Quiz

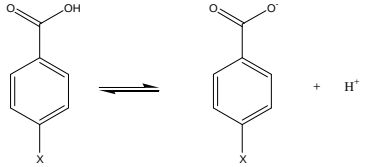
- Select a method you are using for your project and write ~1/2 page discussing the method. Address:
  - What does it do?
  - How does it work?
  - What assumptions are made?
  - Are there particular situations in which it will NOT give good results?

QSAR

QSAR

- QSAR = Quantitative Structure Activity Relationships
- Current Applications
  - Two-dimensional
  - Three-dimensional: requires molecular alignment
- Foundation: Physical Organic Chemistry
  - Relationships between structure and reactivity (equilibrium and rate constants for related structures)
  - Originally formulated by Hammett, extended by Taft and others

Hammett's Standard Reference Reaction



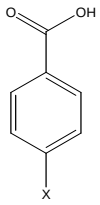
Where X=H at 25°C

Remember: $K_a = \frac{[H^+][X-C_6H_4CO_2^-]}{[X-C_6H_4CO_2H]}$

Substituent Effects on Equilibria

Hammett defined substituent constants

$\sigma_x = \log K_x - \log K_H$



What are your expectations for the values of σ for X=H, X=NH_2 and X=NO_2?

Explain your expectations based on the reference reaction.

The Hammett Equation

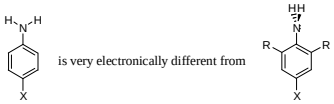
Log k_X = $\rho\sigma_X$ + log K_H or log K_X = $\rho\sigma_X$ + log K_H

ρ (slope) reflects the reaction's sensitivity to the electronic effect of substituents

ρ = 1 for the reference reaction

Important Implications

- σ values implicitly account for the influence of solvation (H-bonding, dipole-dipole)
- No consideration of geometry is included
  - Problematic if steric interactions cause a change in electronic character



  - Problematic for extensions to flexible systems
- Conformation is implicitly included

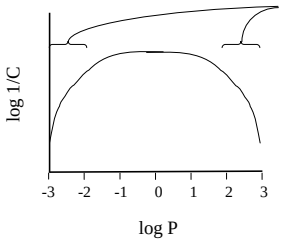
Limitations

- Ortho substituents often interact sterically
- σ values are determined in water, for H-bonding substituents may see problems for non-aqueous phenomenon
- Reactions often change mechanism when substituents with drastically different electronic characteristics (σ) are present
- σ for charged groups is dependent on the ionic strength of the media
- Direct resonance can cause problems

Hansch's Application of the Hammett Equation

- Biological activity of indoleacetic acid-like synthetic hormones
- $\text{Log}(1/C) = -k_1(\log P)^2 + k_2(\log P) + k_3\sigma + k_4$
  - C: Concentration having a standard response in a standard time
  - P: Octanol/water partition coefficient
  - Log P reflects pharmacokinetic influence on activity – does the compound get where it needs to go?
  - σ reflects pharmacodynamic influence on activity – does the electronic nature of the compound induce activity?
- Why is there a squared log P term?

Log (1/C) Versus Log (P)



Poor bioavailability

Compounds that are too polar will not partition into membranes in the first place

Compounds that are not polar enough cannot partition back out

Importance of Hansch's Work

- Demonstrated that biological activities could be quantitatively related to physical and chemical characteristics
- Developed a group-additive method for calculating log P (so that compounds could be predicted prior to their synthesis)
- Utilized a QSAR equation to assist in developing a physical interpretation or generalization about biological activity

Descriptors

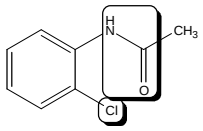
- Descriptor: A numeric representation of structure
- Descriptors used in the Hansch approach (log P, σ) are empirical (derived from experimental observation)
- Limitations
  - σ is a substituent descriptor -> won't be applicable to non-congeneric series
  - Log P is an experimentally determined value -> a computational method is needed before it can be used to make predictions

Computing Log P

- Initial attempt – π method
  - Use measured log P for largest possible substructure
  - Add contributions (π values) for substituents
- More Common – Fragment summation methods
  - Hansch's implementation: CLOGP
    - Defines two hydrophobic fragment types
      - Isolating carbons (ICs) – carbons not double or triple bonded to a heteroatom
      - Hydrogens attached to ICs (ICHs)
    - Contiguous remaining groups are polar fragments

Example Fragmentation

- 2 Polar Fragments
- 7 ICs
- 7 ICHs



Other Considerations

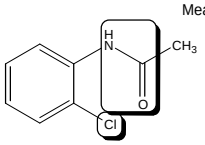
- Fragment environment -> different values stored for fragments in these environments
  - Aliphatic
  - Benzyl
  - Vinyl
  - Styryl
  - Aromatic
- Interactions among fragments
  - Handled by adding correction factors

Example LogP Calculation

Cl amide C_{aliph} C_{arom} H correction factors

0.94 + (-1.51) + 0.2 + 6(0.13) + 7(0.225) –0.12 + 0.30 – 0.84 = 1.34

Measured value = 1.28



Non-Empirical Descriptors

- Topological
  - Descriptors computed from structural formula
  - Conformation independent
- Geometric
  - Descriptors computed from molecular geometry
  - Conformation and stereochemistry dependent
- Electrostatic
  - Descriptors computed from the charges or charge distribution of the molecule
  - Some are conformation/stereochemistry dependent

Class Exercise I

- Build a small molecule containing multiple functional groups
- Perform a conformational search of your choice, with an appropriate forcefield
- Open the resulting database and use Compute->Descriptors to calculate all descriptors implemented in MOE for each of your conformations
  - Which ones do not change with conformation?
  - Which ones do change with conformation?

Weiner's Path Number, w

- An example topological descriptor
- Applied to QSPR of hydrocarbon boiling points in 1947
- Sum of bond distances between carbon atom pairs in the molecule
- Physical meaning: a reflection of size and compactness

Weiner's Path Number (cont'd)

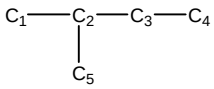
C_1-C_2: 1 C_2-C_3: 1 C_3-C_4: 1 C_4-C_5: 3

C_1-C_3: 2 C_2-C_4: 2 C_3-C_5: 2

C_1-C_4: 3 C_2-C_5: 1

C_1-C_5: 2

Sum = 18



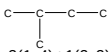
Calculation is simplified by multiplying the number of heavy atoms on each side of every bond and summing

(1x4)+(3x2)+(4x1)+(1x4) = 18

Comparison of Structures

C_5H_{12} Isomers

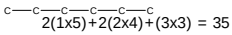
2(1x4)+2(2x3) = 20



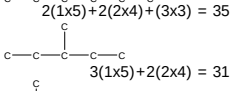
3(1x4)+1(2x3) = 18

4(1x4) = 16

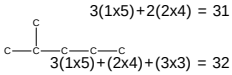
C_6H_{14} Isomers



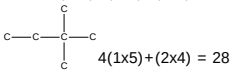
2(1x5)+2(2x4)+(3x3) = 35



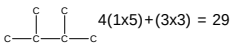
3(1x5)+2(2x4) = 31



3(1x5)+(2x4)+(3x3) = 32



4(1x5)+(2x4) = 28



4(1x5)+(3x3) = 29

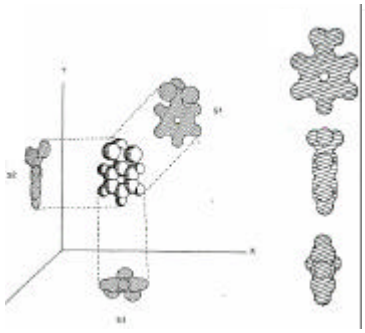
Maximum Negative Charge

- An example electronic descriptor
- A measure of the atom with the greatest partial negative charge
- Physical meaning:
  - Might indicate ability of the molecule to accept a hydrogen bond or interact with a metal ion
  - Conformation dependence varies based on partial charge assignment method
    - Forcefield partial charges are generally conformation and stereochemically independent
    - Quantum mechanical charge distributions are generally conformation and stereochemically variable

Shadow Indices

- An example geometric descriptor
- Calculated from the area of the molecule projected onto the XY, YZ and XZ planes
- Physical meaning:
  - Captures shape and size of molecule
  - Orientation dependent
  - Conformation dependent
  - Stereochemistry independent

Shadow Indices (S1, S2, S3)



Electronic/Geometric

- Common 3D QSAR methods (COMFA, COMSIA…) use electronic descriptors calculated on a grid (thus having geometric dependence)
  - First requires alignment of molecules on the grid
  - Alignment should place groups interacting with common receptor sites in the same location
- This process results in a huge number of descriptors per molecule
- Many of the descriptors are correlated

The Descriptor Explosion

- Most programs used in QSAR can calculate hundreds of standard descriptors + field-based descriptors
- Quantitative models with an overwhelming number of independent variables are over-determined
  - Multiple sets of coefficients exist that reproduce the dependent variables
  - Most of these will not be predictive (fit the data, but without physical meaning)

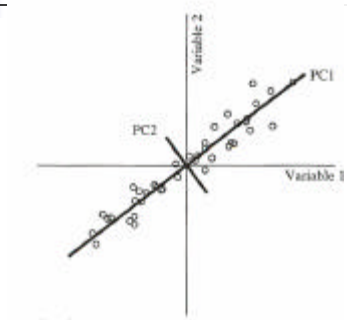
Variable Selection

- Principle Components Analysis
- Elimination of Correlated Descriptors
- Genetic Function Approximation (GFA)
  - Implemented in Cerius^2
  - Evolves models with subsets of possible descriptors to improve the fit of the data
    - Initially develops random population of QSAR models
    - Evaluates fitness (fit) of the models
    - Selects those with better features to create next generation of models from

Principal Components Analysis

- Principal components analysis is a variable reduction method (an alteration of the coordinate system) – allowing visual analysis of multi-dimensional data in fewer dimensions
- The first principal component explains the maximum amount of variation possible in the data set in one direction – the % of variation explained can be precisely calculated

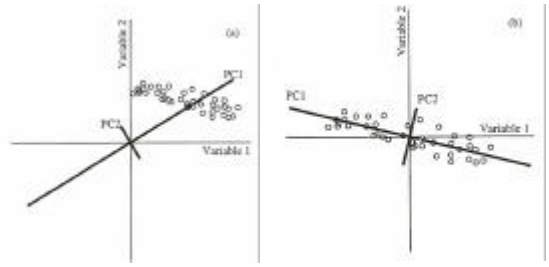
PCA – Example



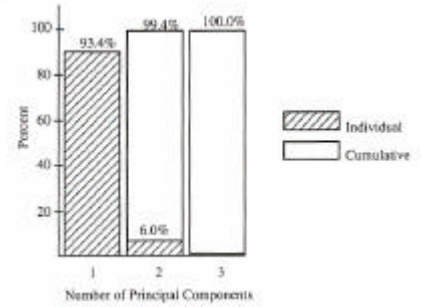
Suggested Preprocessing

- Autoscaling
  - Needed if measurements are of different types with different ranges
- Mean centering
  - Always required for PCA due to orthogonality of the components

Why Mean Centering?



How Many Components (Rank)?



Class Exercise II

- Compute the principle components for your database from the first exercise (you may need to delete fields with identical values for all structures first)
- Generate a principle components report
- How many principle components are needed to describe >75% of the variability in the descriptors? How many for >90%?
- Which descriptors contribute most significantly to the first principle component?

PCA Strengths/Weaknesses

- Strengths
  - Displays highly dimensional data with relatively few plots
  - Can filter noise from data sets
  - Can determine amount of variation contained in each descriptor (loading)
- Weaknesses
  - Inherent dimensionality (rank) must be determined
  - If the dimensionality is greater than three, visualization is still difficult

Reading

- Second Edition – Section 12.12