

A Brief Introduction to RingSystemSuite

**Mark Johnson
Pannanugget Consulting**

**Technical Report
Pannanugget Consulting, Inc.
Kalamazoo, MI
2006**

Table of contents

1. Generating the indices
2. Special considerations when browsing structures using largest ring systems
3. Positioning the chemotypic approach to SA studies amongst other CADD approaches
4. Delineating active regions in HTS data
5. Hierarchically relating the component indices of the RingSystemOrdering
6. Finding critical structural comparisons at the boundary of an active region
7. Interfacing with other ideas and approaches

A Brief Introduction to RingSystemSuite

The preprogrammed Meqi II programming template called RingSystemSuite is designed for CADD problems when interest is directed toward templates as defined by a single ring system. Compounds can have multiple ring systems. To focus on a single ring system of a compound, Meqi provides the option of selecting the first ring system it encounters on a structure that has at least as many bonds as any other of that structure's ring system. Most ring-system templates correspond to one of these "first-largest" ring systems for a sufficient number of interesting structures to merit ones attention. This suite is emphasizes first-largest ring-system indices and includes two resolution-related indices for hierarchically ordering structures by such ring-system indices.

Generally, there is a unique first-largest ring system for most structures. When there are two or more "largest" ring systems, one must be selected. These ties are broken by selecting whichever of the "largest" ring systems Meqi first encounters when reading in the connection table for that structure. The sequence in which Meqi encounters the ring systems depends on the sequencing of the elements in connection table. The latter sequence often reflects the sequence in which the atoms and bonds are drawn when registering the structure, i.e. when entering it into ones compound database. Consequently, in these indices there is only one first-largest ring system.

When there are ties, this method of breaking ties introduces an artifact. It seemed best to avoid this artifact through the use of a canonical ordering method for breaking ties when Meqi was first developed. However, this "first-encounter" method of breaking ties assures that the first-largest ring-system chemotypic identifiers (CIDs) of a structure will always refer to the same ring system in the structure regardless of which first-largest ring-system index is used. In most circumstances, this analytical advantage far outweighs the cost of an occasional artifact when the "largest" ring system is not unique. Moreover, a number of composite ring-system indices are available that provide access to any ring system present on a structure.

This suite also contains selected cyclic-system scaffold indices and bond-deleted functional-group indices. These serve as convenient bridging tools for quickly checking out ideas that might arise when ones attention to attracted to possibly more relevant relationships associated with scaffolds or functional groups.

This tutorial is largely self contained, but does assume some exposure to chemotypic indices, their naming codes and their use in structural browsing along the lines presented in the more basic MeqiSuite tutorial. The tutorial first describes how to compute the RingSystemSuite on a list of structure files. Browsing structures is possibly the most fundamental and straight-forward use of chemotypic indices. This subject, largely covered in the introduction to MeqiSuite, will not be revisited here other than to note some special considerations regarding browsing when using first-largest ring system indices. The core of the tutorial concerns the use of these indices for delineating active regions in lead-finding studies. After positioning this use with respect to what distinguishes the chemotypic approach to SA studies from other CADD approaches, the basic issue of finding and delineating activity regions is presented. The next section, optional on first reading, discusses the logical relationships amongst the indices that underlie their use. These relationships are then employed in the following section to show how to assemble critical structural comparisons for exploring of the boundaries of the active regions. The tutorial ends with a discussion on when and how one might quickly check if another category of chemotypes might provide a more apt chemotypic description of one or more active region. A complete listing of the indices and of their hierarchical relationships is given in Appendix 2.

Section 1. Generating the indices

To compute the RingSystemSuite indices on the files used in this tutorial, open the Meqi preprogrammed command template RingSystemSuite.xls. This template comes with a simple, but time-saving macro listed in Appendix 1. If your Tools|Macro|Security level is set to medium, you will be asked to accept the macro that comes with the template. Higher macro-security levels will disable the macro. The lowest level will accept it.

Once the preprogrammed template is running, check if the Excel 'Save As' command defaults to the Meqi directory. If not, appropriately change **the default** by saving the file to that directory. Now enter

```
MaybridgeTop10Act.sdf MaybridgeSmp.sdf
```

in the Structure-file-name cell A4 if these files have been copied into the Meqi directory. (These files can be downloaded from our website www.pannanugget.com.) Otherwise, you would enter their full path names. Although these two files will be used in the tutorial, any list of *.sdf files can be entered in the Structure-file-name cell A4.

Most *.sdf files have a field listing the registry numbers of the structures in the file. If the names of these fields are listed in the Registry-number-field cell C4, a column of registry numbers will be included in the output headed by the name of the first registry field in the list. In addition, if there is an activity or property table corresponding to each structure file and these tables involve the same columns, their names can be entered in the Input-table cell E4.

Two caveats: Meqi assumes the entries in any list in the command file are space delimited. Consequently, the file and field names in each of the preceding lists cannot include any spaces. Secondly, the input tables can have only one row per structure.

Generally, the rows of an input table will be in a one-to-one correspondence with the structures in the associated structure file. When this is not the case, that input table must have a column, say the j'th column, specifying the ranks of the structures to which its entries correspond and the value of j must be entered in the Selection-column cell F5. This selection column provides a very flexible means of restricting the RingSystemSuite calculations to a specific subset of structures in the structure files.

When multiple files are entered as in the case here, Meqi II creates a 'File' column whose values are '1' for records from the first file, '2' for records from the second file, *et cetera*. In the tutorial, this column is used to distinguish the active from the inactive structures.

Assuming the macro has been enabled, enter ctrl-m (press m while holding down the control key)¹; otherwise follow the instructions in the footnote. Now open or run the Meqi II executable. This generates the output table nuggets.csv used in the analysis presented in this tutorial

¹ The macro creates or updates the CSV Meqi command file meqi.csv. This is the only file that the Meqi II executable reads and from which it gets the commands it is to compute. (It is informative to browse the meqi.csv file in a text editor such as notepad.) After meqi.csv has been created or updated, the macro resaves the file as meqi.xls so that the formatting and annotations are not lost. The macro is easily created by simply recording the keystrokes needed to accomplish these two steps. To assure that the saved files reside in the current directory when the command template is copied to another directory, edit the macro and remove the path information from the file names.

If you wish to view the files in the command window created during the run, right-click its title bar and select properties, set the width of the screen buffer size in the layout option to at least 1000, and make any desired adjustments to the remaining layout options.

Section 2. Special considerations when browsing structures using largest ring systems

The use of hierarchical orderings when browsing structures is covered in the introduction to the MeqiSuite indices. The FirstRingOrdering or the SubstFirstRingOrdering indices are appropriate when ones interest is in largest ring systems. One simply has to play around with them and see which most suits the needs, available information, and constraints of a project. However, there are a couple of caveats when using first-largest indices when browsing structures and their significance is emphasized when looking for a ring-system template amongst a file of structures all of which are confirmed actives in some high-throughput screen (HTS).

If there are only a handful of actives, then any display of all of the actives provides all of the available information in a comprehensible format. But suppose there are a couple hundred actives. Then comprehensibility becomes an issue. The FirstRingOrdering index organizes the structures so that those structures sharing the first-largest ring system are displayed as a single group of structures. In a bar chart of this index, these structures will constitute the tally for a single bar. One such bar chart involving the 210 structures in the two files used in this tutorial is displayed in Figure 1 of the next section. Moreover, if two such groups have closely related first-largest ring systems, these two groups and their associated bars in a bar chart will be positioned close to each other. So far, so good. In addition, should ones interest in a ring-system template generally increase with the proportion of structures having that template, the relative heights of the bars in a FirstRingOrdering bar chart conveniently displays the relative count.

The caveats relate to this count and the concern that an interesting ring system not be overlooked. Only the first-largest ring system of a structure enters into the tallies. The first-largest ring-system tally for a large ring system will usually equal or fall just shy of the number of structures having that ring system. There is little chance of overlooking a large ring system.

However, for small ring systems, the tally will usually equal or be slightly larger than the number of structures in which that is the only ring system. There is little chance of overlooking a small ring system that usually occurs by itself when present on an active structure. However, there is a significant chance of overlooking a smaller ring system that generally occurs in conjunction with a larger ring system when present on an active structure. Observant browsing minimizes this risk by noting any attendant ring systems on the structures that have been grouped together as a consequence of their sharing a common first-largest ring system. When, in this context, an interesting smaller ring system serendipitously springs to mind, one can quickly obtain an unbiased tally of the number of active structures on which it occurs by doing a substring search for its CID in the values of the corresponding composite RngLst***Mqn index.

But suppose most of a large number of active structures involve multiple ring systems and all have roughly the same number of bonds. In this case, one might consider using the ScaffoldSuite of indices rather than the RingSysSuite.

When browsing active structures, there is one ring system with a RngFst***Mqn CID of RYLFV that will usually appear on more active structures than any other—benzene. This ring system may be ignored because it has been exhaustively explored. If not, it is still likely to be ignored due to its general prevalence in most compound libraries. But that prevalence is based on memory. It is not ascertainable from data involving only active structures.

This brings us to the second case in which one takes information on inactive structures into account as well, i.e. SA data. For the remaining part of this tutorial, we have a column indicating whether or not each structure in the file is active or inactive. One doesn't need all of the inactives from a screen, only a random sample of few hundred or so. With this random sample, each ring system has associated with it both the number of actives with that ring system and the number of inactives. The ratio of actives to inactives is likely to be much lower for the benzene ring system than for a ring system of interest.

Section3. Positioning the chemotypic approach to SA studies amongst other CADD approaches

When dealing with SA data, the amount of information a CADD approach takes into account greatly influences what one does and how one goes about it. Because it takes only a handful of indices into account, **QSAR** involves the least amount of information. It can, thereby, form quantitative parametric models of a SA relationship. However, the complexity of receptor-binding interactions often restricts the predictive domains of these models to regions of structure space specified by a single template or a small set of highly similar templates. By increasing the number of indices to a few thousand dimensions, **high-dimensional approaches** embed SA problems in a *single* space informationally rich enough to support similarity searching. Moreover, using fairly standard statistical tools, that space can be clustered and recursively partitioned in a manner that can handle diverse structural libraries. **Chemotypic analysis** uses substructural categories (cyclic systems, ring systems, functional groups, etc.) to form the equivalent of *hundreds* high-dimensional spaces, each informationally rich enough to support chemotypic searching for substructures within these categories. One could quickly get bogged down in an overload of spaces if it were not for the logical relationships that exist between the underlying chemotypic indices, and it is just these logical relationships that enable one to delineate activity regions and set up striking structural comparisons for exploring their boundaries. At the same time, the amount information largely forces one to treat activity regions individually. **Substructure analysis** places no restriction on the number and type of indicator variables that can be generated through substructure searching. However, the number of possible substructures and their lack any a priori organization largely restricts this approach to subject-matter specialists to define relevant hypotheses so as to whittle down the possibilities to a manageable number of queries. **Molecular modeling** takes into account information of such detail that it can resolve intramolecular interactions. However, the amount of information largely restricts attention to single structures and largely negates the types of statistical SA considerations of interest here.

With this general understanding in mind, the goal of chemotypic analysis in SA studies is two-fold. First one seeks a CID that delineates the activity regions of interest. Then one seeks closely related CIDs so as to collect interesting structural comparisons that enable one to explore and adjust the boundaries of that region. Unlike high-dimensional approaches which essentially analyze all of the activity regions with respect to a single molecular space, chemotypic analysis treats the activity regions one at a time with respect to hierarchically related chemotypic indices.

One begins by selecting a category of chemotypes apropos the class of problems one is addressing. If one is addressing a problem in which a single ring system is likely to be a dominating feature of the regions of active structures one expects to encounter, the chemotypic indices in the RingSystemSuite should prove helpful. The first-largest ring-system indices, and especially the FirstRingOrdering and SubstFirstRingOrdering indices, were selected with that focus in mind. If there are two or more active structures sharing a common largest ring system or having largest ring systems that differ only in their hetero atoms, they will be grouped together in a bar chart of either of these two hierarchical ordering indices. If those largest ring systems

constitute dominant structural features of those structures and an unusually high proportion of structures with those ring systems are active, then a ring-system activity region has been delineated. That it happens to be a largest ring system may or may not be critical to the definition of that activity region, but is an ancillary aspect of our use of first-largest ring systems.

Once a ring-system activity region has been delineated, the next step is to find structures having a first-largest ring system that differs in a minor way from those largest ring systems defining the active region. Such structures, if they are represented in the compound library, are often tallied on either side of the bars defining the active region. If an unusually high proportion of the structures are also active, the boundary of the activity region should be broadened. If not, the boundaries as delineated are confirmed and clarified.

Section 4. Delineating active regions in HTS data

We will illustrate the basic logic by means of data graciously provided in conjunction with the McMaster HTS data mining competition (http://hts.mcmaster.ca/competition_1.html). For the training set, roughly 50,000 structures obtained from Maybridge were screened DHFR inhibition activity. The first structure file, MaybridgeTop10Act.sdf, consists of the 10 most active structures. Follow-up work indicates that these actives bind competitively at the receptor site (M. Zolli-Juran, J.C. Cechetto, R. Hartlen, D.M. Daigle and E.D. Brown, *Bioorg. & Med. Chem. Lett.*, 13(2003)2493-2496). The second structure file, MaybridgeSmp.sdf, consists of 200 structures selected at random from the remaining structures, none of which were included in the 10 most active. See the first section on generating the RingSystemSuite for details on forming the data table used in this analysis.

As indicated in the section on generating the indices, the name of the file of 10 active structures was entered first in the A4 structure field followed by the name of the file of 200 randomly selected structures. As a consequence, Meqi generates a 'File' column in which a 1 (for the first file) corresponds to an active structure and a 2 (for the second file) corresponds to an inactive structure. This column will be used to differentially color code the active and inactive structures.

Figure 1 pictorially captures the essence of the chemotypic logic of finding an activity region and exploring its boundary. The rest of the tutorial could be viewed as an explanation of its construction, its origin and its hierarchical logic. It was constructed using Spotfire DecisionSite (www.spotfire.com). Both bar charts have some bars that run off the count scale as a result of the collapsed count slider on the left. The two bar charts are color coded with red for active and blue for inactive. The lower one-sixth of the bar of height 6 toward the left third of the upper bar chart is colored red to indicate one active and the upper five-sixths is colored blue to indicate five inactives. When a bar is "marked" a bar, only the edges of the bar retain the original coloring. The rest of the bar is colored aqua. In Figure 1, the bar of height 8 in the center of the lower bar chart was marked. The subtable corresponding to the marked structures is displayed in the "Details-on-Demand" window here situated at the bottom of the figure. In addition, the proportion of each bar in the other bar chart corresponding to those marked structures is colored aqua. As a final construction note, when the cursor is positioned at a bar, that bar is outlined with a rectangle and the corresponding x-axis coordinate is displayed in the lower-left corner of the figure. In the figure, the cursor is at the aqua bar of height 2 in the upper bar chart.

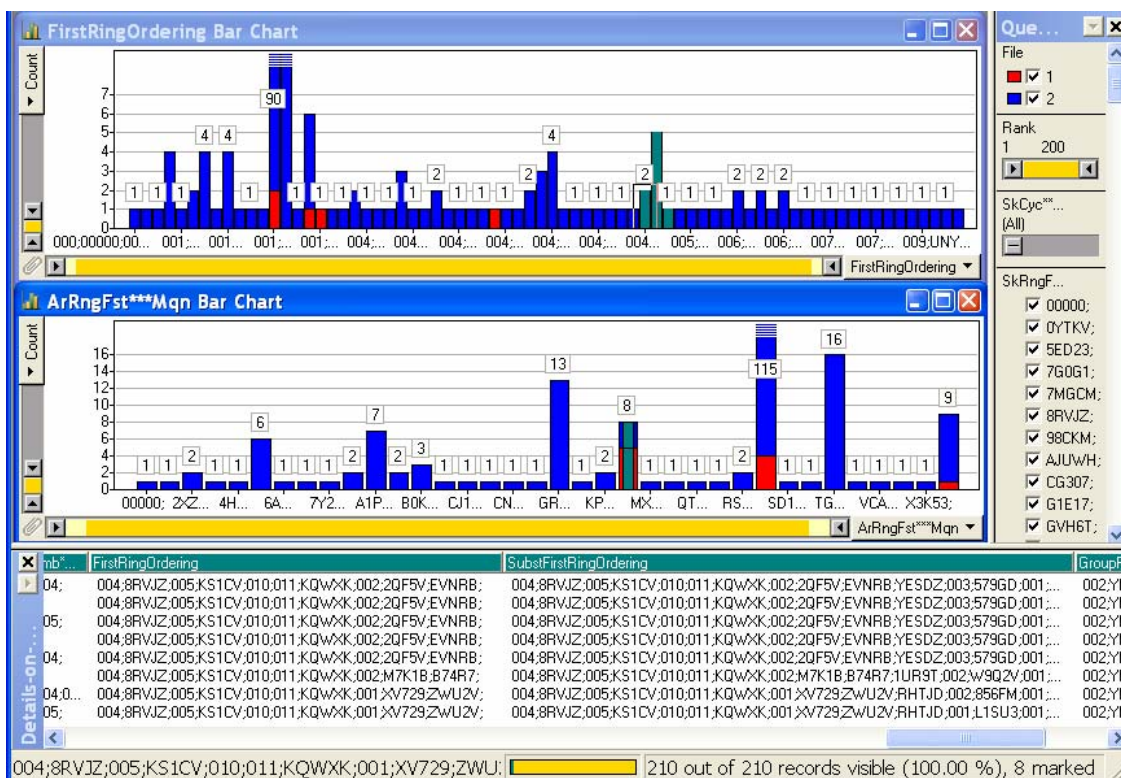


Figure 1. Bar charts of the FirstRingOrdering and ArRngFst***Mqn indices. In structures associated with the bar of height 8 in the lower chart were selected for marking. Contributions of these structures to the bars in the upper chart are indicated with an aqua color. The values for the FirstRingOrdering for the marked structures are given in the table. These values represent the X-axis coordinates of the marked structures in the upper chart. One of these coordinates is given in the lower left-hand corner of the figure and corresponds to the bar of height 1 that is framed by a rectangle in the upper chart.

The upper bar chart hierarchically orders the structures with respect to their largest ring system using the FirstRingOrdering index. It quickly reveals the relative locations of the actives within this hierarchical ordering. The 10 actives are confined to five bars. The tallest bar, cutoff in the figure, contains 88 inactives in addition to the two actives corresponding to the red portion at the bottom of the bar. These 90 structures are all substituted benzenes some of which have other ring systems with six or fewer bonds.

As noted earlier, the first-largest ring system associated with this bar could be ruled out either because benzenes have already been exhaustively considered, or, as is the case here, because of its low proportion of actives. Because of the low proportion of actives, one might expect that the substructures responsible for the activity of these two actives involve something other than or in addition to the benzene ring. A quick glance of their structures, given in Figure 2, suggests this is the case.

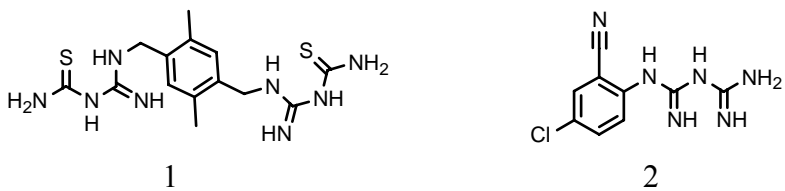


Figure 2. The two active structures associated with the tall “benzene” bar in the upper chart of Figure 1.

I expect the reader quickly formulated a hypothesis as to the substructures that might be responsible for the activities of structures 1 and 2. But how? One certainly didn’t enumerate the thousands of distinct substructures associated with the two structures, a good number of which might very well be incorporated into the representations used by the high-dimensional CADD methods discussed earlier. But one might have restricted his or her attention to the 5 distinct side chains, the one common ring system, and the four distinct functional groups [five if methyl is viewed as a functional group] and unconsciously employed the following **chemotypic principle**:

Critical differences in chemical interactions are correlates of chemotypic changes.

In addition, if the substructure or class of substructures in this hypothesis involves one or both of the two largest functional groups, the reader very likely reasoned that these groups are relatively rare and consequently unlikely to appear on two actives unless they have something to do with that activity. If so, the reader made use of a type of argument that we have attempted to make explicit by including some randomly selected inactives into our analysis.

The next two actives are tallied in two adjacent bars shortly to the right of the “benzene” bar. Our use of the words “adjacent” and “shortly to the right” suggest these activity regions will be associated with simple, 6-membered aromatic ring systems. Their structures are given in Figure 3.

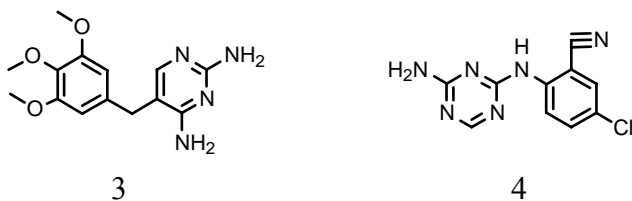


Figure 3. The two actives structures associated with the two adjacent bars slightly to the right of the tall “benzene” bar in the upper chart of Figure 1.

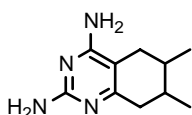
One might ask why these structures are not tallied in the “benzene” bar. It’s strictly a matter of ancillary factors affecting the sequencing of the atoms and bonds in the connection tables of the structures. Had another individual registered these structures, one or both of them might have been tallied in the “benzene” bar. Consequently, we do not really know how many of our 210 structures involve the two hetero ring systems in Figure 3.

But we can find out, and using a dynamic visualization package such as Spotfire DecisionSite, it’s quite easy. One first highlights either one of the two bars and then looks at the values for the FirstRingOrdering in the Details-on-Demand window shown at the bottom of Figure 1. The last entry in that semi-colon delimited list is the CID of the first-largest ring system, i.e. the value of the RngFst***Mqn index. For the bar of height 6 containing structure 3, that value is ‘1C3JF’ corresponding to the pyrimidine ring system. The composite or “list” counterpart, RngLst***Mqn, of that index has a value of ‘1C3JF;RYLFV;’. This composite ring-system

index contains the CIDs of all of the ring systems in a structure. Consequently, if we do a '1C3JF' substring search of the values of this index (In Spotfire DecisionSite, one uses the using the 'Full Text Search' query device.), we obtain all of the structures containing the desired pyrimidine ring system. There turns out to be one additional pyrimidine. It is inactive.

One would now do the same for the triazene ring system of structure 4. In this case there are no additional triazenes. Because of their close relatedness (which is reflected in the adjacency of their corresponding tally bars), it would make sense to think of them as defining a single ring-system based activity region.

Continuing from left to right, the structures increase in size and complexity with respect to their first-largest ring systems. The next active structure constitutes the sole tally in the red bar of height 1 just left of center in the upper chart. Its structure is given in Figure 4.



5

Figure 4. The active structure tallied in the red bar of height 1 just right of center in the upper chart.

The next active bar tallies the remaining five actives. Ordinarily this bar would be all red. However, it and its two neighboring bars of height 2 and 1 are aqua for reasons to be explained shortly. The five structures, close analogs of Trimethoprim, are depicted in Figure 5.

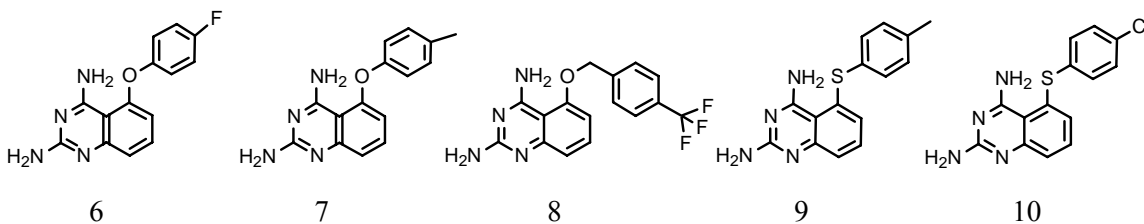


Figure 5. The five structures tallied in the middle aqua bar in the upper bar chart of Figure 1.

This finishes the first phase of isolating the active regions. The first region associated with Figure 2 no longer concerns us as the ring-system category seems inappropriate to the nature of the structures that are active in that region. The second region associated with Figure 3 involves two different ring systems. Treating an active region defined by two different ring systems creates issues that will be more easily addressed at the end of the next section. The latter two regions are delineated by a single ring system. We will address the trimethoprim region associated with Figure 5 as it involves a single ring system represented by multiple actives.

As in an actual lead-template selection, we have triaged the activity regions and selected one. Our basis for triaging the active regions for the purposes of this tutorial is rather idiosyncratic, but the fundamental outcome is the same. Attention is directed to a small number of regions that merit further development. As noted earlier, the amount, detail and structuring of the information restricts our attention to one region at a time. The next section gives a glimpse of the scope and structuring of this information.

Section 5. Hierarchically relating component indices of the RingSystemOrdering

The upper bar chart reflects the information in the FirstRingOrdering index. As noted earlier, when a bar is "marked" in Spotfire DecisionSite, its color is changed to aqua and the data records

for the associated structures are displayed in the details window. In Figure 1, the bar of height 8 in the lower ArRngFst***Mqn chart is marked. We address the reasons for doing so in the next section. Here we only note that the lower table in Figure 1 shows the values for the two columns of the nuggets.csv table corresponding to the FirstRingOrdering and the SubstFirstRingOrdering indices.

We see from Figure 1, that each value for the FirstRingOrdering index is a semi-colon delimited list of 10 terms. Just as names are hierarchically ordered in a phone book by family name and then by first name, the lexicographical ordering of these values along an axis in a plot hierarchically orders the values by their first term, then their second terms, and so on and so forth down to the last term. It was not an accident that the two structures in Figure 3 were tallied in adjacent bars (given that their first-largest ring systems were pyrimidine and triazene). It was a consequence of the construction of this hierarchical ordering. Such orderings are only one aspect of the hierarchical relationships amongst the indices that play such a critical role in this analysis.

What now follows is a rather formal discussion of the logic underlying those relationships. Although it helps to understand this logic, especially for those who wish to construct their own hierarchical ordering indices, the reader can skip to the next section should he or she only be interested in the operational arguments underlying our chemotypic analysis of the boundaries of an activity region described by a ring-system CID.

We'll begin our discussion of these relations in the context of the FirstRingOrdering index. Each of the 10 terms in one of its values is either a count or a CID defined by the *i*'th component index of this hierarchical ordering. The complete listing is given below and can always be found in the Meqi output file summary.txt.

FirstRingOrdering

- 1 SkRngFst**RedAct
- 2 SkRngFst**RedMqn
- 3 ArRngFst**RedB#4
- 4 ArRngFst**RedMqn
- 5 SkRngFst***Act
- 6 ArRngFst***B#4
- 7 ArRngFst***Mqn
- 8 HtArRngFst*Lay*A#Ht
- 9 HtArRngFst***Mqn
- 10 RngFst***Mqn

Much of the construction of this list is based on the idea of resolving structural distinctions. An index *I* is said to resolve another index *J* if whenever two structures have the same value for index *I*, they necessarily have the same value for index *J*. Said another way, if index *I* **resolves** index *J*, then index *I* distinguishes any two structures that index *J* distinguishes. Visually, when going from index *I* to index *J*, one “zooms out”, and conversely, when going from index *J* to index *I*, one “zooms in”. We will write $I < J$ when index *I* resolves index J^2 .

² Mathematically, those structures that have the same value for an index constitute an equivalence class of that index. If index *I* resolves index *J*, i.e. $I < J$, then any equivalence class, say *A*, of index *I* is always a subset of some equivalence class, say *B*, of index *J*, i.e. $A \subset B$.

Clearly, if index I resolves index J and if index J resolves index K, then index I must resolve index K. Because of this critical transitivity property, we can capture all of the hierarchical relationships amongst the indices in the FirstRingOrdering by the following six expressions:

- (1) SkRngFst**RedMqn < SkRngFst**RedAct.
- (2) ArRngFst**RedMqn < ArRngFst**RedB#4
- (3) ArRngFst***Mqn < SkRngFst***Act
- (4) ArRngFst***Mqn < ArRngFst***B#4
- (5) HtArRngFst***Mqn < HtArRngFst*Lay*A#Ht
- (6) RngFst***Mqn < HtArRngFst***Mqn < ArRngFst***Mqn < ArRngFst**RedMqn < SkRngFst**RedMqn.

Note that whenever one index resolves another in the above relationships, the former index follows the latter in the FirstRingOrdering. More importantly, the FirstRingOrdering satisfies the following and much stronger **CID hierarchical ordering property**:

Every CID component index in the FirstRingOrdering index resolves every index that precedes it in the listing of those component indices.

It follows from expression 6 that every CID component index in the FirstRingOrdering index resolves every *CID* index that precedes it in the listing of those component indices³. But how does one know, for example, that the last RngFst***Mqn index resolves the third ArRngFst**RedB#4 *count* index in the listing? That follows from expressions 2 and 6 and repeated use of the transitivity property.

Informative as they are, count indices have very low levels of resolution relative to CID indices, and, consequently, count indices seldom resolve CID indices. For example, the eighth index, HtArRngFst*Lay*A#Ht, in the FirstRingOrdering component listing does not resolve the seventh index, ArRngFst***Mqn. It is easy to think up different ring systems that have the same number of hetero atoms. Still the count indices are important component indices of the FirstRingOrdering, as, unlike CID indices, they provide directions as to where structures may lie along an axis. For example, the first-largest triazenes whose HtArRngFst*Lay*A#Ht value is 3 lie to the right of the first-largest pyrimidines whose HtArRngFst*Lay*A#Ht value is 2.

The first implication of these hierarchical relationships concerns our definition of an activity region. Recall from the upper bar chart that all five structures associated with the middle aqua bar were active. These five structures, depicted in Figure 2, can be viewed as representatives of a first-largest ring-system activity region if there are not a whole lot of other inactive structures in our data set sharing the same CID for one of the first-largest ring-system indices. Of course this is unlikely given the size of the ring system and the diversity of structures screened, but it is

³ This argument begs the question as to the validity of expression 1-6. Much of the intuitive understanding related to the mathematical proof of these expressions can be found in the introduction to the MeqiSuite indices. This validity also relies on the ‘first-encounter’ method of breaking ties that was discussed in the introduction.

informative to see why these five structures are the only ones that can have this ring system for their first-largest ring system.

By the very nature of the bar-chart construction, all five structures have the same ordered value for the FirstRingOrdering index and consequently must share the same value for each of its 10 component indices. As seen in Figure 1, this ordered value is given by

(7) 004;8RVJZ;005;KS1CV;010;011;KQWXK;002;2QF5V;EVNRB;

Now suppose there was another structure having EVNRB for its RngFst***Mqn value that is tallied in some other bar. Its x-axis coordinate would differ from expression 7, and consequently it would differ in value for at least one of the component indices of the FirstRingOrdering index, say, for example, the ArRngFst**RedB#4 count index. But we have just shown that the RngFst***Mqn index resolves the ArRngFst**RedB#4 count index. Since this supposed structure agrees with the other five structures in its RngFst***Mqn value of EVNRB, it must also agree in its ArRngFst**RedB#4 count. Thus, the supposed existence of this structure leads to a contradiction.

There may be other structures having a EVNRB ring-system chemotype, but if so, those structures cannot have EVNRB for their *first-largest* ring-system chemotype. They must also have another ring system as large as or larger than the EVNRB ring system with 11 bonds. We find out that this is not the case by searching the composite RngLst***Mqn index for the substring EVNRB.

Section 6. Finding critical structural comparisons at the boundary of an active region

The preceding section ended with a proof of a critical point. Each bar in the upper chart of Figure 1 corresponds to all, not just some, of the structures having a particular ring system for its first-largest ring system. This is important. The RingSystemOrdering not only grouped together the active regions associated with the two closely related pyrimidine and triazene ring systems in Figure 3, it assured us that the six structures tallied in the “pyrimidine” bar and the one structure tallied in the “triazene” bar were the only structures having those two ring systems as their first-largest ring system. This assurance flows from the **CID hierarchical ordering property** established in the preceding section. This property holds for all of the six hierarchical orderings in the RingSystemSuite and plays a critical role in the following examination of the boundary of the trimethoprim activity region associated with Figure 5.

Every CID defines a boundary separating the structures with that chemotype from those without it. It is natural to seek small structural changes made to the first-largest ring system that will result in structures lying just outside that boundary. A preponderance of active structures so obtained suggests that the boundary of the active region should be extended, whereas a preponderance of inactive structures supports the boundary as defined. Either way, such structural comparisons are helpful to ones understanding of the SA relationships.

One obtains such structural comparisons by ignoring some of the structural detail captured by the RngFst***Mqn index. This is done in the lower bar chart of Figure 1. To see how the index for that chart was selected, expression 6 from the preceding section is reproduced here.

RngFst***Mqn < HtArRngFst***Mqn < ArRngFst***Mqn < ArRngFst**RedMqn < SkRngFst**RedMqn

This expression lists five indices. The indices are ordered from left to right by the amount of information they take into account with the left-most index, RngFst***Mqn, taking the most into

account. Expressed another way, if two structures have the same CID for an index to the left of another index, they will also have the same CID for the latter index, i.e., any index on the left “resolves” any index on the right. The higher-resolution indices lie to the left in this ranking; the lower-resolution indices lie to the right. This ranking is reflected in the ordering of these indices in the listing of the component indices of the FirstRingOrdering in the preceding section and partially accounts for its CID hierarchical ordering property.

The relationship between the upper and lower bar charts in Figure 1 reflects the fact that the index for the lower bar chart is selected from one of the six CID indices in this resolution ranking. As noted in the caption to Figure 1, all of the structures tallied in the bar of height 8 in the lower chart were marked, thus changing the color of that bar to aqua. As a consequence of this marking, Spotfire DecisionSite will color aqua any bar in the upper bar chart by an amount that corresponds to the proportion of the tallied structures that correspond to marked structures. But notice! There are no partially-colored aqua bars. Moreover, the aqua-colored bars constitute a contiguous group. This is by construction. It follows from the CID hierarchical ordering property. Regardless of which of the five indices in the above ranking are selected for the lower chart, when one of its bars are marked, the aqua coloring corresponding to the marked structures in the upper bar chart will consist of a contiguous group of solid aqua bars. Scientifically, what constitutes a single group of structures sharing a common chemotype in the lower chart corresponds to a contiguous region of structures in the upper, hierarchically organized chart.

But how did we know which bar to mark in the lower chart? This takes us back to the end of section 4 when we had just decided to focus on the active region of trimethoprim analogs. Per the discussion of figure 5, that region would have appeared to us as a red bar of height 5 in the upper chart. On marking that bar, Spotfire DecisionSite would change its color to aqua and would proportionally recolor aqua that portion of any bar in the lower chart corresponding to marked structures.

Here again, the relationship between the two charts as a consequence of the CID hierarchical ordering property comes into play. This marking of a single bar in the upper chart will always result in only one bar in the lower chart being recolored aqua when that lower chart was based on any of the indices in the above resolution ranking. In the case of the ArRngFst***Mqn index used in figure 2, five eighths of the bar of height 8 was colored aqua.

This partially colored bar in the lower chart is the bar we seek. When this bar is marked, we obtain Figure 1 with three inactive structures lying just outside of the active region. We’ll take a look at these structures shortly, but it helps to examine the consequences of using some of the other indices in the resolution ranking for the lower chart in that figure.

Had we used the HtArRngFst***Mqn index in the lower chart when marking the trimethoprim bar in the upper chart, our attention would have been directed to a bar of height 5 in the lower chart containing just the five trimethoprim actives. That this would be so can be seen from the values of the FirstRingOrdering in the table of Figure 1. Each is a semi-colon delimited list of 10 terms. The last or tenth term is the value of the RngFst***Mqn index. The first five rows corresponding to the active structures agree in the value EVNRB of this term, but differ from the last three rows with respect to this term. An analogous statement holds with respect to the second to last term or ninth term which is the value for the HtArRngFst***Mqn index. However, all eight rows agree in the value KQWXX of the seventh term which is the value for the ArRngFst***Mqn index. In fact, if one thinks about it, KQWXX must be the x-coordinate of the bar of height 8 that is marked in the lower chart. (Note that this bar lies between ‘KP...’ and ‘MX...’ in the lower chart in Figure 1.) Marking the bar of height 5 in the upper chart and then

chart. The structures in Figures 5 and 6 have two hetero atoms and are tallied in the two following aqua bars. Counts used in this manner in a hierarchical ordering not only help one to locate structures within that ordering, but, not infrequently suggest indices of use in QSAR models.

Suppose we zoomed out one more level of resolution to that defined by the ArRngFst**RedMqn index, we encounter a second relevance boundary that includes the eight structures in the first relevance boundary plus 17 additional inactives. The structures associated with the five bars of height 1 that immediately precede the aqua bar of height 2 in the upper chart of Figure 1 are given in Figure 7.

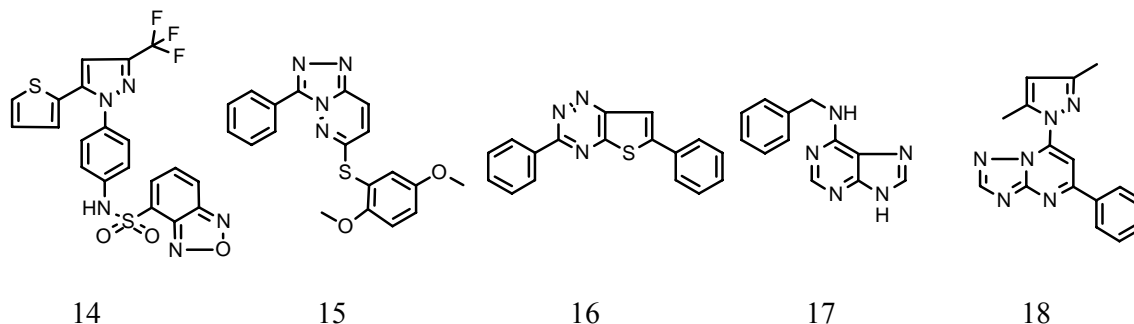


Figure 7. The structures associated with the five bars that immediately precede the aqua bar of height 2 in the upper chart of Figure 1. The structures are sequenced according to the sequencing of their corresponding bars in Figure 1.

We see that the structures tallied in the upper chart of Figure 1 immediately to the left of the aqua bar associated with structures 11 and 12 in Figure 6 have fused 5 and 6-member aryl rings for their first-largest ring systems. The ordering of these ring systems is “explained” by looking at their corresponding values for the FirstRingOrdering. These values are ordered lexicographically in Table 1 exactly as they are ordered in Figure 1.

Table 1. The lexicographical ordering of the values of the FirstRingOrdering of the structures in Figure 7.

Structure number	FirstRingOrdering
14	004;8RVJZ;005;KS1CV;009;010;TG6J3;003;V3ZN3;GBPP1;
15	004;8RVJZ;005;KS1CV;009;010;TG6J3;004;21YT3;EANJ7;
16	004;8RVJZ;005;KS1CV;009;010;TG6J3;004;4P2C1;D7PD5;
17	004;8RVJZ;005;KS1CV;009;010;TG6J3;004;P9MNZ;BNR5K;
18	004;8RVJZ;005;KS1CV;009;010;TG6J3;004;UH215;ZGTQF;

Note that the values of in Table 1 are identical to those at the bottom of Figure 1 up to the fourth term. This fourth term, KS1CV, is the CID assigned by the ArRngFst**RedMqn index to aromatic ring systems comprised of two fused rings. However, they differ in the fifth term which is the value of the SkRngFst***Act index. It is this count index that causes the FirstRingSystem bars associated with the structures in Figure 7 to precede those bars associated with the structures in Figures 5 and 6.

The structures in Figure 7 raise a number of interesting points. Note that the fused ring system can be thought of as being on the “outside” in structures 14 and 17, but on the “inside” in the other three structures. This is a notion captured in the “Layering” option of Meqi. In this option, taking structure 14 as an example, the side-chain bonds are assigned to layer 1. The bonds of the thiophene ring and the fused ring system are assigned layer 2. The bonds of the bridges to the layer-2 ring systems are assigned layer 3. The bonds of the benzene and diazole ring systems are assigned to layer 4, and the bond of their connecting bridge is assigned layer 5.

The structures in Figure 7 are grouped together with the trimethoprim analogs because the `ArRngFst**RedMqn` index does not take distinctions in layering into account. The `ArRngFst*LayRedMqn` index does. This index would assign the same value to structures 14 and 17 that it assigned to the trimethoprim analogs because in these cases the fused ring system at issue here is assigned layer 2. However, it is assigned layer 4 in structures 15, 16 and 18. Consequently, they would receive the same `ArRngFst*LayRedMqn` value, but one that differs from the value assigned the trimethoprim analogs.

The `ArRngFst*LayRedMqn` index is not included in the `RingSystemSuite`. This suite is designed to handle many, but not all issues related to ring system chemotypes. However, it has served a useful purpose whenever it usefully redirects ones attention to possibly more relevant chemotypes.

Section 7. Interfacing with other ideas and approaches

What we have done might be viewed as one suggestion for a first-pass analysis of one activity region. Chemotypic analysis is a new field. Exploration is just beginning on the many issues that are involved. In this final section, we briefly touch a few issues related to how these first-largest ring system indices interface with other chemotypic indices and other methodologies so as to better address the goal of characterizing an activity region and the related need of gathering together the most relevant structural comparisons by which that characterization is facilitated.

Once the EVNRB chemotype was seen to correspond with an activity region, the next step was to zoom out one level using the `HtArRngFst***Mqn` index. However, a search of the corresponding 2QF5V CID didn't turn up any new structures. It might have had we augmented our 10 most active structures with a sample of 5,000 rather than the 200 additional structures included in our data set.

This raises the issue of how many inactive structures should be included in an analysis of this type and how they should be selected. One might say, “The more, the merrier.” We could have included all 49,995 structures in the analysis. An obvious upside would be more structural comparisons at every level of resolution. For example, had we included all of these inactive structures, there would have been 118 inactives tallied in the trimethoprim bar giving rise to the five structures in Figure 5. These would constitute interesting structural comparisons overlooked by our small sample of 200 inactives.

But there are also significant downsides to looking at a very large file of inactives. Meqi processes the Maybridge structures at a rate of 1,200 per minute on a 1.68 GHz IBM ThinkPad. When processing time becomes a factor, that time should be focused on those structures most relevant to the issue.

More importantly, visualization apprehension and physical interaction times can significantly degrade with larger files. The physical interaction times for a visualization package between a user-selected option and displayed result usually increase at least proportionally with the number

of records in the data file. Even a doubling of an interaction time from 10 seconds to 20 seconds can significantly change the nature of what one does and attempts to do with a visualization package.

Visual apprehension is limited by human memory and depreciates with mental fatigue. Consider the 3 structures in Figure 6. Had we included all 49,995 structures in our analysis, there would have been 1428 structures in Figure 6. A hundred structures are a lot to look at and mentally retain when making structural comparisons, and often, only a dozen or so critical comparisons can constitute a solid basis for many SA decisions.

From this writer's perspective, the issue is one of selecting from a large collection of inactive structures those that are particularly relevant. These issues, much less any proposed solutions, lie outside the scope of this introduction other than to make a few general comments.

First, the 200 structures selected at random provide an "unbiased" and sufficiently reliable ranking of the active regions as regard their relative concentration of actives. However, one could have selected a second set of inactives with a high degree of similarity to the actives, especially amongst those structures falling within the relevance boundaries as defined by the various first-largest CID indices provided by RingSystemSuite. The "interesting" structural comparisons would then largely come from this second set of inactives.

Second, a dynamic visualization package such as Spotfire DecisionSite provides sliders and other options for subsetting data based on the values of the variables in the data set. Suppose one creates a relevance boundary that circumscribes an awkwardly large number of structures. RingSystemSuite has a number of scaffolding indices that could then be used to filter out all but a small subset of the most relevant structures.

Third, here we have focused on the first-largest ring system. If that is ones sole interest, no more needs be said. Otherwise, one might also consider looking at some other substructure categories. To this end, RingSystemSuite provides additional hierarchical orderings based on the group of ring systems, the cyclic-system, the group of bond-deleted functional groups, or the first-largest bond-deleted functional group as a means to quickly suggest a better focus. This tutorial will close with a brief look at how these orderings can direct ones attention to other, possibly more relevant, substructure categories.

The component indices of these orderings are listed below:

GroupRingOrdering

- 1 ArRngLst***Cct
- 2 ArRng***Mqn
- 3 HtArRng***Mqn
- 4 Rng***Mqn

CyclicSystemOrdering

- 1 ArRngLst***Cct
- 2 SkCyc***Mqn
- 3 ArCyc***Mqn
- 4 HtArCyc***Mqn
- 5 Cyc***Mqn

GroupFuncGrpOrdering

- 1 BfgLst*Emb*Cct
- 2 Bfg**Emb*Mqn
- 3 Bfg*AaEmb*Mqn

FirstFuncGrpOrdering

- 1 BfgFst*Emb*Mqn
- 2 BfgFstAaEmb*Mqn

If one of the preceding categories or treatments be clearly worth consideration, a bar chart of the relevant hierarchical ordering should quickly make that evident. We will illustrate one such chart using the last of these orderings, but first want to show why the others can be ruled out in this particular case involving the activity regions for DHFR inhibitors.

We first note that both the cyclic system category and the ring systems treated as a group take into account more information than do the corresponding first-largest ring system indices. This will tend to split up the activity regions when ones general goal is to find structural commonalities that define larger activity regions. For example, the structures in Figure 5 involve three different, although closely related, cyclic systems.

Matters are even worse when focusing on the functional groups as a group as is done by the Bfg**Emb*Mqn index. In this case, all five structures involve a different set of functional groups. In fact, all ten structures involve a different set of functional groups.

However, one is immediately struck by the fact that the five structures in Figure 2 all share a large functional group. Consequently, a bar chart of FirstFuncGrpOrdering index is worth exploring. This bar chart is juxtaposed with that of the FirstRingOrdering in Figure 4.

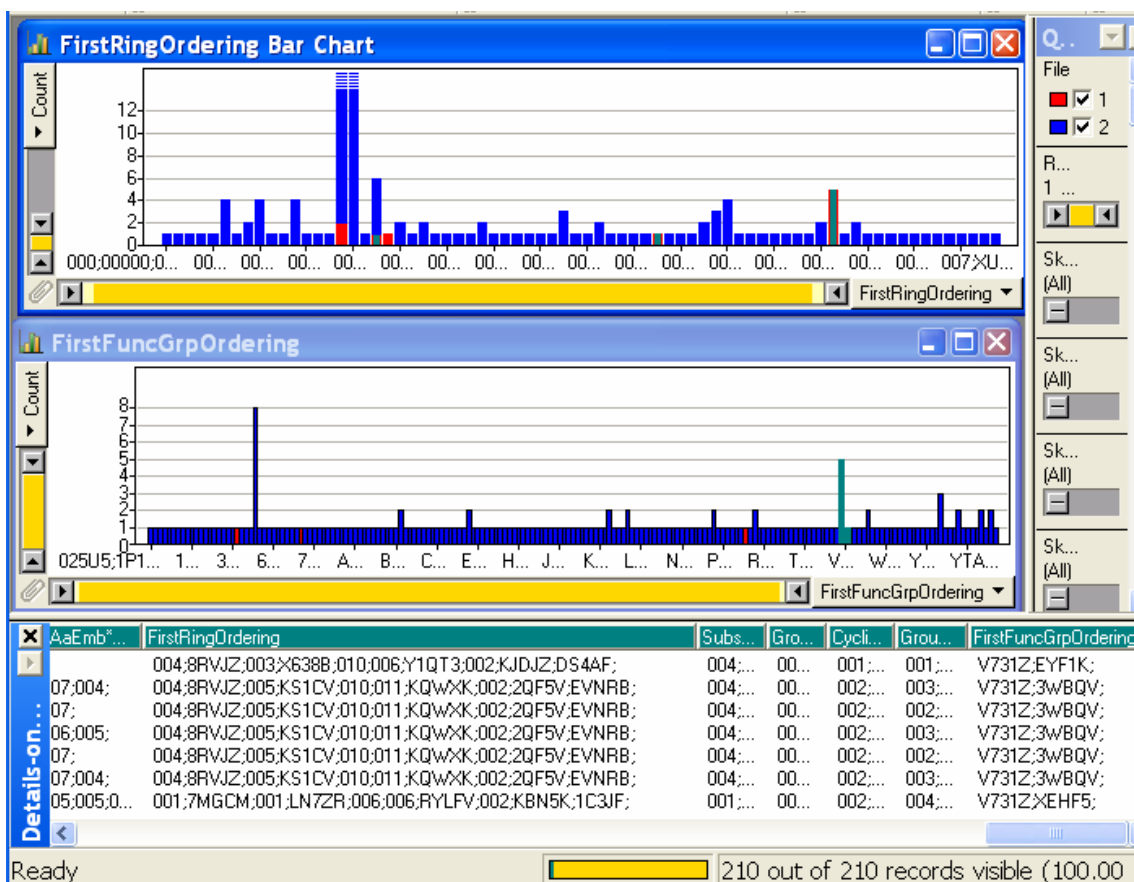
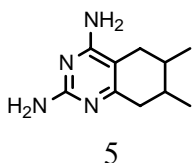


Figure 8. Bar charts of the FirstRingOrdering and FirstFuncGrpOrdering indices. Contributions of active structures signified by the red portion of unmarked bars and by the aqua portion of marked bars. The values for the FirstRingOrdering and FirstFuncGrpOrdering in the lower table are those of the structures tallied in the three contiguous marked aqua bars in right portion of the lower FirstFuncGrpOrdering bar chart, one of height 5 and two of height 1. All of the marked structures are active.

Figure 8 is striking. When the activity proportion is small (10 of 210 structures) and there are a lot of bars, it is unlikely for two or more actives to be tallied in adjacent bars if the structural features being taken into account on those actives are not correlated with the underlying mechanism of activity. The aqua colored region in the lower bar chart was all red prior to marking the structures in that active region.

The adjacency of two bars in a bar chart of a single CID index is purely accidental. Similar, but non-isomorphic substructures from the same substructure category will usually be assigned CIDs quite differently positioned along a lexicographical ordering, and non-similar substructures can be assigned adjacent CIDs along that ordering. So it is worth considering how the contiguity of the aqua bars might arise. This is readily done by looking at the values of the FirstFuncGrpOrdering in the lower right-hand corner. They all have V731Z for their first term, i.e. they share the same V731Z chemotype defined by the component index BfgFst*Emb*Mqn of the FirstFuncGrpOrdering.

But just what is this chemotype? That is best answered by looking at the structure 5 in Figure 4 and reproduced here.



One obtains the “bond-deleted functional groups” (BDFGs) for that structure by deleting all of its carbon-carbon aromatic or single bonds. The fragments that remain are the BDFGs. Structure 5 has only one such group. It has been assigned V731Z for its BfgFst*Emb*Mqn CID. It has the same skeleton as 3-methylhexane. However, this index takes atom and bond typing into account, consequently, the four aromatic implies that this functional group must have four bonds residing on an aromatic ring. A number of other issues come into play when determining which structure have and which structures do not have this V731Z chemotype.

As with the EVNRB chemotype of the RngFst***Mqn index in Figure 1, the V731Z chemotype does not occur on any of the 200 inactives. However, unlike the EVNRB chemotype which is shared by 118 of the remaining 49,995 structures, the V731Z chemotype is shared by only 5 of the remaining structures! This is strong evidence that the DHFR inhibitor region is much more aptly described by a BDFG chemotype than a first-largest ring-system chemotype. These five structures are depicted in Figure 10. They make for some interesting structural comparisons with the seven actives residing in this V731Z activity region.

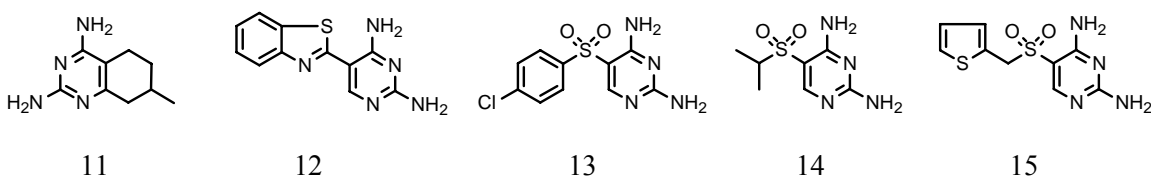


Figure 9. The remaining structures having the V731Z bond-deleted functional-group chemotype.

The FirstFuncGrpOrdering, crude as it is, has served its purpose of directing our attention to another substructure category—that of bond-deleted functional groups. At the time of this writing, a preprogrammed Meqi template for generating a suite of indices devoted to this substructure category along lines similar to the current RingSystemSuite has not yet been developed. Once in place, the natural thing would be to repeat with that suite the types of reasonings and operations presented in this tutorial.

Appendix 1

```
Sub SaveMeqiCommandFile()  
,  
' SaveMeqiCommandFile Macro  
' Macro recorded 6/30/2005 by Mark Johnson  
,  
' Keyboard Shortcut: Ctrl+m  
,  
' ChDir "D:\__programming\MEQI\Meqi_2.x"  
  ActiveWorkbook.SaveAs Filename:= _  
    "meqi.csv", FileFormat:=xlCSV, _  
    CreateBackup:=False  
  ActiveWorkbook.SaveAs Filename:= _  
    "meqi.xls", FileFormat:=xlNormal, _  
    Password:="", WriteResPassword:="", ReadOnlyRecommended:=False, _  
    CreateBackup:=False  
End Sub
```

Appendix 2

The names of all indices in the RingSystemSuite are listed in Table 2 along with a brief description of the information taken into account and any section in which that index is discussed. This is followed by a listing and notes of the hierarchical relationships amongst the indices as regards their resolving power.

Table 2. Listing of the naming codes and descriptions of the 55 indices in RingSystemSuite.

Numbers 50-55 are orderings developed from the preceding 49 indices. The ordering and section columns list the numbers of the ordering indices and the sections in which that index is used or discussed. Indices 1-9, whose names begin with 'Sk', are computed after changing all atom types to carbon and all bond types to single. Indices 10-17, whose names begin with 'Ar', are computed after changing all atom types to carbon and all single, double and triple bond types to single, but leaving aromatic bond types unchanged. Indices 18-25, whose names begin with 'HtAr', are computed after changing N, O, S and P atom types to a hetero-atom type of 'Ht', changing F, Cl, Br, and I to a halogen-atom type of 'Hl' and all remaining atom types to carbon. In addition, all single, double and triple bond types to single, but leaving aromatic bond types unchanged. Indices 26-49 are computed using the original atom and bond types.

#	Code name	Description	Ordering	Section
1	SkCyc****Mqn	Skeleton of the cyclic system	55	
2	SkRngFst**RedMqn	Skeleton of the first-largest, reduced ring system	52,53	5,6
3	SkRngFst**RedAct	Number of atoms in the first-largest, reduced ring system	52,53	5
4	SkRngFst***Mqn	Skeleton of the first-largest ring system		
5	SkRngFst***Act	Number of atoms in the first-largest ring system	52,53	5,6
6	SkRngFstF**Mqn	Attachment-flagged skeleton of the first-largest ring system		
7	SkRngFstF**A#*	Number of attachment flags in the substructure of index 6	53	
8	ArCyc****Mqn	Aryl skeleton of the cyclic system	55	
9	ArRng****Mqn	Aryl-skeleton ring-system group	54	
10	ArRngLst***Mqn	List of aryl-skeleton ring systems		
11	ArRngLst***Cct	Number of ring systems	54,55	
12	ArRngFst**RedMqn	Aryl skeleton of the first-largest, reduced ring system	52,53	5,6
13	ArRngFst**RedB#4	Number of aryl bonds in the substructure of index 14	52,53	5
14	ArRngFst***Mqn	Aryl skeleton of the first-largest ring system	52	4,5,6
15	ArRngFst***B#4	Number of aryl bonds in the substructure of index 16	52	5
16	HtArCyc****Mqn	Hetero-aryl cyclic system	55	
17	HtArRng****Mqn	Hetero-aryl ring system group	54	
18	HtArRng**Lay*Mqn	Hetero-aryl ring system group with layering		
19	HtArRngLst***Mqn	List of hetero-aryl ring systems		
20	HtArRngLst*Lay*Mqn	List of hetero-aryl ring systems with layering		
21	HtArRngFst***Mqn	First-largest hetero-aryl ring system	52,53	5,6,7
22	HtArRngFst*Lay*Mqn	First-largest hetero-aryl ring system with layering		
23	HtArRngFst*Lay*A#Ht	Hetero-atom count for first-largest ring system	52,53	5
24	Cyc****Mqn	Cyclic system	55	

Table 2. Listing of the naming codes and descriptions of the 55 indices in RingSystemSuite.

Numbers 50-55 are orderings developed from the preceding 49 indices. The ordering and section columns list the numbers of the ordering indices and the sections in which that index is used or discussed. Indices 1-9, whose names begin with 'Sk', are computed after changing all atom types to carbon and all bond types to single. Indices 10-17, whose names begin with 'Ar', are computed after changing all atom types to carbon and all single, double and triple bond types to single, but leaving aromatic bond types unchanged. Indices 18-25, whose names begin with 'HtAr', are computed after changing N, O, S and P atom types to a hetero-atom type of 'Ht', changing F, Cl, Br, and I to a halogen-atom type of 'Hl' and all remaining atom types to carbon. In addition, all single, double and triple bond types to single, but leaving aromatic bond types unchanged. Indices 26-49 are computed using the original atom and bond types.

#	Code name	Description	Ordering	Section
25	Rng***Mqn	Ring-system group	54	
26	Rng**Lay*Mqn	Ring-system group with layering		
27	Rng*Aa**Mqn	Alpha-augmented ring-system group		
28	Rng*AaLay*Mqn	Alpha-augmented ring-system group with layering		
29	RngLst***Mqn	List of ring systems		2,4,5
30	RngLst*Lay*Mqn	List of ring systems with layering		
31	RngLstAa**Mqn	List of alpha-augmented ring systems		
32	RngLstAaLay*Mqn	List of alpha-augmented ring systems with layering		
33	RngLst***Act	Atom counts of the ring systems		
34	RngLstAa**Act	Atom counts of the alpha-augmented ring systems		
35	RngFst***Mqn	First-largest ring system	52,53	2,4,5,7
36	RngFst*Lay*Mqn	First-largest ring system with layering	53	
37	RngFstAa**Mqn	First-largest alpha-augmented ring system		
38	RngFstAaLay*Mqn	First-largest alpha-augmented ring system with layering	53	
39	BfgFst*Emb*Mqn	First-largest BDFG with embedding	57	7
40	BfgFstAaEmb*Mqn	First-largest alpha-augmented BDFG with embedding	57	
41	Bfg**Emb*Mqn	Group of BDFGs	56	7
42	Bfg*AaEmb*Mqn	Group of alpha-augmented BDFGs	56	
43	BfgLst*Emb*Mqn	List of BDFGs with embedding		
44	BfgLstAaEmb*Mqn	Group of alpha-augmented BDFGs with embedding		
45	BfgLst*Emb*Cct	Number of BDFGs	56	
46	BfgLst*Emb*Act	Atom counts of the BDFGs		
47	BfgLstAaEmb*Act	Atom counts of the alpha-augmented BDFGs		
48	FirstRingOrdering	Hierarchical ordering of the first-largest ring systems		2,3,4,5,6,7
49	SubstFirstRingOrdering	Hierarchical ordering of the first-largest alpha-augmented ring systems		2,3,5
50	GroupRingOrdering	Hierarchical ordering of the groups of ring-systems		7

Table 2. Listing of the naming codes and descriptions of the 55 indices in RingSystemSuite.

Numbers 50-55 are orderings developed from the preceding 49 indices. The ordering and section columns list the numbers of the ordering indices and the sections in which that index is used or discussed. Indices 1-9, whose names begin with 'Sk', are computed after changing all atom types to carbon and all bond types to single. Indices 10-17, whose names begin with 'Ar', are computed after changing all atom types to carbon and all single, double and triple bond types to single, but leaving aromatic bond types unchanged. Indices 18-25, whose names begin with 'HtAr', are computed after changing N, O, S and P atom types to a hetero-atom type of 'Ht', changing F, Cl, Br, and I to a halogen-atom type of 'HI' and all remaining atom types to carbon. In addition, all single, double and triple bond types to single, but leaving aromatic bond types unchanged. Indices 26-49 are computed using the original atom and bond types.

#	Code name	Description	Ordering	Section
51	CyclicSystemOrdering	Hierarchical ordering of the cyclic systems		7
52	GroupFuncGrpOrdering	Hierarchical ordering of the groups of BDFGs		7
53	FirstFuncGrpOrdering	Hierarchical ordering of the first-largest BDFGs		7

Prefix for the different types of structure representations:

UnSk - Unextended ring and bridge systems (R&BSs) used in conjunction with the hydrogen-reduced graph representation.;

Ar - All atoms are treated as carbons and all bonds are treated as single or aromatic;

HtAr - All atoms are treated as hetero, halogen or carbon and all bonds are treated as single or aromatic;

Code for the substructure category: * - complete structure; Cyc - cyclic system; Rng - ring systems; Bfg - bond-deleted functional groups (BDFGs);

Code for the treatment: * - group; Lst - positioned list; Fst - largest chemotype;

Code for substructure attachment considerations: * - substructure; F - attachments flagged; Aa - alpha-augmented with attachment positions

Code for substructure positioning considerations: * - none; Lay - positioned with respect to layering; Emb - positioned with respect to both layering and linkage;

Code for generalizing the topology: * - none; Red - unsubstituted chains reduced to single bonds

Code variable types: Mqn - CID index; Cct: # of chemotypes; Act - # of atoms; A#Ht - # hetero atoms; A#HI - # of halogens; B#4 - # of aromatic bonds; A#* - # of attachments.

Appendix 3

As discussed in section 5, an index I is said to resolve another index J, and we write $\text{index } I < \text{index } J$, if whenever two structures have the same value for index I, they necessarily have the same value for index J. The resolution relation $<$ is reflexive, i.e. $\text{index } I < \text{index } J$, and transitive, i.e. $\text{index } I < \text{index } J$ and $\text{index } J < \text{index } K$ implies $\text{index } I < \text{index } K$.

Interestingly, a group index and its list counterpart generally assign different values to the same structure, they create the same structural classes, i.e. they both resolve each other. Thus, for example, $\text{Rng}^{***}\text{Mqn} < \text{RngLst}^{***}\text{Mqn}$ and $\text{RngLst}^{***}\text{Mqn} < \text{Rng}^{***}\text{Mqn}$. When this is the case, we shall say the two indices are equivalent, and write, for example, $\text{Rng}^{***}\text{Mqn} \equiv \text{RngLst}^{***}\text{Mqn}$.

This resolution relation enables us to deduce an extremely large number of hierarchical relationships from a few simple expressions involving the code names of the indices. Each code name is a sequence of seven subcodes, the first of which is an optional prefix describing the optional processing of the structure in each section of the preprogrammed RingSystemSuite template. In what follows, the subcode ‘null’ indicates the prefix is omitted. The subcodes are given in the notes to Table 2 in Appendix 2. Any term between the two braces can substitute for the right-hand term of the resolution relation.

- (8) $\text{null}^{????}\text{Mqn} < \text{HtAr}^{????}\text{Mqn} < \text{Ar}^{????}\text{Mqn} < \text{Sk}^{????}\text{Mqn}$
- (9) $?Cyc^{????}\text{Mqn} < ?Rng^{????}\text{Mqn}$
- (10) $?Lst^{????}\text{Mqn} < ??*^{????}\text{Mqn} < ??Fst^{????}\text{Mqn}$
- (11) $???Aa^{??}\text{Mqn} < \{???*^{??}\text{Mqn}, ???F^{??}\text{Mqn}\}$
- (12) $????Emb^{?}\text{Mqn} < ???Lay^{?}\text{Mqn} < ???*^{?}\text{Mqn}$
- (13) $????*^{?}\text{Mqn} < ???Red^{?}\text{Mqn}$

In these expressions, any subcode can be substituted for the appropriate ‘?’ symbol. For example, in expression 6 of Section 6 we find $\text{RngFst}^{***}\text{Mqn} < \text{HtArRngFst}^{***}\text{Mqn}$. This follows from the relation $\text{null}^{????}\text{Mqn} < \text{HtAr}^{????}$ in expression 8. From that same expression 6 we find $\text{ArRngFst}^{***}\text{Mqn} < \text{ArRngFst}^{**}\text{RedMqn}$. This follows from expression 14.

In expression 10, the relation $??*^{????}\text{Mqn} < ??Fst^{????}\text{Mqn}$ may fail to hold when the first-largest substructure is not uniquely determined. Any failure to hold arises from the artifact related to our use of the “first-encounter” method of breaking ties. The relation $?Lst^{????}\text{Mqn} < ??*^{????}\text{Mqn}$ always holds. However, the fact that the reverse relation $??*^{????}\text{Mqn} < ?Lst^{????}\text{Mqn}$ doesn’t hold as well is due to a related artifact that a list of chemotypes for a structure is sequenced by the order in which they are encountered. Our use of this “as encountered” sequencing is justified by some important analytical benefits discussed in the introduction to MeqiSuite. Had the chemotypes been ordered canonically in their listing, the reverse relation would hold as well for the chemotypes when treated as a list adds no structural information that is not already encoded in the chemotypes when treated as a group.

When it comes to the resolution relations that hold between a CID index and one of its count index counterparts, we have the following expressions:

$$(15) \quad ??*???Mqn < \{??*???Cct, ??*???Act, ??*???A\#x, ??*???B\#x\}$$

$$(16) \quad ??Lst???Mqn < \{??Lst???Cct, ??Lst???Act, ??Lst???A\#x, ??Lst???B\#x\}$$

$$(17) \quad ??Fst???Mqn < \{??Fst???Act, ??Fst???A\#x, ??Fst???B\#x\}$$

Again, any term between two braces can be used as the right-hand term of the resolution relation.

The resolution relations that hold between two count indices are subsumed by the following general expression:

$$(18) \quad ??Lst???? < \{??*????, ??Fst????\}.$$