

Meqi II: A Template-Based Programming Environment for Computing Chemotypic Indices

Mark Johnson (mark@pannugget.com), Pannugget Consulting, Kalamazoo, MI

Abstract

Scaffolds, cyclic systems, ring systems, functional groups and side chains are examples of chemotypic categories. A chemotypic index specifies for each structure the substructural components falling in one of these categories along with selected information as regards the positioning of these components.

Meqi II is a template-based programming environment for creating ASCII command files that when operated on by the Meqi executable will generate a CSV table of desired chemotypic indices.

This poster begins with a brief introduction to the notion of a chemotypic index, presents the basic programming template and closes with some simple, but illustrative applications of some typical chemotypic indices.

Chemotypic Indices

Traditional QSAR methods typically provide parametric models on highly related structures. High-dimensional methods provide predictive models on diverse compound collections embedded in a single vector space. **Chemotypic methods provide hierarchically related indices for organizing any compound collection in which each index is the informational equivalent of a high-dimensional vector space.**

The substructural components underlying chemotypic indices are defined with respect to structural properties of atoms and bonds. Two such bond properties and their corresponding substructure components (ring systems and bond-deleted functional groups (BDFGs)) are given in Figure 1 for an illustrative compound.

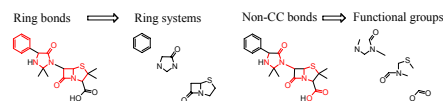


Figure 1. Bonds having the indicated property are highlighted.

A substructural component and its positioning is captured in a short character string called a chemotypic identifier (CID). Two structures that have the same substructural component may or may not share the same CID depending on what aspects of positioning are taken into account. This is illustrated in Figure 2 for four BDFG indices. All five structures are tertiary amines. Yet the sharing of their tertiary amine CIDs, highlighted in red, depends on the index by which those CIDs are defined.

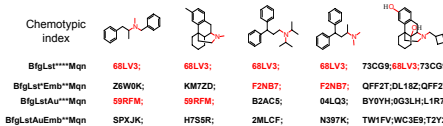


Figure 2. List (Lst) values for four BDFG (Bfg) indices. Two indices involve embedding (Emb) and two involve alpha-augmented (Au) BDFGs. Shared CIDs are highlighted.

The Meqi II Programming Template

Different ways of taking position into account and various useful transformations on the underlying structural representations give rise to hundreds of possible chemotypic indices. Meqi II provides a highly annotated template-based interface for quickly specifying and computing these indices. **The execution of the command file generated through this interface can be incorporated into batch files and other software.**

To illustrate, Bemis and Murcko (J.Med.Chem, 39(1996)2887) dissected molecules "into four units: ring, framework, linker, and side chains" as part of an effort to tabulate their occurrence in pharmaceuticals. Their frameworks and linkers correspond to our cyclic-system skeletons and bridge-system skeletons. The programmed template for generating the corresponding chemotypic indices is given in Figure 3. Distinctions in atom and bond types were ignored (See cells C15 and C16).

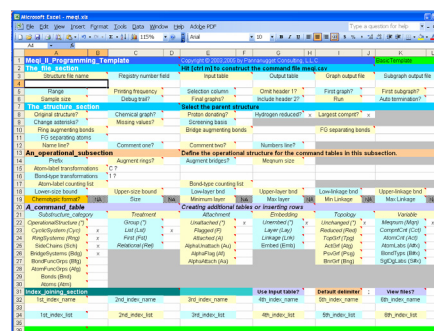


Figure 3. A Preprogrammed Meqi II Programming Template.

The programmed template in Figure 1 can be applied to any SD structure file or files simply by entering their name or names in cell A4 before generating the command file and running the Meqi executable. MeqiSuite is a programmed template designed to generate 62 "command" indices and 4 indices that hierarchically order structures using combinations of the command indices. **Meqi II computes the 66 MeqiSuite indices at roughly 500 structures per minute on a 1.86 GHz Pentium processor.** The remaining plots, created using Spofire DecisionSite (www.spofire.com), involve only MeqiSuite indices.

Applications

Although there are hundreds of chemotypic indices, any given structure has only a handful of substructural components. Over half of the important substructural components of the structure in Figure 1 are displayed. Adding its complete structure, its cyclic system and its three side chains contributes the remaining five. Yet, the general applicability of chemotypic analysis is premised on the chemotypic principle: **Critical differences in chemical interactions are correlates of chemotypic changes.** The applicability of this principle is illustrated in the following applications abstracted from our upcoming publication "Introduction to MeqiSuite."

The figures involve four sets of 200 structures selected at random from the Lancaster catalog of chemicals for synthetic chemists (Figs. 4 & 5), the Merck Index of pharmaceuticals (Fig. 4), the Maybridge (Figs. 4 & 6) and Chembridge (Fig. 4) screening libraries used in the McMaster data-mining competition along with the 10 DHFR-inhibitory hits (Fig. 6) that bound competitively (http://hts.mcmaster.ca/HTSDataMiningCompetition.htm) Figures 4 and 7-9 also involve the 50 oxime analogs of the Acton and Stone sweetener study (Science, 193(1976)2887).

Figure 4 compares five compound libraries with respect to their cyclic-system aryl skeletons in which aromatic and aliphatic bonds are distinguished. **Note how the Maybridge (dark blue) and Chembridge (light blue) training and test data sets may represent quite different regions of structure space.** As expected, the oxime analog series (red) is concentrated in a small region of structure space.

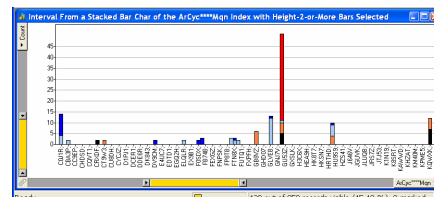


Figure 4. Stacked bar chart comparing five compound libraries

The values for a list index can contain multiple CIDs as is seen for the last structure in Figure 2. The Meqi table operations allows the user to output a list index as a high-dimensional array. **Selecting the most relevant index to output is facilitated by plots involving "first-largest" indices.** The value of such an index for a structure is the CID of the first chemotype in that structure's chemotypic list that has as many atoms as any other chemotype in that list. In Figure 2 the first-largest BDFG for the last structure is the tertiary amine with four atoms.

Figures 5 and 6 are first-largest BDFG plots. The upper red bar of height four in Figure 5 confirms that isothiocyanates are lachrymatory while the lower red bars of heights 2 and 4 confirm that thiols and thioethers, respectively, are malodorous. Figure 6 shows that 7 of the 10 competitive binding hits in the McMaster training data set share a common BDFG.

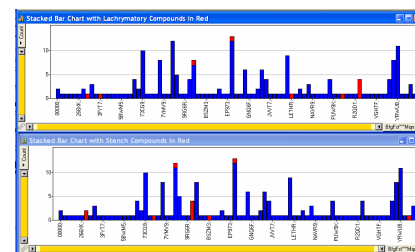


Figure 5. Stacked bar charts of a first-largest BDFG index.

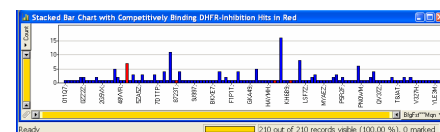


Figure 6. Stacked bar chart of a first-largest BDFG index.

A small structural change accompanied by a large difference in activity signals a structure-activity cliff so often critical to lead discovery and optimization. **Hierarchical orderings and skeletal indices expose structure-activity cliffs.** Besides making clear that the cyclic-system greatly influences percent sweetness, the dramatic drop off in percent sweetness amongst the seven marked structures in Figures 7 and 8 directs attention to the meta vs para-substituted structure-activity cliff. Side-chain and cyclic-system cliffs based on structures differing only in atom or bond labeling are signaled in five of the skeletal groupings in Figure 9.

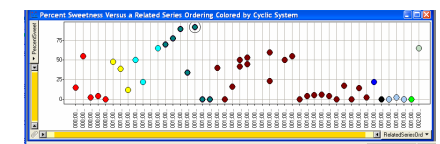


Figure 7. Hierarchical ordering of the oxime analogs.

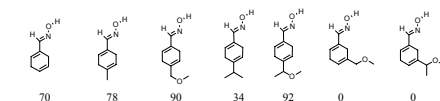


Figure 8. The sequence of seven structures marked in Figure 7 accompanied by their percent sweetness values.

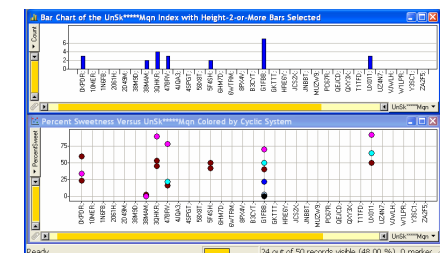


Figure 9. Structures are grouped by common skeleton.

Summary & Comment

Chemotypic analysis opens up many new opportunities for formal analysis methods in drug discovery. The underlying Chemotypic indices incorporate a wide-variety of chemical functionality. They are intuitive and eminently interpretable. They are part of an emerging methodology in chemometrics rich with possibilities. **Meqi II is a simple and flexible template-based programming language that puts the discovery of the relevant chemotypic indices in the hands of the chemometrician, and, as the preceding applications suggest, "The hits just keep on coming!"**

For more information, contact mark@pannugget.com or visit www.pannugget.com.