Comparison of Predicted pKa Values for Some Amino-Acids, Dipeptides and Tripeptides, Using COSMO-RS, ChemAxon and ACD/Labs Methods

O. Toure, C.-G. Dussap, A. Lebert

Dossier

This paper is a part of the hereunder thematic dossier published in OGST Journal, Vol. 68, No. 2, pp. 187-396 and available online here

Cet article fait partie du dossier thématique ci-dessous publié dans la revue OGST, Vol. 68, n°2, pp. 187-396 et téléchargeable ici

DOSSIER Edited by/Sous la direction de : **Jean-Charles de Hemptinne**

InMoTher 2012: Industrial Use of Molecular Thermodynamics
InMoTher 2012 : Application industrielle de la thermodynamique moléculaire

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 68 (2013), No. 2, pp. 187-396

IFP Energies nouvelles

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 68 (2013), No. 2, pp. 281-297

DOI: 10.2516/ogst/2012094

Dossier

InMoTher 2012 - Industrial Use of Molecular Thermodynamics
InMoTher 2012 - Application industrielle de la thermodynamique moléculaire

Comparison of Predicted pK_a Values for Some Amino-Acids, Dipeptides and Tripeptides, Using COSMO-RS, ChemAxon and ACD/Labs Methods

O. Toure*, C.-G. Dussap and A. Lebert

Institut Pascal (Axe GePEB), Université Blaise Pascal, Polytech'Clermont-Ferrand,
24 avenue des Landais, BP 206, 63174 Aubière Cedex - France
e-mail: oumar.toure@polytech.univ-bpclermont.fr - claude-gilles.dussap@polytech.univ-bpclermont.fr - andre.lebert@univ-bpclermont.fr

* Corresponding author

Résumé — **Comparaison des valeurs de pK_a de quelques acides aminés, dipeptides et tripeptides, prédites en utilisant les méthodes COSMO-RS, ChemAxon et ACD/Labs** — Les valeurs de constantes d'acidité (pK_a) jouent un rôle très important, en particulier dans l'industrie alimentaire. Les propriétés chimiques des molécules dépendent significativement de leurs états d'ionisation. La plupart des molécules sont capables de gagner et/ou perdre un proton dans les solutions aqueuses. Ce transfert de proton apparaît la plupart du temps entre l'eau et un atome ionisable de la molécule organique. La réponse de la molécule à la protonation ou à la déprotonation dépend significativement du site concerné par le transfert de proton. La distribution partielle des charges dans la molécule varie également en fonction des sites actifs pour la protonation du couple acide/base. Par conséquent on peut l'utiliser pour déterminer le pK_a d'une molécule.

Dans un premier temps, nous avons utilisé la méthode COSMO-RS, une combinaison du modèle de solvatation diélectrique (COSMO) et d'un traitement de thermodynamique statistique pour des solvants plus réels (RS), pour prédire les constantes de dissociation de 50 molécules environ (des acides aminés, des dipeptides et des tripeptides). Les résultats de pK_a obtenus ont été comparés aux valeurs expérimentales, ainsi qu'aux valeurs de pK_a prédites par deux autres méthodes. Nous avons utilisé respectivement la méthode ChemAxon, utilisant un programme basé sur le calcul des charges partielles des atomes d'une molécule, et la méthode ACD/Labs qui permet de déterminer des valeurs de pK_a pour chaque centre de dissociation en considérant que le reste de la molécule est neutre, en utilisant une base de données internes contenant des structures chimiques ainsi que leurs valeurs expérimentales de pK_a.

L'écart-type moyen des valeurs prédites vaut respectivement 0,596 pour la méthode COSMO-RS, 0,445 pour la méthode ChemAxon et 0,490 pour la méthode ACD/Labs. Au vu de ces résultats, la méthode COSMO-RS apparaît comme une méthode prometteuse pour prédire les valeurs de pK_a de molécules d'intérêt dans l'industrie alimentaire pour lesquelles peu de données de pK_a sont disponibles comme les peptides, d'autant plus que les méthodes ACD/Labs et ChemAxon ont été paramétrées en utilisant un grand nombre de données expérimentales (incluant certaines des molécules étudiées dans cet article) alors que la méthode COSMO-RS a été utilisée d'un point de vue purement prédictif.

L'objectif final de cette étude est d'utiliser ces valeurs de pK_a dans un modèle thermodynamique prédictif pour des produits d'intérêt dans l'industrie alimentaire. Pour ce faire, les effets de

certains facteurs (comme le traitement des conformations dans les calculs COSMO-RS, l'influence de la force ionique) pouvant influencer la comparaison entre les données expérimentales et les données prédites, seront discutés.

Abstract* — *Comparison of Predicted pK_a Values for Some Amino-Acids, Dipeptides and Tripeptides, Using COSMO-RS, ChemAxon and ACD/Labs Methods — *Liquid-phase pK_a values play a key role in food science. Chemical properties of molecules depend largely on whether they are ionized or not. Most organic molecules are capable of gaining and/or losing a proton in aqueous solutions. Proton transfer most frequently occurs between water and any ionizable atom of the organic molecule. The molecule's response to protonation or deprotonation depends significantly on the site that was disturbed by proton transfer. Partial charge distribution in the molecule also varies with protonation of the acid/base active sites. Then it can be used to determine the pK_a of a molecule. First, we use the COSMO-RS method, a combination of the quantum chemical dielectric continuum solvation model COSMO with a statistical thermodynamics treatment for more Realistic Solvation (RS) simulations, for the direct prediction of pK_a constants of about 50 molecules (amino-acids, dipeptides and tripeptides). Then, we compare our results with experimental data and the pK_a values predicted using two other methods. We used respectively the ChemAxon method using a program based on the calculation of partial charge of atoms in the molecule and the ACD/Labs method that enables to calculate single pK_a values for all possible dissociation centers when the rest of the molecule is considered neutral, using an internal database containing chemical structures and their experimental pK_a values. The averaged Root Mean Square Error (RMSE) of the predicted pK_a values for each method compared to experimental results were respectively 0.596 for COSMO-RS, 0.445 for ChemAxon and 0.490 for ACD/Labs. While ACD/Labs and ChemAxon are parameterized using a large set of experimental data (including several of the studied molecules), the COSMO-RS method was used in a fully predictive way. Regarding these results, COSMO-RS appears as a promising method to predict the pK_a values of molecules of interest in food science with scarce available pK_a values such as peptides. The final goal of this study is to use the pK_a values in a predictive thermodynamics model for products of interest in food industry. For this purpose, the effects of several factors (like conformations set treatment in COSMO-RS calculations, ionic strength effect) that can affect the comparison between observed and predicted pK_a data are discussed.*

INTRODUCTION

Foods and biochemical media are generally treated as aqueous mixtures that can be very complex, containing mainly water and other varieties of components that can:

- have different molecular sizes (organic acids and minerals, amino-acids, peptides, proteins, etc.) at the temperature (T) and pressure (p) of the system;
- be liquid (*e.g.* alcohols, polyols), solid (*e.g.* sugars, salts) or gaseous (*e.g.* aromatic volatile compounds) at T, p;
- be charged (*e.g.* ions, carboxyl radicals, amines) or neutral (*e.g.* sugars, polyholosides).

A large number of the molecules of interest in food science and biochemistry contains acidic and or basic groups which govern many of their chemical, physical and biological properties. So, most of these organic molecules are capable of gaining and/or losing a proton in aqueous solutions [1].

As defined by Brönsted, an acid is 'a species having a tendency to lose a proton' while a base is 'a species having a tendency to add on a proton'. Hence for every acid, AH, there is a conjugated base, A^- and for every base, B, there is a conjugated acid, BH^+.

If AH (or BH^+) is a strong acid, *i.e.* it has a great tendency to lose protons, it follows that its conjugated base A^- (or B), is a weak base, *i.e.* has only a small tendency to accept protons.

In aqueous solution, acids react with water acting as a base:

$$AH\,(\text{or } BH^+) + H_2O\,(\text{base}) \rightleftarrows H_3O^+ + A^-\,(\text{or B}) \quad (1)$$

and bases react with water acting as an acid:

$$A^-\,(\text{or B}) + H_2O\,(\text{acid}) \rightleftarrows AH\,(\text{or } BH^+) + OH^- \quad (2)$$

In dilute aqueous solution where almost all measurements are made, water is thus the solvent and its activity

is taken as unity [1, 2]. Then, the acidic dissociation (or ionization) constant K_a is calculated with the following equation:

$$K_a = \prod_i a_i^{v_i} = \frac{(H^+ \text{ or } H_3O^+)(B)}{(A)} \quad (3)$$

where, parentheses denote activities, B and A represent respectively base and acid species. This equation can be written in the form:

$$pK_a = pH + \log\left(\frac{(A)}{(B)}\right) \quad (4)$$

where pK_a is the negative logarithm of K_a, and is equal to the pH at which the activities of A and B are equal [2].

The molecule's response to protonation or deprotonation depends significantly on the site that was disturbed by proton transfer. Partial charge distribution in the molecule also varies with protonation of the acid/base active sites. Since the partial charge distribution is very sensitive to the protonation-deprotonation process (both near and far from the disturbed site) [3, 4], it can be used to determine the pK_a of a molecule, which is a measure of the tendency of a molecule or ion to keep a proton, at its ionization center(s).

The more likely ionization occurs, the more likely a species will be taken up into aqueous solution, because water is a very polar solvent [2] (macroscopic dielectric constant $\varepsilon_r = 80$). If a molecule does not readily ionize, then it will tend to stay in a non-polar solvent such as cyclohexane ($\varepsilon_r = 2$) or octanol ($\varepsilon_r = 10$).

Due to the fact that proton transfer most frequently occurs between water and any ionizable atom of the organic molecule, dissociation constants (pK_a) values play a key role in food science and other process industries.

Indeed, the pK_a of a compound is an important property [5] in both life sciences and chemistry since the propensity of a compound to donate or accept a proton is fundamental for the understanding of chemical and biological processes. In biological terms, pK_a is thus an important concept in determining whether a molecule will be taken up by aqueous tissue components or the lipid membranes. It is also closely related to the concepts of pH (acidity of solution) and log(P) (the partition coefficient between immiscible liquids) [2]. As the pK_a value of a molecule also determines the amount of protonated and deprotonated species at a specific pH, for example at physiological pH, knowing the pK_a of a molecule gives insight into pharmacokinetic properties. The latter includes the rate at which a molecule will diffuse across membranes and other physiological barriers, such as the blood brain barrier. More often, phospholipid membranes easily absorb neutral molecules, while ionized molecules tend to remain in the plasma or the gut before being excreted. Many biological systems also use proton-transfer reactions to communicate between the intra- and extracellular media, and the rate of the proton-transfer reaction depends, in part, on the pK_a values of the species involved.

In another area, microorganisms are inhibited by the non-dissociated forms of weak organic acids. The knowledge of pK_a values is then of great importance in microbiology previsional models [6].

Furthermore, in efforts to take greater control over the 'design-make-test' cycle typically implemented in modern drug discovery efforts [7], considerable attention has been given to providing accurate pK_a measurements with good throughput. Increasing attention given to pK_a during drug discovery is evidenced by the development of high-throughput methods for rapid pK_a determination. While experimental methods continue to become more sophisticated and refined, it is often desirable to predict dissociation constants for "virtual compounds", *i.e.*, those that have been described by a compound designer (chemist or modeler) but that have not yet been synthesized [7].

Many different algorithms [5, 7-10] for predicting pK_a values have been developed, and a few have been packaged into commercial computer software applications.

The main objective of the present study is to look for one (or more) reliable pK_a prediction method(s) that can enable to determine the dissociation constants of some components of interest in food sciences. For this purpose, we used a training set of molecules (amino-acids, dipeptides and tripeptides) having known experimental pK_a values. Then, we used 3 different predictions methods namely ChemAxon [3, 4] (Marvin version 5.4.1.1), ACD/Labs [11] (version 10.01, Release 10.00) and COSMO-RS [12, 13] (COSMOtherm [8, 9], version C2.1, Release 01.11) to predict the pK_a values of these molecules, and compare each predicted value to the corresponding experimental value.

1 MATERIALS AND METHODS

1.1 Experimental Data

The experimental pK_a values (at room temperature and atmospheric pressure) used in this study are taken from "*Dissociation constants of organic bases in aqueous solutions*", Perrin [14] (1965) and "*Dissociation constants of organic bases in aqueous solutions – Supplement*", Perrin [15] (1972). In these books, the experimental information related to each pK_a values is given.

1.2 The ChemAxon Method

The ChemAxon method [3, 4, 16, 17] is based on empirically calculated physico-chemical parameters (mainly partial charges) that are obtained from ionization site-specific regression equations. For a given molecule, it uses three types of calculated parameters (intramolecular interactions, partial charges and polarizabilities) to determine the micro ionization constants pK_a of monoprotic molecules [16]:

$$pK_a = a \times Q + b \times P + c \times S + d \qquad (5)$$

where, Q and P denote respectively the partial charge and the polarizability increments, S is the sum of the structures specific (steric strain or/and hydrogen bond) increments; a, b, c and d are regression coefficients specific to the ionization site. All of these pK_a increments are calculated from ionization-site specific regression equations.

Then, the ratio of microspecies is calculated to assign calculated pK_a values to the atoms of the submitted molecule. Finally, macro pK_a values are obtained from the theoretical relations that hold between macro-micro pK_a values. When a molecule contains more than one ionizable atom (*i.e.* multiprotic compound), one has to distinguish between micro and macro acidic dissociation constants. The micro acidic dissociation constant is obtained from the equilibrium concentration of the conjugated acid-base pairs. The macro acidic dissociation constant is obtained from the global mass and charge conservation law. The pK_a of the active groups at a given pH can be calculated according to this relation [16, 17]:

$$K_{a,i} = \frac{\sum_j c_j^i}{\sum_k c_k^{i-1}} [H^+] \qquad (6)$$

where, $[H^+]$ denotes the proton concentration of the aqueous solution, c_j^i is the concentration of the j-th microspecies that released i protons from the fully protonated molecule, c_k^{i-1} is the concentration of the k-th microspecies that released (i-1) protons from the fully protonated molecule. Ratio of c_j^i and c_k^{i-1} also called microspecies distributions are calculated from the micro ionization constants.

As an illustration case, the pK_a calculation for threonine (pT) is described below. This molecule has 4 different ionic forms shown below (*Fig. 1*).

In the present study, only the two first pK_a values for this molecule are calculated (the corresponding experimental dissociation constants of threonine are $pK_{a1} = 2.09$; $pK_{a2} = 8.81$).

By plotting the titration curves *i.e.* the evolutions of the ratio of the ionized and neutral forms *versus* the *pH* (*Fig. 2*), one can identify the ionic species which are present in the mixture at each *pH* values, and because $pK_a = pH$ where the ratio of two different forms are equal, it is possible to identify all the pK_a values predicted in the range of *pH* specified ($0 \leq pH \leq 14$ for this study).

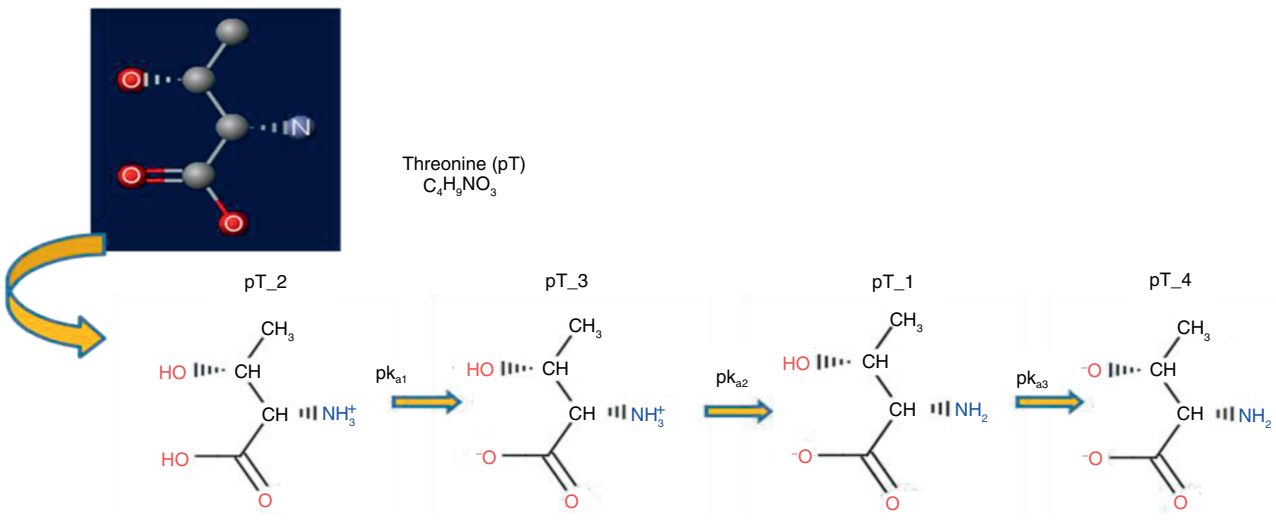

Figure 1

Neutral and ionic forms of threonine (pT). The names of the cationic form (pT_2), the neutral form (pT_3), the anion (pT_1) and the di-anion (pT_4) are evidenced. The two experimental values of this molecule calculated in the present study are $pK_{a1} = 2.09$ and $pK_{a2} = 8.81$.

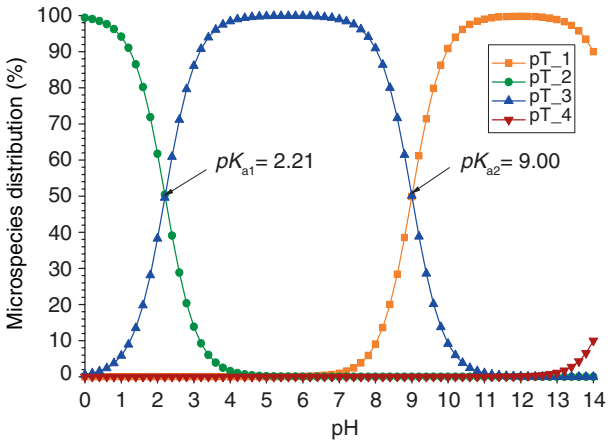

Figure 2

Illustration of the pK_a values of threonine (pT) predicted using the ChemAxon method. This graph shows the evolution of microspecies distribution *vs* pH values, and enlightens the ChemAxon predicted pK_a values that are respectively pK_{a1} = 2.21 and pK_{a2} = 9.00.

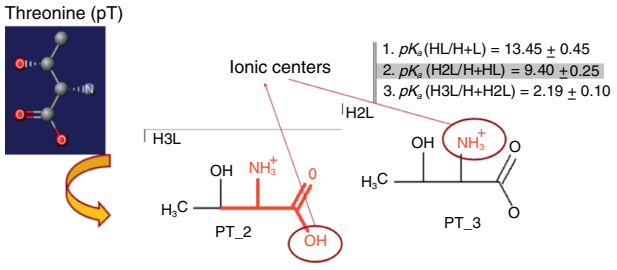

Figure 3

Illustration of the pK_a values of threonine (pT) predicted using the ACD/Labs method. This image shows the different ionic centers in the structures of pT_2 (H3L form) and pT_3 (H2L form) for which ACD predicted pK_a values are respectively pK_{a1} = 2.19 and pK_{a2} = 9.40.

Regarding Figure 2, one can see that the pK_a value pK_{a1} = 2.21 corresponds to the intercept between the microspecies distribution plots of ionic species 2 (pT_2) and 3 (pT_3). Likewise, the pK_a value pK_{a2} = 9.00 corresponds to the intercept between the microspecies distribution plots of ionic species 3 (pT_3) and 1 (pT_1). These values are in very good agreement with the experimental values (respectively pK_{a1} = 2.09 and pK_{a2} = 8.81).

1.3 The ACD/Labs Method

The ACD/Labs pK_a prediction method [11] enables to calculate single pK_a values for all possible dissociation centers when the rest of the molecule is considered neutral, using an internal database containing chemical structures and experimental data.

The algorithm of this calculation mimics the experimental order of protonation of the drawn molecule and determines the pK_a values which can be experimentally measured in aqueous solution.

This model is based on Linear Free Energy Relationships (LFER), applying the Hammett equation [2, 7]:

$$pK_a = pK_a^0 + \Delta(pK_a) \tag{7}$$

where pK_a^0 is the ionization constant for the parent molecule, and $\Delta(pK_a) = \sum \Delta(pK_a)^i$ is the sum over the influence of all other functional groups on the respective pK_a of the $\Delta(pK_a)^i$ values of each reaction center;

$$\Delta(pK_a)^i = -\sum \rho^j \sigma^j \tag{8}$$

where ρ^j is the constant for a particular class j of molecules, and σ^j is the electronic effect of the jth substituent on the ionization constant of the parent molecule.

For this purpose, every ionizable group is characterized by several Hammet-type equations that have been parameterized to cover the most popular ionizable functional groups. The ACD/Labs internal training set contains more than 2 000 derived experimental electronic constants (σ^j). When the required substituent constant is not available from the experimental database; the ACD/Labs method uses another algorithm to describe electronic effect transmissions through the molecular system. The flaw in this method is that the parent molecules inherently carry the majority of the chemical information and without training on a particular parent, predictions for such compounds are impossible. That's why an Internal Reaction Centers Database is used for pK_a prediction using this method. The ACD/Labs internal training set contains more than 31 000 experimental values for 15 932 structures. These data are taken from various articles published in peer-reviewed scientific journals [18].

As an illustration case, the pK_a calculation for threonine (pT) is described in Figure 3.

The ACD/Labs predicted pK_a values corresponding to the two ionic centers studied earlier (in *Sect. 1.4*, using the ChemAxon method) are respectively pK_{a1} = 2.19 and pK_{a2} = 9.40 (when the corresponding experimental values were respectively pK_{a1} = 2.09 and pK_{a2} = 8.81).

1.4 The COSMO-RS Method

COSMO-RS [13, 19, 20] is a predictive method for thermodynamic equilibrium of fluids and liquid mixtures that is a combination of the quantum chemical dielectric continuum solvation model COSMO [12] (acronym of Conductor-like Screening Model) with a statistical thermodynamics treatment for more Realistic Solvation (RS) simulations.

The equilibrium thermodynamic properties derived from the COSMO-RS theory are computed in COSMOtherm, a command line/file driven program which can be run directly from a UNIX or DOS shell. In the present study, we used the C21_0111 version of COSMOthermX [8], a Graphical User Interface to the COSMOtherm [9] command line program.

In this software, the pK_a of a solute j can be estimated from the Linear Free Energy Relationship (LFER) [21, 22]:

$$pK_a = A\left(\frac{\Delta G^j_{neutral} - \Delta G^j_{ion}}{RT \ln 10}\right) + B \qquad (9)$$

where $\Delta G^j_{neutral}$ and ΔG^j_{ion} are respectively the free energies of the neutral and ionic compounds, in the solvent (water in our case) at infinite dilution; A and B denote LFER parameters that were determined for example by correlating calculated free energies of dissociation with the experimental aqueous pK_a for a set of 64 organic and inorganic acids (not including any peptide) [21].

Equation (9) should also be rewritten as:

$$pK_a = c_0 + c_1\left(\Delta G^j_{neutral} - \Delta G^j_{ion}\right) \qquad (10)$$

Thus to obtain a pK_a value it is necessary to do quantum COSMO calculations of a molecule in its neutral and in its ionic state. All the pK_a calculations (of the present study) were done at the large TZVP basis set, in the following denoted BP-TZVP, using a full Turbomole BP-RI-DFT COSMO optimization of the molecular structure.

The LFER parameters $c_0 = B$ and $c_1 = \frac{A}{RT \ln 10}$ used to predict the pK_a values of interest in this study were read from the COSMOtherm parameter file. At ambient temperature, their values are respectively $c_0 = -120.29804$ and $c_1 = 0.10927$ mol/kJ.

pK_a prediction by COSMOtherm is not restricted to aqueous acid pK_a. However, both aqueous base pK_a prediction and pK_a in non-aqueous solvents require reparametrization of the pK_a LFER parameters. Likewise, to compute pK_a at non-ambient temperature, a reparametrization of the LFER parameters is required.

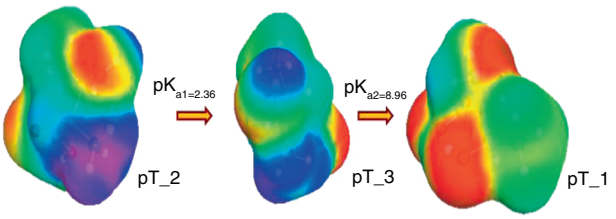

Figure 4

Illustration of the pK_a values of threonine (pT) predicted using the COSMO-RS method. This image shows the different ionic structures involved in the dissociation reactions for which COSMO-RS predicted pK_a values are respectively $pK_{a1} = 2.36$ and $pK_{a2} = 8.96$.

TABLE 1

Summary of the observed (exp.) and predicted pK_a values for threonine

Method	pK_{a1}	pK_{a2}
ChemAxon	2.21	9.00
ACD/Labs	2.19	9.40
COSMO-RS	2.36	8.96
Exp.	2.09	8.81

As an illustration case the pK_a calculation for threonine (pT) at ambient temperature is described in Figure 4.

Regarding this image, we can see that the predicted pK_a values of threonine, using the COSMO-RS method, are respectively $pK_{a1} = 2.36$ and $pK_{a2} = 8.96$ (while the experimental values were respectively $pK_{a1} = 2.09$ and $pK_{a2} = 8.81$). All the predicted pK_a values of the illustrative case are shown in Table 1.

One has to note that all the 3 methods tested (ChemAxon, ACD/Labs, and COSMO-RS) enable the user to input (in their software package) structures in SMILES format [23], and optionally to incorporate "local" data in order to bias predictions. However these optional facilities were not studied, since we are looking for a predictive tool for complex structures.

1.5 Data Analysis

To compare the predicted values *versus* observed ones, we first perform a graphical analysis of pK_a results. Then statistical tests to compare the three prediction methods (ChemAxon, ACD/Labs, and COSMO-RS) are performed.

1.5.1 Indices of Performance of Models

The bias factor (B_f) and the accuracy factor (A_f) are two indices of performance that enable to compare the goodness-of-fit of competing models [24, 25].

The bias factor provides an indication of the average deviation between the model predictions and the average deviation between the model predictions and observed results, it is defined as:

$$B_f = 10^{\frac{\sum \log\left(\frac{pK_a^{\text{cal}}}{pK_a^{\text{exp}}}\right)}{n}} \tag{11}$$

where pK_a^{cal} is the predicted pK_a value, pK_a^{exp} and is the experimental pK_a value and n is the number of observations. A bias factor of 1 indicates perfect agreement between observed and predicted pK_a values. Because over- and under-predictions may cancel out, the bias factor provides no indication of the range of the deviation between predictions and observations. A bias factor greater (resp. lower) than 1 indicates that the model predicts, on average, pK_a values higher (resp. lower) than experimental ones [24, 25].

The accuracy factor (A_f) seeks to provide an estimate of the average deviation between prediction and observation, and is defined as:

$$A_f = 10^{\frac{\sum \left|\log\left(\frac{pK_a^{\text{cal}}}{pK_a^{\text{exp}}}\right)\right|}{n}} \tag{12}$$

1.5.2 Statistical Analysis of Prediction Errors

We calculate the RMSE of the predicted pK_a values for each method:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\Delta pK_a)^2}{n}} \tag{13}$$

Then, a normalization of the residuals (noticing that 95% of these points must be located between -2 and $+2$) is performed on the predicted values using the following equations:

$$\text{Normalized residual} = \frac{\Delta pK_a - \overline{\Delta pK_a}}{\sigma_{\Delta pK_a}} \tag{14}$$

where $\underline{\Delta pK_a} = pK_a^{\text{calc}} - pK_a^{\text{exp}}$ is the error on predicted value; $\overline{\Delta pK_a}$ and $\sigma_{\Delta pKa}$ are respectively the average and the standard deviation of the prediction errors on pK_a values; and are given by:

$$\begin{cases} \overline{\Delta pK_a} = \frac{\sum_{i=1}^{n} \Delta pK_a}{n} \\ \sigma_{\Delta pK_a} = \frac{\sum_{i=1}^{n} \left(\Delta pK_a - \overline{\Delta pK_a}\right)^2}{n} \end{cases} \tag{15}$$

To have a more general comparison of the errors on predicted pK_a, we perform several normality tests on error bars using the statistical tools [26-27] available in the R software [28]. For this purpose, we mainly compare the cumulative distribution functions, the Quantile *versus* Quantile (Q-Q) plots, and the Percentile *versus* Percentile (P-P) plots, of the prediction errors distributions of the competing methods. The CFD describes the probability of "hitting" a value x or less in a given distribution (a normal or Gaussian distribution in our test). The Q-Q plot represents the quantiles of the theoretical fitted distribution (x-axis) against the empirical quantiles of the sample data (y-axis). Likewise, for each value of the data set the P-P plot represents the cumulative density function of the fitted distribution (x-axis) against the empirical cumulative density function of the sample data (y-axis).

1.5.3 Factors that Can Affect the Comparison Between Predicted and Observed pK_a Values

The almost of experimental pK_a values [14, 15] used in the present study, were fitted to zero ionic strength unless otherwise indicated under "Remarks", in which cases, an ionic strength correction of the experimental values should be necessary to perform a better comparison to predicted values. Indeed, if the solutions were no more concentrated than 0.01 M, the corrections of experimental data would be small, and the author may choose to neglect them for his purposes [1]. However there are circumstances in which they must not be neglected because ionization constants change with dilution (although there is seldom a detectable change below 0.001 M) [1]. Since for experimental pK_a measurements methods like potentiometric titrations or spectrometric determinations, the corrections for diluter solutions involve the ionic strength, written as I and define as:

$$\text{I} = \frac{1}{2}\sum_i \left(C_i \eta_i^2\right) \tag{16}$$

where, C_i is the molar concentration of an ion, and η_i is the charge of the ion.

The ionization constant yielded directly by potentiometric titration is appropriately denoted as K_a' (or K_a^M) because it is a mixed constant [1], partly thermodynamic and yet partly concentration-dependent. This mixed character arises from the fact that a pH set is calibrated in terms of hydrogen ion activity (not hydrogen ion concentration), whereas the ionic term is a concentration (not an activity). Thus,

$$K_a'(\text{or } K_a^M) = \frac{(\text{H}^+ \text{ or } \text{H}_3\text{O}^+)[\text{B}]}{[\text{A}]} \tag{17}$$

A relation between pK_a and pK'_a can be derived beginning from:

$$a_i = c_i\gamma_i \tag{18}$$

where γ_i is the activity coefficient (molar scale) of an ion of activity a_i and molar concentration c_i.

For an ion of charge η_i, the activity coefficient is given for dilute solutions by activity coefficients models [1, 2, 29]. In the present study, we use the correction proposed by Ould-Moulaye (in his PhD Thesis [29]) to take into account the influence of ionic strength on pK_a values, using a simplified Goldberg model [30], and assuming that the acidic form is not very concentrated ($\gamma^m_{AH} = 1$). It was demonstrated that (at 25°C):

$$pK_a = pK'_a + 0.51065\sum_i \left(v_i{\eta_i}^2\right)\frac{\sqrt{I_m}}{1+1.6\sqrt{I_m}} \tag{19}$$

where v_i represents the stoechiometric coefficient of the species in dissociation reaction, $v_i > 0$ for products and $v_i < 0$ for reactants; I_m represents the ionic strength in molality scale.

Due to the fact that ChemAxon and ACD/Labs use 2D structures to predict pK_a values, while COSMO-RS performs a geometrical optimization of the 3D structure in its Quantum Chemistry (QC) calculations, it would be interesting to have a look on the influence of conformations set treatment in the COSMO-RS pK_a predictions. For this purpose, the influence of conformations treatment for 2 molecules namely histidine (pH) and histidylglutamic acid (pHE) in COSMO-RS calculations is studied. Sometimes, there are several different reported pK_a values for the same dissociation reaction (ideally, a reliable prediction method would be able to choose the right experimental value). So by using all these points with only one predicted value and 2 or more experimental values, one can influence the global RMSE of the prediction methods. This point is discussed later.

2 RESULTS AND DISCUSSION

2.1 Graphical Comparisons

The predicted pK_a values are plotted *versus* the experimental values in Figure 5. A linear fit of the predicted pK_a data enables to determine the coefficient of determination (r^2) of each competed model. As results, we got respectively $r^2 = 0.98$ for ACD/Labs, $r^2 = 0.98$ for ChemAxon and $r^2 = 0.96$ for COSMO-RS. However, it is very difficult to compare the three prediction methods from this graph, as there are lots of points for which one method gives better predicted pK_a values compared to the others. However, one can already see that the maximum error on predicted pK_a values is observed for the 60th data-point ($pK_a^{exp} = 10.98$) corresponding to one of the dissociation constants of hystidylglutamic acid (pHE). For this data-point, ACD/Labs and ChemAxon methods give erroneous prediction results ($pK_a = 12.86$ and 14.03 respectively) while COSMO-RS predicts a $pK_a = 11.44$ which is significantly closer to the experimental value. One justification of these differences on predicted pK_a should be the fact that the pK_a calculated for this structure (on a dissociation center which is a cyclic nitrogen atom) that was probably not used in the internal database of ACD/Labs and ChemAxon, while COSMO-RS overcome this problem when performing a geometrical optimization during the QC calculations preceding its pK_a prediction.

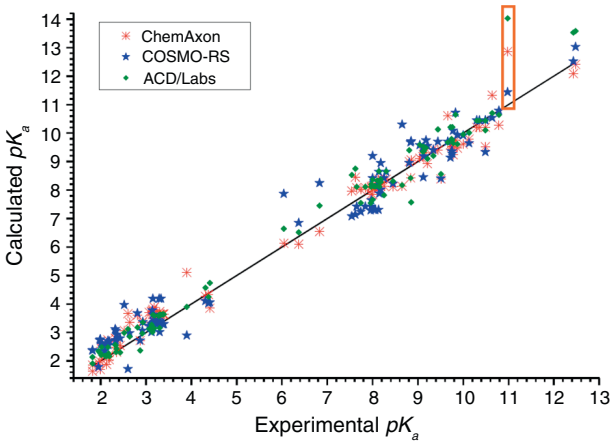

Figure 5

pK_a results for several amino-acids, dipeptides and tripeptides using the ChemAxon (red stars), the COSMO-RS (blue filled stars) and the ACD/Labs (green diamonds) methods. The line represents the equality between predicted values and experimental data and the orange box shows the predicted pK_a data corresponding to the maximum error (data-point number 60).

2.2 Bias Factor and Accuracy Factor

The calculated values of the bias and accuracy factors (using respectively *Eq. 11* and *Eq. 12*) are shown in Tables 2 and 3 for each family, and on the overall data. These values were very close to 1 that means each of the 3 studied methods performs a good prediction of the pK_a values.

Using the bias factor criterion, ACD/Labs and ChemAxon give slightly better pK_a predictions compared to COSMO-RS. For amino-acids the bias factor of COSMO-RS (1.09) is slightly greater than those of ChemAxon (1.02) and ADC/Labs (1.04), while for

TABLE 2
Values of the bias factor (B_f) for each of the 3 methods used for pK_a prediction

	B_f ChemAxon	B_f COSMO-RS	B_f ACD/Labs	Number of points
Amino-acids (AA)	1.02	1.09	1.04	45
Dipeptides	1.07	1.03	1.04	37
Tripeptides	1.04	1.01	1.05	25
Overall data	1.04	1.05	1.04	107

TABLE 3
Values of the accuracy factor (A_f) for each of the 3 methods used for pK_a prediction

	A_f ChemAxon	A_f COSMO-RS	A_f ACD/Labs	Number of points
Amino-acids (AA)	1.06	1.13	1.06	45
Dipeptides	1.11	1.11	1.06	37
Tripeptides	1.05	1.07	1.05	25
Overall data	1.07	1.11	1.06	107

dipeptides and tripeptides COSMO-RS has the smallest bias factor (resp. 1.03 and 1.01), followed by ACD/Labs (1.04 and 1.05) and ChemAxon (1.07 and 1.04). Using the accuracy factor criterion, ACD/Labs and ChemAxon give better pK_a predictions compared to COSMO-RS. For each family, the accuracy factor of COSMO-RS is slightly greater than those of ChemAxon and ADC/Labs. Since ACD/Labs and ChemAxon are parameterized using experimental data of molecules, the slightly higher values in the A_f and B_f values for COSMO-RS is quite normal as discussed later in Section 2.5.

2.3 Statistical Analysis of Prediction Errors

The RMSE of the overall predicted pK_a values for each method compared to experimental results were respectively 0.596 for COSMO-RS, 0.445 for ChemAxon and 0.490 for ACD/Labs (*Tab. 4*).

These RMSE results are in good agreement with literature values. Indeed, for complex molecular structures a RMSE of 0.50 is expected for ACD/Labs [11] (version 10). In a recent study, on 211 drug like-compounds, Manchester *et al.* (2010) [7] got a RMSE of 0.6 for ACD/Labs method (version 10), and 0.8 for the ChemAxon method (Marvin, version 5.2). Their results agree well with the RMSE values given in the present study since we used a more recent version of Marvin (version 5.4.1), that should explain the difference in RMSE.

Then one can plot Figure 6 which compares the normalized residuals (that should ideally be equal to zero) of the 3 prediction methods, using Equation (13). One has to note that the maximum normalized error is got for data-point number 60 (pHE) discussed earlier (6.27 for ACD/Labs; 4.10 for ChemAxon, and 0.54 for COSMO-RS). But, when analyzing this plot, it becomes also clearer that there is not a distinguishable difference between the predictions methods using this comparison criterion, since for data-point number 19 (pH) COSMO-RS has the largest normalized residual (2.90 *versus* −0.09 for ChemAxon and 0.91 for ACD/Labs), and for data-point number 9 (pD) ChemAxon has the largest normalized error (2.53 *versus* −1.97 for COSMO-RS and −0.40 for ACD/Labs). Furthermore, 6 points for COSMO-RS, 5 points for ChemAxon and 6 points for ACD/Labs are located outside the range of [−2, +2] for normalized residual values. That are in good agreement with the 95% points expected to be in the same range, since 107 data-points were used for each of the studied models. These examples illustrated that all the studied methods are undistinguishable in that each can sometimes give large errors. This statement is confirmed in another study [7], in the case of ACD/Labs and ChemAxon.

Figure 7 represents the comparison between the theoretical and observed cumulative distribution functions (CFD) [27] of the prediction errors for each pK_a prediction method. One can notice that the COSMO-RS method's sample CFD is the closest to its theoretical CFD, followed by ChemAxon, while the ACD/Labs method's

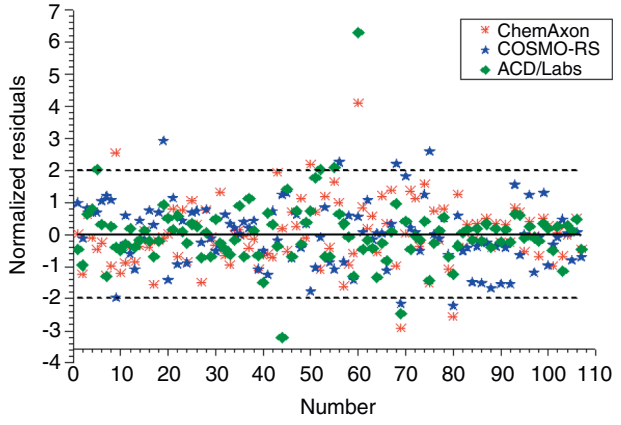

Figure 6

Normalized residuals of the pK_a predicted values for several amino-acids, dipeptides and tripeptides using the ChemAxon, the COSMO-RS and the ACD/Labs methods.

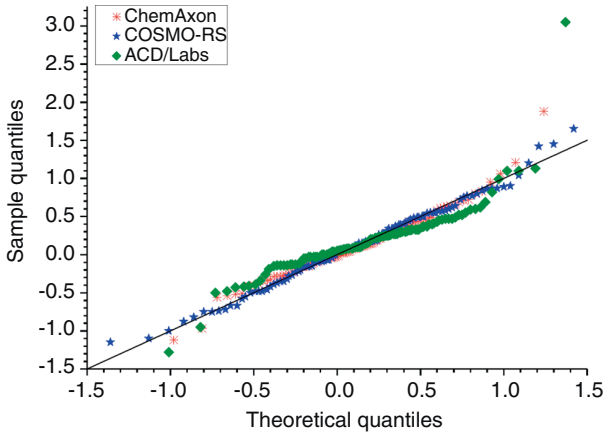

Figure 8

Quantile-Quantile (Q-Q) plots of errors distribution for each of the 3 pK_a prediction methods.

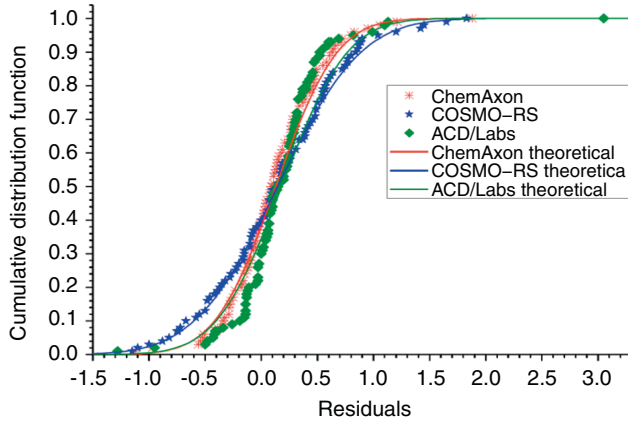

Figure 7

CFD of errors distribution for each of the 3 pK_a prediction methods.

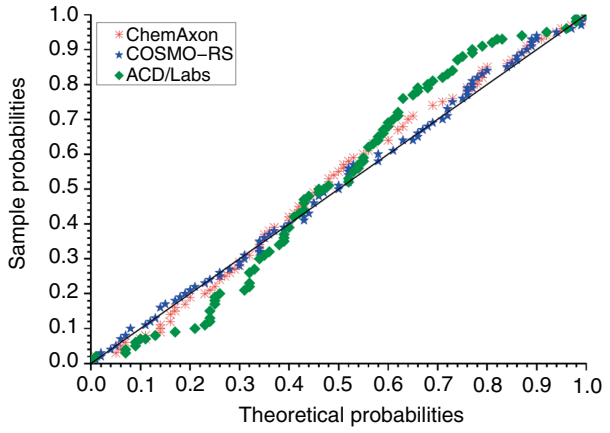

Figure 9

Percentile-Percentile (P-P) plots of errors distribution for each of the 3 pK_a prediction methods.

sample CFD is relatively different compared to its theoretical CFD.

Figures 8 and 9 show respectively the Q-Q and the P-P plots [27] of the prediction errors distribution for each method.

These graphical normality tests (CFD, Q-Q and P-P plots) show that COSMO-RS prediction errors have the closest distribution to the Gaussian distribution. The ChemAxon prediction errors have a slightly less normal distribution, while ACD/Labs method's errors distribution is significantly different from the normal distribution. Several normality tests (*Tab. 5*) confirmed these results.

For each of the studied pK_a prediction methods, several statistical tests on the distribution shown were performed to get the Cullen and Frey graph that enables to have an idea about the position of the observed distribution (red circle point) compared to different theoretical distributions (normal, log-normal, exponential, etc.) and the results of a bootstrap (performed on sample points by randomly taking off one or more points) done 100 times (hollow blue circles) (*Fig. 10-12*).

These confirmed that the nearest theoretical distribution should be the normal distribution, and that

TABLE 4
Values of the average RMSE (Root Mean Square Errors of the differences ΔpK_a between the predicted values and the experimental ones, $\Delta pK_a = pK_a^{calc} - pK_a^{exp}$) for each of the 3 methods used for pK_a prediction (ChemAxon, COSMO-RS and ACD/Labs)

	RMSE ChemAxon	RMSE COSMO-RS	RMSE ACD/Labs	Number of points
Amino-acids (AA)	0.361	0.577	0.407	45
Dipeptides	0.612	0.668	0.669	37
Tripeptides	0.239	0.510	0.259	25
Overall data	0.445 ($r^2 = 0.98$)	0.596 ($r^2 = 0.96$)	0.490 ($r^2 = 0.98$)	107

TABLE 5
Summary of the results of several normality tests on errors distribution for each pK_a prediction method

	Kolmogorov-Smirnov statistic	Cramer-von Mises statistic	Anderson-Darling statistic	Conclusion
ChemAxon	0.06307036 (Not rejected)	0.09646038 (Not rejected)	0.6444582 (Not rejected)	Gaussian distribution
COSMO-RS	0.03971827 (Not rejected)	0.02223085 (Not rejected)	0.2031585 (Not rejected)	Gaussian distribution
ACD/Labs	0.1329101 (Rejected)	0.5802089 (Rejected)	3.523414 (Rejected)	Non-Gaussian distribution

TABLE 6
Values of the average RMSE (Root Mean Square Errors of the differences ΔpK_a between the predicted values and the experimental ones: $\Delta pK_a = pK_a^{calc} - pK_a^{exp}$) for each of the 3 methods used for pK_a prediction (ChemAxon, COSMO-RS and ACD/Labs), after taking into account the influence of ionic strength. The values in parenthesis are those found in literature without any correction to get zero ionic strength pK_a values (see *Tab. 4*)

	RMSE ChemAxon	RMSE COSMO-RS	RMSE ACD/Labs	Number of points
Amino-acids	0.355 (0.361)	0.560 (0.577)	0.424 (0.407)	45
Dipeptides	0.708 (0.612)	0.691 (0.668)	0.830 (0.669)	37
Tripeptides	0.178 (0.239)	0.456 (0.510)	0.208 (0.259)	25
Overall data	0.484 (0.445)	0.588 (0.596)	0.569 (0.490)	107

COSMO-RS errors distribution are the closest to this theoretical distribution.

2.4 Analysis of Several Factors that Can Influence Predicted pK_a Values

2.4.1 Influence of Ionic Strength on pK_a Values

The influence of ionic strength on the experimental pK_a values used has been taken into account using (*Eq. 19*). This correction changes the values of the RMSE of the different prediction methods as shown in Table 6.

Table 6 shows that the correction of ionic strength influence on pK_a decreased the RMSE of the COSMO-RS pK_a values by 0.01 while the ChemAxon and ACD/Labs RMSE increased respectively by about 0.04 and 0.08. That should probably due to the fact that the *ab-initio* calculation performed in the COSMO-RS method enables it to reach a better treatment of electrostatic interactions when compared to other methods using purely empirical parameters. Indeed, the ionic strength influence should only affect the experimental pK_a value, not the performance of a given model. But because ACD/Labs and ChemAxon models are parameterized on a large set of experimental pK_a (probably including non zero ionic strength pK_a values), a small effect of ionic strength should bias their respective RMSE values as illustrated in Table 6.

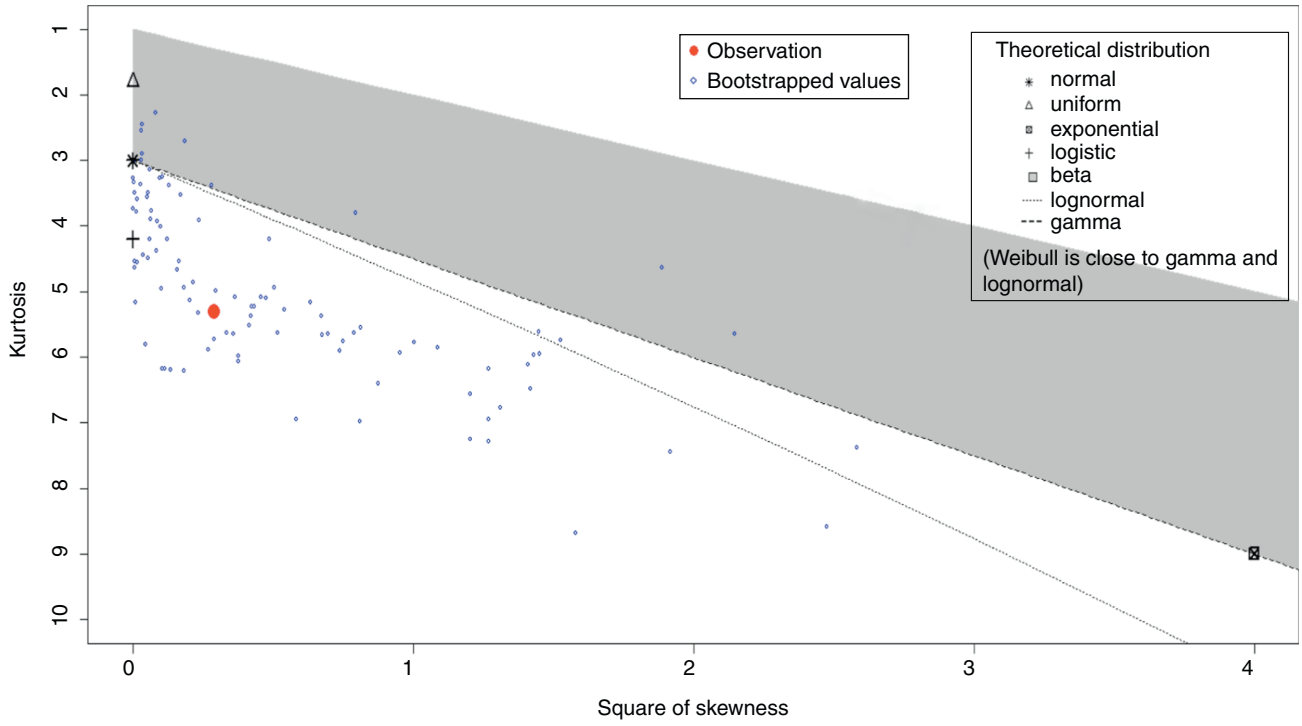

Figure 10

Cullen and Frey graph for the errors distribution of the ChemAxon method.

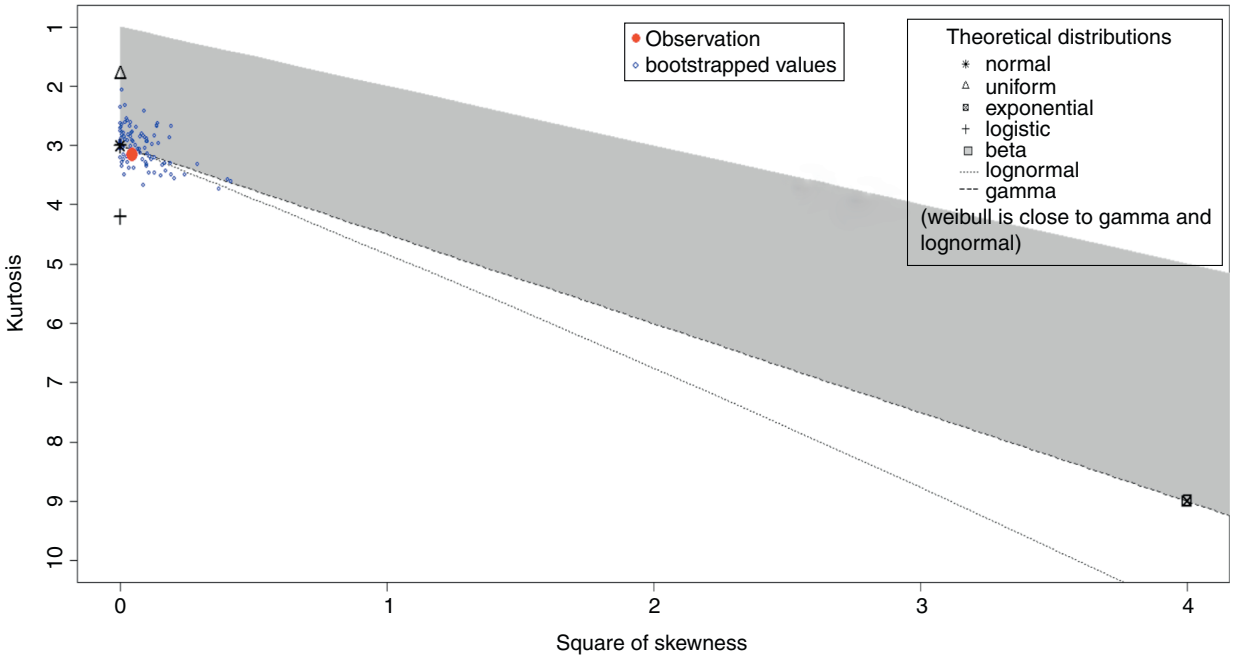

Figure 11

Cullen and Frey graph for the errors distribution of the COSMO-RS method.

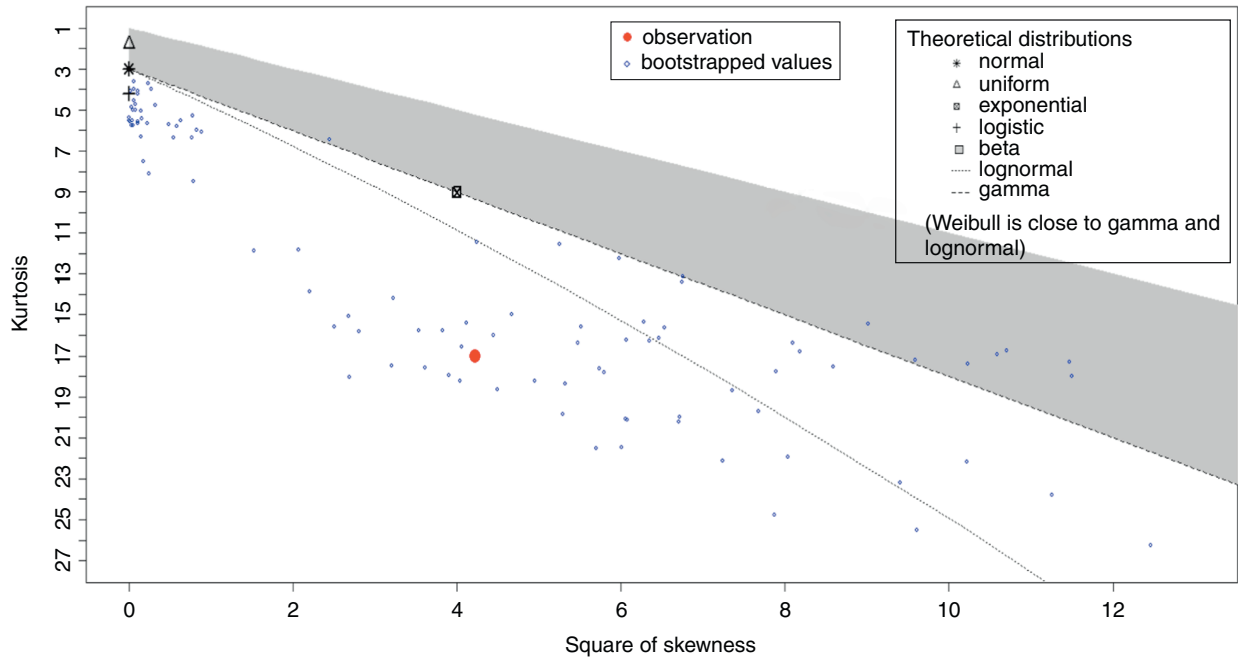

Figure 12

Cullen and Frey graph for the errors distribution of the ACD/Labs method.

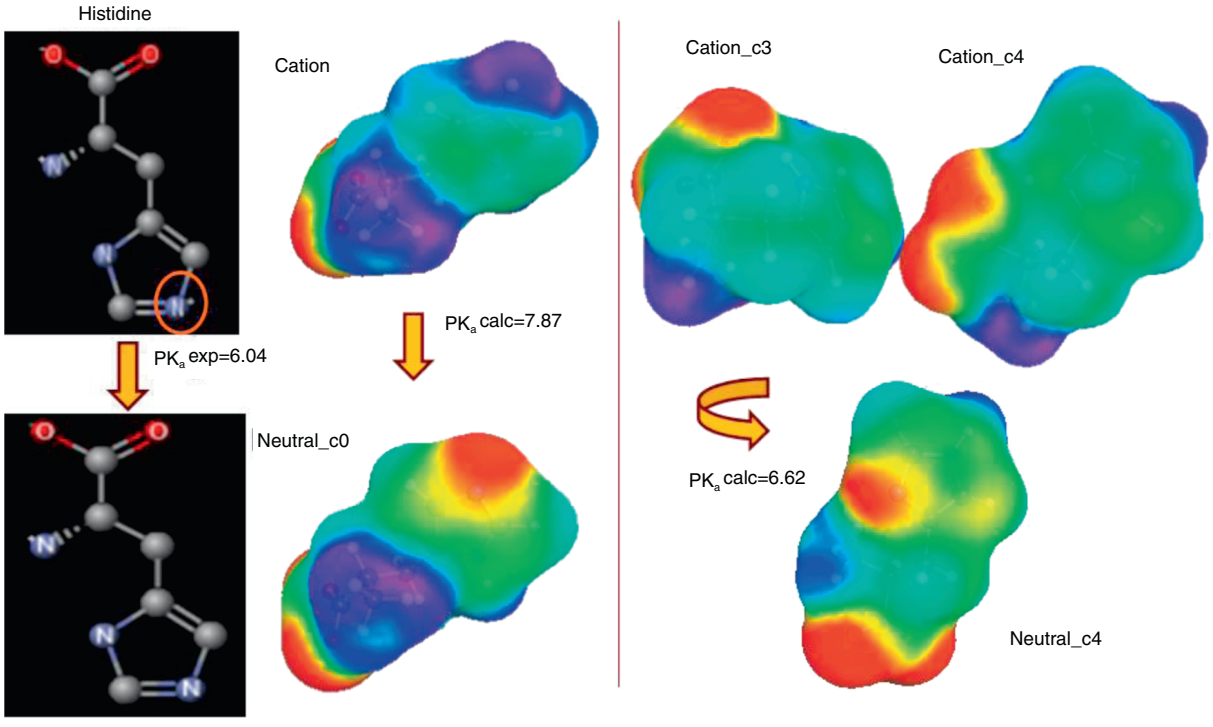

Figure 13

Illustration of the conformers set treatment in the case of histidine molecule. Only the most stable conformations in water are shown on the right (the COSMO-geometries used in the first prediction are shown on the left).

TABLE 7

Values of the average RMSE (Root Mean Square Errors of the differences ΔpK_a between the predicted values and the experimental ones: $\Delta pK_a = pK_a^{calc} - pK_a^{exp}$) for each of the 3 methods used for pK_a prediction (ChemAxon, COSMO-RS and ACD/Labs), after taking into account the conformations of histidine and histidyl-glutamic acid in COSMO-RS calculations. The values in parenthesis show the COSMO-RS pK_a results before this conformers set treatment

	RMSE ChemAxon	RMSE COSMO-RS	RMSE ACD/Labs	Number of points
Amino-acids (AA)	0.361	0.536 (0.577)	0.407	45
Dipeptides	0.612	0.633 (0.668)	0.669	37
Tripeptides	0.239	0.510 (0.510)	0.259	25
Overall data	0.445	0.566 (0.596)	0.490	107

TABLE 8

Values of the average RMSE (Root Mean Square Errors of the differences ΔpK_a between the predicted values and the experimental ones: $\Delta pK_a = pK_a^{calc} - pK_a^{exp}$) for each of the 3 methods used for pK_a prediction (ChemAxon, COSMO-RS and ACD/Labs), after taking into account the conformations of histidine and histidyl-glutamic acid in COSMO-RS calculations and averaging the multiple experimental pK_a data for the same dissociation reaction. The values in parenthesis are those given in Table 4

	RMSE ChemAxon	RMSE COSMO-RS	RMSE ACD/Labs	Number of points
Amino-acids (AA)	0.361 (0.361)	0.536 (0.577)	0.407 (0.407)	45 (45)
Dipeptides	0.622 (0.612)	0.646 (0.668)	0.684 (0.669)	35 (37)
Tripeptides	0.219 (0.239)	0.430 (0.510)	0.273 (0.259)	11 (25)
Overall data	0.468 (0.445)	0.570 (0.596)	0.520 (0.490)	91 (107)

2.4.2 Conformations Treatment Influence on COSMO pK_a Results

The averaged RMSE of the predicted pK_a values for each method compared to experimental results were respectively 0.596 for COSMO-RS, 0.445 for ChemAxon and 0.490 for ACD/Labs.

But, while the ChemAxon and ACD/Labs use a 2D structure to predict pK_a values, the COSMO-RS approach uses a 3D geometry. So the conformations treatment can have a non-negligible influence on COSMO-RS pK_a predicted values. For instance, using the most stable (in water) conformation sets of two of the studied molecules (histidine (*see Fig. 13*) and histidyl-glutamic acid), the RMSE of the COSMO-RS is improved by 0.02 as shown in Table 7.

2.4.3 Multiple Experimental Data Points Influence on pK_a Results

Furthermore, in the experimental pK_a values used, there were some cases where we got several different experimental values listed for the same dissociation center. By averaging these "multiple" experimental data to compare the prediction methods, we got 91 different data points (instead of 107) and the RMSE were slightly different than those got in Table 1.

The combination of a conformations set treatment for 2 molecules (histidine and histidylglutamic acid) and the averaging of multiple pK_a values (for the same dissociation center) decreased the RMSE of the COSMO-RS pK_a values by 0.02 while the ChemAxon and ACD/Labs RMSE increased by approximately 0.03 for each method (*Tab. 8*).

2.5 Discussion

It is well established that Group Contributions Methods (GCM) like ACD/Labs and ChemAxon method are in general more accurate in pK_a calculations. However, they are limited to some chemical families and their respective accuracy depends on the availability of experimental data. Compared to these methods, the main advantage of the COSMO-RS prediction method is that it is fully predictive. Indeed the LFER parameters of the COSMO-RS method were determined on a training set which does not include any of the molecules studied in this paper and the pK_a predicted are quite accurate.

Moreover, ACD/Labs and ChemAxon methods both use a 2D description of molecular structure in their respective algorithms for pK_a calculation. Thus these methods are not able to distinguish conformation treatment effect (which is a 3D effect) while COSMO-RS is able to perform this task (as shown in *Sect. 2.4.2*).

As mentioned earlier, each of the studied methods includes optionally parameterization tools to incorporate "local" data in order to bias predictions. This will increase for sure the prediction results; however these optional facilities were not studied in this paper since we are looking for a predictive tool that is able to predict pK_a of complex structures (that are ubiquitous in foods and biological systems). The COSMO-RS method seems very promising to determine the pK_a of a given molecule in a fully predictive way (especially when the conformations are well treated in calculations) with no available experimental data.

Ideally, a benchmarking of pK_a prediction models would require a universal training set to train all models, and a universal disjoint and similarly the use of diverse test set to compare their prediction. However it is very difficult to perform such task on all of the commercial pK_a prediction utilities used in the present study. With no true benchmarks for pK_a prediction utilities, the only way to identify a superior model is by trusting statistics. The statistics for both training and test data should be separate; unfortunately this is not the case in the present study since we are not able to distinguish the molecules that were used to parameterized ChemAxon and ACD/Labs methods.

All empirically based models should have r^2 closed to 1.0 and RMSE as close to 0.0 as possible over a wide range of compounds. Regarding the r^2 values, each of the studied models should be considered as accurate (r^2 = 0.98 for ChemAxon and ACD/Labs and r^2 = 0.96 for COSMO-RS). The averaged RMSE of the predicted pK_a values for each method compared to experimental results were respectively 0.596 for COSMO-RS, 0.445 for ChemAxon and 0.490 for ACD/Labs. Since, all of these RMSE values are close to 0.0; one can conclude that each model is suitable for pK_a prediction. This statement is confirmed by other statistical analysis (normality tests, bias factor and accuracy factor that are ubiquitous in comparing models in food science).

Moreover, it has been reported that a successful evaluation of seafood spoilage models present a bias factor (B_f) in the range 0.75-1.25. More drastically, a bias factor in the range 0.90-1.05 is considered as good for models dealing with pathogens growth (no 'fail-dangerous' predictions) [25]. The B_f values obtained (on the overall data) in this study are not greater than 1.05 which indicates a very good prediction of the pK_a values by all the models.

Likewise, it is also reported that the best performance that might be expected from a kinetic model encompassing the effect of temperature, pH and aw on growth rate, is $\sim$30%, or an accuracy factor A_f of 1.3. This value is greater than all the A_f values determined within this study, confirming again that all the studied models are reliable.

However, it is difficult to determine the best model because we were not able to distinguish the data used to train each model (especially ChemAxon and ACD/Labs). As suggested by Lee and Crippen [10], it should be interesting to study the performance of a consensus model based on these 3 methods. Since the statistics obtained from a consensus model may not reflect its performance on new data, this kind of study is out of the scope of this paper in which we are looking for a fully predictive model able to treat other products of interest in foods and biological systems.

CONCLUSION

The results presented in this study indicate that all the 3 methods (ChemAxon, ACD/Labs and COSMO-RS) are effective in predicting pK_a values (with a RMSE of about 0.5 pK_a-unit) for compounds of interest in food sciences like amino-acids, dipeptides and tripeptides. Furthermore, it appears that ACD/Labs, ChemAxon and COSMO-RS each can sometimes give large errors. A statistical study of the prediction errors (for this training set) showed that COSMO-RS method has the closest distribution to a normal (or Gaussian) distribution followed by ChemAxon, and that ACD/Labs pK_a errors do not follow a normal law. This was confirmed when analyzing the influence of ionic strength. Since COSMO-RS performs a Quantum Chemistry (QC) calculation on a 3D geometry while the two other methods are using a 2D structure (generated directly from SMILES file), one can expect that conformations treatment has to be taken into account to have a better pK_a prediction. This effect was studied for the case of 2 molecules of our training set and reduced the RMSE of the COSMO-RS method by about 0.02. All these packages include the ability to bias predictions using "local" data but these facilities were not evaluated. ChemAxon's Marvin and ACD/Labs are the fastest tools in term of computer-time. Due to the time-consuming QC calculations preceding its thermodynamics calculations, the COSMO-RS is less fast. But when this calculation is done once for each molecule and ion of interest, the COSMO-RS thermodynamics algorithm takes the same amount of time as ChemAxon and ACD/Labs to perform pK_a predictions.

Regarding these results, COSMO-RS appears as a promising method to predict the pK_a values of molecules of interest in food science with scarce available pK_a values such as peptides.

The final goal of this study is to use the pK_a values in a predictive thermodynamics model for products of interest in food industry.

ACKNOWLEDGMENTS

This work was funded by the Na$^-$ integrated programme (ANR-09-ALIA-013-01) financed by the French National Research Agency. Thanks to Dr. Andreas Klamt and Dr. Fabrice Audonnet for helpful discussions.

REFERENCES

1 Katritzky A.R. (1963) *Physical Methods in Heterocyclic Chemistry*, Academic Press, New York, USA.

2 Perrin D.D. (1981) *pKa Prediction for Organic Acids and Bases*, Dempsey B., Serjeant E.P. (Eds), London, Great Britain, ISBN 0 412 22190 X.

3 Marvin (2011): "Marvin was used for drawing, displaying and characterizing chemical structures, substructures and reactions, *Marvin 5.4.1.1, 2011*, ChemAxon (http://www.chemaxon.com)".

4 Calculator Plugins (2011): "Calculator Plugins were used for structure property prediction and calculation, *Marvin 5.4.1.1, 2011*, ChemAxon (http://www.chemaxon.com)".

5 Harding A.P., Wedge D.C., Popelier P.L.A. (2009) pK_a Prediction from "Quantum Chemical Topology" Descriptors, *J. Chem. Inf. Model* **49**, 1914-1924.

6 Lebert I., Lebert A. (2006) Quantitative prediction of microbial behaviour during food processing using an integrated modelling approach: a review, *Int. J. Refrig.* **29**, 968-984.

7 Manchester J., Walkup G., Rivin O., You Z. (2010) Evaluation of *pKa* Estimation Methods on 211 Druglike Compounds, *J. Chem. Inf. Model.* **50**, 565-571.

8 COSMOthermX (7 December 2011) A Graphical User Interface to the COSMOtherm Program, Tutorial for version C21_0111, *COSMOlogic GmbH & Co. KG*, Leverkusen, Germany.

9 Eckert F., Klamt A. (2010) COSMOtherm, Version C2.1. Release 01.11, *COSMOlogic GmbH & Co. KG*, Leverkusen, Germany.

10 Lee A.C., Crippen G.M. (2009) Predicting pK_a, *J. Chem. Inf. Model* **49**, 2013-2033.

11 ACD/ChemSketch (2006) version 10.01 (Release 10.00), *Advanced Chemistry Development, Inc.*, Toronto, ON, Canada, www.acdlabs.com.

12 Klamt A., Schüürmann G. (1993) COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expression for the Screening Energy and its Gradients, *J. Chem. Soc. Perkin Trans.* **2**, 799.

13 Klamt A., Jonas V., Bürger T., Lohrenz J.W.C. (1998) Refinement and Parameterization of COSMO-RS, *J. Phys. Chem. A* **102**, 5074.

14 Perrin D.D. (1965) *Dissociation constants of organic bases in aqueous solution*, International Union of Pure and Applied Chemistry, Butterworths, London, England.

15 Perrin D.D. (1972) *Dissociation constants of organic bases in aqueous solution*, supplement 1972, International Union of Pure and Applied Chemistry, Butterworths, London, England, ISBN 0 408 70408 X.

16 Szegezdi J., Csizmadia F. (2004) Prediction of dissociation constant using microconstants, *27th ACS (American Chemical Society) National Meeting*, Anaheim, California, 28 March-1 April.

17 Szegezdi J., Csizmadia F. (2007) Method for calculating *pKa* values of small and large molecules, *233rd ACS (American Chemical Society) National Meeting*, IL, Chicago, 25-29 March.

18 ACD/pK_a DB (2006) version 10.0 for Microsoft Windows, Reference Manual, Comprehensive Interface Description, *Advanced Chemistry Development, Inc.*, Toronto, ON, Canada, www.acdlabs.com.

19 Eckert F., Klamt A. (2002) Fast Solvent Screening *via* Quantum Chemistry: COSMO-RS Approach, *AIChE J.* **48**, 369.

20 Klamt A., Eckert F. (2000) COSMO-RS: a novel and efficient method for the *a priori* prediction of thermophysical data of liquids, *Fluid Phase Equilib.* **172**, 43.

21 Klamt A., Eckert F., Diedenhofen M. (2003) First principles calculations of aqueous pK_a values for organic and inorganic acids using COSMO-RS reveal an inconsistency in the slope of the pK_a scale, *J. Phys. Chem. A* **107**, 9380-9386.

22 Eckert F., Diedenhofen M., Klamt A. (2010) Towards a first principles prediction of *pKa*: COSMO-RS and the cluster-continuum approach, *Molec. Phys.* **108**, 3-4, 229-241.

23 Weininger D. (1988) SMILES, A Chemical Language and Information System, 1. Introduction to Methodolody and Encoding Rules, *J. Chem. Inf. Model.* **28**, 31-36.

24 Neumeyer K., Ross T., Thomson G., McMeekin T.A. (1997) Validation of a model describing the effects of temperature and water activity on the growth of psychrotrophic pseudomonads, *Int. J. Food Microbiol.* **38**, 55-63.

25 Mellefont L.A., McMeekin T.A., Ross T. (2003) Performance evaluation of a model describing the effects of temperature, water activity, pH and lactic acid concentration on the growth of *Escherichia coli*, *Int. J. Food Microbiol.* **82**, 45-58.

26 Delignette-Muller M.L., Pouillot R., Denis J.-B., Dutang C. (2010) Fitdistrplus: help to fit of a parametric distribution to non-censored or censored data, R package version 0.1-3, http://CRAN.R-project.org/package=fitdistrplus.

27 Dalgaard P. (2002) *Introductory Statistics with R (Statistics and Computing)*, Springer-Verlag, New York, USA, ISBN 0-387-95475-9.

28 The R software: R Development Core Team (2011) R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/.

29 Ould-Moulaye C.B. (1998) Calcul des propriétés de formation en solution aqueuse des composés impliqués dans les procédés microbiologiques et alimentaires - prédiction et réconciliation de données - modélisation des équilibres chimiques et des équilibres entre phases, *PhD Thesis/Thèse*, Université Blaise Pascal.

30 Goldberg R.N. (1981) Evaluated activity and osmotic coefficients for aqueous solutions: Thirty-six uni-bivalent electrolytes, *J. Phys. Chem. Ref. Data* **10**, 671.

Final manuscript received in December 2012
Published online in May 2013