PCCP

PAPER

Cite this: *Phys. Chem. Chem. Phys.*, 2017, **19**, 17094

WatAA: Atlas of Protein Hydration. Exploring synergies between data mining and *ab initio* calculations†

Jiří Černý, Bohdan Schneider and Lada Biedermannová *

Received 10th January 2017,
Accepted 9th June 2017

DOI: 10.1039/c7cp00187h

rsc.li/pccp

Water molecules represent an integral part of proteins and a key determinant of protein structure, dynamics and function. WatAA is a newly developed, web-based atlas of amino-acid hydration in proteins. The atlas provides information about the ordered first hydration shell of the most populated amino-acid conformers in proteins. The data presented in the atlas are drawn from two sources: experimental data and *ab initio* quantum-mechanics calculations. The experimental part is based on a data-mining study of a large set of high-resolution protein crystal structures. The crystal-derived data include 3D maps of water distribution around amino-acids and probability of occurrence of each of the identified hydration sites. The quantum mechanics calculations validate and extend this primary description by optimizing the water position for each hydration site, by providing hydrogen atom positions and by quantifying the interaction energy that stabilizes the water molecule at the particular hydration site position. The calculations show that the majority of experimentally derived hydration sites are positioned near local energy minima for water, and the calculated interaction energies help to assess the preference of water for the individual hydration sites. We propose that the atlas can be used to validate water placement in electron density maps in crystallographic refinement, to locate water molecules mediating protein–ligand interactions in drug design, and to prepare and evaluate molecular dynamics simulations. WatAA: Atlas of Protein Hydration is freely available without login at www.dnatco.org/WatAA.

Introduction

Water constitutes a large proportion of the cells and tissues of living organisms and it is nowadays accepted that a detailed description of how proteins and nucleic acids interact with water represents an indispensable component of our effort to understand the molecular basis of life.[1–3] In practical terms, hydration properties are often used to parametrize biomolecular force fields and to assess their performance.[4] In molecular modeling and simulation, accounting for water-mediated interactions improves protein folding[5] and structure predictions[6] as well as protein–protein docking.[7,8] In recent years, the subject of biomolecular hydration resulted in numerous review articles,[9–13] books[14] and special journal issues.[15–17]

From the experimental methods available to study molecular details of hydration, crystallography stands out as it provides unique information about the atomic details of the structure of the biomolecule and its ordered hydration layer. Averaging over many crystallographic structures provides clear patterns of preferred hydration sites for individual biomolecular building blocks.[18] The rapidly growing number of biomolecular structures solved with high crystallographic resolution enabled to substantially increase the data-sets used in data mining studies of biomolecular hydration – from tens or few hundreds of structures in the pioneering studies conducted in 1980's[19] and in 1990's[20] to the thousands of structures analyzed in the more recent studies,[21–25] thus enabling for a much more detailed information to be derived.

As others[26] and we[27] have shown, amino-acid (AA) hydration is dependent not only on the type of the AA residue, but also on its conformation, so that for a given residue, the positions and occupancies of its hydration sites can differ greatly between its various conformers. We have observed that the side-chain main-chain group additivity model cannot be applied in case of small to medium sized polar AA (Ser, Thr, Asn) due to frequent water-bridged interactions between the polar backbone and polar functional groups of the side chain. Moreover, we observed numerous potentially "non-canonical" contacts between water molecules and protein atoms, such as carbon-donor[28] hydrogen bonds, OH–π interactions and off-plane interactions with

Laboratory of Biomolecular Recognition, Institute of Biotechnology CAS, BIOCEV, Prumyslova 595, Vestec 252 50, Prague-West, Czech Republic.
E-mail: Lada.Biedermannova@ibt.cas.cz; Tel: +420 325-87-3737

† Electronic supplementary information (ESI) available: File F1, Table S1 and Fig. S1–S6 show further details of the results. See DOI: 10.1039/c7cp00187h

aromatic hetero-atoms.[29] However, the crystallographic data did not allow to establish the exact orientation of water molecules including hydrogen atoms and thus the exact nature of these interactions.

Another aspect that was missing in our previous analysis is the information on the energetics, or the interaction strength of the interactions. These could, in principle, be estimated in a knowledge-based approach from the propensities or ''pseudo-occupancies'', as derived from the analysis of crystal structures. However, the spatial arrangement of any two atoms or molecular fragments in the crystal structure reflects not only their pairwise interaction, but it is also affected by the neighboring residues, ligands *etc.*, as well as by long-range interactions within the whole system. Therefore, the hydration site (HS) pseudo-occupancy can at best be only an indirect and inaccurate measure of the interaction strength. Both the structural and energetic aspects missing from the previous study can be most accurately addressed through quantum mechanics (QM) calculations. QM calculations can provide important insights into the actual strength of the interaction and a more accurate estimate of the preferential placement of the molecular fragments compared to the geometries and statistics based on crystal data.[30]

Given the importance of protein hydration and the wealth of information available in crystallographic databases, it is surprising that there are no online tools dedicated to the topic, while websites covering related areas, such as the SWS website on DNA hydration[21] and the Atlas of protein side-chain interactions[31] exist. Thus, we decided to corroborate our previous crystal-derived data on AA hydration by computational analysis at the *ab initio* QM level by optimizing the position of all hydration sites representing mostly water molecules and calculate their interaction energies. The usage of the DFT-D/RI-TPSS/TZVP method seems especially suited to this task, since it has been shown to provide interaction energies (E_{int}) of biomolecular fragments comparable to the much more computationally intensive benchmark CCSD(T)/CBS method at a significantly reduced cost.[32–34] In this article, we describe the WatAA website, the newly developed atlas of hydration of AA residues in proteins. It provides an easy access to a large amount of information obtained from data mining of protein crystal structures and from QM calculations. We also describe interesting examples of water–AA interactions.

Methods

WatAA server hardware and software

The WatAA server is hosted as a Linux-based virtual machine within the environment provided by the ELIXIR CZ infrastructure. This ensures high availability and professional maintenance as well as easy scaling of the resources if necessary. The presented version of the server ran about a year internally and had been tested over a year as a publicly available service accessible at the www.dnatco.org/WatAA address. The software part employs Apache web server and CSS and JavaScript for the client-side scripting. The interactive visualization of 3D molecular structures, hydration site positions and water probability distributions was implemented using JSmol applet,[35] a JavaScript-based molecular viewer running within a browser. This solution allows for the web service to run across various devices and platforms, including mobile devices, without the need to install additional plug-ins or other software. The JSmol performance is known to depend on the browser version and the computer operation system used. The complete web service was successfully tested with the major web browsers under Linux, OS X and Windows, with the Firefox browser having currently the best performance regarding the JSmol part.

Crystal-derived data

The statistical and structural data that are presented in the atlas were obtained in our previous work.[27] The data mining analysis of a non-redundant set of 2818 high-resolution protein crystal structures from the Protein Data Bank yielded statistical data on hydration of AA as function of their conformational categories. For each of the 20 standard AAs, all the AA residues together with water molecules within 3.2 Å were extracted from the crystal structures and their conformations were classified according to secondary structure and χ_1 rotameric state (torsion angle of the CA–CB bond). In each resulting category (defined by residue type, secondary structure and rotameric state of χ_1 torsion angle), conformational clustering was applied and the largest cluster (Conformer1) was selected for further analysis. In each of these clusters, Fourier averaging[20] of water densities was performed and the hydration sites were obtained as maxima in the resulting pseudo-electron density. The probability (pseudo-occupancy) of a given hydration site was estimated from the water density value at the site position, and considered significant if greater than 0.1, *i.e.*, if a crystallographically ordered water molecule is present at the hydration site position in $>10\%$ of occurrences of the given conformer. The calculations presented in his study and the WatAA website include also additional seven hydration sites of Gly and Ala residues with occupancies <0.1, which were included as a model for the hydration of protein main chain.

In silico data

A computational analysis was performed separately for each crystal-derived hydration site, *i.e.*, each analyzed system consisted of only one water molecule placed at the position of the hydration site and one AA residue. The AA residue was capped with the standard N-terminal acetyl and C-terminal *N*-methyl amide capping groups using LEaP program of AmberTools14,[36] and hydrogen atoms were added to the system using the Open Babel program, version 2.3.1.[37] Acidic AAs (Asp and Glu) were modeled negatively charged (deprotonated), and basic AAs (Arg and Lys) were modeled positively charged (protonated). Histidine was modeled neutral, with protonation on the NE2 atom. Geometry optimization was performed in two stages. In the first one, positions of all added atoms (*i.e.*, hydrogens on both the water molecule and the AA, as well as all atoms of the capping groups) were optimized, while the rest of the system was fixed. In the second stage of energy minimization, also the

position of the water oxygen was allowed to move with respect to the AA. After the each geometry optimization stage the interaction energy between the water molecule and the capped AA was computed. All calculations were performed in program TurboMole v6.4[38,39] at the level of DFT-D/RI-TPSS/TZVP, *i.e.*, DFT augmented with an empirical dispersion term;[33] the solvation effects were described using the conductor-like screening model[40,41] (COSMO) with the dielectric constant for water (ε_r = 78.4). The DFT-D/RI-TPSS/TZVP method was shown to provide very accurate results for biomolecular complexes, comparable to CCSD(T), while being computationally less demanding.[32,33] Statistical analysis of the data was performed using RStudio Version 0.99.903,[42] with ggplot2 package[43] used to generate the histograms and scatter plots.

Results and discussion

Website

The WatAA homepage at www.dnatco.org/WatAA introduces the purpose of the atlas and provides links to all other sections. In all sections, the web-page interface is divided into three columns, the Navigation menu (left), the Text and Table panel (middle), and the interactive 3D structure Visualization panel containing the JSmol applet[35] (right).

The data on the hydrated AA conformers obtained from both the data-mining study and from *ab initio* QM calculations can be accessed in the Atlas section (Fig. 1). The middle panel in this section contains three elements. The first one is the Selection box where the user chooses one of the twenty standard AAs and its conformation, *i.e.*, its secondary structure (either H: α-helix or E: extended β-sheet) and χ_1 rotameric state (torsion angle of the CA–CB bond, either *gauche*(+), *gauche*(−), or *trans*, or in the case of Ala and Gly the only rotamer, denoted NA). The last menu in the Selection box allows to choose a particular hydration site of the selected AA conformer.

The second element is the Display menu that controls the visualization and allows to display or hide any combination of the following three types of data: (1) the crystal-derived atomic positions of the AA conformer with hydration site positions, (2) the crystal-derived 3D map of water distribution around the particular conformer at a chosen pseudo-occupancy level (default 10%) and (3) QM-optimized position of a water molecule at the selected hydration site, together with optimized atomic positions of hydrogens and capping groups on the selected AA conformer. A slider allows to adjust the width of the middle panel relative to the Visualization panel on the right.

The last element are the Result tables, which present the data for the largest conformational cluster (Conformer1) of an AA residue in the selected conformational category. The first table lists how many water molecules and AA residues were analyzed in the data-mining study for the selected AA conformer. The second table summarizes the data obtained from both data mining and from QM calculations for each hydration site on a separate row; the row for the selected hydration site is highlighted in light blue. For each of the HS the results table lists its pseudo-occupancy, data obtained for the crystal-derived and for optimized water position (nearest heavy atom, distance to

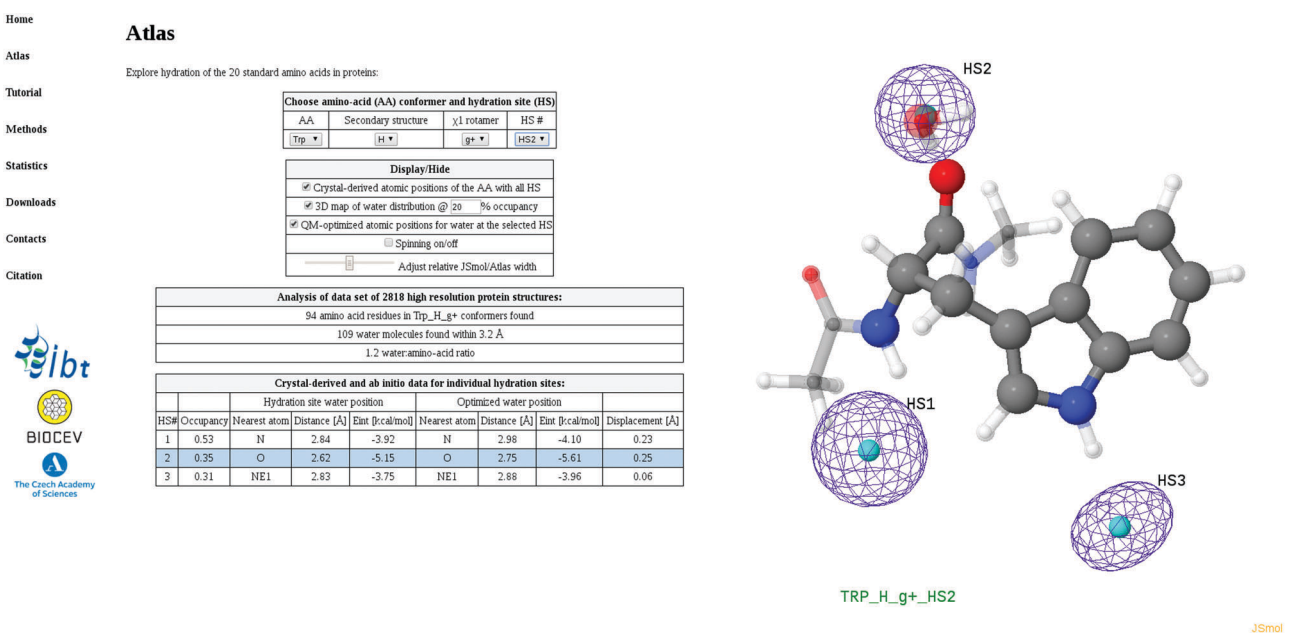

Fig. 1 Screenshot of the Atlas section of the WatAA website. Left column contains the Navigation menu; middle column contains the Selection box, Display menu and Result tables; right column contains the interactive 3D structure Visualization panel. Hydration of the largest conformational cluster in Trp_H_g+ category is visualized. The crystal-derived atomic positions of the selected AA conformer are shown in a non-transparent ball-and-stick representation with hydration site positions shown as cyan spheres; the crystal-derived 3D map of water distribution is contoured at a selected occupancy level of 20% using a blue mesh; the QM-optimized atomic positions of a water molecule in the selected hydration site (HS2), together with the amino-acid conformer (including QM-optimized positions of hydrogens and capping groups) is shown in transparent ball-and-stick representation.

nearest heavy atom, interaction energy) and the displacement (Δx) between the two positions.

The Tutorial section describes the user interface of the atlas, introduces the typical work-flow, explains which selections can be made using the selection boxes as well as what display options are available and provides description of the data in the Results tables as well as the Visualization panel. The Methods section describes the procedures applied to obtain the presented data. A detailed description of the data mining of protein crystal structures and the QM calculations is provided. Most of the data presented in the atlas is available for download and offline usage in the Downloads section. Additional data can be obtained from the authors by email or post as specified in the Contacts section. Information on how to cite the work is given in the Citation section.

Structures of hydrated AA residues.

The crystal-derived data presented in the Atlas were obtained in our previous data mining study.[27] This analysis yielded a total of 323 hydration sites for the 20 standard AAs, in their most populated conformers in each *gauche*(+), *gauche*(−) and *trans* side-chain rotamers in both alpha-helical and extended beta-sheet conformation of the main chain. The hydrated AA conformers obtained in the previous study represent the starting point for the subsequent QM calculations described hereafter. First, we investigated how closely the crystal-derived hydration sites correspond to energy minima at the potential energy surface of the AA fragments. For this purpose, we analyzed separately each hydration site of each AA conformer, performing two stages of energy minimization. In the first minimization, only the positions of atoms which were added to the crystal-derived structures (*i.e.*, hydrogens and capping groups) were allowed to move, while the water molecule's oxygen atom was kept in the crystal-derived position. This set of structures and their corresponding parameters were denoted with "(cryst)" in this article. In the second stage of energy minimization, the water oxygen was allowed to move (in addition to the hydrogens and capping groups), and its optimized position with respect to the AA residue was thus found. The corresponding geometries and parameters were denoted with "(opt)". All data for the 323 hydration sites are summarized in File F1, ESI.†

Water displacement upon optimization

The spatial distance between the crystal-derived position of the water molecule and its optimized position, termed the displacement, or Δx, was small in majority of the systems. For over 90% hydration sites, Δx was below 0.7 Å (Fig. 2a), showing that indeed the majority of the hydration sites positions is close to the local energy minimum. Out of the 323 analyzed hydration sites only 22 of them had Δx greater than 1.0 Å. These were mostly the polar (6× Arg, 5× Asp, 2× Glu) or moderately polar residues (3× Ser, 2× Thr, 2× Trp, 1× Gln), which in a protein molecule would either be hydrated by more than one water, or its hydrogen-bonding capacities would be saturated by other polar atoms. Thus, the occurrences of large Δx of water molecules can be explained by the absence of the broader polar environment in these simulated systems. Besides that, there were 11 hydration sites where the water's positional displacement upon optimization lead to a change in its nearest heavy atom (Table S1, ESI†), mostly from backbone carbonyl oxygen to the nearby amide nitrogen in extended beta-sheet conformations. These 11 systems, as well as the seven hydration sites of Gly and Ala main chain with occupancies below 0.1, were removed from further statistical analysis of the obtained data.

Water – AA interaction distance

For the 305 hydrated AA conformers with occupancy >0.1 and no change in nearest heavy atom upon optimization, we analyzed how the interaction distance, d, between water oxygen and its nearest heavy atom changed upon the geometry optimization. In the set of crystal-derived structures, the distribution of interaction distances was very broad, with d(cryst) values extending from 2.29 to 3.21 Å. In the optimized set of structures, the range of values d(opt) is diminished to 2.64–3.02 Å, which are values typical of a protein–water H-bond. We divided the hydration sites into groups based on the AA and nearest heavy atom type (Table 1).

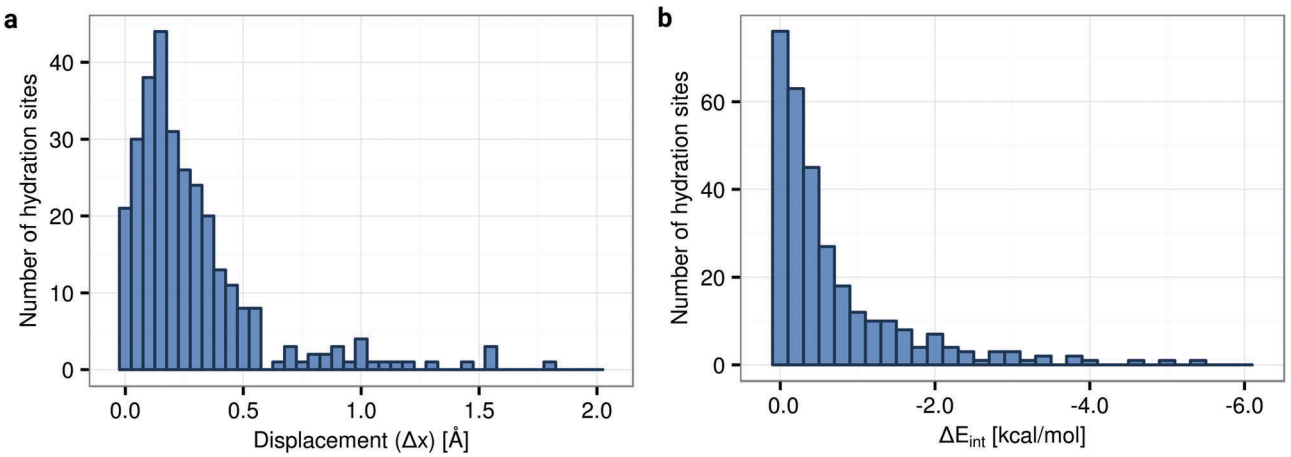

Fig. 2 Change in hydration site properties upon geometry optimization. (a) Position displacement Δx, and (b) interaction energy change ΔE_{int}.

Table 1 Parameters of the water–AA interaction within groups of hydration sites based on their nearest heavy atom type

Nearest heavy atom	Count	d(cryst)[a] Mean	sd	d(opt)[b] Mean	sd	Occupancy[c] Mean	sd	E_{int}(opt)[d] Mean	sd
bb_N[e]	48	2.87	0.11	2.94	0.03	0.19	0.10	−3.9	0.4
Arg_NE	6	2.74	0.12	2.85	0.02	0.21	0.05	−5.7	1.0
Arg_NH1	10	2.97	0.17	2.86	0.05	0.27	0.11	−5.6	0.9
Arg_NH2	5	2.78	0.15	2.84	0.02	0.20	0.10	−4.6	0.5
Asn_ND2	12	2.86	0.08	2.96	0.02	0.23	0.11	−3.1	0.3
Gln_NE2	11	2.85	0.11	2.95	0.01	0.16	0.04	−3.0	0.2
His_ND1	6	2.73	0.16	2.76	0.03	0.34	0.12	−8.2	1.3
His_NE2	6	2.77	0.07	2.87	0.02	0.29	0.05	−3.9	0.1
Lys_NZ	18	2.75	0.14	2.78	0.01	0.17	0.03	−6.0	0.1
Trp_NE1	6	2.85	0.13	2.90	0.03	0.30	0.04	−3.7	0.3
bb_O[f]	68	2.74	0.14	2.78	0.02	0.22	0.11	−4.7	0.8
Asn_OD1	9	2.64	0.07	2.80	0.02	0.15	0.03	−3.9	0.6
Asp_OD1	12	2.61	0.14	2.69	0.02	0.16	0.04	−7.0	1.2
Asp_OD2	11	2.58	0.11	2.67	0.01	0.22	0.05	−7.6	0.5
Gln_OE1	13	2.65	0.08	2.77	0.01	0.18	0.04	−4.6	0.3
Glu_OE1	13	2.68	0.20	2.69	0.02	0.17	0.06	−6.6	0.3
Glu_OE2	12	2.58	0.16	2.66	0.01	0.17	0.05	−8.0	0.7
Ser_OG	14	2.67	0.08	2.80	0.04	0.24	0.08	−4.9	0.9
Thr_OG1	13	2.68	0.12	2.80	0.04	0.21	0.09	−4.8	0.7
Tyr_OH	12	2.64	0.13	2.81	0.07	0.25	0.04	−4.5	1.3
All	305	2.75	0.16	2.82	0.09	0.21	0.09	−5.0	1.5

[a] Interaction distance in the crystal-derived systems in Å. [b] Interaction distance in geometry-optimized systems in Å. [c] Probability of occurrence of water molecule at the hydration site, obtained from data-mining. [d] Interaction energy after geometry optimization in kcal mol^{-1}. [e] Backbone amide. [f] Backbone carbonyl.

In these groups, the mean interaction distance increased slightly upon optimization in all cases except Arg_NH1. The standard deviation of the interaction distance diminished significantly in all groups.

We examined the hydration sites whose crystal-derived interaction distance lied outside the range typical for hydrogen bonding, 2.6–3.0 Å. There were 56 crystal-derived water positions with d(cryst) shorter than 2.6 Å, and 20 water positions with d(cryst) > 3.0 Å. The deviations from the typical distances can be explained by two independent phenomena. First, a large portion of the waters with very short distances was found to interact with the carboxylate group of a negatively charged AA (13× Asp, 13× Glu). This can be interpreted in terms of the carboxylate functional groups' propensity to interact not only with water molecules, but also with alkali-metal ions (Na^+, K^+) and alkaline-earth metal ions (Mg^{2+}), some of which have shorter interaction distances than water. These light metal cations may be in some cases incorrectly identified as water molecules due to their similar electron density, thus shifting the mean distance to shorter values. Secondly, the crystal-derived water positions were based on Fourier averaging over a cluster of hydrated AA residue conformers, and thus contain an inherent imprecision caused by the variation of the AA conformations within the cluster. Besides that, the crystallographic structures used for the analysis, despite having been selected for high resolution (better than 1.8 Å), inevitably varied in many factors, including the methodology used in determining water positions in the hydration layer. Taken together, the QM calculations represent a valuable contribution in both confirming the existence of the crystal-derived hydration sites and refining their position.

Interaction energies

For each hydration site, the interaction energy before and after geometry optimization of the water position was calculated (File F1, ESI†). For 78% of the hydration sites, the change in interaction energy upon optimization (ΔE_{int}) is relatively small, below 1.0 kcal mol^{-1} in absolute value (Fig. 2b). If we look at the relationship between position displacement Δx and interaction energy change ΔE_{int} (Fig. S1, ESI†), it can be seen that for $\Delta x < 0.2$ Å, the interaction energy change is almost always below 1.0 kcal mol^{-1}, while for larger Δx the range of ΔE_{int} raises sharply. This phenomenon can be explained by the shape of the energy well surrounding the local minimum on the potential energy surface.

In the following text we will focus on the of interaction energy after geometry optimization, E_{int}(opt). The largest interaction energies were typically found for water interaction with the side-chains of charged AAs, especially for the acidic residues Asp (atoms OD1, OD2) and Glu (atoms OE1, OE2), but also for basic residues Arg (atoms NE, NH1, NH2) and Lys (atom NZ), with mean E_{int}(opt) of −7.3, −7.0, −5.5 and −6.0 kcal mol^{-1}, respectively. Very large interaction energies were observed also for interaction of water with the uncharged His residue ND1 atom, while the interaction energies for the NE2 atom of His were much lower (mean E_{int}(opt) −8.2 and −3.9 kcal mol^{-1}, respectively). For a detailed discussion of specific examples, see below.

We were interested to see if the crystal-derived occupancies of hydration sites correlate with the calculated interaction energies. A summary of the data for groups of hydration sites based on the AA and atom type of their nearest heavy atom is in Table 1. Rather surprisingly, we have found that there is no significant correlation between the interaction energy, E_{int}(opt) and the occupancy. Despite a broad range of occupancies, the variation of interaction energies within the groups is usually very small (Fig. S2, ESI†). Although it would be logical to assume that a stronger interaction between water molecule and a particular atom type would manifest itself also in higher occupancies of the hydration sites, the relationship between the two quantities seems to be masked by other variables. While the interaction energy is largely determined by local factors – the AA residue and the nearest heavy atom type (see below), the occupancy reflects also broader aspects of the protein architecture, including the average solvent accessibility of the given AA conformer, polarity of its solvent-accessible surface and side-chain packing. Additionally, the occupancy of the crystal-based hydration site is an average over many structures, whereas the DFT-derived interaction energy is calculated for a single structure of an amino acid–water complex. Another factor one has to bear in mind is the accuracy of the DFT calculations.

The particular atom types, however, showed correlations between the interaction energy, E_{int}(opt) and interaction distance, d(opt) (Fig. S3, ESI†), with shorter interaction distances corresponding to lower (more favorable) interaction energies.

Not surprisingly, the water–AA interactions involving nitrogen tended to have larger distances and weaker interaction energies than interactions involving oxygen. Also, interactions with charged residues resulted in shorter distances and lower (more favorable) interaction energies than interactions with neutral residues. The exception was His residue's ND1 atom, which showed the shortest interaction distances and lowest interaction energies of all nitrogen-involving groups (more detailed discussion below).

Browsing the Atlas: noteworthy cases

In the following paragraphs, we examine stereochemistry and energetics of hydration sites of representatives of all types of residues – small and large, charged, polar, hydrophobic and aromatic: Ala, Asp, His, Leu, Thr, Trp, and Tyr. We concentrate on discussing of some interesting positions of hydration sites that we detected in the previous study. The QM calculations allowed us to investigate these interactions in more detail, namely verify whether these HS positions represent minima on the potential energy surface, detect the preferred hydrogen atom orientation and calculate the energy of the interaction.

Alanine

Alanine residue was selected as a model for the hydration of the least hindered chiral main chain. In the helical form, two clearly localized hydration sites were identified, one for the carbonyl group and one for the amide, the carbonyl one has a high occupancy of 38% and an asymmetrical position relative to the carbonyl group. In the extended form, water distribution is more delocalized, with no hydration sites identified for the default occupancy threshold of 10%. However, lowering the threshold allowed to identify three HS, one for amide and two for carbonyl group, located in asymmetrical positions. For all of the alanine hydration sites, Δx upon optimization was small, below 0.5 Å. The QM calculations thus confirm the hydration site positions identified previously for Ala, despite the low occupancies of some of them. The interaction energies for amide HS in Ala (Ala_H_NA_HS2 and Ala_E_NA_HS1, E_{int}(opt) = −3.8 and −3.5 kcal mol^{-1}, respectively) are comparable to the average value for backbone amide (−3.9 kcal mol^{-1}, Table 1), while interaction energies of carbonyl HS (Ala_H_NA_HS1, Ala_E_NA_HS2 and Ala_E_NA_HS3, E_{int}(opt) = −4.2, −3.9 and −2.8 kcal mol^{-1}, respectively) are weaker than the average value for backbone carbonyl group (−4.7 kcal mol^{-1}, Table 1). The interaction energy is most favorable for the HS with highest occupancy (Ala_H_NA_HS1) and least favorable for the one with lowest occupancy (Ala_E_NA_HS3).

Aspartic acid

Asp residue conformers are, together with Gln and Glu, those that have the highest number of hydration sites. In most conformers, water distribution around the side chain carboxyl group is asymmetric, with HS of OD1 often located out of plane of the carboxyl group,[27] and with some HS forming bridges between side chain and main chain. Distances between some HS suggest water can form interaction networks around the hydrated residue (Asp_E_g−, Asp_H_g−). The value of Δx was small for most of the HS, which confirmed the crystal-derived positions, except few cases of HS which showed very large Δx (*e.g.*, Asp_H_t_HS3, Asp_E_g−_HS1).

The large Δx displacements were most pronounced in the Asp_E_g− conformer, in which three out of six HS showed Δx greater than 1.0 Å, namely HS1, HS4 and HS6. A possible explanation for this behavior might be the missing broader structural context of the modeled system, particularly the cooperativity between water molecules in the hydration layer. This might be especially critical in a highly hydrophilic residue such as Asp; in the concerned Asp_E_g− conformer the water:AA ratio is 2.5, one of the highest observed for any of the studied conformers. Therefore, we experimented with geometry optimizations involving more than one water molecule. In the Asp_E_g− conformer we included water molecules in all six HS and performed 400 simultaneous optimizations with differing water hydrogen atom starting positions, to allow for the formation of different H-bond networks. For each of the three HS that showed large Δx in the original optimization, we were able to find a geometry minimum in which the water molecule moved less than 0.7 Å from the crystal-derived position; the position was typically stabilized by an H-bond to a nearby water molecule (data not shown). In the case of HS1, water displacements were the smallest, (down to 0.05 Å), and showed a negative correlation (R^2 = 0.72) with the displacement of HS2, in other words, water molecule in HS2 had to move in to stabilize HS1 water molecule. We conclude that for the stabilization of water molecules in some HS, a larger structural environment is required.

The interactions of Asp side chain hydration sites are very strong, with mean E_{int}(opt) for Asp OD1 and OD2 hydration sites −7.3 kcal mol^{-1}. Interestingly, hydration sites of OD2 have on average stronger interactions than hydration sites of OD1 (*cf.* Table 1). This is reflected also in individual conformers, where even the symmetrically located HS of OD1 and OD2 have different energy, despite similar geometrical arrangement relative to the carboxyl group (*e.g.*, Asp_H_t HS4 and HS5; Asp_E_t HS2 and HS3).

Histidine

The interpretation of the data for this residue is complicated due to the fact that His can be present in different protonation states at a near-neutral physiological pH with protonation on either ND1 or NE2, or on both nitrogens and thus positively charged. Thus, the crystal-derived data are inevitably an average over these different states of His present in the set of crystallographic structures. Although we decided to model His in all cases in the Atlas as protonated on the NE2 atom (which is the most likely state for neutral His under physiological pH[44]), one should keep in mind that in a particular structure the His residue can be in a different protonation state.

The Δx values were generally small for HS of His conformers, the largest (0.84 Å) was found for HS3 in His_H_t, which moved closer to position of the nearby unoccupied HS1. His residue is the only one where we observed correlation between occupancy

and interaction energy of the HS, albeit only for the ND1 atom (R^2 = 0.72, Fig. S2, ESI†). As mentioned above, significantly stronger interaction was found for ND1 hydration sites, where water acts as H-bond donor, than for those of NE2, where it acts as H-bond acceptor (Table 1). Since the ND1 atom is in closer proximity to the residues' backbone than the NE2 atom, water molecules interacting with ND1 in certain conformers have the possibility to interact favorably also with the polar backbone atoms, forming main-chain–side-chain bridging interactions (His_H_g+_HS1, His_E_g+_HS1), which lowers their interaction energy even further. A similar phenomenon was observed also in His_H_t_HS3, which is a hydration site of backbone carbonyl. Due to its off-plane bridging interaction with the heteroaromatic ND1 atom, its interaction energy is much lower (E_{int}(opt) = −7.7 kcal mol^{-1}) than the average value for backbone carbonyl hydration site (−4.7 kcal mol^{-1}), in fact the lowest observed for any backbone carbonyl HS in the data set.

Leucine

Due to the hydrophobic nature of the Leucine side chain, only backbone hydration sites were detected. Their Δx were very small, while other properties are strongly conformationally dependent. Hydration sites of carbonyl oxygen were only found in helical conformations, their strength depends also on side-chain orientation. In Leu_H_g+ it has high occupancy and also E_{int}(opt) above the average. In Leu_H_g− its occupancy is lower and also E_{int}(opt) is less favorable. In Leu_H_t, water occupancy at this position was only about 3%, *i.e.*, below the threshold for hydration site detection. Hydration sites of backbone amide, in conformers where they were detected (Leu_H_t, Leu_E_g−, Leu_E_t), have similar occupancy and interaction energy. The other conformations seem to be sterically unsuitable for amide hydration.

Threonine

In Threonine, as well as in Serine and Tyrosine, the orientation of the side-chain hydroxyl group can vary, and therefore water molecule in a given HS can potentially act either as H-bond donor or H-bond acceptor of the OG1 atom, depending on a broader structural context. In our calculations, only one of these two possible H-bonding states was modeled. The hydration sites of Thr side chain, as modeled in this study, can thus be divided into two groups – those where the geometry optimization resulted in water acting as H-bond donor or acceptor, respectively. The H-bond acceptor mode is more energetically favorable, and results in slightly shorter interaction distances (Fig. S4, ESI†). For both interaction modes, the displacements were generally small (average Δx values of 0.21 and 0.52 Å for H-bond donor and acceptor modes, respectively).

Tryptophan

In almost all Trp conformers investigated in the previous study, we observed main-chain hydration sites in positions where water could interact additionally with the side chain *via* unconventional interactions, such as carbon-donor[28] hydrogen bonds or off-plane heteroaromatic interactions.[29] The QM calculations corroborated some of these presumed interactions, such as the carbon-donor H-bond to CD1 atom in Trp_H_g+_HS1 and Trp_H_g−_HS2, as well as the off-plane heteroaromatic interaction in Trp_H_t_HS2 and Trp_E_g+_HS2. On the other hand, other previously supposed interactions were not confirmed (carbon-donor H-bonds in Trp_H_g+_HS2 and Trp_E_g−_HS1), due to an unsuitable orientation of the water molecule's lone pair away from the CH group (the example of Trp_H_g+_HS2 is shown in Fig. 1). Nevertheless, the Δx value in both these systems was small, and the interaction energy in Trp_H_g+_HS2 was significantly higher than the average value for backbone carbonyl HS. In the literature, there has been some debate concerning the nature of the off-plane heteroaromatic interactions. While the article by Stollar *et al.*,[29] which described this interaction for water–His and water–Trp contacts in proteins, suggested lone-pair–π interaction, a recent study[45] showed OH–π interaction to be more favorable. Our DFT-D calculations are consistent with the latter interpretation.

Tyrosine

The phenomenon of different interaction energy and distance depending on the H-bond donor or acceptor mode of the modeled water molecule, as described above for Threonine, is even more pronounced in Tyrosine. Due to the size of the Tyr side chain, the OH atom hydration sites cannot interact with the main-chain polar atoms, and the effect is thus isolated. In the plot of interaction energy *versus* interaction distance (see Fig. S4, ESI†), the two groups of hydration sites are clearly separated. The H-bond donor and H-bond acceptor groups have d(opt) values from 2.73 to 2.76 Å and from 2.84 to 2.90 Å, respectively; their E_{int}(opt) values range from −6.0 to −5.4 and from −3.4 to −3.0 kcal mol^{-1}, respectively.

In our previous work, we suggested a number of possible interactions between main-chain HS of Tyr and the phenyl ring of its side chain, either *via* a carbon-donor hydrogen bond or *via* an OH–π interaction. The calculations performed in the present study showed a lower than average E_{int}(opt) values (*cf.* Table 1) for two of the backbone carbonyl HS, namely Tyr_H_t_HS3 and Tyr_H_g−_HS1, thus confirming the presence of an additional favorable interaction. In Tyr_H_t_HS3, the water molecule creates an OH–π interaction with the aromatic ring, resulting in E_{int}(opt) value of −6.2 kcal mol^{-1}; in Tyr_H_g−_HS1, the water molecule's additional carbon-donor H-bond to the CD1 atom resulted in E_{int}(opt) value of −5.2 kcal mol^{-1}.

Conclusions

Hydration of biomolecules has increasingly been recognized as an important topic in molecular biology and biophysics.[12] Detailed knowledge of the mutual interactions between proteins, nucleic acids and their aqueous environment is required for a thorough understanding of the structural properties of these systems, and hence also their functions. To remedy the lack of resources providing an overview of protein hydration, we decided to create the web-based atlas, WatAA, presented in this article.

In this atlas, we used synergies between data mining of experimental X-ray structures and *ab initio* QM calculations. The WatAA website thus gathers a substantial volume of information obtained from analysis of protein crystal structures, with crystal-derived hydration sites of a large number of AA conformers. This data is complemented for each hydration site with information about hydrogen atom orientations and energetics of the interaction obtained from QM calculations. The website interface combines these two aspects in an intuitive and illustrative way and provides tools for their direct comparison and visualization. The user can get a quick overview of the physical and structural properties of the ordered hydration layer of AAs in proteins.

The analysis of the data provided in WatAA can lead to interesting new insights, as we illustrated in this article *e.g.*, for the hydration of histidine. Furthermore, we suggest that the application of the conformation-specific hydration patterns of AA residues could lead to more precise water placement algorithms in structural bioinformatics applications such as crystallographic refinement, protein structure prediction, computational drug design as well as in protein–protein and protein–DNA docking and recognition. A proof of concept application of the data is described in the ESI.† WatAA: Atlas of Protein Hydration is available free of charge and without any login requirement at www.dnatco.org/WatAA.

Acknowledgements

This study was supported by Institutional Research Project of the Institute of Biotechnology, CAS [RVO 86652036], by ERDF – Project BIOCEV [CZ.1.05/1.1.00/02.0109], by Czech National Infrastructure for Biological Data – ELIXIR CZ [LM2015047], and by the Czech Science Foundation, GA CR [grant No. P205/12/P729]. Access to the computing and data storage facilities of MetaCentrum [LM2010005] is greatly appreciated.

References

1 M. Chaplin, *Nat. Rev. Mol. Cell Biol.*, 2006, **7**, 861–866.
2 P. Ball, *Chem. Rev.*, 2008, **108**, 74–108.
3 T. Wyttenbach and M. T. Bowers, *Chem. Phys. Lett.*, 2009, **480**, 1–16.
4 M. G. Wolf and G. Groenhof, *J. Comput. Chem.*, 2012, **33**, 2225–2232.
5 G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 3352–3357.
6 L. Jiang, B. Kuhlman, T. Kortemme and D. Baker, *Proteins: Struct., Funct., Bioinf.*, 2005, **58**, 893–904.
7 P. L. Kastritis, K. M. Visscher, A. D. J. van Dijk and A. M. J. J. Bonvin, *Proteins: Struct., Funct., Bioinf.*, 2013, **81**, 510–518.
8 H. I. Parikh and G. E. Kellogg, *Proteins: Struct., Funct., Bioinf.*, 2014, **82**, 916–932.
9 D. Zhong, S. K. Pal and A. H. Zewail, *Chem. Phys. Lett.*, 2011, **503**, 1–11.
10 A. C. Fogarty, E. Duboué-Dijon, F. Sterpone, J. T. Hynes and D. Laage, *Chem. Soc. Rev.*, 2013, **42**, 5672–5683.
11 V. Conti Nibali and M. Havenith, *J. Am. Chem. Soc.*, 2014, **136**, 12800–12807.
12 L. Biedermannová and B. Schneider, *Biochim. Biophys. Acta, Gen. Subj.*, 2016, **1860**, 1821–1835.
13 S. Khodadadi and A. P. Sokolov, *Biochim. Biophys. Acta, Gen. Subj.*, 2017, **1861**, 3546–3552.
14 B. Bagchi, *Water in Biological and Chemical Processes*, Cambridge University Press, Cambridge, 2013.
15 G. Hummer and A. Tokmakoff, *J. Chem. Phys.*, 2014, **141**, 22D101.
16 K. Morgenstern, D. Marx, M. Havenith and M. Muhler, *Phys. Chem. Chem. Phys.*, 2015, **17**, 8295–8296.
17 S. K. Kim, T. Ha and J.-P. Schermann, *Phys. Chem. Chem. Phys.*, 2010, **12**, 10145–10146.
18 V. Makarov, B. M. Pettitt and M. Feig, *Acc. Chem. Res.*, 2002, **35**, 376–384.
19 N. Thanki, J. M. Thornton and J. M. Goodfellow, *J. Mol. Biol.*, 1988, **202**, 637–657.
20 B. Schneider, D. M. Cohen, L. Schleifer, A. R. Srinivasan, W. K. Olson and H. M. Berman, *Biophys. J.*, 1993, **65**, 2291–2303.
21 P. Auffinger and Y. Hashem, *Bioinformatics*, 2007, **23**, 1035–1037.
22 X. Chen, I. Weber and R. W. Harrison, *J. Phys. Chem. B*, 2008, **112**, 12073–12080.
23 D. Matsuoka and M. Nakasako, *J. Phys. Chem. B*, 2009, **113**, 11274–11292.
24 E. Nittinger, N. Schneider, G. Lange and M. Rarey, *J. Chem. Inf. Model.*, 2015, **55**, 771–783.
25 S. Hong and D. Kim, *Proteins: Struct., Funct., Bioinf.*, 2016, **84**, 43–51.
26 G. König and S. Boresch, *J. Phys. Chem. B*, 2009, **113**, 8967–8974.
27 L. Biedermannová and B. Schneider, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2015, **71**, 2192–2202.
28 R. J. Petrella and M. Karplus, *Proteins: Struct., Funct., Genet.*, 2004, **54**, 716–726.
29 E. J. Stollar, J. L. Gelpí, S. Velankar, A. Golovin, M. Orozco and B. F. Luisi, *Proteins: Struct., Funct., Bioinf.*, 2004, **57**, 1–8.
30 P. D. Renfrew, G. L. Butterfoss and B. Kuhlman, *Proteins: Struct., Funct., Genet.*, 2008, **71**, 1637–1646.
31 Atlas of Protein Side-Chain Interactions, www.biochem.ucl.ac.uk/bsm/sidechains, accessed December 2016.
32 P. Jurečka, J. Černý, P. Hobza and D. R. Salahub, *J. Comput. Chem.*, 2007, **28**, 555–569.
33 J. Černý, P. Jurečka, P. Hobza and H. Valdés, *J. Phys. Chem. A*, 2007, **111**, 1146–1154.
34 K. Berka, R. Laskowski, K. E. Riley, P. Hobza and J. Vondrášek, *J. Chem. Theory Comput.*, 2009, **5**, 982–992.
35 R. M. Hanson, J. Prilusky, Z. Renjian, T. Nakane and J. L. Sussman, *Isr. J. Chem.*, 2013, **53**, 207–216.
36 D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, I. T. E. Cheatham, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus,

I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu and P. A. Kollman, *AMBER 14*, University of California, San Francisco, 2014.

37 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.

38 R. Ahlrichs, M. Bär, M. Häser, H. Horn and C. Kölmel, *Chem. Phys. Lett.*, 1989, **162**, 165–169.

39 *TURBOMOLE (V6.2), TURBOMOLE GmbH*, University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 2010.

40 A. Klamt and G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799–805.

41 A. Schäfer, A. Klamt, D. Sattel, J. C. W. Lohrenz and F. Eckert, *Phys. Chem. Chem. Phys.*, 2000, **2**, 2187–2193.

42 RStudio Team, RStudio: Integrated Development for R (version 0.99.903), RStudio, Inc., Boston, MA, 2015.

43 H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York, New York, 2009.

44 I. K. McDonald and J. M. Thornton, *Protein Eng.*, 1995, **8**, 217–224.

45 J. Novotný, S. Bazzi, R. Marek and J. Kozelka, *Phys. Chem. Chem. Phys.*, 2016, **18**, 19472–19481.

