PERKIN 2

A computational model to predict clathration of molecules with cephradine †

Gerardus J. Kemperman,[a] René de Gelder,[b] Ron Wehrens,[c] Frederik J. Dommerholt,[a] Antonius J. H. Klunder,[a] Lutgarde M. C. Buydens[c] and Binne Zwanenburg *[a]

[a] *Department of Organic Chemistry, NSR Institute for Molecular Structure, Design and Synthesis, University of Nijmegen, Toernooiveld 1, 6525, ED Nijmegen, The Netherlands*
[b] *Department of Inorganic Chemistry, NSR Institute for Molecular Structure, Design and Synthesis, University of Nijmegen, Toernooiveld 1, 6525, ED Nijmegen, The Netherlands*
[c] *Department of Analytical Chemistry, NSR Institute for Molecular Structure, Design and Synthesis, University of Nijmegen, Toernooiveld 1, 6525, ED Nijmegen, The Netherlands*

Received (in Cambridge, UK) 30th November 2000, Accepted 13th March 2001
First published as an Advance Article on the web 27th April 2001

The prediction of inclusion of molecules in the hosting framework of the cephalosporin antibiotic cephradine has been investigated. For this purpose linear discriminant analysis has been applied on molecular similarity data. The way the molecular similarity is calculated appeared to be extremely important. The charge distribution similarity of two molecules calculated at optimal shape overlap, is most appropriate to derive a computational model. The predictive power of the model increases when the molecular similarity of a molecule with respect to only a limited number of guiding compounds is used. The ultimate outcome is that complexing behaviour of a molecule can be predicted using a simple equation containing the similarity indices of that molecule with respect to three guiding compounds only. The model eventually obtained can predict the complexing behaviour of independent test sets of molecules with an average score of 86%.

Introduction

During the last decade the importance of computational chemistry and chemometrics has increased significantly. These techniques have had a substantial impact on the developments in several fields of chemistry. In particular, the field of medicinal chemistry has benefited immensely from the predictive power of techniques such as docking and quantitative structure activity relationships (QSAR).

Docking is especially useful for fitting molecules into cavities, such as active sites of biological targets. An essential requirement to perform a docking search is knowledge of the molecular structure of the binding site. Docking has been successfully employed for studying the interaction energies in protein–ligand complexes,[1] and also plays an important role in *de novo* drug design.[2]

Whereas docking makes use of the molecular structure of the binding site, QSAR ignores the binding site and focuses entirely on the molecular structure of the ligands. In contrast to docking QSAR only requires knowledge of the molecular structure of a series of active and non-active ligands for a given binding site. Starting with a set of known ligands QSAR utilizes various statistical methods to derive relationships between the molecular structure of the ligands and their affinity for the binding site. QSAR is based on the assumption that similar molecules exhibit similar properties.

This paper deals with the problem of identifying new molecules that form clathrates with the cephalosporin antibiotic cephradine. Clathration of cephalosporins is a valuable methodology for the isolation of these important antibiotics from aqueous solutions.[3,4] A drawback of the currently known complexants, which are all naphthalene derivatives, is their toxicity and the inherent environmental image problem associated with these compounds.[5,6] Hence, it is highly relevant from an industrial point of view to identify novel complexing agents having environmentally more acceptable properties. Moreover, complexants will be used in enzymatic cephalosporin synthesis. Due to possible enzyme inhibition by complexants, access to a variety of clathrating agents is desirable. Thus, a predictive model for identifying new complexants will be very helpful in addressing the above-mentioned problem. In these clathrates, the cephradine molecules form the hosting framework in which the aromatic compounds are hosted. In addition, a number of water molecules are accommodated in these complexes as well. Based on the crystal structure of the clathrate type complex of cephradine and 2-naphthol,[7] which is depicted in Fig. 1, a large series of novel complexing agents for the cephalosporin antibiotics could be identified, using chemical intuition as the main lead. However, due to the capricious behaviour of molecules regarding complex formation with cephradine, the successful selection of a new molecule for a complexation experiment remains a matter of trial and error. Therefore, it is desirable to

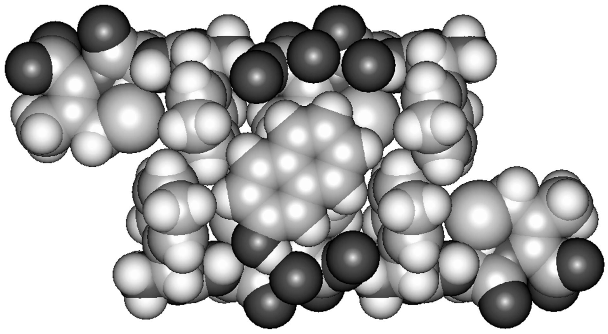

Fig. 1 The complex of cephradine and 2-naphthol.

† The IUPAC name for cephradine is 7-[D-2-amino-2-(cyclohexa-2,4-dienyl)acetamido]-3-methyl-8-oxo-1-aza-5-thiabicyclo[4.2.0]oct-2-ene-2-carboxylic acid.

DOI: 10.1039/b009629f

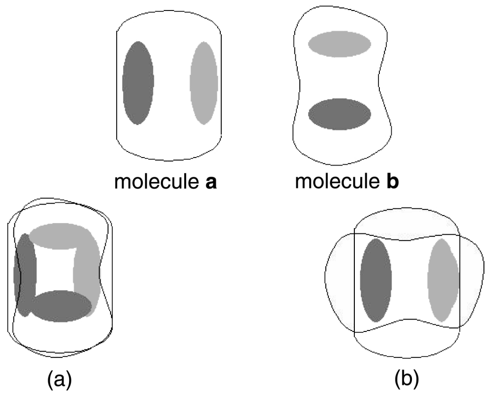

Fig. 2 Shape *versus* charge distribution similarity. The contours represent the shapes of molecules **a** and **b**, and the grey surfaces illustrate their charge distributions. (a) Optimal shape alignment; (b) optimal charge distribution alignment.

replace this empirical method by a more rational, preferably unbiased, design and prediction of new complexing agents.

At first sight docking seems to be an appropriate technique for fitting a guest molecule in the cavity formed by cephradine molecules. However, the hosting framework formed by cephradine is rather flexible and exhibits a remarkable adaptability.[6] In addition, the available space for the guest molecule is dependent on the number of water molecules incorporated in the complex.[6] As a consequence, an exact cavity for conceivable guest molecules cannot be defined, which is an essential requirement for the docking method. As QSAR ignores the structure of the binding site and only makes use of the structure of the guest molecules, this method is suitable to derive a model that can predict the complexing behaviour of a molecule. Application of QSAR for the prediction of clathration has no precedent except our preliminary report on this subject.[8]

Theory and strategy

QSAR is based on the assumption that similar molecules exhibit similar properties. The molecular property of interest may vary from the acidity of a carboxylic acid or the reactivity of a molecule in a certain reaction to the affinity for a biological target. This paper deals with the ability of a molecule to form a complex with cephradine. More difficult than defining the molecular property is to define and quantify molecular similarity. Basically, the question is how to represent molecules. This strongly depends on which property of a molecule is the object of study. As QSAR has not been used for the prediction of clathration previously, no information about suitable parameters could be derived from the literature. For the problem of fitting a molecule into a cavity formed by cephradine molecules, the shape and charge distribution of the molecule intuitively seem to be of prime importance. In addition, 19 physical properties were selected as global descriptors of the complexing agents in order to identify other molecular parameters that are of importance for complex formation with cephradine. It appeared, however, that from these 19 physical properties no predictive model could be derived by QSAR, neither did they contribute any essential information to the molecular similarity data.[8]

The molecular similarities and the physical properties of the guest molecules were calculated using the program Tsar.[10] A set of molecules was selected comprising both complexing agents and compounds that do not form complexes with cephradine. The molecular similarities of these molecules were calculated based on the shape and the charge distribution of the molecules. One way to do this is to optimise shape overlap of two molecules and subsequently calculate the shape similarity index in the orientation as depicted in Fig. 2(a). Similarly, the charge distribution overlap can be optimised followed by calculation of the charge similarity index as is depicted in Fig. 2(b). It has to be emphasized, however, that two molecules are most similar when both the shape and the charge similarity are simultaneously high. Therefore, the following procedure was followed for the calculation of the similarity matrices. Two molecules were first aligned with respect to their shapes as is depicted in Fig. 2(a). At optimal shape alignment their similarities with respect to both shape and charge were calculated. In this way two similarity matrices were obtained, one containing the shape similarities (Xshape) and the other containing the charge similarities at optimal shape overlap (Xcharge).[11] This strategy was preferred over calculating shape similarity at optimal shape overlap and charge similarity at optimal charge overlap as described above, which can result in misleading data as is explained in Fig. 2. Although molecules **a** and **b** are very different, calculation of their shape and charge distribution similarities as depicted in Fig. 2 would result in high similarity indices. However, when the charge distribution similarity is calculated at optimal shape alignment (situation (a) in Fig. 2) the charge similarity index will be much lower.

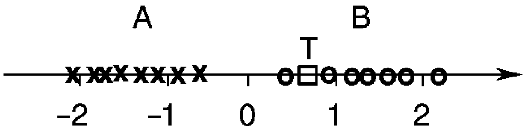

Fig. 3 Projection of objects on the discriminant line.

In order to derive a predicting model from the data consisting of molecular similarity matrices and physical properties, the supervised clustering technique "*linear discriminant analysis*" (LDA) was employed.[12,13] LDA requires that the class type or property of each object (molecule) is known. This condition is fulfilled in the present case. Classification by LDA works particularly well when the relationship between the variables of the objects and the corresponding class type is of linear nature. For optimal results, the data subjected to LDA have to fulfil two important requirements. They should be characterized by a normal distribution with each cluster having a different mean value and equal variance. Whether the second requirement is fulfilled by a data set consisting of a cluster of complexing molecules and a cluster of non-complexing molecules may be argued, since the latter can contain molecules which are much more diverse. However, the molecules selected for the data set all seem very similar from the point of view of chemical intuition.

The LDA algorithm is forced to search for those variables (properties) that are able to discriminate between the objects (complexing and non-complexing agents) in such a way that the desired clustering can be obtained. As such, it searches for a line (the discriminant line) in the multivariate space on which the objects are projected. Meanwhile, the algorithm minimizes the variance within a cluster and maximizes the variance between different clusters. As a result an optimal classification of the individual objects within the clusters is obtained. In case there are two clusters, only one discriminant function is required. The principle is illustrated in Fig. 3, where the two classes A and B have been projected on the discriminant line. When a new object is projected on the discriminant line, it is classified in cluster B if the discriminant function is positive, whereas it is classified in cluster A if the discriminant function is negative. In Fig. 3 it is shown that the test object T is classified in cluster B when projected on the discriminant line. Hence, the discriminant function can be utilized as a model for the prediction of a property of new objects.

The LDA models derived from the similarity matrices were validated by three methods, *i.e.* the LOOM test, the *p*-test and the test on an independent test set. The leave-one-out method (LOOM) removes one compound from the data set, subsequently a new model is calculated based on the remaining compounds and the removed molecule is predicted. This method yields an estimate of the prediction error of the derived model. However, as the columns with the similarities of the removed molecule remain in the matrices, this method may give a slightly positive biased result.

A second validation method is the permutation test (p-test). Also in this test one molecule is removed from the data set and a new model is derived based on the remaining molecules. However, now the desired clustering of molecules is chosen randomly, instead of demanding the complexing and non-complexing molecules to be clustered. The resulting model gives a fully arbitrary prediction of the removed molecule, which has a 50% chance of being correct. If in this case the prediction is much better, then the corresponding model predicts nonsense. The value of p can be regarded as the probability that a classification with such an error rate occurs by chance. Hence, a predictive score of 60% with a low p-value may be very significant, whereas a predictive score of 90% with $p = 0.2$ may be completely useless. In practice, a model for which the outcome of the p-test gives $p < 0.05$ is regarded as significant and $p < 0.01$ as very significant.

For the third validation the data are divided into a training set and a test set. The training set is used to derive a model. The predictive error of the resulting model on the independent test set is a measure of the predictive value of this model.

Results

The QSAR study started off by using a data set of 99 molecules, comprising 56 complexing agents and 43 non-complexing molecules. These 99 molecules were subjected to calculations of the molecular similarity matrices Xshape and Xcharge, whereby Xshape contains the shape similarities at optimal shape overlap and Xcharge contains the molecular similarities with respect to charge distribution at optimal shape overlap. The combined matrix of Xshape and Xcharge is named Xshch. Whereas unsupervised methods, such as principal component analysis and cluster analysis, are often successful in structure–activity applications, it appeared impossible to obtain separation of the complexing and non-complexing molecules by employing these methods on Xshape, Xcharge and Xshch.[8]

The results of the first LDA experiment, performed on the complete matrices Xcharge, Xshape and Xshch (**I** in Table 1), show that the LOOM predictions are only slightly above 50%. Moreover, from the p-tests it must be concluded that the models are not significant as in all cases $p > 0.05$. There are two possible explanations for these results. Either the relationship that is searched for (the relationship between the structure of a molecule and its complexing behaviour) is not present or the number of degrees of freedom is too large. The second explanation means that too many data are used for the modelling procedure. As a consequence, the LDA model is under-determined. If under-determination is the problem, this can be circumvented by using fewer columns to describe the data. The first approach to solve this problem was to reduce the amount of data by principal component analysis (PCA).[14] The models derived in this approach were based on 15 to 30 principal components, which explained 91.9 and 96.1% of the data, respectively. However, the results were disappointing. Hence, another approach was considered.

In a second attempt to solve the problem of under-determination, a set of 20 guiding compounds was chosen based on chemical intuition. Subsequently, the LDA model was derived based on the similarities of all the molecules with these guiding compounds. This means that the sizes of Xshape and Xcharge are reduced from 99 × 99 to 99 × 20 matrices. The results of this approach are also shown in Table 1 (**II**). From these results it can be concluded that the predictions based on Xcharge are very significant, with a reasonable correctness of 67.3%. The matrix Xshape seemingly contains no relevant information and confuses the information of Xcharge in the combined matrix Xshch.[11]

Table 1 The results of LDA on the complete data set (**I**) and on a data set with 20 guiding compounds (**II**)

	I		**II**	
	LOOM (%)	p-Test (p)	LOOM (%)	p-Test (p)
Xshape	53.5	0.160	55.5	0.376
Xcharge	55.5	0.056	67.3	0.000
Xshch	55.5	0.136	62.4	0.020

Once a reasonably successful method for the development of a model was found, the following step was to determine the predictive value of the model on a test set of independent molecules. From the results in Table 1 (**II**) it was concluded that the matrix Xcharge contains the most relevant information, hence Xshape and Xshch were ignored in the next experiments. First, 100 arbitrarily chosen divisions of the complete set of molecules in training and test sets were made. For all 100 training sets a model was derived by LDA, which was subsequently used to predict the test set. This resulted in an average correct prediction of 62.5%. Although this is slightly lower than the LOOM prediction based on Xcharge, it is still well above 50%.

The above-mentioned results seem promising, especially considering the fact that the set of guiding compounds used was not optimised but chosen on chemical intuition only. Perhaps even better results can be obtained with an optimised set of guiding compounds. In order to optimise the set of guiding compounds a genetic algorithm was used. Genetic algorithms (GA's) have been applied to solve complex optimisation problems in several fields, including chemistry.[15,16,17] The principle of a genetic algorithm is easy to understand. It is based on the evolutionary phenomena found in nature, such as 'reproduction', resulting in species with chromosomes formed by combination of those from the parents, and selection by the 'survival of the fittest' criterion. A GA is initiated by generating a population of trial solutions for the subsequent problem. Each individual of the population is characterized by a chromosome. In the present case the individual is a set of guiding compounds with a 'chromosome' consisting of a string of compound ID's corresponding to the column numbers of the whole set of molecules (see Table 2). All individuals are subjected to LDA modelling and the resulting models are evaluated by LOOM. After the evaluation, the individuals are ranked by their LOOM scores. Next the best individuals are selected and subsequently reproduced. The new individuals are then subjected to LDA and evaluated by LOOM, followed by the above described sequence until the termination criterion is reached.

To assess whether even smaller sets of guiding compounds could be used to predict complexation, the GA was allowed to select a maximum number of either 10 or 20 guiding compounds. Again, only the columns of the data set Xcharge were used, which after the addition of 21 molecules contained 120 molecules in total resulting in a 120 × 120 matrix (see Table 2; the extra 21 compounds were predicted by the first models and then experimentally tested). The LOOM error was used as an evaluation function. The complete data matrix was divided into five combinations of training sets with 90 compounds and test sets with 30 compounds. For each of the five training set–test set combinations, five GA runs were performed each with a different initial population. Table 3 shows both the variation in the size of the guiding sets found by the GA and the variation in the LOOM results. The best LOOM results are approximately 20% higher than those obtained with the non-optimised guiding set. In fact, even the lowest LOOM scores in Table 3 are much higher than those in Table 1 (**II**). Furthermore, the models described in Table 3 are highly significant according to the permutation tests performed with 250 permutations. In all but one case the permutation test gave $p = 0.000$ and in one case gave $p = 0.004$, which still points to a very significant model. Remarkably, although all guiding sets selected exhibit similar performance, their compositions are very different. Apparently,

Table 2 The data set used to derive the QSAR models

Row ID	Name	Complex formation
1	1,2,3-Trihydroxybenzene	Yes
2	1,3,5-Trihydroxybenzene	Yes
3	1,2-Dihydroxybenzene	Yes
4	1-Chloronaphthalene	Yes
5	1-Naphthol	Yes
6	2-Aminobenzoic acid	Yes
7	2-Aminophenol	Yes
8	2-Hydroxyacetophenone	Yes
9	2-Hydroxybenzoic acid methyl ester	Yes
10	2-Methoxyacetophenone	Yes
11	2-Naphthol	Yes
12	3-Aminophenol	Yes
13	3-Hydroxybenzoic acid	Yes
14	4-Hydroxybenzoic acid	Yes
15	4-Methylbenzoic acid methyl ester	Yes
16	5-Methoxy-1*H*-indole	Yes
17	Anisole	Yes
18	Benzamide	Yes
19	Benzoic acid	Yes
20	Bipyridine	Yes
21	1*H*-Indene	Yes
22	1*H*-Indole	Yes
23	Quinoline	Yes
24	Salicylic acid	Yes
25	1,2,3,4-Tetrahydro-1-naphthol	Yes
26	Toluene	Yes
27	Vanillin	Yes
28	1,3-Dihydroxybenzene	Yes
29	1,4-Dihydroxybenzene	Yes
30	Indan-1-one	Yes
31	3-Hydroxyacetophenone	Yes
32	4-Aminoacetophenone	Yes
33	4-Methoxyphenol	Yes
34	4-Methylacetophenone	Yes
35	Acetophenone	Yes
36	Phenol	Yes
37	4-Phenylphenol	Yes
38	9*H*-Carbazole	Yes
39	9*H*-Fluorene	Yes
40	4-Aminobenzoic acid ethyl ester	Yes
41	4-Aminobenzoic acid methyl ester	Yes
42	3-Hydroxybenzoic acid methyl ester	Yes
43	4-Hydroxybenzoic acid methyl ester	Yes
44	1-Nitronaphthalene	No
45	4-Aminobenzoic acid	No
46	4-Aminophenol	No
47	4-Hydroxybenzaldehyde	No
48	4-Hydroxypropiophenone	No
49	4-Methylbenzoic acid	No
50	4-Nitrophenol	No
51	Benzaldehyde	No
52	Benzoic acid glycol ester	No
53	Benzoic acid propanglycol ester	No
54	Benzyl alcohol	No
55	*N*,*N*-Dimethylbenzamide	No
56	Propiophenone	No
57	2-Bromonaphthalene	No
58	2-Nitronaphthalene	No
59	*o*-Acetyloxybenzoic acid	No
60	Naphthalene-1-carboxylic acid	No
61	Nitrobenzene	No
62	1-Phenylethanol	No
63	2-Methylcyclohexanone	No
64	3-Hydroxybenzyl alcohol	No
65	4-Hydroxyacetophenone	No
66	4-Hydroxybenzoic acid ethyl ester	No
67	4-Methoxyacetophenone	No
68	Anthracene	No
69	Anthraquinone	No
70	Benzoin	No
71	Cyclohexanecarboxylic acid	No
72	Cyclohexanol	No
73	Isophthalic acid	No
74	Cinnamic acid	No
75	Terephthalic acid	No
76	1,2,3,4-Tetrahydronaphthalene	No
77	α-Tetralone[a]	No

Table 2 The data set used to derive the QSAR models

Row ID	Name	Complex formation
78	Trimesic acid[b]	No
79	2,4-Dihydroxybenzoic acid	Yes
80	2,4-Dihydroxybenzoic acid methyl ester	Yes
81	2-Aminobenzamide	Yes
82	2-Aminobenzoic acid methyl ester	Yes
83	2-Methylbenzoic acid	Yes
84	3,4,5-Trihydroxybenzoic acid methyl ester	Yes
85	3,4-Dihydroxybenzoic acid	Yes
86	3,4-Dihydroxybenzoic acid methyl ester	No
87	3,5-Dihydroxybenzoic acid	Yes
88	3,5-Dihydroxybenzoic acid methyl ester	Yes
89	3-Aminobenzoic acid	No
90	3-Aminobenzoic acid methyl ester	Yes
91	3-Methylbenzoic acid	Yes
92	3-Methylbenzoic acid methyl ester	Yes
93	Aniline	Yes
94	1,2-Diaminobenzene	No
95	1,3-Diaminobenzene	No
96	1,4-Diaminobenzene	No
97	2-Aminopyridine	No
98	2-Hydroxypyridine	No
99	4-Aminopyridine	No
100	1,2,7-Trihydroxynaphthalene	Yes
101	1,6-Dihydroxynaphthalene	Yes
102	1-Acetonaphthone	Yes
103	2,4-Dihydroxy-6-methylpyrimidine	No
104	Indan-2-ol	No
105	2-Methoxybenzaldehyde	Yes
106	2-Methoxybenzoic acid	Yes
107	4-Chlorobenzoic acid	No
108	4-Hydroxyphenylacetic acid	No
109	6-Methoxy-1*H*-indole	Yes
110	Benzylamine	No
111	Coumarin	Yes
112	Phthalic acid	No
113	*m*-Anisidine	Yes
114	Methyl benzoate	Yes
115	*N*,*N*-Dimethyl-4-nitroaniline	No
116	4-Fluorobenzoic acid	No
117	*o*-Vanillin[c]	No
118	*p*-Cresol	Yes
119	1-Naphthylamine	Yes
120	2-(*N*,*N*-Dimethylamino)benzoic acid	Yes

[a] The IUPAC name for α-tetralone is 3,4-dihydronaphthalen-1(2*H*)-one.[b] The IUPAC name for trimesic acid is benzene-1,3,5-tricarboxylic acid.[c] The IUPAC name for *o*-vanillin is 2-hydroxy-3-methoxybenzaldehyde.

it is not important which individual molecules are in the guiding set. This is due to the fact that most molecules are highly similar to at least one or more other molecules within the complete set of 120 molecules. It is, however, of importance which combination of guiding molecules is selected for LDA modelling. As can be concluded from Table 3, there are more combinations comprising different molecules but exhibiting a comparable performance during LDA modelling.

By comparing the results in Table 1 with those in Table 3, it is clear that guiding set selection by a GA has improved both the LOOM and test set predictions significantly. However, the LOOM results in Table 3 are still much better than the test set results, which triggered doubts about the robustness of the model. Possibly too many guiding compounds were allowed in the GA optimisation. This has the consequence that the guiding set is specifically optimised to predict the training set correctly, while it should be trained for a broad range of molecules. By allowing a lower number of guiding compounds, the GA has to select those molecules that are most important for the prediction of the training set. Hence, a less specialized guiding set is obtained and prediction of a broader range of molecules can be achieved.

The use of smaller guiding sets was investigated using the same 120 × 120 Xcharge matrix. This matrix was again divided into training sets consisting of 90 compounds and independent test sets of 30 compounds. The GA was used to select guiding sets containing a maximum of four compounds. Five GA runs were performed for each training set–test set division. It was found that the LOOM predictions were generally worse than those in Table 3, obtained by using larger guiding sets, but the test set predictions were better and moreover much more in agreement with the LOOM predictions. All models were evaluated by a permutation test which in all cases gave a *p*-value of 0.000. A graph of which compounds were selected as guiding compounds showed that compounds 113, 86 and 83 were prominent, as is depicted in Fig. 4.

It appeared that a prediction based on only these three guiding compounds led to prediction results of 80.0, 93.3, 86.7, 86.7, and 83.3% for the five test sets respectively (86% on average). The permutation test gave a *p*-value of 0.000 thus indicating that the model is very significant. From a chemical point of view, it is difficult to see why the GA selected these three molecules. A plot of the differences in class mean between the complexing agents and the non-complexing agents for each of the 120 compounds in the data set does provide some insight. This plot is depicted in Fig. 5. Evidently, molecules 86 and 113 belong to the seven compounds with the largest class mean difference. Despite the fact that molecule 83 is not among these seven compounds it is still selected by the LDA algorithm, this in contrast to several other molecules that have much larger class mean differences. This is because the LDA algorithm

Table 3 The results obtained by using the GA optimised guiding sets

GA run	No. selected[a]	LOOM (%)[b]	Test set (%)[c]
Maximum 10 guiding compounds	4–10	83.3–92.2	71.4
Maximum 20 guiding compounds	8–15	87.8–94.4	70.3

[a] The variation in the number of guiding compounds selected in 5 GA runs. [b] The variation in the LOOM results of the guiding sets resulting from 5 GA runs. [c] The average prediction result from 5 independent test sets.

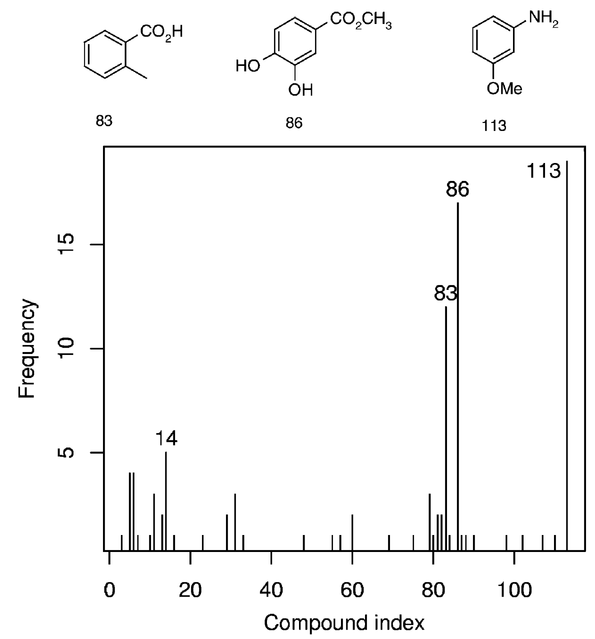

Fig. 4 The frequency by which the molecules were selected as a guiding compound.

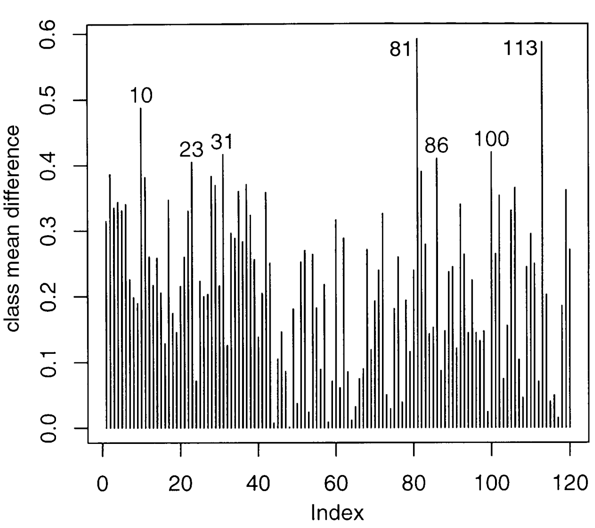

Fig. 5 A plot of the class mean difference for the 120 compounds of Xcharge.

selects those compounds that contain complementary information, which is apparently the case for compounds 83, 86 and 113.

As complexation can be predicted by using the similarities of the molecule with only three guiding compounds, the equation for calculating the similarity indices is very concise, *viz.* eqn. (1). After computation of the similarity indices of a new molecule with compounds 83, 86 and 113 (S83, S86 and S113 respectively), complexation can be predicted with a certainty of approximately 86% *via* eqn. (1). Table 4 illustrates how the model can be used to predict the complexing behaviour of molecules. From the similarity indices S83, S86 and S113, the discriminant function is calculated according to eqn. (1). The outcome of this prediction has an 86% chance of being in agreement with the experimental outcome, which is shown in the last column.

$$D = -0.35 \times S83 - 0.44 \times S86 - 0.83 \times S113 + 0.36 \quad (1)$$

$$D > 0 \text{ no complexation}$$
$$D < 0 \text{ complexation}$$

Conclusion

A model to predict clathrate formation of molecules with the cephalosporin antibiotic cephradine has been derived. For this purpose, linear discriminant analysis (LDA) was employed on molecular similarity data of a set of molecules comprising both complexing agents and molecules that do not form a complex with cephradine. The success of this method strongly depends on how the molecular similarity indices are calculated. Furthermore, the amount of similarity data subjected to LDA should not be too large as this may lead to under-determination of the model. This problem can be avoided by using the similarities of the molecules of the data set with a limited number of guiding compounds only. The ultimate result of this study is a simple equation to predict whether a compound will be able to form a clathrate with cephradine or not. The procedure developed may also be applicable for other molecular clathration problems, or more generally for other problems in host–guest chemistry. The method described in this paper requires that for the host–guest system under investigation a series of suitable guest molecules are known. Such a series of suitable guest molecules can be found by trial and error experimentation. In case the structure of the host–guest complex is known, docking in the hosting cavity can be a useful tool to assist the search for suitable guest molecules. When a series of suitable guest molecules has been discovered, the following procedure may lead to the prediction of a model equation.

1. *Composition of a data set.* Select a series of molecules comprising both suitable guests and molecules that do not form a complex with the host under investigation.

2. *Calculation of the similarity matrix.* Calculate a molecular similarity matrix using an appropriate software package such as Tsar.[10] Optimise the shape overlap of the molecules, followed by calculation of their charge distribution similarity index in the orientation found.

3. *Guiding set optimisation and deduction of a model.* Initialise a genetic algorithm (GA) by making a random selection of 'chromosomes' containing compound ID's corresponding to the similarity matrix calculated in step 2. Start the GA cycle, which consists of the following steps. First, models based on the molecular similarities of the molecules in the data set with the molecules of the selected guiding sets are derived using linear discriminant analysis (LDA) and subsequently evaluated by LOOM. The guiding sets of the best performing models are reproduced for the next cycle, which is repeated until a desired LOOM result is obtained or after the maximum allowed number of cycles has been reached.

4. *Evaluation of the model using a test set.* Select molecules for an independent test set, which is predicted by the model and tested experimentally for complexation with the host under investigation. In case the complete data set is sufficiently large, it can also be divided in a training set and test set prior to the GA optimisation (step 3). The model can then be evaluated by the corresponding test set.

Table 4 Prediction of complexation for 15 arbitrarily chosen compounds

ID[a]	Name	S83[b]	S86[b]	S113[b]	*D*[c]	Model[d]	Exp.[e]
5	1-Naphthol	0.400	−0.132	0.400	−0.054	y	y
12	3-Aminophenol	0.313	0.041	0.400	−0.099	y	y
19	Benzoic acid	0.535	0.208	0.2000	−0.085	y	y
25	1,2,3,4-Tetrahydro-1-naphthol	0.399	0.019	0.310	−0.046	y	y
45	4-Aminobenzoic acid	0.413	−0.024	0.205	0.056	n	n
57	2-Bromonaphthalene	0.138	0.013	0.338	0.025	n	n
61	Nitrobenzene	0.618	−0.052	−0.232	0.359	n	n
71	Cyclohexanecarboxylic acid	0.401	0.481	0.187	−0.148	y[f]	n
74	Cinnamic acid	0.214	0.057	0.139	0.145	n	n
81	2-Aminobenzamide	0.587	0.461	0.179	−0.197	y	y
88	Methyl 3,5-dihydroxybenzoate	−0.291	−0.244	0.143	0.236	n[f]	y
95	1,3-Diaminobenzene	0.042	−0.148	0.457	0.031	n	n
101	1,6-Dihydroxynaphthalene	0.235	−0.098	0.472	−0.071	y	y
108	4-Hydroxyphenylacetic acid	−0.205	−0.238	0.327	0.122	n	n
119	1-Naphthylamine	0.257	0.086	0.519	−0.199	y	y

[a] The ID refers to the row ID of the molecule in the matrix (see Table 2). [b] The similarity indices are rounded to three decimal places. [c] *D* is the discriminant parameter calculated according to eqn. (1). [d] Prediction of the model, y means clathrate formation, while n means no clathrate formation. [e] Experimental outcome, y means clathrate formation, while n means no clathrate formation. [f] Incorrect prediction.

Experimental

Calculation of data matrices for QSAR

Generation of the 3-dimensional molecular structures for the data set was achieved by first drawing them using the program SYBYL and subsequent energy minimization of the structure by molecular mechanics using Gasteiger–Hückel charges.[9] After that, the structures were exported as.mol2 files which can be imported in the program Tsar.[10] Prior to calculation of the similarity matrices, for each structure a second energy minimization was performed using the semi-empirical method included in Tsar (the VAMP module). The similarity matrices were calculated using the program Tsar.[10] Similarity indices have been calculated as follows. Initial alignment of two molecules was made by overlaying their centres of mass. Alignment was then optimised based on shape overlap. The optimisation consisted of two steps. First a full rigid search was performed using angular increments of 18°. Next a simplex optimisation was applied to fine-tune the optimal alignment. At optimal alignment both the shape similarity index and the charge similarity index were calculated. The resulting matrices were Xshape, containing the shape similarities at optimal shape overlap, and Xcharge, containing the charge similarities at optimal shape overlap. The similarity indices were calculated as is shown below.

Also the charge similarity and shape similarity at optimal charge distribution overlap have been calculated.[11] The resulting matrices revealed virtually the same information as Xshape and Xcharge according to principal component analysis, cluster analysis and linear discriminant analysis.[8] Hence, only Xshape and Xcharge were discussed in this paper to avoid confusion.

QSAR modelling

Statistical calculations were performed using R[18] version 0.63 on a Linux Pentium II machine (266 MHz). Genetic algorithms for guiding compound selection were performed on SUN workstations using the PGAPack library.[19]

References

1 H. J. Böhm and G. Klebe, *Angew. Chem., Int. Ed. Engl.*, 1996, **35**, 2588; T. Lengauer and M. Rarey, *Curr. Opin. Struct. Biol.*, 1996, **6**, 402; B. K. Shoichet and I. D. Kuntz, *Chem. Biol.*, 1996, **3**, 151; B. Kramer, G. Metz, M. Rarey and T. Lengauer, *Med. Chem. Res.*, 1999, **9**, 463.

2 I. D. Kuntz, E. C. Meng and B. K. Shoichet, *Acc. Chem. Res.*, 1994, **27**, 117; Y. Yamamoto, Y. Ishihara and I. D. Kuntz, *J. Med. Chem.*, 1994, **37**, 3141; P. D. J. Grootenhuis, D. C. Roe, P. A. Kollman and I. D. Kuntz, *J. Comput. Aided Mol. Des.*, 1994, **8**, 731; N. C. J. Strynadka, M. Eisenstein, E. Katchalski-Katzir, B. K. Stoichet, I. D. Kuntz, R. Abagyan, M. Totrov, J. Janin, J. Cherfils, F. Zimmerman, A. Olson, B. Duncan, M. Rao, R. Jackson, M. Sternberg and M. N. G. James, *Nat., Struct. Biol.*, 1996, **3**, 233; C. L. M. J. Verlinde and W. G. J. Hol, *Structure*, 1994, **2**, 577.

3 Novo Nordisk, Denmark, USP 4003896 (*Chem. Abstr.*, 1977, **86**, 171490m).

4 Gist-brocades, The Netherlands, WO 93/12250, EP 618979, USP 5470717 (*Chem. Abstr.*, 1993, **119**, 137533w); A. Bruggink, E. C. Roos and E. de Vroom, *Org. Process Res. Dev.*, 1998, **2**, 128; A. Bruggink, *Chimia*, 1996, **50**, 431.

5 G. J. Kemperman, R. de Gelder and P. C. Raemaker-Franken, NL 1007828, 1997; WO 99/31109.

6 G. J. Kemperman, F. J. Dommerholt, R. de Gelder, P. C. Raemakers, A. J. H. Klunder and B. Zwanenburg, *J. Chem. Soc., Perkin Trans. 2*, 2000, 1425.

7 G. J. Kemperman, R. de Gelder, F. J. Dommerholt, P.C. Raemakers-Franken, A. J. H. Klunder and B. Zwanenburg, *Chem. Eur. J.*, 1999, **5**, 2163.

8 R. Wehrens, R. de Gelder, G. J. Kemperman, B. Zwanenburg and L. M. C. Buydens, *Anal. Chim. Acta*, 1999, **400**, 413.

9 A program of Tripos Associated Inc. for performing modelling studies and calculations on molecules. The program suite SYBYL consists of a number of computational chemistry modules to describe and predict molecular behaviour.

10 Tsar version 3.2, Oxford Molecular Group PLC, Oxford, UK.

11 *Note*: Besides the similarity matrices Xshape and Xcharge also two other matrices have been calculated. These are the charge distribution similarity at optimal charge distribution overlap and the shape similarity at optimal charge distribution overlap. In this case the shape similarity matrix has the same penalty function as Xcharge in this paper. According to PCA, cluster analysis and LDA these two matrices contain virtually the same information as Xshape and Xcharge. Hence, these results are not discussed explicitly in this paper.

12 D. Livingstone, *Data analysis for chemists*, Oxford University Press, London, 1995, p. 139.

13 W. H. A. M. van den Broek, D. Wiencke, W. J. Melssen, R. Feldhoff, T. Huth-Fehre, T. Kantimm and L. C. M. Buydens, *Appl. Spectrosc.*, 1997, **51**, 856; G. Harris, N. C. Andreasen, T. Cizadlo, J. M. Bailey, H. J. Bockholt, V. A. Magnotta and S. Arndt, *J. Comput. Assisted Tomogr.*, 1999, **23**, 144.

14 S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37.

15 R. Wehrens and L. C. M. Buydens, *Trends Anal. Chem.*, 1998, **17**, 193.

16 D. E. Goldberg, *Genetic algorithms in search, optimisation and machine learning*, Addison-Wesley, Wokingham, 1989.

17 *Handbook of genetic algorithms*, ed. L. Davis, Van Nostrand Reinhold, New York, NY, 1991.

18 R: A Language for data analysis and graphics, R. Ihaka and R. Gentleman, *J. Comput. Graphic. Statist.*, 1996, **5**, 299. The main R site is: http://www.r-project.org.

19 This library is available from ftp://mcs.ano.gov/pub/pgapack/pgapack.tar.Z.