Estimation of Ionization Constants (pK_a) for Drug-Like Compounds

Stephen Jelfs, Peter Ertl and Paul Selzer

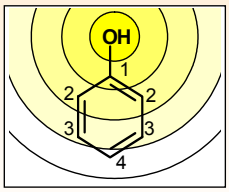
Novartis Institutes for BioMedical Research, Basel, Switzerland

Abstract

A knowledge of the ionization constants (pK_a) of compounds is important for much of the work carried out in the drug discovery process. These constants can have a profound affect on the physicochemical properties of a compound and, in rational drug discovery, are essential for the optimization of ADME characteristics. Notably, compounds in their unionized form tend to be less soluble but can more easily penetrate lipophilic barriers existing between them and a biological target of interest. Furthermore, the correct ionization state of a compound involved in ligand-receptor binding is required prior to docking studies and for the development of reliable SAR. We have therefore developed a pragmatic approach for the estimation of pK_a[1]. Importantly, in-house computational resources and experimental data were exploited to provide reliable predictions for drug-like compounds via the Novartis intranet.

Methodology

The approach utilizes a stepwise algorithm to assign pK_a values to the acidic and basic groups in a compound. At each step, the algorithm finds the next most basic (least acidic) group based on a consensus of predictive models. These models were derived using descriptors including 2D molecular tree structured fingerprints[2] and semi-empirical *quantum chemical* (QC) properties[3]. The molecular tree derived for phenol, for example, comprises of an integer-vector encoding the frequency of atom-type occurrences at through-bond distances moving away from the ionizable oxygen atom of interest:

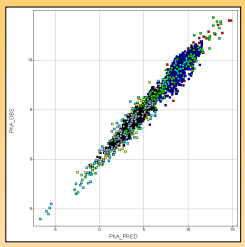


distance (bonds), *d* =	0	1	2	3	4
atom types, *t* =	O.3	C.ar	C.ar	C.ar	C.ar
fingerprint (*t*,*d*) =	**1**	**1**	**2**	**2**	**1**

The QC properties included the partial charge and *electrophilic superdelocalizability* (SE) of the atoms undergoing (de)protonation. For multiprotic compounds, the absolute SE values were replaced by the SE of the neutral structure followed by a series of relative values calculated with combinations of the more basic/acidic groups appropriately ionized. This procedure overcame limitations highlighted in the literature to allow multiprotic, drug-like compounds to be modeled effectively.

Results

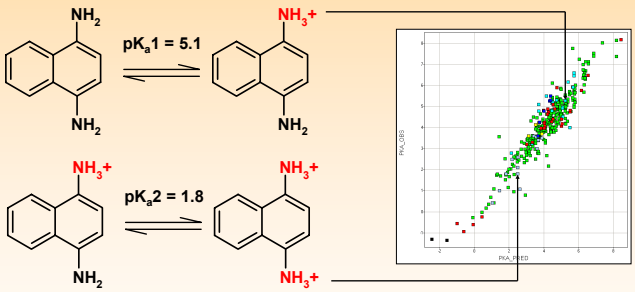
Predictive models were successfully derived using *partial least-squares* (PLS), and validated using 7-fold cross-validation, for a variety of ionizable groups:



Model	N_{cmpds}	N_{pls}	R^2	*RMSE*	R_{CV}^2
Alcohols[a]	246	3	0.91	0.50	0.81
Amines	1453	4	0.88	0.52	0.83
Anilines[b]	361	3	0.88	0.53	0.76
Carboxylic Acids	732	4	0.90	0.33	0.86
Imines[c]	134	4	0.97	0.67	0.90
Pyridines	447	4	0.95	0.58	0.88
Pyrimidines	141	4	0.94	0.46	0.85
Combined	**3514**			**0.51**	

[a] inc. phenols. [b] inc. aminopyridines.
[c] inc. amidines, guanidines & imidazoles

As mentioned, the predictive models are also applicable to multiprotic compounds:



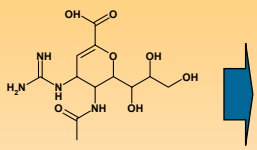
The stepwise algorithm was able to reproduce the correct ionization order of groups found within the 331 multiprotic compounds included in the training set. Independent studies have shown the estimates to be at least comparable to those provided by the commercial package ACD/pK_a and superior to PipelinePilot for the groups studied.

Summary

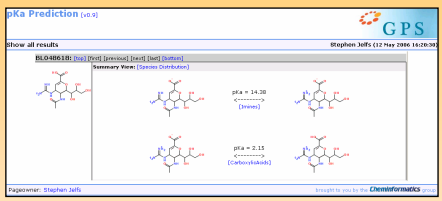
The descriptors described were found to complement one another particularly well, resulting in highly predictive and extrapolative models. We envisage the development of further predictive models and their continued refinement as new experimental data becomes available.

Web-Application

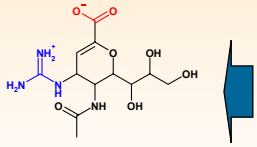
The resultant performance of the approach has enabled the development of a web-application for pK_a prediction (available on the Novartis intranet). The tool can be used to identify important acidic and basic groups within structures, such as the carboxylic acid and guanidine groups in Zanamivir:



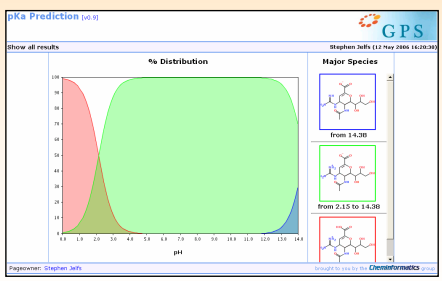
Zanamivir



A species distribution plot is also provided, indicating the major chemical species present in aqueous solution at varying pH-levels. For instance, Zanamivir is shown to be present in its zwitterion form under typical physiological conditions:



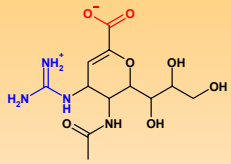
Major Species
ca. pH, 2.15 to 14.38



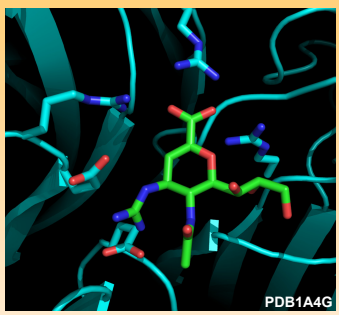
Further information about the web-application is provided via a Wiki discussion forum which summarizes the current status of the predictive models and highlights problematic structures, such as tautomers, that may be encountered. We have also recently extended the tool to include pH-dependent distribution constants (log*D*).

Molecular Recognition

An important application for pK_a constants is the determination of the major chemical species involved in ligand-receptor binding. This is essential if the interactions responsible for molecular recognition are to be correctly identified. For example, the zwitterion form of Zanamivir is involved in binding to the influenza B virus Salidase:



Species involved in ligand-receptor binding ca. pH, 7



Where, the charged carboxylic acid and guanidine groups are essential for binding-site recognition, forming important interactions with the complementary arginine (ca. pK_a, 13.0) and glutamic (ca. pK_a, 4.3) amino acids, respectively.

Prior to docking studies, it is common to enumerate all potential chemical species that could possibly to be involved in binding (ca. pH, 7). We are therefore looking to improve the performance of our approach for the pre-treatment of compounds prior to high-throughput docking. The adverse affects of using the incorrect protonation state or tautomer form of compounds in docking studies may also form the bases of future work.

Acknowledgments

The authors would like to thank Joerg Muehlbacher and Wolfgang Zipfel for their technical support, Bernard Faller and Frederique Loeuillet for providing experimental data, and Ansgar Schuffenhauer, John Priestle, and Bernhard Rohde for their insightful discussions. We are also grateful to all the medicinal chemists involved in testing the beta-version of the web-application.

[1] Jelfs, S. P.; Ertl, P.; Selzer, P. Estimation of Ionization Constants (pK_a) Using Semiempirical and Information-Based Descriptors, *J. Chem. Inf. Model.* **2007**, *47*, 450-459.
[2] Tehan, B. G. et al. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457-472.
[3] Xing, L.; Glen, R. C. Novel Methods for the Prediction of logP, pK_a, and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796-805.