7-7-2012

Automated identification of binding sites forphosphorylated ligands in protein structures

Dario Ghersi
University of Nebraska at Omaha, dghersi@unomaha.edu

Roberto Sanchez
Mount Sinai School of Medicine

Automated identification of binding sites for phosphorylated ligands in protein structures

By: Dario Ghersi and Roberto Sanchez

Abstract

Phosphorylation is a crucial step in many cellular processes, ranging from metabolic reactions involved in energy transformation to signaling cascades. In many instances, protein domains specifically recognize the phosphogroup. Knowledge of the binding site provides insights into the interaction, and it can also be exploited for therapeutic purposes. Previous studies have shown that proteins interacting with phosphogroups are highly heterogeneous, and no single property can be used to reliably identify the binding site. Here we present an energy-based computational procedure that exploits the protein three-dimensional structure to identify binding sites involved in the recognition of phosphogroups. The procedure is validated on three datasets containing more than 200 proteins binding to ATP, phosphopeptides, and phosphosugars. A comparison against other three generic binding site identification approaches shows higher accuracy values for our method, with a correct identification rate in the 80–90% range for the top three predicted sites. Addition of conservation information further improves the performance. The method presented here can be used as a first step in functional annotation or to guide mutagenesis experiments and further studies such as molecular docking.

INTRODUCTION

Phosphorylated molecules play a vital role in a wide range of biological processes, both in prokaryotic and eukaryotic organisms. The phosphate group is used with remarkable versatility by the cell to store energy and to reversibly modify proteins in signaling cascades. Besides proteins and nucleotides, another class of biomolecules that can undergo phosphorylation is represented by sugars, either as intermediates in metabolic processes or as signaling tags that are attached to proteins.

Despite the fact that no rigid classification is possible, we can approximately distinguish between phosphorylation as a means to energetically activate metabolic intermediates or products and phosphorylation as a marker or switch in cell signaling. In the latter case, the addition of the phosphogroup is in some instances capable by itself of inducing conformational changes in proteins or otherwise autonomously driving biochemical processes, but in many cases, a specific decoding process has to take place. Protein domains that specifically recognize the phosphogroup in proteins, sugars, or nucleotides, usually carry out the decoding process.

Significant effort has gone into the characterization of these phospholigand recognition domains, because of their importance for understanding fundamental biological processes coupled with their potential therapeutic exploitation. Historically, the SH2 domain was the first to be discovered[1] as a protein module capable of binding to its cognate ligand in a phosphorylation-dependent manner, with other new domains being identified over the years. As phosphorylation occurs in such a diverse range of contexts, it is not surprising that the domains involved in its selective recognition are oftentimes unrelated from an evolutionary or structural standpoint. Despite this diversity, studies have tried to

identify some of the properties that may be common among all the domains that recognize their cognate ligands in a phosphorylation-dependent manner. A method that can reliably pinpoint the region of a binding site where the phosphate recognition takes place, can be useful to guide mutagenesis experiments, or as a step in functional annotation.

A study by Joughin *et al.***2** focused on recognition of phosphorylated residues in peptides. The authors collected three-dimensional (3D) structures of seven phosphopeptide-binding domains, extracting properties such as amino acid identity, surface curvature, and electrostatic potential, in order to characterize the phosphopeptide-binding region with respect to the whole of the protein surface. The propensities for each of these properties were combined into one joint propensity that was then mapped back on the protein surfaces and used for visual identification of the regions likely to be involved in binding. An important conclusion from this work was that the process of phosphate recognition could not be fully captured by a single property such as the electrostatic potential (in fact in many instances the binding site was not the region of most positive electrostatic potential on the protein surface) or aminoacidic composition.

In the present study, the recognition of all three classes of phosphomodifications (on peptides, nucleotides, and sugars) is considered regardless of whether the phosphogroup has been added for metabolic purposes or as a signaling switch, and we investigate whether it is possible to find a single structure-derived property that has sufficient discriminative power to confidently identify most of the binding sites. This problem is tackled here by detecting energetically favorable regions on the protein surface, along the lines of what has been previously done in binding site identification for drug-like ligands.**3**, **4** An important difference with the identification of binding sites for drug-like ligands is that, in the case of phosphate recognition most of the interaction energy does not come from the van der Waals term, which is what most energy-based approaches for binding site identification exploit. Therefore, different energy maps have to be employed in order to precisely identify the region of the binding site responsible for the selective recognition of the phosphogroup. Furthermore, we investigate whether including evolutionary information in the form of a per-residue conservation score derived from multiple sequence alignments of protein families can further improve the identification of the residues involved in recognition of the phosphogroup. More than 200 phospholigand-binding proteins were included in the study.

MATERIALS AND METHODS

Datasets

Phosphopeptide binding sites

All the crystal structures in the Protein Data Bank (PDB)**5** downloaded on November 28, 2008, were collected, and filtered for the presence of at least one phosphoresidue as indicated by the names “PTR” (phosphotyrosine), “SEP” (phosphoserine), or “TPO” (phosphothreonine). To determine whether these phosphoresidues were participating in protein–protein interactions (i.e., the PDB entry contains at least one phosporesidue binding site) all residues within 5.0 Å of a phosphoresidue were extracted and their chain identifier was recorded. All the cases where the chain identifier of the phosphoresidue and that of all interacting residues were identical were discarded. The sequences corresponding to the chains

interacting with the phosphoresidue were designated as containing a phosphopeptide binding site. Redundancy in this set was removed by clustering the sequences using BLASTCLUST**6** with a sequence identity cutoff of 50%. From each cluster, the highest resolution structure was selected for the final dataset. This procedure yielded a total of 48 different chains. Five of the peptides bound to the clustered chains (PDB codes 1j4x, 1p22, 1u7f, 2oq1, and 2z8p) contained two phosphorylated residues, and each residue was treated independently in the analysis, for a total of 53 different phosphoresidue binding sites. We note that the choice of the BLASTCLUST sequence identity cutoff represents a compromise between the number of sequences that pass the filter and the diversity of the dataset. The distribution of Pfam**7** domains (Supporting Information Table S1) suggests that the dataset is fairly balanced and diverse.

A corresponding dataset of unbound phosphopeptide-binding proteins was also generated by carrying out a BLAST search (with standard parameters and an expectation value of 10^{-6}) using the bound chain sequences as queries against the entire PDB. Hits with sequence identity or coverage less than 95% with respect to the query sequence were excluded. The structures that did not have a "TPO," "SEP," or "PTR" residue were retained. Finally, the crystal structures with an empty binding site, with the highest coverage and the highest resolution (in this order of preference) were retained. This protocol yielded a total of 29 unbound proteins. Four of these proteins corresponded to structures bound to double-phosphorylated peptides, and each binding site was treated independently as for the bound forms, resulting in a total of 33 different phosphoresidue binding sites.

ATP binding sites

To build a diverse dataset of ATP binding proteins we resorted to the sc-PDB**8** database (2008 version), a collection of biologically relevant protein-small molecules complexes. We selected all the proteins in complex with ATP whose binding site was made of a single chain and clustered the sequences at 50% identity to remove redundancy. From each cluster the highest resolution structure was extracted. This protocol yielded a total of 70 different proteins containing a single ATP binding site each. To build a corresponding dataset of unbound structures we followed the protocol described above for the phosphopeptides, yielding a total of 33 proteins with a single ATP binding site each. Complexes with other phosphorylated nucleotides (CTP, GTP, TTP) yielded a much smaller number of cases and were not included in the dataset (See Supporting Information Table S5). Supporting Information Figure S4 and Table S2 show the amino acid distribution in the binding sites and the Pfam**7** domain distribution of the ATP binding dataset.

Phosphosugar binding sites

All the ligands in the sc-PDB**8** database were compared against the phosphosugar α-D-glucose-6-phosphate using the Tanimoto coefficient, computed using Pybel.**9** The ligands with a Tanimoto coefficient = 0.7 were manually inspected to ensure they were phosphorylated sugars, and the corresponding PDB entries were retained. To remove redundancy and gather a corresponding unbound dataset, the same procedure outlined above for the phosphopeptides and ATP datasets was followed, yielding a total of 29 bound and 17 unbound proteins (see Supporting Information Table S4) containing a single phosphosugar binding site each. Supporting Information Figure S4 and Table S3 show the amino acid distribution in the binding sites and the Pfam**7** domain distribution of the phosphosugar binding dataset.

Molecular interaction field calculations

The Molecular Interaction Field (MIF) calculations were performed with our program EasyMIFs,**10** which uses the nonbonded component of the GROMOS force field (G43b1, in vacuo) as made available in the GROMACS**11** package with a distance-dependent dielectric derived from Solmajer and Mehler.**12** The program computes the potential energy between a chemical probe (represented by a particular atom type) and the protein on a regularly spaced grid, using the following equation:

$$V_i = \sum (V_{\mathrm{LJ}}(r_{ij}) V_{\mathrm{E}}(r_{ij}))$$

where the potential energy calculated for a probe at a point *i* in the grid is equal to the sum of a Lennard-Jones and an electrostatics term over all the atoms of the protein. *r ij*represents the distance between the probe at point *i* in the grid and an atom *j* of the protein. The Lennard-Jones and the electrostatics term are expressed by the following two equations:

$$V_{\mathrm{LJ}}(r_{ij}) = \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \quad \text{and} \quad V_{\mathrm{E}}(r_{ij}) = \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}}$$

The $C^{(12)}$ and $C^{(6)}$ parameters in the Lennard-Jones term depend on the chosen probe and the particular atom type and are taken from a matrix of LJ-parameters distributed with the GROMACS package. The dielectric constant $\frac{1}{4\pi\varepsilon_0}$ has been set to 138.935485, as done in the GROMACS package and reported in the GROMACS manual.**13** The distance-dependent dielectric sigmoidal function has been taken from Solmajer and Mehler**12** and has the following form:

$$\varepsilon(r_{ij}) = A + \frac{B}{1 + ke^{-\lambda B r_{ij}}}$$

where *A* = 6.02944; *B* = *e*0 – *A*; *e*0 = 78.4; λ = 0.018733345; *k* = 213.5782. When the distance between the probe and an atom becomes less than 1.32 Å, a dielectric constant of 8 is used. The parameters reported above for the distance-dependent dielectric have been taken from Cui *et al*.**14**

Binding site identification

The binding site identification protocol identifies regions near the protein surface where the interaction with the phosphate oxygen (OP) is particularly favorable, as defined by very negative values of the interaction energy. In order to identify those favorable regions the program SiteHound**10** was used. The program carries out the following four steps (Fig. 1): (i) The MIF generated by EasyMIFs is read in and filtered by retaining only the points that are below a predefined energy threshold (*e*). (ii) The remaining points are clustered based on their position in space with an agglomerative hierarchical clustering algorithm using average linkage. (iii) The resulting dendrogram is cut into non-overlapping clusters by applying a distance cutoff (*d*). (iv) Finally, the clusters are ranked by Total Interaction Energy (TIE), the sum of the energy values of all the points that belong to the same cluster. Only two parameters must be chosen, the energy threshold *e* and the distance cutoff *d*. A grid search for values of *e* and *d* on 25 randomly selected bound structures from the dataset was carried out, with *e* ranging from -9 to -7.5 KJ/mol and *d* from 6.5 to 8 Å, with incremental steps of 0.1. The combination *e* = −8.5 kJ/mol and *d* = 7.8 Å, yielded a good compromise between coverage of cases and accuracy of the prediction.

The software for all computations is freely available at http://sitehound.sanchezlab.org. A web server**15** that automatically carries out the binding site identification procedure, using standard parameters, is also available.

Conservation-based reranking of putative binding sites

The sequences corresponding to the structures in the datasets were extracted from the PDB files. For each sequence a BLAST**6** search was run on the "nr" (nonredundant) database, downloaded in November 2008 from http://www.ncbi.nlm.nih.gov/, and the hits with an *E*-value = 10^{-4} and a coverage = 90% were retained. Subsequently, a multiple sequence alignment (MSA) was constructed for each set of homologs (as defined by the BLAST *E*-value cutoff) using ClustalW**16** with default parameters. The conservation of each column in the MSA was measured using the Jensen-Shannon divergence score (JSD), as described in Capra and Singh.**17** The calculations of the JSD score were performed using the Python program named "conservation_code" available at http://compbio.cs.princeton.edu/conservation. The top five sites identified by SiteHound were sorted using the average of their per-residue conservation scores from the most conserved to the least conserved site.

Assessment of the prediction accuracy

The clusters generated in the binding site identification step are used to identify the residues that are in contact with them by applying an arbitrarily chosen distance cutoff of 4 Å. The groups of residues that contribute to each cluster make up the predicted binding sites and are directly compared with the residues that are within 5 Å of any phospholigand atom in the complexes (or the corresponding residues in the unbound form). The subset of residues in contact with the phosphogroup only were also considered as a more specific test set. Binding site identification can therefore be converted into a classification problem, where the task is to determine whether a given residue is involved in binding or not. The accuracy of the predictions was measured using the Pearson correlation coefficient between the Prediction (P) and the Reference (R). As shown by Baldi *et al.*,**18** the Pearson correlation coefficient for a classifier can be conveniently expressed by using the total number of residues (*N*), the True Positives (TP), the True Negatives (TN), the False Positives (FP), and the False Negatives (FN) with the following equation:

$$C(\mathrm{P},\mathrm{R}) = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

C(P,R) is better known as the Matthews correlation coefficient (MCC).**19** The MCC can be directly related to a χ^2 test applied to the 2 × 2 contingency matrix containing the TP, TN, FP, and FN by using the following equation**18**:

$$\chi^2 = N \times \mathrm{MCC}^2$$

The average size of the proteins in the dataset is 218, 346, and 380 residues for phosphopeptide, ATP, and phosphosugar binding proteins, respectively. An MCC of 0.3 in this case would correspond to p-values smaller than 10^{-4}. Thus a binding site with an MCC = 0.3 was considered as correctly identified.

Comparison with other binding site identification approaches

Two energy-based approaches (Q-SiteFinder**4** and i-Site**20**), a well-established pocket identification program (LigSite**21**), a peptide binding site detection method (PepSite**22**), and a recent phospholigand-binding site identification approach (Phosfinder**23**) were chosen for comparison against SiteHound.

Q-SiteFinder, LigSite, PepSite, and Phosfinder were run from their respective webservers with default parameters, whereas i-Site was run locally following the directions provided in the package.

As Q-SiteFinder and i-Site output clusters ranked by interaction energy (as SiteHound does), the extraction of the putative binding residues was performed in the same manner as for SiteHound. In contrast, LigSite outputs the centers of the predicted pocket, from which the residues within a sphere of predefined radius were extracted and considered as putative binding residues (as done on the LigSite webserver). The default radius used on the webserver (5.0 Å) does yield a very limited number of residues (less than five on average) and in several cases no residues at all (with corresponding very low values of MCC). To ensure a fair comparison against SiteHound, increasing values of the radius were tested and the resulting MCC values recorded. A radius equal to 8 Å yielded the best results. The same approach was taken for PepSite and Phosfinder, yielding an optimal radius of 6 Å and 7 Å, respectively.

RESULTS

Phospholigand-binding site identification

The procedure for identifying phospholigand binding sites relies on the detection of protein surface regions with favorable interaction energy for a OP probe. The output is a list of interaction energy clusters corresponding to putative phospholigand binding sites (Fig. 1), which are ranked according to their total interaction energy. Application of this procedure to the set of phospholigand-binding proteins showed that, depending on the type of phospholigand, the top ranking cluster corresponds to a known phospholigand binding site in 50–69% of the cases for bound structures (protein structures solved with the ligand in the binding site) and 41–47% of the cases for unbound structures (proteins solved without a ligand in the binding site; Fig. 2 and Table I). These numbers increase significantly, to 85–97% for bound structures and 79–88% for unbound structures, when the top three ranking clusters are considered (Table I). In a few cases, known phospholigand sites ranked outside the top three clusters (Supporting Information Fig. S1). In some cases, these may correspond to very weak interactions; in other cases, this suboptimal prediction can be recovered by adding evolutionary information (see Supporting Information Materials and conservation-based reranking below). The clusters tend to be focused on the region of the ligand that contains the phosphate group (Fig. 2); for example, identifying the position of the phosphorylated residue within a peptide [Fig. 2(A,B)] or the location of the phosphate groups within an ATP-binding site [Fig. 2(E,F)]. This observation already distinguishes this approach from some of the more general methods. Such methods identify binding sites based on the location of clefts or pockets in the protein structure,**30**, **31** but without including chemical information that may distinguish different types of binding sites or different regions within one binding site (this is discussed in more detail in the probe selectivity section). As already pointed out by Joughin *et al.*,**2** the electrostatic potential plays an important role in the interaction between proteins and phospholigands but in a nontrivial way. In other words, the binding site does not necessarily correspond to the most positive patch on the protein surface. A similar situation was observed here in some of the test cases,

confirming that a pure electrostatics-based approach would be ineffective in identifying the phospholigand binding sites.

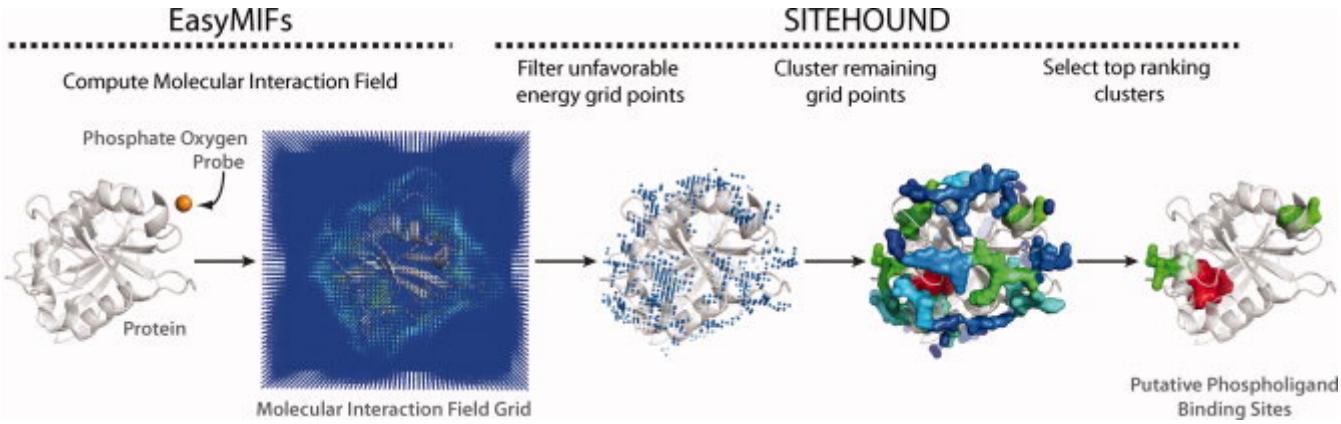

Figure 1.

Procedure for the identification of phospholigand binding sites. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

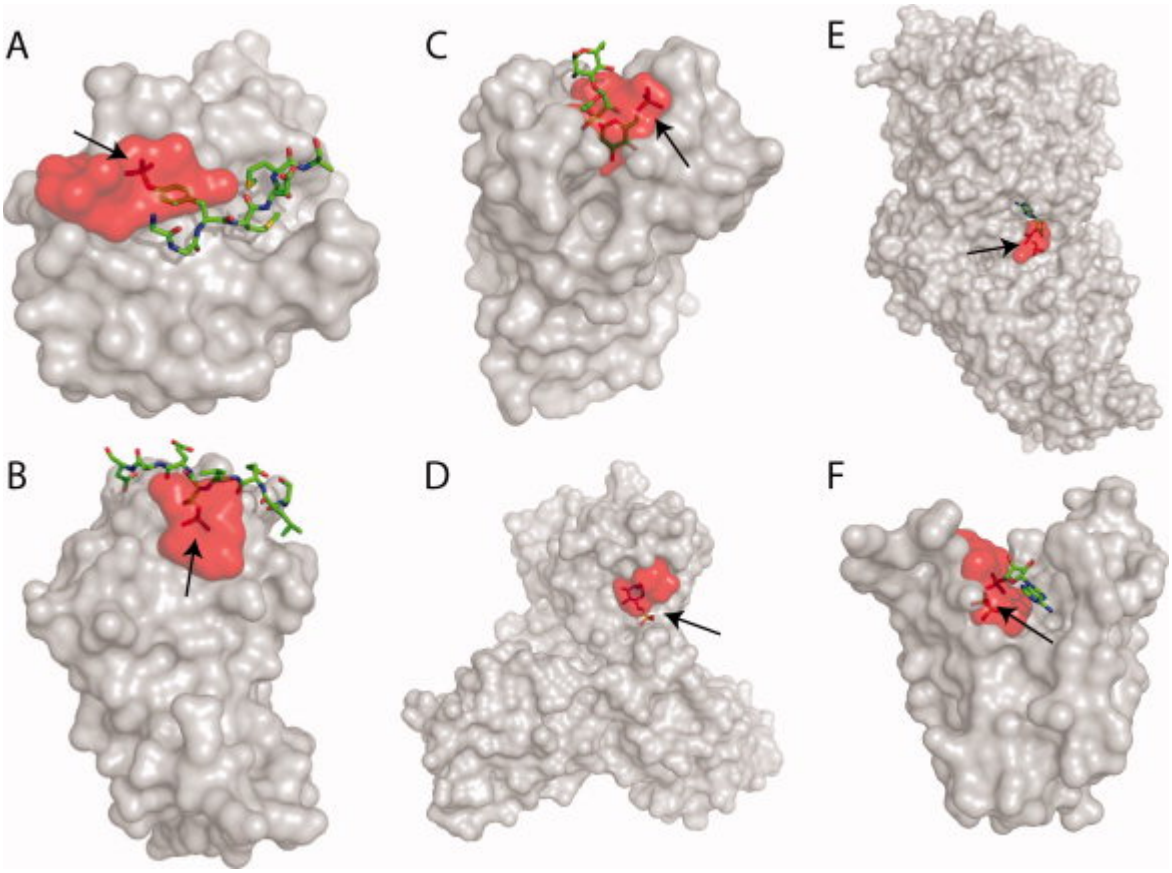

Figure 2.

Examples of phospholigand-binding site identification. Arrows indicate the location of phosphate groups and the top-ranking cluster is shown as a red surface. **A**: SH2 domain bound to phosphopeptide (1ayc).**24 B**: FHA domain bound to phosphopeptide (1g6g).**25 C**: Cation-dependent mannose-6-phosphate receptor bound to pentamannosyl phosphate (1c39).**26 D**: Mannose-6-phosphate receptor bound to mannose-6-phosphate (1sz0).**27 E**: Motor domain of dictyostelium myosin II bound to ATP (1fmw).**28 F**: ATP:corrinoid adenosyltransferase from *Salmonella typhimurium* bound to ATP (1g5t).**29** PDB codes are indicated in parentheses. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table I. Performance for the Top Three Ranking Clusters and for the Top Ranking Cluster Only (With and Without Applying the Conservation-Based Reranking) is Shown as the Number of Correctly Identified Sites Over the Total Number of Proteins

	Without conservation (top 3 sites)		**Without conservation (top site)**		**With conservation (top site)**	
1. The percentage of correctly identified sites is also shown.						
Phosphopeptides (bound)	36/48	75%	24/48	50%	30/48	63%
Phosphopeptides (unbound)	22/29	76%	12/29	41%	21/29	72%
Phosphosugars (bound)	25/29	86%	20/29	69%	22/29	76%
Phosphosugars (unbound)	14/17	82%	8/17	47%	10/17	59%

	Without conservation (top 3 sites)		Without conservation (top site)		With conservation (top site)	
ATP (bound)	58/70	83%	39/70	56%	39/70	56%
ATP (unbound)	27/33	82%	15/33	45%	17/33	52%

Statistically, the performance of the binding-site identification was assessed by treating it as a classification problem, where the objective is to discriminate between the residues that are in contact with the phospholigand versus the ones that are not. The well-established MCC was used to quantify the agreement between the predictions and the actual interacting residues derived from the crystal structures. In this way, it is possible to use the same performance measure to compare bound and unbound forms, even in the presence of conformational changes, without the need for superposition. An MCC value of 0.3 was chosen as the limit for discriminating partially correct predictions from wrong ones, as this corresponded to a good visual match with the known binding sites, and is equivalent to *P*-values smaller than 0.0001 for the typical protein size in our datasets (see Materials and Methods). The distribution of the MCC and the rank of the best prediction (cluster with highest MCC out of the top five clusters) shows that most sites are detected among the top three ranking clusters (Fig. 3 and Table I), with the majority of them ranking first and with an accuracy significantly above the threshold of MCC = 0.3. On average, the approach seems to perform better on the phosphosugar and ATP datasets than on the phosphopeptide dataset. Computing the ratio between the average interaction energy on a 5-Å shell in the binding site and the average interaction energy on a 5-Å shell surrounding the entire protein surface identified a potential explanation for this behavior. As expected, both phosphosugars and ATP binding sites showed larger values, and, therefore, a stronger signal than the one deriving from the phosphopeptides binding sites (Supporting Information Fig. S2). This observation could be related to the fact that many protein–peptide interactions correspond to complexes with relatively low affinity, while ATP or phosphosugars will be found interacting more often with receptors and enzymes, which tend to be stronger interactions.

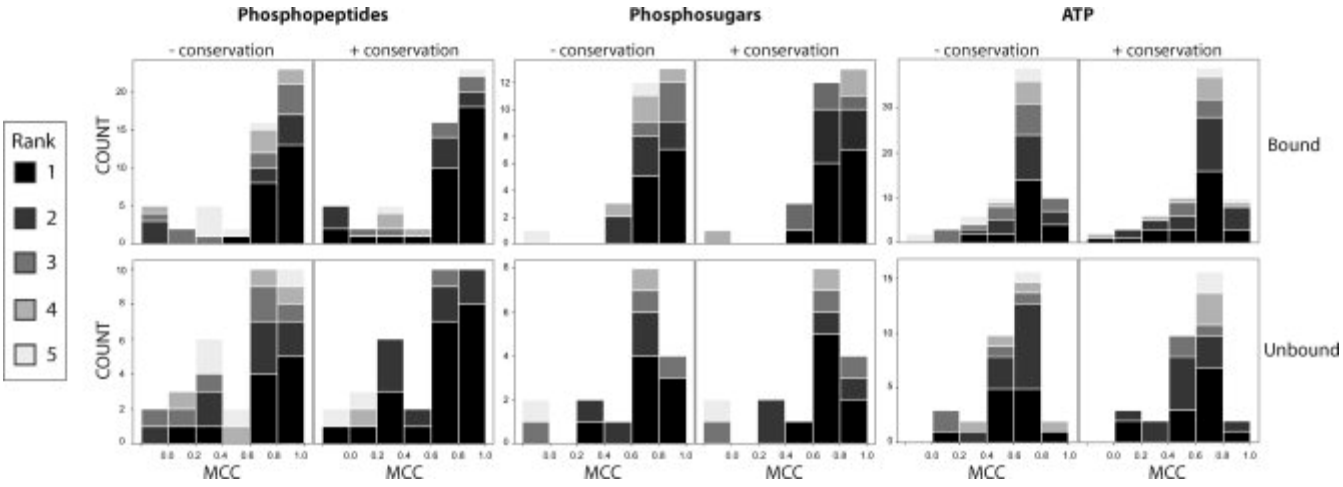

Figure 3.

Accuracy of binding-site identification. MCC distribution for the phospholigand datasets for bound and unbound proteins with and without conservation-based reranking. The stacked bars show the rank of the best prediction, color coded from 1 to 5.

Bound versus unbound structures

Binding-site identification methods tend to be evaluated on protein structures solved in the presence of the ligand (bound structures). However, by definition, binding site identification will be used in practice only in cases where the binding site is unknown, and, therefore, the structure of the protein has been solved in the absence of the ligand (unbound structures). Therefore, it is important to compare the performance of a binding-site identification method on bound and unbound structures. Table I and Figure 3 show that while there is a drop in performance when going from bound to unbound structures it is not large. The largest drop in performance was observed when attempting to identify the binding site using only the top-ranking cluster. In this case, 68–82% of the performance obtained with bound structures is retained, with the phosphosugar set being affected the most. When using the top three clusters, 84–99% percent of the performance obtained with bound structures is retained, again the phosphosugars being the most affected. In both cases the ATP and phosphopeptides sets were affected to a much smaller degree (80% and 82% retention for top site and 99% and 93% retention for top three sites, respectively). This loss of accuracy may be partially explained by conformational changes occurring on ligand binding. When comparing the bound and unbound structures in the three datasets, the largest differences were observed in the phosphosugar set, followed by the ATP set, and then the phosphopeptides (data not shown).

Conservation-based reranking of the putative sites

The fact that a few known sites are missed by the energy-based approach and the effect of ligand-induced conformational changes described in the previous section, prompted us to explore the possible complementarity between the physical information derived from the energy-based approach and the evolutionary information that can be derived from the conservation of residues in a multiple sequence alignment. Approaches based on conservation have been shown to be able to identify functionally important residues**17**, **32** and can complement structure-derived information.**21**, **33** Conservation-based

reranking of the top five clusters identified using the energy-based approach resulted in an increase in the number of known phospholigand binding sites identified by the top cluster (Table I). In general, the accuracy and ranking of the cluster prediction improved (Fig. 3), with the largest effect being observed for the sets of unbound structures and particularly for phosphopeptides. In other words, the confidence on the top predictions is increased when conservation-based reranking is applied. While conservation-based reranking provided a clear boost to the binding site identification approach, conservation information by itself cannot identify the residues that are specifically involved in binding with the same level of accuracy afforded by the energy-based approach, since residues tends to be conserved both for structural and functional reasons. In other words, the residues that are specifically involved in binding usually are a subset of the conserved surface residues (Supporting Information Fig. S3). Hence, overall conservation-based identification of phospholigand binding sites is less accurate than the energy-based approach, but it provides complementary information (Fig. 4), especially for weaker binding sites such as those of phospholigands.

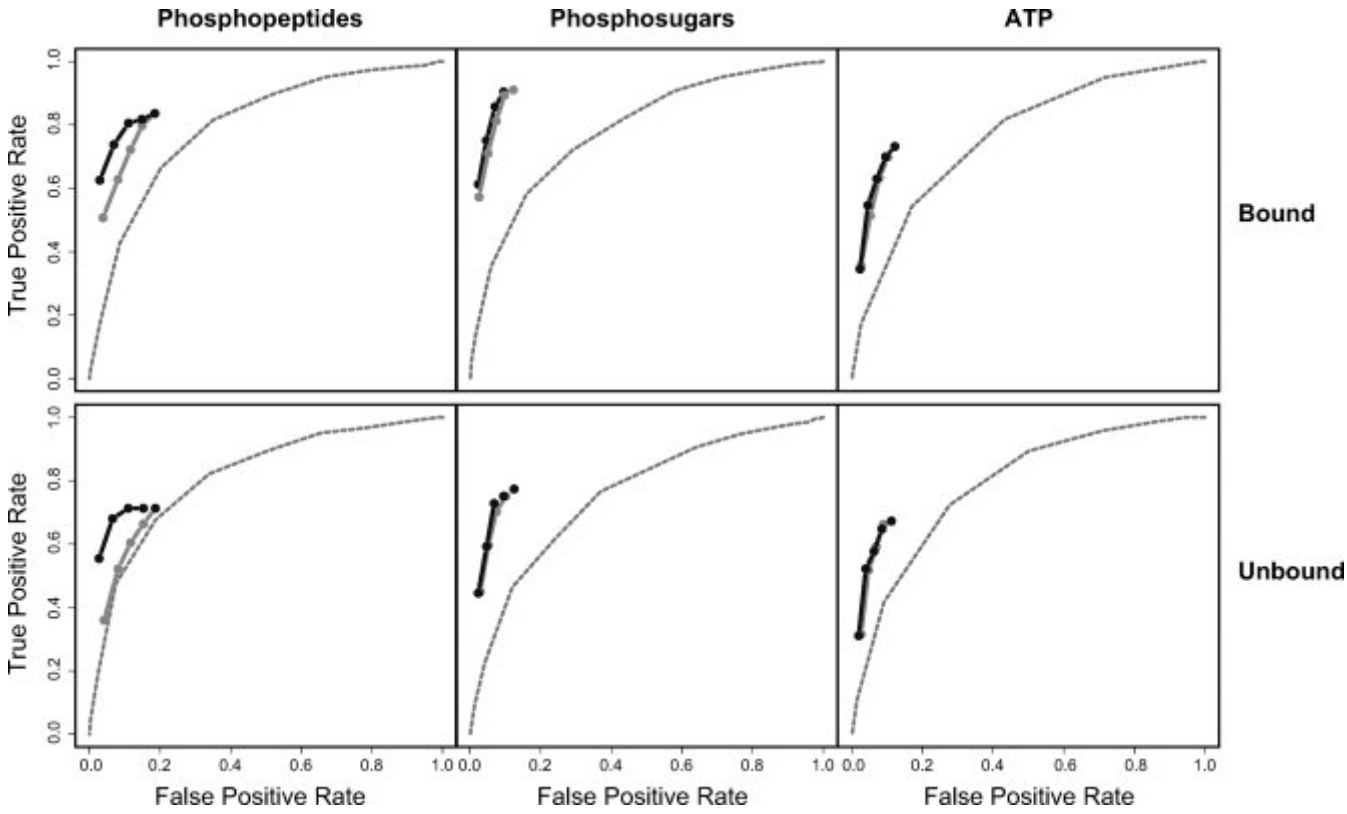

Figure 4.

Summary of the performance. Receiver operator characteristic curves comparing the performance of the energy-based approach used in isolation (solid gray curve), the conservation-based approach used in isolation (dashed gray curve), and the combination of the energy-based and conservation-based approaches (black curve). A curve closer to the upper left corner indicates better performance.

Probe selectivity

One of the claimed advantages of an energy-based approach to the identification of ligand-binding sites is that chemical information can be built into the procedure by selecting different probes for binding-site identification.**10** Hence, the performance of the OP probe was compared with the performance of a chemically different reference probe to examine whether it really provides an advantage in phospholigand-binding site identification. The methyl (CMET) probe was used as a reference since we and others have used it before for general binding site identification.**3**, **4** The CMET probe is also an interesting reference since it mimics geometrical approaches to binding site identification in which the dominant component of the energy comes from van der Waals interactions.**3**, **4** The comparison showed that the OP probe does indeed provide better accuracy than the CMET probe for phospholigand-binding site identification in all three sets (Table II). The largest difference was observed in the phosphopeptides set where the OP probe improves binding-site detection by 72% and 46% for bound and unbound structures, respectively. This is probably due to the fact that peptide-binding sites are more extended and less curved than ATP and phosphosugar binding sites, thus making them more difficult to identify by methods that rely purely on van der Waals interactions or geometry. This advantage is also illustrated by the fact that in the ATP and phosphosugars sets the gains due to the use of the OP probe are greater in the unbound structures than the bound ones, probably due to the larger effect that conformational changes will have on a purely van der Waals-based approach.

As ATP has a large fraction of the molecule composed of phosphogroups, the subset of residues that are in contact with the phosphogroups were also considered as a subsite (thereby excluding the region of the binding site that binds to the nucleotide part of the ligand). In this way, one can directly assess whether it is possible to discriminate the part of the binding site that binds to the phosphogroups versus the one that binds to the nucleotide part of the ligand. The gain provided by the OP probe is in fact larger when restricting the comparison to the phosphogroup subset of ATP (Table II), thus suggesting that the CMET and OP probes can be used in a complementary fashion, as illustrated in Figure 5. Overall, the results indicate that the performance with the CMET probe is inferior to the one achieved by using the OP. In other words, there is an advantage in using the more selective OP probe when studying proteins that are known to bind to phosphorylated ligands.

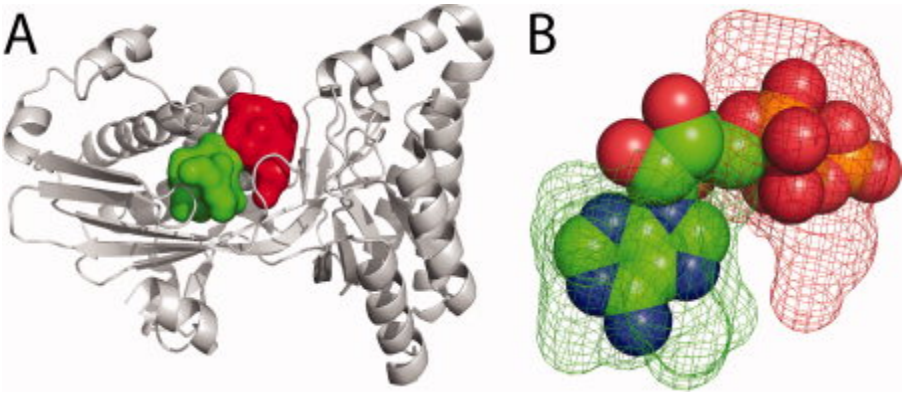

Figure 5.

Table II. Comparison of OP and CMET Probes for Phospholigand-Binding Site Identification

	OP		CMET		OP vs. CMET enrichment	MCC OP (Median)	MCC CMET (Median)
1. The number of cases with an MCC ≥ 0.3 in at least one of the top three clusters and the median value of the MCC are listed for all three datasets.							
Phosphopeptides (bound)	36/53	68%	14/53	26%	157%	0.83	0.65
Phosphopeptides (unbound)	22/33	67%	11/33	33%	100%	0.74	0.54
Phosphopeptides phosphogroup (bound)	35/53	66%	13/53	25%	169%	0.72	0.59

	OP		CMET		OP vs. CMET enrichment	MCC OP (Median)	MCC CMET (Median)
Phosphopeptides phosphogroup (unbound)	22/33	67%	9/33	27%	144%	0.68	0.52
Phosphosugars (bound)	25/29	86%	22/29	76%	14%	0.80	0.72
Phosphosugars (unbound)	14/17	82%	6/17	35%	133%	0.75	0.70
Phosphosugars phosphogroup (bound)	22/29	76%	19/29	66%	16%	0.65	0.55

	OP		CMET		OP vs. CMET enrichment	MCC OP (Median)	MCC CMET (Median)
Phosphosugars phosphogroup (unbound)	12/17	71%	6/17	35%	100%	0.56	0.48
ATP (bound)	58/70	83%	49/70	70%	18%	0.69	0.73
ATP (unbound)	27/33	82%	21/33	64%	29%	0.62	0.63
ATP phosphogroup (bound)	56/70	80%	44/70	63%	27%	0.76	0.59

	OP		CMET		OP vs. CMET enrichment	MCC OP (Median)	MCC CMET (Median)
ATP phosphogroup (unbound)	27/33	82%	18/33	55%	50%	0.63	0.53

Table III. Comparison of SiteHound (OP probe) Against Other Binding-Site Identification Approaches

Phospholigand			Method											
			SiteHound (OP)		I-Site		QSiteFinder		LigSite		PepSite		Phosfinder	
			Coverage, MCC ≥ 0.3	Median, MCC	Coverage, MCC ≥ 0.3	Median, MCC	Coverage, MCC ≥ 0.3	Median, MCC	Coverage, MCC ≥ 0.3	Median, MCC	Coverage, MCC ≥ 0.3	Median, MCC	Coverage, MCC ≥ 0.3	Median, MCC
Peptides	Bound (53 cases)	Ligand	**68%**	0.83	57%	0.83	47%	0.81	48%	0.69	28%	0.60	49%	0.81
		P-group	**66%**	0.72	51%	0.79	42%	0.75	46%	0.65	21%	0.54	49%	0.80
	Unbound (33 cases)	Ligand	**67%**	0.74	61%	0.75	26%	0.68	58%	0.61	18%	0.42	45%	0.78
		P-group	**67%**	0.68	58%	0.81	39%	0.80	55%	0.56	9%	0.54	52%	0.74
Sugars	Bound (29 cases)	Ligand	**86%**	0.80	72%	0.75	72%	0.75	76%	0.76	—	—	41%	0.74
		P-group	**76%**	0.65	66%	0.59	62%	0.55	62%	0.56	—	—	38%	0.70

1. The percentage of cases with an MCC ≥ 0.3 in at least one of the top three clusters (coverage) and the median value of the MCC for those cases are listed for all three datasets. The results include both the full ligands (ligand) and the phosphogroups only (P-group). The highest binding site identification performance is highlighted for each group of ligands.

Phospholigand			Method											
			SiteHound (OP)		I-Site		QSiteFinder		LigSite		PepSite		Phosfinder	
			Coverage, MCC ≥ 0.3	**Median, MCC**	**Coverage, MCC ≥ 0.3**	**Median, MCC**	**Coverage, MCC ≥ 0.3**	**Median, MCC**	**Coverage, MCC ≥ 0.3**	**Median, MCC**	**Coverage, MCC ≥ 0.3**	**Median, MCC**	**Coverage, MCC ≥ 0.3**	**Median, MCC**
	Unbound (17 cases)	Ligand	82%	0.75	53%	0.67	41%	0.70	**88%**	0.67	—	—	35%	0.66
		P-group	**71%**	0.56	47%	0.49	35%	0.53	**71%**	0.59	—	—	29%	0.69
ATP	Bound (70 cases)	Ligand	83%	0.69	**86%**	0.69	77%	0.71	81%	0.67	—	—	69%	0.60
		P-group	**80%**	0.76	67%	0.61	63%	0.57	77%	0.60	—	—	67%	0.78
	Unbound (33 cases)	Ligand	**82%**	0.62	76%	0.67	79%	0.59	73%	0.57	—	—	58%	0.77
		P-group	**82%**	0.63	61%	0.47	58%	0.48	67%	0.47	—	—	58%	0.77

Example of the combined use of multiple probes. Mevalonate kinase in complex with ATP (pdb code: 1kvk).**34** The first-ranking cluster obtained with the OP probe (red) correctly identifies the part of the site involved in binding to the phosphogroup, whereas none of the top 10 clusters obtained with the CME probe identifies that moiety. However, the third-ranking CMET probe cluster correctly identifies the nucleotidic part of the ligand, illustrating the advantage of using multiple probes to characterize heterogeneous binding sites.

Comparison with other binding site identification approaches

Two energy-based approaches (Q-SiteFinder**4** and i-Site**20**), a well-established pocket identification program (LigSite**21**), a peptide-binding site detection method (PepSite**22**), and a recent phospholigand-binding site identification approach (Phosfinder**23**) were chosen for comparison against SiteHound (OP probe). Binding-site identification coverage, defined as the percentage of known binding sites identified among the top three clusters with an MCC of at least 0.3, was used again as the measure of performance. With the exception of the unbound phosphosugars set, where LigSite achieved the highest performance (with SiteHound immediately following), and the ATP bound dataset, where i-Site outperformed the other methods, SiteHound was able to identify more binding sites than the other methods (Table III). The advantage of using chemically specific information becomes more evident when only the residues in contact with the phosphogroup are considered. In this case, the advantages of LigSite for unbound phosphosugars, and of I-site for bound ATP, disappear (Table III). It is interesting to note that while phosfinder shows a much lower phospholigand-binding site identification performance than SiteHound (OP probe), the median MCC of the binding sites that it is able to identify tends to be slightly higher than that of SiteHound. This is likely due to phosfinder relying on matching to existing phospholigand binding sites to identify new sites. Hence, while this seems to present a disadvantage for the identification of binding sites as a whole, it provides some advantage in the exact identification of the residues participating in those binding sites, since they are predefined in the template-binding sites used for matching. This suggests that an approach that combines energy-based methods for binding-site identification, with binding-site templates for the exact delineation of the binding site, may improve the overall performance. While many other binding site identification methods exist,**35** most of them are not tailored for the detection of sites with the special characteristics of phospholigand binding sites, hence we expect the differences in performance shown between SiteHound (OP probe) and the other methods to persist in the context of phosholigand binding site identification. One method that could potentially provide improved performance in the detection of phospholigand binding sites (although at a higher computational cost) is computational solvent mapping,**36** which uses multiple molecular probes to identify druggable binding sites in protein structures. It is conceivable, that a special set of molecular probes could be used to tailor the binding site identification to phospoligand binding sites. However, this idea has not yet been tested.

DISCUSSION AND CONCLUSIONS

We presented a computational approach to identify regions of a protein structure where a specific interaction with the phosphogroup(s) takes place. The testing on three independent datasets comprising 152 bound complexes and 83 unbound proteins involved in phospholigand recognition shows that by using a specific phosphate probe to compute interaction energy maps it is possible to reliably identify the phospholigand binding sites.

While known binding sites are not always identified as the top-ranking cluster, in a majority of the cases they are found among the top-three clusters. While ideally a method would always identify the known binding site as the top-ranking cluster, this is very difficult in practice for several reasons. The existence and the relative strength of a site is dependent on the ligand being considered, and it is possible that more than one ligand-binding site exists in a protein structure. Ultimately, a binding site can only be unequivocally defined when the ligand is also known. However, the main reason for the lack of perfect performance is probably the fact that a simple single-atom chemical probe is being used to achieve generality in the identification of binding sites for different phospholigands. Our previous studies with binding sites for drug-like ligands showed that once the identity of the ligand is added in a docking approach, the number of times that the correct binding site is identified among the top-three candidates increases with respect to the a priori prediction based on the single top-ranking cluster.**3** Hence, providing a small number of alternatives for the binding site location is not only of practical use but also may be necessary when the exact identity of the ligand is unknown.

Despite the variability in the electrostatic potential or the amino acidic composition of the binding sites, the signal derived from the interaction energy with a phosphate probe is invariably higher in the binding site as compared to the rest of the protein and the approach successfully exploits this property (Supporting Information Fig. S2). This seems to be true even in the case of small conformational changes, since the results indicate that the approach is relatively insensitive to the changes that occur upon ligand binding. In part, this is probably due to the method not relying solely on shape or van der Waals interactions. It is likely that the chemical identity of the binding site is retained in a more robust way than the shape alone, thus making it easier to recognize the energetic signature of the binding site in the unbound structure, even if it is geometrically distorted, since it will still tend to stand out from the background.

An optional step involving the reranking of the top predicted sites by conservation score further improved the predictions where the energy-based signal is relatively weak (as in some of the phosphopeptide cases). Adding conservation information seems to provide a way to reduce the noise coming from decoy sites, given that a sizeable number of homologous sequences are available to accurately compute conservation scores. On the other hand, conservation alone cannot be used to precisely pinpoint the residues involved in the specific recognition of the phosphogroup, since they generally form a proper subset of all the conserved residues in a protein family. Hence the structure-based approach described here, while more accurate, is also highly complementary to conservation-based approaches and would be particularly useful when it is necessary to identify residues that are conserved for binding to specific chemical groups, such as the phospholigands. Such chemical information is absent in conservation-based approaches.

It is important to mention that this method cannot be used directly to identify, a priori, proteins that could be involved in phosphate recognition. However, it can guide mutagenesis experiments to confirm specific binding or guide further computational studies such as molecular docking. On the other hand, the more challenging problem of binding site classification (i.e., assigning the possible class of ligands to a binding site) can be considered as an extension of the problem of binding site identification. The results presented here, in particular the combination of different probes, indicate that an energy-based approach is well suited to provide an integrated approach for binding site identification and classification. This study highlights the advantages of using simple chemical information in the process of identifying binding sites with different properties, and thus provides a framework on which to exploit

existing forcefields**37**, **38** for the identification of a variety of binding site types. We envisage that this method will be useful in the context of structure-based functional annotation; for example, using the many structures for proteins of unknown function produced by structural genomics projects and also in the context of rational drug design and the general analysis of newly determined protein structures. The method is well suited for large-scale analysis of many structures or many conformations of one structure, because it is fast and can be fully automated. It is also a good complement to other structural analysis methods that, being more detailed, require more computational power.**36**

Acknowledgements

The authors thank Dr. Mihaly Mezei for helping with the implementation of electrostatic term, and the members of the Sanchez Lab for useful suggestions and discussions.

References

1. YaffeMB.Phosphotyrosine-binding domains in signal transduction.*Nat Rev Mol Cell Biol*2002;**3**:177–186.
2. JoughinBA,TidorB,YaffeMB.A computational method for the analysis and prediction of protein:phosphopeptide-binding sites.*Protein Sci* 2005;**14**:131–139.
3. GhersiD,SanchezR.Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites.*Proteins* 2009;**74**:417–424.
4. LaurieAT,JacksonRM.Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.*Bioinformatics* 2005;**21**:1908–1916.
5. BermanHM,BattistuzT,BhatTN,BluhmWF,BournePE,BurkhardtK,FengZ,GillilandGL,IypeL,JainS,FaganP,MarvinJ,PadillaD,RavichandranV,SchneiderB,ThankiN,WeissigH,WestbrookJD,ZardeckiC.The Protein Data Bank.*Acta Crystallogr D Biol Crystallogr* 2002;**58**(Part 6):899–907.
6. AltschulSF,MaddenTL,SchafferAA,ZhangJ,ZhangZ,MillerW,LipmanDJ.Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*Nucleic Acids Res* 1997;**25**:3389–3402.
7. PuntaM,CoggillPC,EberhardtRY,MistryJ,TateJ,BoursnellC,PangN,ForslundK,CericG,ClementsJ,HegerA,HolmL,SonnhammerEL,EddySR,BatemanA,FinnRD.The Pfam protein families database.*Nucleic Acids Res* 2012;**40**:D290–D301.
8. KellenbergerE,MullerP,SchalonC,BretG,FoataN,RognanD.sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank.*J Chem Inf Model* 2006;**46**:717–727.
9. O'BoyleNM,MorleyC,HutchisonGR.Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit.*Chem Cent J* 2008;**2**:5.
10. GhersiD,SanchezR.EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures.*Bioinformatics* 2009;**25**:3185–3186.
11. Van Der SpoelD,LindahlE,HessB,GroenhofG,MarkAE,BerendsenHJ.GROMACS: fast, flexible, and free.*J Comput Chem* 2005;**26**:1701–1718.
12. SolmajerT,MehlerEL.Electrostatic screening in molecular dynamics simulations.*Protein Eng*1991;**4**:911–917.
13. van der SpoelD,LindahlE,HessB,van BuurenA,ApolE,MeulenhoffP,TielemanP,SijbersA,FeenstraA,van DrunenR,BerendsenH.Gromacs User Manual version 3.3;Nijenborgh Groningen, The Netherlands, 2005.
14. CuiM,MezeiM,OsmanR.Prediction of protein loop structures using a local move Monte Carlo approach and a grid-based force field.*Protein Eng Des Sel* 2008;**21**:729–735.
15. HernandezM,GhersiD,SanchezR.SITEHOUND-web: a server for ligand binding site identification in protein structures.*Nucleic Acids Res* 2009;**37**:W413–W416.
16. LarkinMA,BlackshieldsG,BrownNP,ChennaR,McGettiganPA,McWilliamH,ValentinF,WallaceIM,WilmA,LopezR,ThompsonJD,GibsonTJ,HigginsDG.Clustal W and Clustal X version 2.0.*Bioinformatics* 2007;**23**:2947–2948.
17. CapraJA,SinghM.Predicting functionally important residues from sequence conservation.*Bioinformatics* 2007;**23**:1875–1882.
18. BaldiP,BrunakS,ChauvinY,AndersenCA,NielsenH.Assessing the accuracy of prediction algorithms for classification: an overview.*Bioinformatics* 2000;**16**:412–424.
19. MatthewsBW.Comparison of the predicted and observed secondary structure of T4 phage lysozyme.*Biochim Biophys Acta* 1975;**405**:442–451.

20. MoritaM,NakamuraS,ShimizuK.Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures.*Proteins* 2008;**73**:468–479.
21. HuangB,SchroederM.LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.*BMC Struct Biol* 2006;**6**:19.
22. PetsalakiE,StarkA,Garcia-UrdialesE,RussellRB.Accurate prediction of peptide binding sites on protein surfaces.*PLoS Comput Biol* 2009;**5**:e1000335.
23. ParcaL,MangoneI,GherardiniPF,AusielloG,Helmer-CitterichM.Phosfinder: a web server for the identification of phosphate-binding sites on protein structures.*Nucleic Acids Res*2011;**39**:W278–W282.
24. LeeCH,KominosD,JacquesS,MargolisB,SchlessingerJ,ShoelsonSE,KuriyanJ.Crystal structures of peptide complexes of the amino-terminal SH2 domain of the Syp tyrosine phosphatase.*Structure*1994;**2**:423–438.
25. DurocherD,TaylorIA,SarbassovaD,HaireLF,WestcottSL,JacksonSP,SmerdonSJ,YaffeMB.The molecular basis of FHA domain:phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms.*Mol Cell* 2000;**6**:1169–1182.
26. OlsonLJ,ZhangJ,LeeYC,DahmsNM,KimJJ.Structural basis for recognition of phosphorylated high mannose oligosaccharides by the cation-dependent mannose 6-phosphate receptor.*J Biol Chem*1999;**274**:29889–29896.
27. OlsonLJ,DahmsNM,KimJJ.The N-terminal carbohydrate recognition site of the cation-independent mannose 6-phosphate receptor.*J Biol Chem* 2004;**279**:34000–34009.
28. BauerCB,HoldenHM,ThodenJB,SmithR,RaymentI.X-ray structures of the apo and MgATP-bound states of Dictyostelium discoideum myosin motor domain.*J Biol Chem* 2000;**275**:38494–38499.
29. BauerCB,FonsecaMV,HoldenHM,ThodenJB,ThompsonTB,Escalante-SemerenaJC,RaymentI.Three-dimensional structure of ATP:corrinoid adenosyltransferase from Salmonella typhimurium in its free state, complexed with MgATP, or complexed with hydroxycobalamin and MgATP.*Biochemistry*2001;**40**:361–374.
30. HendlichM,RippmannF,BarnickelG.LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.*J Mol Graph Model* 1997;**15**:359–363,389.
31. LevittDG,BanaszakLJ.POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids.*J Mol Graph* 1992;**10**:229–234.
32. del SolA,PazosF,ValenciaA.Automatic methods for predicting functionally important residues.*J Mol Biol* 2003;**326**:1289–1302.
33. CapraJA,LaskowskiRA,ThorntonJM,SinghM,FunkhouserTA.Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure.*PLoS Comput Biol*2009;**5**:e1000585.
34. FuZ,WangM,PotterD,MiziorkoHM,KimJJ.The structure of a binary complex between a mammalian mevalonate kinase and ATP: insights into the reaction mechanism and human inherited disease.*J Biol Chem* 2002;**277**:18134–18142.
35. ChenK,MiziantyMJ,GaoJ,KurganL.A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds.*Structure* 2011;**19**:613–621.
36. LandonMR,LanciaDR,Jr,YuJ,ThielSC,VajdaS.Identification of hot spots within druggable binding regions by computational solvent mapping of proteins.*J Med Chem* 2007;**50**:1231–1240.
37. GoodfordPJ.A computational procedure for determining energetically favorable binding sites on biologically important macromolecules.*J Med Chem* 1985;**28**:849–857.

38. ScottWRP,HunenbergerPH,TironiIG,MarkAE,BilleterSR,FennenJ,TordaAE,HuberT,KrugerP,van GunsterenWF.The GROMOS biomolecular simulation program package.*J Phys Chem A*1999;**103**:3596–3607.