ELSEVIER

Journal of Molecular Graphics and Modelling 20 (2001) 123–131

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Fast prediction of hydration free energies from molecular interaction fields

Robert Jäger[1], Stefan M. Kast*

Institut für Physikalische Chemie, Technische Universität Darmstadt, Petersenstr. 20, 64287 Darmstadt, Germany

Received 30 November 2000; received in revised form 24 April 2001; accepted 1 May 2001

Abstract

A novel empirical model is presented that allows the fast computation of hydration free energies with high accuracy. The linear model is based upon the separation of the free energy of hydration into a cavity and an interaction term. The cavity contribution is modeled as a linear combination of molecular volume and surface terms. The interaction part is derived from the statistical three-dimensional (3D) free energy density and is modeled approximately as a molecular interaction field using the program GRID. A compression scheme is employed to represent this 3D information on the molecular surface by means of a linear combination of surface functions. A set of 81 small organic molecules with known experimental hydration free energies is used to determine the coefficients of the linear model by least squares regression. The fit is statistically significant yielding a correlation coefficient of 0.99, a root mean square error of 0.27 kcal/mol for the 81 molecules belonging to the training set, and 0.63 kcal/mol for an independent test set of 10 molecules.

Keywords: Solvation; Hydration free energy; GRID; QSAR; Molecular surface

1. Introduction

Fast and accurate prediction of solvation free energies is the focus of various modern research fields, such as drug design, drug disposition, and many molecular recognition phenomena. For instance, the standard free energy of hydration is a key quantity for the construction of thermodynamic cycles to compute binding constants. Moreover, it is directly related to the Henry's law constant allowing (together with activity coefficients) for the estimation of solubility which is an important descriptor for many models predicting bioavailability.

Several computational approaches with different levels of complexity for calculating solvation free energies have been established over the past several decades. These approaches comprise methods such as molecular dynamics or Monte Carlo simulation [1–5], mixed quantum/molecular mechanics techniques [6], and the broad field of implicit (continuum) solvent models that use either classical, quantum-mechanical, or hybrid methods [7–12]. The quantum-mechanical self-consistent reaction field method [13] was applied by Luque et al. [14] in an attempt to devise a fractional description of the free energy of solvation based on the solvent exposed surface of the solute. The benefit of this approach is the insight one gains regarding the influence of structural properties on the solvation process. Correction terms required in many fractional treatments of the transfer free energy [15] now become unnecessary.

Still the computational effort of the above methods is substantial. This has triggered the development of more empirical approaches, particular in the context of screening and processing vast compound libraries. Several methods based on empirical models already exist [16,17]. No et al. [16] have used the concept of a hydration free energy density modeled as a linear combination of physical molecular properties such as atomic charge, atomic polarizability, and a dispersion energy coefficient to predict successfully the free energy of hydration. Viswanadhan et al. [17] find a higher predictive capability from a group contribution approach compared to a quantum-mechanical solvation model.

In this work, we establish a linear model based on a condensed surface representation of hydration free energy density by combining both the concept of a hydration free energy density and the group contribution method. As an intermediate step toward this model description we generate a molecular interaction field to approximate the interaction part of the hydration free energy density. Such molecular interaction fields play an essential part in many 3D-QSAR

* Corresponding author. Tel.: +49-6151-165397; fax: +49-6151-164298.
E-mail addresses: robertm.jaeger@aventis.com (R. Jäger), kast@pc.chemie.tu-darmstadt.de (S.M. Kast).
[1] Present address: Aventis Pharma Deutschland GmbH, Geb. 838, 65926 Frankfurt, Germany.

1093-3263/01/$ – see front matter
PII: S1093-3263(01)00108-5

[15,18,19] and CoMFA [20] applications. One of the most popular approaches to generate such molecular fields with respect to steric and electrostatic interactions is the program GRID by Goodford [21]. Its concept is based on the definition of a variety of molecular fragments, termed *probes*, resembling small organic functional groups, whose interaction potential is computed for all points of a grid that encloses the target structure under investigation. GRID has proved valuable for many ligand and de novo design applications as well as molecular docking strategies [22–25].

This article presents a parameterization strategy that combines local statistical and global thermodynamic properties of the molecule. An optimal balance between these reference properties is found, yielding an empirical model with strong predictive power. Certain preliminary investigations together with a detailed description of the model itself are presented. After successful parameterization, the resulting small set of parameters allows for the rapid calculation of hydration free energies with the input of the three-dimensional (3D) structure of the compound.

2. Statistical basis: three-dimensional hydration free energy density

The thermodynamics of solvation processes in terms of enthalpic and entropic contributions is described by the Gibbs or Helmholtz free energy of solvation. In what follows, we make the distinction between Gibbs and Helmholtz free energies of solvation only where explicitly needed. In the absence of an external field, i.e. for a uniform fluid, the chemical potential or free energy of solvation of a particle at infinite dilution is related to the three-dimensional free energy density (3D-FED), ρ_{solv}, as

$$\Delta G_{\mathrm{solv}} = \int_V \rho_{\mathrm{solv}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \tag{1}$$

where V is the volume and $\boldsymbol{r}$ a spatial vector. This relation has its origin in early work on statistical density functional theory [26]. For a simple liquid with pair potentials $V(\lambda)$ and density ρ, we define

$$\rho_{\mathrm{solv}}(\boldsymbol{r}) = \rho \int_0^1 g(\lambda, \boldsymbol{r}) \left(\frac{\partial V(\lambda, \boldsymbol{r})}{\partial \lambda} \right) \mathrm{d}\lambda \tag{2}$$

where g is the pair distribution function and λ a coupling parameter governing $V(\lambda = 0) = 0$ and $V(\lambda = 1) = V$ (extension to multi-component and molecular fluids is straightforward). Inserting this expression into Eq. (1), we recover the usual thermodynamic integration formula [1]

$$\Delta G_{\mathrm{solv}} = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{3}$$

Eq. (2) can be evaluated by a variety of means. Direct molecular simulation would require to compute free energy derivatives along with 3D distribution functions for a number of values of the coupling coordinate. The computational effort is, however, huge. Less time consuming albeit approximate schemes can be employed like for instance integral equation theory. In particular, the three-dimensional reference interaction site model (3D-RISM) equations yield the spatial correlation functions together with the corresponding 3D free energy density [27]. This approach is still too slow for screening large numbers of compounds; furthermore the systematic error of the approximate treatment is difficult to assess although significant progress in this direction has been made recently [28]. Another attempt to reduce numerical work for estimating 3D solvation properties around molecular solutes has been presented by Hummer and Soumpasis [29,30]. These authors approximate the solvent structure with remarkable success by truncating the potential of mean force expansion at the triplet level in the site–site correlation functions, augmented with simulation data for numerical evaluation. To use this method for estimating free energies densities would, however, require extensive molecular simulations for establishing solute–solvent correlation function databases.

3. Model I: the GRID approximation to the 3D-FED

In an alternative approach to simplify the computation, the solvation process can be understood in terms of two successive steps, namely (a) the creation of a cavity according to the size of the solute molecule and (b) the insertion of the solute by switching on interactions with its environment (charging). This leads to a subdivision of the free energy of solvation in terms of two parts

$$\Delta G_{\mathrm{solv}} = \Delta G_{\mathrm{cav}} + \Delta G_{\mathrm{int}} \tag{4}$$

where the subscripts 'cav' and 'int' denote cavity and interaction contribution, respectively. Most of the difference between Gibbs and Helmholtz free energies, the pressure volume work, is contained in the cavitation term, the interaction part is, therefore, almost independent of the chosen ensemble.

As a consequence of such a partitioning of the 3D-FED, one must carefully divide the total volume V_{total} into the two regions pertaining to the cavity and to the interaction region, respectively. Choosing the solvent accessible surface [31] as a reasonable border separating cavity and interaction regions the volume integral of the free energy density becomes, in analogy to Eq. (4)

$$\int_{V_{\mathrm{total}}} \rho_{\mathrm{solv}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} = \int_{V_{\mathrm{cav}}} \rho_{\mathrm{cav}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} + \int_{V_{\mathrm{int}}} \rho_{\mathrm{int}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \tag{5}$$

In the following, the term *molecular surface* is taken as the solvent accessible surface described by Conolly [31]. The volume enclosed by this solvent accessible surface corresponds to the cavity volume, V_{cav}, in Eq. (5). The volume outside the molecular surface defines the interaction region V_{int}.

In our approach, the interaction part of the 3D-FED, ρ_{int}, is approximated by the molecular interaction field generated from GRID. This program is well suited for two reasons: the fast generation of interaction fields, i.e. 40 CPU seconds are needed on a MIPS R5000 processor (150 MHz) for a molecule with 145 atoms, and the reliable internal force field that contains both enthalpic and entropic contributions. For the modeling of the cavity contribution a linear combination of molecular surface and volume terms is applied [16,32]. Therefore, ΔG_{cav} and ΔG_{int} become

$$\Delta G_{\text{cav}} = \int_{V_{\text{cav}}} \rho_{\text{cav}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \approx \beta S + \gamma V + \text{const.} \tag{6}$$

$$\Delta G_{\text{int}} = \int_{V_{\text{int}}} \rho_{\text{int}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \approx \alpha \sum_m \varepsilon_m \tag{7}$$

where S is the molecular surface, V the volume enclosed by this surface, ε_m the energy value of grid point m, α, β, γ are the parameters, and const. is a residual contribution due to a small cavity. We now focus on the *hydration* free energy, ΔG_{hyd}. In this case, the radius of the solvent is much smaller than the radius of the solute, so the constant term in the linear combination is neglected. The total linear model, therefore reads

$$\Delta G_{\text{hyd}} = \alpha \sum_m \varepsilon_m + \beta S + \gamma V \tag{8}$$

We used the water probe supplied by GRID for a set of 81 molecules with known experimental hydration free energies taken from No et al. [16]. The total energy for a grid point m, ε_m, is calculated by GRID from the sum of Lennard–Jones, hydrogen bonding, and Coulomb potential functions as well as an additional entropy term that accounts for hydrophobic hydration effects. Although the entropy estimation is rather simple in nature, it allows us to interpret the resulting interaction field, divided by the volume element the grid point occupies, as a free energy density.

The clearance of the grid was set to 4 Å with a spacing of 0.5 Å for the grid points. This gives an average number of approximately 17,000 grid points per molecule. The summation according to Eq. (7) was carried out for each molecule over all grid points m with a *negative* energy value. The molecular surface and volume were computed [31] for each molecule, and the parameters α, β, and γ were then determined by linear least squares regression. The fit is significant ($F = 285.16$) yielding a correlation coefficient $R^2 = 0.92$ with a mean error of $\sigma = 1.15$ kcal/mol. Fig. 1 shows the predicted response values plotted against the experimental free energies of hydration.

After successfully applying the molecular interaction field to approximate the interaction part of the hydration free energy, further optimization of the model was pursued in two steps. First, the grid information can be compressed to a surface-based representation thereby improving computational speed. Then, the parameters of the compressed form can be optimized with respect to both local grid data and global thermodynamic information. The first aspect is described in the next part of this paper, the second is presented in Section 5.

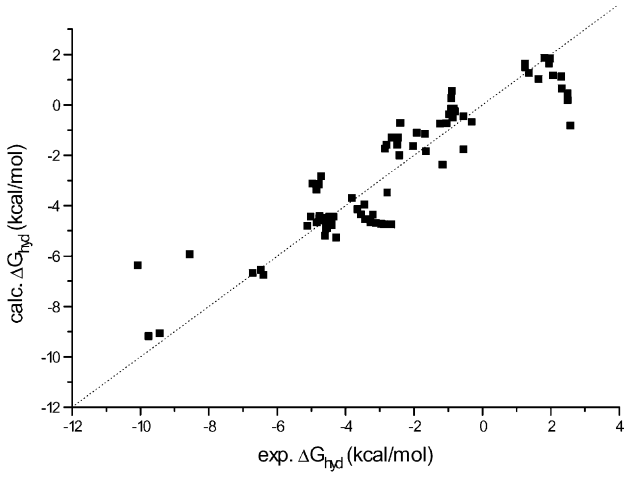

Fig. 1. Calculated free energies of hydration vs. experimental values, as derived from the simple GRID model I (Eq. (8)).

4. Model II: expansion of the 3D-FED in surface-based functions

The GRID-based interaction part of the hydration free energy is used as input in a compression scheme allowing for even faster computation retaining as much information as possible. Starting from the molecular interaction field, the 3D-FED is represented as a linear combination of a set of basis functions, f_j, located on distinct patches, i, on the molecular surface. These functions are chosen in a manner that combines an exponential distance dependence with a spherical harmonics expansion, i.e.

$$f_j(\boldsymbol{r} - \boldsymbol{r}_i) = P^j(\cos\vartheta)\mathrm{e}^{-k|\boldsymbol{r}-\boldsymbol{r}_i|} \tag{9}$$

with P^j being the Legendre polynomial of degree j

$$P^j(\cos\vartheta) = \frac{1}{2^j\, j!}\frac{\mathrm{d}^j}{\mathrm{d}\cos\vartheta^j}(\cos^2\vartheta - 1)^j \tag{10}$$

where $\boldsymbol{r}_i$ is a reference point on patch i. The reference point is the intersection of the mean normal vector of a patch, $\boldsymbol{n}_i$, and its surface, ϑ denotes the angle between the normal vector $\boldsymbol{n}_i$ and vector $\boldsymbol{r} - \boldsymbol{r}_i$. The interaction part of the 3D-FED is then written as

$$\rho_{\text{int}}(\boldsymbol{r}) = \alpha\frac{\varepsilon(\boldsymbol{r})}{\Delta V} \approx \alpha \sum_i \sum_j a_{ij} f_j(\boldsymbol{r} - \boldsymbol{r}_i) \tag{11}$$

where $\varepsilon(\boldsymbol{r})$ is the grid value from the water probe at position $\boldsymbol{r}$, ΔV the associated volume element of that grid point, and a_{ij} represents the expansion coefficients with indices i and j as defined above. The arrangement of a particular surface patch and the determination of the distance $|\boldsymbol{r} - \boldsymbol{r}_i|$ and angle ϑ of a given point in space is illustrated in Fig. 2.

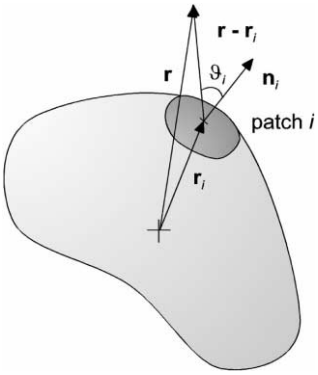

Fig. 2. Schematic representation of the vectors corresponding to a surface patch and the basis functions. Also indicated is the normal vector of a patch i, $\boldsymbol{n}_i$, and the angle ϑ_i between normal vector and distance vector $\boldsymbol{r} - \boldsymbol{r}_i$.

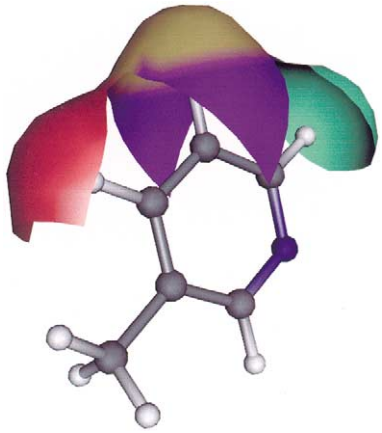

Fig. 3. Patch distribution of 3-methylpyridine. The green, red, and brown patches belong to the underlying hydrogen atoms, the two blue patches correspond to the carbon atom in 5-position.

The interaction part of the free energy is then given by integrating Eq. (11)

$$\Delta G_{\text{int}} = \int_{V_{\text{int}}} \rho_{\text{int}}(\boldsymbol{r})\,\mathrm{d}r = \alpha \int_{V_{\text{int}}} \sum_i \sum_j a_{ij} f_j(\boldsymbol{r} - \boldsymbol{r}_i)\,\mathrm{d}\boldsymbol{r}$$
$$= \alpha \sum_i \sum_j a_{ij} \int_{V_{\text{int}}} f_j(\boldsymbol{r} - \boldsymbol{r}_i)\,\mathrm{d}\boldsymbol{r} \tag{12}$$

The integral is solved analytically yielding a "structural" constant c_j for each order of the Legendre polynomials j, thus simplifying the equation to

$$\Delta G_{\text{int}} = \int_{V_{\text{int}}} \rho_{\text{int}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} = \alpha \sum_j c_j \sum_i a_{ij} \tag{13}$$

The hydration free energy then becomes

$$\Delta G_{\text{hyd}} = \alpha \sum_j c_j \sum_i a_{ij} + \beta S + \gamma V \tag{14}$$

4.1. Parameterization

The implementation of the compression strategy as described above requires a sequence of three steps: (i) the 3D structure of the molecule in the PDB-format [33], (ii) the corresponding molecular surface and its division into patches, and (iii) the appropriate molecular interaction field for each molecule.

4.1.1. Three-dimensional structures

The 3D structures of the 81 molecules were created with the SYBYL program package [34]. The molecular editor was used to build the structures and an energetically favorable conformation was calculated based on SYBYLs internal force field.

4.1.2. Molecular surface and patch generation

In the second step, for each molecule the solvent accessible surface and volume were computed [31] and each molecular surface was divided into a certain number of surface patches. A "nearest atom" algorithm is applied to accomplish the assignment of surface patches. This algorithm measures the distance of a particular surface point to all atoms of the molecule, subtracts the van der Waals radius [35] of the atom, and assigns the PDB serial number of the atom being closest to the surface point. Such an assignment accounts for the stronger interaction of surface exposed atoms with the solvent. The applied algorithm will not assign surface patches to atoms which are buried in the molecule. Thus, the set of points belonging to the same atom forms a patch on the surface. A resulting patch distribution is illustrated in Fig. 3. For clarity only three hydrogen atoms and one carbon atom are considered.

It becomes necessary to label the patches in a consistent manner to build a regression model for the interaction part of the 3D-FED as indicated in Eq. (11). Therefore, the atom serial number from the PDB file was used as a distinctive feature for the patch assignment on the molecular surface. Each patch is now associated with a particular atom of the molecule, which is then combined with a classification scheme based on the atomic structure of the molecule. The atom associated with a particular patch is identified to be a certain structural atom type so that there are as many distinguishable surface patches as there are atom types in the classification scheme.

The structure analysis we used in our procedure follows that of Ghose et al. [36], which consists of roughly 120 atomic structure types characterized by their first and second neighbors. As one can see from Fig. 3, some atoms have associated with them more than one surface patch, most often the benzene carbons, and obviously there are many separate surface patches that belong to the same atom type. To account for these situations, all patches related to the same atom type are treated equally. Thus, for all patches i belonging to the same atom type l the corresponding a_{ij} are

assigned an identical value a_{lj}^{T},

$$\frac{\varepsilon(\boldsymbol{r})}{\Delta V} = \sum_i \sum_j a_{lj}^{\mathrm{T}}(i)\, f_j(\boldsymbol{r} - \boldsymbol{r}_i) \tag{15}$$

where $a_{lj}^{\mathrm{T}}(i)$ denotes the parameter of patch i that corresponds to atom type l. Finally, we get for the hydration free energy

$$\Delta G_{\mathrm{hyd}} = \alpha \sum_{l=1}^{120} h_l \sum_j a_{lj}^{\mathrm{T}} c_j + \beta S + \gamma V \tag{16}$$

where h_l is the number of occurrences of a surface patch associated with atom type l.

The number of different atomic types occurring in a chosen training set of molecules is multiplied by the order of the Legendre polynomial to give the total number of independent variables, i.e. the number of columns of the design matrix, for an ordinary least squares fit. The number of *dependent* variables, i.e. the number of rows of the design matrix, is determined in the next section.

4.1.3. Molecular interaction field

Since the number of grid points is equal to the number of dependent variables for a least squares fit, further manipulation in order to reduce the number of rows of the design matrix is advised. Cubic grids with much smaller dimensions, or *patch grids*, were used to reduce the number of grid points. Each patch grid, therefore, resembles a subset of the entire grid surrounding the target molecule. The center of the patch grid has the same coordinates as the reference point of the associated surface patch. From the grid surrounding the entire molecule we consider for regression only those grid points that lie within one of the patch grids possessing attractive interaction. In addition, the angle ϑ_i between the surface normal of the patch and the vector $\boldsymbol{r} - \boldsymbol{r}_i$ had to be smaller than 110°. Fig. 4 shows such a scenario giving an example of the arrangement of the patch grid, the patch surface, and its normal vector.

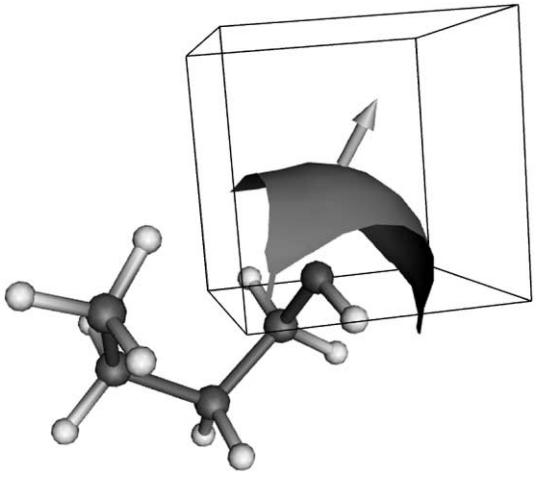

Fig. 4. The volume of the patch grid, the patch surface of the oxygen of 1-butanol, and the corresponding normal vector are shown. The edges of the grid surrounding the whole molecule are omitted for clarity.

Table 1
Results of the least squares fit of the model density with respect to the reference GRID values according to model II[a]

Order j of Legendre polynomial	0	1	2
c_j (Å^3)	58.6593	19.2983	−6.60042
R^2	0.901	0.919	0.922
σ (kcal/mol)	0.295	0.267	0.264
F	142627	90321	62073

[a] The correlation coefficient, the mean error, and the F statistic are listed depending on the maximum order of the Legendre expansion.

This sort of treatment reduces the average number of relevant grid points substantially, although it is clearly a function of the dimensions of the patch grids. The reduction of grid points is done for technical purposes only and accounts for the fact that the value of many grid points outside the patch grids are close to zero and thus are considered irrelevant for the fit. For the set of 81 molecules the resulting optimal side length of the cubic patch grids was 3.5 Å giving an average number of grid points of about 12,000 per molecule. The parameter k controlling the exponential part of the distance function (see Eq. (9)) was similarly optimized leading to a value of $k = 0.66\,\text{Å}^{-1}$. The resulting values for the structural constant c_j in Eq. (13) are listed in Table 1.

4.2. Least squares fit

The optimized values for both the size of the patch grids and the exponential decay of the distance function modeling the 3D-FED allows us to fit the values of a_{lj}^{T} in Eq. (15) with respect to reference GRID data. Given a_{lj}^{T}, the undetermined parameters of model II (see Eq. (16)) are obtained by fitting α, β, γ to experimental values of ΔG_{hyd}. The results from fitting a model grid density to reference GRID data (divided by the volume element) for all 81 molecules simultaneously are summarized in Table 1. Considering the number of data points, the correlation is excellent.

One can see from the slightly increasing correlation coefficient with larger j that the exponential distance dependence alone is able to capture a substantial part of the molecular interaction field computed by GRID. The improvement of the fit statistics by adding angle dependent terms to the mapping function is anticipated but angular terms turn out to have little influence and are almost negligible beyond $j = 1$.

Consequently, the results of the ΔG_{hyd} fit are reported for $j = 0$ and 1 only. The correlation coefficient is $R^2 = 0.88$ ($F = 187.55$) with a mean error of $\sigma = 1.38$ kcal/mol. Although both the correlation coefficient and mean error are as expected slightly better for the simple GRID model I according to Eq. (8), the quality of the patch-based fit is a clear indication of the validity of the compression approach. The correlation between experimental and calculated hydration free energies is depicted in Fig. 5.

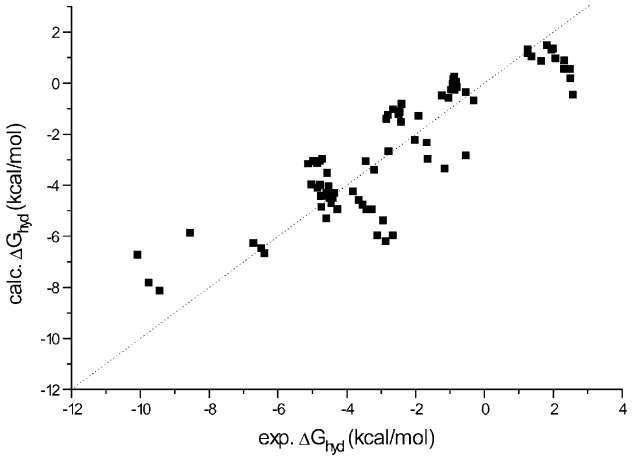

Fig. 5. Calculated free energies of hydration vs. experimental as derived from the compressed model II, Eq. (16), by fitting the parameters in two steps.

5. Final model: simultaneous use of local and global information

The investigations described in the previous sections were used to validate our model for the hydration free energy. First, it was shown that the interaction part of the 3D-FED can be well approximated by a molecular interaction field generated from the water probe of the GRID force field. Second, modeling the approximated 3D-FED by a small amount of surface related parameters does not alter significantly the quality of the fit results.

Given the success of the simple GRID/patch model, one might ask how to further improve the correlation while retaining the predictive capabilities. On one hand, the physics of the solvation process is approximately captured in the GRID model (or its compressed representation) and the surface and volume term. On the other hand, deficiencies of the approximation are accounted for by scaling the contributions to yield experimental data in an independent step. It is, therefore, conceivable that the model (Eq. (16)) can be significantly improved if global information, i.e. experimental ΔG_{hyd}, enters the parameterization process already in the patch parameter fit. One must seek ways to use local information (GRID data) and global information (ΔG_{hyd}) *simultaneously* to determine a_{lj}^{T} and α, β, γ of Eq. (16).

We see from Eq. (16) that α is nonlinearly coupled to the a_{lj}^{T}. An iteration scheme is, therefore, needed to find solutions of the minimization problem. We keep the functional form of Eq. (16) and rearrange in such a way as to be compatible with the design matrix of model II

$$\frac{\Delta G_{\mathrm{hyd}}^{\mathrm{exp}} - \beta S - \gamma V}{\alpha} = \sum_{l=1}^{120} h_l \sum_{j=0}^{1} a_{lj}^{\mathrm{T}} c_j \tag{17}$$

Again, the Legendre polynomial is considered up to $j = 1$ only. The number of rows of the design matrix resulting from Eq. (15) increases by the number of molecules in the training set. Thus, the total design matrix consists of two blocks, *A* and *B*, representing the grid energies divided by the volume element and the LHS of Eq. (17), respectively. Although these dependent variables correspond to different physical properties, the parameter basis is the same. The design matrix is properly scaled and a weight factor, λ, is introduced to find the optimal balance between the two blocks of the matrix. This weight factor is defined in such a way that a value of $\lambda = 1$ corresponds to the situation where the purely linear model (Eq. (17)) is used while $\lambda = 0$ represents Eq. (15) from model II

$$\lambda = \frac{w}{1 + w} \tag{18}$$

where w is the weight factor for the block *B* of the design matrix. If both blocks are weighted equally then $w = 1$ and $\lambda = 0.5$. The function χ^2 to be minimized then is

$$\chi^2 = \sum_m (\mathrm{obs}_{m,A,\mathrm{ref}} - \mathrm{obs}_{m,A})^2 + w \sum_p (\mathrm{obs}_{p,B,\mathrm{ref}} - \mathrm{obs}_{p,B})^2 \tag{19}$$

The summation of the block *A* is taken over all relevant grid points *m*. The reference observable, $\mathrm{obs}_{m,A,\mathrm{ref}}$, is the grid value of point *m* divided by the volume, whereas the modeled observable, $\mathrm{obs}_{m,A}$, is given by Eq. (15). Block *B* is summed over the number of molecules in the training set *p*. Here, $\mathrm{obs}_{p,B,\mathrm{ref}}$ and $\mathrm{obs}_{p,B}$ are the LHS and RHS of Eq. (17), respectively.

5.1. Iteration cycle

For the initial design matrix the values of the parameters α, β, and γ from the previous least square regression according to model II serve as starting values to calculate the LHS of Eq. (17). Subjecting this design matrix to a least squares algorithm, a new set of parameters a_{lj}^{T} is found. From these parameters the interaction part of the 3D-FED was calculated, inserted into Eq. (16), and a new set of parameters α, β, and γ is found again by linear regression. These three parameters are used to recalculate the LHS of Eq. (17) to build a new design matrix and thus closing the iterative cycle. The convergence of the mean square error of the calculated free energies of hydration of the training set was used as a stopping criterion. Several iteration cycles were performed with varying values of λ to find the model with the best predictive power. To estimate the predictive capability of the model we used an independent test set of 10 molecules of known experimental hydration free energies with the parameters derived from the training set.

The value of λ corresponding to the minimum deviation is 2.44×10^{-4}. Upon further reduction, the fit results approach the same statistics (e.g. rms error, correlation coefficient, etc.) as obtained from model II where only the grid values

Table 2
Atom types according to the classification scheme by Ghose et al. [36] and the parameters of the final model for $j = 0$ and 1 terms in the Legendre expansion

Atom type (l)	a_{l0}^{T} (kcal/mol Å^3)	a_{l1}^{T} (kcal/mol Å^3)
1	0.598	1.555
2	−1.880	2.311
3	−1.124	3.827
5	−18.968	16.393
6	−1.395	−9.582
8	−0.156	−6.085
15	−3.978	5.015
16	−4.211	0.903
24	−4.500	1.200
25	−10.471	6.820
27	−2.220	8.069
36	−32.929	7.481
38	−25.772	45.315
39	−2.680	24.452
40	−26.808	44.656
46	−2.668	2.245
47	−6.114	4.122
48	−14.899	−0.274
49	10.920	1.274
50	−41.451	15.919
51	−7.080	5.887
52	−6.155	6.537
56	−122.558	93.539
57	−17.818	−41.028
58	−111.788	46.367
59	−43.361	−5.998
60	40.258	−58.330
66	−57.875	20.781
72	−71.096	24.861
75	−93.476	−13.826

were considered for the linear regression. This behavior is consistent with our expectations and assures the validity of the iterative approach.

5.2. Results

Only 30 of the possible 120 atom types are present in the training set for which the parameter values are reported in Table 2. Since the Legendre expansion was considered up to $j = 1$, each atom type has two associated parameters. The corresponding c_j values are reported in Table 1. The final model equation together with the values for the parameters α, β, and γ is then

$$\Delta G_{\mathrm{hyd}} = 7.99 \times 10^{-4} \sum_l h_l \sum_j a_{lj}^{\mathrm{T}} c_j + \left(\frac{0.06176}{\text{Å}^2} S - \frac{0.04083}{\text{Å}^3} V \right) \text{kcal/mol} \qquad (20)$$

The correlation coefficient obtained for the training set is $R^2 = 0.99$ and the rms error turns out to be $\sigma = 0.27$ kcal/mol. The small magnitude of λ (see above) reflects the inherent, different amount of data points between blocks *A* and *B*. In comparison to model II, we find that

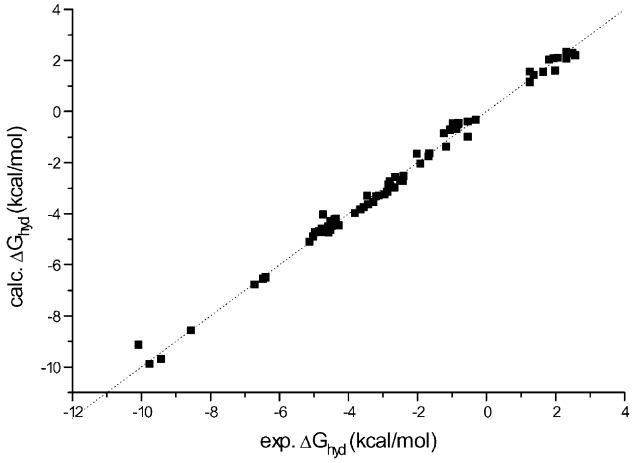

Fig. 6. Values of ΔG_{hyd} obtained from the iterative procedure plotted against experimental data.

incorporation of global information is essential for the improvement of the final model. Calculated hydration free energies are plotted against the experimental values in Fig. 6.

Since we used the same reference data as No et al. [16], it is legitimate to assess the quality of our model by a direct comparison of the rms error. No et al. [16] reported value of 0.43 kcal/mol is slightly higher than our value of 0.27 kcal/mol. Since both approaches use a 3D-FED as statistical basis, and both use different formulations for the empirical model, the concept of the 3D-FED and surface-compressed forms thereof is certainly valid.

For all training set members the sum $\beta S + \gamma V$ is always positive. This is another confirmation of the validity of our approach because the sum directly measures the cavity contribution. The individual surface and volume terms have no apparent physical meaning. In the work of No et al. [16], the surface contribution appears with a negative sign while the volume term is positive and also here the authors refrain from any interpretation of these individual contributions. The volume is a convenient descriptor, but one could always describe the cavity contribution purely in terms of surface functions alone using the divergence theorem [37,38]. To further improve our model we anticipate surface curvature as a better descriptor to capture the physics of the cavity formation.

The mean square error for predictions of ΔG_{hyd} of the test set is $\sigma = 0.63$ kcal/mol thus being within range of experimental uncertainty ensuring the models predictive capability. The calculated free energies of hydration for the test set together with the experimental values are listed in Table 3.

Table 3 shows excellent agreement of calculated and experimental hydration free energies except for the value of the hexanal molecule. The poor prediction of ΔG_{hyd} for hexanal can be attributed to the dominant conformation or the conformation mixture of hexanal in solution which might be very different to the one generated by the SYBYL force field. Similar predictive restrictions due to an increase of

Table 3
Experimental [17] and calculated hydration free energies from the final model (kcal/mol) for the test set

Compound	Experimental ΔG_{hyd}	Calculated ΔG_{hyd}
1-Hexene	1.73	1.50
2-Methylpentane	2.56	2.29
3-Hexanol	−3.73	−4.30
4-Ethylpyridine	−4.66	−4.68
Cis-1,2-dimethylcyclohexane	1.60	1.68
Ethylpropionate	−2.83	−3.26
Hexanal	−2.85	−4.30
Isopentylacetate	−2.24	−2.91
Methylformate	−2.82	−2.21
Tert-butylbenzene	−0.44	−0.54

conformational degrees-of-freedom were also obtained by No et al. [16].

Two further aspects of our model should be discussed. One is the assignment of surface patches, since it represents a degree-of-freedom for the entire parameterization procedure. Our initial approach uses a pure distance criterion, i.e. each surface point is associated with the closest atom. More elaborate assignment strategies could allow contributions to a surface element by several atoms, and, therefore might change results. Secondly, the parameterization quality strongly depends on the quality of the underlying GRID force fields. Albeit GRID has proved its value in many applications where the focus was on energetic considerations, interpretation of the molecular interaction fields as an approximation for the interaction part of the free energy of hydration is justifiable. The GRID program estimates a solvent entropy contribution from the preferential orientation of water molecules in the first solvation shell. As it was recently shown [39], this contribution is indeed the dominant part in a quantitative description of the hydrophobic hydration. Although quantitative improvement of the GRID entropy estimate is feasible, it certainly captures the essential physics.

Charged species would pose a problem, not for the basic model, but rather with respect to the parameterization process since the absolute chemical potential of an ion in infinite dilution is difficult to measure. The parameterization would, therefore, benefit from highly accurate theoretical free energy data that is hard to get for such a large number of compounds so that the empirical model reaches sufficient significance. This is of particular importance for tautomeric equilibria that usually involve charged species. Once a model is parameterized, the difference of chemical potentials of the different tautomeric forms that are estimated separately, can be used to compute the equilibrium constant.

A promising approach towards a chemical understanding of solvation phenomena would be the identification of relevant surface patches. With a given patch distribution application of principal component or partial least squares algorithms [40] offers a promising route towards a further condensed representation of surface patches. Inspection of the underlying molecular structure of the newly formed patches is expected to render identification of structural fragments relevant to the solvation process. Work in this direction is currently underway.

6. Conclusions

An empirical model was presented to allow for the fast prediction of free energies of hydration. The statistical basis of this model was derived from the concept of a 3D free energy density which is accessible through both explicit and implicit solvent models. In our approach the interaction part of the 3D-FED is approximated by an appropriate molecular interaction field generated by the program GRID. In a second step, the molecular interaction field is modeled as a linear combination of surface functions. A standard atom classification scheme was used to associate the surface functions to particular atom types and to establish a model based on group contributions.

The final model was parameterized with respect to the simultaneous prediction of local grid data as well as global hydration free energies. The resulting rms deviation for a training set of 81 molecules is 0.27 kcal/mol and for an independent test set of 10 molecules 0.63 kcal/mol. Our simultaneous and optimally balanced use of local and global data represents a significant departure from other empirical work. Because a maximum of physical information is retained, the predictive capability is optimized. One might get the impression that the model presented is overly complicated and suffers from overfitting. The latter issue can be safely ruled out given the vast amount of reference data (the grid points and the experimental hydration free energies) compared to the number of parameters. Admittedly, the parameterization process is complex, whereas the application of the final model itself is certainly not. There are exactly two atom-type specific parameters plus three global ones for interaction part, surface and volume term, respectively. The slowest part in the evaluation of a single compound is the computation of the molecular surface and the patch generation. Starting from the spatial structure, the determination of the free energy of hydration is a matter of a few seconds.

Although the main focus of this work was the description of the model and its parameterization, it is desirable to extend the training set towards a complete coverage of all structural types of the atom classification scheme.

Acknowledgements

We thank Jürgen Brickmann and Bernd Schilling for their valuable input and extensive discussions regarding this work. Many helpful comments on the use of the program GRID by Peter Goodford are greatly appreciated. We also thank Robert Ivkov for many beneficial remarks on the manuscript. This work has been supported by the Deutsche Forschungsgemeinschaft, Bonn.

References

[1] P.A. Kollman, Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules, Acc. Chem. Res. 29 (1996) 461–469.

[2] T.P. Straatsma, Free energy by molecular simulation, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, Vol. 9, VCH Publishers, New York, 1996, pp. 81–127.

[3] P.M. King, Free energy via molecular simulation: a primer, in: W.F. van Gunsteren, P.K. Weiner, A.J. Wilkinson (Eds.), Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Escom Science Publishers, Leiden, The Netherlands, 1993, pp. 267–314.

[4] W.L. Jorgensen, Computation of free energy changes in solution, in: P.V.R. Schleyer (Ed.), Encyclopedia of Computational Chemistry, Vol. 2, Wiley, New York, 1998, pp. 1061–1070.

[5] E.M. Duffy, W.L. Jorgensen, Prediction of properties from simulations: free energies of solvation in hexadecane, octanol, and water, J. Am. Chem. Soc. 122 (2000) 2878–2888.

[6] R.H. Wood, E.M. Yezdimer, S. Sakane, J.A. Barriocanal, D.J. Doren, Free energies of solvation with quantum-mechanical interaction energies from classical mechanical simulations, J. Chem. Phys. 110 (1999) 1329–1337.

[7] S.A. Best, K.M. Merz Jr., C.H. Reynolds, Free energy pertubation study of octanol/water partition coefficients: comparison with continuum GB/SA calculations, J. Phys. Chem. B 103 (1999) 714–726.

[8] B. Roux, T. Simonson, Preface, Biophys. Chem. 78 (1999) 1.

[9] A. Papazyan, A. Warshel, Continuum and dipole-lattice models of solvation, J. Phys. Chem. B 101 (1997) 11254–11264.

[10] C.J. Cramer, D.G. Truhlar, Implicit solvation models: equilibria, structure, spectra, and dynamics, Chem. Rev. 99 (1999) 2161–2200.

[11] M. Orozco, F.J. Luque, Theoretical methods for the description of the solvent effect in biomolecular systems, Chem. Rev. 100 (2000) 4187–4225.

[12] L. Shao, H.-A. Yu, J. Gao, XSOL, a combined integral equation (XRISM) and quantum-mechanical solvation model: free energies of hydration and applications to solvent effects on organic equilibria, J. Phys. Chem. A 102 (1998) 10366–10373.

[13] J. Tomasi, M. Persico, Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent, Chem. Rev. 94 (1994) 2027–2094.

[14] F.J. Luque, X. Barril, M. Orozco, Fractional description of free energies of solvation, J. Comput. Aided Mol. Design 13 (1999) 139–152.

[15] T.I. Oprea, C.L. Waller, Theoretical and practical aspects of three-dimensional quantitative structure–activity relationships, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, Vol. 11, VCH Publishers, New York, 1997, pp. 127–182.

[16] K.T. No, S.G. Kim, K.H. Cho, H.-A. Scheraga, Description of hydration free energy density as a function of molecular physical properties, Biophys. Chem. 78 (1999) 127–145.

[17] V.N. Viswanadhan, A.K. Ghose, U.C. Singh, J.J. Wendoloski, Accurate prediction of solvation free energies of small organic molecules: additive–constitutive models based on molecular fingerprints and atomic constants, J. Chem. Inf. Comput. Sci. 39 (1999) 405–412.

[18] H.-J. Böhm, G. Klebe, H. Kubinyi, Wirkstoffdesign Spektrum, Akad. Verl., Heidelberg, 1996, Chapter 21, pp. 381–397.

[19] G. Greco, E. Novellino, Y.C. Martin, Approaches to three-dimensional quantitative structure–activity relationships, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, Vol. 11, VCH Publishers, New York, 1997, pp. 183–240.

[20] R.D. Cramer III, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, J. Am. Chem. Soc. 110 (1988) 5959–5967.

[21] P. Goodford, Multivariate characterization of molecules for QSAR analysis, J. Chemometrics 10 (1996) 107–117.

[22] M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems, Quant. Struct. Act. Relat. 12 (1993) 9–20.

[23] G. Cruciani, P. Crivori, P.-A. Carrupt, B. Testa, Molecular fields in quantitative structure–permeation relationships: the VolSurf approach, J. Mol. Struct. (Theochem.) 503 (2000) 17–30.

[24] L.H. Alifrangis, I.T. Christensen, A. Berglund, M. Sandberg, L. Hovgaard, S. Frokjaer, Structure–property model for membrane partitioning of oligopeptides, J. Med. Chem. 43 (2000) 103–113.

[25] H.-J. Böhm, G. Klebe, H. Kubinyi, Wirkstoffdesign Spektrum, Akad. Verl., Heidelberg, 1996 (Chapter 25).

[26] J.-P. Hansen, I.R. McDonald, Theory of Simple Liquids, 2nd Edition, Academic Press, New York, 1990 (Chapter 6.7).

[27] A. Kovalenko, F. Hirata, Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model, J. Chem. Phys. 110 (1999) 10095–10112.

[28] K. Du, D. Beglov, B. Roux, Solvation free energy of polar and nonpolar molecules in water: an extended interaction site integral equation theory in three dimensions, J. Phys. Chem. B 104 (2000) 796–805.

[29] G. Hummer, D.M. Soumpasis, Computation of the water density distribution at the ice–water interface using the potentials-of-mean-force expansion, Phys. Rev. E 49 (1994) 591–596.

[30] G. Hummer, D.M. Soumpasis, Statistical mechanical treatment of the structural hydration of biological macromolecules: results for B-DNA, Phys. Rev. E 50 (1994) 5085–5095.

[31] M.L. Connolly, Solvent-accessible surfaces of proteins and nucleic acids, Science 221 (1983) 709–713.

[32] A. Ben-Naim, R. Lovett, Solvation free energy of a hard sphere solute in a square well solvent as a function of solute size, J. Phys. Chem. B 101 (1997) 10535–10541.

[33] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[34] SYBYL 6.6, Tripos Inc., 1699 South Hanely Road, St. Louis, MO 63144, USA.

[35] A. Bondi, van der Waals volumes and radii, J. Phys. Chem. 68 (1964) 441–451.

[36] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and LOGP methods, J. Phys. Chem. A 102 (1998) 3762–3772.

[37] J.D. Jackson, Classical Electrodynamics, Wiley, New York, 1975.

[38] R. Jäger, F. Schmidt, B. Schilling, J. Brickmann, Localization and quantification of hydrophobicity: the molecular free energy density (MolFESD) concept and its application to sweetness recognition, J. Comput. Aided Design 14 (2000) 631–646.

[39] T. Lazaridis, Solvent reorganization energy and entropy in hydrophobic hydration, J. Phys. Chem. B 104 (2000) 4964–4979.

[40] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.