Structural bioinformatics

Small-molecule inhibitor starting points learned from protein–protein interaction inhibitor structure

David Ryan Koes* and Carlos J. Camacho
Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260, USA

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 predict 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 Wells and McLeod (2007) have provided a number of reasons why PPIs are important. 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 (Rodriguez et al. 2009). 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 energy of a PPI is contributed by just a few “hot spot” residues (Jackson and Well 1998; Moreira et al. 2004; Rajmani et al. 2009). 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 (Meireles et al. 2004). Indeed, the chemical mimicry of small clusters of key residues, typically deeply buried “anchoring” residues (Rajmani et al. 2009), has produced small-molecule inhibitors for a number of PPI targets (Christ et al. 2010; Liu et al. 2005; 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; Kooten et al. 2004) attempt to directly compute the energetic contributions of residues. Solvent accessible surface area (SASA) (Meireles et al. 2004; Rajmani et al. 2004) and sequence conservation (Bromberg and Rost, 2004; Keskin et al. 2005; Luchte 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; Barnett et al. 2004) to provide starting points for smaller molecules. A consensus scoring mechanism is derived using machine learning techniques such as decision trees (Barnett et al. 2004; Gcoe et al. 2004; Lise et al. 2006; Tuncbag et al. 2004) or support vector machines (Cho et al. 2004; Lise et al. 2006; Tuncbag et al. 2004), and achieve accuracies of 60–80%. Often a consensus scoring scheme is effective on small molecules, although no single consensus scheme is effective on all PPIs (Gcoe et al. 2004; Tuncbag et al. 2004).

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 energetic hot spots 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. 2005; Pott et al. 2011), 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 Sali 2010). This important insight suggests that it may be possible to automatically identify these 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 [Henrich et al., 2010], either through shape descriptors [Weisel et al., 2007] or chemical probes [Brenke et al., 2006; 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 computational alanine scanning [Kontemme et al., 2009] was used to identify co-located hot spots on a helix that provided a significant portion of the free energy of the helical interaction. This method is only partially successful at identifying SMISPs corresponding to known PPI inhibitors. In addition to being limited to helical interfaces, we find the energy criteria used to be less informative in characterizing SMISPs than SASA and alternative energy calculations. Our approach, which uses a consensus score, is more successful at recovering SMISPs of known inhibitors.

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 exhaustive rule classifier for filtering SMISPs using 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., 2004]. 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, that overlap the atoms of the ligand in the ligand-bound structure in this aligned receptor chain and that are bound to a standalone ligand (i.e., not a modified ligand). 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

![Fig. 1. The complete workflow of our method. A benchmark of small molecule inhibitor starting points (SMISPs) is extracted from the structural information of protein-protein and protein-ligand complexes in the PDB. A training set is generated from this benchmark and used to train both a rule-based and SVM classifier which then back-annotate an entire non-redundant subset of PPI complexes from the PDB.](https://academic.oup.com/bioinformatics/article-abstract/28/6/784/309806)
We decompose PPI structures into clusters of multiple interface residues. The SMISPs sterically delineate part of the binding site of the ligand in the PPI. The LEU-63 residue of the SMISP fills a deep hydrophobic pocket that is likely to affect the function of the PPI. Since they bind to the fully formed complex, but we include them in the benchmark set since a significant perturbation of the interface structure will likely affect the function of the PPI. We further refine our collection of SMISPs derived from structure by incorporating binding affinity data from the PDBbind benchmark set since a significant perturbation of the interface structure should be more than 1Å. This distance is roughly equal to two turns of an α-helix and encompasses a volume larger than the 500Å^3 usually observed in protein-ligand interactions. 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, ΔG^E and ΔG^P, the absolute (ΔSASA) and relative (ΔSASA/ΔV) 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 3.1b1. CHARMM is used to add missing atoms, including hydrogens, and to quickly minimize the resulting structure to optimize hydrogen bonding.

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, ΔG^E and ΔG^P, the absolute (ΔSASA) and relative (ΔSASA/ΔV) 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 3.1b1.

We construct a training set from the benchmark set that is evenly distributed among the PPI structures: for each PPI we generate 100 positive examples and 100 negative examples. 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, ΔG^E and ΔG^P, the absolute (ΔSASA) and relative (ΔSASA/ΔV) 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 3.1b1.

We further refine our collection of SMISPs derived from structure by incorporating binding affinity data from the PDBbind benchmark databases. We restrict our benchmark SMISPs to only those where the ligand bound structure (or a homologue with at least 95% sequence identity) has an affinity of 100 μM or better to exclude spurious or low-affinity ligand-protein interactions. The resulting collection of SMISPs includes 135 distinct SMISPs derived from 267 ligands targeting 51 PPI complexes. There are nearly twice as many ligands as SMISPs since in most cases ligands targeting the same protein define identical or nearly identical SMISPs. Two SMISPs from this set, both derived from known PPI inhibitors, are shown in Figure 2. We construct a training set from the benchmark set that is evenly distributed among the PPI structures: for each PPI we generate 100 positive examples and 100 negative examples. 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, ΔG^E and ΔG^P, the absolute (ΔSASA) and relative (ΔSASA/ΔV) 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 3.1b1.
SMISPs learned from PPI inhibitor structure

Table 1. Validation set of PPIs with known inhibitors

<table>
<thead>
<tr>
<th>Description</th>
<th>PPI PDB Ch Lig. PDB</th>
<th>#SMISPs</th>
<th>#Clust</th>
</tr>
</thead>
<tbody>
<tr>
<td>p53/MDM2</td>
<td>1YCR 3LBK</td>
<td>100</td>
<td>311</td>
</tr>
<tr>
<td>p53/MDM4</td>
<td>3DAB 3LBJ</td>
<td>45</td>
<td>79</td>
</tr>
<tr>
<td>Bak/Bcl-2</td>
<td>1EUE 31U3</td>
<td>56</td>
<td>512</td>
</tr>
<tr>
<td>Bcl-xL/Bax</td>
<td>2J3G 12F4</td>
<td>36</td>
<td>123</td>
</tr>
<tr>
<td>ZipA/FtsZ</td>
<td>1TNR 1O01</td>
<td>36</td>
<td>123</td>
</tr>
<tr>
<td>XIAP/Smac</td>
<td>1G73 2JK7</td>
<td>117</td>
<td>1793</td>
</tr>
<tr>
<td>Cat1</td>
<td>1EUE 31U3</td>
<td>56</td>
<td>512</td>
</tr>
<tr>
<td>TGF/α</td>
<td>2J3G 12F4</td>
<td>36</td>
<td>123</td>
</tr>
</tbody>
</table>

PPIs were identified from the literature. Samudrala et al. (2002) and only inhibitors 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 unbalanced training data, we re-sample the data to create a balanced training set. 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 protein of the PPI is a short peptide where all, or nearly all, of the interface residues make up a SMISP. Since the entire peptide defines a SMISP, these 12 PPIs are trivial to identify as SMISPs and are excluded from the training set. There are a total of 7800 training examples from 39 PPI complexes. The composition of the training set is further described in the Supplementary Data.

Leave-one-complex-out validation is performed by removing all training examples of a PPI complex from the training set and then using these examples as a test set resulting in 39 unique train-test datasets. The cross-validation accuracy (correct predictions divided by total predictions) is evaluated by removing all training examples of a PPI complex from the training set and then using these examples as a test set resulting in 39 unique train-test datasets. The cross-validation accuracy is calculated as the average number of correctly classified examples as a function of the threshold. The average cross-validation accuracy of 74% is obtained with a score threshold of 0.55. Higher score thresholds increase the specificity, but at the cost of substantially reduced recall. The advantage of the cross-validation receiver operating characteristic (ROC) curves is 0.82.

Table 2. Most informative features ranked by the information gain

<table>
<thead>
<tr>
<th>Feature</th>
<th>IG</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SASA%</td>
<td>0.5</td>
<td>Average SASA%</td>
</tr>
<tr>
<td>SASA%</td>
<td>0.4</td>
<td>Average SASA%</td>
</tr>
<tr>
<td>SASA%</td>
<td>0.3</td>
<td>Average SASA%</td>
</tr>
<tr>
<td>SASA%</td>
<td>0.2</td>
<td>Average SASA%</td>
</tr>
<tr>
<td>SASA%</td>
<td>0.1</td>
<td>Average SASA%</td>
</tr>
</tbody>
</table>

The advantage of our exhaustive rule classifier is that the thresholds for different attributes are determined simultaneously and optimally, in contrast to methods, such as decision trees, that greedily select and set attributes and thresholds. The C/C++ source code of our implementation is available in the Supplementary Data.

3 RESULTS

We analyze the information theoretic properties of our cluster features to highlight the most useful features for classification. We show that we can recover known PPI inhibitor sites and assess the general accessibility of PPIs to small-molecule inhibitors.

3.1 Information gain analysis

A selection of cluster features are shown ranked by the information gain of the corresponding single-attribute rule in Table 2. The single most informative feature is the average SASA%. The average SASA% and ΔGPC are also informative and have good classification accuracies. These average values, which represent the entire cluster, are more informative than extrema values of the same criteria.
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 ± 3.6</td>
</tr>
<tr>
<td>0.128</td>
<td>Ave ΔSASA%</td>
<td>39.6 ± 3.7</td>
</tr>
<tr>
<td>0.103</td>
<td>Ave ΔΔΔG&lt;</td>
<td>&lt; -2.3</td>
</tr>
<tr>
<td>0.081</td>
<td>Ave ΔΔΔG&lt;</td>
<td>&gt; -2.3</td>
</tr>
<tr>
<td>0.070</td>
<td>Max ΔΔΔG&lt;</td>
<td>&gt; 0.427</td>
</tr>
<tr>
<td>0.042</td>
<td>Min ΔΔΔG&lt;</td>
<td>&lt; -1.3</td>
</tr>
<tr>
<td>0.038</td>
<td>Total Rate&lt;</td>
<td>&gt; 1.6</td>
</tr>
<tr>
<td>0.032</td>
<td>Min Cons</td>
<td>&lt; 0.086</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 computed using the entire training set. Although interface residues have been found to have a distinguishable conservation profile (bioinformatics.btr717.tex), 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 (Rost and Sander, 2004). 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 ΔASASA 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 ΔG< and total ΔASASA 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

---

**Fig. 3.** SMISPs predictions for some of the PPIs from Table 1. The PPI is represented by a receptor protein (surface) and a ligand protein (transparent magenta). A small-molecule inhibitor (green) is posed by aligning the corresponding receptors. The single largest SMISP ranked in the top three is shown as magenta sticks. PDB access codes are provided in Table 1. In Figures (a-f) the predicted SMISPs overlap the inhibitor and at least partially delineate the binding pocket(s). In Figures (g-h) the SMISPs only marginally overlap the inhibitor and identify a nearby, but distinct, binding pocket. (a) p53/MDM2. (b) XIAP-BIR3/Caspase-9. (c) HIV gp41. (d) Bcl-xL/Beclin 1. (e) HIV-1 Integrase/p75. (f) ZipA/FtsZ. (g) HPV E1/E2. (h) TNF-α.

---

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 ± 3.6</td>
</tr>
<tr>
<td>0.175</td>
<td>Ave ΔASASA%</td>
<td>39.6 ± 3.7</td>
</tr>
<tr>
<td>0.167</td>
<td>Max ΔΔΔG&lt;</td>
<td>&gt; 0.427</td>
</tr>
<tr>
<td>0.156</td>
<td>Ave ΔASASA</td>
<td>46.1 ± 4.1</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.
of the classifiers, we find that 95% of the protein chains in our both classifiers. Unsurprisingly, given the cross-validation accuracy. A cluster is only classified as a SMISP if it is predicted by decomposed into more than 95 million clusters. Since true SMISP analyze our entire set of PPIs. The interface residues of these PPIs are 3.3 PDB-Wide analysis

Three SMISPs is shown in Figure 3 in order to illustrate the diversity in Table 1 and ranges from 30 to 762. The largest SMISP in the top training set. The number of identified SMISPs for each PPI is shown from Table 3 and the 0.55 SVM score threshold. The SMISPs are filtered using the most informative two-attribute rule predicted SMISP in Figure 3 and the three remaining complexes, Eight of the eleven PPIs of our validation set are shown with a hot spot residues (Chan et al., 1998; Lin et al., 1998). 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 and b 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, 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 4 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 4 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. 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 inhibition. Our automated structural bioinformatics method learns as a unit, identify a binding site that is susceptible to small-molecule inhibition. 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 methods. Our method naturally identifies groups of co-located residues, SMISPs, 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.

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. 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](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.


SMISPs learned from PPI inhibitor structure


