Beilstein-Institut

STRUCTURAL BROWSING INDICES AS HIGH-THROUGHPUT SAR ANALYSIS TOOLS

MARK JOHNSON[a] AND YONG-JIN XU[b]

[a] Computer-Assisted Drug Discovery, Pharmacia Inc., 301 Henrietta Street, Kalamazoo, MI 490006, USA.

[b] Discovery Medicinal Chemistry, Pharmacia Inc., 800 North Lindbergh Blvd., Creve Coeur, MO 63167, USA.
E-mail: mark.a1.johnson@am.pnu.com

ABSTRACT

Structural browsing indices (SBIs) have been proposed as tools for organizing and exploring large sets of chemical structures in a manner complementary to that addressed by substructure and similarity-based methodologies. Molecular equivalence indices (MEQIs) comprise a special subclass of SBIs that play a central role in constructing a suite of SBIs appropriate to a variety of browsing, chemical-diversity, and SAR tasks. After presenting a general definition of a molecular equivalence index, three different ways of constructing SBIs based on MEQIs will be illustrated. The first index uniquely identifies the chemical graph of a compound and will be used to identify the sets of geometric and stereoisomers in a compound collection as well as to visually assess the overlap of two compound collections. The second index identifies a largest set of nonoverlapping functional groups of a compound and will be used to visually identify a functional-group-based receptor-relevant subspace associated with ACE inhibitors. The third index provides a hierarchical ordering of compounds whose use will be illustrated in the context of browsing structures and SAR relationships.

INTRODUCTION

The problem of organizing collections of molecular structures has been with us in one form or another since the dawn of modern chemistry. The development of substructure-searching algorithms was one of the initial pursuits in the creation of databases specifically structured for chemists and reflects the natural partial ordering of compounds with respect to the substructure-relationship. The last 15 years has seen the development of sophisticated algorithms for similarity searching, another way of exploring the compounds in a large collection based on the computation of a distance relationship between them. However, neither of these two methods provides a systematic way of assuring that all of the compounds in a collection have been examined.

Clustering and projection methods have long been available as statistical tools for organizing objects embedded in a high-dimensional space that does facilitate systematic browsing. Projection methods organize the objects into a low-dimensional space, usually the plane, so that distances in the points reflect distances between the points in the high-dimensional space. Clustering methods traditionally organize the objects along a line so that related clusters tend to occur together.

Long predating chemistry, humankind faced the problem of organizing large collections of hand-made objects as market places evolved. Modern department stores now display millions of items. Yet in one day you browse a large department store

for a sense of what it sells, and if you wish to buy a shirt, you are likely to find the shirts that interest you in reasonably close proximity of each other. This is an organizational feat that merits study by investigators in cheminformatics.

How do the department store managers do it? Basically they form a hierarchy of equivalence classes. Appliances, clothes, cosmetics…. Within appliances: stoves, refrigerators, washing machines…. Stoves might then be organized by size or manufacturer. Lastly the items are ordered along aisles in a manner consistent with this organizational hierarchy.

The organizational hierarchy can be distinguished from the clustering and projection methods we have just mentioned in that the equivalence classes are in some sense inherent in the nature of the object. We needn't see a stove in a cluster of other stoves, refrigerators, and washing machines to recognize that stove as an appliance and not a cosmetic.

The molecular equivalence indices presented here were developed with this department store analogy in mind, only, in this case, the equivalence classes are entities such as the chemical graph, the cyclic system, and chemical formula of a molecule, its side chains, ring systems and functional groups. Rouvray [1] reviews a number of the notions of structural equivalence that have played an important role in the development of chemistry. The formal perception of various integral components of a molecule has its origin in the dawning of cheminformatics, [2] as does the perception of an exhaustive set of a particular genre of components. [3] The idea of looking at a formalized notion of molecular equivalence and studying the resulting equivalence classes is more recent [4] as is the notion of hierarchically organizing structures by means of numbers. [5] The notion of systematically incorporating various notions of molecular equivalence into browsing

indices whose values essentially serve as names for the resulting equivalence classes [6] forms the subject of this study.

After describing the set of structures that will serve to illustrate the concepts, a general definition of a molecular equivalence index (MEQI) will be given. A simple, yet fundamental, MEQI that assigns each chemical graph a unique code [7] will be presented and used to find the sets of geometric and stereoisomers in collection of compounds and to illustrate a simple mechanism for determining which structures occur in each of two collections. A more general MEQI identifies a largest set of nonoverlapping functional groups of a compound and will be used to visually identify a functional-group-based receptor-relevant subspace associated with ACE inhibitors. Finally, a MEQI specifically designed to hierarchically order compounds with respect to their cyclic systems and arrangement of their side chains will be illustrated in the context of browsing structures and SAR relationships.

AN ACE-INHIBITOR DATASET

In a recent paper, Pearlman and Smith [8] develop the concept of a receptor-relevant subspace using 78 angiotensin-converting enzyme (ACE) inhibitors. In Figure 3 of that study, these 78 compounds are positioned in a localized area of a three-dimensional BCUT space when viewed against a backdrop of a "5% diverse subset of the total MDDR [9] population." Bob Pearlman graciously sent us the structures of those 78 ACE inhibitors and Veer Shanmugasundaram kindly provided us with a similar diverse subset of 3932 compounds based on a comparable three-dimensional BCUT space from the MDDR collection at Pharmacia. Choosing a "comparable" subset of the MDDR compounds to serve as a backdrop was thought to increase our chances of finding a receptor-relevant subspace using MEQIs,

a concept that will be discussed in the section on the alpha-augmented functional group ensemble MEQI. No attempt has been made to verify the suitability of this expectation.

DEFINING A MOLECULAR EQUIVALENCE INDEX

If a chemical descriptor is viewed broadly enough to include any function that maps the space of compounds to a linearly ordered set, a MEQI can be viewed as a special case of a chemical descriptor. However, in the case of a MEQI, this mapping can always be viewed as a composite mapping in that it first maps the space of compounds to a space of visually interpretable representations and then maps this intermediary space to a linearly ordered set.

This decomposition of a MEQI is illustrated in Figure 1. A few comments are needed to explain the figure. For computationally purposes, one must replace the compounds by some approximate mathematical representation. In Figure 1, we use a slight generalization of the chemical graph in which both the vertices and the edges are labeled. Mathematicians call this a colored or labeled graph. By allowing for loops and multiple edges, one

obtains a labeled pseudograph. Thus, in our case, the equivalencing function always maps the space of labeled pseudographs onto itself. The particular equivalencing function in Figure 1 deletes all single-degree vertices labeled 'H' for hydrogen. In particular, it converts all chemical graphs to their hydrogen-reduced counterparts, but note that our definition of this equivalencing function is operationally defined for any labeled pseudograph.

The second mapping assigns each labeled pseudograph a unique code. This code depends only on the labeled pseudograph on which it is computed, and not on the compound mapped to that pseudograph by the equivalencing function. This code could be a number base 10, a number base 36, such as is used in car license plates, or a character string. However, the resulting values must be linearly ordered. In some cases, these assigned values depend on the sequence in which graphs are presented to the naming algorithm with the first graph labeled number 1, the second 2, *et cetera*. [6] *A priori* naming procedures [5,7] depend only on the labeled pseudograph and will consequently be independent of the time and place in which the naming is carried out. Obviously, the utility of a MEQI diminishes rapidly if this naming function is
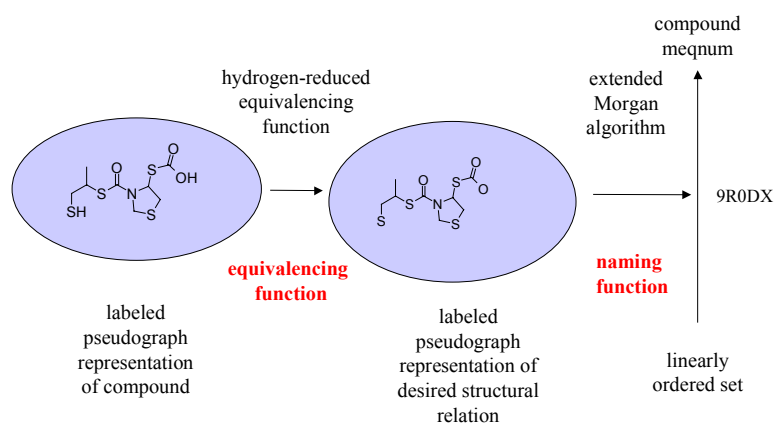

Figure 1. Two basic components of a molecular equivalence index mapping a compound to its compound meqnum.

Beilstein-Institut

not unique for all practical purposes, i.e. nonisomorphic labeled pseudographs are assigned distinct values. (It remains an open question if there exists a one-to-one naming function that lies outside the NP-completeness class. [10])

In this study we will be using an extension [11] of the Morgan algorithm [12] to compute an *a priori* naming function. We have yet to encounter a case of nonuniqueness. This algorithm assigns a number base 34. (I's and O's are not used because of their possible confusion with 0's and 1's.) We refer to this number as a molecular equivalence number or meqnum for short.

For every distinct equivalencing function, we obtain a distinct MEQI. When the equivalencing function maps a compound to its hydrogen-reduced graph as in Figure 1, we call the resulting assigned numbers "compound meqnums." In an analogous way, we obtain compound-skeleton meqnums, cyclic-system meqnums, cyclic-system skeleton meqnums, *et cetera*.

THE COMPOUND MEQNUM

Finding Geometric and Stereoisomers

The compound meqnum identifies a compound up

to geometric and stereoisomerism. Even this simple meqnum has interesting uses. For example, the pharmacological activity of a compound is often stereospecific, whereas most chemical descriptors are not. This would seriously diminish the utility of most chemical descriptors in lead-optimization contexts if it were not for the fact that lead optimization efforts in drug discovery quickly focus on those compounds with the desired handedness at the critical stereocenters. However, there are often cases in which both stereoisomers are present and one must remove the compound with the undesired handedness before proceeding further. This is easily done by computing the compound meqnums for all of the compounds and then constructing the histogram given in Figure 2. We will assume that the compound with the desired handedness will be synthesized whenever the compound with the undesired handedness is synthesized. Consequently, the compound meqnum of any compound with the undesired handedness will occur twice since the corresponding stereoisomer will also be present and have the identical chemical graph.

Emerging graphical capabilities are enabling us to visualize relationships involving high-content
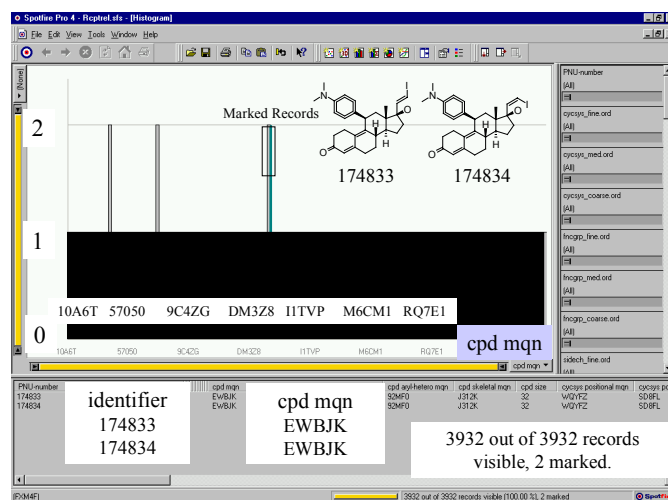

Figure 2. Histogram for finding geometric and stereo isomers with with the compound meqnums along the x-axis. The two geometric isomers associated with the marked bar of height two are displayed.
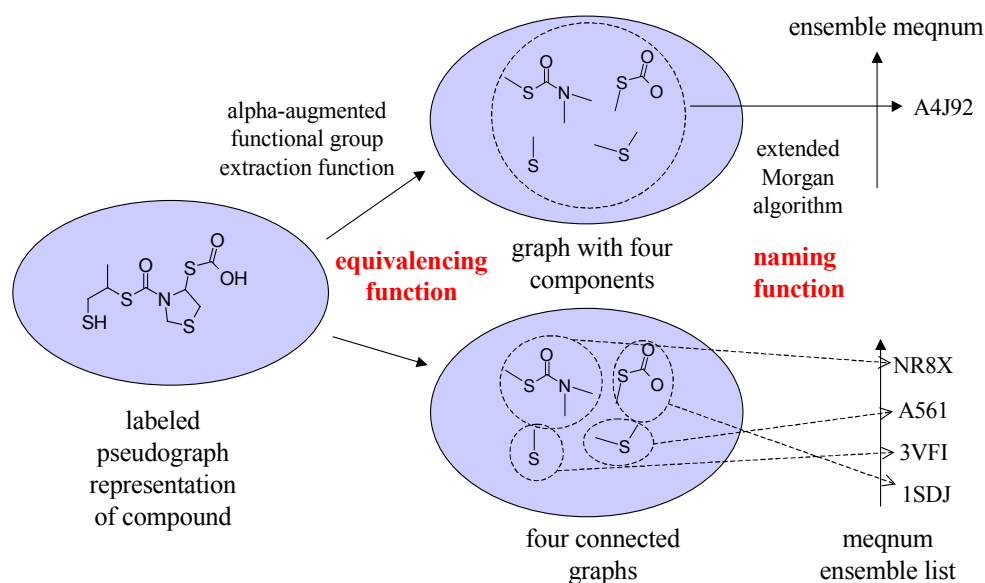
Figure 3. Construction of two alpha-augmented functional group MEQIs using a naming function that generates a single meqnum and a list of meqnums for multicomponent graphs, respectively.

variables such as MEQIs. Spotfire [13] allows the use of string-valued variables for the axes of a plot and provides many of the navigational aids required for efficient browsing. By simply selecting the compound meqnum variable for the x-axis in the histogram view in Spotfire, Figure 2 pops into view.

Out of a data set of roughly 4000 compounds, one quickly and visually isolates all the pairs of geometric and stereoisomers. These pairs correspond to the three thin bars of height 2 representing 6 compounds. The structures can be seen by moving the mouse diagonally across its top to form an enclosing rectangle which "marks" the compounds. One of the bars of height 2 in Figure 2 is marked. The details window gives the identifiers of the two tallied compounds as **174833** and **174834** and gives EWBJK for the common compound meqnum. The remaining 3926 compounds are represented by corresponding bars of height 1 compressed so tightly as to give the visual impression of a solid black horizontal bar of that height.

Comparing Two Compound Collections

A similar logic allows one to quickly find the intersection in two compound collections. Again, compounds that occur in both collections would be represented by bars of height 2 or greater. These can be marked appropriately and the other compounds deleted. The remaining bars can then be proportionally colored by source. Multicolored bars would reflect chemical graphs found in both collections. Monocolored bars would represent isomers and other compounds with the same chemical graph found in only one collection.

AN ALPHA-AUGMENTED FUNCTIONAL GROUP MEQNUM ENSEMBLE

The concept of a receptor-relevant subspace as developed by Pearlman and Smith [8] can be viewed generally as any formal specification of a class of compounds in which compounds with the desired receptor affinity are highly concentrated. In this section, we would like to illustrate another group of MEQIs by developing one that provides a
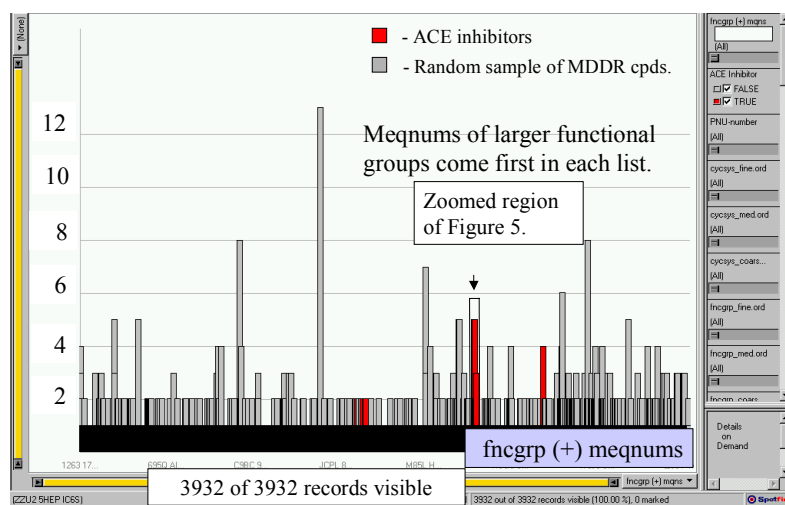
Beilstein-Institut

Figure 4. Histogram of alpha-augmented functional group list.

simple means of specifying a receptor-relevant subspace for the 78 ACE inhibitors in our data set.

Figure 3 shows two distinct MEQIs involving the same equivalencing function, but two different, yet related naming functions. To define the equivalencing function, divide the atoms of a chemical graph into separating atoms and non-separating atoms. Call a largest-connected subgraph consisting only of non-separating atoms a maximal group. By letting the separating vertices be any carbon atom that does not share a double bond with

any oxygen, nitrogen, or sulfur or share a triple bond with nitrogen, we obtained the maximal functional groups. By augmenting these maximal functional groups with their adjacent alpha carbon atoms, we obtain the alpha-augmented functional groups (AFGs) that form the disconnected graph of four components depicted Figure 3.

We now have a choice of naming functions. We can use the one in Figure 1 which always assigns a single number to a graph whether connected or not. This gives the ensemble meqnum A4J92 in the
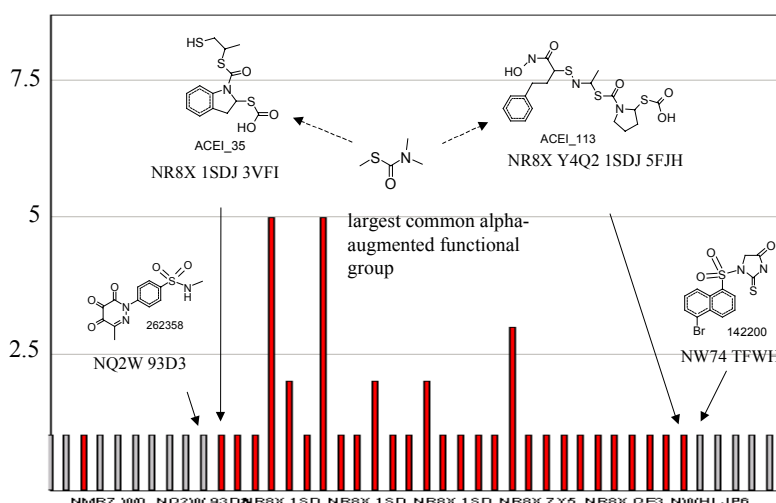

Figure 5. Zoomed region of histogram in Figure 4 of the alpha-augmented functional-group meqnum-ensemble list showing a grouping of ACE inhibitors with respect to their largest perceived functional group.
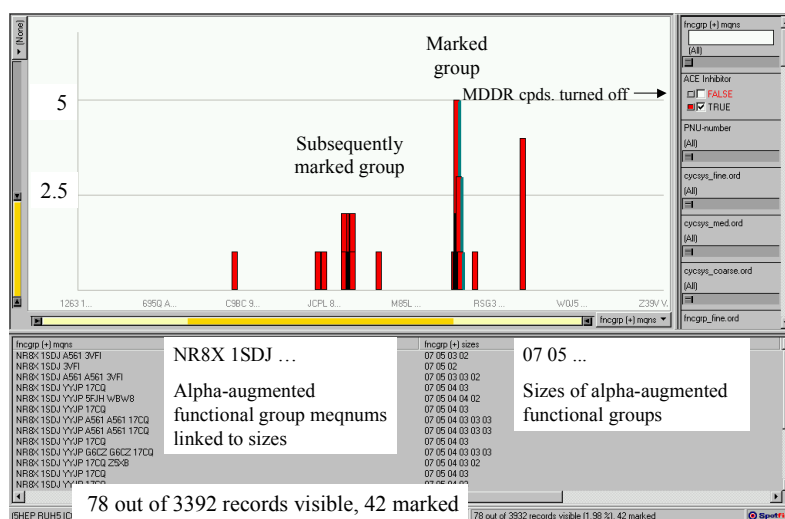
Figure 6. Marked region of ACE inhibitors suggesting alpha-augmented functional groups associated with ACE activity.

upper portion of Figure 3. Alternatively, we could apply the naming algorithm to each of the components and, in that way, obtain a list of numbers. This is illustrated in the lower portion of Figure 3, in which the outcome of the naming function is a meqnum ensemble list. There are k! ways of ordering a list of k numbers. To order the AFG lists canonically, we order the names first by the number of atoms in the corresponding component. When two or more components have the same number of atoms, the numbers are ordered lexicographically. The ensemble meqnum is nice when a short number is required. The meqnum ensemble list gives us substring access to its components and will be used here.

Figure 4 is obtained by simply selecting the AFG meqnum ensemble list variable for the x-axis of the histogram and coloring the bars to indicate the proportion of ACE inhibitors amongst the compounds with a particular set of alpha-
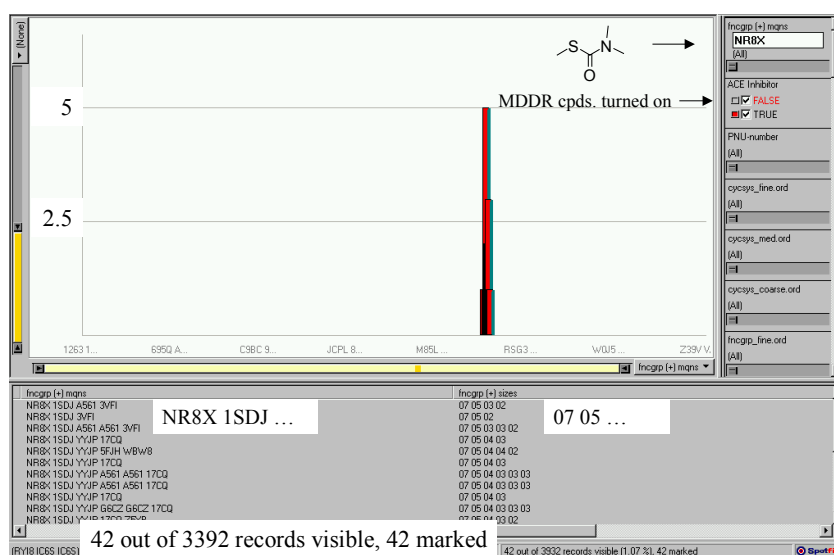

Figure 7. Substring search demonstrating the specificity of a suggested alpha-augmented functional group with meqnum NR8X.
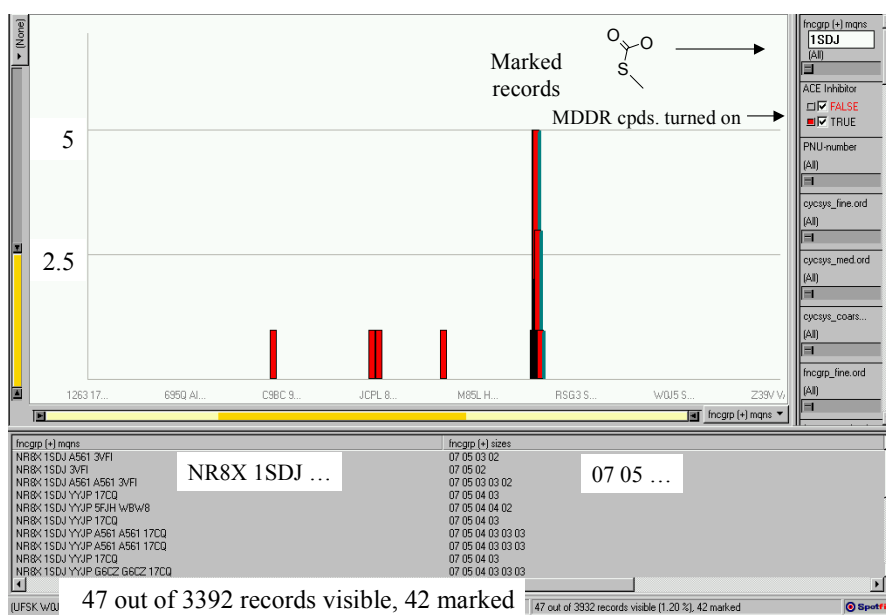
Figure 8. Substring search demonstrating the specificity of a co-occurring
alpha-augmented functional group with meqnum 1SDJ

augmented functional groups.

We immediately see that one combination of AFGs is shared by 13 non-ACE inhibitors, and another combination of AFGs is common to 5 ACE inhibitors. However, most of the compounds have a unique combination of AFGs, and consequently, we obtain the black horizontal bar of height 1 along the bottom. The importance of using an meqnum ensemble list rather than a ensemble meqnum is revealed when we use the x-axis slider to zoom in on the narrow region on either side of the red bar corresponding to the 5 ACE inhibitors. This gives rise to Figure 5. Since the AFG meqnums in each ensemble list are ordered first by size, and since the carbamothioate AFG with meqnum NR8X is the largest AFG in quite a few ACE inhibitors, but is not the largest AFG in any non-ACE inhibitors, we obtain a very interesting interval of uninterrupted ACE inhibitors.

Zooming back out and turning off the non-ACE inhibitors, we obtain Figure 6. One can now easily mark the interval of ACE inhibitors displayed in Figure 5. This reveals the AFG lists for each of the

marked compounds. Again we note that each begins with NR8X.

To check if the associated AFG occurs on any other compounds, which would necessarily contain another AFG of 7 or more atoms, one enters NR8X in the substring search window for the AFG slider as indicated in the upper-right portion of Figure 7. When finished, all compounds without that AFG are removed from view. In Figure 7, we see that the non-ACE inhibitors have been turned back on! Consequently, we see that all compounds containing the NR8X functional group are ACE inhibitors.

But Figure 6 also reveals that the thiocarbonate AFG 1SDJ is present whenever NR8X is present. Searching for those compounds that contain 1SDJ, we obtain Figure 8. There are 47 such compounds, all ACE inhibitors. The data are inadequate to determine if only one or both of these functional groups is critical to activity in this subseries of the ACE inhibitors.

It is informative to repeat this logic by marking the compound in the "subsequently marked region" in
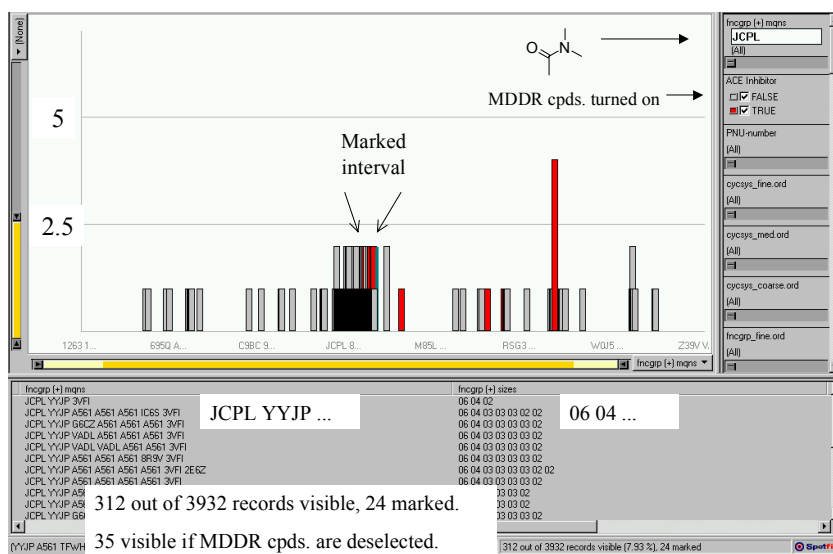
Figure 9. Substring search demonstrating the nonspecificity of a suggested
alpha-augmented functional group with meqnum JCPL.

Figure 6. The results are summarized in Figure 9. We see that there were 24 marked compounds whose largest AFG is the amide JCPL. The corresponding substring search reveals a total of 312 compounds with that AFG, 35 of which are ACE inhibitors. Consequently, we conclude that this amide AFG is not ACE-receptor specific, even though it may still contribute to activity when other more receptor-specific structural features are present in a particular arrangement.

A DESIGNED CYCLIC SYSTEM-ORDERING Browsing Structures

Efficient systematic browsing requires that structures be linearly ordered. If we are to look at every structure m in a collection of n structures without looking at any one more than once, we would necessarily encounter them in some sequence. One of the most common sequences is defined by the registry number of compounds. Figure 10 shows the first 12 structures one would encounter when lexicographically ordering the 3854 MDDR structures in our data set by their registry number. Although very useful for finding

particular compounds when the registry number is known, this ordering does not facilitate our finding a particular cyclic system or getting a good sense of its representatives.

Now suppose the structures were ordered by a MEQI that maps each structure to its cyclic system. Then, for each cyclic system, there would be a single largest interval of compounds comprised of all the compounds with that particular cyclic system. Long/short intervals would represent cyclic systems represented by many/few compounds, respectively. However, adjacent intervals would generally represent compounds coming from entirely unrelated cyclic systems. For example, an interval of steroids might be adjacent to an interval of indoles.

This raises the question as to how one gets closely related cyclic systems to be associated with closely positioned intervals. The natural solution is to develop a hierarchical ordering so that, for example, the compound intervals associated with cyclic systems sharing the same cyclic skeleton are grouped together. Such groupings are easily obtained as follows:
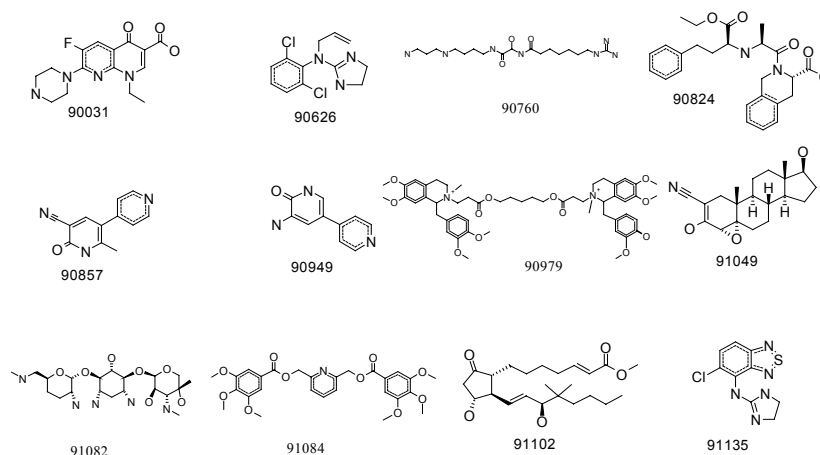
Figure 10. First 12 of 3854 random MDDR structures as traditionally ordered by registration number.

Let SBI$_j$, j = 1,…,J, be any finite sequence of SBIs. These will usually be a combination of MEQIs associated with the cyclic system skeleton, the set of component ring systems, *et cetera* and suitably chosen counts of the number of atoms, number of component ring systems, *et cetera*. If m denotes an arbitrary structure, then 'SBI$_1$(m) SBI$_2$(m) … SBI$_J$(m)' is a keyword list. A variable taking such keyword lists as values hierarchically orders structures when its values are lexicographically ordered. For example, if J were 2 and SBI$_1$ and SBI$_2$ were MEQIs representing a cyclic-skeleton meqnum and cyclic-system meqnum, respectively, we would immediately accomplish our purpose of assuring that compound intervals associated with cyclic systems sharing the same cyclic skeleton were grouped together.

The proof of the relevance of a particular sequence of SBIs in constructing a hierarchical ordering lies in the relevance of the compound orderings that emerge. Such relevance is best demonstrated though numerous examples in a variety of contexts. Space restrictions allow only a rather superficial demonstration of a rather involved cyclic system ordering we are exploring.

The first SBI in the construction of this ordering is the number of ring systems. Since this number is 0 for acyclic structures, all acyclic structures precede all non-acyclic structures in our ordering. Consequently, to extract a short section of the 3854 MDDR structures in our data set that shows that our cyclic system ordering groups related cyclic systems, we list structures 1001 – 1012 in our ordering. The list, given in Figure 11, consists of 12 aromatic, single-ring-system structures beginning with 6 quinoxalinediones, followed by a 1,2,3,4-tetrahydropyrido[4,3-d]pyrimidine-2,4-dione, and then 5 1,2,4-benzotriazin-3-ones. Our perception program currently treats a ketone as an acyclic group. Consequently, the first quinoxalinedione has 3 acyclic groups, the next three have 4, and the last two have 5. Because of this ordering of the number of acyclic groups within a cyclic system, we know there are exactly 3 and 2 single-ring-system quinoxalindiones with 4 and 5 acyclic groups, respectively. Similarly, the interval of 1,2,4-benzotriazin-3-ones begins with two compounds with 2 acyclic groups. The last three compounds have 3 such groups. Consequently, we know there are exactly 2 single-ring-system 1,2,4-benzotriazin-3-ones with 2 acyclic groups in this subcollection of the MDDR.

Beilstein-Institut
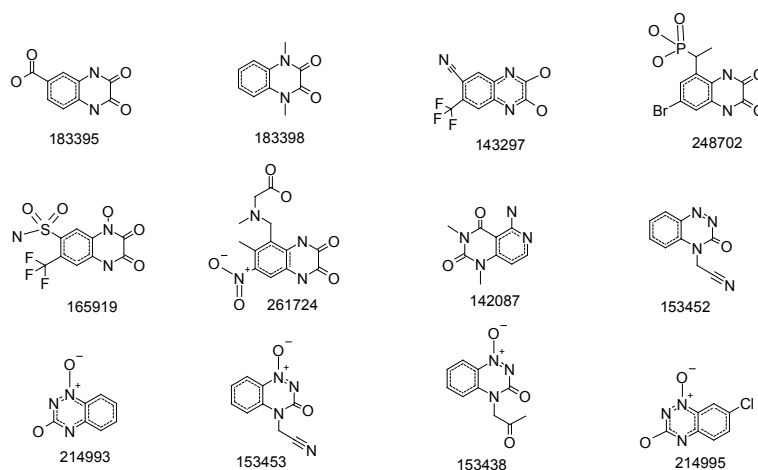
Figure 11. Compounds 1001-1012 in the fine-grained cyclic-system
rrdering of 3854 random MDDR structures

Browsing a Structure-Activity Relationship

A visual analysis of a structure-activity relationship (SAR) provides an intuitive feel for the structures on which it is based and roughly determines which structural features are critical to activity. There are many aspects to a comprehensive visual analyses of an SAR. One aspect that is repeatedly encountered is to find a group of compounds with a common cyclic system and similarly positioned side-chains.

medium and fine-grained cyclic system ordering. The medium-grained ordering only distinguishes between compounds with different cyclic systems. The fine-grained ordering further distinguishes the compounds by the number of side-chains, how they are positioned, and the particular set of side chains. Figure 12 illustrates how the two levels of resolution work together. The figure is restricted to the 78 ACE inhibitors. The upper histogram has the medium-grained cyclic-system ordering along the

Figure 12. Linked histogram and scatter plot of 78 ACE inhibitors with medium and fine-grained cyclic-system orderings for the x-axes.

This is easily facilitated with the joint use of a

x-axis. The lower scatter plot has the fine-grained

Figure 13. An interval in a fine-grained cyclic-system ordering which uncovered
a structure-activity cliff based on small side-chain difference.

cyclic-system ordering along the x-axis and minus the log of the IC50 concentration for the y-axis. The tallest bar in the histogram indicates the presence of a cyclic system represented by 11 compounds. When one "marks" this tallest bar, the corresponding points in the scatterplot are marked as well. These 11 marked points form an interval of contiguous marked points because the fine-grained ordering is simply a further elaboration of the medium-grained ordering.
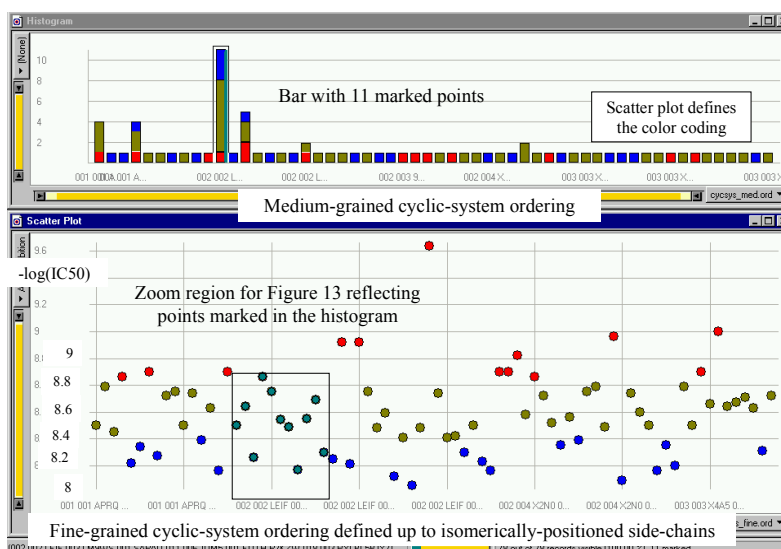
Because the cyclic-system orderings are based purely on structure, one has no guarantee or even expectation that a particular activity will relate to that ordering. However, one can expect to see closely related structures positioned close to one another. Should these similarly positioned structures differ markedly in activity, we will have found a "structure-activity cliff" where a small structural change is accompanied by a large change in activity. Such an occurrence identifies a critical position in the SA analysis. Figure 13, a blow-up of the marked region in the lower scatter plot of Figure 12, illustrates such an occurrence. Notice that ACE inhibitors **62**, **64**, and **72** have side-chains at the same position and that the number of atoms

in the side-chains increases as we move along this particular part of the ordering. As we go from the propyl group to the aminopropyl group, a marked increase in activity is observed, revealing a structure-activity cliff.

POSITIONING MOLECULAR EQUIVALENCE INDICES IN CHEMINFORMATICS

MEQIs are another tool in a long line of tools for organizing and browsing structures. Figure 14 is an attempt to put these tools into a comparative perspective, not with respect to the pros and cons of the possible uses to which such tools have been put, but wit h respect to their mathematical and inferential structure. The major categories along the

Figure 14. Positioning molecular equivalence indices as natural tools for visually organizing large compound collections.

Beilstein-Institut

first row of the figure groups these tools by the underlying mathematical space.

Although complex and difficult to navigate, the space of chemical graphs, partially ordered by the substructure relation, is arguably the most fundamental of the three representations. Because a chemical graph is such a rich storage vehicle, substructure searching gives the user exquisite control in retrieving specific subsets of structures. On the other hand, manually specifying such subsets is too time consuming and the resulting inferential structure too restrictive for most purposes of SAR analysis.

High-dimensional chemical-descriptor spaces have become increasingly important with the advent of similarity searching and the development of data-mining software, especially recursive partitioning programs. The component chemical descriptors usually have very limited structural content by themselves, but taken all together, they can encode a very significant amount of the structural information in a molecule. These spaces are arguably the most simple in that one can define an algebra over them. Consequently, one can "automate" analyses. On the other hand, these high-dimensional spaces are visually unintuitive and often what actually takes place in this automation can differ significantly from what one believes is taking place. (See the paper of this Beilstein workshop by Stanley Young for recent developments along these lines.)

Structural browsing indices are variables whose values are linearly ordered, but there is no restriction that they behave as numbers admitting algebraic operations. The only requirement is that intervals along this linear ordering represent some type of structural commonality. The more such intervals there are, the arguably more rich is the information content of the corresponding index. (One could think of a single fragment chemical descriptor that might be a component of a high-dimensional descriptor space as a browsing index, but it would be a relatively uninformative one. One of its intervals would represent the compounds with the structural fragment and the other would represent the remaining compounds.)

Structural browsing indices have been around for a long time, and have always played an important role in visualizing chemical space. The idea of capturing in a few variables much of the distance information in a high-dimensional point cloud has a long history in statistics and in cheminformatics. Often two principal components suffice. Although some information is sacrificed, much is gained by being able to visualize the captured information in a two-dimensional point cloud. Hierarchically clustering objects and then correspondingly ordering the objects along a line also has a long history, but is receiving renewed interest from the scientific visualization community. (See the papers of this Beilstein workshop by Jeff Saffer for recent developments in visualization methods based on projection and hierarchical clustering.) Meqnum orderings provide a third alternative.

The three types of orderings can be operationally distinguished four ways. First, a MEQI is distinguished from the other two indices in that it can be computed on a single object. The other two types of clustering and projection indices only make sense with respect to a collection of compounds. Their values change with changes in that collection.

Second, the visual grouping of structures is hierarchically organized for MEQI and clustering-based methods whereas these groupings are spatially distinguished in projection methods. This distinction leads naturally into the third distinguishing criteria. Since spatial distinctions rely upon the eye to say whether or not a particular point is or is not in a cluster, the user has

considerable freedom in deciding which groupings of points are clusters and which are not. Operationally, the structural groupings are exactly set fourth when using MEQI and clustering-based methods.

MEQIs are again distinguished from the other two visualization categories when it comes to interpreting the clusters. The interpretation of a meqnum is set forth by the equivalencing function. Moreover, the labeled pseudograph to which a compound is mapped by that function serves as a visual specification of its class with respect to that equivalencing function. This contrasts markedly with the groupings set up via the other two visualization methods. Sometimes these methods generate clusters which admit obvious specifications that distinguish the clusters, but it would seem to be a rare instance where this would be the case if all possible structures were represented in the collection of compounds that was clustered.

SUMMARY AND CONCLUSION

In this study we have attempted a rather broad overview of the types of MEQIs that can be generated and the variety of uses to which they can be put. Our overview is far from exhaustive, and the examples invite further development. Hopefully, this brief sketch of some of the directions we are pursuing in delineating roles MEQIs might play in cheminformatics and structure-activity analysis will suggest areas of interest to others.

REFERENCES AND NOTES

[1]   Rouvray, 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, NY, 1990; pp 15-42.

[2]   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 I*, **1973**, 2071-2076.

[3]   Randić, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for all Self-Avoiding Paths for Molecular Graphs. *Comput. & Chem.* **1979**, 3, 5-13.

[4]   Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, 39, 2887-2893.

[5]   Lawson, A. J. The Lawson Similarity Number (LN): Offline Generation and Online Use. In *The Beilstein Online Database*; Heller, S., Ed.; ACS Symp. Ser **1990**, 436, 143-155.

[6]   Johnson, M. A. Browsable Structure-Activity Datasets. In *Advances in Molecular Similarity*; Carbó-Dorca, R.; Mezey, P., Eds.; JAI Press Inc., 1998, 2, 153-170.

[7]   Randić, M. Molecular ID Numbers: By Design. *J. Chem. Inf. Comput. Sci.* **1986**, 26, 134-136.

[8]   Pearlman, R. S.; Smith, K. S. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 28-35.

[9]   Modern Drug Data Report database is distributed by MDL Information Systems, San Leandro, CA.

[10]   Read, R. C.; Corneil, D. G. The Graphs Isomorphism Disease, *J. Graph Theory* **1977**, 1, 339-363.

[11]   Xu, Y.-j.; Johnson, M. Extending the Morgan Sequence Algorithm for the Needs of Structural Browsing Indices, to appear in *J. Chem. Inf. Comput. Sci.*, **2001**.

[12]   Morgan, H. L. The Generation of a Unique Machine Description for Chemical Graphs - A Technique Developed at Chemical Abstracts Service, *J. Chem. Doc.* **1965**, 5, 107-113.

[13]   Spotfire is a product of Spotfire, Inc. Cambridge, MA (www.spotfire.com).