International Journal of Pharmaceutics 360 (2008) 122–147

Contents lists available at ScienceDirect

International Journal of Pharmaceutics

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

Comparison of aqueous solubility estimation from AQUAFAC and the GSE

Parijat Jain[a,*], Kia Sepassi[a,b], Samuel H. Yalkowsky[a]

[a] University of Arizona, College of Pharmacy, 1703 E. Mabel Street, Tucson, AZ 85721, United States
[b] Johnson & Johnson Pharmaceutical Research and Development, L.L.C. 3210 Merryfield Row, San Diego, CA 92121, United States

ARTICLE INFO

Article history:
Received 9 January 2008
Received in revised form 14 April 2008
Accepted 15 April 2008
Available online 23 April 2008

Keywords:
Solubility
GSE
AQUAFAC
Activity coefficient
Melting point
Entropy
Cross-validation

ABSTRACT

The GSE (General Solubility Equation) and AQUAFAC (Aqueous Functional Group Activity Coefficients) are two empirical models for aqueous solubility prediction. This study compares the aqueous solubility estimation of a set of 1642 pharmaceutically and environmentally related compounds, using the two methods. The average absolute errors in the solubility prediction are 0.543 log units for AQUAFAC and 0.576 log units for the GSE. About 88.0% of the AQUAFAC solubilities and 83.0% of the GSE molar aqueous solubilities are predicted within one log unit of the observed values. The marginally greater accuracy of AQUAFAC is due to the fact that it utilizes fitted-parameters for many structural fragments and is based on experimental solubility data. The GSE on the other hand is a simpler, non-regression based equation which uses two parameters for solubility prediction.

Published by Elsevier B.V.

1. Introduction

Aqueous solubility is of fundamental importance in a large number of scientific disciplines and practical applications. It helps us understand the drug behavior in an aqueous environment. Therefore, it is essential to have an accurate means for predicting aqueous solubility. The GSE (General Solubility Equation) and the AQUAFAC (Aqueous Functional Group Activity Coefficients) are two such empirical models for aqueous solubility prediction.

The GSE (Jain and Yalkowsky, 2001; Ran et al., 2002) is based on the fact that the aqueous solubility of a solute depends upon its crystallinity and its polarity, which are determined by the melting point (MP) in Celsius, and the octanol–water partition coefficient (ClogP or $\log K_{ow}$), respectively, using the following expression:

$$\log S_w = 0.5 - 0.01(\mathrm{MP} - 25) - \log K_{ow} \quad (1)$$

where, S_w is the molar aqueous solubility. If the solute has a melting point less than 25 °C, i.e., if it is a liquid, the term (MP − 25) is set to zero. The following assumptions are used in the derivation of the GSE for non-electrolyte compounds:

1. The melting point of the solute does not change in the presence of water.
2. The ideal solubility (or the crystal–liquid solubility ratio) of a solid solute is described by the van't Hoff equation, with the entropy of melting being described by the Walden's rule (entropy of melting, $\Delta S_m = 56.7$ J/mol K). This is applicable to most uncharged organic molecules.
3. The heat capacity of melting (Cp_m) is negligible.
4. The effect of mutual saturation of water and octanol is negligible.
5. Organic non-electrolyte liquids are completely miscible with octanol so that the mole fraction is 0.5 (i.e., a molar solubility of 3.15). The logarithm of 3.15 is approximately 0.5, which is the intercept in the GSE.
6. The solubility of the solute is low enough so that the molarity of water is close to 55.5 M.

It should be noted that the GSE is strictly applicable to the non-electrolytes and that no fitted-parameters are used in its derivation.

On the other hand, AQUAFAC model, originally developed by Myrdal et al., is based on group contribution values (q-values) which are based on experimental aqueous solubilities (Myrdal et al., 1992, 1993, 1995; Pinsuwan et al., 1997). The aqueous activity coefficient of a compound is determined using a summation of simple additive constitutive group values, i.e.,

$$\log \gamma_w = \sum n_i q_i \quad (2)$$

where n_i is the number of times group i appears in the compound and q_i is the contribution of group i to the total activity coefficient.

* Corresponding author. Tel.: +1 520 626 4309; fax: +1 520 626 4063.
E-mail address: jainp@pharmacy.arizona.edu (P. Jain).

0378-5173/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.ijpharm.2008.04.028

Table 1
Results of aqueous solubility estimations

	GSE (predicted)	AQUAFAC (fitted)
N	1642	1642
Intercept	−0.259	−0.146
AAE (log unit)	0.576	0.465
RMSE (log unit)	0.773	0.653
Slope	0.867	0.949
R^2	0.882	0.912
Error ≤1 log unit	83.19%	88.06%

N, total number of compounds; AAE, average absolute error; RMSE, root mean-square error; Intercept, slope and R^2 for the plot between experimental and predicted, log molar aqueous solubilities.

The molar aqueous solubility (S_w) for poorly soluble solutes at high dilution can be calculated using

$$S_w = \frac{X_i^c}{\gamma_w} \cdot 55.5 \quad (3)$$

where, X_i^c is the ideal crystalline mole fraction solubility, γ_w is the aqueous activity coefficient and 55.5 mol/L is the molarity of water. Eq. (3) can be combined with the Van't Hoff equation to give

$$\log S_w = 1.74 - \log \gamma_w - \frac{\Delta S_m (T_m - T)}{2.303 \cdot R \cdot T} \quad (4)$$

where, R is the gas constant, ΔS_m is the entropy of melting, T_m is the melting point and T is the ambient temperature, both in Kelvin. Thus, the aqueous solubility can be determined using Eq. (4) where, γ_w is obtained from the AQUAFAC model.

The entropy of melting is calculated using the following equation developed by Dannenfelser and Yalkowsky (1999) and revised by Jain et al. (2004):

$$\Delta S_m = 50 - R \ln \phi + R \ln \sigma \quad (5)$$

where, σ is the molecular symmetry number (the number of positions into which a molecule can be rotated that are identical to a reference position) which accounts for the likelihood of the molecule being in the proper orientation. The molecular flexibility (ϕ) accounts for the likelihood of the molecule being in the proper conformation, for incorporation into the crystal lattice. It is calculated using the following equation:

$$\phi = 2.435^{[SP3+0.5\cdot SP2+0.5\cdot RR-1]} \quad (6)$$

where SP3 is the number of nonring, nonterminal sp^3 atoms like CH_2, CH, C, NH, N, O, S, etc., SP2 is the number of nonring, nonterminal sp^2 atoms like =CH, =C, =N, etc. and RR is the number of rigid single or fused conjugated aromatic ring systems.

In this study we compare aqueous solubilities determined from the GSE and the AQUAFAC model, for 1642 organic compounds.

2. Methods

2.1. Data collection

The experimental molar aqueous solubilities of 1642 pharmaceutically and environmentally related organic compounds were collected from WATERNT™ v 1.0 EPA and AQUASOL databases. The experimental entropies of melting were obtained from Chickos et al., 1999. Eq. (5) was used to calculate the entropies of the compounds for which the experimental values are not available. The calculated partition coefficients for all the compounds were obtained from ClogP software. The melting point data was collected from several electronic and print literature (Merck Index; Chemfinder website; AQUASOL database; EPI Suite, 2000; Howard and Meylan, 1997; Chickos and Nichols, 2002; Jain and Yalkowsky, 2006).

2.2. Solubility prediction

Compounds with observed solubilities of greater than 1 M are not included in the study owing to the fact that the solvent cannot be regarded as pure water. Also, long chain compounds with a flexibility number of 15 or greater are not included due to the possibility of self-association.

2.2.1. GSE

The partition coefficients and the melting points were used to calculate the predicted aqueous solubility using the GSE, Eq. (1).

2.2.2. AQUAFAC

All compounds were fragmented using the group breakdown scheme of Jain and Yalkowsky, 2006, into 147 groups. The predicted aqueous activity coefficients (γ_w) and the entropies of melting (ΔS_m) values were calculated using Eqs. (2) and (5), respectively. AQUAFAC-predicted aqueous molar solubility data (S_w) were calculated from the melting point (MP) and entropy of melting parameters (σ and Φ) and group activity coefficients (γ_w) using Eq. (4). The latter are based on the experimental solubility data.

A ten-fold cross-validation experiment was performed on the complete data set. For each validation, approximately 1/10th of the data were randomly selected using the RAND function, from Microsoft Excel 2000, and used as the test set. The remaining data were used as the training set to generate the group contribution values. Each compound was included in only one test set.

2.3. Statistical analysis

The group contribution values for calculating aqueous activity coefficients were generated by multiple linear regressions using SPSS (version 10.0). All other data analysis was performed using Microsoft Excel (2002). The average absolute error (AAE) for each calculation was determined by

$$AAE = \frac{\sum |\log S_{pred} - \log S_{exp}|}{N} \quad (7)$$

The root mean-square error (RMSE) was determined by

$$RMSE = \sqrt{\frac{\sum (\log S_{pred} - \log S_{exp})^2}{N}} \quad (8)$$

where, $\log S_{pred}$ and $\log S_{exp}$ are the logarithms of molar solubilities, predicted and experimental, respectively, and N is the total number of organic compounds.

3. Results and discussion

A complete alphabetical list of 1642 organic compounds along with their melting points, partition coefficients, experimental as well as AQUAFAC- and GSE-predicted aqueous molar solubility values, is provided in Appendix A. The compounds range from 10^{-13} to 10^{0}, in experimental molar solubility.

Fig. 1 shows the relationship between the experimental and the GSE-predicted aqueous solubilities. As shown in Table 1, the regression line has a slope of 0.867 and an R^2 of 0.882. The AAE in the prediction of aqueous solubilities using the GSE for all the 1642 compounds was observed to be 0.576 log units.

On the other hand, the plot of the experimental versus the AQUAFAC-fitted aqueous solubilities, for 1642 organic compounds