

How much chemistry does *your* computer know?

John Bradshaw

Daylight CIS Inc.

johnb@daylight.com

Outline

- History of QSAR
- Designing 'well-spread' sets
- *de novo* design
- Increasing screening capacity
- Increasing compound supply
- Normalizing structures
- Chemoinformatics, the rebirth of QSAR?

Underlying assumption

“Since the development of structural theory and stereochemistry from the mid to late 19th century, to understand a molecular compound and to account for its physical, chemical and biological properties, has been very much a matter of knowing and understanding its structure.”

E. Francour *HYLE*, 6(1), 2000

Everything we do in (Q)SAR assumes that there *is* a relationship between the “S” and the “A”.

The big step forward was realizing that knowledge of structure *per se* was not sufficient.

Around 1850, von Liebig, Laurent and Dumas proposed the idea of radicals from which a larger structure could be built up. Whilst this was manifestly years ahead of its time, it is still the basis of how (organic) chemistry is taught and also forms a rational way for complex structures to be broken down into pieces, about which information is known.

In the 1930's Hammett and others showed that unrelated properties of compounds with the same 'radicals' could be correlated quantitatively. E.g. the pK_a of substituted benzoic acids and the hydrolysis rates of the corresponding esters.

These **Linear Free Energy Relationships** (LFER's) could be used to assign quantitative property values to radicals which were transportable across systems, within reason.

This allowed properties of unknown molecules to be predicted.

The Hammett World

- Small data set
- Unequivocal assignment of substituents
- Unequivocal assignment of parameters
- Compounds pure, sampled from a large bulk sample of known provenance.
- Analytical assay carried out repetitively to high degree of accuracy.

Typical Hammett Data Set

| Substituent | pK _a | σ |
|-------------------|-----------------|-------|
| H | 10.17 | 0.00 |
| 3-Me | 10.32 | -0.07 |
| 3-OMe | 9.86 | 0.12 |
| 3-Cl | 9.26 | 0.37 |
| 3-Br | 9.17 | 0.39 |
| 3-NO ₂ | 8.55 | 0.71 |

Steric effects

In the 1940's and 50's the ideas of LFER, which had been aimed at electronic effects were extended to steric effects by Taft *et al.*

This led to not only to correlations between property and structure but to a greater understanding of the processes involved, as it was occurring in a climate of intense study of mechanism in organic chemistry.

The next major advance was in the 1960's when Hansch and Leo at Pomona College realized that the idea of LFER's could be extended to biological systems by introducing a free energy term for transport through membranes i.e. logP.

The CLOGP algorithm used to calculate logP values for unknown molecules was developed by Weininger and Leo, just as described earlier. Taking the contribution values for the 'radicals' and correcting for topological proximity.

The Hansch World

- Small data set
- Unequivocal assignment of substituents
- Unequivocal assignment of parameters
- Compounds pure, sampled from a large bulk sample of known provenance.
- Biological assay carried out repetitively to ‘reasonable’ degree of accuracy.

Typical Hansch Data Set

| Substituent | Log 1/C | logP |
|-------------------|---------|------|
| H | 1.24 | 1.46 |
| 2-Me | 2.00 | 1.95 |
| 4-OMe | 1.50 | 1.34 |
| 4-Cl | 2.52 | 2.35 |
| 4-Br | 2.82 | 2.59 |
| 4-NO ₂ | 1.92 | 1.38 |

By interposing this step of property between the structure and the activity, compounds of *different structural classes*, but having *equivalent property values* were shown to have similar activities.

The well-spread set

- The Hansch methodology was almost synonymous with multiple linear regression.
- In order to have an interpretable, meaningful relationship, the x-parameter sets needed to be orthogonal and exhibit reasonable variance.
- Sets of compounds were chosen which maximized their separation in space.

The down side was you did not know *a priori* what the correlation space would be.

Other techniques such as PCA solved the statistical problem but lost the interpretability and the key test proposed by Hansch viz:

The result must make physicochemical sense

So what happened?

- N was never big enough to satisfy the statisticians.
- Required too high a precision
 - Better be precisely wrong than approximately right.
- Came to be thought of as a *post-facto* rationalization tool rather than as a guide for optimization.

The hedgehogs take over.....

At this point in his talk John showed an amusing cartoon depicting pre-historic hedgehogs turning a newly invented stone wheel on into a stone square.

The caption read “Hedgehogs tried their hardest to un-invent the wheel”.

This cartoon has been omitted to save space.

“Let’s make N=1”: The rise of molecular modelling.

- Obsession with what things *were* rather than what they *did*.
- Receptors were named by their sequence, rather than their response to small molecules, no sense that the differences in structure were relevant.
- Shape is all.
- Free energy is dead -- long live enthalpy.

The well-spread set

- There was no need for the well-spread set.
- The aim of modelling of this type was to go directly for the final compound.
- Unfortunately, quite often there was no 'Plan B'

“We have the technology....”

- “Let’s screen everything”
 - Forget chemistry completely.
- Computers will sort out the data.
 - Finance houses do data mining
 - Medics have used regression trees for ages
 - It’s all the same - isn’t it?
- The birth of chemoinformatics

Compounds 'R' Us

- Combinatorial chemistry
- Array synthesis
- External suppliers
 - 3-4 million compounds available for screening.
- Need to make choices again.

The HTS World

- Large data set
- Equivocal assignment of substituents
- Equivocal assignment of parameters
- Compounds may not be pure, not sampled from a large bulk sample of known provenance.
- Biological assay carried out rapidly to lower degree of accuracy.

Diversity and choosing compound sets

- As in the early days of Hansch analysis, a well-spread set of compounds is required.
 - Need to be parameterized in a particular multidimensional space
- The numbers now are large, you cannot rely on the chemist to choose individually appropriate parameters.
- Large scale parameterization based on structure.

Representing compounds as numbers.

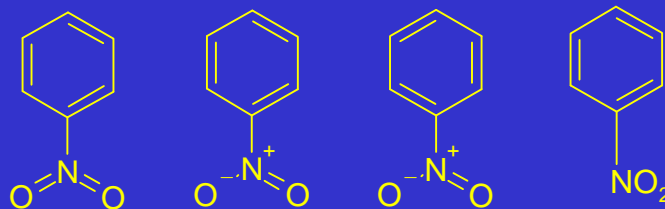
- Compound hierarchy
 - Parent
 - Parent Isomer
 - Version
 - » Sample/Preparation
- Ensure you know what your chosen parameter is about.

3-D examples

- In the NCI database, probably the largest public domain database of structures, 3D coordinates are added using Corina
 - NSC 624151 (120586-49-4) and NSC 624152 (120586-50-7) are given the same 3D structure, despite being optical isomers.
 - guanidine sulphate NSC 7296 is converted faithfully into a 3D structure with the guanidine and sulphate entities separated by 10Å.

The vagaries of valence bond representation.

Along with the development of the structural theory that the understanding of structure has required, non-textual graphical techniques have developed to represent these three dimensional objects and the concepts surrounding them in a two-dimensional graphic space. These techniques are now so entrenched that “it is difficult to imagine that we could talk, write or even think about molecular structures without recourse to them.”



A chemist will automatically recognize that the above structures represent the same chemical. Only when you have ‘taught’ your computer to recognize identity, can you approach the more difficult problem of diversity.



Tautomers

- Three schools of thought
 - Chemistry driven.
 - Represent it how it is (!) within the constraints of a VB representation.
 - Informatics driven.
 - Be consistent then retrieve the associated experimental data.
 - ‘CAS’ approach
 - try and do both of the above by bending VB representation.

These give very different answers when you move into say a 3-D world.

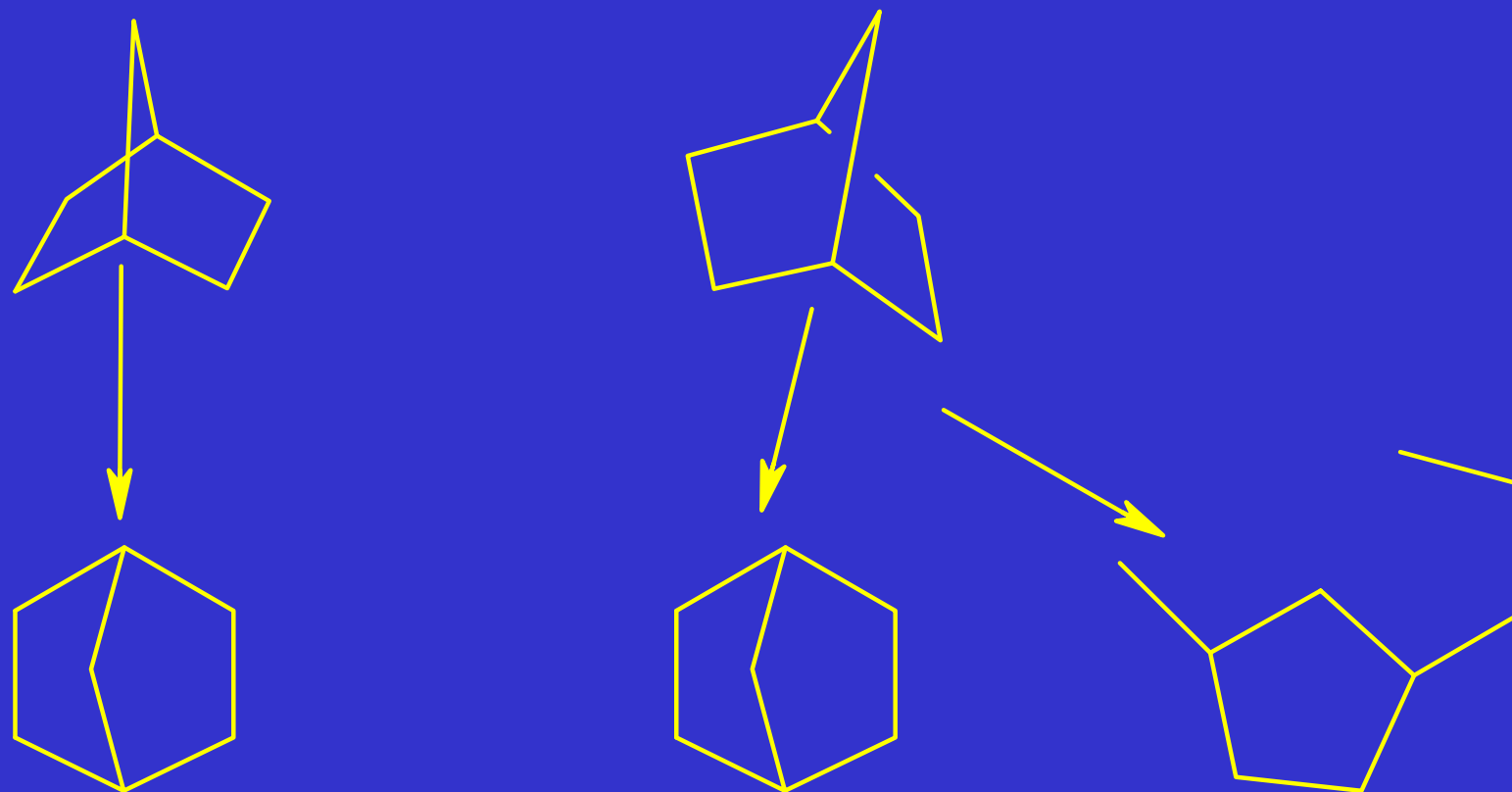
H-donors and acceptors can move around a molecule, depending on the underlying 2D-structure from which the 3-D was derived..

Graphical input

With the advent, of drawing packages it has become increasingly common to input structures into a database using a drawing package such as ISISDraw™ or ChemDraw™

Unfortunately these are optimised to present the visual clues to the viewing chemist rather than to interpret a 2-D representation of a 3-D object for a computer.

Norbornanes



02/06/00

Filters and “appropriateness”

- Need to be cognizant of the limits imposed by ‘body chemistry’.
- Given that a large number of compounds are known to affect ‘body chemistry’, and a substantially greater number to have no effect. It is likely that any new drug will be more ‘like’ the former group than the latter.
 - V.J. Gillet et al *JCICS* **1998**, 38,165-179
- This imposes limits on the space in which to search for diversity.

Teaching your computer chemistry

- In order to calculate the appropriate numbers to parameterize your compounds for diversity, subset selection, clustering etc. The following need to happen:-
 - All the compounds must be represented consistently, i.e. valence bond structures, tautomers, hierarchy.
 - You must be able to communicate chemical concepts e.g H-bond donors/acceptors accurately so that the computer can detect their presence in molecules consistently.

- Ensure that the programs you use to calculate the parameters also speak the same language/dialect.
- Ensure that your database system associates the appropriate data with the appropriate level of hierarchy. E.g. ClogP is data about the Parent. A molecular weight from a supplier may be about the Version, even though it was supplied with the Sample, whereas you may be comparing Parents as your biological test is in solution.

Instant QSAR, the real value of HTS

- One of the difficulties with traditional QSAR was that events overtook the completion of the experiment.
- “Good enough, soon enough” philosophy often meant not all compounds were tested.
- With HTS and related technologies it should be possible to get all the results at once.

Handling sets of compounds

- Traditionally there has been an urge to classify the biological response of compounds.
 - Nitro-compounds are toxic
 - Lipophilic amines are absorbed intranasally
- QSAR allowed these relationships to be quantified
 - TOPKAT, MetabolExpert etc

- What we did not handle well, were the informatics. Tools were still designed to handle compounds one at once.
- The value of QSAR is that it derives information about a *set of compounds* from the data about the individual components.
- This information can be compared with information about another *set of compounds*.
- Ironically this is what Hammett did originally.

- Tools are now available to handle information about sets of compounds.
- At the trivial level we can say how “like” one library is to another, or one plate to another.
- More to the point, techniques are becoming available to compare the information extracted from sets of compounds, i.e to compare QSARs.

The future

- After a few years in the wilderness we are getting the right balance between the biology, chemistry and information handling.
- Hopefully we have learned that technology is just a means to an end.
- Hopefully, too, there is a rosy future for the chemist in informatics.

“However, it’s really all in the
genes.....”

- It’s the hedgehogs
again...

The cartoon used earlier was repeated
again here. It has been removed to save space.