Journal of Computer-Aided Molecular Design **18:** 75–87, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods

Aixia Yan[1,2], Johann Gasteiger[1,*], Michael Krug[3] & Soheila Anzali[3]
[1]*Computer-Chemie-Centrum and Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany;* [2]*Present address: Department of Chemistry, Central Chemistry Laboratory, South Parks Road, Oxford OX1 3QH, UK;* [3]*Merck KGaA, Global Technology BCI, D-64271 Darmstadt, Germany*

Received 16 May 2003; accepted in revised form 18 February 2004

Key words: backpropagation, Kohonen's self-organizing neural network, multi-linear regression, neural network, radial distribution function, solubility

Summary

Several quantitative models for the prediction of aqueous solubility of organic compounds were developed based on a diverse dataset with 2084 compounds by using multi-linear regression analysis and backpropagation neural networks. The compounds were described by two different structure representation methods: (1) with 18 topological descriptors; and (2) with 32 radial distribution function codes representing the 3D structure of a molecule and eight additional descriptors. The dataset was divided into a training and a test set based on Kohonen's self-organizing neural network. Good prediction results were obtained for backpropagation neural network models: with 18 topological descriptors, for the 936 compounds in the test set, a correlation coefficient of 0.92, and a standard deviation of 0.62 were achieved; with 3D descriptors, for the 866 compounds in the test set, a correlation coefficient of 0.90, and a standard deviation of 0.73 were achieved. The models were also tested by using another dataset, and the relationship of the two datasets was examined by Kohonen's self-organizing neural network.

Abbreviations: BPG – backpropagation; KNN – Kohonen's self-organizing neural network; MLRA – multilinear regression analysis; MMP – mean molecular polarizability; RDF – radial distribution function.

Introduction

The aqueous solubility of organic compounds is an important molecular property that plays a significant role in pharmaceutical, environmental, and other physical and biological processes. Several models for the prediction of the solubility have been reported [1–20]. Most methods use *in silico* prediction tools where the organic compounds are represented by molecular descriptors derived from molecular structures [3–15], the models were constructed by techniques such as neural networks [4–14] or genetic algorithm and neural network [15]. Other models from experimental data [16–18] and the molecular group contribution method [19, 20] have also been generated.

Since any quantitative structure–property relationship (QSPR) model is constructed on a particular dataset, it is important to build a model by using a large, diverse dataset, and test the model by using other different datasets. In this work, we build solubility prediction models on a larger dataset than in our former studies [12, 13]. The compounds are represented by the same topological descriptors [12] and the 3D descriptors [13] as used before. We had carried out studies with the two structure representation methods on the widely used Huuskonen dataset [9]. Use of a larger dataset compiled at Merck KGaA (for prediction) showed the limited applicability of the Huuskonen dataset [12]. We therefore use the Merck

[*] To whom correspondence should be addressed. Fax: +49-9131-8526566; E-mail: Gasteiger@chemie.uni-erlangen.de

dataset to produce a model with hopefully a wider applicability.

The structures of the organic compounds were represented by two methods: (1) using 18 topological descriptors; (2) using 32 radial distribution function (RDF) codes [21, 22] and eight additional descriptors. The topological descriptors are proven to have a clear physicochemical interpretation, quite familiar to organic and medicinal chemists. The selection of descriptors was based on statistical analysis. The 3D descriptors contain more information than 2D descriptors, and the tedious descriptor selection can be avoided.

The following studies were performed with both structure representations: the relationship of structures and aqueous solubility of organic compounds was examined by Kohonen's self-organizing neural network (KNN), and the set of compounds was split into a training and a test set based on their distribution in a KNN map; the quantitative models were built by multi-linear regression analysis (MLRA) and back-propagation (BPG) neural network, and the models were also tested by using another dataset. The overlap of the two datasets was also investigated by their projection into a two-dimensional KNN map [23].

Datasets

Dataset A:

The dataset used for building the models was obtained from Merck KGaA Company, Germany, and contained 2827 compounds. The dataset had been compiled from the Beilstein database, different papers, and the Merck catalog. After removing duplicates, wrong structures, and restricting the data to those measured in the temperature range of 20–25 °C, 2084 compounds were left (in case of 3D descriptors an RDF code could be obtained for 2083 compounds). The aqueous solubility values are expressed as logS, where S is the solubility in mol/l, with a minimum value of -10.83, a maximum value of 2.35, and a mean value of -2.29. No information on the pH value was given. The average molecular weight and average logP for the Merck dataset were 217.43 and 1.73, respectively.

Dataset B:

The other dataset is the same as that from Huuskonen's work [9]. It consisted of a dataset of 1297 diverse compounds taken from the AQUASOL database of the University of Arizona [24] and the SCR PHYSPROP Database [25]. Using this dataset, we had developed some solubility prediction models [12, 13], and some other groups also derived new prediction models using different kinds of input descriptors and methods [10, 11]. However, the number of molecules in this work is different from that in Ref. 9 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. 9, and thus we removed these duplicates. The compound cyhexatin includes the element tin (Sn), and another compound, oryzalin, that could not be converted by the PETRA program [26–29], was also excluded. Then, a set of 1293 compounds resulted. The aqueous solubility values were measured at a temperature of 20–25degree and are expressed as logS, where S is the solubility in mol/l, with a minimum value of -11.62, a maximum value of 1.58, and a mean value of -2.72. The average molecular weight and average logP for the Huuskonen dataset were 199.0 and 2.55, respectively.

Representation of structures

Representation of structures by 2D descriptors

The molecules were represented by different kinds of topological descriptors. With a method similar to our previous work [12], in total 64 descriptors were calculated.

LogP (P is the partition coefficient of a solute between l-octanol and water) was calculated by a method based on the Ghose/Crippen approach [30–33].

All other descriptors were calculated with the program package PETRA (Parameter Estimation for the Treatment of Reactivity Applications) [26–29]. PETRA is a program package comprising various methods for the calculation of physicochemical properties in organic molecules. All methods are empirical in nature and have been developed and published over the last 20 years in our group.

Using PETRA, the following properties for each atom of every compound were calculated: σ charge, π charge, partial atomic charges, σ electronegativity, π electronegativity, lone-pair electronegativity and atomic polarizability. Based on this, autocorrelation vectors [34] were computed by the program AUTOCORR. In the autocorrelation vectors calculation, the hydrogen atoms were excluded. Topological autocorrelation vectors for each of the above seven physicochemical atomic properties were calculated for each molecule by using the following equation:

$$A\,(d) = \sum_{ij} p_i p_j \tag{1}$$

$A(d)$ is the topological autocorrelation coefficient referring to atom pairs i, j which are separated by d bonds. p_i is an atomic property, e.g. the σ charge on atom i. Thus, for each compound, a series of coefficients for different topological distances d, a so-called autocorrelation vector is obtained; seven distances from $d = 0$ to $d = 6$ were considered.

In statistical analyses, it was found that the seven 2D autocorrelation coefficients are highly correlated for the properties σ electronegativity, π electronegativity, and atomic polarizability; the first component of 2D autocorrelation coefficients has the highest standard deviation for the properties: σ charge, π charge, partial atomic charges. In the seven 2D autocorrelation coefficients for the property of lone-pair electronegativity, the first component has the second highest standard deviation. (The standard deviation of the seventh component is slightly larger than that of the first component. But the values of the seventh component for more than half of the compounds are zero, while the values of the first component for a few compounds are zero.) Thus, the first component ($d = 0$) of the autocorrelation coefficients for each property was selected for the following analysis. This value corresponds to the sum of the squares of each atomic property for a molecule.

Using PETRA, the other 14 descriptors were calculated. The polarizability of a molecule was represented by the mean molecular polarizability [35, 36]. The molecular weight is correlated to the size of the molecule. The relative aromatic and aliphatic degree of a molecule was characterized by aromatic and aliphatic indicator values. The aromatic indicator of a molecule (*i_aro*) is equal to the number of aromatic atoms divided by the total number of atoms (excluding hydrogen atoms) in the molecule. The aliphatic indicator of a molecule (*i_ali*) 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 to participate in hydrogen bonding was described by the number of hydrogen bond donor groups, the highest hydrogen bond acceptor potential, the highest hydrogen bond donor potential, and the number of atoms of the elements fluorine, nitrogen, and oxygen. The highest hydrogen bonding acceptor potential (M_H_ACC) is equal to the maximum lone-pair electronegativity on an atom considering all N, O, or F atoms in a compound. The highest hydrogen bonding donor potential (M_H_DON) is equal to the most positive charge on the hydrogen atom in the groups -OH, -NH, and -SH of a compound. In addition, the number of atoms of the elements hydrogen, carbon, sulfur, and chlorine were also calculated.

This first component of the autocorrelation coefficient of the seven physicochemical properties were put together with the other 14 descriptors, and with logP. Pairwise correlation analysis was then done; a descriptor was eliminated if the correlation coefficient was equal to or higher than 0.90. This left 18 descriptors, as shown in Table 1.

It should be mentioned that the selected 18 topological descriptors in this work were not only used for MLRA models, but also for neural network models. Neural networks could model the complicated nonlinear relationships between the input descriptors and the output vector and could provide better results. The 18 descriptors are all important because they reflect the different physicochemical properties of the compounds.

Representation of structures by 3D descriptors

As in our previous work [13], the compounds were also described by a set of 32 RDF (radial distribution function) code values representing the 3D structure of a molecule and eight additional descriptors. The 3D coordinates were obtained using the 3D structure generator CORINA [37].

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 [21, 22]:

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

with,

$$f = \frac{1}{\sqrt{\sum_r [g(r)]^2}} \tag{3}$$

f is a scaling factor and N the number of atoms. By including characteristic atomic properties A of atoms i and j, the RDF code can be used in different tasks to fit the requirements of the information to be represented.

Table 1. Multilinear regression models with all the 18 topological descriptors or 12 topological descriptors.

Descriptors	Coefficients[a]	t-score[b]	Coefficients[c]	t-score[d]
logP	−0.589	−14.675	−0.587	−16.956
enlp_1 = $\sum \chi_{LP}^2$ (χ_{LP}: lone-pair electronegativity)	−0.00137	−0.840		
enpi_1 = $\sum \chi_{\pi}^2$ (χ_{π}: π-electronegativity)	0.00115	1.338		
ensig_1 =$\sum \chi_{\sigma}^2$ (χ_{σ}: σ-electronegativity)	−0.00295	−5.112	−0.0025	−10.930
qpi_1 = $\sum q_{\pi}^2$ (q_{π}: π-charge)	−5.257	−6.968	−5.302	−7.545
qtot_1 =$\sum q_p^2$ (q_p: partial atomic charges)	0.575	1.364	0.793	2.796
Mean molecular polarizability (MMP)	0.0248	1.300	0.0252	2.701
Aliphatic indicator of molecule (i_ali)	0.112	0.557		
Aromatic indicator of molecule (i_aro)	−0.233	−1.136	−0.169	−1.482
Highest hydrogen bond acceptor potential (M_H_ACC)	0.0432	1.696	0.0428	1.747
Highest hydrogen bond donor potential (M_H_DON)	0.990	2.351	0.989	2.373
Hydrogen bond donor groups (#H_donors)	−0.0978	−2.806	−0.0933	−2.876
Number of atoms of hydrogen (#H-atoms)	0.0116	0.710		
Number of atoms of nitrogen (#N-atoms)	0.441	6.226	0.388	8.474
Number of atoms of oxygen (#O-atoms)	0.507	5.092	0.416	6.755
Number of atoms of fluorine (#F-atoms)	0.367	3.339	0.245	4.003
Number of atoms of sulphur (#S-atoms)	0.0194	0.285		
Number of atoms of chlorine (#Cl-atoms)	0.0892	1.340		

[a] Regression coefficients of the MLR model with all the 18 topological descriptors.
[b] The t-score values of the MLR model with all the 18 topological descriptors.
[c] Regression coefficients of the MLR model with the 12 topological descriptors.
[d] The t-score values of the MLR model with the 12 topological descriptors.

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.

The additional eight descriptors were also calculated by PETRA [26–29], including: mean molecular polarizability [35, 36], 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 the elements nitrogen and oxygen.

It should be pointed out that the selected 40 3D descriptors in this work were not only used for MLR models, but also for neural network models. Neural networks could model the complicated nonlinear relationships between the input descriptors and the output vector and could provide better results. For each molecule, a set of the 32 RDF codes reflects the complete three-dimensional information in the defined distance r of [1.0–7.4[Å, and the other eight descriptors correspond to the molecular hydrogen bonding ability and some other important physicochemical properties of the compounds.

Training and test set selection method

A Kohonen's self-organizing Neural Network (KNN) includes an input layer and an active layer of output neurons [23]. The output layer with competitive neurons performs an unsupervised learning. During the learning process, the neuron in the output layer, whose weight vector matches the input pattern most closely (typically based on minimum Euclidean distance), is chosen as the winner. Afterwards, the weights are updated to improve the response.

The Kohonen's self-organizing neural network has the special property of effectively creating a spatially organized internal representation of various features of input signals and their abstractions. The perception of similarity of objects is an essential feature. In a self-organizing neural network the neurons in the output layer 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 closely together in the two-dimensional map. The Kohonen's self-organizing neural network has been widely applied for classification and pattern recognition studies such as modeling polymers [38], clustering water samples [39, 40] and searching for surrogates in drug design [41].

A Kohonen's self-organizing neural network was applied to separate the dataset into a training set and a test set. The division based on a KNN map is superior to random selection. The advantage of such a procedure was already shown in previous work [42]. This method for splitting a dataset into a training and a test set assures that both sets cover the information space as good as possible. As the test set was not used during training of the MLRA or BPG model, it still can be considered as an external dataset.

Results and discussion

Splitting the dataset into a training and test set by KNN map with 2D descriptors

In this work, the Kohonen's self-organizing neural network was generated by the program SONNIA (formerly KMAP) [43].

A toroidal KNN with 52×44 neurons was utilized with the 18 descriptors used as input vectors. The initial learning spans were 26 and 22, with an initial learning rate of 0.7 and a rate factor of 0.95. The initial weights were randomly initialized, and training was performed for a period of 1600 epochs in an unsupervised manner. A map was depicted as a rectangle feature map by cutting the torus twice along any two orthogonal main circles on the torus [23]. The map was formed according to the ranges of solubility of the most frequently occupied neuron. From Figure 1, one can see that compounds with a different range of solubility are projected into different areas. There is an island of molecules with high solubility, surrounded by a sea of intermediate solubility with some depressions of low solubility. This indicates that the descriptors chosen can quite well be used for modeling solubility.

In the Kohonen map, 1148 of a total of 2288 neurons are occupied. Afterwards, one object of each neuron was taken for the training set, the other objects represented the test set. Thus, the 2084 compounds were divided into a training set of 1148 compounds and a test set of 936 compounds after the KNN classification. The empty neurons in the Kohonen map do not contain any objects, so they were not treated. But when producing the Kohonen map, we tried to adjust the size of the map and the corresponding parameters of the Kohonen network so that the minimum percentage of empty neurons was contained in the Kohonen map.

Splitting the dataset into a training and test set by KNN map with 3D descriptors

Similarly, with 32 RDF codes and additionally eight descriptors, the whole set of 2083 compounds was split into a training and a test set by Kohonen Neural Network. A toroidal KNN with 48×46 neurons was utilized with the 40 descriptors used as input vectors. The initial learning spans were 24 and 23, with an initial learning rate of 0.7 and a rate factor of 0.95. The initial weights were randomly initialized, and training was performed for a period of 1600 epochs in an unsupervised manner. A four-fold map as shown in Figure 2 was formed according to the ranges of solubility of the most frequently occupied neuron. In the Kohonen map, 1217 of a total of 2208 neurons are occupied. A certain clustering of compounds according to solubility can be seen, but this clustering is not as pronounced as with the 18 topological descriptors. Afterwards, one object of each neuron was selected for the training set, the other objects represented the test set. Thus, the 2083 compounds were divided into a training set of 1217 compounds and a test set of 866 compounds after the KNN classification.

Modeling of solubility by multilinear regression analysis

Modeling of solubility by MLRA with 2D descriptors

A multi-linear regression analysis (MLRA) was performed with the SPSS software [44] using 18 descriptors as input variables for the combined dataset of 2084 compounds. The 1148 compounds in the training set were used to build a model, and the 936 compounds were used for the prediction of solubility. The following equation was achieved and the corresponding regression coefficients are shown in Table 1.

$$\log S = \sum (c_i D_i) - 0.160 \quad (4)$$

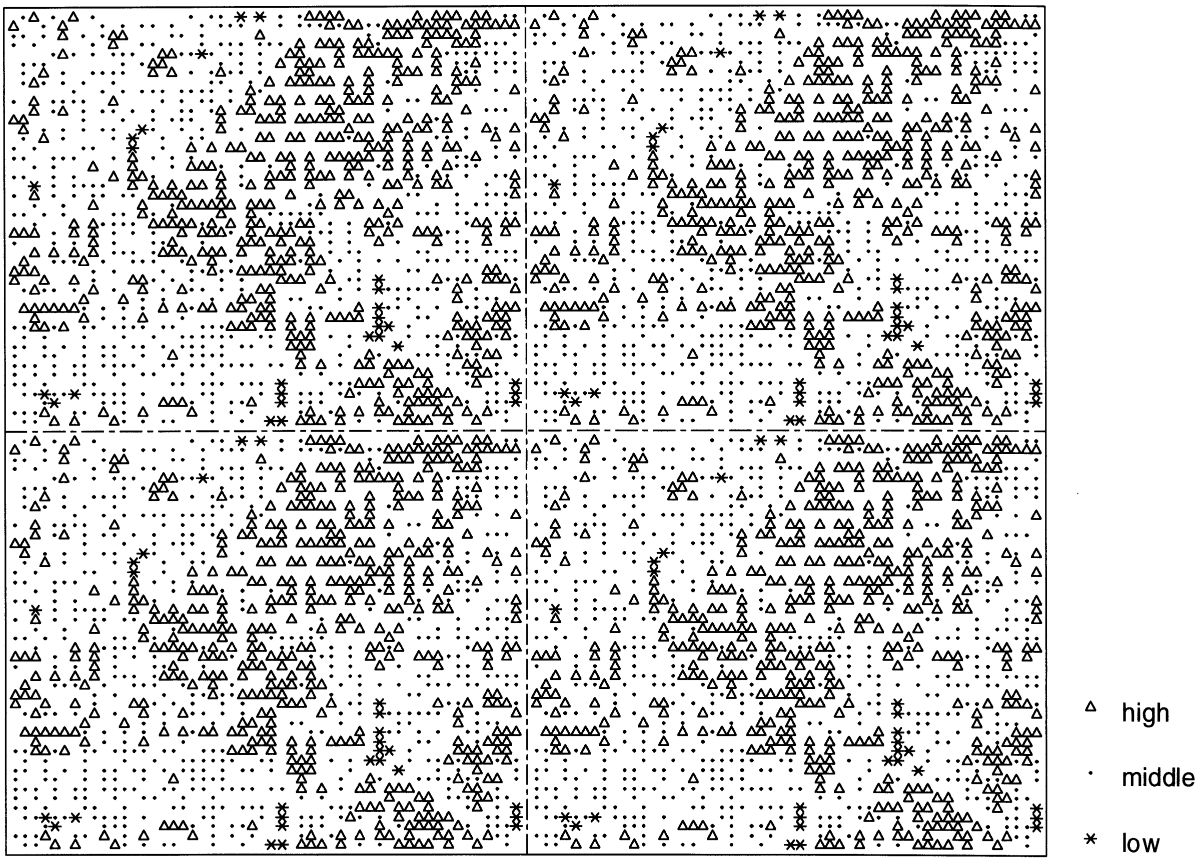

Figure 1. A four-fold toroidal KNN map for 2084 compounds by using the 18 topological input descriptors. 'High' means compounds with high solubility where logS is in the range of [−2.045 ∼ 2.346], 'middle' means compounds with intermediate solubility where logS is in the range of [−6.436 ∼ −2.046], and 'low' means compounds with low solubility where logS is in the range of [−10.830 ∼ −6.437]. Four identical KNN maps were put together like tiles in order to show the closed topology of the surface of a torus.

In the equation, D_i is a topological descriptor, and c_i is its corresponding regression coefficient in the MLRA model. For the training set, $r = 0.86$, $r^2 = 0.74$, $s = 0.94$, $MAE = 0.71$, $F = 173.64$, and $n = 1148$ and for the test set $r = 0.86$, $r^2 = 0.74$, $s = 0.73$, $MAE = 0.67$, and $n = 936$. (r is correlation coefficient, r^2 is the square of the correlation coefficient, s is standard deviation, and MAE is mean absolute error.)

After analyzing the above MLRA model, it was found that t-score values for several descriptors are very small. From the linear model standpoint, several non-significant descriptors can be omitted. So another MLRA model was built after omitting six descriptors (enlp_1, enpi_1, i_ali, #H-atoms, #S-atoms and #Cl-atoms). The following equation has been obtained and the corresponding regression coefficients are also shown in Table 1.

$$\log S = \sum (c_i D_i) - 0.086 \tag{5}$$

From this model, similar regression results were obtained. For the training set, $r = 0.86$, $r^2 = 0.73$, $s = 0.94$, $MAE = 0.71$, $F = 260.89$, and $n = 1148$ and for the test set $r = 0.86$, $r^2 = 0.74$, $s = 0.73$, $MAE = 0.67$, and $n = 936$.

Modeling of solubility by MLRA with 3D descriptors

A multi-linear regression analysis was performed with the SPSS software using 40 3D descriptors as input variables. The 1217 compounds in the training set were used to build a model, and the 866 compounds in the test set were used for the prediction of solubility. The following equation was obtained and the corresponding regression coefficients are shown in Table 2.

$$\log S = \sum (c_i T_i) - 2.88 \tag{6}$$

In the equation, T_i is a 3D descriptor, and c_i is its corresponding regression coefficient in the MLRA model.

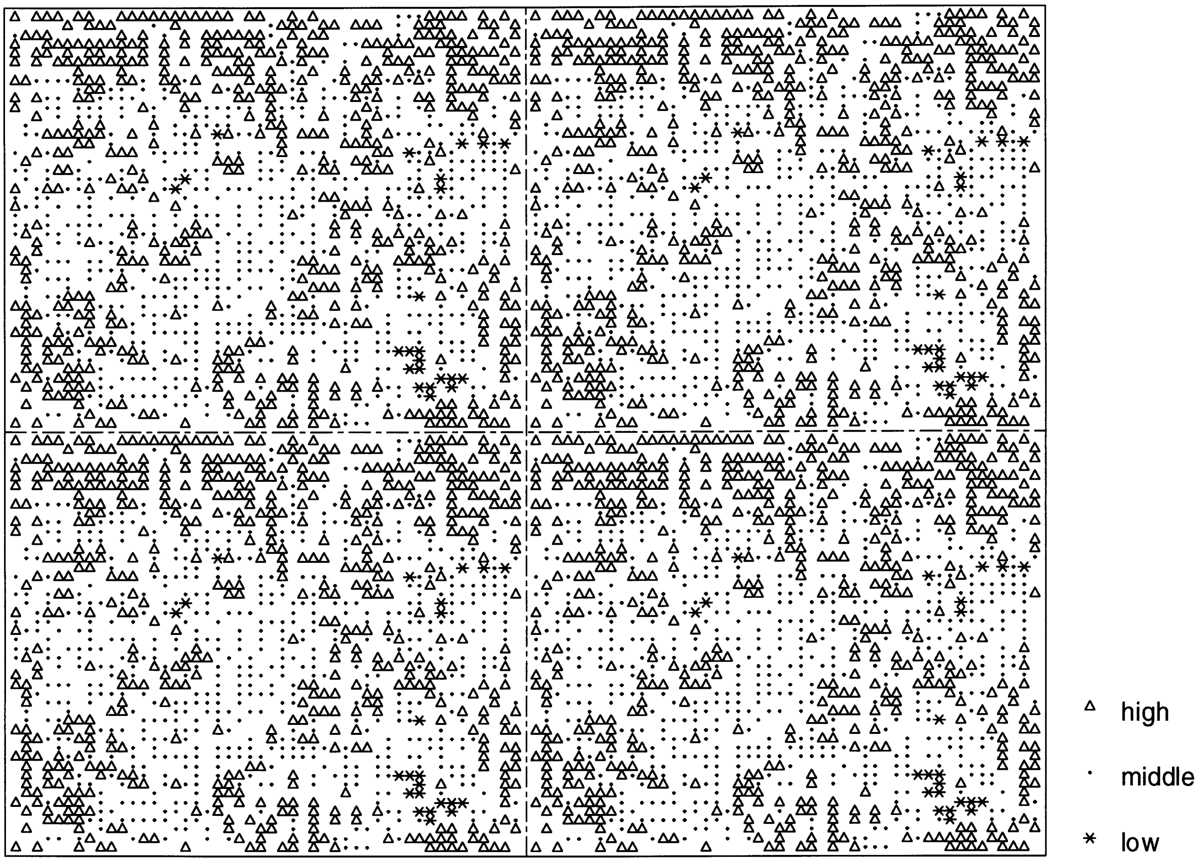

Figure 2. A four-fold toroidal KNN map for 2083 compounds by using the 40 3D descriptors. High means compounds with high solubility where logS is in the range of [−2.045 ∼ 2.346], middle means compounds with intermediate solubility where logS is in the range of [−6.436 ∼ −2.046], and low means compounds with low solubility where logS is in the range of [−10.830 ∼ −6.437]. As in Figure 1, four identical maps were put together to reflect the toroidal nature of the map.

For the training set, $r = 0.81$, $r^2 = 0.65$, $s = 1.06$, $MAE = 0.78$, F = 54.63, and $n = 1217$, and for the test set $r = 0.81$, $r^2 = 0.66$, $s = 0.80$, $MAE = 0.77$, and $n = 866$.

Once again, the MLR model was rebuilt after omitting 17 linear non-significant descriptors. The following equation was achieved and the corresponding regression coefficients are also shown in Table 2.

$$\log S = \sum (c_i T_i) - 2.80 \qquad (7)$$

From this model, similar results were obtained. For the training set, $r = 0.81$, $r^2 = 0.65$, $s = 1.06$, $MAE = 0.79$, F = 95.37, and $n = 1217$, and for the test set $r = 0.81$, $r^2 = 0.66$, $s = 0.80$, $MAE = 0.77$, and $n = 866$.

Modeling of solubility by a backpropagation neural network

Modeling of solubility by a backpropagation neural network with 2D descriptors

The SNNS program [45] was used for a Backpropagation Neural Network. A standard backpropagation network was applied to estimate solubility. An input layer with 18 input units, an output layer with one neuron representing 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. The network was trained following the ‘standard backpropagation’ algorithm as implemented in SNNS, employing a learning rate of

Table 2. Multilinear regression models with all 40 3D descriptors or 23 3D descriptors.

Descriptors	Coefficients[a]	t-score[b]	Coefficients[c]	t-score[d]
RDF_1	5.025	2.408	5.380	4.423
RDF_2	0.155	0.143		
RDF_3	3.393	4.415	3.908	8.577
RDF_4	0.887	0.939		
RDF_5	1.015	0.747	1.482	1.726
RDF_6	0.343	0.393		
RDF_7	1.201	2.045	1.161	2.786
RDF_8	−1.905	−3.433	−1.837	−3.881
RDF_9	2.069	3.533	1.767	3.692
RDF_10	−0.717	−1.158		
RDF_11	0.358	0.483		
RDF_12	0.523	0.582		
RDF_13	−2.202	−2.724	−1.773	−3.721
RDF_14	0.424	0.508		
RDF_15	−2.599	−3.472	−2.341	−4.338
RDF_16	−0.309	−0.377		
RDF_17	0.748	0.873		
RDF_18	−0.946	−0.984		
RDF_19	1.710	1.857	1.191	1.821
RDF_20	−3.062	−3.167	−2.820	−3.469
RDF_21	2.077	1.961	1.846	1.963
RDF_22	−1.569	−1.394	−1.321	−1.747
RDF_23	−0.444	−0.323		
RDF_24	1.190	0.865		
RDF_25	−0.383	−0.334		
RDF_26	−0.730	−0.660	−1.155	−1.514
RDF_27	−3.019	−2.500	−2.355	−3.029
RDF_28	0.988	0.691		
RDF_29	−0.437	−0.330		
RDF_30	−6.532	−4.256	−6.486	−5.980
RDF_31	6.475	4.131	6.490	4.618
RDF_32	−3.220	−2.542	−3.239	−2.743
MMP	−0.085	−13.075	−0.082	−14.619
i_ali	0.931	4.116	0.815	4.550
i_aro	0.173	0.808		
M_H_ACC	−0.00658	−0.238		
M_H_DON	1.459	3.087	1.518	3.367
#H_donors	0.169	4.779	0.168	5.206
#N-atoms	0.245	7.253	0.242	8.507
#O-atoms	0.357	11.533	0.348	15.437

[a] Regression coefficients of the MLR model with all the 40 3D descriptors.

[b] The t-score values of the MLR model with all the 40 3D descriptors.

[c] Regression coefficients of the MLR model with the 23 3D descriptors.

[d] The t-score values of the MLR model with the 23 3D descriptors.

0.2. Each input and output value was scaled between 0 and 1 by using the following equation:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{8}$$

in which x' is the normalized value, x is any one of the descriptor vectors, $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the descriptor vector in the dataset.

Again, 1148 compounds were used as training set and the other 936 compounds as test set. In the process, the architecture of the neural network was optimized. The number of hidden layer neurons was varied from 5 to 13. The optimized neural network architecture was 18-11-1. The best number of training epochs was selected by the early stopping method [46] in order to avoid overtraining, and it was found to be 12,000. In the neural network, training was stopped when a minimum error on the test set was reached. Training was ceased at this point and the corresponding network was taken as reference.

For the training set, $r = 0.93$, $s = 0.61$, $MAE = 0.50$, and $n = 1148$, and for the test set, $r = 0.92$, $s = 0.62$, $MAE = 0.49$, and $n = 936$.

Test the model with an additional dataset

We used the Huuskonen dataset for testing the models. The dataset comprises 1293 compounds in total. After excluding those compounds that overlap with the Merck dataset, 799 compounds remained and were used for testing. Input and output values were scaled between 0 and 1, according to the larger ranges of descriptors in the Husskonen and Merck datasets.

It was observed that the ranges of most of the 18 descriptors in the Merck dataset are larger than those of the Huuskonen dataset, except for three descriptors: logP, MMP, and the number of hydrogen bond donor groups (#H_donors). In the Huuskonen dataset, the ranges of the descriptors logP, qtot_1, and #H_donors are −6.9060∼8.9110, 0.0050∼2.5864, and 0∼17, while in the Merck dataset, they are −3.2240∼8.9110, 0.0081∼2.7615, and 0∼13. The distributions of the 18 normalized topological descriptors and the logS values of the Huuskonen and the Merck dataset are shown in the histogram of Figure 3.

With the best architecture of the backpropagation neural network as derived above, the solubility for this dataset was estimated. For this dataset, $r = 0.94$, $s = 0.72$, MAE $= 0.56$, and $n = 799$.

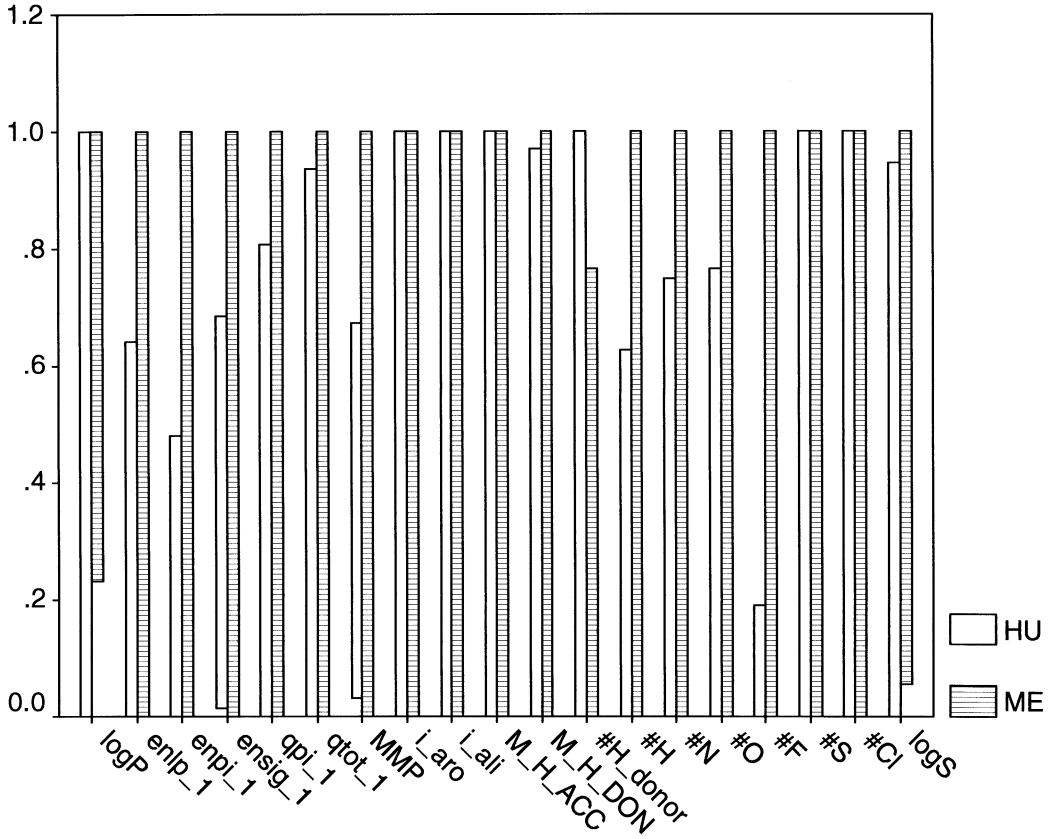

Figure 3. The distributions of the 18 normalized topological descriptors and logS of the Huuskonen dataset and the Merck dataset. 'HU' is the Huuskonen dataset, and 'ME' is the Merck dataset.

Modeling of solubility by a backpropagation neural network with 3D descriptors

Again, 1217 compounds were used as training set and the other 866 compounds as test set. In the process, the architecture of the neural network was optimized. The number of hidden layer neurons was varied from 5 to 10. The optimized neural network architecture was 40-9-1. The best number of training epochs was selected by the early stopping method in order to avoid overtraining, and it was found to be 13,500. For the training set, $r = 0.93$, $s = 0.60$, $MAE = 0.49$, and $n = 1217$, and for the test set, $r = 0.90$, $s = 0.73$, $MAE = 0.58$, $n = 866$.

Test the model with an additional dataset

As before, the 799 compounds in the Huuskonen dataset (excluding those also present in the Merck dataset) were used for testing the models. With the best architecture of the backpropagation network as derived above, the solubility for this dataset was estimated. For this dataset, $r = 0.91$, $s = 0.88$, $MAE = 0.66$, and $n = 799$.

The overall performances for our models are shown in Table 3. From Table 3, it was found that the prediction results of both BPG neural network models with 2D or 3D descriptors are good and acceptable. The prediction results of the 2D MLRA and BPG neural network models with 18 topological descriptors are slightly better than the corresponding 3D MLRA and BPG neural network models with 32 RDF code values and eight other descriptors. The neural networks provide better prediction results than MLRA.

The molecules were represented by two methods of using 18 topological descriptors or 40 3D descriptors. Their computation time for the different descriptors was compared. As to 2D descriptors, under the Linux 2.4 computer server (PIII 600MHZ), it needed about five minutes for calculating all the possible descriptors for the entire Merck dataset (2084 compounds) by using the PETRA program package.

Table 3. The overall performances of different models by using multilinear regression (MLR), and backpropagation (BPG) neural network when the compounds were described by two different methods (r is correlation coefficient, s is standard deviation, and MAE is mean absolute error).

		Training set				Test set				Additional test set[a]			
		n	r	sd	MAE	n	r	sd	MAE	n	r	sd	MAE
2D model	MLR	1148	0.86	0.94	0.71	936	0.86	0.73	0.67	799	0.91	0.73	0.68
(18 descriptors)	BPG	1148	0.93	0.61	0.50	936	0.92	0.62	0.49	799	0.94	0.72	0.56
3D model	MLR	1217	0.81	1.06	0.78	866	0.81	0.80	0.77	799	0.87	0.84	0.82
(40 descriptors)	BPG	1217	0.93	0.60	0.49	866	0.90	0.73	0.58	799	0.91	0.88	0.66

[a] Additional test set is the Huuskonen dataset excluding the overlap with the Merck dataset.

About 14 seconds were consumed for computing the topological autocorrelation coefficient of each atomic property for all the Merck dataset by the program AUTOCORR. As to 3D descriptors, the 3D coordinates of a compound can be rapidly generated by CORINA. For the 2083 compounds, their 3D coordinates can be generated by CORINA in 30 seconds, their atomic property of atomic number can be calculated by PETRA in 74 seconds, and then their RDF code values can be converted by the RCODE program in six seconds.

From the number of input descriptors, for the 2D models, only 18 descriptors were used. But the method involved a tedious descriptor selection process. For the 3D models, 32 RDF codes and eight additional descriptors were used, but it needed no descriptor selection process. For the prediction results, the 2D models give better results.

Comparing the overlap of the two datasets

Comparing the overlap of the two datasets by using 2D descriptors as input

The Kohonen's self-organizing neural network has the characteristics of mapping the objects according to their similarity. In order to investigate the similarity of the two datasets further, the 2084 compounds in the Merck dataset and the 799 compounds in the Huuskonen dataset (excluding those already contained in the Merck dataset) were put together, and were fed into a 41×35 KNN map by using 18 topological descriptors as input descriptors. All the compounds were classified into three groups: only the Merck dataset (1588 compounds) as shown by triangles, only the Huuskonen dataset (799 compounds) represented by squares, and the overlap (the compounds existing in both datasets) as indicated by asterisks.

The map was formed according to different datasets of the most frequently occupied neuron. Their distribution is shown in Figure 4. It was observed that, apart from the removed compounds, which were identical for the two datasets, 310 compounds of the reduced Huuskonen dataset occupied the same neurons as those of 376 compounds in the Merck dataset.

Figure 4 shows that large areas of the map are occupied by Merck compounds (indicated by triangles) only. Thus, it can be concluded that the Merck dataset is more diverse than the Huuskonen dataset.

Comparing the overlap of the two datasets by using 3D descriptors as input

Similarly, the overlap of the two datasets according to their compounds' distribution in a KNN map was examined when the compounds were described by 40 descriptors (32 RDF code values and eight additional descriptors). The 2083 compounds in the Merck dataset and the 799 compounds (excluding the same compounds as in the Merck dataset) in the Huuskonen dataset were put together, and were projected into a 39×37 KNN map as shown in Figure 5. All the compounds were classified into three groups: only the Merck dataset (1587 compounds), indicated by triangles, only the Huuskonen dataset (799 compounds), represented by squares, and the overlap (the compounds existing in both datasets) as shown by asterisks.

The map was formed according to different datasets of the most frequently occupied neuron. It was found that 359 compounds of the reduced Huuskonen dataset cover the same neurons as 407 compounds in the Merck dataset. And one can see that the Merck dataset covers a large area and contains more information than the Huuskonen dataset.

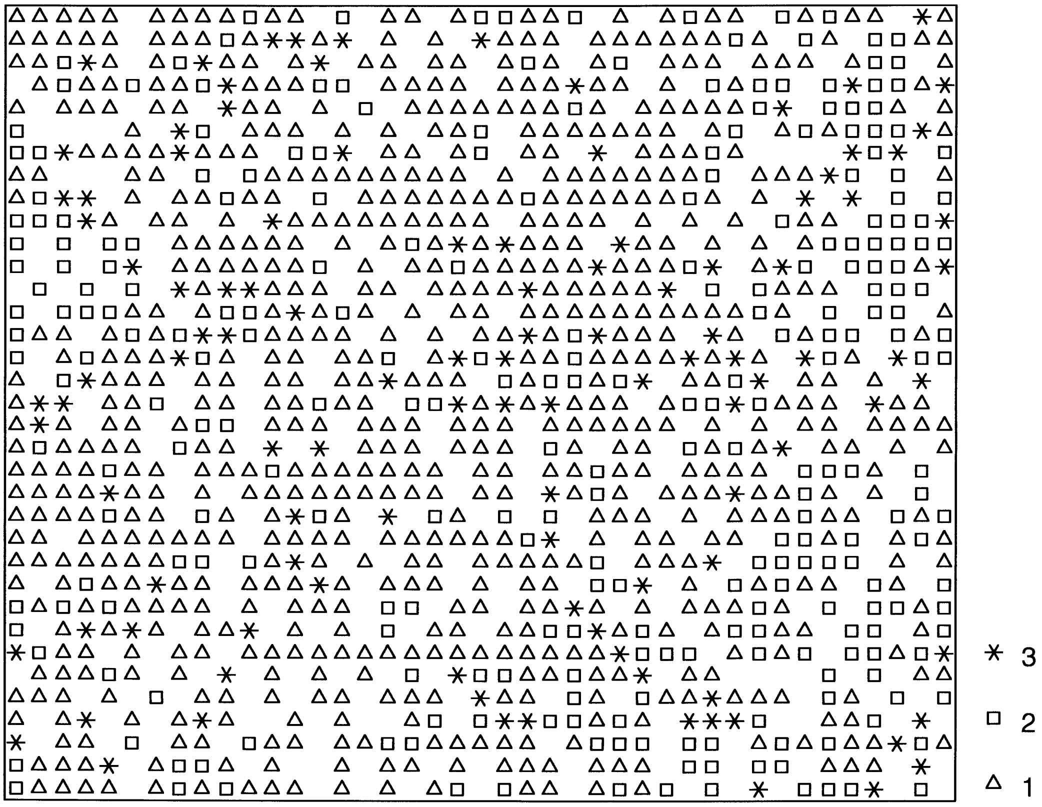

Figure 4. The distribution of the compounds of the two datasets in the same KNN (Kohonen's self-organizing neural network) map by using 18 topological descriptors as input descriptors. Triangles represent the most frequently appearing 1588 compounds in the Merck dataset (excluding the same compounds with the Huuskonen dataset); squares represent the most frequently appearing 799 compounds in the Huuskonen dataset (excluding the same compounds with the Merck dataset), asterisks represent the most frequently appearing overlap of the Huuskonen and the Merck dataset.

Conclusions

Kohonen's self-organizing neural network is a valuable tool for monitoring the selection of structure descriptors. In the training process, an unsupervised method is adopted; the projection position for a certain compound in a KNN map was determined by its structure characteristics and the relationship (similarity and dissimilarity) with other compounds. If a set of descriptors is good for the prediction of a certain property (such as solubility), the compounds with a similar property will be closer in the corresponding KNN map with this set of descriptors.

Kohonen's self-organizing neural network is also an excellent tool for splitting a large dataset into a training and a test set. The training set is used for building a model and the test set for testing the model. This kind of training and test set selection method can ensure that both the training set and test sets contain as much information as possible.

Kohonen's self-organizing neural network can also be used for the comparison of the similarity and dissimilarity of different datasets. That is to say, it can be used for the comparison of the distribution of the objects in different datasets in a KNN map. Such an approach is also useful for comparing different compound libraries in order to assign their similarity or diversity.

The structure of organic compounds can be described by 2D or 3D descriptors. 2D descriptors are simple and easy to understand while the 3D descriptors contain more information. The descriptors of both methods are derived from the molecular structures, which are suitable to *in silico* data screening and library design.

Models for the prediction of aqueous solubility of organic compounds were built by using multi-linear regression analysis and backpropagation neural networks based on 2D and 3D descriptors. Both mod-

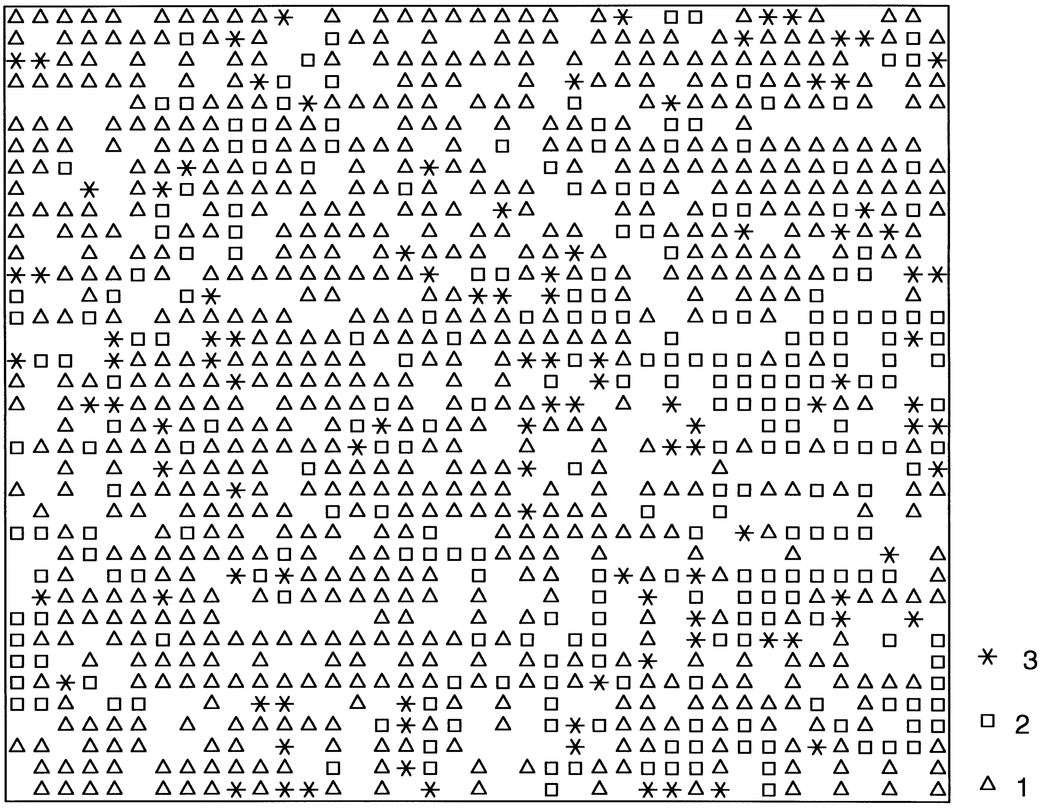

Figure 5. The distribution of the compounds of the two datasets in the same KNN (Kohonen's self-organizing neural network) map by using 40 3D descriptors as input descriptors. Triangles represent the most frequently appearing 1587 compounds in the Merck dataset (excluding the same compounds with the Huuskonen dataset); squares represent the most frequently appearing 799 compounds in the Huuskonen dataset (excluding the same compounds with the Merck dataset), asterisks represent the most frequently appearing overlap of the Huuskonen and the Merck dataset.

els provide good prediction results. The models developed for the prediction of solubility can be applied to large datasets with rapid calculation speed such that a wide range of compounds can be processed. The non-linear neural network models provide better predictions than a multilinear regression analysis.

The models built with the larger, more diverse Merck dataset also have a fairly good prediction capability for those compounds of the Huuskonen dataset not contained in the Merck dataset. This attests to the more global validity of the solubility models developed here compared to our previous studies with the entire Huuskonen dataset [12, 13].

A message to take home from this study is to first have a look at the descriptor space by an unsupervised learning method such as KNN. Then a linear model should be built by MLRA, followed by an investigation of whether a non-linear model with a backpropagation neural network provides better results.

Comparison of our results with the quality of prediction obtained by other authors suggests that an error of prediction of solubility of 0.50 to 0.60 logS units is the best that one can hope for from such datasets that have been collected from a variety of sources of different experimental quality. It is hoped that in the future datasets on aqueous solubility of organic compounds become available that have been measured under strictly controlled experimental conditions such as temperature and pH value. Furthermore, it should be ensured that indeed thermodynamic conditions are maintained.

Acknowledgements

Dr. Aixia Yan appreciates a Research Fellowship from the Alexander von Humboldt Foundation. The research obtained financial support from the Bundesministerium für Bildung und Forschung. We thank Dr. J. Maerz (Merck KGaA) for making the Merck dataset available to us. We also thank Dr. J. Huuskonen and Dr. I.V. Tetko for providing us with the other dataset.

References

1. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., Adv. Drug Deliv. Rev., 23 (1997) 3.
2. Jorgensen, W.L. and Duffy, E.M., Adv. Drug Deliv. Rev., 54 (2002) 355.
3. Gao, H., Shanmugasundaram, V. and Lee, P., Pharmaceut. Res., 19 (2002) 497.
4. Bodor, N. and Huang, M.J., J. Am. Chem. Soc., 113 (1991) 9480.
5. Sutter, J.M. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 36 (1996) 100.
6. Mitchell, B.E. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 38 (1998) 489.
7. McElroy, N.R. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 41 (2001) 1237.
8. Bruneau, P., J. Chem. Inf. Comput. Sci., 41 (2001) 1605.
9. Huuskonen, J., J. Chem. Inf. Comput. Sci., 40 (2000) 773.
10. Tetko, I.V., Tanchuk, V.Y., Kasheva, T.N. and Villa, A.E.P., J. Chem. Inf. Comput. Sci., 41 (2001) 1488.
11. Liu, R.F. and So, S.S., J. Chem. Inf. Comput. Sci., 41 (2001) 1633.
12. Yan, A.X. and Gasteiger, J., QSAR Comb. Sci., 22 (2003) 821.
13. Yan, A.X. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 43 (2003) 429.
14. Engkvist, O. and Wrede, P., J. Chem. Inf. Comput. Sci., 42 (2002) 1247.
15. Wegner, J.K. and Zell, A., J. Chem. Inf. Comput. Sci., 43 (2003) 1077.
16. Peterson, D.L. and Yalkowsky, S.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1531.
17. Ran, Y.Q., Jain, N. and Yalkowsky, S.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1208.
18. Yang, G., Ran, Y.Q. and Yalkowsky, S.H., J. Pharm. Sci., 91 (2002) 517.
19. Kuhne, R., Ebert, R.-U., Kleint, F., Schmidt, G. and Schuurmann, G., Chemosphere, 30 (1995) 2061.
20. Klopman, G. and Zhu, H., J. Chem. Inf. Comput. Sci., 41 (2001) 439.
21. Hemmer, M.C., Steinhauer, V. and Gasteiger, J., Vibrat. Spectrosc., 19 (1999) 151.
22. Hemmer, M.C. and Gasteiger, J., Anal. Chim. Acta, 420 (2000) 145.
23. Zupan, J. and Gasteiger, J., Neural Networks in Chemistry and Drug Design, Second edn. Wiley-VCH, Weinheim, Germany, 1999.
24. Yalkowsky, S.H. and Dannefelser, R.M., The ARIZONA dATAbASE of Aqueous Solubility. College of Pharmacy, University of Arizona, Tucson, AZ, 1990.
25. Syracuse Research Corporation. Physical/Chemical Property Database (PHYSPROP), SRC Environmental Science Center, Syracuse, NY, 1994.
26. Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.
27. Gasteiger, J. and Saller, H., Angew. Chem. Int. Ed. Engl., 24 (1985) 687.
28. Gasteiger J., Empirical methods for the calculation of physicochemical data of organic compounds. In: Jochum, C., Hicks, M.G. and Sunkel, J. (Eds.), Physical Property Prediction in Organic Compounds. Springer Verlag, Heidelberg, Germany, 1988, pp. 119–138.
29. PETRA can also be accessed on the web: http://www2.chemie.uni-erlangen.de/software/petra/index.html, see also http://www.mol-net.de
30. Ghose, A.K. and Crippen, G.M., J. Comput. Chem., 7 (1986) 565.
31. Ghose, A.K. and Crippen, G.M., J. Chem. Inf. Comput. Sci., 27 (1987) 21.
32. Ghose, A.K., Pritchett, A. and Crippen, G.M., J. Comput. Chem., 9 (1988) 80.
33. Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K., J. Chem. Inf. Comput. Sci., 29 (1989) 163.
34. Wagener, M., Sadowski, J. and Gasteiger, J., J. Am. Chem. Soc., 117 (1995) 7769.
35. Gasteiger, J. and Hutchings, M.G., J. Chem. Soc. Perkin 2, (1984) 559.
36. Miller K.J., J. Am. Chem. Soc., 112 (1990) 8533.
37. Sadowski, J. and Gasteiger J., Chem. Rev., 93 (1993) 2567. http://www2.chemie.uni-erlangen.de/software/corina/index.html
38. Harrison, R.W., J. Math. Chem., 26 (1999) 125.
39. Aguilera, P.A., Frenich, A.G., Torres, J.A., Castro, H., Vidal, J.L.M. and Canton M., Water Res., 35 (2001) 4053.
40. Brodnjak-Voncina, D., Dobcnik, D., Novic, M. and Zupan, J., Anal. Chim. Acta, 462 (2002) 87.
41. Anzali, S., Mederski, W.W.K.R., Osswald, M. and Dorsch, D., Bioorg. Med. Chem. Lett., 8 (1998) 11.
42. Simon, V., Gasteiger, J. and Zupan, J., J. Am. Chem. Soc., 115 (1993) 9148.
43. Terfloth, L. and Gasteiger, J., Screening-Trends Drug Discov., 2 (2001) 49. http://www2.chemie.uni-erlangen.de/software/kmap/ and http://www.mol-net.de
44. SPSS v. 10.0, SPSS Inc., Chicago, IL. http://www.spss.com
45. 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/
46. Tetko, I.V., Livingstone, D.J. and Luik, A.I., J. Chem. Inf. Comput. Sci., 35 (1995) 826.