Prediction of sub-cavity binding preferences using an adaptive physicochemical structure representation

Izhar Wallach\textsuperscript{1,2,*} and Ryan H. Lilien\textsuperscript{1,2,3,*}

\textsuperscript{1}Department of Computer Science, \textsuperscript{2}Donnelly Centre for Cellular and Biomolecular Research and \textsuperscript{3}Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada

Abstract

Motivation: The ability to predict binding profiles for an arbitrary protein can significantly improve the areas of drug discovery, lead optimization and protein function prediction. At present, there are no successful algorithms capable of predicting binding profiles for novel proteins. Existing methods typically rely on manually curated templates or entire active site comparison. Consequently, they perform best when analyzing proteins sharing significant structural similarity with known proteins (i.e. proteins resulting from divergent evolution). These methods fail short when used to characterize the binding profile of a novel active site or one for which a template is not available. In contrast to previous approaches, our method characterizes the binding preferences of sub-cavities within the active site by exploiting a large set of known protein–ligand complexes. The uniqueness of our approach lies not only in the consideration of sub-cavities, but also in the more complete structural representation of these sub-cavities, their parametrization and the method by which they are compared. By only requiring local structural similarity, we are able to leverage previously unused structural information and perform binding inference for proteins that do not share significant structural similarity with known systems.

Results: Our algorithm demonstrates the ability to accurately cluster similar sub-cavities and to predict binding patterns across a diverse set of protein–ligand complexes. When applied to two high-profile drug targets, our algorithm successfully generates a binding profile that is consistent with known inhibitors. The results suggest that our algorithm should be useful in structure-based drug discovery and lead optimization.

Contact: izhaw@cs.toronto.edu; lilien@cs.toronto.edu

1 INTRODUCTION

The ability to identify and exploit patterns of protein–small-molecule interaction is a critical component of protein function prediction, pharmacophore inference, molecular docking and protein design (Halperin et al., 2002; Langer and Hoffmann, 2006; Powers et al., 2006; Sousa et al., 2006). In most cases, the protein–ligand interface is characterized by a number of geometric and/or chemical features (Dret et al., 2006). This characterization is facilitated by mining high-resolution experimental structures where, ideally, a single protein would be observed interacting with several different bound ligands. Unfortunately, for most proteins, this type of multiple-binding information is not available (Berman et al., 2000). To avoid this problem, we chose to focus on sub-cavities within an active site with the assumption that structurally similar sub-cavities are likely to exhibit similar binding profiles. It is important to emphasize the definition of sub-cavity utilized in this work. We define a sub-cavity to be a small region of the traditionally described active site capable of interacting with a single chemical group (e.g. phenyl, hydroxyl and carboxyl). That is, an active site is generally composed of 5–20 sub-cavities. By considering protein–ligand interactions at the sub-cavity level, we can utilize binding information from structurally and functionally distinct proteins. A pair of proteins whose active sites differ significantly when compared in their entirety may still share similarity at the sub-cavity level. In this work, we decompose a target active site into a set of sub-cavities, identify structurally similar sub-cavities within other proteins and then use this information to construct a binding profile. This approach enables inference when no global receptor similarity is available.

There are several existing approaches to analyzing an active site’s protein–ligand binding preference. In most cases, these methods aim to predict protein function which differs from our aim of identifying the local binding patterns of sub-cavities. A result of these different goals is that a direct comparison between our work and the described methods using a common dataset is not feasible. State-of-the-art methods can be classified into three groups:

Template-based methods: these methods (Laskowski et al., 2005; Stark and Russell, 2003; Wallace et al., 1997) accept a template structure as input (e.g. catalytic triad in serine proteases) and query a given structural database for matching patterns. Their reliance on the provided input template makes them particularly useful for studying or predicting relations to an already known pattern. The strength of the template-based methods is their speed. The lightweight template representation (usually, a set of amino acid residues) allows rapid queries against large databases; however, the templates are often overly simplistic and are incapable of capturing the rich physicochemical variations among binding pockets. Further, templates are often manually generated and are therefore restricted in their complexity. These limitations make general binding site analysis with template-based methods difficult.

Binding site similarity methods: these methods compare entire binding sites and retrieve sets of proteins which share globally similar binding patterns. They are commonly used for protein function prediction where similar binding patterns imply similar catalytic functionality. Several approaches utilize active site geometry (i.e. amino acid backbone, solvent accessible surface or active site volume) (Kuhn et al., 2006; Morris et al., 2005), while other methods exploit both geometry and the chemical function of the amino acid residues flanking the binding site (Najmanovich et al., 2008; Schmitt et al., 2002; Shatsky et al., 2005). A slightly different approach combines sequence alignment with
subsequent spatial pattern matching of the defined active site surface (Binkowski et al., 2003). These methods infer patterns using a variety of techniques to identify similarity such as sub-graph isomorphism, geometric hashing, multiple structural alignment and clique detection (Kinoshita and Nakamura, 2003; Leibowitz et al., 2001; Perona and Ayache, 1998; Shulman-Peleg et al., 2008). Unlike the template-based methods, the binding site similarity methods often utilize an elaborate model that includes both geometrical constraints and chemical profile. While these methods are suitable for comparing whole binding sites, they are less effective when considering functionally similar proteins which only share local ‘hot-spots’ within their binding pockets. They are also less useful when analyzing proteins with novel structure and function. These novel proteins are unlikely to match any patterns derived from known active sites. Local binding site similarity methods: recently methods that search for local similarities within binding sites have emerged. These methods allow inference when global active site similarity is poor. They characterize local interaction patterns within the binding site cavity. The latter is particularly useful for structure-based drug discovery, where the interaction between a ligand’s chemical group and the flanking amino acids can be separately studied outside the context of the binding site. In (Kupas et al., 2007), the Klebe group defines a set of functional pseudo-centers for each amino acid and pairs each pseudo-center with an associated surface-patch (Schmitt et al., 2002). Each surface-patch reflects the chemical function and location of its corresponding residue. A wavelet representation of the surface-patches allows fast structural comparison with adjustable resolution. A slightly different approach (Ramensky et al., 2007) associates each sub-cavity with its occupying ligand fragment. Using the ligand fragments as anchor points, the sub-cavities are aligned and a similarity is computed using the spatial and chemical overlap of the neighboring residues. Local similarity methods suffer from two major shortcomings. First, existing methods define sub-cavities using an arbitrary fixed distance threshold from either the associated residue or the ligand fragment. Second, the inclusion of apo proteins potentially reduces the quality of identified templates. Selecting a threshold that is too large results in the inclusion of irrelevant and potentially confusing residues; conversely, a threshold that is too small results in a partial model that ignores relevant residues; proteins frequently undergo moderate structural changes upon ligand binding. Methods which ignore the holo structure are likely to infer a biased view of the binding profile.

2 APPROACH
We have developed a sub-cavity-based approach to characterizing protein-small-molecule binding patterns. Our algorithm deconstructs the active site into a set of potentially overlapping sub-cavities and then infers the binding profile of each sub-cavity. The deconstruction allows us to exploit the sub-cavity similarity that often exists between structurally diverse proteins. The binding profile of the entire active site can then be constructed by joining the information gleaned from each sub-cavity. The approach differs from previous work in several important ways: first, we analyze only protein-small-molecule complexes. The current abundance and diversity of holo structures allows us to avoid inclusion of apo structures during learning. This design decision removes binding site localization from the inference problem and ensures that analyzed sub-cavities are indeed involved in binding. We discuss the possibility of relaxing this restriction in Section 4.4. Second, we divide each binding site into sub-cavities according to the chemical groups of the bound ligand. This separation enables us to identify sub-cavities that are likely to form interactions, and more importantly, to label each sub-cavity with the chemical group to which it is bound (i.e. its functionality). Third, we model sub-cavities by combining the shape of the binding site (i.e. its solid 3D volume) with the chemical profile of its flanking residues to form a single physicochemical representation. This allows us to benefit from the accuracy of modeling the shape of the active site while still accounting for the chemical properties of the surrounding residues. Furthermore, this representation allows us not only to avoid matching the flanking residues directly but also to account for their cumulative effects at any location within the sub-cavity. Finally, we allow the algorithm to iteratively cluster sub-cavities with the same function and to reshape sub-cavities. The iterative sub-cavity reshaping procedure is unique to our approach and provides an advantage over simply including all residues within a distance cutoff. Reshaping increases the within-class similarity (i.e. sub-cavities with the same function become more similar) while reducing the between-class similarity. This procedure not only improves the classification results (Section 4) but also produces optimized sub-cavity structures.

In the context of this article, we define the following terms: (i) a chemical group is a group of atoms that characterize a chemical moiety. Like a set of building blocks, a limited set of chemical groups can specify the structure of virtually all small molecules. We utilize a set of 47 chemical groups (e.g. phenyl, hydroxyl, carboxyl) inspired by (Chen et al., 1999). (ii) A function type is the mapping of a chemical group to its corresponding chemical interaction type. We utilize six functional types: hydrophobic, aromatic, acidic, basic, hydrogen-bond donor (HBD) and hydrogen-bond acceptor (HBA) (McGregor, 2007). Because some chemical groups correspond to more than one function type (i.e. HBD and HBA) the mapping is one-to-many. In practice, each of our 47 chemical groups can be described by one of only 9 sets of function types.

In this framework, a small molecule can be considered a spatial arrangement of active chemical groups connected by inert bridging fragments. We propose that within a sub-cavity, the function types of the protein residues specify a set of preferred ligand fragments (i.e. chemical groups) with which to bind. Recent analysis of the binding site variations (Kahraman et al., 2007) as well as local binding site similarity methods (Kupas et al., 2007; Ramensky et al., 2007) provide evidence that such patterns do exist. Using the structures of solved protein complexes we can learn which sub-cavities (parametrized by shape and function type) preferentially associate with which ligand fragments (parametrized by chemical group). Using the occupying ligand fragments to define sub-cavities not only isolates properly configured sub-cavities but also pairs each sub-cavity with the bound chemical group. This information can then be used to infer a binding profile for a query potentially apo protein.

3 METHODS
The training of our algorithm can be divided into a sub-cavity generation stage (steps 1–4) and a sub-cavity comparison stage (steps 5–6). The process is summarized below and in Figure 1.
After generating and clustering the sub-cavities (steps 1–6), we can perform functional inference on a novel sub-cavity with unknown function by assigning the sub-cavity the function type of its most similar cluster. Functional inference on a novel sub-cavity with unknown function by assigning the sub-cavity the function type of its most similar cluster.

**3.1 Dataset generation**

We developed the PSMDB database (Wallach and Lilien, 2009) explicitly to enable non-biased inference of binding patterns. The PSMDB provides sets of protein–small-molecule complexes that are ideally suited for the analysis in the present work. It includes only high-quality crystal structures that contain at least one non-covalently bound ligand. Furthermore, the PSMDB handles structural redundancy at both the protein and ligand levels and thus contains highly similar ligands interacting with different proteins and different ligands interacting with highly similar proteins. In contrast to other small-molecule databases, this feature provides multiple instances of sub-cavity–ligand interactions while reducing the bias toward any single pattern.

**3.2 Ligand chemical analysis**

The chemical groups contained within each ligand are used as anchoring points to define the sub-cavities. We identify chemical groups by first defining a template library of chemical groups (modeled after (Chen et al., 1999) and encoded using SMARTS patterns (James et al., 2000)) and then matching these templates against each ligand structure. Similar to the pseudo-center approach described in (Schmitt et al., 2002), we assign a location for each identified chemical group and label the group with up to six features (e.g. HB, HBA, ...) corresponding its functional type. This reduced ligand representation captures the functional type and arrangement of each chemical group while removing structural scaffolding that is not likely to form a significant interaction.

**3.3 Shape identification and characterization of binding sites**

We identify the binding site cavities and generate structural models that include the shape of the cavity and the function type of the amino acid residues that flank it. Since all input structures contain a bound ligand, we are able to easily identify the location of the binding site. We place the protein structure over a grid of 0.75 Å (half the length of a carbon–carbon bond) consisting of grid cells with the center of each cell defined as a grid point. We then assign one of three labels (inner volume, binding surface and inner surface) to each grid cell. Starting with one of the ligand’s atoms, we select the atom’s corresponding grid cell and label it as inner volume. We then iterate over the neighboring grid cells. Inspired by the POCKET and SURFNET algorithms (Laskowski, 1995; Levitt and Banaszak, 1992), we define the solvent accessible surface by first checking for van der Waals clashes of protein atoms with a pseudo water molecule located at each grid point. If there is no clash, we mark the grid cell as an inner volume and recursively proceed to its neighbors. When a grid point clashes with a protein atom, we mark the cell as a binding surface cell, tentatively mark all its neighbors that have not yet been explored as inner surface and then backtrack. The inner surface points are either be relabeled to binding surface or inner volume as they or their neighbors are visited. In order to remain within the cavity, we stop and backtrack when a grid point reaches a distance of >5 Å from all ligand atoms. We use a burial degree measurement to locate the mouth of the binding pocket (Schmitt et al., 2002). We then define the solvent accessible surface by first checking for van der Waals clashes of protein atoms with a pseudo water molecule located at each grid point. If there is no clash, we mark the grid cell as an inner volume and recursively proceed to its neighbors. When a grid point clashes with a protein atom, we mark the cell as a binding surface cell, tentatively mark all its neighbors that have not yet been explored as inner surface and then backtrack. The inner surface points are either be relabeled to binding surface or inner volume as they or their neighbors are visited. In order to remain within the cavity, we stop and backtrack when a grid point reaches a distance of >5 Å from all ligand atoms. We use a burial degree measurement to locate the mouth of the binding pocket (Schmitt et al., 2002). We then define the solvent accessible surface by first checking for van der Waals clashes of protein atoms with a pseudo water molecule located at each grid point. If there is no clash, we mark the grid cell as an inner volume and recursively proceed to its neighbors. When a grid point clashes with a protein atom, we mark the cell as a binding surface cell, tentatively mark all its neighbors that have not yet been explored as inner surface and then backtrack. The inner surface points are either be relabeled to binding surface or inner volume as they or their neighbors are visited. In order to remain within the cavity, we stop and backtrack when a grid point reaches a distance of >5 Å from all ligand atoms. We use a burial degree measurement to locate the mouth of the binding pocket (Schmitt et al., 2002). We then define the solvent accessible surface by first checking for van der Waals clashes of protein atoms with a pseudo water molecule located at each grid point.

**3.4 Binding site division into sub-cavities**

Having identified the extent of the binding site as well as the chemical groups of the ligand, we then define sub-cavities that are likely to participate in binding. We define a sub-cavity to be the region of the binding site that surrounds a chemical group of the bound ligand (Fig. 5) and we label the sub-cavity with the chemical group's function type. We define an initial sub-cavity to be the set of binding site grid points that are closer than 3 Å to a ligand chemical group. We eliminate sub-cavities which have <20% of their...
more optimal conserved structure. At this point in the process, we find the grid points that occupy its shape (i.e. the shape defined by an eigenvalue decomposition of the covariance matrix). We then translate the sub-cavity center of mass to the origin and rotate the sub-cavity in Section 3.5. The sub-cavity is rigidly transformed such that its principle axes are aligned with the x and y axes and its center of mass is at the origin. The cumulative effect of each flanking functional group (illustrated by the shaded regions) is computed for every grid point in the new coordinate frame that overlaps the sub-cavity.

The comparison of two sub-cavities requires a similarity function that accounts for both their shape and chemical features. We regard one sub-cavity as a reference and the other as a query. We place a grid over the reference and find the grid points that occupy its shape (i.e. the shape defined in Section 3.4). The three principal axes of the occupied grid points are determined by an eigenvalue decomposition of the covariance matrix. We then translate the sub-cavity center of mass to the origin and rotate the sub-cavity to align the three principal axes with the x, y and z axes, respectively (Fig. 2B). We next calculate a feature vector, defined as a chem-vector, for each grid point occupied by the reference structure. The chem-vector reflects the local cumulative effect of six functional types of the flanking residues (Fig. 2B). We pose the query sub-cavity over the reference's coordinate frame and the chem-vector for each grid point occupied by the reference structure. The chem-vector reflects the local cumulative effect of six functional types of the flanking residues. The chem-vector for each grid point occupied by the reference structure.

### 3.5 Sub-cavity similarity

We now have a set of sub-cavities which are observed to participate in binding, extracted from a large diverse set of protein–ligand complexes and have specified a similarity function which accounts for sub-cavity shapes and chemical profiles. Using the similarity function, we can determine the degree of overlap between the shapes as well as the similarity in chemical features. We regard one sub-cavity as a reference and the other as a query. We place a grid over the reference and find the grid points that occupy its shape (i.e. the shape defined by an eigenvalue decomposition of the covariance matrix). We then translate the sub-cavity center of mass to the origin and rotate the sub-cavity in Section 3.5. The sub-cavity is rigidly transformed such that its principle axes are aligned with the x and y axes and its center of mass is at the origin. The cumulative effect of each flanking functional group (illustrated by the shaded regions) is computed for every grid point in the new coordinate frame that overlaps the sub-cavity.

The comparison of two sub-cavities requires a similarity function that accounts for both their shape and chemical features. We regard one sub-cavity as a reference and the other as a query. We place a grid over the reference and find the grid points that occupy its shape (i.e. the shape defined in Section 3.4). The three principal axes of the occupied grid points are determined by an eigenvalue decomposition of the covariance matrix. We then translate the sub-cavity center of mass to the origin and rotate the sub-cavity to align the three principal axes with the x, y and z axes, respectively (Fig. 2B). We next calculate a feature vector, defined as a chem-vector, for each grid point occupied by the reference structure. The chem-vector reflects the local cumulative effect of six functional types of the flanking residues (Fig. 2B). We pose the query sub-cavity over the reference's coordinate frame and the chem-vector for each grid point occupied by the reference structure.

### 3.6 Sub-cavity clustering and reshaping

We now have a set of sub-cavities which are observed to participate in binding, extracted from a large diverse set of protein–ligand complexes and have specified a similarity function which accounts for sub-cavity shapes and chemical profiles. Using the similarity function, we can determine the degree of overlap between the shapes as well as the similarity in chemical features. We regard one sub-cavity as a reference and the other as a query. We place a grid over the reference and find the grid points that occupy its shape (i.e. the shape defined in Section 3.4). The three principal axes of the occupied grid points are determined by an eigenvalue decomposition of the covariance matrix. We then translate the sub-cavity center of mass to the origin and rotate the sub-cavity to align the three principal axes with the x, y and z axes, respectively (Fig. 2B). We next calculate a feature vector, defined as a chem-vector, for each grid point occupied by the reference structure. The chem-vector reflects the local cumulative effect of six functional types of the flanking residues (Fig. 2B). We pose the query sub-cavity over the reference's coordinate frame and the chem-vector for each grid point occupied by the reference structure.

The comparison of two sub-cavities requires a similarity function that accounts for both their shape and chemical features. We regard one sub-cavity as a reference and the other as a query. We place a grid over the reference and find the grid points that occupy its shape (i.e. the shape defined in Section 3.4). The three principal axes of the occupied grid points are determined by an eigenvalue decomposition of the covariance matrix. We then translate the sub-cavity center of mass to the origin and rotate the sub-cavity to align the three principal axes with the x, y and z axes, respectively (Fig. 2B). We next calculate a feature vector, defined as a chem-vector, for each grid point occupied by the reference structure. The chem-vector reflects the local cumulative effect of six functional types of the flanking residues (Fig. 2B). We pose the query sub-cavity over the reference's coordinate frame and the chem-vector for each grid point occupied by the reference structure.

The comparison of two sub-cavities requires a similarity function that accounts for both their shape and chemical features. We regard one sub-cavity as a reference and the other as a query. We place a grid over the reference and find the grid points that occupy its shape (i.e. the shape defined in Section 3.4). The three principal axes of the occupied grid points are determined by an eigenvalue decomposition of the covariance matrix. We then translate the sub-cavity center of mass to the origin and rotate the sub-cavity to align the three principal axes with the x, y and z axes, respectively (Fig. 2B). We next calculate a feature vector, defined as a chem-vector, for each grid point occupied by the reference structure. The chem-vector reflects the local cumulative effect of six functional types of the flanking residues (Fig. 2B). We pose the query sub-cavity over the reference's coordinate frame and the chem-vector for each grid point occupied by the reference structure.
4 RESULTS AND DISCUSSION
The current work describes experiments using both simulated and real experimental data. In Section 4.1, we describe an experiment on simulated sub-cavities. The experiments described in Section 4.2 were performed with real experimental data and explore sub-cavity label recovery for a set of functionally similar proteins. These experiments demonstrate the performance of our approach on real sub-cavities. In Section 4.3, we describe the results of an evaluation of a non-redundant set of over 300 protein-small-molecule complexes taken from the PSMD database. We analyze the predictive performance of the resulting clusters and highlight functional inference on the binding sites of HIV-1 Protease and Thrombin. Finally, in Section 4.4 we discuss design decisions, limitations and possible extensions of the presented work.

4.1 Simulated data
We began by evaluating the clustering and reshaping stages of our algorithm (Section 3.6). We tested the ability of our procedure to recover the structure of a sub-cavity template given a set of noisy instances of a template structure. We first randomly generated 20 sub-cavity templates according to the distribution of sub-cavity sizes, number of pseudo-centers and function types found in the sub-cavities extracted from 1754 non-redundant protein-ligand complexes in the PSMD database. For each template, we generated approximately 20 noisy instances or observed sub-cavities. This was done by randomly adding structure grid points adjacent to the sub-cavity surface and altering its pseudo-center locations. The number of random grid points that were added varied from 10% to 40% of the size of the initial sub-cavity. Each experiment was evaluated using the following parameters: (i) homogeneity—the percent of TP within each cluster; (ii) cluster coherence—a measure similar to homogeneity but which penalizes false negatives (FN) (missed relevant sub-cavities); and (iii) net cluster similarity—a measure of the compactness of each cluster. After refinement, 26 clusters (mean cluster size of 15) and 27 clusters (mean cluster size of 13) were found for the 10% and 40% noise experiments, respectively. The additional clusters were formed by splitting a single cluster with consistent labeling into two smaller clusters (data not shown). This is a consequence of Affinity Propagation’s automatic selection of the number of clusters—a feature that is considered both an advantage and a disadvantage of the approach.

Homogeneity: we define the homogeneity score of a recovered cluster to be its precision, TP/(TP+FP). This measurement is particularly important when structurally distinct sub-cavities with identical labels are split into multiple clusters. Figure 3A shows the changes in homogeneity gained by each iteration of clustering and reshaping. We found the algorithm to yield a high homogeneity score (0.97–0.99) regardless the amount of noise. We saw a small decrease in the homogeneity of the clusters (<0.02) which can be attributed to the appearance of additional small clusters after reshaping. A single error in a small cluster significantly effects the homogeneity score thereby slightly pulling down the reported averages. This is observed for experiments run with higher noise levels.

Net cluster similarity: we evaluated the net cluster similarity of the clustering process after each reshaping iteration. Using the cluster exemplar identified by Affinity Propagation, the net similarity is the sum of all similarities between each sub-cavity and its cluster’s exemplar plus the sum of all exemplar preferences (Frey and Dueck, 2007). This indicated how well the objective function had been maximized. For example, if we had recovered the optimal structures, the net similarity of every cluster would be maximal. Figure 3C shows the change in net similarity with respect to the reshaping iterations. We see that the net similarity increased with the iterative reshaping process. This implies that the sub-cavities within each cluster became more similar to their exemplar and also to each other.

4.2 Clustering different protein classes
In the second experiment, we evaluated our algorithm on a set of experimentally solved protein-ligand complexes. We assembled a dataset of 80 protein-ligand complexes from the PSMD database (482 sub-cavities) spanning 6 different enzyme classes. Each protein in the same enzyme class shares the same 3-number prefix of their Enzyme Classification (EC) number (Bairoch, 2000) and catalyzes a similar chemical reaction. We evaluated the recovered clusters by the homogeneity (Section 4.1) of the cluster’s most common function type. Our algorithm returned 39 clusters with a mean cluster size
of 12 sub-cavities. Figure 4 shows the distribution of cluster scores and demonstrates that most clusters have very high homogeneity.

4.3 PDB sub-cavity analysis and inference

In order to establish the suitability of our algorithm for inferring protein–ligand interactions, we analyzed large non-redundant datasets of sub-cavities derived from complexes in the PSMDB. In these experiments, we trained our system using one set of protein–ligand complexes and then evaluated the performance of the system using a novel, independent test set. Although the complexes in the training and testing set were different, our prediction was that the complexes would share local or sub-cavity similarity which could then be used to infer sub-cavity binding preferences. Our experiments using two well-characterized protein systems, HIV Protease and Thrombin, confirmed this prediction and highlight the utility of this approach.

4.3.1 Sub-cavity analysis

We extracted 6573 sub-cavities from the PSMDB-25% dataset which contains protein–small-molecule complexes having no >25% protein sequence identity. From this initial set, we created five different random subsets each containing 650 sub-cavities. Each of the five sets was run through our algorithm to obtain clusters and corresponding exemplars. For each cluster we calculated a homogeneity score (similar to Section 4.2) that is the ratio between the highest number of occurrences of any function type and the size of the cluster. We labeled each cluster by the most frequent function type.

For each experiment, we randomly sampled a test set of 300 sub-cavities not included in the training set. Using the similarity function of Section 3.5, we identified the cluster with the most similar exemplar to the query. The homogeneity of the most similar cluster was compared to a set threshold. If the homogeneity score was above a threshold, we made a prediction and compared the label of the cluster to the known label of the sub-cavity. The percent of correctly predicted sub-cavities was reported as the precision of the experiment (Table 1). If a prediction is not made, it is not counted against the precision. Table 1 shows that the best predictions were made by clusters with high homogeneity. This effect became more pronounced if reshaping had been performed. Visual inspection revealed that some of the false predictions occurred when two different chemical groups were located in close proximity. In this situation, the function types’ associated cavities were extremely similar and the algorithm had difficulty teasing the two apart.

Table 1. Inference success rate by homogeneity score cutoffs

| Reshaping iteration | <0.5   | 0.5–0.75 | 0.75–1.00 |
|---------------------|--------|----------|-----------|
| 0                   | 0.61   | 0.73     | 0.72      |
| 1                   | 0.62   | 0.73     | 0.75      |
| 2                   | 0.61   | 0.73     | 0.84      |

Clusters having higher homogeneity scores demonstrate better inference precision. The iterative clustering-reshaping process increases the precision for clusters with a higher homogeneity score. It supports the assumption that the similarity within a cluster comes from the sharing of a common substructure—the reshaping process uncovers this structure and increases prediction accuracy. Clusters with lower homogeneity scores are less likely to share common substructure and benefit less from the reshaping process.

For example, a phenol group shows two proximate function types—a hydrophobic group arising from the benzene ring and a HBD group arising from the associated hydroxyl group. In this case, it is possible for the algorithm to confuse one binding pocket for another. This suggests that a single cluster may support interactions with multiple different chemical groups. A larger training set may provide disambiguation of these cases.

4.3.2 Inference

Toward the goal of automatic sub-cavity-based pharmacophoric inference, we characterized the function type of each sub-cavity within a protein binding site. This was a challenging task for several reasons: any bound ligand may not exploit all possible interaction sites, different bound ligands may satisfy the same interaction site in alternate ways and the presence of a ligand chemical group in a sub-cavity does not necessarily imply that the chemical group is the ideal interaction partner for the associated sub-cavity. Despite these caveats, the set of known binding ligands provided arguably the best source of information for learning sub-cavity binding preferences.

We performed binding inference for two protein systems, HIV-1P and Thrombin. These systems were selected because they are high-profile drug targets for which multiple complex structures had been solved. Starting from the known ligand binding site, we predicted the function type of each sub-cavity and compared these predictions to a set of known binding ligands. We constructed a training set of approximately 650 sub-cavities in a manner similar to Section 4.3.1; however, most importantly, we removed any sub-cavity which originated from either test protein or their homologs.

We applied the algorithm over the training set to produce a set of clustered sub-cavities. Using the function prediction method described in Section 4.3.1 with homogeneity thresholds of 0.65, 0.75 and 0.85 the sub-cavities of the two test proteins were predicted. The predictions were compared to a set of ligands known to bind the specified sub-cavities.

Binding site prediction for HIV-1 Protease: HIV-1 Protease is an aspartic protease responsible for the cleavage of two HIV polyprotein chains into several distinct proteins including the protease itself. The PDB contains structures for a total of nine HIV protease inhibitors which use the structural scaffold illustrated in Figure 5A. For example, the ligand of 1HWR (Ala, 1998) contains two phenyl groups (hydrophobic and aromatic) (R3, R7), two 1-butene groups (hydrophobic) (R4, R6), two hydroxyl groups (HBD and HBA) (R1, R2) and one carbonyl group (HBA) (R5). In our experiment, the 1HWR PDB complex was used as the query protein. We extracted seven sub-cavities using three homogeneity score
Table 2. Inference results for HIV-1 Protease active site

| Prediction | HBD | HBD | Arom. | Arom. | – | – |
|------------|-----|-----|-------|-------|---|---|
| 0.65       | 8/8 | 9/9 | 9/9   | 8/9   | – | 9/9|
| 0.75       | 8/9 | 9/9 | 9/9   | 9/9   | – | – |
| 0.85       | 8/9 | –   | –     | –     | – | – |

Sub-cavity inference results for the binding site of HIV-1 Protease. Using three different homogeneity score thresholds (0.65, 0.75, and 0.85), the predicted sub-cavity labels were compared to a set of nine ligands. R-groups refer to Figure 5A. The predictions made by our algorithm appear in the ‘Prediction’ row (Arom., aromatic). No prediction was made when the cluster’s homogeneity score did not pass the set threshold (indicated by ‘−’). An entry ‘X/Y’ indicates X correct predictions made for the Y ligands with a corresponding R-group (i.e. not all ligands have a substituted chemical group at each R position).

Sub-cavity inference results for the binding site of HIV-1 Protease. Using three different homogeneity score thresholds (0.65, 0.75, and 0.85), the predicted sub-cavity labels were compared to a set of nine ligands. The predictions made by our algorithm appear in the ‘Prediction’ row. No prediction was made when the cluster’s homogeneity score did not pass the set threshold (indicated by ‘−’). An entry ‘X/Y’ indicates X correct predictions made for the Y ligands with a corresponding R-group (i.e. not all ligands have a substituted chemical group at each R position).

Table 3. Inference results for Thrombin active site

| Prediction | R1 | R2 | R3 | R4 | R5 | R6 | R7 |
|------------|----|----|----|----|----|----|----|
| 0.65       | HBD + HBA| – | – | – | – | – |
| 0.75       | HBD + HBA| – | 9/9| 9/9| – | – | – |
| 0.85       | 6/6 | – | – | – | – | – | – |

Sub-cavity inference results for the binding site of Thrombin. Using three different homogeneity score thresholds (0.65, 0.75, and 0.85), the predicted sub-cavity labels were compared to a set of nine ligands. The predictions made by our algorithm appear in the ‘Prediction’ row. No prediction was made when the cluster’s homogeneity score did not pass the set threshold (indicated by ‘−’). An entry ‘X/Y’ indicates X correct predictions made for the Y ligands with a corresponding R-group (i.e. not all ligands have a substituted chemical group at each R position).

Sub-cavity inference results for the binding site of Thrombin. Using three different homogeneity score thresholds (0.65, 0.75, and 0.85), the predicted sub-cavity labels were compared to a set of nine ligands. The predictions made by our algorithm appear in the ‘Prediction’ row. No prediction was made when the cluster’s homogeneity score did not pass the set threshold (indicated by ‘−’). An entry ‘X/Y’ indicates X correct predictions made for the Y ligands with a corresponding R-group (i.e. not all ligands have a substituted chemical group at each R position).
We have presented an algorithm that infers binding site patterns may also allow protein function prediction or provide support for well-studied scaffold. Detailed knowledge of a pharmacophore map all members of the Lilien lab for helpful discussions and comments lead optimization and generated pharmacophore map could be used for virtual docking, prediction. More specifically, we believe that an automatically binding inference for proteins that do not share significant structural which they are compared. We demonstrated the algorithm's ability sub-cavities, but also in the more complete structural representation of the uniqueness of our approach lies not only in the consideration of computational challenges may arise when our approach is scaled recovery. Second, although not explicitly discussed in this article, inference, it is possible we have been too conservative with reporting the performance of our independent training and testing experiments. A TP by our definition requires a known holo structure involving the specified chemical group. In practice, the inferred interaction may indeed be correct; however, there may simply be no experimentally solved structure that can verify it.

First, we demonstrated the importance of sculpting sub-cavities toward their maximal common substructure; however, our results suggest that a more sophisticated MCS algorithm (Shatsky et al., 2006) may improve both classification accuracy and structure recovery. Second, although not explicitly discussed in this article, the consideration of sub-cavities, but also in the more complete structural representation of these sub-cavities, their parametrization and the method by which they are compared. We demonstrated the algorithm’s ability to leverage previously unused structural information to perform binding inference for proteins that do not share significant structural similarity with known systems. Using HIV-1 Protease and Thrombin as test cases, we have taken the first step toward sub-cavity-based pharmacophore inference. We intend to extend our work toward fully automatic pharmacophore inference and protein function prediction. More specifically, we believe that an automatically generated pharmacophore map could be used for virtual docking, lead optimization and de novo drug design. An example of a lead optimization effort using this approach would be to apply predicted binding preferences to the replacement of chemical groups on a well-studied scaffold. Detailed knowledge of a pharmacophore map may also allow protein function prediction or provide support for human-generated binding hypotheses.

5 CONCLUSION

We have presented an algorithm that infers binding site patterns by utilizing local similarity among active site sub-cavities. The uniqueness of our approach lies not only in the consideration of sub-cavities, but also in the more complete structural representation of these sub-cavities, their parametrization and the method by which they are compared. We demonstrated the algorithm’s ability to leverage previously unused structural information to perform binding inference for proteins that do not share significant structural similarity with known systems. Using HIV-1 Protease and Thrombin as test cases, we have taken the first step toward sub-cavity-based pharmacophore inference. We intend to extend our work toward fully automatic pharmacophore inference and protein function prediction. More specifically, we believe that an automatically generated pharmacophore map could be used for virtual docking, lead optimization and de novo drug design. An example of a lead optimization effort using this approach would be to apply predicted binding preferences to the replacement of chemical groups on a well-studied scaffold. Detailed knowledge of a pharmacophore map may also allow protein function prediction or provide support for human-generated binding hypotheses.

ACKNOWLEDGEMENTS

We thank Mr Abraham Heifets, Mr Satyam Merja, Ms Maria Safi and all members of the Lilien lab for helpful discussions and comments on drafts.

Funding: Bill and Melinda Gates Foundation (Grand Challenges Explorations) (to R.H.L.)

Conflict of Interest: none declared.

REFERENCES

Ala-Pietila, T. et al. (1998) Molecular recognition of cyclical urea HIV-1 protease inhibitors. J. Biol. Chem., 273, 12325–12331.
Banerchi, A. (2000) The ENZYME database in 2000. Nucleic Acids Res., 28, 304–305.
Berman, H.M. et al. (2000) The Protein Data Bank. Nucleic Acids Res., 28, 235–242.
Biskupski,T. et al. (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. J. Mol. Biol., 332, 505–526.
Chen,X. et al. (1999) Automated pharmacophore identification for large chemical data sets. J. Chem. Inf. Comput. Sci., 39, 887–896.
Dox et al. (2006) Predicting molecular interactions in silico: an updated guide to pharmacophore identification and its applications to drug design. Front. Med. Chem., 3, 551–564.
Fischler,M.A. and Bolles,R.C. (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM, 24, 381–395.
Frey.B.J. and Dueck,D. (2007) Clustering by passing messages between data points. Science, 315, 972–976.
Geilich,C. et al. (2007) Thermodynamic inhibition profiles of a cyclodextrin and a cyclodextrin derivative towards thrombin: the same but for different reasons. Angew. Chem. Int. Ed., 46, 8511–8514.
Hajjaj et al. (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins, 47, 409–443.
Hendlich,M. et al. (1997) Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. J. Mol. Graph. Model., 15, 239–251.
James,A.C. et al. (2006) Daylight Theory Manual-Daylight 4.71. Daylight Chemical Information Systems. Available at www.daylight.com.
Kahraman,A. et al. (2007) Shape variation in protein binding pockets and their ligands. J. Mol. Biol., 368, 283–301.
Koninghi,K. and Nakamura,H. (2003) Identification of protein biochemical functions by similarity searching using the molecular surface database e-site. Protein Sci., 12, 1589–1595.
Kuhn,D. et al. (2006) From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. J. Mol. Biol., 369, 1023–1044.
Kups-Bastian,K. et al. (2007) Large scale analysis of protein-binding cavities using self-organizing maps and wavelet-based surface patches to describe functional properties, selectivity discrimination, and putative cross-reactivity. Proteins Struct. Funct. Bioinform., 71, 1288–1306.
Langer,G. and Hoffmann,R.D. (2006) Pharmacophore and Pharmacophore Searches. 1st edn, Wiley-VCH, Palo Alto.
Laskowski,R.A.A. et al. (2005) Protein function prediction using local 3D templates. J. Mol. Biol.
Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J. Mol. Graph., 13, 323–330.
Ledvina,N. et al. (2001) Automated multiple structure alignment and detection of a common substructural motif. Proteins Struct. Funct. Genet., 43, 235–245.
Levi,D.G. and Banaszak,L.J. (1992) Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J. Mol. Graph., 10, 229–234.
McGeough,M. (2007) Pharmacophore map of small molecule protein kinase inhibitors. J. Chem. Inf. Model., 47, 2374–2382.
Morris,R.J. et al. (2009) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics, 25, 2347–2355.
Nopens, N. et al. (2008) Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. Bioinformatics, 24, i105–i111.
Nelder,J.A. and Mead,R. (1965) A simplex method for function minimization. Comput. J., 7, 308–313.
Pellecchia,M. et al. (2002) NMR in drug discovery. Nat. Rev. Drug Discov., 1, 211–219.
Prosen, X. and Aschauer,N. (1998) A geometric algorithm to find small but highly similar 3D substructures in proteins. Bioinformatics, 14, 516–522.
Powell,R. et al. (2006) Comparison of protein active site structures for functional annotation of protein and drug design. Proteins Struct. Funct. Bioinform., 65, 124–135.
I.Wallach and R.H.Lilien

Ramensky, V. et al. (2007) A novel approach to local similarity of protein binding sites substantially improves computational drug design results. Proteins, 69, 349–357.

Schmitt, S. et al. (2002) A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol, 323, 387–406.

Shatsky, M. et al. (2005) Recognition of binding patterns common to a set of protein structures. In Research in Computational Molecular Biology (RECOMB). Springer, Berlin/Heidelberg, pp. 440–455.

Shatsky, M. et al. (2006) The multiple common point set problem and its application to molecular binding pattern detection. J Comput Biol, 13, 407–428. PMID: 16597249.

Shulman-Peleg, A. et al. (2008) Prediction of interacting single-stranded rna bases by protein-binding patterns. J Mol Biol, 379, 299–316.

Sousa, S.F.F. et al. (2006) Protein-ligand docking: current status and future challenges. Proteins, 65, 15–26.

Stark, A. and Russell, R.B. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res, 31, 3341–3344.

Wallach, I. and Lilien, R. (2009) The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. Bioinformatics, 25, 615–620.