

MIT Open Access Articles

A computational tool to predict the evolutionarily conserved protein-protein interaction hot-spot residues from the structure of the unbound protein

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Agrawal, Neeraj J., Bernhard Helk, and Bernhardt L. Trout. "A Computational Tool to Predict the Evolutionarily Conserved Protein-Protein Interaction Hot-Spot Residues from the Structure of the Unbound Protein." FEBS Letters 588, no. 2 (November 12, 2013): 326–333.

As Published: <http://dx.doi.org/10.1016/j.febslet.2013.11.004>

Publisher: Elsevier

Persistent URL: <http://hdl.handle.net/1721.1/101385>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-NonCommercial-NoDerivs License



**A computational tool to predict the evolutionarily conserved
protein-protein interaction hot-spot residues from the structure of
the unbound protein**

Neeraj J Agrawal¹, Bernhard Helk², Bernhardt L Trout^{1*}

¹ 77 Massachusetts Avenue, E19-502b, Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

² Novartis Pharma AG, CH-4057, Basel, Switzerland.

* Author for correspondence: trout@mit.edu, Tel: +1-617- 258-5021, Fax: +1-617-253-2272

ABSTRACT

Identifying hot-spot residues - residues that are critical to protein-protein binding - can help to elucidate a protein's function and assist in designing therapeutic molecules to target those residues. We present a novel computational tool, termed spatial-interaction-map (SIM), to predict the hot-spot residues of an evolutionarily conserved protein-protein interaction from the structure of an unbound protein alone. SIM can predict the protein hot-spot residues with an accuracy of 36-57%. Thus, the SIM tool can be used to predict the yet unknown hot-spot residues for many proteins for which the structure of the protein-protein complexes are not available, thereby providing a clue to their functions and an opportunity to design therapeutic molecules to target these proteins.

INTRODUCTION

It is estimated that a human protein-protein interaction (PPI) interactome is composed of as many as 650,000 different PPIs, and understanding these interactions is expected to lead to new therapeutic targets.¹ Proteins are the work-horse of the cellular machinery, and the formation of specific protein complexes led by specific PPIs underpins many cellular processes. Aberrant PPIs, either through the loss of a function or through the formation and/or stabilization of a protein-protein complex at an inappropriate time or location, are implicated in many diseases such as cancer and autoimmune diseases. Elucidating the regions of the protein that drive the PPI helps in understanding the protein function and in designing drugs that target the regions that are involved in the PPI.^{2,3}

Over the past decade, a large number of protein structures have been solved, and the number of solved structures of protein-protein complexes has been also increasing. These structures of the complexes yield information on the residues that are present in the protein-protein binding regions. These residues constitute the structural epitope of the protein. However, not all of the residues that are present in the binding region contribute equally to the binding energy of the complex. In pioneering work on the binding of human growth hormone (GH) to its receptor, Cunningham et al. identified a region of energetically important residues on the protein surface that were critical to the binding.⁴ Following their work and other experiments, it became evident that only a few of the binding-region residues contribute a major component of the binding energy. These residues, which constitute the functional epitope, are termed hot-spot residues. Although a qualitative definition of hot-spot residues is straightforward, consensus on the quantitative definition of hot-spot residues is still lacking. One of the definitions of a hot-spot

residue can be construed as the residue that contributes more than a certain threshold (e.g., 2.5 kcal/mol⁵) to the binding energy of the PPI. Because direct experimental measurements of the contributions of individual residues to the protein-protein binding free energy are currently very tedious, an operational definition of a hot-spot residue is often used. Operationally, a hot-spot residue can be defined as a residue that, when mutated to alanine, leads to at least some given increase (e.g., 10-fold) in the protein-protein dissociation constant (K_D).

Experimentally, site-directed mutagenesis has been widely used to analyze how protein-protein interfaces function. In this method, subsets of the protein residues are systematically mutated, mostly one at a time, and the effect of each mutation on the protein-protein binding energy is analyzed. The preferred residue to mutate to is alanine because the alanine amino acid lacks a side chain beyond the β -carbon. Hence, the binding assays performed in conjunction with (alanine) mutagenesis identify hot-spot residues as defined by the operational definition. In these experiments, it is tacitly assumed that the mutation of a residue to alanine does not lead to structural perturbations of the protein. In fact, Rao et al. have aptly demonstrated the limitation of such an assumption.⁶ In their experiments, although the mutation F19A led to a significant reduction in the binding strength of human Prolactin to its receptor, residue F19 cannot be considered to be a hot-spot residue because the F19A mutation is accompanied with significant structural changes.⁶ In experiments in which site-directed mutagenesis is restricted to only surface-exposed residues, as identified from the protein structure, the chances of protein structure perturbation upon mutation greatly diminishes.

On the computational front, a few tools have been developed to identify hot-spot residues. All of these bioinformatics tools, which have been trained over a dataset, can be broadly classified into two categories: tools that are based on the structure of the protein-protein complex and tools that are based on the sequence/structure of the unbound protein. The first category includes energy-based tools⁷⁻¹¹, and machine learning-based tools such as PCRPI¹², KFC¹³, MINERVA¹⁴, HotPoint¹⁵, and others¹⁶. While these tools can identify hot-spot residues with great accuracy, the requirement of the protein-protein complex structure severely limits the application of such tools, and these tools cannot be employed to predict hot-spot residues when the structure of the protein-protein complex is unavailable. The other category of computational tools identifies hot-spot residues by using the sequence or structure of the unbound protein alone. Tool such as ISIS^{5,17} is designed to identify protein-protein interaction hot-spot residues using an unbound protein structure and/or sequence. The majority of other sequence-based computational tools, e.g., PredUs¹⁸, meta-PPISP¹⁹ and ConSurf²⁰, are designed to identify protein-protein binding-region residues. Another tool, called FTMAP²¹, has been designed to predict hot-spot residues of small molecule ligand interactions with a protein by using the structure of the protein. Readers are directed to reviews^{22,23} from the laboratory of Nussinov on the available computational tools for predicting the binding-region residues. In this article, we present a new method for the prediction of the hot-spot residues from the structure of the unbound protein. We also compare our method to other methods (ISIS^{5,17}, meta-PPISP¹⁹, PredUs^{18,24} and ConSurf²⁰), which also use the sequence/structure information of only the unbound protein to predict the hot-spots/binding region residues of the protein.

Recently, our group developed a tool that was called the spatial-aggregation-propensity (SAP) to identify aggregation-prone regions in proteins.²⁵ SAP is a measure of the dynamic exposure of hydrophobic patches on the protein surface. The SAP tool can also predict, using the unbound protein structure, the binding regions in a protein.²⁶ Thus, a patch of exposed hydrophobic residues that is indicated by a high SAP value of the region is a good indicator of a protein binding region. Furthermore, previous work on the detection of hydrophobic patches on the surfaces of proteins has also shown the utility of finding hydrophobic patches for identifying protein binding regions.²⁷ Recently, Kozakov et al. also demonstrated that protein hot-spots are characterized by regions that are patterned with hydrophobic and polar residues.²⁸ With this background in mind, we developed a computational tool called the spatial-interaction-map (SIM).

We apply the SIM tool to a number of proteins, to predict their hot-spot residues. By design, the SIM tool can be applied to a single (i.e., static) structure of the protein and to multiple structures of the protein. When the SIM tool is applied to a static structure, we refer to it as sSIM; when the SIM tool is applied to multiple structures, we refer to it as dSIM. We compare the SIM-predicted residues with the experimentally known hot-spot residues and the experimentally known binding-region residues; we also compare ISIS, PredUs, meta-PPISP and ConSurf in terms of their ability to predict hot-spot and binding-region residues for these proteins. Because a few previous studies on the characterization of protein-protein interfaces have cast doubt on the utility of hydrophobicity for the prediction of the protein-protein interface²⁹⁻³¹, we also compare our predictions obtained by using SIM against predictions obtained by performing simple hydrophobic analysis. For benchmarking purposes, we also report

the results that were obtained when all of the exposed residues were considered to be hot-spot and binding-region residues.

For validation of our computational method, we resort to the experimentally known hot-spot residues and binding-region residues of evolutionarily conserved protein-protein interactions. Publicly available databases such as ASEdb³² and BID³³ contain a repository of experimentally known hot-spot residues, and the HotSprint³⁴ database contains a repository of computational hot-spot residues. However, quite a large number of protein-protein interactions contained in ASEdb and BID belong to an antigen-antibody interaction, which is an evolutionarily non-conserved interaction. Furthermore, these databases do not necessarily provide information on the known binding-region residues. For most of the protein-protein interactions, a number of binding-region residues still lack experimental data that can be used for classifying them as hot-spot or non-hot-spot residues. This lack of information can affect the performance of a computational method when the reported method's accuracy is based on the ratio of the correctly predicted hot-spot residues to the total number of predicted residues. To account for this lack of experimental information, we calculate the accuracy and the theoretical maximum accuracy (defined as the accuracy when all of the binding region residues that have not been experimentally tested are assumed to be hot-spot residues) for each method. Hence, in this work, we study only those proteins for which the protein-protein interaction is evolutionarily conserved and for which both the hot-spot residues and the binding-region residues are experimentally known.

In this report, we show the results for IL-13 protein. The results for other proteins, specifically IL-2, growth hormone receptor, IL-15, growth hormone, Fc-domain of an IgG1, erythropoietin, IL-13R α 1 and EGFR, are given in the Supporting Information, Section S4. For all of these proteins, we report any concerns on the quality of the experimental data in the Supporting Information tables.

METHODS

Spatial-Interaction-Map (SIM)

The input to the spatial-interaction-map (SIM) tool is a fully atomistic three-dimensional structure of the protein (see Supporting Information, Sections S1.1 for the details on the methods used to obtain protein structure and perform molecular simulations). sSIM indicates SIM computed on a single protein structure, and dSIM indicates SIM computed over multiple structures of the protein. These multiple structures of the protein are generated using molecular-dynamics simulations. Calculations to perform SIM analysis can be divided into four steps. Step I: using the structure of the protein, we assign an effective-hydrophobicity value to each of the residues of the protein. The effective hydrophobicity Φ_{eff} of the i^{th} residue is defined as:²⁵

$$\Phi_{\text{eff},i} = \sum_{\text{Simulation Average}} \left(\frac{\text{SAA of side chain atoms of residue } i}{\text{SAA of side chain atoms of fully exposed residue}} \right) \times \Phi_i.$$

SAA, the solvent-accessible area of the side-chain atoms of residue i , is computed at each simulation snapshot (for sSIM, the summation is over only one structure); the SAA of the side-chain atoms of fully exposed residues (e.g., for amino acid X) is obtained by calculating the SAA of the side-chain atoms of the middle residue in the fully extended conformation of the tripeptide Ala-X-Ala, and the hydrophobicity of each residue Φ_i is obtained from the hydrophobicity scale of Black and Mould.³⁵ The SAA is the area of the surface that is obtained from rolling a probe sphere on the surface of the protein. A probe sphere of radius 1.4 Å, which is equivalent to that of the water molecule, is used. The van der Waals radii of each of the atoms of the protein are taken from the CHARMM22 force-field.³⁶ We normalize the hydrophobicity scale in such a way that glycine has a hydrophobicity of zero. Thus, residues that are more hydrophobic than glycine (Ala, Cys, Pro, Met, Val, Trp, Tyr, Ile, Leu, and Phe) have positive values, while residues that

are less hydrophobic than glycine (Thr, Ser, Lys, Gln, Asn, His, Glu, Asp, and Arg) have negative values for Φ_i . Furthermore, we normalize the hydrophobicity scale in such a way that the most hydrophobic residue (Phe) has a value of 0.5 while the least hydrophobic residue (Arg) has a value of -0.5. Step II: the second step of the SIM tool identifies the clusters of highly hydrophobic residues that are present on the protein surface. We define a cutoff value for the effective hydrophobicity (Φ_{cutoff}); for each Φ_{cutoff} value, we identify all of the residues that have $\Phi_{\text{eff},i} > \Phi_{\text{cutoff}}$ as highly hydrophobic residues. A cluster of highly hydrophobic residues is defined as two or more highly hydrophobic residues being present in the vicinity of each other (see also the Supporting Information, Section S2). In our work, the distance between two residues is defined as the least distance between any two atoms of these residues. We use a (Euclidian) distance of 10 Å between two residues as a cutoff for defining the vicinity. The distance of 10 Å (i.e., the patch size of $\sim 320 \text{ \AA}^2$) corresponds approximately to the lower limit of the size of the protein-protein interface.³⁷ We then implement the reverse Cuthill-McKee algorithm to identify the clusters of highly hydrophobic residues.³⁸ For computing the dSIM, the SAA is averaged over the simulation, while the distances between the residues are computed for a representative frame (in our work, we use the last frame from the MD simulation). Step III: the third step of the SIM identifies solvent-exposed charged-residues (Arg, Lys, Asp, Glu) in the vicinity of these hydrophobic clusters. Any solvent-exposed charged-residue within a (Euclidian) distance of 5 Å from any of the highly hydrophobic residues is selected as belonging to the cluster as well. A SAA cutoff of 10 \AA^2 is used to distinguish between solvent-exposed and buried residues. Step IV: the fourth step of SIM further narrows down the number of predicted residues by discarding all but the most highly conserved residues. We use a ConSurf score of less than 0.5 as an

indicator of high evolutionary conservation (see Supporting Information, Section S3). Any other sequence-conservation algorithm can be used as well.

Exposed Residues

For the multiple structures of the protein obtained using MD simulations, we use the VMD software³⁹ to compute the solvent-accessible area of the side-chain atoms (including hydrogen atoms) of each residue. The van der Waal radius of each atom was assigned using the CHARMM22³⁶ force field, and a probe radius of 1.4 Å was used to represent the water molecule. Any residue with a SAA of its side-chain atoms of greater than 10 Å² is identified as an exposed residue.

Simple Hydrophobic Analysis

We use the above-mentioned method to identify all of the exposed residues on the protein surface. All of the exposed hydrophobic residues (i.e., TRP, TYR, VAL, MET, PHE, PRO, ILE, LEU, CYS and ALA) are considered to be predicted residues when using this method. For brevity, this method is referred as “Hydrophobic” in all of the figures.

Bioinformatics Tools

The details on the bioinformatics tools can be found in the Supplementary Information, Section S1.2.

Identification of binding-region residues and hot-spot residues

From the protein-protein complex structure, residues that are present in the binding region are identified from their loss of solvent accessibility upon binding by using the PDBePISA tool (<http://pdbe.org/pisa>).⁴⁰ If the protein of interest binds to multiple partners, then we identify all of

the residues that are involved in binding to all of its partners as binding-region residues. We also identify all of the experimentally known hot-spot residues for each of the proteins. Only the residue that upon mutation to alanine leads to at least a 10-fold increase in the dissociation constant K_D of the protein-protein binding (i.e., $\Delta\Delta G > 1.37$ kcal/mol) are retained as hot-spot residues. To discount the allosteric effects of mutations, only the hot-spot residues that are present in the binding region are considered.

Evaluation of Performance

We evaluate the performance of each method for each protein in terms of its accuracy and coverage. The accuracy is calculated as the ratio of the number of correctly predicted residues to the total number of predicted residues, whereas the coverage is calculated as the ratio of the number of correctly predicted residues to the number of experimentally observed residues. The accuracy and coverage is calculated for both the binding-region residue prediction and the hot-spot residue prediction. Let P be the set of all of the residues that are predicted by a given method for a given protein. Let B be the set of all of the experimentally known binding-region residues, and let H be the set of all of the experimentally known hot-spot residues for a given protein. For each protein, we also generate the set NH of all of the experimentally known binding-region residues that are experimentally known to not be hot-spot residues. Then, the accuracy (ACC) and coverage (COV) of a method for a protein are given as:

$$ACC_B = |P \cap B| / |P|,$$

$$COV_B = |P \cap B| / |B|,$$

$$ACC_H = |P \cap H| / |P|,$$

$$COV_H = |P \cap H| / |H|,$$

where $|\cdot|$ represents the cardinality of the set, \cap represents the intersection of the sets, and the superscript c denotes the complement of the set. Subscripts B and H indicate the performance of a method for predicting the binding-region and hot-spot residues, respectively. In most of the instances, only a few of the residues that are involved in protein binding have been mutated experimentally to identify their contribution to the protein binding. Hence, we also compute the theoretical maximum accuracy, $\max\text{ACC}$, of each method for the prediction of hot-spot residues as the ratio of the number of predicted residues that lie in the binding region and are not non-hot-spot residues to the total number of residues predicted. Here, we have assumed that whenever experimental information is unavailable for a binding-region residue, we count that residue as a hot-spot residue.

$$\max\text{ACC}_H = |\mathbf{P} \cap \mathbf{B} \cap \mathbf{N}^c| / |\mathbf{P}|.$$

True positives (the number of predicted residues that are also experimentally known hot-spot residues), false positives (the number of predicted residues that are not experimentally known hot-spot residues), and false negatives (the number of experimentally known hot-spot residues that are not predicted to be hot-spot residues) for each of the methods are also reported in the Supporting Information, Tables S2-S7.

RESULTS

Identification of hot-spot residues using SIM

The sSIM tool is applied to a protein structure that is obtained from either x-ray or NMR studies. Whenever the structure of the protein is not available, the SIM tool can also be applied to protein structures that are obtained from any other method, such as homology modeling. The dSIM tool is applied to multiple structures of the protein; these multiple structures can be generated by performing fully atomistic molecular-dynamics simulations on the protein. First, SIM computes the effective hydrophobicity, Φ_{eff} , of each residue in the protein. Φ_{eff} normalizes the hydrophobicity of each residue by its fractional solvent-accessible-area (SAA); thus, all buried (including hydrophobic) residues have Φ_{eff} equal to zero. SIM then generates a contact-map matrix C of dimensions $N \times N$, where N is the total number of residues in the protein. Figure 1A depicts the contact-map matrix for protein IL-13. An element C_{ij} of this matrix is one if the residues i and j are within 10 Å of each other; otherwise, it is zero. By design, the matrix C is symmetric. SIM then applies a high-hydrophobicity filter to set all of the entries of row and column i to zero if the Φ_{eff} of residue i is less than Φ_{cutoff} . In Figure 1B, we show the results that were obtained by using $\Phi_{\text{cutoff}} = 0.15$ to filter out the residues with low Φ_{eff} from the matrix C . The reverse Cuthill-McKee algorithm (as implemented in MATLAB) is then applied to reorder this sparse matrix in such a way as to identify the clusters of highly hydrophobic residues.³⁸ Figure 1C shows that the four clusters C1, C2, C3, and C4, which are composed of highly hydrophobic residues, are present on the surface of IL-13. This procedure selects clusters that are composed of exposed hydrophobic residues. We discard the clusters (e.g., cluster C3) that have only one highly hydrophobic residue (see the Supporting Information, Section S2). Furthermore, the clusters that are very close to the N- and C-termini are also discarded (e.g., cluster C4).

Surface-exposed charged residues in the vicinity of the residues in clusters C1 and C2 are then identified from the protein structure. We further reduce the number of the predicted residues by eliminating the residues that are less conserved along evolution. Figure 2A maps these clusters onto the IL-13 surface. Each cluster is composed of exposed conserved charged residues and exposed conserved hydrophobic (i.e., $\Phi_{\text{eff}} > \Phi_{\text{cutoff}}$) residues. For comparison, we have also mapped the hydrophobicity of each residue (Φ) and the ConSurf score of each residue onto the IL-13 surface in Figures 2C and 2D, respectively. Figure 2C shows that this protein has many exposed hydrophobic residues and that these regions are distributed over its surface. Thus, it becomes very difficult to pick a certain hydrophobic region that is involved in binding compared to other regions. Similarly, many conserved residues are exposed on the protein surface, as seen in Figure 2D, which makes the selection of a certain conserved region over other regions difficult.

The number of predicted residues using SIM can be controlled by varying the value of Φ_{cutoff} . At a very large value of Φ_{cutoff} , a small number of residues are predicted, while at moderate values of Φ_{cutoff} , a large number of residues are predicted. For $\Phi_{\text{cutoff}} = 0$, even the buried (and conserved) residues will be predicted, and for $\Phi_{\text{cutoff}} = -0.5$, all of the conserved residues in the protein will be predicted. Hence, preferably, Φ_{cutoff} should be set to values that are greater than 0.1.

Interleukin-13 (IL-13)

Human IL-13 is a ~12 kDa cytokine and is important for the development of the T-helper cell type 2 (Th2) response. Dysregulation of the IL-13-mediated response has been linked to asthma and allergic diseases. Structurally, IL-13 belongs to the four-helix bundle superfamily. IL-13-

mediated hetero-dimerization of receptors IL-13R α 1 and IL-4R α initiates the downstream signaling via recruitment and activation of STAT6. IL-13 first binds to IL-13R α 1 (with K_D =1.69 nM) followed by the binding of this complex to the IL-4R α receptor. IL-13 can also bind to another receptor, IL-13R α 2, with a very high affinity (K_D =15.5 fM).⁴¹ Lupardus et al. have characterized the binding energetic of the IL-13 – IL-13R α 1 and IL-13 – IL-13-R α 2 interactions by mutating the residues on the surface of IL-13 to alanine.⁴¹ The resulting change in the interaction energy upon mutation was measured by isothermal titration calorimetry and surface plasmon resonance. Their experiments identify nine hot-spot residues on IL-13; eight of these hotspots are crucial for binding to IL-13R α 1, while three are crucial for binding to IL-13R α 2. The residues K104 and F107 are two of the most crucial residues for binding to both of the receptors. Indeed, the mutation K104A or F107A leads to more than a 5000-fold increase in the K_D of IL-13 binding to IL-13R α 2. To identify the binding-region residues of IL-13, we use the available x-ray structures of IL-13 bound to IL-13R α 1 (PDB ID: 3BPO⁴²) and IL-13 bound to IL-13R α 2 (PDB ID: 3LB6⁴¹).

We predict the hot-spot and binding-region residues by sSIM, meta-PPISP, PredUs and ConSurf, using the available NMR structure (PDB ID: 1IJZ⁴³) of unbound IL-13. For ISIS, we use the sequence of IL-13. We also perform a 20 ns MD simulation of IL-13 and apply the dSIM tool to the last 15 ns of the simulation. As we decrease the value of Φ_{cutoff} from 0.2 to 0.1, we identify more and larger clusters by sSIM. A similar trend is observed for dSIM; however, no cluster is identified by dSIM when $\Phi_{\text{cutoff}} = 0.2$ is used. Figure 3A shows that, for predicting a binding-region residue, both meta-PPISP and ISIS fare no better than randomly selecting an exposed residue on the surface of IL-13. Similarly, selecting a conserved exposed residue or an

exposed hydrophobic residue of IL-13 does not have any advantage over the random selection of exposed residues. Thus, structural and sequence-conservation information alone suffers from low accuracy and cannot be used to identify binding-region residues. PredUs can predict binding-region residues with an accuracy of 50%, which is almost twice the accuracy of tools such as meta-PPISP and ISIS. The SIM tool performs much better at predicting binding-region residues than all of these tools, and the coverage of the SIM prediction increases at the cost of the accuracy as we decrease the value of Φ_{cutoff} . The SIM tool, even at a low Φ_{cutoff} , has almost twice the chance of correctly predicting a residue to be in the binding region compared to meta-PPISP, PredUs and ISIS.

Moreover, the SIM tool can predict preferentially the hot-spot residues in the binding region (see Figure 3B and the Supporting Information, Section S4.1). The SIM tool can predict more than 1/3rd of the hot-spot residues correctly and with a considerably higher accuracy. The SIM analysis at a high value of Φ_{cutoff} correctly predicts the hot-spot residues that are important for binding to the high-affinity receptor IL-13R α 2, and reducing the value of Φ_{cutoff} identifies the hot-spot residues for binding to the low-affinity receptor IL-13R α 1. Importantly, both sSIM and dSIM can identify correctly K104 and F105, which are the two most important hot-spot residues of IL-13. The tools meta-PPISP, PredUs and ISIS can also identify the hot-spot residues for binding to IL-13R α 1 and IL-13R α 2, respectively, although with a low accuracy. Moreover, both meta-PPISP and ISIS fail to predict the K104 and F105 residues. The lack of experimental data on the energetic contribution of all of the residues that are present in the protein-binding interface is highlighted in Figure 3B by large error bars on the accuracy of each method.

DISCUSSION

A large amount of structural information has been accumulated over the years on proteins and protein-protein complex structures. Whereas protein-protein complex structures yield information on the residues that are present in the binding interface, additional subsequent experiments or computational studies must be performed to determine the contributions of each of these residues to the protein-protein binding configurations. Alanine-scanning mutagenesis experiments have been the key driver on the experimental front to identify precisely the role of each of the binding-region residues. On the computational front, applications of computational alanine-scanning mutagenesis (in which the energy functional is parameterized by using available experimental alanine-mutagenesis data) on the protein-protein complex structure has been shown to be promising in determining the role of these binding-region residues. While in general, a large number of residues are buried in the protein-protein complex interface, only a fraction of these residues, termed hot-spot residues, are critical to the PPIs. The presence of these hot-spot residues has been confirmed experimentally by alanine mutagenesis experiments in which the mutation of only a few of the binding-region residues to alanine has abrogated the binding of the proteins to a large extent. Although a plethora of computational tools are available to determine the hot-spot residues from the protein-protein complex structure, there is a general lack of computational tools to identify hot-spot residues by using the sequence / structure of the unbound protein alone.

In this work, we have shown that a new computational tool, called SIM, can be used to predict the hot-spot residues of an evolutionarily conserved protein-protein interaction by using the structure of the unbound protein alone. The SIM tool is devised to identify clusters of

exposed hydrophobic residues along with the exposed charged residues because both hydrophobic and electrostatic interactions are expected to contribute greatly to protein-protein binding energy. To identify the exposed hydrophobic residues, the SIM tool uses the normalized (with respect to its fractionally exposed surface area) hydrophobicity value for each residue. In the previous studies from our laboratory, normalized hydrophobicity values for residues have been shown to be superior to non-normalized hydrophobicity values (where the hydrophobicity value of a residue depends only on its residue type irrespective of its exposed area in the protein structure) for the prediction of protein binding-region residues in the laboratory.^{25,26} Moreover, sequence conservation is used as an additional criterion to improve the quality of the SIM predictions because the conservation of residues over evolution is often considered to be an indicator of the importance of the residue for either the protein structure or protein interaction. The SIM tool can be applied either directly to the static structure of the protein or to the multiple conformations generated via the MD simulations. While the requirement of the protein structure limits the applicability of the SIM tool to the proteins with known structure, advances in the structure modeling of proteins using homology modeling can be used to alleviate this limitation. The SIM tool based on molecular simulations to some extent accounts for the contribution of the protein flexibility and dynamic exposure of the residues.

In this work, we validate the predictions of hot-spot residues by the SIM tool for 43 experimentally known hot-spot residues of six proteins: IL-13, IL-2, GHR, Fc-domain, IL-15 and GH. For these experimentally known hot-spot residues, we show that SIM predicts hot-spot residues with an average accuracy of 36-57% for $\Phi_{\text{cutoff}} = 0.2$ and 23-45% for $\Phi_{\text{cutoff}} = 0.15$ (see Supporting Information, Section S4.12; the lower bound represents the average accuracy, while

the upper bound represents the average theoretical maximum accuracy). The hot-spot residue prediction accuracy of the SIM (see Figure 4B) is superior compared to meta-PPISP (3-26%), ISIS (2-26%) and ConSurf (8-26%). PredUs (8-43%) can predict hot-spot residues with a comparable average theoretical maximum accuracy, compared to the SIM tool. Furthermore, from the comparison of SIM predictions with the hydrophobicity predictions, it also becomes evident that the SIM tool, which identifies the clusters of residues that have high effective hydrophobicity and neighboring charged residues, is more accurate for the prediction of hot-spot residues compared to a simple hydrophobic analysis, which identifies all of the exposed hydrophobic residues as hot-spot residues. The average accuracy of SIM for the prediction of binding-region residues (69% for $\Phi_{\text{cutoff}} = 0.2$ and 61% for $\Phi_{\text{cutoff}} = 0.15$), as seen in Figure 4A, is also better than the average accuracy of meta-PPISP (32%), PredUs (51%), ISIS (32%) and ConSurf (33%).

It should be noted that the observed performance of SIM and other computational tools for hot-spot residue prediction is affected by a number of factors. Most importantly, the quality of the experimental data can be dubious. We have observed a number of experiments in which a mutation of a residue that is not a binding-region residue leads to a substantial loss of binding. This allosteric effect of a mutation might be from a protein-structure perturbation that occurs when the mutation occurs. Unless there is an available structure of the protein-protein complex, it becomes difficult to determine, a priori, whether the mutated residue is a binding-region residue or a non-binding region residue. Hence, in the absence of the structure of the protein-protein complex, the experimentalist might report these non-binding region residues as hot-spot residues. These experimental false positives will cause the observed coverage of the predictive

computational methods to be lower than the true coverage. Second, a lack of exhaustive experimental data on the identification of all hot-spot residues of a protein leads to lower observed accuracy of the predictive computational methods than their true accuracy. In fact, for the proteins that we studied, experimental mutagenesis data were lacking for a number of binding-region residues, and this lack of data is reflected by a large difference between the accuracy and the theoretical maximum accuracy of hot-spot residue prediction in our analysis. Third, any computational tool that is based on the structure of the unbound protein will fail to account for the conformational change in the structure of the protein upon binding to its receptor. Although simple molecular-dynamics simulations can account for some of the protein flexibility around the unbound conformation, advanced molecular simulation techniques must be used to observe large protein conformational changes in these simulations.

While the performance of the SIM tool can be hampered due to the above limitations, the SIM tool nevertheless offers some unique advantages. SIM offers flexibility in predicting hot-spot residues with either high accuracy or high coverage. This flexibility of SIM can be applied in systematic mutagenesis experiments to identify hot-spot residues. We suggest that initially a high value of Φ_{cutoff} be used for predicting a small number of residues. The mutagenesis experiments can then be performed on these residues. The value of Φ_{cutoff} can be further lowered in a stepwise fashion to identify a larger number of residues, which can then be tested experimentally. One of the limitations of the SIM tool is its inability to associate the identified hot-spot residues with a binding partner. If the protein binds to two receptors, the SIM tool cannot predict whether the predicted hot-spot residues are involved in binding to the first receptor or the second receptor or both. Hence, when the SIM-identified hot-spot residues test as

negative in experiments in which the binding of the mutated protein to one of its receptors is measured, this test does not conclusively indicate that the predictions are incorrect. The predicted hot-spot residues might be important for binding to its other receptor, and hence, caution should be exercised when comparing the SIM predictions with the experimental results. The strategy of rational mutagenesis by combining experimental mutagenesis with *in silico* hot-spot residue prediction can lead to identification of hot-spot residues by performing a smaller number of experiments. The chances of success from a mutagenesis experiment that attempts to correctly identify hot-spot residues are much higher when guided by the SIM tool.

ACKNOWLEDGMENTS

The authors thank Elise Champion, Fabienne Courtois and Raghvendra Hosur for various helpful discussions.

The authors acknowledge funding support from Novartis Pharma AG and the Singapore-MIT Alliance. Computing resources were made available in part by the National Center for Supercomputing Applications grant MCB100085.

Dr. Bernhard Helk is an employee of Novartis Pharma AG. The authors have filed patent applications on aspects of the disclosed technology.

The paper was generated in a collaboration between Novartis Pharma AG as sponsor and the Massachusetts Institute of Technology. The authors contributed to the study design, the collection, analysis, and interpretation of the data, and the writing of the report. The sponsor contributed to aspects of the decision to submit the paper for publication.

REFERENCES

1. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. *Proc Natl Acad Sci USA* 2008;105(19):6959-6964.
2. Thanos CD, DeLano WL, Wells JA. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc Natl Acad Sci USA* 2006;103(42):15422-15427.
3. Bullock BN, Jochim AL, Arora PS. Assessing helical protein interfaces for inhibitor design. *J Am Chem Soc* 2011;133(36):14220-14223.
4. Cunningham BC, Wells JA. High-resolution epitope mapping of high-receptor interactions by alanine-scanning mutagenesis. *Science* 1989;244(4908):1081-1085.
5. Ofra Y, Rost B. Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol* 2007;3(7):1169-1176.
6. Rao GV, Brooks CL. Functional epitopes for site 1 of human prolactin. *Biochemistry* 2011;50(8):1347-1358.
7. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci USA* 2002;99(22):14116-14121.
8. Lise S, Buchan D, Pontil M, Jones DT. Predictions of hot spot residues at protein-protein interfaces using support vector machines. *PLoS One* 2011;6(2).
9. Xia JF, Zhao XM, Song JN, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics* 2010;11.
10. Tuncbag N, Gursoy A, Keskin O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics* 2009;25(12):1513-1520.
11. Guharoy M, Chakrabarti P. Empirical estimation of the energetic contribution of individual interface residues in structures of protein-protein complexes. *J Comput Aid Mol Des* 2009;23(9):645-654.
12. Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N. PCRPI: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res* 2010;38(6).
13. Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins* 2007;68(4):813-823.
14. Cho KI, Kim D, Lee D. A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res* 2009;37(8):2672-2687.
15. Tuncbag N, Keskin O, Gursoy A. HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res* 2010;38:W402-W406.
16. Lise S, Archambeau C, Pontil M, Jones DT. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics* 2009;10.
17. Ofra Y, Rost B. ISIS: interaction sites identified from sequence. *Bioinformatics* 2007;23(2):E13-E16.
18. Zhang QC, Petrey D, Norel R, Honig BH. Protein interface conservation across structure space. *Proc Natl Acad Sci USA* 2010;107(24):10896-10901.
19. Qin S, Zhou HX. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics* 2007;23(24):3386-3387.

20. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 2003;19(1):163-164.
21. Brenke R, Kozakov D, Chuang G-Y, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. Fragment-based identification of druggable hot spots of proteins using Fourier domain correlation techniques. *Bioinformatics* 2009;25(5):621-627.
22. Nussinov R, Schreiber G, editors. *Computational Protein-Protein Interactions*: CRC Press; 2009.
23. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform* 2009;10(3):217-232.
24. Zhang QC, Deng L, Fisher M, Guan JH, Honig B, Petrey D. PredUS: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res* 2011;39:W283-W287.
25. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci USA* 2009;106(29):11937-11942.
26. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Prediction of protein binding regions. *Proteins* 2011;79(3):888-897.
27. Lijnzaad P, Argos P. Hydrophobic patches on protein subunit interfaces: Characteristics and prediction. *Proteins* 1997;28(3):333-343.
28. Kozakov D, Hall DR, Chuang G-Y, Cencic R, Brenke R, Grove LE, Beglov D, Pelletier J, Whitty A, Vajda S. Structural conservation of druggable hot spots in protein-protein interfaces. *Proc Natl Acad Sci USA* 2011;108(33):13528-13533.
29. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci USA* 1996;93(1):13-20.
30. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol* 1999;285(5):2177-2198.
31. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struc Biol* 2002;12(1):14-20.
32. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol* 1998;280(1):1-9.
33. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics* 2003;19(11):1453-1454.
34. Guney E, Tuncbag N, Keskin O, Gursoy A. HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res* 2008;36:D662-D666.
35. Black SD, Mould DR. Development of hydrophobicity parameters to analyze proteins which bear posttranslational or cotranslational modifications. *Analytical Biochemistry* 1991;193(1):72-82.
36. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 1998;102(18):3586-3616.

37. Fiorucci S, Zacharias M. Prediction of protein-protein interaction sites using electrostatic desolvation profiles. *Biophys J* 2010;98(9):1921-1930.
38. Cuthill E, McKee J. Reducing the bandwidth of sparse symmetric matrices. *Proceedings of the 1969 24th national conference*; 1969. p 157-172.
39. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* 1996;14(1):33-&.
40. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 2007;372(3):774-797.
41. Lupardus PJ, Birnbaum ME, Garcia KC. Molecular basis for shared cytokine recognition revealed in the structure of an unusually high affinity complex between IL-13 and IL-13R alpha 2. *Structure* 2010;18(3):332-342.
42. LaPorte SL, Juo ZS, Vaclavikova J, Colf LA, Qi XL, Heller NM, Keegan AD, Garcia KC. Molecular and structural basis of cytokine receptor pleiotropy in the interleukin-4/13 system. *Cell* 2008;132(2):259-272.
43. Moy FJ, Diblasio E, Wilhelm J, Powers R. Solution structure of human IL-13 and implication for receptor binding. *J Mol Biol* 2001;310(1):219-230.

FIGURE LEGENDS

Figure 1: (A) The contact-map matrix for IL-13 generated using its structure (PDB ID: 1IJZ). The indices represent the residue numbers. Green indicates 0 while yellow indicates 1. (B) The contact-map matrix for IL-13 after the application of a high-hydrophobicity filter. Here, we used $\Phi_{\text{cutoff}} = 0.15$. (C) The contact-map matrix is clustered using the reverse Cuthill-McKee algorithm. Cluster C3 has only one element and hence is discarded. All of the residues in cluster C4 are very close to the N-terminal and, hence, C4 is also discarded. The row and column of this matrix do not represent the residue number.

Figure 2: (A) The sSIM map of IL-13 for $\Phi_{\text{cutoff}} = 0.15$. The red region indicates residues that were predicted by sSIM. (B) Experimental hot spots (C) Hydrophobicity scale mapped onto the IL-13 structure. The red (value > 0) indicates hydrophobic residues. (D) ConSurf scores mapped onto the IL-13 structure. The red (value < 0.5) indicates conserved residues.

Figure 3: (A) Accuracy and coverage of various methods for the predictions of binding-region residues of IL-13. The results for sSIM (green) and dSIM (red) are also shown for various values of Φ_{cutoff} . (B) Accuracy and coverage of various methods for the predictions of hot-spot residues of IL-13. The results for sSIM (green) and dSIM (red) are also shown for various values of Φ_{cutoff} . The error bars indicate the theoretical maximum accuracy. Because the experimentally known hot spot K105 is not conserved (ConSurf score = 2.83), the exposed and conserved criteria have a coverage of less than 100%.

Figure 4: (A) Average accuracy and average coverage of various methods for the prediction of the binding-region residues of six proteins (IL-13, IL-2, GHR, Fc, IL-15, and GH). The results for sSIM (green) are shown for two values of Φ_{cutoff} . (B) Average accuracy and average coverage of various methods for the prediction of the hot-spot residues of six proteins. The results for sSIM (green) are shown for two values of Φ_{cutoff} . The error bars indicate the theoretical maximum accuracy. Note that the y-axis scale is 0-60%.

Figure 1

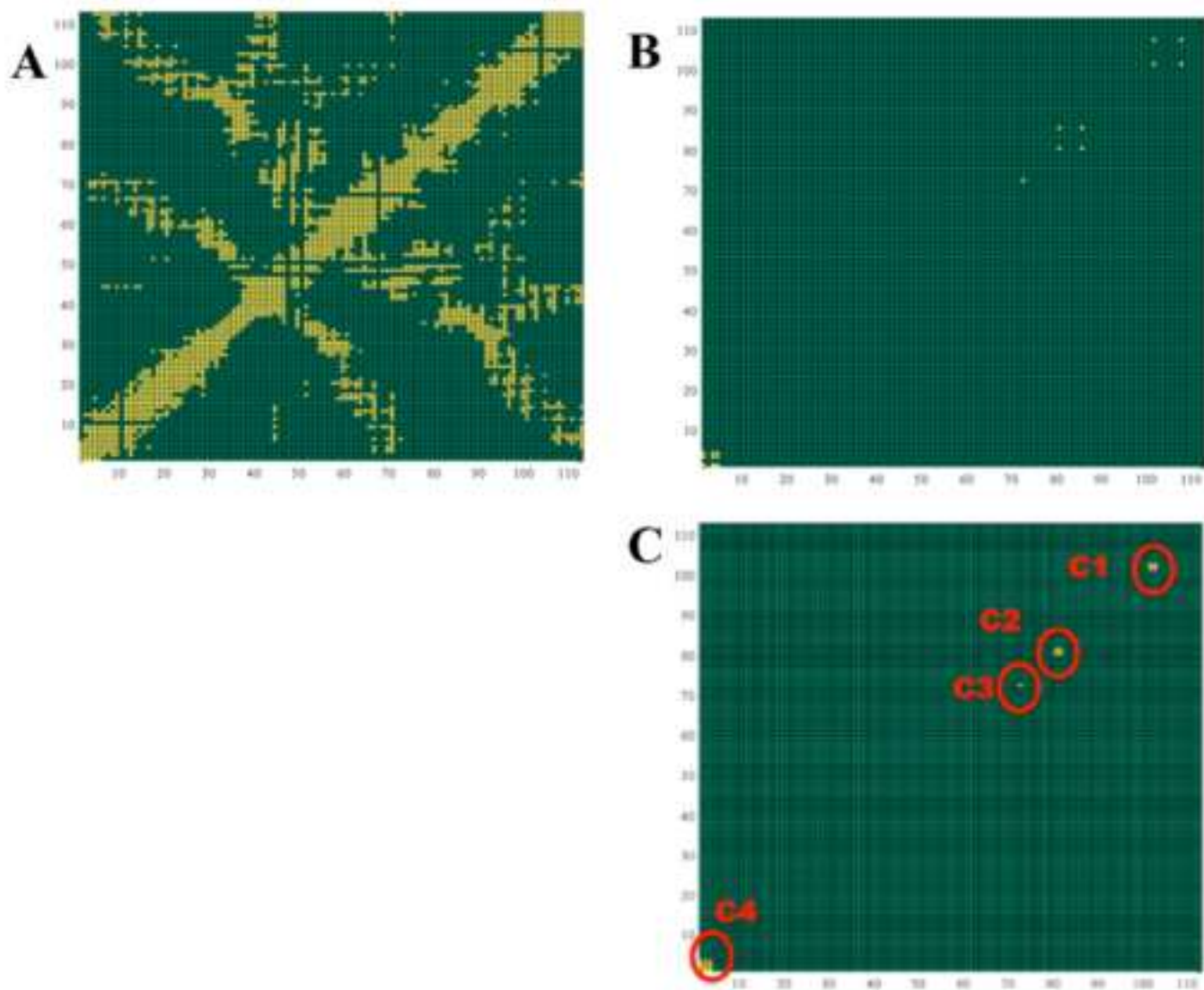


Figure 2

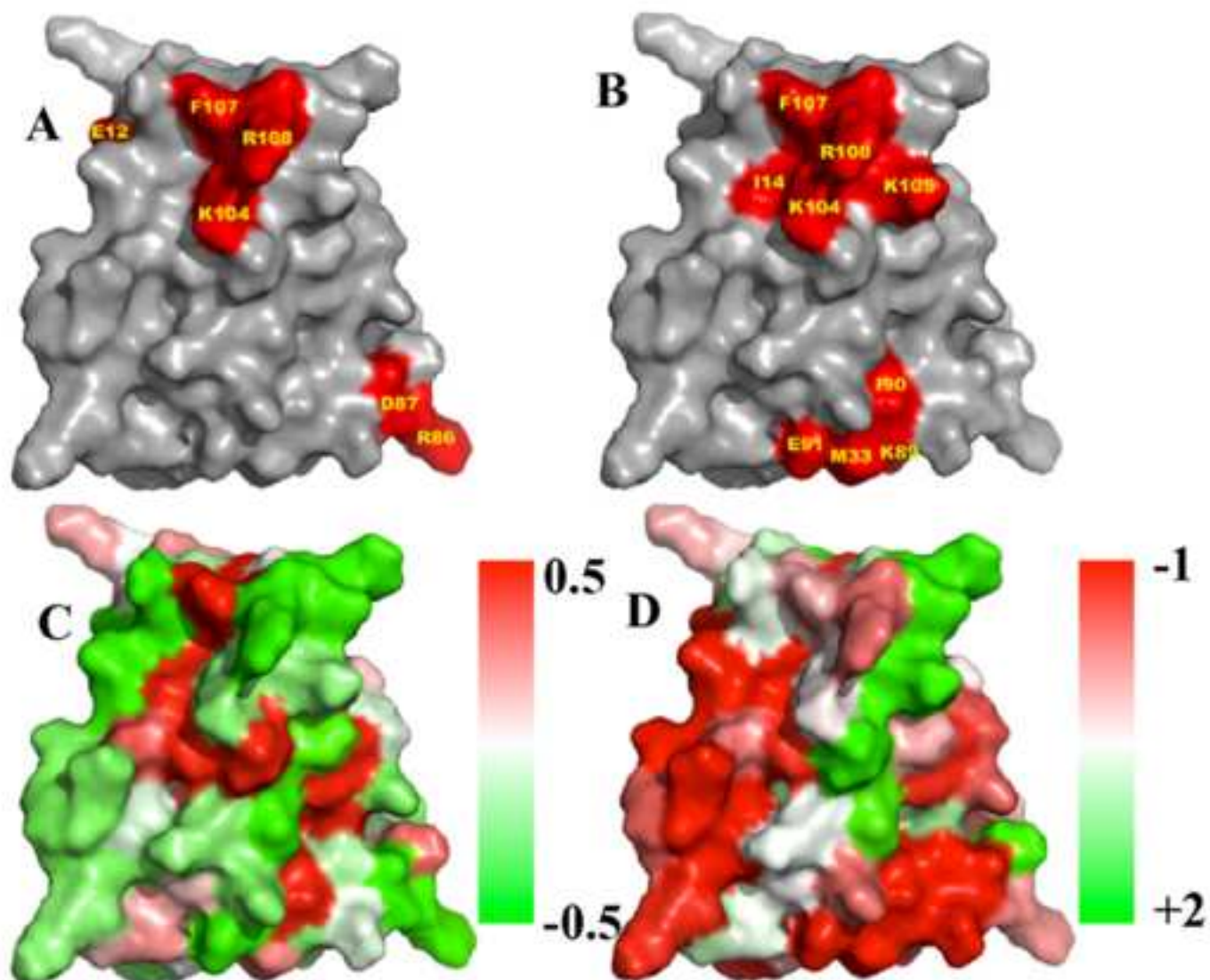
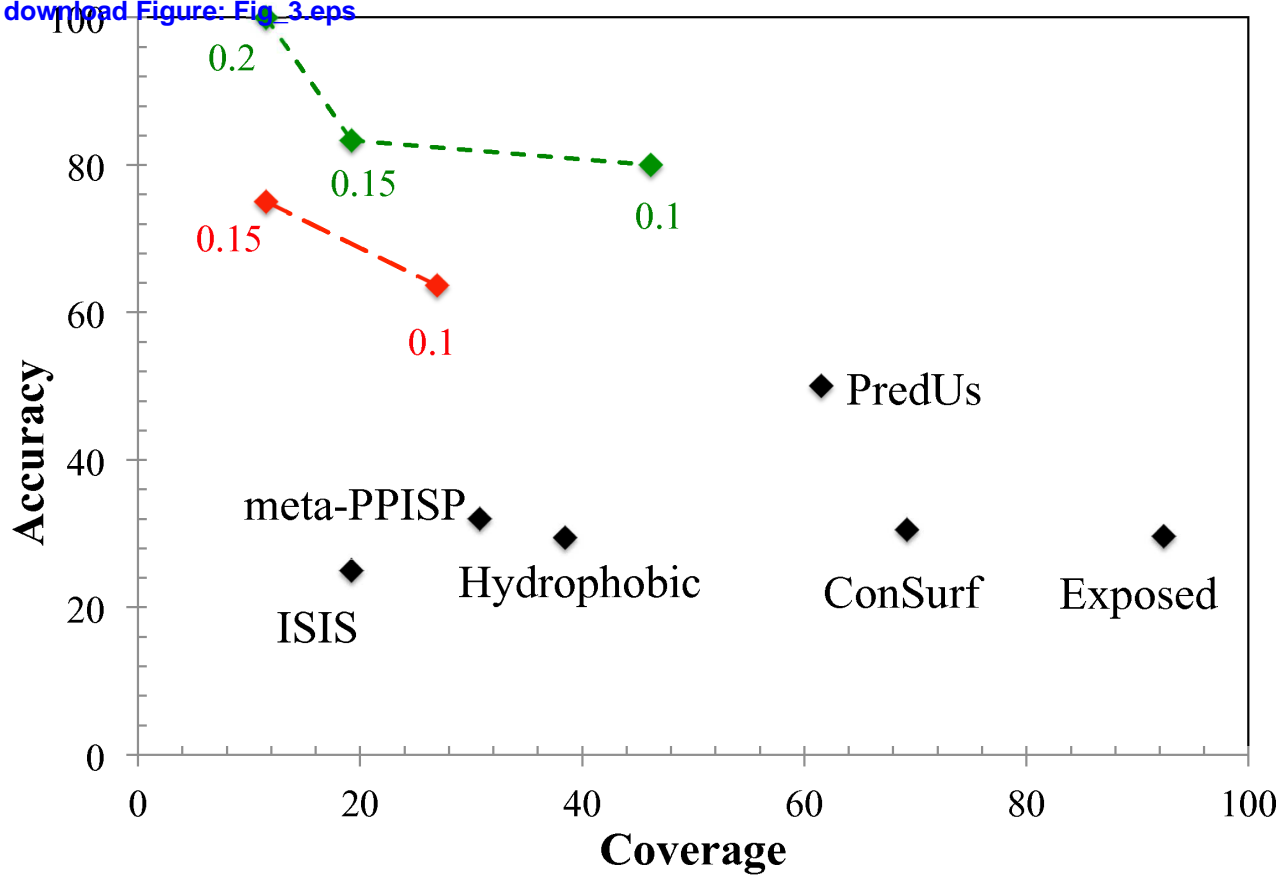


Figure 3

[Click here to download Figure: Fig_3.eps](#)



B)

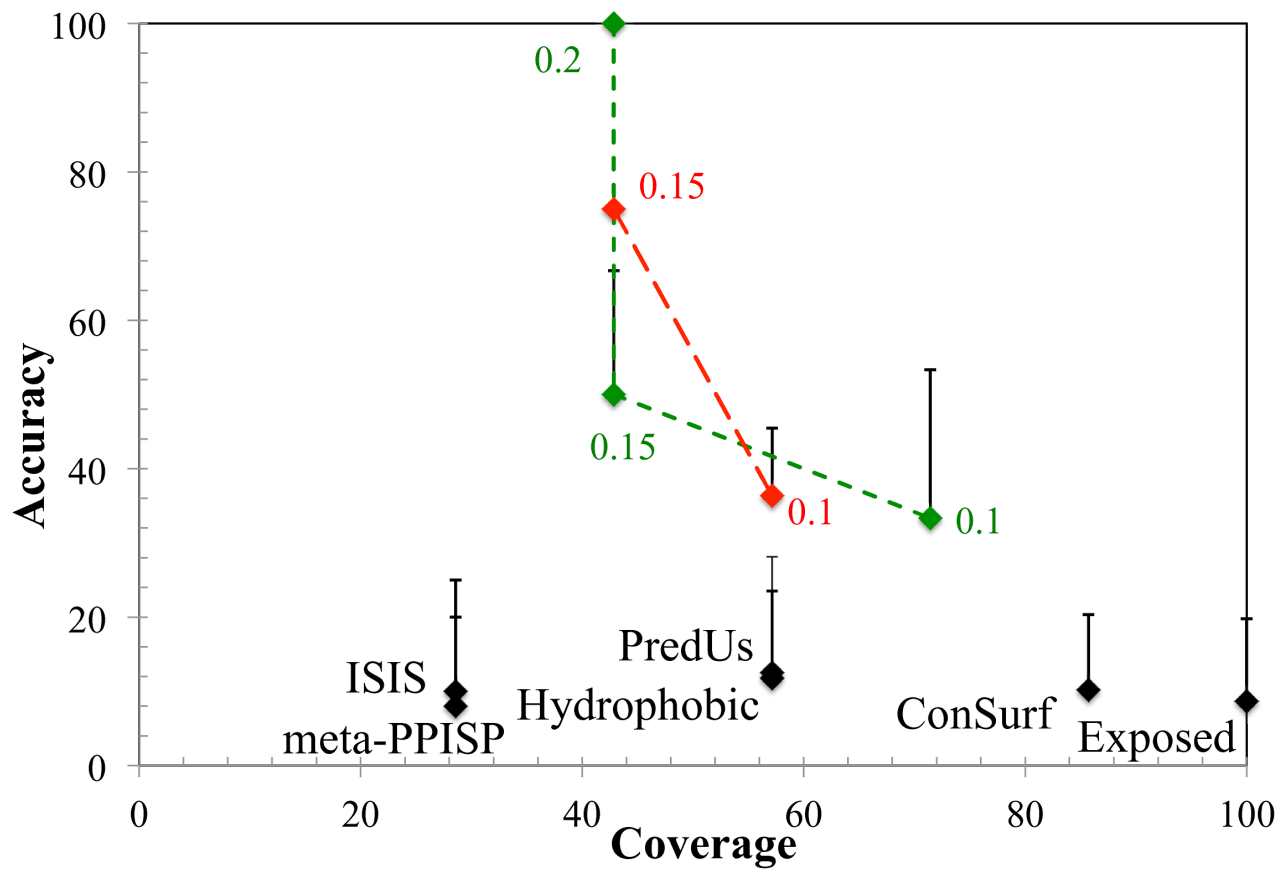
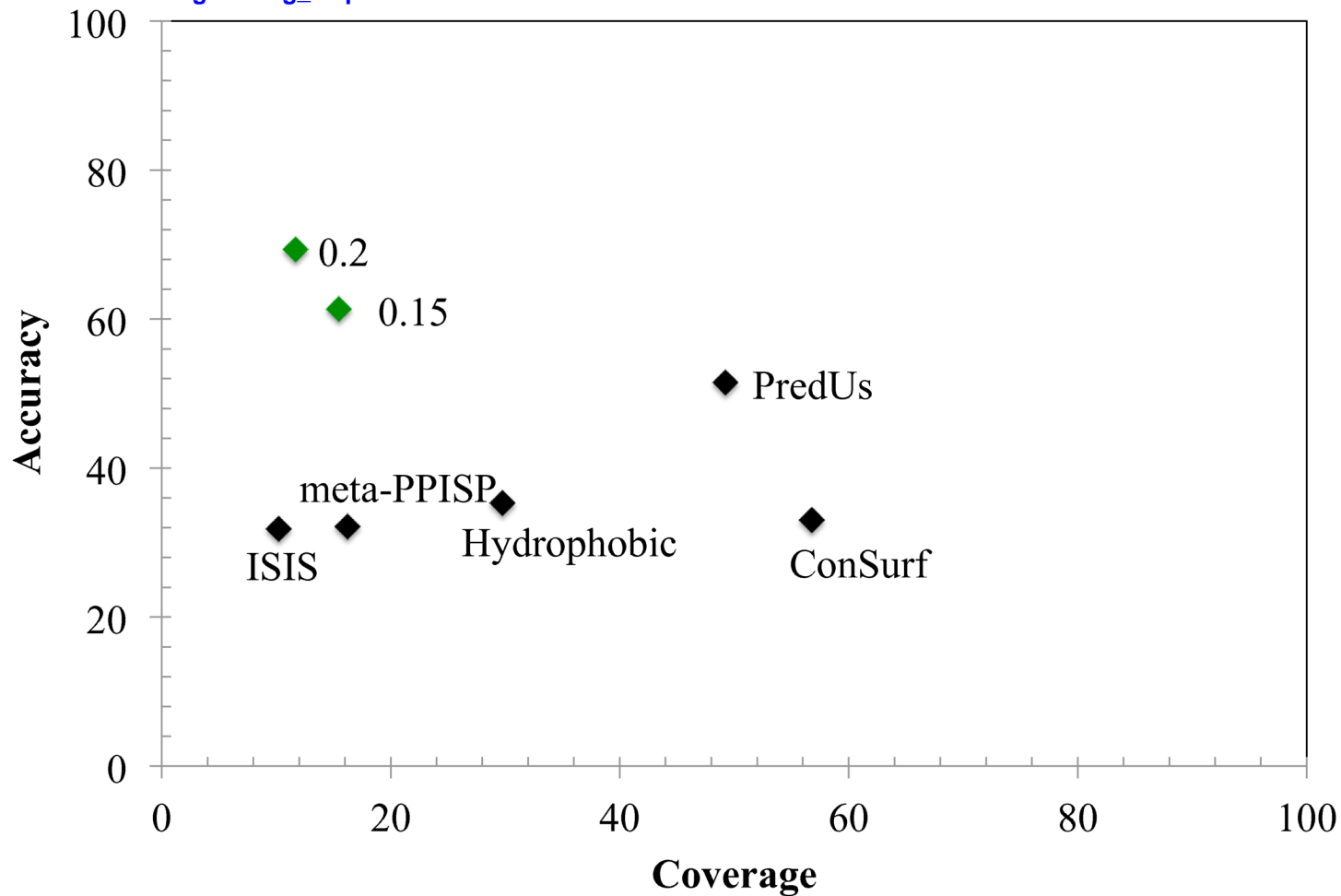
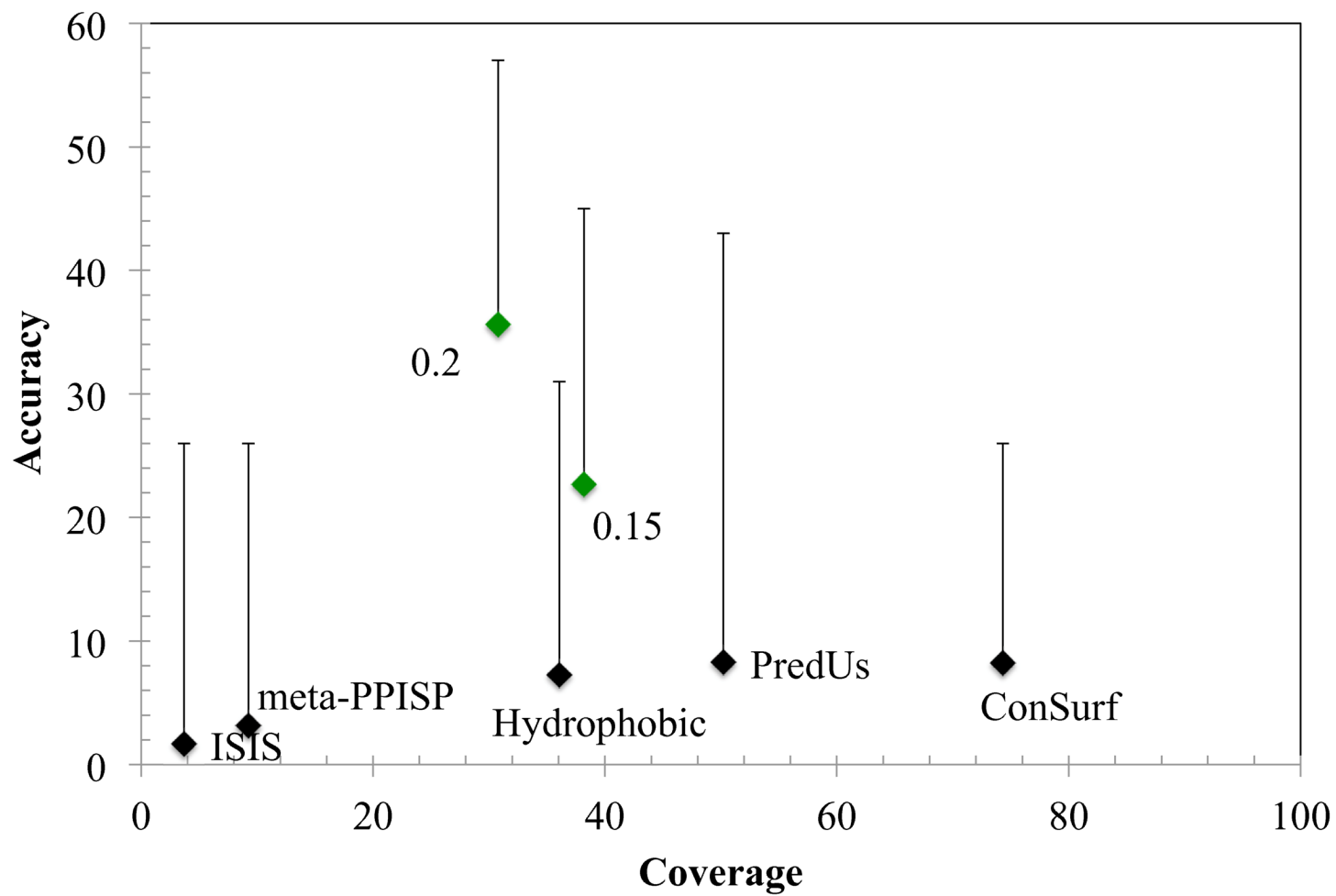


Figure 4
A) [Click here to download Figure: Fig_4.eps](#)



B)



Supplementary material for online publication only

[Click here to download Supplementary material for online publication only: SUPPORTING INFORMATION.pdf](#)