Partition of protein solvation into group contributions from molecular dynamics simulations

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · OCTOBER 2004

Impact Factor: 2.63 · DOI: 10.1002/prot.20292 · Source: PubMed

Xavier de la Cruz

VHIR Vall d’Hebron Research Institute

Tim Meyer

Universitätsmedizin Göttingen

Josep L Gelpí

University of Barcelona

Partition of Protein Solvation Into Group Contributions From Molecular Dynamics Simulations

Antonio Morreale,[1] Xavier de la Cruz,[1,2] Tim Meyer,[1] Josep Lluís Gelpí,[1,3] F. Javier Luque,[4*] and Modesto Orozco[1,3*]

[1]*Molecular Modeling and Bioinformatics Unit, Institut de Recerca Biomèdica, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain*
[2]*Institució Catalana per la Recerca i Estudis Avançats (ICREA), Lluís Companys, 23, Barcelona 08028, Spain*
[3]*Departament de Bioquímica i Biologia Molecular, Facultat de Química, Universitat de Barcelona, Martí i Franquès 1, Barcelona 08028, Spain*
[4]*Departament de Fisicoquímica, Facultat de Farmacia, Universitat de Barcelona, Avgda Diagonal 643, Barcelona 08028, Spain*

ABSTRACT **Linear response theory coupled to molecular dynamics simulations with an explicit solvent representation is used to derive fractional contributions of amino acid residues to the solvation of proteins. The new fractional methods developed here are compared with standard approaches based on empirical 1D and 3D statistical potentials, as well as with estimates obtained from the analysis of classical molecular interaction potentials. The new fractional methods, which have a clear physical basis and explicitly account for the effects due to protein structure and flexibility, provide an accurate picture of the contribution to solvation of different regions of the protein. Proteins 2005;58:101–109. © 2004 Wiley-Liss, Inc.**

Key words: free energy of hydration; protein-solvent interations; molecular dynamics; linear response theory; fractional solvation

INTRODUCTION

Solvation modulates the properties of solutes, changing their nuclear and electronic structures, dynamics, and spectroscopic and reactive properties. In biological systems water is crucial to understand the structure, flexibility, and function of biomolecules, particularly those exhibiting a large charge density like proteins or nucleic acids. Accordingly, the free energy of hydration is an important parameter to gain insight into the behavior of biomolecules.

Solvation is usually described as the combination of three processes involved in the transfer of a solute from the gas phase to the bulk solvent. First, a solute-shape cavity needs to be created in the solvent to accommodate the solute, breaking then the network of interactions between solvent molecules. The work required to generate this cavity is named “cavitation” free energy, which is always an unfavorable contribution to solvation. A second term arises from the dispersion-repulsion interactions between solute and solvent molecules. Finally, a third contribution arises from the change in solute-solvent and solvent-solvent electrostatic interactions upon insertion of the solute charge distribution in the solvent (for a more detailed analysis see refs. 1–5).

The solvation free energy measures the interaction of the entire solute with the surrounding solvent. However, not all the regions of the solute interact with the same strength with the solvent. Therefore, knowledge of the contributions to solvation due to different fragments in the solute is of interest, which explains the popularity of methods for partitioning the transfer or hydration free energy in *structure-activity relationships* studies in drug design projects.[6–10] The partition of the hydration free energy into residue contributions is also valuable to characterize proteins and their pattern of interactions. Thus, exposed hydrophobic regions are putative regions of protein-protein interactions.[11,12] Indeed, certain distribution patterns of hydrophobic-hydrophilic residues are indicative of transmembrane protein segments.[13] In turn, the existence of small hydrophilic patches surrounded by larger hydrophobic areas in partially buried cavities might indicate the existence of ligand binding sites.[14] Finally, the occurrence of large polar-cationic patches in the surface of nucleic acid binding proteins can signal regions that interact with the nucleic acids.[15]

Most methods used to predict partitioning properties in proteins are monodimensional and exploit empirical parameters that characterize, for example, the average transfer free energy of a residue (or its side-chain) between gas phase or an apolar solvent and water.[16–20] Though these monodimensional profiles have proven to be very valuable, they cannot provide an accurate picture of the fractional distribution of hydration around the protein, which also depends on the conformation of the protein.

Grant sponsor: Spanish Ministry of Science; Grant numbers: SAF2002-04282 and BIO2003-06848.

*Correspondence to: Modesto Orozco, Molecular Modeling and Bioinformatics Unit, Institut de Recerca Biomèdica, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain. E-mail: modesto@mmb.pcb.ub.es or F. Javier Luque, Departament de Fisicoquímica, Facultat de Farmacia, Universitat de Barcelona, Avgda Diagonal 643, Barcelona 08028, Spain. E-mail: javier@far1.far.ub.es

Received 12 May 2004; Accepted 30 June 2004

Published online 29 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20292

There are a few empirical fractional methods that roughly account for 3D effects.[21–23] These methods assume that the residue contribution to the solvation is equal to its "intrinsic" hydration (or transfer) free energy (i.e., that of the fully exposed residue) weighted by the accessibility of the residue in the protein (i.e., the ratio between the accessible surface of the residue in the protein and free in solution). These methods are widely used to evaluate protein stability or to signal the presence of residues in unfavorable protein environments.[21,24] However, they also have intrinsic shortcomings, because solvation is a complex, subtle process,[1–5] and many effects other than the solvent exposure modulate the contribution of a given residue in a protein.[25–27]

In this article we analyze the use of linear response theory (LRT) coupled to molecular dynamics (MD) simulation of hydrated proteins as a source of fractional contributions to hydration. Results obtained with this approach are compared with those obtained using empirical 1D and 3D potentials, as well as with those derived from classical molecular interaction potential calculations.

METHODS

Calculation of Solvation Free Energy

The solvation free energy can be determined (see Introduction) from the addition of electrostatic and nonelectrostatic (cavitation and dispersion-repulsion) terms [Eq. (1)]. Generally, the steric contribution is assumed to depend on the solvent accessible surface [Eq. (2)], and a common tension factor is used for all the atoms of the protein (for detailed explanation see ref. 4):

$$\Delta G_{\mathrm{solv}} = \Delta G_{\mathrm{ele}} + \Delta G_{\mathrm{ster}} \tag{1}$$

where $\Delta G_{\mathrm{ster}} = \Delta G_{\mathrm{cav}} + \Delta G_{\mathrm{disp\text{-}rep.}}$

$$\Delta G_{\mathrm{ster}} = \sum_{i} \xi_i \mathrm{SAS}_i \tag{2}$$

where ξ_i and SAS_i stand for the tension factor and solvent accessible surface of atom i.

The electrostatic free energy of solvation is more difficult to compute rigorously.[1–5] However, for a classical nonpolarizable solute and assuming a linear response of the solvent to the generation of the solute charge distribution (LRT), it can be calculated from Eq. (3):

$$\Delta G_{\mathrm{ele}} = \int_0^1 \lambda \sigma V = \frac{1}{2}\left[\lambda^2\right]_0^1 \sigma V = \frac{1}{2}\sigma V \tag{3}$$

where λ is a coupling parameter that varies from 0 (pure solvent) to 1 (fully charged solute), σ is the solute charge distribution, and V is the solvent reaction potential.

LRT lies in the foundations of continuum solvation models,[1–5,28] but it can also be easily implemented in the context of MD or Monte Carlo simulations[4,28–43] combined with a discrete treatment of solvent. In this context, the electrostatic free energy is computed as shown in Eq. (4), where Q are atomic charges of solute (i) and solvent (j), and r_{ij} are interatomic solute-solvent distances. The brackets in Eq. (4) indicate that the calculation is performed in a Boltzmann's ensemble of solute-solvent configurations:

$$\Delta G_{\mathrm{ele}} = \frac{1}{2}\left\langle \sum_{i,j} \frac{Q_i Q_j}{r_{ij}} \right\rangle \tag{4}$$

As explained in detail in a previous communication,[43] the use of Eq. (4) presents many advantages to standard continuum models for the representation of protein solvation. Particularly, the consistency between the force-field parameters used to collect the trajectory and the calculation of ΔG_{ele} is guaranteed. Furthermore, there is not arbitrariness in the choice of the cavity size or the dielectric constants, and specific solute-solvent interactions are explicitly considered. Finally, the temporal coupling between solute and solvent distributions, which is ignored in continuum models, is also accounted for in discrete LRT calculations.

Partition of the Solvation Free Energy in Group Contributions

Though solvation is a property of the entire molecule, it is useful to divide it into atom or group contributions. The partitioning of the steric term is simple, because the solvent accessible surface can be easily integrated into atoms or groups [see Eq. (2)]. Unfortunately, the partition of the electrostatic component is more complicated, because the solvent reaction field is generated by the entire solute charge distribution and any partitioning scheme is arbitrary. We have recently proposed two fractional methods that can be implemented both in the context of continuum and discrete calculations,[25–28] which are denoted surface-based and atom-based methods.[25]

Within the classical LRT framework, the surface-based partitioning scheme determines the electrostatic contribution of a given atom i ($\Delta G_{\mathrm{ele},i}$) from the electrostatic interaction energy of the entire solute with the solvent molecules located in the vicinities of such an atom. This definition leads to Eq. (5), where a solvent molecule S is assigned to atom i based on a simple distance-proximity criterion. The surface-based contribution of a given residue is then determined by adding the individual contributions of its constituting atoms.

$$\Delta G_{\mathrm{ele}} = \sum_{i=1}^{N} \Delta G_{\mathrm{ele},i} = \sum_{i=1}^{N} \frac{1}{2}\left\langle \sum_{i=1}^{N} \sum_{\substack{S=1\\ S\in i}} \sum_{s=1}^{ns} \frac{Q_i Q_s}{\left|r_s - r_i\right|} \right\rangle \tag{5}$$

where ns is the number of atoms in each solvent molecule, N is the number of atoms in the solute, r_S and r_i are the positions of atoms in solvent and solute molecules, and the brackets mean an average using Boltzmann's samplings.

The atom-based partitioning method is perhaps more intuitive and determines the fractional contribution to solvation of atom i from the electrostatic interaction energy of such an atom with all the solvent molecules that surround the protein, then leading to Eq. (6). Again, the fractional contribution of a given residue is computed by adding the individual contributions of its constituting atoms.

TABLE I. Spearman's Correlation Coefficients Between the Hydration Fractional Contributions of Residues Computed with Different Methods

	LRT-ab	CMIP-sb	CMIP-ab	PROSA	RGLLZ	KD	WACS	FP	ESKW
LRT-sb	0.85 ± 0.06	0.78 ± 0.06	0.74 ± 0.09	0.35 ± 0.13	0.73 ± 0.08	0.76 ± 0.06	0.75 ± 0.04	0.73 ± 0.09	0.75 ± 0.06
LRT-ab		0.83 ± 0.07	0.68 ± 0.08	0.38 ± 0.15	0.74 ± 0.06	0.74 ± 0.06	0.75 ± 0.07	0.73 ± 0.05	0.74 ± 0.06
CMIP-sb			0.73 ± 0.07	0.41 ± 0.12	0.69 ± 0.08	0.67 ± 0.09	0.67 ± 0.09	0.68 ± 0.09	0.67 ± 0.08
CMIP-ab				0.29 ± 0.09	0.61 ± 0.13	0.62 ± 0.10	0.59 ± 0.10	0.59 ± 0.12	0.62 ± 0.09
PROSA					0.52 ± 0.04	0.50 ± 0.08	0.38 ± 0.16	0.49 ± 0.06	0.47 ± 0.07
RGLLZ						0.92 ± 0.02	0.72 ± 0.05	0.93 ± 0.03	0.94 ± 0.02
KD							0.83 ± 0.03	0.91 ± 0.03	0.96 ± 0.01
WACS								0.74 ± 0.07	0.81 ± 0.03
FP									0.95 ± 0.02

Values were averaged for the six proteins examined here (values expressed as value ± SD). SB and AB denote fractional contributions obtained using surface-based and atom-based partitioning schemes (see text for details).

$$\Delta G_{\mathrm{ele}} = \sum_{i=1}^{N} \Delta G_{\mathrm{ele},i} = \sum_{i=1}^{N} \frac{1}{2} \left\langle \sum_{j=1}^{S} \sum_{s=1}^{ns} \frac{Q_i Q_s}{|r_s - r_i|} \right\rangle \qquad (6)$$

Previous analysis for small drugs[25] showed that the two LRT-based partitioning schemes provide similar results. However, such a good agreement is not a priori expected for proteins, because exposed residues are not expected to be always those showing the largest contribution to solvation. For example, the surface-based contribution will not assign any contribution to solvation for a charged residue completely buried in the interior of the protein, while depending on the protein environment it might have a non-negligible contribution according to the atom-based approach.

Simulation Details

Systems selected for simulations

The fractional methods were tested using six small soluble proteins (Table I) representative of all α-proteins (1CEI, 4ICB), all β-proteins (1ARK, 1SRO), and α+β proteins (2GB1, 3CI2). For these systems, our previous work[43] demonstrated that MD/LRT methods provide an accurate representation of the solvation free energy.

As described elsewhere,[43,44] the structures of the proteins were built up using either crystallographic or NMR coordinates taken from the Protein Data Bank. Structural ions were included in the calculations in their crystallographic positions, but all the structural waters were removed. Additional ions needed to achieve electroneutrality were added using the titration (TIT) module of CMIP.[45] All the systems were then hydrated using boxes containing between 3204 and 4737 water molecules. These were then optimized, heated, and finally equilibrated for 1 ns of unrestrained MD simulation. MD trajectories were continued for 10 ns at constant pressure (1 atm) and temperature (298 K) with an integration step of 2 fs using SHAKE[46] and Particle Mesh Ewald.[47] AMBER parm95[48] and TIP3P[49] force-fields were used in all cases. Additional details of these simulations are explained elsewhere.[43]

Furthermore, hydration free energy of the 20 free amino acids was determined by means of LRT/MD calculations. For this purpose each amino acid was capped by acetyl and *N*-methyl groups, and hydrated using boxes containing between 596 and 825 water molecules. The system was heated and equilibrated for 200 ps, and MD trajectories were continued for 1 ns at constant pressure (1 atm) and temperature (298 K) with an integration step of 2 fs, SHAKE[46] and Particle Mesh Ewald.[47] AMBER parm95[48] and TIP3P[49] force-fields were used in all cases. The hydration free energy of the free amino acids [ΔG_{sol}^{res}(free)] was then determined by subtracting the fractional contribution of the capping groups to the solvation free energy of the simulated residue. [ΔG_{sol}^{res}(free)] values determined for the 20 residues were compared to the "theoretical" solvation values [ΔG_{sol}^{res}(res)] obtained by assuming a linear relationship between the solvent-accessible surface of the residue and its solvation contribution in the protein [Eq. (7)]. The [ΔG_{sol}^{res}(res)] values determined for each residue type were obtained by averaging the individual estimates in Eq. (7) for the six proteins considered in this study.

$$\Delta G_{sol}^{res}(\mathrm{res}) = \frac{\mathrm{SAS}^{res}(\mathrm{free})}{\mathrm{SAS}^{res}(\mathrm{protein})} \times \Delta G_{sol}^{res}(\mathrm{prot}) \qquad (7)$$

Calculations based on statistical potentials

Hydrophobic 1D profiles were computed using a variety of empirical methods available at the Expasy website (http://us.expasy.org/). Particularly, we used the following methods: ESKW,[16] FP,[17] WACS,[18] KD,[19] and RGLLZ,[20] which exploit slightly different approaches to obtain residue hydrophobicity. In all cases a window of one residue was used to enable comparison with MD/LRT calculations. The hydrophobic potential in PROSA-II[21,24] was also used to obtain 3D hydrophobic profiles adapted to the conformation of the protein. To be consistent with the method, calculations were carried out using the experimental structures and a C_β description of proteins.

Molecular interaction potentials calculations

The TIT utility of the CMIP[45] method was used to create a drop of water around each protein (in its experimental conformation). For this purpose, interaction (electrostatic + van der Waals) energy grids were computed to position a single water molecule at the most favorable grid point. A new grid was then computed for the protein + 1 water system, a new water molecule was added, and so on, until a total of 800 water molecules were added. Based on the interaction energies between these highly structured water molecules and the protein atoms, solvation fractional contributions were determined using the two approaches outlined above. In all the cases the Mehler-Solmajer electrostatic potential was used to account for the dielectric environment.[50] AMBER-95 parameters[48] were used to describe the molecular interactions.

RESULTS AND DISCUSSION

3D Solvation Profiles

The use of fractional methods based on physical potentials allowed us to obtain a 3D distribution pattern of polar/apolar regions in the protein. Particularly, the LRT/MD approach used here provided a time-averaged view of solvation, which takes into account the intrinsic structural flexibility of the protein. 3D profiles were computed from the analysis of a huge number of configurations taken along 10 ns of the MD simulation. Figure 1 contains examples of these profiles, which allowed us to detect and quantify those regions of strong and weak hydration. In all cases the residues exposed to the exterior showed good or very good interaction with water, as expected for soluble cytoplasmatic proteins. However, the ability of different regions of different exposed parts of the protein to interact with water was not the same, but spots of very strong solvation coexisted with regions of less favorable interaction with solvent. Interestingly, the resolution of the 3D profiles was high, which allowed us to detect differences in hydration of residues of the same type but immersed in different chemical environments. Clear examples are Glu^{7}, Glu^{14}, and Glu^{26} of 3CI2, which showed the same percentage of solvent accessible surface displaying fractional solvation free energies (surface-based partitioning scheme) ranging from -18 to -50 kcal/mol. The same situation was found in many other cases, like the triad Glu^{6}, Glu^{12}, and Glu^{52} for 4ICB, which, displaying the same level of exposure, had fractional contribution to solvation free energies ranging from -30 to -70 kcal/mol. Other unexpected results evident from 3D plots are the important contributions or quite buried residues, which, during the dynamics, have transitory interactions with water that yield to non-negligible fractional contributions, even when the surface-based partitioning scheme is used. Examples are Asp^{26} in 1ARK or Arg^{48} in 3CI2.

There are local differences between atom- and surface-based fractional contributions derived from MD/LRT calculations, which are especially important in the buried regions (see below). However, for exposed areas, where the solute-solvent interaction is dominated by close interactions, general profiles obtained from the two partitioning schemes are very similar, as shown in Figure 1. Statistical analysis shows Spearman's correlation in the range 0.8–0.9 between atom- and surface-based partitioning schemes, confirming that in general, at least for exposed residues, atom- and surface-based MD/LRT approaches provide similar results. A more detailed comparison between atom- and surface-based partitioning schemes considering also internal regions is presented below.

Monodimensional Solvation Profiles

Most bioinformatic studies exploit a 1D representation of the solvation properties along the protein sequence. We computed the 1D hydration profile for the six proteins considered here. Results obtained with the linear-response method coupled to MD simulations (MD-LRT) were first compared with those derived from TIT by using molecular interaction potentials, as well as with those derived from PROSA-II and 1D empirical methods. For the sake of clarity only two representative 1D empirical methods (KD and RGLLZ; see Methods) are displayed in the normalized-comparative profiles (see Fig. 2), although all five (ESKW, FP, WACS, KD, RGLLZ, see Simulation Details section) were considered in the statistical analysis (see below).

Hydration profiles (Fig. 2) obtained with the two MD/LRT (surface-based and atom-based) methods are quite similar, as can be seen from the Spearman's correlation coefficients, which range from 0.80 to 0.94 (average 0.85 ± 0.05). Thus, the two MD/LRT fractional approaches provide similar representations of the hydration properties in typical proteins even when buried regions are considered (see Table I). In fact, no major differences are found in the quality of the correlations when all, or only exposed residues (see above), are considered in the comparisons. The largest quantitative discrepancies between atom- and surface-based partitioning schemes are typically found for charged residues, and to a lesser extent for residues close to those charged groups. In some cases, especially for acidic residues, the obtained hydration contribution is larger (in absolute value) for the atom-based approach than for the surface-based approach. This indicates that the residue is partially buried inside the protein, but has, despite its small solvent-accessible surface, intense long-range electrostatic interactions with the solvent. This is the situation found for residues $Asp^{16,18}$ and Glu^{19} in 1ARK, $Asp^{29,60}$, $Glu^{21,23,69}$, Arg^{59}, and Lys^{68} in 1CEI, Asp^{46} in 2GB1, and Asp^{55} and Glu^{61} in 4ICB. In other cases the atom-based contribution is smaller (in absolute value) than that obtained using the surface-based approach. This situation typically corresponds to basic residues with long side chains, especially when they are close to acidic residues, which distort the solvent reaction field in the vicinity of the cationic site. This is the case of Lys^{68} and Arg^{59} in 1CEI, $Lys^{10,31,50}$ in 2GB1, $Lys^{2,53}$ in 2CI2, and $Lys^{26,56}$ in 4ICB. Clearly, averaging of fractional contributions into structure- or sequence-based windows would drastically reduce the discrepancy found between MD/LRT surface- and atom-based schemes. This discrepancy, however, should not be considered an artifact, but a logical consequence of the different partitioning schemes.

The agreement between atom- and surface-based approaches obtained using the CMIP fractional contributions is slightly worse relative to the MD/LRT results (Spearman's correlation coefficients ranging from 0.66 to 0.82, with an average value of 0.73 ± 0.07), suggesting that the use of dynamical information reduces the dependence of the hydration contributions on the partitioning scheme. MD/LRT solvation distributions correlate reasonably well with those obtained from a simple TIT CMIP

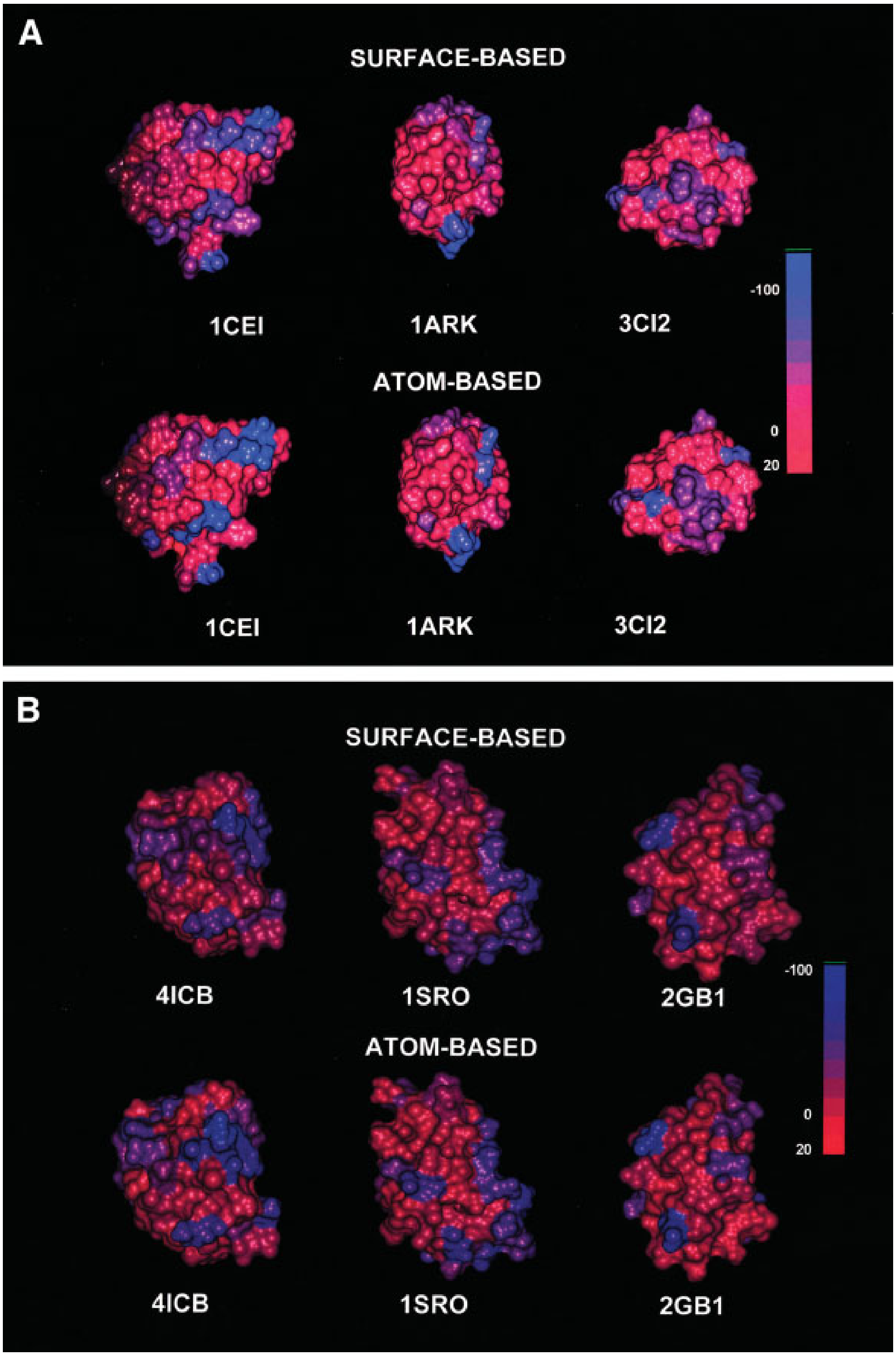

Fig. 1. MD/LRT 3D fractional profiles (the more negative the value, the better the fractional solvation) for the six proteins considered in the study. Calculations were performed using both atom-based and surface-based partitioning schemes.

calculation (see Fig. 2 and Table I). Thus, Spearman's correlation coefficients of 0.78 ± 0.06 and 0.68 ± 0.08 are obtained between MD/LRT and CMIP results when surface- and atom-based partitioning schemes are used, respectively. Reasonable Spearman's coefficients are also obtained for the cross-correlations of the MD/LRT atom-based results to the CMIP surface-based results (0.74 ± 0.09) and vice versa (0.83 ± 0.07; see Table I). In summary, although the dynamical information is necessary to characterize accurately the hydration properties of proteins, simple TIT CMIP calculations performed for a single protein configuration might provide qualitatively acceptable results at a reduced

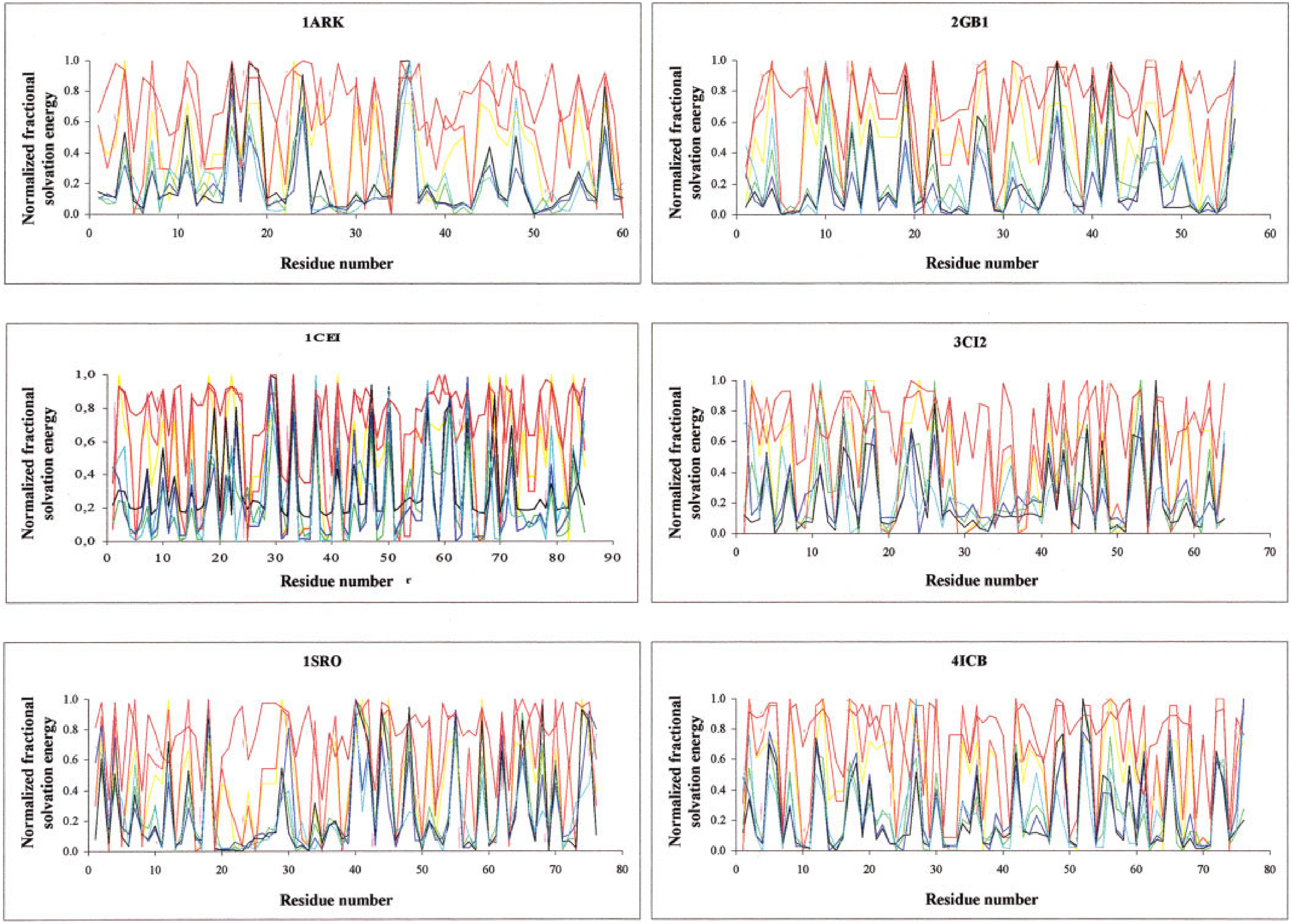

Fig. 2. 1D fractional profiles of the six proteins examined here obtained from different methods using a common window of one residue. RGLLZ yellow; KD red; LRT/MD (surface-based) green; LRT/MD (atom-based) black; CMIP (surface-based) dark blue; CMIP (atom-based) light blue; PROSA brown. All values have been normalized to the range from 0 (worst interaction with solvent) to 1 (best interaction with solvent).

computational cost. This suggests that CMIP calculations might be a useful computational tool to provide qualitative information about fractional solvation properties in coarse-grained genomic-scale projects.

Large discrepancies exist between the hydration profiles obtained with physical models and with PROSA-II (see Fig. 2 and Table I), as reflected in the average Spearman's correlation coefficients obtained when PROSA-II results are compared to MD/LRT (surface-based: 0.35 ± 0.13; atom-based: 0.38 ± 0.14) and CMIP (surface-based: 0.41 ± 0.12; atom-based: 0.29 ± 0.09) results. Accordingly, the representations of solvation obtained from atomic simulations and those derived from the statistical potential implemented in PROSA-II are quite different. The fact that solvation potentials in PROSA-II were not designed to represent the solvation properties of proteins, but to make fast estimations of the relative stability of different conformations of a protein can explain, at least in part, these discrepancies.

Most sequence-based empirical methods rely on the intrinsic hydrophilicity assigned to each residue, which is assumed to be independent of the protein environment. Such an assumption limits the accuracy of these methods. However, the agreement between MD/LRT results and most empirical methods is quite reasonable, with Spearman's correlation coefficients in the range 0.73–0.75 in most cases (see Table I). The agreement is slightly worse when CMIP calculations are considered, and very poor when comparison is made with PROSA-II potentials. Not surprisingly, 1D empirical profiles agree very well amongst them (Spearman's correlation coefficients around 0.9 in most cases), reflecting their formal similarities. These findings suggest that in general the residue's hydrophobicity determines its position within the protein. Overall, our results support the use of simple empirical measures of hydrophobicity/hydrophilicity to obtain a general picture of protein solvation. However, we cannot ignore that a more detailed representation implies the use of accurate 3D approaches like the MD/LRT methods presented here.

Relationship Between Fractional Solvation and Accessible Surface

It is widely assumed that there is a linear relationship between the hydration contribution of a given residue and its exposure to the solvent. Though this assumption is very practical from a computational point of view, it is unclear whether or not it is valid in proteins, where more subtle effects modulate the interaction of a given residue with the solvent molecules. To investigate this point, we used the MD/LRT (surface- and atom-based) fractional scheme to determine the solvation free energies of the 20 free amino acids in water. These values were then compared to the corresponding hydration contributions obtained by extrapolation of the values determined in the different proteins by using Eq. (7).

The results shown in Table II demonstrate that the hydration contribution of the different residues depends strongly on the protein environment, as reflected in the large standard deviations associated with the average $[\Delta G_{sol}^{res}(\text{prot})]$ values. The protein environment largely decreases the strength of the interaction between solvent and residues. As expected, this effect is of particular relevance for apolar residues, which are typically buried inside the protein. However, interaction was cut in half even for charged residues, although they were expected to be exposed on the protein surface.

TABLE II. Average MD/LRT Hydration Fractional Contributions of the 20 Amino Acids in Proteins [$\Delta G_{\mathrm{sol}}^{\mathrm{res}}$ (prot)] and the Values [$\Delta G_{\mathrm{sol}}^{\mathrm{res}}$ (res)] Obtained for Free Amino Acids Derived by Extrapolation of the Values in the Proteins Obtained by Using Eq. (7)

Residue type	$\Delta G_{\mathrm{sol}}^{\mathrm{res}}$ (prot)	$\Delta G_{\mathrm{sol}}^{\mathrm{res}}$ (res)	$\Delta G_{\mathrm{sol}}^{\mathrm{res}}$ (free)
ALA	−9.07 (6.57)	−20.82 (8.57)	−19.09
	−4.74 (2.24)	*−21.22 (38.74)*	
ARG	−33.37 (9.64)	−100.29 (65.02)	−69.40
	−31.64 (19.79)	*−88.04 (97.73)*	
ASN	−19.02 (7.19)	−42.90 (23.79)	−33.26
	−12.45 (3.89)	*−28.74 (14.36)*	
ASP	−40.30 (20.49)	−76.57 (23.40)	−86.76
	−60.08 (24.11)	*−131.39 (88.89)*	
CYS	N.D.	N.D.	N.D.
GLN	−15.92 (3.76)	−32.55 (12.74)	−32.01
	−12.47 (4.45)	*−23.43 (6.63)*	
GLU	−37.73 (16.29)	−78.42 (43.06)	−85.72
	−52.82 (18.54)	*−116.41 (77.28)*	
GLY	−6.32 (3.19)	−14.66 (6.64)	−19.27
	−3.89 (2.15)	*−9.69 (7.42)*	
HIS	−7.03 (2.56)	−55.11 (42.41)	−29.70
	−8.70 (9.04)	*−52.96 (42.17)*	
ILE	−4.64 (3.35)	−26.50 (23.09)	−17.56
	−2.89 (2.16)	*−20.37 (21.54)*	
LEU	−4.56 (3.26)	−23.47 (14.80)	−19.70
	−2.58 (2.96)	*−17.27 (32.59)*	
LYS	−37.38 (10.47)	−79.72 (15.09)	−75.68
	−31.09 (15.55)	*−63.41 (33.22)*	
MET	−10.76 (6.83)	−23.30 (14.80)	−20.58
	−5.48 (3.10)	*−12.46 (9.34)*	
PHE	−3.68 (2.89)	−32.49 (21.74)	−23.34
	−1.26 (1.46)	*−13.94 (28.02)*	
PRO	−8.15 (5.16)	−17.55 (10.30)	−17.74
	−3.82 (2.48)	*−16.94 (38.55)*	
SER	−10.12 (4.80)	−32.09 (14.11)	−25.27
	−6.75 (2.64)	*−22.16 (11.25)*	
THR	−9.83 (4.36)	−25.51 (6.60)	−23.38
	−6.43 (2.83)	*−17.72 (6.60)*	
TRP	−16.02 (1.73)	−54.70 (13.97)	−25.65
	−7.00 (5.04)	*−24.28 (15.72)*	
TYR	−12.36 (8.49)	−44.06 (31.83)	−24.62
	−6.10 (3.75)	*−20.48 (20.20)*	
VAL	−7.33 (6.00)	−25.92 (18.20)	−19.13
	−3.08 (2.56)	*−13.20 (17.66)*	

The MD/LRT hydration free energies determined for the free amino acids ($\Delta G_{\mathrm{sol}}^{\mathrm{res}}$ (free)) are also given. To avoid numerical uncertainties, residues with a fraction of solvent- accessible surface below 2% were not included in the analysis. Values obtained using the surface-based and atom-based partitioning schemes are in roman and in italics, respectively. Standard deviations are displayed in parentheses. All values are in kcal/mol. N.D. = no data.

There is a reasonable relationship between the solvent-accessible surface of a residue and its average contribution to the hydration free energy (see Fig. 3). Thus, the

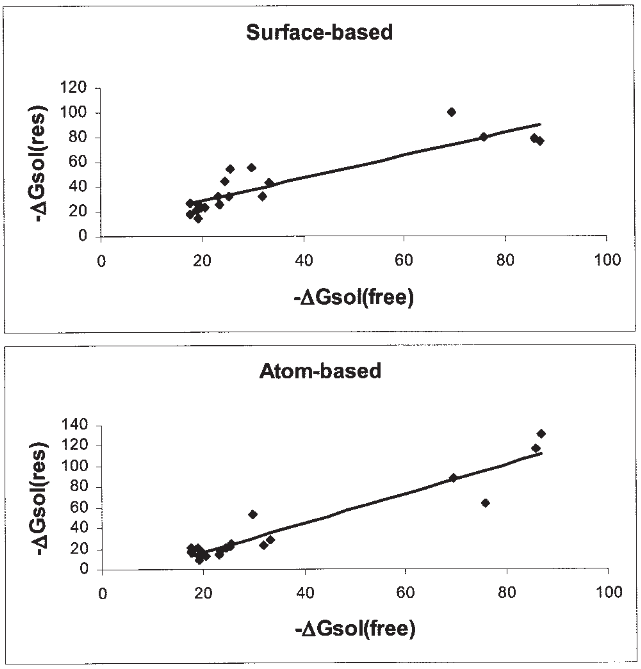

Fig. 3. Comparison between the hydration contribution determined from MD/LRT simulations for 20 amino acids free in water [ΔG_{sol}^{res}(free)] and from extrapolation of the hydration contribution obtained for the residues in the protein environment by using Eq. (7) [ΔG_{sol}^{res}(res)]. Values are in kcal/mol.

theoretical estimate of the hydration free energy of a free residue obtained by using Eq. (7) [ΔG_{sol}^{res}(res)] and the value obtained from simulation of the free residue [ΔG_{sol}^{res}(free)] correlate with Pearson's coefficients of 0.89 (surface-based) and 0.95 (atom-based). However, we cannot ignore that such correlations are probably dominated by the pairing division between charged and neutral residues, and partial correlations between the charged or neutral groups are not as good. [ΔG_{sol}^{res}(res)] values obtained from Eq. (7) are generally larger than those computed directly for the free amino acid [ΔG_{sol}^{res}(free)], as noted in scaling coefficients of 0.9 (surface-based) and 0.85 (atom-based) approach, suggesting that the reduction in the solvent-accessible surface is not the only factor that explains the decrease in the solvation contribution of residues in the protein.

The preceding discussion warns against the assumption of general validity of linear relationships as those given in Eq. (7) to provide accurate hydration free energies of residues in a protein environment. Thus, the averaged [ΔG_{sol}^{res}(prot)] values have large standard deviations, indicating that two residues of a certain type equally buried in the protein might exhibit very different solvation contributions depending on the electrostatic environment. Furthermore, the correlations between the [ΔG_{sol}^{res}(res)] and [ΔG_{sol}^{res}(free)] values computed separately for any of the six proteins are not good ($r <$ 0.6 in most cases), irrespective of the partitioning scheme (atom- or surface-based) used to determine [ΔG_{sol}^{res}(prot)]. In summary, although a qualitative relationship between the hydration contribution of a residue and its solvent exposure is found when a large amount of data is averaged, this relationship does not hold to quantitatively predict the contribution of a specific residue in a specific protein.

CONCLUSIONS

MD simulations were used in conjunction with LRT to examine two different partitioning (atom- and surface-based) schemes to derive hydration contributions of residues in proteins. This procedure provides a very detailed 3D picture of how the solvent interacts with the protein. This information can be condensed to 1D profiles that qualitatively agree with models based on classical molecular interaction calculations and with sequence-based empirical approaches. The agreement is less satisfactory with empirical estimates derived from solvent-accessible surfaces like PROSA-II.

Our results demonstrate the existence of a rough linear relationship that allows us to estimate the hydration contribution of a residue as the product of its intrinsic solvation (i.e., the hydration free energy of the free residue) and its solvent exposure in the protein. However, this relationship is too crude to provide accurate measures, and caution is necessary in using this approach for estimating the hydration contribution of specific residues.

ACKNOWLEDGMENTS

We thank the CIRI-CEPBA Consortium for computational facilities. A.M. and T.M. are fellows of the Spanish Ministry of Science.

REFERENCES

1. Böttcher CJF. Theory of Electrostatic Polarisation. Amsterdam: Elsevier; 1952.
2. Tomasi J, Persico M. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. Chem Rev 1994;94:2027–2094.
3. Cramer CJ, Truhlar DF. Implicit Solvation Models: Equilibria Structure Spectra and Dynamics. Chem Rev 1999;99:2161–2200.
4. Orozco M, Luque FJ. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. Chem Rev 2000;100: 4187–4225.
5. Rivail JL, Rinaldi D. Liquid-state quantum chemistry: computational applications of the polarizable continuum models. In Leszczynski J, editor. Computational chemistry reviews of current trends. Singapore: World Scientific; 1995. p 139.
6. Martin YC. Quantitative Drug Design. A Critical Introduction. New York: Marcel Dekker; 1978.
7. Horvath AL. Molecular Design. Amsterdam: Elsevier; 1992.
8. Hansch C, Leo A. Exploring QSAR: Fundaments and Applications in Chemistry and Biology. Washington, DC: American Chemical Society; 1995.
9. Leo A, Hansch C, Elkins D. Partition coefficients and their uses. Chem Rev 1971;71:525–616.
10. Rekker RF. The hydrophobic fragmental constant. New York: Elsevier; 1977.
11. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
12. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
13. von Heijne G. Membrane proteins: From sequence to structure. Ann Rev Biophys Biomol Struct 1994;23:167–192.
14. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–121.
15. Jones S, Shanahan HP, Berman HM, Thornton JM. Using electro-

static potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 2003;31:7189–7198.

16. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 1984;179:125–142.
17. Fauchere JL, Plisca V. Hydrophobic parameters π of amino acid side chains from the partitioning of N-acetyl-amino acid amides. Eur J Med Chem 1983;18:369–375.
18. Wolfenden R, Andersson L, Cullis PM, Sothgate CCB. Affinities of amino acid chains for solvent water. Biochemistry 1981;20:849–855.
19. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
20. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus H. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229: 834–838.
21. Sippl MJ. Boltzmann's Principle, Knowledge-Bases Mean Fields and Protein Folding. An Approach to the Computational Determination of Protein Structures. J Comput-Aided Mol Des 1993;7:473–501.
22. Eiseberg D, McLachan AD. Solvation Energy in Protein Folding and Binding. Nature 1986;319:199–203.
23. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1999;288:477–487.
24. Sipple MA. ProSa 2003. PROtein Structure Analysis. Version 4.0;17:355–362.
25. Morreale A, Gelpí JL, Luque FJ, Orozco M. Continuum and Discrete Calculation of Fractional Contributions to Solvation Free Energy. J Comput Chem 2003;24:1610–1623.
26. Muñoz J, Barril X, Hernández B, Orozco M, Luque FJ. Hydrophobic similarity between molecules: A MST-based hydrophobic similarity index. J Comput Chem 2002;23:554–563.
27. Luque FJ, Bofill JM, Orozco M. New strategies to incorporate the solvent polarization in self-consistent reaction field and free-energy perturbation simulations. J Chem Phys 1995;103:10183–10191.
28. Luque FJ, Curutchet C, Muñoz-Muriedas J, Bidon-Chanal A, Soteras I, Morreale A, Gelpí JL, Orozco M. Continuum solvation models: Dissecting the free energy of solvation. Phys Chem Chem Phys 2003;5:3827–3836.
29. Lee FS, Chu ZT, Warshel A. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs. J Comput Chem 1993;14:161–185.
30. Lee FS, Chu ZT, Bolger MB, Warshel A. Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603. Prot Eng 1992;5:215–228.
31. Warshel A, Rusell ST. Calculations of electrostatic interactions in biological systems and in solutions. Quart Rev Biophys 1984;17: 283–422.
32. Roux B, Yu HA, Karplus M. Molecular basis for the Born model of ion solvation. J Phys Chem 1990;94:4683–4688.
33. King U, Warshel A. Investigation of the free-energy functions for electron-transfer reactions. J Chem Phys 1990;93:8682–8692.
34. Kong YS, Warshel A. Linear Free Energy Relationships with Quantum Mechanical Corrections: Classical and Quantum Mechanical Rate Constants for Hydride Transfer between NAD+ Analogs in Solutions. J Am Chem Soc 1995;117:6234–6242.
35. Aqvist J. Ion water interaction potentials derived from free-energy perturbation simulations. J Phys Chem 1990;94:8021–8024.
36. Aqvist J, Medina C, Samuelsson JE. A new method for predicting binding affinity in computer-aided drug design. Prot Eng 1994;7: 385–391.
37. Aqvist J, Hansson T. Validity of Electrostatic Linear Response in Polar Solvents. J Phys Chem 1996;100:9512–9521.
38. Orozco M, Luque FJ, Habibolahzadeh D, Gao J. The polarization contribution to the free energy of hydration. J Chem Phys 1995;102:6145–6152.
39. Carlson HA, Jorgensen WL. An Extended Linear Response Method for Determining Free Energies of Hydration. J Phys Chem 1995;99:10667–10673.
40. Pierce AC, Jorgensen WL. Estimation of Binding Affinities for Selective Thrombin Inhibitors via Monte Carlo Simulations. J Med Chem 2001;44:1043–1050.
41. Wesolowski SS, Jorgensen WL. Estimation of Binding Affinities for Celecoxib Analogues with COX-2 via Monte Carlo-Extended Linear Response. Bioorg Med Chem Lett 2002;12:267–270.
42. Orozco M, Luque FJ. Generalized Linear Response Approximation in Discrete Methods. Chem Phys Lett 1997;265:473–480.
43. Morreale A, de la Cruz X, Meyer T, Gelpi JL, Luque FJ, Orozco M. Linear Response Theory: An alternative to PB and GB methods for the analysis of molecular dynamics trajectories? Proteins 2004. Forthcoming.
44. Meyer T, de la Cruz X, Luque FJ, Orozco M. To be published.
45. Gelpi JL, Kalko SG, Barril X, Cirera J, de la Cruz X, Luque FJ, Orozco M. Classical Molecular Interaction Potentials: Improved Setup Procedure in Molecular Dynamics Simulations of Proteins. Proteins 2001;45:428–437.
46. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 1977;23:327–341.
47. Darden TA, York DM, Pedersen LG. Particle mesh Ewald: an N×log(N) method for Ewald sums in large systems. J Chem Phys 1993;98:10089–10092.
48. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Fergurson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A Second Generation Force Field for the Simulation of Proteins Nucleic Acids and Organic Molecules. J Am Chem Soc 1995;117:5179–5197.
49. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79:926–935.
50. Mehler EL, Solmajer T. Electrostatic Effects in Proteins: Comparison of Dielectric and Charge Models. Protein Eng 1991;4:903–910.