Published in final edited form as:

Proteins. 2014 April ; 82(4): 679–694. doi:10.1002/prot.24450.

Protein Function Annotation By Local Binding Site Surface Similarity

Russell Spitzer, Ann E. Cleves, Rocco Varela, and Ajay N. Jain*

Dept. of Bioengineering and Therapeutic Sciences, Univ. of California, San Francisco, CA

Abstract

Hundreds of protein crystal structures exist for proteins whose function cannot be confidently determined from sequence similarity. Surfex-PSIM, a previously reported surface-based protein similarity algorithm, provides an alternative method for hypothesizing function for such proteins. The method now supports fully automatic binding site detection and is fast enough to screen comprehensive databases of protein binding sites. The binding site detection methodology was validated on apo/holo cognate protein pairs, correctly identifying 91% of ligand binding sites in holo structures and 88% in apo structures where corresponding sites existed. For correctly detected apo binding sites, the cognate holo site was the most similar binding site 87% of the time. PSIM was used to screen a set of proteins that had poorly characterized functions at the time of crystallization, but were later biochemically annotated. Using a fully automated protocol, this set of 8 proteins was screened against approximately 60,000 ligand binding sites from the PDB. PSIM correctly identified functional matches that pre-dated query protein biochemical annotation for five out of the eight query proteins. A panel of twelve currently unannotated proteins was also screened, resulting in a large number of statistically significant binding site matches, some of which suggest likely functions for the poorly characterized proteins.

Keywords

binding site; similarity; PSIM; protein function; Surfex; annotation

Introduction

We have previously shown that a local, surface-based, protein binding site similarity metric can identify biochemically relevant relationships between proteins having varying levels of sequence similarity.¹ PSIM distinguished subtle differences between proteins sharing significant sequence similarity but which bind different ligands, and it also identified common functional motifs shared by heterogeneous proteins. For example, the pocket similarity was higher for kinase pairs known to share ligand binding preferences than for kinase pairs with divergent ligand specificity. At the other end of the scale, among highly diverse ATP-binding proteins, functional motifs were placed into an accurate taxonomy using the PSIM metric. Sensible segregation was seen between proteins with divergent functions (e.g. serine threonine kinases and tRNA synthetases) as well as among functionally distinct variants within related classes. Extensive comparison to other structural similarity algorithms was presented (e.g. SiteBase,² spherical harmonics,³ sup-CK,⁴ PocketMatch,⁵ and SMAP⁶), illustrating the complementary of the PSIM local surface similarity approach.

*To whom correspondence should be addressed: ajain@jainlab.org.

Figure 1 illustrates the PSIM computation, showing the optimal alignment of *E. coli* lysyl-tRNA synthetase and human glycyl-tRNA synthetase, proteins with 30% sequence identity. The method computes a local comparison of surface shape and electrostatic properties from the vantage point of a putative ligand, irrespective of residue identities or protein backbone correspondence. The two proteins in Figure 1 both possess an ATP-binding motif that is characteristic of type II tRNA synthetases. The PSIM taxonomy distinguished this motif from the related, but functionally distinct, type I tRNA synthetase motif. Even in cases where local sequence similarity was non-existent, as observed in some functionally related binding-site pairs from organisms in different kingdoms, PSIM's local similarity computation produced the correct groupings and structural alignments. The ability to make coarse-grained classifications as well as fine-grained distinctions using only binding site similarity suggested the possibility to use PSIM for functional protein annotation.

In the absence of obvious sequence similarity or biochemical experiments, accurate functional annotation requires the ability to directly relate two proteins based on their physical characteristics. PSIM addresses this challenge by evaluating the similarity of binding site surfaces between proteins of interest. The original PSIM algorithm required either a bound ligand or a manually identified location to define a binding site. PSIM has been augmented with an automatic binding site detection feature, removing the need for bound ligands or manual preparation. Screening for functional annotation requires an efficient approach for querying against tens of thousands of protein structures. To this end, the PSIM algorithm has been improved, resulting in a 100-fold speed increase while maintaining performance accuracy. Screening against large libraries also requires the ability to calculate the significance of results given that a large set of purely random comparisons may yield some apparently high scores. A new empirical framework for attributing p-values to protein pair-wise similarity scores provides statistical confidence in PSIM annotations.

The binding site detection methodology was validated on a set of 304 apo/holo crystal structure pairs from the LigASite database.⁷ Valid binding sites on apo structures were identified and accurately matched to corresponding sites on holo structures. Site detection performance was equivalent on both holo and apo structures: 91% and 88%, respectively. Further, when querying the holo sites using a detected apo site, we recovered the cognate holo site in 87% of cases. Additionally, similarity scores for apo/holo cognate pairs were statistically separable from scores of random structure pairs (p-value < 1×10^{-10}).

Functional annotation was demonstrated with a screen using a temporally segregated data set. A set of eight proteins whose functions were unknown at the time of deposition were screened against the fraction of the PDB available prior to their definitive biochemical annotation. Using PSIM, the correct function for five of the eight proteins was identified. With respect to functional annotation, PSIM offered complementary information to other established sequence and structural comparison methods (BLASTP,⁸ CE,⁹ and SMAP¹⁰). Finally, a screen of twelve uncharacterized proteins against a large database of binding sites resulted in several suggested annotations that merit further investigation.

We believe that the PSIM surface-based approach to functional annotation provides a novel avenue to predicting protein function. PSIM is complementary to sequence-based and backbone-structure-based approaches and may be applied for confirmation or discovery of function. The results provided here demonstrate that the PSIM method may be productively applied on comprehensive repositories of protein binding sites.

Methods and Data

Functional annotation screening required several algorithmic enhancements including: automatic binding site detection, throughput improvements, and a method for determining the statistical significance of a raw similarity score. The new methods along with the data sets utilized for validation and screening are detailed in the following. Data, software, and computational protocols are available by request (see www.jainlab.org for details).

Molecular Data Sets

Curated LigASite Database—The LigASite Database contains proteins which have been crystallized in both the presence and absence of a bound ligand. When retrieved, the set contained 383 non-redundant pairs of apo/holo structures. Pairs were removed if their ligand did not meet several criteria: a minimum of 7 heavy atoms, at least 3 heavy atoms within 1 angstrom of the protein, and sufficient “buriedness.” The buriedness of the ligand was measured by taking the ratio of near-ligand protein atoms to the total number of heavy atoms on the ligand. A ratio of 4.0 was used to filter the LigASite database. These various criteria were set such that the ligands present in a standard small-molecule docking data set would be preserved (the Astex85 Set¹¹). The majority of ligands that did not pass the filter were ions and crystallization agents with mostly hydrophobic interactions with protein surfaces. The filtered LigASite set contained 304 apo/holo pairs and 606 ligand binding sites.

Temporally Segregated JCSG Query Set—The temporally segregated query set was created to validate PSIM’s ability to annotate proteins of unknown function without ligands bound. Proteins were identified which had been structurally determined prior to their biochemically annotation. To accomplish this, we queried the Joint Center for Structural Genomics (JCSG) for structures characterized as “Unknown,” “Uncharacterized,” or “Hypothetical.” These queries resulted in 608 matches representing 605 unique PDB identifiers. This set was then reduced to only those structures without bound ligands, eliminating all of the proteins with obvious ligand binding activity or confirmed binding sites. Manual inspection produced 8 unique proteins which were biochemically annotated subsequent to their crystallization.

Un-annotated JCSG Query Set—Annotation of currently uncharacterized proteins was performed on an additional set from the JCSG. The un-annotated JCSG query set was created in a similar manner to the temporally segregated set. The 605 structures characterized as “Unknown,” “Uncharacterized,” or “Hypothetical” were filtered to remove those that had an annotated PFAM domain or an EC number. The 264 structures which remained were filtered again to keep only structures that were determined using X-ray crystallography. The remaining 248 PDB structures came from a variety of sources: bacteria (213), archaea (26), eukaryota (6), viruses/other (3). From this pool, 12 structures from model organisms were selected for screening.

Curated RCSB PDB Library—Screening was performed against a library of known binding sites derived from the RCSB PDB.¹² All available PDB structures which possessed a bound ligand were procured. This set was then filtered in an identical manner to the LigASite Database, removing those nominal binding sites with very small ligands or which did not form a clear cavity. In total, 30,999 PDB structures with 63,669 binding sites were present in the final library. At the time of curation, this represented roughly 50% of the crystal structures present in the PDB.

Computational Methods

Surface Based Binding Site Similarity—The core technology of PSIM is a local surface-based molecular similarity computation, which itself was based upon a similarity method for the comparison of small molecules.^{1,13} Figure 2 depicts the alignment procedure as applied to lysyl-tRNA synthetase and glycyl-tRNA synthetase. Structures typically begin out of alignment with their binding sites in different locations and orientations. The scope of the comparison must be defined by either a bound ligand, a manual selection, or by an automatic binding site detection routine. In this example, the bound ATP molecules are used to define the binding sites on both structures. Each binding site is tessellated with observation points, illustrated by colored spheres in Figure 2A. Each observation point measures the distances to the nearest steric, positively charged and negatively charged surfaces. A transformation is found which moves each observation set to the opposing protein and minimizes the differences in corresponding surface measurements (Figure 2B). For example, observation points which are initially near a positively charged surface on lysyl-tRNA synthetase are placed as close as possible to similarly charged surface on glycyl-tRNA synthetase. After optimizing the positions of observation points, the transformation is applied to the proteins, placing their binding sites in alignment, as in Figure 2C. The similar regions of the protein surface are then depicted using the colored sticks, as in Figure 2D. The two synthetases shown have extremely similar binding sites, as indicated by the numerous colored sticks, despite having only 30% sequence identity. The binding site surfaces similarity is amalgamated into a single score ranging from 0 to 1 with this example producing 0.66, representing a high level of similarity. Further details about the computation of protein site similarity scores can be found in our earlier work.¹

Improvements to Surflex-PSIM—The original PSIM algorithm,¹ while accurate, left several opportunities for enhancements to aid in screening applications. While the original computation optimized the alignment of proteins by moving protein atoms, the new method moves the observation points instead. This significantly reduces the number of calculations required, because the average protein has on the order of 10,000 atoms in contrast to the average 400 points in an observation set. Using observation points instead of protein atoms also allows for the geometric hashing of protein coordinates, crucial for the quick lookup of nearby surfaces. In total, these enhancements reduce execution time from 60 seconds per comparison to 0.5 seconds. The results of the new method correlated nearly perfectly with those produced by the previous version (Pearson's $r = 0.99$).

The definition of the binding-site region has also been improved to allow for two simultaneous and independent sets of observation points. The original PSIM alignment was largely dependent on a single set of observation points. This occasionally allowed for degenerate solutions when working with two binding sites with significantly different geometry or widely varying size. PSIM now optimizes the placement of both feature sets simultaneously. Using two sets of observation points allows for the geometries of both binding sites to be utilized throughout the optimization process, producing more accurate alignments when comparing divergent sites.

Automatic Binding Site Detection—The PSIM algorithm now includes a procedure for automatically detecting binding sites on protein structures, which is critical for comparing binding sites on structures without bound ligands. An example with Bothropstoxin-I (PDB 3CXI) is shown in Figure 3A. The protein is divided into 1 Angstrom voxels marked as either solvent-exposed or internal to the protein. Voxels are ranked by their level of concavity, with concavity measured by counting the number of intersecting paths between pairs of atoms in the region of interest. Paths are assigned a score based on their lengths and the number of internal voxels intersected. Shorter paths that intersect with few internal

voxels receive higher scores, and the sum of a voxel's path scores is defined as its concavity. This method parallels previous algorithms in Surflex-Dock¹⁴ and Hammerhead^{15,16} for binding site detection and characterization. A fraction of the highest scoring voxels are retained and organized into discrete regions, using a straightforward procedure for connected-components enumeration. In Figure 3B, the results of this procedure are shown with each set of colored spheres representing a different concave region.

Each detected connected region is characterized by constructing an optimal pseudo-ligand for the site, again paralleling procedures developed previously for docking.^{14–16} Each pseudo-ligand is built from three types of molecular fragments: positive polar (N-H), methane (CH₄), and negative polar (C=O). All such fragments are placed in candidate binding sites so as to optimize their intermolecular interaction with the protein according to the Surflex-Dock scoring function.¹⁴ Pseudo-ligands with low protein interaction scores or with too few atoms are removed. In Figure 3C pseudo-ligands have been built for Bothropstoxin-I; the yellow pseudo-ligand on the right side of the protein occupies the same cavity as the cognate ligand vitamin E (orange).

All parameters for concavity detection, region definition, and binding site filtering were selected so that the ligand binding sites within the Astex85 docking set were identified and retained. Note that the procedure does not identify *only* those sites within the Astex85 set that contain ligands used in docking benchmarking. There are roughly three-fold more nominal binding sites detected than contained bound ligands, and it is not known what proportion of these other sites may have a biological function. The intention for the PSIM algorithm has been to ensure detection of all plausible binding sites rather than to optimize for detection of only those sites where bound ligands have been crystallographically determined.

Converting Raw Similarity Scores to P-Values—PSIM raw similarity scores are converted to p-values using an empirical statistical framework and a randomly selected set of unrelated protein binding sites. The random set was created by reducing the PDB library detailed above to proteins with less than 30% sequence identity and by removing proteins which bound the same ligand. This produced 1556 binding sites. A candidate binding site is profiled against this set, producing raw scores, which were observed to be normally distributed. Using a normal distribution to fit the empirical score population, a score resulting from comparison of the candidate site to a protein of interest is converted into a p-value using the analytical integral of the fitted distribution. A typical protein binding site had a mean of 0.48 and standard deviation of 0.026 using PSIM with this random background set. So, a PSIM score of 0.60 would yield a p-value of 1.0×10^{-6} . The same approach, employing the same background set of protein binding sites, was used to compute p-values for CE¹⁷ and SMAP⁶ in the results that follow. The comparison methods CE and SMAP were used as comparators due to their maturity, wide use, and availability.

Binding Site Screening

Functional annotation of binding sites was performed in a screen against the previously described library of roughly 60,000 PDB binding sites. Following automatic binding site detection on each query protein, all of the predicted binding sites were compared to the entire binding site library. The resultant scores were transformed into p-values, and those scores passing a threshold were labeled as hits (Z-score > 5.0, equivalent to $p < 2.9 \times 10^{-7}$). This threshold was chosen such that if no binding site within the library was related to the query, there would be a less than 2.0% chance of returning a nominal hit.

Results and Discussion

The overall aim of this study was to establish the feasibility of a high-throughput protocol utilizing local protein binding site similarity for functional annotation of poorly characterized proteins. Given a high-resolution structure of such a protein from a project such as the JCSG, putative binding site detection yields candidates for screening against a library of well-characterized sites from the PDB. The following details experiments for validation of the binding site detection algorithm, making use of matched apo and holo protein structures. Next, the feasibility of functional annotation is presented, making use of 8 proteins from the JCSG, using a strategy of temporal segregation. Functional annotation evidence for each query protein was derived from protein structures deposited in the PDB *before* the query structure. Finally, the results of a screen for function of 12 proteins that are, as yet, poorly understood, will be presented, providing putative new annotations for these proteins.

Binding Site Detection and Apo/Holo Matching

To assess whether the automatic binding site detection and characterization approach would be potentially useful for functional annotation, three aspects were quantified: 1) the detection rate of ligand binding sites in holo protein structures (with ligands removed); 2) the detection rate in corresponding apo structures; and 3) whether the detected sites on apo structures could be correctly matched to their cognate holo binding sites. Performance in the last two assessments was the most important in establishing feasibility for PSIM for the functional annotation application.

Detection of Binding Sites on Structures with Ligands Bound—Binding site discovery was performed on 304 holo structures (containing a total of 606 bound ligands), using the parameters derived from initial experiments done on a ligand docking benchmark (see Methods for details). After removal of ligands, the detection algorithm produced a list of putative binding sites, each characterized by a pseudo-ligand that defined its extent. If such a binding site overlapped the volume of the known bound ligand by more than $20\text{\AA}^3$, the binding site was considered found. Figure 4 illustrates a successful case of detection. Overall, the site detection procedure correctly identified the locations of 554 out of 606 ligands, a discovery rate of 91%. In total, slightly more than 4,000 putative binding sites were proposed, yielding a ratio of detected sites to crystallographically observed ligands of about 7:1. It is not possible to know what proportion of the “excess” sites are false positives and which may have an as yet uncharacterized function. However, as mentioned earlier, for a functional annotation application, the critical feature is a high true positive detection rate. Nominal false positive sites contribute to increased computational burdens, but the presence of “decoy” sites on query protein structures can be ameliorated by using conservative cutoffs for defining screening hits.

Detection of Binding Sites on Structures without a Ligand Bound—The detection problem in apo structures is more challenging than for holo structures, in some cases becoming impossible due to binding site collapse. Binding sites which were occluded by molecular motion between apo and holo forms were omitted from this analysis. Overall, 11 protein pairs of the original 304 exhibited very significant conformational movement between apo and holo forms. For example, the pair of riboflavin kinase structures, shown in Figure 5, demonstrated a significant side chain motion which completely obscured the ligand binding site. In the holo form, ATP (depicted in orange skin) occupies a tight pocket which, in the apo form, is filled by the rotation of a proline and a phenylalanine (depicted with sticks). Obscured binding sites were identified by aligning the apo and holo structures and determining the amount of protein from the apo structure which overlapped with a

ligand from a holo structure. If more than 50% of the ligand volume was occluded, the binding site was considered obscured.

Of the 606 binding sites from the holo detection experiments, 584 were unobscured in the apo set. Within this set of 584 detectable apo binding sites, the algorithm discovered 514 sites, a discovery rate of 88% (criteria were analogous to the previous experiment). Within the apo set, the ratio of total discovered binding sites to established ones was just slightly lower than in the holo case, roughly 6:1. Figure 6 presents an example of an accurate recovery of the bothropstoxin-I binding site. Upon alignment of the apo and holo structures of bothropstoxin-I, the holo structure ligand Vitamin E overlapped with a detected apo binding site. Note that the overall extent of the apo site, as defined by the green molecular fragments, was smaller than that occupied by vitamin E. Comparison of sites with asymmetrical sizes, which had been a challenge for the original PSIM algorithm, is now addressed with the use of bi-directional site comparison (see Computational Methods). Tolerance to such site variations was critical for efficient recovery of matching sites, as described below. If we consider the full set of apo binding sites (including those 22 that were collapsed relative to their holo form), the analysis changes relatively little. The success rate for detection becomes 85% (514 detected sites from 606 total).

Recovery of Holo Structures using the Apo Structure as a Query—The last aspect of feasibility for protein function annotation requires that a properly detected binding site on an apo protein be correctly matched to a protein binding site with known function. Here, we assessed the likelihood of correctly identifying the cognate partner of each detected apo binding site (514 total) as the top-scoring match in a screen against *all* of the ligand binding sites from the holo structures (606 sites). In 448/514 cases, based on local surface similarity alone, the detected binding sites were correctly matched (an 87% success rate). Further, there was a significant separation in the overall distributions of scores when considering cognate apo/holo pairs and randomly assigned pairs ($p < 1.0 \times 10^{-10}$ by Wilcoxon rank sum test).

We assessed the degree to which the presence of “excess” detected binding sites would obscure the matching of an apo protein binding site to its correct cognate holo partner. The distribution of scores for matching cognate/apo pairs from the preceding analysis was treated as the positive distribution. We then identified all pairings of the excess binding sites (3362 in total) from the apo proteins to the holo binding sites (606 in total). From this set of pairings, we randomly selected 10,000 pairs, including only those where the apo and holo proteins were cross-pairs. This set forms a pool of potential false positives that may pose a problem for accurate matching of the true apo/holo pairs. The optimal PSIM alignments and scores were computed for these putative false positives and treated as a negative distribution. The separation between these distributions, measured by ROC area, was 0.999 (the 95% confidence interval was 0.999–1.0). Clearly, the presence of the excess apo sites does not pose a substantial problem in terms of overshadowing the scores of the true matches between the apo binding sites and the correct holo sites.

Last, we assessed the direct matching of entire apo proteins, considered as collections of all detected binding sites, against holo proteins, considered as collections of all liganded sites. The score between each apo/holo pair was simply the maximum PSIM value for all pairings of their respective binding sites. In making an ROC analysis, positive pairs were the matching apo/holo proteins from the LigASite database,⁷ and all other pairs were considered to be negatives. The ROC area was 0.90 (95% confidence interval 0.87–0.93). This final analysis paralleled the previous two. The presence of the “excess” binding sites was not a significant factor, with scores arising from their matches almost never reaching the magnitude seen between true positive binding site matches. However, as seen above, apo

proteins occasionally undergo binding site collapse for the critical site, thus reducing the chance to make a correct high-scoring match. Further, the “true” match for an apo protein scores less well than “false” matches roughly 13% of the time (see the previous analysis regarding top match success rate). Last, because all scores are considered as a pool, secondary matches for one protein (nominally false positives) may score better than the correct match for another. An example of this last case is 1BKZ (apo galectin-7), which was correctly matched to 5GAL (galectin-7 bound to N-acetyl-lactosamine, score = 0.815), but which also produced a high score against 2ZGM (0.780), which is an *Agrocybe aegerita* lectin in complex with lactose. The latter “false positive” is a reasonable match, but it is labeled as a negative, and it scores better than a number of the exact apo/holo cognate pairs. Even with this complication, recovery of 80% of the true positive matches was obtained at a nominal false positive rate of 0.2%.

Overall, considering binding site collapse in apo proteins (4% loss), detection performance (88% success), and matching accuracy (87% correct top match), the results suggest a probability of roughly 73% for correctly finding and matching a binding site to its cognate ligand-bound variant. This represents an upper bound on the effectiveness of the functional annotation pipeline. In this scenario, the cognate match will be close to maximally similar to the query under normal conformational variation. In a prospective functional annotation scenario, where no database proteins have significant sequence similarity to the query, it should be expected that the structural similarity of the query binding site to any database member would be lower, thus reducing the chances of identifying a strongly related site.

Analysis of Temporally Segregated JCSG Proteins

Figure 7 shows the structures of eight proteins from the JCSG chosen to validate the functional annotation capability of the PSIM algorithm (see Methods and Data for details on the selection protocol). For each protein, site detection was performed using the default parameterization described earlier, and the locations of those putative binding sites are depicted with colored surfaces in Figure 7. For each identified protein site, a screen was carried out using PSIM against a library of roughly 60,000 protein binding sites extracted from the PDB. In each case, database proteins were considered whose deposition date was *earlier* than the date the query protein was biochemically characterized. Using this approach, top-scoring matches were returned for 5 of the 8 query structures, providing matches that elucidated biological function. The remaining 3 structures (1NC5, 1PB6, and 2GBO) yielded no matches exceeding the chosen significance threshold. Further analysis showed that 1NC5 and 1PB6 matched variations of the same proteins deposited in the PDB subsequently to the original structures. The following provides details for the five positive cases: 2FB5, 1PV1, 1DUS, 2AS0, and 1YVO.

2FB5: A Novel Nucleotide Binding Fold—The *Bacillus cereus* protein structure 2FB5 was deposited in 2005 without any knowledge of function. Both the function of the protein and its binding partners remained elusive until homologues of the protein were studied in *Bacillus subtilis* and *Thermotoga maritima* in 2008. These homologues were shown to possess di-adenylate cyclase activity.¹⁸ The ATP co-crystal structures exhibited a novel nucleotide binding fold, which defined a new PFAM motif, designated DisA_N (PF02457, “DisA bacterial checkpoint controller nucleotide-binding”).

The PSIM screen for 2FB5 returned two matches to structures that had been deposited *prior* to the 2008 paper (1DMA from 1995, and 2UVX from 2007), correctly identifying the nucleotide binding site (see Table I for additional details). Figure 8 shows these two hits aligned to the 2FB5 binding site. Figure 8A highlights a strong similarity in the region corresponding to the hinge of CDK2. This region possesses a characteristic alternating set of

hydrogen bond donor and acceptor surfaces essential to binding the nucleotide head of ATP. This similarity clearly identifies the presence of a nucleotide binding motif in 2FB5 and is further reinforced by the presence of the same motif in 1DMA (Exotoxin A), where the motif is also used to bind nucleotides.

The SMAP and CE methods yielded weaker p-values than PSIM for these protein comparisons (see Table I), and this is explained by their dependence on congruence of protein backbones. Figure 8B shows that the secondary structure of 2UVX is quite distinct from that of 2FB5. There exists almost no secondary structure or atom to atom correspondences in the optimal binding site alignment. However, the local surface similarity depicted in Figure 8C shows nearly identical geometry and surface charges in the region around the nucleotide. The green sticks portray a tight cavity formed in the shape of the adenosine, with alternating blue and red sticks indicating the similar “hinge” shared by all 3 structures. For 2FB5, the PSIM approach would have suggested a nucleotide binding function at the time of structure deposition in 2005 through highly significant similarity to 1DMA (which had been deposited in 1997). This may have hastened characterization of this protein.

1PV1: Hypothetical Esterase—The *S. cerevisiae* protein structure 1PV1 was deposited in 2003, having been hypothesized to have esterase activity based on 40% sequence similarity with the known human esterase hEstD. Uncertainty regarding the catalytic mechanism (serine- or cysteine-based) was resolved at that time (its unusual fold made sequence-based determination of this aspect impossible).¹⁹ However, the specific structural reasons for pronounced resistance to organophosphate (OP) inhibition was not understood until further studies, including mutagenesis, were published in 2008.¹⁹ The PSIM database screen of 1PV1 returned four binding site matches deposited prior to the 2008 publication: chloroperoxidase T (PDB 1A8U, $p = 7.1 \times 10^{-12}$), a still unnamed *Thermotoga maritima* esterase (PDB 3DOI, $p = 4.4 \times 10^{-11}$), butyrylcholinesterase (PDB 1XLU, $p = 4.8 \times 10^{-10}$), and valacyclovir hydrolase (PDB 2OCI, $p = 3.7 \times 10^{-8}$). The presence of a mixture of haloperoxidases and esterases owes to their established mechanistic inter-relatedness²⁰ and also to the unusual structural features of 1PV1.

The 1A8U match (chloroperoxidase T, a co-factor free haloperoxidase) was deposited in the PDB in 1998,²¹ well before deposition of 1PV1 and thus could have been used to help guide the functional understanding of this protein. The alignment shown in Figure 9 shows the near-perfect alignment of the catalytic triad common to both enzymes (lower left, Asp-Ser-His residues). The single stark difference between the two binding sites is the intrusion of an indole from 1PV1 into the binding site of 1A8U where the benzoate ligand is bound. OP pesticides are known to be substrates of chloroperoxidases,²² and the structural alignment seen in Figure 9 implicates W197 (numbering from 1PV1) as an interesting feature to investigate. The W197I mutant exhibits over 1000-fold more sensitivity to paraoxon inhibition than the wild-type enzyme.¹⁹ While it is speculative, we believe that the group of alignments produced by PSIM, all of which were identifiable in 2003, may have accelerated the structural understanding of OP resistance. Unlike an inference based upon sequence identity, the alignment shown in Figure 9 both confirms the presence and geometry of key functional groups as well as exposing a particular tryptophan as an unusual feature. The secondary structures of 1A8U and 1PV1 (Figure 9B) show relatively little local similarity, again explaining weak p-values from backbone and sequence based methods (CE 1.9×10^{-2} SMAP: 2.0×10^{-2} and BLASTP: 1.6×10^{-1}).

1DUS: Subtleties with Structural Comparisons—The *Methanococcus jannaschii* protein structure 1DUS was deposited into the PDB in 2000, and had no sequence similarity to any annotated proteins. Structural similarity methods available at the time suggested

similarity to a diverse group of methyltransferase proteins. Further testing and analysis ultimately lead to 1DUS's annotation as a DNA methyltransferase.²³

PSIM returned nine matches that were deposited prior to the deposition of 1DUS. The results are shown in Table II, clearly indicating that this binding site is a methyltransferase. All of the most significant matches were DNA and RNA methyltransferases. These were correctly ranked above proteins such as PRMT3 and ChOMT. The results for CE yielded weaker p-values and also placed PRMT3, an arginine methyltransferase, among the top three hits. For this case, the SMAP method, while producing p-values less extreme than for PSIM, correctly ranked the nucleotide methyltransferases above the others.

The oldest structure which could have provided the correct annotation for 1DUS is 1AQJ, deposited 4 years prior to the deposition of 1DUS. The alignment of 1AQJ and 1DUS (Figure 10A) shows a clear correspondence between residues of the uncharacterized protein and the confirmed DNA methyltransferase. Here, in contrast to the previous examples, both the local similarity at the surface and atomic level *and* the local backbone congruence (see Figure 10B) are high. Interestingly, though these features are apparent to all of the tested structural methods, BLASTP was unable to detect the similarity between 1DUS and 1AQJ (E-value, 1.3).

2AS0: Strong Confirmation of Activity—The *Pyrococcus horikoshii* protein structure 2AS0 was deposited in the PDB in 2005. Structure and sequence based evidence suggested its function as a hypothetical RNA methyltransferase.²⁴ The hypothesis was confirmed in 2008, demonstrating methyltransferase activity specific to 23S rRNA.²⁵ The PSIM screen of 2AS0 returned nine hits, all indicating methyltransferase activity. All matches were RNA or DNA specific methyltransferases, strongly indicating nucleotide-specific methyltransferase activity. The earliest match was deposited in 1998 (shown in Figure 11A). The nearly identical fold structure depicted in Figure 11B suggests this to be a particularly easy case for structural comparison methods, and both the CE and SMAP p-values were less than 1.0×10^{-5} . Of note, a later variant of the protein represented in the previously mentioned 1DUS example was among the matches (PDB code 2YX1, PSIM p-value: 1.2×10^{-9}).

1YVO: Alternative Characterization of Putative Function—The *Pseudomonas aeruginosa* protein structure 1YVO was deposited into the PDB in 2005. At that time, it had been thought of as belonging to the GCN5-related family of acetyltransferases (GNAT), with BLAST-based sequence searching identifying similar enzymes in a wide range of bacteria.²⁶ The enzymes included hypothetical proteins with no known function or were putative phosphinothricin acetyltransferases. Such acetyltransferases catalyze the acetylation of phosphinothricin (a widely used herbicide), and these enzymes are interesting due to the potential for engineering resistance into plants treated with phosphinothricin.

Shortly after the original deposition of the structure, two protein structures were identified as being similar, based on a DALI search: 1VHS and 1ON0, the former being a putative phosphinothricin acetyltransferase and the latter a simply a putative acetyltransferase.²⁶ Late in 2006, the same authors showed that the enzyme was *not* a phosphinothricin acetyltransferase, as suggested based on their initial structural investigation.²⁷ Rather, it was shown that it was an acetyltransferase, but that both L-methionine sulfone and L-methionine sulfoxide were included among substrates with measurable turnover. The PSIM screen of 1YVO returned six matches deposited prior to definitive functional annotation of the protein. All matches were to proteins with confirmed acetyltransferase activity. The strongest match, 2C27 (an acetyltransferase that is a mycothiol synthase), was deposited in 2005 (see Figure 12A). The similarity exists within the binding site and also at the scale of secondary structure (see Figure 12B). As with the case of the clear catalytic triad motif of

1PV1, the structural similarities here reflect a physicochemical environment in the binding site that is congruent to examples with known acetyltransferase activity. Additionally, 2I79, 2BEI, and 1OZP were also identified by PSIM. Little is known about the 2I79 protein, but the 2BEI protein is a human thialysine acetyltransferase, and the 1OZP protein is a bacterial mycothiol synthase (as with 2C27).

The family of proteins represented by those just discussed have widely varying substrate specificity, and their functional characterization has been furthered very little by sequence-based modeling. In cases such as that of 1YVO, where sequence based annotation is non-definitive, structural methods can be utilized as an orthogonal annotation approach. Here, while some structural methods pointed toward specific phosphinothricin activity, PSIM did not. The PSIM hits included acetyltransferases that acetylate amines that are beta to sulfur atoms (mycothiol synthase and thialysine acetyltransferase). The substrates identified by Davies et al.²⁷ have sulfur atoms at the gamma position. It is not clear, however, whether identification of these protein structures by PSIM is likely to be of help in further characterization of the enzyme, or, indeed, would have helped during the initial biochemical investigations. However, the clear complementarity of the PSIM hits to those identified by other methods, both sequence- and structure-based, offers the advantage of an additional information source in developing hypotheses.

Library Search of Currently Unknown JCSG Proteins

In the foregoing, the PSIM algorithms for binding site detection and database screening were validated using retrospective experiments. Analysis of performance on the apo/holo detection and matching exercise suggested an upper bound of roughly a 73% chance of correctly identifying a structural match given an un-liganded query protein structure. For five of eight cases (63%), the PSIM method was able to provide correct matches to identify or disambiguate biochemical functions of proteins using structures publicly available *prior* to definitive functional annotation of the query proteins. Given this degree of success, PSIM was also tested on 12 JCSG proteins with no current functional annotation. None of these proteins has either known PFAM domains or sequence similarity to any known protein (see Figure 13).

A PSIM database screen analogous to the validation experiment just described returned numerous statistically significant matches. Table III summarizes the screening hits obtained for all twelve proteins. The vast majority of nominal hits were matches to shallow surface pockets of proteins within the PDB binding site database or where the bound ligand within the matched site was notably promiscuous (e.g. ethylene glycol, a common crystallization reagent). It is not clear that such sites lack important biological functions, but because their potential functions are not clear, their relevance is difficult to assess. For the query structure 2GNX, which included twelve separate putative binding sites, direct inspection of the matches to identify non-degenerate ones yielded four related hits. Figure 14 illustrates the match to the uric acid binding site of 2YZD, with the same coloring scheme as used earlier. The protein is a urate oxidase from *Arthrobacter globiformis*, and it exhibits folds typically found in purine- and pterin-binding enzymes.²⁸ Other matching structures included three proteins bound to hadacidin (1CG1, 1CG0, and 1KKB, not shown). The three structures are all variants of the same purine biosynthesis enzyme (adenylosuccinate synthetase) from *E. coli*, and hadacidin is an inhibitor of multiple enzymes involved in purine biosynthesis.²⁹

Based on the related functions of the matches to proteins found in *Arthrobacter globiformis* and *E. coli*, we suggest that the mouse protein whose structure was deposited as 2GNX (Uniprot ID Q6P1I3) is likely an enzyme involved in some aspect of purine biosynthesis or metabolism. Clearly, this suggestion of potential function is not conclusive. However, it has been included to show that highly significant hits exist for proteins of current interest and to

stimulate experimental investigation of the putative functions. All of the data for the database screen of the twelve unknown proteins is included as part of the data archive associated with this paper. Automated methods for detecting deeply buried high-content ligand matches where multiple matched structures show high mutual similarity at the ligand level will supplant manual inspection of hit lists.

Conclusions

We have presented a fully automatic high-throughput pipeline for protein function annotation through local binding site surface similarity comparisons. The PSIM method is capable of detecting known binding sites in un-liganded protein structures approximately 90% of the time and of matching such sites to their cognate ligand-bound variant with roughly 90% success. Given the occasional occurrence of protein binding site collapse when a binding partner is absent, overall we estimate that 70% success is the upper bound for PSIM to correctly identify a match for a protein site in a large database screen. A query site can be screened against a comprehensive database of ligand binding sites (roughly 60,000 members) in less than one day on a single computing core using standard desktop hardware. Of course, such an application is “embarrassingly parallel” making large screens of many sites against large databases easily prosecuted on computing clusters.

Our initial feasibility studies here on eight JCSG proteins whose functions were definitively assigned subsequent to their structure deposition yielded success in five cases. The most challenging case was one in which a novel nucleotide binding fold was evident only using the local surface similarity method implemented within PSIM. Other structural comparison methods and sequence methods yielded insufficiently strong matches to provide confident functional annotation. In a number of cases, structural comparison methods that relied upon protein backbone similarities yielded solid matches. However, the PSIM approach offered generally stronger evidence in a statistical sense, and it certainly provided an orthogonal means by which to rank the matching protein binding sites. Currently the Protein Structure Initiative³⁰ contains at least 264 proteins which contain a PFAM domain of “unknown function” or no PFAM domain at all, and these present an opportunity to apply novel similarity metrics for annotation. Our hope is that investigators will be able to make use of the PSIM method along with the PDB ligand binding site database developed for this work in order to aid in protein characterization.

A related area of inquiry involves studies of proteins in dynamic simulation environments, where groupings of related structures may be used for improvements in sampling and for elucidation of transient binding sites.³¹ The PSIM approach may offer a different picture from that seen through atomistic clusterings of related protein variants. Apart from questions involving protein function, the PSIM method has clear applications in molecular docking, both for protein alignment and conformational variant selection.³² This is an area in which many aspects of protein binding site curation and analysis can benefit from fine-grained local similarity computations. For example, it should be the case that clever selection of conformational variants for a particular binding site that *cover* the space of known variations offer a close to optimal choice of protein structures for multi-structure docking.³³

Acknowledgments

The authors gratefully acknowledge NIH for partial funding of the work (grants GM070481 and GM101689). Dr. Jain has a financial interest in BioPharmics LLC, a biotechnology company whose main focus is in the development of methods for computational modeling in drug discovery. Certara L.P. has exclusive commercial distribution rights for the Surflex platform, licensed from BioPharmics LLC. The authors are also indebted to Marguerite Sheffer for assisting in the final manuscript review.

References

1. Spitzer R, Cleves AE, Jain AN. Surface-Based Protein Binding Pocket Similarity. *Proteins: Structure, Function, and Bioinformatics*. 2011; 79:2746–2763.
2. Kinnings SL, Jackson RM. Binding site similarity analysis for the functional classification of the protein kinase family. *Journal of Chemical Information and Modeling*. 2009; 49:318–329. [PubMed: 19434833]
3. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape variation in protein binding pockets and their ligands. *Journal of Molecular Biology*. 2007; 368:283. [PubMed: 17337005]
4. Hoffmann B, Zaslavskiy M, Vert JP, Stoven V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC bioinformatics*. 2010; 11:99. [PubMed: 20175916]
5. Yeturu K, Chandra N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*. 2008; 9:543. [PubMed: 19091072]
6. Xie L, Xie L, Bourne PE. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*. 2009; 25:i305–i312. [PubMed: 19478004]
7. Dessailly BH, Lensink MF, Orengo CA, Wodak SJ. LigASite: A database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Research*. 2008; 36:D667–D673. [PubMed: 17933762]
8. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009; 10:421. [PubMed: 20003500]
9. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*. 1998; 11:739–747. [PubMed: 9796821]
10. Xie L, Bourne P. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics*. 2007; 8:S9. [PubMed: 17570152]
11. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *Journal of Medicinal Chemistry*. 2007; 50:726–741. [PubMed: 17300160]
12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research*. 2000; 28:235–242. [PubMed: 10592235]
13. Jain AN. Morphological Similarity: A 3D Molecular Similarity Method Correlated with Protein-Ligand Recognition. *J Comput Aided Mol Des*. 2000; 14:199–213. [PubMed: 10721506]
14. Jain AN. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem*. 2003; 46:499–511. [PubMed: 12570372]
15. Welch W, Ruppert J, Jain AN. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol*. 1996; 3:449–62. [PubMed: 8807875]
16. Ruppert J, Welch W, Jain AN. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci*. 1997; 6:524–33. [PubMed: 9070435]
17. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*. 1998; 11:739–747. [PubMed: 9796821]
18. Witte G, Hartung S, Buttner K, Hopfner KP. Structural Biochemistry of a Bacterial Checkpoint Protein Reveals Diadenylate Cyclase Activity Regulated by DNA Recombination Intermediates. *Molecular Cell*. 2008; 30:167–178. [PubMed: 18439896]
19. Legler PM, Kumaran D, Swaminathan S, Studier FW, Millard CB. Structural Characterization and Reversal of the Natural Organophosphate Resistance of a D-Type Esterase, *Saccharomyces cerevisiae* S-Formylglutathione Hydrolase. *Biochemistry*. 2008; 47:9592–9601. [PubMed: 18707125]
20. Pelletier I, Altenbuchner J. A bacterial esterase is homologous with non-haem haloperoxidases and displays brominating activity. *Microbiology (Reading, Engl)*. 1995; 141(2):459–468.
21. Hofmann B, Tölzer S, Pelletier I, Altenbuchner J, Van Pee K, Hecht H. Structural investigation of the cofactor-free chloroperoxidases. *Journal of molecular biology*. 1998; 279:889–900. [PubMed: 9642069]

22. Boškin A, Tran CD, Franko M. Oxidation of organophosphorus pesticides with chloroperoxidase enzyme in the presence of an ionic liquid as co-solvent. *Environmental Chemistry Letters*. 2009; 7:267–270.
23. Huang L, Hung L, Odell M, Yokota H, Kim R, Kim SH. Structure-based experimental confirmation of biochemical function to a methyltransferase, MJ0882, from hyperthermophile *Methanococcus jannaschii*. *Journal of Structural and Functional Genomics*. 2002; 2:121–127. [PubMed: 12836702]
24. Sun W, Xu X, Pavlova M, Edwards AM, Joachimiak A, Savchenko A, Christendat D. The crystal structure of a novel SAM-dependent methyltransferase PH1915 from *Pyrococcus horikoshii*. *Protein Science*. 2007; 14:726–741.
25. Purta E, O'Connor M, Bujnicki JM, Douthwaite S. YccW is the m5C Methyl-transferase Specific for 23S rRNA Nucleotide 1962. *Journal of Molecular Biology*. 2008; 383:641–651. [PubMed: 18786544]
26. Davies AM, Tata R, Agha R, Sutton BJ, Brown PR. Crystal structure of a putative phosphinothricin acetyltransferase (PA4866) from *Pseudomonas aeruginosa* PAC1. *Proteins: Structure, Function, and Bioinformatics*. 2005; 61:677–679.
27. Davies AM, Tata R, Beavil RL, Sutton BJ, Brown PR. l-Methionine Sulfoximine, but Not Phosphinothricin, Is a Substrate for an Acetyltransferase (Gene PA4866) from *Pseudomonas aeruginosa*: Structural and Functional Studies. *Biochemistry*. 2007; 46:1829–1839. [PubMed: 17253769]
28. Juan ECM, Hoque MM, Shimizu S, Hossain MT, Yamamoto T, Imamura S, Suzuki K, Tsunoda M, Amano H, Sekiguchi T. Structures of *Arthrobacter globiformis* urate oxidase-ligand complexes. *Acta Crystallographica Section D: Biological Crystallography*. 2008; 64:815–822.
29. Shigeura HT, Gordon CN. Hadacidin, a new inhibitor of purine biosynthesis. *Journal of Biological Chemistry*. 1962; 237:1932–1936. [PubMed: 13911878]
30. Gifford LK, Carter LG, Gabanyi MJ, Berman HM, Adams PD. The Protein Structure Initiative Structural Biology Knowledgebase Technology Portal: A Structural Biology Web Resource. *Journal of Structural and Functional Genomics*. 2012; 13:57–62. [PubMed: 22527514]
31. Bowman GR, Beauchamp KA, Boxer G, Pande VS. Progress and challenges in the automated construction of Markov state models for full protein systems. *The Journal of Chemical Physics*. 2009; 131:124101. [PubMed: 19791846]
32. Spitzer R, Jain AN. Surflex-Dock: Docking benchmarks and real-world application. *Journal of Computer-Aided Molecular Design*. 2012; 26:687–699. [PubMed: 22569590]
33. Jain AN. Effects of Protein Conformation in Docking: Improved Pose Prediction Through Protein Pocket Adaptation. *Journal of Computer-Aided Molecular Design*. 2009; 23:355–374. [PubMed: 19340588]

Figure 1.

Aligned binding sites of two tRNA synthetases. *E. coli* lysyl tRNA synthetase (PDB code 1E24) is shown in pink, with *H. sapien* glycyl-tRNA synthetase (2ZT7) in teal. The binding site surfaces were aligned using PSIM, with the bound ATP molecules (purple and green) defining the binding site scopes. The thin sticks highlight regions of positive, negative, and steric similarity between the two proteins (blue, red, and green respectively).

Figure 2.

The alignment procedure of Surfex-PSIM. *A*: 1E24-LysU in purple and 2ZT7-GlyRS in teal, out of alignment with colored spheres representing the observation points used to measure their binding site surfaces. *B*: A transformation is found which optimally transfers the observations points from their initial locations to similar environments on the other protein. *C*: The alignment used to transform the observation points is applied to the proteins, aligning their binding sites. *D*: Red, blue, and green sticks depict regions of negative, positive, and steric surface similarity, respectively.

Figure 3. PSIM binding site detection. *A*: Bothropstoxin-I (PDB 3CXI) is shown with the bound ligand removed. *B*: Colored spheres indicate discrete regions which passed the concavity filter. *C*: Automatically constructed pseudo-ligands are shown in the concavities identified. Vitamin E, the cognate ligand, is depicted in orange in its bound pose overlapped by the yellow pseudo-ligand (right side).

Figure 4. Holo protein binding site detection. PDB structure 3CXI (Bothropstoxin-I), shown in blue, is bound to vitamin E, shown in orange sticks. A predicted binding site overlapping vitamin E is shown as a transparent surface.

Figure 5.

Apo binding site collapse. The holo (purple) and apo (teal) structures of riboflavin kinase are depicted in ribbons (PDB codes 1N08 and 1N05, respectively). The area occupied by the ligand (ATP) is depicted in orange. Residues which overlap with the binding site region are depicted as sticks in the apo structure.

Figure 6. Correspondence of apo and holo binding sites. The holo form if Bothropstoxin-I bound to vitamin E is shown (3CXI in blue) along with the apo form (3I3I, in purple). The green molecular fragments on the apo structure depict the binding site detected by PSIM in terms of both location and extent.

Figure 7. JCSG protein structures deposited before they were biochemically annotated. The regions identified as putative binding sites by PSIM are shown in colored surfaces.

Figure 8.

Query structure 2FB5. *A*: 2FB5 is shown in purple, along with two screening hits 1DMA (Exotoxin A) and 2UVX (CDK2) in blue. ATP from a later crystallized structure is shown in orange as a guide. *B*: The secondary structure of 2FB5 and 2UVX. *C*: The similarity in the surfaces of 2FB5 and 2UVX is depicted with sticks.

Figure 9. Query structure 1PV1. *A*: 1PV1 is shown in purple, aligned with 1A8U (chloroperoxidase T), in blue. The ligand of 1A8U (benzoic acid) is shown in orange, and the catalytic triad (Asp-Ser-His) is depicted in balls and sticks. *B*: The secondary structures of 1PV1 and 1A8U are depicted with the binding site occupied by an orange surface.

Figure 10.

Query structure 1DUS. *A*: 1DUS is shown in purple aligned with 1AQJ (DNA-methyltransferase TaqI) in blue. The ligand of 1AQJ (sinefungin) is shown in orange sticks. *B*: The secondary structures of 1DUS and 1AQJ are shown with their binding sites identified (orange surface).

Figure 11.

Query structure 1AS0. *A*: 2AS0 is shown in purple aligned to 2DPM (adenine-specific methyltransferase DpnII) in blue. *B*: The secondary structures of 2AS0 and 2DPM are shown with the binding sites identified (orange surface).

Figure 12.

Query structure 1YVO. *A*: 1YVO is shown in purple aligned to 2C27 (acetyltransferase, specifically a mycothiol synthase) in blue. *B*: The secondary structures of 1YVO and 2C27 are shown with the binding site identified (orange surface).

Figure 13. Twelve proteins retrieved from the JCSG which are currently uncharacterized. Binding sites predicted by PSIM appear as colored surfaces.

Figure 14. Possible function for PDB structure 2GNX. *A*: 2GNX (purple) is shown aligned to 2YZD (blue). The ligand of 2YZD is 8-azaxanthine (orange), and enzyme involved in purine biosynthesis. *B*: The ligand is surrounded by PSIM similarity sticks, depicting a similar steric (green sticks) and polar (red and blue sticks) environment.

Table 1

Screening results for 2FB5

Screen Hit	PSIM p-value	CE p-value	SMAP p-value	BLASTP p-value	Depos. Date	Protein Name	Protein Function
IDMA	1.1e-07	5.1e-01	7.9e-02	1.6e+00	1995	Exotoxin A	ADP ribosylation
2UVX	8.0e-11	9.4e-01	8.2e-01	-	2007	CDK2	ATP Dep. Phosphorylation

Table II

Screening results for 1DUS organized by PSIM p-value

Screen Hit	PSIM p-value	CE p-value	SMAP p-value	BLASTP p-value	Depos. Date	Protein Name	Protein Function
1G55	4.0e-11	1.3e-03	1.6e-04	2.6e-02	2000	DNMT2	DNA methyltransferase
1QAO	6.3e-11	8.0e-04	1.9e-05	2.0e-05	1999	ERMC	RNA methyltransferase
1QAQ	1.0e-10	8.0e-04	2.5e-06	2.0e-05	1999	ERMC	RNA methyltransferase
1AQJ	9.2e-10	1.3e-03	7.2e-06	1.3E+00	1996	TAQI	RNA methyltransferase
2HMY	1.0e-09	3.1e-03	5.7e-03	2.5E+00	1999	HHAI	DNA methyltransferase
1QAN	1.1e-09	3.1e-03	3.0e-05	2.0e-05	1999	ERMC	RNA methyltransferase
1F3L	5.8e-08	3.1e-04	9.3e-04	8.0e-05	2000	PRMT3	ARG methyltransferase
1EG2	9.2e-08	1.7e-02	8.8e-04	-	2000	RSRI	DNA methyltransferase
1FP1	2.6e-07	2.0e-03	1.1e-02	2.6e-01	2000	ChOMT	Chalcone O-methyltransferase

Table III

Summary of screening results for 12 unknowns.

Query Protein	Number of Sites	Hit Count $p < 1.0e-07$	Hit Count $p < 1.0e-03$
2GNX	12	34	1652
1G2R	2	1	233
1QZ4	7	18	860
1RYL	3	2	266
1ZSO	5	1	461
2ICU	7	7	539
2IN5	14	4	787
2Q0X	13	12	1207
2QZB	6	15	599
3C4R	10	115	1536
3L6I	7	8	796
3O12	4	1	695