BMC Bioinformatics

RESEARCH ARTICLE Open Access

Integrating water exclusion theory into β contacts to predict binding free energy changes and binding hot spots

Qian Liu[1,2], Steven CH Hoi[2], Chee Keong Kwoh[2], Limsoon Wong[3] and Jinyan Li[1*]

Abstract

Background: Binding free energy and binding hot spots at protein-protein interfaces are two important research areas for understanding protein interactions. Computational methods have been developed previously for accurate prediction of binding free energy change upon mutation for interfacial residues. However, a large number of interrupted and unimportant atomic contacts are used in the training phase which caused accuracy loss.

Results: This work proposes a new method, βACV_{ASA}, to predict the change of binding free energy after alanine mutations. βACV_{ASA} integrates accessible surface area (ASA) and our newly defined β contacts together into an atomic contact vector (ACV). A β contact between two atoms is a direct contact without being interrupted by any other atom between them. A β contact's potential contribution to protein binding is also supposed to be inversely proportional to its ASA to follow the water exclusion hypothesis of binding hot spots. Tested on a dataset of 396 alanine mutations, our method is found to be superior in classification performance to many other methods, including Robetta, FoldX, HotPOINT, an ACV method of β contacts without ASA integration, and ACV_{ASA} methods (similar to βACV_{ASA} but based on distance-cutoff contacts). Based on our data analysis and results, we can draw conclusions that: (i) our method is powerful in the prediction of binding free energy change after alanine mutation; (ii) β contacts are better than distance-cutoff contacts for modeling the well-organized protein-binding interfaces; (iii) β contacts usually are only a small fraction number of the distance-based contacts; and (iv) water exclusion is a necessary condition for a residue to become a binding hot spot.

Conclusions: βACV_{ASA} is designed using the advantages of both β contacts and water exclusion. It is an excellent tool to predict binding free energy changes and binding hot spots after alanine mutation.

Background

A binding hot spot is a small area in a protein binding interface whose mutation can lead to a big change in binding free energy. The determination of its accurate location in the interface is a fundamental problem in structural biology, and is useful for applications such as rational drug design and protein engineering [1]. In wet labs, a residue's contribution to binding free energy can be determined through mutation experiments. For example, alanine scanning mutagenesis [2] mutates interfacial residues individually into an alanine, and then measures the change of binding free energy ($\Delta\Delta G$) to quantify the contribution of the side chain of the mutated residue. Based on these wet-lab experimental outcomes and databases [3-6], it has been reported that binding free energy is unevenly distributed in protein interfaces [7]. In fact, there are always a small fraction of interfacial residues—the binding hot spot—which make major contribution to the binding [7,8] with $\Delta\Delta G \geq 2$ kcal/mol [3]. But wet-lab experiments are both time and cost expensive. Reliable computational methods are thus needed for accurate prediction of binding free energy change.

FoldX [9,10], Robetta [11,12] and CC/PBSA [13] are some well-known physics-based methods for this prediction problem. These methods use empirical terms (such as hydrogen bonds), the van der Waals terms and

*Correspondence: jinyan.li@uts.edu.au
[1] Advanced Analytics Institute and Center for Health Technologies, Faculty of Engineering and IT, University of Technology, Sydney, Australia
Full list of author information is available at the end of the article

Coulomb electrostatics to learn a linear function for estimating the effect on the change of binding free energy after residue mutations. However, the predicted energy by these methods has a large discrepancy from experimentally measured $\Delta\Delta G$ [14]. Thus, other methods have been proposed to qualitatively identify binding hot spots. For example, protein sequences are used by [15] and ISIS [16], while protein tertiary structures are used together with docking techniques by [17]. Protein quaternary structures have been also widely used [18]. For example, Hotsprint [19] and HotPOINT [20] generate rules to identify binding hot spots from features such as conservation, accessible surface area (ASA), residue propensity and/or residue pairwise potentials. Machine learning models are also widely used for predicting binding hot spots. Decision trees are used in MINERVA [14] to induce rules at different levels of protein information including structure, sequence and molecular interactions. Later, machine learning algorithms SVM and its ensemble are employed to combine energetic terms such as van der Waals potentials, solvation energy, hydrogen bonds and Coulomb electrostatics, and/or other protein sequences and structure information for a better hot spot prediction performance. Recently, Bayesian Networks are used to combine three main sources of information related to conservation, FoldX-calculated $\Delta\Delta G$ and atomic contacts for a novel probabilistic model of binding hot spots prediction [21]. Very recently, random forests have been proposed to predict hot spots [22] by using structural neighborhood properties of mutated residues and other conventional physicochemical features [23,24]. Besides alanine mutations, hot spots after mutations to any other type of residues are also investigated [6] and their binding free energy changes can be predicted [13,25] with good performance. Several of these methods are also assessed in a community-wide test for predicting mutation effects on protein-protein interaction affinity [26].

In spite of intensive research, the prediction still needs a big improvement. The existing methods usually used those atomic contacts based on Voronoi diagram or simply defined by a distance threshold with little consideration on the local atomic organization of the contacts. If the distance threshold is too large, e.g., larger than 6 Å, an atomic contact between two atoms i and j may have no direct contact area, because the space between i and j can accommodate other atoms. Such interrupted contacts constitute a large proportion of the traditionally used contacts. It is highly questionable whether they are really important to protein binding. In fact, important contacts in hot spot prediction [10,11] or those closely related to binding hot spots [14] are generally not interrupted, such as hydrogen bonds, salt bridges and $\pi - \pi$ contacts.

To overcome these drawbacks, we propose a novel classifier βACV_{ASA} for predicting $\Delta\Delta G$ and binding hot spots. The main idea of βACV_{ASA} is to use atomic contact vector (ACV) of β contacts (that's why our classifier is named βACV for short) instead of distance-cutoff contacts. β contact, found on β-skeletons [27], is our newly defined contact [28]. A β contact between two atoms restricts that there is no other atoms between these two atoms, and requires that the two atoms should have enough direct contact area to form an interaction. The definition of β contacts can filter out a lot of unimportant and interrupted distance-cutoff contacts. Our analysis has found that β contacts are only a small fraction number of those contacts based on a distance threshold [28], but they are effective to distinguish crystal packing from homodimers [28] and to predict protein-ligand binding affinity [29].

Another important idea is that the relative ASA properties are integrated by our βACV classifier based on the water exclusion hypothesis of binding hot spots. The water exclusion hypothesis states that the topological shape of a binding hot spot and its surrounding residues can be characterized as an O-ring structure [3]. Few residues on the O-ring, which are largely exposed to bulk solvent water, can contribute significantly to the protein binding. Thus in βACV, the energy contribution of a β contact to protein binding is required to be inversely proportional to its ASA.

Our βACV_{ASA} was tested on a dataset of 396 alanine mutations to show its superior performance. We compared βACV_{ASA} with the following methods: (i) ACV methods using distance-cutoff contacts to reveal the importance of β contacts to protein binding; (ii) a βACV method without ASA integration to confirm whether the water exclusion theory is necessary for binding hot spots; and (iii) several widely-used state-of-the-art methods such as Robetta, FoldX, HotPOINT and KFC to show the overall better prediction capability of βACV_{ASA}.

Methods

Dataset

The data stored in the ASEdb database [3] and the mutations in BID [4] having $\Delta\Delta G$ measurements are both used for evaluating our method. In total, our dataset contains 22 protein-protein complexes (detailed in Additional file 1: Table S2). All of them have quaternary structures in PDB and meet the following three requirements. First, no redundancy exists among these protein complexes. Given two protein complexes (e.g., interacted pair A and B, and interacted pair C and D), a sequence identity is calculated through BLAST with the default setting for A and C, A and D, B and C, and B and D, denoted by $S(A, C)$, $S(A, D)$, $S(B, C)$ and $S(B, D)$ respectively. These two protein complexes are redundant if $S(A, C) \geq 40\%$ and $S(B, D) \geq 40\%$, or $S(A, D) \geq 40\%$ and $S(B, C) \geq 40\%$. According to this criterion, most of the protein complexes in our dataset are non-redundant. For those redundant complexes, the mutations in the similar proteins must be in different positions. Our requirement on this sequence identity is reasonable, since atomic contacts and ASA used in this work are derived from complexes only, rather than from sequences (such as required by conservation scores). Secondly, only alanine mutations are considered. Thirdly, mutated atoms before mutation must have at least one distance-cutoff atomic contact with the partner proteins. The mutated atoms are those atoms except N, CA, C, O and CB. Under these requirements, our dataset has 396 alanine mutations (detailed in Additional file 1: Table S3). Of these mutations, 86 are binding hot spot residues having $\Delta\Delta G \geq 2$ kcal/mol.

Atomic β contacts in protein binding interfaces

Atomic β contact is a recently proposed notion of atomic contacts for modeling the well-organized protein 3D structures [28]. Its detail can be found in [28]. For easy reference, we give a brief description of β contacts and a simple method to produce β contacts from a protein quaternary structure.

Atomic β contacts: a definition

Given a quaternary structure of a protein complex p, a β contact between two atoms i and j in p requires that (i) the spatial distance between i and j is less than a threshold T_d plus the sum of their van der Waals radii defined by [30] (distance-cutoff contacts for short), (ii) i and j share a Voronoi facet in p's Voronoi diagram, and (iii) the contact cannot break p's β-skeleton. The β-skeleton [27] of a discrete set p is an undirected graph in computational geometry. In this graph, two points i and j have an edge if angle ikj is sharper than a threshold determined by β, $\forall k \in p, k \neq i, j$. This angle threshold is denoted as $\angle\beta$, which actually defines a forbidden region *fr* of the contact between i and j, e.g., the gray regions in Figure 1. When $\beta = 1$, namely $\angle\beta = 90°$, *fr* is the sphere with the midpoint of i and j as the center and with c's length as the diameter as shown in Figure 1(b). This sphere is similar to van der Waals radii of atoms. The forbidden region *fr* of a β contact usually does not cover any other atoms. Otherwise, if there is an atom k in *fr*, for example as shown in Figure 1(a) when $\angle\beta = 90°$ or in Figure 1(c) when $\angle\beta = 75°$, the contact between i and j is not a β contact. A β contact suggests that its two atoms should have enough direct contact area to form an important interaction. The number of atomic β contacts in protein binding interfaces is only a small fraction number of distance-based contacts or less than half the number of contacts in the Voronoi diagrams when $T_d = 3.3$ as found by [28]. Interestingly, the use of β contacts can achieve better prediction performance for distinguishing false binding of crystal packing from homodimers.

A method to produce β contacts

A protein complex p can be represented as an atomic β contact graph $b(p)$, if all of the heavy atoms are represented by nodes, and the β contacts are represented by edges. To produce $b(p)$ for p, Qhull is first used to obtain the Delaunay triangulation [31] for all nodes. After that, the distance threshold T_d is used to remove those atomic contacts whose distances are too large. T_d is set as 3.3 Å (the diameter of a water molecule 2.8 Å plus 0.5 Å). This threshold is an insensitive factor to β contacts when it is large enough. Please refer to the Additional file 1 for an analysis of β contacts under several different T_ds. Thirdly, each atomic contact is checked to guarantee that it satisfies the β skeleton requirements. To sharpen the difference of those mutations with higher $\Delta\Delta G$ and those with lower $\Delta\Delta G$, the angle threshold $\angle\beta$ is set as 75° in this work, whose forbidden region *fr* is larger than that of $\angle\beta = 90°$ as shown in Figure 1(b) and (c). That is, $\angle\beta = 75°$ is a stricter condition than using $\angle\beta = 90°$ to produce β contacts. The rationale to choose the stricter condition $\angle\beta = 75°$ is illustrated in the following situation. Assume (i) A, B and C are the center points of three atoms with van der Waals radii 1.8 Å (for example, the radius of common Carbon atoms in protein structures), (ii) there is a covalent bond between A and B (AB for short) with spatial

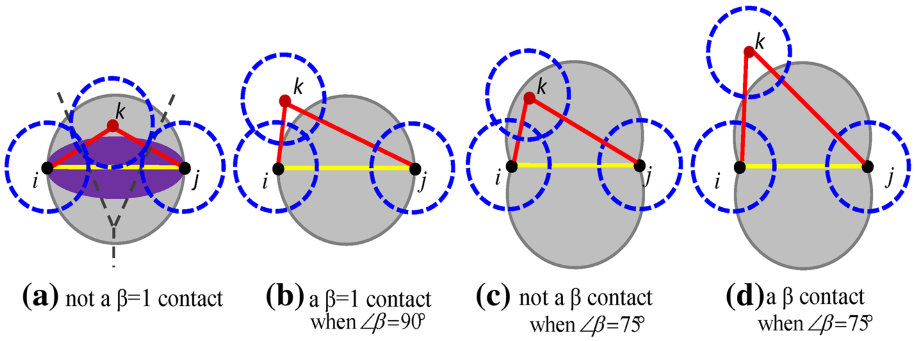

k
i
j
(a) not a β=1 contact
k
i
j
(b) a β=1 contact when ∠β=90°
k
i
j
(c) not a β contact when ∠β=75°
k
i
j
(d) a β contact when ∠β=75°

Figure 1 Examples of β contacts and non-β contacts. Three points, denoted by i, j and k, represent atoms. The dashed circles represent the van der Waals spheres in 2D space. The lines in yellow are of interest.

distance 1.5 Å, and a non-covalent bond between A and C (AC for short) and one between B and C (BC for short), and (iii) the van der Waals sphere of A and that of C are circumscribed each other with the spatial distance 3.6 Å, and the same for the van der Waals sphere of A and that of C. Then, the angle $\angle ABC = 78°$. A stricter threshold than 78° is 75°, which was chosen by this work.

Our prediction methods

This section describes how to construct our $\beta\mathrm{ACV}_{ASA}$ classifier, including how to define βACV and how to integrate ASA into βACV.

βACV: a vector representation for interfacial alanine mutations

We use an atomic contact vector (ACV) [32] of β contacts to represent an interfacial alanine mutation. We also use β contacts of the interfacial bound water molecules to update the vector, and use the atomic environment of the mutation to expand the basic vector.

Constructing a basic βACV vector: To produce a basic βACV for an alanine mutation mut_r^{ala} of residue r in a protein complex p, we build the β contact graph $b(p)$ for p. We then remove the coordinates of the mutated atoms in r to get a quaternary structure resulted from the mutation, denoted by p_r^{mut}. Then, we produce another β contact graph $b(p_r^{mut})$ for p_r^{mut}.

By the mutation mut_r^{ala}, some contacts in $b(p)$ may disappear in $b(p_r^{mut})$, namely those mutated contacts, while some new contacts can be formed in $b(p_r^{mut})$, called new contacts. Both of those mutated and new contacts are represented in a βACV vector. As the heavy atoms from the 20 standard residues are grouped into 8 atomic types (shown in Additional file 1: Table S4) by this work, our βACV vector has 36 pairs of atomic types as elements. The value $v(Ti, Tj)$ of each element in βACV with atomic types Ti and Tj is calculated, using Equation 1.

$$v(Ti, Tj) = \sum_{(x,y)\in M(Ti,Tj)} \frac{1}{d^2_{(x,y)}} - \sum_{(x',y')\in N(Ti,Tj)} \frac{1}{d^2_{(x',y')}} \tag{1}$$

where x and x' are of the atomic type Ti, y and y' are of the atomic type Tj, (x, y) and (x', y') are two atomic pairs, $d(*, *)$ is the spatial distance between a pair of atoms, and $M(Ti, Tj)$(or $N(Ti, Tj)$) is the set of all those mutated (or the set of all those new) contacts whose atomic types are Ti and Tj. Here term $d^2(*, *)$ is specially used to follow the same idea as Coulomb's law which also uses the inverse of squared distance. Note that the other common contacts between $b(p)$ and $b(p_r^{mut})$ are not used in βACV. Alanine mutations of Ala are assumed to have insignificant $\Delta\Delta G$ and alanine mutations of Gly are not considered.

It can be seen that a basic βACV considers all mutated contacts and new contacts, including both across-interface contacts and those contacts from the same proteins or same biological units. However, atomic contacts between covalent-bond nearby atoms are not used in Equation 1. The covalent-bond nearby atoms of a given atom i are those atoms that have not more than three covalent-bond steps from i. For example, suppose $i - j - k - l - m$, where $-$ indicates a covalent bond. From i, the covalent-bond step is 0 to i, is 1 to j, is 2 to k, is 3 to l and is 4 to m, respectively. Thus, i, j, k and l are covalent-bond nearby atoms of i, while m is not. In βACV, the contacts between i and its covalent-bond nearby atoms are excluded from M or N in Equation 1. This is reasonable, because spatially close distances between i and its covalent-bond nearby atoms are more likely due to the rigidity of their covalent bonds.

Bound water molecules in protein interfaces: Protein folding and binding occur in a solvent environment *in vivo*. Water molecules are heavily involved in protein binding and sometimes they can form a compulsory part of the protein interfaces. In this work, a water molecule in PDB is considered as a part of a binding interface if (i) it has at least 3 potential hydrogen-bonds contacts, or (ii) it has 2 potential hydrogen-bond contacts and also has at least 2 other contacts with spatial distances less than 4 Å. A potential hydrogen-bond contact is required to have a spatial distance less than 3.2 Å between a hydrogen donor (such as a nitrogen atom) and a hydrogen acceptor (such as an oxygen atom). Water molecules under this requirement, named bound water molecules, are such closely involved in protein folding and binding that they can play an integral part. Bound water molecules are then grouped into the Oxygen atomic type with more than one hydrogen atom (shown in Additional file 1: Table S4) to update the values of the elements in the basic βACV vector. We did not consider the contacts between any two water molecules.

The neighbourhood atoms of mutated residues: Information of neighbourhood atoms of mut_r^{ala} is used to expand the basic βACV vector. Assume that S is a set of atoms which have β contacts with the mutated atoms under $\angle\beta = 90°$. For each atom in S including the mutated atoms, its nearby atoms are added into S. Then, an atomic vector with the above 8 atomic types is also used to represent those atoms in S in the bound state. The value of its element Tk is calculated using $v(Tk)^b$ in Equation 2. Similarly, $v(Tk)^u$ in Equation 2 is used to calculate another atomic vector for representing the atoms in S in the unbound state.

$$v(Tk)^b = \sum_{j\in S, t_j = Tk} E_j^{loc}; \qquad v(Tk)^u = \sum_{j\in S, t_j = Tk} E_j^{loc}(u) \tag{2}$$

where E_j^{loc} (or $E_j^{loc}(u)$) is the relative local ASA of atom j in the bound (or unbound) state calculated via Equation 6 below. Water molecules were not considered here. Thus, each basic βACV vector is now expanded by another 16 atomic types for representing surrounding information of mutated atoms. So, the expanded βACV is a vector representation with 52 elements.

βACV$_{ASA}$: integrating the water exclusion hypothesis into βACV

Solvent water is compulsory for protein binding, but water exclusion—small accessible surface area (ASA)—is a necessary condition for a residue to become binding hot spot [3,33,34]. Few highly exposed residues can make significant contribution to protein binding strength [34]. Thus, we integrate ASA information into each atomic pair of βACV in Equation 1, and name the method βACV$_{ASA}$. We note that except Equation 1, the other definitions in βACV$_{ASA}$ are the same as those in βACV.

Given a protein complex p, we take the following steps to integrate the water exclusion theory into Equation 1. The first step is to use NACCESS [30] to produce ASA for all of the atoms and residues in both bound and unbound states. For p in the bound state, we then define special ASA terms for an atom i using Equation 3, and for a residue R_i using Equation 4 and Equation 5.

$$E_i = \sqrt{\frac{ASA_i}{50}} \qquad B_i = max(0, 1 - E_i) \qquad (3)$$

$$E_{R_i}^{bb} = \sqrt{\frac{ASA_{R_i}^{bb}}{max(ASA_{R_i}^{bb})}} \qquad B_{R_i}^{bb} = max(0, 1 - E_{R_i}^{bb}) \qquad (4)$$

$$E_{R_i}^{sc} = \sqrt{\frac{ASA_{R_i}^{sc}}{max(ASA_{R_i}^{sc})}} \qquad B_{R_i}^{sc} = max(0, 1 - E_{R_i}^{sc}) \qquad (5)$$

In Equation 3, ASA_i is accessible surface area of atom i, while E_i is its relative ASA, and B_i is its relative ASA burial compared to the maximum ASA, where number 50 is roughly half of NACCESS-calculated ASA of a single water molecule without any neighbor atoms (the ASA of a water molecule is $98.47 = 4 \times 3.14 \times 2.8^2$ Å^2). In Equations 4 and 5, $ASA_{R_i}^{bb}$ and $ASA_{R_i}^{sc}$ are accessible surface area of backbone atoms (i.e., *bb*) and of side-chain atoms (i.e., *sc*) for a residue R_i, while $E_{R_i}^{*}$ and $B_{R_i}^{*}$ are the relative ASA and the relative ASA burial of $* \in \{bb, sc\}$. $max(ASA_{R_i}^{bb})$ and $max(ASA_{R_i}^{sc})$ are the maximum ASA of backbone atoms and of side-chain atoms for the residue type of R_i, which are calculated in a triplet of ALA-R_i-ALA by NACCESS. These backbone atoms and side-chain atoms are defined in the same way as those in [30].

We compute the local ASA E_i^{loc} and local ASA burial B_i^{loc} of an atom i via Equation 6.

$$E_i^{loc} = \begin{cases} E_i \times E_{R_i}^{bb} & \text{if } i \text{ is a backbone atom of } R_i \\ E_i \times E_{R_i}^{sc} & \text{if } i \text{ is a side-chain atom of } R_i \end{cases}$$

$$B_i^{loc} = \begin{cases} B_i \times B_{R_i}^{bb} & \text{if } i \text{ is a backbone atom of } R_i \\ B_i \times B_{R_i}^{sc} & \text{if } i \text{ is a side-chain atom of } R_i \end{cases} \qquad (6)$$

where the multiplication of relative ASA burial of both atom i and its residue is used to calculate local ASA burial B_i^{loc}. This is because relative ASA of both an atom and its residue are critical in describing the accessibility of an atom. For example, an atom may be buried with small ASA but its covalent-bond atoms might be exposed. When relative ASA of atoms or residues are used individually, the performance was worse (data not shown).

To integrate water exclusion theory into Equation 1, we determine the value $v(Ti, Tj)$ of each element in βACV$_{ASA}$ through Equation 7 instead of Equation 1.

$$v(Ti, Tj) = \sum_{(x,y) \in M(Ti,Tj)} \frac{B_x^{loc} \times B_y^{loc}}{d_{(x,y)}^2} - \sum_{(x',y') \in N(Ti,Tj)} \frac{B_{x'}^{loc} \times B_{y'}^{loc}}{d_{(x',y')}^2} \qquad (7)$$

where T_*, x, y, x', y', M and N have the same meaning as those in Equation 1.

Comparison of β contacts with distance-cutoff contacts: To compare the performance of β contacts with distance-cutoff contacts for predicting $\Delta\Delta G$, ACV$_{ASA}$ based on distance-cutoff contacts is constructed in a similar way to constructing βACV$_{ASA}$. To further show the importance of β contacts in protein binding interfaces, nonβACV$_{ASA}$ is also constructed for alanine mutations at the setting of $\angle\beta = 90°$. In nonβACV$_{ASA}$, the values of its elements are the difference of the values of the 52 elements between ACV$_{ASA}$ and βACV$_{ASA}$. To highlight the advantage of β contacts, ACV$_{ASA}$ is also evaluated with different spatial distance thresholds (from 2.9 Å to 5 Å) for defining atomic contacts across interfaces and within binding sites.

Our βACV$_{ASA}$ classifier and its variants described above are summarized in Table 1.

Table 1 Description of βACV$_{ASA}$ and its variant methods

Methods	Description (of the representation for an alanine mutation)
βACV	An ACV of β contacts without ASA integration
βACV_{ASA}	An ACV of β contacts with ASA integration
ACV$_{ASA}$	An ACV of distance-cutoff contacts with ASA integration
nonβACV_{ASA}	The difference of βACV_{ASA} and ACV$_{ASA}$

Ridge regression: predict $\Delta\Delta G$ and binding hot spots

Ridge regression in Matlab is used here to learn a relation between atomic contact vectors and $\Delta\Delta G$. By this regression, values in each column are normalized for the training dataset. Ridge regression minimizes average square error SE between the experimental ($\Delta\Delta G_i^e$) and the predicted $\Delta\Delta G_i^p$ in the training data with N mutations where $SE = \frac{\sum_i (\Delta\Delta G_i^e - \Delta\Delta G_i^p)^2}{N-1}$.

In our evaluation, leave-one-out cross-validation is used for all of the 396 mutations, and then the correlation coefficient R and average standard deviation $\delta = \sqrt{SE}$ are calculated. Under this evaluation framework, there is one outlier prediction by βACV_{ASA} and one outlier by βACV for the whole dataset with 396 mutations. These outliers have less than -3 kcal/mol predicted $\Delta\Delta G$, or more than 11 kcal/mol predicted $\Delta\Delta G$, as shown in Additional file 1: Table S3. This may be due to limited alanine mutations of a high $\Delta\Delta G$ in the dataset.

In this work, predicted hot spot residues are those residue mutations with a predicted $\Delta\Delta G$ $\geq$2 kcal/mol, same as the true hot spot definition.

Hot spot prediction and evaluation measures

βACV$_{ASA}$ is also assessed by applying to the classification problem of binding hot spots. Classification performance is measured by *precision*(*p*.), *recall*(*r*.), *accuracy*(*acc*.) and *F*1 whose definitions are given in Equation 8.

$$\begin{aligned} precision(p.) &= \frac{TP}{TP+FP} \\ recall(r.) &= \frac{TP}{TP+FN} \\ accuracy(acc.) &= \frac{TP+TN}{TP+TN+FP+FN} \\ F1 &= \frac{2 \times precision \times recall}{precision + recall} \end{aligned} \tag{8}$$

where binding hot spots are considered as the true cases, while non-hot spots as the false cases; TP, FP, TN and FN are true positives, false positives, true negatives and false negatives, respectively. Hence, *precision* is the number of correct hot spot predictions divided by the number of positive predictions, *recall* is the fraction of correct hot spot predictions over all hot spots, while *accuracy* is the number of correctly predicted hot spots and non-hot spots divided by the number of all mutations. These measures are also used in [14,20,35] with the same definitions.

Results and discussion

β contacts are better than distance-cutoff contacts for predicting $\Delta\Delta G$

Our βACV_{ASA} classifier is compared with ACV$_{ASA}$ and with nonβACV$_{ASA}$ to show the importance of β contacts in the prediction of $\Delta\Delta G$ under alanine mutations. The prediction results are presented in Figure 2(a), (b) and (c).

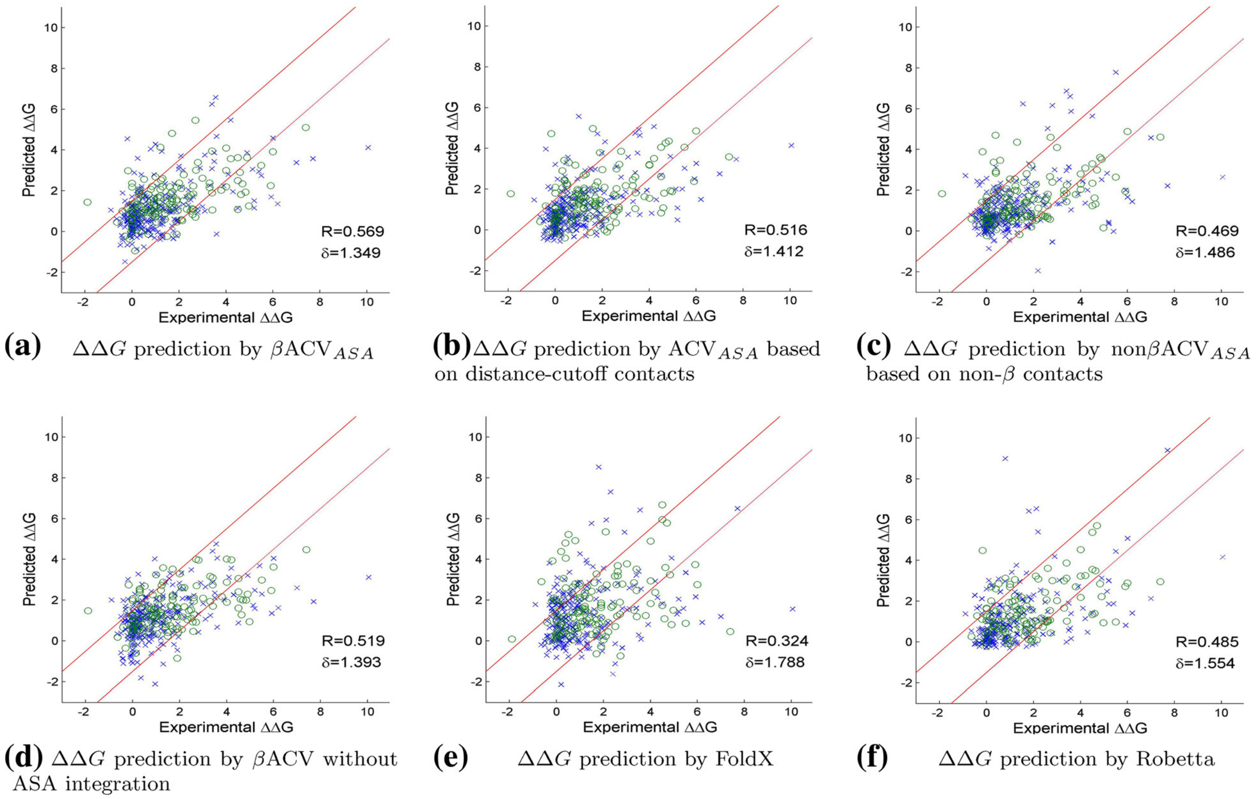
It can be seen from Figure 2(a) and (b) that βACV$_{ASA}$ has a better $\Delta\Delta G$ prediction performance according to both correlation coefficient R and average standard deviation δ. The number of β contacts used by βACV$_{ASA}$ is only a small fraction of the number of distance-cutoff contacts used by ACV$_{ASA}$. For example, there are 54,286 distance-cutoff contacts across binding interfaces for the 22 protein complexes, but there are only 9,830 β contacts across the binding interfaces ($\angle\beta = 90°$), and 4,096 β contacts under the setting of $\angle\beta = 75°$ which is actually used by βACV$_{ASA}$. So, βACV$_{ASA}$ uses only 7.55% number of distance-cutoff atomic contacts but it achieves a better prediction performance.

The comparison between βACV$_{ASA}$ and nonβACV$_{ASA}$ (Figure 2(a) and (c)) further suggests the importance of β contacts in $\Delta\Delta G$ prediction. In Figure 2(c), nonβACV$_{ASA}$ has much lower R (0.469) and a higher δ (1.486) than βACV$_{ASA}$, but nonβACV$_{ASA}$ uses all non-β contacts of $\angle\beta = 90°$, that is, 81.8% number of distance-cutoff atomic contacts.

A lot of alanine mutations are not binding hot spot residues, having a small $\Delta\Delta G$, i.e., <2 kcal/mol. These mutations heavily affect the calculation of R and δ. On the other hand, the prediction of residue mutations with a high $\Delta\Delta G$ is more important. Thus, the classification performance for these binding hot spots is also assessed. The results are shown in Table 2. It is noted that F1 is not the objective function to be optimized in the regression process.

In Table 2 among βACV$_{ASA}$, ACV$_{ASA}$ and non βACV$_{ASA}$, βACV$_{ASA}$ has the highest precision, recall, F1 and accuracy, while nonβACV$_{ASA}$ has the lowest. For example, βACV$_{ASA}$'s F1 is 0.604, 0.122 higher than nonβACV$_{ASA}$'s F1. However, ACV$_{ASA}$ and nonβACV$_{ASA}$ show quite similar performances. The reason would be that non-β contacts (used by nonβACV$_{ASA}$) are often dominant in distance-cutoff contacts (used by ACV$_{ASA}$).

The performances of ACV$_{ASA}$ with different spatial distance thresholds are shown in Table 3. We can see that the performance has a growing tendency when the threshold increases. Nevertheless, the best performance of ACV$_{ASA}$ ($F1$ = 0.5 in Table 2) is much lower than that of βACV$_{ASA}$ ($F1$ = 0.604). Of special interest, when the spatial distance threshold is set at 3.6 Å, ACV$_{ASA}$ has a number of distance-cutoff contacts nearly the same as the number of our used β contacts. In this special case of almost the same number of contacts used, ACV$_{ASA}$ has much worse performance than βACV$_{ASA}$, and only about half of the distance-based contacts are β contacts across the 22 protein-protein binding interfaces. These results affirm that β contacts are advantageous over distance-based contacts for predicting binding hot spot residues.



(a) $\Delta\Delta G$ prediction by βACV$_{ASA}$

(b) $\Delta\Delta G$ prediction by ACV$_{ASA}$ based on distance-cutoff contacts

(c) $\Delta\Delta G$ prediction by nonβACV$_{ASA}$ based on non-β contacts

(d) $\Delta\Delta G$ prediction by βACV without ASA integration

(e) $\Delta\Delta G$ prediction by FoldX

(f) $\Delta\Delta G$ prediction by Robetta

Figure 2 $\Delta\Delta G$ predicted by different methods. In **(a)-(d)**, 'o' represents non-polar residues, while 'x' represents polar residues. *R* is the Pearson correlation coefficient and δ is the average standard deviation. *R* is specially calculated for **(a)** and **(d)** based on the data set after removing the one outlier prediction. *R* is slightly changed to 0.543 or 0.515 respectively when the outlier is not removed. The two diagonal red lines represent $y = x \pm 1.5$.

Water exclusion is a necessary condition of hot spot binding

Literature work has reported that water exclusion is a necessary condition for an interfacial residue to become a hot spot residue [3,33]. To confirm the importance of water exclusion in the prediction of $\Delta\Delta G$, the performance by βACV when ASA is not integrated is assessed.

Table 2 Prediction performances by different methods for the same set of binding hot spots

Methods	Precision	Recall	F1	Accuracy
βACV$_{ASA}$	0.615	0.593	0.604	0.830
ACV$_{ASA}$	0.526	0.477	0.500	0.793
nonβACV_{ASA}	0.513	0.454	0.482	0.788
βACV	0.564	0.616	0.589	0.813
FoldX	0.400	0.488	0.440	0.730
Robetta	0.526	0.465	0.494	0.793
HotPOINT	0.439	0.547	0.487	0.750
KFC2a	0.443	0.767	0.562	0.740
KFC2b	0.521	0.570	0.544	0.793

Table 3 Prediction performance and the numbers of used contacts by βACV$_{ASA}$ and ACV$_{ASA}$

Methods	Distance[1]	#contacts[2]	Precision	Recall	F1	Accuracy
βACV$_{ASA}$		**2,881**	**0.615**	**0.593**	**0.604**	**0.830**
ACV$_{ASA}$	2.9	347	0.486	0.419	0.450	0.778
	3.0	513	0.465	0.382	0.420	0.770
	3.1	715	0.394	0.302	0.342	0.747
	3.2	966	0.487	0.442	0.463	0.778
	3.3	1,293	0.450	0.419	0.434	0.763
	3.42	1,884	0.438	0.372	0.403	0.760
	3.5	2,394	0.494	0.442	0.466	0.780
	3.55	2,789	0.443	0.407	0.424	0.760
	3.6	3,123	0.437	0.360	0.395	0.760
	4	7,542	0.463	0.430	0.446	0.768
	4.5	15,389	0.482	0.465	0.473	0.775
	5	26,752	0.488	0.465	0.476	0.778

[1]: The spatial distance threshold of two atoms.
[2]: The number of atomic contacts involving in the 396 mutations, including mutated contacts and new contacts.

This prediction performance is shown in Figure 2(d) and Table 2. Comparing Figure 2(d) with Figure 2(a), βACV$_{ASA}$ has a better regression performance than βACV indeed. It also has a better F1 performance as seen in Table 2. These results confirm that water exclusion plays an important role in hot spot prediction, and it should be a necessary condition for an interfacial residue to become a binding hot spot residue in protein-protein complexes.

Our method βACV$_{ASA}$ is superior to several widely-used methods

Our βACV$_{ASA}$ classifier is also assessed against the state-of-the-art methods FoldX [9,10], Robetta [11], HotPOINT [20] and KFC [36]. The prediction performances of these previous methods were obtained through their web servers (Robetta, HotPOINT and KFC) or the standalone executable program (FoldX with default settings).

Comparison results

Figures 2(e) and 2(f) show that the prediction performance of FoldX and Robetta are much worse than our βACV$_{ASA}$. These two methods have a R of 0.324 or 0.485, much smaller than βACV$_{ASA}$'s 0.569; their δ is 1.788 or 1.554, much larger than βACV$_{ASA}$'s 1.349. Table 2 also shows their classification performance on the 396 mutations: FoldX's F1 is 0.44, while Robetta's F1 is 0.494, both worse than βACV$_{ASA}$'s 0.604. From Table 2, our method also has better classification performance than HotPOINT and KFC. Other performance comparison results are provided in the Additional file 1 when tested on BID (including protein-peptide interfacial residues) or under the leave-one-complex-out cross-validation framework.

An example of hot spot predictions

We use 3HFM as a case study to illustrate the difference of the binding hot spot prediction results by βACV$_{ASA}$, FoldX and Robetta. The 3HFM complex is an antibody-antigen binding between HyHEL-10 and hen egg white lysozyme. According to ASEdb, a total of 25 alanine mutations were experimented, 11 of which have $\Delta\Delta G$ more than 2 kcal/mol.

Our βACV$_{ASA}$ correctly identified 9 binding hot spot residues with a recall of 0.818, but made 3 false positive predictions with a precision of 0.75 (Figure 3). This gives an F1 of 0.783. However, FoldX made only one hot spot prediction which is correct with a recall of 0.091, and Robetta has a recall of 0.455 (5 out of 11) and a precision of 0.833 (5 out of 6), namely an F1 of 0.588. Both of these methods have a lower prediction performance than our βACV$_{ASA}$. What is more important is that the four positive predictions correctly made only by our βACV$_{ASA}$, not by FoldX or Robetta, have a high $\Delta\Delta G$, such as Trp in position 98 of Chain H ($\Delta\Delta G = 5.5$ kcal/mol) and Tyr in position 50 of Chain L ($\Delta\Delta G = 4.6$ kcal/mol). Please refer to Additional file 1: Table S3 for detail.

Conclusion

A new classifier βACV$_{ASA}$ has been proposed to predict $\Delta\Delta G$ and binding hot spot residues. The novel idea of this classifier is to integrate the water exclusion theory into β contacts. Tested on a data set of 396 alanine mutations, βACV$_{ASA}$ has been found to outperform many other methods. This confirms that β contacts are truly better than traditional distance-cutoff contacts to capture the energetic characteristics of protein folding and binding. This also confirms that water exclusion is a necessary condition for a residue to become a binding hot spot residue.

Availability of supporting data

All the supporting data are included as additional files.

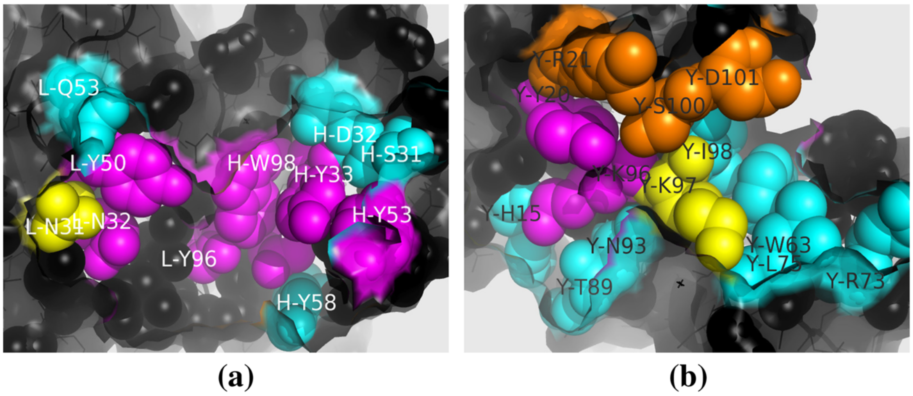


Figure 3 Prediction results by βACV$_{ASA}$ for the residues in the interface between Chain Y and Chain HL(together) in 3HFM. In **(a)** and **(b)**, the true predicted hot spot residues are in magenta, the false predicted non-hot spot residues are in yellow, the false predicted hot spot residues are in orange, while the true predicted non-hot spots are in cyan; X-YZZ stands for residue Y in position ZZ of Chain X.

Additional file

Additional file 1: This additional file covers an analysis on β contacts of different T_ds, more evaluation results and related discussions (including the statistical significance of the difference among Figure 2(a) to 2(d), the dataset and evaluation on BID, the evaluation under leave-one-complex-out cross-validation, and a discussion on using the 396 mutations), the groups of atomic types, and the detail of our dataset.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

QL designed the methods and performed the experiments. JL and SH supervised the study. JL, LW and CK participated in the analysis. QL drafted the manuscript. QL, SH, LW, CK and JL read and revised the manuscript. All authors approved the final version.

Acknowledgements

This work was partially supported by an ARC Discovery Project (DP130102124).

Author details

[1]Advanced Analytics Institute and Center for Health Technologies, Faculty of Engineering and IT, University of Technology, Sydney, Australia. [2]School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore. [3]School of Computing, National University of Singapore, Singapore 117417, Singapore.

Received: 30 July 2013 Accepted: 19 February 2014
Published: 26 February 2014

References

1. Schreiber G, Fleishman SJ: **Computational design of protein-protein interactions.** *Curr Opin Struct Biol* 2013, **23**(6):903–910.
2. Wells J: **Systematic mutational analyses of protein-protein interfaces.** *Methods Enzymol* 1991, **202:**390–411.
3. Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces.** *J Mol Biol* 1998, **280:**1–9.
4. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J: **The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces.** *Bioinformatics* 2003, **19**(11):1453–1454.
5. Kumar MDS, Gromiha MM: **PINT: protein-protein interactions thermodynamic database.** *Nucleic Acids Res* 2006, **34**(suppl 1):D195–D198.
6. Moal IH, Fernandez-Recio J: **SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models.** *Bioinformatics* 2012, **28**(20):2600–2607.
7. Clackson T, Wells J: **A hot spot of binding energy in a hormone-receptor interface.** *Science* 1995, **267:**383–386.
8. Thorn KS, Bogan AA: **ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions.** *Bioinformatics* 2001, **17**(3):284–285.
9. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.** *J Mol Biol* 2002, **320**(2):369–387.
10. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L: **The FoldX web server: an online force field.** *Nucl Acids Res* 2005, **33**(Web Server issue):W382–W388.
11. Kortemme T, Baker D: **A simple physical model for binding energy hot spots in protein-protein complexes.** *Proc Natl Acad Sci USA* 2002, **99**(22):14116–14121.
12. Kortemme T, Kim DE, Baker D: **Computational alanine scanning of protein-protein interfaces.** *Sci STKE* 2004, **2004**(219):pl2.
13. Benedix A, Becker CM, de Groot BL, Caflisch A, Bockmann RA: **Predicting free energy changes using structural ensembles.** *Nat Methods* 2009, **6:**3–4.
14. Cho KI, Kim D, Lee D: **A feature-based approach to modeling protein-protein interaction hot spots.** *Nucl Acids Res* 2009, **37**(8):2672–2687.
15. Chen P, Li J, Wong L, Kuwahara H, Huang J, Gao X: **Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences.** *Proteins: Struct Funct Bioinformatics* 2013, **81**(8):1351–1362.
16. Ofran Y, Rost B: **Protein-protein interaction hotspots carved into sequences.** *PLoS Comput Biol* 2007, **3**(7):e119.
17. Grosdidier S, Recio JF: **Identification of hot-spot residues in protein-protein interactions by computational docking.** *BMC Bioinformatics* 2008, **9:**447.
18. Zhu X, Ericksen SS, Demerdash ONA, Mitchell JC: **Data-driven models for protein interaction and design.** *Proteins: Struct Funct Bioinformatics* 2013, **81**(12):2221–2228.
19. Guney E, Tuncbag N, Keskin O, Gürsoy A: **HotSprint: database of computational hot spots in protein interfaces.** *Nucl Acids Res* 2008, **36**(Database-Issue):662–666.
20. Tuncbag N, Gursoy A, Keskin O: **Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy.** *Bioinformatics* 2009, **25**(12):1513–1520.
21. Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N: **PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces.** *Nucl Acids Res* 2010, **38**(6):e86.
22. Pallara C, Jimenez-Garcia B, Perez-Cano L, Romero-Durana M, Solernou A, Grosdidier S, Pons C, Moal IH, Fernandez-Recio J: **Expanding the frontiers of protein-protein modeling: from docking and scoring to binding affinity predictions and other challenges.** *Proteins: Struct Funct Bioinformatics* 2013, **81**(12):2192–2200.
23. Deng L, Guan J, Wei X, Yi Y, Zhou S: **Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties.** In *Research in Computational Molecular Biology, Volume 7821 of Lecture Notes in Computer Science*. Edited by Deng M, Jiang R, Sun F, Zhang X. Springer: Berlin, Heidelberg; 2013:333–344.
24. Wang L, Liu ZP, Zhang XS, Chen L: **Prediction of hot spots in protein interfaces using a random forest model with hybrid features.** *Protein Eng Design Sel* 2012, **25**(3):119–126.
25. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D: **BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations.** *Nucleic Acids Res* 2013, **41**(Web Server issue):W333–W339.
26. Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, Rodrigues JPGLM, Trellet M, Bonvin AMJJ, Cui M, Rooman M, Gillis D, Dehouck Y, Moal I, Romero-Durana M, Perez-Cano L, Pallara C, Jimenez B, Fernandez-Recio J, Flores S, Pacella M, Praneeth Kilambi K, Gray JJ, Popov P, Grudinin S, Esquivel-Rodriguez J, Kihara D, Zhao N, Korkin D, Zhu X, et al.: **Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions.** *Proteins: Struct Funct Bioinformatics* 2013, **81**(11):1980–1987.
27. Kirkpatrick DG, Radke JD: **A framework for computational morphology.** In *Computational Geometry, Machine Intelligence and Pattern Recognition. Volume 2*; 1985:217–248.
28. Liu Q, Kwoh CK, Hoi SCH: **Beta atomic contacts: identifying critical specific contacts in protein binding interfaces.** *PLoS ONE* 2013, **8**(4):e59737.
29. Liu Q, Kwoh CK, Li J: **Binding affinity prediction for protein-ligand complexes based on β contacts and B factor.** *J Chem Inf Model* 2013, **53**(11):3076–3085.
30. Hubbard SJ, Thornton JM: **'NACCESS', computer program.** Tech. rep., Department of Biochemistry Molecular Biology, University College London, 1993.
31. Barber BC, Dobkin DP, Huhdanpaa H: **The Quickhull algorithm for convex hulls.** *ACM Trans Math Softw* 1996, **22**(4):469–483.
32. Mintseris J, Weng Z: **Atomic contact vectors in protein-protein recognition.** *Proteins* 2003, **53**(3):629–639.
33. Li J, Liu Q: **'Double water exclusion': a hypothesis refining the O-ring theory for the hot spots at protein interfaces.** *Bioinformatics* 2009, **25**(6):743–750.

34. Martins JM, Ramos RM, Pimenta AC, Moreira IS: **Solvent-accessible surface area: how well can be applied to hot-spot detection?** *Proteins: Struct Funct Bioinformatics* 2013, **82**(3):479–490.
35. Xia JF, Zhao XM, Song J, Huang DS: **APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility.** *BMC Bioinformatics* 2010, **11:**174.
36. Zhu X, Mitchell JC: **KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density and plasticity features.** *Proteins* 2011, **79**(9):2671–2683.

doi:10.1186/1471-2105-15-57
Cite this article as: Liu *et al.*: Integrating water exclusion theory into β contacts to predict binding free energy changes and binding hot spots. *BMC Bioinformatics* 2014 **15**:57.