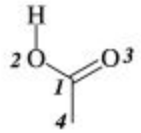
Development of Methods for the Determination of pKa Values

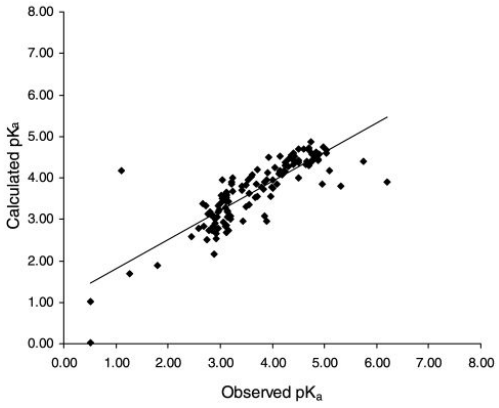
1. **Linear regression to search for relationships between pKa (experimental data) and calculated Q properties.** [1]

- Experimental pKa values were taken from the Physprop database, which reports pKa values for over 1,800 compounds
- All the structures were initially extracted as 2D SDfiles and converted into 3D models using *Corina*.
- The QM calculations were carried out with a modified version of *Mopac*
- Each structure was fully optimised (EF routine, PRECISE),prior to any parameter calculation using the *AM1 Hamiltonian within Mopac*.
- A set of properties, derived from frontier electron theory [22], were computed from the eigenvectors cj and the eigenvalues j where refers to the atomic orbital (i.e. s, px, py and pz) and j to the molecular orbital. Given a molecule with N molecular orbitals, whose levels from 1 to m are occupied, and an atom p with q atomic orbitals.
- QM descriptors (equations included in paper):
    - Electrophilic frontier electron density (FE)
    - Nucleophilic frontier electron density (FN)
    - Electrophilic superdelocalisability (SE)
    - Nucleophilic superdelocalisability (SN)
    - Radical superdelocalisability (SR)
    - Atom self-polarisability (ALP)
    - Partial atomic charges (AQ)
    - Energies of the HOMO and LUMO



aliphatic carboxylic acids

- Results(Aliphatic Carboxylic Acids):
    - (all - 185 molecules)
        - pKa= 4.29*ALP1 - 41.77*AQ2 - 30.04*AQ3+0.71*FE3 +56.06 **(eq 11)**
        - r =0.83, r^2 = 0.69, r_{cv}^2 =0.67, F= 101.28, s= 0.564
    - (excluding amino acids - 143 molecules)
        - pKa =3.24*ALP3 -2.80*SE3+ 19.43 (**eq 12)**
        - r= 0.84, r^2 = 0.70, r_{cv}^2 0.69, F =165.80, s=0.510



Plot of calculated vs. observed pK_a for the aliphatic acid dataset, using Eq. 12 from Table 7.

- (excluding amino acids - 141 molecules)
  - pKa =3.53*ALP3 -2.65*SE3+ 25.01 (**eq 13)**
  - r= 0.90, r^2 = 0.70, r_{cv}^2 =0.69, F =165.80, s=0.510

2. **Thermodynamic manipulation after calculating dissociation energies of 16 aliphatic carboxylic acids [2]**
   - 3D structures of compounds using *SYBYL*
   - Preliminary semi empirical geometry optimization at AM1 level using *MOPAC*.
   - Geometric optimization at ab-initio SCF level using 4 basis sets.
   - Free energy of solvation of gas phase molecular structures calculated using **PCM-UAHF** in *Gaussian 94 software* **(eq 7)**
   - Proton transfer energy calculated at SCF and MP2 levels in gas phase and solution phase **(eq 11 and 12)**.
     - Equations can be read in the theory section of this paper[2]
   - Results (unsure about my interpretation of the results)
     - pKa range (0.89-5.05)
     - The precision of calculated free energies of dissociation is <u>not sufficient</u> for the prediction of <u>absolute pKa values.</u>
     - The results with 16 aliphatic carboxylic acids suggest that, within chemical classes, <u>experimental trends</u> of pKa can be <u>well reproduced</u> when using PCM-UAHF for the solvation contribution to compound acidity
     - The gas-phase portion of the proton-transfer energy is apparently better described with 6-31G** and 6-31+G** than with the considerably greater basis sets 6-311G(2d,2p) and 6-311+G(2d,2p), respectively, and better at the SCF and SCFfree energy level than with MP2.

3. **Artificial intelligence algorithm to identify molecular sub structures that have high probability of being relevant to observed property/behaviour [3]**
    - Uses *MULTICASE* program to analyse the relationship with organic acid structures and their (first) pka value.
    - There are no implementation details about the MULTICASE AI algorithm, just an explanation of what it does. Couldn't locate a source code either. Although the paper say s that they are available from Professor Klopman.
    - Data
        i. 2464 molecules
        ii. Contains 242 aliphatic and alicyclic carboxylic acids
        iii. More details in database section of paper[3]
        iv. The program can only handle neutral species therefore zwitter ions are entered into their training data in neural form.
        v. Only most stable tautomers were used - usually with the weaker acidic proton.
    - Input
        i. Structural formula
        ii. Activities (pK_b in this paper)
    - All molecular structures are fragmented into all possible substructures (descriptors) of 2 to 10 linearly connected heavy atoms, which are labeled as active or inactive depending on whether the parent molecule is acidic.
    - Fragments with highest probability of being responsible for acidity are considered to perform multivariate regression analysis.
    - pKb is calculated using the coefficients obtained in the QSAR (Quantitative Structure Activity Relationship) equation.

Modulator	r_i	n_i	M_i	$r_i(n_iM_i)$
F—C—CO—OH	+8.8	2	1	17.6
Cl—C—CO—OH	+8.3	1	1	8.3
log *P*	−1.6	1	1.05	−1.7
Hardness, $(\varepsilon_{HOMO} - \varepsilon_{LUMO})/2$	+5.1	1	1.14	−5.8
Total				18.4

Constant = 56.8;
$pK_b = 56.8 + 18.4 = 75.2$ Multi-CASE units = 13.4 p*K* units;
$pK_a = 14.0 - pK_b = 0.6$;
cf. experimental $pK_a = 0.46$;
error = 0.14 p*K* units.

SCHEME 1. The multi-CASE prediction of the acidity in the case of chlorodifluoroacetic acid.

    -
    - Results

TABLE I.
Statistical Parameters of the Multi-CASE Predictions of the Acidity.

Set	Training / Test	Phi^2	OC, %	Sens. %	Spec. %	*r*	SD
1	616 / 1848	0.861	97.0	90.5	99.7	.925	0.907
2	1232 / 1232	0.902	97.9	93.8	99.6	.942	0.831
3	1848 / 616	0.897	97.8	94.3	99.3	.952	0.774

Phi^2 measures the accuracy of the predictions with respect to expectations from acidity randomness and equals 1 for a perfect fit. Observed concordance (OC) is the ratio of the sum of true positives and true negatives divided by the total number of predictions. Sensitivity (sens) represents the probability of an experimentally active compound to be predicted active. Specificity (spec) renders the probability of an experimentally inactive compound to be predicted inactive. The *r* value is the correlation coefficient between the experimental and predicted acidities and is 0 for no correlation and 1 for a perfect one. The SD value is the standard deviation.

References

[1] Prediction of the pKa of Carboxylic Acids Using the ab Initio Continuum-Solvation Model PCM-UAHF

[2] Estimation of pKa Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids

[3] Application of the multiple computer automated structure evaluation methodology to a quantitative structure–activity relationship study of acidity

[4] MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program