The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Identifying Protein Binding Hot Spots by Using Deeply Buried Atomic Contacts

Zhenhua Li[1] Jinyan Li[1,*]

[1]Bioinformatics Research Center, School of Computer Engineering,
Nanyang Technological University, Nanyang Avenue, Singapore 639798

Abstract Solvent accessible surface area (SASA) is an important descriptor of protein residues and atoms. It is widely used as an outstanding feature in hot spot prediction by many computational methods. However, SASA is not capable of distinguishing slightly buried residues, of which most are non hot spots, and deeply buried ones that are usually inside a hot spot. In this work, we propose a new descriptor for residues, atoms and for atomic contacts, namely "burial level", which can capture the depth the residues are buried. We identified the number of different kinds of the deeply buried atomic contacts at different burial level that are directly broken in alanine substitution, and we used these values as input for SVM to classify between hot spot or non hot spot residues. We got an F measure of 0.6237 under the leave-one-out cross-validation on a data set containing 258 mutations. This performance is better than other computational methods.

Keywords hot spot, burial level, SASA

1 Introduction

Protein-protein interactions which play a key role in life are generally dominated by hydrogen bonds, salt bridges and hydrophobic contacts across the interface [18, 20]. These local interactions have to be desolvated, densely packed and hence deeply buried to make contribution to the binding free energy [40, 34, 14, 32]. This is why the energetically important hot spot residues in the interface tend to cluster into local regions with low solvent accessible surface area (SASA) values [3, 19].

Identifying these energetically important residues, which can offer useful information to protein engineering and better understanding on protein-protein interaction [28], is usually done by site-directed alanine mutagenesis. This experimental method mutates the target residue into alanine which only but still has a C^{β} heavy atom on its side-chain [9, 38]. Residues whose mutation results in large binding free energy change ($\geq$2.0 kcal/mol, for example) are defined as hot spot residues [8].

Many feature-based [16, 11, 6, 36, 39] and energy-based [27, 21, 22, 25] computational approaches have been proposed to address the hot spot prediction problem. Almost all of these feature-based methods use the SASA information of the residue as a critical feature in the prediction. A low SASA is necessary for a residue to be a hot spot residue, however, it is not sufficient as a large number of non hot spot residues also have low SASA

*Corresponding author. Email: jyli@ntu.edu.sg

values. Therefore, the SASA is not effective to tell the difference between slightly buried residues, a large part of which are non hot spot residues, and deeply buried residues that are very likely to be hot spot residues.

In this work, we introduce a new descriptor for protein atoms and residues. It is named burial level, a more informative concept than SASA. In the definition of burial level, the buried immobilized water molecules are treated as an integral part of the protein complex. We show that our definition of residue burial level is nicely correlated to $\Delta\Delta G$. A high burial level is not only in general necessary for hot spot residues but also more sufficient for them in comparison to SASA. In other words, most hot spot residues tend to have high burial level while most non hot spot residues are exposed or just slightly buried. We also define the burial level of atomic contacts and we calculate the number of three types of buried interfacial atomic contacts that are directly broken when the residue is substituted by alanine at different burial level. The number of those deeply buried atomic contacts together with the burial level of the residue itself are further fed into SVM as features to classify interfacial residues into hot spot residues or non hot spot residues. By applying our method to a data set of 258 mutations, we achieved an F measure of 0.6237 under the hot spot definition of $\Delta\Delta G \geq 2.0$ kcal/mol, which is better than other computational methods. We also conducted a detailed analysis of the features used in this work, and we found that the hot spot residues tend to have significantly more deeply buried atomic contacts than non hot spot residues.

2 Materials and Methods

2.1 Data set

Our data set is collected by retrieving the experimental alanine mutagenesis data from the alanine scanning energetics database (ASEdb) [35] and some previously published works [10, 13, 7, 30, 31]. The requirement for the data is that: the 3D structure of the wild-type protein complex is solved by X-ray crystallography and is reported in PDB [2], and the solvent information is included in the PDB file. We do not consider protein-ligand interaction or protein-peptide interaction in this work thus those interactions that do not have an extended interface are excluded. The structural similarity of the complexes are tested by the CE algorithm [33]. If the two chains of the two complexes have a significant similarity at the same time, their binding interfaces are further checked to ensure that there is no redundancy in the data set. Only the mutations in the interface are considered.

Our data set consists of 258 mutations distributed in 13 protein complexes. Hot spot residues are usually defined by setting $\Delta\Delta G \geq 1.0$ kcal/mol or $\Delta\Delta G \geq 2.0$ kcal/mol. We prefer to the second choice, as only a higher $\Delta\Delta G$ threshold can reflect the direct influence of the mutation. That is, the interfacial atomic contacts that are directly broken by the mutation are taken into consideration. Some researchers even suggested that a residue should have a $\Delta\Delta G$ higher than 4.0 kcal/mol so as to have a strong impact on the binding of the two proteins [28]. In practice, a lower value is taken to get enough data for statistical analysis [28].

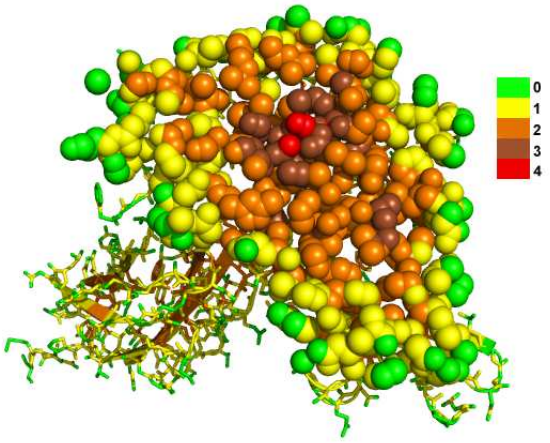


Figure 1: Cross-section of a growth hormone and growth hormone receptor complex (PDB id: 1A22) showing the burial level pattern. A layer of atoms in the front are shown as spheres while those at the back are shown as sticks.

2.2 Feature generation

2.2.1 Burial level of atom, residue and atomic contact

Our definition of burial level is based on atomic contact graph. The atomic contact graph of a protein complex is an undirected graph with heavy atoms as nodes and atomic contact as edges. The atoms in this graph is labeled as exposed or buried according to its SASA. If the SASA of an atom is not less than 10.0Å^2, it is exposed, otherwise it is buried. The SASA is calculated by the NACCESS software which is based on the Lee and Richards algorithm [23]. All the *exposed* water molecules, which we consider as part of the bulk solvent, are removed in advance while the buried water molecules are kept as a part of the complex, thus their oxygen atoms are a part of the atomic contact graph. The atomic contact is defined by a distance threshold and the Voronoi diagram. Two atoms are considered to be in contact if they have a distance less than their Van der Waals radius plus the diameter of a water molecule (2.75Å) and they share a Voronoi facet. This distance threshold is based on a water-free idea and it has been used in [24].

In an atomic contact graph, the burial level of an atom is defined as the length of the shortest path from this atom to its nearest exposed atom. For example, the burial level of exposed atoms is 0 and the burial level of their immediate buried neighbors is 1. We calculate the burial levels by adding a pseudo node, which represents bulk solvent, to the atomic graph. This node is connected to all of the exposed nodes directly. Then the burial level of any atom equals to the length of the shortest path from this atom to the pseudo node minus 1. This is exactly the single-source-shortest-path problem and it can be simply solved by using Dijkstra's algorithm[12].

Figure 1 shows a burial level pattern inside a growth hormone and growth hormone receptor complex. As can be seen from the figure, atom burial level is indeed a good indicator to describe the extent to which an atom is buried inside a protein or a protein

complex. It is clear that the burial level of any two neighboring atoms can have a difference of at most 1. Because the complex is not perfectly global, it can be noted that burial level 2 is "thicker" with more atoms.

The burial level of a residue is the average value of the burial levels of all atoms in the residue. For an atomic contact, if the burial levels of the two atoms are the same, the burial level of the atomic contact equals to the burial level of the two atoms, otherwise it equals to the smaller one of the two burial levels.

2.2.2 Directly broken atomic contact, atomic contact types and features

When a residue is mutated into alanine, some interfacial atomic contact are directly broken because of the removal of certain atoms; Some other interfacial atomic contacts may also be broken or distorted due to the conformational change in the local region [31]. Given a residue, we just consider its directly broken interfacial atomic contacts, which are the contacts that are made by atoms other than C, N, O, C^{α} and C^{β} and their contact partners are from the other chain.

We classify the atomic contacts into three types. If a contact is between a positively charged atom and a negatively charged atom, which usually corresponds to a salt bridge, it is called a Type I contact. If a contact is between a hydrogen bond donor and a hydrogen bond acceptor, which usually is a hydrogen bond, it is classified as Type II contact. Contacts that are neither Type I nor Type II are classified as Type III. Here, the definitions for positively charged atoms, negatively charged atoms, hydrogen bond donors and hydrogen bond acceptors are followed from those in [19]. We do not further divide the Type III contacts into subtypes such as other polar contact, hydrophobic contact and so on because they are all not that specific as Types I and II contact. Note that the definitions for Type I and Type II contacts are not exactly the same as salt bridges and hydrogen bonds in terms of the geometrical requirements, yet they can be still very important [29].

In this work, we use those deeply buried atomic contacts whose burial level is not less than 2. We refer the atomic contacts at burial level 0 as exposed atomic contacts and those at burial level 1 as slightly buried atomic contacts. Let $C(i, j)$ denote the number of Type i directly broken interfacial atomic contacts at burial level j of a residue. Then our model contains 6 features to describe a residue: $C(I,\geq 2)$, $C(II,2)$, $C(II,\geq 3)$, $C(III,2)$, $C(III,\geq 3)$ and the burial level of the residue. SVM model based on this feature set is named DBAC (**D**eeply **B**uried **A**tomic **C**ontacts). For comparison, we have also built another model named AC (**A**tomic **C**ontacts) which is based on another feature set containing $C(I,0)$, $C(I,1)$, $C(I,\geq 2)$, $C(II,0)$, $C(II,1)$, $C(II,2)$, $C(II,\geq 3)$, $C(III,0)$, $C(III,1)$, $C(III,2)$, $C(III,\geq 3)$ and the burial level of the residue. The maximum value of burial level depends on the size of the protein complexes and the size of the interfaces. In general, very few contacts have burial level larger than 3, so we do not distinguish further when the burial level are larger than 3. For Type I contact, there are very few cases that have burial level larger than 2, thus we do not use $C(I,\geq 3)$ as another feature but merge it with $C(I,2)$ into $C(I,\geq 2)$.

2.3 Support Vector Machine training-testing protocol

Support Vector Machines (SVMs) are widely used in many classification and regression problems. They have also been adopted in the hot spots prediction problems [6, 39, 25] with various feature sets and training-testing protocols. In this work, we take

the LIBSVM software [5] which is a tool for SVM model training and testing available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. The kernel we used is the radial basis function (RBF). We did not conduct feature selection because our method is straightforward, and the number of features is not large. Meanwhile a feature selection on the whole data set will introduce over-fitting. However, we do have evaluated the performance on two different feature sets: using the deeply buried atomic contacts only (DBAC) and using all the atomic contacts (AC), where the latter feature set is evaluated just for comparison.

The performance is evaluated under the leave-one-out cross-validation. To avoid the over-fitting, we have strictly followed a nested-loop cross-validation procedure. There are 258 mutations in our data set, each time one mutation is taken as the test data and the remaining 257 mutations are used to train the model. The two parameters, namely cost and gamma, are optimized on the training data by a grid search. The grid search evaluates the performance, F measure, of SVMs with different parameter values on the training data under the 5-fold cross-validation, and the parameter values with the best performance are chosen to build a training model on the training data. This training model is then applied to the test data, that is, the mutation held out in advance. This procedure is repeated 258 times till every mutation in the data set is tested.

2.4 Metrics in performance evaluation and statistical analysis

The performance is measured by sensitivity, precision, specificity, accuracy and F measure (F1) that are defined as follows: $sensitivity = \frac{TP}{TP+FN}$, $precision = \frac{TP}{TP+FP}$, $specificity = \frac{TN}{TN+FP}$, $accuracy = \frac{TP+TN}{TP+FP+TN+FN}$, and $F1 = \frac{2 \times sensitivity \times precision}{sensitivity+precision}$, where TP, FP, TN and FN are the number of true positives, false positives, true negatives and false negatives, respectively. A better classifier should predict the hot spot residues with less false positives and less false negatives, thus the F measure which combines sensitivity and specificity is used to indicate the overall performance.

We also test the significance of difference in $\Delta\Delta G$ values of the predicted hot spot and non hot spot residues. Basically, a classifier divides the mutations in the data set into two groups: computational hot spot residues and computational non hot spot residues. The significance of $\Delta\Delta G$ value difference in this two groups are tested by Mann-Whitney test [26]. The result of a classifier with higher F1 value can be less significant when its false positives have very low $\Delta\Delta G$ values (near 0 kcal/mol or even negative) and false negatives have high $\Delta\Delta G$ values.

We have also checked the value distribution of individual features in hot spot and non hot spot residues. The significance of difference in the two classes is also tested by Mann-Whitney test.

3 Results and Discussion

3.1 Performance of hot spot residue prediction

As introduced, 5 new features are derived from those deeply buried interfacial atomic contacts which are directly broken by the alanine substitution. The feature values of a residue are then fed into SVM together with the overall residue burial level to predict whether this residue is a hot spot residue or not. The performance under the leave-one-out cross-validation is shown in the second row of Table 1. We achieved an F measure of

Table 1: Performance of our method (DBAC) in comparison with using all atomic contact (AC) and Robetta

Method	Sensitivity	Precision	Specificity	Accuracy	F1	p-value
DBAC	0.58	0.6744	0.9327	0.8643	0.6237	3.0280×10^{-12}
AC	0.32	0.5333	0.9327	0.8140	0.4	1.2849×10^{-5}
Robetta	0.44	0.3667	0.8173	0.7442	0.4	5.3817×10^{-8}

Table 2: Comparison of our method with MINERVA

Method	Sensitivity	Precision	Specificity	Accuracy	F1	p-value
DBAC	0.5833	0.7	0.9366	0.8652	0.6364	2.4011×10^{-9}
MINERVA	0.5	0.6667	0.9366	0.8483	0.571	1.3731×10^{-7}

0.6237, when $\Delta\Delta G \geq 2.0$ was used a threshold to define hot spot residues. The precision of our method is higher than the recall, which means that there are fewer false positives than false negatives. A reason for this is that our model emphasizes the contribution of directly broken atomic contacts. The contacts that are broken or newly formed by the conformational change during the mutation are hard to be defined quantitatively. The $\Delta\Delta G$ of some hot spot residues whose mutation may result in large conformational change, cannot be fully explained by its directly broken atomic contacts. This is reflected in the lower sensitivity value. The non hot spot residues, whose $\Delta\Delta G$ is low, tend to have fewer directly broken atomic contacts, leading to a smaller number of false positives and hence a higher precision.

We have also evaluated the performance by using the AC feature set which takes those exposed and slightly buried atomic contacts into consideration as well. The performance is shown in the third row of Table 1. It can be noted that the performance is not improved although the extra exposed and slightly buried atomic contacts are added to the feature set; rather the F measure is driven down to 0.4 and the statistical significance is reduced a lot as well. The reason will be presented later.

We have compared our method with Robetta [21, 22], a widely recognized gold standard for benchmark comparison in the field. Robetta server can predict the $\Delta\Delta G$ value of interfacial residues by computational alanine scanning based on an energetic function. Its performance on our data set is shown in the forth row of Table 1. Our performance is remarkably better than that of Robetta in terms of both F1 and p-value.

We have also compared our method with another machine learning method, MINERVA [6], which uses SVM as well and is based on a larger feature set containing various aspects of information of target residue such as weighted atomic packing density, relative surface area burial, weighted hydrophobicity and so on. MINERVA has a good performance in terms of the F1 value as reported in comparison with other previous machine-learning methods. Because its source code and software are not available, we compare the performance on the common data that is reported in the MINERVA paper and that is also covered in our data set. This common data set contains 178 mutations with 36 hot spot residues and 142 non hot spot residues. The comparative performances is shown in Table 2, from which it can be noted that MINERVA does not outperform our method in terms of

Table 3: Statistical analysis on the features. The features in the feature set DBAC are emphasized in italics.

Feature	*RBL**	C(I, 0)	C(I, 1)	*C(I, ≥2)*
p-value	5.0897×10^{-10}	0.2008	0.8204	*0.0050*
Feature	C(II, 0)	C(II,1)	*C(II, 2)*	*C(II, ≥3)*
p-value	0.0013	0.4133	1.2419×10^{-9}	3.5031×10^{-6}
Feature	C(III, 0)	C(III,1)	*C(III,2)*	*C(III,≥3)*
p-value	0.0034	0.0945	1.9061×10^{-16}	1.1621×10^{-9}

*:residue burial level

F measure and p-value. Besides, MINERVA conducted a feature selection based on the whole data set, which selects only 12 out of 54 features. This procedure can introduce much over-fitting to the performance. MINERVA also has been tested on an independent data set derived from BID [15] where the importance of a residue is labeled as “strong”, “intermediate”, “weak” or “insignificant”. As in MINERVA, a residue is regarded as a hot spot residue only when its label is “strong”. We have tested our model (trained on our data set, 258 mutations) on a subset of this independent data set of which the corresponding PDB files have the solvent information reported, containing 111 mutations. The performance of our model in terms of F1 on this data set is 52%, same as that of MINERVA. Note that the label of of a residue is not perfectly correlated to its $\Delta\Delta G$, which is an inconsistency between the training data and the test data, however it still can indicate the contribution of a residue to the binding.

3.2 Feature analysis

We tested the significance of difference for the values of a feature in the hot spot and non hot spot residues. These p-values are reported in Table 3. It is clear that the DBAC features have very low p-values, indicating that their values are significantly different between the two classes. The p-value of $C(I, \geq 2)$ is not as low as that of other DBAC features, because residues that have salt bridges are fewer. The numbers of slightly buried atomic contacts are not that significant as those of the deeply buried ones. Thus our feature set can indeed reflect the contrast between hot spot residues and non hot spot residues, and the idea of excluding slightly buried and exposed atomic contacts and instead using deeply buried atomic contacts is statistically reasonable.

3.2.1 Residue burial level

Residue burial level is a very important feature to predict hot spot residues. Its p-value shows the most significant difference between hot spot residues and non hot spot residues as can be seen in Table 3. Here, we explain that residue burial level is more sufficient than SASA in hot spot prediction.

Bogan and Thorn [3] found that the hot spot residues tend to have low SASA values. Based on this observation, they suggested the existence of a ring of energetically less important residues that are responsible of protecting the hot spot. Generally a low SASA value is a necessary condition for a residue to become a hot spot residue, thus it is usually used someway in the hot spot prediction. For example, the HotSprint database [16] defines

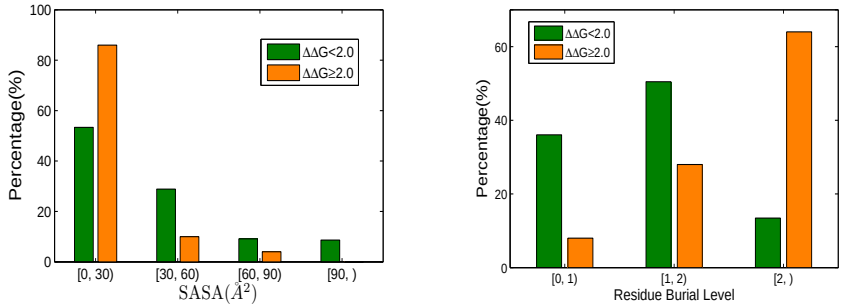


Figure 2: Distribution of the SASA values (a), and of the residue burial levels (b), in the hot spot and non hot spot residues.

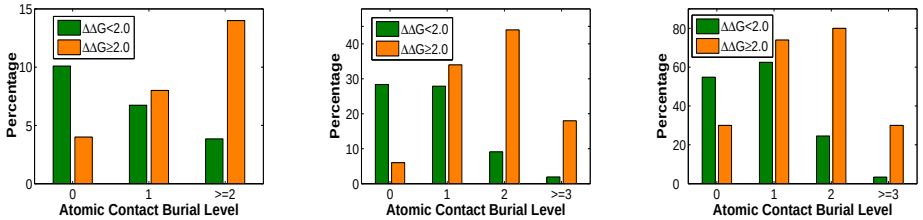


Figure 3: Percentage of hot spot residues and non hot spot residues that have at least one Type I (a), Type II (b) or Type III (c) directly broken interfacial atomic contact at different burial level.

the computational hot spots as those conserved residues that have large SASA change in complex formation and low SASA in the complex. However, a low SASA value is not a sufficient condition of hot spot. As observed in Figure 2(a), in our data set, the hot spot residues tend to have low SASA values with more than 80% of the hot spot residues having SASA less than 30 $\AA^2$. But the non hot spot residues also follow such a tendency (55%) in a less remarkable yet observable way.

In contrast, as shown in Figure 2(b), the hot spot residues tend to have a high burial level, while non hot spot residues do not. More than 60% of the hot spot residues have a burial level no less than 2.0, whereas the non hot spot residues with such burial levels are less than 20%. Thus, we can conjecture that a high burial level is not only necessary but also more sufficient than a low SASA value for a hot spot residue.

3.2.2 Deeply buried atomic contacts

Type I atomic contacts roughly correspond to salt bridges. Some researchers believe that buried salt bridges provide neutral or even negative contribution to protein stability [1, 37] because the desolvation of charged groups requires more energy than the interaction

energy of the formation of the salt bridge [17]. But in protein-protein interaction, it is found that interfacial salt bridges are more buried than intra-chain salt bridges, and the salt bridges are found favorable across the interface [40]. This may be the reason that the two proteins are folded independently with more charged residues exposed and the conformation change of them during the complex formation is very restricted, thus the two proteins prefer to interact in an electrostatic complementary manner.

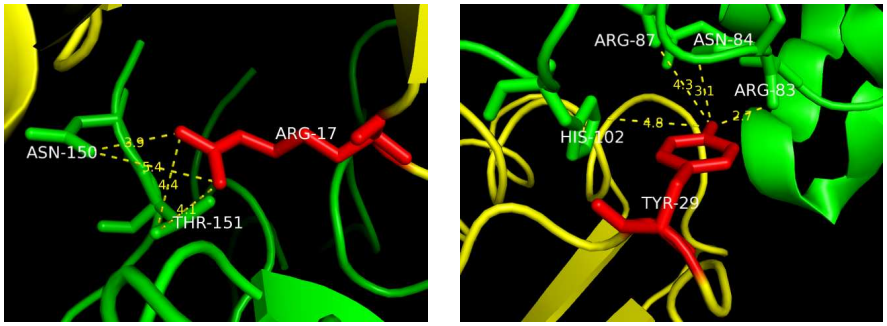
Percentages of hot spot residues and non hot spot residues whose $C(I,x)$ are larger than 0 are plotted in Figure 3(a). Generally the hot spot residues tend to have their salt bridges buried while non hot spot residues do not. Two adjacent exposed oppositely charged groups may not form stable salt bridges at all [34], thus some exposed Type I contacts, most of which are possessed by non hot spot residues, may not be stable salt bridges.

The hydrogen bonds indeed play a key role in protein-protein interaction [20]. Most interfacial hydrogen bonds are extremely buried and the more buried a hydrogen bond donor/acceptor is, the more likely it is to form a hydrogen bond [40]. Thus being buried is favorable for interfacial hydrogen bonds. Figure 3(b) shows the percentages of hot spot residues and non hot spot residues whose $C(II,x)$ are larger than 0. It can be seen that nearly 30% of the non hot spot residues have exposed Type II atomic contacts, but very few of them have deeply buried hydrogen bonds. There are more residues have deeply buried hydrogen bonds while a few of them have exposed hydrogen bonds. The number of residues that have extremely buried (burial level$\geq$ 3) atomic contacts is limited by the size of the protein complexes.

Type III contacts contain all other kinds of contacts that are neither salt bridges nor hydrogen bonds, including hydrophobic contacts and other polar contacts. Actually hydrophobic contacts are not specific contact between atoms but are the packing of groups of hydrophobic side chains. The contribution of hydrophobic contacts to bonding free energy is correlated with the buried surface area [32]. Thus energetically important hydrophobic contacts are those buried ones. Generally the protein-protein interface are dominated by salt bridges, hydrogen bonds and hydrophobic contacts, but sometimes other contacts also make contribution to the binding [40]. A hot spot is usually a densely packed region in the interface, thus the number of buried contacts of a hot spot residue tend to be large, which can be reflected by deeply buried Type III contacts. As shown in Figure 2(c), more than 80% of the hot spot residues have Type III contact at burial level 2 and only about 20% non hot spot residues have Type III contact at this burial level.

3.3 Case study: two residues that are difficult to classify

Figure 4 shows the structure of two residues that are difficult to classify. ARG-17 of BPTI shown in Figure 4(a) is well buried in the interface of the complex with a very low SASA of 8.0Å^2, a small SASA value even not enough to define an exposed atom. Argnines are actually very likely to be hot spot residues [3, 28] especially when they have such a low SASA. However, this ARG-17 is a non hot spot residue, having a $\Delta\Delta$G of only 0.5 kcal/mol. Its burial level is 1.55, which is not a high value and more importantly, we found that almost all its atomic contacts with bovine chymotrypsin are just slightly buried or even exposed. There are 4 Type II contacts as shown in the figure with 2 exposed and 2 slightly buried. It also has another 15 Type III contacts, with 13 slightly buried, 1 exposed and 1 deeply buried. We successfully classified this residue as non hot spot residue.



(a) (b)

Figure 4: Two residues that are mutated in our data set. (a) ARG-17 of BPTI (in yellow) in bovine chymotrypsin (in green)-BPTI complex (PDB id: 1CBW). (b) TYR-29 of barstar (in yellow) in barnase (in green)-barstar complex (PDB id: 1BRS).

Another example as shown in Figure 4(b) is TYR-29 of barstar. This residue is exposed with an SASA of 67.11Å^2, however it is a hot spot residue with a $\Delta\Delta$G of 3.4 kcal/mol. There are only two hot spot residues that have an SASA larger than 60Å^2 in our data set. We can still successfully identify it as a hot spot residue by using its deeply buried atomic contacts. The side-chain of tyrosine which contains an aromatic ring and a hydroxyl group is capable of forming aromatic π-interactions and hydrogen bonds [3]. As can be seen from the figure, although TYR-29 of barstar is partially exposed, its side-chain stretches into the complex and forms many deeply buried atomic contacts. For the 4 Type II interfacial atomic contacts shown in the figure, 3 are deeply buried and 1 is slightly buried. There are another 8 deeply buried Type III contacts, of which 7 are made by the aromatic ring and 5 are atomic contacts with HIS-102, an active site residue of barnase [4].

4 Conclusion

We have proposed a feature-based method to predict protein-binding hot spots by using the deeply buried interfacial atomic contacts that are directly broken during alanine substitution. The method is based on a graph theoretical definition of burial level of residues, atoms and atomic contacts. We achieved an F measure of 0.6237 when $\Delta\Delta$G$\geq$2.0 is used as the threshold to define hot spot residues.

The burial level of a residue is more intuitive than the concept of SASA, and it is nicely correlated with the $\Delta\Delta$G of a residue. We have shown that a high residue burial level is in general necessary for a residue to be a hot spot residue. And more importantly, it is more sufficient than SASA that is frequently used in hot spot prediction. Our results also reveal that the hot spot residues tend to have deeply buried atomic contacts while the non hot spots tend to have exposed and slightly buried ones. This is consistent with previous studies that emphasize the energetic contribution of buried salt bridges, hydrogen bonds and hydrophobic contacts.

Acknowledges

This research work was funded by a Singapore MOE ARC Tier-2 grant (T208B2203).

References

[1] S. Albeck, R. Unger, and G. Schreiber. Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. *Journal of Molecular Biology*, 298(3):503 – 520, 2000.

[2] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[3] Andrew A. Bogan and Kurt S. Thorn. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280(1):1–9, July 1998.

[4] Ashley M. Buckle, Gideon Schreiber, and Alan R. Fersht. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0Å resolution. *Biochemistry*, 33(30):8878–8889, August 1994.

[5] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] Kyu-Il Cho, Dongsup Kim, and Doheon Lee. A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Research*, 37(8):2672–2687, May 2009.

[7] T. Clackson, M. H. Ultsch, J. A. Wells, and A. M. de Vos. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *Journal of Molecular Biology*, 277(5):1111–1128, April 1998.

[8] T. Clackson and J. A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, January 1995.

[9] BC Cunningham and JA Wells. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, 244(4908):1081–1085, 1989.

[10] Brian C. Cunningham and James A. Wells. Comparison of a structural and a functional epitope. *Journal of Molecular Biology*, 234(3):554 – 563, 1993.

[11] Steven J. Darnell, David Page, and Julie C. Mitchell. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins: Structure, Function, and Bioinformatics*, 68(4):813–823, September 2007.

[12] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, December 1959.

[13] DA Dougan, RL Malby, LC Gruen, AA Kortt, and PJ Hudson. Effects of substitutions in the binding surface of an antibody on antigen affinity. *Protein Engineering*, 11(1):65–74, 1998.

[14] A. Fernández and H. A. Scheraga. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc Natl Acad Sci U S A*, 100(1):113–118, January 2003.

[15] T. B. Fischer, K. V. Arunachalam, D. Bailey, V. Mangual, S. Bakhru, R. Russo, D. Huang, M. Paczkowski, V. Lalchandani, C. Ramachandra, B. Ellison, S. Galer, J. Shapley, E. Fuentes, and J. Tsai. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, 19(11):1453–1454, July 2003.

[16] E. Guney, N. Tuncbag, O. Keskin, and A. Gursoy. HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Research*, 36(Database):D662–D666, December 2007.

[17] Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein science : a publication of the Protein Society*, 3(2):211–226, February 1994.

[18] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, January 1996.

[19] O. Keskin, B. Ma, and R. Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, 345(5):1281–1294, February 2005.

[20] Ozlem Keskin, Attila Gursoy, Buyong Ma, and Ruth Nussinov. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chemical reviews*, 108(4):1225–1244, April 2008.

[21] Tanja Kortemme and David Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, 99(22):14116–14121, October 2002.

[22] Tanja Kortemme, David E. Kim, and David Baker. Computational alanine scanning of protein-protein interfaces. *Sci. STKE*, 2004(219):pl2+, February 2004.

[23] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, 55(3):379–400, February 1971.

[24] Jinyan Li and Qian Liu. 'double water exclusion': a hypothesis refining the O-ring theory for the hot spots at protein interfaces. *Bioinformatics*, 25(6):743–750, March 2009.

[25] Stefano Lise, Cedric Archambeau, Massimiliano Pontil, and David Jones. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics*, 10(1):365+, October 2009.

[26] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[27] Irina Massova and Peter A. Kollman. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *Journal of the American Chemical Society*, 121(36):8133–8143, September 1999.

[28] Irina S S. Moreira, Pedro A A. Fernandes, and Maria J J. Ramos. Hot spots-A review of the protein-protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, 68(4):803–812, June 2007.

[29] C. Nick Pace, Gerald R. Grimsley, and J. Martin Scholtz. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *Journal of Biological Chemistry*, 284(20):13285–13289, May 2009.

[30] A. O. Pineda, A. M. Cantwell, L. A. Bush, T. Rose, and E. Di Cera. The thrombin epitope recognizing thrombomodulin is a highly cooperative hot spot in exosite I. *The Journal of Biological Chemistry*, 277(35):32015–32019, August 2002.

[31] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym, and G. Schreiber. The modular architecture of protein-protein binding interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 102(1):57–62, January 2005.

[32] K. A. Sharp, A. Nicholls, R. F. Fine, and B. Honig. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science*, 252(5002):106–109, April 1991.

[33] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–747, September 1998.

[34] D. P. Sun, U Sauer, H. Nicholson, and B. W. Matthews. Contributions of engineered surface salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. *Biochemistry*, 30(29):7142–53, July 1991.

[35] K. S. Thorn and A. A. Bogan. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–285, March 2001.

[36] Nurcan Tuncbag, Attila Gursoy, and Ozlem Keskin. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25(12):1513–1520, June 2009.

[37] C. D. Waldburger, J. F. Schildbach, and Sauer R. T. Are buried salt bridges important for protein stability and conformational specificity? *Nature Structural Biology*, 2:122 – 128, 1995.

[38] James A. Wells. Systematic mutational analyses of protein-protein interfaces. 202:390 – 411, 1991.

[39] Jun F. Xia, Xing M. Zhao, Jiangning Song, and De S. Huang. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*, 11(1):174, 2010.

[40] D. Xu, C. J. Tsai, and R. Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng*, 10(9):999–1012, September 1997.