Burial level change defines a high energetic relevance for protein binding interfaces

Zhenhua Li, Ying He, Limsoon Wong and Jinyan Li*

Abstract—Protein-protein interfaces defined through atomic contact or solvent accessibility change are widely adopted in structural biology studies. But, these definitions cannot precisely capture energetically important regions at protein interfaces. The burial depth of an atom in a protein is related to the atom's energy. This work investigates how closely the *change* in burial level of an atom/residue upon complexation is related to the binding. Burial level change is different from burial level itself. An atom deeply buried in a monomer with a high burial level may not change its burial level after an interaction and it may have little burial level change. We hypothesize that an interface is a region of residues all undergoing burial level changes after interaction. By this definition, an interface can be decomposed into an onion-like structure according to the burial level change extent. We found that our defined interfaces cover energetically important residues more precisely, and that the binding free energy of an interface is distributed progressively from the outermost layer to the core. These observations are used to make predictions for binding hot spots. Our approach's F-measure performance on a benchmark data set of alanine mutagenesis residues is much superior or similar to those by complicated energy modeling or machine learning approaches.

Index Terms—Protein interface, protein binding, hot spot, O-ring

1 Introduction

The definition of protein binding interfaces is a fundamental basis in structural bioinformatics studies. There are two common approaches to the definitions. One is based on the change in solvent accessible surface area ($\Delta SASA$) upon binding [1]. An atom or a residue that loses its $SASA$ exceeding a threshold after complex formation [2], [3], [4], [5] is considered to be interfacial. The second approach is through the use of distance or atomic contact [6], [7], [8], [9], [10], [11], [12], [13], [14], sometimes combined with Voronoi diagram or other geometric structures [15], to define protein interfaces. Both of them have many variants with respect to threshold settings and minor changes in implementation. Nevertheless, they all follow one concept: residues/atoms spatially close to their binding partner are defined to be a part of the interface. In fact, interfaces defined by these two approaches can be very similar when the parameters are tuned accordingly [16].

Using only spatial proximity or only $\Delta SASA$ to define protein binding interfaces does not always capture the real contribution of the atoms/residues to the binding free energy—the most important property of binding. On one hand, some supporting residues or atoms relevant to the binding free energy are not covered by these interface definitions. On the other hand, some partially or even fully exposed residues or atoms at the rim region which contribute little to the binding free energy are included in the interface. To address the first problem, Keskin's group [17], [7], [18], [19] proposed an idea to include some "nearby residues" in their interface model. Nearby residues are not in direct contact with the interaction partner but are in contact with those directly contacting residues. That is a partial solution to the first problem. But, it leaves the second problem unaddressed and has a risk of including even a higher number of irrelevant atoms/residues.

Ideally, a definition of binding interfaces should cover all and only those residues/atoms that contribute significantly to binding free energy. However, the development of a good definition for this purpose is still difficult. First, the actual contribution of individual atoms/residues to the overall binding affinity is hard to determine and quantify. The most popular way of quantifying the importance of interfacial residues is by site-directed alanine mutagenesis with which the change in binding free energy ($\Delta\Delta G$) upon mutating into alanine is taken as the importance of a residue. Second, the energetic contribution of residues to binding is not uniformly distributed. Only a small fraction of the residues—hot spot residues—account for the most of binding free energy [20], [21]. A past insightful observation is that hot spot residues, residing at the core of interfaces, are surrounded by energetically less important residues [20], [21]. Those surrounding residues form a regular structure

- *Zhenhua Li and Ying He are with the School of Computer Engineering, School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798. E-mail: 2lizhenhua@gmail.com, YHe@ntu.edu.sg*
- *Limsoon Wong is with the School of Computing, National University of Singapore, Singapore 117417. Email: wongls@comp.nus.edu.sg*
- *Jinyan Li is with the Advanced Analytics Institute, FEIT, University of Technology Sydney, Australia. * Corresponding author. Email: jinyan.li@uts.edu.au*

named "O-ring". This is an influential hypothesis in theory to understand the organization and topology of energetically important residues. However, it does not provide any additional structural information to computationally determine the hot spot residues from a protein quaternary structure. Theoretical energy-based investigations of protein interfaces, e.g., the one by Kortemme and Baker [22], heavily depend on complicated energy terms, such as Lennard-Jones potential, hydrogen bond potential, Coulomb electrostatics, solvation, etc. Thus, they lack mathematical simplicity and geometrical comprehension.

We introduce a new definition for protein binding interfaces through the use of *burial level change* of atoms in proteins. Burial level change is a notion different from burial level itself. Burial level (BL) is a metric to measure the extent an atom/residue is buried in a protein or protein complex as studied by our earlier work [23], [24]. The burial level of an atom is measured as the length of its shortest path to the nearest exposed atom of the structure. Burial level shares some essential idea of effective Born radius [25], Euclidean depth [26] and the Voronoi shelling order [15]. The burial depth of an atom inside a protein was found to be related to its contribution to the energy of the system about 20 years ago [25]. However, a deeply buried atom/residue can be far away from the binding interface, and it may not be necessarily relevant to any binding free energy. Instead, we propose here that the change in burial level is better at capturing this relevance.

We define an atom as an interfacial atom if its *burial level change* upon binding (ΔBL) is bigger than or equal to 1. An interface is thus a set of those atoms whose ΔBL is 1, 2, 3, etc. Under this definition, it can be conceived that the group of atoms of ΔBL equal to $1, 2$, or $3, \ldots$, each forms a shell structure. Therefore, an interface by our definition builds a nested layer structure according to ΔBL measurements, with those atoms of high ΔBL placed at the center and atoms of low ΔBL at the outer areas.

In comparison with the traditional definitions, ours consists of a different set of atoms. Especially, our definition excludes many energetically insignificant atoms but covers those essential supportive atoms, indicating that the new definition captures the energetic relevance more precisely than the traditional models.

It is particularly interesting that atoms with a bigger ΔBL contribute more significantly in a progressive manner to binding free energy in general. Thus, an interface under our definition can be regarded as a structure of multilayer O-rings each shelling its inner layers (onion-like). From the energetic point of view, we name this interface structure a "layered O-ring", generalizing the famous "O-ring" theory [21] to characterize the topological organization of energetically important atoms/residues with finer granularity.

To validate that our layered O-ring structure mimics the hot spot and O-ring structures in protein binding interfaces, we use our model to predict hot spot residues. Previously, hot spots are predicted either by energy-based [22], [27], [28], [29] or machine learning [30], [31], [32], [33] methods, which lack either good performance or interpretability or both. A benchmark data set containing 471 alanine mutations is used in our evaluation. It has 180 or 86 hot spot residues under the hot spot criteria $\Delta\Delta G \geq 1.0$kcal/mol or $\Delta\Delta G \geq 2.0$kcal/mol, respectively. If we consider a onion-like interface itself as the hot spot under the first definition criteria ($\Delta\Delta G \geq 1.0$kcal/mol), a F measure of 0.68 is achieved. When $\Delta\Delta G \geq 2.0$kcal/mol is used to define hot spot, we can simply consider the outmost layer ($\Delta BL = 1$) as O-ring and other inner layers as hot spot.This simple criteria achieves a F measure of 0.56. These performances are better than or similar to other energy-based or machine learning models, indicating that the energy distribution is well captured by our ΔBL-based interfaces.

This work is fundamental to many structural studies on protein-protein interactions, as there is always an underlying definition/model for the protein binding interfaces. For example in protein interface prediction [34], [35], [36], [37] or docking applications [38], [39], [40], protein interface models are explicitly used to distinguish interfacial residues/atoms from those irrelevant to binding. In protein interface characterization, where the size, shape, residue composition, evolution, conservation and other properties of protein recognition sites are examined, the models of protein binding interfaces are always used to scale the scope of the study [1], [41], [6]. Some other studies involving overall binding affinity [42], specificity [43] or interchain co-evolution [44], [45], [46] also need interface models as used implicitly. Our method can be applied to all the aforementioned studies to gain new insight into protein-protein interactions.

2 METHODS

2.1 Data set

The protein docking benchmark 4.0 [47] data set is used in this study to compare the size and region of different interface models. This data set is of high structural diversity and is non-redundant at the family-family level. The original bound structures were downloaded from the protein data bank (PDB) [48], except one interaction (PDB:1ML0) where the PDB entry does not match with the chains specified in the docking data set. For this one, the structure from the benchmark data set is used.

To evaluate the energetic importance of the residues in our newly defined interfaces and to test the performance of hot spot prediction, a site-directed alanine mutagenesis data set is used, taken from ASEdb [49] and other previous publications [50], [51], [52],

[53], [54], [55]. There are a total of 471 mutations in this data set involved in 20 protein-complexes. These 20 protein interaction complexes are shown in Table 1. Of the 471 mutations, 180 or 86 are hot spot mutations, if a residue of $\Delta\Delta G \geq 2.0$kcal/mol or $\Delta\Delta G \geq 1.0$kcal/mol is considered as a hot spot residue, respectively. The full list of mutated residues can be found in Table S1.

2.2 Definition for our layered O-ring interfaces and definitions for traditional interfaces

Preliminaries: Only heavy atoms are considered by this study. Buried water is considered as a part of the protein monomer or complex when water information is available in a structure. To distinguish them from highly exposed water molecules in the bulk solvent, a water molecule with a $SASA$ larger than 10Å^2 are removed iteratively until no water molecule has a $SASA$ larger than 10Å^2 in the remaining structure. Water molecules not removed by this procedure are "buried" water molecules and they form an integral part of the protein monomer or complex.

2.2.1 Burial level of atoms in protein monomers or complexes

The burial level of an atom a, denoted as $BL(a)$, measures how deep it is buried or how far away it is from the bulk solvent. Its precise calculation is based on an atomic contact graph of a protein monomer or a protein complex. An atomic contact graph is a graph with its nodes representing the atoms of the protein and its edges representing the atomic contacts between the atoms. Two atoms are in contact with each other if and only if they share a Voronoi facet and the distance between them is less than the sum of their radius plus the diameter of a water molecule, 2.75Å. The nodes in an atomic contact graph are labeled as 'exposed' or 'buried' depending on a $SASA$ threshold of 10Å^2. Then, the burial level of an atom is defined as the length of the shortest path from it to the nearest exposed atom. We add a pseudo node into the atomic contact graph, and connect it to all and only exposed atom. This problem is then transformed into a single-source shortest path problem, as the burial level of an atom equals to the length of the shortest path from this atom to the pseudo node minus one.

The burial level of an atom a in the atomic contact graph of a protein *monomer* is denoted as $BL_m(a)$, while its burial level in a protein *complex* is denoted by $BL_c(a)$. More details for constructing an atomic contact graph are available in our previous work [23]. An intuitive view of burial level in protein monomers and protein complexes is shown in Figs. 1a and 1b.

2.2.2 Using burial level change to define layered O-ring (onion-like) interfaces

Our layered O-ring (onion-like) [1] interface of a protein complex is a set of atoms I_{LO}:

$$I_{LO} = \{a \mid \Delta BL(a) \geq 1\} \quad (1)$$

where a is an atom in this protein complex, and $\Delta BL(a)$ stands for the change of burial level of atom a upon binding, namely $\Delta BL(a) = BL_c(a) - BL_m(a)$.

To get the monomer structure, the two binding partners in a protein binding complex are simply separated from each other. The ΔBL of buried water molecules in a protein complex is determined as follows. If a water molecule is already buried in one of the protein monomers, its ΔBL is similarly calculated as for the other regular atoms in the protein monomers. If a water molecule is not buried in any of the two monomers but it is buried in the protein complex, its BL_m is set to 0.

ΔBL values shown in Fig. 1c are computed according to the atomic graph of the unbound state of the two proteins in Fig. 1a and their bound state in Fig. 1b. It can be noticed that if an atom is buried deeper in the complex than in its unbound protein, then it is an interface atom. Atoms with the highest ΔBL are buried in the core of the interface surrounded by other atoms with lower ΔBL. For these high ΔBL atoms, moving them to any direction from the core will either reduce their burial levels in the complex ($BL_c \downarrow$) or increase their burial levels in their monomers ($BL_m \uparrow$). Either of these changes will reduce their ΔBL values ($\Delta BL \downarrow$). Thus, according to the ΔBL, the atoms in an interface defined by the new model are organized into a nested onion-like layer structure with atoms' ΔBL equal to 1, 2, 3, ..., each in a shell structure.

Intuitively, ΔBL is always non-negative, since any atom in a protein complex can only be buried deeper than in the monomer. In fact, this proposition is true under special but reasonable conditions.

Theorem 1: ΔBL is always non-negative under the following conditions: (i) the atomic contact structures in the protein monomers are maintained in the protein complex, and (ii) inter-protein atomic contacts are between exposed atoms and inter-protein protein-water-protein contacts are between exposed protein atoms and water molecules.

Proof: Let $G_1(V_1^e, V_1^b, E_1)$ and $G_2(V_2^e, V_2^b, E_2)$ denote the atomic contact graphs of the two binding monomers, where V_i^e and V_i^b represent the exposed and buried atoms, respectively, and E_i represents atomic contacts within a monomer. Similarly, let the atomic contact graph of the complex be denoted as $G(V^e, V^b, E)$. In the interface formation process, we have $V^e \subseteq V_1^e \bigcup V_2^e$ and $V^b \supseteq V_1^b \bigcup V_2^b$. This is due to that some exposed atoms in the unbound

1. We use onion-like interface or layered O-ring interface interchangeably in the rest of this paper.

TABLE 1
The 20 complexes used in this study to assess the energetic relevance of protein binding interfaces.

PDB id	Partner 1	Partner 2	Mutations	$\Delta\Delta G \geq$ 1kcal/mol	$\Delta\Delta G \geq$ 2kcal/mol
1A22	Growth hormone	Growth hormone receptor	66	17	8
1A4Y	Ribonuclease inhibitor	Angiogenin	28	6	3
1AHW	Tissue factor	FAB 5G9	8	5	1
1BRS	Barnase	Barstar	14	11	9
1BXI	Colicin E9 immunity protein	Colicin E9	28	10	6
1CBW	BPTIb	Bovine chymotrypsin	9	1	1
1DAN	Soluble tissue factor	Blood coagulation factor VIIA	77	8	3
1DFJ	Ribonuclease A	Ribonuclease inhibitor	14	11	4
1DVF	FV D1.3	FV E5.2	25	22	9
1DX5	Thrombinb	Thrombomodulin	17	11	5
1FC2	Protein A	IgG	3	2	1
1FCC	Protein G	IgG	8	5	4
1GC1	CD4	Envelope protein GP120	48	3	0
1JCK	Staphylococcal enterotoxin C3	TCR vβ	29	14	7
1JRH	Interferon-γ receptor	Antobody A6	32	19	9
1JTG	Beta-lactamase	Beta-lactamase inhibitory protein	10	7	2
1NMB	FAB NC10	N9 Neuraminidase	1	1	0
1VFB	IGG1-Kappa D1.3 FV	Hen egg white lysozyme	29	11	3
2PTC	Trypsin inhibitorb	Beta-trypsin	1	1	1
3HFM	Hen egg white lysozyme	HYHEL-10 IGG1 FAB	24	15	10
Total			471	180	86

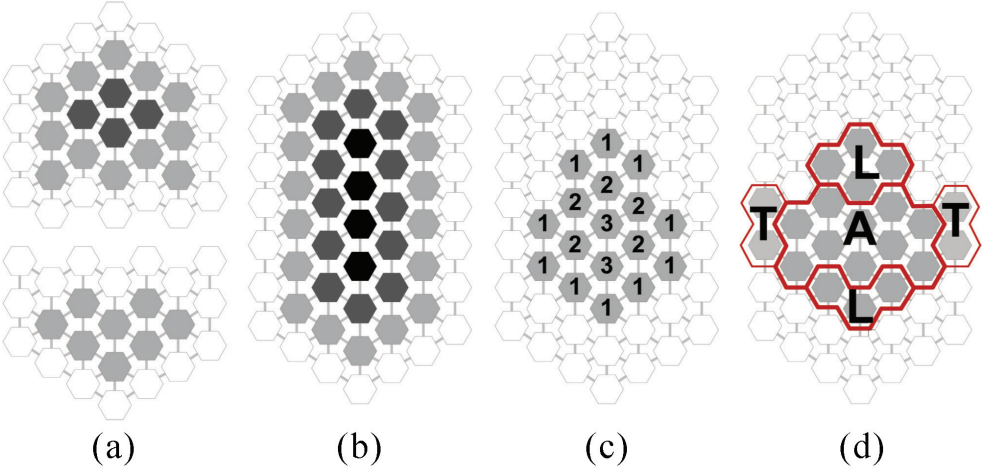


Fig. 1. Atomic contact graphs and burial level patterns of two protein monomers (a) and their binding complex (b). Each hexagon represents an atom and the grey lines between hexagons represent atomic contacts. Atoms with burial level 0 (exposed), 1, 2 and 3 are colored in white, light grey, dark grey and black, respectively. In (c), the onion-like interface is shown in light grey and non-interfacial atoms in white. The number inside a hexagon is the ΔBL of the atom. In (d), atoms in either onion-like or traditional interfaces are in light grey, which are divided into three regions: atoms in both onion-like interface and traditional interface (region A), atoms in onion-like interface but not in traditional interface (region L) and atoms in traditional interface but not in onion-like interface (region T).

monomers are buried in the complex and some water molecules are buried and appear only in the complex. As those intra-protein edges are maintained in the complex graph, we have $E = E_1 \bigcup E_2 \bigcup E_o$, where E_o represents inter-protein and protein-water-protein contacts. From the second condition, we have $E_o \subseteq V_1^e \times V_2^e \bigcup V_1^e \times W \bigcup V_2^e \times W$. Here W are those water molecules only buried in the complex.

To transform the monomer graphs into protein complex graphs, we first consider an intermediate graph $G'(V^{e'}, V^{b'}, E)$, where $V^{e'} = V_1^e \bigcup V_2^e$ and $V^{b'} = V_1^b \bigcup V_2^b \bigcup W$. Comparing with monomer graphs, the only difference is W and E_o are added, while the label (exposed/buried) is unchanged. The length of the shortest path from any nodes in $V_1^b \bigcup V_2^b$ to nodes in $V^{e'}$ is also unchanged. In G', for any vertex in V_1^b, its nearest vertex in $V^{e'}$ can only be in V_1^e, because a path from this vertex to any vertex in V_2^e has to go through a vertex in V_1^e. It is obvious that vertices in W have shortest path to their nearest nodes in $V^{e'}$

is 1 as they all are directly connected to at least one vertex in $V^{e'}$, under the second condition.

We can then adjust G' to G by changing the label of the nodes. We have $V^e \subseteq V^{e'}$ and $V^b \supseteq V^{b'}$. Consider another pseudo node in G and G' that directly connected to all vertices in V^e and $V^{e'}$, respectively. The burial level of a vertex equals to its length of the shortest path to this pseudo node minus 1. As $V^e \subseteq V^{e'}$, transform G' into G is equivalent to delete edges $(V^{e'} - V^e) \times \{p\}$ from G', where p is the pseudo node. Deleting edges from a graph can increase some inter-vertex distances as fewer routes are available. Thus the burial level of a vertex is either unchanged or increased. □

In real cases, the two conditions in the theorem may not hold sometimes, especially when the binding undergoes conformational changes. However, these two conditions are sufficient although not necessary for $\Delta BL \geq 0$. Thus, to define an interface from the static 3-dimensional structure, these two conditions can be fulfilled easily.

In some extreme cases the interface under our definition may not consist of atoms from both sides. When one binding partner is extremely small and none of its atoms is buried in the binding complex, no interface atoms will be defined in this binding partner. This rarely happens in protein-protein binding interfaces where the partners usually have adequate size and shape. Another issue is that, just like other interface models, our model may also define multiple connected regions together as an interface between two proteins.

2.2.3 Traditional definitions of protein binding interfaces

A $\Delta SASA$-based and an atomic contact-based definition of protein binding interfaces are presented for comparison to our onion-like interfaces. According to the $\Delta SASA$ principle, an interface is defined as a set of atoms $I_{\Delta SASA}$ which loses $SASA$ after binding:

$$I_{\Delta SASA} = \{a \mid \Delta SASA(a) < 0\} \tag{2}$$

Here a is an atom in the complex, and $\Delta SASA(a)$ denotes the change of solvent accessible surface area of atom a upon binding, i.e. $\Delta SASA(a) = SASA_c(a) - SASA_m(a)$, where $SASA_c(a)$ and $SASA_m(a)$ are the $SASA$ of atom a in the binding complex and the monomer, respectively. $\Delta SASA$ is always negative or zero when calculated from a static structure. A buried water molecule in the protein complex but not in any of the two monomers is also in the interface under this definition.

The atomic contact-based interface (I_{AC}) we used here is not purely based on a distance threshold. It is defined in the same manner as the one used to calculate burial level. Water is also included in this model, making it a tripartite interface. Interfacial water molecules are those in contact with both proteins. Interface atoms are those in contact with the other side or with interfacial water. If there are no water molecules reported in a structure, the model degenerates into a bipartite. Detailed information of the atomic contact-based tripartite interface model can be found in our previous work [23].

2.3 Hot spot residue prediction and performance evaluation

All side chain atoms other than C^β of a residue R will be removed or directly affected when it is mutated into alanine under alanine mutagenesis experiments to investigate its energetic importance in terms of $\Delta\Delta G$. This part of a residue is named the short side chain, and denoted by $\hat{R}$. For example, a threonine has three heavy atoms in its side chain, C^β, $O^{\gamma 1}$ and $C^{\gamma 2}$, so its short side chain $\hat{R}$ contains two atoms: $O^{\gamma 1}$ and $C^{\gamma 2}$. Specially, a glycine does not have any heavy atom in its side chain, its C^α is then defined to be in the short side chain $\hat{R}$.

We use the layer structured onion-like interface to mimic the hot spot and o-ring structure in a protein binding complex. We vary the number of innermost layers to form the hot spot, and all the outer layers are defined as the O-ring. If any atom of an residue's short side chain is included in the hot spot defined this way, this residue is predicted as a hot spot residue.

Other features are also constructed to compare with ΔBL based predictions. These features include the average burial level of $\hat{R}$ denoted by $BL_c(\hat{R})$, the $SASA$ of $\hat{R}$ denoted by $SASA_c(\hat{R})$, and the $\Delta SASA$ of $\hat{R}$ denoted by $\Delta SASA(\hat{R})$, defined as

$$BL_c(\hat{R}) = \frac{\sum_{a \in \hat{R}} BL_c(a)}{|\hat{R}|} \tag{3}$$

$$SASA_c(\hat{R}) = \sum_{a \in \hat{R}} SASA_c(a) \tag{4}$$

$$\Delta SASA(\hat{R}) = \sum_{a \in \hat{R}} \Delta SASA(a) \tag{5}$$

A residue is predicted as a hot spot residue if the value of the feature BL is bigger than a threshold cutoff. For the $SASA$- or $\Delta SASA$-based prediction method, a residue is predicted as a hot spot residue if the value of the feature is less than a threshold. The performance is evaluated by leave-one-out cross validation. Each time one mutation is held out as test data and the best threshold that optimizes the F measure is found in the remaining training data. The optimal threshold is then applied to the test data to predict whether it is a hot spot residue or not. This process is repeated for every mutation in the data set.

The performance is measured by precision ($precision = \frac{TP}{TP+FP}$), recall ($recall = \frac{TP}{TP+FN}$) and F measure ($F1 = \frac{2 \times TP}{2 \times TP + FP + FN}$), where TP, FP and FN are the number of true positives, false

positives and false negatives, respectively. F measure indicates the overall performance.

3 Experimental Results

Our results start with a real example of onion-like interface, and then we present size, region and energy relevance comparison results against the traditional models. The most important results on the layer-wise energy distribution and tendency of the atoms and residues in our newly defined interfaces are described in the third part.

3.1 Onion-like interfaces: an example

Our definition of protein binding interfaces has been schematically illustrated in Fig. 1. Given the structure of two monomers (Fig. 1a) and their binding complex (Fig. 1b), the atoms' burial level changes (ΔBLs) upon binding can be derived, and those atoms with positive ΔBLs are interfacial atoms (Fig. 1C) by our definition. Such an interface actually consists of multiple layers of atoms corresponding to different ΔBL measurements. Atoms of a high ΔBL are placed at the core and those with a low ΔBL are located at the outer layers. That is an intuitive reason why we call our model the onion-like model. A real onion-like interface is presented in Fig. 2 which is extracted from the interaction between a trypsin and a CMTI-1 squash inhibitor.

In this interface, there are three layers of atoms. Atoms with $\Delta BL = 1$ form the outmost layer. The second layer is sandwiched between the outmost layer and the innermost layer, and the innermost layer is double shelled by the two outer layers. There are 9, 157 and 217 atoms in the innermost, middle and outmost layer, respectively. At the residue level, there are 5, 21 and 67 residues having at least one atom in the inner most, middle and outmost layer, respectively. As atoms in the same residue can differ in terms of their ΔBL measurements, the 5, 21 and 67 residues mentioned above are actually 68 distinct residues. One of them does not have any atom in the outmost layer. This residue, ARG-5 from the inhibitor, penetrates deeply into the trypsin and obtains a very high ΔBL. Three of its side chain atoms have a ΔBL of 3 and all other atoms have a ΔBL of 2. A groove environment accommodating this residue is clearly shown in the binding site of the trypsin (Fig. 2b).

The O-ring theory [20], [21] suggested a dichotomy of binding free energy—energetically important hot spot residues are surrounded by energetically less important residues. Our onion-like interface, as demonstrated in Fig. 2, generalizes this long influential hypothesis by decomposing an interface into a nested multi-layer structure where the energy importance of the atoms/residues can be sorted layer by layer, exhibiting a regular tendency as described later.

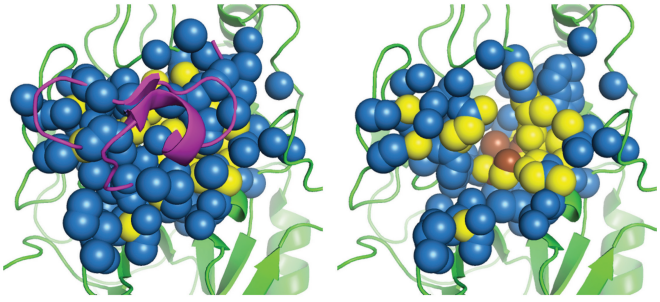


Fig. 2. An onion-like interface between a bovine beta-trypsin (green) and CMTI-I (magenta) (PDB: 1PPE). In (a) both proteins are shown, and in (b) only the enzyme. The atoms with ΔBL of 0, 1 and 2 are indicated by color skyblue, yellow and brown, respectively.

3.2 Size, region and energy relevance comparison with traditional interfaces

3.2.1 *Onion-like interfaces differ in physical region*

The comparison between interfaces under our definition and those under traditional definitions ($\Delta SASA$- or atomic contact-based models) has been briefly demonstrated in Fig. 1d. The grey part in that figure is a union of atoms covering interfacial atoms by our definition and those by the traditional models. This atom union is divided into three regions: (i) region A—the overlapping atoms common to both the onion-like interface and the traditional interfaces, (ii) region L—the atoms only in the onion-like interface, and (iii) region T—the atoms only in the traditional interface. Usually, region L has two clusters of atoms, each from one protein, located at the back of the traditional interface core. Although they are not directly in contact with the other side, they provide the interaction scaffold [7]. Region T is a hoop of atoms. They are very close to the bulk solvent and also near the interaction partner.

In Fig. 3, the sizes of interfaces (defined by the number of interfacial atoms) under different definitions are compared. In both subfigures, it can be noted that when the interface is small, onion-like interfaces tend to contain fewer atoms than those defined by traditional models, indicating that region L is smaller than region T. However, when the interface is large, region L can contain more atoms than region T, so that the overall size of onion-like interface can be larger than traditional interfaces. This tendency is more obvious when comparing with $\Delta SASA$-based interfaces (Fig. 3b).

The distribution of $SASA$ and burial level (BL) of the atoms of the three different interface models are summarized in Fig. 4. Our onion-like interfaces have a higher number of low-$SASA$-and-high-BL atoms, and a lower number of high-$SASA$-and-low-BL atoms. In particular, as shown in Fig. 4a, our

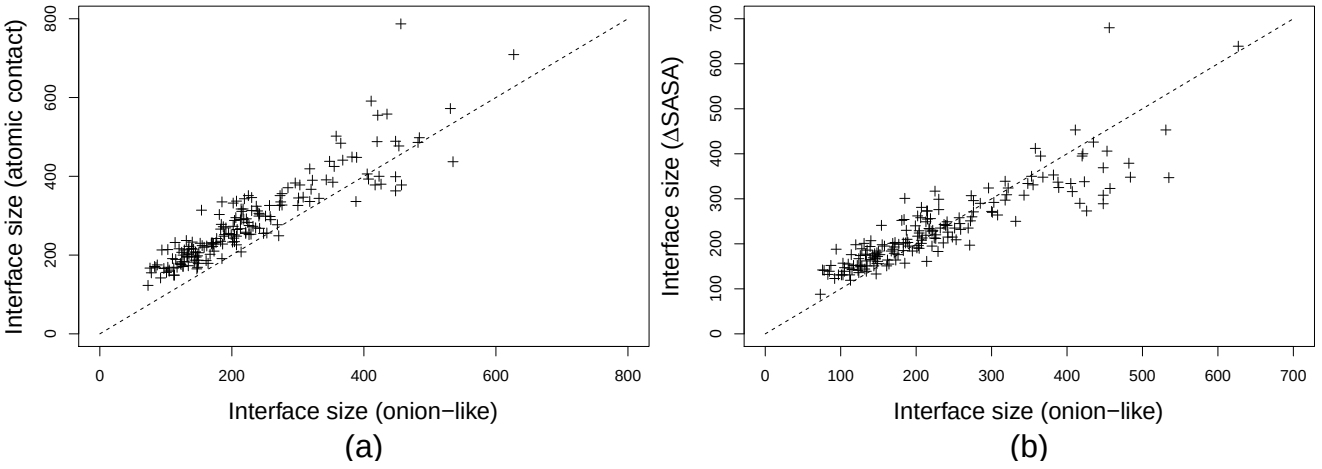


Fig. 3. The sizes of onion-like interfaces versus those of atomic contact-based (a) and $\Delta SASA$-based (b) interfaces.

onion-like interfaces cover a lot of completely buried atoms ($SASA = 0\text{Å}^2$), all of which are located at region L but not at T. In fact, the atoms in region L have a bound state buried level of minimum 2. As they are not in direct contact with the other side or losing any $SASA$ in binding, the burial level in the unbound state should be at least 1. Moreover, their burial level must be increased by at least 1 in the bound state for them to be in the onion-like interface, thus the bound state burial level is at least 2. This can be also confirmed by the occurrence results shown in Fig. 4b. Comparing between the atomic contact-based model and $\Delta SASA$ model, although atomic contact-based interfaces have much higher number of atoms than $\Delta SASA$ interfaces, their difference in the distribution of atoms' $SASA$ and BL is not that significant. An atomic contact-based interface always has more atoms in every interval of $SASA$ and BL than a $\Delta SASA$-based model. This indicates that the two traditional models differ from each other only in the thickness of the interface. If a smaller distance threshold is used in the atomic contact model, the resulting interface can be very similar to that defined by $\Delta SASA$.

3.2.2 *Atoms in onion-like interfaces are energetically more relevant*

The alanine mutagenesis data set of 471 mutations is used here to investigate the energetic importance of the interfacial atoms/residues defined by the three models. Fig. 5 shows the numbers of mutations (broken down into categories by the $\Delta\Delta G$ values) that are covered by the three interface models. A mutation is covered by a model means that at least one of its short side chain atoms is in the interface defined by the model. It can be seen that our model successfully excludes more number of energetically insignificant residues, as indicated by the two leftmost groups of bars where $\Delta\Delta G$ is lower than 1.0kcal/mol. Actually,

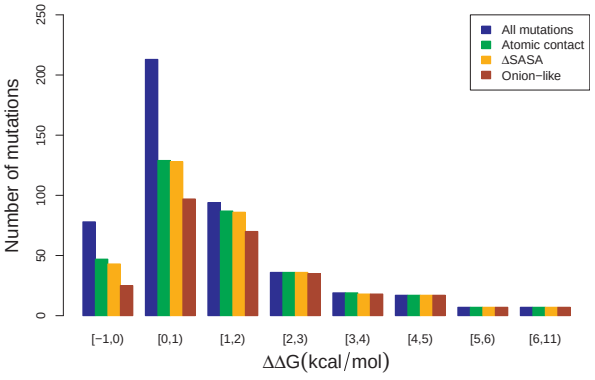


Fig. 5. Distribution of the $\Delta\Delta G$ of the mutations in the data set and the coverage of the mutations with different $\Delta\Delta G$ values by the three interface models.

only 42% of the residues with $\Delta\Delta G < 1.0$kcal/mol are involved in the onion-like model, while for the atomic contact-based and $\Delta SASA$-based models, the percentage is as high as 60% and 59%, respectively. For the 32 residues that are extremely important with a $\Delta\Delta G \geq 4.0$kcal/mol, our interfaces cover all of them, as the traditional interfaces do, despite of much smaller size. Most of the mutations that are covered by traditional models but not by our new model are in region T (see Fig. 1d). Their $\Delta\Delta G$ when mutated to alanine is low in general, indicating the sound rationale of excluding this region from interface.

It has been discussed in literature that an interface should include directly contacting residues/atoms as well as those "support" residues/atoms that provide the interacting scaffolds [7]. In our model, region L can be considered as the support region. The energetic importance of this region is illustrated by comparing the mutations that lie in different regions. Here, the $\Delta SASA$-based model is used as the traditional method so that we can get more atoms in region L for statistical analysis. The distribution of $\Delta\Delta G$

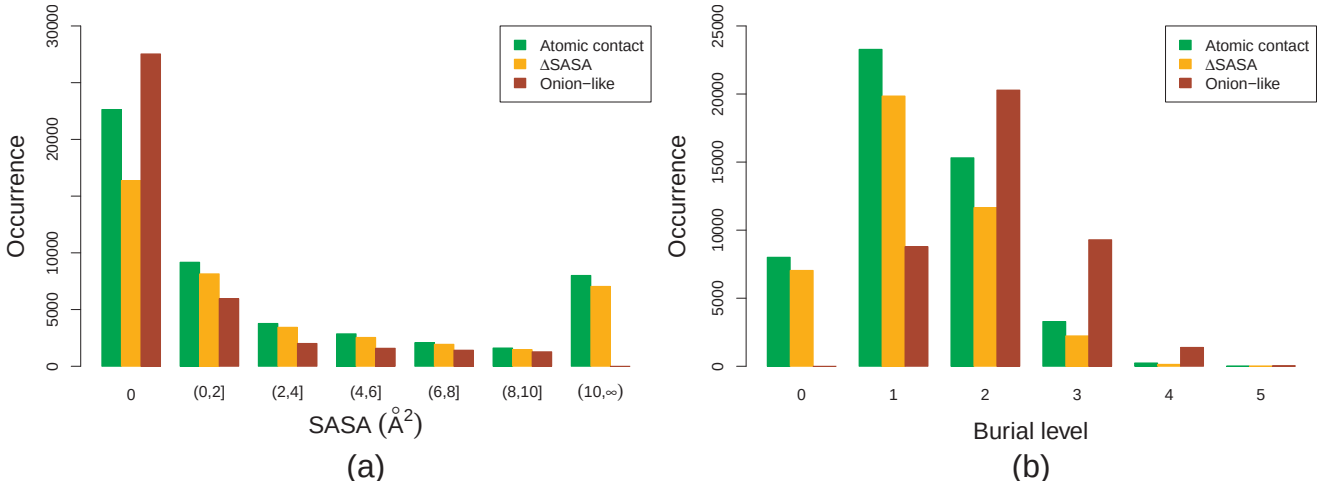


Fig. 4. Distribution of SASA (a) and burial level (b) of atoms in interfaces defined by the three models.

of mutations whose short side chain atoms are in L (short side chain atoms in region L), $A+L$ (short side chain atoms located at both A and L), A, $A+T$ or T are shown in Fig. 6. For Table 2, we tested the significance of the difference in $\Delta\Delta G$ of mutations in these four regions (Wilcoxon rank-sum test [56]).

Since region L is a novel interface region that is not covered by traditional interface models, previous interface alanine scanning did not consider this region at all, resulting in very few mutations lying only in this region in the data set. As indicated in Fig 6, there are only 3 mutations in this group. The energetic importance of these 3 mutations in region L is not found to be significantly different from any other region in this data set; see column 2 of Table 2. However, many residues stretch in multiple regions, and the importance of region L can be assessed by investigating the mutations in $A+L$.

Two conclusions regarding the importance of region L can be made. (i) $\Delta\Delta G$ of mutations in region $A+L$ is significantly higher than that of mutations in region $A+T$ (p-value: 2.9×10^{-4}). (ii) Although the difference between region $A+L$ and region A is not significant, when comparing with region T or $A+T$, the difference between mutations in $A+L$ and those in region T or in region $A+T$ is more significant (p-values: 2.9×10^{-4} versus 0.011 when comparing with $A+T$, 2.8×10^{-8} versus 1.9×10^{-6} when comparing with T). These facts suggest that region L is indeed a "support" region. So, extension of residues from region A into region L is beneficial in terms of binding free energy contribution, while extension of residues from region A into region T is harmful.

3.3 Binding hot spots in onion-like interfaces

3.3.1 Progressive energy tendency from outer layer to inner layer

As introduced, our onion-like model decomposes an interface into a multi-layer structure according to ΔBL of the atoms. It is interesting to utilize this layered model to predict protein binding hot spots, following the O-ring hypothesis which believes that an interface can be divided into a hot spot region and a protective O-ring. In fact, the multi-layer structure can be regarded as several layers of O-rings, each shelling its inner layers as "hot spot".

We vary the number of outmost layers to construct the O-ring, and assume that all the rest inner layers form the hot spot. A residue is predicted as a hot spot residue, if at least one of its short side chain atoms is in the hot spot region of the layered O-ring structure.

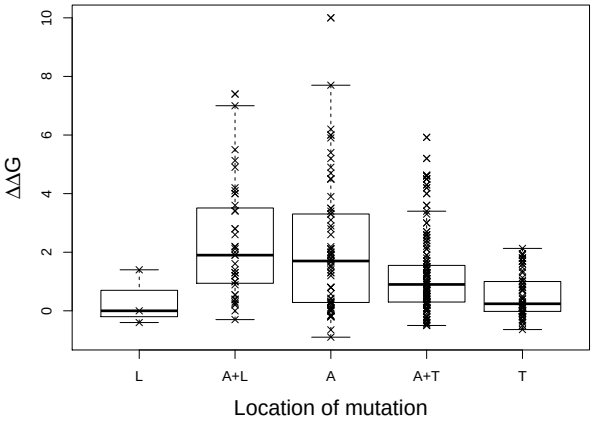


Fig. 6. Distribution of $\Delta\Delta G$ of mutations at different regions.

TABLE 2
Statistical significance (p-values) of the difference in $\Delta\Delta G$ between mutations in different regions by Wilcoxon rank-sum test.

	L	A+L	A	A+T
T	0.55	2.8×10^{-8}	1.9×10^{-6}	1.6×10^{-4}
A+T	0.19	2.9×10^{-4}	0.011	
A	0.084	0.29		
A+L	0.051			

TABLE 3
Performances of hot spot prediction under hot spot definition $\Delta\Delta G \geq 1.0$ kcal/mol.

Row#	Method	Precision	Recall	F1
2	$\Delta BL \geq 1$	0.56	0.86	0.68
3	$\Delta BL \geq 2$	0.70	0.42	0.52
4	$\Delta BL \geq 3$	0.89	0.09	0.16
5	$\Delta SASA(\hat{R})$	0.51	0.81	0.63
6	$BL_c(\hat{R})$	0.46	0.79	0.58
7	$SASA_c(\hat{R})$	0.53	0.77	0.63
8	Robetta	0.69	0.58	0.63
9	FoldX	0.65	0.57	0.61

TABLE 4
Performances of hot spot prediction under hot spot definition $\Delta\Delta G \geq 2.0$ kcal/mol.

Row#	Method	Precision	Recall	F1
2	$\Delta BL \geq 1$	0.30	0.98	0.46
3	$\Delta BL \geq 2$	0.50	0.64	0.56
4	$\Delta BL \geq 3$	0.72	0.15	0.25
5	$\Delta SASA(\hat{R})$	0.48	0.52	0.50
6	$BL_c(\hat{R})$	0.42	0.69	0.52
7	$SASA_c(\hat{R})$	0.36	0.60	0.45
8	Robetta	0.43	0.48	0.45
9	FoldX	0.51	0.42	0.46
10	KFC2a	0.49	0.72	0.58
11	KFC2b	0.56	0.51	0.54
12	HotPoint	0.47	0.50	0.48

Table 3 and Table 4 shows the hot spot prediction performances of this method under hot spot definitions of $\Delta\Delta G \geq 1.0$kcal/mol and $\Delta\Delta G \geq 2.0$kcal/mol, respectively. Rows 2, 3 and 4 in Tables 3 and 4 show the performance using the prediction criteria of ΔBL no less than one (the whole layered O-ring structure as hot spot), no less than two (the first outmost layer as O-ring and other inner layers as hot spot), and no less than three (the first and second outmost layers as O-ring and the other inner layers as hot spot), respectively. It can be seen that in both tables the precision becomes higher when a larger ΔBL cutoff is used. In fact, the average $\Delta\Delta G$ of the residues in the data set that have short side chain atoms with ΔBL no less than 1, 2, or 3 is 1.60kcal/mol, 2.38kcal/mol or 3.50kcal/mol, respectively. This suggests that the inner layers are energetically more important than outer layers with a progressively increasing average binding free energy contribution as ΔBL goes up.

When $\Delta\Delta G \geq 1.0$kcal/mol is used to define hot spots, the best F measure is achieved by the prediction criterion $\Delta BL \geq 1$, i.e. the whole onion-like interface is considered as the hot spot. Under hot spot definition of $\Delta\Delta G \geq 2.0$kcal/mol, $\Delta BL \geq 2$ has the best performance. In this case the outmost layer can be considered as the O-ring structure hypothesized by [21].

3.3.2 Performance comparison with other features

The performances of our method is compared with those by using $\Delta SASA$, burial level, and $SASA$ as prediction criteria. As shown in rows 5, 6, and 7 of Tables 3 and 4, their performances under the two hot spot definitions are worse.

For these four features, ΔBL is the only one that is capable of capturing the two important characteristics of hot spot residues: close to the interaction partner and far away from the bulk solvent. $\Delta SASA$ is related to the closeness to the interaction partner, but it cannot describe the distance to the bulk solvent. Burial level of a residue can describe the distance to bulk solvent, but a deeply buried residue may be far from the actual interface. For $SASA$, a small $SASA$ does not indicate the residue is deeply buried, or indicate it is close to the partner. ΔBL is capable of describing both because if an atom has a high ΔBL, it will be deeply buried in the complex and it will be very close to the interaction partner.

As introduced, the concept BL is different from ΔBL. BL cannot measure how much an atom/residue is relevant to the binding. However, if BL is combined with other properties that guarantee the relevance to binding, for example dense or specific inter-subunit contacts, it has been demonstrated to be useful in hot spot prediction [23], [24].

3.3.3 Performance comparison with existing methods

We have presented the performances of our method in row 2 of Table 3 and row 3 of Table 4. This simple prediction idea can actually performs better than or similar to previous methods FoldX [27], Robetta [22], [57], KFC2 [33] and HotPoint [28], [58] as shown in rows 8 and 9 of Table 3 and rows 8 to 12 of Table 4.

Robetta and FoldX are energy-based methods. They are capable of predicting the $\Delta\Delta G$ measurements for alanine mutations. Robetta is published very early, and it has become a widely used hot spot prediction method. We use its online service to generate its predictions. FoldX is an algorithm uses an empirical force field. Stand-alone FoldX program version 3.0 beta 4 for Windows is used. We first use FoldX to repair the PDB structures. Then interface alanine scannings were carried out on the repaired structures at room temperature (298K) and neutral PH (PH=7). If water molecules are present in the PDB record, their water bridges are considered. These two methods do not have attractive performances, as shown in Rows 8 and 9 of Tables 3 and 4. KFC2 and HotPoint only used hot spot definition criterion $\Delta\Delta G \geq 2$kcal/mol. KFC2 is a recently proposed method based on machine learning techniques. It uses several features related to $SASA$, neighboring residues, atomic density, local contacts and plasticity. Two variants, KFC2a and KFC2b, are built using different feature sets and trained in support vector machines. As shown in Rows 10 and 11 of Table 4, KFC2a has a slightly better performance than our method, and KFC2b has a slightly worse performance than our method. However, note that most of the protein complexes in our data set were already used in their method for training. It is thus not

surprising that it has a good performance on this data set. HotPoint combines the sequence conservation, $SASA$ and $\Delta SASA$, and applies some thresholds on these features to predict hot spot residues. Its average performance is mainly attributed to the use of $SASA$ and $\Delta SASA$, and to the unclear relation between sequence conservation and energetic hot spots [31]. Its performance is worse than our method.

We would like to point out that all these earlier methods are much more complicated and less interpretable than our method—there is essentially only one feature, namely ΔBL, in our method. Meanwhile, our method uses a natural decomposition of the interface according to the O-ring theory.

3.3.4 Case study on the energetic tendency of interfacial residues

The barstar side of a barnase-barstar interface is shown in Fig. 7. Six residues were experimentally mutated into alanine in this side of the interface [59]. Three of them are confirmed to be hot spot residues and three are non hot spot residues. ASP-39 is a hot spot residue located at the center of the interface. It has three atoms in its short side chain, two with $\Delta BL = 2$ and one with $\Delta BL = 3$. Its C^β also has a very high ΔBL of 3. All of its atoms including those in the backbone are in the onion-like interface. This residue has a $\Delta\Delta G$ as high as 7.7kcal/mol. Another hot spot aspartate residue, ASP-35, is also located in the core of the interface and the two oxygens in its short side chain have $\Delta BL = 2$. Comparing with ASP-39, it does not contain any atom with $\Delta BL = 3$, its $\Delta\Delta G$ is then accordingly lower at 4.5kcal/mol. TYR-29 has a $\Delta\Delta G$ of 3.4kcal/mol which is the lowest among these three hot spot residues. It has two atoms with $\Delta BL = 2$, but a part of its short side chain is not covered by the onion-like interface.

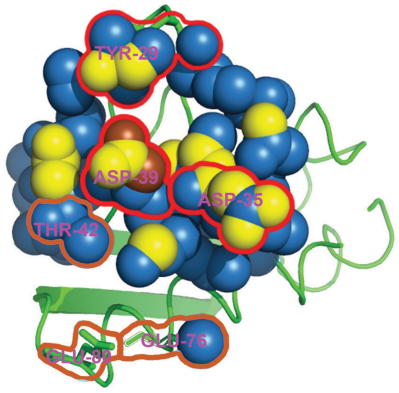
For the three non hot spot residues, none of them is completely covered by the onion-like interface. These three residues can be sorted according to their engagement in the interface as: THR-42 (two main chain atoms and C^β in onion-like interface), GLU-76 (one short side chain atom in onion-like interface) and GLU-80 (no atom in onion-like interface). It is interesting that this order is matched perfectly with their decreasing order of $\Delta\Delta G$: 1.8kcal/mol (THR-42), 1.3kcal/mol (GLU-76) and 0.5kcal/mol (GLU-80). Taking the three hot spot residues together, a regular $\Delta\Delta G$ decreasing trend is observed from the core to outer layers of this onion-like interface and further to the non-interfacial residues.

4 Discussion

Definition of protein binding interfaces is fundamental to almost all structural studies of protein-protein interactions. Due to its wide usage, a model of protein binding interface has to be very simple and straightforward. Previous models, either based on $\Delta SASA$ or inter-chain contacts, are all very intuitive and easy to implement, which is partly why they are so widely accepted for decades. However, another important requirement of an interface model has previously been overlooked. That is, an interface defined has to be "relevant" to the binding. One might argue that a protein interacts with other proteins as a whole, so the whole protein complex should be defined as the protein interface. This is of course a trivial definition and it loses the whole point of defining a protein interface. On the other hand, one can also use only spatial proximity—the most obvious and straightforward definition of binding relevance, which is exactly what traditional models do. Tsai *et al.* [17] made an attempt to define protein binding interface based on van der Waals energy, and they concluded that such interfaces are similar to those under contact-based models. Considering only van der Waals energy is obviously incomplete. Furthermore, even if energy terms are also considered in interface modeling, current energy functions are complicated and computationally intensive. They lose mathematical elegance and cannot be applied easily. In this work we show that our onion-like model of protein binding interface is more accurate in capturing energetically important residues than traditional models and yet it is simple, mathematically elegant and easy to use.

There are some alternative ways to define the BL of an atom, other than the definition used in this work. An immediate idea would be using the Euclidean distance to the surface, which was proposed previously [26], [60]. In this work we define it with a graph model based on the contacts between atoms. The reason is that, first, the contacts between atoms are biologically more meaningful than Euclidean distances. Also, such a definition of burial level yields the onion-like structures with nested layers for protein interfaces, which is effective to define hot spots in an interface. Bouvier *et al.* [15] used the burial depth of atoms to model protein interfaces. Their model, called Voronoi shelling order (VSO), measures the number of interfacial contacts between interfacial atoms and the surface, where the contacts are defined based on Voronoi diagram. As that model also uses the burial depth of atoms and Voronoi diagram, it may easily cause confusion to the correct understanding of our model. Actually, our model is totally different. VSO is a traditional inter-chain contacts based model where the VSO value is but an annotation of the interfacial atoms.

The proposal of the onion-like interface model transforms the understanding of protein binding interfaces in several ways. First, an interface is not just an interface consisting of two patches of binding surfaces any more, rather it consists of two "clusters" of binding atoms/residues. In this sense, we may say



(a)

Residue	$\Delta\Delta G$ (kcal/mol)	ΔBL (all atoms)	ΔBL (short side chain)
ASP-39	7.7	4 2 2	0 2 1
ASP-35	4.5	4 4 0	1 2 0
TYR-29	3.4	3 2 0	2 2 0
THR-42	1.8	3 0 0	0 0 0
GLU-76	1.3	1 0 0	1 0 0
GLU-80	0.5	0 0 0	0 0 0

(b)

Fig. 7. (a) The onion-like interface structure of the barstar side of a barnase-barstar interface (PDB: 1BRS). Atoms in the onion-like interface are shown in spheres. Color skyblue, yellow and brown indicate ΔBL value 1, 2, and 3, respectively. Experimentally identified hot spot and non hot spot residues are surrounded by red and orange curves, respectively. (b) $\Delta\Delta G$ and ΔBL of the six mutated residues. The three numbers in each cell in column 3 and 4 correspond to the numbers of atoms with $\Delta BL = 1$, 2 or 3.

traditional models are two-dimensional definitions but our model is three-dimensional. In other words, to measure the size of an interface, the overall $\Delta SASA$ should be used with caution. Furthermore, an appropriate measure for the size of protein interfaces should be the number of atoms/residues, the "volume" of the interface, the sum of ΔBL, or even the volume integral of ΔBL. Second, the atoms/residues in the interface are no longer equally important. The ΔBL always comes along with an interfacial atom/residue. The internal onion-like structure and organization of an interface is as important as the interface itself, as one can easily define an "interface" inside an interface by setting a higher ΔBL threshold. For some applications, such as analyzing the interface residue/atom composition, new methods taking ΔBL into consideration would be potentially beneficial.

Our onion-like model can be applied to other protein binding studies. For example, in protein binding interface prediction, usually all the interfacial residues are equally important. However, it is obvious that misclassifying a residue in the core is a worse mistake than misclassifying a residue in the outer layers. In this case the residues can be weighted by ΔBL, or more simply, one can just predict the ΔBL values instead of predicting whether a residue is in the interface or not. In protein docking applications, similarly as in protein interface prediction, the onion-like model can be used to define interface residues and, more importantly, it can be used to weight the residues when calculating the root-mean-square deviation RMSD. The onion-like model can also be used in the scoring functions to weight different cases. In general protein binding interface analyses, the properties of protein binding interfaces can be revisited under the onion-like interface definition to gain new knowledge. The size, shape, residue composition, hydrophobicity and other properties of a protein interface can be re-evaluated. Moreover, as our model is a layered model, different layers of interfaces can be analyzed separately. This would deliver new and detailed knowledge of the principles of protein-protein interactions.

5 Conclusion

A new definition of protein binding interfaces is proposed that can capture the energetically important regions more precisely than the traditional atomic contact- or $\Delta SASA$-based models. This newly defined interface is named an onion-like interface as it is based on the burial level change of atoms which can be used to form different levels of atoms. The prediction of binding hot spots is made by transforming the multilayer interface structure into the dichotomy of a hot core and an O-ring. We also proposed to predict hot spots by using the number of short side chain atoms with high ΔBL. The results have verified that our simple prediction ideas perform better than or similar to previous energy-based methods and machine learning methods.

Acknowledgments

This work was supported by a Singapore MOE Tier-2 funding grant (T208B2203) and an Australian Research Council Discovery Project (ARC DP130102124).

References

[1] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc Natl Acad Sci U S A*, vol. 93, no. 1, pp. 13–20, 1996.

[2] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces." *Proteins Struct Funct Bioinf*, vol. 43, no. 2, pp. 89–102, May 2001.

[3] P. Chakrabarti and J. Janin, "Dissecting protein-protein recognition sites," *Proteins Struct Funct Bioinf*, vol. 47, no. 3, pp. 334–343, 2002.

[4] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, and E. S. Huang, "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?" *Protein Sci*, vol. 13, no. 1, pp. 190–202, Jan. 2004.

[5] E. D. Levy, "A simple definition of structural regions in proteins and its use in analyzing interface evolution," *J Mol Biol*, vol. 403, no. 4, pp. 660–670, Nov. 2010.

[6] Y. Ofran and B. Rost, "Analysing six types of protein-protein interfaces." *J Mol Biol*, vol. 325, no. 2, pp. 377–387, Jan. 2003.

[7] O. Keskin, C. J. Tsai, H. Wolfson, and R. Nussinov, "A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications." *Protein Sci*, vol. 13, no. 4, pp. 1043–1055, Apr. 2004.

[8] W. K. K. Kim, A. Henschel, C. Winter, and M. Schroeder, "The many faces of protein-protein interactions: A compendium of interface geometry." *PLoS Comput Biol*, vol. 2, no. 9, pp. e124+, Sep. 2006.

[9] J. L. Chung, W. Wang, and P. E. Bourne, "Exploiting sequence and structure homologs to identify protein-protein binding sites," *Proteins Struct Funct Bioinf*, vol. 62, no. 3, pp. 630–640, Mar. 2006.

[10] Q. Xu, A. A. A. Canutescu, G. Wang, M. Shapovalov, Z. Obradovic, and R. L. L. Dunbrack, "Statistical analysis of interface similarity in crystals of homologous proteins." *J Mol Biol*, vol. 381, no. 2, pp. 487–507, Jun. 2008.

[11] K. Yura and S. Hayward, "The interwinding nature of protein-protein interfaces and its implication for protein complex formation," *Bioinformatics*, vol. 25, no. 23, pp. 3108–3113, Dec. 2009.

[12] J. von Eichborn, S. Gunther, and R. Preissner, "Structural features and evolution of protein-protein interactions," *Genome Inform*, vol. 22, pp. 1–10, Jan 2010.

[13] A. R. Kinjo and H. Nakamura, "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, Jun. 2010.

[14] M. E. Johnson and G. Hummer, "Interface-Resolved network of Protein-Protein interactions," *PLoS Comput Biol*, vol. 9, no. 5, pp. e1 003 065+, 2013.

[15] B. Bouvier, R. Grünberg, M. Nilges, and F. Cazals, "Shelling the voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics, and composition," *Proteins Struct Funct Bioinf*, vol. 76, no. 3, pp. 677–692, 2009.

[16] L. Lo Conte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites." *J Mol Biol*, vol. 285, no. 5, pp. 2177–2198, Feb. 1999.

[17] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov, "A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique." *J Mol Biol*, vol. 260, no. 4, pp. 604–620, Jul. 1996.

[18] O. Keskin, B. Ma, and R. Nussinov, "Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues." *J Mol Biol*, vol. 345, no. 5, pp. 1281–1294, 2005.

[19] O. Keskin and R. Nussinov, "Similar binding sites and different partners: Implications to shared proteins in cellular pathways," *Structure*, vol. 15, no. 3, pp. 341–354, Mar. 2007.

[20] T. Clackson and J. A. Wells, "A hot spot of binding energy in a hormone-receptor interface." *Science*, vol. 267, no. 5196, pp. 383–386, 1995.

[21] A. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces," *J Mol Biol*, vol. 280, no. 1, pp. 1–9, 1998.

[22] T. Kortemme and D. Baker, "A simple physical model for binding energy hot spots in protein-protein complexes," *Proc Natl Acad Sci U S A*, vol. 99, no. 22, pp. 14 116–14 121, 2002.

[23] Z. Li and J. Li, "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, 2010.

[24] Z. Li, L. Wong, and J. Li, "DBAC: A simple prediction method for protein binding hot spots based on burial levels and deeply buried atomic contacts," *BMC Syst Biol*, vol. 5, no. Suppl 1, p. S5, 2011.

[25] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *J Am Chem Soc*, vol. 112, no. 16, pp. 6127–6129, 1990.

[26] A. Pintar, O. Carugo, and S. Pongor, "Atom depth as a descriptor of the protein interior," *Biophys J*, vol. 84, no. 4, pp. 2553–2561, April 2003.

[27] R. Guerois, J. E. Nielsen, and L. Serrano, "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, 2002.

[28] N. Tuncbag, A. Gursoy, and O. Keskin, "Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy," *Bioinformatics*, vol. 25, no. 12, pp. 1513–1520, Jun. 2009.

[29] A. Benedix, C. M. Becker, B. L. de Groot, A. Caflisch, and R. A. Bockmann, "Predicting free energy changes using structural ensembles," *Nat Methods*, vol. 6, no. 1, pp. 3–4, 2009.

[30] S. J. Darnell, D. Page, and J. C. Mitchell, "An automated decision-tree approach to predicting protein interaction hot spots," *Proteins Struct Funct Bioinf*, vol. 68, no. 4, pp. 813–823, 2007.

[31] K. Cho, D. Kim, and D. Lee, "A feature-based approach to modeling protein-protein interaction hot spots." *Nucleic Acids Res*, vol. 37, no. 8, pp. 2672–2687, 2009.

[32] J. F. Xia, X. M. Zhao, J. Song, and D. S. Huang, "APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," *BMC Bioinf*, vol. 11, no. 1, pp. 174+, 2010.

[33] X. Zhu and J. C. Mitchell, "KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density and plasticity features," *Proteins Struct Funct Bioinf*, vol. 79, pp. 2671–2683, 2011.

[34] H. X. Zhou and Y. Shan, "Prediction of protein interaction sites from sequence profile and residue neighbor list," *Proteins Struct Funct Bioinf*, vol. 44, no. 3, pp. 336–343, 2001.

[35] H. Neuvirth, R. Raz, and G. Schreiber, "ProMate: A structure based prediction program to identify the location of Protein-protein binding sites," *J Mol Biol*, vol. 338, no. 1, pp. 181–199, Apr. 2004.

[36] H. Chen and H. X. X. Zhou, "Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data." *Proteins Struct Funct Bioinf*, vol. 61, no. 1, pp. 21–35, Oct. 2005.

[37] H. X. Zhou and S. Qin, "Interaction-site prediction for protein complexes: a critical assessment," *Bioinformatics*, vol. 23, no. 17, pp. 2203–2209, Sep. 2007.

[38] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin, "Haddock: A protein-protein docking approach based on biochemical or biophysical information," *Journal of the American Chemical Society*, vol. 125, no. 7, pp. 1731–1737, 2003.

[39] A. M. Bonvin, "Flexible protein-protein docking," *Current Opinion in Structural Biology*, vol. 16, no. 2, pp. 194 – 200, 2006.

[40] B. Jiménez-García, C. Pons, and J. Fernádez-Recio, "pydockweb: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring," *Bioinformatics*, 2013.

[41] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci U S A*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[42] I. H. Moal, R. Agius, and P. A. Bates, "Protein-protein binding affinity prediction on a diverse set of structures," *Bioinformatics*, vol. 27, no. 21, pp. 3002–3009, Nov. 2011.

[43] E. L. Humphris and T. Kortemme, "Design of Multi-Specificity in protein interfaces," *PLoS Comput Biol*, vol. 3, no. 8, pp. e164+, Aug. 2007.

[44] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia, "Correlated mutations contain information about protein-protein interaction," *J Mol Biol*, vol. 271, no. 4, pp. 511–523, Aug. 1997.

[45] L. Hakes, S. C. C. Lovell, S. G. G. Oliver, and D. L. L. Robertson, "Specificity in protein interactions and its relationship

with sequence diversity and coevolution." *Proc Natl Acad Sci U S A*, vol. 104, no. 19, pp. 7999–8004, May 2007.
[46] M. G. Kann, B. A. Shoemaker, A. R. Panchenko, and T. M. Przytycka, "Correlated evolution of interacting proteins: looking behind the mirrortree." *J Mol Biol*, vol. 385, no. 1, pp. 91–98, Jan. 2009.
[47] H. Hwang, T. Vreven, J. Janin, and Z. Weng, "Protein-protein docking benchmark version 4.0," *Proteins Struct Funct Bioinf*, vol. 78, pp. 3111–3114, 2010.
[48] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res*, vol. 28, no. 1, pp. 235–242, Jan. 2000.
[49] K. S. Thorn and A. A. Bogan, "ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions." *Bioinformatics*, vol. 17, no. 3, pp. 284–285, 2001.
[50] T. Clackson, M. H. Ultsch, J. A. Wells, and A. M. de Vos, "Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity." *J Mol Biol*, vol. 277, no. 5, pp. 1111–1128, 1998.
[51] B. C. Cunningham and J. A. Wells, "Comparison of a structural and a functional epitope," *J Mol Biol*, vol. 234, no. 3, pp. 554 – 563, 1993.
[52] L. Cedergren, R. Andersson, B. Jansson, M. Uhln, and B. Nilsson, "Mutational analysis of the interaction between staphylococcal protein a and human igg1," *Protein Engineering*, vol. 6, no. 4, pp. 441–448, 1993.
[53] D. A. Dougan, R. L. Malby, L. C. Gruen, A. A. Kortt, and P. J. Hudson, "Effects of substitutions in the binding surface of an antibody on antigen affinity," *Protein Eng*, vol. 11, no. 1, pp. 65–74, 1998.
[54] D. J. Sloan and H. W. Hellinga, "Dissection of the protein g b1 domain binding site for human igg fc fragment," *Protein Science*, vol. 8, no. 8, pp. 1643–1648, 1999.
[55] J. Pons, A. Rajpal, and J. F. Kirsch, "Energetic analysis of an antigen/antibody interface: Alanine scanning mutagenesis and double mutant cycles on the hyhel-10/lysozyme interaction," *Protein Science*, vol. 8, no. 5, pp. 958–968, 1999.
[56] F. Wilcoxon, "Individual comparisons by ranking methods," *Biom Bull*, vol. 1, no. 6, pp. 80–83, 1945.
[57] T. Kortemme, D. E. Kim, and D. Baker, "Computational alanine scanning of protein-protein interfaces," *Sci STKE*, vol. 2004, no. 219, pp. pl2+, 2004.
[58] N. Tuncbag, O. Keskin, and A. Gursoy, "HotPoint: hot spot prediction server for protein interfaces," *Nucleic Acids Res*, vol. 38, no. suppl 2, pp. W402–W406, 2010.
[59] G. Schreiber and A. R. Fersht, "Energetics of protein-protein interactions: Analysis ofthe barnase-barstar interface by single mutations and double mutant cycles," *J Mol Biol*, vol. 248, no. 2, pp. 478 – 486, 1995.
[60] A. Pintar, O. Carugo, and S. Pongor, "DPX: for the analysis of the protein core," *Bioinformatics*, vol. 19, no. 2, pp. 313–314, 2003.

PLACE PHOTO HERE

Zhenhua Li received his PhD in Bioinformatics and Computational Biology from Nanyang Technological University in 2013. Before that, he studied computer science in Wuhan University where he was awarded BEng and MEng degrees in 2007 and 2009, respectively. His current research interests include medical data analysis, bioinformatics and data mining..

PLACE PHOTO HERE

Ying He received the BS and MS degrees in electrical engineering from Tsinghua University, and the PhD degree in computer science from Stony Brook University. He is currently an associate professor at the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include the broad area of visual computing. He is particularly interested in the problems that require geometric computation and analysis..

PLACE PHOTO HERE

Limsoon Wong is KITHCT Professor of Computer Science and Professor of Pathology at the National University of Singapore. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He serves/served on the editorial boards of Information Systems, Journal of Bioinformatics and Computational Biology, Bioinformatics, Biology Direct, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Drug Discovery Today, Journal of Biomedical Semantics, and Methods. He is a scientific advisor to Semantic Discovery Systems (UK), Molecular Connections (India), and CellSafe International (Malaysia). He received his BSc(Eng) in 1988 from Imperial College London and his PhD in 1994 from University of Pennsylvania.

PLACE PHOTO HERE

Jinyan Li is an Associate Professor and core member at Advanced Analytics Institute and Center for Health Technologies, Faculty of Engineering and IT, University of Technology, Sydney, Australia. His research is focused on fundamental data mining algorithms, machine learning, gene expression data analysis, structural bioinformatics, and information theory. He is known for the notion of emerging patterns in data mining, and is known for 'double water exclusion hypothesis in bioinformatics. Jinyan obtained his PhD from the University of Melbourne, Master degree of Engineering from Hebei University of Technology, and Bachelor degree of Science from National University of Defense Technology.