ISSN 0027-1314, Moscow University Chemistry Bulletin, 2007, Vol. 62, No. 5, pp. 269–272. © Allerton Press, Inc., 2007.
Original Russian Text © N.I. Zhokhova, E.V. Bobkov, I.I. Baskin, V.A. Palyulin, A.N. Zefirov, N.S. Zefirov, 2007, published in Vestnik Moskovskogo Universiteta. Khimiya, 2007, No. 5, pp. 329–332.

Calculation of the Stability of β-Cyclodextrin Complexes of Organic Compounds Using the QSPR Approach

N. I. Zhokhova, E. V. Bobkov, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, and N. S. Zefirov

Department of Organic Chemistry

e-mail: zhokhova@org.chem.msu.ru

Received April 9, 2007

Abstract—The QSPR model for predicting the Gibbs free energy of formation of complexes of different organic compounds with β-cyclodextrin has been constructed by the multiple linear regression method with the use of the double cross-validation procedure. The model has good predictive power.

DOI: 10.3103/S0027131407050112

The development of pharmaceutical formulations with desired pharmacokinetic and therapeutic properties based on the incorporation of drug molecules into the cyclodextrin (CD) cavity is a promising direction for solving the problem of targeted drug delivery. Complexation with CDs prevents the biodegradation of an organic compound, decreases its volatility, enhances its solubility and bioavailability, and reduces the risk of undesirable side effects [1–4]. The principle advantages of natural CDs as carriers of drug molecules are their well-known structure, the possibility of chemical modification, low toxicity, and pharmacological activity [1]. β-CD and its derivatives are the most promising for practical use due to their relative availability and low cost.

The importance of the development of a reliable theoretical method of computation and prediction of the stability of organic complexes with CDs is caused, in particular, by the difficulties faced by experimenters in determining the corresponding parameters due to the low solubility of guest molecules in aqueous solutions. Computational methods of quantum chemistry and molecular dynamics have limitations associated with the large sizes and conformational flexibility of CD molecules in an aqueous solution [5]. The QSPR methodology (quantitative structure–property relationships) is alternative to these methods. QSPR is successfully used for predicting the properties of chemical compounds [6, 7]. Previously [8–11], some QSPR models were suggested for calculation of the thermodynamic stability of different organic complexes with CDs. These models have a common disadvantage: their predictive power has been overestimated. Traditional methods used for prediction in most studies involve either the use of one fixed set of compounds for independent prediction or the use (after selection of descriptors) of a standard cross-validation procedure, which often leads to overestimation of the predictive power of models since the information from the test set can be indirectly taken into account in constructing a model. This disadvantage can be avoided by using the double cross-validation procedure. Previously, we used this procedure for constructing models by the linear regression and artificial neural network methods for prediction of the physicochemical properties and biological activity of organic compounds [12]. In the present work, we used the multiple linear regression method and the double cross-validation procedure for calculation of the stability of host–guest complexes of different organic compounds with β-CD.

Experimental Gibbs free energies of formation (ΔG, kJ/mol) of complexes of 218 organic compounds (aromatic hydrocarbons, alcohols, phenols, ethers, esters, aldehydes, ketones, acids, sulfur-containing compounds, nitriles, anilines, heterocyclic compounds, steroids, and barbiturates) with β-CD in water at 25°C [8] were used. QSPR models were constructed and descriptors (physicochemical, topological, charge-based, steric, etc.) were calculated with the NASAWIN program package developed at the Chemistry Department, Moscow State University [13–16]. Fragmental descriptors were generated by an algorithm implementing the scheme of hierarchical classification of structural fragments and atoms [15]. The scheme makes it possible to generate a large number of sets of fragments of chemical structures and ensures a versatile classification of atomic types. For constructing multivariate (i.e., including a great number of descriptors) linear regression models with the use of the double cross-validation procedure, the initial database of chemical compounds is divided into three parts in the 3 : 1 : 1 ratio: the training, internal test, and external test sets, respectively. Each compound is successively included in all three sets. The information of the internal test set is used for control and optimization of the predictive power of models: the process of stepwise selection of descriptors and construction of models is terminated as the smallest prediction error is achieved for this set. The information

of the external test set is not used in constructing and selecting models and serves for independent evaluation of the resulting linear regression model. The predicted property for each compound is calculated as the average of all predicted values obtained for all partitions when this compound falls into the external test set. At the output, based on averaging statistical parameters of a large number (in this work, 20) of partial linear regression models, we obtained a multivariate model that contains (1) information on the statistical parameters of 20 linear regression models and the averaged multivariate model, (2) data of independent prediction of the modeled property for each of the compounds from the database, and (3) data on the type and contribution of each descriptor included in the multivariate model. This methodology ensures correct estimate of the predictive power of a multivariate linear model.

RESULTS AND DISCUSSION

At the first step of construction of QSPR models, we used descriptors that take into account the number of occurrences of structural fragments in the chemical structure of guest molecules. Previously, we studied the fragment approach for estimating the physicochemical properties of organic compounds, in particular, chromatographic retention indices [17]; boiling point [17]; flash point [18]; magnetic susceptibility [19]; polarizability [20]; density (for liquids), viscosity, and saturated vapor pressure of organic compounds [16]; etc. An advantage of fragmental descriptors is their clear meaning and the possibility of rapid automatic generation of these descriptors on the basis of only the structural formula of the compound, without taking into account the information on the spatial structure or electronic structure of molecules [15]. To select fragments that provide the most adequate description of the geometry and topology of the organic structures under consideration, we calculated the number of occurrences of the following fragments: chains containing from one to 15 non-hydrogen atoms, three- to five-membered rings; branched fragments containing from four to six non-hydrogen atoms, bicyclic rings containing from six to 15 atoms, and tricyclic rings containing from 12 to 15 atoms.

The best prediction characteristics were obtained for the multivariate model that considers fragments containing from one to four non-hydrogen atoms.

For external test sets:

the parameter Q^2 for double cross-validation, $Q^2_{\text{predict}} = 0.704$;

root-mean-square error, $\text{RMSE}_{\text{predict}} = 2.87$ kJ/mol;

mean absolute error, $\text{MAE}_{\text{predict}} = 1.98$ kJ/mol.

For the training sets:

averaged squared correlation coefficient, $R^2_{\text{train}} = 0.876$;

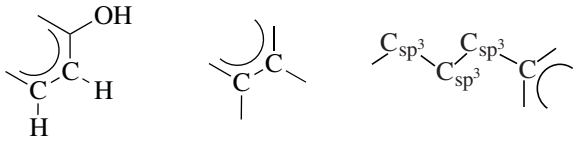
root-mean-square error,

$\text{RMSE}_{\text{train}} = 1.50$ kJ/mol;

Mean absolute error,

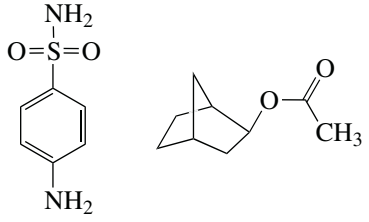
$\text{MAE}_{\text{train}} = 1.11$ kJ/mol.

To enhance the quality of the models based only on fragmental descriptors, it is expedient to include other types of descriptors that can describe the modeled property. As is known, according to the thermodynamic equilibrium conditions, hydrophobic organic molecules capable of entering into the CD ring form inclusion complexes with CDs in aqueous solutions. The free energy change is caused by the fact that the hydrophobic molecule is pushed out of the aqueous medium and that additional hydrogen bonds are formed between the water molecules being in contact with the hydrophobic molecule [1]. Among the factors that have an effect on complex formation are hydrophobic interactions, formation of hydrogen bonds between polar groups of the guest and hydroxyl groups of the host, steric hindrance, conformational changes of the CD–water adduct, and the change (release) in the Gibbs free energy when the substrate enters into the CD cavity and replaces water. Therefore, to construct models, we used, along with fragmental descriptors, various sets of the following descriptors: (1) lipophilicity ($\log P$); (2) hydrogen bond descriptors, which characterize the ability of an organic compound to act as a hydrogen-bond donor or acceptor; (3) descriptors taking into account steric factors and characterizing the molecular volume and molecular surface area; (4) descriptors characterizing the charge distribution in molecules; (5) descriptors related to the distribution of atomic properties in fragments (chains containing from one to five atoms) and based on the number of electrons, atomic radius, electronegativity, ionization potential, etc. More than 500 linear regression models were constructed based on combinations of different descriptors. The following criteria were applied to the selection of descriptors in the course of construction of the multivariate model: a given descriptor should increase the correlation coefficient for prediction and have a high index of occurrence in partial linear models. The best results were obtained when, in addition to the fragmental descriptors, the lipophilicity parameter $\log P$ and descriptors characterizing the distribution of atomic properties in structural fragments were used. It is worth noting that the lipophilicity parameter has the largest statistical significance for all models in which it was used. This is evidence of the importance of consideration of the hydrophobic component of the Gibbs free energy of interaction of an organic molecule with the CD since the latter has the hydrophobic cavity. The smallest contribution, in combination with the fragmental descriptors, is made by the descriptors characterizing steric features of guest molecules and by the hydrogen bond descriptors. As a result, we obtained a multivariate model constructed on the basis of fragmental descriptors, lipophilicity, and descriptors characterizing the properties of the atoms in



1 **2** **3**

Scheme 1.



4 **5**

Scheme 2.

the fragments. This model has the following statistical characteristics:

For external test sets:

the parameter Q^2 for double cross-validation, $Q^2_{\text{predict}} = 0.828$;

root-mean-square error, $\text{RMSE}_{\text{predict}} = 2.16$ kJ/mol;

mean absolute error, $\text{MAE}_{\text{predict}} = 1.56$ kJ/mol.

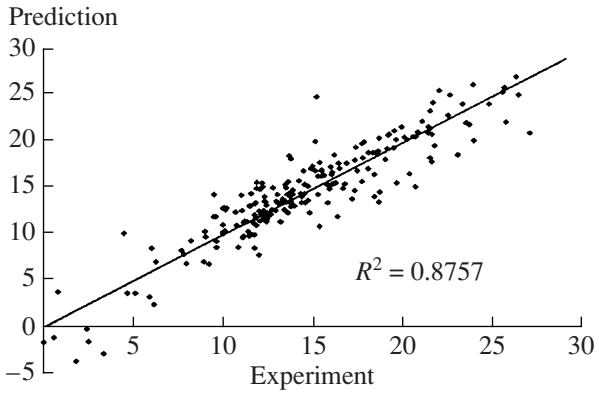
For the training set:

averaged squared correlation coefficient, $R^2_{\text{train}} = 0.936$;

root-mean-square error,

$\text{RMSE}_{\text{train}} = 1.19$ kJ/mol;

Mean absolute error,



Scatter diagram of the independent prediction of the Gibbs free energy (Δ*G*, kJ/mol) of formation of the complexes of the studied organic compounds with β-CD, obtained by the model, vs. the experimental data.

$\text{MAE}_{\text{train}} = 0.91$ kJ/mol.

For this combined model, the following descriptors are most significant (in order of decreasing contribution): lipophilicity; the average product of atomic radii in positions 1–5 for all five-atom chains; the number of occurrences of fragments of type **1**, **2**, and **3** (Scheme 1); and the sum of the differences in electronegativity for all X–H bonds (where X is any non-hydrogen atom) in a molecule.

Descriptors that characterize the fragment composition of guest molecules are necessary for constructing predictive models and compensate for the limitations of other types of descriptors. For this model, the largest errors of the independent prediction of the property (Δ*G*) are observed for the β-CD complexes with the following compounds: streptocid (the difference between the experimental and predicted values $\Delta G_{\text{exp}} - \Delta G_{\text{calcd}} = -8.22$ kJ/mol) and 2-norbornyl acetate (–10.21 kJ/mol) (Scheme 2, structures **4** and **5**, respectively). Exclusion of these compounds from the database does not noticeably change the statistical characteristics of the model. The figure shows the scatter diagram of the independent prediction of the Gibbs free energy (Δ*G*, kJ/mol) of formation of the complexes of the studied organic compounds with β-CD, obtained by the model, versus the experimental data.

Thus, using stepwise multiple linear regression and the double cross-validation procedure, we obtained the multivariate QSPR model based on the fragmental descriptors taking into account the number of occurrences of fragments containing up to four non-hydrogen atoms, as well as the lipophilicity parameter $\log P$ and descriptors based on the properties of the atoms in the fragments. The model has high predictive power for the Gibbs free energy Δ*G* of formation of β-cyclodextrin complexes with different organic compounds. The use of the double cross-validation procedure in the course of construction of the model provides a reliable and unbiased assessment of its predictive power.

REFERENCES

1. Uekama, K., Hirayama, F., and Irie, T., *Chem. Rev.*, 1998, vol. 98, p. 2045.
2. Astakhova, A.V. and Demina, N.B., *Khim.-Farm. Zh.*, 2004, vol. 38, no. 2, p. 46.
3. *Cyclodextrins and Their Industrial Uses,* Duchene, D., Ed., Paris, 1987.
4. *Cyclodextrins in Pharmacy*, Dordrecht, 1994.
5. Lipkowitz, K.B., *Chem. Rev.*, 1998, vol. 98, p. 1829.
6. Hansch, C. and Leo, A., *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*, Washington: ACS Professional Reference Book, 1995, vol. 1.
7. Zefirov, N.S. and Palyulin, V., *J. Chem. Inf. Comput. Sci.*, 2002, vol. 42, p. 1112.
8. Katrizky, A.R., Fara, D.C., Yang, H., Karelson, M., Suzuki, T., Solov'ev, V.P., and Varnek, A., *J. Chem. Inf. Comput. Sci.,* 2004, vol. 44, p. 529.

9. Klein, C.Th., Polheim, D., Viernstein, H., and Wolschann, P., *J. Incl. Phen. Macr. Chem.*, 2000, vol. 36, no. 4, p. 409.
10. Suzuki, T.A., *J. Chem. Inf. Comput. Sci.*, 2001, vol. 41, p. 1266.
11. Suzuki, T., Ishida, M., and Fabian, W.M., *Comput.-Aided Mol. Design*, 2000, vol. 14, p. 669.
12. Baskin, I.I., Zhokhova, N.I., Palyulin, V.A., Ivanova, A.A., Zefirov, A.N., and Zefirov, N.S., *XVI European Symposium on Quantitative Structure-Activity Relationships and Molecular Modeling. Mediterranean Sea, Italy. September*, 2006, p. 206.
13. Baskin, I.I, Palyulin, V.A, and Zefirov, N.S, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*, Sanz, F., Giraldo, J., and Manaut, F., Eds., Barcelona, 1995, p. 30.
14. Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., *J. Chem. Inf. Comput. Sci.*, 1997, vol. 37, p. 715.
15. Artemenko, N.V., Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., *Dokl. Akad. Nauk*, 2001, vol. 381, no. 2, p. 203.
16. Artemenko, N.V., Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., *Izv. Ross. Akad. Nauk, Ser. Khim.*, 2003, no. 1, p. 19.
17. Zefirov, N.S. and Palyulin, V.A., *J. Chem. Inf. Comput. Sci.*, 2001, vol. 41, p. 1022.
18. Zhokhova, N.I., Baskin, I.I., Palyulin, V.A., Zefirov, A.N., and Zefirov, N.S., *Izv. Ross. Akad. Nauk, Ser. Khim.*, 2003, no. 9, p. 1787.
19. Zhokhova, N.I., Baskin, I.I., Palyulin, V.A., Zefirov, A.N., and Zefirov, N.S., *Zh. Strukt. Khim.*, 2004, vol. 45, no. 4, p. 660.
20. Zhokhova, N.I., Baskin, I.I., Palyulin, V.A., Zefirov, A.N., and Zefirov, N.S., *Izv. Ross. Akad. Nauk, Ser. Khim.*, 2003, no. 5, p. 1005.