High-Throughput, *In Silico* Prediction of Aqueous Solubility Based on One- and Two-Dimensional Descriptors

Ola Engkvist and Paul Wrede*

CallistoGen AG, Neuendorfstrasse 24b, D-16761 Hennigsdorf, Germany

Received March 20, 2002

An aqueous solubility model has been developed. The model is based solely on one- and two-dimensional descriptors and an artificial neural network to ensure fast execution. 63 descriptors expressing physicochemical and topological properties were used. The final model consisted of a training set of 3042 molecules, a test set of 309 molecules and an independent validation set of 307 molecules. The squared correlation coefficients were 0.91 for the training set, 0.89 for the test set and 0.86 for the independent validation set.

INTRODUCTION

The rapid development of combinatorial chemistry and high throughput screening (HTS) has increased the number of identified lead candidates. However, due to pharmacokinetic liabilities, it has been difficult to turn all these candidates into drugs.[1,2] One of the most important pharmacokinetic problems is low aqueous solubility, which leads to low bioavailability. Therefore a lot of effort has been invested in developing in vitro solubility measuring equipment and *in silico* based solubility prediction tools. For a review of the present status of experimental work see ref 3. Several *in silico* prediction tools have been developed during the last years. Most of the prediction tools are based on molecular descriptors.[4−7] Methods based on molecular group contributions[8] and experimental data[9,10] have also been developed.

METHODS

The development of an *in silico* aqueous solubility model can be divided into three consecutive steps: I) compilation of an accurate data set, II) selection of a method to approximate the functional relationship between the molecular descriptors and the aqueous solubility, and III) selection of a suitable set of descriptors. In this study two different data sets were used: the first consisted of 1318 molecules (set A), which was used in several published studies,[4−7] the second and larger set was derived from the database PHYSPROP (set B).[11] Since the PHYSPROP database was gathered from literature, a lot of compounds not relevant for drug discovery were included. The data set was therefore filtered to enhance the “druglikeness”. The applied filters are described in the reference.[12] 5350 compounds remained after filtering. The data set was further tailored by keeping only compounds where the solubility was measured between 20 and 25 °C. 3351 compounds fulfilling this selection criterion were found. There is some overlap between the two data sets, since the data set A consisted partially of molecules from the PHYSPROP database. A total of 647 identical molecules was found in the two databases, which correspond to 60% of set A and to 19% of set B.

Supervised artificial neural networks (ANNs) were used to approximate the functional relation between the molecular descriptors and the solubility. ANNs have successfully been applied to predicting druglikeness.[13,14] The main advantage of ANN is the inclusion of nonlinear relations in the model.[15] In a previous study ANN were shown to be superior to multilinear regression methods for aqueous solubility prediction.[4] All our ANN calculations were performed with the SNNS program package.[16] The training of the feed-forward neural networks was performed using the back-propagation momentum method.[17] All trained nets had 63 input neurons, 5 hidden neurons and 1 output neuron. All layers were fully connected resulting in 320 weights. The training was performed over 4000 cycles with a learning rate of 0.2 and a momentum term of 0.1. For technical reasons all input and output values were linearly scaled to between 0.1 and 0.9. During each epoch the training patterns were presented to the network in randomized order.

The most critical issue is the descriptor selection. Aqueous solubility is a complex process, which ultimately depends on the intermolecular interactions in the crystal and water phase. It is important that the descriptor set describes both the breaking of intermolecular bonds in the crystal and the solute−water interactions. This has been exemplified by Ran et al.,[9] who successfully applied a simple equation, where the solubility is correlated to the experimentally measured melting point and logP of the compounds. We implemented 63 descriptors for various topological and physicochemical properties that we considered might be of importance in modeling aqueous solubility. The descriptors express the composition, topology, hydrogen bonding capacity and lipophilicity of the molecules. These descriptors are related both to the crystal energy and to the solute water interactions. The descriptors are given in the Supporting Information. We plan to use the solubility prediction tool together with virtual synthesis[18] and virtual screening[19] tools. It is therefore preferable to use descriptors that are fast to calculate, like simple one-dimensional (1D) and two-dimensional (2D) instead of more complicated three-dimensional (3D) descriptors. 3D descriptors are based on molecular geometry, while

* Corresponding author phone: +49-(0)-3302-202 4500; fax: +49-(0)-3302−202 4555; e-mail: paul.wrede@callistogen.com.

10.1021/ci0202685 CCC: $22.00
Published on Web 08/10/2002

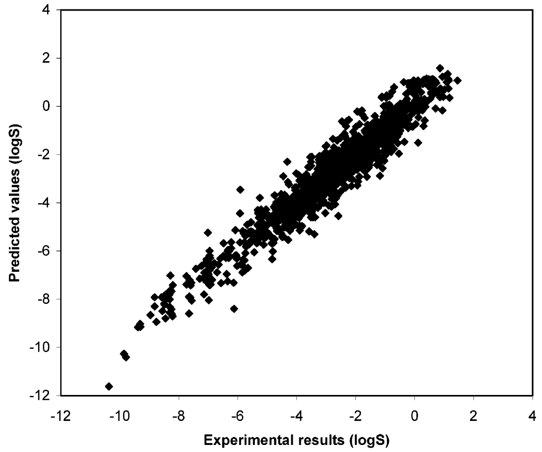

Figure 1. Correlation between predicted and experimentally measured logS for the test set in the first model.

Table 1. Results for the Two Developed Models, the Data Set Size (N), Squared Correlation Coefficient (r^2), and the Standard Deviation (SD) in logS Units Are Given

	model 1			model 2		
	N	r^2	SD	N	r^2	SD
training set	1160	0.96	0.56	3042	0.91	0.84
test set	130	0.95	0.57	309	0.89	0.87
validation set	2767	0.79	1.18	307	0.86	0.80

1D descriptors are based on molecular composition, and 2D descriptors on atomic connectivity.

RESULTS

Two different models were developed for solubility prediction. In the first model, 90% of data set A was randomly chosen as a training set, and the rest was used as a test set. The molecules of data set B that were not included in set A were used as an independent validation set. The procedure was repeated in a 10-fold cross validation to avoid chance correlations. The results are presented in Figure 1 and Table 1. The average squared correlation coefficient (r^2) was 0.96 (standard deviation 0.56 logS units) for the training set and 0.95 (standard deviation 0.57 logS units) for the test set. However, the result for the validation set was rather poor, yielding a squared correlation coefficient of 0.79 (standard deviation 1.18 logS units).

Therefore, a second model was developed, this time with the data set B derived from the PHYSPROP database. 90% of the compounds were randomly selected as the training set and the rest were used as the test set. The procedure was repeated in a 10-fold cross validation to avoid chance correlations. The molecules of set A but not of set B were used as validation data. The average squared correlation coefficient (r^2) was 0.91 (standard deviation 0.84 logS units) for the training set and 0.89 (standard deviation 0.87) for the test set (Table 1 and Figure 2). For the independent validation set, a squared correlation coefficient of 0.86 (standard deviation 0.80 logS units) was obtained. The data are presented in Figure 3. This squared correlation coefficient is significantly better than for the validation set in the first model, indicating that the second model is the one for predicting aqueous solubility. In the second data set there are a few outliers; visual inspection did not reveal any reason they are outliers. However, it is not unlikely that, out of several thousand experimentally measured values, a few of them might be wrong.

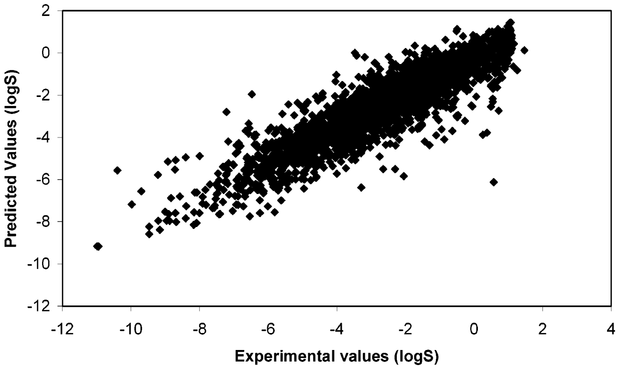

Figure 2. Correlation between predicted and experimentally measured logS for the test set in the second model.

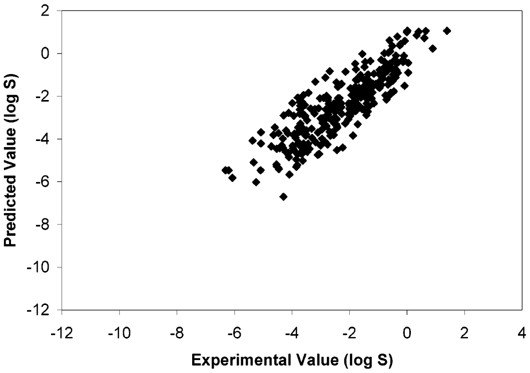

Figure 3. Correlation between predicted and experimentally measured logS for the validation set in the second model.

The first data set (set A) used in our study was used in several studies.[4−7] All studies successfully predicted the solubilities for this data set. Squared correlation coefficients between 0.92 and 0.94 were reported. However, only in the study by Bruneau[6] was an additional data set used. He used a proprietary data set of 522 molecules to validate his model. His conclusion agrees with ours, that set A might give poor predictions of solubilities outside the data set. His model is based on 3D descriptors. 3D descriptors are more time-consuming to calculate than the simpler 1D and 2D descriptors and therefore are not suitable when large virtual databases are profiled. In our opinion the reason that the second model performs better than the first is that data set B is larger and more diverse than data set A, since a model from data set A is not successful in predicting data set B. However, a model derived from data set B successfully predicts the molecules in data set A not included in data set B. The validation set in the second model has a surprisingly low standard deviation (0.80 logS units), in fact slightly lower than the standard deviation for the training set (0.84 logS units) and the test set (0.87 logS units). This indicates that the validation set consists of molecules that can easily be predicted by the model. Thus the validation set resides in a part of the chemical space that is well covered and described by the training set. It is noted that the validation set does not include any molecules with lower logS than −6, i.e., there are no molecules in the validation set with very low aqueous solubility. Another possible explanation is that data set A is of higher quality than data set B. However, there is nothing in the method of deriving the data sets indicating this case. The data set overlaps partly as well.

In the first model there are 3.6 training examples per weight and in the second model 9.5 training examples per weight. The first model could therefore have a higher natural tendency to overfit the training data and be weaker for prediction than the second model. However, we did not observe any overfitting while training the first model.

To assess the usefulness of the new model, the solubilities for the World Drug Index[20] (WDI) and Available Chemical Directory[21] (ACD) databases were calculated. The databases were filtered according to the same rules as for set B.[12] After filtering, the WDI database was reduced to 50256 molecules and the ACD database to 215209 molecules. Calculating the solubility for all the molecules took less than 5 min on a Pentium III (1 GHz) processor. Thus the method is very fast and can be used to screen very large virtual libraries. To achieve oral absorption, a compound with medium intestinal permeability and a projected human potency of 1 mg/kg needs a minimum aqueous solubility of 52 μg/mL.[22] All molecules in the databases with solubility less than 52 μg/mL were therefore flagged as having potential absorption problems. In WDI 35.5% and in ACD 48.8% of the molecules were flagged, indicating that molecules used as drugs have higher aqueous solubility than a set of random organic molecules.

CONCLUSIONS

A data set of 3351 molecules was extracted from the PHYSPROP database. The data set was used to train an ANN with 63 1D and 2D descriptors. An independent data set of 307 compounds derived from a different source was used to validate the model. The trained net successfully predicted the solubilities of the independent validation set. The model includes only 1D and 2D descriptors, which makes it suitable for use with virtual synthesis and virtual screening. However, as with all models based on experimental data, the accuracy can never be better than the accuracy of the data used to parametrize it. It is expected that a better model could be developed with a data set not obtained from published sources but measured in-house. Another source of improvement would be to include descriptors that distinguish between cis/trans isomers.

ACKNOWLEDGMENT

We are grateful to Prof. Jarmo Huuskonen for providing us with his data set in a computer readable format and to Prof. Gisbert Schneider and Dr. Ulrich Rester for many fruitful discussions.

Supporting Information Available: Sixty-three molecular descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Smith, D. A.; van de Waterbeemd, H.; Walker, D. K. *Pharmacokinetics and Metabolism in Drug Design*; WILEY−VCH, 2001.

(2) *Pharmacokinetic Optimisation in Drug Research*; Testa, B., van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; WILEY−VCH, 2000.

(3) Avdeef, A. Physicochemical Profiling (Solubility, Permeability and Charge State). *Curr. Top. Med. Chem.* **2001**, *1*, 277−351.

(4) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(5) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compound Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

(6) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605−1616.

(7) Liu, R.; Sun, H.; So, S.-S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633−1639.

(8) Klopman, G.; Zhu, H.; Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 439−445.

(9) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208−1217.

(10) Peterson, D. L.; Yalkowsky, S. H. Comparison of Two Methods for Predicting Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1531−1534.

(11) PHYSPROP (Physical/Chemical Property database), Syracuse Research Corporation, SRC Environmental Research Centre, Syracuse, New York, 1994.

(12) The following rules for the elimination of undesired compounds were used: MW below 80 and above 800 are removed, the compound must contain at least one carbon. The compound must also contain at least one nitrogen, or oxygen or sulphur. The compound can have at most 6 Cl, Br and I. Only molecules consisting of the organic subset are allowed, i.e., H, C, N, P, C, S, O, F, Cl, Br and I. The following reactive groups are excluded: acylhalides, phosphanes, peroxides, isocyanates, anhydrides, acylhalides, acylcyanides, azides, cyanides and diazonium compounds. Compounds crystallized as salts are also excluded.

(13) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem*. **1998**, *41*, 3325−3329.

(14) Ajay; Bemis, G. W.; Murcko, M. A. Can We Learn To Distinguish between “Drug-like” and “Nondrug-like” Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(15) Schneider, G.; Wrede, P. Artificial Neural Networks for Computer-based Molecular Design. *Prog. BioPhys. Mol. Biol.* **1998**, *70*, 175−222.

(16) SNNS (Stuttgart Neural Network Simulator), Version 4.2; University of Tübingen, 1998.

(17) *Neural Networks in Chemistry and Drug Design;* 2nd ed.; Zupan, J., Gasteiger, J., Eds.; Wiley-VCH: Weinheim, 1999.

(18) Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. De novo Design of Molecular Architectures by Evolutionary Assembly of Drug-derived Building Blocks. *J. Comput. Aid. Drug. Des*. **2000**, *14*, 487−494.

(19) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. “Scaffold Hopping“ by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem*. **1999**, *111*, 3068−3070; *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.

(20) WDI: (Derwent World Drug Index), Version 2001/01 (Issue 18), Derwent Information Ltd., 2001. http://www.derwent.com/.

(21) ACD: (Available Chemical Directory), Version 2001/01, MDL Information Systems, 2001. http://www.mdl.com/.

(22) Lipinski, C. A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharm. Tox. Methods* **2000**, *44*, 235−249.

CI0202685