Conservation of water molecules in protein binding interfaces

Zhenhua Li and Ying He

School of Computer Engineering,
Nanyang Technological University,
Block NS4-04-33, 50 Nanyang Avenue,
Singapore 639798
E-mail: lizh0021@ntu.edu.sg
E-mail: YHe@ntu.edu.sg

Longbing Cao

Faculty of Engineering and Information Technology,
Advanced Analytics Institute,
University of Technology Sydney,
CC02.01.207 (Level 1, Building 2),
Blackfriars Street, Chippendale, Blackfriars Campus,
P.O. Box 123 Broadway, NSW 2007, Australia
E-mail: longbing.cao@uts.edu.au

Limsoon Wong

School of Computing,
National University of Singapore,
Level 3, Building COM1, 13 Computing Drive,
Singapore 117417
E-mail: WongLS@Comp.NUS.EDU.SG

Jinyan Li*

Advanced Analytics Institute,
University of Technology Sydney,
P.O. Box 123, Broadway NSW 2007, Australia
Fax: +61 2 9514 9276
E-mail: Jinyan.Li@uts.edu.au
*Corresponding author

Abstract: The conservation of interfacial water molecules has only been studied in small data sets consisting of interfaces of a specific function. So far, no general conclusions have been drawn from large-scale analysis, due to the challenges of using structural alignment in large data sets. To avoid using structural alignment, we propose a solvated sequence method to analyse water conservation properties in protein interfaces. We first use water information to label the residues, and then align interfacial residues in a fashion similar to normal sequence alignment. Our results show that, for a water-contacting interfacial residue, substituting it into hydrophobic residues tends to desolvate the local area. Surprisingly, residues with short side chains also tend not to lose their contacting water, emphasising the role of water in shaping binding sites. Deeply buried water molecules are found more conserved in terms of their contacts with interfacial residues.

Keywords: protein-protein interface; water; interface; substitution; conservation; burial level.

Reference to this paper should be made as follows: Li, J., He, Y., Cao, L., Wong, L. and Li, J. (xxxx) 'Conservation of water molecules in protein binding interfaces', *Int. J. Bioinformatics Research and Applications*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Zhenhua Li received his BEng and MEng Degrees from Wuhan University, P.R. China, in 2007 and 2009, respectively. Since 2009, he has been a PhD student in the school of Computer Engineering, Nanyang Technological University, Singapore. His current research interests include bioinformatics and data mining.

Ying He received the PhD in Computer Science from Stony Brook University in 2006. Since then, he joined Nanyang Technological University as an Assistant Professor. His research interests fall into the general areas of visual computing and he is particularly interested in the problems which require geometric analysis and computation.

Dr. Longbing Cao was awarded a PhD in Intelligent Sciences in Chinese Academy of Sciences, China and PhD in Computing Sciences at the University of Technology Sydney, Australia. He is a Professor and the Founding Director of the Advanced Analytics Institute at the University of Technology Sydney, Australia, and the Data Mining Research Leader of the Australian Capital Markets Cooperative Research Centre. His research interests include data mining and machine learning and their applications, behaviour informatics, multi-agent technology, and agent mining. He has prodigious experience in practical innovation in enterprise data mining and analytics in many different domains including the public sector, social welfare, capital markets, banking, insurance, telecommunication and education.

Limsoon Wong is a provost's chair professor of computer science and a professor of pathology at the National University of Singapore (NUS). He is currently head of the computer science department at NUS. Before joining NUS, he was the Deputy Executive Director for Research at A*STAR's Institute for Infocomm Research. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He has/had served on the editorial boards of Information Systems, Journal of Bioinformatics and Computational Biology, Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Drug Discovery Today, and Journal of Biomedical Semantics.

Jinyan Li received his PhD Degree in computer science from the University of Melbourne in 2001. He is an associate professor in the University of Technology Sydney, Australia. His research is focused on protein structural bioinformatics, statistically important discriminative patterns, interaction subgraphs, and classification methods. Jinyan has published over 100 research articles. One of his most interesting works was a cancer diagnosis technique for childhood leukemia disease through the discovery of emerging patterns from the gene expression data, and currently he is very interested in infectious disease studies and water bioinformatics by exploring graph theories and biological water exclusion principles.

1 Introduction

Protein-protein binding interfaces are wet places involving many water molecules that shape the binding sites by filling cavities, bridging the local inter-chain contacts and networking some specific contacts by hydrogen bonding (Ikura et al., 2004; Li and Lazaridis, 2007; Reichmann et al., 2008; Ahmed et al., 2011). Interfacial water molecules are also capable of adjusting binding specificity and affinity (Wilson, 1989; Meenan et al., 2010). They are an integral part of binding interfaces rather than a separate binding environment.

Structural, functional and energetic roles of water molecules in an interface were previously investigated using structural analysis (Urakubo et al., 2006; Knight et al., 2009), molecule dynamic simulation (Lu et al., 2006) or statistical analysis (Rodier et al., 2005). Importance of water molecules was earlier quantified using site-directed mutagenesis (Ikura et al., 2004; Urakubo et al., 2006; Reichmann et al., 2008). Theoretical energetic contribution of water molecules to protein binding was examined by incorporating the water molecules to energy models implicitly (Benedix et al., 2009) or explicitly (Jackson et al., 1998; Guerois et al., 2002).

Water conservation at protein-binding sites is another interesting topic. It was investigated previously by focusing on small data sets of one or two families or functional groups (Poornima and Dean, 1995; Shaltiel et al., 1998; Bottoms et al., 2002; Bairagya et al., 2009). For example, the authors of Shaltiel et al. (1998) looked at only seven different crystal structures of protein kinase A; and they detected six conserved and structured water molecules near the active site. However, no large-scale analysis has been undertaken on this topic; and no high-level pattern of interfacial water conservation has been drawn yet. With data of protein complexes growing fast in the Protein Data Bank (PDB) (Berman et al., 2000), we can significantly enlarge the scope by analysing water conservation properties across protein-binding interfaces in PDB.

Alignment is a necessary step in such a conservation study. Unlike residues in a protein structure, water molecules are not a part of the protein sequence. Thus sequence alignment cannot be used directly. Structural alignment is a good idea. However, it is difficult to apply it to large-scale data sets due to its high computational cost. Another challenge is that structural alignment is not yet as standard as sequence alignment. Different structural alignment algorithms may lead to opposite observations and conclusions. An example can be seen from the following two works done separately by Kinjo and Nakamura (2010) and Gao and Skolnick (2010b) on the structural space of protein-protein interfaces. Kinjo and Nakamura (2010) used a structural alignment algorithm called GIRAF (Kinjo and Nakamura, 2009); they found that the structural similarity between protein-binding interfaces could be observed only in homologous families. In another work done by Gao and Skolnick (2010b), the authors used another algorithm named iAlign (Gao and Skolnick, 2010a), and concluded that the structural space of protein-protein interfaces is highly connected with significant similarities observed universally even within randomly generated interfaces.

Although water molecules are not a part of protein amino acid sequences, they are closely engaged in protein interactions via their contacts with residues. This type of contact is always biased; i.e., water prefers to contact to polar or hydrophilic residues, and hence the conservation of these contacts is biased. The more specific the binding is, the more reproducible it is. The conservation patterns of water can be inferred by investigating the hydration patterns of the residues. In this work, we conduct an analysis on the conservation of interfacial water molecules based on a tripartite interface model and a solvated sequence idea. First, interfacial water molecules are identified by using the tripartite interface model. They are then assigned to their neighbouring residues of a protein (the solvated sequence idea). With the interfacial hydration properties of interfacial residues in hand, sequence alignment is followed in the database of sequences to align the interfacial residues. In particular, we are interested in the solvation change of residues in a residue substitution. We find that substitutions to hydrophobic residues tend to desolvate the residue. We also find that deeply buried residues are more likely to preserve their water-contacting sites, indicating that water molecules at the core of protein interfaces are organised more sophisticatedly so that they are harder to be removed, added or shifted.

2 Methods

2.1 Data set of interface overlaps

A search against PDB (Berman et al., 2000) was done in August 2010 to obtain our data, with the following requirements:

i the structure is solved by X-ray crystallography;

ii the resolution is no worse than 3.0Å;

iii there are at least 2 protein chains in the biological assembly

iv the chain length is larger than 30.

PDB entries with artificial mutations are further removed. As water molecules tend to be under reported (Carugo and Bordo, 1999), in order to guarantee the quality of water information, the protein structures used are required to simultaneously contain more than 20% and 1% water molecules in number and proportion respectively. Only heavy atoms are used. The biological interfaces are then determined within the PDB structures obtained. First, we extract those biological assemblies that can be obtained without coordinate transformation. Then, bipartite interfaces are detected within these biological assemblies. Only those interfaces having no less than 100 atoms are kept, since two spatially-nearby protein chains with too few nearby atoms are not likely to be in a real interaction. We obtained 30,635 protein-protein interfaces distributed in 10,009 PDB entries.

Interface alignments are needed to capture the conservation of interfacial water molecules. However, structural alignment will be very expensive for such a large data set (in the order of 10^4). Therefore, to capture the evolutionary relationship between protein interfaces, and also to remove the redundancy, we define the similarity and alignment between interfaces based on sequence alignment. Given the sequences of protein interactions that we obtained from PDB, an all-against-all BLAST search is performed with an E-value threshold of 0.0001.

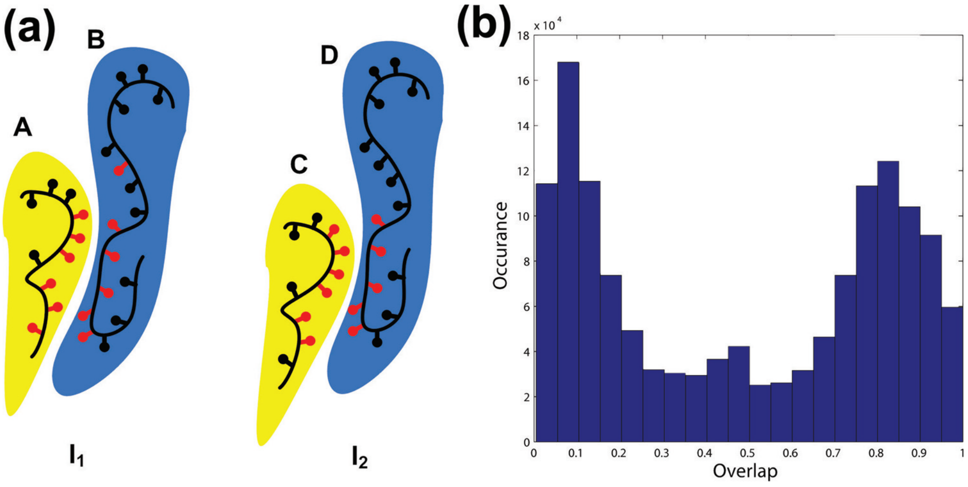
To measure the similarities between two interfaces, we define a concept of 'overlap' between two interfaces. Suppose we are given two interfaces $I_1 = \{A, B\}$ (the interface between chains A and B), and $I_2 = \{C, D\}$ (the interface between chains C and D). If each of the two chains in one protein interaction is aligned significantly with one of the two chains of the other interface (for example, A and C are aligned, and B and D are aligned), the 'overlap' between these two interfaces is defined as:

$$overlap(I_1, I_2) = \min\left\{\frac{2 \times |f_{A\to C}(I_1^A) \bigcap I_2^C|}{|I_1^A| + |I_2^C|}, \frac{2 \times |f_{B\to D}(I_1^B) \bigcap I_2^D|}{|I_1^B| + |I_2^D|}\right\} \qquad (1)$$

where $|X|$ is the cardinality of set X, I^A is the set of interfacial residues of interface I that belong to chain A, and $f_{A\to B}$ is the mapping function from the amino acids in chain A to chain B according to the sequence alignment. If chains A and B can also be aligned with chains D and C respectively at the same time, then whichever of the two cases leads to a higher overlap is used. Residue identity is not considered here. A residue is considered at the overlapping region of the two interfaces as long as it is aligned to an interfacial residue in the other interface. If there is no such alignment, the overlap is 0. A schematic diagram of the overlap calculation is shown in Figure 1(a).

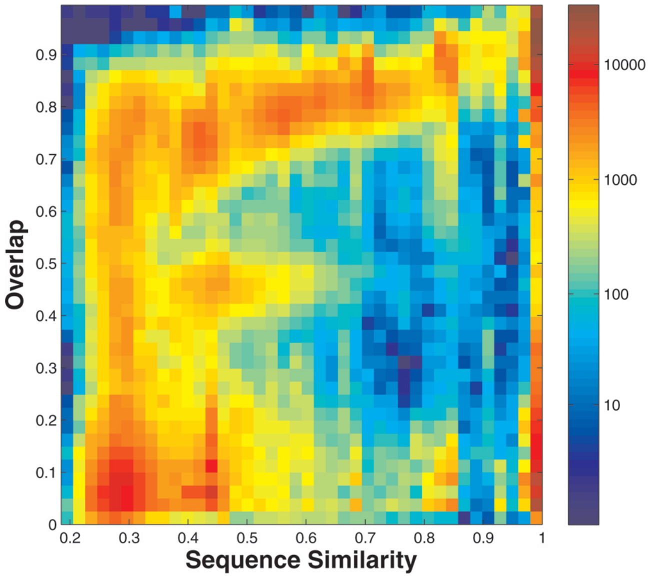
The distribution of the overlap between interfaces in our data set is shown in Figure 1(b), excluding those interface pairs without significant sequence alignment. A bimodal distribution is observed. The reason for such a distribution is that, two protein chains may interact with each other in multiple binding sites, resulting in lowly overlapped interface pairs. This is especially common in multi-oligomers where multiple chains are organised together with many binary interfaces inside. An example is the interface between α and β subunits of hemoglobin. In a hemoglobin tetramer, the α and β chains can interact in two different ways with two distinct binary interfaces. The overlap between them is very low. To better understand the distribution of overlap with respect to sequence similarity, we plot the joint distribution of interface overlap and sequence similarity in Figure 2. The sequence similarity of two interfaces is defined as the minimum of the similarities of the two sequence alignments. From Figure 2 it can be noted that, when the sequence similarity is very high, the interfaces can be of either high overlap or

Figure 1 (a) Calculating the overlap between interfaces. In chains A and C, 5 aligned positions occur in both interfaces; in chains B and D, 5 aligned positions occur in both interfaces. The overlap between I_1 and I_2 is then $min\{2 \times 5/(6+6), 2 \times 5/(5+6)\} = 5/6$. (b) Distribution of overlap in our data – a bimodal distribution (see online version for colours)



extremely low overlap, indicating the different binding sites between same proteins. When sequence similarity is low, three dense region of overlap is observed. The top part indicates that the interface can be very conserved ($overlap > 0.7$), when the sequence similarity is low. The middle part indicates that some interfaces are changed, but not completely – a part of the original interface remains. The dense bottom part consists the overlap between different binding sites. Recall that, we are not considering the residue identity in the overlap calculation. If there are not too many indels in interfaces, interfaces at the same sites of two proteins will highly overlap with each other. If there are many indels between two interfaces, we can actually consider them as different interfaces that are not aligned, as many indels can change the local water arrangement dramatically. Thus we removed those alignment pairs with low overlap. The threshold is set to 0.5, which is roughly the minimum between the two maxima; see Figure 1(b). This threshold roughly eliminates those overlap between different and shifted interfaces; see Figure 2. Redundancy is also removed, with another overlap threshold 0.8; that is, if two interfaces have an overlap larger than 0.8 (corresponding to the second peak), one of them is removed. The reason we remove these alignments is that PDB is a biased data base, where some proteins are reported repeatedly and some others are not. Some commonly seen proteins/protein complexes are, for example, antibodies and hemoglobins. If the redundancy is not removed, the results will be dominated by only a few types of protein-protein interactions. Two two thresholds can also remove most of those interface pairs with very high sequence similarity. When interface similarity is near 1.0 (Figure 2), most overlaps are either lower than 0.4 or higher than 0.8. Finally there are 4310 interface alignment pairs between 3481 interfaces, in which there are 178,251 pairs of residue-residue alignments.

Figure 2 Joint distribution of sequence similarity and interface overlap. Each cell is coloured according to the number of occurrence in the data set (see online version for colours)



2.2 Burial level and tripartite interface model

Given a protein complex, exposed water molecules are first removed repeatedly with a solvent accessibility threshold of 10 Å^2. In the remaining structure, the atoms are labelled as exposed or buried based on the threshold of 10 Å^2. Also, an atomic contact graph is built based on the remaining structure, in which nodes are atoms and edges are atomic contacts. Two atoms are defined to be in contact if they share a Voronoi facet and their distance is less than their radius plus 2.75Å, which is the diameter of a water molecule. A pseudo node representing bulk solvent is added to the graph, connecting to all the exposed atoms. The burial level of an atom is defined as the length of the shortest path from it to its nearest exposed neighbour. It is equal to the length of the shortest path from the pseudo node to it minus one.

We model protein interfaces as tripartite graphs. To calculate the interface tripartite graph, first a residue contact graph of all the residues (including water) is constructed. Two residues (water is also referred to as residue here) are defined to be in contact if at least one pair of atoms, each from one residue, is in contact. Based on the residue contact graph, interfacial water molecules are identified as those water molecules that are in contact with both sides of the interaction. At last, the interface residues from the two interacting proteins are identified as those residues that are in contact with the other protein or with interfacial water. The nodes in the interface tripartite graph consists of interfacial water molecules and interfacial residues from both proteins. Based on this definition, interfacial residues are those residues that contact directly with the other chain (dry residue), with interfacial water molecules only (wet spot) or with both the other chain and interfacial water at the same time (dual residue). More detailed information of the concept of burial level and tripartite interface model can be found in another previous work (Li and Li, 2010).

2.3 Substitutional conservation of water

We simply label an interface residue as water-contacting or non-water-contacting if it is or is not in contact with at least one water molecule. We define the conservation of water in the substitution from one residue type to another as the probability of the water-contacting state preserved after the substitution:

$$C_{substitution}(X \to Y) = P(L_Y = 1 \mid L_X = 1) \tag{2}$$

where X and Y belong to the 20 standard amino acid types and L_X is the label of X. $L_X = 1$ when X is water-contacting and $L_X = 0$ otherwise. $X \to Y$ stands for the substitution of residue type X to residue type Y, both being interfacial residues. $C_{substitution}(X \to Y)$ can be estimated by:

$$C_{substitution}(X \to Y) = \begin{cases} \dfrac{2 \times C(L_X = 1, L_Y = 1)}{2 \times C(L_X = 1, L_Y = 1) + C(L_X = 1, L_Y = 0)}, & \text{if } X = Y \\ \dfrac{C(L_X = 1, L_Y = 1)}{C(L_X = 1, L_Y = 1) + C(L_X = 1, L_Y = 0)}, & \text{if } X \neq Y \end{cases} \tag{3}$$

where $C(L_X = x, L_Y = y)$ is the number of such residue-residue alignments that residues X and Y are aligned and their labels are x and y, respectively. Note that the count of residue alignments when $X = Y$ is doubled. The reason is that if two aligned residues $X1$ and $X2$ are of the same type X, they contribute twice to the substitution $X \to X$ ($X1 \to X2$ and $X2 \to X1$).

Also, only side chains are considered when the residues are labelled. That is, $L_X = 1$ if and only if its side chain is in contact with water. For residue types other than glycine, their side chains consist of atoms other than C, N, C^{α}, and O. For glycine, C^{α} is defined as its side chain because it does not have other heavy atom.

2.4 Hydration site conservation

When the two residues in a residue-residue alignment $X \to Y$ belong to the same amino acid type, it is interesting to understand the conservation of water contact in more detail at an atomic level. We define the water conservation status of an atom in a residue type as its probability of contacting water in all residue alignments $X' \to Y'$, where X', Y' belong to the same residue type as X, Y. It is calculated by:

$$C_{site}(X_a) = \frac{C(L_{X_a} = 1, L_{Y_a} = 1)}{C(L_{X_a} = 1, L_{Y_a} = 1) + C(L_{X_a} = 1, L_{Y_a} = 0)} \tag{4}$$

where X_a denotes the atom a in residue X, and its label L_{X_a} is 1 if it is in contact with water, or 0 otherwise.

2.5 Water-contacting mode and mode conservation

When the amino acid is not changed in a site, the similarity in their water contact can be measured. For an interfacial residue, its water-contacting mode is defined as a vector of length N, where N is the number of heavy atoms in that residue. For example, in an arginine residue, there are 11 heavy atoms. Each value in the vector is 1 or 0, denoting whether the corresponding atom is water contacting or not. The conservation between the solvation of two residues of the same type is defined as the cosine similarity between their water-contacting modes:

$$C_{mode}(i,j) = \frac{M_i \cdot M_j}{\|M_i\| \times \|M_j\|} = \frac{|S_i \bigcap S_j|}{|S_i| \times |S_j|} \tag{5}$$

where M_i is the water-contacting mode of residue i and S_i is the set of water-contacting atoms of residue i. $\|M\|$ is the Euclidean norm of vector M.

3 Results and discussions

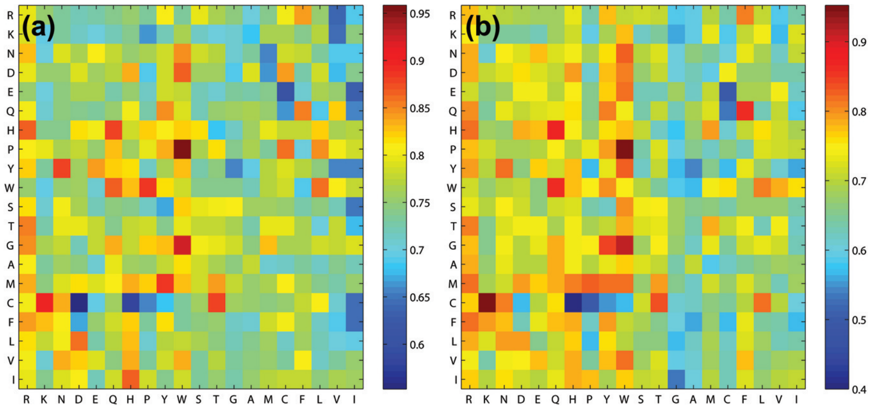
3.1 Solvated substitution

Figure 3 shows substitutional conservation analysis results when a whole residue (a) or when only its side chain (b) is considered. In both cases, it can be noted that substitution to hydrophobic residues (right part) are more likely to desolvate the residue. Let us divide both panels into four parts: the top-left, the bottom-left, the top-right, and the bottom-right. The top-left part corresponds to substitutions within polar or charged residues. As can be seen, their contacting water molecules are very likely to be preserved in the substitution. Large substitutional conservation is observed in the substitution from hydrophobic residues to polar or charged residues (bottom-left part). This indicate that, if a hydrophobic residue is already in contact with water molecules, substitution of it into polar or charged residues is not likely to remove the contacting water. In the case of changing a polar or charged residue to hydrophobic residue (top-right part), water-contacting status is not that conserved. Thus substituting a polar or charged residue to a hydrophobic one is very likely to desolvate it. If the substitution is within hydrophobic residues (bottom-right part), contacting water is also not likely conserved, indicating contacts between hydrophobic residues and water molecules are very unstable and easy to break.

Comparing Figure 2(a) and (b), it can be seen that the bias is clearer when only the side chain is considered. This is reasonable as side chain is the place where the residues differ from each other.

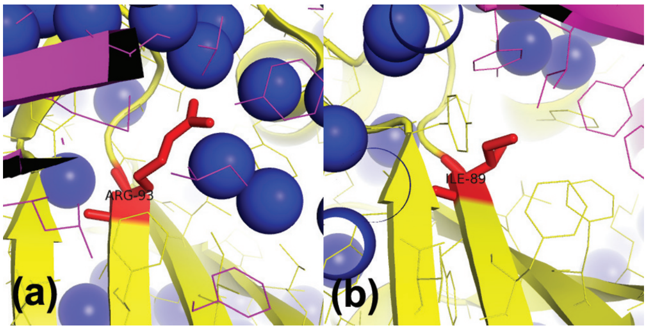
An example of water structural change due to residue substitution is shown in Figure 4. In this figure, the two aligned residues are of different types. For the arginine residue in a ribokinase dimer interface shown in Figure 4(a), a cluster of water molecules is observed around it. However, they are not blocking the direct contact of this residue to the other side, and are just offering a wet environment for its contacts. The local water structure near the isoleucine in a 2-keto-3-deoxygluconate kinase subunit interface shown in Figure 4(b) is dramatically different, although they are aligned very well. Overall, the interface is rich in water

Figure 3 Substitutional conservation of water between residue types when: (a) whole residue is considered or, (b) only side chain is considered. Each cell of the grid corresponds to the substitutional conservation from one type of residue (left) to another (top). The residue types are ordered according to the hydropathy index (Kyte and Doolittle, 1982) (see online version for colours)



molecules. There are 26 interfacial water molecules, contributing more than 6% of the interfacial atoms, yet none of them successfully access the isoleucine. A large area of hydrophobic contact is created. Usually, such large hydrophobic contacting area can contribute a lot to binding free energy (Sharp et al., 1991).

Figure 4 Two aligned residues and their surrounding water arrangement: (a) an arginine (red and sticked) in a ribokinase dimer interface [PDB:1VM7] and (b) an isoleucine (red and sticked) in a 2-keto-3-deoxygluconate kinase subunit interface [PDB:2DCN]. Water molecules are shown in blue spheres (see online version for colours)



3.2 Site conservation

When a residue is not substituted in a residue-residue alignment – i.e., the two residues are of the same type – we can achieve a close and detailed view of the conservation of water down to the atomic level. In Table 1, the site conservation of water contact of the atoms in different amino acids is shown. Generally, water

Table 1 Site conservation

AA	*Site conservation*
ARG	C:0.38 C^{α}:0.53 N:0.48 O:0.53 C^{β}:0.58 C^{γ}:0.56 C^{δ}:0.56 N^{ε}:0.52 C^{ζ}:0.40 $N^{\eta 1}$:0.60 $N^{\eta 2}$:0.61
LYS	C:0.30 C^{α}:0.43 N:0.47 O:0.50 C^{β}:0.46 C^{γ}:0.46 C^{δ}:0.47 C^{ε}:0.50 N^{ζ}:0.55
ASN	C:0.49 C^{α}:0.52 N:0.57 O:0.60 C^{β}:0.62 C^{γ}:0.46 $O^{\delta 1}$:0.62 $N^{\delta 2}$:0.66
ASP	C:0.40 C^{α}:0.59 N:0.55 O:0.61 C^{β}:0.62 C^{γ}:0.50 $O^{\delta 1}$:0.64 $O^{\delta 2}$:0.63
GLU	C:0.39 C^{α}:0.54 N:0.48 O:0.54 C^{β}:0.58 C^{γ}:0.55 C^{δ}:0.43 $O^{\varepsilon 1}$:0.60 $O^{\varepsilon 2}$:0.57
GLN	C:0.38 C^{α}:0.49 N:0.39 O:0.53 C^{β}:0.55 C^{γ}:0.54 C^{δ}:0.42 $O^{\varepsilon 1}$:0.59 $N^{\varepsilon 2}$:0.62
HIS	C:0.50 C^{α}:0.54 N:0.54 O:0.60 C^{β}:0.63 C^{γ}:0.38 $N^{\delta 1}$:0.49 $C^{\delta 2}$:0.50 $C^{\varepsilon 1}$:0.54 $N^{\varepsilon 2}$:0.58
PRO	C:0.42 C^{α}:0.53 N:0.46 O:0.60 C^{β}:0.59 C^{γ}:0.61 C^{δ}:0.62
TYR	C:0.45 C^{α}:0.53 N:0.53 O:0.55 C^{β}:0.54 C^{γ}:0.35 $C^{\delta 1}$:0.46 $C^{\delta 2}$:0.41 $C^{\varepsilon 1}$:0.50 $C^{\varepsilon 2}$:0.51 C^{ζ}:0.40 O^{η}:0.61
TRP	C:0.46 C^{α}:0.53 N:0.49 O:0.51 C^{β}:0.50 C^{γ}:0.31 $C^{\delta 1}$:0.51 $C^{\delta 2}$:0.38 $N^{\varepsilon 1}$:0.56 $C^{\varepsilon 2}$:0.39 $C^{\varepsilon 3}$:0.48 $C^{\zeta 2}$:0.60 $C^{\zeta 3}$:0.60 $C^{\eta 2}$:0.59
SER	C:0.44 C^{α}:0.56 N:0.61 O:0.59 C^{β}:0.63 O^{γ}:0.68
THR	C:0.40 C^{α}:0.51 N:0.61 O:0.61 C^{β}:0.56 $O^{\gamma 1}$:0.62 $C^{\gamma 2}$:0.65
GLY	C:0.52 C^{α}:0.67 N:0.62 O:0.65
ALA	C:0.42 C^{α}:0.60 N:0.61 O:0.62 C^{β}:0.70
MET	C:0.41 C^{α}:0.57 N:0.56 O:0.59 C^{β}:0.59 C^{γ}:0.50 S^{δ}:0.50 C^{ε}:0.61
CYS	C:0.39 C^{α}:0.51 N:0.51 O:0.64 C^{β}:0.57 S^{γ}:0.62
PHE	C:0.38 C^{α}:0.52 N:0.54 O:0.59 C^{β}:0.57 C^{γ}:0.34 $C^{\delta 1}$:0.45 $C^{\delta 2}$:0.43 $C^{\varepsilon 1}$:0.45 $C^{\varepsilon 2}$:0.46 C^{ζ}:0.48
LEU	C:0.38 C^{α}:0.45 N:0.50 O:0.55 C^{β}:0.51 C^{γ}:0.38 $C^{\delta 1}$:0.53 $C^{\delta 2}$:0.55
VAL	C:0.39 C^{α}:0.50 N:0.53 O:0.56 C^{β}:0.46 $C^{\gamma 1}$:0.58 $C^{\gamma 2}$:0.58
ILE	C:0.42 C^{α}:0.55 N:0.54 O:0.62 C^{β}:0.48 $C^{\gamma 1}$:0.53 $C^{\gamma 2}$:0.60 $C^{\delta 1}$:0.57

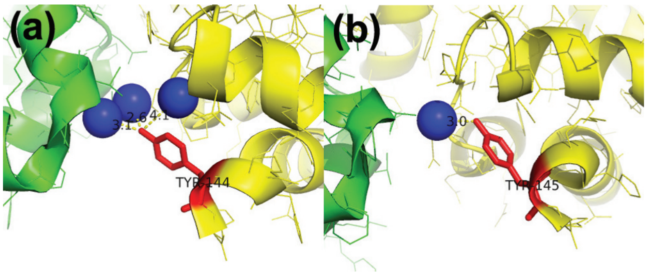
Conservation of water-contact in different sites of residues. AA: amino acid type.

contacts with polar and charged atoms are more conserved. For example, in the side chains of ARG, LYS, ASN, ASP, GLN, GLU, TYR, or SER, the polar or charged groups at the end of their side chains possess the largest water-contact conservation. This emphasises the importance of interfacial water in bridging local contacts. Specific contacts such as hydrogen bonds with water molecules can be formed by these atoms.

A pair of aligned tyrosine residues in hemoglobin subunit interface and their nearby water structures are shown in Figure 5. Both residues have interfacial water molecules, contacting with the polar atom O^{η} in a very close distance (see the numbers in the figure) at the end of the side chain. In both cases, at the back of the water molecules in the other side, a threonine is observed. Threonines and tyrosines can both donate and accept hydrogen bonds (McDonald and Thornton, 1994). The water molecules can bridge the contacts between the two residues by using hydrogen bonds.

It is interesting that the most conserved water contacts are those with C^{β} of alanine. Their site conservation is 0.70, even higher than that of any polar or charged atoms. This can be explained by another important function of interfacial water molecules: filling cavities. Residues with short side chains such as alanine may create small pockets or cavities; water molecules are needed in this case to shape

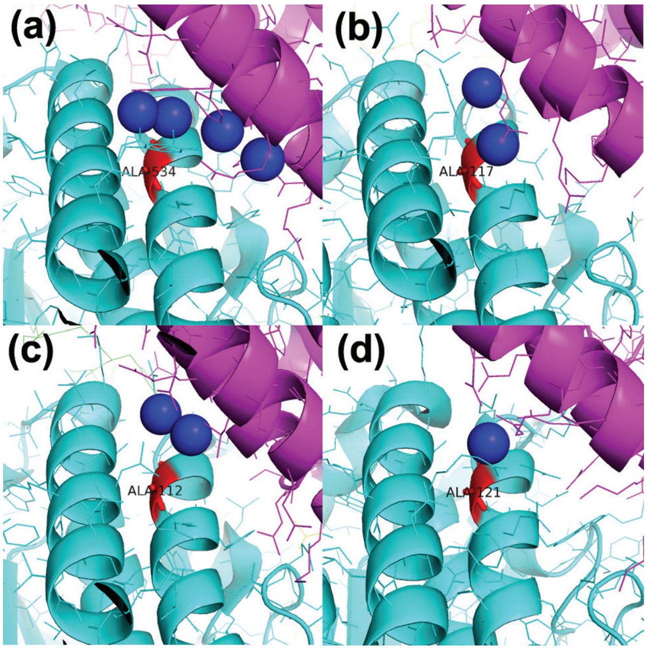
Figure 5 Two aligned tyrosines (red and sticked) in the interface between α and β subunit of hemoglobin of (a) bovine [PDB:2QSP] and (b) human [PDB:3IC0]. Water molecules directly contacting with the two tyrosine residues are shown in blue spheres (see online version for colours)



the binding site. This is also the reason that the site conservation of C$^{\alpha}$ in glycine is very high (0.67).

In Figure 6, four aligned alanines in four dehydrogenase subunit interfaces are shown. At each interface shown in the figure, two stacks of α-helices are observed, making the interface flat. Unlike their nearby residues, the four alanine residues shown in the figure do not have a long side chain and hence they cannot reach

Figure 6 Water-contacting structure of four aligned alanine residues in: a rat formyltetrahydrofolate dehydrogenase subunit interface (a, [PDB:2O2P]), and three betaine aldehyde dehydrogenase subunit interfaces (b[PDB:2WOX], c[PDB:1WNB] and d[PDB:3EK1]). The conserved alanines are shown in red, and their contacting water molecules are shown in blue spheres (see online version for colours)

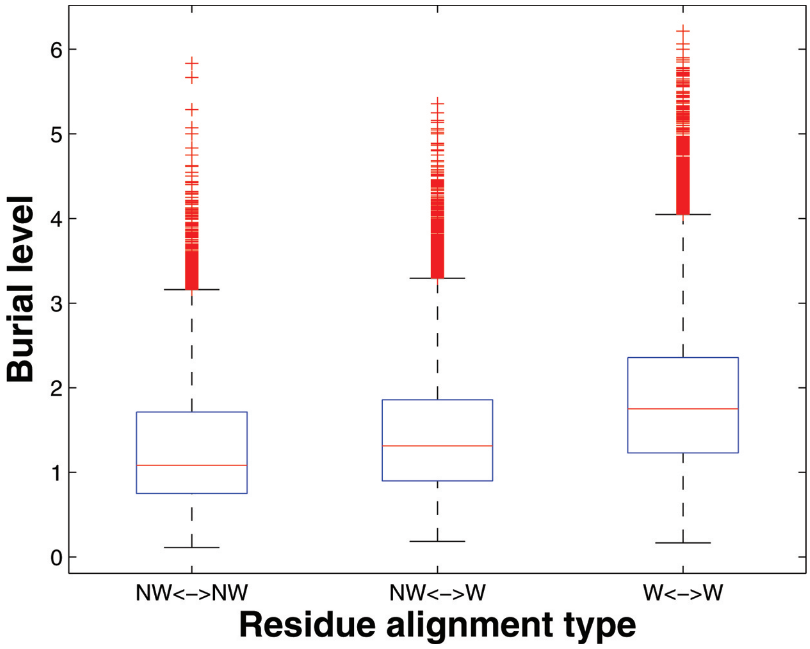


the interacting partner directly. To maintain the local secondary structure, the backbone of the alanine residues cannot be changed too much, thus a small gap is created between the two chains, where the water molecules come and fill in. These alanine residues are typically called wet spots. They contact with the other side indirectly through water molecules.

3.3 Conservation of water at different burial levels

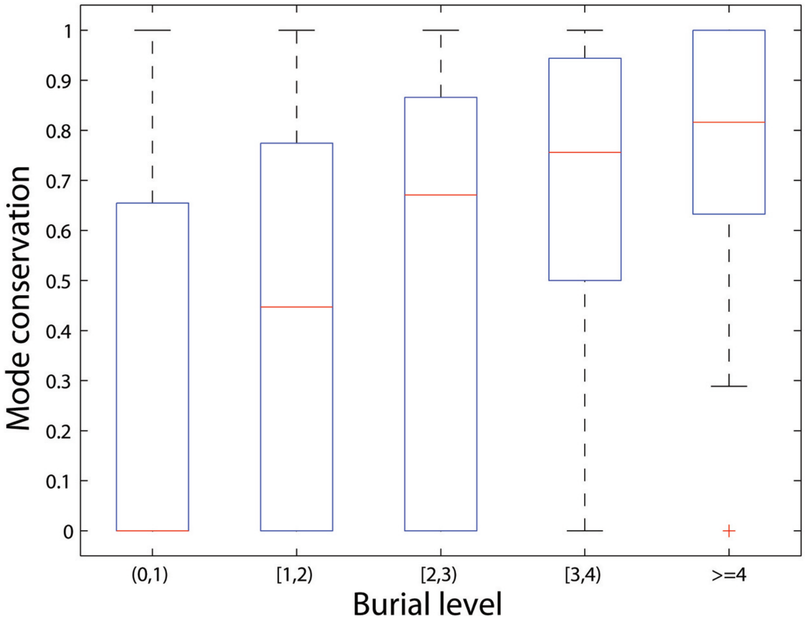
We also investigated the relationship between residue burial level and its hydration. The residue-residue alignments in our data set can be divided into three types according to the hydration of whether the two residues are water-contacting or not: non water-contacting to non water-contacting (NW↔ NW), non water-contacting to water-contacting (NW↔ W), and water-contacting to water-contacting (W↔ W). The distribution of burial level in these three types of residue alignments is shown in Figure 7. We find that residues whose contacting water is conserved are more likely to be those deeply buried ones, indicating that water molecules at the core of the protein-binding interface are more conserved.

Figure 7 Distribution of burial level (as average of the two residues) in three types of interfacial residue-residue alignments. NW↔ NW: both residues are non water-contacting; NW↔ W: only one of the residue is water-contacting and; W↔ W: both residues are water-contacting (see online version for colours)



The conservation of deeply buried water can also be captured by analysing the conservation of water-contacting modes (see Methods) in alignments of identical residue type. The relationship between water-contacting mode conservation and burial level is shown in Figure 8. Deeply buried residues always have high water-contacting conservation. This again suggests that water molecules in the core of the interface are harder to be changed, even in the atoms they are contacting to.

Figure 8 Relationship between water-contacting mode conservation and burial level (as average of the two aligned residues) (see online version for colours)



Immobilised water molecules in PDB structures tend to be under reported, especially buried water molecules near the surface. The quality of water information is also usually correlated with resolution. However, our observation of high conservation of deeply buried water is not due to the uneven quality of deeply and slightly buried water molecules at different resolutions. To confirm this, we repeated the experiment using a higher resolution threshold to process our data set. We calculated the average of the burial level in the three types of residue-residue alignment, using a different resolution threshold of 2.0 Å. The comparison with the results under the original threshold 3.0 Å is shown in Table 2. The difference (tested by Mann-Whitney U test (Mann and Whitney, 1947)) is generally not significant,

Table 2 Influence of resolution threshold used to process the data

Resolution threshold	$BL(NW \leftrightarrow NW)$	$BL(NW \leftrightarrow W)$	$BL(W \leftrightarrow W)$
$\leq$ 3.0Å	1.3006	1.4474	1.8552
$\leq$ 2.0Å	1.3490	1.4500	1.8748
p-value	0.0036	0.0887	0.1680

Comparing the average burial level in the three types of residue alignments under different resolution threshold. NW ↔ NW: both residues are non water-contacting; NW ↔ W: only one of the residue is water-contacting and W ↔ W: both residues are water-contacting.

especially in the case when the water-contacting status is conserved (W↔ W), where the p-value is 0.1680.

Protein interface is usually wetter in the rim and drier in the core, which means that interface rim is more crowded with water molecules. Thus, if water molecules are just randomly choosing their binding residues level by level, the water-contacting status of residues at the rim should be more conserved, due to the higher probability of being in contact with water in the rim. However, the observation here suggests that water molecules at the core are more conserved. Therefore, it can be inferred that deeply buried water molecules are organised more stably in local positions with specific local contacts so that they are harder to be changed.

4 Conclusions

Based on an idea of solvated sequence, a conservation analysis of the hydration of residues is studied. We found that, for a residue that is in contact with water, substituting it into a hydrophobic residue tends to drop its contact with water. A detailed atomic level analysis reveals that polar and charged sites of residues are better at preserving their contacting water molecules. Also, short-side-chained residues are not likely to lose their contacting water molecules, as long as they are not substituted by other residues. Another finding of this work is that deeply burieds residues have very stable local water environment, indicating that water molecules deeply buried in the interface are more conserved.

Acknowledgements

This research work was funded in part by two Singapore MOE Tier-2 grants (T208B2203 and MOE2009-T2-2-004).

References

Ahmed, M.H., Spyrakis, F., Cozzini, P., Tripathi, P.K., Mozzarelli, A., Scarsdale, J.N., Safo, M.A. and Kellogg, G.E. (2011) 'Bound water at protein-protein interfaces: partners, roles and hydrophobic bubbles as a conserved motif', *PLoS ONE*, Vol. 6, No. 9, pp.e24712.

Bairagya, H.R., Mukhopadhyay, B.P. and Bhattacharya, S. (2009) 'Role of the conserved water molecules in the binding of inhibitor to IMPDH-II (human): a study on the water mimic inhibitor design', *Journal of Molecular Structure: THEOCHEM*, Vol. 908, Nos. 1–3, pp.31–39.

Benedix, A., Becker, C.M., de Groot, B.L., Caflisch, A. and Bockmann, R.A. (2009) 'Predicting free energy changes using structural ensembles', *Nature Methods*, Vol. 6, No. 1, pp.3, 4.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) 'The protein data bank', *Nucleic Acids Res.*, Vol. 28, No. 1, pp.235–242.

Bottoms, C.A., Smith, P.E. and Tanner, J.J. (2002) 'A structurally conserved water molecule in rossmann dinucleotide-binding domains', *Protein Science*, Vol. 11, No. 9, pp.2125–2137.

Carugo, O. and Bordo, D. (1999) 'How many water molecules can be detected by protein crystallography?', *Acta Crystallogr. D Biol. Crystallogr.*, Vol. 55, No. Pt 2, pp.479–483.

Gao, M. and Skolnick, J. (2010a) 'iAlign: a method for the structural comparison of protein-protein interfaces', *Bioinformatics*, Vol. 26, No. 18, pp.2259–2265.

Gao, M. and Skolnick, J. (2010b) 'Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected', *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 107, No. 52, pp.22517–22522.

Guerois, R., Nielsen, J. E. and Serrano, L. (2002) 'Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations', *J. Mol. Biol.*, Vol. 320, No. 2, pp.369–387.

Ikura, T., Urakubo, Y. and Ito, N. (2004) 'Water-mediated interaction at a protein-protein interface', *Chem. Phys.*, Vol. 307, Nos. 2–3, pp.111–119.

Jackson, R.M., Gabb, H.A. and Sternberg, M.J.E. (1998) 'Rapid refinement of protein interfaces incorporating solvation: application to the docking problem1', *Journal of Molecular Biology*, Vol. 276, No. 1, pp.265–285.

Kinjo, A.R. and Nakamura, H. (2010) 'Geometric similarities of protein-protein interfaces at atomic resolution are only observed within homologous families: an exhaustive structural classification study', *J. Mol. Biol.*, Vol. 399, No. 3, pp.526–540.

Kinjo, A.R. and Nakamura, H. (2009) 'Comprehensive structural classification of ligand-binding motifs in proteins', *Structure*, Vol. 17, pp.234–246.

Knight, J.D.R., Hamelberg, D., Mccammon, A.J. and Kothary, R. (2009) 'The role of conserved water molecules in the catalytic domain of protein kinases', *Proteins: Structure, Function, and Bioinformatics*, Vol. 76, pp.527–535.

Kyte, J. and Doolittle, R. (1982) 'A simple method for displaying the hydropathic character of a protein', *J. Mol. Biol.*, Vol. 157, pp.105–132.

Li, Z. and Lazaridis, T. (2007) 'Water at biomolecular binding interfaces', *Phys. Chem. Chem. Phys.*, Vol. 9, No. 5, pp.573–581.

Li, Z. and Li, J. (2010) 'Geometrically centered region: A "wet" model of protein binding hot spots not excluding water molecules', *Proteins Struct. Funct. Bioinf.*, Vol. 78, No. 16, pp.3304–3316.

Lu, Y., Yang, C.Y. and Wang, S. (2006) 'Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes', *J. Am. Chem. Soc.*, Vol. 128, No. 36, pp.11830–11839.

Mann, H.B. and Whitney, D.R. (1947) 'On a test of whether one of two random variables is stochastically larger than the other', *The Annals of Mathematical Statistics*, Vol. 18, No. 1, pp.50–60.

McDonald, I.K. and Thornton, J.M. (1994) 'Satisfying hydrogen bonding potential in proteins', *J. Mol. Biol.*, Vol. 238, No. 5, pp.777–793.

Meenan, N.A.G., Sharma, A., Fleishman, S.J., MacDonald, C.J., Morel, B., Boetzel, R., Moore, G.R., Baker, D. and Kleanthous, C. (2010) 'The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction', *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 107, No. 22, pp.10080–10085.

Poornima, C.S. and Dean, P.M. (1995) 'Hydration in drug design. 3. conserved water molecules at the ligand-binding sites of homologous proteins', *Journal of Computer-Aided Molecular Design*, Vol. 9, No. 6, pp.521–531.

Reichmann, D., Phillip, Y., Carmi, A. and Schreiber, G. (2008) 'On the contribution of water-mediated interactions to protein-complex stability', *Biochemistry*, Vol. 47, No. 3, pp.1051–1060.

Rodier, F., Bahadur, R.P., Chakrabarti, P. and Janin, J. (2005) 'Hydration of protein-protein interfaces', *Proteins Struct. Funct. Bioinf.*, Vol. 60, No. 1, pp.36–45.

Shaltiel, S., Cox, S. and Taylor, S.S. (1998) 'Conserved water molecules contribute to the extensive network of interactions at the active site of protein kinase?a', *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 95, No. 2, pp.484–491.

Sharp, K.A., Nicholls, A., Fine, R.F. and Honig, B. (1991) 'Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects', *Science*, Vol. 252, No. 5002, pp.106–109.

Urakubo, Y., Ikura, T. and Ito, N. (2006) 'Crystal structural analysis of protein-protein interactions drastically destabilized by a single mutation', *Protein Science*, Vol. 17, pp.1055–1065.

Wilson, D.K. (1989) 'Substrate specificity and affinity of a protein modulated by bound water molecules', *Nature*, Vol. 304, pp.404–407.