Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Identification of hot spot residues at protein-protein interface

Lei Li, Bing Zhao, Zhanhua Cui, Jacob Gan, Meena Kishore Sakharkar and Pandjassarame Kangueane*

School of mechanical and aerospace engineering, Nanyang Technological University, Singapore - 639798;
Pandjassarame Kangueane* - Email: mpandjassarame@ntu.edu.sg; * Corresponding author

received February 4, 2006; revised March 27, 2006; accepted March 29, 2006; published online April 4, 2006

Abstract:

It is known that binding free energy of protein-protein interaction is mainly contributed by hot spot (high energy) interface residues. Here, we investigate the characteristics of hot spots by examining inter-atomic sidechain-sidechain interactions using a dataset of 296 alanine-mutated interface residues. Results show that hot spots participate in strong and energetically favorable sidechain-sidechain interactions. Subsequently, we describe a novel, yet simple 'hot spot' prediction model with an accuracy that is similar to many available approaches. The model is also shown to efficiently distinguish specific protein-protein interactions from non-specific interactions.

Key words: protein-protein interaction; interface analysis; hot spot residues; inter-atomic interaction

Abbreviations:

BPTI = bovine pancreatic trypsin inhibitor; NA = number of atoms in a residue involving sidechain-sidechain interactions at the interface; NA_l = number of atoms with legitimate (favorable) sidechain-sidechain contacts; NA_{il} = number of atoms with illegitimate (unfavorable) sidechain-sidechain contacts; NC_l = number of legitimate (favorable) sidechain-sidechain contacts; NC_{il} = number of illegitimate (unfavorable) sidechain-sidechain contacts; PDB = protein databank; vdW = van der Waals; NPV = native predictive value; PPV = positive predictive value; SN = sensitivity; SP = specificity

Background:

Many biological processes such as signal transduction, transport, cellular motion and regulatory mechanisms are mediated by protein-protein interactions. The study of protein-protein interactions has gained momentum for deciphering the specificity of protein-protein interfaces. Many parameters (e.g. interface hydrophobicity, residue frequencies and pairing preferences at interface) have been defined to describe interface features. **[1, 2, 3, 4, 5, 6, 7]** Recently, the contribution of individual residues to subunit interactions have been estimated using alanine-scanning mutagenesis, where the mutation of a target residue to alanine is followed by the measure of ΔΔG (change in binding free energies), as described elsewhere. **[8]** The binding free energy is observed to be dominantly contributed by high energy residues, called 'hot spots'. **[9, 10]** For example, at the BPTI-trypsin interface, hot spot Lys15->Ala mutation (ΔΔG = 10 kcal·mol^{-1}) leads to a 200-fold decrease in association rate, while low energy residue ARG17->ALA (ΔΔG < 0.5 kcal·mol^{-1}) has little effect on association rate. **[11]** Therefore, interface specificity is effectively determined by hot spots.

Because hot spots are a good indicator of interface specificity, their characteristics have been widely investigated. **[10, 12, 13, 14, 15, 16, 17, 18]** Hot spots are enriched in TRP, TYR and ARG and are often surrounded by hydrophobic rings to occlude bulk solvent. **[10]** In addition, hot spots statistically correlate with structurally conserved residues in ten protein families. **[12]** Moreover, hot spots from different monomers prefer to interact and their couplings are structurally conserved. **[13]** It has also been found that hot spots are related to central interface residues using the small-world network approach (proteins represented as networks, residues as nodes and interactions as edges). **[14, 15]**

In recent years, a number of computational methods have been developed to predict hot spots. These methods are classified into two types: (1) energy-based; and (2) structure-based. In the energy-based methods, functions are developed to calculate a residue's ΔΔG by simulating residue mutation to alanine. **[19, 20, 21, 22, 23]** These methods give good qualitative prediction results. However, high computational cost and the difficulty in operation (e.g. data processing) make them unsuitable for easy implementation. A good example of structure-based methods is the one described by Gao and colleagues. **[24]** In this method, interface residues are covered by a grid box and the contribution by each residue to binding affinity is estimated by rolling different kinds of probes (representing hydrophobic group, hydrogen bonds) over the grids close to the residue. Thus, residues having high energy contribution are predicted as hot spots. This method is subject to complex structural analysis and comparison. Despite these developments, a simple, robust 'hot spot' prediction model is still unavailable. Here, we describe the analysis and the grouping of 296 alanine-scanned interface residues into three types (hot spots, warm and unimportant residues) towards the development of a novel hot spot prediction model.

Methodology:

Definition of interface residues

ASA (Solvent-accessible surface area) of a residue was calculated using the program NACCESS. **[25]** A residue with an interface area (ΔASA) > 1Å^2 is defined as an interface residue and ΔASA is the change in ASA of the residue upon protein dimer formation from monomer state.

Dataset of alanine-mutated interface residues

A dataset of 296 alanine-mutated interface residues (Supplementary table 1: column 1, 2 and 3) derived from 15 protein-protein complexes (Supplementary table 2) was obtained from ASEdb (Alanine Scanning Energetics database). **[26]** These residues have ΔΔG in the range -0.9 - 10 kcal·mol^{-1}. The dataset was classified into three groups: hot spots (ΔΔG ≥ 1.5 Kcal·mol^{-1}), warm residues (0.5 – 1.5 Kcal·mol^{-1}) and unimportant residues (< 0.5 Kcal·mol^{-1}), as

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

described by Gao *et al.,* **[24]**

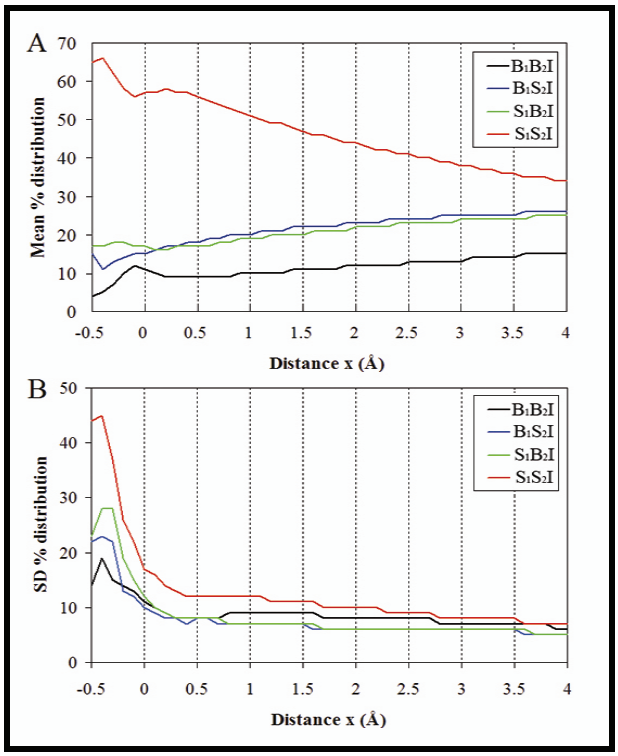

Figure 1: The distributions (A: Mean distributions; B: standard deviation) for four categories of inter-atomic interactions ($B_1B_2I/B_1S_2I/S_1B_2I/S_1S_2I$) at the protein-protein interfaces in a non-redundant dataset described elsewhere. **[5]** B = backbone, S = side-chain, subscript '1' refers to large monomer (e.g. enzymes, antibodies), and subscript '2' refers to small monomer (e.g. inhibitors, antigens). By definition, two atoms from different monomers were considered to interact if the distance between their centers is less than the sum of their van der Waals (vdW) radii plus x (Å). The value of x is varied from -0.5 to 4 (Å) at increments of 0.1 (Å). The vdW radius is taken from. **[37]**

Definition of inter-atomic sidechain-sidechain interactions

Protein-protein complexation is determined by inter-atomic interactions between monomers. Hence, we investigated the three groups of residues (hot spots, warm and unimportant residues) in terms of their contribution to the inter-atomic interactions. The inter-atomic interactions are composed of four categories, namely S_1S_2I, S_1B_2I, B_1S_2I and B_1B_2I (S: sidechain atom, B: backbone atom, subscript 1 and 2 refer to different monomers). The prevalence of these four inter-atomic interactions at the interface of protein-protein complexes (70 non-redundant complexes **[5]**) was examined by calculating their means and standard deviations at varying inter-atomic distances (Figure 1). S_1S_2I dominates at the interface and hence we exclusively selected S_1S_2I for studying hot spots, warm and unimportant residues. By definition, two sidechain atoms from different monomers were considered to interact if the distance between their centers is less than the sum of their van der Waals (vdW) radii plus a cutoff distance of 0.5Å, at which cutoff the mean of S_1S_2I is maximum and the standard deviation is minimum (Figure 1).

Classification of inter-atomic sidechain-sidechain contacts

We classified inter-atomic sidechain-sidechain contacts into two groups (energetically favorable and unfavorable contacts) using the scheme described by Sobolev and colleagues **[27]** (Table 1).

Definition of NA, NA_l-NA_{il} and NC_l-NC_{il}

We investigated each residue in the dataset using (1) NA, (2) NA_l-NA_{il} and (3) NC_l-NC_{il} (Supplementary table 1), illustrated in Figure 2. NA is the number of atoms of a residue participating in sidechain-sidechain contacts. NA_l-NA_{il} is the difference between the number of atoms in favorable contacts (NA_l) and unfavorable contacts (NA_{il}). It was employed to explore energetic contribution for a residue to protein-protein interface in terms of atoms. NC_l-NC_{il} is the difference between favorable contacts (NC_l) and unfavorable contacts (NC_{il}). It was used to explore energetic contribution for a residue to protein-protein interface in terms of inter-atomic contacts.

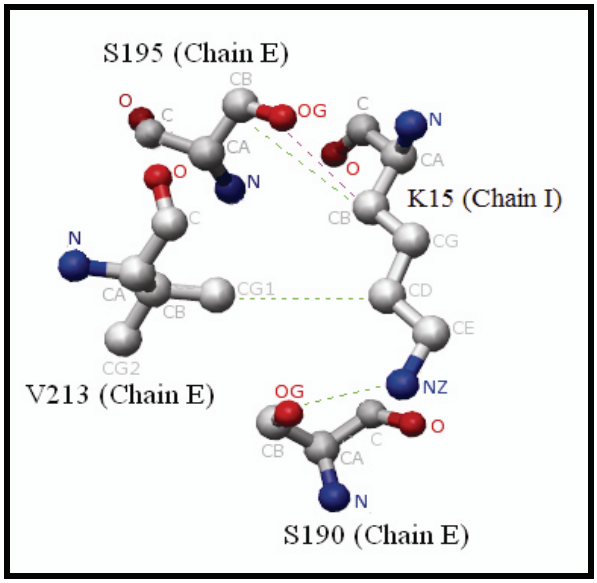

Figure 2: Illustration of NA, NA_l, NA_{il}, NC_l and NC_{il}. The interaction of K15 (PDB ID: BPTI, Chain I) to S190, S195 and V213 (trypsin, Chain E) is shown (PDB ID: 2PTC). K15 has three interacting side-chain atoms (CB, CD and NZ) and therefore the NA value is 3. Therein, the three atoms are all involved in favorable contacts (green line) and only CB participates in unfavorable contacts (red line). Thus, the NA_l value is 3 and NA_{il} is 1. In addition, K15 has three favorable contacts and one unfavorable contact; hence NC_l is 3 and NC_{il} is 1. Carbon atom: white; oxygen atom: red; Nitrogen atom: blue

Results:

Our goal is to investigate the characteristics of hot spots by comparing them with other interface residues using inter-atomic interactions. We collected 296 alanine-scanning interface residues consisting of 83 hot spots, 80 warm residues and 133 unimportant residues. At the interfaces of subunit interactions, S_1S_2I (side chain – side chain interaction) dominates and thus, S_1S_2I was subsequently used in this study. It should be noted that GLY (lacking side

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

chains) was disregarded in this analysis. However, the current dataset contains only two Gly residues and neither of them is a hot spot. Thus, the elimination of Gly did not significantly effect the analysis. For each residue in the dataset, we calculated the number of atoms (NA) participating in S_1S_2I, the number of atoms involved in favorable contacts (NA_l) and unfavorable contacts (NA_{il}). The number of favorable contacts (NC_l) and unfavorable contacts (NC_{il}) were further calculated. We used these values to calculate NA, NA_l-NA_{il} and NC_l-NC_{il} for each residue to compare the difference between hot spots, warm and unimportant residues (Supplementary Table 1: column 5, 6 and 7).

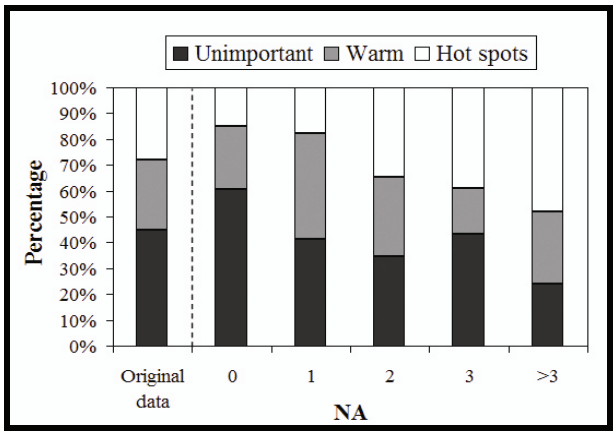

Figure 3: Percentage distribution of hot spots, warm and unimportant residues in 296 interface residues obtained from ASEdb (alanine scanning energetics database) **[26],** based on the value NA (the number of atoms for a residue involved in side-chain--side-chain interactions across protein-protein interface). The first column shows the percentage of the three types in the 296 residues. The number of residues is 114 for NA=0, 34 for NA=1, 52 for NA=2, 46 for NA=3 and 50 for NA>3. White: hot spots; gray: warm residues; black: unimportant residues

NA

Figure 3 shows percentage distributions of the three types of interface residues (hot spots, warm and unimportant residues) based on the value of NA. The percentage of hot spots increases from 15% to 50% with NA, while that of unimportant residues decreases from 60% to 23%. Interestingly, the percentage of warm residues does not significantly change with NA. This suggests that S_1S_2I interactions are prominent among hot spots. When NA = 1, the percentage of warm residues (41%) are larger than that in the original dataset (27%) and when NA > 1 hot spots (>33%) are higher than that in the original dataset (28%). It should be noted that nearly 40% of the residues in the dataset do not participate in inter-atomic sidechain-sidechain contacts (NA = 0). Hence, these residues can not be identified as hot spots, warm and unimportant residues using NA, NA_l-NA_{il} and NC_l-NC_{il} values.

NA_l-NA_{il}

Figure 4A shows percentage distributions of hot spots, warm and unimportant residues based on NA_l-NA_{il}. The percentages of unimportant residues decrease with the increase in NA_l-NA_{il}, and that of hot spots increases. The percentage of warm residues is not significantly affected by NA_l-NA_{il}. We also show that when NA_l-NA_{il} > 1, the percentage of hot spots is larger than the fraction of hot spots in the original dataset (28%).

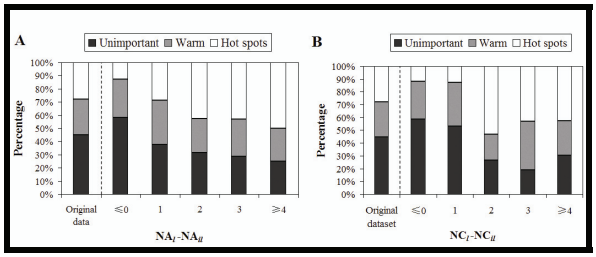

Figure 4: Percentage distribution of the three residue types in 182 interface residues with NA>0 (Unimportant residues: 64; warm residues: 52; hot spots: 66) based on the value of (A) NA_l-NA_{il} (the difference between numbers of sidechain atoms for a residue involved in favorable and unfavorable contacts). The number of residues is for 24 for NA_l-NA_{il} <1, 45 for NA_l-NA_{il} =1, 54 for NA_l-NA_{il} =2, 35 for NA_l-NA_{il} =3 and 24 for NA_l-NA_{il} >4 (B) NC_l-NC_{il} (the difference between numbers of favorable and unfavorable contacts). The number of residues is 17 for NC_l-NC_{il}<1, 32 for NC_l-NC_{il} =1, 30 for NC_l-NC_{il} =2, 21 for NC_l-NC_{il} =3, 82 for NC_l-NC_{il} >3. The first column in each graph shows the percentages of the three types in 296 interface residues. White: hot spots; gray: warm residues; black: unimportant residues

NC_l-NC_{il}

Figure 4B shows percentage distribution of the three types of interface residues types based on NC_l - NC_{il}. The percentage of unimportant residues decreases with the increase in NC_l-NC_{il}, and hot spots increases. The percentage of warm residues does not significantly change with NC_l-NC_{il}. It was also found that the percentage of hot spots is high when NC_l-NC_{il} ≥ 2, in comparison to the fraction (28%) of hot spots in the original dataset.

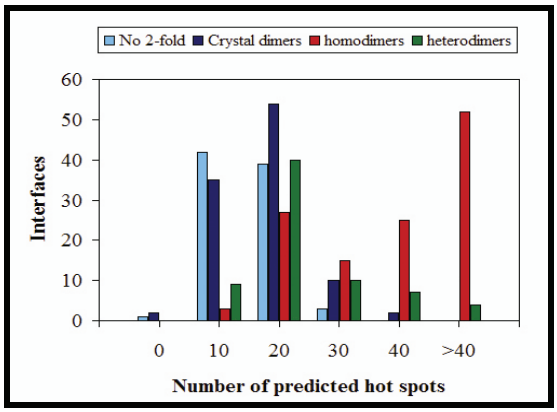

Figure 5: Distribution of hot spots calculated using our 'hot spot' prediction method for non-specific complexes (crystal-packing artifacts) and specific complexes (homodimers and heterodimers). These complexes are derived from the dataset of Bahadur *et al.*,. **[36]** Cyan, large crystal packing interfaces with no 2-fold symmetry; blue, crystal dimers; red, homodimers; green, protein–protein complexes

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Discussion:

A 'hot spot' prediction approach

Results show that the fraction of hot spots increases and unimportant residues decreases with increase in NA, NA_l-NA_{il} and NC_l-NC_{il}. However, the fraction of warm residues is not significantly affected by these three parameters. Thus, hot spots are preferentially involved in strong and energetically favorable sidechain-sidechain interactions, unimportant residues tend to participate in weak and energetically unfavorable sidechain-sidechain interactions. Here, we used NA, NA_l-NA_{il} and NC_l-NC_{il} to develop a method to identify hot spots using interface residues in structural complexes. We classified the residues in our dataset using a combination of three parameters. This is based on the observation that hot spots are prevailing in residues with NA > 1, NA_l-$NA_{i\,l}$ > 1 or NC_l-NC_{il} > 1, and unimportant residues are predominant in those with NA = 0, NA_l-$NA_{il} \leq 1$ or NC_l-$NC_{i\,l} \leq 1$ (Figure 3 and 4). Table 2 shows that the percentages of unimportant residues when (i) NA = 0, (ii) NA = 1 && NA_l-$NA_{il} \leq 1$ && NC_l-$NC_{i\,l} \leq 1$, and (vi) NA > 0 && NA_l-$NA_{il} \leq 1$ && NC_l-$NC_{il} \leq 1$ are larger than that in original dataset; and hot spots in (iii) NA = 1 && NA_l-$NA_{il} \leq 1$ && NC_l-$NC_{il} \geq 2$, (vii) NA > 1 && NA_l-$NA_{il} \leq 1$ && NC_l-$NC_{il} \geq 2$, and (ix) NA > 1 && NA_l-$NA_{il} \geq 2$ && NC_l-$NC_{il} \geq 2$ are higher than original dataset. Thus, the residues with NC_l-$NC_{il} \geq 2$ could be predicted as hot spots and those with NC_l-$NC_{il} \leq 1$ as unimportant residues. Therefore, these observations find application in the development of an expert system for the identification of hot spots from structural complexes.

Atom class*		Favorable (+) or unfavorable (-) contact							
		I	II	III	IV	V	VI	VII	VIII
I	Hydrophilic	+	+	+	-	+	+	+	+
II	Acceptor	+	-	+	-	+	+	+	-
III	Donor	+	+	-	-	+	+	-	+
IV	Hydrophobic	-	-	-	+	+	+	+	+
V	Aromatic	+	+	+	+	+	+	+	+
VI	Neutral	+	+	+	+	+	+	+	+
VII	Neutral-donor	+	+	-	+	+	+	-	+
VIII	Neutral-acceptor	+	-	+	+	+	+	+	-

Table 1: Legitimacy of contacts between side-chain atoms in different classes

*I: Hydrophilic = nitrogen or oxygen atoms that can donate and accept hydrogen bonds. II: Acceptor = nitrogen or oxygen atoms that can only accept a hydrogen bond. III: Donor = nitrogen that can only donate a hydrogen bond. IV: Hydrophobic = carbon atoms that are not in aromatic rings and do not have a covalent bond to a hydrophilic atom. V: Aromatic = carbon atoms in aromatic rings. VI: Neutral = carbon atoms that have a covalent bond to at least one atom of class I or two or more atoms from class II or III; nitrogen atoms if it has covalent bonds with 3 carbons; sulfur atoms in all cases. VII: Neutral-donor = carbon atoms that has a covalent bond with only one atom of class III. VIII. Neutral-acceptor = carbon atoms that has covalent bond with only one atom of class II. The classification is derived from. **[27]**

Comparison with other 'hot spot' predication methods

We evaluated our 'hot spot' prediction method by comparing them with three other methods: (1) PP_SITE **[24],** (2) alanine scanning method developed by Kortemme and coworkers **[20]** and (3) FOLDEF. **[22]** The PP_SITE method is structure-based, while the other two are energy-based. We assessed the performance of the four methods in distinguishing hot spots and unimportant residues. The PP_SITE classified interface residues into three types (hot spots, warm and unimportant residues) and its predicted warm residues include 43% of experimental hot spots. **[24]** In Alanine Scanning and FOLDEF methods, we considered interface residues with calculated $\Delta\Delta G \geq 1$ Kcal·mol^{-1} as predicted hot spots and other residues as predicted unimportant residues.

The four methods were evaluated using our dataset of 296 interface residues. The FOLDEF and our method identified all the 296 residues, while the PP_SITE method identified 226 residues and alanine scanning method identified 261 residues (See supplementary table 1). Then, we retained the identified residues which belong to experimental hot spots and unimportant residues (FOLDEF and our method: 215 residues; PP_SITE: 160 residues; Alanine scanning: 187 residues). Finally, for each method, we calculated sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV)and average successful rate ((TP+TN)/(TP+TN+FN+FP)) for hot spot prediction (Table 3).

Our method and FOLDEF showed high average successful rate (71% - 72%), compared to the other two methods (PP_SITE: 66%; Alanine Scanning: 68%). Thus, the FOLDEF and our method can effectively distinguish between hot spots and unimportant residues. Our method efficiently identified hot spots (SN = 72%; SP = 72%), while the FOLDEF efficiently identified unimportant residues (SN = 45%; SP = 88%). In addition, the PP_SITE correctly identified most host spots (SN = 90%) in these methods. However, it could not effectively differentiate unimportant residues from hot spots (SP = 37%). It agreed with the conclusions drawn by Gao *et al.* that the PP_SITE over-estimated unimportant residues. From these analyses, we can see that our method has remarkable hot spot prediction accuracy relative to the prevailing prediction approaches.

Misidentified hot spots

Out of the 83 hot spots in our dataset, 23 were not predicted, 17 of which do not have sidechain-sidechain interactions (NA = 0) and the remaining five do not make significant energetic contribution to sidechain-sidechain interactions (NC_l-NC_{il} = 1). It seems that energetic contribution of these hot spots to protein-protein interaction could not be reflected by their participation in inter-monomeric sidechain-sidechain interactions. In order to understand how the 23 misidentified hot spots contribute to protein-protein interaction, they were studied in detail and several reasons were found. (1) Some of them interact with interfacial water molecules to enhance the stability of protein-protein interaction. For instance, the residue D51 in the protein Im9 (PDB: 1BXI) hydrogen bonds two interfacial water molecules buried in cavities. **[28]** (2) Some neighbor with hot spots with NC_l - $NC_{il} \geq 2$. The mutations to alanine may influence their neighboring hot spot's conformation which then reduces protein-protein interaction. In human growth hormone-hGH receptor complex (PDB: 3hhr), misidentified

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

hot spots I103 ($\Delta\Delta G$ =1.8 Kcal·mol^{-1}) and I105 ($\Delta\Delta G$ =2.0 Kcal·mol^{-1}) neighbor hot spot W104 ($\Delta\Delta G$ =4.5 Kcal·mol^{-1}). (3) Some have role in stabilizing monomer structure so that their mutations to alanine disrupt monomer conformation which weakens protein-protein interactions, such as the residue D58 in Tissue Factor (PDB: 1DAN). **[29]** (4) Some contribute to protein-protein interaction by participating in backbone-backbone or backbone-sidechain interactions. The residue K15 in protein Basic Pancreatic Trypsin Inhibitor (PDB: 1CBW) are involved in three backbone-backbone hydrogen bonds. **[30]** Similarly, the residues N23 and Q120 in Staphylococcal enterotoxin C3 (PDB: 1JCK) form hydrogen bonds with backbone atoms of interacting monomer T cell antigen receptor V_β. **[31]**

Groups	Interface residues			
	Hot-spot 83 (28%)	**Warm 80 (27%)**	**Unimportant 133 (45%)**	**Total (296)**
(i) NA=0	17 (15%)	28 (25%)	69 (60%)	114
NA=1				
(ii) NA_l-$NA_{il}\leq 1$&NC_l-$NC_{il}\leq 1$	2 (8%)	10 (40%)	13 (53%)	25
(iii) NA_l-$NA_{il}\leq 1$& NC_l-$NC_{il}\geq 2$	4 (44%)	4 (44%)	1 (11%)	9
(iv) NA_l-$NA_{il}\geq 2$& NC_l-$NC_{il}\leq 1$	0	0	0	0
(v) NA_l-$NA_{il}\geq 2$& NC_l-$NC_{il}\geq 2$	0	0	0	0
NA>1				
(vi) NA_l-$NA_{il}\leq 1$& NC_l-$NC_{il}\leq 1$	4 (17%)	6 (25%)	14 (58%)	24
(vii) NA_l-$NA_{il}\leq 1$& NC_l-$NC_{il}\geq 2$	6 (55%)	2 (18%)	3 (27%)	11
(viii)NA_l-$NA_{il}\geq 2$&NC_l-$NC_{il}\leq 1$	0	0	0	0
(ix) NA_l-$NA_{il}\geq 2$&NC_l-$NC_{il}\geq 2$	50 (44%)	30 (27%)	33 (29%)	113

Table 2: Classification of the residues in the datasets using the three parameters (NA, NA_l-NA_{il} and NC_l-NC_{il})

	Our approach	**PP_SITE [24]**	**Alanine Scanning [20]**	**FOLDEF [22]**
SN	72%	90%	60%	45%
SP	72%	47%	74%	88%
PPV	62%	57%	64%	70%
NPV	81%	86%	71%	72%
Averaged successful rate	72%	66%	68%	71%
% of warm residues predicted as hot spots	45%	60%	42%	21%

Table 3: Evaluation of 'hot spot' prediction approaches

The four prediction methods were assessed by comparing their performance on the differentiation between hot spots and unimportant residues. Warm residues were disregarded. SN=sensitivity; SP=specificity; PPV= positive predictive value; NPV= negative predictive value; average successful rate = ((TP+TN) / (TP+TN+FN+FP)). Both predicted warm residues and hot spots by the method PP_SITE were regarded as predicted hot spots here and the evaluation is based on the PP_SITE prediction result with surface punishment. **[24]** For the alanine scanning method and the FOLDEF method, we considered interface residues with calculated $\Delta\Delta G \geq 1$ Kcal·mol^{-1} as predicted hot spots and other residues as predicted unimportant residues.

Distinction between specific and non-specific complexes

Assessing the oligomeric state of a protein from its X-ray structure is not always straightforward and protein subunit interfaces often coexist with 6 to 12 packing interfaces. **[32, 33]** The distinction between oligomers (specific complexes) and crystal-packing artifacts (non-specific complexes) is often made on the basis of interface area and specific interface area is generally larger. **[3, 34, 35]** Recently, Bahadur *et al.* observed that three independent parameters (non-polar interface area, fraction of fully buried atoms and residue propensity score at interface) could distinguish between homo-dimers and non-specific complexes and these are indistinguishable based on interface area. **[36]** Here, we used our 'hot spot' prediction method to distinguish between specific and non-specific complexes, using the dataset of Bahadur *et al.* which contains 188 large crystal-packing artifacts, 122 homo-dimers and 70 hetero-dimers. Figure 5 show that the low abundance of hot spots distinguishes the crystal-packing interfaces from homo-dimeric interfaces. Using the number (23) of hot spots as a cutoff, 179 out of 188 non-specific interfaces and 88 out of 122 homo-dimeric interfaces were identified. In other words, 86% of the proteins are correctly classified as monomers and homo-dimers using hot spots as a criterion. The hot spot cutoff was selected manually in this study and with larger data sets, the cutoff has to be refined to optimized, for the distinction between homo-dimers and monomers. We also calculated the correlations between the number of hot spots and the three parameters observed by Bahadur *et al.* and found a weak correlation (correlation coefficient $R^2 < 0.17$). Thus, the 'hot spot' prediction method could be applied along with these three parameters for homo-dimer identification. However, the prediction method could not efficiently distinguish between hetero-dimers and non-specific complexes. This may be due to the binding mechanism of hetero-dimers, which assemble from preformed protein components. In the free components, the surface patches that form the interface are in contact with the solvent and their physical/chemical properties are not significantly different from the remainder of protein

surface.

Electronic supplementary material
The dataset of 296 alanine-mutated interface residues in this work is available online.

Acknowledgement:
Lei Li acknowledges NANYANG Technological University, Singapore for support.

References:

[1] S. Jones & J. M. Thornton, *Proc Natl Acad Sci.,* 93:13 (1996) [PMID: 8552589]
[2] S. Jones & J. M. Thornton, *J Mol Biol.,* 272:121 (1997) [PMID: 9299342]
[3] H. Ponstingl, *et al., Proteins,* 41:47 (2000) [PMID: 10944393]
[4] L. Lo Conte, *et al., J Mol Biol.,* 285:2177 (1999) [PMID: 9925793]
[5] P. Chakrabarti & J. Janin, *Proteins,* 47:334 (2002) [PMID: 11948787]
[6] R. P. Bahadur, *et al., Proteins,* 53:708 (2003) [PMID: 14579361]
[7] F. Glaser, *et al., Proteins,* 43:89 (2001) [PMID: 11276079]
[8] J. A. Wells, *Methods Enzymol.,* 202:390 (1991) [PMID: 1723781]
[9] T. Clackson & J. A. Wells, *Science,* 267:383 (1995) [PMID: 7529940]
[10] A. Bogan & K. S. Thorn, *J Mol Biol.,* 280:1 (1998) [PMID: 9653027]
[11] M. J. Castro & S. Anderson, *Biochemistry,* 35:11435 (1996) [PMID: 8784199]
[12] B. Ma, *et al., Proc Natl Acad Sci.,* 100:5772 (2003) [PMID: 12730379]
[13] I. Halperin, *et al., Structure,* 12:1027 (2004) [PMID: 15274922]
[14] A. del Sol, *et al., Bioinformatics,* 21:1311 (2005) [PMID: 15659419]
[15] A. del Sol & P. O'Meara, *Proteins,* 58:672 (2005) [PMID: 15617065]
[16] X. Li, *et al., J Mol Biol.,* 344:781 (2004) [PMID: 15533445]
[17] T. Haliloglu, *et al., Biophys J.,* 88:1552 (2005) [PMID: 15596504]
[18] V. Brinda & S. Vishveshwara, *BMC Bioinformatics,* 6: 296 (2005) [PMID: 16336694]
[19] T. Kortemme, *et al., Sci STKE*, 2004:pl2 (2004) [PMID: 14872095]
[20] T. Kortemme & D. Baker, *Proc Natl Acad Sci.,* 99: 14116 (2002) [PMID: 12381794]
[21] I. Massova & P. A. Kollman, *J. Am. Chem. Soc.,* 121: 8133 (1999)
[22] R. Guerois, *et al., J Mol Biol.,* 320:369 (2002) [PMID: 12079393]
[23] G. M. Verkhivker, *et al., Proteins,* 48:539 (2002) [PMID: 12112677]
[24] Y. Gao, *et al., J Mol Model,* 10:44 (2004) [PMID: 14634848]
[25] S. Hubbard & J. Thornton, *NACCESS*, University College London, (1993)
[26] S. Thorn, & A. Bogan, *Bioinformatics,* 17:284 (2001) [PMID: 11294795]
[27] V. Sobolev, *et al., Proteins,* 25:120 (1996) [PMID: 8727324]
[28] U. C. Kuhlmann, *et al., J Mol Biol.,* 301:1163 (2000) [PMID: 10966813]
[29] R. F. Kelley, *et al., Biochemistry,* 34:10383 (1995) [PMID: 7654692]
[30] I. Leder, *et al., J Exp Med.,* 187:823 (1998) [PMID: 9500785]
[31] J. Scheidig, *et al., Protein Sci.,* 6:1806 (1997) [PMID: 9300481]
[32] S. Dasgupta, *et al., Proteins,* 28:494 (1997) [PMID: 9261866]
[33] J. Janin & F. Rodier, *Proteins,* 23:580 (1995) [PMID: 8749854]
[34] J. Janin, *Nat Struct Biol.,* 4:973 (1997) [PMID: 9406542]
[35] O. Carugo & P. Argos, *Protein Sci.,* 6:2261 (1997) [PMID: 9336849]
[36] R. P. Bahadur, *et al., J Mol Biol.,* 336:943 (2004) [PMID: 15095871]
[37] C. Chothia, *J Mol Biol*.,105:1 (1976) [PMID: 994183]

Edited by N. Srinivasan
Citation: Li *et al.*, Bioinformation 1(4): 121-126 (2006)