Information and Organic Molecules:

Structure Considerations Via Integer Statistics

Daniel J. Graham
Loyola University Chicago

Email: dgraha1@luc.edu

Qualitative Organic Analysis = ***"Guess What is in the Bottle"***

Text: *Systematic Identification of Organic Molecules* by Shriner, Fuson, and Curtin.

Primary Reference: *Beilstein Index*

Requirements for Success: Ability to play chemist plus detective.

Hindsight:

1. **Organic Qualitative Analysis is Really Organic Quantitative Analysis!!**

Is there one compound in the bottle? Y/N
Is the compound soluble in water? Y/N
When the compound is burned, is the flame sooty? Y/N
When a little bromine is added to the compound, does the color disappear? Y/N

. Y/N
. Y/N
. Y/N

Is the compound inexpensive? Y/N

☞ **Different compounds express different amounts of information. This information can be quantified in bits.**

2. There are an enormous number of different compounds which express information--

~ 2×10^6 in 1972.

Where are We Now?

1. Still Asking Qs about molecules in order to quantify the information expressed by organic molecules--although the questions are much more sophisticated.

2. ~ 20×10^6 molecules in database libraries--and counting.

***

$20^{100} \approx 10^{130}$ possible 100 residue proteins from the naturally occurring amino acids.

A 100 residue protein expresses roughly

$3^{100} \approx 5 \times 10^{47}$ conformations.

☞**Organic Molecules offer infinite possibilities.**

Now Working On...

Information expressible by organic molecules, given infinite possibilities.

Base Information Content in Organic Molecules, Graham and Schacht, *J. Chem. Information and Computer Sci.* **2000**, *43*, 187.

Approach focuses on molecular descriptors not obvious to qual organic lab courses.

***

☞ What is the total number of complexions Ω_t

(= the number of possible electronic messages)

expressed by the molecule?

Not a Y/N Question, but....

☞ Ω_t provides a handle on the maximum possible regio-information expressible by the molecule.

Why should one care about Ω_t?

$CH_3CH_2CH_2COOH$ OH^-

$CH_2=CHCH_2CH_3$ Br_2

$CH_3CH_2CH_2CH_2CH_3$

Let all information processing, e.g. interaction with a reagent, involve random sensing of complexions/electronic messages.

The probability that *a particular complexion* will be sensed is $1/\Omega_t$. The probability p_1 that a sensed complexion will be allied with region 1 is

$$p_1 = \frac{\Omega_1}{\Omega_1 + \Omega_2}$$

(with analogous expression for p_2).

The Shannon measure μ (= the number of bits) associated with the two complexion groups is

$$\mu = -K \sum_{i=1}^{i=D} p_i \ln (p_i)$$

$$= -K [p_1 \ln(p_1) + p_2 \ln(p_2)]$$

μ-values are enhanced when the Ω_i are nearly equal and are diminished when the Ω_i are markedly disparate. ($K = 1/\ln(2)$)

μ = 0 bits for any molecule lacking in composite behavior.

Models and chemical structure theory provide ways of *estimating* D and Ω_t.

Integer statistics offer insight regarding the maximum possible D and μ, given values of Ω_t.

Integer statistics can tell us what is likely and unlikely in organic molecules, given infinite possibilities.

The fundamental theorem of arithmetic...

holds that any integer can be expressed as a unique product of prime numbers (2, 3, 5, 7, ...).

e.g. 120 = 2•2•2•3•5.

☞ **A molecule expressing Ω_t complexions demonstrates a *maximum* D-value equal to the number of prime factors of Ω_t.**

The maximum possible regio-information μ for a molecule follows from Shannon's formula. It depends exclusively on Ω_t having D number of prime factors f_i:

$$\mu(\Omega_t) = - K [p_1 \ln (p_1) + p_2 \ln(p_2) + ... + p_D \ln(p_D)]$$

$$= - K \sum_{i=1}^{D} p_i \ln p_i$$

where

$$p_i = \frac{f_i}{(f_1 + f_2 + ... + f_D)}$$

Integer statistics based on elementary number theory establish the distribution properties of both ***D*** and $\boldsymbol{\mu}$. In so doing, the traits distinguishing high regio-information molecules are highlighted.

The statistical approach encompasses all possible molecules and conditions, not just those reported to date in chemical databases.

The aim is to view regio-information and organic molecules in an alternative and general way.

Multiple chiral centers assure alliance of a molecule with the highest quantiles.

For example....

If $\Omega_t = 2^N$, N being the number of chiral centers

$\boldsymbol{D} \approx N$ and $\boldsymbol{\mu} \approx K\ln(N)$ bits.

For a protein with $N \approx 100$, $\boldsymbol{\mu} \approx K\ln(100) \approx 6.644$ bits. Such a molecule is allied with the $(1 - 10^{-30})$ quantile.

☞ The statistical perspective reinforces our present conceptions of high information molecules, namely natural products, biopolymers, and crystals with high degrees of chirality.

The lowest information molecules are those lacking functional groups and asymmetric carbons.

☞ Integer statistics offer a framework for molecular regio-information. This framework quantifies the maximum case scenarios along with their distribution properties.

Current efforts are focused on the statistics of information processing in molecular recognition and chemical reaction events.