

Specifying and Using Hierarchically-Organized Structural Categories

Mark Johnson (mark@pannanugget.com), Pannanugget Consulting, Kalamazoo, MI

Abstract

The **molecular structure** and its **cyclic systems**, **ring systems**, **side chains** and **functional groups** are structural categories of common interest in chemistry. A molecular equivalence index or **meqi** is an index whose value for a structure is a list of the names of the substructures of a compound in one of these structural categories.

Each name reflects a number of independent representational features that result in **hundreds of flavors** of meqis in each of the preceding **five basic categories**. Within each representational feature, the choices are hierarchically organized so that one can organize meqis by their levels of resolution. We present and illustrate the five basic categories, the different representational features and the basic hierarchical organization of meqis.

We also illustrate a **new feasibility** in ranking the many meqi flavors regarding their suitability as a basis for **selecting screening libraries** and show that functional-group meqis offer exciting and here-to-fore **unexplored opportunities** in that regard.

Introduction

The notion of the molecular structure and its cyclic systems, ring systems, side chains and functional groups are basic to how a chemist perceives and discusses compounds and has been a part of computational chemistry since its inception. Computationally assigning useful names to these entities has been in the literature since the development of the Morgan algorithm (Morgan, J. Chem. Doc. 5(1965)107). (See Xu and Johnson, J. Chem. Inf. Comp. Sci., 41(2001)181 for more references)

Applications in which these names are becoming part of the design and analysis of structure-activity studies are just starting to emerge. Nilakantan et al. (J. Chem. Inf. Comp. Sci., 30(1990)65; Comb. Chem. & HTS, 5(2002)105) explored ring systems and cyclic systems. Bemis and Mucko explored most of the preceding categories using the skeletal labeling of the atoms and bonds (J. Med. Chem., 39(1996)2887) and studied side-chains using concepts of layering (J. Med. Chem., 42(1999)5095). Johnson and Xu (In Chemical Data Analysis in the Large, ed. M.G.Hicks, Logos Verlag, Berlin, 2000,67; J. Chem. Inf. Comp. Sci. 42(2002)912) developed the notion of a molecular equivalence index and its visual analysis and explored functional groups and cyclic systems as a means of characterizing the difference between two chemical libraries.

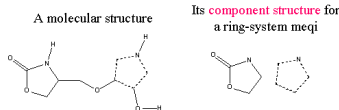
Here we illustrate the basic **concept of a molecular equivalence index**, present the five basic categories and their many flavors, discuss aspects of their hierarchical relationships and then present a pilot study of two meqis from each category with respect to their **relevance in designing screening libraries**.

Molecular Equivalence Indices or Meqis

A meqi is an index whose value for a structure is a list of the names of the substructures of a compound in a "natural" and well-specified category such as its molecular structure, cyclic systems, ring systems, side chains and functional groups. Other natural categories could include its bridge systems or conjugation systems, but we will be focusing on the first five.

A meqi has two basic components: the substructures that define it and the molecular equivalence numbers or **meqnums** assigned to those substructures either as a group to form a unitary meqnum or individually to form a list of meqnums called a composite meqnum.

A molecular equivalence index has an "interpretable" component structure and a molecular-equivalence number (meqnum)



Its **unitary meqnum**: A single name for the ring systems taken jointly

ASP6B

Its **composite meqnum**: A list of the names for the individual ring systems

KYTT5 NBR57

Interconverting a Functional-Group List with a Massively Columned Table of Frequency Variables

Compound number	Functional-group list	Names of functional-group count variables				
		Acid	Amide	Ketone	Sulfide	Amine
1	ketone	0	0	1	0	0
2	sulfide sulfide	0	0	0	2	0
3	amide amine	0	1	0	0	1
4	acid	1	0	0	0	0
5	acid amine amine	1	0	0	0	2

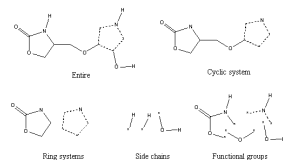
A meqi is a **high-content variable**. Each value is a list of keywords. Each variable is equivalent to a high-dimensional array of molecular descriptors. Because it is a single-columned variable, it can be fully represented by a **single axis of a plot** amenable to **visual analysis**.

Five Basic Categories

Meqis are fundamentally classified by their natural component structures. One speaks of a ring-system or functional-group meqi.

The entire structure and its cyclic system, ring systems, side chains and functional groups constitute five basic categories that should find significant applications in drug discovery.

Meqis are fundamentally classified by their natural component structures

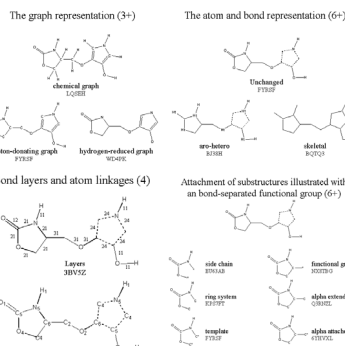


Hundreds of Flavors

The value of the computed meqnum computed on a substructure reflects the underlying graph representation and the atom and bond labels.

However, general layering information (side chain, first layer of ring systems, first layer of bridge systems, next layer of ring systems...) can be incorporated into the bond labels and general linkage information (connected to only side-chain edges, only ring edges, only side-chain and ring edges,...) can be incorporated into the atom labels.

In addition, substructures can be augmented by bonds alpha to the substructure and the attachment atoms appropriately distinguished.



These four aspects are implicit in the calculation of any meqnum. As pictured below, these aspects have at least 3, 5, 4 and 6 options. Thus there are at least 3 x 6 x 4 x 6 or **432 flavors** of ring-system meqis, functional-group meqis, *et cetera*.

Hierarchically Organized

Two compounds that have the same cyclic system necessarily have the same set of ring systems, but not conversely. That's just another way of saying that any cyclic-system meqi resolves the corresponding ring-system meqi, but not conversely.

The concept of hierarchical resolution extends to the basic underlying graph and the preceding four aspects of a meqi's representation. Any meqi based on the chemical graph resolves the corresponding meqi based on the hydrogen-reduced graph. Any meqi using the atom and bond labels of the parent graph resolves the corresponding meqi using those of the skeletal graph. Similar statements can be made regarding each of the remaining three aspects of meqi representation.

New Feasibilities

Although there are hundreds of meqis, each is easily specified by simply selecting an option for each aspect of a meqi's specification. The following table illustrates one possible option array.

Table 1. By simply selecting one option from each category, a single meqi out of hundreds is unambiguously specified.

Graph	Substructure	Atom labels	Bond labels	Embedding	Attachment
chemical graph	entire	original	original	none	side chain
proton-donating	cyclic system	hetero	aromatic	layering	ring system
hydrogen reduced	ring systems	skeletal	skeletal	linkage	template
	side chains			both	functional group
	functional groups				alpha-extended alpha-attached

The Triple Crown

The thirteen contests

Willett and Wintermann's classic study (QSAR, 5(1986)18) compared molecular similarity measures based on their nearest neighbor prediction performance across a number of chemical properties and activities. Here we compare the performance of meqis based on their likelihood of containing the name of a substructure (here called a **recognitional substructure**) clearly associated with a therapeutic mechanism of a compound.

The data: 200 structures were selected at random from the Twelfth Edition of Merck Index and put in a database along with the designation of their therapeutic categories. Thirteen of these therapeutic categories contained at least five compounds.

Each of the thirteen therapeutic categories suggests a contest. To win a contest, it helps a meqi to involve substrings (meqnums) associated with a recognitional substructure in that therapeutic category. The following two conditions constituted a meqi having such a substring:

- the substring must occur on at least 2 compounds,
- over half of the compounds with that substring must be in that therapeutic category.

For each meqi, the number of hits (compounds in that therapeutic category containing at least one of its suggested recognitional substructures) is tallied. Coming in first with the most hits constitutes a **win**. Coming in first or second constitutes a **show**. A meqi wins the **triple crown** if it wins all thirteen contests. There were no triple crown winners this round, but there were good and bad contestants.

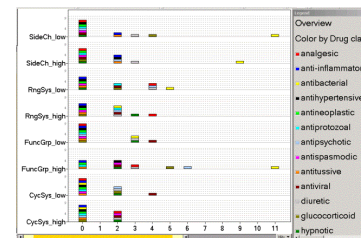
In case of ties, the first condition is relaxed so that a substring need only occur on one compound. The hits are recomputed. Of the tied meqis, that one with the most hits under this relaxed condition wins. If there is still a tie, neither meqi is awarded a win, but both meqis (there were never more than two ties at this point) are awarded a show.

The eight contestants and their performance

The meqi contestants are listed in the y-axis of the figure.

All of the meqis use the hydrogen-reduced graph with the original atom and bond label options in Table 1.

The "low" resolution meqi in each substructure class uses the "none" embedding option and "ring system" attachment option. The "high" resolution meqi in each substructure class uses the "both" embedding option and the "alpha attached" attachment option.



Ties reduced the number of cases of wins to 10. For shows, the number of cases was 12 as no meqi found a recognitional substructure for the antineoplastic category. Under the assumption that each meqi had the same chance of winning any one contest, the probability is only 1/8 that it will. Under this assumption, the number of wins for both the RngSys_high and FuncGrp_high meqis were greater than would be expected by chance factors alone.

When shows are taken into account, the **FuncGrp_high meqi clearly stood out**. Moreover, it was significantly better than its low-resolution counterpart in the head-to-head comparison.

Interestingly, this meqi always had a **hit rate of 75% or better** under the weaker "tie-breaker" conditions.

Table 2. Tallying the wins and shows.

Meqi	Wins ¹	Shows ²	L vs H ³ Wins
SideCh_low	0	3	2
SideCh_high	2	3	5
RngSys_low	1	4	4
RngSys_high	3**	4	4
FuncGrp_low	0	2	2
FuncGrp_high	3**	6**	7**
CycSys_low	0	0	3
CycSys_high	1	1	3

- Significance is based on one-sided binomial test with $p = 1/8$ and $n = 10$.
 - Significance is based on one-sided binomial test with $p = 1/4$ and $n = 12$.
 - Comparisons are between the low resolution meqi (L) with its high resolution counterpart (H). Significance is based on two-sided binomial test with $p = 1/2$ and n being 7, 8, 9 and 6 for the four top to bottom pairings.
- ** : significant at the 0.025 level.

Software used

ISIS from www.mdli.com was used to store and write structures; DecisionSite from www.spotfire.com was used for above figure; Meqi from www.pannanugget.com was used to construct the indices and determine the recognitional substructures.