ABSTRACT

Motivation: Protein–protein interactions (PPIs) are a promising, but challenging target for pharmaceutical intervention. One approach for addressing these difficult targets is the rational design of small-molecule inhibitors that mimic the chemical and physical properties of small clusters of key residues at the protein–protein interface. The identification of appropriate clusters of interface residues provides starting points for inhibitor design and supports an overall assessment of the susceptibility of PPIs to small-molecule inhibition.

Results: We extract Small-Molecule Inhibitor Starting Points (SMISPs) from protein-ligand and protein–protein complexes in the Protein Data Bank (PDB). These SMISPs are used to train two distinct classifiers, a support vector machine and an easy to interpret exhaustive rule classifier. Both classifiers achieve better than 70% leave-one-complex-out cross-validation accuracy and correctly interpret SMISPs of known PPI inhibitors not in the training set. A PDB-wide analysis suggests that nearly half of all PPIs may be susceptible to small-molecule inhibition.


Contact: dkoes@pitt.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on August 22, 2011; revised on October 31, 2011; accepted on December 21, 2011

1 INTRODUCTION

Protein–protein interactions (PPIs) play a key role in nearly every biological function and are a promising new class of biological targets for therapeutic intervention. Dömling (2008) and McClendon, 2007). PPIs present a number of unique challenges compared to targets that have historically dominated pharmaceutical efforts, such as enzymes, G-protein-coupled receptors, and ion-channels. Dömling et al. 2008. Unlike these targets, which have evolved to bind small molecules, PPIs have no convenient natural substrate to serve as a starting point for small-molecule design. However, alanine scanning mutagenesis reveals that most of the evolved to bind small molecules, PPIs have no convenient natural channels (Paolini et al. 2004). A small cluster of co-located interface residues that includes at least one hot spot provides a starting point for the rational design of small molecule inhibitors. Meeres et al. 2011. Indeed, the chemical mimicry of small clusters of key residues, typically deeply buried ‘anchor’ residues (Rajamani et al. 2004), has produced inhibitors for several PPI targets Christ et al. 2010, Liu et al. 2008, Popowicz et al. 2011.

The importance of hot spot residues and the difficulty in performing alanine scanning mutagenesis has led to the development of several computational methods for predicting hot spots. Energy-based methods Camacho and Zhang, 2005; Rost et al. 2004 attempt to directly compute the energetic contributions of residues. Solvent accessible surface area (SASA) Meeres et al. 2010, Rajamani et al. 2004 and sequence conservation (Bromberg and Rost, 2008; Keskin et al. 2005; Lucht et al. 1999) have also been used to predict hot spots. However, the most successful computational methods use some combination of these and other features Cho et al. 2004; Darve et al. 2003; Guney et al. 2006; Lise et al. 2006; Tuncbag et al. 2006; Zhu and Mitchell, 2011 and achieve accuracies of 60–80%. Often a consensus scoring mechanism is derived using machine learning techniques such as decision trees Darve et al. 2003 or support vector machines Cho et al. 2004; Lise et al. 2006; Tuncbag et al. 2006; Guney et al. 2006, although ad hoc consensus schemes are effective as well Guney et al. 2006; Tuncbag et al. 2006.

In order to enhance specificity and affinity, additional interactions beyond those present in a single individual hot spot residue are needed. Nearby residues that do not meet the criteria of a hot spot may also play an important, if not essential, role in the interaction. Additionally, these residues may describe extra pockets or energetics essential for small molecule binding or specificity. Although identifying a minimum set of stereochemical properties consistent with a small-molecule binding site is challenging Hajduk et al. 2006; Pont et al. 2004, the increasing number of ligand-protein and protein–protein structures in the Protein Data Bank (PDB) provides a foundation to explore this important problem with respect to PPIs. In fact, a systematic analysis of such structures reveals that residues that participate in both ligand and protein binding have distinctly different characteristics from other interface residues Davis and Sun, 2010. This important insight suggests that it may be possible to automatically identify those interface residues that are most susceptible to small-molecule intervention.

In this work we describe a novel structural bioinformatics approach that identifies and ranks those clusters of interface residues in a PPI that are most suitable as starting points for rational small-molecule design. We refer to these clusters as Small-Molecule Inhibitor Starting Points (SMISPs). A SMISP is larger than a hot spot, but substantially smaller than the entire collection of interface residues. A SMISP cluster may include both those residues critical to the protein–protein interaction and those with features important for binding specificity, all within a volume accessible to a small molecule.
SMISPs are complementary to approaches that identify binding sites through an analysis of the receptor surface (Arora, 2010), either through shape descriptors (Weis et al., 2007) or chemical probes (Henschel et al., 2008; Fuller et al., 2009). However, a SMISP, as a collection of interface residues, not only defines a binding site, it also defines a binding mode selected by evolution that provides an initial target for rational small-molecule design. More general binding site identification techniques can then provide insight on how to extend this natural site or explore the flexibility of the site.

Previous work has identified clusters of interface residues from helical interfaces as small-molecule starting points (Jochim and Arora, 2010). A manually specified energy criteria based off of starting points in 48% of protein–protein interactions. SMISPs to identify known PPI inhibition sites. Finally, a PDB-wide in defining SMISPs. We demonstrate the ability of our predicted and role of various factors, such as SASA and free energy estimates, in defining SMISPs. We demonstrate the ability of our predicted SMISPs to identify known PPI inhibition sites. Finally, a PDB-wide analysis predicts the existence of suitable small-molecule inhibitor starting points in 48% of protein–protein interactions.

In a departure from analyses that calibrate to free energies from alanine scanning experiments, we identify SMISPs directly from protein–protein and protein-ligand structures. From this structural analysis we develop a consensus score based on physical and evolutionary descriptors for predicting and ranking SMISPs. We develop a methodology for learning two distinct classifiers: an evolutionary descriptors for predicting and ranking SMISPs. We annotate all clusters of interface residues with aggregate indicators of energy, SASA, and sequence overlap with known small-molecule binding sites. We identify hot spot residues and interface residues (Cho et al., 2007; Lise et al., 2009). For these problems, experimental data, in the form of alanine scanning experiments and PPI crystal structures, is readily available. In order to generate a similar benchmark set of SMISPs, we mine the structural data in the PDB by identifying PPI interface residues that overlap with known small-molecule binding sites. We annotate all clusters of interface residues with aggregate indicators of energy, SASA, and sequence conservation. This benchmark set is then used to train both an exhaustive rule classifier, which generates an easy to interpret rule and a support vector machine (SVM) classifier, for ranking SMISPs. Our approach allows us to examine the importance and role of various factors, such as SASA and free energy estimates, in defining SMISPs. We demonstrate the ability of our predicted SMISPs to identify known PPI inhibition sites. Finally, a PDB-wide analysis predicts the existence of suitable small-molecule inhibitor starting points in 48% of protein–protein interactions.

2 METHODS

We use machine learning techniques to learn both filtering and scoring criteria for identifying SMISPs. Similar approaches have successfully been used to identify hot spot residues and interface residues (Cho et al., 2009; Darnell et al., 2009b). For these problems, experimental data, in the form of alanine scanning experiments and PPI crystal structures, is readily available. In order to generate a similar benchmark set of SMISPs, we mine the structural data in the PDB by identifying PPI interface residues that overlap with known small-molecule binding sites. We annotate all clusters of interface residues with aggregate indicators of energy, SASA, and sequence conservation. This benchmark set is then used to train both an exhaustive rule classifier, which generates an easy to interpret filter, and an SVM classifier, which produces a numerical score. The complete workflow of our method is shown in Figure 1.

To support a PDB-wide analysis, we first generate a non-redundant subset of PPI complexes. We use the search feature of the PDB to retrieve all protein-only structures with at least two chains in the biological assembly with a resolution of less than 4.0 Å that have less than 95% sequence similarity resulting in 13,415 structures (downloaded 6/25/2011). After removing misclassified, non-interacting, or otherwise unanalyzable assemblies, 12,456 complexes remain to form our non-redundant set of PPI structures.

2.1 Benchmark set

Protein–protein interactions remain a challenging and relatively untargeted area of pharmaceutical intervention resulting in a scarcity of proven small-molecule starting points. However, an analysis of the structures deposited in the PDB reveals that for many PPIs there exists a structure where at least one chain of the PPI is complexed with a small molecule that binds at the PPI interface site. A sufficiently high-affinity small-molecule targeting such a site will, at the very least, perturb the interaction. Consequently, we utilize such ligand-bound structures to identify benchmark SMISPs in the original PPI. A benchmark SMISP is the collection of all interface residues from a PPI structure that overlap a high-affinity ligand from a protein-ligand structure aligned to the PPI structure. A benchmark SMISP at least partially delineates the binding site of the ligand, thus providing a validated starting point for the design of a small-molecule inhibitor.

For each chain of each complex in our non-redundant set, we identify all structures in the PDB that have 95% or greater sequence similarity to this receptor chain and that are bound to a standalone ligand (i.e., not a modified residue). We consider only ligands with a molecular weight greater than 150 Da to eliminate non-specific interactions such as ions and crystallographic buffers. We then align the ligand-bound structure to the original PPI complex. The collection of at least two PPI interface residues that contain atoms that overlap the atoms of the ligand in the ligand-bound structure in this aligned assembly is marked as a SMISP. Atom centers must be less than 2.5 Å apart for atoms of the ligand and a residue to be considered overlapping (i.e., less than the distance of a hydrogen bond). In some cases the ligand-bound structure is not a single chain protein, but a protein–protein complex that is homologous to the original PPI complex. In this case we impose an additional constraint that the backbone in the
An interface residue is defined as a residue where the change in SASA incorporating binding affinity data from the PDBbind benchmark set since a significant perturbation of the interface structure will be substantially distorted from the original. GLU-29 and ASN-27 delineate a polar region of the receptor that close around the inhibitor in the ligand-bound structure. The backbones of the remaining residues serve to further delineate the inhibitor binding site. The interface-2 cytokine receptor complex (PDB: 2BS1) shown with an 8 μM inhibitor (PDB: 1M48). The positively charged guanidinium group of ARG-36 overlays a guanidinium on the inhibitor while ILE-68 fills a hydrophobic pocket that closes around the inhibitor, while ILE-42 fills a hydrophobic pocket that closes around the inhibitor in the ligand-bound structure.

The change in absolute SASA of a residue is calculated by subtracting the un-complexed SASA. For every cluster of interface residues, we generate a collection of aggregate features for use in SMISP classification. For each residue in the cluster we compute two energy scores, \( \Delta G^\text{conf} \) and \( \Delta G^\text{desolv} \), the absolute (\( \Delta \text{SASA} \)) and the relative (\( \Delta \text{SASA}/\text{SASA} \)) change in solvent accessible surface area, and two measures of sequence conservation, an evolutionary rate (Rate) and a conservation score (Cons). These features are aggregated into the minimum, maximum, average, and total value for each cluster. For all calculations, we use the first biological assembly deposited in the PDB and preprocess the structure with CHARMM version 31b1. CHARMM is used to add missing atoms, including hydrogens, and to quickly preprocess the structure with CHARMM version 31b1 (Brooks et al. 1983).

\( \Delta G^\text{conf} \) FastContact Camacho and Zhang (2006) is used to compute a per-residue estimate of the free energy (kcal/mol) of complexation. It includes both electrostatic (\( \Delta G^\text{e} \)) and desolvation (\( \Delta G^\text{desolv} \)) effects within an atomistic pair-wise potential. More negative values indicate energetically favorable residues.

\( \Delta \text{SASA} \) We use version 3.2.1 of the Rosetta software (Kinnear et al. 2004) to perform computational alanine scanning. The AlaScan filter is used with the \( \Delta G \) optimized scoring function, an interface distance cutoff of 10 Å, five repeats, and an initial local refinement docking. This value represents the change in free energy of the alanine mutation, so a larger, more positive value indicates that the alanine mutation destabilizes the complex and the original residue is a hot spot.

\( \Delta \text{SASA}/\text{SASA} \) The change in absolute SASA of a residue is calculated by subtracting the SASA of the residue in the PPI complex from the SASA of the residue when all other protein chains have been removed from the PPI structure. That is, the bound conformation of the chain of the residue is used to compute the un-complexed SASA.

For every cluster of interface residues, we generate a collection of aggregate features for use in SMISP classification. For each residue in the cluster we compute two energy scores, \( \Delta G^\text{conf} \) and \( \Delta G^\text{desolv} \), the absolute (\( \Delta \text{SASA} \)) and the relative (\( \Delta \text{SASA}/\text{SASA} \)) change in solvent accessible surface area, and two measures of sequence conservation, an evolutionary rate (Rate) and a conservation score (Cons). These features are aggregated into the minimum, maximum, average, and total value for each cluster. For all calculations, we use the first biological assembly deposited in the PDB and preprocess the structure with CHARMM version 31b1. CHARMM is used to add missing atoms, including hydrogens, and to quickly preprocess the structure with CHARMM version 31b1. CHARMM is used to add missing atoms, including hydrogens, and to quickly preprocess the structure with CHARMM version 31b1.

\( \Delta G^\text{conf} \) FastContact Camacho and Zhang (2006) is used to compute a per-residue estimate of the free energy (kcal/mol) of complexation. It includes both electrostatic (\( \Delta G^\text{e} \)) and desolvation (\( \Delta G^\text{desolv} \)) effects within an atomistic pair-wise potential. More negative values indicate energetically favorable residues.

\( \Delta \text{SASA} \) We use version 3.2.1 of the Rosetta software (Kinnear et al. 2004) to perform computational alanine scanning. The AlaScan filter is used with the \( \Delta G \) optimized scoring function, an interface distance cutoff of 10 Å, five repeats, and an initial local refinement docking. This value represents the change in free energy of the alanine mutation, so a larger, more positive value indicates that the alanine mutation destabilizes the complex and the original residue is a hot spot.

\( \Delta \text{SASA}/\text{SASA} \) The change in absolute SASA of a residue is calculated by subtracting the SASA of the residue in the PPI complex from the SASA of the residue when all other protein chains have been removed from the PPI structure. That is, the bound conformation of the chain of the residue is used to compute the un-complexed SASA.
Table 1. Validation set of PPIs with known inhibitors

<table>
<thead>
<tr>
<th>Description</th>
<th>PPI PDB</th>
<th>Ch</th>
<th>Lig. PDB</th>
<th>#SMISPs</th>
<th>#Chem</th>
</tr>
</thead>
<tbody>
<tr>
<td>p53/Mdm2</td>
<td>1YCR</td>
<td>B</td>
<td>3LBK</td>
<td>100</td>
<td>311</td>
</tr>
<tr>
<td>p53/Mdm4</td>
<td>3DAB</td>
<td>B</td>
<td>3LBJ</td>
<td>45</td>
<td>79</td>
</tr>
<tr>
<td>VEGF/Tie2</td>
<td>2VRW</td>
<td>B</td>
<td>3CLX</td>
<td>30</td>
<td>537</td>
</tr>
<tr>
<td>XAP/Smac</td>
<td>1G73</td>
<td>A</td>
<td>2IKK</td>
<td>117</td>
<td>1793</td>
</tr>
<tr>
<td>HIV gp41</td>
<td>1AKC</td>
<td>C</td>
<td>2KPS</td>
<td>66</td>
<td>639</td>
</tr>
<tr>
<td>Bcl-2/Bcl1</td>
<td>2YLF</td>
<td>B</td>
<td>2O2N</td>
<td>1503</td>
<td></td>
</tr>
<tr>
<td>HIV-1 Integase/p75</td>
<td>2XAO</td>
<td>B</td>
<td>2O22</td>
<td>762</td>
<td>1327</td>
</tr>
<tr>
<td>Zmp1/Ank2</td>
<td>2F47</td>
<td>A</td>
<td>113YF</td>
<td>40</td>
<td>263</td>
</tr>
<tr>
<td>HIV-1 p55</td>
<td>1UEC</td>
<td>A</td>
<td>1UEC</td>
<td>56</td>
<td>312</td>
</tr>
<tr>
<td>TNF- α</td>
<td>1TNY</td>
<td>C</td>
<td>2AZ5</td>
<td>80</td>
<td>2979</td>
</tr>
</tbody>
</table>

PPIs were identified from the literature: (Bartlett and Palade, 2010) and only examples with known structure that bind at the PPI interface are considered. The number of SMISPs predicted by our combined classifier is shown with the total number of clusters evaluated. As clusters are all possible collections of co-located interface residues, the number of clusters is combinatorially related to the number of interface residues.

PPIs have large benchmark SMISPs that are decomposed into all possible maximal subsets of residues that fit within a 12Å distance cutoff.

Negative examples are generated by selecting clusters of interface residues from the same chain(s) as the benchmark SMISPs that are equal in size to the SMISPs selected as positive examples and in no way overlap any benchmark SMISPs of the PPI.

Since the number of negative examples is typically much larger than the number of positive examples and SVMs perform poorly on imbalanced training data, we re-sample the data to create a balanced training set. (Bartlett and Palade, 2010) For both positive and negative examples, random selection with replacement is performed. A sample size of 100 was found to produce stable results across multiple trials of random sampling.

Missing values, which are present for the sequence conservation scores when no similar sequences are found, are replaced by average values. Twelve PPIs in the benchmark set contain no negative examples because the ligand when no similar sequences are found, are replaced by average values. Twelve SMISPs selected as positive examples and in no way overlap any benchmark SMISPs from the same chain(s) as the benchmark SMISPs that are equal in size to the positive examples. The number of SMISPs predicted by our combined classifier is shown with the total number of clusters evaluated. As clusters are all possible collections of co-located interface residues, the number of clusters is combinatorially related to the number of interface residues.

Validation set

In order to qualitatively characterize the performance of our classifiers, we construct a validation set of 11 PPI complexes with known inhibitors where both the protein–protein and protein-ligand structures are available. These complexes were identified from recent publications (Stewart et al., 2009). Wells and McClendon, 2007 and PPI databases (Bartlett and Palade, 2010). Blommers et al., 2008) and are shown in Table 1. These complexes are chosen from the training set because the inhibitors were not present in the binding affinity databases or a sufficiently homologous complex was not available in the non-redundant subset. We do not consider three PPIs (PDBs: 3DT7, 1TNR, 2NOD) with known inhibitors and structure where the inhibitor binding site does not overlap the PPI interface. These inhibitors are presumed to function allosterically and are beyond the scope of our SMISP classification. The overlap between this validation set and the training set is with the Bcl-2/Bax/HB3 complex of Figure 2. The validation set includes a Bcl-2/Bax/HB3 complex.
Table 2. Single attribute rules.

<table>
<thead>
<tr>
<th>Info. Gain</th>
<th>Rule</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.137</td>
<td>Ave ΔSASA</td>
<td>≥ 44.6 68 ± 3.6</td>
</tr>
<tr>
<td>0.128</td>
<td>Ave ΔSASA%</td>
<td>≥ 39.6 68 ± 3.7</td>
</tr>
<tr>
<td>0.103</td>
<td>Ave ΔGIC</td>
<td>&lt; -2.3 67 ± 3.9</td>
</tr>
<tr>
<td>0.081</td>
<td>Ave ΔGIC−</td>
<td>&lt; -1.64 61 ± 3.4</td>
</tr>
<tr>
<td>0.070</td>
<td>Max ΔGIC</td>
<td>≥ 0.427 62 ± 5.1</td>
</tr>
<tr>
<td>0.042</td>
<td>Min ΔGIC−</td>
<td>&lt; -1.3 61 ± 3.7</td>
</tr>
<tr>
<td>0.038</td>
<td>Total Rate4</td>
<td>≥ 1.6 52 ± 2.7</td>
</tr>
<tr>
<td>0.032</td>
<td>Min Cons</td>
<td>&lt; 0.086 52 ± 1.5</td>
</tr>
</tbody>
</table>

The optimal single attribute rules for the most informative aggregate statistic of each of the computed properties are shown ranked by information gain. The information gain and rule thresholds are computing using the entire training set. The accuracies are averages across the 39 train-test cross-validation sets and are shown with the standard error of the mean. The complete list of attributes with additional measures of performance are shown in Supplementary Table S2.

Table 3. The three most informative two-attribute rules.

<table>
<thead>
<tr>
<th>Info. Gain</th>
<th>Rule</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.175</td>
<td>Ave ΔASASA</td>
<td>≥ 44.6 72 ± 3.8</td>
</tr>
<tr>
<td>0.171</td>
<td>Ave ΔASASA%</td>
<td>≥ 39.6 71 ± 3.8</td>
</tr>
<tr>
<td>0.167</td>
<td>Max ΔΔGR</td>
<td>≥ 0.425 66 ± 3.8</td>
</tr>
<tr>
<td></td>
<td>Ave ΔGR</td>
<td>≥ 46.1</td>
</tr>
</tbody>
</table>

The information gain and rule thresholds are computed using the entire training set. The accuracies are averages across the 39 train-test cross-validation sets and are shown with the standard error of the mean. The ten most informative two-attribute rules are shown with additional measures of performance in Supplementary Table S3.

Although interface residues have been found to have a distinguishable conservation profile (Bordner and Abagyan, 2003), hot spot prediction using only sequence performs poorly (Ofran and Rost, 2003; Tuncbag et al., 2009) and residues that participate in both ligand and protein binding have been found to be less conserved (Brown and Sali, 2010). Consistent with these results, we find that the conservation metrics are the least informative and that, on average, there is a slight preference for predicted SMISPs to be less conserved than the rest of the interface.

Various forms of ΔSASA are clearly the best indicators of SMISPs. However, as shown in Table 2, energy metrics complement surface area metrics when combined in a two-attribute rule. Seven of the top ten most informative two-attribute rules contain some combination of energy and surface area terms. The combination of average ΔGIC and total ΔSASA% has a cross validation accuracy of 71% and provides greater specificity, 91%, than the most informative surface area only rule. This is similar to rule-based hot spot prediction (Tuncbag et al., 2009), where a manually derived...
rule consisting of an energy pair-potential and relative SASA had an accuracy of 70% on an independent test set.

3.2 Validating predicted SMISPs

Eight of the eleven PPIs of our validation set are shown with a predicted SMISP in Figure 3 and the remaining complexes, which are similar, are shown in Supplementary Figure S2. The SMISPs are filtered using the most informative two-attribute rule from Table 3 and the 0.55 SVM score threshold. The SMISPs are then ranked by score. Both classifiers are trained on the entire training set. The number of identified SMISPs for each PPI is shown in Table 4 and ranges from 30 to 762. The largest SMISP in the top three SMISPs is shown in Figure 3 in order to illustrate the diversity of interactions that are present in a top ranked SMISP (the smaller top ranked SMISPs are typically subsets of the shown SMISP).

In nine of the complexes shown in Figures 3(a–f) and Supplementary Figure S2, the SMISPs clearly identify the binding site and, in many cases, duplicate the hydrophobic and electrostatic interactions of the small-molecule ligands. In fact, eight of these nine complexes possess ligands that were designed specifically to mimic the residues of the predicted SMISP or to otherwise target the pocket described by the predicted SMISP and therefore represent a retrospective validation of the predictions. In two cases, p53/MDM2 and HIV gp41, the predicted SMISPs contain experimentally verified hot spot residues (Chan et al. 1998; Lin et al. 1993). The Supplementary Data elaborates on the properties and significance of these predictions. In addition, all 12 short-peptide PPIs that were omitted from the training set as trivial classifications are correctly identified (not shown).

In the remaining two complexes the predicted SMISPs have little or no overlap with the inhibitor. In both Figure 3(g) and h the inhibitor clashes with the PPI receptor indicating that it binds to a significantly different receptor conformation. Consequently, significant portions of the small molecule do not overlap any PPI residues, preventing the identification of a SMISP. Of course, the SMISPs that are predicted may provide an alternative, as yet unexplored, mechanism of inhibition.

In most cases, our top ranked SMISPs partially or fully delineate the binding site of a known inhibitor, indicating that SMISPs provide an immediate and useful computational hypothesis for the initiation of structure-based design. We provide examples of predicted SMISPs with potential therapeutic application in the Supplementary Data. Since our method provides a ranking of SMISPs, for a specific PPI we can always identify the most promising SMISP. A list of the top five SMISPs predicted for every chain in our non-redundant dataset is provided in the Supplementary Data and the complete set is accessible through an online search interface at http://pocketquery.csb.pitt.edu.

3.3 PDB-Wide analysis

We use our SMISPs classifiers trained on the full training set to analyze our entire set of PPIs. The interface residues of these PPIs are decomposed into more than 95 million clusters. Since true SMISP clusters are expected to be relatively rare, we combine the rule classifier and SVM score threshold with the highest cross-validation accuracy. A cluster is only classified as a SMISP if it is predicted by both classifiers. Unsurprisingly, given the cross-validation accuracy of the classifiers, we find that 95% of the protein chains in our dataset have at least one predicted SMISP. In order to extend our predictions of individual SMISPs to an overall assessment of the susceptibility of PPIs to inhibitor design, we evaluate the distribution of predicted SMISPs within each PPI. We consider the quantity of SMISPs (the percent of interface clusters that are predicted SMISPs) and the quality (the maximum score of the predicted SMISPs).

Figure 4 shows a density plot of the number of PPI chains in our non-redundant set whose predicted SMISPs fall within given quantity and quality thresholds. For example, as indicated by the color plotted at (0%, 0.85), there are 14,252 chains with at least one predicted SMISP with a maximum score that is at least 0.85. PPIs with known inhibitors (from Figures 3 and Table 4) are visibly enriched in the higher quality, higher quantity region of Figure 4. A quantity/quality threshold of (≥6%, ≥0.84), shown shaded in Figure 4, selects 11 of the 13 complexes with known inhibitors. The remaining two complexes, HPV E1/E2 and TNF-α from Figures 3 and h, were the two cases where the predicted SMISP failed to identify the binding site of the inhibitor due to conformational changes of the protein receptor. Since this is the strictest threshold that captures all the PPIs where the predicted SMISPs correctly identify the binding site of the inhibitor, we predict that complexes within this threshold are likely susceptible to small-molecule inhibition through the chemical mimicry of a small set of interface residues. Nearly half, 48%, of all the PPIs in our dataset fall within this threshold. Of course, this estimate does not include the potential to design allosteric inhibitors. Additionally, our analysis is based on existing structures and therefore cannot predict the susceptibility of PPIs to new mechanisms of inhibition.
The quality and quantity thresholds computed for each PPI chain define a partial order, where one chain is better than another if it has both a higher score and a greater quantity. We use this partial order to rank every PPI chain in our dataset and provide this ranking in the Supplementary Data.

4 DISCUSSION

PPIs are an emerging class of biological targets that have been poorly addressed by high throughput screening (HTS). The steadily increasing amount of PPI structural information makes structure-based rational design one attractive alternative. The interactions of the complex itself provide a natural starting point as long as the most favorable interactions can be effectively identified. Previous work has largely focused on identifying individual hotspots, i.e. single residues that contribute a significant portion of the free energy of the complex. We have built upon this previous work to address the problem of finding an entire set of residues that, as a unit, identify a binding site that is susceptible to small-molecule inhibition. Our automated structural bioinformatics method learns both scoring (SVM-based) and filtering (rule-based) classifiers for identifying these sites.

A novel aspect of our method is that we use the structure of PPI complexes and known binders at the PPI interface to create our training set. SMISPs are identified purely from the steric overlap of small-molecule binders and the PPI structure. This results in a less quantitative assessment of training examples compared to using alanine scanning results. However, our structure-based approach has several advantages. The existence of a high-affinity ligand at the protein interface is more relevant to our goal of identifying hot spots, i.e. single residues that contribute a significant portion of protein stability. Additionally, the ΔG does not include any backbone contributions, which may be important for small molecule design. Our structure-based method naturally identifies groups of co-located residues, SMISPs, as opposed to single, potentially isolated hot spot residues. Finally, our method benefits from the rapidly increasing number of available protein structures, exemplified by the PDB.

We use a standard SVM classifier for ranking SMISPs, but also find that simple thresholding rules provide an easily interpretable basis for prioritizing our predictions. However, our structure-based method naturally identifies groups of co-located residues, SMISPs, as opposed to single, potentially isolated hot spot residues. Finally, our method benefits from the rapidly increasing number of available protein structures, exemplified by the PDB.

Our approach correctly predicts the binding sites and, to a lesser extent, the interactions of several known PPI inhibitors. The SMISPs identified using our approach and available from http://pocketquery.csb.pitt.edu provide us with a new and powerful tool to address the problem of finding an entire set of residues that, as a unit, identify a binding site that is susceptible to small-molecule inhibition. Our automated structural bioinformatics method learns both scoring (SVM-based) and filtering (rule-based) classifiers for identifying these sites.

A novel aspect of our method is that we use the structure of PPI complexes and known binders at the PPI interface to create our training set. SMISPs are identified purely from the steric overlap of small-molecule binders and the PPI structure. This results in a less quantitative assessment of training examples compared to using alanine scanning results. However, our structure-based approach has several advantages. The existence of a high-affinity ligand at the protein interface is more relevant to our goal of identifying hot spots, i.e. single residues that contribute a significant portion of protein stability. Additionally, the ΔG does not include any backbone contributions, which may be important for small molecule design. Our structure-based method naturally identifies groups of co-located residues, SMISPs, as opposed to single, potentially isolated hot spot residues. Finally, our method benefits from the rapidly increasing number of available protein structures, exemplified by the PDB.

We use a standard SVM classifier for ranking SMISPs, but also find that simple thresholding rules provide an easily interpretable alternative classifier with accuracy similar to more sophisticated techniques. Unlike previous work that relied on a subjective evaluation of histograms to set rule thresholds, our exhaustive rule classifier uses information theoretic techniques to automatically identify the optimal classification rules from the training set.

Our approach correctly predicts the binding sites and, to a lesser extent, the interactions of several known PPI inhibitors. Perhaps more interestingly, the application of our classifiers to a non-redundant subset of the PDB containing more than 12,000 complexes reveals that nearly half the complexes contain at least one chain that has a similar distribution of predicted SMISPs as complexes with known inhibitors. This suggests that PPIs may be substantially more susceptible to small-molecule inhibition than indicated by the current distribution of pharmaceutical targets [Paolini et al. 2004]. The historical bias against PPI targets present in existing compound libraries makes structure-based rational design a logical paradigm for developing PPI inhibitors. The SMISPs identified using our approach and available from http://pocketquery.csb.pitt.edu provide an immediate computational hypothesis to initiate such efforts which have the potential to ultimately culminate in the design of novel therapeutics.

Funding: National Institute of Health [1R21GM087617, R01GM097082].

Conflict of Interest: none declared.

REFERENCES


Chan, D., Cherkowski, C. and Kim, P. (1998). Evidence that a prominent cavity in the coiled coil of HIV type 1 gp41 is an attractive drug target. PNAS, 95(26), 15613.


