Contents lists available at ScienceDirect

International Journal of Pharmaceutics

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

Pharmaceutical Nanotechnology

Prediction of aqueous solubility from SCRATCH

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

[a] Pharmaceutical and Analytical Development Department, Novartis Pharmaceuticals Corp., East Hanover, NJ 07936, USA
[b] University of Arizona, College of Pharmacy, Tucson, AZ 85721, USA

ARTICLE INFO

Article history:
Received 5 June 2009
Received in revised form
29 September 2009
Accepted 2 October 2009
Available online 9 October 2009

Keywords:
Solubility
Prediction
SCRATCH
GSE
Activity coefficient
Melting point
Model
Cross-validation

ABSTRACT

This study proposes the SCRATCH model for the aqueous solubility estimation of a compound directly from its structure. The algorithm utilizes predicted melting points and predicted aqueous activity coefficients. It uses two additive, constitutive molecular descriptors (enthalpy of melting and aqueous activity coefficient) and two non-additive molecular descriptors (symmetry and flexibility). The latter are used to determine the entropy of melting. The melting point prediction is trained on over 2200 compounds whereas the aqueous activity coefficient is trained on about 1640 compounds, making the model very rigorous and robust. The model is validated using a 10-fold cross-validation on a dataset of 883 compounds for the aqueous solubility prediction.

A comparison with the general solubility equation (GSE) suggests that the SCRATCH predicted aqueous solubilities have a slightly greater average absolute error. This could result from the fact that SCRATCH uses two predicted parameters whereas the GSE utilizes one measured property, the melting point. Although the GSE is simpler to use, the drawback of requiring an experimental melting point is overcome in SCRATCH which can predict the aqueous solubility of a compound based solely on its structure and no experimental values.

1. Background

1.1. Aqueous solubility

Aqueous solubility is one of the most important physico-chemical factors that affect the dissolution, the absorption and in turn the bioavailability of a drug. Poor aqueous solubility is a common problem that frequently poses a challenge to efficient drug design and formulation development. The present work aims to provide a means to predict the aqueous solubility of a compound solely from its chemical structure.

The GSE (general solubility equation) and the AQUAFAC (AQUeous Functional group Activity Coefficients) are two of the most successful empirical models for aqueous solubility prediction (Jain and Yalkowsky, 2001; Ran et al., 2002; Myrdal et al., 1992, 1995; Pinsuwan et al., 1997). Both these methods require experimental melting point data, which could be a limitation at the early drug discovery phase. The proposed model predicts the aqueous solubility without the use of any experimental data. The method uses the predicted aqueous activity coefficients from the AQUAFAC model and the predicted melting points from the estimated enthalpies and entropies of melting. Thus, the aqueous solubilities obtained from this model are truly predicted.

For a crystalline solute in water, the molar aqueous solubility (S_w) at high dilution is given by,

$$S_w = \frac{X_i^c}{\gamma_w} 55.5 \tag{1}$$

where X_i^c is the ideal crystalline mole fractional solubility, γ_w is the aqueous activity coefficient and 55.5 mol/L is the molarity of water. Combining Eq. (1) with the van't Hoff's equation for ideal solubility gives,

$$\log\ S_w = 1.74 - \log\ \gamma_w - \frac{\Delta S_m (T_m - T)}{2.303RT} \tag{2}$$

where ΔS_m is the entropy of melting, T_m and T are the melting point and temperature, respectively in Kelvin and R is the gas constant.

1.2. Aqueous activity coefficients

The molar aqueous activity coefficient (γ_w) is a group additive constitutive property and can be obtained from the AQUAFAC model (Myrdal et al., 1992, 1995, 1993; Pinsuwan et al., 1997), using the following relationship,

$$\log\ \gamma_w = \sum n_i q_i \tag{3}$$

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 aqueous activity coefficient.

[*] Corresponding author. Tel.: +1 862 778 5285; fax: +1 973 781 4554.
E-mail address: jainp@pharmacy.arizona.edu (P. Jain).

0378-5173/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpharm.2009.10.003

1.3. Melting points

In the absence of experimental data, the melting point of a drug must be estimated from its structure. Several methods exist for the prediction of melting points (Abramowitz and Yalkowsky, 1990; Austin, 1930; Constantinou and Gani, 1994; Dearden and Rahman, 1988; Joback and Reid, 1987; Krzyzaniak et al., 1995; Marrero and Gani, 2001; Simmamora and Yalkowsky, 1994). Recently, Jain and Yalkowsky (2006) proposed a model to predict the melting points solely from a group contribution approach and two non-additive, non-constitutive geometric properties. Thermodynamically, the melting point (T_m) may be obtained by the following relationship,

$$T_m = \frac{\Delta H_m}{\Delta S_m} \tag{4}$$

where ΔH_m is the enthalpy of melting and ΔS_m is the entropy of melting. The enthalpy of melting is assumed to be an additive constitutive property and is given by,

$$\Delta H_m = \sum n_i m_i \tag{5}$$

where m_i is the group contribution of group i to the heat of melting and n_i is the same as above. In general, entropy is the measure of the randomness of the molecules in a system. Eq. (4) indicates that a greater entropy corresponds to a lower a melting point for compounds with similar enthalpies. The first successful attempt to estimate the total entropy of melting is known as Walden's rule, which states that,

$$\Delta S_m^{tot} = 56.7\,\text{J/mol K} \tag{6}$$

Walden's rule was modified by Dannenfelser and Yalkowsky (1999) and again by Jain et al. (2004) to give,

$$\Delta S_m^{tot} = 50 - 19.1 \;\log \sigma + 7.4\Phi \tag{7}$$

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

$$\Phi = SP3 + 0.5(SP2 + RR) - 1 \tag{8}$$

where *SP3* is defined as the number of nonring, nonterminal sp^3 atoms (such as CH_2, CH, C, NH, N, O, S, etc.), *SP2* is the number of nonring, nonterminal sp^2 atoms (such as =CH, =C, =N, etc.) and *RR* is the number of rigid single or fused ring systems in the molecule. It is important to realize that both σ and Φ are properties of the whole molecule and are not group additive. Finally, the melting points are estimated from the following equation, which is obtained by substituting Eqs. (5) and (7) into Eq. (4).

$$T_m = \frac{\Delta H_m}{\Delta S_m} = \frac{\sum n_i m_i}{50 - 19.1\;\log \sigma + 7.4\Phi} \tag{9}$$

1.4. Solubility from SCRATCH

Incorporating Eqs. (3) and (9) into Eq. (2) gives

$$\log S_w = 1.74 - \log \sum (n_i q_i) - \frac{\Delta S_m((\sum n_i m_i/(50 - 19.1 \log \sigma + 7.4\Phi)) - 298)}{2.303R298} \tag{10}$$

which can be rearranged to,

$$\log\; S_w = 1.74 - \log \sum (n_i q_i) - \frac{\sum n_i m_i - 298(50 - 19.1 \log \sigma + 7.4\Phi)}{5709} \tag{11}$$

The molar aqueous solubility determined from the above equation is termed as the solubility predicted from SCRATCH, since only the chemical structure of the solute is needed.

1.5. General solubility equation (GSE)

The GSE is based on the fact that the aqueous solubility of a solute depends upon its crystallinity (Eq. (10)) and its octanol–water partition coefficient (log K_{ow}), which is a measure of its polarity, as per the following expression:

$$\log\; S_w = 0.5 - 0.01(MP - 25) - \log\; K_{ow} \tag{12}$$

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 derivation and assumptions of the GSE is described in detail by Ran et al. (2001).

2. Methods

2.1. Data

The data for the estimation of the melting points and aqueous activity coefficients of the 883 compound validation dataset for the SCRATCH model were obtained from Jain and Yalkowsky (2006) and Jain et al. (2008). The experimental molar aqueous solubilities were collected from WATERNT^{TM} v 1.0 EPA and AQUASOL databases. The partition coefficients (log K_{ow}) for the GSE were obtained from EPI Suite (2000). 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. Each compound was broken down into groups using the molecular fragmentation scheme of Jain and Yalkowsky (2006). The datasets containing the enthalpy of melting, entropy of melting, aqueous activity coefficients and group counts were prepared in Microsoft Excel 2000. Multiple linear regressions were performed using SPSS for Windows version 10.0 (SPSS Inc., Chicago, IL). The regression analysis generated the group contribution values, m_i and q_i.

2.2. Solubility from SCRATCH

As explained previously, the molar aqueous solubility (S_w) of an organic compound can be obtained from its activity coefficient (γ_w), enthalpy of melting (ΔH_m) and melting point (T_m) values using Eq. (2). However, Eq. (2) requires the use of experimental melting points. Jain and Yalkowsky (2006) have published a list of molecular fragments and proximity factors for various functional groups along with their enthalpic contributions (m_i). These group coefficients are used to calculate the predicted enthalpy values (Eq. (5)). The entropy values are calculated from the symmetry (σ) and flexibility (Φ) numbers using Eqs. (6) and (7). Thus, in order to over come the limitation of using the experimental values, all of the melting points in this study were predicted from Eq. (4) using the calculated values for ΔH_m and the ΔS_m. The data for the aqueous activity coefficients for the same compounds were obtained using the Jain et al. (2008) model, which uses the sum of their group activity coefficients, i.e., $\sum n_i q_i$. Finally, Eq. (11) is used to calculate the SCRATCH solubilities for the nearly 900 compounds in the validation dataset.

2.3. Statistics

2.3.1. AAE and RMSE

The average absolute error (AAE) for each calculation was determined by

$$\mathrm{AAE} = \frac{\sum |X_{\mathrm{pred}} - X_{\mathrm{exp}}|}{N} \tag{13}$$

where X is either the melting point (T_m), the logarithm of aqueous activity coefficients ($\log \gamma_w$) or the logarithm of the aqueous solubility ($\log S_w$), and N is the total number of organic compounds. Similarly, the root mean-square error (RMSE) was determined by

$$\mathrm{RMSE} = \sqrt{\frac{\sum (X_{\mathrm{pred}} - X_{\mathrm{exp}})^2}{N}} \tag{14}$$

2.3.2. Cross-validation

A 10-fold cross-validation was performed on the complete data set of 883 compounds. It was based on 2230 calculated melting points from Jain and Yalkowsky (2006) and 1641 calculated activity coefficients from Jain et al. (2008). For each validation, approximately 1/10th of the data were randomly selected using the RAND function in Microsoft Excel 2000, and used as a test set. Each compound was included only once in each of the 9 training sets and in one test set. Each round was run in the following manner: The test set compounds were deleted from the complete enthalpy data (2230 compounds) and from the complete aqueous activity coefficient data (1641 compounds). The remaining compounds in both data sets were treated as new training sets. Any compound in the test set with a group or fragment not present in the training set was deleted from the test set. This was done to ensure true prediction from the training set. Regressions were run to obtain the enthalpic, as well as the activity coefficient group contribution values, i.e., m_i and q_i, respectively. These were then used to obtain the solubility values for the test set. The AAEs for the SCRATCH solubility were calculated for each round and averaged.

3. Results and discussion

The complete alphabetical list of the compounds studied, their experimentally determined aqueous solubilities, as well as their SCRATCH, and GSE predicted values is provided in Appendix B. The experimental and predicted melting points, enthalpies of melting and aqueous activity coefficients, as well as the molecular symmetry numbers, molecular flexibility numbers, and the partition coefficients (ClogP) of all the compounds have been published previously (Jain and Yalkowsky, 2006; Jain et al., 2008). The compounds range from 1.49 E-11 M to 0.99 M in their molar aqueous solubilities and from 85.5 K to 710.5 K in their melting points.

3.1. Aqueous activity coefficients prediction

A plot of experimental versus the predicted aqueous activity coefficients yields a regression line with a slope of 0.976 and an R^2 of 0.859 (Fig. 1). The average absolute error (AAE) in the prediction of aqueous activity coefficients is 0.484 log units and the RMSE is 0.674. The near overlap of the regression generated line and the line of identity illustrates the closeness of predicted values to the true values.

3.2. Melting point prediction

Fig. 2 shows the relationship between the experimental and the predicted melting points. The regression line has a slope of 0.989 and an R^2 of 0.841. The average absolute error in the prediction of melting points for all the complete validation dataset compounds is 33.1 K and the RMSE is 43.3 K. As in the case of the activity coefficients, the agreement in the predicted and experimental values of the melting points is evidenced by the near overlap of the regression line and the line of identity.

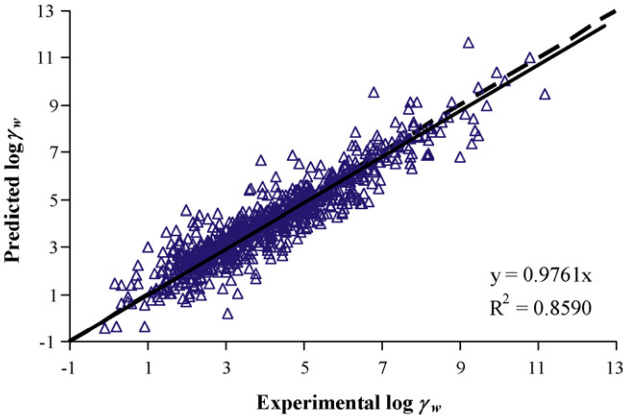

Fig. 1. Plot of predicted and experimental aqueous activity coefficients (γ_w) for 883 compounds (solid line: regression line; dashed line: line of identity).

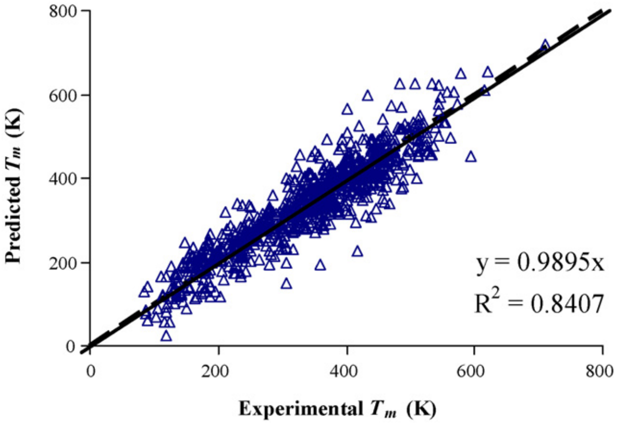

Fig. 2. Plot of predicted and experimental melting points for 883 compounds (solid line: regression line; dashed line: line of identity).

3.3. Aqueous solubility prediction from SCRATCH model

Finally, predicted aqueous activity coefficients, entropies of melting, and melting points were used to calculate the solubilities

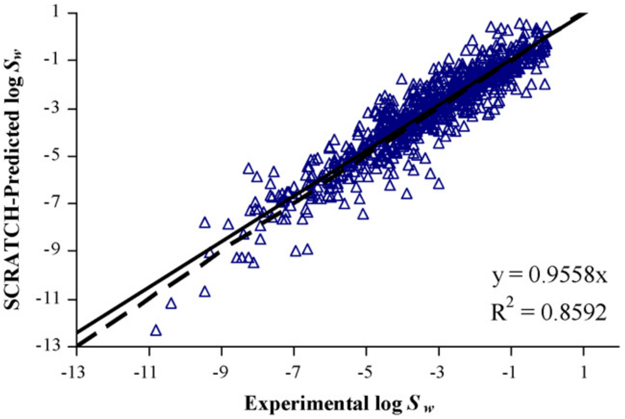

Fig. 3. Plot of logarithm of experimental and SCRATCH-predicted aqueous molar solubilities (solid line: regression line; dashed line: line of identity).

Table 1
Ten-fold cross-validation of the SCRATCH model.

	Training set		Test set				
	Melting point	Activity coefficient	Size	Deleted	AAE	R^2	% Error ≤1 log unit
Round 1	2142	1553	87	1	0.771	0.608	78.4
Round 2	2142	1553	88	0	0.774	0.795	67.04
Round 3	2142	1553	86	2	0.941	0.633	65.12
Round 4	2142	1553	85	3	0.730	0.717	78.82
Round 5	2142	1553	88	0	0.755	0.784	65.91
Round 6	2142	1553	86	2	0.877	0.680	65.12
Round 7	2142	1553	88	0	0.649	0.842	76.14
Round 8	2142	1553	88	0	0.634	0.820	80.68
Round 9	2142	1553	88	0	0.813	0.701	70.45
Round 10	2139	1550	91	0	0.658	0.762	76.92
Average					0.760	0.734	72.46

from the SCRATCH model (Eq. (11)). Fig. 3 shows the relationship between the experimental and the SCRATCH predicted values. The regression line through the origin has a slope of 0.956 and an R^2 of 0.859. About 84% of the molar solubilities were predicted within 1 log unit of their reported values. These results are noteworthy considering the size and diversity of the compounds which span over 10 orders of magnitude in molar aqueous solubility. The solubilities obtained using SCRATCH, are purely predicted with no experimental solubility data used. This is a major advantage over existing models.

3.4. Cross-validation

The result of the cross-validation is shown in Table 1. The AAE of the aqueous solubilities in logarithmic units of each of the ten test sets are given in the last column. The overall mean AAE of the ten rounds is 0.760. This value is the true prediction error of the SCRATCH model for the compounds studied. Also on an average about 72% of the aqueous solubility data can be predicted within 1 log unit of its AAE.

3.5. Comparison with the GSE

Table 2 summarizes the results of the SCRATCH and GSE-estimated solubilities. The average absolute errors of the SCRATCH and the GSE models for the given data set are 0.760 and 0.656 respectively. The factors responsible for this difference are discussed below.

The SCRATCH equation utilizes two sets of group contribution values. The m_i value along with the entropic factors (σ and Φ) gives an estimate of the role of crystallinity in determining solubility. The q_i value is based upon activity coefficients that were calculated from experimental aqueous solubility data for liquids and hypothetical super cooled liquids. Thus, the error in the SCRATCH model could result from the error in the prediction of melting points and/or from the aqueous activity coefficients.

The GSE uses experimentally determined melting points and ClogP values. Since melting point determinations are generally quite accurate, the GSE error results primarily from ClogP estimates. Furthermore, the major assumption in the GSE, that the octanol is an ideal solvent for all the solutes, may not be true for strongly hydrogen bonding and strongly self associating compounds. Therefore, as compared to the SCRATCH equation, the crystal term is more accurate in the GSE because the actual melting point is used. On the other hand the aqueous activity coefficients are more accurate in the SCRATCH equation because they are based directly on solubility data, whereas the GSE utilizes ClogP to estimate the aqueous activity coefficients.

Table 2
Results of aqueous solubility predictions.

Parameter	SCRATCH	GSE
Number of compounds	883	883
Slope	0.907	0.912
R^2	0.734	0.781
AAE (log unit)	0.760	0.656
% Error ≤1 log unit	72.46	77.60

Overall, the GSE is simpler but requires an experimental melting point value, whereas the SCRATCH needs only the structure of a compound in order to predict its aqueous solubility.

4. Conclusion

SCRATCH is a semi-empirical algorithm for the estimation of aqueous solubilities using predicted melting points and predicted aqueous activity coefficients. The AAE for the prediction of melting points and aqueous activity coefficients for a validation dataset are 33.1 °K and 0.484 log units, respectively. These melting points and the activity coefficients are finally used to obtain the SCRATCH aqueous solubilities, with an AAE of 0.760 log units. The prediction ability of the model is confirmed by cross-validation.

The comparison resulted in the GSE being slightly more accurate than the SCRATCH model for the same set of compounds. This could be explained by the fact that the SCRATCH uses predicted melting points as well as predicted aqueous activity coefficients. Since the errors from both the predicted values can propagate, the average error can be expected to be greater than that of the GSE in which only one property (the partition coefficient) is predicted. The GSE requires experimental melting points for the solubility prediction whereas the SCRATCH does not need any experimental values. This is a big advantage in early drug discovery for newly synthesized compounds, when the drug is not completely characterized or the experimental melting points are not available.

The SCRATCH model provides an accurate and widely applicable tool for estimation of aqueous solubility values of organic compounds from their chemical structures, two sets of group contribution values, and two non-additive geometric parameters.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ijpharm.2009.10.003.

References

Abramowitz, R., Yalkowsky, S.H., 1990. Melting point, boiling point and symmetry. Pharm. Res. 7, 942.

AQUASOL, 2001. AQUASOL Database. University of Arizona, Tucson, AZ.

Austin, J.B., 1930. A relation between the molecular weight and melting points of organic compounds. J. Am. Chem. Soc. 52, 1049.

ClogP, 2009. ClogP Software V4.0. Biobyte Corporation.

Constantinou, L., Gani, R., 1994. New group contribution scheme for estimating properties of pure compounds. AIChE J. 40, 1697.

Dannenfelser, R.-M., Yalkowsky, S.H., 1999. Predicting the Total Entropy of Melting: Application to Pharmaceuticals and Environmentally Relevant Compounds. J. Pharm. Sci. 88, 722–724.

Dearden, J.C., Rahman, M.H., 1988. QSAR approach to the prediction of melting points of substituted anilines. Math. Comput. Model 11, 843.

EPI Suite, MPBPWIN V1.41, 2000. U.S. Environmental Protection Agency.

Jain, N., Yalkowsky, S.H., 2001. Estimation of the aqueous solubility. I. Application to organic nonelectrolytes. J. Pharm. Sci. 90, 234–252.

Jain, A., Gang, Y., Yalkowsky, S.H., 2004. Estimation of total entropy of melting of organic compounds. Ind. Eng. Chem. Res. 43, 4376–4379.

Jain, A., Yalkowsky, S.H., 2006. Estimation of melting points of organic compounds-II. J. Pharm. Sci. 95, 12.

Jain, P., Sepassi, K., Yalkowsky, S.H., 2008. Comparison of aqueous solubility estimation from AQUAFAC and the GSE. Int. J. Pharm. 360, 122–147.

Joback, K.G., Reid, R.C., 1987. Estimation of pure component properties from Group contributions. Chem Eng. Commun. 57, 233.

Krzyzaniak, J.F., Myrdal, P.B., Simmamora, P., Yalkowsky, S.H., 1995. Boiling point and melting point prediction for aliphatic, non-hydrogen bonding compounds. Ind. Eng. Chem. Res. 34, 2530.

Marrero, J., Gani, R., 2001. Group-contributino based estimation of pure component properties. Fluid Phase Equil. 183, 183.

Myrdal, P., Ward, G.H., Dannenfelser, R.-M., Mishra, D.S., Yalkowsky, S.H., 1992. AQUAFAC 1: aqueous functional group activity coefficients: application to hydrocarbons. Chemosphere 24, 1047–1061.

Myrdal, P., Ward, G.H., Simamora, P., Yalkowsky, S.H., 1993. AQUAFAC: Aqueous Functional Activity Coefficients. SAR QSAR Environ. Res. 1, 53–61.

Myrdal, P.B., Manka, A., Yalkowsky, S.H., 1995. AQUAFAC 3: aqueous functional group activity coefficients: applications to the estimation of aqueous solubility. Chemosphere 30, 1619–1637.

Pinsuwan, S., Myrdal, P.B., Lee, Y.C., Yalkowsky, S.H., 1997. AQUAFAC 5: aqueous functional group activity coefficients; applications to alcohols and acids. Chemosphere 35, 2503–2513.

Ran, Y., Jain, N., Yalkowsky, S.H., 2001. Prediction of aqueous solubility of organic compounds by the General Solubility Equation (GSE). J. Chem. Inf. Comput. Sci. 41, 1208–1217.

Ran, Y., He, Y., Yang, G., Johnson, J.L.H., Yalkowsky, S.H., 2002. Estimation of aqueous solubility of organic compounds by using the general solubility equation. Chemosphere 48, 487–509.

Simmamora, P., Yalkowsky, S.H., 1994. Group contribution methods for predicting the melting point and boiling point of aromatic compounds. Ind. Eng. Chem. Res. 33, 1405.