Quantitative Structure−Property Relationship Estimation of Cation Binding Affinity of the Common Amino Acids

Jesús Jover, Ramón Bosque, and Joaquim Sales*

Departament de Química Inorgànica, Universitat de Barcelona, Martí i Franquès, 1, 08028-Barcelona, Spain

Received: November 26, 2008; Revised Manuscript Received: February 10, 2009

The quantitative structure−property relationship (QSPR) methodology is applied to estimate the binding affinity of lithium, sodium, potassium, copper, and silver cations to the 20 common amino acids. The proposed model, nonlinearly derived from computational neural networks (CNN), contains seven descriptors and was validated by an external prediction set. Good results are obtained with correlation coefficients, R^2, and root-mean-square errors (rms) (kJ/mol) of 0.998 (3.89), 0.999 (2.86), and 0.997 (3.90) for the training, prediction, and validation sets, respectively. Five of the descriptors of the model correspond to the amino acids and the other two to the cations; they encode information clearly related to the cation−amino acid interactions responsible for the binding affinity values analyzed. A detailed analysis of results shows that, despite the different nature of the bonding between the metal cations and the amino acids, the neural networks used are capable of predicting accurately the property studied.

1. Introduction

The complexation of metal cations to amino acids is one of the most important processes in bioinorganic chemistry; for instance, all metalloenzymes present at least an amino acid coordinated to a metal cation. Consequently, great amounts of experimental and theoretical work have been devoted to the determination of the binding affinity, cation basicity, and dissociation energy involved in these reactions.[1]

Several experimental methods to obtain affinities and cation basicities are known, the kinetic method being the most widely used.[2,3] This method gives only relative values, so to obtain absolute values it must be calibrated with known affinities or basicities of one reference compound, and depending on the accuracy of the anchor value selected, the absolute values can be significantly different. According to the literature, this method tends to underestimate binding energies.[4] Another widely used technique is threshold collision-induced dissociation (TCID).[5] This method allows the determination of dissociation energies at 0 K, although by applying thermal corrections the absolute bond dissociation enthalpies at 298 K can be obtained.[6] A third technique is the measurement of ligand exchange equilibrium in trapping devices.[6] This technique allows the determination of Gibbs free energies corresponding to the ligand exchange reactions. The use of anchor values, as well as the subtraction of the $T\Delta S$ terms, leads to the determination of bond dissociation enthalpies at 298 K.[7] The continuous improvement in both hardware and software has made possible the enhancement of the accuracy with which theoretical calculations can be performed on these systems. In fact, they have been frequently used to transform the experimental relative cation basicities and affinities into absolute values. However, the number of amino acid−ion affinities and basicities calculated is not very large. A great portion of the published work deals only with the simplest amino acid, glycine, and to a lesser extent alanine; some of the ions that have been studied are those from the alkaline and alkaline earth groups, and some of the first transition series as well as silver derivatives.[8−10] Moreover, it is not easy to compare the results for the calculations from different authors, because of the differences in the level of calculation (HF, DFT with different functionals or MP2), basis set used, and treatment of the basis set superposition error.

Recently, we have calculated the binding affinity (BA) of the 20 common amino acids and the cations Li^+, Na^+, K^+, Cu^+, and Ag^+, defined as the increment of enthalpy at 298 K for the dissociation of the amino acid−cation complexes, $[M(AA)]^+$, into the free amino acid and the cation. To obtain comparable and reliable values of this magnitude, we have used rigorous and high level computational methods based on DFT. A complete and homogeneous set of BA values of the interaction of the 20 common amino acids with the five mentioned cations is reported.[11]

The quantitative structure−activity relationship approach (QSPR) has become a very useful tool in the prediction and interpretation of several physical and chemical properties of families of compounds. The basis of such a relationship is the assumption that the variation of behavior of the compounds, as expressed by any measured property, can be correlated with changes in molecular features of the compounds termed descriptors. Descriptors are numerical values used to describe different characteristics of a certain structure to yield information about the studied property. QSPR methods are based on statistically linear and nonlinear functional forms that relate the property with descriptors. Its development involves the selection of descriptors to satisfactorily characterize the sets of compounds and the application of algorithms, such as multiple linear regression or computational neural networks (CNN) to build the QSPR model. The advantage of CNN is their inherent ability to incorporate nonlinear relationships in the derivation of the QSPR models. Both types of QSPR approaches, linear and nonlinear, have been applied to the correlation of many diverse physicochemical properties of chemical substances.[12] More recently, QSPR methodology has been successfully applied to more complex systems, the so-called multicomponent systems, where the property studied depends not only on the chemical nature of the compound, but also on the medium or the experimental conditions, such as temperature, pressure, solvent, etc. In these cases, besides the habitual molecular descriptors of

* Corresponding author. Fax: +34934907725. E-mail: joaquim.sales@ub.edu.

10.1021/jp810391z CCC: $40.75
Published on Web 03/19/2009

the compounds studied, other descriptors associated with the other components of the system have to be used; due to the fact that the studied property depends on all of these conditions, CNN methods turned out to be more appropriate for the derivation of the models. The published QSPR studies on multicomponent systems refer, mainly, to physical properties (density, viscosity, vapor pressure) of hydrocarbons as a function of the temperature,[13−15] to the estimation of the solvation free energy of organic compounds in different solvents,[16] and also to the estimation of pK_a of several families of organic compounds (phenols, benzoic and carboxylic acids, anilines) in different solvents.[17−19]

The QSPR methodology has been scarcely applied in the estimation of cation basicities. Two papers have appeared on the prediction of lithium basicities of a wide set of organic compounds.[20,21] Linear and nonlinear methods have been used, and the predicted values are good with $R^2 \approx 0.8-0.9$, and rms errors of around 5−8 kJ/mol. Recently, the cation affinities of the 20 common amino acids with H^+, Na^+, Cu^+, and Ag^+ cations have been analyzed by QSPR methods.[22] Using molecular descriptors of the amino acids, several models were derived by multiple linear regressions, partial least-squares analysis, and artificial neural networks. For each cation studied, the authors propose a different model, containing two or three molecular descriptors of the amino acids; however, no intent has been made to derive a single model useful for the prediction of the BA for the four cations simultaneously.

In this Article, we apply the QSPR methodology to derive a model capable of estimating simultaneously the binding affinity of the Li^+, Na^+, K^+, Cu^+, and Ag^+ cations to the 20 common amino acids. The values of this property depend clearly on the nature of both components, amino acids and cations; thus it must be considered as a multicomponent system, and CNN methods are used to derive the corresponding model. The set of descriptors used to characterize the $[M(AA)]^+$ complexes studied is composed of the habitual molecular descriptors of the amino acid and of several physicochemical properties, such as atomic number, ionization energy, electron affinity, radii, electronegativity, etc., of the cations.

2. Methodology

2.1. Data Set. The data set comprises 100 values of binding affinity corresponding to the interaction, in gaseous phase, of the 20 common amino acids with the cations Li^+, Na^+, K^+, Cu^+, and Ag^+. These values are those calculated previously using high level DFT methods,[11] and, accordingly, they will be called DFT calculated values. These BA values spread over a range of 365 kJ/mol, from 115.6 kJ/mol for the $[K(Gly)]^+$ to 479.8 kJ/mol for the $[Cu(Arg)]^+$ complex, the mean value being 242.4 kJ/mol. Table 1 contains all of the DFT calculated and QSPR estimated BA values.

This full set was divided in three subsets, training, prediction, and validation sets, containing 70, 20, and 10 BA values, respectively. The selection of these three subsets was done to guarantee that they all contain values of the different amino acids and cations in the same proportion as the full set; for example, each amino acid is used once in the prediction set, and also four values of each cation are included in this subset. The training set is used exclusively to derive the model; the prediction set, formed by BA values that were not employed for the model development, is used to test the predictive capacity of the model; and the validation set is used to train the neural networks (see below), to avoid its overtraining and to make sure that they have good and general predictive capacity.

2.2. Amino Acid Descriptors. The structural descriptors of the amino acids were calculated using the CODESSA program.[23] In our previous study,[11] we performed a systematic search of the amino acid conformers using molecular mechanics and HF calculations. The lowest energy conformers were reoptimized with the Gaussian 03 package, at the B3LYP/DZVP level, to find the most stable conformation, which was used therefore to derive the binding affinity values. The Gaussian 03 output files were sent to the CODESSA program to calculate several hundreds of molecular descriptors, which can be classified in six general classes: constitutional, topological, geometrical, electrostatic, quantum-chemical, and charge partial surface area (CPSA). To encode the effect of the side chains of the amino acids molecules in the coordination to the metals, we have added two descriptors to the descriptor pool, the number of heteroatoms present and the polarizability of these side chains, which has been calculated using an additive method.[24]

2.3. Cation Descriptors. To characterize the cation metals, we have employed several physicochemical properties such as atomic number, atomic weight, ionization energy, electron affinity, several electronegativity scales (Pauling, Mulliken, Sanderson, Allred−Rochow), hardness and softness, cation radii, and charge/radius ratios, etc. Their values have been taken mainly from Emsley's handbook.[25] A total of 31 descriptors were considered for each cation, and they were imported to the CODESSA program.

The heuristic routines of CODESSA were used to make the first reduction in the pool of descriptors with the compounds forming the training set. This process eliminates all of the incomplete and invariant descriptors, as well as the ones with the correlation coefficient below 0.01. Descriptors correlated (above 0.95) with another descriptor, which has a higher regression coefficient, were also deleted. Following these procedures, the initial pool of 360 descriptors formed by the molecular descriptors of the amino acids and the specific descriptors of the metals was reduced to 164.

2.4. CNN Methods. The CNN treatment was performed with the ADAPT (Automated Data Analysis and Pattern recognition Toolkit) program,[26,27] including feature selection routines (genetic algorithm[28] and simulated annealing[29]) and CNN procedures.[30] The CNNs used are three-layer, fully connected, feed-forward networks and have been described in detail by Jurs et al.[30,31] The number of neurons of the input layer corresponds to the number of descriptors in the model. The number of hidden neurons controls the flexibility of the network and has to be adjusted until the optimal network architecture is achieved. The optimization of the number of neurons in these two first layers is done by means of a building up procedure that consists of starting with a low number of neurons and increasing it one unit until the results achieved with the new architecture are not better than those obtained with the previous one. The output layer contains one neuron representing the BA value. In this case, the best architecture found was 7−5−1.

The descriptors retained after the heuristic procedures of CODESSA were imported to ADAPT, where they were subjected to the objective feature selection routines of this program for the compounds in the training set. In this case, only identical test and intercorrelation of descriptors are taken into account, both with a cutoff value of 0.9. The new reduced pools contain 79 descriptors (10 of the cations), which were used as the starting point in the nonlinear selection of the models.

Fully CNNs were developed using an automatic genetic algorithm descriptor selection routine with a CNN for evaluating the fitness of each subset of descriptors selected. The fitness of descriptor subsets was calculated as COST = TSET + 0.4|TSET

TABLE 1: DFT Calculated and QSPR Estimated BA (kJ/mol) Values for the Training, Prediction, and Validation Sets

no.	complex	DFT calculated	QSPR estimated
1	[Ag(Ala)]$^{+a}$	212.8	214.8
2	[Ag(Asp)]$^{+a}$	232.4	233.0
3	[Ag(Gln)]$^{+a}$	256.7	262.6
4	[Ag(Glu)]$^{+a}$	234.1	237.0
5	[Ag(Gly)]$^{+a}$	206.1	205.6
6	[Ag(His)]$^{+a}$	281.7	281.3
7	[Ag(Ile)]$^{+a}$	219.0	214.2
8	[Ag(Lys)]$^{+a}$	290.9	288.5
9	[Ag(Phe)]$^{+a}$	236.2	231.8
10	[Ag(Pro)]$^{+a}$	230.3	234.9
11	[Ag(Ser)]$^{+a}$	224.9	226.0
12	[Ag(Thr)]$^{+a}$	227.8	224.2
13	[Ag(Trp)]$^{+a}$	260.0	257.3
14	[Ag(Val)]$^{+a}$	216.5	215.7
15	[Ag(Asn)]$^{+c}$	245.8	250.3
16	[Ag(Met)]$^{+c}$	260.0	256.9
17	[Ag(Arg)]$^{+b}$	316.0	317.9
18	[Ag(Cys)]$^{+b}$	228.2	226.3
19	[Ag(Leu)]$^{+b}$	221.1	222.7
20	[Ag(Tyr)]$^{+b}$	240.4	240.0
21	[Cu(Arg)]$^{+a}$	479.8	477.1
22	[Cu(Asn)]$^{+a}$	348.7	350.1
23	[Cu(Cys)]$^{+a}$	329.8	326.5
24	[Cu(Gln)]$^{+a}$	372.9	369.8
25	[Cu(Glu)]$^{+a}$	336.5	333.7
26	[Cu(Gly)]$^{+a}$	302.6	302.3
27	[Cu(Ile)]$^{+a}$	317.3	313.0
28	[Cu(Leu)]$^{+a}$	318.9	315.5
29	[Cu(Lys)]$^{+a}$	439.8	441.1
30	[Cu(Met)]$^{+a}$	371.0	371.2
31	[Cu(Phe)]$^{+a}$	343.9	345.7
32	[Cu(Ser)]$^{+a}$	314.7	323.7
33	[Cu(Thr)]$^{+a}$	319.2	321.4
34	[Cu(Tyr)]$^{+a}$	349.4	352.6
35	[Cu(Asp)]$^{+c}$	330.0	329.8
36	[Cu(Trp)]$^{+c}$	376.2	375.1
37	[Cu(Ala)]$^{+b}$	308.4	308.6
38	[Cu(His)]$^{+b}$	410.9	411.6
39	[Cu(Pro)]$^{+b}$	316.9	317.6
40	[Cu(Val)]$^{+b}$	314.4	312.4
41	[K(Arg)]$^{+a}$	192.4	186.8
42	[K(Asn)]$^{+a}$	150.3	147.0
43	[K(Cys)]$^{+a}$	115.8	123.3
44	[K(Glu)]$^{+a}$	136.9	141.3
45	[K(Gly)]$^{+a}$	115.6	118.7
46	[K(His)]$^{+a}$	151.7	161.3
47	[K(Ile)]$^{+a}$	122.1	127.0
48	[K(Leu)]$^{+a}$	127.1	126.5
49	[K(Met)]$^{+a}$	135.1	137.9
50	[K(Pro)]$^{+a}$	135.9	127.0
51	[K(Ser)]$^{+a}$	131.6	131.0
52	[K(Thr)]$^{+a}$	133.1	135.7
53	[K(Tyr)]$^{+a}$	134.6	132.2
54	[K(Val)]$^{+a}$	124.7	123.9
55	[K(Ala)]$^{+c}$	120.7	118.0
56	[K(Lys)]$^{+c}$	149.9	151.0
57	[K(Asp)]$^{+b}$	139.9	139.5
58	[K(Gln)]$^{+b}$	154.1	150.8
59	[K(Phe)]$^{+b}$	135.0	127.0
60	[K(Trp)]$^{+b}$	144.6	144.3
61	[Li(Ala)]$^{+a}$	240.1	235.3
62	[Li(Arg)]$^{+a}$	345.9	347.9
63	[Li(Asp)]$^{+a}$	285.5	282.8
64	[Li(Cys)]$^{+a}$	247.8	251.4
65	[Li(Glu)]$^{+a}$	282.3	286.8
66	[Li(Gly)]$^{+a}$	235.6	233.7
67	[Li(His)]$^{+a}$	325.1	319.6
68	[Li(Leu)]$^{+a}$	249.1	251.0
69	[Li(Met)]$^{+a}$	278.3	276.0
70	[Li(Phe)]$^{+a}$	263.8	263.1
71	[Li(Thr)]$^{+a}$	274.2	274.1

TABLE 1 Continued

no.	complex	DFT calculated	QSPR estimated
72	[Li(Trp)]$^{+a}$	283.8	284.7
73	[Li(Tyr)]$^{+a}$	264.9	271.0
74	[Li(Val)]$^{+a}$	245.3	247.2
75	[Li(Gln)]$^{+c}$	313.5	309.8
76	[Li(Pro)]$^{+c}$	256.4	252.5
77	[Li(Asn)]$^{+b}$	303.2	299.4
78	[Li(Ile)]$^{+b}$	248.6	253.4
79	[Li(Lys)]$^{+b}$	311.3	309.7
80	[Li(Ser)]$^{+b}$	269.1	266.8
81	[Na(Ala)]$^{+a}$	175.9	173.3
82	[Na(Arg)]$^{+a}$	266.1	266.2
83	[Na(Asn)]$^{+a}$	224.8	222.7
84	[Na(Asp)]$^{+a}$	211.0	207.5
85	[Na(Cys)]$^{+a}$	182.2	186.3
86	[Na(Gln)]$^{+a}$	231.7	232.0
87	[Na(Ile)]$^{+a}$	181.4	186.9
88	[Na(Leu)]$^{+a}$	182.7	188.7
89	[Na(Lys)]$^{+a}$	231.7	235.2
90	[Na(Phe)]$^{+a}$	200.1	194.6
91	[Na(Pro)]$^{+a}$	201.5	193.6
92	[Na(Trp)]$^{+a}$	215.2	220.4
93	[Na(Tyr)]$^{+a}$	200.6	202.7
94	[Na(Val)]$^{+a}$	179.7	183.5
95	[Na(His)]$^{+c}$	239.9	248.3
96	[Na(Ser)]$^{+a}$	200.6	197.0
97	[Na(Glu)]$^{+b}$	210.5	211.1
98	[Na(Gly)]$^{+b}$	172.3	170.6
99	[Na(Met)]$^{+b}$	209.6	207.0
100	[Na(Thr)]$^{+b}$	203.6	199.5

[a] Training set. [b] Prediction set. [c] Validation set.

TABLE 2: Descriptors of the Model

relative negative charged surface area, RNCS (amino acid)
maximum nucleophilic reactivity index for an oxygen atom (amino acid)
maximum nucleophilic reactivity index for a nitrogen atom (amino acid)
number of heteroatoms in the side chain (amino acid)
polarizability of the side chain (amino acid)
charge/Shannon radii (metal)
Pauling electronegativity (metal)

− VSET|, where TSET and VSET denote rms errors for the training and validation sets, respectively. Models chosen with this quality factor performed better than models chosen with just training set rms error as the quality factor. That is, CNNs that produce training and validation set errors that are low and similar in magnitude tend to perform well in predicting properties of interest for compounds not used in the training process. A quasi-Newton method BFGS (Broyden−Flectcher−Golfarb−Shanno)[31] is used to train the network. Ten models were obtained from ADAPT, and the statistical analysis was performed for the training and the validation sets; as the coefficients obtained with all of the models were very similar, the one containing the descriptors that encode more related information to the nature of the interactions between the amino acids and the cations was selected.

3. Results and Discussion

According to the procedure indicated, the proposed model contains the seven descriptors given in Table 2: five descriptors are molecular descriptors of the amino acids, and the other two belong to the metal cations. The descriptor relative negative charged surface area, RNCS, is a CPSA descriptor; this type of descriptors is constructed from the surface area of the whole molecule or its fragments and in terms of the charge distribution in the molecule;[32] they encode features responsible for polar intermolecular interactions. The charges used to derive this descriptor have been calculated by the method proposed by Zefirov,[33] based on the Sanderson scale of electronegativity. This descriptor is calculated as the product of solvent-accessible surface area of the most highly negative atom and the relative negative charge. It can be associated with the ion−dipole interactions established between the cations and the amino acid molecules. The two quantum-chemical descriptors, maximum nucleophilic reactivity index for the oxygen and nitrogen atoms,[34] are derived from the Fukui frontier molecular orbital theory. They are defined as:

$$N'_A = \sum_{i \in A} c^2_{i\mathrm{HOMO}}$$

where the summation is performed over all atomic orbitals i of a given atom A, and $c_{i\mathrm{HOMO}}$ denotes the ith atomic orbital coefficient for the HOMO. These indexes give an estimate of the relative reactivity of the atoms in the molecule and are related to the activation energy of the corresponding chemical reaction. Thus, they are clearly related to the ability of the amino acid to bond to the cation through the nitrogen and oxygen atoms. The maximum values of these descriptors correspond to the aminic nitrogen atom and the carboxylic oxygen atom of the amino acid backbone, which are the atoms that are likely to coordinate to the metal cations.[11] The other two structural descriptors, number of heteroatoms and polarizability of the side chain, describe the capacity of the amino acid to coordinate to the metal cations by means of the side chain, and both encode information about the strength of these bonds. The descriptors of the model corresponding to the metal cations are the Pauling

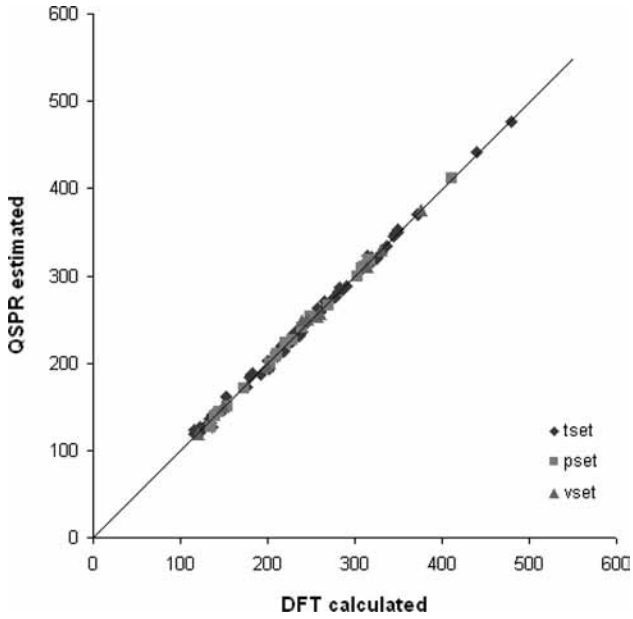

Figure 1. Plot of QSPR estimated versus DFT calculated BA (kJ/mol) values for the training, prediction, and validation sets.

electronegativity and the charge/Shannon radii ratio; both are evidently associated with the strength of the bond between the cation and the amino acid. The dissociation energy of a chemical bond depends on the difference of electronegativities between the two atoms implicated; however, the charge/radii ratio of an atom is a measure of its polarizing power and consequently of the covalent-ionic degree of the involved bond between the cation and the heteroatoms of the amino acid and therefore of its bond energy. It is interesting to note that in one recent QSPR study on the enthalpies of formation of organometallic compounds,[35] the proposed model contains, besides seven constitutional molecular descriptors, these same descriptors of the metals, showing the contribution of the metals to the enthalpy of formation of the organometallic compounds.

The statistical results obtained with this model are very good; the coefficient of determination, R^2, is 0.998, 0.999, and 0.997 for the training, prediction, and validation sets, respectively, and the rms error is 3.89, 2.86, and 3.90 kJ/mol for these three sets, respectively. Figure 1 shows the plot of predicted versus DFT calculated values of BA for the 100 complexes studied. To ensure that these results are not due to chance effects, Y-randomization experiments were conducted, and the average R^2 and the rms error for the 10 cases performed are 0.235 and 68.47 kJ/mol, respectively, proving that chance correlations are not present in this study. The neural network analysis was repeated five times, varying in each case the splitting of the original data set into training, prediction, and validation sets; the statistical values obtained in each case are very similar, the average rms error being 3.99, 3.85, and 3.70 kJ/mol for the training, prediction, and validation sets, respectively; the average R^2 is 0.998 for each one of the three subsets, showing that the model has a good statistical stability and validity. This procedure allowed also the calculation of the 5-fold cross-validation correlation coefficient (Q^2) and rms error for the predictions; in this case, these two parameters are 0.997 and 3.90 kJ/mol, respectively, showing the good prediction ability of the proposed model.

Table 3 shows the statistical parameters for the subsets of BA values for each metal; these results, with R^2 and rms errors very similar between them and also to ones of the complete set, prove the robustness of the model. A simple method that measures the relative importance of the descriptors in CNN-derived models is known;[36] the first descriptor is randomly scrambled and the property is calculated again, and obviously the rms error for the new prediction will be larger than the model error, the so-called base rms error. The difference between the scrambled rms error and the base rms error represents the importance of the descriptor into the model: bigger differences are associated with more important descriptors. In the present case, the two most significant descriptors are those of the cations, charge/Shannon radii ratio and Pauling electronegativity, showing the strong dependence of the BA on the nature of the metal in the interaction with the amino acid molecule; the third most important descriptor is the RNCS of the amino acids.

TABLE 3: Statistics of Binding Affinity Estimations for Each Metal Derivatives Subset

subset	n	R^2	rms
$[Li(AA)]^+$	20	0.998	3.36
$[Na(AA)]^+$	20	0.969	4.28
$[K(AA)]^+$	20	0.925	4.63
$[Cu(AA)]^+$	20	0.996	2.96
$[Ag(AA)]^+$	20	0.998	3.00

The complexes studied, $[M(AA)]^+$, can adopt different structures depending on the conformation of the amino acid and the number and type of coordination modes to the metal. The metallic cations can be coordinated to the amino acid molecules by different sites, aminic nitrogen, carboxylic oxygen, hydroxylic oxygen, other heteroatoms present in the side chain, and also by the rings that several amino acids have. Another option to be considered is the possibility of the amino acid to adopt the zwitterionic form. In our previous study, we optimized very accurately the geometry of the free amino acids molecules and analyzed, very carefully, the possible different modes of coordination. We have concluded that the 100 complexes analyzed adopt one of the following four possibilities (Figure 2). The Li^+, Na^+, Cu^+, and Ag^+ derivatives of the most simple amino acids (Gly, Ala, Val, Leu, Ile) adopt the form A, where the cations are coordinated by the aminic nitrogen and the carboxylic oxygen atoms; while the K^+ cation is coordinated by the two oxygen atoms of the acid group (form B). The other amino acids contain more heteroatoms (oxygen, nitrogen, or sulfur) or rings in the side chain, and therefore the more stable configurations found are the tridentate complexes (form D), where the metal is bonded to the nitrogen and carboxylic oxygen of the amino acid backbone and to the side chain through one heteroatom or ring. The complexes of the Li^+, Na^+, K^+, and Ag^+ cations with the amino acid proline (Pro) present the

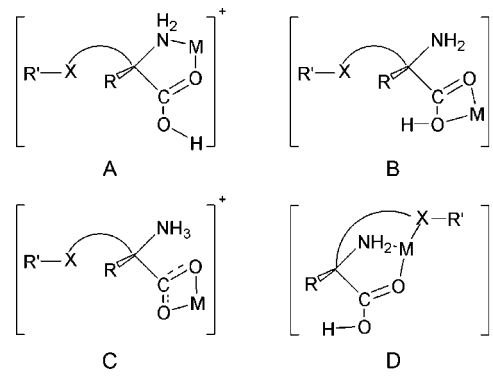

Figure 2. Coordination modes between the cation and the amino acids.

TABLE 4: Predictions with Residuals Larger than Twice the rms Error

complex	entry	DFT calc.	QSPR est.	residual
$[K(Cys)]^+$	43	115.8	123.3	−7.5
$[Na(Pro)]^+$	91	201.5	193.6	7.9
$[K(Phe)]^+$	59	135.0	127.0	8.0
$[Na(His)]^+$	95	239.9	248.3	−8.4
$[K(Pro)]^+$	50	135.9	127.0	8.9
$[Cu(Ser)]^+$	32	314.7	323.7	−9.0
$[K(His)]^+$	46	151.7	161.3	−9.6

zwitterionic form (form C), with the cations coordinated to both oxygen atoms of the acid functional group. On the other hand, the Cu^+ derivative of this amino acid adopts the neutral form coordinated through the aminic nitrogen and the carboxylic oxygen (form A). Even though the cations studied are apparently very similar, they differ in some basic aspects such as the electronic configuration, the radii, and consequently the polarizing power; thus, the interactions with the amino acids can have different degrees of ionic or covalent bonding. To sum, the nature and strength of the chemical bonds in the complexes studied are rather diverse. Despite all of these facts, the model proposed has very good statistical parameters, showing a high capacity to predict the binding affinity of this kind of system. The neural networks used have demonstrated their capability to interpret the different types of interactions between the cations and the amino acids in the $[M(AA)]^+$ complexes analyzed.

The values of Table 1 give one rms error for the 100 entries of 3.71 kJ/mol; there is not any entry with a residual larger than 3 times this error, and only seven complexes have residuals larger than twice the rms error (Table 4). Among these values, three are positive while the other four are negative; they correspond to complexes containing three different cations and five amino acids, and consequently no structural reason is apparent to explain these larger errors. However, four complexes are K^+ derivatives, and as it was stated before, the cation K^+ differs from the other cations studied in the sense that it coordinates to the amino acid in a different way. With the simplest amino acids, it takes form C, when the other four cations adopt form A. Also, of the five amino acids present in these seven complexes, three have rings in the corresponding side chains, and their coordination to the metals is more complex, mainly in the case of the proline. Although the goodness of results obtained proves undoubtedly that these differences in the modes of coordination have been recognized by the neural networks, it is not unexpected that the larger residuals in the predicted values correspond to complexes of the cation K^+ and/or of the amino acid proline. Accordingly, the statistical parameters of the subset formed by the K^+ complexes in Table 3 show a little lower R^2 value and a little higher rms error than do the other four subsets.

4. Conclusions

The QSPR methodology has been successfully applied to the estimation of the binding affinity of the Li^+, Na^+, K^+, Cu^+, and Ag^+ cations to the 20 common amino acids. The proposed model, nonlinearly derived, contains seven descriptors: five of them are molecular descriptors of the amino acids, and the other two correspond to the metal cations. The statistical parameters found and the validation methods used show the goodness, robustness, and the predictive capacity of the model. The descriptors contained in the model encode information clearly related to the factors that affect the bonds between the cation and the amino acid responsible for the BA values analyzed. The neural networks used in the derivation of the model are able to recognize the different nature of the interactions present in the complexes $[M(AA)]^+$, a consequence of the conformations adopted by the amino acids and the atoms that coordinate to the metals. The QSPR-derived model was demonstrated to be a simple, efficient, and rapid method of estimating binding affinities of the amino acids with different cations. The obtained results show that QSPR methods can be useful in the estimation of binding properties of cations and amino acids not reported here and also of more complex systems such as metal cation/peptides.

Acknowledgment. We are grateful to the Catalan Government for financial support (grant 2005 SGR 00814), and J.J. is grateful to Universitat de Barcelona for a BRD grant.

References and Notes

(1) Rodgers, M. T.; Armentrout, P. B. *Acc. Chem. Res.* **2004**, *37*, 989.
(2) Cooks, R. G.; Patrick, J. S.; Kotiaho, T.; Mcluckey, S. A. *Mass Spectrom. Rev.* **1994**, *13*, 287.
(3) Cooks, R. G.; Wong, P. S. H. *Acc. Chem. Res.* **1998**, *31*, 379.
(4) Ryzhov, V.; Dunbar, R. C.; Cerda, B.; Wesdemiotis, C. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 1037.
(5) Rodgers, M. T.; Armentrout, P. B. In *Comprehensive Coordination Chemistry II: From Biology to Nanotechnology. Volume 2: Fundamentals*; Lever, A. B. P., Ed.; Elsevier: New York, 2004.
(6) Gal, J. F.; Maria, P. C.; Massi, L.; Mayeux, C.; Burk, P.; Tammiku-Taul, J. *Int. J. Mass Spectrom.* **2007**, *267*, 7.
(7) Gapeev, A.; Dunbar, R. C. *Int. J. Mass Spectrom.* **2003**, *228*, 825.
(8) Kish, M. M.; Ohanessian, G.; Wesdemiotis, C. *Int. J. Mass Spectrom.* **2003**, *227*, 509.
(9) Lau, J. K. C.; Wong, C. H. S.; Ng, P. S.; Siu, F. M.; Ma, N. L.; Tsang, C. W. *Chem.-Eur. J.* **2003**, *9*, 3383.
(10) Shoeib, T.; Siu, K. W. M.; Hopkinson, A. C. *J. Phys. Chem. A* **2002**, *106*, 6121.
(11) Jover, J.; Bosque, R.; Sales, J. *J. Chem. Soc., Dalton Trans.* **2008**, 6441.
(12) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.
(13) Chalk, A. J.; Beck, B.; Clark, T. A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1053.
(14) Suzuki, T.; Ebert, R.-U.; Schüürman, G. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 776.
(15) Halberstam, N. M.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. A. *Dokl. Chem.* **2002**, *384*, 140.
(16) Kravtsov, A. A.; Karpov, P. V.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. *Dokl. Chem.* **2007**, *41*, 128.
(17) Jover, J.; Bosque, R.; Sales, J. *QSAR Comb. Sci.* **2007**, *26*, 385.
(18) Jover, J.; Bosque, R.; Sales, J. *QSAR Comb. Sci.* **2008**, *27*, 563.
(19) Jover, J.; Bosque, R.; Sales, J. *QSAR Comb. Sci.* **2008**, *27*, 1204.
(20) Jover, J.; Bosque, R.; Sales, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1727.
(21) Tämm, K.; Fara, D. C.; Katritzky, A. R.; Burk, P.; Karelson, M. *J. Phys. Chem. A* **2004**, *108*, 2812.
(22) Siu, F.-M.; Che, C.-M. *J. Phys. Chem. A* **2006**, *110*, 12348.
(23) Katritzky, A. R.; Lovanov, V. S.; Karelson, M. *CODESSA, Reference Manual v 2.13*; Semichem and the University of Florida, 1997.
(24) Bosque, R.; Sales, J. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1154.
(25) Emsley, J. *The Elements*; Claredon Press: Oxford, 1989.
(26) Jurs, P. C.; Chow, J. T.; Yuan, M. *Computer-Assisted Drug Design*; The American Chemical Society: Washington, DC, 1979.
(27) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Functions*; Wiley: New York, 1979.
(28) Luke, B. T. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 179.
(29) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77.
(30) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. *Environ. Toxicol. Chem.* **1994**, *13*, 841.
(31) Wessel, M. D.; Jurs, P. C. *Anal. Chem.* **1994**, *66*, 2480.
(32) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323.
(33) Zefirov, N. S.; Kirpichenock, M. A.; Izmailov, F. F.; Trofimov, M. I. *Dokl. Akad. Nauk. SSSR* **1987**, *296*, 883.
(34) Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984.
(35) Jover, J.; Bosque, R.; Martinho Simoes, J. A.; Sales, J. *J. Organomet. Chem.* **2008**, *693*, 1261.
(36) Guha, R.; Jurs, P. C. *J. Chem. Inf. Comput. Model.* **2005**, *45*, 800.

JP810391Z