JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

A Protein Solvation Model Based on Residue Burial

Nicoletta Ceres, Marco Pasi, and Richard Lavery*

Bases Moléculaires et Structurales des Systèmes Infectieux, Université Lyon I/CNRS UMR 5086, IBCP, 7 Passage du Vercors, 69367 Lyon, France

Supporting Information

ABSTRACT: The influence of solvent on the individual amino acids of a protein depends not simply on their surface exposure but rather on the degree of their burial within the structure. This property can be related to a simple geometrical measure termed circular variance. Circular variance depends on the spatial distribution of neighboring residues and varies from zero to one as a residue becomes buried. Its only adjustable parameter is a cutoff distance for selecting neighbors. Here, we show that circular variance can be used to build a fast and effective model of protein solvation energies. For this, we combine a coarse-grain protein representation with statistical potentials derived by Boltzmann inversion of circular variance probability distributions for different classes of pseudoatom within a large protein structure database. The method is shown to work well for distinguishing native protein structures from decoy structures generated in a variety of ways. It can also be used to detect specific residues in unfavorable solvent environments. Compared to surface accessibility, circular variance calculations are faster, less sensitive to small conformational changes, and able to account for the longer-range interactions that characterize the electrostatic component of solvent effects. The resulting solvation energies can be used alone or as part of a more general coarse-grain protein model.

INTRODUCTION

Coarse graining is developing rapidly in the field of biomolecular simulations. By grouping sets of atoms into single pseudoatoms, it is possible to reduce the number of degrees of freedom in a macromolecular system, to speed up energy and force calculations, and also to accelerate movement through conformational space, thanks to the smoothing of the energy hypersurface associated with coarse-grain representations. However, since biological systems depend on the presence of water, coarse graining needs to be applied not only to the solute molecules but also to the solvent. This can be done in a way analogous to the treatment of the solute, by grouping together several water molecules (typically four) to form pseudosolvent particles,[1] but if this is still too slow, then it is possible to pass to continuum solvent representations. These can be based on classical electrostatics via the Poisson−Boltzmann equation, with a solute envelope separating the high dielectric solvent from the low dielectric interior of the biomacromolecule.[2] Since the numerical solutions of the Poisson−Boltzmann equation are still slow to compute, a number of approximate models, based on the generalized Born equation, have also been developed.[3−5] Alternatively, very fast approaches based on interparticle distances[6] or on the exposure of particles to solvent can be used.

In the latter case, calculations generally involve solvent-accessible surface areas (SASA). Building on the pioneering work of Eisenberg and McLachlan,[7] similar methods are still widely used as a component of coarse-grain models of both small molecules[8] and biomacromolecules and are also a common component of continuum electrostatic solvent models where they are used to represent nonpolar contributions.[9,10] For proteins, SASA-based models are widely used for both structure prediction and protein−protein docking.[11,12] Although analytic derivatives of surface accessibility can be calculated[13−15] and progress continues to be made in this area,[16] surface accessibility itself has the disadvantage of being sensitive to the definition of the solute envelope and to small changes in conformation, which can lead a given atom, or residue, to become abruptly inaccessible to the solvent. This has fuelled an interest in the concept of residue depth[17] and attempts to integrate this property into SASA-like solvation models.[18,19]

We would like to propose an alternative approach for treating solvent interactions that replaces accessibility calculations with circular variance (CV), a simple measure based on the vectorial distribution of the neighbors around a point.[20] This concept, used in directional statistics to describe 2D angular distributions, can be extended to 3D and, in the case of molecules, can quantify the extent to which an atom is buried within the structure and thus protected from solvent.[21] CV has simple analytical derivatives and can be applied equally well to atomic or coarse-grain representations. Unlike SASA, CV does not change abruptly when an atom drops below the solute surface but rather changes smoothly from 0.0 to 1.0 in passing from a fully exposed atom to a fully buried one (see Figure 1). In addition, CV is calculated for neighbors within a chosen cutoff distance (chosen here to be 10 Å), allowing it to take into account not only atoms that are in direct steric contact but also those that could have an electrostatic impact on the solvation of the target atom (note that 10 Å is close to the Debye length at physiological ionic strength).

In this work, we develop a simple protein solvation model based on CV and using a coarse-grain protein representation. We have chosen to represent protein conformations using the coarse-grain model proposed by Zacharias,[22] where each amino acid has one backbone pseudoatom placed at Cα and either one or two pseudoatoms representing the side chain (excepting

Received: February 22, 2012
Published: April 19, 2012

dx.doi.org/10.1021/ct3001552 | J. Chem. Theory Comput. 2012, 8, 2141−2144

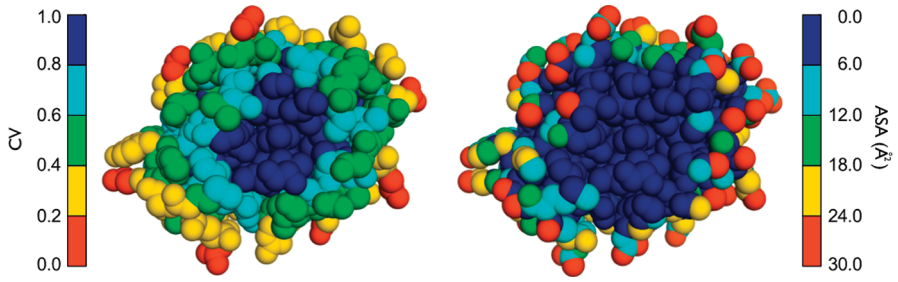

Figure 1. Circular variance (CV, left) versus accessible surface area (ASA, right) for a heat shock protein (PDB 3FH2). The view illustrates a 12 Å slice through the center of the protein. The coloring goes from orange for exposed atoms (CV ≈ 0, ASA ≈ 30 Å^2) to blue for buried atoms (CV ≈ 1, ASA ≈ 0 Å^2). The CV values calculated for the pseudoatom model of the heat shock protein have been mapped onto the all-atom representation. The ASA values were computed using NACCESS with standard parameters.[23]

glycine, which has none). Smaller side chains have a single pseudoatom (termed SC1) at the geometrical center of the side chain heavy atoms, while larger side chains (Arg, Gln, Glu, His, Lys, Met, Phe, Trp and Tyr) have one pseudoatom (SC1) at the midpoint of the Cβ−Cγ bond and a second (SC2) at the geometrical center of the heavy atoms beyond Cβ. This representation generates a total of 48 pseudoatom classes (assuming Cα is residue dependent).

COMPUTATIONAL METHODS

The CV of pseudoatom i (CV_i) is calculated as one minus the modulus of the vector sum of the unit vectors $r_i/|r_i|$ from i to all neighbors within a cutoff distance r_c (here we set r_c to 10 Å), divided by the number of these vectors n_i:

$$CV_i = 1 - \frac{1}{n_i}\left|\sum_{j\neq i, r_i\leq r_c}\frac{r_i}{|r_i|}\right|$$

In an earlier publication, Mezei somewhat surprisingly found that a modified formula (termed CV^W), which gives more weight to distant neighbors, performed better in some cases:[21]

$$CV_i^W = 1 - \frac{\left|\sum_{j\neq i, r_i\leq r_c} r_i\right|}{\sum_{j\neq i, r_i\leq r_c}|r_i|}$$

We were able to show that this formulation actually works not by giving more weight to distant neighbors but by giving less weight to bound neighbors, whose spatial distribution around the center i is virtually constant. Calculating CV_i without taking bound neighbors into account gives results that are indistinguishable from CV_i^W (see Figure S1 in the Supporting Information). This formulation was found to the best for our purposes and is used here.

In order to obtain an effective solvation energy (S_i) for pseudoatom i, we first calculate the probability distribution (P_i) of CV_i values for each class of pseudoatom using a large database of experimental protein structures. This database contains a total of 1202 structures of globular, soluble proteins sharing less than 20% sequence identity and having no missing atoms, selected from the corresponding PISCES data set.[24] Figure S2 in the Supporting Information shows the CV probability distributions for all classes of pseudoatom. As expected, hydrophilic pseudoatoms are generally exposed to solvent with CV distributions weighted toward low values, while hydrophobic pseudoatoms are generally buried and have distributions weighted to high values.

We can now calculate an effective solvation energy S_i for each pseudoatom class using Boltzmann inversion (where k_BT = 0.593 kcal mol^{-1} at 25 °C):

$$S_i = -k_BT\ln\left(\frac{P_i}{P_R}\right)$$

Note that the observed CV probability distributions P_i are weighted by a reference probability P_R calculated by averaging the P_i over all structures and all pseudoatom classes (see Supporting Information). In order to simplify calculations, the solvent energies can be fitted accurately using third-degree polynomials (see Table S1 in the Supporting Information). The total solvation energy (CVSE) for a molecule with N pseudoatoms then becomes

$$CVSE = \sum_{i=1}^{N} S_i = \sum_{i=1}^{N}(a_k CV_i^3 + b_k CV_i^2 + c_k CV_i + d_k)$$

where k is the class of pseudoatom i. Because of the finite range of CV, each pseudoatom will have a minimum value of its solvation energy, termed S_i^M (see Table S1, Supporting Information). This corresponds to the pseudoatom being in its ideal environment with respect to the solvent. More positive values within a given protein structure, S_i^P, quantify unfavorable environments, and the difference $S_i^P - S_i^M$, which we term a solvation energy deficit (SED), can be used to analyze specific protein conformations on a residue-by-residue basis.

In terms of efficiency, CV calculations only require distance calculations between neighboring pseudoatoms (typically of the order of 10 distances per center). This procedure can be accelerated by maintaining a pair list with an appropriate cutoff distance. The total number of distance calculations is roughly 100 times less than for 2D slice[25] or surface grid methods[26] and the corresponding analytic derivatives are easily obtained.

RESULTS

We can now ask whether CV has been able to capture the basic physics underlying solvation energies. In order to test this, we calculated the total CV solvation energy (CVSE) for three sets of protein decoys created using different strategies.[27] The solvation energy alone would not normally be expected to correctly distinguish native states from decoys, since it ignores changes in conformational energy (hydrogen bonding, steric or torsional strain, etc.). However, the results we obtain show that poor solvation is a dominant factor. These calculations use only side chain solvation energies since the Cα and SC1 probability distributions were generally very similar. We also checked that

changing the CV cutoff from 10 Å to either 8 or 12 Å had little effect.

The first test set (termed "misfold") contains a single decoy for each of 26 proteins, created by threading the native sequence onto another protein structure and then optimizing the conformation using Monte Carlo simulated annealing.[28] In all cases, the CV solvation score was more negative for the native protein than for the decoy with an average energy difference of −33 kcal mol^{-1} (Figure 2).

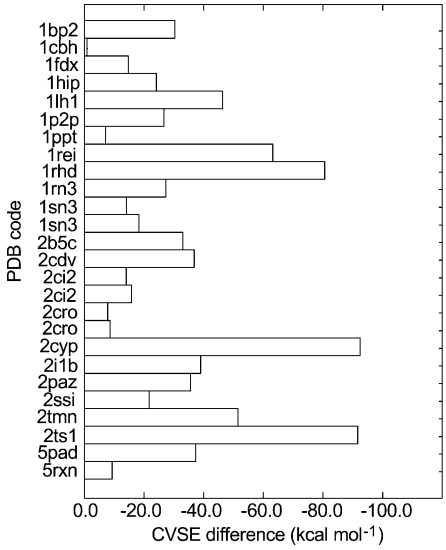

Figure 2. CV solvation energy differences between the native and the decoy structure of the 26 proteins belonging to the "misfold" decoy set.[28] The PDB code for each protein is shown on the left of the figure.

The second and third decoy sets (termed "4state_reduced" and "lattice_ssfit") contain between 600 and 2000 decoys for each protein. The former set involved a combinatorial generation of multiple substates for 10 residues within each protein followed by energy minimization,[29] while the latter set used conformations initially generated on a tetrahedral lattice and then optimized in off-lattice space.[30]

Figure 3 shows that, with one notable exception, the CV solvation energy places the native structure at, or close to, the optimal solvation energy compared to the decoys. For the 4state_reduced set, the native structure is within the best 5% of energies in all but two cases (1sn3 7% and 2cro 8%), and for the lattice_ssfit set, the native structure has the optimal energy in all but one case (1dtk-A, in the top 0.2%).

The notable exception involves the protein 1trl-A, a C-terminal fragment of thermolysin. The results in Figure 3 refer to this protein fragment in a monomeric state, although it has been shown to exist in solution as a dimer.[31] Using the minimal solvation energy for each class of pseudoatom, we can calculate the solvation energy deficit for each pseudoatom in a given protein structure, as described above. Significantly positive SED values indicate pseudoatoms that are in unfavorable solvent environments. The results for 1trl are shown in Figure 4.

We can see that several residues on the surface of 1trl-A are indeed poorly solvated. Taking into account the dimeric structure improves the solvation energy of the monomer by −9.4 kcal mol^{-1} and also explains why the native monomer is relatively unstable compared to many of its decoy structures.

The last example uses more than 6000 decoy conformations of the thermostable subdomain of the chicken villin headpiece

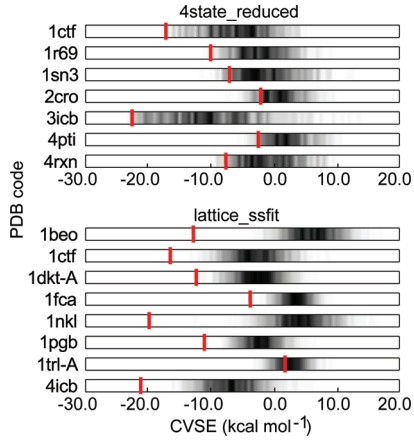

Figure 3. Distribution of CVSE for protein decoys belonging to the 4state_reduced (above) and lattice_ssfit (below) sets. The PDB codes for each protein are shown on the left. Darker shading indicates a higher density of decoy solvation energies, and the red bars indicate the solvation energy of the native protein.

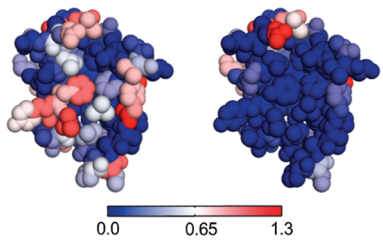

Figure 4. SED for each pseudoatom at the dimerization interface of the protein fragment 1trl, calculated using the monomer structure (left) or using the dimer (right). The blue to red scale (kcal mol^{-1}) indicates increasingly positive SED values (poorer solvent environments). The values for these all-atom representations have been mapped from calculations with the corresponding pseudoatom model. Terminal residues are omitted for clarity.

generated during molecular dynamics simulations in explicit solvent, starting either from the native structure or from various non-native structures. Each decoy conformation extracted from the simulations was energy minimized using an implicit solvent model.[32] This set includes decoys up to 12 Å rmsd (calculated using only Cα atoms) away from the native conformation.

Figure 5 shows that the best CV solvation energies are obtained for snapshots from the simulation beginning with the native structure and also shows a reasonable correlation between the CV solvation energies and the rmsd values of each simulation. Quantitatively, the CV solvation energies perform as well as much more costly methods involving both molecular mechanics and continuum electrostatic terms (MM/PBSA or MM/GBSA) using the quality measures applied in the publication by Fogolari et al.[32] (see Table S2 in the Supporting Information).

CONCLUSIONS

In conclusion, the results presented here suggest that CV is indeed a useful starting point for modeling protein solvation. It has the advantage of measuring the degree of burial, rather than simple surface accessibility, taking into account more distant neighbors that can contribute electrostatically and being sufficiently fast to make it an interesting option in coarse-grain approaches. We have also found that excluding bound

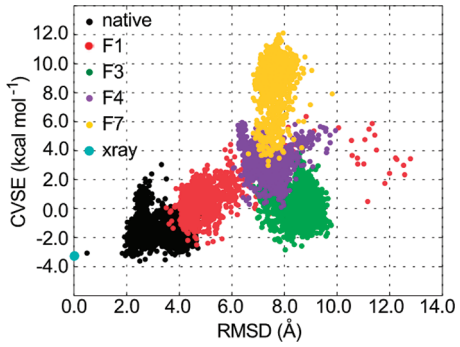

Figure 5. Cα rmsd values with respect to the native structure of the thermostable domain of the villin headpiece (PDB 1VIII) versus CV solvation energies (CVSE). The points correspond to energy-minimized snapshots from molecular dynamics simulations using different starting conformations, native or non-native (F1, F3, F4, F7). Each simulation is color coded. The native X-ray structure is indicated by the large blue circle (bottom left).

neighbors from CV calculations leads to the better results for all the decoy tests we performed (typically improving the scores of the worst cases by several percent and also improving the momomer solvation energy for dimeric 1trl by 3 kcal mol^{-1}).

Our model could be adapted in a number of ways. The CV-based solvation term could be derived for an all-atom representation or fitted to data other than experimental structures. It could equally well be integrated into a more general coarse-grain force field, using an iterative refinement procedure to distribute energetic contributions between one- and two-body terms. We are indeed using this approach in the PaLaCe model currently under development.

Finally, as shown by the 1trl protein fragment, looking at departures from optimal solvation energy on a residue-by-residue basis may be a useful way of detecting anomalies with protein structures.

■ ASSOCIATED CONTENT

Supporting Information

Details of the protein structural data set, the calculation of CV reference probabilities, CV probability distributions and solvation energies for the different classes of pseudoatoms, and complete results for the "misfold" set of protein decoys. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: richard.lavery@ibcp.fr

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors wish to aknowledge funding from CNRS, the Rhône-Alpes project CIBLE, and the ANR project EXPE-NANTIO.

■ REFERENCES

(1) Yesylevskyy, S. O.; Schäfer, L. V.; Sengupta, D.; Marrink, S. J. *PLoS Comput. Biol.* **2010**, *6*, e1000810.
(2) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137−143.
(3) Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275−291.
(4) Bashford, D.; Case, D. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.
(5) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265−84.
(6) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59−107.
(7) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199−203.
(8) Knight, J. L.; Brooks, C. L. *J. Comput. Chem.* **2011**, *32*, 2909−2923.
(9) Cerutti, D. S.; Ten Eyck, L. F.; McCammon, J. A.; et al. *J. Chem. Theory Comput.* **2005**, *1*, 143−152.
(10) Lopes, A.; Alexandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T. *Proteins* **2007**, *67*, 853−867.
(11) Wang, T.; Wade, R. C. *Proteins* **2003**, *50*, 158−169.
(12) Li, L.; Guo, D.; Huang, Y.; Liu, S.; Xiao, Y. *BMC Bioinformatics* **2011**, *12*, 36−44.
(13) Fraczkiewicz, R.; Braun, W. *J. Comput. Chem.* **1998**, *19*, 319−333.
(14) Sridharan, S.; Nicholls, A.; Sharp, K. A. *J. Comput. Chem.* **1995**, *16*, 1038−1044.
(15) Perrot, G.; Cheng, B.; Gibson, K. D.; Vila, J.; Palmer, K. A.; Nayeem, A.; Maigret, B.; Scheraga, H. A. *J. Comput. Chem.* **1992**, *13*, 1−11.
(16) Klenin, K. V.; Tristram, F.; Strunk, T.; Wenzel, W. *J. Comput. Chem.* **2011**, *32*, 2647−2653.
(17) Chakravarty, S.; Varadarajan, R. *Structure Fold Des.* **1999**, *7*, 723−732.
(18) Liu, S.; Zhang, C.; Liang, S.; Zhou, Y. *Proteins* **2007**, *68*, 636−645.
(19) Allison, J. R.; Boguslawski, K.; Fraternali, F.; van Gunsteren, W. F. *J. Phys. Chem. B* **2011**, *115*, 4547−4557.
(20) Mardia, K. V. Jupp, P. E. *Directional statistics*; John Wiley & Sons Inc: Chichester, U.K., 2000.
(21) Mezei, M. *J. Mol. Graphics Modell.* **2003**, *21*, 463−472.
(22) Zacharias, M. *Protein Sci.* **2003**, *12*, 1271−1282.
(23) Hubbard, S. J. Thornton, J. M. *NACCESS*; University College London: London, U.K., 1993.
(24) Wang, G.; Dunbrack, L. *Bioinformatics* **2003**, *19*, 1589−1591.
(25) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379−400.
(26) Lavery, R.; Pullman, A. *Biophys. Chem.* **1984**, *19*, 171−181.
(27) Samudrala, R.; Levitt, M. *Protein Sci.* **2000**, *9*, 1399−401.
(28) Holm, L.; Sander, C. *J. Mol. Biol.* **1992**, *225*, 93−105.
(29) Park, B.; Levitt, M. *J. Mol. Biol.* **1996**, *258*, 367−392.
(30) Samudrala, R.; Xia, Y.; Levitt, M.; Huang, H. S. *Pac. Symp. Biocomput. '99* **1999**, *4*, 505−516.
(31) Rico, M.; Jimenez, M. A.; Gonzalez, C.; De Filippis, V.; Fontana, A. *Biochemistry* **1994**, *33*, 14834−14847.
(32) Fogolari, F.; Tosatto, S.; Colombo, G. *BMC Bioinformatics* **2005**, *6*, 301−313.