Journal of Inclusion Phenomena and Macrocyclic Chemistry **36:** 409–423, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

A Method for Predicting the Free Energies of Complexation between β-Cyclodextrin and Guest Molecules

CHRISTIAN TH. KLEIN[1,⋆], DIETER POLHEIM[2], HELMUT VIERNSTEIN[3] and PETER WOLSCHANN[1]
[1]*Institut für Theoretische Chemie und Strahlenchemie, Währinger Straße 17, A-1090 Vienna;* [2]*F. Joh. Kwizda Ges.m.b.H, Vienna;* [3]*Institut für Pharmazeutische Technologie, Althanstraße 14, A-1090 Vienna*

(Received: 3 December 1998; in final form: 14 February 1999)

Abstract. In the present work, linear regression models for the prediction of the free energies of complexation between guest molecules and β-cyclodextrin are deduced. For 70 compounds (mostly pharmaca), the experimentally determined 1 : 1 stability constants are transformed into the respective free energies, which are then correlated with molecular descriptors.

The statistically significant descriptors, which lead to models with remarkable predictive power, indicate that besides volume, shape and lipophilicity, which have the largest contribution to the complexation energy, complexation is also significantly influenced by the flexibility and the hydrogen bonding capacity of the guest molecule.

Key words: host-guest systems, linear regression, molecular descriptors, statistical models, correlation analysis

1. Introduction

Cyclodextrins (CDs) are cyclic macromolecules, obtained by the degradation of starch by α-1,4-glucan-glycosyltransferase. They are composed of 6 (α-CD), 7 (β-CD) or 8 (γ-CD) $\alpha(1 \rightarrow 4)$ linked glucose units [1]. The molecular shape of CDs resembles that of cones, having a hydrophobic cavity.

One of the most important properties of CDs is their ability to include small organic molecules (guests) in the cavity. The driving force of the complexation seems to be the hydrophobic effect, but, nevertheless, the complexes are also stabilized by van der Waals forces and hydrogen bonds [2].

The applications of cyclodextrins in pharmacy, environmental and technical chemistry, or other branches are wide, due to the multiple effects inclusion can have on the guest molecules: complexation with hydrophobic molecules makes the latter more water soluble and may be used for selective extraction, avoiding

[⋆] Author for correspondence: Tel.: +43 1 4277-52774; Fax: +43 1 4277-9527; E-mail: christian.klein@tbi.univie.ac.at organic solvents [3]. Light-, temperature- or oxidation-sensitive substances may be protected by complexation with cyclodextrins [3]. Pharmacon-cyclodextrin complexes often increase the bioavailability of the active substance and permit its controlled release [3]. Being chiral (D-glucose units), CDs interact differently with the enantiomers of the same compound, and can thus be used in enantioselective chromatography [3].

From these considerations it is clear, that a knowledge of the complexation abilities of guest molecules with CDs is necessary to decide whether or not a host-guest complexation is useful in a particular application.

On the other hand, experimental determination of the complexation constants is often difficult, mainly due to the low solubility of the guest molecules. A method for theoretical prediction of the complexation properties of guest molecules, would thus be desirable.

In the present work, we present a modality, based on multiple regression analysis, of theoretically estimating the free energies of complexations of β-CD:guest systems. Although most of the data available express the binding affinity of guest molecules to CDs in terms of stability constants, we correlate free energies (-RT ln K_{complex}) of complexation with molecular descriptors, from similar reasons like those in QSAR: since energy is an additive quantity, it can be described by an additive function of individual descriptors, each reflecting a certain contribution to the total energy. In contrast, individual contributions to the overall stability constant are multiplicative quantities.

2. Methods

As mentioned above, one of the driving forces of the host-guest inclusion is the hydrophobic interaction, van der Waals forces and hydrogen bonding also being important. It is obvious that the magnitude, shape and flexibility of the guest molecules should also be crucial factors in the complexation process. Thus, the following descriptors for modeling the host-guest complexation have been considered:

(a) the molecular surface (S), and the molecular volume (V) which is proportional to the size of the molecule;

(b) the ovality (O) [4], defined as the ratio of the actual surface (S) and the minimum surface S_{min}, that is the surface the molecule would have if it was a perfect sphere. It can be calculated from the actual molecular surface (S) and the corresponding molecular volume (V):

$$O = \frac{S}{S_{\text{min}}} = \frac{S}{4\pi\left(\frac{3V}{4\pi}\right)^{2/3}}. \tag{1}$$

Hence the ovality is a descriptor of the molecular shape;

(c) the partition coefficient (log P), being proportional to the hydrophobicity of a molecule, is used for the description of the hydrophobic interaction;

(d) the molecular refractivity, (MR), which is proportional to the volume and to the polarizability (α) of a molecule. Since dispersion (van der Waals) forces between two interacting moieties *1* and *2* are themselves proportional to their polarizabilities (α_1 and α_2), MR can give information on whether or not dispersion forces are important in the host-guest interaction;

(e) the ${}^3\kappa$-shape index and the flexibility (ϕ) of a molecule, as defined by Kier and Hall [5]; ϕ is directly related to the degree of linearity and the presence of rings and/or branching.

(f) the electrotopological index [6, 7], S_i, of an atom i based on the electronegativity of that atom and its local topology. S_i is calculated from an intrinsic state value (I_i), and a perturbation (ΔI_i) on atom i by all other skeletal atoms:

$$S_i = I_i + \Delta I_i. \tag{2}$$

I_i is a function of the principal quantum number of the atom i, the count of s electrons and the count of valence electrons in the skeleton. Because we deal with different classes of compounds, the comparison (i.e., correlation) of S_i values of certain atoms does not make sense (as it would in the case of homologous compounds within the same class). Therefore, the sum of S_i values over all atoms (E) is used as a descriptor; E describes whether the molecular surface is hydrophilic or hydrophobic, and thus models, together with log P, the lipophilicity of the guest molecule (see Discussion).

(g) the number of hydrogen bond donors present in the guest molecule (n_{HB}); donor groups considered are -OH, -NH, -SH;

(h) the number of heteroatoms (N, Cl, S, O) as indicator variables.

All properties excepting the ovality are calculated using the TSAR (tools for structure-activity relationship) program [8] from the Oxford Molecular Simulation Package. The ovality is calculated from the molecular surface (S) and the molecular volume (V) using Equation (1).

The regression models are also derived with the TSAR program, using multiple linear regression (MLR) and partial least squares (PLS). In the case of MLR, the two way stepping algorithm, which selects statistically significant variables via their partial F-test, is employed. The quality of the models is estimated considering the regression coefficient r, the overall F value which indicates whether the regression is significant or not, and the t values for testing the significance of individual regression coefficients. The F- and t-tests are briefly discussed in the Appendix. The predictive ability of a regression model is reflected by the cross-validation r^2 (r^2_{cv}) and is obtained as follows: after a required group of data is deleted, the remaining data are used to produce a new model y, which is then employed to predict the value that has been excluded. A model is produced for each group of data (n) and the *predictive residual sum of squares (PRESS)* is calculated:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{3}$$

where $\hat{y}_i$ is the predicted, and y_i the actual value. The value for the predictive r^2_{cv} is

$$r^2_{\text{cv}} = 1 - \frac{\text{PRESS}}{\text{SSY}}. \tag{4}$$

SSY is the sum of squares of the observations, and measures the total variability in the observations:

$$\text{SSY} = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{5}$$

The closer the value of r^2_{cv} is to 1, the better is the predictive power. For a good model r^2_{cv} should be close to r^2.

3. Results

The 70 compounds used for deducing the regression models are listed in Table I. From the references indicated in the table, the experimentally determined stability constants (at 25 °C) for 1 : 1 complexes are converted into free energies (-RT ln K).

There are included only compounds for which the stability constants have been determined by the same method, namely the solubility method [3].

The best model with respect to the correlation coefficient r, the standard error s, the F-value and the predictive r^2_{cv} is given by the following equation:

$$\begin{aligned} \Delta G \;=\; & -0.0186 \cdot S - 0.1767 \cdot \log P + 7.109 \cdot O + 0.3305 \cdot \phi - 0.2924 \cdot {}^3\kappa \\ & +0.0443 \cdot E - 0.1442 \cdot n_{\text{HB}} + 0.3921 \cdot n_N + 0.9257 \cdot n_{\text{Cl}} - 12.794 \end{aligned} \tag{6}$$

$$r = 0.927, \quad s = 0.377, \quad F_0 = 40.98, \quad r^2_{\text{cv}} = 0.812.$$

The F-test (see Appendix) indicates the significance of the regression model: $F_{0.05,9,60} = 2.04 \ll F_0$, in other words the regression equation is statistically significant at the 95% level. The predictive r^2_{cv} is high and fairly close to r^2. r^2_{cv} is obtained by successively leaving out *one* compound from the model building. The quality of Equation (6) can be underlined also by the fact that r^2_{cv} is stable, when a different number of groups is left out in cross validation: for 5, 10, 15 omitted compounds, the respective r^2_{cv} values are 0.819, 0.821 and 0.812.

Table I. Experimental and predicted values of free energies of complexation (kcal/mol). (a) Prediction with best MLR model (Equation (6)), (b) prediction with PLS model (Table IV).

Compound	$\Delta G_{\text{experimental}}$	$\Delta G^{\text{a}}_{\text{predicted}}$	$\Delta G^{\text{b}}_{\text{predicted}}$
1 Prostaglandin E_1^3	−4.388	−4.357	−4.351
2 Prostaglandin F_2^3	−4.202	−4.185	−4.152
3 Prostacyclin[3]	−4.013	−4.182	−4.102
4 Hydrocortisone[3]	−4.918	−5.040	−5.067
5 Beclomethasone dipropionate[3]	−4.142	−4.020	−4.187
6 Fludiazepam[3]	−3.182	−3.038	−3.006
7 Indomethacin[3]	−3.365	−3.360	−3.314
8 Flurbiprofen[3]	−5.036	−4.405	−4.419
9 Fenbufen[3]	−3.591	−3.949	−3.904
10 Ketoprofen[3]	−3.897	−4.702	−4.635
11 Piroxicam[3]	−2.654	−2.458	−2.414
12 Phenobarbital[3]	−4.441	−3.803	−3.743
13 Thiopental[3]	−4.472	−4.106	−4.118
14 Phenythoin[3]	−4.168	−4.389	−4.355
15 Sulphaphenazole[3]	−3.208	−3.075	−3.082
16 Acetohexamide[3]	−4.007	−3.610	−3.611
17 Clofibrate[3]	−4.252	−3.981	−3.989
18 Menadion[3]	−3.095	−3.344	−3.332
19 *p*-Ethylaminobenzoate[3]	−3.666	−3.700	−3.736
20 *p*-Butylaminobenzoate[3]	−4.360	−4.149	−4.188
21 *p*-Ethylhydroxybenzoate[3]	−4.109	−4.185	−4.181
22 *p*-Butylhydroxybenzoate[3]	−4.630	−4.648	−4.652
23 Medazepam[3]	−3.280	−3.412	−3.418
24 Prednisolone acetate[3]	−5.109	−4.841	−4.882
25 Cortisone[3]	−4.566	−4.903	−4.901
26 Cortisone acetate[3]	−4.915	−4.599	−4.589
27 Triamicinolone acetonide[3]	−4.767	−4.886	−4.932
28 Dexamethasone[3]	−4.980	−4.831	−4.855
29 Fluocinolone acetonide[3]	−4.723	−4.582	−4.627
30 Hydrocortisone acetate[3]	−4.771	−4.695	−4.699
31 Sulfapyridine[3]	−3.687	−3.907	−3.919
32 Sulfadimethoxine[3]	−3.080	−2.967	−3.008
33 Pentobarbital[3]	−4.115	−3.758	−3.744
34 Cyclobarbital[3]	−3.697	−3.947	−3.919
35 Hexobarbital[3]	−4.206	−3.95	−3.937
36 Mephobarbital[3]	−4.315	−3.649	−4.586
37 Triamicinolone diacetate[3]	−4.584	−4.007	−3.991
38 Nitrazepam[3]	−2.692	−2.699	−2.742

Table I. Continued.

Compound	$\Delta G_{\text{experimental}}$	$\Delta G^{\text{a}}_{\text{predicted}}$	$\Delta G^{\text{b}}_{\text{predicted}}$
39 Nimetazepam[3]	−2.364	−2.648	−2.718
40 *m*-Methylcinnamic acid[3]	−4.003	−3.709	−3.812
41 *p*-Hydroxycinnamic acid[3]	−3.859	−3.609	−3.705
42 Griseofulvine[9]	−2.006	−2.603	−2.523
43 Hydrochlorothiazide[10]	−2.400	−2.160	−2.150
44 Hydroflumethiazide[10]	−2.052	−2.557	−2.467
45 Mefenamic acid[11]	−3.403	−3.620	−3.604
46 Picotamide[12]	−2.395	−2.831	−2.842
47 Progabide[13]	−3.461	−2.840	−2.849
48 Proscillaridin[14]	−4.922	−5.029	−4.971
49 Prostaglandine A_1[15]	−4.274	−4.332	−4.317
50 Prostaglandine B_1[15]	−3.928	−4.308	−4.284
51 Sulfanilamide[16]	−3.774	−2.998	−3.047
52 Sulfamethomidine[17]	−3.182	−2.845	−2.877
53 Furosemide[18]	−2.435	−2.731	−2.741
54 Digitoxigenin[19]	−5.588	−5.601	−5.643
55 Cinnarizine[20]	−4.964	−5.044	−4.079
56 Dehydrocholic acid[19]	−5.179	−5.252	−5.271
57 Chlorothiazide[10]	−1.548	−1.927	−1.891
58 Carbutamide[17]	−3.126	−2.662	−2.729
59 Betamethasone valerate[3]	−4.722	−4.672	−4.685
60 Paramethasone[3]	−4.625	−4.828	−4.858
61 Sulfamonomethoxine[17]	−3.384	−3.196	−3.233
62 Sulfisomidine[17]	−2.872	−3.126	−3.159
63 Sulfisoxazole[17]	−3.167	−3.299	−3.297
64 Tolnaftate[21]	−5.235	−5.353	−5.330
65 Digitoxin[3]	−5.747	−5.713	−5.712
66 Acenocumarol[22]	−3.744	−3.366	−3.382
67 Allobarbital[23]	−2.705	−3.292	−3.259
68 Amobarbital[23]	−3.735	−3.905	−3.915
69 Bendroflumethiazide[10]	−2.589	−2.988	−2.952
70 Barbital[23]	−2.435	−3.547	−3.494

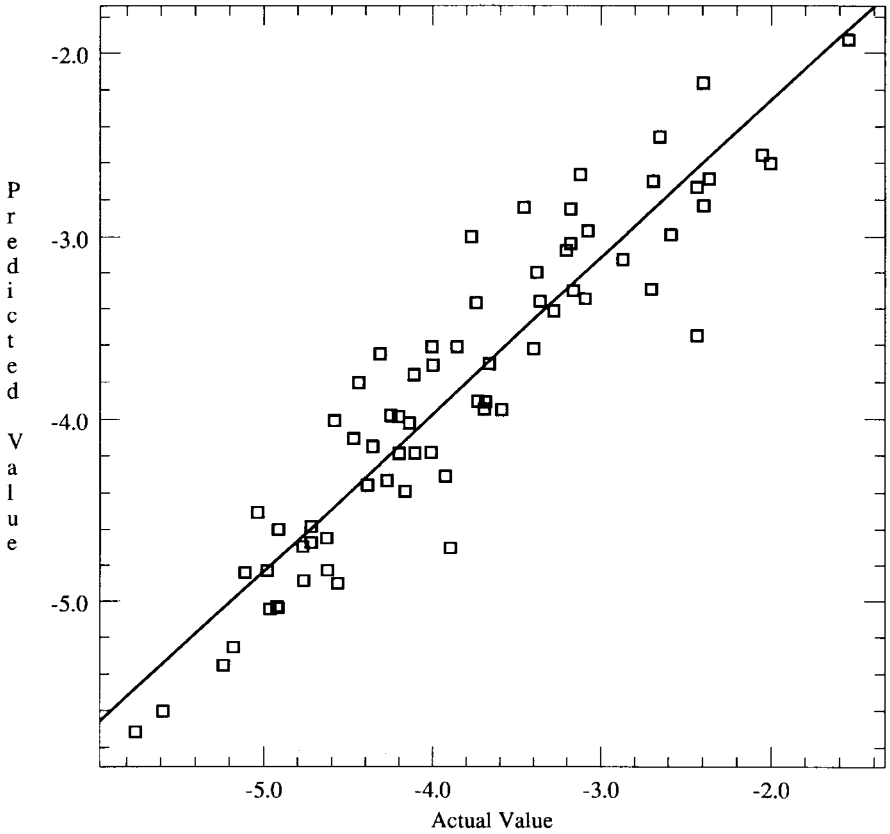

Figure 1. Predicted free energies of complexation (kcal/mol), plotted versus experimental values.

In Table I measured and predicted values are presented. They show mostly a good agreement, i.e., in 80% of the cases, the residual value between the experimental and predicted value is below the standard error. The experimental values are plotted versus the predicted ones in Figure 1.

The statistics of the individual regression coefficients is shown in Table II. Since the $t_{\alpha/2,60} = 2.0$ for $P = 0.95$, the condition $|t_0| > t_{0.025,60}$ is fulfilled for each of the regression coefficients, i.e., each of them is statistically significant (see Appendix).

In Table III the correlation matrix of the main descriptors, considered as candidates for the regression models, is presented.

It shows, that some of them are strongly correlated: S with V, MR with V, (per definition- and implicitly to S) and the sum E of electrotopological indices S_i with V. It is therefore not surprising, that a good model is also obtained with the molecular refractivity MR instead of S:

Table II. Statistics and significance of the coefficients of Equation (6).

Descriptor	Coefficient	SE	t_0-Value	t-Probability
S	−0.0186	0.0030	−6.200	0
$\log P$	−0.1767	0.0537	−3.290	1.6×10^{-3}
O	7.1096	1.1181	6.358	0
φ	0.3305	0.1100	3.004	3.8×10^{-3}
E	0.0443	0.0089	4.977	0
n_{HB}	−0.1442	0.0699	−2.063	4.3×10^{-2}
n_{N}	0.3921	0.0465	8.432	0
n_{Cl}	0.9257	0.1384	6.688	0
$^3\kappa$	−0.2925	0.1082	−2.702	8.9×10^{-3}
Constant	−12.7497	1.2558		

$$\begin{aligned}\Delta G &= -0.0344 \cdot MR - 0.2164 \cdot \log P + 5.374 \cdot O + 0.0872 \cdot \phi + 0.0317 \cdot E \\ &\quad -0.1962 \cdot n_{\mathrm{HB}} + 0.3868 \cdot n_{\mathrm{N}} + 0.9583 \cdot n_{\mathrm{Cl}} - 0.1722 \cdot^3 \kappa - 11.015 \quad (7)\end{aligned}$$

$$r = 0.917, \quad s = 0.401, \quad F_0 = 35.35, \quad r_{\mathrm{cv}}^2 = 0.784.$$

From the definition, MR is proportional to the molecular volume but also to the polarizability of a molecule. Hence it can be regarded also as a measure of how important dispersion forces are for the complexation process. The fact, that substituting S with MR does not improve the quality of the model, suggests that for the given data set MR does not contain any additional information compared to S, i.e., that the contribution to the variation stemming from MRs is due to the sizes of the molecules. Omission of the indictor variables (n_{N}, n_{Cl}, n_{HD}) substantially affects the quality of the regression equations (see Discussion):

$$\begin{aligned}\Delta G &= -0.0087 \cdot S - 0.3298 \cdot \log P + 2.153 \cdot O - 0.1427 \cdot \phi \\ &\quad +0.0193 \cdot E + 0.2548 \cdot^3 \kappa - 5.525 \qquad (8)\end{aligned}$$

$$s = 0.668, \quad r = 0.734, \quad F_0 = 12.26, \quad r_{\mathrm{cv}}^2 = 0.438.$$

To get a clearer insight into the importance of the individual contributions to complexation, MLR and PLS models are deduced, using variables scaled to zero mean and unity variance; thus the absolute values of the regression coefficients directly indicate the importance of the respective variable. PLS is a variant of principal components regression (PCR) and therefore permits analysis of highly correlated

Table III. Correlation matrix of the main descriptors used in the deduction of the regression models. Strong correlations ($r > 0.9$) are printed in bold.

	S	V	log P	O	ϕ	E	n_{HD}	n_{N}	n_{Cl}	$^3\kappa$	MR
S	1										
V	**0.97**	1									
log P	0.28	0.22	1								
O	0.67	0.47	0.37	1							
ϕ	0.82	0.69	0.35	0.83	1						
E	0.87	**0.93**	−0.01	0.36	0.57	1					
n_{HD}	0.49	0.47	−0.36	0.35	0.44	0.56	1				
n_{N}	−0.36	−0.34	−0.37	−0.36	−0.31	−0.25	−0.005	1			
n_{Cl}	−0.04	−0.02	0.09	−0.12	−0.05	−0.03	−0.34	0.8	1		
$^3\kappa$	0.54	0.36	0.39	0.85	0.87	0.26	0.29	−0.12	−0.05	1	
MR	**0.96**	**0.98**	0.26	0.53	0.68	0.86	0.42	−0.32	−0.003	0.38	1

Table IV. Coefficients and pseudo-regression coefficients of the best MLR (Equation (6)) and PLS model, obtained from variables scaled to zero mean and unity variance.

Descriptor	Coefficient	
	MLR	PLS
V		-1.0370
S	-1.5612	-0.4868
$\log P$	-0.2521	-0.2382
O	0.8518	0.6311
ϕ	0.6910	0.6861
E	0.7791	0.9072
n_{HD}	-0.1451	-0.1712
n_{N}	0.5459	0.5347
n_{Cl}	0.3262	0.3301
${}^{3}\kappa$	-0.5397	-0.5784
Constant	-3.8443	-3.8117
r^2_{cv}	0.812	0.829

variables; it is particularly useful, when the number of descriptors exceeds the number of experimental values. Table IV shows the pseudo-regression coefficients for the best PLS model (highest r^2_{cv}) , together with the scaled coefficients of Equation (6): they testify the high contribution of steric and lipophilicity parameters (see Discussion).

The good predictive ability of Equation (6) and of the PLS equation is also exemplified by a *de novo* prediction, using compounds not included in the deduction of the model, as shown in Table V.

4. Discussion

In Table IV the importance of individual contributions to the complexation process is directly related to the absolute value of the respective regression coefficients. It can be remarked that the major contribution to the complexation energy stems from steric descriptors (V and/or S, O ${}^{3}\kappa$), followed by descriptors of the lipophilicity (E, $\log P$). The fact that the highest contribution comes from the volume (in the PLS model), from the molecular surface (in the MLR model) and from the ovality of the guest molecules, indicates that the considered host-guest systems are well defined inclusion complexes. The ovality is a relative quantity and must therefore be interpreted together with the volume (or the surface). O and S have opposite contributions to the complexation energy and are in balance: if the

Table V. De novo prediction of free energies of complexation (kcal/mol), using (a) Equation (6) and (b) the PLS model from Table IV.

Compound	$\Delta G_{\text{experimental}}$	$\Delta G_{\text{predicted}}{}^{a}$	$\Delta G_{\text{predicted}}{}^{b}$
Diazepam	−3.182	−3.226	−3.217
Prostaglandine E_2	−4.216	−4.120	−4.057
Sulfadiazine	−3.441	−3.293	−3.316
Sulfamerazine	−2.696	−3.322	−3.350
Benzidine	−4.567	−3.674	−3.722
m-HO-cinnamic acid	−3.497	−3.643	−3.734
p-Methyl-cinnamic acid	−3.615	−3.665	−3.772
Scilliroside	−4.132	−4.179	−4.086
Triflumizole	−3.630	−2.720	−2.750

volume is large (favorable contribution to ΔG), the ovality must be also large (non-favorable contribution to ΔG), otherwise the compound cannot enter the cavity of the cyclodextrin, which in the case of β-CD has an average diameter of 0.78 nm [1]. The ${}^3\kappa$ shape index is related to the degree and centrality of branching in the guest molecule: it is larger when branching is nonexistent, or when it is located at the extremities of the molecular graph. The negative sign of the regression coefficient indicates that non-branched or terminally branched molecules should have increased complexation ability.

The larger the partition coefficient (log P), i.e., the more hydrophobic a compound is, the higher the stabilization of the complex due to the hydrophobic effect will be. This is in agreement with experimental findings, that the hydrophobic effect is a major force in the complexation process [2].

The sum of electrotopological indices, E, is decreased by less electronegative atoms, buried in the skeleton, and increased by terminal (i.e., generally more exposed) atoms of high electronegativity. An increased E-value will thus occur in molecules with a rather hydrophilic molecular surface; this aspect is not considered in the calculation of log P, since the corresponding increments do not depend on topological features. Hence, the destabilizing effect of increased E-values on the host-guest complexation can be explained from the above considerations. Being a measure of how polar the molecular surface of a molecule is, E is related to the hydrophilicity of the molecule, and thus represents a supplement to log P.

The flexibility of the guest also contributes substantially to the complexation. The larger ϕ is, the more flexible is a molecule. Because the coefficient of ϕ is positive, an increased value will lower the complexation energy: more rigid guests will have better complexation ability than more flexible ones, because the host-guest interaction is better defined in the first case.

MR does not improve the quality of a model, when it substitutes S; moreover, if both, MR and S, are considered in the deduction of models, as in Equation (6), only S is selected by the stepping algorithm, MR being statistically insignificant. These facts suggest that the variation in the MRs is stemming basically from the variation of the molecular size.

Some of the descriptors discussed above are rather strongly correlated with each other (Table III); this makes an interpretation of individual contributions somewhat difficult. On the other hand, excluding correlated descriptors from the regression equations substantially affects the predictive quality of the model, indicating that they contain also complementary information. However, the interpretation is facilitated by analyzing to what extend the information is complementary or similar. Consider, for example, the molecular surface, S, and the sum of electrotopological indices, E. The high correlation of 0.87 results essentially from the fact that both descriptors depend on the size (on the number of atoms) of a molecule. On the other hand, E reflects also the polarity of the molecular surface: by omitting E from Equation (6) the cross-validation r^2 (r^2_{cv}) drops from 0.812 to 0.741, which is a significant decrease of the predictive power. Evidently, in these cases of strong correlation conclusions on the relative importance of individual contributions cannot be inferred from the regression coefficients.

Indicator variables, which may be regarded as correction factors, are of great importance, as demonstrated by comparing Equations (6) and (8). In the case of n_{HB} the justification for such a correction is obvious: compounds with hydrogen bond donors have an additional possibility to interact with the guest's hydroxyl groups. Thus, if hydrogen bonding is significant in the complexation process, n_{HB} will directly reflect the contribution of the energy of hydrogen bonds to the overall complexation energy.

For the indicator variables n_{N} and n_{Cl} the justification is not as straightforward, but one has to keep in mind that in structure-activity studies series of homologous compounds are usually used; in the present study substances with a large variability (stemming from different classes) are employed to deduce the regression models: it is obvious that correction factors are necessary to improve the correlation. Considering the variability of the studied compounds, the obtained models are surprisingly good. A similar attempt to our approach has been made to correlate experimentally determined complex formation constants with measured quantities [27], resulting in a simple regression model. The authors used a dipolarity/polarizability parameter, solute and solvent hydrogen bonding terms and an intrinsic molecular volume of the solute in their study, which comprised 20 organic molecules. Their results are somewhat in variance with our findings, as the polarizability of the solute molecules appears to be important for the stability of the CD-guest complexes, while the hydrogen bonding capacity of the solute appears to be insignificant. However, one has to keep in mind that the size of and structural variance in our data set is much larger than in the respective paper. Moreover, the statistical significance of individual regression coefficients sensitively depends on the complexity of the model. A significant descriptor in a simple model may loose its significance in a more complex model, while other descriptors, formerly insignificant, become important. The success of correlation analysis applied to such different compounds as in the present study can be explained by the fact, that the host-guest interaction in the case of CDs is not as specific as it is for receptor-ligand interaction: while e.g., in enzyme-substrate docking *local* (steric, hydrophobic and electrostatic) properties of the ligand are crucial for an optimal interaction, for the CD-guest systems *overall* molecular properties like volume, shape, hydrophobicity, ovality or flexibility appear to be sufficient for a good description of complexation.

5. Conclusions

In the present work regression models are presented which permit a good prediction of the free energies of complexation between β-CD and guest molecules. The statistically significant descriptors are molecular surface (S), ovality (O), ${}^3\kappa$ shape index, flexibility (ϕ), partition coefficient ($\log P$), the sum of the electrotopological indices E and indicator variables. Log P describes the lipophilicity and E the polarity of the molecular surface, while S, O, ${}^3\kappa$ are descriptors of magnitude and molecular shape of the guest, respectively. Rigid molecules (low ϕ) appear to have higher complexation ability than flexible ones, since the host-guest complex is better defined. Molecules with hydrogen bond donors have the additional possibility to stabilize the complex, by forming hydrogen bonds with the host's hydroxyl groups, reflected in n_{HB}. The predictive abilities of the model is good, reflected in high predictive r^2_{cv} values at different leave-out levels and good *de novo* predictions. Moreover, the used descriptors can be easily and rapidly calculated from commercial modeling packages.

Acknowledgements

The present work has been performed within a project of the European Union (FAIR-CT96-1436). The authors thank Luckhana Lawtrakul for reading the paper carefully and for useful comments.

Appendix

5.1. SIGNIFICANCE OF THE REGRESSION: THE f-TEST

If two variances, s_1^2 and s_2^2 of a random variable are consistent, i.e., if their differences are not significant, the ratio $F_0 = s_1^2/s_2^2$ is also a random variable and is described by the F (Fisher)-repartition. The probability density of the F-distribution depends on the degrees of freedom, ν_1 and ν_2 of the two variances: $q = q(F; \nu_1, \nu_2)$. Since the analytical expression of q is known, F_{α,ν_1,ν_2} values can be calculated for any probability P ($\alpha = 1 - P$) from the definition of the repartition function as the probability, that a certain random variable is lower than a given value:

$$P(F < F_{\alpha,\nu_1,\nu_2}) = Q(F_{\alpha,\nu_1,\nu_2}) = \int_0^{F_{\alpha,\nu_1,\nu_2}} q(F;\nu_1,\nu_2)\,\mathrm{d}F. \tag{A1}$$

The F_{α,ν_1,ν_2} values are usually found in appropriate tables. For example $F_{0.05,2,4} = 6.94$, i.e., a F-distributed random variable with 2 and 4 degrees of freedom will be lower than 6.94 with a probability of 0.95 (at 95% level). The F_{α,ν_1,ν_2}-values are used by the *F-test* to accept or reject the hypothesis, whether two variances are consistent or not. Thus, if $F_0 > F_{\alpha,\nu_1,\nu_2}$, the hypothesis is rejected, and the rejection is the more significant, the larger F_0 is.

In the concrete case of testing the significance of a regression, the *F-test* is applied as follows: Consider n dependent variables (observations) y_i, each being related to (nxk) independent variables, $x_{i1}, x_{i2}, \ldots, x_{ik}$, by the regression coefficients a_j:

$$y_i = a_0 + \sum_{j=1}^{k} a_j x_{ij}. \tag{A2}$$

The smaller the sum of squares of the residuals (SSE), i.e., of the differences between the observed values and predicted ones, compared to the sum of squares of the regression, SSR, obtained from the differences between predicted values and the mean value, the better is the regression model.

One can show [25] that SSR has k degrees of freedom, whereas the sum of squares of the residuals, SSE, has $(n-k-1)$ degrees of freedom, and, moreover, that SSR and SSE are independent. Hence the variation due to the regression and due to the residuals are, respectively

$$s_R^2 = \frac{SSR}{k} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}, \quad s_E^2 = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}. \tag{A3}$$

Since SSE should be as small as possible, the ratio $F_0 = s_R^2/s_E^2$ should be large and the hypothesis that s_R^2 and s_E^2 are consistent should be rejected. This is the case if $F_0 > F_{\alpha,k,n-k-1}$ and thus the regression is significant. Some authors suggest [26], that for a satisfactory predictive power, the F_0-ratio should exceed at least four times the percentage point $F_{a,k,n-k-1}$.

5.2. SIGNIFICANCE OF INDIVIDUAL REGRESSION COEFFICIENTS: THE t-TEST

Following a similar argumentation as above, the significance of individual regression coefficients is usually tested employing the *t (Student)-test*. The variable

$t_0 = a_j/se(a_j)$, where $se(a_j)$ denotes the standard error of the regression coefficient a_j, is t-distributed. If $|t_0| > t_{\alpha/2,n-k-1}$, the coefficient will be considered significant.

References

1. J. Szejtli: *Topics in Inclusion Science-Cyclodextrin Technology*, Kluwer Academic Publisher (1988).
2. Y. Kotake and E. G. Janzen: *J. Am. Chem. Soc.* **111**, 8551 (1989).
3. D. Duchêne (ed.): *Cyclodextrins and their Industrial Uses*, Editions de Santé Paris (1987).
4. N. Bodor, Z. Gabanyi, and C-K. Wong: *J. Am. Chem. Soc.* **111**, 3783 (1989).
5. L. B. Kier and L. H. Hall: in K. B. Lipkowitz and D. B. Boyd (eds.), *Reviews in Computational Chemistry*, Chap. 9, 1992, p. 367.
6. L. B. Kier, and L. H. Hall: *Pharm. Res.* **7**, 801 (1990).
7. L. H. Hall, B. Mohney, and L. B. Kier: *J. Chem. Inf. Comput. Sci.* **31**, 76 (1991).
8. TSAR3.1 User Guide: Oxford Molecular Limited, 1997.
9. M. Kata and B. Tüske: *Pharmazie* **43**, 52 (1988).
10. O. I. Corrigan and C. T. Stanley: *J. Pharm. Pharmacol.* **34**, 621 (1982).
11. Y. Hamada, N. Nambu, and T. Nagai: *Chem. Pharm. Bull.* **23**, 1205 (1975).
12. P. Mura, A. Liguori, G. Bramanti, and L. Poggi: *Acta Pharm. Technol.* **34**, 77 (1988).
13. N. F. Farraj, S. S. Davis, G. D. Parr, and H. N. E. Stevens: *Int. J. Pharm.* **52**, 11 (1989).
14. K. Uekama, T. Fujinaga, M. Otagiri, N. Matsuo, and Y. Matsuoka: *Acta. Pharm. Sued.* **20**, 287 (1983).
15. K. Uekama, F. Hirayama, and T. Irie: *Chem. Lett.* 661 (1978).
16. K. Uekama, Y. Uemura, F. Hirayama, and M. Otagiri: *Chem. Pharm. Bull.* **31**, 3284 (1983).
17. K. Uekama, F. Hirayama, M. Otagiri, Y. Otagiri, and K. Ikeda: *Chem. Pharm. Bull.* **26**, 1162 (1978).
18. E. Fenyvesi, K. Takayama, J. Szejtli, and T. Nagai: *Chem. Pharm. Bull.* **32**, 670 (1984).
19. Y. Okada, Y. Kubota, K. Koizumi, S. Hizukuri, T. Ohfuji, and K. Ogata: *Chem. Pharm. Bull.* **36**, 2176 (1988).
20. T. Tokumura, Y. Tsushima, M. Kayano, Y. Machida, and T. Nagai: *J. Pharm. Sci.* **74**, 496 (1985).
21. J. Szeman, E. Fenyvesi, B. Zsadon, M. Szilasi, and L. Decsei, Europ. Patent No. 0119453 A3 (1984).
22. M. Otagiri, J. G. Fokkens, G. E. Hardee, and J. H. Perrin: *Pharm. Acta Helv.* **53**, 241 (1978).
23. K. Koizumi, H. Miki, and Y. Kubota: *Chem. Pharm. Bull.* **28**, 319 (1980).
24. K. Uekama, F. Hirayama, K. Esaki, and M. Inoue: *Chem. Pharm. Bull.* **27**, 76 (1979).
25. D. C. Montgomery and E. A. Peck: *Introduction to Linear Regression Analysis*, J. Wiley & Sons (1992), p. 135.
26. N. R. Drapper, and H. Smith: *Applied Regression Analysis*, J. Wiley & Sons (1981), p. 93.
27. J. H. Park and T. H. Nah: *J. Chem. Soc. Perkin Trans. 2*, 1359 (1994).