4-2013

Beta Atomic Contacts: Identifying Critical Specific Contacts in Protein Binding Interfaces

Qian LU
Nanyang Technological University

Chee Keong KWOH
Nanyang Technological University

Steven C. H. HOI
Singapore Management University, chhoi@smu.edu.sg

Citation

LU, Qian; KWOH, Chee Keong; and HOI, Steven C. H.. Beta Atomic Contacts: Identifying Critical Specific Contacts in Protein Binding Interfaces. (2013). *PLoS ONE*. 8, (4), e59737-1-10.
Available at: https://ink.library.smu.edu.sg/sis_research/2296

Beta Atomic Contacts: Identifying Critical Specific Contacts in Protein Binding Interfaces

Qian Liu*, Chee Keong Kwoh, Steven C. H. Hoi*

BIRC, School of Computer Engineering, Nanyang Technological University, Singapore, Singapore

Abstract

Specific binding between proteins plays a crucial role in molecular functions and biological processes. Protein binding interfaces and their atomic contacts are typically defined by simple criteria, such as distance-based definitions that only use some threshold of spatial distance in previous studies. These definitions neglect the nearby atomic organization of contact atoms, and thus detect predominant contacts which are interrupted by other atoms. It is questionable whether such kinds of interrupted contacts are as important as other contacts in protein binding. To tackle this challenge, we propose a new definition called beta (β) atomic contacts. Our definition, founded on the β-skeletons in computational geometry, requires that there is no other atom in the contact spheres defined by two contact atoms; this sphere is similar to the van der Waals spheres of atoms. The statistical analysis on a large dataset shows that β contacts are only a small fraction of conventional distance-based contacts. To empirically quantify the importance of β contacts, we design βACV, an SVM classifier with β contacts as input, to classify homodimers from crystal packing. We found that our βACV is able to achieve the state-of-the-art classification performance superior to SVM classifiers with distance-based contacts as input. Our βACV also outperforms several existing methods when being evaluated on several datasets in previous works. The promising empirical performance suggests that β contacts can truly identify critical specific contacts in protein binding interfaces. β contacts thus provide a new model for more precise description of atomic organization in protein quaternary structures than distance-based contacts.

Citation: Liu Q, Kwoh CK, Hoi SCH (2013) Beta Atomic Contacts: Identifying Critical Specific Contacts in Protein Binding Interfaces. PLoS ONE 8(4): e59737. doi:10.1371/journal.pone.0059737

Editor: Franca Fraternali, King's College London, United Kingdom

Received January 9, 2012; **Accepted** February 21, 2013; **Published** April 22, 2013

Funding: This work was supported by Singapore MOE (Ministry Of Education) ARC (Academic Research Council) funding project (Tier-2 grant T208B2203). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: liuq0011@e.ntu.edu.sg (QL); chhoi@ntu.edu.sg (SCH)

Introduction

Specific binding between proteins plays a fundamental role in molecular functions and biological processes. The discovery of governing principles behind specific protein interactions is thus an essential issue in proteomics. Protein interactions are generally considered to be governed by their binding interfaces which consist of interfacial residues/atoms and their contacts. In order to uncover contributing factors to specific interactions, binding interfaces and their contacts are firstly quantified according to several widely-used criteria: some definitions consider atomic distance between atoms each from one protein [1–7], while another more complicated definition takes into account Voronoi diagrams of entire complexes [8–13]; the other criterion defines binding interfaces using the change of solvent accessible surface area (ΔASA) upon the formation of protein complexes [14–21]. Under these definitions, protein binding is found to be driven by forces from those atomic contacts such as hydrogen bonds, electrostatic interactions, van der Waals forces, salt bridges, hydrophobic attractions, etc.

However, these criteria define an interface simply as a cluster of spatially close atoms and their contacts but pay little attention to the local surroundings of its defined contacts. Thus, a lot of non-specific contacts are detected, which makes it still very difficult, if not completely impossible, to pinpoint the governing principles according to these existing contact definitions. One piece of evidence for the non-specific contacts is that these definitions will detect larger 'binding interfaces' in crystal packing, and it is very hard to distinguish these crystal-packing 'binding interfaces' from true ones. Here, crystal packing is the artifact of the crystallographic packing environments and is randomly formed during the crystallization process; but they do not occur in solution or in their physiological states [22]. With crystal packing as the reference state, a perfect contact definition is expected to satisfy that no or fewer contacts are detected in crystal packing; based on this definition, crystal packing should be easily distinguished from specific biological binding of proteins using a simple learning algorithm.

In this work, we propose a new definition: β atomic contacts. A β atomic contact β of atoms β and j must satisfy β-skeletons [23] where c's forbidden region contains no other atom. This forbidden region is defined by the parameter β. In this work, β is set to 1, defining a sphere with the midpoint of i and j as the center and with the spatial distance between i and j as the diameter (similar to the van der Waals spheres of atoms). Thus, our definition only detects "perceptually meaningful" contacts. We expect β atomic contacts to provide a more precise model of atomic organization in protein 3D structures than the previous definitions.

To demonstrate the efficacy of our β contacts in identifying critical atomic contacts in protein binding interfaces, we adopt β atomic contacts to define protein interfaces and then investigate the difference between homodimeric interfaces and crystal-packing interfaces. Many previous works also endeavored to detect distinguishing characteristics of crystal packing and specific biological binding. Some works have revealed a significant difference between protein surfaces and interfaces in amino acid composition, as well as a high similarity of protein surfaces to crystal packing [16,18,24–26]. Several other methods have been proposed to identify biological protein complexes from crystal packing. Both Weng's group [27] and Klebe's group [28] represented interfaces using atomic contact vectors (ACV), and then took them as inputs of machine-learning algorithms to construct efficient classifiers for distinguishing different types of protein binding, such as permanent and transient interactions and crystal packing [27,28]. PITA scored crystal packing using their contact size and chemical complementarity [29]. Zhu *et al.* [15] extracted six properties from interfaces, such as interface size, amino acid composition and gap volume, and then fed them into an SVM to train their NOXclass classifier to discriminate obligate and non-obligate interactions and crystal packing [15]. Using residue-based Voronoi tessellations of protein structures, Bernauer *et al.* constructed an SVM classifier DiMoVo for identifying biological protein interactions [11]. Taking the advantage of the hypothesis that energetically important residues are generally protected by the O-ring [30], Liu and Li designed the propensity vector of residue contacts within the O-ring to develop OringPV for the distinction between crystal packing and biological interactions and between two different types of biological interactions [31]. However, almost all of them use knowledge extracted from the simple definitions, such as defining interfaces using ASA change or defining interfacial contacts using a threshold of atomic distance.

In this work, we use β atomic contacts in interfaces to classify homodimers from crystal packing. In this classification, we represent an interface by an ACV [27] based on β contacts, and then design a new classifier, called β atomic contact vector (βACV). βACV is a linear SVM classifier with selected distinguishable types of β atomic contacts by SVM-RFE as input. Evaluated on several previous datasets, βACV achieves better classification than the ACV classifier simply based on the distance-based contacts, although β atomic contacts are only a small fraction of the contacts under the latter definition. Our βACV is also compared with several existing methods in the literature, including PISA, DiMoVo and NOXclass. The results demonstrate that β contacts are superior to these methods in most cases. All these comparisons suggest that β contacts are more capable of capturing specific binding contacts than the other definitions. A web server of the proposed βACV solution is also available at http://sunim1.sce.ntu.edu.sg/liuqian/bacv/index.py.

Materials and Methods

Datasets

Three datasets in the literature are used to comprehensively evaluate β atomic contacts.

The first *Bahadur* dataset contains 178 crystal packing and 113 biological homodimers from the previous works [16,19]. This dataset has been used to develop DiMoVo [11].

The second *Ponstingl* dataset has 95 crystal packing and 76 homodimers [32]. This dataset has been used in several existing works [27,28], including PITA [29] and PISA [33].

The third *non-redundant* dataset is compiled by [11] from the Bahadur, Ponstingl and NOXclass datasets [15]. In this non-redundant dataset, two proteins have no more than 30% sequence identity. This dataset includes 314 crystal packing and 144 homodimers after preprocessing [11].

What are β atomic contacts

As presented in the File S1, the various definitions of atomic contacts proposed in previous works have several limitations. To tackle these limitations, we propose a new definition–-β atomic contacts. Given a protein complex p, **an atomic contact, denoted as $c(i,j)$, between two atoms i and j is called a β contact if and only if** $d(i,j) \leq T_d \wedge c(i,j) \in e(VD(p)) \wedge c(i,j) \in e(b(p))$. Specifically, these three requirements are described as follows:

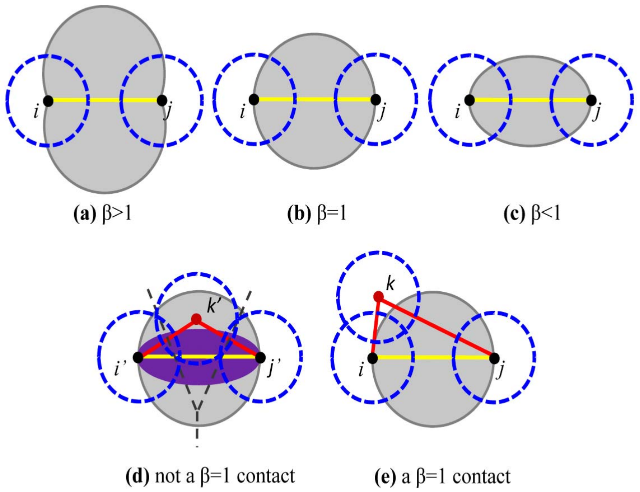
$d(i,j) \leq T_d$: This first requirement states that the surface distance between i and j, $d(i,j)$, must be less than or equal to T_d. Here, a surface distance of two heavy atoms is their Euclidian distance minus the sum of their van der Waals radii as defined in [34]. A similar surface distance definition is also used in [5,35,36]. For a simple description, the 'distance' in our definition and method is always calculated in this way except when otherwise specified. These contacts under this requirement are called distance-based contacts.

$c(i,j) \in e(VD(p))$: For the second requirement, i and j must share a Voronoi facet in $VD(p)$ where $VD(p)$ is the Voronoi diagram of p and $e(VD(p))$ is a set of edges in $VD(p)$. These contacts under the two requirements above are called Voronoi-based contacts.

$c(i,j) \in e(b(p))$: The last requirement indicates that $c(i,j)$ cannot break β-skeletons [23]. That is, $c(i,j)$ is an edge of the β-skeleton $b(p)$ of p where $e(b(p))$ is a set of edges in $b(p)$. These contacts under the three requirements above are called β atomic contacts.

A β-skeleton of a discrete set p is an undirected graph in computational geometry where two points i and j have an edge if any angle ikj is sharper than a threshold determined by $\beta, \forall k \in p, k \neq i,j$. β actually defines a forbidden region for the contact between i and j, just like the gray regions in Figure 1(a)–(c) with different β values. In Figure 1, two atoms i and j have an edge in β-skeletons if there are no other atoms k whose center is in their forbidden region. In other words, if there is any atom whose center is in the gray region, the atomic contact between i and j is interrupted, and the contact should not exist in β-skeletons. For example when $\beta = 1$, two atoms i' and j' do not have a β contact in Figure 1(d) because there is a k' in their forbidden region, while the contact between two atoms i and j is a β contact in Figure 1(e) since any atom k is outside the forbidden region. It is also interesting to note that in Figure 1(d), two atoms i' and j' have a Voronoi-based contact, but they only share a smaller-size facet (e.g., the dash gray line down in Figure 1(d)) which is also far away from the center region of the contact–-the center region of the contact between two atoms i' and j' is a small arch region very close around the contact, that is, the magenta region in Figure 1(d). Compared to this contact in the Voronoi-based definition, our β criterion assumes that two atoms should have enough contact area in their center region to form an important interaction.

In β-skeletons, with different values of β from bigger to smaller, the forbidden gray regions decrease as shown in Figure 1 from (a) to (c), and thus the number of atomic contacts in β-skeletons increases. When β is small enough, the contacts defined on β-skeletons are similar to those on Voronoi diagrams or even to those on distance-based definitions. In this work, β is set to 1, and this β-skeleton is also called the Gabriel graph [37,38]. We would like to emphasize that (i) β-skeletons are a totally different concept from β shape [39], a generalization of the α shape [40], which is also commonly used in the analysis of protein structures, such as



(a) β>1 (b) β=1 (c) β<1

(d) not a β=1 contact (e) a β=1 contact

Figure 1. β skeletons and β contacts. Three points, *i*, *j* and *k*, represent the atoms. The dash circles in blue represent van der Waals spheres in 2D space. The lines in yellow are of interest. In the first row, if *i* and *j* have a β contact, their surface distance is less than a threshold T_d and the gray regions are required to contain no other atom by β skeletons when $\beta > 1$ (left), $\beta = 1$ (center) and $\beta < 1$ (right), respectively. In (d), a region in magenta is the center area which is very close around the line in yellow, and the dash lines represent Voronoi facets.
doi:10.1371/journal.pone.0059737.g001

ASA calculation and protein shape detection; and (ii) β contacts can be either long-range contacts or short-range contacts in tertiary structures which are different contact definitions based on both sequence separation and spatial distance(please refer to the references [41–43] for the definitions of long-range contacts); but short- and long-range contacts focus on sequence separation while β contacts emphasize the spatial organization of atomic interactions.

Detecting β atomic contacts in protein 3D structures

A protein 3D structure can be modeled as a β atomic contact graph $b(p)$ where heavy atoms are considered as points, and β atomic contacts as edges. Given a protein 3D structure, its $b(p)$ can be produced by the following process.

First, Qhull is used to produce Delaunay triangulation [44] for all points. Second, a surface-distance threshold T_d is used to remove those atomic contacts whose distances are too large. T_d is set to 3.3 Å (the diameter of a water molecule 2.8 Å plus 0.5 Å). T_d is the maximum surface-distance between the van der Waals spheres of two atoms, as discussed in the β contact definition. Thirdly, each atomic contact is checked to guarantee that it satisfies β skeletons, that is, the Gabriel graph here. Since we are interested in atomic contacts between proteins, all atoms which have no contact across binding interfaces are removed.

β atomic contact vectors in protein interfaces

An atomic contact vector (ACV for short) [27] is adopted to represent an interface. In this vector representation, all heavy atoms of the twenty standard residues in proteins are grouped according to twelve atomic types in the File S1. These atomic types are similar to those in [45]. Hence, the atomic contact vector for an interface has 78 atomic pairs ($78=\frac{12\times 11}{2}+12$). The value for each pair is its occurrence in a β atomic contact graph when T_d is set to 3.3 Å. Since disulfide bonds are almost as strong as covalent bonds, disulfide bonds whose spatial distance of two sulfur atoms across interfaces is less than 2.6 Å are also considered as an atomic pair in the vector. Finally, the vector for a protein interface has 79 pairs, called βACV1a (β $\underline{a}$tomic $\underline{c}$ontact $\underline{v}$ectors) for short. Similarly, we also construct βACV1 in which the surface-distance threshold of two contact atoms is as small as 0.5 Å, that is, $T_d = 0.5$ Å.

In addition to atomic types, the distance between contact atoms is also an important factor in protein binding. Given two atomic pairs with the same types, one pair has a small distance between atoms, while atoms in the other pair have a much larger distance; the first pair generally has different importance to protein binding from the second pair. One example with this property is hydrogen bonds in interfaces: if a Nitrogen atom and an Oxygen atom have less than 3.5 Å spatial distance, their contact may be a hydrogen bond; but these two atoms cannot form a hydrogen bond directly if their spatial distance is too large, for example, more than 5 Å. Therefore, we take into account the surface-distance information of atomic contacts and split βACV1a into three sub-vectors: each of them contains atomic contacts whose surface-distance falls in one of the three regions: $\leq$0.5, (0.5,1.9] and (1.9,3.3], and they are named as $\leq$0.5 contacts, (0.5,1.9] contacts and (1.9,3.3] contacts for short. Here$1.9=0.5+2.8/2$. This vector representation has 235 pairs ($235=(\frac{12\times 11}{2}+12)\times 3+1$), which is referred to as βACV3 for short.

Meanwhile, to enable a fair comparison, distance-based ACV1 (dACV1), ACV1a (dACV1a) and ACV3 (dACV3) are also constructed in a similar way.

Our proposed classifier βACV and evaluation measures

We want to evaluate β contacts in classifying homodimers and crystal packing. In our classification task, a dataset of crystal packing and homodimers is represented by$D=\{(x_i,y_i)|y_i\in\{-1,1\}_{i=1}^{n}\}$, where $y_i=-1$ indicates that this vector is from crystal packing or $y_i=1$ indicates it is from homodimers; n is the total number of crystal packing and homodimers; x_i is the β atomic contact vector (βACV1, βACV1a or βACV3) for interface i. SVM with a linear kernel in LIBSVM [46] (the freely available SVM library) is then employed to train our classifier for identifying homodimers from crystal packing. A short description of SVM is provided in the File S1.

In our βACV classifier, SVM-RFE (a short description of RFE is provided in the File S1) is firstly used to find the feature set S_f with the best accuracy for the features βACV1 or βACV1a or βACV3. SVM-RFE uses SVM learning to obtain feature weight w_i and then removes features with the lowest value $\|w_i\|^2$ step-by-step until the predefined criteria are satisfied. This process uses a five-fold cross-validation. Then, two established ways are used to evaluate classification performance. One is feature-selection classification performance by using a leave-one-out cross-validation on the learning datasets. The other is the independent-dataset testing. That is, a βACV classifier with features S_f is constructed for predicting those complexes whose proteins have low sequence similarity to the complexes in the dataset for feature selection. The independent-testing datasets for the Bahadur, Ponstingl and NOXclass datasets can be found in [11].

Finally, *recall*(*r*.), *specificity*(*sp*.), *accuracy*(*acc*.), and Mathew's correlation coefficient *MCC* are adopted to evaluate the classification performance of β atomic contact vectors. Their definitions are provided in the File S1. MCC is more meaningful in a dataset which has a significant imbalance between the numbers of positive and negative samples.

Results and Discussion

Comparison of β atomic contacts with distance-based atomic contacts

To demonstrate that β contacts are better than distance-based contacts in describing protein binding interfaces, we compare these two types of definitions from the following aspects. We firstly measure the numbers of distance-based contacts and β atomic contacts to see their difference. We then compare their prediction performance on the three datasets. Following that, we provide a detailed comparative analysis of the selected features by RFE, especially of the top 10 atomic-contact features, for distance-based contacts and β contacts.

***β* atomic contacts are a small fraction of distance-based atomic contacts.** We perform statistical analysis of the number of atomic contacts under the three different definitions, distance-based, Voronoi-based, and β contacts. The analysis is based on the non-redundant dataset where $T_d = 3.3$. The result is shown in Table 1.

From Table 1, we find that β atomic contacts are only a small fraction of distance-based atomic contacts, i.e., 10.5% of distance-based contacts in homodimers, and 11.3% in crystal packing.

We also show the comparison of the atoms involved in β contacts with those in the complement of β contacts with respect to distance-based contacts in Figure 2. These atoms and their ASA are from the non-redundant dataset with $T_d = 3.3$. It seems that a lot of contact atoms in distance-based contacts are not defined to be in direct interfaces according to β contacts; Figure 2 also clearly shows that atoms only in distance-based contacts mostly have small ASA change after complex formation, while the atoms with larger ASA change are mostly in β contacts.

***β* atomic contacts and distance-based atomic contacts in the classification of crystal packing and homodimers.** We then compare β contacts with distance-based contacts in classifying crystal packing and homodimers. The results are shown in Tables 2 and 3.

It is clearly seen from Table 2 that β contacts aggregately have better performance than distance-based contacts, although β contacts are only a small fraction of distance-based contacts. In the nine pair-wise comparisons in Table 2 (three pairs of βACV1 versus dACV1, βACV3 versus dACV3, and βACV1a versus dACV1a on the three datasets), β contacts are superior in seven times, and distance-based contacts are superior in only one (there is one tie). For example on Bahadur, β-contact βACV3 has much better MCC, with 7.7 percent points higher than distance-based dACV3.

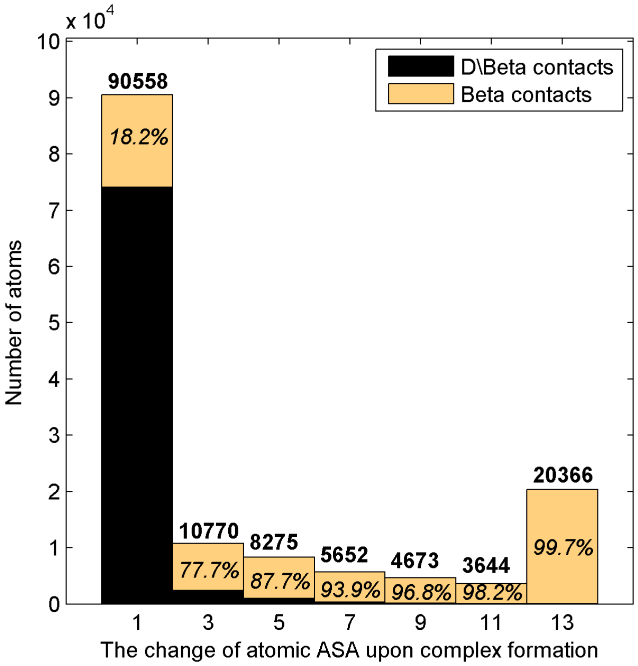


Figure 2. The ASA change (Å) of atoms in β contacts and in the complement of β contacts with respect to distance-based contacts (referred to as 'D\Beta contacts' for short in the figure). The integer in bold on a bar is the number of atoms whose ASA change falls in the region of the bar, while the percent in *italics* is the corresponding percentage of atoms only in β contacts.
doi:10.1371/journal.pone.0059737.g002

Table 3 presents the classification performance of β contacts and distance-based contacts on the independent datasets. In these nine comparisons, β contacts perform better in six times. Again, β contacts aggregately demonstrate better classification performance than distance-based contacts.

The difference between β contacts and distance-based contacts are also evaluated by $D\backslash\beta$ACV1 and $D\backslash\beta$ACV3 which only use the corresponding complement of β contacts with respect to distance-based contacts, that is, those atomic contacts not in βACV1 and βACV3 but in dACV1 and dACV3, respectively. The results are shown in Table 3. β contacts still achieve better performance on the independent datasets of NOXclass and Ponstingl and similar performance on the independent dataset of Bahadur. This similar performance should be due to the fact that this independent dataset is easily distinguished with high

Table 1. The difference of the numbers of distance-based, Voronoi-based and β atomic contacts for 114 homodimers and 314 crystal packing.

	in homodimers			**in crystal packing**		
	Distance-based	**Voronoi-based**	**β contacts**	**Distance-based**	**Voronoi-based**	**β contacts**
Voronoi-based	*508,792*	0	*71,126*	*265,293*	0	*49,425*
β contacts	*579,918*	*71,126*	0	*314,718*	*49,425*	0
Number of contacts	**647,878**	**139,086**	**67,960**	**354,652**	**89,359**	**39,934**
	(4,499.2±2,822.3)	(965.9±572.8	(471.9±286.4)	(1,129.5±533.7)	(284.6±125.0)	(127.2±57.2)

The numbers in *Italics* are the difference of the numbers of the contacts under the definitions of its column and row.
The number of the last two rows in **bold** is the number of atomic contacts under the definition of its column.
X±Y in last row: X is the mean of the number of atomic contacts in interfaces of a dataset while Y is the standard deviation.
doi:10.1371/journal.pone.0059737.t001

Table 2. The comparison of feature-selection classification performance achieved by distance-based and β atomic contacts and other methods in the literature.

Dataset		distance-based contacts			β contacts			DiMoVo	Ref [28]
		dACV1	dACV3	dACV1a	βACV1	βACV3	βACV1a		
Ponstingl	r.	0.895	0.895	0.908	0.895	0.934	0.921	-	-
	sp.	0.947	0.947	0.947	0.937	0.968	0.989	-	-
	acc.	**0.924**	0.924	0.93	0.918	**0.953**	**0.959**	-	0.948
	MCC	**0.846**	0.846	0.858	0.834	**0.905**	**0.918**	-	-
Bahadur	r.	0.840	0.877	0.821	0.887	0.925	0.877	0.890	-
	sp.	0.955	0.955	0.955	0.983	0.983	0.943	0.980	-
	acc.	0.911	0.926	0.904	**0.947**	**0.961**	**0.918**	0.945	-
	MCC	0.810	0.840	0.795	**0.887**	**0.917**	**0.825**	0.884	-
Nonredundant	r.	0.882	0.917	0.861	0.896	0.958	0.868	-	-
	sp.	0.978	0.971	0.978	0.984	0.994	0.975	-	-
	acc.	0.948	0.954	0.941	**0.956**	**0.983**	0.941	-	-
	MCC	0.877	0.893	0.862	**0.898**	**0.959**	0.862	-	-

r., sp., acc. and MCC represent recall, specificity, accuracy, and Matthew's correlation coefficient respectively. The **bold numbers** are the larger values in each of three pair-wise comparisons of βACV1 vs dACV1, βACV3 vs dACV3, and βACV1a vs dACV1a. In this table, significant features are selected on a dataset, and then a method with the selected features is evaluated on this dataset under a leave-one-out cross-validation process, as discussed in **Materials and Methods**.

doi:10.1371/journal.pone.0059737.t002

classification performance for both distance-based and β contacts (as shown in Table 3). Meanwhile, $D\backslash\beta$ACV3 has similar performance to dACV3, because the distance-based contacts predominate in their used contacts.

From Tables 2 and 3, we observe that (i) βACV1 can achieve good classification performance, which suggests that atomic contacts with smaller distance play a vital role in protein binding; (ii) βACV3 aggregately outperforms βACV1 and βACV1a, indicating that atomic contacts with relatively larger distance can also contribute to protein binding and atomic contacts with different distances have various contributions to protein interactions. In our web server, βACV3 has been made available to the scientific community.

In conclusion, β contacts are generally more capable in capturing critical specific binding contacts than distance-based contacts.

Analysis of the selected features in the classifications

Top 10 selected features in β atomic contacts and in distance-based atomic contacts. To deepen our understand-

Table 3. The comparison of classification performance on the independent datasets achieved by distance-based and β atomic contacts and other methods in the literature.

Dataset		distance-based contacts			β contacts			$D\backslash\beta$ contacts[1]		DiMoVo	PISA	PITA	NOXclass
Training on		dACV1	dACV3	dACV1a	βACV1	βACV3	βACV1a	$D\backslash\beta$ACV1[1]	$D\backslash\beta$ACV3[1]				
Ponstingl	r.	0.908	0.934	0.934	0.947	0.947	0.921	0.803	0.908	0.710	0.920	0.840	-
	sp.	0.934	0.929	0.898	0.938	0.934	0.938	0.863	0.929	0.920	0.760	0.910	-
	acc.	0.927	0.930	0.907	**0.940**	**0.937**	**0.934**	0.848	0.924	0.868	0.802	0.894	-
	MCC	0.815	0.827	0.780	**0.851**	**0.844**	**0.832**	0.627	0.808	0.643	0.602	0.730	-
Bahadur	r.	0.868	0.974	0.947	0.842	0.947	0.974	0.816	0.921	0.840	-	-	-
	sp.	0.986	0.978	0.986	1.000	0.971	0.971	1.000	0.986	0.950	-	-	-
	acc.	0.960	**0.977**	**0.977**	**0.966**	0.966	0.972	0.960	0.972	0.929	-	-	-
	MCC	0.880	**0.935**	**0.933**	**0.898**	0.902	0.920	0.881	0.915	0.780	-	-	-
NOXclass	r.	0.861	0.924	0.937	0.911	0.937	0.810	0.759	0.924	0.790	-	-	0.950
	sp.	0.909	0.841	0.868	0.909	0.882	0.877	0.714	0.836	0.970	-	-	0.680
	acc.	0.896	0.863	**0.886**	**0.910**	**0.896**	0.860	0.726	0.860	0.920	-	-	0.751
	MCC	0.745	0.702	**0.747**	**0.784**	**0.765**	0.659	0.424	0.697	0.790	-	-	0.556

r., sp., acc. and MCC represent recall, specificity, accuracy, and Matthew's correlation coefficient respectively. The **bold numbers** have the same meaning as those in Table 2. Here, significant features and a prediction method on them are trained on a dataset, and the evaluation is performed on another dataset, as discussed in **Materials and Methods**. [1] $D\backslash\beta$ contacts are the complement of β contacts with respect to distance-based contacts, while $D\backslash\beta$ACV1 and $D\backslash\beta$ACV3 are ACV vectors based on $D\backslash\beta$ contacts.

doi:10.1371/journal.pone.0059737.t003

Table 4. The top 10 features of distance-based and β atomic contacts when βACV3 and dACV3 are trained on the non-redundant dataset.

Rank	distance-based contacts		β contacts	
	types of atomic contacts	T_d range	types of atomic contacts	T_d range
1st	N_3H_1_$\mathbf{C_4H_3}$	(1.9,3.3]	$\mathbf{C_3\ (S_2)H_0}$_$\mathbf{C_4H_3}$	≤0.5
2rd	$\mathbf{C_4H_1}$_$\mathbf{C_4H_1}$	(1.9,3.3]	N_3H_2_$\mathbf{C_3\ (S_2)H_1}$	≤0.5
3th	$\mathit{O_2H_1}$_$\mathbf{C_4H_2}$	(1.9,3.3]	$\mathbf{C_4H_1}$_$\mathbf{C_4H_3}$	≤0.5
4th	N_3H_1_$\mathit{O_1H_0}$-	(1.9,3.3]	$\mathit{O_1H_0}$-_$\mathit{O_2H_1}$	(1.9,3.3]
5th	$\mathbf{C_3\ (S_2)H_0}$_$\mathbf{C_3\ (S_2)H_1}$	(1.9,3.3]	N_4H_3/2+_$\mathbf{C_4H_1}$	(1.9,3.3]
6th	N_3H_1_$\mathbf{C_4H_1}$	(1.9,3.3]	$\mathit{O_1H_0}$-_$\mathit{O_1H_0}$-	(1.9,3.3]
7th	$\mathit{O_1H_0}$_$\mathbf{C_4H_3}$	(1.9,3.3]	$\mathit{O_2H_1}$_$\mathbf{C_4H_2}$	≤0.5
8th	$\mathit{O_1H_0}$_$\mathit{O_1H_0}$	(1.9,3.3]	N_3H_1_$\mathit{O_1H_0}$	≤0.5
9th	$\mathbf{C_3\ (S_2)H_1}$_$\mathbf{C_3\ (S_2)H_1}$	(0.5,1.9]	$\mathbf{C_4H_3}$_$\mathbf{C_4H_3}$	(1.9,3.3]
10th	$\mathbf{C_4H_3}$_$\mathbf{C_4H_3}$	(1.9,3.3]	$\mathbf{C_4H_1}$_$\mathbf{C_4H_1}$	(0.5,1.9]

X_Y means atomic contacts between X and Y, while X and Y are atomic types in the File S1. Carbon atoms are in **bold**, while Oxygen atoms are in *italics*.

doi:10.1371/journal.pone.0059737.t004

ing of the difference between β atomic contacts and distance-based atomic contacts, we show the top 10 selected features of βACV3 and dACV3 by SVM-RFE in Table 4 when βACV3 and dACV3 are trained on the non-redundant dataset.

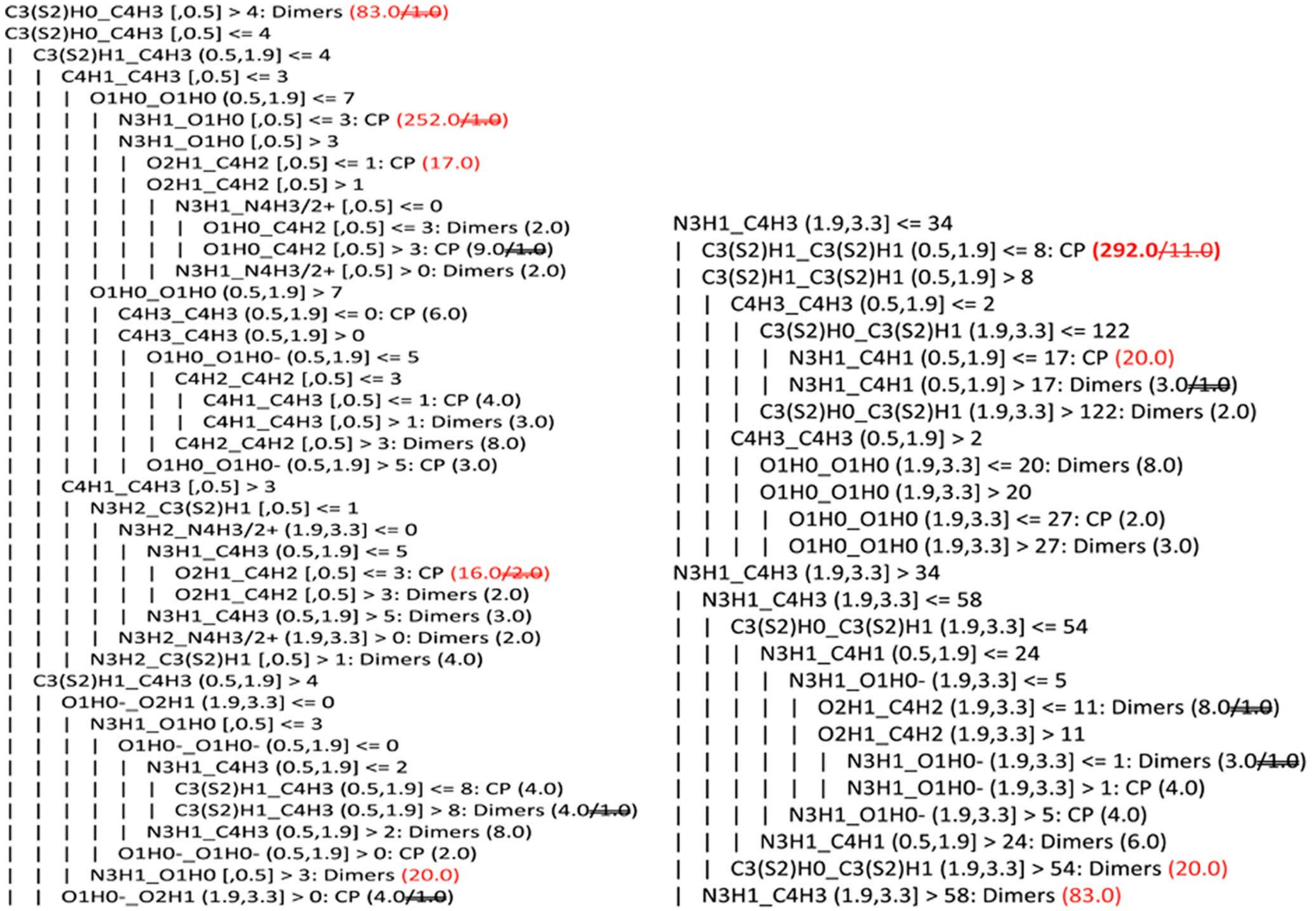
From Table 4, these top 10 features of β atomic contacts indicate two interesting phenomena. One is the hydrophobic effect--the contacts among Carbon atoms are chosen to be significantly important to biological binding, although their atomic types are different and their surface distances are ≤0.5 Å, or in (0.5,1.9] Å, or in (0.5,1.9] Å. The other is that hydrogen bonds, ≤0.5 contacts between N_3H_1 and O_1H_0 in β contacts, also play an important role in classifying biological binding from crystal packing.

The top 10 selected features of distance-based contacts capture contacts among Carbon atoms, but miss hydrogen bonds. Furthermore, most of the top features of distance-based contacts are those atomic contacts with relatively larger surface-distance, in (1.9,3.3] Å in Table 4. The main reason is the spatial constraint: given a sphere with an atom as the center, the larger the radius is, i.e., the larger surface-distance threshold T_d here, the more other atoms can be covered without atomic clashes. That is, in distance-based contacts, the number of contacts with the bigger surface-distance (1.9,3.3] Å is generally much larger than the number of contacts with the smaller surface-distance ≤0.5 Å. However, this does not hold in β contacts. For example on the non-redundant dataset, the number of ≤0.5 contacts in β contacts is 43,063, and the number of (1.9,3.3] contacts is 14,007. (1.9,3.3] contacts are about one-third of ≤0.5 contacts in β contacts. However in distance-based contacts, the number of ≤0.5 contacts and that of (1.9,3.3] contacts are 57,286 and 635,254; (1.9,3.3] contacts are over ten times more than ≤0.5 contacts in distance-based contacts. This misleads SVM and RFE to prefer the (1.9,3.3] contacts in distance-based contacts, since they have a higher occurrence.

With the discussion above in mind, one argument in distance-based contacts is: when the (1.9,3.3] contacts mislead SVM and RFE, why the SVM classifier does not have much worse performance. There are at least two helpful factors contributing to classifiers based on distance-based contacts. One is that distance-based contacts can easily represent atomic density in interfaces; the other is that interface contact size can also greatly help the classification performance of distance-based contacts. Both atomic density and contact size should be distinguishable features and a possibly necessary condition for biological binding; they are easily but indirectly implied in distance-based ACV3 vector, although the contacts are divided into different types. However, both of them should not be sufficient conditions for specific protein binding.

Decision trees of distinguishing features in β atomic contacts and in distance-based atomic contacts. To visualize the selected features by RFE and to provide some clues of governing principles underlying protein binding, we show the decision trees in Figure 3(a) for β contacts and in Figure 3(b) for distance-based contacts only using these selected features when βACV3 and dACV3 are trained on the non-redundant dataset. The details of how to construct decision tree are provided in the File S1. With these selected features, βACV3 and dACV3 achieve accuracy of 0.983 and 0.954; in the decision trees with 5-fold cross-validation, β contacts have accuracy of 0.891, and distance-based contacts have accuracy of 0.915. Since the important features in SVM cannot be guaranteed to be the same as those in the decision trees, we do not pay more attention to the similar performances of the two trees. Instead, we would like to see whether easily interpretable knowledge can be derived from these two decision trees, because SVM-based βACV3 has much better classification performance but poor interpretability.

Figure 3(a) suggests three interesting rules. One is about contacts between two Carbon atoms, called R1 in the first line of Figure 3(a). R1 suggests that if an interface has more than four ≤0.5 contacts between the atomic types $C_3\ (S_2)H_0$ and C_4H_3 ($\mathit{C_3\ (S_2)H_0}$_C_4H_3 for short), it has a probability of 98.8%(82/83) to be a biological binding. This rule is consistent with the hydrophobic effect. The other two interesting rules are closely related to hydrogen bonds. One hydrogen-bond-involving rule is: given an interface with less ≤0.5 $C_3\ (S_2)H_0$_C_4H_3, it can still be biological binding if this interface has: **(i)** more than four (0.5,1.9] $C_3\ (S_2)H_1$_C_4H_3 contacts, and **(ii)** more than three hydrogen bonds (≤0.5 N_3H_1_O_1H_0 contacts), and **(iii)** no (1.9,3.3] O_1H_0-_O_2H_1 contacts. In the non-redundant dataset, 20 biological binding interfaces and none of the crystal packing satisfy this rule. In contrast, 251 crystal packing and only one biological interface satisfy the other hydrogen-bond-involving rule (called R2 for



```
C3(S2)H0_C4H3 [,0.5] > 4: Dimers (83.0/~~1.0~~)
C3(S2)H0_C4H3 [,0.5] <= 4
|   C3(S2)H1_C4H3 (0.5,1.9] <= 4
|   |   C4H1_C4H3 [,0.5] <= 3
|   |   |   O1H0_O1H0 (0.5,1.9] <= 7
|   |   |   |   N3H1_O1H0 [,0.5] <= 3: CP (252.0/~~1.0~~)
|   |   |   |   N3H1_O1H0 [,0.5] > 3
|   |   |   |   |   O2H1_C4H2 [,0.5] <= 1: CP (17.0)
|   |   |   |   |   O2H1_C4H2 [,0.5] > 1
|   |   |   |   |   |   N3H1_N4H3/2+ [,0.5] <= 0
|   |   |   |   |   |   |   O1H0_C4H2 [,0.5] <= 3: Dimers (2.0)
|   |   |   |   |   |   |   O1H0_C4H2 [,0.5] > 3: CP (9.0/~~1.0~~)
|   |   |   |   |   |   N3H1_N4H3/2+ [,0.5] > 0: Dimers (2.0)
|   |   |   O1H0_O1H0 (0.5,1.9] > 7
|   |   |   |   C4H3_C4H3 (0.5,1.9] <= 0: CP (6.0)
|   |   |   |   C4H3_C4H3 (0.5,1.9] > 0
|   |   |   |   |   O1H0_O1H0- (0.5,1.9] <= 5
|   |   |   |   |   |   C4H2_C4H2 [,0.5] <= 3
|   |   |   |   |   |   |   C4H1_C4H3 [,0.5] <= 1: CP (4.0)
|   |   |   |   |   |   |   C4H1_C4H3 [,0.5] > 1: Dimers (3.0)
|   |   |   |   |   |   C4H2_C4H2 [,0.5] > 3: Dimers (8.0)
|   |   |   |   |   O1H0_O1H0- (0.5,1.9] > 5: CP (3.0)
|   |   C4H1_C4H3 [,0.5] > 3
|   |   |   N3H2_C3(S2)H1 [,0.5] <= 1
|   |   |   |   N3H2_N4H3/2+ (1.9,3.3] <= 0
|   |   |   |   |   N3H1_C4H3 (0.5,1.9] <= 5
|   |   |   |   |   |   O2H1_C4H2 [,0.5] <= 3: CP (16.0/~~2.0~~)
|   |   |   |   |   |   O2H1_C4H2 [,0.5] > 3: Dimers (2.0)
|   |   |   |   |   N3H1_C4H3 (0.5,1.9] > 5: Dimers (3.0)
|   |   |   |   N3H2_N4H3/2+ (1.9,3.3] > 0: Dimers (2.0)
|   |   |   N3H2_C3(S2)H1 [,0.5] > 1: Dimers (4.0)
|   C3(S2)H1_C4H3 (0.5,1.9] > 4
|   |   O1H0-_O2H1 (1.9,3.3] <= 0
|   |   |   N3H1_O1H0 [,0.5] <= 3
|   |   |   |   O1H0-_O1H0- (0.5,1.9] <= 0
|   |   |   |   |   N3H1_C4H3 (0.5,1.9] <= 2
|   |   |   |   |   |   C3(S2)H1_C4H3 (0.5,1.9] <= 8: CP (4.0)
|   |   |   |   |   |   C3(S2)H1_C4H3 (0.5,1.9] > 8: Dimers (4.0/~~1.0~~)
|   |   |   |   |   N3H1_C4H3 (0.5,1.9] > 2: Dimers (8.0)
|   |   |   |   O1H0-_O1H0- (0.5,1.9] > 0: CP (2.0)
|   |   |   N3H1_O1H0 [,0.5] > 3: Dimers (20.0)
|   |   O1H0-_O2H1 (1.9,3.3] > 0: CP (4.0/~~1.0~~)
```

(a) Decision tree of β contacts

```
N3H1_C4H3 (1.9,3.3] <= 34
|   C3(S2)H1_C3(S2)H1 (0.5,1.9] <= 8: CP (292.0/~~11.0~~)
|   C3(S2)H1_C3(S2)H1 (0.5,1.9] > 8
|   |   C4H3_C4H3 (0.5,1.9] <= 2
|   |   |   C3(S2)H0_C3(S2)H1 (1.9,3.3] <= 122
|   |   |   |   N3H1_C4H1 (0.5,1.9] <= 17: CP (20.0)
|   |   |   |   N3H1_C4H1 (0.5,1.9] > 17: Dimers (3.0/~~1.0~~)
|   |   |   C3(S2)H0_C3(S2)H1 (1.9,3.3] > 122: Dimers (2.0)
|   |   C4H3_C4H3 (0.5,1.9] > 2
|   |   |   O1H0_O1H0 (1.9,3.3] <= 20: Dimers (8.0)
|   |   |   O1H0_O1H0 (1.9,3.3] > 20
|   |   |   |   O1H0_O1H0 (1.9,3.3] <= 27: CP (2.0)
|   |   |   |   O1H0_O1H0 (1.9,3.3] > 27: Dimers (3.0)
N3H1_C4H3 (1.9,3.3] > 34
|   N3H1_C4H3 (1.9,3.3] <= 58
|   |   C3(S2)H0_C3(S2)H1 (1.9,3.3] <= 54
|   |   |   N3H1_C4H1 (0.5,1.9] <= 24
|   |   |   |   N3H1_O1H0- (1.9,3.3] <= 5
|   |   |   |   |   O2H1_C4H2 (1.9,3.3] <= 11: Dimers (8.0/~~1.0~~)
|   |   |   |   |   O2H1_C4H2 (1.9,3.3] > 11
|   |   |   |   |   |   N3H1_O1H0- (1.9,3.3] <= 1: Dimers (3.0/~~1.0~~)
|   |   |   |   |   |   N3H1_O1H0- (1.9,3.3] > 1: CP (4.0)
|   |   |   |   N3H1_O1H0- (1.9,3.3] > 5: CP (4.0)
|   |   |   N3H1_C4H1 (0.5,1.9] > 24: Dimers (6.0)
|   |   C3(S2)H0_C3(S2)H1 (1.9,3.3] > 54: Dimers (20.0)
|   N3H1_C4H3 (1.9,3.3] > 58: Dimers (83.0)
```

(b) Decision tree of distance-based contacts

Figure 3. Decision tree of β contacts on the non-redundant dataset. Each line is a branch in decision trees; '|' and an indent represent a sub-branch; 'Dimers' indicates a class label of biological binding, while 'CP' refers to a class label of crystal packing; means the number of misclassified complexes in a branch; the format of a line is: types of atomic contacts with their surface-distance information, followed by a splitting rule and a class label if possible. For example, the rule 'C_3 $(S_2)H_0$_C_4H_3 [,0.5] $>$4: Dimers (83.0)' in the first line of Figure 3(a) suggests the following prediction: 83 interfaces have more than four $\leq$0.5 contacts of C_3 $(S_2)H_0$ and C_4H_3 in the non-redundant dataset; among these interfaces, only one is crystal packing.
doi:10.1371/journal.pone.0059737.g003

short). As shown in Figure 3(a), R2 requires an interface with **(i)** less than five $\leq$0.5 C_3 $(S_2)H_0$_C_4H_3 contacts, **(ii)** less than five (0.5,1.9] C_3 $(S_2)H_1$_C_4H_3 contacts, **(iii)** less than four $\leq$0.5 C_4H_1_C_4H_3 contacts, **(iv)** less than eight (0.5,1.9] O_1H_0_O_1H_0 contacts, and **(v)** not more than three hydrogen bonds ($\leq$0.5 N_3H_1_O_1H_0 contacts). This rule is a little more complicated, but it is reasonable: there should be more than one type of specific important contacts to protein binding, and enough occurrence of several of them in an interface should produce a true biological binding; hence, a false biological binding should not have enough critical specific contacts, which must exclude all potential combinations of specific important contacts. This makes a complicated rule unique to crystal packing.

According to these three rules of β contacts in Figure 3(a), we believe that the following contacts should be closely related to specific important contacts to protein binding: the $\leq$0.5 contacts of C_4H_3 with C_3 $(S_2)H_0$ and with C_4H_1, (0.5,1.9] C_4H_3_C_3 $(S_2)H_1$ contacts, and $\leq$0.5 N_3H_1_O_1H_0 contacts. These contacts are consistent with previous observations: $\leq$0.5 N_3H_1_O_1H_0 contacts generally can be hydrogen bonds; C_4H_3 are mostly in the side chains of such hydrophobic residues as Val, Ile and Leu, indicating the hydrophobic effect; C_3 $(S_2)H_1$ are almost in the aromatic side chains, providing π-involving interactions in binding interfaces. In contrast, it is crystal packing, not biological binding, which prefer (1.9,3.3] O_1H_0-_O_2H_1 contacts. Further in Figure 3(a), higher occurrence of contacts involving O_1H_0- almost suggests crystal packing prediction. O_1H_0- should play a destructive role in binding interfaces unless it can form salt bridges.

Similarly, distance-based contacts in Figure 3(b) also suggest two interesting rules. One is that only 83 biological binding have more than fifty-eight (1.9,3.3] N_3H_1_C_4H_3 contacts; the other is that in those interfaces which have not more than thirty-four (1.9,3.3] N_3H_1_C_4H_3 contacts, 281 crystal packing and 11 biological binding have not more than eight (0.5,1.9] C_3 $(S_2)H_1$_C_3 $(S_2)H_1$ contacts. These two rules are simple. However, the second rule has 11 false negative predictions, while the first rule is much hard to interpret according to our current biological knowledge.

Finally, we would like to note that β contacts are all in distance-based contacts, and then those contacts in the interesting rules of β contacts are in fact also in distance-based contacts. But distance-based contacts have so many non-β contacts, masking the detection of critical specific atomic contacts in the interesting rules of β contacts.

Two misclassified examples in the decision tree of β contacts. In the decision tree of β contacts, each of the two interesting rules, R1 and R2, has a misclassified interface. According to R1 in the first line of Figure 3(a), 82 biological interfaces have more than four $\leq$0.5 C_3 $(S_2)H_0\_C_4H_3$ contacts, and only one out of 314 crystal packing interfaces satisfies R1 in the non-redundant dataset. This misclassified crystal packing interface is 1RB3 as shown in Figure 4(a). This interface has five $\leq$0.5 C_3 $(S_2)H_0\_C_4H_3$ contacts. However, there are contradictory conclusions on this interface. On one hand, the authors who determined 1RB3 in PDB recommended it as a dimeric unit; it seems that the decision tree of β contacts shares the conclusion with the original authors of 1RB3. On the other hand, the non-redundant dataset labels 1RB3 as 'crystal packing'; the existing classifiers in the literature, such as NOXclass, DiMoVo and PISA, also predict it as a crystal packing. However, 1RB3's interface size is as small as 615 Å^2, which is much smaller than the cut-off of interface ASA 856 Å^2 suggested by the previous work [32] to distinguish crystal packing from homodimers. Meanwhile, almost all of the classifiers in the literature have the same bias–they heavily rely on interface size. That is, 1RB3's smaller interface size can easily mislead their predictions: these classifiers are more likely to have wrong predictions for 1RB3, and tend to consider 1RB3 as a crystal packing according to the same bias of 1RB3's smaller interface size. Thus, these predictions provide nothing more than 1RB3's smaller interface size. In summary, 1RB3 may be a potential dimer, just as R1 and the original authors suggest, but whether it is actually a dimer remains a question until it can be verified in wet-lab experiments.

The second rule R2 is in the sixth line of Figure 3(a), which covers 251 crystal packing and only one biological binding in the non-redundant dataset. This biological binding is 3SDH whose interface is shown in Figure 4(b). This 3SDH interface has plenty of interfacial water molecules, and large-size non-standard residues as shown in Figure 4(b). These two kinds of molecules are not evaluated in β contacts so far, which may be the reason why the decision tree of β contacts misclassifies 3SDH.

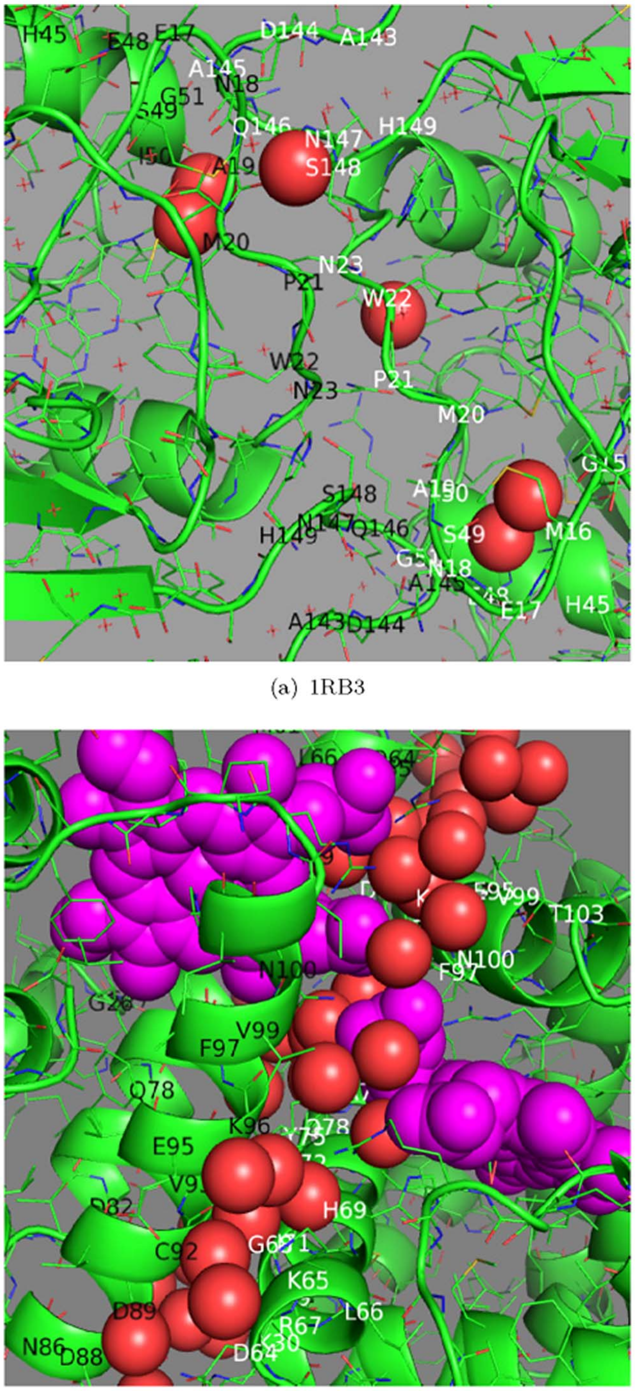
β contacts outperform previous methods in the classification of crystal contacts and homodimers

Feature-selection classification performance. The feature-selection classification performance of β atomic contacts is evaluated against those achieved by DiMoVo [11], and Block's method [28]. The best classification performance of the Block's method and the DiMoVo prediction are shown in Table 2.

In Table 2, our β contacts trained on the Ponstingl dataset have better classification performance than the best performance (accuracy 0.948) of the Block's method. When the Block's method uses SVM, its best accuracy is 0.919 with an RBF kernel, which is much less than our accuracy 0.959.

On the Bahadur dataset, DiMoVo's performance is recalculated by using its recalls for homodimers and crystal packing in [11]. β contacts have at least comparable, if not better, performance with DiMoVo. However, our βACV1a and βACV3 on β contacts adopt a linear kernel which is simpler than the RBF kernel used in DiMoVo.

Classification performance on the independent datasets. In addition, the classification performance of β atomic contacts on the independent datasets is also compared with those achieved by other methods in the literature, including DiMoVo, PISA, PITA [29], and NOXclass. Their classification results are shown in Table 3 where the performance of DiMoVo, PISA, PITA, and NOXclass on the independent datasets is recalculated by using the recall and specificity and the datasets in [11]. Here,



(a) 1RB3

(b) 3SDH

Figure 4. Two misclassified examples in the decision tree of β contacts (better viewed in color). (a) The crystal packing in 1RB3 is misclassified as biological binding by the biological rule R1. (b) The biological interface in 3SDH follows the crystal packing rule R2. In (a) and (b), the residues labeled in black (chain a) and white (chain b) form an interface; interfacial waters whose spatial distances to both chains are less than 3.5 Å are in the red sphere view; non-standard residues are in the magenta sphere view; Carbon: green; Oxygen: red; Nitrogen: blue.

doi:10.1371/journal.pone.0059737.g004

the independent dataset has protein complexes whose proteins have less than 30% sequence similarity to those proteins of complexes in the training dataset.

Trained on the Ponstingl dataset, our βACV3 of β contacts have accuracy of 0.937, and MCC of 0.844, which are much higher than those achieved by DiMoVo, PISA and PITA. For example, PITA achieves the best accuracy of 0.894 and MCC of 0.73 among DiMoVo, PISA and PITA; its accuracy is 4.3 percent points lower than βACV3's accuracy, and its MCC is 11.4 percent points lower than βACV3's MCC.

When βACV1a and βACV3 of β contacts are trained on the Bahadur dataset, they again achieve better performance than DiMoVo. Our βACV3 has 0.966 accuracy, 3.7 percent points higher than DiMoVo's, and βACV3 has 0.902 MCC, 12.2 percent points higher than DiMoVo's. In this case, MCC is a better metric than accuracy to compare β contacts with DiMoVo, since the independent-testing dataset of Bahadur is quite unbalanced: crystal packing is about four times larger than homodimers. Hence, the great improvement of MCC suggests that βACV3 is much better than DiMoVo to capture protein specific binding.

When the NOXclass dataset is the training dataset, β contacts have much better performance than NOXclass, although β contacts cannot achieve better performance than DiMoVo. A reason for this is that β contacts can easily distinguish crystal packing from homodimers in the NOXclass dataset. Removal of several features does not change training accuracy significantly, which misleads SVM and SVM-RFE into choosing all features or fewer features as the best feature set. However, the samples in the NOXclass's independent dataset are much harder to distinguish.

In conclusion, β contacts demonstrate its superior classification power to the other methods in the literature under non-β contact definitions. This partially, if not entirely, results from the fact that the new β contact definition can capture specific binding patterns in homodimers and then benefit the classification of homodimers from crystal packing.

Conclusion

The main contribution of this work is to propose the novel concept of β atomic contacts to identify critical specific contacts across protein binding interfaces. To evaluate the efficacy of the proposed β contacts, we design a new classification scheme βACV for classifying crystal packing and homodimers. We compare βACV's classification performance with those achieved by the existing methods on the three datasets. The promising performance achieved by βACV demonstrates that β contacts can truly identify a compact set of critical specific contacts in protein binding interfaces which are only a small fraction of conventional distance-based contacts. Thus, β atomic contacts provide a new fundamental and precise unit for atomic organization in computational structural analysis. In future, β atomic contacts have many other applications, such as the estimation of folding and binding free energy, the prediction of binding hot spots, protein docking as well as other structural analyses for folding and binding of proteins and RNA/DNA. In these potential applications, one should pay more attention to repacking, as the exact positioning of residues is particularly important to β contacts.

Supporting Information

File S1 Introduction of methods and measures, and more discussion of beta contacts.
(PDF)

Acknowledgments

We thank Piyushkumar A. Mundra for discussion on SVM-RFE.

Author Contributions

Conceived and designed the experiments: QL. Performed the experiments: QL. Analyzed the data: QL. Contributed reagents/materials/analysis tools: QL SCH CKK. Wrote the paper: QL SCH CKK.

References

1. Ofran Y, Rost B (2003) Analysing six types of protein-protein interfaces. J Mol Biol 325: 377–387.
2. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 260: 604–620.
3. Lawrence MC, Colman PM (1993) Shape complementarity at protein/protein interfaces. J Mol Biol 234: 946–950.
4. Larsen TA, Olson AJ, Goodsell DS (1998) Morphology of protein-protein interfaces. Structure 6: 421–427.
5. Preissner R, Goede A, Frommel C (1998) Dictionary of interfaces in proteins (DIP). data bank of complementary molecular surface patches. J Mol Biol 280: 535–550.
6. Korkin D, Davis FP, Sali A (2005) Localization of protein-binding sites within families of proteins. Protein Sci 14: 2350–2360.
7. Davis FP, Sali A (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics 21: 1901–1907.
8. Headd JJ, Ban YEA, Brown P, Edelsbrunner H, Vaidya M, et al. (2007) Protein-protein interfaces: Properties, preferences, and projections. J Proteome Res 6: 2576–2586.
9. Cazals F, Proust F, Bahadur RP, Janin J (2006) Revisiting the Voronoi description of proteinprotein interfaces. Protein Sci 15: 2082–2092.
10. Bouvier B, Grünberg R, Nilges M, Cazals F (2009) Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics, and composition. Proteins 76: 677–692.
11. Bernauer J, Bahadur RPP, Rodier F, Janin J, Poupon A (2008) DiMoVo: a voronoi tessellationbased method for discriminating crystallographic and biological protein-protein interactions. Bioinformatics 24: 652–8.
12. Li Z, Li J (2010) Geometrically centered region: A "wet" model of protein binding hot spots not excluding water molecules. Proteins 78: 3304–3316.
13. McConkey B, Sobolev V, Edelman M (2002) Quantification of protein surfaces, volumes and atomatom contacts using a constrained Voronoi procedure. Bioinformatics 18: 1365–1373.
14. Cho KI, Kim D, Lee D (2009) A feature-based approach to modeling protein-protein interaction hot spots. Nucl Acids Res 37: 2672–2687.
15. Zhu H, Domingues FS, Sommer I, Lengauer T (2006) NOXclass: prediction of protein-protein interaction types. BMC Bioinformatics 7.
16. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 336: 943–955.
17. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N (2001) Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 43: 89–102.
18. Jones S, Thornton JM (1997) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272: 121–132.
19. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Dissecting subunit interfaces in homodimeric proteins. Proteins 53: 708–719.
20. Chakrabarti P, Janin J (2002) Dissecting protein-protein recognition sites. Proteins 47: 334–343.
21. Gong S, Park C, Choi H, Ko J, Jang I, et al. (2005) A protein domain interaction interface database: InterPare. BMC Bioinformatics 6: 207.
22. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. Brief Bioinform 10: 217–232.
23. Kirkpatrick DG, Radke JD (1985) A framework for computational morphology. Computational Geometry, Machine Intelligence and Pattern Recognition, 2: 217–48.
24. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285: 2177–2198.
25. Janin J, Miller S, Chothia C (1988) Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 204: 155–164.
26. Carugo O, Argos P (1997) Protein-protein crystal-packing contacts. Protein science 6: 2261–3.
27. Mintseris J, Weng Z (2003) Atomic contact vectors in protein-protein recognition. Proteins 53: 629–639.
28. Block P, Paern J, Hullermeier E, Sanschagrin P, Sotriffer CA, et al. (2006) Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. Proteins 65: 607–622.

29. Ponstingl H, Kabir T, Thornton JM (2003) Automatic inference of protein quaternary structure from crystals. Journal of Applied Crystallography 36: 1116–1122.
30. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280: 1–9.
31. Liu Q, Li J (2010) Propensity vectors of low-ASA residue pairs in the distinction of protein interactions. Proteins 78: 589–602.
32. Ponstingl H, Henrick K, Thornton JM (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 41: 47–57.
33. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372: 774–797.
34. Hubbard SJ, Thornton JM (1993) 'NACCESS', computer program. Technical report, Department of Biochemistry Molecular Biology, University College London.
35. Jain AN (1996) Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. Journal of Computer-Aided Molecular Design 10: 427–440.
36. Trott O, Olson AJ (2010) AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31: 455–461.
37. Gabriel KR, Sokal RR (1969) A new statistical approach to geographic variation analysis. Systematic Zoology (Society of Systematic Biologists) 18: 259–270.
38. Matula DW, Sokal RR (1980) Properties of gabriel graphs relevant to geographic variation research and clustering of points in the plane. Geogr Anal 12: 205–222.
39. Kim DS, Seo J, Kim D, Ryu J, Cho CH (2006) Three-dimensional beta shapes. Computer-Aided Design 38: 1179–1191.
40. Edelsbrunner H, Mücke EP (1994) Three-dimensional alpha shapes. ACM Trans Graph 13: 43–72.
41. Selvaraj S, Gromiha M (2003) Role of hydrophobic clusters and long-range contact networks in the folding of (alpha/beta)$_8$ barrel proteins. Biophys J 84: 1919–1925.
42. Wu S, Zhang Y (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 24: 924–931.
43. Chen P, Li J (2010) Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. BMC Structural Biology 10: S2.
44. Barber BC, Dobkin DP, Huhdanpaa H (1996) The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software 22: 469–483.
45. Tsai J, Taylor R, Chothia C, Gerstein M (1999) The packing density in proteins: standard radii and volumes. J Mol Biol 290: 253–266.
46. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines.