J Comput Aided Mol Des (2010) 24:357–360
DOI 10.1007/s10822-010-9354-4

Blind prediction test of free energies of hydration with COSMO-RS

Andreas Klamt · Michael Diedenhofen

Received: 16 November 2009 / Accepted: 30 March 2010 / Published online: 10 April 2010

Abstract The COSMO-RS method, a combination of the quantum chemical dielectric continuum solvation model COSMO with COSMO-RS, a statistical thermodynamics treatment of surface interactions, simulations, has been used for the direct, blind prediction of free energies of hydration within the SAMPL challenge. Straight application of the latest version of the COSMO*therm* implementation in combination with a rigorous conformational sampling yielded a predictive accuracy of 1.56 kcal/mol (RMSE) for the 23 compounds of the blind prediction dataset. Due to the uncertainties of the extrapolations and assumptions involved in the derivation of the experimental data, the accuracy of the predicted data may be considered to be within the noise level of the experimental data.

Keywords Solvation COSMO-RS

Introduction

Blind tests are important for the evaluation of the predictive power of computational models. Therefore we take every opportunity to participate in blind tests and the second SAMPL challenge [1] was a welcome opportunity to validate the predictive capabilities of our COSMO*therm* [2] implementation of the COSMO-RS method [3–5] with respect to the free energy of hydration $\Delta G^X_{\mathrm{hydr}}$ and with respect to tautomerization free energies, i.e. $\Delta G^X_{\mathrm{taut}}$. COSMO-RS is a combination of the dielectric continuum solvation model COSMO [6] with a statistical thermodynamics treatment of surface interactions. As a result, COSMO-RS predicts the chemical potential of almost arbitrary compounds in almost any pure or mixed solvent in a wide temperature range, just based on quantum chemical COSMO calculations for solutes and solvent molecules. Hence the subject of the present challenge, i.e. the prediction of free energies in the aqueous phase at infinite dilution of the solute is just a special case of the much broader applicable COSMO-RS method.

The SAMPLE2 challenge consisted of two parts, one for gas–water transfer energies, and the other addressing tautomerization equilibria in aqueous solution. The COSMO-RS predictions for the latter part will be discussed in a forthcoming paper in a special issue on tautomerization [7]. Hence here we only discuss the first part of the SAMPL2 challenge, the XFER part, which is focussed on the free energies of hydration $\Delta G^X_{\mathrm{hydr}}$, or equivalently on water–air Henry's law constants. This is the most considered quantity of solvation models. Quantum chemically based solvation models meanwhile claim to be able to predict $\Delta G^X_{\mathrm{hydr}}$ for small and medium sized, neutral organic compounds with an accuracy of ~1 kcal/mol (RMSE) down to 0.5 kcal/mol for SMx [8] or COSMO-RS [3–5], as was recently proven [9] on the large training dataset of the SM8-model containing overall 2346 solvation free energies, 284 of which being hydration energies. Typically quantum chemically based solvation methods nowadays use DFT as basis, which have proven to yield rather robust electrostatic properties. While DFT is known to much less accurate for reaction energies, solvation energies can be calculated with a much higher accuracy because usually the molecule stays rather the same during solvation with no major electronic

A. Klamt (✉) · M. Diedenhofen
COSMOlogic GmbH&CoKG, Burscheider Str. 515, 51381
Leverkusen, Germany
e-mail: klamt@cosmologic.de

A. Klamt
Institute of Physical and Theoretical Chemistry, University
of Regensburg, 93040 Regensburg, Germany

changes or bond re-organizations. Nevertheless, it may well be important to search for different conformations in the gas phase and in the liquid phase, since the a change from less polar conformations to more polar conformation may well happen during hydration and may be of significant influence on the solvation free energy.

Data

The datasets for both parts of the SAMPL2 challenge, XFER2 and TAUT2, have been collected and assembled by the organizers of the challenge are described in detail in the introductory article of this special issue [1].

The XFER2 dataset is split into 3 parts. 8 compounds were given as explanatory data together with experimental data, which may be used—but have not in our case—for calibration of the method, 23 compounds, called obscure data set, were given for blind prediction and 10 compounds, the so-called investigatory subset, were given for blind prediction without any availability of experimental data.

Methods

For the XFER part of SAMPL2, submitted as entry 319, we applied exactly the same protocol and methods as in the SAMPL08 challenge [10], with the exception of using the latest COSMO*therm* parameterization, BP-TZVP_C21_0108 [2]. Hence we only give a short description of the method here. Starting with a conformational search for low-energy conformations in the gas phase and in the conductor reference state (COSMO) we finally perform BP-TZVP [11–14] gas phase and COSMO calculations with the TURBOMOLE program [15, 16] for the complete set of conformations with full geometry optimization in both reference phases. The gas phase energies and the COSMO files of the solutes, together with the COSMO file of water, are then post processed by the COSMO*therm* program. COSMO*therm* calculates independently Boltzmann weighted free energies for the gas phase and for the aqueous phase and returns the difference of these as the free energy of hydration of the solute. It might be noted that the gas phase energies and COSMO files for about 50% of the XFER09 compounds were already available in the COSMO*base* database [17]. For the others the average total calculation time per compound was about 1 CPU hour on a 2.4 GHz processor.

Results and discussion

While we have predicted the hydration free energies for all compounds or the XFER09 part, we only consider here the explanatory and obscure subsets, because only for these experimental data are available for comparison. Table 1 and Fig. 1 give the experimental data vs. the COSMO*therm* predictions.

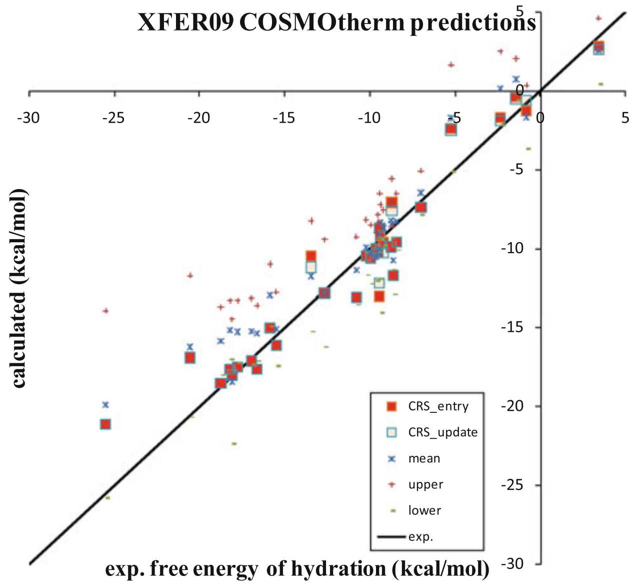

Fig. 1 COSMOtherm predictions vs. exp. Data for the XFER2 data set. The mean value of all entries is marked as *. The spread of the predicted values of all entries, i.e. the mean ± the standard deviation, are marked by the + and − signs

The root mean square deviation (RMSD) of the COSMO*therm* predictions from the experimental values is 1.69 (1.54/2.05) kcal/mol for the entire set of the 31 compounds. The partial results for the obscure dataset, i.e. blind test, and explanatory dataset respectively, are given in brackets. Due to some missing gas phase energy files, which were not correctly copied into the project directory, some of the gas phase energies had been estimated by the COSMO*therm* program based on the COSMO solvation energies, instead of using independent gas phase energies. Therefore we have added the correctly updated COSMO*therm* results as CRS_update. The RMSD performance thus changed to 1.59 (1.56/1.69) kcal/mol. Fortunately the data of the blind test are only slightly effected, while a significant improvement for the explanatory data is achieved by this update. Thus at the end we have an overall RMSD accuracy of about 1.6 kcal/mol, while the RMSD of the other entries ranges from 2.0 to more than 7 kcal/mol, with an average of ~3 kcal/mol.

From Fig. 1 it is apparent that the RMSD does substantially result from the large deviations for the sugar molecules glucose and xylose, and the structurally somewhat similar glycerol. COSMO*therm* predicts substantially less negative values for the hydration free energies of these compounds than the experimental values recommended by the organizers. It is striking, that COSMO*therm* in these

Table 1 Recommended experimental, COSMO*therm* predictions, and mean values and ranges of all submitted predictions for the explanatory (*italics*) and obscure datasets of XFER09

Compound	Exp.	CRS_entry	CRS_update	Mean	Upper	Lower
4-nitroaniline	*−9.45*	*−13.0*	*−12.2*	*−10.1*	*−8.3*	*−11.9*
Trimethyl_phosphate	*−8.70*	*−7.1*	*−7.6*	*−8.5*	*−5.6*	*−11.3*
Glycerol	*−13.43*	*−10.5*	*−11.2*	*−11.8*	*−8.3*	*−15.3*
Pentachloronitrobenzene	*−5.22*	*−2.4*	*−2.5*	*−1.8*	*1.6*	*−5.2*
Hexachlorobenzene	*−2.33*	*−1.7*	*−1.9*	*0.1*	*2.5*	*−2.2*
Hexachloroethane	*−1.41*	*−0.4*	*−0.5*	*0.7*	*2.0*	*−0.6*
Trimethyl_orthotrifluoroacet	*−0.80*	*−1.3*	*−0.6*	*−1.7*	*0.3*	*−3.7*
Octafluorocyclobutane	*3.43*	*2.8*	*2.6*	*2.5*	*4.6*	*0.4*
Cyanuric_acid	−18.06	−18.0	−18.0	−18.4	−14.5	−22.3
D-glucose	−25.47	−21.1	−21.1	−19.9	−13.9	−25.8
D-xylose	−20.52	−16.9	−16.9	−16.2	−11.7	−20.6
5-iodouracil	−18.72	−18.5	−18.5	−15.8	−13.7	−18.0
5-bromouracil	−18.17	−17.6	−17.6	−15.1	−13.3	−17.0
5-chlorouracil	−17.74	−17.5	−17.5	−15.2	−13.3	−17.2
5-flurouracil	−16.92	−17.1	−17.1	−15.2	−13.2	−17.3
Uracil	−16.59	−17.6	−17.6	−15.3	−13.6	−17.1
6-chlorouracil	−15.83	−15.0	−15.0	−13.0	−11.0	−14.9
5-trifluoromethyluracil	−15.46	−16.1	−16.1	−15.1	−12.8	−17.4
Caffeine	−12.64	−12.8	−12.8	−12.8	−9.4	−16.2
Ketoprofen_(racemic)	−10.78	−13.1	−13.1	−11.4	−9.3	−13.5
Naproxen	−10.21	−10.5	−10.5	−9.9	−8.2	−11.7
Acetylsalicylic_acid	−9.94	−10.6	−10.6	−10.4	−8.5	−12.2
Phthalimide	−9.61	−10.0	−10.0	−10.5	−9.0	−12.0
Methyl_paraben	−9.51	−10.3	−10.3	−9.1	−7.9	−10.2
Diflunisal	−9.40	−8.7	−8.7	−10.3	−6.5	−14.0
Propyl_paraben	−9.37	−9.2	−9.7	−8.4	−7.2	−9.5
Ethyl_paraben	−9.20	−9.6	−10.3	−8.7	−7.6	−9.8
Butyl_paraben	−8.72	−9.9	−9.9	−8.2	−7.1	−9.3
Sulfolane	−8.61	−11.7	−11.7	−10.8	−8.6	−12.9
Flurbiprofen_(racemic)	−8.42	−9.6	−9.6	−8.3	−6.5	−10.1
Ibuprofen_(racemic)	−7.00	−7.4	−7.4	−6.5	−5.1	−7.9

cases is in very good agreement with the mean value of all predicted values, while the recommended values are at the lower end of the variance of the predictions. Although it cannot be granted that there is no reason for a systematic error in all the prediction methods, this seems to be unlikely. Since the experimental values of these compounds are derived by extrapolation from measurements of very low vapor pressures at elevated temperatures, combined with hardly to quantify solubility or activity data of these very soluble compounds in water, it appears to be an open question, whether the deviation in these cases arises from prediction error or from the uncertainties in the experimental data. The same is true for pentachloronitrobenzene, although there is no structural relation to the sugars in this case. Nevertheless it should be noted, that a dimensionless Henry's law constant of 0.025, corresponding to a free energy of solvation of −2.17 kcal/mol is reported on the web site of GSI Environmental Inc. This value would be in best agreement with our predictions and with the mean of all predictions.

The doubts in the accuracy of the recommended experimental data are further supported by the changes of the recommendations for cyanuric acid from initially −6.44 kcal/mol to −18.06 kcal/mol, and for glycerol from −8.26 kcal/mol to now −13.43 kcal/mol between the original recommendations and a later update arising from a revision of the experimental data, which was caused by a discussion based on the large deviation of the recommended values from the predictions. A detailed analysis of available experimental data for sulfolane [18] also suggests that the experimental value most likely is by about 0.7 kcal/mol more negative than the recommendation and by that at least closer to the mean of the predictions and to the COSMO*therm* prediction.

The solvation energies in the explanatory dataset were considered as unusual or unexpected. From the perspective of our COSMO*therm* predicitions these data are reproduced with almost the same accuracy as those from the remaining dataset, and hence they do not appear as unusual or unexpected.

Conclusions

COSMO-RS in its COSMO*therm* implementation, combined with a careful conformational sampling for gas phase and aqueous phase, did achieve the most accurate predictions for the free energies of hydration in the SAMPL2 challenge. Since it was recently shown in a very broad validation study on 2346 compounds, accuracies of ~0.5 kcal/mol can be achieved by COSMO*therm* for this quantity, it may be that the considerably larger deviations of about 1.5 kcal/mol found in this challenge is partly caused by uncertainties of the experimental data, which to a large portion are derived from downward extrapolations of vapour pressures and/or solubility data over large temperature ranges.

References

1. For details see introductory paper of this special issue: Skillman, Geballe, Guthrie, JCAMD, 2010, Volume 24
2. Eckert F, Klamt A (2008) COSMO*therm*, Version C2.1-Revision 01.08; COSMOlogic GmbH&CoKG, Leverkusen, Germany (2008); see also URL: http://www.cosmologic.de
3. Klamt A (1995) J Phys Chem 99:2224
4. Klamt A, Jonas V, Bürger T, Lohrenz JCW (1998) J Phys Chem 102:5074
5. Klamt A (2005) COSMO-RS: from quantum chemistry to fluid phase thermodynamics and drug design. Elsevier, Amsterdam
6. Klamt A, Schüürmann G (1993) J Chem Soc Perkin Trans 2:799
7. Klamt A, Eckert F, Diedenhofen M (in print) JCAMD 2010, 24 (in print)
8. Marenich AV, Olson RM, Kelly CP, Cramer CJ, Truhlar DG (2007) J Chem Theory Comput 3:2011–2033
9. Klamt A, Mennucci B, Tomasi J, Barone V, Curutchet C, Orozco M, Luque FJ (2009) Acc Chem Res 42:489–492
10. Klamt A, Eckert F, Diedenhofen M (2009) J Phys Chem B 113:4508–4510
11. Becke AD (1988) Phys Rev A 38:3098
12. Perdew JP (1986) Phys Rev B 33:8822
13. Schäfer A, Huber C, Ahlrichs R (1994) J Chem Phys 100:5829
14. Eichkorn K, Weigend F, Treutler O, Ahlrichs R (1997) Theor Chem Acc 97:119
15. Eichkorn K, Treutler O, Öhm H, Häser M, Ahlrichs R (1995) Chem Phys Lett 242:652
16. TURBOMOLE, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007; see also URL: http://www.turbomole.com
17. Eckert F, Klamt A (2008) COSMO*base*, Version C2.1-Revision 01.08; COSMOlogic GmbH&CoKG, Leverkusen, Germany (2008); see also URL: http://www.cosmologic.de
18. Peter Guthrie, University of Western Ontario, private communication