Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation

Aixia Yan and Johann Gasteiger*

Computer-Chemie-Centrum and Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received August 23, 2002

Two quantitative models for the prediction of aqueous solubility of 1293 organic compounds were developed by a Multilinear Regression (MLR) analysis and a Back-Propagation (BPG) neural network. The molecules were described by a set of 32 values of a Radial Distribution Function (RDF) code representing the 3D structure and eight additional descriptors. The 1293 compounds were divided into a training set of 797 compounds and a test set of 496 compounds based on a Kohonen self-organizing neural network map. The obtained models show a good predictive power: for the test set, a correlation coefficient of 0.96 and a standard deviation of 0.59 were achieved by the back-propagation neural network approach.

INTRODUCTION

The solubility of organic compounds in water is an important property to be considered in the design of a drug, as it influences the uptake, distribution, transport, and eventually the bioactivity of a drug at the site of its actions. Thus, it is of high interest to estimate the aqueous solubility of new drug candidates at an early stage of the drug design process.[1]

Methods for solubility prediction have been reviewed recently,[1,2] and several models have been built. The models for the prediction of solubility are mainly based on the following: (1) experimentally determined physicochemical properties such as melting point and partition coefficient;[3−5] (2) group contribution schemes;[6,7] and (3) theoretically calculated molecular descriptors such as clog P, molecular topological indices, and so forth.[8−17] With the latter descriptors, the solubility of a compound can be estimated from its molecular structure directly. The models based on calculated molecular descriptors are suitable for general virtual screening and library design.

The main objectives of this study were the following: (1) to choose from the very beginning a set of descriptors that need no further selection and optimization, (2) the method should be fast enough to apply it also to a large data set and (3) a wide range of compounds should be processed, and (4) the prediction results should be at least as good as those obtained from the models of other authors.

The 3D structure of a molecule contains more chemical information than the 2D structure and plays an important role in drug design.[18] To examine the relationship between the solubility of a compound and its 3D structure, we propose a new method to predict solubility by using a set of 32 values of a Radial Distribution Function (RDF) code[19,20] representing the 3D structures, and eight additional descriptors, which characterize molecular polarizability, relative aromatic and aliphatic degree, and the ability of atoms in participating in hydrogen bonding.

The relationship between the structure of molecules and their solubility was investigated by a Kohonen self-organizing network, and two quantitative models were developed by a Multilinear Regression (MLR) analysis and a Back-Propagation (BPG) neural network.[21] The 1293 compounds were divided into a training set of 797 compounds and a test set of 496 compounds, according to their distribution in a Kohonen's self-organizing Neural Network (KNN) map. The architecture of the neural networks and the training epochs were optimized.

DATA SETS

Recently, a new promising method for the prediction of aqueous solubility based on molecular topology and neural networks was proposed by Huuskonen.[15] The method was applied to a data set of 1297 diverse compounds taken from the AQUASOL database of the University of Arizona[22] and the PHYSPROP database.[23] Using this data set, some other groups derived new prediction models using different kinds of input descriptors and methods.[16,17]

In our work, the set of diverse compounds from ref 15 is investigated. The aqueous solubility values were measured at a temperature of 20−25 °C and are expressed as logS, where S is the solubility in mol/l. However, the number of molecules in this work is different from that in ref 15 because four compounds were eliminated for the following reasons: after checking the original file, it was found that the compounds saccharin and karbutilate were contained twice in ref 15, and thus we removed these doubles. The compound cyhexatin includes the element tin (Sn), and another compound, oryzalin, that could not be converted by the PETRA program,[24] were also excluded. Then, a set of 1293 compounds resulted.

METHODS

Structure Representation. We built our models by using a structure representation methods developed in our group.

* Corresponding author phone: +49-9131-8526570; fax: +49-9131-8526566; e-mail: Gasteiger@chemie.uni-erlangen.de.

10.1021/ci025590u CCC: $25.00
Published on Web 12/21/2002

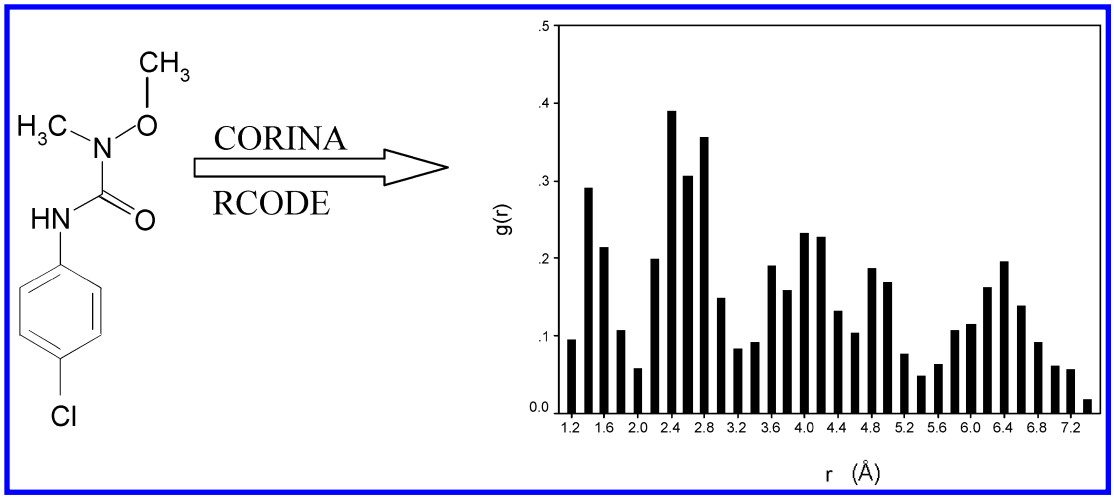

Figure 1. The example of RDF value for the compound monolinuron.

The molecules were described by a set of 32 Radial Distribution Function (RDF) code values[19,20] representing the 3D structure and eight additional descriptors. The 3D coordinates were obtained using the 3D structure generator CORINA.[25] CORINA requires only the connection table and optionally available stereochemical information to produce the Cartesian coordinates of the atoms. If stereochemical descriptors are missing, CORINA makes reasonable assumptions on the configuration at the stereocenters. Furthermore, CORINA tries to find a conformation with low energy. The 3D coordinates of the atoms are rapidly transformed into a set of RDF code by the program RCODE. Under the Linux 2.4 computer sever (PIII 600MHZ), for the 1314 compounds (Huuskonen's data set and the 21 drugs and agrochemicals), their 3D coordinates can be generated by CORINA program in 20 s, and their RDF code values can be converted by RCODE program in 4 s.

The RDF code has been proven to be a good representation for the 3D structure, which has several merits such as (1) independence from the number of atoms; (2) unambiguity regarding the three-dimensional arrangement of the atoms; (3) invariance against translation and rotation of the entire molecule; and (4) providing valuable information of the molecule, e.g. interatomic distances, ring conformation, and atom types. The RDF code has been successfully used for simulating the infrared spectra and deriving the 3D structure of organic molecules from their infrared spectra.[19,20]

The radial distribution function of an ensemble of N atoms can be interpreted as the probability distribution to find an atom in a spherical volume of radius r. The RDF function used in this work is as follows:

$$g(r) = f\sum_{i}^{N-1}\sum_{j>i}^{N} A_i A_j \cdot e^{-B(r-r_{ij})^2} \qquad (1)$$

With

$$f = \frac{1}{\sqrt{\sum_{r}[g(r)]^2}} \qquad (2)$$

f is a scaling factor and N is the number of atoms. By including characteristic atomic properties A of atom i and j, the RDF code can be used in different tasks to fit the requirements of the information to be represented. The exponential term contains the distance r_{ij} between the atoms i and j and the smoothing parameter B, which defines the probability distribution of the individual distances. $g(r)$ was calculated at a number of discrete points with defined intervals.

Each molecule was represented by a vector of length 32. The parameter B was set to 25 Å^{-2} corresponding to a total resolution of 0.2 Å in the defined distance r of [1.0−7.4[Å. That means the 32 defined distance intervals are [1.0−1.2[, [1.2−1.4[, ... [7.2−7.4[. The RDF for the structure derivations was calculated with the atomic number as atomic property. For example, the RDF value of the compound monolinuron is shown in Figure 1. Other atomic properties such as partial charges have also been investigated with no increase in predictive power. Thus, the simplest property, atomic number was used.

Additionally, eight descriptors were calculated by PETRA,[24] including the following: mean molecular polarizability,[26] aromatic indicator of a molecule, aliphatic indicator of a molecule, highest hydrogen bond acceptor potential, highest hydrogen bond donor potential, number of hydrogen bond donor groups, and the number of atoms of element nitrogen and oxygen.

The relative aromatic and aliphatic degrees of a molecule were characterized by the aromatic and aliphatic indicator. The aromatic indicator of a molecule is equal to the numbers of aromatic atoms divided by the total number of atoms (excluding hydrogen atoms) in the molecule. The aliphatic indicator of a molecule is equal to the number of sp^3 carbons divided by the total number of carbon atoms in this molecule.

The ability of a molecule in participating in hydrogen bonding was described by the highest hydrogen bond acceptor potential, highest hydrogen bond donor potential, the number of hydrogen bond donor groups, and the number of atoms of elements nitrogen, and oxygen. The highest hydrogen bonding acceptor potential is equal to the maximum lone-pair electronegtivity on the atoms of N, O, or F in a compound. The highest hydrogen bonding donor potential is equal to the most positive charge on the hydrogen atom in the groups −OH, −NH, and −SH in one compound.

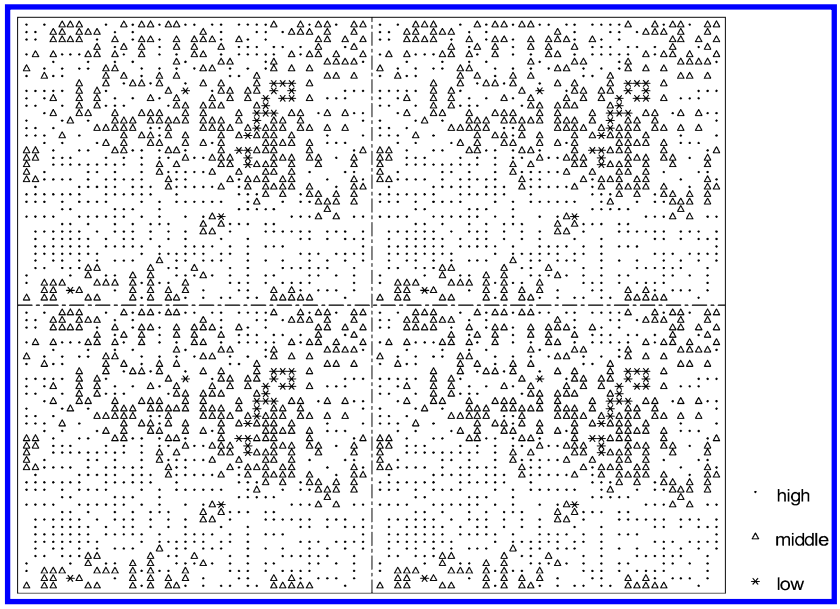

Figure 2. A 4-fold toroidal KNN map for 1293 compounds by using the 40 input descriptors. High means compounds with high solubility where *logS* is in the range of [−2.82 ∼ 1.58], middle means compounds with middle solubility where *logS* is in the range of [−7.21 ∼ −2.83], and low means compounds with low solubility where *logS* is in the range of [−11.62 ∼ −7.22].

In total, each compound was represented by 40 descriptors.

Training/Test Set Selection by Kohonen Self-Organizing Neural Network. The Kohonen's self-organizing Neural Network (KNN) has the special property of effectively creating a spatially organized internal representation of various features of input signals and their abstractions.[21] The perception of the similarity of objects is an essential feature. In a self-organizing neural network the neurons are arranged in a two-dimensional array to generate a two-dimensional feature map such that similarity in the data is preserved. In other words, if two input data vectors are similar, they will be mapped into the same neuron or into neurons close together in the two-dimensional map.

A Kohonen self-organizing neural network was applied to separate the data set into a training set and a test set. The division based on a KNN map is superior over the random selection. The advantage of such a procedure was shown in previous work.[27] This method for splitting a data set into training and test set ensures that both sets cover the information space as well as possible. As the test set was not used during training of the MLR or BPG model, it still can be considered as an external data set.

In this work, the following programs and software packages were applied. The CACTVS system was used for structure management, editing, comparing, and data extracting.[28] CORINA was employed for rapid generation of 3D coordinates.[25] The PETRA program was applied for the calculation of physicochemical properties in organic molecules.[24] RCODE program was used for the calculation of Radial Distribution Function (RDF) codes from the 3D coordinates. SONNIA (formerly KMAP) was utilized for building Kohonen self-organizing neural network.[29] SPSS software was used for multilinear regression analysis.[30] SNNS was used for constructing the Back-Propagation (BPG) neural network.[31]

RESULTS AND DISCUSSION

A toroidal KNN with 39 × 38 neurons is utilized with the 40 descriptors used as input vectors. The initial learning spans are 19.5 and 19, with an initial learning rate of 0.7 and a rate factor of 0.95. The initial weights are randomly initialized, and training was performed for a period of 1600 epochs in an unsupervised manner. A map was formed according to the ranges of solubility of the most frequently occupied neuron. From Figure 2, one can see that compounds with a different range of solubility are projected into different areas.

In the Kohonen map, 797 of a total of 1482 neurons are occupied. Afterward, one object of each neuron was taken for the training set, and the other objects represented the test set. Thus, the 1293 compounds were divided into a training set of 797 compounds and a test set of 496 compounds after the KNN classification.

Build a Model by MLR (Multilinear Regression) Analysis. The multilinear regression analysis was performed with the SPSS software using 40 descriptors as input variables. The 797 compounds in the training set were used to build a model, and the 496 compounds were used for the prediction of solubility. The prediction results are shown in Table 1 and Figures 3 and 4. For the training set, $r = 0.89$ ($r^2 = 0.79$), $s = 0.93$, $MAE = 0.70$, and $n = 797$, and for the test set $r = 0.90$ ($r^2 = 0.82$), $s = 0.79$, $MAE = 0.68$, and $n = 496$. (r is the correlation coefficient, s is the standard deviation, and MAE is the mean absolute error, which equals to the mean value of the absolute errors.)

Build a Model by BPG (Back-Propagation Neural Network). The SNNS program[31] was used for Back-Propagation Neural Network. A standard back-propagation network was applied to estimate the solubility. An input layer with 40 neurons, an output layer with one neuron representing the *logS*, and a hidden layer of several neurons were used. All layers were completely connected. The initial weights were randomly initialized between −0.1 and 0.1.

Table 1. Comparison of the Prediction Power of the Models We Present Here with Other Published Models Based on Huuskonen's Data Set and a Data Set of Another 21 Compounds by Multilinear Regression (MLR) Analysis and Artificial Neural Network (ANN)[d]

		training set			test set			additional test set		
models		*n*	r^2	*s*	*n*	r^2	*s*	*n*	r^2	*s*
our model	MLR	797	0.79	0.93	496	0.82	0.79	21	0.56	1.20
	ANN	797	0.93	0.50	496	0.92	0.59	21	0.85	0.77
Huuskonen's model[a]	MLR	884	0.89	0.67	413	0.88	0.71	21	0.83	0.88
	ANN	884	0.94	0.47	413	0.92	0.60	21	0.91	0.63
Tetko's model[b]	MLR2	879	0.86	0.75	412	0.85	0.81	21	0.77	0.99
	ANN4	879	0.95	0.47	412	0.92	0.60	21	0.90	0.64
Liu's model[c]	ANN (7:2:1)	1033	0.86	0.70	258	0.86	0.71	21	0.79	0.93
	ANN (7:4:1)	1033	0.86	0.70	258	0.86	0.70	21	0.79	0.91

[a] Results from ref 15. [b] Best results from ref 16. [c] Best results from ref 17. [d] *n*: number of compounds; r^2: square of correlation coefficient; *s*: standard deviation.

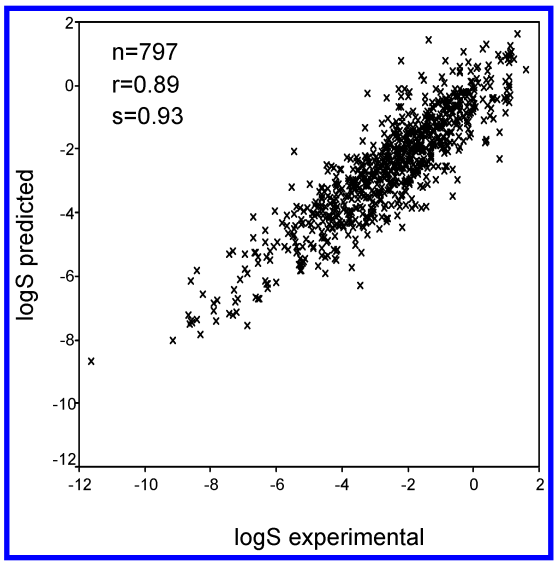

Figure 3. Predicted vs experimental solubility values of 797 compounds in the training set by multilinear regression analysis.

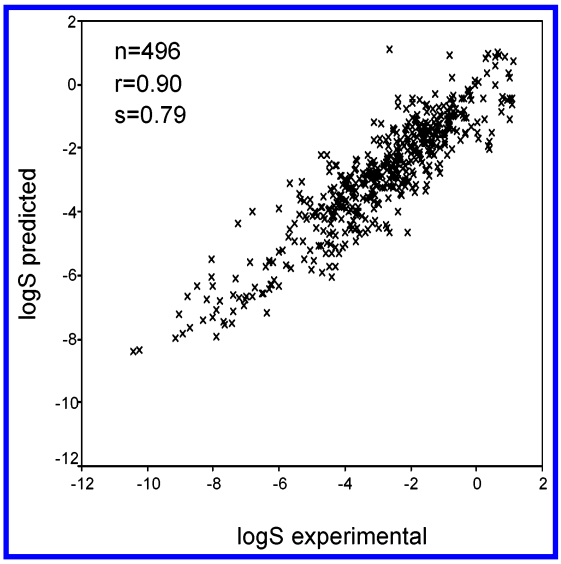

Figure 4. Predicted vs experimental solubility values of 496 compounds in the test set by multilinear regression analysis.

Each input and output value was scaled between 0 and 1. The net was trained following the "standard back-propagation" algorithm as implemented in SNNS, employing a learning rate of 0.2.

Again, 797 compounds were used as training set and the other 496 compounds as test set. In the process, the architecture of neural network was optimized. The number of hidden layer neurons was varied from 5 to 10. The optimized neural network architecture was 40-8-1. The best number of training epochs was selected by the early stopping method in order to avoid overtraining, and it was 6000. For the training set, $r = 0.97$ ($r^2 = 0.93$), $s = 0.50$, $MAE = 0.41$, and $n = 797$, and for the test set, $r = 0.96$ ($r^2 = 0.92$), $s = 0.59$, $MAE = 0.49$, and $n = 496$. The results are shown in Figures 5 and 6 and Table 1.

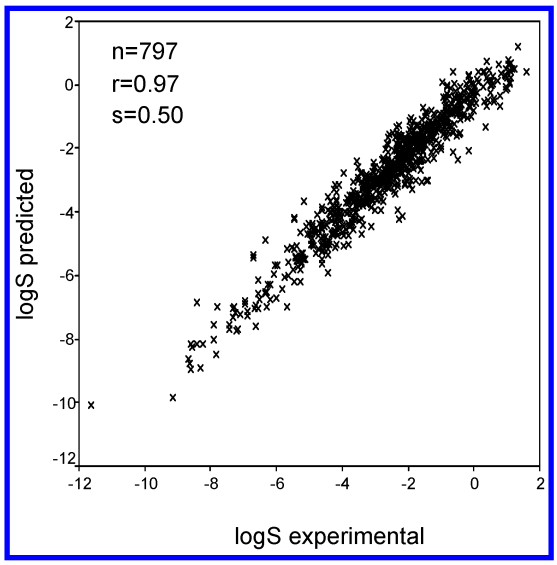

Figure 5. Predicted vs experimental solubility values of 797 compounds in the training set by back-propagation neural network.

Additionally, another test set designed by Yalkowsky,[15,16] that comprises 21 compounds of drugs and agrochemicals, was used for testing the derived models. The prediction results are shown in Tables 1 and 2.

Table 1 shows the predicted results of aqueous solubility of our models compared with other works.[15−17] The prediction results of neural network of our model are similar to those of Huuskonen's and Tetko's models.

We used another data set from Merck KGaA for testing the models. The data set comprises 2743 compounds in total. After excluding the overlap with the Huuskonen data set and selecting only those values that had been measured at a temperature of 20−25 °C, 1587 compounds were remaining and used for testing. Input and output values were scaled between 0 and 1, according to the larger ranges of descriptors in Huuskonen and Merck data set. With the best architecture of the BPG network as derived above, the solubility for this

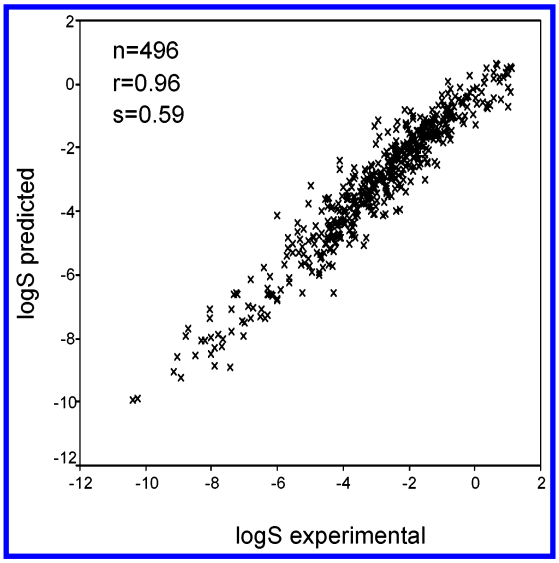

Figure 6. Predicted vs experimental solubility values of 496 compounds in the test set by back-propagation neural network.

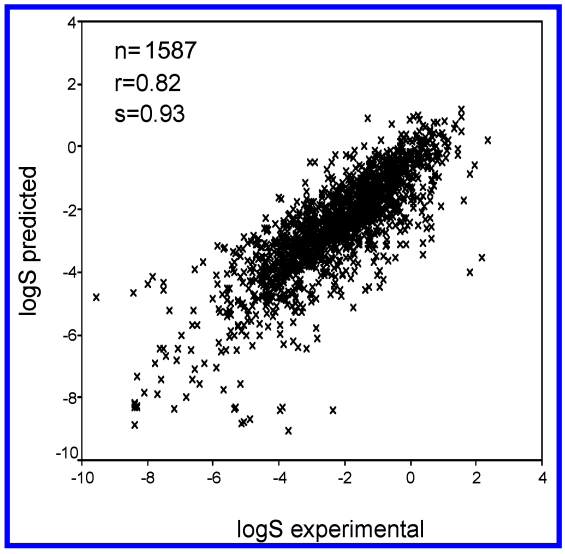

Figure 7. Predicted vs experimental solubility values of 1587 compounds of the Merck data set by back-propagation neural network.

Table 2. Predicted and Experimental Aqueous Solubility for 21 Compounds[15,16] by Multilinear Regression (MLR) Analysis and Back-Propagation (BPG) Neural Network

no.	CAS number	name	logS_exp	MLR	BPG
1	37680-73-2	2,2′,4,5,5′-PCB	−7.89	−7.12	−7.85
2	94-09-7	benzocaine	−2.32	−1.81	−2.19
3	50-78-2	aspirin	−1.72	−1.44	−1.87
4	58-55-9	theophylline	−1.39	−1.07	−1.27
5	60-80-0	antipyrine	0.39	−2.99	−1.31
6	1912-24-9	atrazine	−3.85	−2.44	−3.83
7	50-06-6	phenobarbital	−2.32	−2.81	−2.80
8	330-54-1	diuron	−3.80	−3.26	−3.70
9	67-20-9	nitrofurantoin	−3.38	−0.45	−2.52
10	57-41-0	phenytoin	−3.90	−2.99	−3.18
11	439-14-5	diazepam	−3.76	−5.19	−4.81
12	58-22-0	testosterone	−4.09	−4.11	−4.52
13	58-89-9	lindane	−4.64	−3.93	−5.04
14	56-38-2	parathion	−4.66	−2.64	−3.66
15	333-41-5	diazinon	−3.64	−3.06	−2.66
16	77-09-8	phenolphthalein	−2.90	−4.28	−4.62
17	121-75-5	malathion	−3.37	−3.45	−2.79
18	2921-88-2	chlorpyriphos	−5.49	−4.33	−4.79
19	363-24-6	prostaglandin_E2	−2.47	−4.06	−3.07
20	50-29-3	p,p′-DDT	−8.08	−6.60	−7.86
21	57-74-9	chlordane	−6.86	−6.41	−7.66

data set was estimated. The prediction results for this data set are represented in Figure 7 ($r = 0.82$, $s = 0.93$, $MAE = 0.77$, and $n = 1587$).

It was found that most descriptors obtained from the Merck data set have a larger range than those for the Huuskonen's data set. For instance, the mean molecular polarizability in Huuskonen's data set ranges from 5.17 to 65.80 Å^3, while in the Merck data set, it ranges from 2.34 to 96.62 Å^3. Obviously, the Merck data set contains more diverse compounds, that are not effectively represented in the training data set. Neural networks do not show the ability to extrapolate if the test set contains more information than the training set. Thus, we will build neural network models with the more diverse data set in the future.

CONCLUSIONS

The radial distribution function (RDF) code is a good method for representing the 3D structure of molecules. It describes in detail the 3D characteristics of molecules. This kind of structure representation method can be used for the prediction of molecular properties, such as solubility.

In this work, the 3D coordinates of a compound can rapidly be generated by CORINA, then Radial Distribution Function (RDF) codes can quickly be calculated by RCODE program. Under the Linux 2.4 computer sever (PIII 600MHZ), for the 1314 compounds (Huuskonen's data set and the 21 drugs and agrochemicals), their 3D coordinates can be generated by CORINA program in 20 s, their atomic property of atomic number can be calculated by PETRA program in 40 s, and then their RDF code values can be converted by RCODE program in 4 s. Afterward, a set of 32 RDF codes fully describes the detailed 3D information of a molecule, together with eight additional descriptors, which have low pairwise correlation, were used for building a model. The advantage of our method is that a cumbersome selection of descriptors can be avoided. The approach needs no experimental data for the description of compounds, and thus the method is suitable to virtual screening and library design.

The neural network approach provides better models than multilinear regression analysis. The models developed for the prediction of solubility can be applied to large data sets with rapid calculation speed, a wide range of compounds can be processed, and the prediction results of neural networks are as good as other models.

The Huuskonen data set is relatively limited in diversity, and the models based on a data set with larger structural variety will be built in the future.

ACKNOWLEDGMENT

Dr. Aixia Yan appreciates a Research Fellowship from the Alexander von Humboldt Foundation, financial support from the Bundesministerium Fuer Bildung und Forschung, and helpful discussion with M. C. Hemmer and Dr. L. Terfloth. We also thank Dr. J. Huuskonen, Dr. I. V. Tetko, and Merck KGaA for providing us with data sets.

Supporting Information Available: The name of compounds used in the test set in this study with their predicted and experimental aqueous solubility values (Table S). This

material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(2) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355−366.

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

(4) Ran, Y. Q.; 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.

(5) Yang, G.; Ran, Y. Q.; Yalkowsky, S. H. Prediction of the Aqueous Solubility: Comparison of the General Solubility Equation and the Method Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **2002**, *91*, 517−533.

(6) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.

(7) Kuhne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schuurmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.

(8) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601−609.

(9) Bodor, N.; Huang, M. J. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480−9483.

(10) Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.

(11) Mitchell, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.

(12) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247.

(13) Gao, H.; Shanmugasundaram, V.; Lee, P. Estimation of Aqueous Solubility of Organic Compounds with QSPR Approach. *Pharm. Res.* **2002**, *19*, 497−503.

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

(15) 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.

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

(17) Liu, R. F.; 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.

(18) Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical Information in 3D Space. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030−1037.

(19) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra. *Vibrat. Spectrosc.* **1999**, *19*, 151−164.

(20) Hemmer, M. C.; Gasteiger, J. Prediction of Three-Dimensional Molecular Structures Using Information from Infrared Spectra. *Anal. Chim. Acta* **2000**, *420*, 145−154.

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

(22) Yalkowsky, S. H.; Dannelfelser, R. M. *The ARIZONA dATAbASE of Aqueous Solubility*; College of Pharmacy, University of Arizona: Tucson, AZ, 1990.

(23) Syracuse Research Corporation. *Physical/Chemical Property Database (PHYSPROP)*; SRC Environmental Science Center: Syracuse, NY, 1994.

(24) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Compounds*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Heidelberg, 1988; pp 119−138. http://www2.chemie.uni-erlangen.de/software/petra/index.html.

(25) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581. http://www2.chemie.uni-erlangen.de/software/corina/index.html.

(26) Miller, K. J. Additivity methods in molecular polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533−8542.

(27) Simon, V.; Gasteiger, J.; Zupan, J. A Combined Application of Two Different Neural Network Types for the Prediction of Chemical Reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148−9159.

(28) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116. http://www2.chemie.uni-erlangen.de/software/cactvs/index.html.

(29) http://www2.chemie.uni-erlangen.de/software/kmap/.

(30) SPSS v 10.0, SPSS Inc., Chicago, IL, http://www.spss.com.

(31) SNNS: Stuttgart Neural Network Simulator, Version 4.2, Developed at University of Stuttgart, Maintained at University of Tübingen, 1995. http://www-ra.informatik.uni-tuebingen.de/SNNS/.

CI025590U