consistent with the more detailed findings regarding drug-likeness that have been incorporated into the Lipinski rules.[12]

Trends such as that seen in Figure 7, unless further restricted, are applicable across the whole range of compounds over which the defining indices are defined. Because a simple MEQI is a purely categorical variable whose values have no inherent scientific ordering, any SAR patterns defined with respect to them are categorical in nature in that all compounds within a category are predicted to behave in the same manner and no prediction is made for those compounds falling outside of that category. A composite MEQI lies somewhere in between. It has two distinct notions of ordering potentially relevant to an SAR pattern. We have already discussed notions of order arising out of the various ways one might permute the component meqnums. The second ordering notion is embodied in the number of component values in the composite meqnum. We shall now illustrate how both of these ordering notions come into play when comparing the MDDR and ACD libraries.

Figure 8 is the histogram based on the aryl-hetero functional-group (AHFG) MEQI, again proportionally colored by library. Again there are a large number of categories whose associated histogram bars vary immensely in height. As the null graph (no vertices or edges) is assigned 0 by our naming algorithm, those compounds with no maximal functional groups (i.e. all atom types of the hydrogen-reduced chemical graph are C) are assigned a composite meqnum value of 0. The bar for this category is labeled 1, and, as

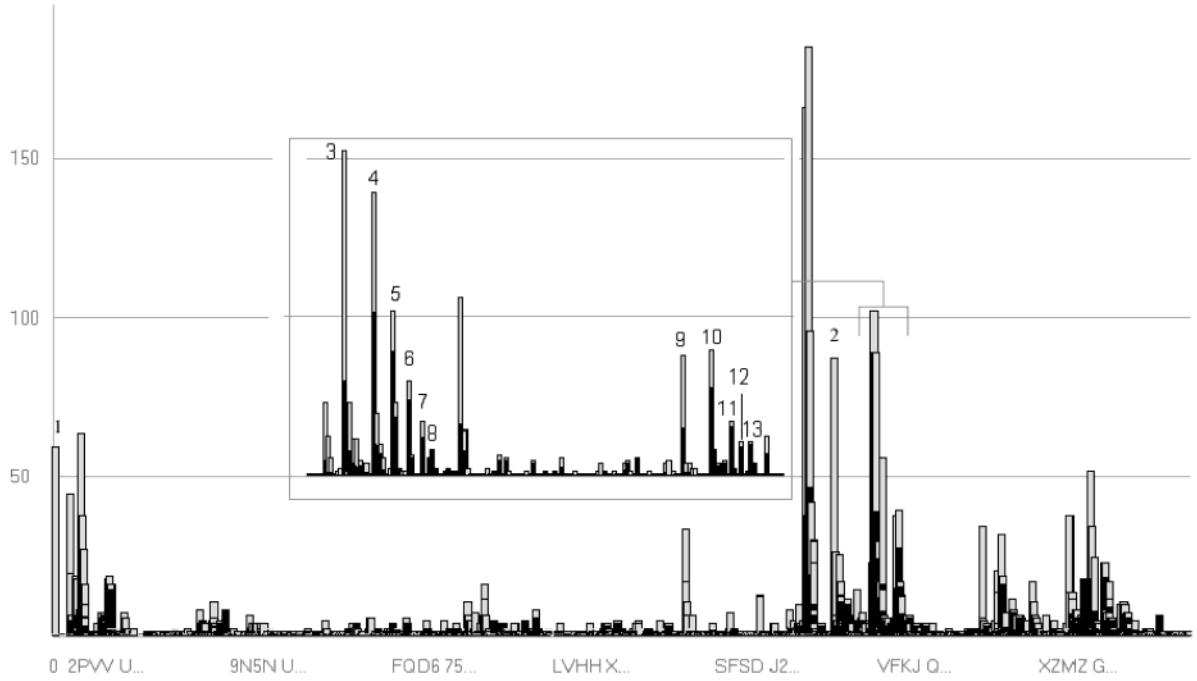

Figure 8. Histogram giving the count of the number of compounds for each value of the composite functional-group aryl-hetero MEQI. The shaded portion corresponds to the corresponding number coming from the MDDR library.

Table 1. Selected Aryl-Hetero Functional Group (AHFG) Classes Suggested by Figure 8

Class No.	Compound count	MDDR%	AHFG composite meqnum	meqnum pseudograph
1	60	0	0	
2	87	12.6	VE9C	
3	102	29.4	VE9C UKR6	
4	86	57.3	VE9C UKR6 UKR6	UKR6 Ht*
5	52	75	VE9C UKR6 UKR6 UKR6	
6	30	80	VE9C UKR6 UKR6 UKR6 UKR6	
7	17	70.6	VE9C UKR6 UKR6 UKR6 UKR6 UKR6	
8	9	100	VE9C UKR6 UKR6 UKR6 UKR6 UKR6 UKR6	
9	38	39.5	VE9C VE9C UKR6	
10	40	70	VE9C VE9C UKR6 UKR6	
11	17	88.2	VE9C VE9C UKR6 UKR6 UKR6	VE9C Ht* * Ht
12	11	81.8	VE9C VE9C UKR6 UKR6 UKR6 UKR6	
13	11	90.9	VE9C VE9C UKR6 UKR6 UKR6 UKR6 UKR6	

Figure 8 and Table 1 indicate, there are 60 compounds in this equivalence class. All are ACD compounds.

With the histogram bars so densely placed, one must zoom in on regions to see what is going on. The inset in Figure 8 is one such blowup. The locally relative height and coloring trends in the subsequences of annotated bars 3−8 and 9−13 caught our attention. Their composite meqnum meqnums and library proportionalities are given in Table 1 along with the two types of labeled-pseudograph specifications giving rise to the component meqnums.

Note that the VE9C component meqnums arise from a graph with three vertices and, consequently, always precede the UKR6 meqnums in our primarily size-based ordering of the component meqnums. Moreover, the composite meqnums are ordered lexicographically in Table 1 in necessary agreement with their relative ordering in Figure 8. This relative ordering is indicated with the class number in Table 1 and annotated on the corresponding bar in Figure 8. The presence in Figure 8 of histogram bars intervening between the bars associated with adjacent rows in Table 1 indicates the existence of an AHFG equivalence classes containing compounds with VE9C as one of its largest maximal-functional groups but also having maximal functional groups with meqnums other than UKR6 and VE9C. Finally, it might be noted that any sulfide, alcohol, ether, or amine (primary, secondary, or tertiary) would have UKR6 as one of its AHFG component meqnums, and any amide or ester would have VE9C as one of its AHFG component meqnums. However, no ketones, sulfoxides, sulfones, or carboxylic acids will be found in any of these classes for those functional groups would generate AHFG component meqnums other than VE9C and UKR6.

The results in Table 1 provide a strong affirmation of the initial rise in the MDDR proportion with the number of functional groups. For classes 2−8, which contain only one VE9C component meqnum, there is an obvious increase in that proportion with the number of UKR6 groups. A similar statement can be made for the second group of classes, 9−13, containing two VE9C component meqnums. When, as in Figure 4, a general trend across a diverse set of compounds is affirmed to hold for a more circumscribed subset of those compounds, the trend often becomes more pronounced. For example, in Table 1 compounds with 5, 6, and 7 functional groups show MDDR percentages of 83, 75, and 95, significantly higher than that seen for comparable functional group counts in Figure 4.

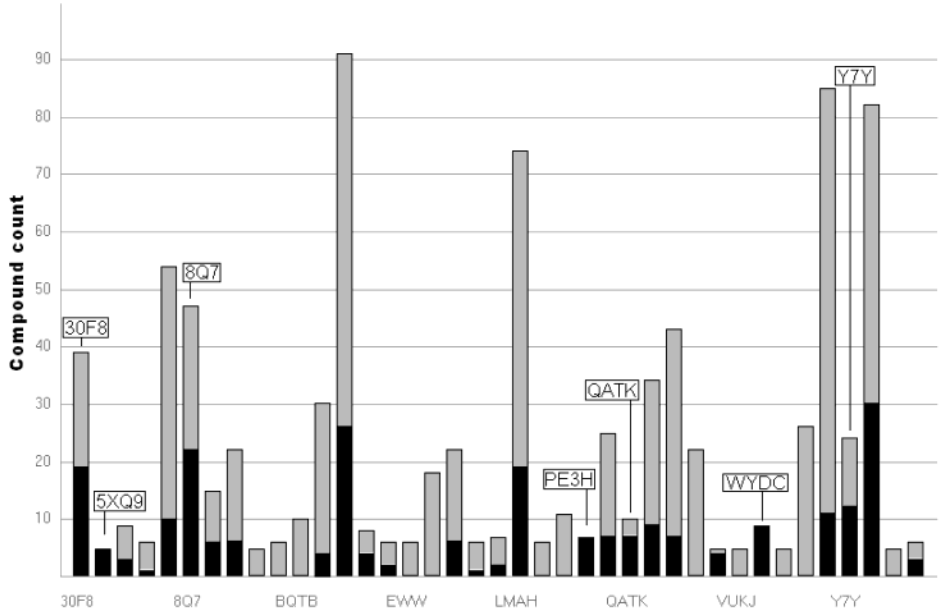

Figure 9. Histogram giving the number of compounds for each value for the largest maximal-functional group MEQI of those compounds having J703P (see Figure 6) for a reduced-aryl cyclic-system MEQI. Only classes with five or more compounds are shown.

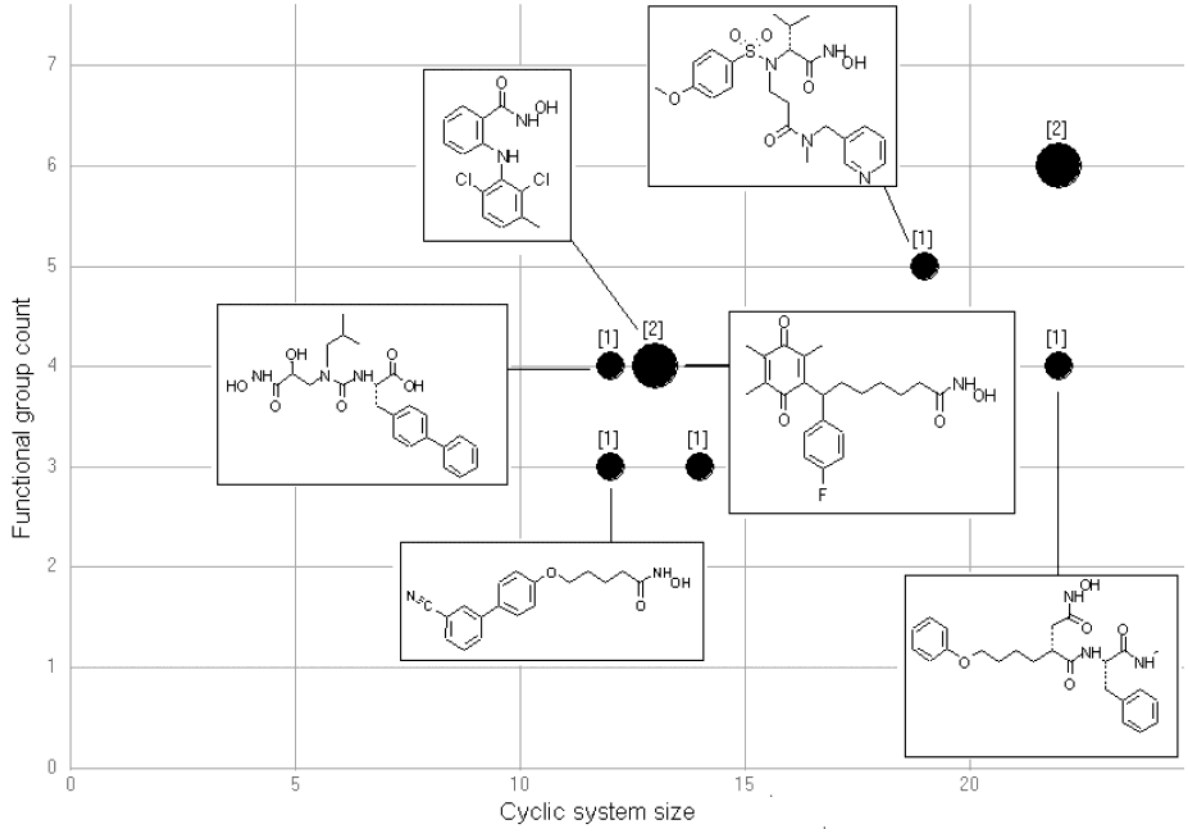

Figure 10. Plot of cyclic system size and functional group count for the nine compounds with a reduced-aryl cyclic-system value of J703P and a largest maximal-functional group value of WYDC (the hydroxamic acid group). The size and annotation of the marker gives the number of compounds associated with that marker. Representatives associated with some markers are displayed.

JOINT ANALYSIS OF THE CYCLIC SYSTEM AND FUNCTIONAL GROUP INDICES

The preceding results suggest that both cyclic-system and functional-group features are associated with distinctions between the regions of compound space represented by the MDDR and ACD libraries. However, with the exception of our observation regarding aromaticity, our conclusions all regarded correlates of compound size. Moreover, these trends with size were rather crude. This may reflect our ignoring functional-group features when examining cyclic-system features and our ignoring cyclic-system features when examining functional-group features. This leads to a confounding of results that is often unavoidable because of small sample sizes and high correlations that often exist among ones predictive descriptors. However, as we have just seen, MEQI-based histograms naturally direct ones attention to

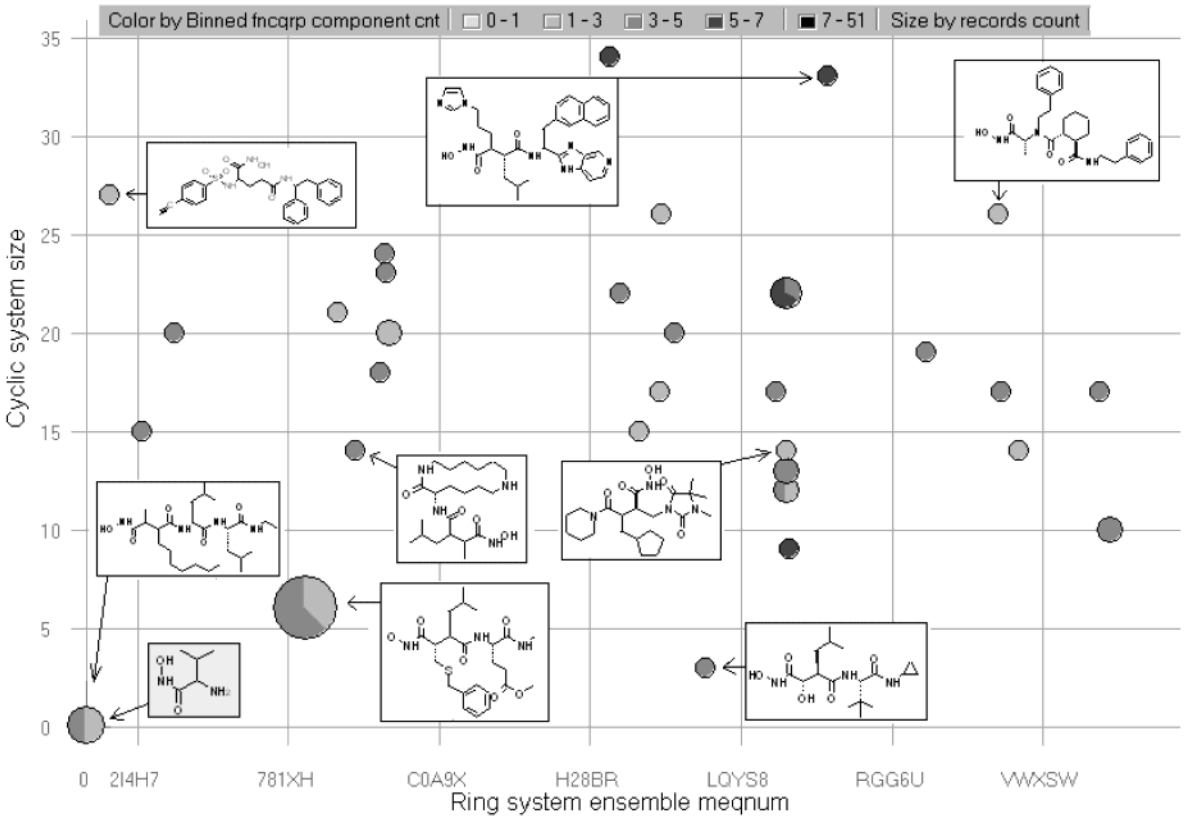

Figure 11. Plot of the ring system ensemble meqnum versus the cyclic system size for the 47 compounds with WYDC for its largest maximal-functional group value. Each arrow points from a representative structure to its associated pie marker. The size of the pie marker increases with the number of compounds in its associated class. Each pie is shaded to indicate the proportion of the compounds in that class having the number of functional groups indicated at the top of the figure. The structural drawing of the lone ACD compound is shaded.

large equivalence classes that can be further broken down into interesting subclasses. Such a breakdown led to the interesting observation from Figure 6 regarding aromaticity. Figure 9 further breaks down the compounds in the RACS (reduced-aryl cyclic-system) class J703P in Figure 6 by their largest maximal-functional group.

Interestingly, although the RACS class J703P consisting of two linked simple aromatic rings is relatively nondrug-like (25.6% MDDR compounds), some subclasses of compounds containing particular largest maximal-functional groups are quite drug-like. Seven subclasses that have significant improvement on drug-likeness are marked in Figure 9. Table 2 lists the detailed information about those classes. The one-sided binomial statistical test[13] was used to estimate the likelihood that purely chance factors could give rise to the enrichment of MDDR compounds in these classes that exceeded the 0.256 proportion for the reduced-aryl cyclic-system class J703P in Figure 6.

Subclasses that have significant enrichment become more interesting if the source of the enrichment can be established. Consider the hydroxamic-acid subclass WYDC. First of all, we must show that the functional group WYDC rather than other structural features in these nine compounds leads to the drug-likeness of this class. This is evident from Figure 10 where the cyclic-system sizes of the compounds making up this group are plotted against the functional group counts. The diversity of the group, despite their common reduced-aryl cyclic system, is evident from the variation in both variables. Moreover, no obvious additional commonalty exists in the group other than the hydroxamic-acid group.

Table 2. Classes Marked in Figure 9 with Significant Enrichment in Drug-Likeness[a]

Meqnum	Count	MDDR%	Enrichment	Significant level	Pseudograph
30F8	39	49%	1.9	0.058	
5XQ9	5	100%	3.9	0.042	
8Q7	47	47%	1.8	0.052	
PE3H	7	100%	3.9	0.003	
QATK	10	70%	2.7	0.15	
WYDC	9	100%	3.9	0.0002	
Y7Y	24	50%	2.0	0.29	

[a] The significance level is based on the one-sided binomial test.

The diversity of structures in Figure 10 suggests that this SAR pattern might hold against a larger domain of compounds. A search across both data sets for the substring WYDC turned up 46 MDDR compounds and only one ACD compound. Figure 11 demonstrates that this group of 47 compounds represents a strikingly broad range of ring system, cyclic system sizes, and functional group counts. Thus we feel functional group WYDC, the hydroxamic acid group, contributes strongly to drug-likeness.

Continuing our effort to generalize this SAR pattern, we checked if it held across the broader class of compounds

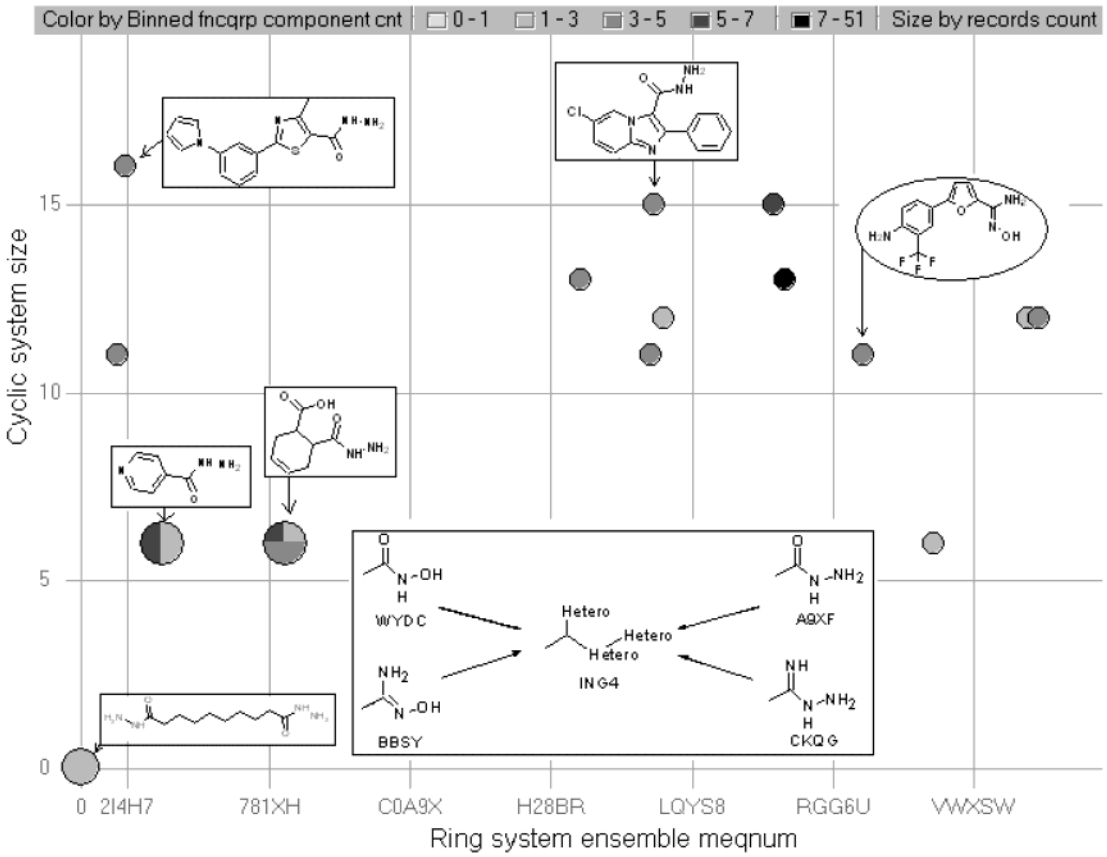

Figure 12. Plot of the ring-system ensemble meqnum versus the cyclic system size for the 23 compounds with A9XF, CKQG, or BBSY for their aryl-hetero largest maximal-functional group. Each arrow points from a representative structure to its associated pie marker. The size of the pie marker increases with the number of compounds in its associated class. Each pie is shaded to indicate the proportion of the compounds in that class having the number of functional groups indicated at the top of the figure. The lone MDDR compound is circled. The large box depicts the functional groups associated with a largest AHFG component value of ING4.

that have for their largest functional group one that falls in the aryl-hetero functional-group (AHFG) class ING4 which contains the functional-group class WYDC as a subclass. A search for those records having ING4 for their largest AHFG component meqnum turned up an additional 23 compounds consisting of 19 acid hydrazides, 3 *N*-hydroxy carboximidamides, and 1 carboximidohydrazide. The pseudograph and meqnum specifications of these functional groups are given in the inset in Figure 12 along with the pseudograph specification of the aryl-hetero functional group ING4. We see from this diversity plot that the compounds vary considerably in their ring-system ensembles, the size of their cyclic system, and their functional group counts. The data are insufficient to say much about the contribution to drug-likeness of the carboximidohydrazide and the *N*-hydroxy carboximidamide functionalities. However, the simple conversion of the hydroxyl of the hydroxamic acid to the amine of the acid hydrazide moves one from a region of chemical space with a drug-like to nondrug-like estimated odds of 46/1 to a region of chemical space with drug-like to nondrug-like odds estimated to be less 1/19 (as there were no MDDR compounds in this group).

This finding of a drug-like structure−activity cliff based on the simple change from the hydroxamic acid to the acid hydrazide should be viewed as only a first step in a comprehensive analysis that would incorporate a mechanistic explanation characterizing the full extent of the cliff. The next step would be to capture that mechanistic explanation into the set of chemical descriptors underlying a formal quantitative SAR model of drug-likeness. Thus, this type of visual analysis could contribute to the early phases of SAR modeling by suggesting types of chemical descriptors that should be incorporated into that modeling.

The finding of such structure−activity cliffs also plays a role in critically analyzing existing SAR models. Models of drug-likeness that predict the drug-likeness of a number of hydroxamic acids but does not predict the nondrug-likeness of their acid hydrazide counterparts are clearly ignoring some of the detailed differences in the ACD and MDDR libraries.

Finding this drug-like structure−activity cliff involved an examination of only a part of the information in the inset of Figure 8. A comprehensive search for other structure−activity cliffs would involve a systematic visual analysis of all the information in that figure at a comparable level of detail using larger drug-like and nondrug-like libraries more carefully screened for that purpose. However, for our purpose of illustrating the use of MEQIs in visual structure browsing and the types of serendipitous discoveries that thereby arise, these two rather limited and unscreened libraries were more than adequate.

SUMMARY

We have shown that molecular equivalence indices offer a new means of converting molecular structural information into codes amenable to formal structural and SAR analysis. Moreover, these codes can be visually interpreted and communicated. Their hierarchical relationships make them

ideal for structurally browsing large compound collections and for visually analyzing complex SARs both globally and locally. In demonstrating the general use of these tools, we have shown how cyclic-system and functional group features can be important in analyzing drug-likeness. We have also shown how molecular equivalence indices provide a facile means of drilling down and browsing very specific subclasses of compounds. This capability sets up a context for serendipitous SAR discoveries that was well illustrated in our finding the drug-like contributions of the hydroxamic acid functionality and then going back up the cyclic-system hierarchy to broaden the domain over which that conclusion was valid. We then showed that the hydroxamic-acid SAR did not extend to the next level up the chemical functionality hierarchy. In so doing, we made the serendipitous discovery that the acid hydrazide functionality was a strong contributor to nondrug-likeness. Thus, we found a drug-like structure−activity cliff where a small change in functionality corresponds to a movement from a drug-like to a nondrug-like region of structural space. Such discoveries can be expected when the methods presented here are incorporated into a much more extensive and systematic analysis of much larger libraries more carefully screened to the purpose fully characterizing the drug-like regions of chemical space. Finally, we have shown how molecular equivalence indices provide a new, but largely unexplored, means of visually depicting the diversity of the compounds over which a particular SAR conclusion has been established.

ACKNOWLEDGMENT

This manuscript was significantly revised and improved in response to the carefully considered criticisms by one of the reviewers. We thank that reviewer for his or her very helpful and thought-provoking effort.

REFERENCES AND NOTES

(1) (a) Johnson, M. A. In *Advances in Molecular Similarity*; JAI Press Inc.: 1998; Vol. 2 Browsable Structure−Activity Datasets, p 153. (b) Johnson, M. A.; Xu, Y.-J. In *Chemical Data Analysis in the Large: The Challenge of the Automation Age*; Hicks, M. G., Ed.; 2001; www.Beilstein-institut.de/bozen2000/proceedings. (c) Rouvary, D. H. The Evolution of the concept of Molecular Similarity. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley Inter-Science: New York, 1990; p 15.

(2) (a) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of 'drug-likeness'. *Drug Discovery Today* **2000**, *5*, 49. (b) Ajay, W. P. W.; Murcko, M. A. Can We Learn to Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314. (c) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci*. **1998**, *39*, 165. (d) Sadoeski, J.; Kubinyi, H.; A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325. (e) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165. (f) Sadowski, J.; Kubinyi, H. A scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325. (g) Wagener, M.; Van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280.

(3) Behzed, M.; Chartrand, G.; Lesniak-Foster, L. *Graphs & Digraphs*; Wadsworth Inc.: Belmont, 1979.

(4) (a) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887. (b) Bemis, G. W.; Murcko, M. A. The properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095.

(5) Balaban, A. T.; Filip, P.; Balaban, T.-S.; Computer Program for Finding All Possible Cycles in Graphs. *J. Comput. Chem.* **1985**, *6*, 316.

(6) (a) Adamson, G. W.; Creasey, S. E.; Eakins, J. P.; Lynch, M. F. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part V. More Detailed Cyclic Fragments. *J. Chem. Soc., Perkin Trans. I* **1973**, 2071. (b) Nilakantan, R.; Bauman, N.; Haraki, K.; Venkatarahavan, R. A. Ring-Based Chemical Structural Query System: Use of a Novel Ring-Complexity Heuristic. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 65.

(7) Read, R. C.; Corneil, D. G. The Graphs Isomorphism Disease. *J. Graph Theory* **1977**, *1*, 339.

(8) Xu, Y.-J.; Johnson, M. A. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181.

(9) Johnson, M.; Schulz, M.; Cheng, C. Visualizing Keyword Lists and Other High-Dimensional Binary Response Variables. *Computing Sci. Statistics* **1999**, *31*, 341.

(10) Lawson, A. The Lawson Similarity Number (LN): Offline Generation and Online Use. In *The Beilstein Online Database*; ACS Symp. Ser.; 1990; Vol. 436, p 143.

(11) Use Spotfire software package, a commercial data visualization package marketed by Spotfire AB, Göteberg, Sweden, http://www.spotfire.com.

(12) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Freeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3.

(13) Conover, W. J. *Practical Nonparametric Statistics*; John Wiley & Sons: New York, 1980.

CI025535L