Journal of Computer-Aided Molecular Design, **13:** 611–623, 1999.
KLUWER/ESCOM

Application of multivariate data analysis methods to Comparative Molecular Field Analysis (CoMFA) data: Proton affinities and pKa prediction for nucleic acids components

Raimundo Gargallo*, Christoph A. Sotriffer, Klaus R. Liedl & Bernd M. Rode
Department of Theoretical Chemistry, Institute of General, Inorganic and Theoretical Chemistry, University of Innsbruck, Innrain 52a, A-6020 Innsbruck, Austria

Received 21 July 1997; Accepted 11 December 1998

Key words: CoMFA, multivariate analysis, proton affinity, protonation constant

Summary

Multivariate data analysis methods (Principal Component Analysis (PCA) and Partial Least Squares (PLS)) are applied to the analysis of the CoMFA (Comparative Molecular Field Analysis) data for several nucleic acids components. The data set includes nitrogenated bases, nucleosides, linear nucleotides, 3′, 5′-cyclic nucleotides and oligonucleotides. PCA is applied to study the structure of the CoMFA data and to detect possible outliers in the data set. PLS is applied to correlate the CoMFA data with either calculated AM1 proton affinities or with experimental pKa values. The possibility of making a prediction of pKa values directly from 3D structures of the monomers for polynucleotides is also shown. The influence of the superposition criteria and of conformational changes along the glycosidic bond on the pKa prediction are studied as well.

Introduction

Comparative Molecular Field Analysis (CoMFA [1]) is a 3D Quantitative Structure–Activity Relationship (3D QSAR) approach which has been widely applied to predict activity coefficients and biological activities [2]. Despite the usefulness of CoMFA to predict a wide range of activities and thermodynamical constants, few papers have reported its use in the prediction of protonation constant values. To our knowledge, CoMFA has been applied to the prediction of protonation constants of clonidine-like imidazolines, 2-substituted imidazoles and 1-methyl-2-substituted-imidazoles [3], phenols and anilines [4] and to the prediction of the acidity of benzoic acids in the gas phase [5].

CoMFA has not been used to predict any thermodynamic constants concerning the chemical equilibria of nucleic acids components. The present work shows the application of the CoMFA analysis to the prediction of protonation thermodynamic constants (pKa and proton affinities) for different components of nucleic acids, including nitrogenated bases, nucleosides, linear and cyclic nucleotides and oligonucleotides. The correct knowledge of the protonation equilibria is an important step in the study of hydrogen-bonding equilibria in the field of modified nucleic acids [6].

The structure of the data set and the similarity between compounds are important aspects which should be considered before making predictions for test compounds. Outliers and other inhomogeneities can easily be detected from the deviations in the calculated values of the model or by examining the corresponding residual plot. Principal Component Analysis (PCA) is a powerful multivariate data analysis method which can be used to study the structure of the data set and the possible presence of outliers [7,8]. Multivariate data methods, like PCA or Partial Least Squares (PLS), have already been applied to the analysis of data of non-bonded and electrostatic interactions in DNA sequences [9].

*On leave from the Department of Analytical Chemistry, University of Barcelona, Diagonal 647, E-08028 Barcelona, Spain. Correspondence should be addressed to R. Gargallo there. E-mail: raimon@zeus.qui.ub.es

In this work, PCA is applied to the previous exploratory analysis of the CoMFA data. Calculated proton affinities and experimental pKa values are correlated to the CoMFA data by means of PLS.

Materials and methods

Data set description

Table 1 shows all the compounds studied in this work. The data set includes nitrogenated bases, nucleosides, linear nucleotides, 3′, 5′-cyclic nucleotides and oligonucleotides. The pKa values (at 37 °C and 0.15 M, which can be considered to be the physiological conditions of temperature and ionic strength) were extracted from published works (see references in Table 1).

The nitrogenated bases are those present in naturally occurring nucleic acids, except for hypoxanthine, which is the 2-deamino analogue of guanine. For the purinic bases (hypoxanthine and adenine) and derivatives the protonation takes place at the N1 atom (Figure 1). On the other hand, for the pyrimidinic bases (cytosine, uracil and thymine) and derivatives the protonation takes place at the N3 atom. For adenine and cytosine bases and derivatives the protonation takes place in the pH range 3–5 and for uracil, thymine and hypoxanthine derivatives the protonation takes place at pH values below 8 [6,10]. The CoMFA analysis was performed over all the compounds showing the nitrogenated bases in the deprotonated form, i.e., the nitrogen atoms studied (N1 or N3) did not have any hydrogen atom bonded.

Four oligonucleotides (each with a length of three nucleotides) were included in the data set as model compounds for four larger polynucleotides: poly(uridylic) acid (poly(U)), poly(inosinic) acid (poly(I)), poly(cytidylic) acid (poly(C)) and poly(adenylic) acid (poly(A)).

The pKa values for the polynucleotides included in Table 1 were obtained in different ways, according to the polyelectrolytic behaviour of the polynucleotides. Thus, the pKa values for poly(U) and poly(I) were determined from potentiometric and spectrophotometric data obtained from acid-base titrations [11,12]. The acid-base equilibria of these polynucleotides do not show any kind of polyelectrolytic effect, i.e., the pKa value for one of the monomers in the polynucleotide does not depend on the protonation of the neighbour monomers.

On the other hand, the pKa values for poly(C) and poly(A) were determined from spectrophotometric data [13,14]. The acid-base equilibria of these polynucleotides showed the presence of important polyelectrolytic effects, which were reflected in the dependence of the pKa value on the degree of protonation of the polynucleotide. These important polyelectrolytic effects are related to the conformational change from a single stranded helical conformation at neutral pH values to a double stranded helical conformation at pH values near 5. The pKa value given for poly(A) and poly(C) is not a thermodynamic constant, but an intrinsic constant defined as the extrapolated apparent pKa value for a protonation degree equal to zero, i.e., for the theoretical point where no effects of neighbouring protonated sites are present.

Building the molecular structures

All the structures of the monomers and polymers were built with the Biopolymer module of the SYBYL 6.3 software [15]. The geometry optimization for the monomers was first performed with molecular mechanics calculations using the Tripos force field and the Gasteiger–Hückel charges. In order to establish the proper conformation of each molecule, it was required to sample a large conformational space, which was done by means of grid searches on the χ angle (defined as the torsion angle of the nitrogenated base in relation to the sugar ring, i.e., the glycosidic bond). In general, the final optimized structures showed the *anti* conformation to be the energetically lowest structure. However, in a few cases, the *anti* conformation had to be achieved manually [6,16] in order to have the same final conformation for all the compounds [2]. Finally, the structures were optimized and the atomic charges were calculated by application of the AM1 method [17] as implemented in Gaussian 94 [18].

The influence of the conformation along the glycosidic bond on the pKa prediction has been studied in two sets of monomers (cytidine and cIMP). For these structures, the χ angle was kept fixed at the values 0, 60, 120, 180, 240 and 300 degrees and the charges were calculated by means of the AM1 method.

The conformational space of the oligonucleotides was first explored by means of successive runs of simulated annealing, namely by three cycles consisting of heating to 600 K (1000 fs) and cooling to 50 K (1000 fs). The final geometry optimization and atomic charges calculation were performed using the AM1 method.

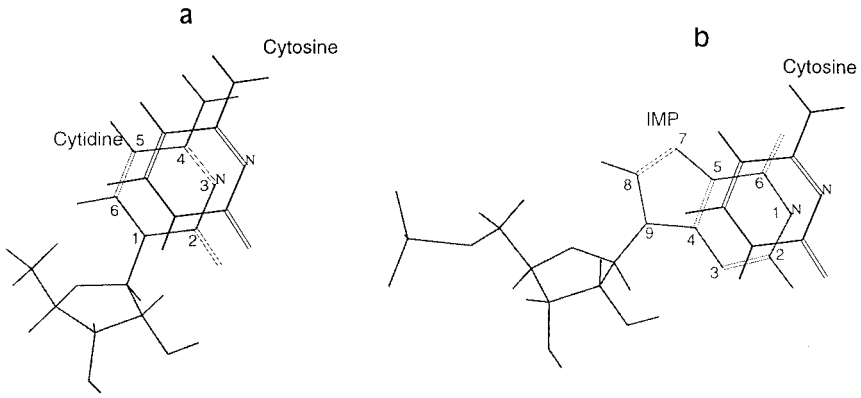

Figure 1. Alignment of the molecules in relation to the six-membered ring for (a) pyrimidinic bases and (b) purinic bases. A small shift of the compounds has been included to clarify the figure. The nitrogen atoms N3 (pyrimidinic bases) and N1 (purinic bases), where the protonation takes place, are also shown.

Table 1. Data set studied in this work. CMP, IMP and UMP are 5′-nucleotides

Number	Compound	Type	Experimental pKa (37 °C and 0.15 M ionic strength)	Reference
1	Adenine	Base	4.07[a]	14
2	Adenosine	Nucleoside	–	–
3	2′-AMP	Linear nucleotide	3.60	34
4	3′-AMP	Linear nucleotide	3.50	34
5	5′-AMP	Linear nucleotide	3.62[b]	34
6	ADP	Linear nucleotide	3.73[b]	34
7	ATP	Linear nucleotide	3.87[b]	34
8	Cytosine	Base	4.56	[d]
9	Cytidine	Nucleoside	4.13	[d]
10	CMP	Linear nucleotide	–	–
11	cCMP	Cyclic nucleotide	3.87	13
12	Hypoxanthine	Base	8.51	12
13	Inosine	Nucleoside	8.50	12
14	IMP	Linear nucleotide	–	–
15	cIMP	Cyclic nucleotide	8.37	33
16	Uracil	Base	8.93	33
17	Uridine	Nucleoside	8.86	33
18	UMP	Linear nucleotide	8.76	33
19	cUMP	Cyclic nucleotide	8.52	33
20	Thymine	Base	9.40	34
21	Thymidine	Nucleoside	9.57	34
22	poly(A)	Oligonucleotide	4.87[a,c]	14
23	poly(C)	Oligonucleotide	4.20[c]	13
24	poly(I)	Oligonucleotide	8.86	12
25	poly(U)	Oligonucleotide	9.36	11

[a]pKa value at 25 °C and 0.1 M ionic strength.
[b]pKa value at 40 °C and 0.1 M ionic strength.
[c]Extrapolated pKa value obtained for the whole protonation processes of poly(C) and poly(A).
[d]Unpublished values from potentiometric titrations.

Superposition of the molecules

After energy minimization, the molecules were aligned by superimposing the six-membered pyrimidine ring on the six-membered ring of both the purinic as well as the pyrimidinic bases (Figure 1). In this alignment process, the N1 purinic atom and the N3 pyrimidinic atom were always located in the same position. In the case of the oligonucleotides, this alignment was done for the nitrogenated base located in the central nucleotide. The influence of the superposition criteria on the pKa prediction was also studied for some monomers (uridine and 5′-AMP).

Calculation of the interaction energies

The steric and electrostatic potential energy fields of each molecule were calculated with the SYBYL software at various lattice points surrounding the molecule using an H^+ as probe atom. The interaction energies between each molecule and the probe atom were calculated at a total of 6900 grid points with 2 Å spacing in a lattice of 44 × 38 × 28 Å using the default Lennard-Jones and Coulomb potential functions. The steric and electrostatic energy values were truncated to 30 kcal/mol. At lattice intersections 'inside' the molecules (as determined by a corresponding steric interaction energy value of more than 30.0 kcal/mol) no electrostatic energies were calculated. A distance dependent dielectric constant and the block-scaling to constant group variance (CoMFA standard scaling option in SYBYL) were also applied.

Proton affinity calculation

Proton affinity (PA) is defined as the negative of the enthalpy change of the reaction $M + H^+ \rightarrow MH^+$, thus:

$$PA = \Delta H_f^0(M) + \Delta H_f^0(H^+) - \Delta H_f^0(MH^+)$$

The computation of PA required the calculation of the energies of formation for the deprotonated and protonated forms. The experimental value of the heat of formation of H^+ was assumed to be 367.2 kcal/mol [19,20]. All the calculations were performed using AM1, which provides the standard enthalpies of formation at 298 K.

The calculation of ΔH_f^0 of the protonated compounds requires the protonation site to be clearly defined because little information is available on that topic. The protonation sites considered here were the nitrogen atom N1 in the purinic bases and the nitrogen atom N3 in the pyrimidinic bases, whose pKa had already been determined experimentally. Thus, the protonated compounds (MH^+) were defined as those which have this nitrogen atom protonated and, on the other hand, the deprotonated compounds (M) were defined as those which have this nitrogen atom deprotonated.

Data matrix and multivariate analysis

The CoMFA data obtained were arranged in an $m \times n$ matrix, where m is the number of compounds included in the data set and n is the number of lattice intersections where the energy interactions were calculated ($n = 6900$). So, each row of CoMFA descriptors corresponds to the field of the corresponding molecule. Multivariate data analysis methods were applied to the CoMFA data matrix with the SYBYL 6.3 QSAR module.

PCA. No factor rotation was applied for the PCA analysis and the CoMFA standard scaling was applied for PCA. A column filtering of 2.0 kcal/mol was applied in order to reduce the initial number of CoMFA columns.

PLS. The regression equations were derived by means of the PLS module [21,22] based on the calculation of orthogonal scores and the Nipals algorithm. The latent variables were subjected to both leave-one-out and leave-more-out cross-validation tests in order to evaluate the number of components yielding the highest r_{cv}^2 without overfitting [23,24]. In the leave-one-out method, each compound was excluded once from the data set and predicted by the sub-model generated from the remaining molecules. In the leave-more-out method, the data set was divided into 5 groups, meaning that each target property value is predicted by a model based on about 80% of the available data. The standard deviation threshold (minimum sigma) was set to 2.0. All rows had the same weight and the CoMFA standard scaling was applied.

Results

Exploratory analysis of the data set

PCA was applied to the CoMFA data matrix containing the interaction energies of all the monomers ($m = 21$ compounds). The eigenvalues obtained for each principal component (PC) are plotted in Figure 2. From this plot, the main importance of the first two PCs can be affirmed. All others are of lower importance.

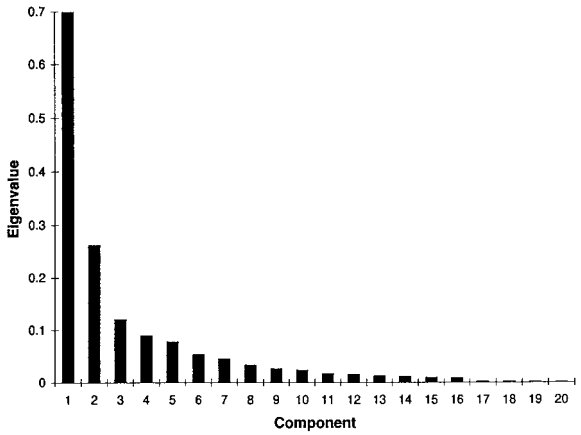

Figure 2. Eigenvalues plot obtained from the PCA analysis of the CoMFA data matrix containing 21 monomers.

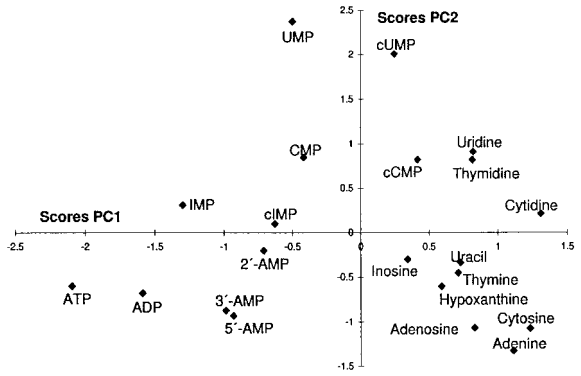

Figure 3. Scores plot of the second principal component versus scores of the first principal component obtained from the PCA analysis of the CoMFA data matrix containing 21 monomers.

The score plots obtained for the first and second PCs are shown in Figure 3. According to the first PC scores, the compounds can be classified in two main groups: those including the phosphate group (left side) and those which do not include this group in their structure (right side). The second PC could explain the influence of the different types of nitrogenated bases. So, the compounds including pyrimidinic bases (cytosine or uracil) in their structures are grouped at the top of the plot and the compounds including purinic bases (hypoxanthine or adenine) are grouped at the bottom. The group of the nitrogenated bases does not obey this rule. On the other hand, the values for the second PC of the compounds containing the same kind of base are different according to the type of the nitrogen where the protonation takes place. So, UMP is located over CMP, IMP is located over 5′-AMP and uridine is located over cytidine. According to this fact, both UMP and cUMP cannot be considered as possible outliers.

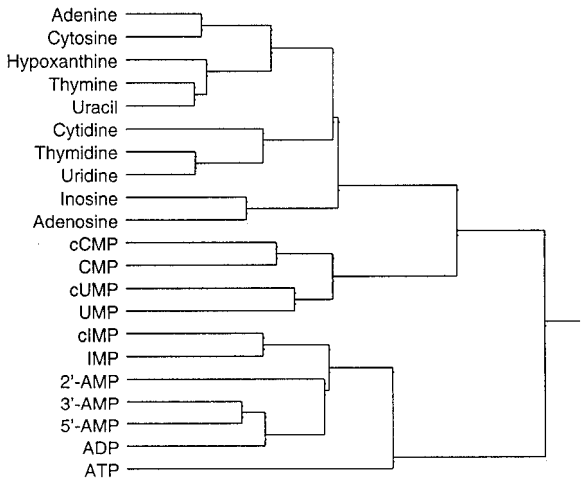

Figure 4. Dendrogram obtained from the PCA of the CoMFA data matrix containing 21 monomers.

A classification of the compounds into families was performed using the SYBYL hierarchical clustering procedure of the obtained PCs; the obtained clustering dendrogram is shown in Figure 4. The analysis of the dendrogram from top to bottom shows the data set divided in two main groups. These groups are further divided in four groups (from top to bottom): bases and nucleosides, nucleotides containing pyrimidinic bases, nucleotides containing purinic bases and ATP. It is worth mentioning how the nitrogenated bases are classified in Figures 3 and 4, because it is different from the way the other compounds are classified. Thus, the nitrogenated bases are grouped according to the type of the nitrogen atom, i.e., according to the net charge in the protonation site. This fact could be related to the small size of the nitrogenated bases, which could make the influence of this net charge more important than in the case of nucleosides or nucleotides. On the other hand, ATP could be a possible outlier because it is the only nucleotide which is not grouped with any other nucleotide.

Proton affinity calculation

The calculated PAs are shown in Table 2. The calculated PAs are strongly correlated with the pKa values and with the presence of phosphate groups in the structure. Thus, the PA values for the adenine and cytosine derivatives are usually lower than the values calculated for the hypoxanthine, uracil and thymine derivatives, whose pKa is higher. On the other hand, the presence of the phosphate group increases the PA value as can be observed in the series adenine $\approx$ adeno-

sine < 5′-AMP or in the series cytosine ≈ cytidine < CMP. In general, this fact can be related to the lower pKa value found for the compounds including the phosphate group in their structure.

Table 2 also shows the experimental PA values [25–28] reported for some of the compounds studied in this work. The calculated PA values for the adenine and cytosine bases and nucleosides agree with the experimental ones.

Proton affinity prediction

The predicted PA values obtained with the PLS analysis and the relative error (in percentage of the calculated PA) are included in Table 2. The results of the leave-one-out cross-validation were:

$$m = 21;\ r^2_{cv} = 0.972;$$
$$\text{optimum number of PLS components} = 4$$

It is worth to mention that Figure 3 shows clustered structures and that with this situation a leave-one-out cross-validation method cannot be sufficient to select the optimal number of components in PLS. In this case, the leave-more-out method can give a more reliable estimation of the optimal PLS component number [29]. Figure 5a shows the variation of the Standard Error in the Prediction (SEP) and of the r^2_{cv} with the number of PLS components. When the appropriate number of PLS components is chosen, SEP is lower and r^2_{cv} is higher. However, the prediction power of the PLS analysis will be worse when the number of PLS components chosen is greater than the optimum number. Accordingly, the results of the leave-more-out cross-validation were:

$$m = 21;\ r^2_{cv} = 0.932;$$
$$\text{optimum number of PLS components} = 3$$

In this case, the leave-more-out method gives a number of components lower than the leave-one-out method, which could lead to an overfitted model. The final results for the whole PLS analysis, using 3 as the number of PLS components were:

$$r^2 = 0.971;\ \text{standard error} = 14.47;$$
$$\text{Steric (\%)} = 30.6;\ \text{Electrostatic (\%)} = 69.4$$

The PLS analysis shows a good recovery for the PA values for the compounds in the data set. The fractions of steric and electrostatic field contribution indicate that in this model electrostatic effects contribute more significantly than steric effects. This fact is related to the presence of electronegative atoms in the structure of these compounds and to the presence of the phosphate group, whose importance was already pointed out in the PCA results.

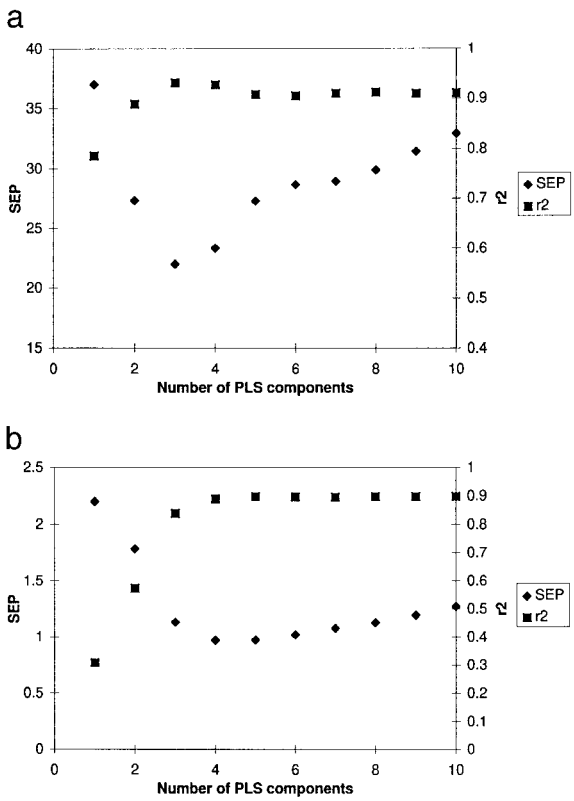

Figure 5. Variation of the Standard Error in the Prediction (SEP) and of r^2_{cv} with the number of PLS components. (a) Cross-validation analysis of CoMFA and PA data. (b) Cross-validation analysis of CoMFA and pKa data.

pKa prediction

pKa prediction for monomers

The data set studied comprised only 18 monomers because the pKa values for adenosine, CMP and IMP were not available at the temperature and ionic strength conditions chosen. Figure 5b shows the variation of SEP and r^2_{cv} with the number of PLS components. Accordingly, the results of the leave-more-out cross-validation were:

$$m = 18;\ r^2_{cv} = 0.890;$$
$$\text{optimum number of PLS components} = 4$$

One PLS component more is needed for this PLS analysis than in the CoMFA-PLS analysis of PA. This fact could be related to the different source of the

Table 2. AM1 calculated, experimental and predicted proton affinity values for the monomers data set

Number	Compound	AM1 calculated PA (kcal/mol)	Experimental PA (kcal/mol)	Ref.	Predicted PA (kcal/mol)	Relative error (%)
1	Adenine	214.4	224.2	25	235.2	−9.7
2	Adenosine	216.9	236.6	26, 27	212.7	2.0
3	2′-AMP	374.8			372.4	0.6
4	3′-AMP	433.6			391.5	9.7
5	5′-AMP	381.1			385.9	−1.3
6	ADP	432.8			447.4	−3.4
7	ATP	483.4			487.1	−0.8
8	Cytosine	219.9	225.9	25	229.7	−4.4
9	Cytidine	221.2	234.8	26, 27	224.1	−1.3
10	CMP	360.4			352.6	2.2
11	cCMP	270.5			271.8	−0.5
12	Hypoxanthine	328.7			334.4	−1.7
13	Inosine	324.5			308.2	5.0
14	IMP	421.5			440.5	−4.5
15	cIMP	372.5			378.7	−1.7
16	Uracil	338.8			332.3	1.9
17	Uridine	332.5			321.2	3.4
18	UMP	419.6			422.9	−0.8
19	cUMP	378.5			376.9	0.4
20	Thymine	337.2			340.8	−1.1
21	Thymidine	332.5			328.8	1.1

dependent variables. So, while PAs were calculated theoretically in an uniform way, pKa's are values obtained under slightly different experimental conditions. Then, a higher number of components may thus be needed to explain the experimental data, while the theoretically calculated values can be explained by a simpler model.

The final results for the whole PLS analysis, using 4 as the number of PLS components were:

$$r^2 = 0.996;\ \text{standard error} = 0.193;$$
$$\text{Steric (\%)} = 46.0;\ \text{Electrostatic (\%)} = 54.0$$

The r^2_{cv} value can be considered to be significant and shows a good prediction power for this analysis and for the calibration data set. However, cross-validation is not infallible and the result obtained for this analysis could be misleading [30,31]. It is already known that strongly grouped data with much fewer real degrees of freedom than the number of compounds may give an improper result and an over-optimistic cross-validation [32]. Thus, any data analysis works better and the final result is more reliable when the data are fairly symmetrically distributed around their mean, and the precision is evenly distributed over its range of variation. Thus, a wide range of the target property (biological activity, pKa, etc.) should be preferred for the analysis and should cover a range much larger than the standard deviation of the data [2]. This is not the case for the data set studied here because the experimental pKa values are grouped in two sets, near 4 and 9, according to the different nature of the protonated nitrogen atom. Under these conditions, two separate PLS models could be built for the two data sets. However, we believe that it is possible to carry out the global PLS analysis for the whole data set studied despite the differences in the experimental pKa values because the compounds included in this data set share similar structures, like the ribose ring or the phosphate group, and the inclusion of all the compounds could improve the global data analysis. This situation is graphically shown in Figure 6, where the Y scores versus the X scores for the four components are plotted. The scores plots for the first (Figure 6a) and second (Figure 6b) PLS components clearly show the

existence of data clusters. However, the scores plots for the third (Figure 6c) and fourth (Figure 6d) PLS components show linearity.

The Y scores versus X scores plot is a way to detect possible outliers in the proposed PLS model. From the visual analysis of the scores plot for the first PLS component (Figure 6a), ATP can be defined as a possible outlier because it lies far away from the others. However, this does not necessarily mean that it should be removed from the data set. In order to assess that ATP is a real outlier, the model was re-built after the ATP was removed. The new PLS model obtained did not get better than the previous one and that would mean that ATP is really only an extreme end-member compound. Then, ATP was put back again, since it obviously contains information and does not harm the model. This kind of extreme compounds can actively help to span the model.

Table 3 shows the experimental pKa values, the PLS predicted pKa values and the absolute error. The predicted pKa values for the whole data set agree with the experimental ones for all the bases, nucleosides, linear nucleotides and cyclic nucleotides. Three compounds (adenosine, CMP and IMP) of unknown pKa value at 37 °C were also included in the data set in an attempt to predict their pKa value under these experimental conditions. The predictions for these compounds should be reliable because their main features are already contained in a region which had been explored by the training set [31], as shown in the previous exploratory analysis of the data set. The pKa values obtained for these compounds are within the expected range.

A very useful result of the CoMFA procedure are the plots which indicate those regions which have the highest influence on the property studied. Figure 7 shows the three-dimensional coefficient contour maps obtained for the CoMFA-PLS analysis, together with the ADP molecule. The black color in Figure 7a indicates the regions where the inclusion of positive charge increases the pKa value. This means that a more negative charge in this region decreases the pKa value. The figure shows that this region is located around the phosphate groups. In fact, the calculated pKa values decrease when a phosphate group is added to the nucleosides.

The grey color indicates the regions where the inclusion of negative charge increases the pKa value. Negative charge in this region is mainly due to the different nature of the deprotonated nitrogen. So, hypoxanthine, uracil and thymine derivatives show a negative charge at the nitrogen where protonation takes place. In fact, the pKa values for the hypoxanthine derivatives are higher than those obtained for the adenine derivatives and similar effects are observed for cytosine and uracil derivatives.

The regions with large influence of the steric field are shown in Figure 7b. The black color indicates the regions where the inclusion of bulky groups increases the pKa value and the grey color indicates the regions where the inclusion of bulky substituents does not increase the pKa value. The figure shows that both regions are located around the ribose and phosphate groups.

Influence of the superposition criteria

The results of any multivariate analysis method based on CoMFA data depend strongly on the superposition criteria used to align the database. The robustness of the method was tested by predicting the pKa for two compounds (uridine and 5′-AMP) shifted from the initial position described above. Figure 8 shows the uridine monomer and the six directions along which the compounds were shifted 1 Å from the starting position. The pKa values predicted with the previous PLS model are shown in Table 4. According to these results, the proposed PLS model depends slightly on the superposition criteria and moderate shifts can affect the pKa prediction.

Influence of the conformational changes along the glycosidic bond

The final PLS model was applied to study the influence of the conformational changes along the glycosidic bond on the pKa prediction. Table 5 shows the pKa predicted for some conformers of the two monomers studied (cytidine and cIMP). It is worth to mention that the experimental pKa values are for an equilibrium of conformational states around the minimum state. The values of the torsion angles in the AM1 optimized structures are 175.2° (cytidine) and 144.8° (cIMP). For both monomers, the best predictions are obtained when the torsion angle is close to the value in the AM1 optimized structure. On the other hand, worse predictions are obtained when the nitrogenated base is located on the ribose ring (*syn* conformation).

pKa prediction for polymers

Table 6 shows the experimental pKa values for four polynucleotides and the calculated pKa values for the corresponding oligonucleotides. The PLS model applied to predict the pKa of the polymers was the

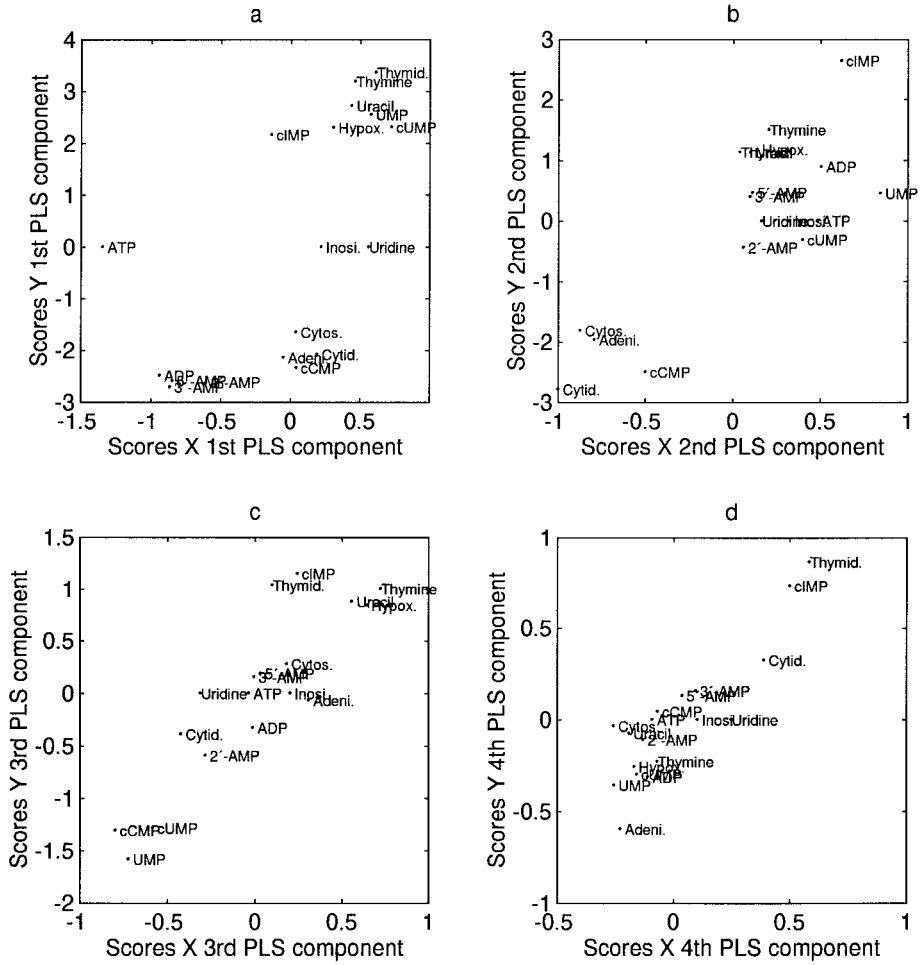

Figure 6. Plot of Y scores versus X scores obtained in the PLS analysis of CoMFA and pKa data for the monomers data set. (a) Y scores for the first PLS component versus X scores for the first PLS component; (b) Y scores for the second PLS component versus X scores for the second PLS component; (c) Y scores for the third PLS component versus X scores for the third PLS component; (d) Y scores for the fourth PLS component versus X scores for the fourth PLS component.

previously developed one for the monomers (data set of 18 monomers). The predicted pKa values are quite different from the actual ones. The differences found are always higher than 0.7 pKa units.

A second model was tested in order to improve the results obtained with the first model. This new model includes, in addition to the CoMFA data matrix, the atomic charge of the N1 (purinic bases) or N3 (pyrimidinic bases) positions, which are those related to the protonation process. So, the inclusion of this column tries to add extra information about the chemical nature of the protonation site, which is different in purinic bases (amide-type nitrogen) and in pyrimidinic bases (aromatic ring nitrogen). The predicted pKa values and the residuals obtained after the CoMFA-PLS analysis are also shown in Table 6. The results of the leave-more-out cross-validation for the new PLS model containing the atomic charge as an extra descriptor are:

$m = 18$; $r^2_{cv} = 0.960$;

optimum number of PLS components = 4

and the results for the whole PLS analysis, using 4 as the number of PLS components is:

$r^2 = 0.994$; standard error = 0.230;

Steric (%) = 26.5; Electrostatic (%) = 22.9;

Charge (%) = 50.6

The initial contribution of the electrostatic field is now split in two terms. The most important contribution is now that of the atomic charge on the protonation

Table 3. Experimental and predicted pKa values for the monomers data set

Number	Compound	Experimental pKa	Predicted pKa	Difference
1	Adenine	4.07	4.35	−0.28
2	Adenosine	–	4.52	
3	2′-AMP	3.60	3.58	0.02
4	3′-AMP	3.50	3.36	0.14
5	5′-AMP	3.62	3.64	−0.02
6	ADP	3.73	3.90	−0.17
7	ATP	3.87	3.83	0.04
8	Cytosine	4.56	4.19	0.37
9	Cytidine	4.13	4.38	−0.25
10	CMP	–	4.91	
11	cCMP	3.87	3.75	0.12
12	Hypoxanthine	8.51	8.61	−0.10
13	Inosine	8.50	8.73	0.23
14	IMP	–	7.58	
15	cIMP	8.37	8.25	0.12
16	Uracil	8.93	8.78	0.15
17	Uridine	8.86	8.74	0.12
18	UMP	8.76	8.76	0.00
19	cUMP	8.52	8.64	−0.12
20	Thymine	9.40	9.40	0.00
21	Thymidine	9.57	9.47	0.10

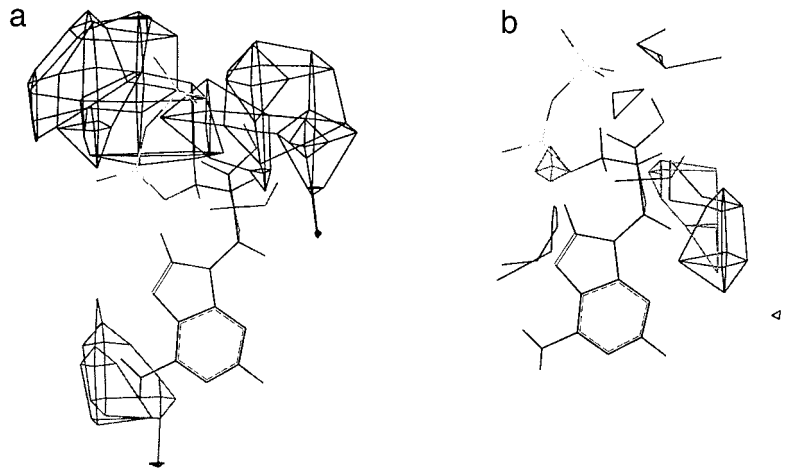

Figure 7. Views of the CoMFA contour map (standard deviation* coefficients) for the CoMFA analysis of the monomers data set. (a) Electrostatic fields. Areas where negatively charged substituents decrease the pKa value are marked in black. Areas where negatively charged substituents increase the pKa value are marked in grey. (b) Steric fields. Areas where bulky substituents increase the pKa value are marked in black. Areas where bulky substituents decrease the pKa value are marked in grey.

Table 4. Influence of the superposition criteria on the prediction of the pKa values for some monomers

Compound	Experimental pKa	Position	Predicted pKa	Difference
Uridine	8.86	1	8.66	0.20
		2	8.30	0.56
		3	8.79	0.07
		4	8.28	0.58
		5	8.60	0.26
		6	8.82	0.04
5′-AMP	3.62	1	4.55	−0.93
		2	3.74	−0.12
		3	4.38	−0.76
		4	3.90	−0.28
		5	4.51	−0.89
		6	3.62	0.00

Table 5. Influence of the χ torsion angle on the prediction of the pKa values for some monomers

Compound	Experimental pKa	χ angle	Predicted pKa	Difference
Cytidine	4.13	0	3.78	0.35
		60	3.82	0.31
		120	4.05	0.08
		180	4.33	−0.20
		240	3.30	0.83
		300	3.51	0.62
cIMP	8.37	0	7.45	0.92
		60	8.24	0.13
		120	8.67	−0.32
		180	8.63	−0.26
		240	7.22	1.15
		300	6.86	1.51

[a] χ angle in cytidine in the calibration data set: 175.2°.
[b] χ angle in cIMP in the calibration data set: 144.8°.

Table 6. Experimental and predicted pKa values for the oligonucleotides data set

Number	Compound	Experimental pKa	Predicted pKa[a]	Difference[a]	Charge N	Predicted pKa[b]	Difference[b]
22	poly(A)	4.87	4.15	0.72	−0.324	3.77	1.10
23	poly(C)	4.2	4.98	−0.78	−0.367	4.56	−0.36
24	poly(I)	8.86	7.59	1.27	−0.448	8.75	0.11
25	poly(U)	9.36	10.59	−1.23	−0.468	9.31	0.05

[a] PLS model without inclusion of the atomic charge on the nitrogen atom.
[b] PLS model with inclusion of the atomic charge on the nitrogen atom.

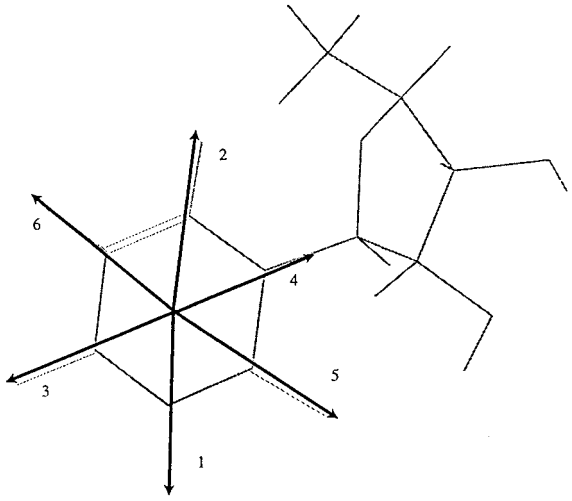

Figure 8. Influence of the superposition criteria on the prediction of the pKa values for uridine. The arrows indicate the direction along which the compound was shifted 1 Å from its original position.

site. The overall electrostatic contribution is now of minor importance and could be mainly related to the influence of the phosphate groups, as can be observed from Figure 7a where only two important electrostatic regions are observed. The degree of correlation of both electrostatic terms should be weak because of the large distance existing between the electrostatic regions responsible for these terms.

The pKa values predicted for the oligonucleotides with the new model are more similar to the experimental values than those predicted with the model which only includes the CoMFA data. The better predictions are for those polynucleotides whose pKa was determined directly from the experimental results. On the other hand, the worst predictions are those for the polynucleotides whose pKa was determined by extrapolation of the experimental results.

Conclusions

The application of multivariate data analysis methods like PCA to a CoMFA data matrix has been shown to be a useful tool in order to assess the homogeneity of the data set. Thus, both the possible presence of outliers and the main sources of variance in the data set can be identified with these methods.

This work has also included the calculation of PA for the data set by means of the AM1 method. Even though AM1 is not the best method for thermodynamic calculations, the selection of this semi-empirical method has been a compromise between the computational resources available and the accuracy of the calculated PA values. Thus, these values have to be considered critically, and only qualitative information can be extracted from the calculated values.

This work has shown the CoMFA approach to be a suitable method to predict proton affinity and pKa values for nitrogenated bases and derivatives. The limited possibility of making pKa predictions for homopolymers has also been shown. The PLS model derived from the study of the monomer data set seems suitable for the prediction of pKa values for homopolynucleotides, although a satisfactory prediction for these pKa values could only be achieved when the atomic charge of the protonation site is also included in the PLS model.

Acknowledgements

R.G. thanks the Spanish Ministerio de Educacion y Cultura for a post-doctoral grant. Thanks are due to Dr. Romà Tauler (Department of Analytical Chemistry, University of Barcelona) for helpful discussions.

References

1. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
2. Kim, K.H., ACS Symposium Series, 606 (1995) 302.
3. Kim, K.H. and Martin, Y.C., J. Med. Chem., 34 (1991) 2056.
4. Martin, Y.C., Wu, J., Curley, J.F. and Kim, K.H., Abstracts of Papers of the ACS, 211 (1996) 149.
5. Kim, K.H. and Martin,Y.C., J. Org. Chem., 56 (1991) 2723.
6. Saenger, W. (Ed.) Principles of Nucleic Acid Structure, Springer-Verlag, New York, NY, 1988.
7. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
8. Massart, D.L., Vandeginste, B.G.M., Deming, Y., Michotte, L. and Kaufman, L., Chemometrics. A Textbook, Elsevier, Amsterdam, 1988.
9. Norinder, U. and Jonsson, J., Quant. Struct.-Act. Relat., 13 (1994) 295.
10. Burger, K. (Ed.) Biocoordination Chemistry, Ellis Horwood Ltd, Chichester, U.K., 1990.
11. Casassas, E., Gargallo, R., Gimenez, I., Izquierdo-Ridorsa, A. and Tauler, R., Anal. Chim. Acta, 283 (1993) 538.
12. Tauler, R., Izquierdo-Ridorsa, A., Gargallo, R. and Casassas, E., Chemometr. Intel. Lab. Syst., 27 (1995) 163.
13. Casassas, E., Gargallo, R., Izquierdo-Ridorsa, A. and Tauler, R., Reactive Polymers, 27 (1995) 1.
14. Casassas, E., Marques, I. and Tauler, R., Macromolecules, 27 (1994) 1729.
15. SYBYL 6.3, Tripos Associates, St. Louis, MO.
16. Visvanadham, V., Reddy, R.M., Bacquet, R.J. and Erion, M.D., J. Comput. Chem., 14 (1993) 1019.

17. Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P., J. Am. Chem. Soc., 107 (1985) 3902.
18. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Gill, P.M.W., Johnson, B.G., Robb, M.A., Cheeseman, J.R., Keith, T., Petersson, G.A., Montgomery, J.A., Raghavachari, K., Al-Laham, M.A., Zakrzewski, V.G., Ortiz, J.V., Foresman, J.B., Cioslowski, J., Stefanov, B.B., Nanayakkara, A., Challacombe, M., Peng, C.Y., Ayala, P.Y., Chen, W., Wong, M.W., Andres, J. L., Replogle, E.S., Comperts, R., Martin, R.L., Fow, D.J., Binkley, J.S., Defrees, D.J., Baker, J., Stewart, J.P., Head-Gordon, M., Gonzalez, C. and Pople, J.A., Gaussian 94, Gaussian Inc., Pittsburgh, PA, 1995.
19. Pointet, K., Milliet, A., Hoyau, S. and Renou-Gonnord, F.R., J. Comput. Chem., 18 (1997) 629.
20. Stull, D.R. and Prophet, H. (Eds) JANAF Thermochemical Tables, Vol. 37, National Reference Data Service, National Bureau of Standards NSRDS-NBS, U.S. Government Printing Office, Washington, DC, 1971.
21. Geladi, P., J. Chemometr., 2 (1988) 231.
22. Wold, S., Albano, C., Dunn, W.J., Edlung, U., Esbenson, K., Geladi, P., Hellberg, S., Lindberg, W. and Sjöström, M., In Kowalski, B.R. (Ed.) Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, The Netherlands, 1984, pp. 17–94.
23. Diaconis, P. and Efron, B., Sci. Am., 116 (1984) 96.
24. Wold, S., Technometrics, 4 (1978) 397.
25. Greco, F., Liguori, A., Sindona, G. and Uccella, N., J. Am. Chem. Soc., 112 (1990) 9092.
26. Hunter, E.P. and Lias, S.G., J. Phys. Chem. Ref. Data, in press.
27. Hunter, E.P., Lias, S.G., Brown, R.L. and Stein, S.E., In Mallard, W.G. and Linstrom, P.J. (Eds.) NIST Standard Reference Database Number 69, February 1997, National Institute of Standards and Technology, Gaithersburg, MD (http://webbook.nist.gov).
28. Del Bene, J.E., J. Phys. Chem., 87 (1983) 367.
29. Forina, M., Drava, G., Boggia, R., Lanteri, S. and Conti, P., Anal. Chim. Acta, 295 (1994) 109.
30. Kroemer, R.T., Ettmayer, P. and Hecht, P., J. Med. Chem., 38 (1995) 4917.
31. Kroemer, R.T., Hecht, P., Guessregen, S. and Liedl, K.R., In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in Drug Design, Vol. 3: Recent Advances, Kluwer, Dordrecht, The Netherlands, 1998, pp. 41–56.
32. Wold, S., Johansson, E. and Cocchi, M., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design. Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993.
33. Casassas, E., Gargallo, R., Gimenez, I., Izquierdo-Ridorsa, A. and Tauler, R., J. Inorg. Biochem., 56 (1994) 187.
34. Izatt, R.M., Christensen, J.J. and Rytting, J.H., Chem. Rev., 71 (1971) 439.