A fast method for the determination of fractional contributions to solvation in proteins

David Talavera, Antonio Morreale, Tim Meyer, Adam Hospital, Carles Ferrer-Costa, Josep Lluis Gelpi, Xavier de la Cruz, Robert Soliva, F. Javier Luque and Modesto Orozco

Protein Sci. 2006 15: 2525-2533; originally published online Sep 25, 2006;
Access the most recent version at doi:10.1110/ps.062406706

A fast method for the determination of fractional contributions to solvation in proteins

DAVID TALAVERA,[1,9] ANTONIO MORREALE,[2,9] TIM MEYER,[1] ADAM HOSPITAL,[1] CARLES FERRER-COSTA,[1] JOSEP LLUIS GELPI,[1,3,8] XAVIER DE LA CRUZ,[1,4] ROBERT SOLIVA,[1,5] F. JAVIER LUQUE,[6] AND MODESTO OROZCO[1,3,7,8]

[1]Molecular Modeling and Bioinformatics Unit, Institut de Recerca Biomèdica, Barcelona 08028, Spain
[2]Unidad de Bioinformática, Centro Biología Molecular Severo Ochoa, Campus de Cantoblanco, Madrid 28049, Spain
[3]Departament de Bioquímica i Biologia Molecular, Facultat de Biologia, Universitat de Barcelona, Barcelona 08028, Spain
[4]Institució Catalana per la Recerca i Estudis Avançats (ICREA), Barcelona 08018, Spain
[5]Drug Discovery, J. Uriach y Compañía, Barcelona 08184, Spain
[6]Departament de Fisicoquímica, Facultat de Farmacia, Universitat de Barcelona, Barcelona 08028, Spain
[7]Nodo de Estructura y Modelización, Instituto Nacional de Bioinformática, Fundación Genoma España, Madrid 28020, Spain
[8]Computational Biology Program, Barcelona Supercomputing Center, Barcelona 08034, Spain

(RECEIVED June 20, 2006; FINAL REVISION June 20, 2006; ACCEPTED July 19, 2006)

Abstract

A fast method for the calculation of residue contributions to protein solvation is presented. The approach uses the exposed polar and apolar surface of protein residues and has been parametrized from the fractional contributions to solvation determined from linear response theory coupled to molecular dynamics simulations. Application of the method to a large subset of proteins taken from the Protein Data Bank allowed us to compute the expected fractional solvation of residues. This information is used to discuss when a residue or a group of residues presents an uncommon solvation profile.

Keywords: protein solvation; protein structure; molecular dynamics; protein stability

Supplemental material: see www.proteinscience.org

The solvation free energy is the reversible work needed to transfer the solute from the gas phase to solution at constant pressure, temperature, and concentration (Böttcher 1952; Tomasi and Persico 1994; Rivail and Rinaldi 1995; Cramer and Truhlar 1999; Orozco and Luque 2000). From a computational point of view, it is convenient to partition such process into three steps: (1) creation of a solute-shaped cavity in the solvent, (2) switching on the steric properties of the solute inside the cavity, and (3) building up of the charge distribution of the solute. According to this partitioning scheme, the solvation free energy (ΔG_{solv}) is determined as the addition of electrostatic (ΔG_{ele}), cavitation (ΔG_{cav}), and van der Waals (ΔG_{vw}) terms (Equation 1), though these two latter contributions are often grouped into the "steric" contribution (ΔG_{ster}).

$$\Delta G_{solv} = \Delta G_{cav} + \Delta G_{vw} + \Delta G_{ele} = \Delta G_{ster} + \Delta G_{ele} \quad (1)$$

The experimental determination of ΔG_{solv} is difficult even for small molecules and impossible for large macromolecules. This has fueled the development of accurate theoretical approaches to calculate ΔG_{solv} of small and medium molecules in water with an error <1 kcal/mol (see discussion in Tomasi and Persico 1994; Rivail and

[9]These authors contributed equally to this work.

Reprint requests to: Modesto Orozco, Molecular Modeling and Bioinformatics Unit, Institut de Rercerca Biomèdica, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain; e-mail: modesto@mmb.pcb.ub.es; fax: +34-93-4037158.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.062406706.

Rinaldi 1995; Cramer and Truhlar 1999; Orozco and Luque 2000). The application of these methods to macromolecular systems such as proteins is, nevertheless, limited mainly due to the large size of these molecules and the complexity of their conformational space. However, calculation of the "effective" ΔG_{solv}, i.e., the negative value of the reversible work needed to transfer the ensemble of conformations sampled by the protein in solution to the gas phase, is less complex.

The direct calculation of the "steric" terms is difficult (Orozco and Luque 2000), and for macromolecules it is generally assumed that they can be captured from empirical relationships with the solvent-accessible surface (SAS). The electrostatic component is usually computed using continuum models such as the Generalized Born or Poisson-Boltzman methods (Gilson and Honig 1988; Davis and McCammon 1990; Still et al. 1990; Tomasi and Persico 1994; Cramer and Truhlar 1999; Orozco and Luque 2000). Continuum calculations can be performed for either a single structure (i.e., the experimental one) or more accurately for an ensemble of configurations collected from molecular simulations.

Discrete linear response theory coupled to molecular dynamics simulations with explicit solvent (MD/LRT) (Warshel and Russell 1984; Aqvist 1990; King and Warshel 1990; Roux et al. 1990; Still et al. 1990; Lee et al. 1992, 1993; Aqvist et al. 1994, Carlson and Jorgensen 1995; Kong and Warshel 1995; Orozco et al. 1995; Aqvist and Hansson 1996; Orozco and Luque 1997; Pierce and Jorgensen 2001; Wesolowski and Jorgensen 2002; Morreale et al. 2004) is a good alternative to continuum methods to determine the effective ΔG_{solv} of proteins (for the sake of brevity, hereafter the term "effective" is omitted). The MD/LRT approach avoids the uncertainties implicit to the partition of the system into solvent and solute regions. Furthermore, ΔG_{solv} can be easily decomposed into group contributions (Morreale et al. 2003), which in turn might be used to identify residues that are crucial for the physical properties of proteins or that display unusual solvation, thus unraveling their potential role in molecular recognition and signaling.

Different authors (for detailed discussion, see Orozco and Luque 2000; Morreale et al. 2005) have suggested that the fractional solvation contribution of a residue i in a protein (ΔG^i_{sol}) can be estimated from a linear function (Equation 2) that relates (1) the intrinsic solvation free energy of the free residue, i.e., the solvation free energy of a fully exposed residue, $\Delta G^i_{sol}(free)$; and (2) the fraction of the residue surface that remains accessible to the solvent in the protein, ρ^i_{prot} (Equation 3). However, Equation 2 fails to reproduce the fractional solvation in certain cases such as bulky charged residues, where the solvation penalty arising upon burying of the polar and apolar moieties is very different.

$$\Delta G^i_{sol} = \Delta G^i_{sol}(free)\rho^i_{prot} \tag{2}$$

$$\rho^i_{prot} = \frac{\langle SAS^{prot}_i \rangle}{\langle SAS^{free}_i \rangle} \tag{3}$$

where $\langle SAS^{prot}_i \rangle$ and $\langle SAS^{free}_i \rangle$ stand for the average solvent-accessible surface of the residue in the protein or free (fully exposed) in solution.

Recently we have developed an intuitive fractional solvation scheme where ΔG^i_{sol} is determined from the magnitude of the protein–solvent interaction in the vicinities of residue i. The method can be applied within the continuum SCRF formalism (Luque et al. 1995, 2003; Muñoz et al. 2002; Morreale et al. 2003), and also within discrete MD/LRT calculations (Morreale et al. 2005). In this latter case, ΔG^i_{sol} is computed using a simple expression over the ensemble of structures collected along an equilibrated trajectory of the solvated protein (Equation 4)

$$\Delta G^i_{sol} = \langle \xi SAS_i \rangle + \left\langle \sum_{n=1}^{N} \sum_{k=1}^{M} \sum_{j=1}^{3} \frac{Q_n Q_j}{r_{jn}} \right\rangle \tag{4}$$

where ξ is a tension parameter that usually adopts the same value for all the heavy atoms of the protein and zero for hydrogen atoms, SAS_i is the solvent-accessible surface of residue i, N is the total number of atoms in the protein, M is the number of water molecules closer to residue i than to any other residue, r_{jn} is the distance between atoms in water molecules and in the protein, and Q stands for the atomic charge; finally, the bracket means that these values are averaged along the structures sampled in the MD trajectory.

Despite recent advances in computer power, Equation 4 is still of little use in massive genomic-scale studies. In this context, we explore here the suitability of a simple fractional approach, which assumes that the solvation free energy of each residue can be determined considering (1) the intrinsic solvation free energy of the residue, $\Delta G^i_{sol}(free)$, and (2) the fraction of its apolar and polar surface that remains accessible to the solvent in the protein. In particular, the method accounts for the different (de)solvation of polar and apolar fragments of a given residue. This is achieved through the introduction of three adjustable parameters (Equation 5), which account for the intrinsic solvation free energy of each residue (α_i), and the desolvation penalty arising from the reduction of apolar (β_i) and polar (δ_i) solvent-accessible surfaces of the residue

$$\Delta G^i_{sol} = \alpha_i + \beta_i\left(1 - \rho^i_{prot}(ap)\right) + \delta_i\left(1 - \rho^i_{prot}(pol)\right) \tag{5}$$

where $\rho^i_{prot}(ap)$ and $\rho^i_{prot}(pol)$ stand for the fraction of the apolar and polar solvent-accessible surface of the residue in the protein relative to the fully solvent-exposed state, respectively.

In principle, the three adjustable parameters (α_i, β_i, and δ_i) depend on the nature of the residue and on the protein. Here we assume that averaging data for different proteins can be useful to derive consensus values for each residue. Accordingly, the method is parametrized using MD/LRT data obtained from a small database of MD trajectories (see Materials and Methods) and applied to a large nonredundant database of monomeric proteins to determine the normal solvation profiles of the 20 amino acids.

Table 1. *Fitted parameters used in Equation 5 to compute the fractional solvation free energy of the 20 residues*

Residue	α (kcal/mol)	β (kcal/mol · Å^2)	δ (kcal/mol · Å^2)
ALA	−19.1	9.6	9.3
ARG	−57.9	0.9	46.8
ASN	−26.9	4.0	17.2
ASP	−74.1	−6.1	82.2
CYS	−5.7	3.7	2.7
GLN	−25.8	3.5	17.4
GLU	−71.8	5.6	68.5
GLY	−11.7	4.5	6.2
HIS	−27.0	2.7	21.1
ILE	−17.2	9.5	6.6
LEU	−21.0	11.0	9.3
LYS	−62.2	13.3	35.2
MET	−17.9	10.4	6.4
PHE	−23.4	12.9	9.6
PRO	−17.6	9.6	7.5
SER	−15.8	7.8	3.3
THR	−22.2	8.0	11.6
TRP	−40.5	26.0	11.0
TYR	−25.2	1.2	20.1
VAL	−17.3	14.1	2.8

Results

Despite its formal simplicity, Equation 5 reproduces quite reasonably the solvation free energy of different types of residues determined from MD/LRT calculations (see Fig. 1 for randomly selected examples; the complete set is available upon request). The agreement is better than that obtained using Equation 2 (data not shown), especially for charged residues, where the assumption that burying polar and apolar moieties has the same desolvation cost can introduce important errors.

The set of fitted parameters (see Table 1) reproduce well the fractional solvation free energies for the whole set of proteins included in the training set (see Fig. 2). When the fitted model is applied to the proteins in the validation set, the performance decreases, but it is still quite remarkable ($r^2 = 0.81$, with a systematic deviation of only 4%) (see Fig. 3). A similar performance is obtained when the analysis is restricted to neutral residues ($r^2 = 0.80$). In contrast, when the validation set is computed using Equation 2, the systematic deviation amounts to 11% and the agreement between predicted and MD/LRT results is worst ($r^2 = 0.71$). Finally, surface-dependent methods (Equations 2, 5) reproduce better the

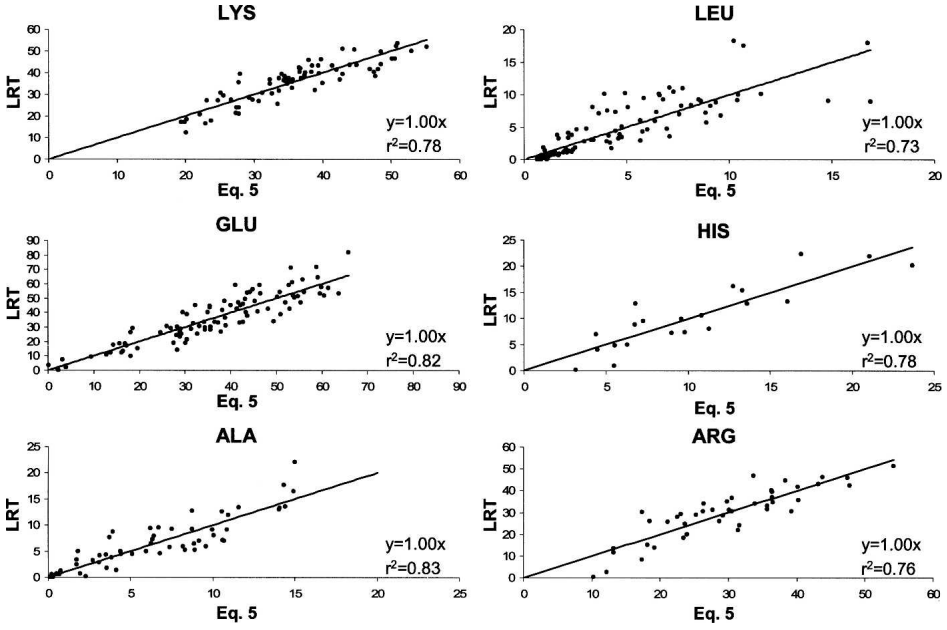

Figure 1. Comparison of fractional solvation free energies (in kilocalories per mole) determined from MD/LRT calculations or from Equation 5 for selected residues. (The complete list of plots is available at http://mmb.pcb.ub.es/aminoacids/MainFrame.html.)

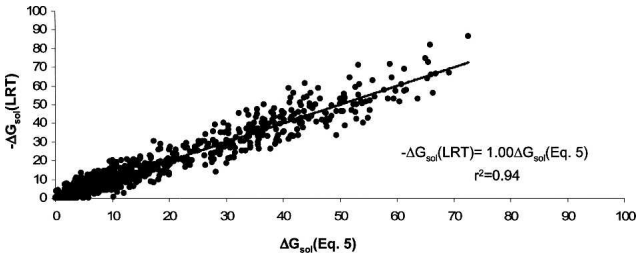

Figure 2. Comparison of fractional solvation free energies (in kilocalories per mole) determined from MD/LRT calculations or from Equation 5 for the 11 proteins included in the training set.

MD/LRT results than sequence-based fractional solvation methods, which assume that the solvation contribution of each residue depends only on the chemical nature of the amino acid (systematic deviation of 60% and $r^2 = 0.60$) (data not shown).

To further test the reliability of Equation 5, we compared the magnitude of the parameter α_i, which accounts for the intrinsic solvation free energy of each residue, with the solvation free energy of the free residue determined from MD/LRT simulations $\left(\Delta G^i_{sol}(free)\right)$ (data taken from Morreale et al. 2005). Let us note that these latter values were determined using a different set of simulations and were not introduced at any stage in the parametrization process. The results in Figure 4 demonstrate that the set of α_i parameters are very close to the $\Delta G^i_{sol}(free)$ values, which supports the physical meaning of the fractional model explored here.

As noted previously (Morreale et al. 2005), 1D solvation profiles determined from different partitioning schemes can be compared using Spearman's correlation test (see Fig. 5 for selected examples). In general, the MD/LRT solvation profiles correlate well with those obtained from Equation 5 (Spearman's coefficient of 0.90 for the set of 11 proteins in the training set), the worse correlation being detected for 1TBP (0.82) and the best for 4ICB (0.96). The value obtained for the proteins in the validation set was 0.92. The solvation profiles determined using Equation 2 are acceptable but less accurate than those obtained from Equation 5 (data not shown). Finally, none of the empirical 1D or 3D methods considered in the comparison showed a similar agreement with the MD/LRT profiles (Morreale et al. 2005).

In summary, after calibration with discrete MD/LRT data, Equation 5 provides a fast estimate of the fractional solvation of a given residue from the 3D structure of a protein. The root mean square deviation (RMSD) between predicted and MD/LRT fractional estimates is ~4 kcal/mol, a non-negligible error in absolute terms, but very small in relative terms when considering that the solvation free energy of a given protein might amount to several thousands of kilocalories per mole. Equation 5 can then be used to obtain rough, but qualitatively correct estimates of the fractional solvation at the genomic scale, something unfeasible using more accurate methods. In this context, we computed the fractional solvation free energies for a large (~1250 structures) subset of nonredundant proteins taken from the Protein Data Bank (see Materials and Methods) using Equation 5. The results in Table 2 demonstrate that fractional solvation free energies are always smaller than those obtained for the free residue (Table 2), indicating that even those residues generally considered to be "exposed" are, in fact, partially buried by the protein. The largest variation in the free energy of solvation of residues among proteins is found for charged residues, but in relative terms, hydrophobic residues are those that are more strongly desolvated in proteins (80%–90% of their solvation free energy is lost when embedded in proteins). The desolvation of polar and charged residues upon insertion in proteins represents 30%–70% of the original free energy of solvation (i.e., the average fractional solvation free energy is 20%–60% that of the free residue) (see Table 2). Then the assumption that charged residues are exposed to the solvent in the protein surface, showing no penalty to their intrinsic solvation, is not supported by the present calculations.

Analysis of the normalized distributions of fractional solvation free energies for the different residues (Fig. 6) shows that the most populated bins for apolar residues correspond to highly desolvated residues, which are likely to be buried in the protein matrix. In turn, bins corresponding to solvation free energies close to the intrinsic solvation free energy are poorly populated. The result is an exponential decay, which permits us to define "oversolvated" apolar residues as those having a fractional solvation free energy greater (in absolute terms) than the threshold values given in Table 3, which were obtained as the limit for integration of 80%, 90%, and 95% population. We could expect that "oversolvated" apolar residues have a small number of inter-residue contacts and are located in apolar regions that can mediate, for instance, protein–protein interaction.

The histograms for polar residues, especially the charged ones, are more complex and make it necessary

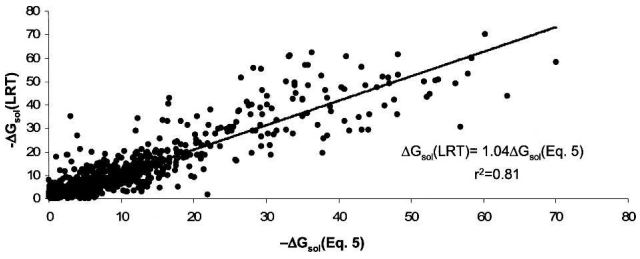

Figure 3. Comparison of fractional solvation free energies (in kilocalories per mole) determined from MD/LRT calculations or from Equation 5 for the proteins in the validation set.

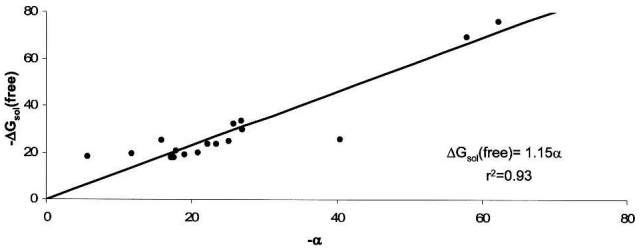

Figure 4. Representation of the solvation free energies $\left(\Delta G^{i}_{sol}(free)\right)$ of the 20 free residues computed from MD/LRT calculations with the intrinsic solvation free energy given by the α_i parameter in Equation 5. Values are in kilocalories per mole.

to define upper and lower thresholds (Fig. 6), since there is a sizable population of bins corresponding to either very large (in absolute terms) or very small fractional solvation. These profiles can be rationalized by considering (1) the preference of these residues to be well solvated in the exterior of the protein and (2) the fact that for globular proteins,the probability of a residue to be desolvated in the interior of the protein is greater than that of being exposed on the surface (by a factor roughly dependent on protein radii). This makes it necessary to derive corrected distribution profiles in order to take into account the different probability of residues to be in the interior of the protein matrix or on the surface of proteins (see Materials and Methods).

Figure 7 shows the corrected distribution profiles for a series of representative residues (the complete set of background and corrected plots is available at http://mmb.pcb.ub.es/aminoacids/MainFrame.html). These profiles provide valuable information on the impact that the anisotropy in residue distribution has on the residue solvation. We can distinguish three main kinds of profiles (Fig. 7): (1) flat curves covering most of the bins with a population between 0.9 and 1.1 relative to that expected from random models (i.e., Gly), (2) profiles with a maximum at low (in absolute terms) fractional solvation and an exponential decay for better solvation (i.e., Cys, Ile, Val, and other apolar residues), and (3) Gaussian-like profiles with a maximum corresponding to bins with large or very large fractional solvations (i.e., Lys, Glu, and most polar residues). Finally, there are few residues, like Ala and Met, with mixed profiles, which are intermediate between those of types 1 and 2, and His, Pro, Ser, and Tyr, with profiles intermediate between those of types 2 and 3.

The corrected distribution profiles can be fitted to a set of five Gaussians and one exponential (see Supplemental Material), which, in turn, can be used to predict the probability to find any fractional solvation free energy for a residue. Table 4 reports the threshold solvation values where the population of the corrected bins is <0.9 (i.e., the relative population of the bin is 90% of that expected for a random model). For apolar residues, the threshold

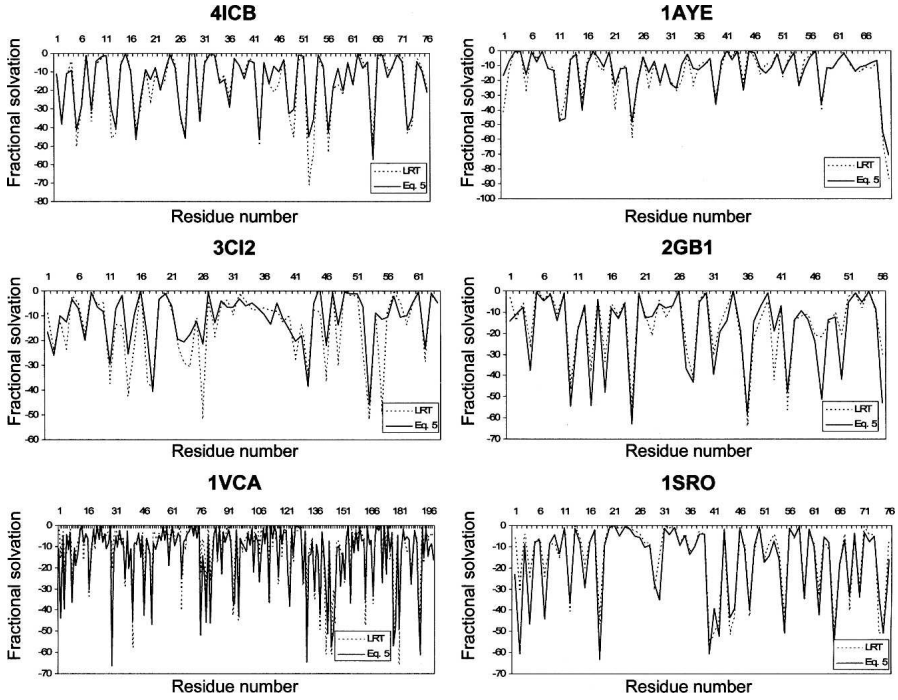

Figure 5. Solvation profiles for randomly selected proteins obtained using Equation 5 (solid lines) and from MD/LRT calculations (dotted lines). Values are in kilocalories per mole.

Table 2. *Average fractional solvation free energy, desolvation cost (loss of solvation upon incorporation of the residue into a protein), and fraction of the solvation free energy of the free residue retained in the protein*

Residue	Fractional solvation free energy (kcal/mol)	Desolvation cost (kcal/mol)	Ratio
ALA	−4.2	14.9	0.2
ARG	−29.1	28.9	0.5
ASN	−13.6	13.4	0.5
ASP	−24.8	49.3	0.3
CYS	−0.1	5.6	0.0
GLN	−13.0	12.8	0.5
GLU	−29.1	42.7	0.4
GLY	−4.1	7.6	0.4
HIS	−9.5	17.5	0.4
ILE	−3.0	14.2	0.2
LEU	−3.3	17.6	0.2
LYS	−36.1	26.2	0.6
MET	−4.0	13.8	0.2
PHE	−3.8	19.6	0.2
PRO	−6.1	11.5	0.4
SER	−8.6	7.2	0.6
THR	−8.2	14.1	0.4
TRP	−9.2	31.3	0.2
TYR	−8.6	16.6	0.3
VAL	−2.9	14.5	0.2

values are similar to those determined from uncorrected solvation profiles (Table 3) with a cutoff of 80% (see above). There are some polar residues (Arg, Asn, Gln, and Ser) for which the ''upper limits'' almost correspond to the α values (see Table 1), and other polar and even charged residues (Asp, Glu, Lys) have upper limits of solvation smaller (in absolute terms) than the α values, suggesting that the protein structure rarely allows the full interaction of these residues with the solvent. Finally, and quite surprisingly, it is not rare to have largely desolvated polar residues, as shown in the very small values (in absolute term) for the anomalous desolvation threshold found for polar and charged residues.

The information derived from the analysis of the solvation profiles can be valuable to detect when a residue has unusual solvation properties and can therefore be involved in functional roles, such as oversolvated apolar residues that mediate interactions with other molecules or anomalously desolvated charged residues in the interior of the protein implicated in salt bridges or ligand binding. This will be analyzed in more detail in a further paper, but as a preliminary application here, we have examined the solvation contributions of residues Glu11, Asp20, Ser117, and Asn132 in the T4 lysozyme (PDB code 3LZM), which play critical roles in both catalysis (Glu11 and Asp20) and binding (Ser117 and Asn132) with the peptide part of the cell wall substrate. In fact, replacement of both Glu11 and Asp20 by a variety of polar and apolar residues yielded mutants with negligible enzymatic activity and a substantial reduction (between five-fold and 200-fold) in the catalytic activity (Shoichet et al. 1995). In the native protein, Glu11 forms a salt-bridge interaction with Arg145, and the carboxylate group is partially surrounded by water molecules. Similarly, one

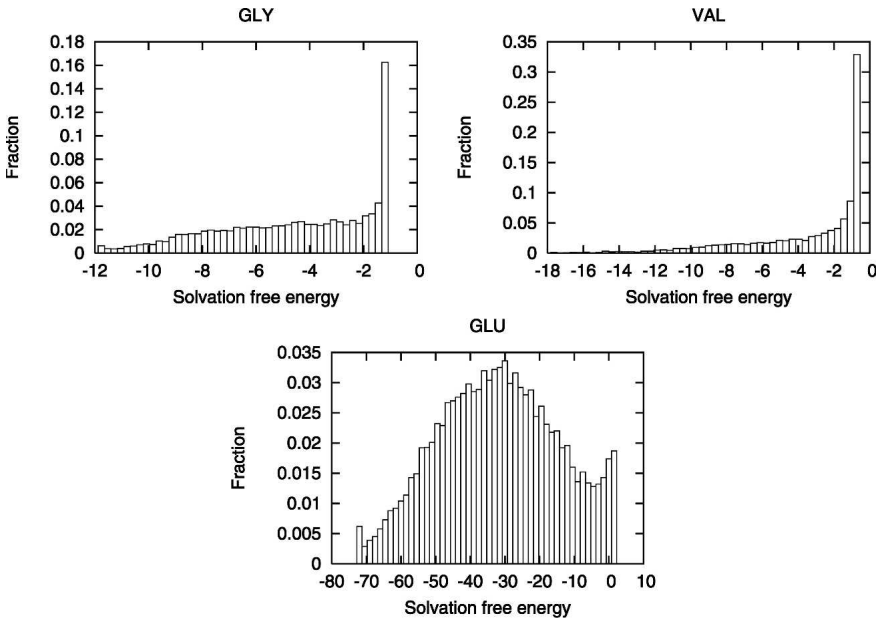

Figure 6. Normalized distributions of uncorrected fractional solvation free energies (in kilocalories per mole) for selected residues in the set of 1253 monomeric-cytoplasmatic proteins.

Table 3. *Threshold values for "oversolvation" of apolar residues (those showing a pure exponential profile in histograms of fractional solvation free energies) corresponding to the integration values of 80%, 90%, and 95% population*

Residue	80%	90%	95%
ALA	−9.6	−12.4	−14.4
CYS	−0.8	−1.7	−2.4
ILE	−5.8	−7.9	−10.1
LEU	−7.1	−10.1	−13.0
MET	−8.6	−11.9	−14.8
PHE	−7.7	−11.0	−13.9
TRP	−16.0	−21.5	−26.2
TYR	−13.5	−16.4	−18.3
VAL	−6.5	−8.6	−10.8

Residues showing fractional solvation free energies larger (in absolute terms) to these values are "oversolvated."

of the carboxylate oxygens of Asp20 forms hydrogen-bond contacts with the NH backbone units of residues 22 and 24, but the other remains partially exposed to the solvent. Finally, Ser117 and Asn132 form a short hydrogen-bond interaction and contribute to defining the groove occupied by the ligand. The fractional solvation contributions of residues Glu11, Asp20, Ser117, and Asn132 amount to −13, −15, −4, and −11 kcal/mol, values that are much smaller (in the border of the 90% threshold value) than those of normal Glu, Asp, Ser, and Asn residues (−22, −24, −9, and −17 kcal/mol, respectively). This would indicate that these residues are placed in an unstable environment, suggesting that their presence should obey to functional roles.

Discussion

Linear response theory coupled to molecular dynamics simulations (MD/LRT) is valuable to gain insight into the interaction between individual protein residues and solvent water under physiological conditions. However, MD simulations are still too expensive and require some degree of expertise, which precludes the use of MD/LRT at the genomic scale, thus making it necessary to develop efficient methods to examine the solvation properties of residues in a protein. This task has been accomplished by using different approaches, such as the definition of 1D solvation profiles, which simply rely on the chemical nature of the residue's side chain, or strategies that take into account the fraction of solvation related to the reduction of the solvent-exposed surface of the residue in the protein relative to the free residue in solution.

In this study, we have presented a more elaborate surface-based approach, which combines (1) the intrinsic solvation free energy of the residue, $\Delta G^{i}_{sol}(free)$, and (2) the fraction of its apolar and polar surface that remains accessible to the solvent in the protein. This is achieved through the introduction of three adjustable parameters (Equation 5), which account for the intrinsic solvation free energy of each residue (α_i), and the desolvation

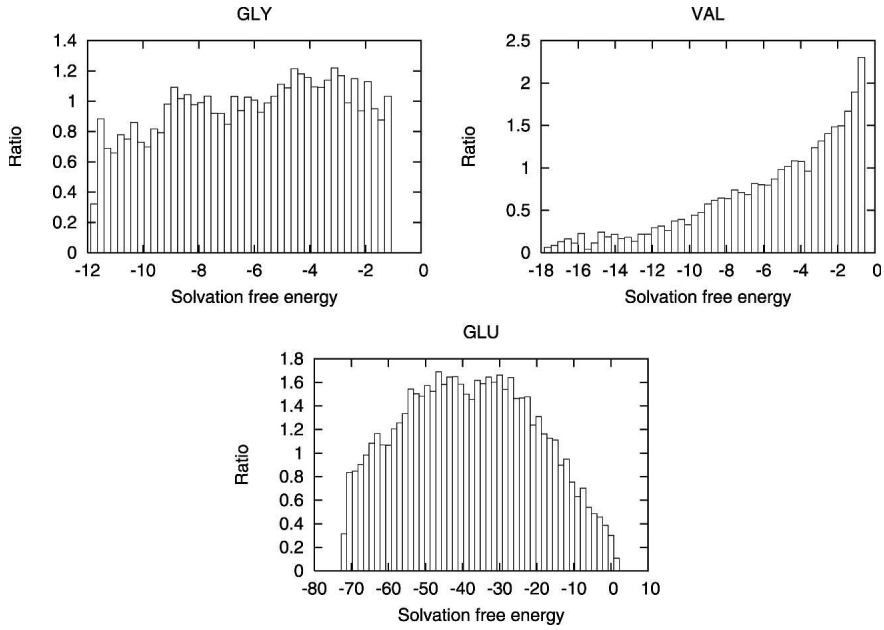

Figure 7. Normalized distributions of corrected fractional solvation free energies (in kilocalories per mole) for selected residues in the set of 1253 monomeric-cytoplasmatic proteins. Profiles were obtained by dividing the bin populations in Figure 6 by those obtained by assuming a random distribution of residues in the proteins (see Materials and Methods).

Table 4. *Cutoff values (kcal/mol) determined to detect when the solvation of a given residue is above or below the 90% threshold relative to that expected for random models*[a]

Residue	Upper limit (kcal/mol)	Lower limit (kcal/mol)
ALA	−8.0	—
ARG	−56.0	−17.8
ASN	−23.8	−8.6
ASP	−60.6	−8.4
CYS	−1.0	—
GLN	−24.6	−9.0
GLU	−67.5	−12.0
GLY	−9.1	—
HIS	−17.5	−4.0
ILE	−5.4	—
LEU	−7.1	—
LYS	−55.9	−25.3
MET	−7.6	—
PHE	−8.2	—
PRO	−14.4	−2.2
SER	−15.5	−5.6
THR	−14.2	−4.5
TRP	−17.6	—
TYR	−16.4	—
VAL	−5.0	—

[a]See Figure 7.
Note that for exponential decay profiles, no lower limit exists.

penalty arising from the reduction of apolar (β_i) and polar (δ_i) solvent-accessible surfaces of the residue. Since these parameters depend on the nature of the residue and on the protein, a fitting procedure to the fractional solvation contributions of the residues determined from MD/LRT calculations for a training set of proteins is necessary. The results presented here demonstrate that this approach suffices to provide quite accurate estimates of the MD/LRT values, keeping the physical sense of the computed properties.

The analysis of ~1250 representative structures corresponding to soluble globular proteins taken from the Protein Data Bank allowed us to derive the distribution of fractional solvation contributions for the 20 natural residues. This information is of particular interest, as the corrected solvation profiles provide a simple way to identify when a given residue in a protein is under- or oversolvated with respect to the normal distribution in proteins. Due to the large number of data, corrected histograms are rather smooth, but to further reduce discontinuities between bins and provide nonzero values for very unlikely populated bins, we fitted the histograms to a series of exponentials and (up to four) Gaussian functions using standard nonlinear fitting routines.

The surface-based method (Equation 5) is efficient enough to provide genomic-scale information on the solvation profile of residues in normal monomeric proteins. The direct solvation profiles, or those corrected by considering random models, provide a physically based, fast, and intuitive technique to detect residues with an anomalous interaction with the surrounding solvent. Thus, inspection of the fractional solvation contributions for a given protein might be an efficient tool to provide evidence about the potential implication of certain residues in functional roles, such as protein signaling or ligand binding.

Materials and methods

The solvation parameters were calibrated from data collected from trajectories of 11 proteins for which high-resolution X-ray experimental structures were available (1ARK, 1CEI, 1SRO, 2GB1, 3CI2, 4ICB, 1A9U, 1VCA, 1TBP, 1AYE, and T0139) in the Protein Data Bank (Berman et al. 2000). Another set of unrelated proteins (1JHS, 1JZB, 1K40, 1KEX, 1KOE, 1LIT, 1LSY) was used for validation of the model. The X-ray structures were immersed in large rectangular boxes containing $3–5 \times 10^3$ water molecules, and ions were also added with the CMIP procedure (Gelpi et al. 2001) to achieve electroneutrality. They were then minimized, thermalized, and equilibrated using the protocol noted elsewhere (Morreale et al. 2005). At this point, MD simulations were run at constant temperature (298 K) and pressure (1 atm) for at least 5 nsec. SHAKE (Ryckaert et al. 1977) was used to constrain all the bonds at their equilibrium values in conjunction with an integration time step of 2 fsec. Periodic Boundary Conditions and the Particle Mesh Ewald methods were used to treat long-range electrostatic effects (Darden et al. 1993). AMBER-95 (Cornell et al. 1995) and TIP3P (Jorgensen et al. 1983) force fields were used in all cases. All the trajectories were run using AMBER 6.0 (Pearlman et al. 1995; D.A. Case, D.A. Pearlman, J.W. Caldwell, T.E. Cheatham III, W.S. Ross, C.L. Simmerling, T.L. Darden, K.M. Marz, R.V. Stanton, A.L. Cheng, et al. University of California, San Francisco) computer program. Snapshots used for solvation calculations were taken every 20 psec.

Total, polar, and apolar solvent-accessible surfaces were determined using the NACCESS program (S.J. Hubbard and J.M. Thornton, University College London) with all its default parameters for the ensembles collected for the 11 proteins in the training set. No major differences were found in the fitted model by considering the X-ray structures (data not shown). Once the parameters were fitted for each residue type, the predictive power of the model was tested using the seven proteins of the validation set.

A massive analysis was performed by taking the Cluster-90 of PDB, which contains the structure of nonredundant proteins selected, one for each 90% identity cluster. This set was further reduced by eliminating membrane proteins, those forming oligomers, and structures with gaps, missing or anomalous residues, yielding a final database with 1253 proteins. If present, noncovalent ligands were eliminated, though prosthetic groups and structural ions were retained since they are considered to be constituent parts of the structure. By using Equation 5 on this large database, we derived the expected solvation profile (computed as normalized 50-bin histograms) for a given residue type. These profiles are biased toward small (in absolute terms) fractional free energies due to the fact that in globular proteins there are always more residues buried in the interior of the protein than exposed on the surface. Since this bias increases with the size of the protein, a correction term accounting for the different sizes was considered. To this end, the uncorrected solvation profiles were grouped into six categories depending on

the protein size (1–100, 101–200, 201–300, 301–400, 401–500, and >500 residues). For each category, ''background'' residue profiles were determined by placing each residue in all the positions for all the proteins in that category and computing the corresponding fractional solvation by using Equation 5. This calculation was repeated for all the 1253 proteins yielding to 6 × 20 background profiles, which were used to correct the original 6 × 20 profiles (the population in each bin of the corrected profile is equal to that in the uncorrected profile divided by that found in the background one). The final 20 corrected profiles (one for each residue) were generated by scaling the six corrected profiles according to the population of proteins in each group.

Electronic supplemental material

Fitted parameters (α, β, γ) for the function

$$f(x) = \sum_{i=1}^{5} \alpha_i * e^{-\beta_i * (x-\gamma_i)^2} + \alpha_6 * e^{-\beta_6 * |x-\gamma_6|}$$

that describes the probability above background for the existence of a given residue with a certain fractional solvation free energy are given in the Supplemental Material.

Acknowledgments

This work was supported by the Instituto Nacional de Bioinformática (INB-Genoma España), Fundación Ramón-Areces, and the Spanish Ministry of Education and Science (BIO2003-06848, CTQ2005-09365, and Structural Genomic Project). The Molecular Dynamics Extended Library (MoDEL) is supported by the Barcelona Supercomputing Center.

References

Aqvist, J. 1990. Ion water interaction potentials derived from free-energy perturbation simulations. *J. Phys. Chem.* **94:** 8021–8024.

Aqvist, J. and Hansson, T. 1996. Validity of electrostatic linear response in polar solvents. *J. Phys. Chem.* **100:** 9512–9521.

Aqvist, J., Medina, C., and Samuelsson, J.E. 1994. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **7:** 385–391.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Böttcher, C.J.F. 1952. *Theory of electrostatic polarisation.*. Elsevier, Amsterdam.

Carlson, H.A. and Jorgensen, W.L. 1995. An extended linear response method for determining free energies of hydration. *J. Phys. Chem.* **99:** 10667–10673.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. 1995. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **117:** 5179–5197.

Cramer, C.J. and Truhlar, D.F. 1999. Implicit solvation models: Equilibria, structure, spectra and dynamics. *Chem. Rev.* **99:** 2161–2200.

Darden, T.A., York, D.M., and Pedersen, L.G. 1993. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98:** 10089–10092.

Davis, M.E. and McCammon, J.A. 1990. Electrostatics in biomolecular structure and dynamics. *Chem. Rev.* **90:** 509–521.

Gelpi, J.L., Kalko, S.G., Barril, X., Cirera, J., de la Cruz, X., Luque, F.J., and Orozco, M. 2001. Classical molecular interaction potentials: Improved setup procedure in molecular dynamics simulations of proteins. *Proteins* **45:** 428–437.

Gilson, M.K. and Honig, B. 1988. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies and conformational analysis. *Proteins* **4:** 7–18.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79:** 926–935.

King, U. and Warshel, A. 1990. Investigation of the free-energy functions for electron-transfer reactions. *J. Chem. Phys.* **93:** 8682–8692.

Kong, Y.S. and Warshel, A. 1995. Linear free energy relationships with quantum mechanical corrections: Classical and quantum mechanical rate constants for hydride transfer between NAD+ analogs in solution. *J. Am. Chem. Soc.* **117:** 6234–6242.

Lee, F.S., Chu, Z.T., Bolger, M.B., and Warshel, A. 1992. Calculations of antibody–antigen interactions: Microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603. *Protein Eng.* **5:** 215–228.

Lee, F.S., Chu, Z.T., and Warshel, A. 1993. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs. *J. Comput. Chem.* **14:** 161–185.

Luque, F.J., Bofill, J.M., and Orozco, M. 1995. New strategies to incorporate the solvent polarization in self-consistent reaction field and free-energy perturbation simulations. *J. Chem. Phys.* **103:** 10183–10191.

Luque, F.J., Curutchet, C., Muñoz-Muriedas, J., Bidon-Chanal, A., Soteras, I., Morreale, A., Gelpí, J.L., and Orozco, M. 2003. Continuum solvation models: Dissecting the free energy of solvation. *Phys. Chem. Chem. Phys.* **5:** 3827–3836.

Morreale, A., Gelpí, J.L., Luque, F.J., and Orozco, M. 2003. Continuum and discrete calculation of fractional contributions to solvation free energy. *J. Comput. Chem.* **24:** 1610–1623.

Morreale, A., de la Cruz, X., Meyer, T., Gelpí, J.L., Luque, F.J., and Orozco, M. 2004. Linear response theory: An alternative to PB and GB methods for the analysis of molecular dynamics trajectories? *Proteins* **57:** 458–467.

Morreale, A., de la Cruz, X., Meyer, T., Gelpí, J.L., Luque, F.J., and Orozco, M. 2005. Partition of protein solvation into group contributions from molecular dynamics simulations. *Proteins* **58:** 101–109.

Muñoz, J., Barril, X., Hernández, B., Orozco, M., and Luque, F.J. 2002. Hydrophobic similarity between molecules: A MST-based hydrophobic similarity index. *J. Comput. Chem.* **23:** 554–563.

Orozco, M. and Luque, F.J. 1997. Generalized linear response approximation in discrete methods. *Chem. Phys. Lett.* **265:** 473–480.

Orozco, M. and Luque, F.J. 2000. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.* **100:** 4187 4225.

Orozco, M., Luque, F.J., Habibolahzadeh, D., and Gao, J. 1995. The polarization contribution to the free energy of hydration. *J. Chem. Phys.* **102:** 6145–6152.

Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P.A. 1995. AMBER: A package of computer programs for applying molecular mechanics normal mode analysis molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **91:** 1–41.

Pierce, A.C. and Jorgensen, W.L. 2001. Estimation of binding affinities for selective thrombin inhibitors via Monte Carlo simulations. *J. Med. Chem.* **44:** 1043–1050.

Rivail, J.L. and Rinaldi, D. 1995. Liquid-state quantum chemistry. Computational applications of the polarizable continuum models. In *Computational chemistry reviews of current trends* (ed. J. Leszczynski), pp. 139–174. World Scientific, Singapore.

Roux, B., Yu, H.A., and Karplus, M. 1990. Molecular basis for the Born model of ion solvation. *J. Phys. Chem.* **94:** 4683–4688.

Ryckaert, J.P., Ciccotti, G., and Berendsen, H.J.C. 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **23:** 327–341.

Shoichet, B.K., Baase, W.A., Kuroki, R., and Matthews, B.W. 1995. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci.* **92:** 452–456.

Still, W.C., Tempczyk, A., Hawley, R.C., and Hendrickson, T. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112:** 6127–6129.

Tomasi, J. and Persico, M. 1994. Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. *Chem. Rev.* **94:** 2027–2094.

Warshel, A. and Russell, S.T. 1984. Calculations of electrostatic interactions in biological systems and in solutions. *Q. Rev. Biophys.* **17:** 283–422.

Wesolowski, S.S. and Jorgensen, W.L. 2002. Estimation of binding affinities for celecoxib analogues with COX-2 via Monte Carlo-extended linear response. *Bioorg. Med. Chem. Lett.* **12:** 267–270.