Predicting Value of Binding Constants of Organic Ligands to Beta-Cyclodextrin: Application of MARSplines and Descriptors Encoded in SMILES String

Piotr Cysewski and Maciej Przybyłek *

Department of Physical Chemistry, Faculty of Pharmacy, Collegium Medicum of Bydgoszcz, Nicolaus Copernicus University in Toruń, Kurpińskiego 5, 85-950 Bydgoszcz, Poland
* Correspondence: m.przybylek@cm.umk.pl

Received: 29 June 2019; Accepted: 12 July 2019; Published: 15 July 2019

Abstract: The quantitative structure–activity relationship (QSPR) model was formulated to quantify values of the binding constant (lnK) of a series of ligands to beta-cyclodextrin (β-CD). For this purpose, the multivariate adaptive regression splines (MARSplines) methodology was adopted with molecular descriptors derived from the simplified molecular input line entry specification (SMILES) strings. This approach allows discovery of regression equations consisting of new non-linear components (basis functions) being combinations of molecular descriptors. The model was subjected to the standard internal and external validation procedures, which indicated its high predictive power. The appearance of polarity-related descriptors, such as XlogP, confirms the hydrophobic nature of the cyclodextrin cavity. The model can be used for predicting the affinity of new ligands to β-CD. However, a non-standard application was also proposed for classification into Biopharmaceutical Classification System (BCS) drug types. It was found that a single parameter, which is the estimated value of lnK, is sufficient to distinguish highly permeable drugs (BCS class I and II) from low permeable ones (BCS class II and IV). In general, it was found that drugs of the former group exhibit higher affinity to β-CD than the latter group (class III and IV).

Keywords: beta-cyclodextrin; QSPR; binding constant

1. Introduction

Molecular complexes, such as inclusion adducts, clathrates, cocrystals and solvates, have been widely used in many fields, including pharmacy [1–6], agriculture [7], the food industry [1,8,9] and explosives [10,11]. In the past two decades, cyclodextrins (CDs) have been one of the most extensively studied complexation agents, especially as pharmaceutical excipients [12–15]. The CDs’ adducts with active pharmaceutical ingredients (APIs) are mainly used for solubility and bioavailability enhancement [12–17], stability improvement [13,18], stomach, skin and eye irritation reduction [13,17,19], and prevention of unpleasant odor and bitter taste [13,20]. Probably the most commonly used compounds in pharmaceutical formulations belonging to this class are alpha- (α-CD), beta- (β-CD), gamma- (γ-CD) cyclodextrins and their analogues such as (2-hydroxypropyl)-beta-cyclodextrin (HP-β-CD), sulfobutylether beta-cyclodextrin (SBE-β-CD) or randomly methylated-β-cyclodextrins (RM-β-CDs) [21]. The main criterion used for distinguishing different CDs (α-CD, β-CD and γ-CD) corresponds to six, seven and eight D-glucopyranose units, respectively. These excipients have been used for all main types of drug delivery systems (oral, nasal, rectal, dermal, ocular, parenteral) [13,21].
Apart from pharmaceutical applications, cyclodextrins are widely used in personal care products and perfumes manufacturing [20,22,23], which relies on their high stability, solubilizing abilities and vapor pressure reduction (fragrances industry). The unique properties of cyclodextrins are related to their specific structural features. These compounds are characterized by a symmetric toroidal shape. Due to the relatively hydrophobic cavity, cyclodextrins play the role of hosts in molecular inclusion complexes. On the other hand, they are hydrophilic on the outside surface, which results in strong interactions with water molecules. This characteristic structure is somewhat similar to biocatalysts [24–26]. An interesting example of the use of such cyclodextrin-based artificial enzymes is asymmetric and stereospecific synthesis (halogenation, hydrohalogenation, oxidation, reduction, photolysis, aldol reactions, hydrogenation, substitution and addition reaction) [25]. Noteworthy, the stereoselectivity of CDs was utilized for separation of racemic mixtures [27,28].

Quantitative structure–activity relationship (QSPR) methodology has been extensively used for evaluating the formation abilities of molecular complexes, and characterizing their properties [4,29–38]. Cyclodextrin binding constant modeling deserves special attention due to its practical importance. In recent years, several interesting approaches have appeared, such as application of molecular docking [38], conductor like screening model for real solvents (COSMO-RS) and quantum chemical-based descriptors [31], and topological indices [32,34,39]. Most of these models are simple regression equations. In general, better accuracy can be achieved when non-linear methods are applied. In our previous work, a novel approach of combining the non-linear MARSplines (multivariate adaptive regression splines) [40] methodology with common molecular descriptors calculated from simplified molecular input line entry specification (SMILES) code was applied for solubility modeling [41,42]. The major advantage of this procedure is its good predictive power and relatively simple model, which is a regression equation of new factors. The aim of this study was to apply a similar methodology for β-CD stability binding modeling.

2. Materials and Methods

The experimental values used for model development and validation were obtained from the datasets published by Suzuki et al. [43] and Mirrahimi et al. [38]. This collection comprises binding constants of 1:1 β-CD complexes with different organic compounds. In case of experimental values of the same compounds from different sources the mean value was taken into account. The list of all data is provided in the Supplementary Materials (Table S1).

2.1. Molecular Descriptors

Currently, a variety of molecular parameters are freely available for potential applications, which ensures that formulated models can be readily applied for predictions of compounds’ properties. Here two online tools were used for collecting the set of descriptors, namely ChemDes [44] and the BioCCI module of the BioTriangle platform [45]. The former provides a direct and integrated way of retrieving the sets of descriptors catalogued as Chemopy Descriptors (1135), CDK Descriptors (275), RDKit Descriptors (196), Pybel Descriptors (24), BlueDesc Descriptors (174) and PaDEL Descriptors (1875). The number of potential parameters is provided in parenthesis. All these indices can be calculated on-line [46]. The second source also offers a limited number of descriptors but offers sets suited for intermolecular interactions, which is the key advantage of this software. This is available using the BioTriangle webserver [47]. Since the BioTriangle allows for different geometrical transformations of descriptors calculated for pairs of molecules, the β-CD-ligand pairs descriptors were included.

2.2. Data Pre-Treatment

After completing the datasets of all descriptors, the standard pre-treatment procedure was implemented. It comprised elimination of descriptors not computable for the whole set of ligands and the remaining content across the whole population. Then, highly correlated and low-variance descriptors were also removed. Data curating was undertaken by taking advantage of the Data
PreTreatment 1.2 module relying on the variable reduction Wootton, Sergent and Phan-Tan-Luu’s (V-WSP) algorithm [48,49]. Then, the dataset was divided into training and test sets using the activity division approach implemented in the Dataset Division 1.2 tool [50,51]. Both programs are written in Java and are freely available [52].

2.3. Model Development Using MARSplines

In this work, a MARSplines [40] methodology was applied as implemented in STATISTICA 12 [53]. This methodology leads to the following general regression formula, where $F_i$ is the regression factor and $a_i$ are the regression parameters:

$$\ln(K_{est}^{\beta CD}) = a_0 + \sum_{i=1}^{n} a_i F_i$$  \hspace{1cm} (1)

The left-hand side of this equation stand for the response variable, which is confronted with experimental values. Here it is defined by the value of the natural logarithm of a binding constant quantifying affinity of a ligand toward $\beta$–CD determined experimentally. The MARSplines method is the procedure designated for finding the analytical formula based on descriptors and so-called knots. Such relationships are termed basis functions. The values represent splitting of the set of values into sub-regions treated with alternative mathematical formula. The number of basis functions and factors in the model is controlled at an arbitrary level for balancing between accuracy and complexity of the model. To avoid model overfitting, the final model undergoes inspection of the regression coefficients by removing such factors for which statistical significance is not reached ($p > 0.05$). Additionally, the contribution to the model of each factor is inferred from the values of standardized regression coefficients ($\beta_i$). Only such factors are included in the final model for which $|\beta_i| > 0.09$. Furthermore, the model was refined, internally validated and characterized in terms of fitting criteria using QSARINS software [54–56]. As a result of this procedure, the model was simplified by selecting the most important variables using a genetic algorithm (GA).

The simplest factor generated by the MARSplines procedure has a form identical to the classical QSPR approach and is expressed simply as multiplication of descriptor values by a coefficient, whose value is optimized for maximizing correlations between computed and estimated response values. The main improvement, however, comes from accounting for non-linearity by direct inclusion of more complex basis functions combined into factors. Hence, an advantage of the QSPR model formulation using the MARSplines procedure is the benefit of formally being in the multiple linear regression (MLR) format by including non-linear properties of considered datasets. Hence, the golden standard QSPR model development and validation procedures can be directly applied [41,42,57].

3. Results and Discussion

There are two main reasons that justify the efforts of the obtained model building. The first is obviously of substantive nature for deriving a model that is as accurate as possible and characterized by a low cost of applications. Hence, the screening of new potential ligands, or comparing a leading compound of an API and derivatives suggested by a drug design procedure, represent the immediate value of the obtained model. There is also a methodological reason for exploring the landscape of potential application in the chemistry domain of the MARSplines procedure. This is not explored deeply enough bearing in mind its high potential, effectiveness and ease of use.

3.1. Findings

Based on the MARSplines algorithm, the following descriptors were included in the model (Table 1): XLogP and Wlambda2.unity (source: BlueDesc); carbonTypes.8 (source: CDK), MLFER_A, AATS6m, AATS4i and PNSA-3 (source: PADEL); the tensor product of PEOEVSA9 descriptor vectors denoted as PEOEVSA9*PEOEVSA9; and, the vector sum of Chiv1 parameter denoted as Chiv1plusChiv1 (source:...
Taking into account the relatively large training set population \( n = 187 \), the number of variables seems to be reasonable fulfilling the general rules of acceptable QSPR model complexity, as documented in Table 1 and Figure 1. Internal validation, fitting criteria and external validation parameters, including \( R^2 \) (determination coefficient), \( R_{adj}^2 \) (adjustment determination coefficient), \( F \) (Fisher ratio), SD (standard deviation), MAE (mean absolute error), MAPE (mean absolute percentage error), RMSE (root-mean-square error), PRESS (predicted residual error sum of squares) and \( K_{xx} \) (descriptors’ global correlation measure) [58,59], suggest that the model is well fitted to the training set and, most importantly, the external test set examples were well predicted. The results of external validation are presented in Figure 1. As one can see, the proposed model is characterized by high determination coefficients. Interestingly, \( R^2 \), MAE and MAPE values are even slightly better for the external test set (0.936, 0.44, 9.3%, respectively) than for the training set (0.907, 0.49, 15.4%, respectively). This suggests that the model complexity is optimal. It is worth mentioning that the over-fitting problem should be taken into account when analyzing the quality of QSPR models, especially those that are non-linear. In the case of overly complex models, the training set data are exceptionally well fitted, but the test set prediction quality is far inferior. It is worth mentioning that the MARSplines protocol implemented in the STATISTICA software prevents overfitting by taking advantage from the generalized cross validation (GCV) algorithm, which reduces the model to be as simple as possible.

### Table 1. Multivariate adaptive regression splines (MARSplines) model parameters along with the validation results.

<table>
<thead>
<tr>
<th>Factor</th>
<th>( \beta_i )</th>
<th>( \alpha_i )</th>
<th>Basis Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>5.4277</td>
<td></td>
<td>Max(0; 28.194-Chiv1+Chiv1)</td>
</tr>
<tr>
<td>F1</td>
<td>-0.3991</td>
<td>-1.2679</td>
<td>Max(0; XLogP+0.1340)</td>
</tr>
<tr>
<td>F2</td>
<td>0.5652</td>
<td>1.1090</td>
<td>Max(0; carbonTypes.8)</td>
</tr>
<tr>
<td>F3</td>
<td>0.3772</td>
<td>1.3356</td>
<td>Max(0; Wlambda2.unity-1.2400)</td>
</tr>
<tr>
<td>F4</td>
<td>-0.1559</td>
<td>-1.4658</td>
<td>Max(0; 0.5620-MLFER_A)</td>
</tr>
<tr>
<td>F5</td>
<td>-0.1613</td>
<td>-1.3482</td>
<td>Max(0; Wlambda2.unity-1.2400)</td>
</tr>
<tr>
<td>F6</td>
<td>-0.1391</td>
<td>-0.3182</td>
<td>Max(0; 0.5620-MLFER_A)</td>
</tr>
<tr>
<td>F7</td>
<td>-0.2130</td>
<td>-2.5385</td>
<td>Max(0; XLogP+0.1340)-Max(0; 15.8078-PNSA-3)</td>
</tr>
<tr>
<td>F8</td>
<td>-0.1372</td>
<td>-0.0120</td>
<td>Max(0; 66.0412-AATS6m)-Max(0; MLFER_A-0.5620)</td>
</tr>
<tr>
<td>F9</td>
<td>-0.0977</td>
<td>-0.0003</td>
<td>Max(0; PEOEVSA9*PEOEVSA9-9.8878-370)-Max(0; XLogP+0.1340)</td>
</tr>
<tr>
<td>F10</td>
<td>-0.1258</td>
<td>-0.0002</td>
<td>Max(0; 988.378-PEOEVSA9*PEOEVSA9)-Max(0; XLogP+0.1340)</td>
</tr>
<tr>
<td>F11</td>
<td>0.0910</td>
<td>0.0257</td>
<td>Max(0; Wlambda2.unity-1.2400)-Max(0; AATS4i-154.1756)</td>
</tr>
<tr>
<td>F12</td>
<td>0.0944</td>
<td>0.2092</td>
<td>Max(0; Wlambda2.unity-1.2400)-Max(0; PEOEVSA9*PEOEVSA9-9.8878-370)</td>
</tr>
</tbody>
</table>

Model statistics: internal validation (MAE_CV = 0.51, RMSE_CV = 0.65, \( Q^2_{LOO} = 0.90, Q^2_{LMO} = 0.90, \) PRESS_CV = 99.00), fitting criteria (\( N = 187, R^2 = 0.91, \) \( R_{adj}^2 = 0.91, \) MAE = 0.61, RMSE = 0.63, \( F = 189.45, SD = 0.63, K_{xx} = 0.35 \)) and external validation (training set: MAE = 0.49, MAPE = 15.4%, test set: MAE = 0.44, MAPE = 9.3%).

Some of the parameters used in the model, such as like XLogP and MLFER_A, are quite intuitive and their physical meaning can be easily explained. The appearance of the hydrophilic measure, namely the group contribution logP parameter (XlogP), confirms the role of the hydrophobic nature of the cyclodextrin cavity, while MLFER_A is the Abraham solubility parameter expressing the acidity. The role of polarity in \( \beta \-CD \) molecular complexes formation was emphasized by PNSA-3 (charged partial surface area index [60]) and BioTriangle interaction descriptor PEOEVSA9*PEOEVSA9. This latter feature was calculated based on the MOE-type parameter involving the contributions of surface area and partial charge [61]. Another feature calculated using the BioTriangle platform, namely Chiv1plusChiv1, is associated with the Chiv1 descriptor belonging to the atomic valence connectivity indices class [62,63]. Of note, these descriptors were widely used in solving quite similar QSAR problems associated with target-ligand binding [64–68]. In the MARSplines model, there was also one topological descriptor characterizing carbon type (carbonTypes.8) [69] and the appearance of two autocorrelation indices, AATS6m and AATS4i [69]. Autocorrelation descriptors are probably one of the most extensively used quantitative structure-activity relationship/quantitative structure property
relationship (QSAR/QSPR) descriptors. Although the physical meaning of these parameters is not straightforward, our previous studies showed that this broad class of descriptors was found to be useful in the modelling of the affinity of compounds in the solid state [29,57].

![Figure 1. The relationship between experimental and calculated values of binding constants.](image)

### 3.2. Comparison to Existing Models

Comparison of the determination coefficient calculated for the obtained model with two regression models reported in recent years is presented in Table 2. Although these models were generated using similar datasets, it should be taken into account that depending on the validation procedure, different results can be obtained. Nevertheless, the correlation coefficients are lower or approximately equal to the MARSplines model. This suggests that the proposed approach is a good alternative for β-CD calculation. The major advantage of calculating molecular descriptors from the SMILES code is low computational cost. However, the proposed QSPR model has some limitations associated with ignoring the geometrical features of molecular complexes, such as conformation and solvation effects. These effects can be included using optimized 3D structures. Furthermore, it should be taken into account that in some cases, the stoichiometry of β-CD complexes is not 1:1 [70–72]. In such cases, molecular modelling methods such as molecular-dynamics docking or quantum-chemical binding constant calculations are more appropriate than the proposed approach.

<table>
<thead>
<tr>
<th>Model Description</th>
<th>R²</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Set</td>
<td>Test Set</td>
</tr>
<tr>
<td>MARSplines</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td>Molecular docking-based descriptors</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>Monte Carlo optimised topological descriptors</td>
<td>0.92</td>
<td>0.93</td>
</tr>
</tbody>
</table>
3.3. Exemplary Model Applications

The obvious application of the model provided by Equation (1) and Table 1 relates to its predictive power. Hence, it is possible to anticipate, before actual measurement, the probable affinity of the considered API toward β-CD. There is, of course, a limitation due to the applicability domain. For example, there are no organic and metalo-organic salts in the model. Hence, it is very unlikely that the model helps in situations where drugs are prepared in such forms. However, many drugs are, in principle, treatable by the model and at least the rational selection of the candidates for experimental measurements can be advised.

It is also possible to suggest alternative, less obvious applications of the formulated MARSplines model. For example, in the Biopharmaceutical Classification System (BCS) it is assumed that two measures such as solubility and permeability can be used for grouping drugs in respect of their bioavailability. In Table 3 this classification is shown [73]. Of note, cyclodextrins and their solubilizing abilities have been discussed in the context of BCS classification [74,75].

Table 3. Biopharmaceutical Classification System (BCS) [73] using solubility and permeability as qualitative criterions.

<table>
<thead>
<tr>
<th>High Solubility</th>
<th>Low Solubility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>Class II</td>
</tr>
<tr>
<td>This class comprise compounds characterized by good absorption profiles.</td>
<td>The bioavailability is directly related to the dissolution behavior.</td>
</tr>
<tr>
<td>Class III</td>
<td>Class IV</td>
</tr>
<tr>
<td>The active pharmaceutical ingredient (API) is soluble, however absorption profile is dependent on limited permeation behavior.</td>
<td>The API is characterized by very low bioavailability.</td>
</tr>
</tbody>
</table>

Hence, for proper bioavailability assessment, both water solubility and permeability must be known. It is interesting to see if there is any correlation between the BCS class of a given drug and its estimated affinity toward β-CD. For this purpose, information about the BCS classification was collected for 300+ drugs [76]. Those that are found to be outside of the applicability domain were excluded from the analysis. For the remaining drugs, the values of the molecular descriptors were collected. This, in turn, allowed for application of the MARSplines model and prediction of lnK values. The obtained results are presented in Figure 2 and Table S2. As can be seen from Figure 2, those APIs exhibiting good permeability (Class I and II) are characterized by higher affinity to cyclodextrin. This is understandable since the cyclodextrin cavity is rather hydrophobic, like for lipid biological barriers. The most important message coming from Figure 2 is that Class I and II have very similar distributions to each other and, at the same time, are distinct from Class III and IV. Indeed, application of a statistical non-parametrical test revealed that the medians are statistically the same (p = 0.27) for Class I and II but either combination with remaining classes reached statistical significance (p < 0.001). Similarly, the analysis of Classes III and IV versus the other two classes consistently confirms that low permeability can be distinguished from high values by predicted drug affinity to β-CD. In order to turn this qualitative conclusion into a practically useful formula, a second MARSplines model was formulated. However, the target of the modeling this time was the classification into low and high permeability cases. Hence, only one quantitative parameter was used for classification model formulation, namely, computed values of lnK. As a dependent value, the binary flag for permeability was declared. The obtained formulae are provided below:

ClassA = 0.1655 + 0.2510 × max(0; LnK – 4.8148) + 0.0734 × max(0; 4.8148 – LnK) – 0.2455 × max(0; LnK – 8.0157)  \( (2) \)

ClassB = 0.8345 – 0.2510 × max(0; LnK – 4.8148) – 0.0734 × max(0; 4.8148 – LnK) + 0.2455 × max(0; LnK – 8.0157)  \( (3) \)
was developed based on the MARSplines methodology, and molecular descriptors were derived from the SMILES code. The internal and external validation indicated good accuracy of the model. The appearance of polarity-related descriptors, such as XlogP, indicated the hydrophobic nature of the cyclodextrin cavity, which is consistent with the nature of cyclodextrins. It is well known that the hydrophilicity/hydrophobicity of a drug can be used for evaluation of the drugs’ permeability. Therefore, the model was used for predicting affinity to β-CD of exemplary compounds belonging to different BCS classes. As was established, APIs exhibiting high permeability (I and II BCS Class) are generally characterized by higher lnK values than compounds revealing low permeability (class III and IV). This shows that β-CD complexation seems to offer an alternative for complex and expensive experimental permeability modeling studies.

4. Conclusions

Compounds exhibiting high symmetry, such as fullerenes or nanotubes, have been used in various branches of medicine and pharmacy, including drug delivery [77–79]. This also applies to cyclodextrins, which due to their specific shape features, have been widely used to increase API solubility. In this work, the QSPr model of the binding constant of different compounds to β-CD was developed based on the MARSplines methodology, and molecular descriptors were derived from the SMILES code. The internal and external validation indicated good accuracy of the model. The appearance of polarity-related descriptors, such as XlogP, indicated the hydrophobic nature of the cyclodextrin cavity, which is consistent with the nature of cyclodextrins. It is well known that the hydrophilicity/hydrophobicity of a drug can be used for evaluation of the drugs’ permeability. Therefore, the model was used for predicting affinity to β-CD of exemplary compounds belonging to different BCS classes. As was established, APIs exhibiting high permeability (I and II BCS Class) are generally characterized by higher lnK values than compounds revealing low permeability (class III and IV). This shows that β-CD complexation seems to offer an alternative for complex and expensive experimental permeability modeling studies.

Supplementary Materials: The following are available online at http://www.mdpi.com/2473-8994/11/7/922/s1,
Table S1: Experimental and calculated LnK values, Table S2: LnK values predicted for compounds belonging to different classes according to Biopharmaceutical Classification System (BCS).

Author Contributions: Both authors contributed equally to the manuscript.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References


25. Macaev, F.; Boldescu, V. Cyclodextrins in asymmetric and stereospecific synthesis. Symmetry 2015, 7, 1699–1720. [CrossRef]


Symmetry 2019, 11, 922


56. QSAR Research Unit in Environmental Chemistry and Ecotoxicology. Available online: http://www.qsar.it/ (accessed on 1 June 2019).


Mol. Divers. 2013, 17, 421–434. [CrossRef]
70. Wen, X.; Tan, F.; Jing, Z.; Liu, Z. Preparation and study the 1:2 inclusion complex of carvedilol with β-cyclodextrin. 
Chem. Commun. 2002, 2, 2876–2877. [CrossRef]
74. Loftsson, T. Cyclodextrins and the biopharmaceutics classification system of drugs. 
75. Loftsson, T. Drug permeation through biomembranes: Cyclodextrins and the unstirred water layer. 
76. Dahan, A.; Wolk, O.; Kim, Y.H.; Ramachandran, C.; Crippen, G.M.; Takagi, T.; Bermejo, M.; Amidon, G.L. 
Purely in silico BCS classification: Science based quality standards for the world’s drugs. 
Mol. Pharm. 2013, 10, 4378–4390. [CrossRef]
Fullerenes Nanotub. Carbon Nanostruct. 2009, 17, 361–377. [CrossRef]
79. Szefler, B. Nanotechnology, from quantum mechanical calculations up to drug delivery. 
Int. J. Nanomed. 2018, 13, 6143–6176. [CrossRef]