BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 6 2009, pages 743–750
doi:10.1093/bioinformatics/btp058

Structural bioinformatics

'Double water exclusion': a hypothesis refining the O-ring theory for the hot spots at protein interfaces

Jinyan Li* and Qian Liu
Bioinformatics Research Center, School of Computer Engineering, Nanyang Technological University, Singapore 639798

Received on November 03, 2008; revised on January 02, 2009; accepted on January 23, 2009
Advance Access publication January 29, 2009
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: The O-ring theory reveals that the binding hot spot at a protein interface is surrounded by a ring of residues that are energetically less important than the residues in the hot spot. As this ring of residues is served to occlude water molecules from the hot spot, the O-ring theory is also called 'water exclusion' hypothesis. We propose a 'double water exclusion' hypothesis to refine the O-ring theory by assuming the hot spot itself is water-free. To computationally model a water-free hot spot, we use a *biclique pattern* that is defined as two *maximal* groups of residues from two chains in a protein complex holding the property that every residue contacts with all residues in the other group.

Methods and Results: Given a chain pair A and B of a protein complex from the Protein Data Bank (PDB), we calculate the interatomic distance of all possible pairs of atoms between A and B. We then represent A and B as a bipartite graph based on these distance information. Maximal biclique subgraphs are subsequently identified from all of the bipartite graphs to locate biclique patterns at the interfaces. We address two properties of biclique patterns: a non-redundant occurrence in PDB, and a correspondence with hot spots when the solvent-accessible surface area (SASA) of a biclique pattern in the complex form is small. A total of 1293 biclique patterns are discovered which have a non-redundant occurrence of at least five, and which each have a minimum two and four residues at the two sides. Through extensive queries to the HotSprint and ASEdb databases, we verified that biclique patterns are rich of true hot residues. Our algorithm and results provide a new way to identify hot spots by examining proteins' structural data.

Availability: The biclique mining algorithm is available at http://www.ntu.edu.sg/home/jyli/dwe.html.

Contact: jyli@ntu.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Anatomy, characterization and statistical analysis of protein–protein interfaces have been broadly and extensively studied under the principles (Chothia and Janin, 1975; Jones and Thornton, 1996) of protein–protein recognition and interaction. A protein binding interface composes of two relatively large, spatially close protein surfaces with good geometric shape and chemical complementarity. The formation of protein chain interfaces is driven by various natural forces such as van der Waals contacts and electrostatic interactions, resulting in the removal of water molecules from the binding sites (Fernandez and Scott, 2003; Privalov *et al.*, 2007). This water-removal-then-binding has been well characterized by the influential 'O'-ring theory (Bogan and Thorn, 1998; Chakrabarti and Janin, 2002; DeLano, 2002b; Moreira *et al.*, 2007; Thorn and Bogan, 2001), which is also called 'water exclusion' hypothesis. It highlights that the stability of a complex is determined by only a small number of energetically outstanding residues; it also reveals that these 'hot-spot' residues are usually located at the center of the interface and surrounded by energetically less important residues that shape like an O-ring to occlude bulk water molecules from the hot spot. A lab-verified example of the O-ring theory was reported in the pioneering work (Clackson and Wells, 1995) in which alanine-scanning mutagenesis was used to determine the binding hot spots between human growth hormones and human growth hormone-binding proteins. The O-ring theory is profounding. However, the organizational topology of the ring-inside, energetically more important hot residues is uncertain and not specified by the O-ring theory.

To refine this long-standing theory, we suggest a 'double water exclusion' hypothesis to characterize the topological organization of residues in a hot spot and their neighboring residues. On one side, we agree with the O-ring theory that there should exist a ring of residues surrounding the hot spot for avoiding the invasion of water molecules after the complex formation; on the other hand, we suppose that the hot spot itself is water-free. The water-free assumption may be too strict, so, our another hypothesis is whether the amount of water molecules inside a hot spot is proportional to its phylogenetic evolution progression toward to the perfect water-free binding. In this work, we focus on the investigation of the 'double water exclusion' hypothesis only, while the latter (and bigger) hypothesis is left as an open question for interested readers and ourselves in future research.

We propose to use contact residues that are *densely* interacted in a compact region to model a hot spot. [Here, a pair of residues contact to each other if there exists a pair of atoms whose distance is below the sum of their corresponding van der Waals radii plus the diameter (2.75 Å) of a water molecule.] We take the classical graph term 'biclique' (Eppstein, 1994) to denote this molecular topology, and call it 'biclique pattern'. Formally, a biclique pattern between two chains in a protein complex consists of two *maximal* groups of residues (one from each chain) such that every residue contacts with

*To whom correspondence should be addressed.

all residues in the other group. Thus, if a biclique pattern is observed from a protein complex, then all possible pairs of the residues from the two sides are spatially so close that there is no sufficient room between them to accommodate any water molecule. That is, this is an interface structure exhibiting a zero-water tolerance property and forming a collective force of multiple, dense atom-atom pairs. Such a structure is therefore useful for lowering the local dielectric constant and enhancing specific electrostatic and hydrogen bond interactions to strengthen the stability of the binding. On the other hand, by the maximality requirement of biclique pattern, any neighbor residue at one side of the pattern disassociates with at least one residue at the other side. This makes the neighbor residues flexible to form a fence preventing the biclique pattern residues from solvent partially or entirely. Thus, if every residue in a biclique pattern has a small solvent accessible surface area (SASA) in the complex, then this biclique pattern reflects best the spirit of 'double water exclusion'—all the inner hot residues are organized as a biclique without holding any water molecule, while there exist a ring of residues blocking the solvent accessibility to the hot spot.

Our biclique pattern concept integrates ideas not only from the O-ring theory, but also from the 'coupling' hypothesis (Halperin *et al.*, 2004) which is another insightful proposition about hot-spot residues. As said in the O-ring theory, energetically important hot residues are often clustered, and usually located at the center of lesser important residues. While, the 'coupling' hypothesis bridges the relation between hot residues from the two sides of an interface, stating that experimental hot residues tend to couple across a two-chain interface. Combining these two observations, it can be seen that the central hot residues of an O-ring structure are 'coupled and interacting' across the interface. Our biclique pattern concept exactly refines this hybrid idea by specifying that the coupled and interacting hot residues are a maximal cluster of residues that have a full 'water exclusion' adjacency.

The biclique pattern concept can also be supported by the influential work by Keskin *et al.* (2005), where the organization of structurally conserved hot-spot residues has been studied. Keskin *et al.* (2005) used the term 'hot region' to describe assemblies of hot residues that are located within densely packed regions with a network of interactions; their analysis over 44 interface clusters with 568 hot-spot residues shows that 79% of the hot residues were found to be in the hot regions. By definition, it can be seen that the all-versus-all full adjacency required by a biclique pattern is the most optimal network of interactions for a region having two groups of residues. Therefore, our biclique pattern concept theoretically strengthens and signifies the hot region proposition. The maximality requirement also guarantees that the size of a biclique pattern is as large as possible, increasing the binding stability as much as possible. These enhancements and consistency suggest that our theoretical top-down 'biclique' approach meets the empirical bottom-up 'region' approach at a point where both demonstrate a spatially well-organized topology of hot-spot residues.

This work is also related to previous studies on contact residues and binding interfaces. Wolfson, Nussinov and their coworkers (Halperin *et al.*, 2004; Keskin *et al.*, 2004, 2005; Li *et al.*, 2004; Mintz *et al.*, 2005; Tsai *et al.*, 1996, 1997) defined a protein interface as two parts: contact residues and the neighbor residues of the contact residues. A pair of residues (one from each chain) is defined as 'contacting' across the interface if the distance between any two atoms of the two residues is less than the sum of their corresponding van der Waals radii plus 0.5 Å; a residue from the same chain of a contact residue is defined to be a 'nearby' residue of this contact residue if the distance between their two C^{α} atoms is <6 Å. This category of definitions also include several variations as proposed in Davis and Sali (2005); Gong *et al.* (2005); Korkin *et al.* (2005); Larsen *et al.* (1998); Lawrence and Colman (1993); Ofran and Rost (2003); Preissner *et al.* (1998). The key difference from these existing work is that we investigate clusters of contact residues that are densely interacted like a biclique in a compact region where the compactness is measured by the water exclusion principle. Besides, we use the small complex SASA constraint to narrow down the scope of biclique patterns, so that the biclique patterns have a good correspondence to hot spots, and have the best spirit of the 'double water exclusion' hypothesis.

We address two properties of biclique patterns: (i) their occurrence and statistical significance in the protein data bank (PDB) protein complexes, and (ii) their relation with verified hot spots in literature. The former is an important indicator to see whether biclique patterns are statistically occurring in protein–protein interactions significantly; while the latter is an important evaluation way to see whether our 'double water exclusion' truly refines the O-ring theory. In our method, we sum up a non-redundant occurrence in the PDB complexes for every biclique pattern. We also break the sum into obligate and transient interactions. This distinction is meaningful because obligate and transient interactions are characterized by distinct physico-chemical properties (Mintseris and Weng, 2005; Ofran and Rost, 2003; Sprinzak *et al.*, 2006). Therefore, it is possible that a biclique pattern occurs in only obligate interactions, or in only transient interactions, or in both types of interactions. If a biclique pattern occurs frequently, we calculate its complex SASA by using the NACCESS (Hubbard and Thornton, 1993) software (http://www.bioinf.manchester.ac.uk/naccess). For those with a small SASA, we believe that they are likely to be a hot spot.

2 DATA AND METHOD

We downloaded X-ray crystallographic protein structures having resolution better than 2.5 Å from a March 2008 version of PDB (http://www.rcsb.org). We considered only PDB entries with two or more polypeptide chains that each have more than 30 amino acids. In total, we obtained 17 248 PDB complexes for this study.

It is computationally challenging to efficiently identify biclique patterns due to the maximality constraint and the all-versus-all adjacency. In this work, we take an efficient approach to the localization of biclique patterns at every pair of interacting protein chains. Our algorithm consists of three main computational steps: (i) construction of bipartite graphs based on the residues' 3D-coordinate information of interacting chain pairs within protein complexes, (ii) discovery of maximal bicliques from every bipartite graph and (iii) calculation of exact non-redundant occurrence for every maximal biclique that meets a size threshold.

Step 1: a bipartite graph G is a graph whose vertices can be divided into two disjoint sets V_1 and V_2 such that no two vertices within V_1 or within V_2 are adjacent (Asratian *et al.*, 1998). It is usually denoted by $G=(V_1, V_2, E)$, where E is the set of edges of G. To transform an interacting chain pair into a bipartite graph, we represent every residue as a vertex, and an edge is assigned between a residue x_1 in one chain and a residue x_2 in the other chain if and only if there exists a pair of atoms between x_1 and x_2 whose distance is less than a threshold. In this study, we set the threshold as the sum of the corresponding van der Waals radii plus the diameter (2.75 Å) of a water molecule.

Step 2: for a bipartite graph $G=(V_1,V_2,E)$ representing a pair of interacting polypeptide chains, we identify the complete set of maximal biclique subgraphs from G. A biclique subgraph H of G is a graph consisting of two sets of vertices $X_1 \subseteq V_1$ and $X_2 \subseteq V_2$ such that every vertex in X_i is adjacent to all vertices in X_j where $j \neq i$. A biclique subgraph H is maximal in G if there is no other biclique in G that contains H. Observe that this maximal all-versus-all adjacency emulates the biological 'water-exclusion' hypothesis very well. Maximal biclique subgraphs are identified through our earlier Linear time Closed pattern Mining algorithm for Maximal BiClique subgraphs (LCM-MBC) algorithm (Li *et al.*, 2007). The LCM-MBC algorithm has two input parameters p and q that can control the minimum number of vertices in each side of a maximal biclique.

Step 3: we identify bicliques that occur in protein complexes with a high non-redundant occurrence; while bicliques that occur in only a few interacting chain pairs are not of our primary interests. By 'occur', we mean that not only all the residues of the biclique are matched, but also the biclique structure is maintained. In other words, two biclique patterns, $P_1=(Y_1,Y_2,E_1)$ and $P_2=(Y_3,Y_4,E_2)$, are considered as the same if and only if $Y_1=Y_3$, $Y_2=Y_4$ and $E_1=E_2$ (or $Y_1=Y_4$, $Y_2=Y_3$ and $E_1=E_2$). Assume m number of chain pairs are used, written as CHAINPAIRS$=\{chainPair^{(i)}=(SR_1^{(i)},SR_2^{(i)}) \mid i=1,2,\ldots,m\}$, where $SR_1^{(i)}$ and $SR_2^{(i)}$ represent the string of residues for the two chains. Transform every $chainPair^{(i)}=(SR_1^{(i)},SR_2^{(i)})$ into a bipartite graph $G^{(i)}=(V_1^{(i)},V_2^{(i)},E^{(i)})$. Let BiC be the set of all maximal bicliques in these bipartite graphs $G^{(i)}$, $i=1,2,\ldots,m$. For a maximal biclique $H^{(j)}=(X_1^{(j)},X_2^{(j)}) \in BiC$, its occurrence is determined as follows. We go through $\{G^{(i)} \mid i=1,2,\ldots,m\}$ to count the number of those containing $H^{(j)}$; if the number exceeds a pre-defined threshold *sup*, then $H^{(j)}$ is a maximal biclique subgraph that corresponds to a biclique pattern of significance. To guarantee a non-redundant occurrence for $H^{(j)}$, actually, we check whether those $chainPair^{(i)}$ containing $H^{(j)}$ have a high BLAST similarity to each other. We remove the redundant chair pairs in the final occurrence counting for $H^{(j)}$. Here, $chainPair^{(a)}$ and $chainPair^{(b)}$ are redundant to each other if $similarity(SR_1^{(a)},SR_1^{(b)}) \geq 90\%$ and $similarity(SR_2^{(a)},SR_2^{(b)}) \geq 90\%$, or $similarity(SR_1^{(a)},SR_2^{(b)}) \geq 90\%$ and $similarity(SR_2^{(a)},SR_1^{(b)}) \geq 90\%$. The protein similarity score was calculated under the default parameter setting of the BLAST software that was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml). The BLAST database is from the ftp site ftp://ftp.ncbi.nlm.nih.gov/blast/db/.

Our algorithm is termed as `biBonder`. Its pseudo code is shown in Algorithm 1. This algorithm was implemented in Python and our experiments were conducted in Linux environment with CPU of 2589.845 MHz and RAM of 31 GB. The program takes ~0.375 s on average to locate all maximal bicliques from a pair of interacting protein chains.

Algorithm 1 Algorithm `biBonder`

Input:

CHAINPAIRS $=\{chainPair^{(i)}, i=1,2,\ldots,m\}$;

p is the size threshold for one side of a biclique pattern, q the size threshold for the other side; *sup* is the minimum non-redundant occurrence;

Description:

1: convert $\{chainPair^{(i)}, i=1,2,\ldots,m\}$ into a set of bipartite graphs $\{G^{(i)}, i=1,2,\ldots,m\}$;
2: use LCM-MBC to mine maximal biclique subgraphs from all $G^{(i)}$. Let BiC be the set of all maximal biclique subgraphs discovered;
3: **for all** $H^{(j)} \in BiC$ **do**
4: $count=0$;
5: **for all** $G^{(i)}$ **do**
6: **if** $H^{(j)}$ is a subgraph of $G^{(i)}$ **then** $count$++;
7: remove the redundancy and update $count$;
8: **if** $count \geq sup$ **then** $H^{(j)}$ is a highly conserved biclique pattern;
9: output all highly conserved biclique patterns;

We note that maximal biclique or its generalized form quasi-biclique subgraph has recently emerged in bioinformatics studies for the characterization of protein–protein interactions and their networks (Li *et al.*, 2006; Morrison *et al.*, 2006; Suryani *et al.*, 2008). Here, we have explored a novel use of bicliques in structural bioinformatics.

In our method, one important post-processing step is a statistical evaluation of these biclique patterns and the identification of obligate and transient interactions that contain a given biclique pattern. We use the highly accurate NOXclass algorithm (Zhu *et al.*, 2006) to differentiate between obligate and transient interactions, and also distinguish crystal packing.

3 OCCURRENCE AND STATISTICAL SIGNIFICANCE OF BICLIQUE PATTERNS

There are many examples of hormones that bind multiple receptors, or receptors that bind multiple hormones (DeLano *et al.*, 2000) where their 'consensus' sites have been shown useful to understand biological functions. Similarly to those consensus information, the occurrence information of a biclique pattern in many interacting protein–protein interfaces is of our interest. This information can be obtained by using the `biBonder` algorithm. As shown in its pseudo code, three parameters are required to specify, namely, the minimum number of residues in one side of a biclique pattern (i.e. p), the minimum number of residues in the other side of the biclique pattern (i.e. q), and the minimum occurrence in the pairs of interacting protein chains (i.e. *sup*).

Many choices are available for these parameters. Here, we present in Table 1 the numbers of biclique patterns when we set $p=2$, $q=4$ and $sup=5$. Column 1 of this table specifies the size of biclique patterns where 'x residues' means that *exactly* x residues are in one side of the biclique pattern, and 'y residues' means exactly y residues in the other side; column 2 shows the total number of biclique patterns in each size category; column 3 the maximum non-redundant occurrence of the biclique patterns in each category; columns 4–6 the average (and maximum, minimum) complex SASA per residue of the residues in each category. In total, we identified 1293 biclique patterns which have a non-redundant occurrence of at least five and which each have a minimum two residues in one side and minimum four residues at the other side. Most of these biclique patterns contain uneven numbers of residues at their two sides, in particular those consisting of only two residues in one side. There are no biclique patterns whose total number of residues exceeds 10.

Table 1. Biclique patterns of different sizes, all with a minimum non-redundant occurrence 5 ($sup=5$), their total number for each size category, the maximum occurrence and the complex SASA information

Size of a biclique pattern (x resides versus y residues)	Total number	Max occ.	Complex SASA per residue		
			Ave.	Max.	Min.
2–4	916	87	15.66	75.62	0.33
2–5	173	36	17.79	55.39	0.64
2–6	53	31	15.07	48.87	1.73
2–7	21	30	18.09	45.46	1.91
2–8	5	16	9.64	15.57	1.83
3–4	93	33	17.02	47.05	0.42
3–5	12	7	19.38	43.12	3.48
3–6	5	8	11.93	24.27	7.38
3–7	2	7	15.49	22.23	8.75
4–4	12	8	12.57	27.72	5.13
4–5	1	5	7.34	7.34	7.34

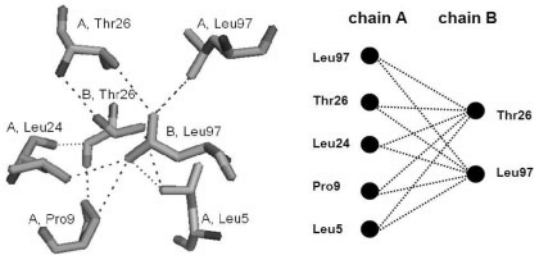


Fig. 1. A biclique pattern identified in the PDB protein complex 1A8G that also occurs in other four PDB complexes. Dotted lines between two residues indicate that their distance in 3D space is not larger than the sum of their corresponding van der Waals radii plus the diameter of a water molecule.

We describe in detail an example of biclique pattern using Figure 1. This biclique pattern has a non-reduntant occurrence of four in obligate interactions, and a non-redundant occurrence of one in transient interactions. Specifically, this pattern occurs in the interaction between chain A of PDB 1A8G at Leu5-Pro9-Leu24-Thr26-Leu97, and chain B at Thr26-Leu97. See the left panel of Figure 1, where a dotted line represents that the distance between some two atoms of the two residues is not larger than the sum of their corresponding van der Waals radii plus the diameter of a water molecule; the right panel of this figure shows a virtual graph representation of this biclique pattern used in our computational algorithm. One interesting observation is that the residues can be far away in sequence, but they are close to each other in 3D space, exhibiting a full 'water exclusion' adjacency.

This pattern also occurs in the obligate interactions between chain A and B of 1CWQ, between chain A and B of 1E0P and between chain A and B of 6UPJ; and in the transient interaction between chain A and B of PDB 1HVH. Note that these five protein complexes are diverse: 1A8G is about 'HIV-1 protease in complex with SDZ283-910'; 1CWQ is about 'M intermediate structure of the wild type Bacteriorhodopsin in combination with the ground state structure'; 1E0P is about 'L intermediate of Bacteriorhodopsin'; 6UPJ is about 'HIV-2 protease/U99294 complex'; while 1HVH is about 'non-peptide cyclic Cyanoguanidines as HIV protease inhibitors'. By our definition and algorithm, the geometric arrangement of the residues and the shape of the bicliques are similar across the five chain pairs. We also found that the global sequence positions and the relative positions between the constituent residues of this biclique are different from one PDB entry to another as detailed below:

PDB ID (no. of residues in the two chains)	Residues of the biclique	
	Chain A	Chain B
1A8G (99–99)	LEU5-PRO9-LEU24-THR26-LEU97	THR26-LEU97
1CWQ (248–248)	LEU92-LEU93-LEU94-THR89-PRO91	THR90-LEU93
1E0P (228–228)	LEU92-LEU93-LEU94-THR90-PRO91	THR89-LEU94
6UPJ (99–99)	LEU24-LEU5-LEU97-PRO9-THR26	LEU97-THR26
1HVH (99–99)	LEU5-PRO9-LEU24-THR26-LEU97	THR26-LEU97

Let H be a biclique pattern in a bipartite graph G, we assess a statistical significance ρ of H in G as a ratio between the observed occurrence of H in G over its expected occurrence, written as

$$\rho(H,G)=\frac{occ^{(H,G)}_{observed}}{occ^{(H,G)}_{expected}}.$$

Let $H=(X_1,X_2)$ be a biclique pattern contained in a bipartite graph $G=(V_1,V_2,E)$. To determine $occ^{(H,G)}_{expected}$, we assume that the probability $p(r)$ of every residue r locating at a position in a protein is estimated by the frequency of r in the Swiss-Prot database. Let H' be a random biclique with the same size and structure as H. Then, the probability of $H'=H$ is $\prod_{r\in X_1}p(r)*\prod_{r\in X_2}p(r)$. Suppose G contains n number of bicliques having the same size and structure as H, thus, $occ^{(H,G)}_{expected}$ equals to $p(H'=H)*n$.

As an example, the statistical significance of the biclique pattern $H=$(LEU-PRO-LEU-THR-LEU, THR-LEU) of Figure 1 is as high as 3.0×10^6, 4.5×10^3, 5.2×10^3, 3.9×10^6, and 1.9×10^6, respectively, in the five chair pairs 1A8Gab, 1CWQab, 1E0Pab, 6UPJab, and 1HVHab. The significance difference is attributed to the different number n of H' in the 5 chain pairs. If a biclique pattern occurs in multiple chain pairs in PDB, in this case, we take the average as the significance of this pattern in PDB. We found that: the 916 biclique patterns of the size '2–4' in Table 1 have an average significance value of 1.38×10^7 with the minimum of 33.4 and the maximum of 2.6×10^9; and the 12 biclique patterns of the size '4–4' have an average significance value of 3.38×10^9 with the minimum of 153909.3 and maximum of 1.1×10^{10}.

4 BICLIQUE PATTERNS ARE RICH OF HOTSPOT RESIDUES

A hot spot was defined as a cluster of residues that are energetically important in protein complex formation (Clackson and Wells, 1995). A hot spot typically constitutes only a small subset of interfacial residues that are buried in the middle of the interface. As hot spots contribute most to the binding affinity and strength of protein interactions, understanding hotspot residues is a fundamental problem in molecular biology [see two surveys DeLano (2002b); Moreira *et al.* (2007)]. Alanine-scanning mutagenesis (Clackson and Wells, 1995) is a widely used experimental method to determine whether a residue is in a hot spot. A residue is in a hot spot if its free energy of binding is significantly changed ($\Delta\Delta G\geq$ 2 kcal/mol) upon the mutation to alanine. ASEdb [Alanine Scanning Energetics database (Thorn and Bogan, 2001)] is a searchable database containing hotspot residues that have been verified by the experiment of alanine-scanning mutagenesis. Although this is a small database due to experimental difficulties (very slow and labor-intensive), the data quality is high.

To evaluate our hypothesis, we make queries first to the ASEdb database to test whether our biclique pattern residues are hotspot residues, and then to the HotSprint (Guney *et al.*, 2008) database which is the latest database storing *computational* hot spots.

The ASEdb database contains 28 protein complexes. However, only 13 of them are matched with PDB entries: 1a4y, 1ahw, 1brs, 1bxi, 1cbw, 1dan, 1dvf, 1gc1, 1jck, 1vfb, 2ptc, 2hfm and 3hhr. The other 15 protein complexes' structure information is not available from PDB, or not well matched (two of them). This pre-processing result is consistent with a literature work (Gao *et al.*, 2004). In these 13 protein complexes, 439 alanine-mutated residues have been

Table 2. Numbers of *very warm* and *hot* residues of the 13 protein complexes stored in ASEdb in comparison to those contained in our biclique patterns

PDB	$\Delta\Delta G \geq 1.5$ kcal/mol		$\Delta\Delta G \geq 2.0$ kcal/mol	
	ASEdb	Biclique	ASEdb	Biclique
1A4Y	4	4	3	3
1AHW	1	1	1	1
1BRS	10	9	9	8
1BXI	7	7	6	6
1CBW	1	1	1	1
1DAN	6	6	3	3
1DVF	19	18	8	8
1GC1	1	0	0	0
1JCK	6	5	4	4
1VFB	6	5	3	3
2PTC	1	1	1	1
3HFM	6	5	5	4
3HHR	14	11	8	7
Total	82	73	52	49
sensitivity		73/82 = 89%		49/52 = 94%

The larger $\Delta\Delta G$ the residues are, more likely those are in our biclique patterns.

tested, including 268 residues whose $\Delta\Delta G < 0.5$, 89 residues whose $\Delta\Delta G$ is between 0.5 and 1.5 kcal/mol (called *warm* residues), 30 residues whose $\Delta\Delta G$ is between 1.5 and 2.0 kcal/mol (called *very warm* residues) and 52 residues whose $\Delta\Delta G \geq 2$ kcal/mol (called *hot* residues). Our biclique patterns contain a total of 576 residues that are identified from the same 13 protein complexes by using Algorithm 1, under the setting $p = q = 2$, $sup = 1$ and SASA threshold is 36%. (We note that under this setting the biclique residues constitute around 10% of the total residues in a chain pair on average over the 13 protein complexes.)

Of the 52 hot residues of ASEdb, 94.2% of them (49) are also our biclique residues; of the 30 very warm residues of ASEdb, 80% of them (24) are also our biclique residues; of the 89 warm residues of ASEdb, 64% of them (57) are covered by our biclique residues; and of the 268 residues whose $\Delta\Delta G < 0.5$, 31.7% of them (85) are contained in our biclique residues. Therefore, we can note that lab-verified *hot* residues are almost all (49 out of 52) contained in our biclique patterns; thus we can conjecture that almost all hot residues of ASEdb satisfy the property of double water exclusion. When residues are energetically becoming less important from the range *very warm* to *warm* and to the range $\Delta\Delta G < 0.5$, their likelihood to be contained in our biclique patterns becomes smaller and smaller from 80 to 64% and to 31.7. Therefore, we can also conjecture that our biclique patterns are less likely to contain energetically less important residues. Table 2 details the break down of this result for the 13 protein complexes.

We are unable to determine the accuracy of our biclique residues that are also true hot residues, as there are 361 out of our 576 biclique residues that have not been tested by the alanine-scanning mutagenesis—the 361 biclique residues may be or may not be true hot residues. However, we have at least verified based on ASEdb that if a residue is a hot residue, then it can be covered by our biclique pattern. That means, double water exclusion is a necessary condition for a true hot residue. However, our biclique patterns also contain some energetically less important residues. This leaves us room to refine the hypothesis of double water exclusion again to further narrow down biclique patterns so that those energetically less important biclique residues can be filtered.

If the total 268 residues ($\Delta\Delta G < 0.5$) are taken as a negative set, the specificity of our method is $1 - 85/268 = 68.3\%$; if the total 133 residues ($\Delta\Delta G \leq 0$) are taken as a negative set, then our biclique residues contain 41 of them, thus the specificity of our method is $1 - 41/133 = 69.1\%$. These results indicate that our biclique residues did not entirely exclude energetically less important residues. However, we would like to note that the double water exclusion hypothesis and biclique patterns are still biologically acceptable. Alanine-scanning mutagenesis, the main biotechnology adapted in ASEdb, examines the binding free energy change of a single residue upon its mutation to alanine. This experimental method does not solve the cooperativity and additivity problems of energetically important residues. Recently, alanine-shaving technologies (Moreira *et al.*, 2007) have been proposed. Alanine shaving is a process of making multiple simultaneous alanine mutations, thus cooperativity and additivity can be measured by comparing the simultaneous free-energy change to the sum of the free-energy changes attributed to single mutations. Therefore, some single mutations that are less important under alanine-scanning mutagenesis can be energetically important under alanine shaving. Thus, some energetically less important single mutants can be considered as an integral part of a hot spot. We would also like to note that as the 'double water exclusion' mechanism underlines our biclique patterns, residues in a biclique pattern (especially the one having high occurrence and with multiple residues in each side) are therefore excellent candidate residues for the simultaneous shaving.

HotSprint (Guney *et al.*, 2008) is a database storing computational hot spots for 35 776 protein interfaces among 49 512 protein interfaces extracted from the multi-chain structures in PDB (as of February 2006). Those computational hot spots were derived based on residues' conservation score, propensity, and SASA, and they are highly correlated with the experimental hot spots with a sensitivity of 76%. HotSprint is a very new database which does not provide automatic query service yet over large number of PDB entries. Therefore, we report our manual query search results and present examples to illustrate the 'double water exclusion' mechanism that refines the 'O'-ring theory.

We continue to use the simple example of biclique pattern described in Figure 1 which is detected from PDB 1a8g, consisting of five residues (Leu5-Pro9-Leu24-Thr26-Leu97) from chain A (total 99 residues), and two residues (Thr26-Leu97) from chain B (total 99 residues). According to the HotSprint database (by the default setting), only Leu5 was not in the hot spot, the other six residues are all hot residues. The conservation score (Keskin *et al.*, 2004), monomer SASA, and complex SASA for these seven residues are listed in Table 3.

We can see from this table that these seven residues are all buried in the complex with a very small SASA ranging from 0.0 to 2.63 Å^2. Therefore, this compact cluster of residues is almost entirely not accessible to solvent, namely, there exists a ring of neighbor residues constructing a shelter to prevent the hot residues from the water molecules. This is the necessary condition for a residue to become a hot residue (Bogan and Thorn, 1998; DeLano, 2002b) by the O-ring theory. The neighbor residues from 1a8g chain A are Pro1,

Table 3. The conservation score and SASA information of seven residues in a biclique pattern in the interface between chain A and B of PDB entry 1a8g

Pos.	Name	Cons. score	SASA in chain	SASA in complex
Biclique residues in 1a8g chain A				
5	LEU	6	126.88	1.27
9	PRO	7	22.39	2.63
24	LEU	7	23.91	0
26	THR	7	69.56	0.39
97	LEU	7	127.3	0
Biclique residues in 1a8g chain B				
26	THR	7	63.71	0.13
97	LEU	7	121.82	0.03

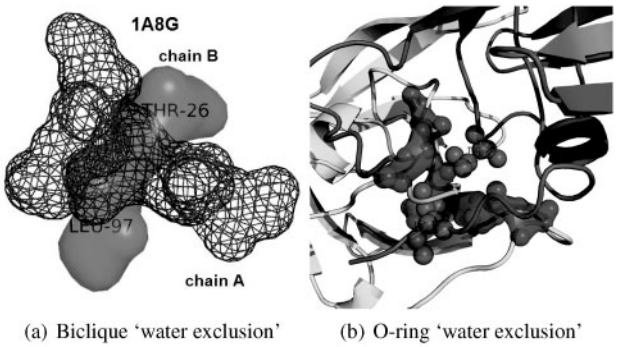
Table 4. The conservation score and SASA information of the ring residues surrounding a biclique pattern (neighbouring residues in 1a8g chain A)

Pos.	Name	Cons. score	SASA in chain	SASA in complex
1	PRO	7	149.94	94.77
2	GLN	7	165.3	122.92
3	ILE	6	77	31.69
4	THR	6	85.36	67.53
25	ASP	7	36.03	13.31
27	GLY	7	74.11	32.99
95	CYS	7	59.46	2.97
96	THR	7	101.82	28.45
98	ASN	7	126.99	26.36

Gln2, Ile3, Thr4, Asp25, Gly27, Cys95, Thr96 and Asn98. Their conservation score, monomer SASA and complex SASA are shown at Table 4. These neighbor residues (except Cys95) all have a large complex SASA, implying that, together with the neighbor residues from chain B, they indeed form a residue wall to prevent the central biclique pattern from any water molecule. This is a nice picture showing an O-ring structure with a 'double water exclusion' mechanism: the central biclique residues form a compact water-free hot spot, while the neighboring residues form an O-ring to occlude any solvent accessibility.

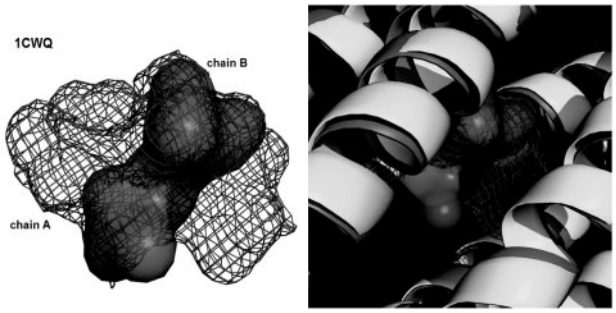
This nice picture is depicted in Figure 2 which was plotted by the PyMOL software (DeLano, 2002a). A lock-and-key architecture can be clearly observed in the left panel: the five hot residues from chain A form the lock (or groove), while the two hot residues from chain B act as a key. This is in agreement with the mechanism of 'anchoring residues' in protein–protein interactions (Rajamani *et al.*, 2004), which explains the kinetically low structural rearrangement of the residues during the formation of complex.

As mentioned, this biclique pattern also occurs between chain A of PDB 1CWQ at LEU92-LEU93-LEU94-THR89-PRO91 and chain B at THR90-LEU93. (1CWQ has 248 residues in each of its two chains.) Though this protein complex is quite different from the 1A8G complex in both structure and sequence, the shape of this biclique pattern is very similar in these two protein complexes as expected. See Figure 3 and compare it to Figure 2.



(a) Biclique 'water exclusion' (b) O-ring 'water exclusion'

Fig. 2. The biclique pattern in PDB 1A8G with the five hot residues—Leu5-Pro9-Leu24-Thr26-Leu97—in chain A, and two hot residues—Thr26-Leu97—in chain B (best viewed in color). (**a**) The biclique shaped in 3D space like a groove-anchor, exhibiting an inner 'water exclusion'. (**b**) The biclique as a hot spot embedded in the binding interface between chain A and chain B, surrounded by neighbor residues of large SASA.



(a) Biclique 'inner water exclusion' (b) O-ring 'water exclusion'

Fig. 3. The biclique pattern in PDB 1CWQ with the five hot residues—LEU92-LEU93-LEU94-THR89-PRO91—in chain A, and two hot residues—THR90-LEU93—in chain B (best viewed in color).

We manually searched HotSprint and found that all or almost all residues of this biclique pattern are hot residues in 1CWQ, 1E0P and 6UPJ; and also in many redundant obligate PDB entries such as 1bdq, 1kzk, 1lzq, 1rl8, 1sgu, 1sh9, 1sp5, 1u8g 1w5v, 1w5w, 1w5x, 1w5y, 1wbk, 1wbm, 1xl2, 1xl5, 1ytg, 1yth, 1ztz 2bqv, etc.

Our second example of biclique pattern is a highly frequent pattern, occurring non-redundantly in 12 chain pairs. Interestingly, all these interactions are obligate; also, the average complex SASAs per residue of this pattern in these chain pairs are all small ∼2–3 $Å^2$. Table 5 shows the protein complex chain pairs that contain this biclique pattern. According to the available results provided by the HotSprint database, almost all the residues of this biclique pattern in these 12 chain pairs are hotspot residues.

These protein chair pairs are all about the study of photosynthetic reaction center, though they have <90% BLAST similarity to each other. For example, the H chains between 1DXR and 1OGV have a similarity score of only 39%, while their L chains have a similarity score of only 56%. (Refer to Section 2 for the BLAST software.) The amazing result here is that this biclique pattern occurs exactly in the same position across 11 chain pairs (2–12 of Table 5) despite of their structural difference. Furthermore, the last eight protein complexes of Table 5 have been previously investigated by Koepke *et al.* (2007);

Table 5. A biclique pattern located in 12 obligate interactions that share a homologous structure

PDB ID	Biclique pattern	
	Residues in chain L	Residues in chain H
1DXR	LYS8 TYR9 VAL11	GLY113 LEU90 PRO114 VAL112
1OGV	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2BOZ	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2GNU	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2J8C	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2J8D	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2UWT	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2UWU	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2UWV	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2UX3	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2UXJ	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109
2UXK	LYS8 TYR9 VAL11	GLY110 LEU87 PRO111 VAL109

all of them are X-ray structure of the Photosynthetic Reaction Center from Rhodobacter Sphaeroides under neutral or charge separated states of different pH levels. The amino acids of these protein chains are the same, but their secondary structure and domain assignment are different under different pH conditions. We can conjecture that these different charge states and pH levels do not change the hot spot at LYS8, TYR9 and VAL11 of chain L, binding with GLY110, LEU87, PRO111 and VAL109 of chain H.

5 CONCLUDING REMARKS

We have proposed a 'double water exclusion' hypothesis to refine the influential O-ring theory for the study of hot spots at protein binding interfaces. We take the classical graph term 'maximal biclique' to model the molecular topology of contact residues tightly packed in close vicinity between a pair of interacting protein chains. We term this molecular topology as 'biclique pattern' to emphasize the collective force of the multiple, dense atom-atom pairs. Based on PDB structural data, we have addressed two properties of biclique patterns: the non-redundant occurrence and statistical significance of biclique patterns, and the relation with computational and experimental hotspot residues. We have observed that biclique patterns can have very high occurrence in both obligate and transient interactions. This indicates that the biclique topology commonly exists in interacting residues, and they are not random patterns as evaluated by their statistical significance. We have verified that many biclique residues are hotspot residues through queries to the HotSprint and ASEdb databases when the complex SASA of the residues is small. This result supports our 'double water exclusion' mechanism; and this result also strongly suggests that the influential 'O'-ring theory, the 'coupling' tendency, together with our biclique pattern concept can provide a road map to deeply understand the binding affinity of protein interactions.

ACKNOWLEDGEMENTS

We thank G.Liu for the source code of the LCM-MBC algorithm, and S.Lukman for technical discussion on some biological concepts.

Funding: Singapore MOE ARCTier-2 (grant T208B2203).

Conflict of Interest: none declared.

REFERENCES

Asratian,A.S. *et al* (1998) *Bipartite Graphs and their Applications*. Cambridge University Press.

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.

Chakrabarti,P. and Janin,J. (2002) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343.

Chothia,C. and Janin,J. (1975) Principles of protein-protein recognition. *Nature*, **256**, 705–708.

Clackson,T. and Wells,J. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.

Davis,F.P. and Sali,A. (2005) Pibase: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.

DeLano,W. (2002a) *The Pymol User's Manual*. Delano Scientific, San Carlos, CA.

DeLano,W.L. (2002b) Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.*, **12**, 14–20.

DeLano,W.L. *et al*. (2000) Convergent solutions to binding at a protein-protein interface. *Science*, **287**, 1279–1283.

Eppstein,D. (1994) Arboricity and bipartite subgraph listing algorithms. *Inf. Process. Lett.*, **51**, 207–211.

Fernandez,A. and Scott,R. (2003). Dehydron: a structurally encoded signal for protein interaction. *Biophys. J.*, **85**, 1914–1928.

Gao,Y. *et al*. (2004) Structure-based method for analyzing protein–protein interfaces. *J. Mol. Model.*, **10**, 44–54.

Gong,S. *et al*. (2005) A protein domain interaction interface database: interpare. *BMC Bioinformatics*, **6**, 8.

Guney,E. *et al*. (2008) Hotsprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res.*, **36**, D662–D666.

Halperin,I. *et al*. (2004) Protein-protein interactions: coupling of structurally conserved residues and of hot spots across interfaces—implications for docking. *Structure*, **12**, 1027–1038.

Hubbard,S. and Thornton,J. (1993) *Naccess Computer Program*. Department of Biochemistry and Molecular Biology, University College London.

Jones,S. and Thornton,J.M. (1996) Principles of protein–protein interactions. *Proc. Natl Acad. Sci.*, **93**, 13–20.

Keskin,O. *et al*. (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.

Keskin,O. *et al*. (2005) Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, **345**, 1281–1294.

Koepke,J. *et al*. (2007) Ph modulates the quinone position in the photosynthetic reaction center from rhodobacter sphaeroides in the neutral and charge separated states. *J. Mol. Biol.*, **371**, 396–409.

Korkin,D. *et al*. (2005) Localization of protein-binding sites within families of proteins. *Protein Sci.*, **14**, 2350–2360.

Larsen,T.A. *et al*. (1998) Morphology of protein–protein interfaces. *Structure*, **6**, 421–427.

Lawrence,M.C. and Colman,P.M. (1993) Shape complementarity at protein/protein interfaces. *J. Mol. Biol.*, **234**, 946–950.

Li,H. *et al*. (2006) Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, **22**, 989–996.

Li,J. *et al*. (2007) Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: a one-to-one correspondence and mining algorithms. *IEEE T. Knowl. Data En.*, **19**, 1625–1637.

Li,X. *et al*. (2004) Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states–implications for docking. *J. Mol. Biol.*, **344**, 781–795.

Mintseris,J. and Weng,Z. (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl Acad. Sci.*, **102**, 10930–10935.

Mintz,S. *et al*. (2005) Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions. *Proteins*, **61**, 6–20.

Moreira,I.S. *et al*. (2007) Hot spots–a review of the protein-protein interface determinant amino-acid residues. *Proteins*, **68**, 803–812.

Morrison,J. *et al*. (2006) A lock-and-key model for protein-protein interactions. *Bioinformatics*, **22**, 2012–2019.

Ofran,Y. and Rost,B. (2003). Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, **325**, 377–387.

Preissner,R. *et al*. (1998) Dictionary of interfaces in proteins (dip): data bank of complementary molecular surface patches. *J. Mol. Biol.*, **280**, 535–550.

Privalov,P.L. *et al*. (2007) What drives proteins into the major or minor grooves of dna? *J. Mol. Biol.*, **365**, 1–9.

Rajamani,D. *et al*. (2004) Anchor residues in protein–protein interactions. *Proc. Natl Acad. Sci.*, **101**, 11287–11292.

Sprinzak,E. *et al*. (2006) Characterization and prediction of protein–protein interactions within and between complexes. *Proc. Natl Acad. Sci.*, **103**, 14718–14723.

Suryani,L. *et al*. (2008) Interacting amino acid preferences of 3d pattern pairs at the binding sites of transient and obligate protein complexes. In *Proceedings of APBC*, Kyoto, Japan, pp. 69–78.

Thorn,K.S. and Bogan,A.A. (2001) Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.

Tsai,C.J. *et al*. (1996) A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.*, **260**, 604–620.

Tsai,C.J. *et al*. (1997) Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.*, **6**, 53–64.

Zhu,H. *et al*. (2006) Noxclass: prediction of protein–protein interaction types. *BMC Bioinformatics*, **7**, 27.