Quantitative Structure−Property Relationship (QSPR) Prediction of Solvation Gibbs Energy of Bifunctional Compounds by Recursive Neural Networks†

Luca Bernazzani,‡ Celia Duce,‡ Alessio Micheli,§ Vincenzo Mollica,‡ and Maria Rosaria Tiné*,‡

Dipartimento di Chimica e Chimica Industriale, Università di Pisa, Via Risorgimento 35, 56126, Pisa, Italy, and Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, 56127, Pisa, Italy

In this paper we apply a recursive neural network (RNN) model to the prediction of the standard Gibbs energy of solvation in water of mono- and polyfunctional organic compounds. The proposed model is able to directly take as input a structured representation of the molecule and to model a direct and adaptive relationship between the molecular structure and the target property. A data set of 339 mono- and polyfunctional acyclic compounds including alkanes, alkenes, alkynes, alcohols, ethers, thiols, thioethers, aldehydes, ketones, carboxylic acids, esters, amines, amides, haloalkanes, nitriles, and nitroalkanes was considered. As a result of the statistical analysis, we obtained for the predictive capability estimated on a test set of molecules a mean absolute residual of about 1 kJ·mol^{-1} and a standard deviation of 1.8 kJ·mol^{-1}. This results is quite satisfactory by considering the intrinsic difficulty of predicting solvation properties in water of compounds containing more than one functional group.

Introduction

During the past decade a quantitative structure−activity (or quantitative structure−property) QSAR/QSPR method[1,2] on the basis of the direct treatment of the molecular structure by a recursive neural network (RNN) has been developed by our group. Standard QSAR/QSPR approaches correlate, by a proper mapping function, a target property to the structure of the molecule encoded by problem specific molecular descriptors. The innovative feature of RNN is its ability to directly deal with hierarchically structured representations of molecules in the form of labeled graphs which are a vehicle of richer information than the flat vectors of descriptors used in the traditional QPSR/QSAR models. The RNN itself maps the representation of the molecule to the target property. Of course, the possibility of processing structured information is particularly appealing in the context of prediction tasks in chemistry where the compounds can naturally be represented as labeled graphs.

In this context, we proposed a rational approach to the representation of chemical structures by using a limited number of fundamental atomic groups ordered as the corresponding two-dimensional graph.[3] The problems tackled by our earlier RNN-QSAR/QSPR studies range from the prediction of simple physical properties, such as the boiling points of alkanes[1,5] and the standard Gibbs energy of solvation of monofunctional compounds,[3] to the prediction of the pharmacological activity of substituted benzodiazepines.[1,2,5] More recently, we applied this technique to the prediction of the glass transition temperature (T_g) of (meth)acrylic polymers and copolymers,[4,6,7] the melting point (T_{fus}) of ionic liquids,[7,8] and the acute toxicity of simple aromatic molecules.[9]

In this paper we focus our attention on the prediction of the standard Gibbs energy of solvation in water, $\Delta_{solv}G°$, of polyfunctional compounds. The prediction of Gibbs energies of solvation for small organic molecules is of considerable interest in drug design,[10] in the analysis of protein folding and binding,[11] and in the development of force fields by computer simulation.[12,13] In a previous paper[3] we applied our RNN-QSPR model to the analysis of the $\Delta_{solv}G°$ in water of 179 monofunctional acyclic organic compounds. Our model showed a better descriptive and predictive ability than group contribution methods or usual QSPR approaches using multilinear regression analysis and matched the performances obtained by standard neural network based QSPR methods, which are tuned by the background knowledge carried by known molecular descriptors.

More extensive experiments would refine the comparison with group contribution methods. Therefore we decided to extend the study to sets of compounds spanning over a widespread survey of chemical structures and functionalities and including polyfunctional compounds where it is known that the theoretical limitations of the group contributions method are stronger and the method requires the introduction of corrective terms. In fact, the ultimate goal is the extension of the applicability of the prediction methods to compounds and problems not covered by the standard methods. A new research was carried out in existing literature, collecting all available material on $\Delta_{solv}G°$, obtained by using either different experimental methods or data derived from different thermodynamic quantities related to $\Delta_{solv}G°$ (i.e., Henry's law constants).

Method

In this section we give an outline of the main aspect of our RNN model for QSAR/QSPR analysis. A detailed description of the model is reported elsewhere.[1−9]

RNNs directly deal with a hierarchically structured representation of molecules in the form of labeled rooted ordered trees. Moreover, RNNs can adaptively encode the input structures by learning from the given structure−property training examples. The learning algorithm allows the model to tune the free parameters of the encoding process. The use of an adaptive

† Part of the "Workshop in Memory of Henry V. Kehiaian".

* Corresponding author. E-mail: mrt@dcci.unipi.it; phone: +390502219268; fax: +390502219260.
‡ Dipartimento di Chimica e Chimica Industriale.
§ Dipartimento di Informatica.

10.1021/je100535p
Published on Web 07/21/2010

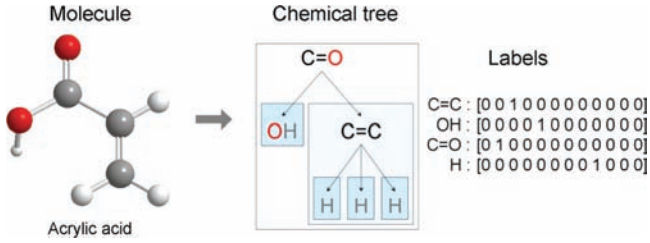

Figure 1. Representation of molecular graph for acrylic acid, as a chemical tree and an input data file. Boxes on the chemical tree exemplify the recursive processing of RNN.

recursive model does not need an a priori definition of descriptors and does not require a similar size of the structured data.

Chemical compounds are represented as labeled rooted ordered trees by a 2D graph that can be easily obtained from their structural formula. Molecules are fragmented into atomic groups that constitute the vertices of the tree (see Figure 1). The rules for the representation of the molecular structure are described in detail in ref 3 and shortly sketched below. A priority scale is defined among the fragments to determine the root and the relative ordering of the subtrees. The tree root is always placed on the highest priority group, and its children were ordered according to the group priority rules. A label is assigned to each vertex, that is, a tuple of variables categorically distinguishing the symbol of the atomic group. Though labels are conventionally defined, they can convey chemical information through orthogonality or similarity to other labels. The tree representation, though based on a 2D graph, can also describe 3D features by defining appropriate rules, for example, the children ordering can be used to indicate chirality in analogy with Fisher's projections. The direction of the ordering (i.e., clockwise or counter-clockwise) is the one that assigns the lowest position to the group with highest priority. In the case of two or more groups with identical priority, the direction is determined by the subsequent group in the priority scale until the ambiguity is solved.

For each structure, the neural model encodes, through a recursive process, the substructures according to their molecular topology and to the content of each vertex label (see Figure 1). Finally, the code developed by the model is mapped to the property values by the output part of the neural network. It is worth it to stress that the training of the output mapping part is made together with the training of the encoding part to realize a direct and adaptive structure−property relationship. An incremental process progressively adds neural hidden units, HU, until the error among the outputs of the training examples and their target values is below a previously determined threshold or training error tolerance, TET.

The chemical groups used in this work are: CH_3, CH_2, C, H, C═C, C≡C, OH, O, C═O, NH_2, NH, N, SH, S, CN, NO_2, F, Cl, Br, and I. They coincide, even though not completely, with the functional groups identifying the various classes of organic compounds. In particular the groups COOH, COO−, $CONH_2$, CO−N<, and COX (X = Cl, Br, I) are represented as subtrees constituted by two atomic groups. Moreover, the C−H group is formed by C and H. In this way we maintain the same approach in describing the C−H bond independently of the hybridization of the carbon atom.

As exemplified above, this kind of structure-based representation has the advantage of generality, as it can adequately represent any sort of chemical compound.[4] Moreover, its flexibility allows the tuning of the level of structural detail to the characteristics of the investigated molecular data set.

Results and Discussion

In this work we deal with a data set of 339 mono- and poly functional acyclic compounds including alkanes, alkenes, alkynes, alcohols, ethers, thiols, thioethers, aldehydes, ketones, carboxylic acids, esters, amines, amides, haloalkanes, nitriles, and nitroalkanes. The experimental data were taken from refs 14 to 19. Seven tasks were performed in this study, and 16 trials were carried out in each task. As the data set includes data of $\Delta_{solv}G°$, derived from different thermodynamic quantities, it is affected by high noise. As a consequence, the system's TET was set accordingly to avoid overtraining phenomena, namely, to 1 kJ·mol^{-1} in tasks 1 to 4 and 6 and to 5 kJ·mol^{-1} in tasks 5 and 7 to assess the tolerance and fitting conditions. The target $\Delta_{solv}G°$ values in the training set ranged from (−40.63 to 14.94) kJ·mol^{-1}.

For each task the whole data set was divided into three disjoint training, test, and guess test sets. The training set (236 compounds) and test set (60 compounds) were used for the learning and validation processes, respectively, while the guess test set (12 compounds) was settled on to test the performance of the RNN model in some challenging conditions. The compounds were selected so that the test set was representative of the different molecular sizes, topologies, and functional groups. The complete list of investigated compounds, the corresponding values of the target property, $\Delta_{solv}G°$, and the mean residual for each performed task are reported in the Supporting Information. To have a significant outcome, in each task 16 trials were carried out on the same training/test split, and the results were averaged over the different trials. The main statistics computed over all of the tasks are shown in Table 1, which indicates the number of recursive HU, the mean absolute residual, MAR, the correlation coefficients, *R*, and the standard deviation, *S*, as obtained by computing the mean output over the performed trials.

The experimental Gibbs energies of solvation of the molecules included in the guess test set are reported in Table 2 together with the mean residuals, δ_i, evaluated as the difference between the mean predicted values over the 16 trials and the experimental one. In fact, for this guess test set the results have to be evaluated individually and not statistically. The guess test set was formed by 12 compounds selected within a list of 13 (1-buten-3-yne, 2-propen-1-ol, *trans*,*trans*-2,4-hexadienal, 3-buten-2-one, 1-methylethenyl ethanoate, propenoic acid, trimethoxymetane, 1,1,1-trimethoxyethane, di(2-chloroethyl)sulphide, trichloronitromethane, *N*,*N*-dimethylformamide, 1,2,3-propanetriol, 1,2-ethanediol), with only one polyalcohol being chosen at a time in the experiments. The structures of these compounds were scarcely represented in the whole data set. In fact they contain two or more atomic groups whose combination is not represented in any molecule of the training set. In the following we will analyze the statistical results of every task as well as the specific residuals for the guess test set compounds.

Only two polyalcohols (1,2-ethanediol and 1,2,3-propanetriol) are present in the whole data set, beside a significant number of monoalcohols. In tasks 1 and 3 we put 1,2,3-propanetriol in the training set and 1,2-ethanediol in the guess test set, while in tasks 2 and 4 to 7 we moved 1,2-ethanediol to the training and 1,2,3-propanetriol to the guess test set. As the triol solvation properties in water are much more strongly affected by the interactions among the OH groups than those of alcohols and diols, we planned task 1 to test the influence of the triol on the

Table 1. Mean and Maximum Absolute Residual (MAR, MaxAR), Standard Deviation of Residuals (*S*), Correlation Coefficient (*R*), Number of RNN HU, and Number of the Molecules (*N*) in the Training and Test Set of the Different Tasks

			training set				test set			
			MAR	MaxAR	*S*	R^b	MAR	MaxAR	*S*	R^b
task	HU[a]	*N*	kJ·mol^{-1}				kJ·mol^{-1}			
1	57	236 tr 60 ts	0.09	0.55	0.14	0.999	1.17	10.2	2.03	0.982
2	58	236 tr 60 ts	0.09	0.64	0.14	0.999	1.20	8.51	1.86	0.985
3	59	236 tr 60 ts	0.09	0.65	0.14	0.999	0.99	8.38	1.68	0.987
4	65	236 tr 60 ts	0.09	0.54	0.14	0.999	1.00	6.46	1.62	0.988
5	13	236 tr 60 ts	0.61	3.52	0.84	0.997	1.08	9.39	1.82	0.985
6	61	267 tr 60 ts	0.09	0.63	0.14	0.999	0.97	8.78	1.69	0.987
7	14	267 tr 60 ts	0.58	3.38	0.79	0.997	1.12	7.28	1.86	0.985

[a] Number of HUs calculated as the average of the number of HUs over the trials. [b] Linear correlation coefficient between experimental and calculated values.

Table 2. Experimental $\Delta_{solv}G^\circ$ Values and Mean Residuals[a], δ_i, of the Guess Set Compounds for Tasks 1 to 7

	$\Delta_{solv}G^\circ$	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
compound	kJ·mol^{-1}							
1-buten-3-yne	0.17	−0.04	0.44	−0.64	0.84	−0.76	−0.73	−0.64
2-propen-1-ol	−21.06	−1.85	−1.88	−1.71	−1.96	−3.69	−1.19	−2.31
trans,*trans*-2,4-hexa-dienal	−19.39	4.79	4.52	2.38	3.24	3.48	3.65	2.01
3-buten-2-one	−20.77	2.93	4.84	5.79	6.46	4.71	5.40	5.14
1-methylethenyl ethanoate	−11.69	−3.27	−4.94	−1.12	−3.93	−3.29	11.72	−2.72
propenoic acid	−25.96	0.25	1.14	−5.32	−0.76	−0.98	−0.75	−0.93
trimethoxymethane	−18.47	14.78	10.33	−1.67	9.99	11.60	12.12	10.23
1,1,1-trimethoxyethane	−18.30	14.60	10.66	12.52	10.42	12.35	−1.04	11.59
di(2-chloroethyl) sulphide	−16.40	3.29	5.63	13.22	3.91	2.69	0.40	−1.17
trichloronitromethane	−6.17	−4.97	−6.02	4.10	−5.22	−2.91	−5.31	−4.44
N,*N*-dimethyl- formamide	−32.70	8.57	9.23	1.00	1.02	0.93	1.26	1.93
1,2,3-propanetriol	−38.60	nd	11.76	nd	12.76	12.92	11.84	12.24
1,2-ethanediol	−32.03	−2.48	nd	−2.67	nd	nd	nd	nd

[a] Calculated as the difference between predicted and experimental values; nd = not determined because the compound is not included in the guess set of the corresponding task.

RNN performance, whereas task 2 was performed to assess the RNN capability to extrapolate the solvation properties of this compound from those of monoalcohols and diols. If we compare the test set statistics of tasks 1 and 2 reported in Table 1, we can observe that the substitution of 1,2,3-propanetriol in the training set with 1,2-ethanediol improves the regression parameters of the test set even though slightly increasing the MAR (less than 0.03 kJ·mol^{-1}).

The same training and test sets of tasks 1 and 2 were used in tasks 3 and 4, respectively, to investigate the influence of the N label modification on the RNN performances. To improve the RNN prediction of the $\Delta_{solv}G^\circ$ of *N*,*N*-dimethylformamide, in tasks 3 and 4 we modified the label of the N group by considering it similar, instead that orthonormal, to the NH_2 and NH groups. This modification resulted in a sharp decrease of the amide residuals from about (10 to 1) kJ·mol^{-1} (see Table 2). Furthermore, their standard deviation calculated over the 16 trials decreased more than 1 kJ·mol^{-1}, remaining almost constant from task 3 to 7 (see Supporting Information). If we compare the statistics of test set results in tasks 1 and 3 and in tasks 2 and 4, we can see that the modification of the N label improves the MAR of 0.2 kJ·mol^{-1} by also improving the regression parameters. Moreover, the stability of the mean predicted output of the *N*,*N*-dimethylformamide, $\Delta_{solv}G^\circ$, and of its standard deviation throughout very different learning levels shows that in this case the RNN learning process was successfully precise and accurate. This is a very important result by considering that the amide group is of great biological significance, being characteristic of the peptide bond in proteins. Finally, a comparison of test set results in tasks 3 and 4 indicates that the similarity imposed to N, NH, and NH_2 labels reduces the influence of the triol on the RNN prediction performance.

On the basis of the results obtained in tasks 1 to 4 we decided to use the same data splitting of tasks 2 and 4 and the new N label in the last three tasks.

Tasks 6 and 7 were planned to investigate the robustness of the model with respect to the noise of the data. In these tasks we put into the training set 31 new molecules with a greatest uncertainty in the target values. Surprisingly, we obtained a better statistics in the test set, while that of the training set remained almost unchanged. Apparently, despite the increased uncertainty on the target values, the benefits introduced by the greater the number of data in the training set prevail for the improvement of the RNN performances. The robustness of the model to the noise and the uncertainty of the experimental data is thus confirmed.

In tasks 5 and 7 we used the same data split as in tasks 4 and 6, but the RNN learning process was stopped when a higher error threshold (5 kJ·mol^{-1}) was attained and a corresponding lower number of HUs was involved in the calculation. The test set MAR of these tasks are only slightly worse than those of the corresponding task where TET was 1 kJ·mol^{-1}. Hence, an early stopping of training does not improve the general performance of the model, showing that overtraining conditions have been avoided.

The analysis of the residuals of the guess test set compounds, reported in Table 2, shows that the residuals of the first six compounds are quite satisfactory, proving that the RNN is able to give good prediction on the $\Delta_{solv}G^\circ$ of molecules containing conjugate double bonds even in the absence of similar compounds in the training set (Table 2). Also in these cases, the MAR and the standard deviations are not affected by different learning conditions.

We obtain good results also for 1,2-ethanediol, while 1,2,3-propanetriol shows residuals of about 12 $kJ \cdot mol^{-1}$. We ascribe, in this case, the disagreement between experimental and calculated value to the need of a better RNN training because of the scarce sampling of polyalcohols in the training set. On the other hand, the stability of the mean predicted output and standard deviation of 1,2,3-propanetriol $\Delta_{solv}G^\circ$ throughout very different learning conditions (tasks 2 and 4 to 7) is an indication of the robustness of our RNN model.

Insufficiency of sampling also affect the $\Delta_{solv}G^\circ$ prediction for orthoesthers. Moreover, in this case, the large errors (Table 2) could reflect the uncertainty of $\Delta_{solv}G^\circ$ experimental data. The measurements can actually be affected by the hydrolysis of the orthoesther in water.

Conclusions

There is an intrinsic difficulty in the prediction of solvation properties of polyfunctional compounds in water. In fact, water enhances the existing intramolecular correlation among functional groups pertaining to the same structure, probably by setting up structured extramolecular domains involving a variable number of water molecules. As a consequence, standard group contribution methods that usually reproduce well the solvation properties of monofunctional compounds fail when treating polyfunctional molecules, unless specific parameters are introduced to account for interactions among two or more functional groups in the same structure. This, of course, reduces their generality and their predictive ability. The advantage of our RNN is that it is able to automatically take into account the interactions among groups provided that a sufficient number of training examples are present in the data set. Moreover, we proved that the performance of our method can be greatly improved by extending the data set and that it is slightly affected by the noise of experimental data. This offers some opportunity in the prediction task, by allowing the exploitation, even if with great care, of also less accurate experimental data.

Supporting Information Available:

The complete list of investigated compounds, the corresponding values of the target property, $\Delta_{solv}G^\circ$, and the mean residual for each performed task are reported. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Bianucci, A. M.; Micheli, A.; Sperduti, A.; Starita, A. Application of Cascade Correlation Networks for Structures to Chemistry. *Appl. Int. J.* **2000**, *12*, 117–146.

(2) Micheli, A.; Sperduti, A.; Starita, A.; Bianucci, A. M. Analysis of the Internal Representations Developed by Neural Networks for Structures Applied to Quantitative Structure-Activity Relationship Studies of Benzodiazepines. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 202–218.

(3) Bernazzani, L.; Duce, C.; Micheli, A.; Mollica, V.; Sperduti, A.; Starita, A.; Tiné, M. R. Predicting Physical-Chemical Properties of Compounds from Molecular Structures by Recursive Neural Networks. *J. Chem. Inf. Model.* **2006**, *46*, 2030–2042.

(4) Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tiné, M. R. Prediction of the glass transition temperature of (meth)acrylic polymers containing phenyl groups by recursive neural network. *Polymer* **2007**, *48*, 7121–7129.

(5) Micheli, A.; Sperduti, A.; Starita, A.; Bianucci, A. M. A Novel Approach to QSPR/QSAR Based on Neural Networks for Structures. In *Soft Computing Approaches in Chemistry*; Sztandera, L. M., Cartwright, H. M., Eds.; Springer-Verlag: Heidelberg, Germany, 2003; pp 265−296.

(6) Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tiné, M. R. Recursive neural networks prediction of glass transition temperature from monomer structure: an application to acrylic and methacrylic polymers. *J. Math. Chem.* **2009**, *46*, 729–755.

(7) Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tiné, M. R. Evaluation of hierarchical structured representations for QSPR studies of small molecules and polymers by recursive neural networks. *J. Mol. Graphics Modell.* **2009**, *27*, 797–802.

(8) Bini, R.; Chiappe, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tiné, M. R. Ionic liquids: prediction of their melting points by a recursive neural network model. *Green Chem.* **2008**, *10*, 306–309.

(9) Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Tiné, M. R. Toxicity prediction by a chemoinformatic approach. *Ind. Chim. (Rome)* **2009**, *27*, 136–143.

(10) Kollman, P. A. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc. Chem. Res.* **1996**, *29*, 461–469.

(11) Eisenberg, D.; McLachlan, A. D. Solvation Energy in Protein Folding and Binding. *Nature (London)* **1986**, *319*, 199–203.

(12) Duffy, E. M.; Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.

(13) Jorgensen, W. L.; Tirado-Rives, J. Free Energies of Hydration for Organic Molecules from Monte Carlo Simulations. *Perspect. Drug Discovery Des.* **1995**, *3*, 123–138.

(14) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. Group Contribution to the Thermodynamic properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *J. Solution Chem.* **1981**, *10*, 563–595.

(15) Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. The Factors that Influence The Solubility of Gases and Vapours in Water at 298 K, and a New Method for Its Determination. *J. Chem. Soc., Perkin. Trans.* **1994**, *2*, 1777–1791.

(16) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A Fuzzy ARTMAP-Based Quantitative Structure-Property Relationship (QSPR) for the Henry's Law Constant of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 85–112.

(17) Guthrie, J. P. A group equivalents scheme for free energies of formation of organic compounds in aqueous solution. *Can. J. Chem.* **1992**, *70*, 1042–1054.

(18) Playsunov, A. V.; Shock, E. L. Group Contribution Values of the Infinite Dilution Thermodynamic Functions of Hydration for Aliphatic Noncyclic Hydrocarbons, Alcohols, and Ketones at 298.15 K and 0.1 MPa. *J. Chem. Eng. Data* **2001**, *46*, 1016–1019.

(19) Playsunov, A. V.; Shock, E. L. Thermodynamic functions of hydration of hydrocarbons at 298.15 K and 0.1 MPa. *Geochim. Cosmochim. Acta* **2000**, *64*, 439–468.

Received for review May 21, 2010. Accepted July 7, 2010.

JE100535P