Chemosphere 182 (2017) 399–405

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Solubility of organic compounds in octanol: Improved predictions based on the geometrical fragment approach

Didier Mathieu

CEA, DAM, Le Ripault, 37260, Monts, France

HIGHLIGHTS

- Two very simple models to predict octanol solubility are reported.
- In the lack of experimental data, the first one performs as well as current state-of-the-art models.
- Using melting point data, the second one yields some improvement through simple corrections.
- Good performance is achieved through a physically-motivated fragmentation of the molecules.
- Some limitations of popular applicability domain definitions are demonstrated.

ARTICLE INFO

Article history:
Received 7 April 2017
Received in revised form
5 May 2017
Accepted 8 May 2017
Available online 10 May 2017

Handling Editor: I. Cousins

Keywords:
Solubility
Octanol
Molecular modeling
Quantitative structure-property relationships

ABSTRACT

Two new models are introduced to predict the solubility of chemicals in octanol (S_{oct}), taking advantage of the extensive character of $\log(S_{oct})$ through a decomposition of molecules into so-called geometrical fragments (GF). They are extensively validated and their compliance with regulatory requirements is demonstrated. The first model requires just a molecular formula as input. Despite an extreme simplicity, it performs as well as an advanced random forest model involving 86 descriptors, with a root mean square error (RMSE) of 0.64 log units for an external test set of 100 molecules. For the second one, which requires the melting point T_m as input, introducing GF descriptors reduces the RMSE from about 0.7 to <0.5 log units, a performance that could previously be obtained only through the use of Abraham descriptors. A script is provided for easy application of the models, taking into account the limits of their applicability domains.

1. Introduction

The solubility of liquid and solid compounds in dry 1-octanol (S_{oct}) is of considerable interest due to its importance in pharmacology and environmental chemistry. It is particularly useful to describe the transport and fate of organic pollutants in the environment (Li et al., 2003) or to characterize the storage capacity of chemicals in lipids (Anliker and Moser, 1987). A typical error of 0.4 log unit for experimentally derived $\log(S_{oct})$ values has been suggested by Raevsky et al. (2007).

For several years, predictive models have been developed to estimate this property (Li et al., 1995; Sepassi and Yalkowsky, 2006; Raevsky et al., 2007; Raevsky and Schaper, 2008). In the lack of experimental data for the compound under consideration, a recent random forest model based on 86 input descriptors may be used to predict $\log(S_{oct})$ with a fair accuracy, characterized by a root mean square error (RMSE) of 0.66 log units for an external test set (Buonaiuto and Lang, 2015). A similar accuracy with RMSE ≃0.7 log units can be obtained much more easily if the experimental melting point T_m of the compound is available:

$$\log(S_{oct}) = p - m(T_m - 298\ \mathrm{K}) \quad (1)$$

where $p = 0.5$ and $m = 0.01$, T_m is the melting point in Kelvin, or taken to be 298 K for compounds liquid at room temperature (Admire and Yalkowsky, 2013). Finally, significantly more accurate predictions can be obtained in the favorable case where Abraham descriptors have been measured for the compound studied, in addition to T_m. In this case, a RMSE as small as 0.47 log units was

E-mail address: didier.mathieu@cea.fr.

http://dx.doi.org/10.1016/j.chemosphere.2017.05.045
0045-6535/

recently reported (Abraham and Acree, 2014).

The present work introduces two new models with some advantages over previous ones. A first model not requiring any experimental data predicts $\log(S_{oct})$ with the same accuracy as obtained from the random forest of Buonaiuto and Lang (2015). However, it is much simpler and easier to apply, involving only back-of-the-envelope calculations. A second model taking advantage of experimental T_m values yields some improvement over Eq. (1). While present approaches introduce a dose of empiricism with respect to Eq. (1), some physical insight is still used to restrict the search space of mathematical relationships between $\log(S_{oct})$ and molecular structure, as detailed below.

2. Modeling approach

2.1. Purely additive scheme

The solubility of any compound in octanol is obviously determined by solute-solvent interactions, which may be split into additive contributions arising from the various fragments that make up the molecule under consideration, hence the potential interest of additivity methods to predict this property. More specifically, the quantity that lends itself to a representation in term of additive contributions is $\log(S_{oct})$. Indeed, for any fixed temperature T, it is proportional to the free energy change accompanying the dissolution of a compound in octanol. As such, it is an extensive property that should be reasonably approximated in term of additive contributions arising from the fragments that make up the solute molecule:

$$\log(S_{oct}) = s_0 + \sum_j n_j s_j \tag{2}$$

where the sum runs over all kinds of fragments, s_0 is an empirical constant, s_j the contribution of any fragment of type j and n_j the number of such fragments in the molecule. This approach was already used over 20 years ago in the OCTASOL approach of Li et al. (1995). However, somewhat surprisingly in view of the extremely encouraging results obtained, it was not developed any further. Possible issues include:

- the use of group contributions arising from a somewhat ad hoc decomposition of the molecules into fragments, which might introduce a model selection bias;
- the relatively large number (36) of adjustable parameters.

Therefore, this attractive approach is presently revisited, using a smaller number of adjustable parameters derived from a predefined fragmentation scheme (Section 2.3) in an attempt to overcome such difficulties.

2.2. Model based on the melting point

A major difficulty with Eq. (2) stems from the fact that $\log(S_{oct})$ includes a crystal term involving the product of T_m and the melting entropy (Alantary and Yalkowsky, 2016). Like melting points, this quantity is notoriously difficult to predict quantitatively using an additivity scheme (Johnson and Yalkowsky, 2005). As a presumably better alternative to previous equations, the extensive character of $\log(S_{oct})$ and the significant correlation observed with T_m suggests that it might be fruitful to combine them as follows:

$$\log(S_{oct}) = s_0 + \sum_j n_j s_j - m(T_m - 298\ \mathrm{K}) \tag{3}$$

where the melting point T_m is again taken to be 298 K for liquids (as done in Eq. (1)). The corresponding term is indeed irrelevant for a liquid as it does not have to go through a melting transition before mixing with octanol. This model is a straightforward generalization of Eq. (1) where the intercept p is now allowed to depend linearly on constitutive descriptors.

2.3. Geometrical fragments

A critical aspect of any additivity method is the algorithm employed to decompose molecules into fragments. Group contribution (GC) methods are especially popular (Gmehling, 2009). As illustrated by the OCTASOL method for the prediction of $\log(S_{oct})$, they are usually reliable, but their scope is unfortunately limited due to the need for extensive experimental data to fit the many group parameters involved. In recent years, a specially simple fragmentation scheme first introduced to estimate crystal densities (Beaucamp et al., 2007) proved very successful to predict other properties fully or mainly determined by intermolecular interactions, including sublimation enthalpy (Mathieu, 2012), flash point (Mathieu, 2010), flammability limit temperatures (Mathieu, 2013) and liquid density (Mathieu and Bouteloup, 2016). Like long-standing approaches to estimate properties in term of additive contributions (van Krevelen, 1990), this one defines a separate fragment for every non-hydrogen atom in the molecule and its hydrogen neighbors. However, in contrast to well-established methods, it deliberately ignores bond orders, which are assumed to play only a secondary role on inter- (as opposed to intra-) molecular interactions. This is motivated by the view that interactions between a given atom of the molecule under study and surrounding species primarily depends on the availability of this atom, i.e. on the number and size of its neighbors (Beaucamp et al., 2007; Mathieu, 2012). To emphasize the role of these obvious geometric considerations, the present approach is referred to as the geometrical fragment (GF) method (Mathieu and Alaime, 2014).

A custom notation for the GF fragments consistent with the underlying assumptions was previously introduced, where a fragment is referred to as $Xn_c\text{-}n_H$ where X is the symbol of the central atom, n_c its coordination number and n_H the number of hydrogen neighbors (Mathieu, 2012). For instance, C4-3, C4-2 and N3-1 denote respectively a methyl group ($-CH_3$), a methylene group ($>CH_2$) and a secondary amine (>NH). The interest of this bond order agnostic approach may be illustrated in the case of the C2-0 fragment, which encompasses two distinct bonding environments for sp carbon atoms, namely those encountered in dimethylacetylene ($-C\equiv$) and propadiene ($=C=$). Similarly, C3-1/N2-0 are used for sp2 carbon/nitrogen atoms either aromatic (as in benzene or pyridine) or not (as in ethene or methanimine). Nevertheless, the present paper uses a familiar and probably more readable notation where bonds as explicitly shown. The reader should just keep in mind the fact that atoms with distinct bonding environments but sharing common values of n_c and n_H define the same type of fragment.

In view of the simplicity of the present procedure, Eq. (2) exhibits obvious limitations. In particular, any set of compounds sharing a common set of functional groups exhibit a single value of $\log(S_{oct})$. For instance, a common $\log(S_{oct})$ is obtained for the three isomers of any disubstituted benzene, including xylene, nitrophenol … For the latter, this might appear as a specially daring approximation as one of the isomers (2-nitrophenol) exhibits internal hydrogen bonding. In an attempt to account for the role of such interactions, various corrections associated with intramolecular hydrogen bonds have been considered, e.g. OH…O, NH…O, with possible distinctions for 5- and 6-membered H-bonded cyclic structures … Since no significant improvement was obtained, such corrections are not considered further.