Science of the Total Environment 463-464 (2013) 781–789

Contents lists available at SciVerse ScienceDirect

Science of the Total Environment

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

Validation of quantitative structure–activity relationship models to predict water-solubility of organic compounds

Claudia Ileana Cappelli [a], Serena Manganelli [a], Anna Lombardo [a], Andrea Gissi [a,b], Emilio Benfenati [a,*]

[a] Laboratory of Chemistry and Environmental Toxicology, Istituto di Ricerche Farmacologiche Mario Negri, IRCCS, via Giuseppe La Masa 19, 20156 Milan, Italy
[b] Dipartimento di farmacia, Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", via Orabona 4, I-70125 Bari, Italy

HIGHLIGHTS

- The powerful predictive ability of the models could be useful to identify questionable experimental data.
- In regression T.E.S.T. and ADMET gave the best performance, which can be improved if mol/L units are preferred to mg/L.
- All models can predict soluble compounds better than poorly soluble ones.

ARTICLE INFO

Article history:
Received 5 April 2013
Received in revised form 19 June 2013
Accepted 19 June 2013
Available online xxxx

Editor: Eddy Y. Zeng

Keywords:
Water-solubility
Quantitative structure–activity relationships
Comparison study
REACH

ABSTRACT

Water-solubility is an important physicochemical property in pharmaceutical and environmental studies. We assessed the performance of five predictive computer models: ACD/PhysChem History, ADMET Predictor, T.E.S.T., EPI Suite-WSKOWWIN and EPI Suite-WATERNT; two of them are commercial, the others are free. We used more than 4000 compounds with experimental values to evaluate the models, considering the chemicals inside and outside the applicability domain of the models, those used to build up the model (training set) and those not present in it (prediction set). We also evaluated their ability to predict continuous solubility values, and solubility classes. Overall, considering the whole data set, some models gave a good statistical performance, with R^2 up to 0.88.

1. Introduction

The water-solubility is important for a number of reasons. It is a prerequisite for setting up test conditions for studies of environmental fate (e.g. biodegradation, bioaccumulation) and effects (on humans and other living organisms), and it is a basis for other environmental parameters, such as octanol–water partition coefficient, K_{ow}, organic carbon–water partition coefficient, K_{oc}, and Henry's law constant. It is also a regulatory trigger for waiving certain physicochemical and ecotoxicological endpoints.

For an organic solute to dissolve in water, first, the solute molecules must be separate from one another. Second, the solvent molecules must sufficiently separate to create a cavity large enough to accommodate the solute. Once the solute occupies the cavity, there will be new attractive forces between solute and solvent. Finally, the water molecules in the solvation shell will form extra H-bonds to neighboring water molecules. Thus, water-solubility depends not only on the affinity of a solute for water, but also on its affinity for its own structure. Strongly bound molecules require considerable energy to separate them. Such compounds have high melting points (for solids) and generally, solids with a high melting temperature have poor solubility in any solvent.

Removal of a molecule from its crystal lattice means an increase in entropy, and this can be difficult to model accurately. For this reason, and the fact that the experimental error on solubility measurements can be quite high, especially for compounds with very low solubility, the prediction of water-solubility is not as accurate as for other properties, such as octanol–water partitioning (ECHA, 2012). The assessment of water-solubility is complicated by a number of considerations, including ionization and formation of salts. These effects may significantly alter the water-solubility (Cronin and Livingstone, 2004). Furthermore, solubility can vary considerably with temperature, so solubility data must be reported at a given temperature. Water solubilities can be reported in various ways: in pure water, at a specified pH, at a specified ionic strength, as the undissociated species (intrinsic solubility), and so on.

* Corresponding author at: Via La Masa 19, 20156 Milano, Italy. Tel.: +39 02 39014420.
E-mail address: emilio.benfenati@marionegri.it (E. Benfenati).

0048-9697/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scitotenv.2013.06.081

Under the European legislation on chemicals, REACH (EC Regulation No. 1907/2006), there are provisions for the use of data generated by quantitative structure–activity relationship (QSAR) methods. QSAR models seek mathematical correlations between chemical structure and biological activity. Related terms include quantitative structure–property relationships (QSPR) when a physico-chemical property—in the present study water-solubility—or reactivity is modeled. Recently, several studies to predict water-solubility were performed. Das and Roy (2013) developed predictive models on aqueous solubility of a large set of drug, drug like compounds and agrochemicals with two-dimensional descriptors named extended topochemical atom (ETA) indices, other topological, structural, spatial and electronic non-ETA descriptors and lipophilicity parameter ClogP. They employed genetic function approximation (GFA), genetic partial least squares (G/PLS) and stepwise multiple linear regression (MLR) to generate the models.

Toropov et al. (2013) calculated descriptors by means of the Monte Carlo method using the CORAL software (http://www.insilico.eu/coral) to build up QSPRs for water-solubility.

In another paper (Zeng et al., 2012), density functional theory (DFT) method was introduced to explore QSPR models relating to water-solubility of halogenated methyl-phenyl ethers. Using stepwise multiple regression technique, Zeng and co-workers obtained two models which contain three variables: energy of the lowest unoccupied molecular orbital (E_{LUMO}), most positive atomic partial charge in molecule (q^+) and quadrupole moment (Q_{yy} or Q_{zz}).

The density functional theory (DFT) in the Gaussian 03 package and the semi-empirical (PM6) method in the molecular package (MOPAC), based on a strategy of the QSPRs with the Genetic Algorithm-Artificial Neural Network (GA-ANN) predictions, were used for the estimation of water-solubility of 209 congeners of chloro-trans-azobenzene (Wilczyńska-Piliszek et al., 2012).

A study presented by Bhhatarai and Gramatica (2011) was focused on non-ionic perfluorinated chemicals. Their model uses two-dimensional descriptors: T(F..F), that represents the sum of distances between pair of fluorine atoms (it increases with the number and the distance between two fluorine atoms in a molecule), and SIC1 (structural information content) that gives information mainly on the structural symmetry in the molecule.

QSAR models can be useful to minimize time, cost, and resources, as a substitute for experimental data and as a supplement to experimental data in weight-of-evidence approaches. An investigation by Federchimica within the EC funded project ANTARES (http://www.antares-life.eu/) indicated that in Italy the average experimental cost per chemical to observe the obligations of REACH regulation about water-solubility is high, approximately 2350 euro.

Within the ANTARES project the performance of five existing models to predict the water-solubility of organic compounds was evaluated. Three of the models are freely available: EPI Suite-WSKOWWIN and WATERNT modules (US EPA), and T.E.S.T. (US EPA). The other two are commercially available: ADMET Predictor (Simulation Plus) and ACD/PhysChem History (ACD/Labs).

QSAR models employ a training set, which contains the chemical information at the basis of the models. The models can be tested on a different set of compounds, unknown to the models (the prediction set). This enables one to assess whether the model is truly predictive, or has just been overfitted to the training set. We therefore determined the statistics on the prediction set as well as the overall results.

The predictive ability of the models depends heavily on the compounds used in the training set. To this extent, the ability to assess the confidence in the predicted value is crucial for the correct interpretation and application of QSAR models. The concept of the applicability domain (AD) (Sushko et al., 2010) can be used, defined as the response and chemical structure space in which a model makes predictions with a given reliability (Netzeva et al., 2005). Thus, we also evaluated the models considering the AD. Finally, we took account of the information on the AD and the presence in the training set.

2. Materials and methods

2.1. Chemicals used to check models' performance

Five software packages were selected to predict the water-solubility of a large set of 4621 heterogeneous organic compounds. The experimental values originated from the training and validation sets of T.E.S.T. (Toxicity Estimation Software Tool). We considered other experimental values gathered from Toolbox v.2.2, WATERNT v.1.01 and WSKOWWIN v.1.42 data sets. The Toolbox water-solubility values originated from the Phys-chem EPISUITE database and OECD HPVC inventory.

In order to have data of high quality, we only used compounds whose experimental values were in acceptable agreement between those from T.E.S.T. (training and validation sets), Toolbox, WATERNT and WSKOWWIN. When the difference between the experimental values exceeded 1 log unit, the chemical was removed. We deleted duplicates, inorganic chemicals and mixtures. We also excluded chemicals where there was disagreement about their IUPAC name, SMILES notation and CAS number. Some compounds were eliminated because of ambiguous management of the structure format. For cis and trans isomers we calculated the geometric mean of the experimental values, because 2D software cannot distinguish between two isomers. Also in this case, when the difference between the experimental values exceeded 1 log unit, the chemical was removed.

2.2. The five in silico models

2.2.1. ACD/PhysChem history

2.2.1.1. Solubility data source. ACD/Solubility DB is based on an internal database of more than 6000 compounds with their experimental solubility values and literature references. Training data are not directly accessible, but were provided by the company upon request. An in-house program and ChemFinder were used to determine which chemicals were present in the training set for ACD.

2.2.1.2. Description of the model. ACD/Labs algorithm (see also ACD/Labs web site, http://www.acdlabs.com/products/percepta/predictors.php) uses the following parameters in correlation equations: boiling point, logP/logD, MW, MolVol, HydBonding (number of H-acceptors and H-donors), and refractive index.

The general ACD/Solubility DB algorithm does not calculate solubility for some chemical structures, which are mentioned in the guidance.

2.2.1.3. Software availability. We used the version 12.0 commercialized by Advanced Chemistry Development, Inc. (ACD/Labs).

2.2.1.4. Interpretation of the output. ACD can calculate pH-dependent aqueous solubility, intrinsic solubility, and solubility of the chemical dissolved in pure (unbuffered) water at 25 °C and zero ionic strength. Since experimental water-solubility values of our data set were not calculated at specific pH values and to compare results from different models, we selected solubility in pure water at 25 °C and zero ionic strength.

2.2.1.5. Applicability domain. No AD is specified for this version of the model.

2.2.2. ADMET Predictor

2.2.2.1. Solubility data source. ADMET Predictor (http://www.simulations-plus.com/Definitions.aspx?lID=58&pID=13) is based on a data set of 3596 organic compounds, created by appending the Meylan set—containing available melting point information and native water solubilities selected