THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Mapping the Hydropathy of Amino Acids Based on Their Local Solvation Structure

S. Bonella,*[,†,∥] D. Raimondo,[†,∥] E. Milanetti,[†,∥] A. Tramontano,[†,‡,§] and G. Ciccotti[†]

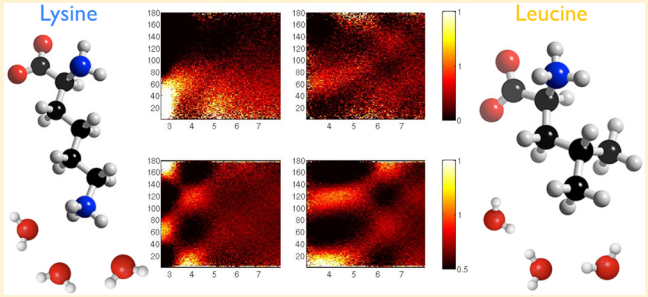
[†]Department of Physics, Sapienza University of Rome, Ple A. Moro 5, 00185 Rome, Italy

[‡]Institut Pasteur Fondazione Cenci-Bolognetti, 00185, Rome, Italy

[§]Center for Life Nano Science @Sapienza, Istituto Italiano di Tecnologia, Sapienza University of Rome, Ple A. Moro 5, 00185, Rome, Italy

Supporting Information

ABSTRACT: In spite of its relevant biological role, no general consensus exists on the quantitative characterization of amino acid's hydropathy. In particular, many hydrophobicity scales exist, often producing quite different rankings for the amino acids. To make progress toward a systematic classification, we analyze amino acids' hydropathy based on the orientation of water molecules at a given distance from them as computed from molecular dynamics simulations. In contrast with what is usually done, we argue that assigning a single number is not enough to characterize the properties of an amino acid, in particular when both hydrophobic and hydrophilic regions are present in a residue. Instead we show that appropriately defined conditional probability densities can be used to map the hydrophilic and hydrophobic groups on the amino acids with greater detail than possible with other available methods. Three indicators are then defined based on the features of these probabilities to quantify the specific hydrophobicity and hydrophilicity of each amino acid. The characterization that we propose can be used to understand some of the ambiguities in the ranking of amino acids in the current scales. The quantitative indicators can also be used in combination with standard bioinformatics tools to predict the location of transmembrane regions of proteins. The method is sensitive to the specific environment of the amino acids and can be applied to unnatural and modified amino acids, as well as to other small organic molecules.



1. INTRODUCTION

An accurate determination of the hydrophilicity/hydrophobicity of amino acids side chains in peptides and proteins is important in a number of relevant biological problems (such as protein folding, protein−protein, and peptide−protein interaction) and for the biophysical understanding of lipid−protein interaction within biological membranes. In the past few years, several computational methods have been suggested to study the structure and topology of transmembrane proteins, many of which rely, directly or indirectly, on the measure of the hydrophobicity of the amino acids. This has led to the development of a rather large number of hydrophobicity scales, either experimentally derived or knowledge based. Recent experimentally derived scales are based on approaches that include experiments to characterize the insertion of transmembrane protein by the Sec translocon system[1] (this is the translocon scale to which we shall refer in the following) and folding of small beta-barrel membrane protein.[2] The majority of the scales based on experimental methods, however, calculate (from measured concentrations) the transfer free energies of species from polar to apolar solvents and estimates the hydrophobicity from this free energy: the larger the gain in going to the non polar solvent, the more hydrophobic the solute.[3] Some common trends can be observed in experimental scales:[4] isoloeucine presents the highest level of hydrophobicity in most scales; glycine is often found at an intermediate level, which means a neutral hydrophobicity; the lowest hydrophobicity is usually assigned to aspartic acid, i.e. this is the most hydrophilic amino acid. In spite of these trends, however, there is a general lack of quantitative agreement between experimentally derived hydrophobicity scales with rankings showing, in some cases, quite marked variations from scale to scale. For example, in the 16 scales tabulated by Wilce et al.,[5] phenylalanine is ranked first (most hydrophobic) by four of the scales, yet one scale places it at the 16th position, close to being the least hydrophobic. Similar variations are reported for tryptophan. The situation is complex also when theoretical analyses are used to derive the scales. This is usually done via knowledge-based methods by selecting a test set of proteins and estimating the probability of finding a given amino acid in a region exposed to the solvent (in which case the residue is considered hydrophilic) or buried in the structure (hydrophobic). These observations can, however, be biased by the chosen test set and assume again that the local structure, the environment, and even the protonation state

Received: January 28, 2014
Revised: May 18, 2014
Published: May 20, 2014

dx.doi.org/10.1021/jp500980x | J. Phys. Chem. B 2014, 118, 6604−6613

do not play a role in determining the hydrophobicity of an amino acid. Finally, another serious limitation common both to experimental and theoretical scales is that, although they claim to identify the hydrophobicity of a single residue, they often use different molecules (polypeptides in experiments and amino acids in proteins, or protein substructures in knowledge based methods) to build the scale. Given the relevant changes in electronic structure and geometry that occur when an amino acid goes from being isolated to the polypeptidic or the proteic environment, the issue of how general and/or relevant are the hydrophobicity values assigned in each scale is still open. Indeed, the notion of context dependence has been demonstrated by several recent studies,[6−12] and different observables (ranging from local compressibility,[13] to water density fluctuations,[14] to free energy of formation of a cavity near the surface[15]) have been suggested to characterize it.

In spite of the fact that a univocal measure (and indeed definition) of hydrophobicity is still lacking, hydrophobicity scales are widely used by the biological community. In this paper, we propose a different approach to study and classify the hydropathy of isolated amino acids. Our purpose is not to derive a new scale *per se*, but to develop a tool to analyze and compare each residue's average solvation structure in water and infer from it the residue hydrophobic and/or hydrophilic features. This study is performed using molecular dynamics to compute two probability densities that describe the orientation of the water molecules at a given distance from the solute. We recently demonstrated that these probabilities can be used to qualitatively identify the correct hydrophobicity trend for a biologically relevant test set, the quaternary ammonium cations.[16] In this work, we show that they can also be effectively used to characterize the hydropathic properties of the amino acids with greater detail than other available techniques, in some cases providing insight on the ambiguous classification of residues in different existing scales. In order to compare our results with the vast available literature and to devise alternative, flexible, and cost-effective input parameters for bioinformatics tools, for example, to identify transmembrane segments of proteins, we also define a quantitative measure of the hydrophobicity/philicity of amino acids. Based on our analysis, we argue that for amino acids, this measure should not be expressed as a single number but rather as a set of three values. The first two gauge, respectively, the hydrophilic and hydrophobic features of the residue, and the third quantifies the relative weight of these features and therefore provides a "global" measure of hydropathy.

The characterization that we propose has several advantages. First, since it focuses on isolated amino acids, it is independent from environmental factors such as the presence of neighbors. Furthermore, since in our model of the amino acid we do not use capping (i.e., the chemical blockage of the charged terminals with appropriate groups), also this bias is absent. These factors affect both experimental scales (that use either capping or polypeptides in the measure) and knowledge-based rankings (that consider data sets of amino acids in proteins), and their influence might account for some of the discrepancies in the literature. Second, depending on the specific experimental or bioinformatics set up, available scales are often selectively sensitive to either the hydrophobic or the hydrophilic properties of a residue. The same side chain, however, often includes groups of different character in its structure. Our probability densities, being simultaneously sensitive to both aspects, provide a more complete description of a residue. Finally, the method can be used to study unnatural amino acids (often present in peptide mimics or in natural antibiotics) and amino acids in different protonation states, as well as other small solutes of biological interest. Perhaps even more interestingly, by an appropriate choice of the waters to be included in the analysis (see section 4.1) it can be used to map the biomolecules' hydropathy either globally or focusing on regions of particular interest.

The paper is organized as follows. In the Theory section we define our hydropathy indicators and summarize their properties. In Methods we provide details on the simulations performed to compute them and to obtain the quantitative parameters for hydrophobicity/hydrophilicity. We also illustrate the indicators used to compare our results to those of the translocon scale for transmembrane segment prediction. This latter analysis is considered to show that our method may provide a useful (and considerably easier to set up) alternative to translocon in cases for which the latter has not been calibrated, e.g., unnatural amino acids. This, in turn, may prove useful since the nontrivial experiments needed to create translocon can be easily applied only to natural amino acids and are difficult to extend beyond the specific environment of the experimental set up. We then present and discuss several results. First, we show how to isolate the features of the probability densities that are due to the residue (the side chain specific to each amino acid) from those due to the charged NH_3^+ and $COOH^-$ terminals (common to all amino acids at physiological pH). We then quantify the hydropathic features of the amino acids and show that, when only one of the indicators is considered, our ranking agrees well with experimental scales sensitive to the same characteristic. This means, for example, that a ranking based only on the "hydrophobic component" of the scale has a high correlation with scales based on experiments that are sensitive mainly to hydrophobic features. We then discuss the probability density profiles for a typical amino acid that is ranked in significantly different ways in different scales, and we show how different features in the probability density can account for this ambiguity. Finally, we use our hydrophobicity scale as a parameter of a commonly adopted bioinformatics tool to determine the location of transmembrane segments of proteins, and show that our ranking leads to results that are comparable to those obtained by employing translocon. In the Supporting Information we summarize the algorithm used to fit the calculated probability densities as a linear combination of Gaussians (this facilitates the quantitative analysis of the simulations; see Methods) and give the parameters for the reconstruction of the most relevant probability for all the amino acids. We also report the data for the transmembrane segment prediction, and define some of the (standard) quantities we adopt to assess our results. Finally, we discuss in more detail the comparison with other, commonly used, scales.

2. THEORY

The different orientations of the water molecules around a solute can be used to investigate its hydropathy. For this purpose, the H_2O molecule is represented as a tetrahedron, where an sp^3-hybridized oxygen atom lies at the center and two hydrogen atoms and two lone pair electrons point to the vertices. This geometry enables the water molecule to form up to four hydrogen bonds (HB). When water shares its hydrogens with other molecules two hydrogen bonds can be formed. The other two can be formed via the negative charge density of the lone pairs accepting a hydrogen atom from neighboring molecules.

dx.doi.org/10.1021/jp500980x | J. Phys. Chem. B 2014, 118, 6604−6613

If the neighbors are other water molecules, these interactions give rise to a hydrogen bond network. In the presence of a heterogeneous solute instead, two main situations arise.[17] When a nonpolar (hydrophobic) molecule is in aqueous solution, the Coulombic interaction among H_2Os is stronger than the interaction with the solute, which only involves van der Waals forces. As a result, water molecules orient themselves to maximize the number of HBs with neighboring water molecules and form a cage around the solute. Such a cage is called a "clathrate". When a charged or neutral but polar (hydrophilic) molecule is solvated, on the other hand, it interacts mainly via Coulomb forces with water. Depending on the charge and polarity of the solute, either one of the hydrogens or one of the lone pair electrons of the hydration water molecules is attracted toward it and the number of HBs among water molecules decreases. The orientation of water molecules in this case is called "inverted". In ref 16 we showed that these different orientations can be characterized via the probability density, $P(\theta_h|R)$, to find any one of the HB directions of a water molecule at an angle θ_h with the vector joining the solute and the H_2O, *conditional* on the H_2O being at a distance R from the solute (see definitions below and Figure 1 to identify these angles and vectors). Additional information on the charge of the solute can be obtained via the probability density $P(\theta_d|R)$ to find the vector $\vec{d}$, $\vec{d} = \vec{OH}_1 + \vec{OH}_2$ (see Figure 1), forming an angle θ_d with the vector joining the solute and the H_2O, *conditional* on the water being at a given distance R from the solute. A more detailed description of the features of these densities is given in the next subsection. While the idea to use information related both to the distance and the orientation of water around a solute to describe hydrophobicity is not new,[17−23] previous work used a *joint*, rather than conditional, probability density. As shown in ref 16, our choice makes it possible to strongly enhance the features of the probability for short distances, improving the resolution of the structure of the first and second solvation shell and providing a finer characterization of the properties of the solute. The distance and angles introduced are defined as follows. R is the modulus of the vector $\vec{R} = \vec{r}_i - \vec{r}_s$, where $\vec{r}_i = (x_i,y_i,z_i)$ is the position of the oxygen atom in water molecule i, and $\vec{r}_s = (x_s,y_s,z_s)$ is the position of atom "s" in the solute. When dealing with polyatomic, nonspherically symmetric molecules, the definition of a distance among the solute and the H_2Os is non trivial. A convenient choice is given by $D_i = \min_s|\vec{r}_i - \vec{r}_s|$. As discussed in ref 16 (see also refs 17 and 22) this distance associates each water molecule univocally to its closest solute's atom and provides a suitable description of the "surface" exposed to the solvent. For solutes made of polyatomic units containing hydrogen, we define $\vec{r}_s$ as the position of the center of mass of the polyatomic unit. θ_h, on the other hand, is constructed as follows. Consider the four HB vectors associated with each H_2O molecule. The first two are the O−H vectors (see Figure 1) that join the oxygen atom with the hydrogen atoms; the other two are the O−LP vectors where LP is the orientation of the lone pair orbitals. Having defined the random variable θ_h as the value of *any one* of the HB directions, each of the four

$$\theta_h^i = \arccos\left(\frac{\vec{R}\cdot\vec{V}_{HB_i}}{|\vec{R}||\vec{V}_{HB_i}|}\right) \quad (1)$$

(where $HB_i = \{O{-}H1, O{-}H2, O{-}LP1, O{-}LP2\}$) is an acceptable realization (and in the histogram that estimates the probability $P(\theta_h, R)$; see subsection 3.1, all θ_h^i contribute). Finally, we have

$$\theta_d = \arccos\left(\frac{\vec{R}\cdot\vec{d}}{|\vec{R}||\vec{d}|}\right) \quad (2)$$

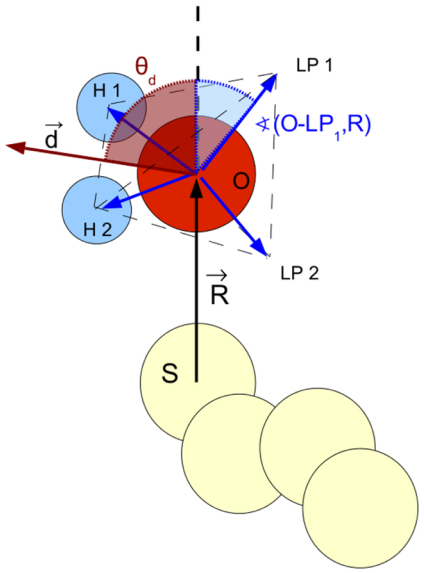


Figure 1. Schematic representation of the vectors and angles used in the conditional probabilities. The solute (S) is represented as a polymer of light yellow spheres, while a representative H_2O is drawn as the red oxygen (O) plus two blue hydrogen atoms (H1 and H2). LP1 and LP2 indicate the directions of the lone pairs. The water's tetrahedron is shown, together with the HB vectors (in blue). The vector $\vec{d}$ is shown in red, while $\vec{R}$ is in black. θ_d and one of the θ_h angles are drawn as the shaded light red and blue areas, respectively.

2.1. Characteristics of the Probability Densities. The hydropathy of a solute is characterized via the different peaks in the conditional probability densities introduced in the previous subsection. Let us consider two situations:

- *Hydrophilic cation/anion*: A water molecule in the vicinity of a positively (negatively) charged solute reorients to point toward the ion with one of its lone pairs (hydrogens). As illustrated in Figure 2 (first sketch from the left), for a cation the orientation of the HB vectors is such that two maxima are expected in $P(\theta_h|R)$: the first for $\theta_h^{\gamma 1} \approx 70°$ (angle formed with $\vec{R}$ by the vectors O−LP2, O−H1, and O−H2 in Figure 2), the second for $\theta_h^{\gamma 2} \approx 180°$ (angle formed with $\vec{R}$ by O−LP1). In the presence of an anion, on the other hand, the neighboring water molecules orient with one of the hydrogen atoms toward the solute. The orientation of the H_2O in this case is represented in the second image of Figure 2. Despite the rotation of the water molecule, the values of the θ_h angles do not change with respect to the previous case. So, also for a negatively charged solute, $P(\theta_h|R)$ will have maxima at $\theta_h^{\gamma 1,2}$.

  The second probability density, $P(\theta_d|R)$, discriminates positively and negatively charged solutes. For an anion, since the water orients with one of the Hs pointing toward the solute (see Figure 2), a single peak at $\theta_d^a \approx 130°$ will be observed. For a cation, it is one of the lone pairs that points to the solute, causing a single peak in the density at $\theta_d^c \approx 75°$.

dx.doi.org/10.1021/jp500980x | J. Phys. Chem. B 2014, 118, 6604−6613

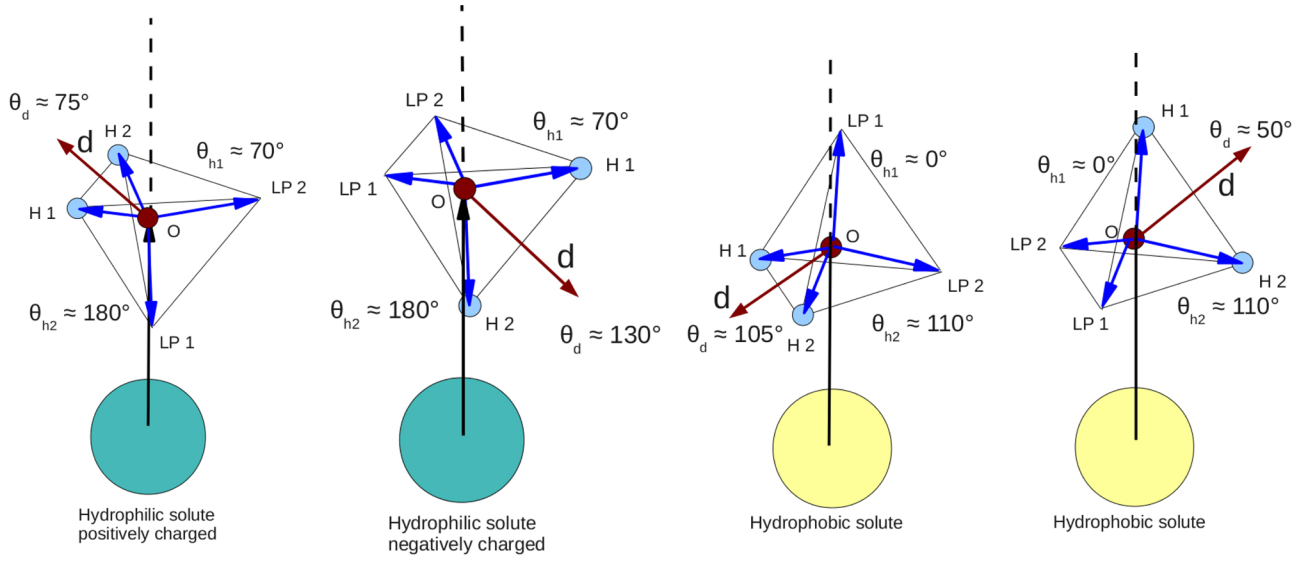


Figure 2. Schematic representation of a H_2O molecule near a hydrophilic or hydrophobic solute (light green and yellow spheres, respectively). The four vectors pointing to the H's or in the direction of the lone pairs are represented in blue, the $\vec{R}$ in black, and the expected values for θ_h are indicated. From left to right we show the orientation in the proximity of a positive and negative ion (first two images), then the two equivalent orientations near a nonpolar solute. We also show as the red arrow the orientation of the vector $\vec{d}$.

- *Hydrophobic solute*: In this case, the H_2O molecules position one of the faces of the tetrahedron toward the solute (Figure 2, two rightmost sketches), leaving all the HB directions available for bonding with other water molecules. As shown in the figure, there are two physically equivalent orientations of water leading to maxima of $P(\theta_h|R)$ at $\theta_h^{n1} \approx 0°$ and $\theta_h^{n2} \approx 110°$. The water orientation also produces two, physically equivalent, peaks in the $P(\theta_d|R)$ at angles $\theta_d \approx 50°$ or 105°.

The discussion above refers to the ideal case of an isolated water molecule interacting with the solute. In more realistic conditions, the structure of the peaks will be less defined due to factors such as, for example, the (many body) solute−solvent and solvent−solvent interactions, thermal motion, and the details of the potentials chosen to model the system. In spite of these factors, however, the main features described above are visible—isolated for solutes with univocal hydrophaticity, or in combination in the case of amphiphilc molecules—and will be used to classify the hydropathic properties of the amino acids.

3. METHODS

3.1. Simulation Set Up and Estimators of the Probabilities. The results presented in this work are based on a set of molecular dynamics simulations performed using Amber.[24] In each case, the simulated system consisted of N_w water molecules surrounding a single uncapped amino acid in a cubic box with periodic boundary conditions. The number of water molecules, which in turn determined the size of the simulation box, was fixed by imposing that, irrespective of its size, the amino acid was surrounded by a 25 Å thick layer of H_2Os. The water molecules were described via the TIP4P/2005 model,[25] and we used the "FF99SB" Amber force field as modified in ref 26 to model both intermolecular forces (Lennard-Jones and Coulomb) and intramolecular forces associated with bond stretching, bond angles, and dihedral angles. Partial charges were generated using the Austin Model 1[27] with bond charge correction (AM1-BCC).[28,29] All simulations were performed at T = 300 K, and the system was thermostated using Langevin dynamics with a time step of 1 fs. SHAKE[30] was used to constrain bonds involving hydrogens. Initial configurations for the amino acids were assigned using the most likely side-chain conformation from Dunbrack's 2010 backbone-independent rotamer library.[31] The solvated system was equilibrated and the relaxed geometry used as initial condition in a MD trajectory of 2 ns. Along the trajectory, we stored configurations every 0.2 ps. The stored configurations were subsequently used to estimate the joint probability densities

$$P(R, \theta_{h,d}) = \frac{1}{N_w}\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}} dt \sum_{j=1}^{N_w} \delta(R - D_j(t))\delta(\theta_{h,d} - \tilde{\theta}_{h,dj}(t)) \quad (3)$$

($\mathcal{T}$ is the length of the trajectory and $D_j(t)$ and $\tilde{\theta}_{h,dj}(t)$ values of the distance and angle at time t along it, respectively) by constructing two histograms, $h(R,\theta_{h,d})$. To that end, the positions and orientations of water molecules within a 20 Å radius from the amino acid were binned in a two-dimensional grid. The bin width in R was 0.05 Å, while the bins for the angles were 1° wide. From these histograms, we estimated the marginal probability $P_m(R)$ via numerical integration over the angle. The conditional probability was then obtained as

$$P(\theta_{h,d}|R) = \frac{P(R, \theta_{h,d})}{P_m(R)} \approx \frac{h(R, \theta_{h,d})}{h_m(R)} = h(\theta_{h,d}|R) \quad (4)$$

where $h_m(R)$ is the estimator of the marginal probability. The expression above will be used for qualitative analysis of the amino acids' properties. To facilitate a quantitative description, we obtained an estimate of the probability smoother than $h(\theta_h,R)$ [not $(h\theta_d,R)$ since this quantity turned out to be less relevant for our analysis] by fitting the probability densitity $P(R,\theta_h)$ with a Gaussian Mixture:[32,33]

$$P(R, \theta_h) \approx P^{GM}(R, \theta_h) = \sum_{\alpha=1}^{\nu} \pi_\alpha G_\alpha(\{R, \theta_h\}; \Sigma_\alpha, \mu_\alpha) \quad (5)$$

dx.doi.org/10.1021/jp500980x | J. Phys. Chem. B 2014, 118, 6604−6613

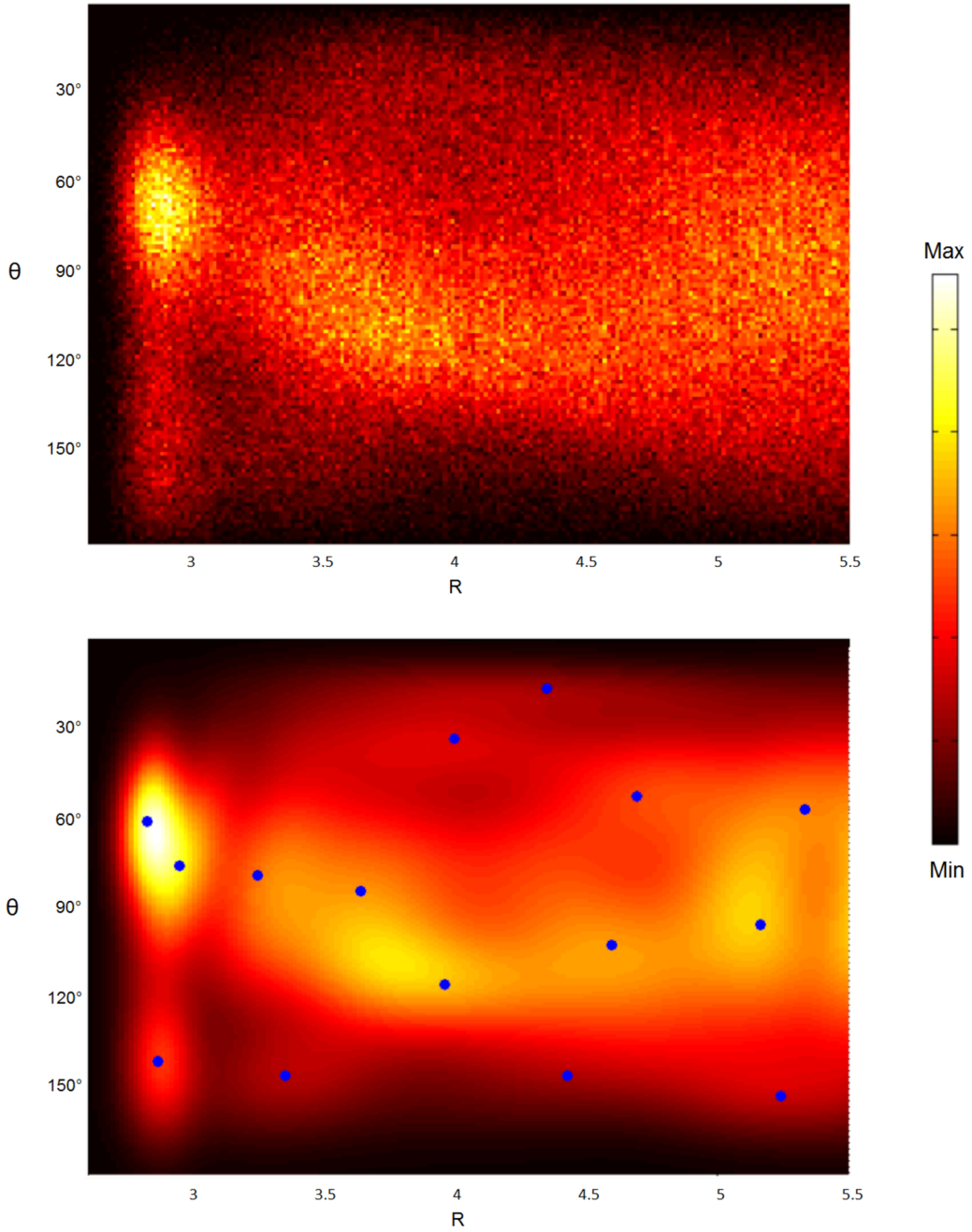


Figure 3. Histogram estimator of $P(\theta_h,R)$ (upper panel) and Gaussian mixtures reconstruction (lower panel) of the joint probability $P(\theta_h,R)$ for lysine. *y*-axis is θ, *x*-axis is *R*. The blue dots in the lower panel are the means of the Gaussians used in the reconstruction (not all centers are show, as some fall out of the range of the plot).

where $\{\pi_\alpha\}$, non negative numbers such that $\sum_{\alpha=1}^{\nu} \pi_\alpha = 1$, is the set of probabilities that (R,θ_h) are attributed to the Gaussian α, and $G_\alpha(\{R,\theta_h\};\Sigma_\alpha\mu_\alpha)$ is a bidimensional Gaussian of mean $\mu_\alpha = (\mu_\alpha^R\mu_\alpha^\theta)$ and covariance matrix Σ_α. The π_α and the parameters in the Gaussians were determined using the Expectation Maximization algorithm.[34,35] The algorithm is summarized in the Supporting Information. All the fits in this work converged with $\nu = 20$ Gaussians, with the exception of asparagine and leucine for which we used $\nu = 25$. A typical example of the agreement of a converged fit with the calculated histogram is shown in Figure 3, where we present the histogram estimating the joint probability (upper panel) and the Gaussian fit (lower panel) for leucine. The blue circles on the figure indicate the means of the Gaussians. The estimate of the conditional probability in eq 4 was then obtained from this fit by first integrating (analytically) eq 5 with respect to θ_h to get the marginal probability density in *R* and then computing the ratio of the joint over the marginal probability. This procedure smooths the signal with very limited loss of accuracy even when starting from an histogram obtained with relatively short trajectories. In particular, it allows one to determine more precisely the position and the intensity of different peaks in the conditional probability density which is the crucial information for characterizing the amino acid's hydropathy. To that end we use three numbers that quantify, respectively, the hydrophilicity, hyrdophobicity, and overall water solvation propensity of the residue. The first number, indicated as I_y, is given by the sum of the intensities of the two "hydrophilic" peaks in $h(\theta_h|R)$

dx.doi.org/10.1021/jp500980x | J. Phys. Chem. B 2014, 118, 6604−6613

(at $\theta_h^{y1,2} \approx 70, 180°$); the second, I_n, is the sum of the intensities of the "hydrophobic" peaks (at $\theta_h^{n1,2} \approx 0, 110°$); the third is given by the ratio $I = I_n/(I_n+I_y)$. These peaks are shown in Figure 4, where, as an example, we present the reconstructed

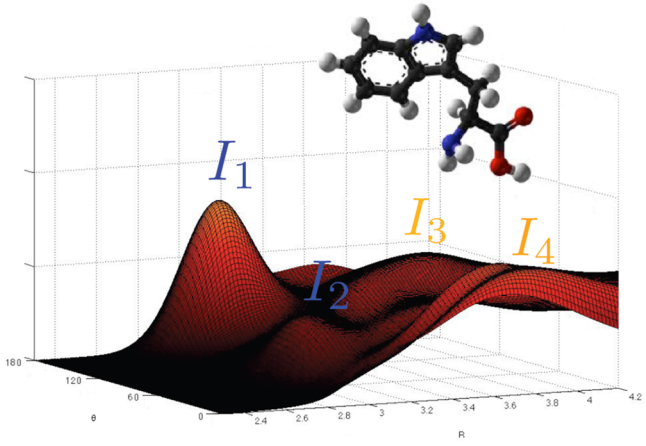


Figure 4. Three dimensional plot of the Gaussian Mixture fit for $P(\theta_h|R)$ for tryptophan. The hydrophilic peaks are indicated as the blue $I_{1,2}$, the hydrophobic ones by the yellow $I_{3,4}$. We also show a ball and stick representation of the amino acid, *C* atoms are in gray, *N* in blue, *O* in red, and *H* in white.

histogram for tryptophan. (The discussion of the features of the probability for this interesting case is postponed to section 4.2.) While more refined indicators (related, for example, to the areas underneath the peaks) could be devised, we found that this simple choice was sufficient and minimized the ambiguity in the calculations (using, for example, the areas as an alternative indicator, one should choose, somewhat arbitrarily, the area associated with each peak).

Since the specific potential model chosen for water influences the features of the probability densities obtained numerically compared to the ideal ones discussed in 2.1, we close this subsection by discussing the one adopted in this work. In the TIP4P model, the lone pairs are not explicitly included but appropriately located partial charges are used to reproduce the HB features and Coulomb interactions: two (equal) positive partial charges are located on the hydrogens, and a negative partial charge is placed along the H1−O−H2 bisectrix. This modifies somewhat the location of the peaks in the densities. While the effect is essentially irrelevant for most cases, the shape of $P(\theta_d|R)$ for positively charged solutes changes significantly: due to the tendency to form a direct bond between the positive charge of the ion and the negative partial charge placed in proximity of the oxygen atom, the peak at $\theta_d \approx 75°$ described in the previous section is shifted and spreads to form a wide region of high probability density located at $\theta_d \in [0,60]°$. However, tests performed in ref 16 and comparisons with ab initio[23] calculations indicate that the peak is still a significant indicator and that TIP4P is sufficiently reliable for our purposes. [The conditional probabilities can of course be constructed for any given potential, so, if a more reliable model becomes available it can be easily incorporated.] Furthermore, the absence of the lone pairs in the model prevents from computing directly the O−LP1 and O−LP2 vectors introduced below eq 1. To include them in our histogram, we used the positions of the H and O atoms (explicitly present in TIP4P) to rebuild, via a geometrical construction, the full tetrahedral structure of the water molecule and identify the directions of the lone pairs via the two vertices of the tetrahedron that do not point toward Hs. Also this procedure was validated in previous calculations.[16]

3.2. Prediction of Transmembrane Segments in Proteins. We also tested, in the spirit discussed in the Introduction, the ability of our method to provide adequate input parameters for bioinformatics tools. We focused, in particular, on methods to determine which segments (contiguous sets of amino acids) in the folded structure of a membrane protein actually span across the membrane. This is certainly the most common use of hydrophobicity values in the life science community. The test was performed using the Membrane Protein Explorer (MPEx) server (http://blanco.biomol.uci.edu/mpex/). This is a tool for determining the topology of membrane proteins based on sequence information combined with an analysis of the hydropathy of the amino acids as quantified by a given scale. MPEx determines the regions that are most likely to be in the (hydrophobic) membrane environment by averaging hydrophobicity values over a 19 reside window (which corresponds to a helical segment of length compatible with typical membrane crossing segments) and assigning the start and end position of (helical) transmembrane regions with the algorithm introduced by Jayasinghe et al.[36] We fed the hydrophobic component of our scale to the server and compared the results with the predictions based on the translocon hydrophobicity scale,[1] which is, as already mentioned, a state of the art scale available on the server and specifically derived for transmembrane helical proteins. The performance of the two scales was compared on a test data set of 125 transmembrane proteins for whom the three-dimensional structure has been determined by X-ray diffraction.[37] The accuracy of the prediction was quantified via two indicators: the two-class classification accuracy criterion (Q2)[38] and the Segment OVerlap value (SOV).[39] The Q2 score evaluates the performance of the prediction method as the percentage of residues predicted correctly to be part of a trans-membrane helix. The widely used per-segment accuracy SOV (defined in the Supporting Information) measures the overlap between the experimental and predicted trans-membrane regions and the relative displacement of their boundaries,[40] assessing the usefulness of the prediction in inferring the topology of the transmembrane protein, i.e., the number of times the protein traverses the membrane. Both Q2 and SOV were computed using codes developed in-house.

4. RESULTS AND DISCUSSION

This section is organized as follows. First we illustrate and solve a difficulty arising from our choice to use uncapped amino acids to build the scale. This choice is motivated by the intent to modify as little as possible the chemical and steric features of the molecules, but it has one drawback: the presence of the charged terminals strongly affects the structure of the conditional probabilities and masks the features due to the side chain. We begin by proposing a solution to this problem in section 4.1. We then compare our characterization of amino acid's hydropathy with alternative rankings. Our focus will be, in particular, on illustrating on a typical example how the detailed reconstruction of the solvation structure can be used to rationalize the origin of the different ranking attributed in the literature to some amino acids which, in turn, reflect the different behavior of the molecules in different environments. Finally, we assess the performance of our scale when it is used

dx.doi.org/10.1021/jp500980x | J. Phys. Chem. B 2014, 118, 6604−6613

as a parameter for the prediction of the location of transmembrane regions of proteins.

4.1. Isolating Relevant Signal in the Probability Densities. Let us consider the characteristics of lysine and leucine. These two amino acids have well-defined, and very different, hydropathy: lysine is known to be a positively charged hydrophilic amino acid, whereas leucine is hydrophobic, neutral, and nonpolar. However, the probability density plots for these two amino acids, calculated as discussed in the Methods section, and shown in Figure 5, appear similar.

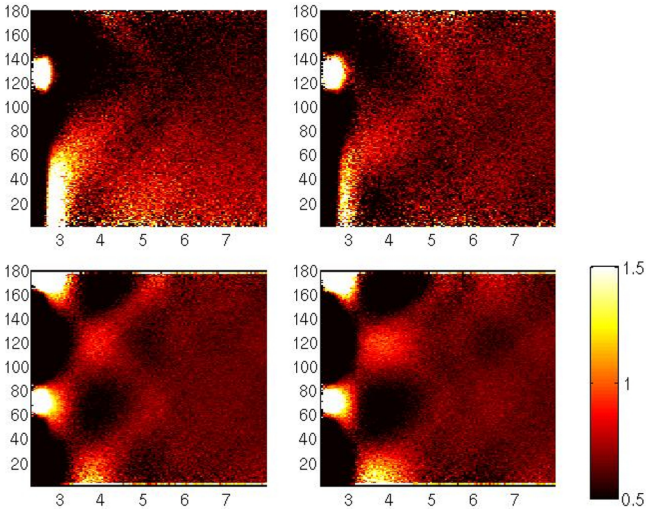


Figure 5. Estimated $P(\theta_d|R)$ (upper panel) and $P(\theta_h|R)$ (lower panel) for Lysine (Lys) (left panel) and Leucine (Leu) (right panel). The horizontal axis is the distance R in Å and vertical axes is the angles in degrees. The histograms were collected for all heavy atoms of the two amino acids.

$P(\theta_d|R)$ (upper panel, lysine on the left, leucine on the right) shows in both cases a narrow peak centered at $\theta_d \approx 130°$ and, at larger distances, a broad peak extending from $\theta_d = 0°$ to $\theta_d = 60°$. As discussed in the previous section, the first peak indicates the presence of a negative charge, the second is associated with a positive one. The $P(\theta_h|R)$s (lower panel of the figure, again with lysine on the left and leucine on the right) show two distinct peaks around $\theta_h \approx 70°$ and $\theta_h \approx 180°$ at small distances and two more peaks located at $\theta_h \approx 0°$ and $\theta_h \approx 120°$ for larger distances. The structure of these peaks is consistent, for both amino acids, with hydrophilic solutes. The origin of the problem can be easily understood and corrected. The signatures of the (dominating) charges appearing in the figure are in fact those associated with the amino acids' terminals, NH_3^+ and COO^-. In the histograms, the effect of these groups overwhelms that of the side chain, which is what differentiates the amino acids among themselves. The presence of the terminals also complicates experiments: indeed this is the reason why experimental scales cannot use isolated amino acids and have to modify them chemically. Our method, however, allows us to filter out the contribution of the terminals using the following observation. The histograms in the figure were calculated using eq 3, with N_w equal to the total number of water molecules in the system. We can, however, also compute the histograms excluding from the sum the water molecules that are nearest to the terminals based on our definition of the distance D_j (see previous subsection), i.e., the H_2Os that are more directly influenced by the terminals. The resulting conditional probability densities are shown in Figure 6 (same arrangement as in the

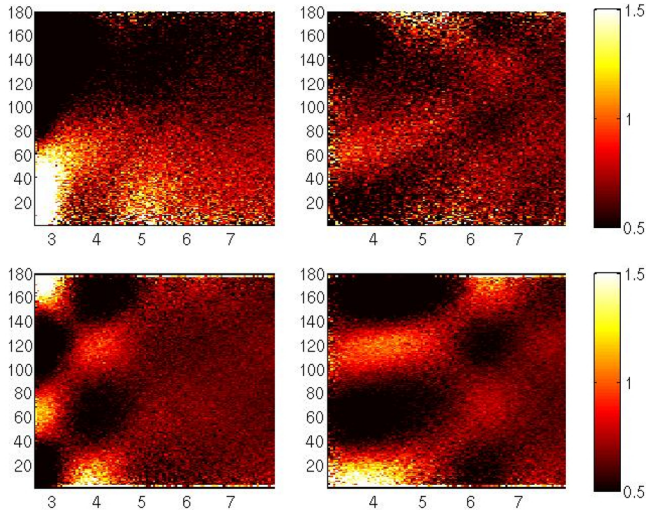


Figure 6. Plots of the estimated $P(\theta_d|R)$ (upper panel) and $P(\theta_h|R)$ (lower panel) for lysine (Lys) (left panel) and leucine (Leu) (right panel), same setting as in Figure 5, but the probability densities are computed excluding the water molecules "pertaining" to the charged terminals.

previous figure). Clear differences can be now be observed in the plots. Lysine produces a broad peak extending from $\theta_d \approx 0°$ to $\theta_d \approx 60°$ in $P(\theta_d|R)$ and two peaks in $P(\theta_h|R)$ around $\theta_h \approx 70°$ and $\theta_h \approx 180°$ for small distances followed by two relatively narrow peaks at $\theta_h \approx 0°$ and $\theta_h \approx 120°$ at larger distances. This identifies it as a positively charged hydrophilic amino acid, consistent with its usual classification. Leucine, on the other hand, originates a broad peak in $P(\theta_d|R)$ centered at $\theta_d \approx 60°$ and two peaks around $\theta_h \approx 0°$ and $\theta_h \approx 110°$ extending from $R = 3$ to $R = 5.5$ Å in $P(\theta_h|R)$. These are the features associated with a noncharged hydrophobic solute, and also in this case, the analysis is now in agreement with the common classification and with the observation that this amino acid is more often found in the hydrophobic core of folded proteins. Note that, since we did not modify the microscopic model of the system, the filtering method just described enhances and isolates the specific features of the residues without changing the overall chemical properties of the amino acids or transforming its geometry. In this sense, the signal still carries the full information on the effects of the environment on the properties of the region that we isolated.

The freedom to select which water molecules are included in the analysis is an interesting feature of the method. It opens the possibility to zoom on specific parts of the atomic structure of the solute by selecting, on the basis of a given question or simply sequentially, waters that are associated with one (or more) chemical group. This permits to isolate its contribution (and specific hydropathy) from the global hydropathy profile of the molecule.

4.2. Benchmarking of the Method. Reconstructing the histograms associated only to the side chains of the amino acids with the Gaussian Mixture technique described in the Methods section, we measured the intensities of the hydrophilic and hydrophobic peaks and obtained the values reported in Table 1. Our data was compared with the other scales appearing in the literature by computing the Pearson correlation coefficient among the values of one of our indicators (e.g., the one for hydrophobicity) and those of the reference scale (see Supporting Information for details). In the following, we will indicate our

dx.doi.org/10.1021/jp500980x | J. Phys. Chem. B 2014, 118, 6604−6613

Table 1. Values of Our Indicators of Hydropathy[a]

amino acid	I_n	I_y	I	I_n rank
Ala	1.33	0.51	0.72	13
Arg	0.97	1.22	0.44	3
Asn	0.92	2.95	0.24	1
Asp	1.13	5.98	0.16	8
Cys	1.26	1.02	0.55	12
Gln	1.02	2.37	0.30	7
Glu	1.10	3.15	0.26	6
Gly	1.35	0.55	0.73	14
His	0.96	0.62	0.61	2
Ile	1.46	0.51	0.74	19
Leu	1.47	0.54	0.73	20
Lys	0.99	1.71	0.37	4
Met	1.38	0.65	0.68	16
Phe	1.36	0.54	0.71	15
Pro	1.38	0.53	0.72	17
Ser	1.00	1.95	0.34	5
Thr	1.13	1.92	0.37	9
Trp	1.25	1.22	0.51	11
Tyr	1.20	0.66	0.64	10
Val	1.45	0.51	0.74	18

[a]The first column indicates the amino acid (listed in alphabetical order and identified via the corresponding three letter code); the second column reports the hydrophobic component (sum of the intensities of peaks 2 and 4 in the probability density); the third column the hydrophilic component (sum of the intensities of peaks 1 and 3); the fourth column is the ratio of the hydrophobic component versus the sum of the intensities of the four peaks; the fifth column reports the ranking from least (1, Asn) to most hydrophobic (20, Leu) based on I_n (in the case of equal I_n value (to the number of digits included), we assigned the more hydrophobic ranking to the residue with larger I).

set of indicators with the acronym WOPHS (Water Orientation Probability Hydropathy Scale). The average correlation of the hydrophobic index with the hydrophobicity scales available in the literature is 78% ± 5%. The correlation values of the hydrophilic WHOPS index with the three scales of hydrophilicity that we found is 73% (p-value 2×10^{-4}), 63% (p-value 3×10^{-3}) and 65% (p-value 2×10^{-3}) for the Grantham,[41] the Hoop,[42] and the Wood and Kuhn[43] scale, respectively. These results are of the same quality of those commonly adopted in the literature to validate hydropathicity scales. Interestingly, a clustering analysis (see Supporting Information) positions WOPHS among the experimental scales. This feature, unique among theoretical scales, indicates that, irrespective of the details and complexity of the experimental scale considered, our method is on average as effective as experiments (but considerably cheaper, faster, and more flexible). In the Supporting Informations, we provide a more detailed comparison with available scales (experimental and theoretical) and illustrate the reasons for closer agreement or larger discrepancies in some interesting cases.

The characterization of hydropathy that we propose is, however, richer than a simple scale. The features of our conditional probability densities can, in particular, be used to analyze the characteristics of different portions of the amino acids and even explain the ambiguous classification of some of them in the scales available in the literature. An interesting example is provided by tryptophan. This amino acid shows one of the rankings with the broader variation between different scales derived from experiment, exhibiting a solvent-dependent behavior that persists with or without the introduction of peptide bonds on either side of the α-carbon atom.[44] Indeed, in the 16 scales tabulated by Wilce et al.,[5] tryptophan is ranked first (most hydrophobic) by five of the scales, yet one scale ranks it as 16th, i.e., close to being the least hydrophobic. These differences originate from the amphiphilic properties of the side chain, which lead to different assessments of its hydrophobicity depending on the method and solute chosen. Considering, for example, two of the most used experimental scales, tryptophan is classified as the most hydrophobic amino acid in the H_2O/octanol scale, while it occupies the sixth position in the H_2O/cyclohexane scale where it ends up being less hydrophobic than phenylalanine, valine, leucine, isoleucine and methionine. In this case, the difference is due to the different nature of the solvents. Octanol is rather polar and can form H bonds with the π electrons and the hydrogens in the tryptophan's indole ring that lower the solvation's free energy and therefore increase the penalty for transfer to water. On the other hand, cyclohexane is a very simple hydrophobic substance (this solvent is completely apolar and proved an excellent choice as a model solvent for the protein interior because it interacts with the solute only via van der Waals forces[45]). Therefore, the local environment of the side chain is better defined in cyclohexane, and the factors discussed above are less likely to play a role. The amphiphilic characteristics of tryptophan are clearly visible in the reconstructed conditional probability $P(\theta_h|R)$, which is shown in Figure 4. The two peaks at short distances (the blue $I_{1,2}$) are the markers of hydrophilicity, while the (broader) features in the region around 3.5 Å are a hydrophobic signature (yellow $I_{3,4}$). Interestingly, the measure of the hydrophilic (sum of the intensity of I_1 and I_2) and hydrophobic contributions (sum of the intensity of I_3 and I_4) is very similar giving $I_y = 1.22$ and $I_n = 1.24$, indicating an essentially amphiphilic molecule with a very slight hydrophobic propensity, which is marginally also shown in the global indicator $I = 0.51$. The combined information from these numbers is fully consistent with the response obtained via different measurements. Furthermore, similarly to what was done to eliminate the signal from the charged terminals in the previous subsection, the groups responsible for the different parts of the signal can be identified by sequentially switching off the contribution of the waters associated, via our definition of distance, to different parts of the solute. Note that, even though in this case the reason for the ambiguous behavior of tryptophan could be predicted also simply from the knowledge of the chemical structure of the molecule, the analysis that we suggest does not rely on any a priori knowledge of the solute and could be repeated for more complex molecules in which the link among structure and hydropathy is less direct.

4.3. Applying the WOPHS Scale to the Prediction of Transmembrane Regions of Proteins. As mentioned in the Introduction, given the difficulties of experiments on these systems, hydrophobicity scales are instrumental in detecting the location of transmembrane regions of proteins. For this reason, more recently developed experimental scales are based on specific systems that mimic the membrane insertion process. Among the most effective ones, there is the translocon scale developed by Hessa et al.[1] Here the authors engaged the endoplasmic reticulum Sec61 translocon with an extensive set of designed polypeptide segments and used the results to derive a hydrophobicity scale explicitly aimed at identifying the features of transmembrane regions of proteins. We tested whether our scale could be effectively used to detect transmebrane regions as well. As mentioned in the Methods section, the accuracy of the predictions was assessed, for both scales, by measuring Q2

dx.doi.org/10.1021/jp500980x | J. Phys. Chem. B 2014, 118, 6604−6613

and SOV values on predictions obtained via the MPEx server on a relevant test set of membrane proteins the structure of which is known experimentally. A complete list of the proteins used and of the values of the indicators is available in the Table in the Supporting Informations. Predictions based on the translocon scale result in an average Q2 value equal to 77 (i.e., 77% of the predicted transmembrane residues match with experiments), and an average SOV of 93 (out of a maximum value of 100 for a perfect prediction). When our scale is used, the average Q2 is equal to 74, while the average SOV is 87. Not surprisingly, the data shows that the specifically derived translocon scale is on average more accurate. However, the difference in the performance is rather small—about 3% and 5% for Q2 and SOV, respectively—and probably not biologically significant when the uncertainties arising from the assignments of the precise location of the residues at the boundaries of the segments from the experimentally determined structures are kept into account. In fact, secondary structure assignments vary by 5−12% even between different crystals of the same protein[46] and it is well-known that C- and N-terminal ends of helices are difficult to define.[47] The similarity of the performance of WHOPS and translocon is somewhat remarkable given that, contrary to translocon, our scale was not built specifically for the purpose of studying transmembrane molecules. Furthermore, while translocon cannot be applied to moieties other than the naturally occurring amino acids (short of modifying the biological system in an experimentally complex way), our method can be used with no modification or additional computational cost to include modified amino acids and/or other small molecules of biological interest (e.g., antibiotics).

5. CONCLUSIONS

In this paper, we have shown that a detailed characterization of the hydropathic features of amino acids can be obtained by simulating the actual environment of the water distribution around the amino acid. This approach has the great advantage of permitting to selectively consider the contribution of its various moieties. For example, one can extract the contribution of the side chain filtering out the contribution of its charged termini. We tested the accuracy of our classification by comparing its values with those provided by the many available scales. Interestingly and importantly, although our results do not depend on complex and time-consuming experiments, they correlate very well with recently experimentally derived scales, thus opening the road to the accurate investigation of the effect of amino acid modifications (for example, post-translational modifications that are often mis-regulated in diseases) as well as to the measure of the hydrophobicity of unnatural amino acids, present for example in most antibiotics. Another interesting application of our method regards the possibility of predicting the bioavailability (i.e., absorption by the biological system) of different drugs. The classification is also effective as an ingredient of one of the most widely used application of hydrophobic scales, the prediction of the location of transmembrane segments in proteins. Also in this case, the possibility of manipulating the input parameters of the method (charge, chirality, etc.) will permit to take into account different cases, not necessarily amenable to experiments. As a final comment, it would be interesting to adapt the method to the case of larger solutes, such as proteins (and in particular their binding sites). However, this method cannot be applied when the solvation structure of water changes with respect to the one described in Section 2.1. [We thank an anonymous referee for pointing this out.] This happens, for example, near hydrophobic areas of low curvature and size around 1 nm. In this case, in fact, water reorients to optimize the overall H-bond network assuming an orientation (the so-called dangling bond structure[19,22,48,49]) very similar to the one in the vicinity of a small negatively charged hydrophilic molecule. In its current form, our scheme would not be able to discriminate between the two situations. Future work will focus on overcoming this limitation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information contains a summary of the Gaussian Mixture scheme employed to reconstruct $P(\theta_h,R)$ and parameters of the reconstruction for the 20 natural amino acids. A detailed comparison with existing hydrophobicity/hydrophilicity scales can also be found in the Supporting Information together with the data employed for transmembrane segments prediction. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sara.bonella@roma1.infn.it.

Author Contributions

∥These authors contributed equally to the work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are grateful to W. Babiaczyk and A. Gebreissilassie for running the MD simulations preliminary to the Gaussian Mixture reconstruction and analysis performed in this work and to G. Csanyi for pointing us in the direction of Gaussian Mixtures. Funding from the IIT SEED project No. 259 SIMBEDD and from award number KUK-I1-012-43 made by King Abdullah University of Science and Technology (KAUST), is acknowledged.

■ REFERENCES

(1) Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G. Recognition of Transmembrane Helices by the Endoplasmic Reticulum Translocon. *Nature* **2005**, *433*, 377−81.

(2) Moon, C. P.; Fleming, K. G. Side-Chain Hydrophobicity Scale Derived from Transmembrane Protein Folding into Lipid Bilayers. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10174−7.

(3) Chan, H. S.; Dill, K. A. Solvation: How to Obtain Microscopic Energies from Partitioning and Solvation Experiments. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 425−59.

(4) Lienqueo, M. E.; Salazar, O.; Henriquez, K.; Calado, C. R. C.; Fonseca, L. P.; Cabral, J. M. S. Prediction of Retention Time of Cutinases Tagged with Hydrophobic Peptides in Hydrophobic Interaction Chromatography. *J. Chromatogr., A* **2007**, *1154*, 460−3.

(5) Wilce, M. C. J.; Aguilar, M.-I.; Hearn, M. T. W. Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Angew. Chem.* **1995**, *67*, 1210−1219.

(6) Rotenberg, B.; Patel, A. J.; Chandler, D. Molecular Explanation for Why Talc Surfaces Can Be Both Hydrophilic and Hydrophobic. *J. Am. Chem. Soc.* **2011**, *133*, 20521−7.

(7) Patel, A. J.; Varilly, P.; Jamadagni, S. N.; Acharya, H. Extended Surfaces Modulate Hydrophobic Interactions of Neighboring Solutes. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 17678−17683.

(8) Daub, C. D.; Leung, K.; Luzar, A. Structure of Aqueous Solutions of Monosodium Glutamate. *J. Phys. Chem. B* **2009**, *113*, 7687−700.
(9) Limmer, D. T.; Willard, A. P.; Madden, P.; Chandler, D. Hydration of Metal Surfaces can be Dynamically Heterogeneous and Hydrophobic. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 4200−4205.
(10) Garde, S.; Patel, A. J. Unraveling the Hydrophobic Effect, One Molecule at a Time. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 16491−2.
(11) Mittal, J.; Hummer, G. Interfacial Thermodynamics of Confined Water near Molecularly Rough Surfaces. *Faraday Discuss.* **2010**, *146*, 341−352.
(12) Johnson, M. E.; Hummer, G. The Topology of Interface Interaction Networks Reflects on Protein−Protein Interaction Specificity and Regulation. *Biophys. J.* **2013**, *104*, 160A.
(13) Sarupria, S.; Garde, S. Quantifying Water Density Fluctuations and Compressibility of Hydration Shells of Hydrophobic Solutes and Proteins. *Phys. Rev. Lett.* **2009**, *103*, 037803.
(14) Patel, A. J.; Varilly, P.; Chandler, D. Fluctuations of Water near Extended Hydrophobic and Hydrophilic Surfaces. *J. Phys. Chem. B* **2010**, *114*, 1632−7.
(15) Patel, A. J.; Garde, S. Efficient Method to Characterize the Context-Dependent Hydrophobicity of Proteins. *J. Phys. Chem. B* **2014**, *118*, 1564−73.
(16) Babiaczyk, W. I.; Bonella, S.; Guidoni, L.; Ciccotti, G. Hydration Structure of the Quaternary Ammonium Xations. *J. Phys. Chem. B* **2010**, *114*, 15018−28.
(17) Cheng, Y. K.; Sheu, W. S.; Rossky, P. J. Hydrophobic Hydration of Amphipathic Peptides. *Biophys. J.* **1999**, *76*, 1734−1743.
(18) Raschke, T. M.; Levitt, M. Nonpolar Solutes Enhance Water Structure within Hydration Shells while Reducing Interactions between Them. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6777−6782.
(19) Lee, C. Y.; McCammon, J. A.; Rossky, P. J. The Structure of Liquid Water at an Extended Hydrophobic Surface. *J. Chem. Phys.* **1984**, *80*, 4448.
(20) Lee, S. H.; Rossky, P. J. A Comparison of the Structure and Dynamics of Liquid Water at Hydrophobic and Hydrophilic Surfaces: A Molecular Dynamics Simulation Study. *J. Chem. Phys.* **1994**, *100*, 3334.
(21) Zichi, D. A.; Rossky, P. J. Solvent Molecular Dynamics in Regions of Hydrophobic Hydration. *J. Chem. Phys.* **1986**, *84*, 2814.
(22) Cheng, Y. K.; Rossky, P. J. Surface Topography Dependence of Biomolecular Hydrophobic Hydration. *Nature* **1998**, *392*, 696−699.
(23) Grossman, J. C.; Schwegler, E.; Galli, G. Quantum and Classical Molecular Dynamics Simulations of Hydrophobic Hydration Structure around Small Solutes. *J. Phys. Chem. B* **2004**, *108*, 15865−15872.
(24) Salomon-Ferrer, R.; Case, D.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198−210.
(25) Abascal, J. L. F.; Vega, C. A General Purpose Model for the Condensed Phases of Water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.
(26) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950−1958.
(27) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(28) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.
(29) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.
(30) Ryckaert, J.; Ciccotti, G.; Herman, J.; Berendsen, J. 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.
(31) Dunbrack, R. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431−440.
(32) McLachlan, G.; Peel, D. *Finite Mixture Models*; Wiley: New York, 2000.
(33) Marin, J.; Mengersen, K.; Robert, C. In *Essential Bayesian Models. Handbook of Statistics: Bayesian Thinking - Modeling and Computation*; Dey, D., Rao, C., Eds.; Elsevier: Amsterdam, 2011.
(34) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1−38.
(35) Wu, J. C. F. On the Convergence Properties of the EM Algorithm. *Ann. Stat.* **1983**, *11*, 95−103.
(36) Jayasinghe, S.; Hristova, K.; White, S. H. Energetics, Stability, and Prediction of Transmembrane Helices. *J. Mol. Bio.* **2001**, *312*, 927−34.
(37) Jayasinghe, S.; Hristova, K.; White, S. H. MPtopo: A Database of Membrane Protein Topology. *Protein Sci.* **2001**, *10*, 455−8.
(38) Rost, B.; Sander, C. Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol.* **1993**, *232*, 584−599.
(39) Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. A Modified Definition of Sov, a Segment-based Measure for Protein Secondary Structure Prediction Assessment. *Proteins* **1999**, *34*, 220−3.
(40) Lee, J. Measures for the Assessment of Fuzzy Predictions of Protein Secondary Structure. *Proteins: Struct., Funct., Bioinf.* **2006**, *462*, 453−462.
(41) Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **1974**, *185*, 862−4.
(42) Hopp, T. P.; Woods, K. R. Prediction of Protein Antigenic Determinants from Amino Acid Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78*, 3824−8.
(43) Kuhn, L. A.; Swanson, C. A.; Pique, M. E.; Tainer, J. A.; Getzoff, E. D. Atomic and Residue Hydrophilicity in the Context of Folded Protein Structures. *Proteins* **1995**, *23*, 536−47.
(44) Wolfenden, R. Experimental Measures of Amino Acid Hydrophobicity and the Topology of Transmembrane and Globular Proteins. *J. Gen. Physiol.* **2007**, *129*, 357−62.
(45) Sharp, K. A.; Nicholls, A.; Friedman, R.; Honig, B. Extracting Hydrophobic Free Energies from Experimental Data: Relationship to Protein Folding and Theoretical Models. *Biochemistry* **1991**, *30*, 9686−97.
(46) Rost, B. In *Protein Structure Determination, Analysis, and Modeling for Drug Discovery*; Chasman, D., Ed.; Dekker: New York, 2003; pp 207−249.
(47) Cubellis, M. V.; Cailliez, F.; Lovell, S. C. Secondary Structure Assignment that Accurately Reflects Physical and Evolutionary Characteristics. *BMC Bioinf.* **2005**, *6* (Suppl 4), S8.
(48) Huang, X.; Margulis, C. J.; Berne, B. J. Do Molecules as Small as Neopentane Induce a Hydrophobic Response Similar to That of Large Hydrophobic Surfaces? *J. Phys. Chem. B* **2003**, *107*, 11742−11748.
(49) Perera, P. N.; Fega, K. R.; Lawrence, C.; Sundstrom, E. J.; Tomlinson-Phillips, J.; Ben-Amotz, D. Observation of Water Dangling OH Bonds around Dissolved Nonpolar Groups. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12230−12234.