Published online 24 October 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 4* ***1231–1242***
doi:10.1093/nar/gkq987

Phosphate binding sites identification in protein structures

Luca Parca, Pier Federico Gherardini, Manuela Helmer-Citterich* and Gabriele Ausiello

Department of Biology, Centre for Molecular Bioinformatics, University of Rome 'Tor Vergata', Via della Ricerca Scientifica snc, 00133 Rome, Italy

Received May 20, 2010; Revised September 17, 2010; Accepted October 5, 2010

ABSTRACT

Nearly half of known protein structures interact with phosphate-containing ligands, such as nucleotides and other cofactors. Many methods have been developed for the identification of metal ions-binding sites and some for bigger ligands such as carbohydrates, but none is yet available for the prediction of phosphate-binding sites. Here we describe Pfinder, a method that predicts binding sites for phosphate groups, both in the form of ions or as parts of other non-peptide ligands, in proteins of known structure. Pfinder uses the Query3D local structural comparison algorithm to scan a protein structure for the presence of a number of structural motifs identified for their ability to bind the phosphate chemical group. Pfinder has been tested on a data set of 52 proteins for which both the apo and holo forms were available. We obtained at least one correct prediction in 63% of the holo structures and in 62% of the apo. The ability of Pfinder to recognize a phosphate-binding site in unbound protein structures makes it an ideal tool for functional annotation and for complementing docking and drug design methods. The Pfinder program is available at http://pdbfun.uniroma2.it/pfinder.

INTRODUCTION

Many important chemical reactions and molecular interactions that occur in the cell involve ligands containing the phosphate group.

More than half of known proteins has been shown to interact with a phosphate group (1). Several of these proteins are involved in essential pathways and their malfunction leads to severe diseases and other abnormalities in humans (2,3). Moreover the affinity for the phosphate group is essential in nucleotide recognition and nucleotide-containing ligands were the earliest cofactors bound to proteins (4).

The ability to bind phosphate has evolved in many non-homologous protein families. There are however some preeminent groups that dominate this distribution such as that of P-loop containing proteins (5) or proteins with a Rossmann-type fold (6).

The possibility to characterize a protein for its ability to interact with a phosphate, or a phosphate-containing ligand, is therefore of paramount importance. Different methods exist for predicting the binding sites of a variety of ligands such as various metal ions or carbohydrates (7–11). However, to the best of our knowledge, no method is yet available for the identification of phosphate binding sites (PbSs) even if the biological relevance of this specific ligand is beyond question.

The methods that predict binding sites for specific ligands in a protein structure can be classified as 'comparative' or 'non-comparative' (12). Comparative methods search for structural similarities between different proteins that interact with similar types of ligands and often benefit from libraries of predefined template motifs. Conversely non-comparative approaches only make use of structural and chemo-physical features, calculated from the structure of interest to identify potential ligand-binding sites.

Many methods have been developed which are specific for the identification of metal ion-binding sites. Fold-X (7) is a force field for the detection of single atom-binding sites and can be applied to metal ions (Mg, Zn, Ca, Mn and Cu). The method searches for the chosen ion-binding site, by superimposing known metal-binding sites onto the query structure. Geometric and energetic criteria are then used to accept or discard candidate solutions. Fold-X is able to identify from 90% to 97% of the binding sites, depending on the nature of the metal, with 21% of overpredictions. The GG algorithm (8) uses geometrical features of the protein structure to derive Ca ions-binding sites through graph theory. The algorithm searches for clusters of surface oxygen atoms whose center determines

*To whom correspondence should be addressed. Tel: +39 06 72594324; Fax: +39 06 2023500; Email: citterich@uniroma2.it the binding site. Atoms different from the oxygen are not allowed inside the sphere described by the cluster. This algorithm has a sensitivity and a selectivity that range from 87% to 91% and from 74% to 77%, respectively.

Some non-comparative methods have been developed for more complex ligands, such as carbohydrates. Taroni *et al.* (9) characterized the structural features of the binding sites in 19 carbohydrates-binding proteins. Six parameters were evaluated in this analysis: solvation potential, residue propensity, hydrophobicity, planarity, protrusion and relative accessible surface area. The authors then used these features to calculate the probability that a surface patch binds a carbohydrate and obtained an accuracy of 65%, considering a binding site as correctly predicted if its overlap with the real binding site is >70%. Kulharia *et al.* (10) developed a method that predicts binding sites location for inositol and carbohydrates, using a methylene probe to derive van der Waals interaction energies from a protein structure and amino acid propensities, derived from a data set composed of protein–carbohydrate complexes. This method, called InCa-SiteFinder, has specificity and sensitivity of 98 and 72%, respectively, but the authors were more permissive in the assignment of correct predictions. A predicted binding site is considered correct if its overlap with the real binding site is >25%. Ghersi and Sanchez (11) used a similar approach, determining, for a protein structure, molecular interaction fields (called MIFs) using a methyl probe and a phosphate oxygen probe. In this way, regions sharing a higher probability to encompass a binding site can be identified. In 95% of the bound protein structures and 79% of the unbound the correct binding site is among the top three predicted binding sites. Joughin *et al.* (13) developed a method for the identification of phosphorylated peptides-binding sites. The method uses propensity values derived from the physical and chemical properties of nine phospho–peptide-binding domains and was tested on the same set of structures. These methods, developed for ligands bigger than metal ions, seem to detect favorable binding regions instead of predicting the position in space of the ligand atoms.

In many PbSs, the phosphate is involved in a network of hydrogen bonds with the backbone atoms of the protein residues. The backbone forms a geometrically and energetically favorable scaffold, which tightly binds the phosphate ion. Glycine residues have a critical role in these binding sites as they allow neighboring residues to assume an optimal conformation. Another common feature is the positive electrostatic potential that promotes the binding of the negatively charged phosphate moiety (14) and the presence of coordination metal ions like Mg and Zn (15).

Fragment-based approaches to docking and drug design (16,17) have shown that a binding pocket can be treated as composed of different, partly independent, subsites interacting with the various molecular 'hooks' composing a ligand. Moreover structural motifs associated with specific ligand 'fragments' (i.e. as opposed to whole ligands) have been identified (18,19). Therefore it is logical to cast the question of predicting binding sites in terms of predicting spots of favorable interactions with chemical fragments that recur in multiple molecules.

Brakoulias and Jackson (20) built a data set of 3D phosphate-binding motifs, (PbMs) by comparing and then clustering a large set of PbSs. Their work identified 476 binding sites for this ligand grouped into 10 main clusters. Kinoshita *et al.* (21) compared 491 protein-binding sites in protein–mononucleotide complexes, identifying four frequent structural motifs, like the P-loop, and analyzing their distribution among protein superfamilies. In a more comprehensive study Ausiello *et al.* (18) built a data set of binding motifs associated with specific chemical fragments that are present in a variety of different ligands. The majority of these binding motifs resulted to bind a phosphate group.

In this study we describe Pfinder, a new comparative method for the identification of PbSs on a protein structure, and its application to an ensemble of apo (unbound ligand) and holo (bound ligand) structures of phosphate-binding proteins. The method performs a local structural comparison between the query protein and a data set of PbSs (18) thus identifying groups of amino acids, which display a similarity to known binding motifs. Pfinder then evaluates all the candidate predictions using several geometric criteria and a sequence conservation score to select the final solutions.

We obtained comparable results with both apo and holo structures, showing that the method is robust with respect to the conformational changes that occur upon ligand binding. Pfinder can therefore be used for the functional annotation of proteins solved in structural genomics projects as well as to further characterize already annotated proteins. Drug design and molecular docking efforts could also take advantage of this method, since the predictions show the exact position in space of the phosphate moiety.

MATERIALS AND METHODS

The Pfinder method

Pfinder is a novel method to predict the position of PbSs on the surface of a target protein structure. Pfinder uses the Query3D local structure comparison algorithm (22,23) to scan a structure with a list of known PbM templates. The PbSs are then evaluated and filtered by considering their position with respect to the solvent accessible surface and the clefts on the surface of the protein.

The PbMs data set

Pfinder uses a previously defined set of PbMs (18). In that study a number of structural motifs shared by at least two folds and associated with specific ligand fragments were identified. From that data set we selected only the motifs interacting with a phosphate group. Our final data set therefore contains 231 motifs, composed by at least three residues, that are present in at least two different SCOP (24) folds and bind at least one phosphate group. Since each motif is represented by a different protein structure for each different fold in which the motif is present we selected a representative structure to be used

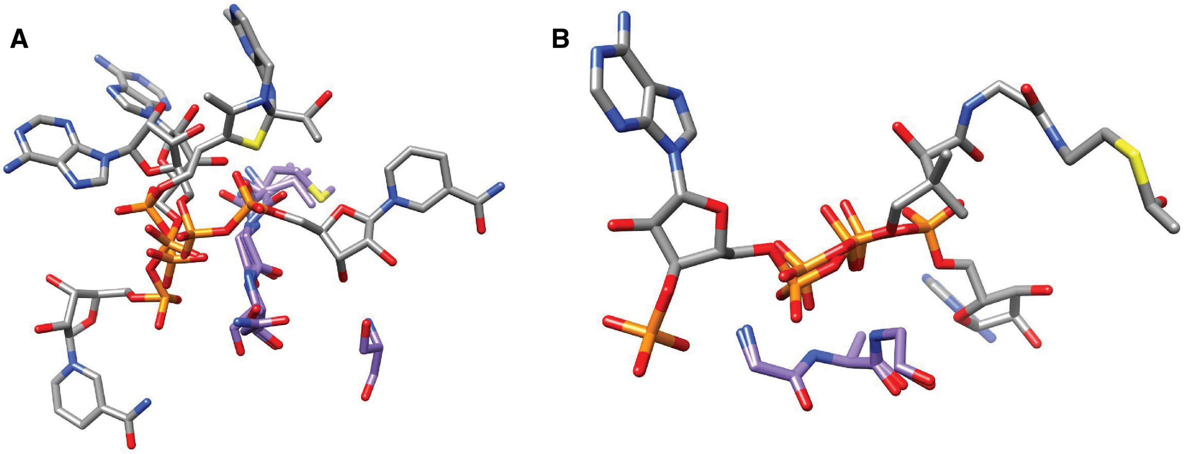

Figure 1. Two examples of PbMs used by the method. Each motif is represented by the residues forming the motif, that belong to different structures, and their corresponding bound ligands. Carbon atoms of the binding residues are in purple while those belonging to the ligand are in white, all the other atoms are colored by type (phosphorus in orange, oxygen in red, nitrogen in light blue, sulphur in yellow). (**A**) PbM (id 598). The motif is [V,M]-G-[N,A,S]-S where the final serine residue is present in only two of the three protein structures with different folds that share the motif. The three structures belong to a inositol-1′-phosphate synthase from *M. tubercolosis* (PDB code 1GR0), to an adenylyltransferase from *Methanobacterium thermoautotrophicum* (PDB code 1M8F) and to the *Klebsiella pneumoniae* acetolactate synthase (PDB code 1OZH). (**B**) PbM (id 1075). Two glycines and an alanine interact with the phosphate groups forming a G-A-G motif in two protein structures with different folds. The two structures belong to a *Thermus thermophilus* kinase (PDB code 1V1B) and to a histone acetyltransferase from *Saccharomyces cerevisiae* (PDB code 1QSM).

as reference (Figure 1). In choosing the representative structure for each motif we adhered to the following criteria:

(i) When one of the structures possesses one or more binding residues that clearly differ in type or position from all the other folds/structures that structure is never chosen as representative;
(ii) The phosphate group must clearly interact with the motif. When multiple structures satisfy this requisite the one with the phosphate position that best corresponds to the average of the others is chosen as representative; and
(iii) We kept a single phosphate group for each motif.

A further manual inspection discarded 16 motifs for which no suitable representative structure could be found. The complete list of motifs we used is listed in Supplementary Table S1. The final data set contains 215 motifs binding 82 different ligands. The 10 most frequently occurring ligands are reported in Table 1.

Table 1. Name, PDB three-letter code and number of occurrences of the 10 most abundant ligands containing phosphate groups found in the PbMs data set

Ligand	PDB code	Occurrence
Flavin-adenine dinucleotide	FAD	17
Phosphate ion	PO4	17
Adenosine-5′-diphosphate	ADP	12
Pyridoxal-5′-phosphate	PLP	12
Adenosine-5′-triphosphate	ATP	11
Phosphoaminophosponic acid-adenylate ester	ANP	10
NADP	NAP	8
NADPH	NDP	8
Nicotinamide-adenine-dinucleotide	NAD	7
Flavin mononucleotide	FMN	7

The structural comparison method

Pfinder uses the Query3D structural comparison algorithm to search for the occurrence of PbMs in the structure under analysis. Query3D identifies local, sequence-independent, similarities between two protein structures: the algorithm searches for the largest subset of amino acids with similar positions in space between two protein structures or between a protein structure and a small set of residues (e.g. a structural motif). Protein chains are represented as ensembles of non-connected residues, each one described using two points: the Cα and the geometric centroid of the side-chain atoms. Every residue is also associated with a list of neighbors (i.e. distance between their Cαs <7.5 Å).

Two sets of residues are deemed similar if they fulfill three criteria: neighborhood, structural similarity [root mean square deviation value (RMSD)] and biochemical similarity. The first criterion requires that all the residues present in the set of matching amino acids are neighbors of at least one of the other residues in the set. The second criterion requires every match to have a RMSD equal or lower than a specified threshold. The similarity criterion allows residues to substitute for each other when their score in a BLOSUM62 (25) matrix is equal or greater than a specified threshold.

The phosphate-binding motifs search

Pfinder compares the protein structure to be analyzed with each phosphate-binding motif in the data set. Whenever Query3D finds a structural match the method predicts a PbS. A structural match implies a 3D transformation (rotation + translation) that superimposes one structure, e.g. the binding motif, onto another, e.g. the protein of interest. Since each phosphate-binding motif includes the coordinates of the interacting phosphate group this is roto-translated in space as well. The new position of the phosphate group defines a predicted PbS.

In order to remove multiple overlapping predictions we used an agglomerative (centroid linkage) hierarchical clustering procedure with a 2.0 Å threshold. The prediction closest to the centroid of the cluster was retained.

The main application of Pfinder is the prediction of PbSs in the difficult case of newly determined protein structures with no homologs in the Protein Data Bank (PDB) (26). In order to verify the ability of the method to work for these proteins we adopted the limitation of not considering structural matches involving PbMs derived from structures potentially homologous to the protein under analysis. To this end we removed all the PbMs that came from a protein belonging to the same homology group using PISCES (27) groups at 30% sequence identity.

The solvent excluded surface filtering

The structural comparison procedure of Pfinder identifies PbMs irrespective of whether they are located on the protein surface or not. This produces a great amount of matches between PbMs and residues in the core of the structure that cannot possibly interact with a phosphate for steric reasons. Pfinder therefore discards all the predicted PbSs that are located inside the Solvent Excluded Surface (28) of the structure, calculated using the UCSF Chimera MSMS package (29).

The solvent excluded surface is described as a triangular mesh and the phosphate is considered inside if the vector joining it to the nearest protein atom intersects the surface mesh an even number of times.

However some predictions, while falling outside the surface, are still located too close to it to be considered biologically relevant. We measured the distance between all the crystallized phosphate groups and the surface in each of the 59 protein structures of the training set and found no phosphorus atom closer than 0.7 Å to any surface point. Therefore we also discarded all the predictions closer than this threshold to the protein surface.

In order to also remove PbSs predicted at the interface between two protein chains we calculated the solvent excluded surface using the complete PDB structures instead of the single protein chains. Accordingly all the PbSs at the interface will appear as inside the structure and discarded.

Conservation of the structural motifs

Pfinder calculates a sequence conservation value for each identified PbM in order to discriminate between true and false predictions. This value represents the relative sequence conservation of the residues forming the structural motif with respect to all the other residues in a multiple alignment of the protein family. To this end we retrieved all the PFAM (30) domains associated with the UniProt (31) id corresponding to the structure under analysis. The PFAM multiple alignments were then mapped on the structure using the UniProt sequence as a link. The correspondences between the residues from the structure and the UniProt sequence were determined with the alignment program Needle from the EMBOSS (32) suite.

The conservation of each residue is defined as the percentage of similar residues in the corresponding multiple alignment column. Two residues are considered similar if their substitution value in a BLOSUM62 matrix is equal or higher than 1.

In order to normalize and compare conservation values from different multiple alignments we calculated the percentile corresponding to each value with respect to the distribution of values in the alignment. The final score of a motif is given by the average of all of its amino acid values. Motifs formed only by amino acids with no value assigned were considered has having a conservation score of zero.

Construction of the training set

The set of protein structures used to train the method was derived from the set of nucleotide-binding proteins used by Zhao *et al.* (33). This is a non-redundant data set (at 30% sequence identity) of 56 high-quality protein structures with a resolution <2.8 Å. The proteins in this set do not bind DNA and have at least four residues that establish contacts with phosphate-containing ligands. The data set includes 16 adenosine complexes, 9 guanosine complexes and 31 dinucleotide cofactors complexes. To make a non-redundant data set of protein chains we removed all the identical protein chains from the data set leaving only a representative structure for each group. We obtained a set of 54 unique protein chains, belonging to 53 complexes.

In some cases the phosphate groups of the ligands do not directly interact with the surface of the protein. Obviously these groups cannot be predicted by Pfinder. In order to remove such proteins from the data set we defined a phosphate group as interacting with a structure if at least three amino acids of the protein have an atom located closer than 4.0 Å from the phosphorus atom. The structures that did not fulfill this criterion were eliminated thus reducing the training set to 40 protein chains.

Since this data set is focused on nucleotides, we added 20 protein structures that bind different types of phosphate-containing ligands in order to have the same ratio of nucleotides/non-nucleotides phosphorylated ligands as the whole PDB (roughly 2:1). These 20 ligands were chosen at random from a set of 1273 phosphate-containing ligands occurring in less than 10 PDB structures. For each of the twenty ligands, a protein structure is randomly chosen from those that bind it and then added to the training set. These proteins bind ligands ranging from phosphorylated amino acids to aliphatic and aromatic phosphorylated compounds.

We also removed from the data set all the structures having a sequence identity >30% with any structure of the test set. The final training set includes 62 protein chains (59 proteins) that are listed in Supplementary Table S2.

Construction of the test set

The set of protein structures used to test the method was obtained from the LigASite database (34). LigASite contains 337 apo-holo pairs of high-quality structures that bind biologically relevant ligands (LigASite v8.0, July 2009). The pairing of apo and holo structures gives information about the conformational changes occurring upon ligand binding. Thanks to these two data sets the method can be tested on the conformation effectively binding the ligand and on the unbound structure as well. Since the data set is composed of complexes that bind all kinds of ligands, we excluded 170 pairs because they did not bind ligands containing phosphate groups. In order to discard structures binding ligands used in crystallization and purification methodologies, four structures that bind single phosphate groups were also removed. Three of these, 1MG2, 2GTE and 2F10 contain phosphate groups not involved in the biological function of the protein. The fourth protein structure, 1X55, contains a phosphate group and a non-hydrolyzable enzyme inhibitor analogue of the biological ligand, the asparaginyl–adenylate; as consequence the structure can not be retained. We also removed from the data set 13 structures containing mutations. Furthermore 48 pairs of proteins were discarded because the binding pocket defined by LigASite was not located in a chain shared by both the apo and the holo structures.

Additionally, as already done for the training set, we discarded 48 structures whose phosphate groups were not directly bound to the protein (phosphorous atom not within 4.0 Å from at least three residues). Two pairs have been discarded since they lack corresponding PFAM alignments. The final test set comprises 52 pairs of structures that are listed in Supplementary Table S3.

The majority of the ligands in the test set contains adenosine and nicotinamide (Table 2).

In order to classify the predicted PbSs in the apo structures as true positives or not we inferred the position of the ligand by comparing the apo and holo structures with Query3d and then roto-translating the bound ligand accordingly.

Table 2. Name, PDB three-letter code and number of occurrences of the 10 most abundant ligands containing phosphate groups found in the test data set

Ligand	PDB code	Occurrence
Adenosine-5′-diphosphate	ADP	10
NADP	NAP	6
Guanosine-5′-diphosphate	GDP	6
Adenosine-5′-triphosphate	ATP	6
NADPH	NDP	5
Uridine- 5′- monophosphate	U5P	4
Nicotinamide–adenine–dinucleotide	NAD	3
Flavin–Adenine Dinucleotide	FAD	3
Adenosine-5′-monophosphate	AMP	3
Cytidine-5′-monophosphate	C5P	2

RESULTS

Overview

We developed Pfinder, an automated method for the identification of phosphate-binding sites in protein structures. The method compares the protein structure to be analyzed with a library of 3D PbMs. This library contains groups of amino acids (in 3D conformation) that bind a phosphate group. Pfinder predicts the PbS by comparing the query protein with the library of PbMs and then uses other geometric criteria and a sequence conservation score to filter and evaluate the predictions. The method has been trained on a manually curated data set of proteins complexed with non-peptide ligands containing at least one phosphate group and then tested on a set of proteins whose structures have been crystallized both with and without the phosphate-containing ligand.

Parameters optimization

We trained our method on a list of 59 proteins structures binding phosphorylated ligands (‘Materials and Methods’ section). This data set was used to tune the parameters of Query3D (22,23) which is the local structure comparison algorithm we used in the first step of our procedure. Two parameters were considered, namely the RMSD threshold and the minimum BLOSUM62 (25) value that allows two residues to match with each other.

In order to find the best combination of such parameters, the entire set of PbMs was used to scan all the proteins in the training set. Three different values were considered for the RMSD threshold, namely 0.7, 0.9 and 1.1 Å. Query3D allows two residues to substitute for each other when their score in a BLOSUM62 matrix is equal or higher than a specified threshold (‘Materials and Methods’ section). Three different values (−1, 0 and 1) for this threshold have been tested.

To verify that Pfinder could also work correctly on structures dissimilar from those already solved we discarded all the matches coming from PbMs belonging to possible homologues of the chain under analysis (‘Materials and Methods’ section). On the training set this limit corresponds to the removal of ~1–2% of the total structural matches obtained.

For each identified structural match, Pfinder positions the phosphate group associated with the matching PbM onto the query protein, using the roto-translations corresponding to the structural match. After this step we apply a first filter by discarding all the predictions that are located inside the solvent excluded surface (SES) (28) of the protein. The remaining predictions are clustered as described in the ‘Materials and Methods’ section.

The predictions are then evaluated according to their distance from the experimentally determined position of the phosphate group belonging to the crystallized ligand. We consider as true positives (TP) the predictions for which the phosphorus atom is placed in a sphere of radius 5.0 Å centered on the crystallized phosphorus atom. This threshold value is low enough to clearly identify the site where the phosphate group can bind. Moreover, since only one phosphate group per motif was considered ('Materials and Methods' section), this threshold is high enough to make a prediction represent all the phosphate groups bound by a motif.

During the training the conservation threshold was selected using the conservation values assigned to each PbM ('Materials and Methods' section). The optimal threshold is defined as the one that better discriminates TP PbMs from false positive (FP) PbMs. We calculated the Matthews Correlation Coefficient (MCC) for each possible threshold and selected the best for each set of different parameters. All the predicted PbMs with a conservation value under this threshold are deemed poorly conserved and are therefore discarded.

Performance with the training set

Figure 2 shows the complete results for all the combinations of parameters we tested. Using the less stringent RMSD and substitution matrix threshold values (1.1 Å and −1, respectively) the method was able to correctly identify at least one TP phosphate group in 51 of the 59 structures of the training set. The conservation threshold, with the highest MCC for this parameters combination, was 74.3 and produced an average of 30.7 ± 2.9 FP per structure. Using the most stringent parameter values (0.7 Å and 1) and a conservation threshold of 70 the method produced a much lower number of FP predictions (1.0 ± 0.1 per structure) but the number of protein structures of the training set without any correct prediction raises from 1 to 26 out of 59. We determined the set of parameters that results in the maximum percentage of protein structures with at least one correctly predicted

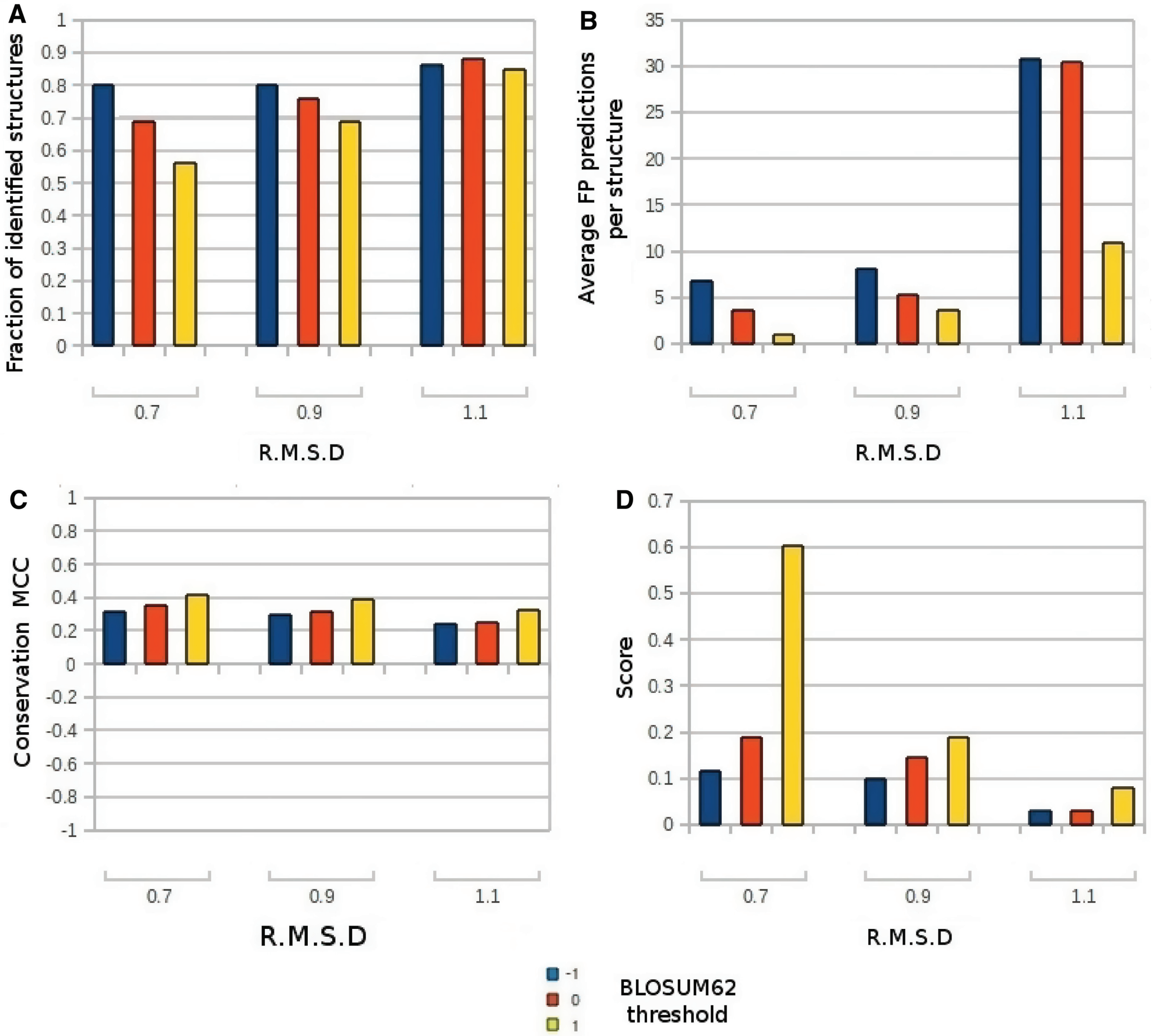

Figure 2. Results obtained on the protein structures of the training set. Each bar in the graphs represents the results for a different combination of RMSD and substitution parameters used. The RMSD threshold is reported on the X-axis, while different colors show the BLOSUM62 threshold. (**A**) Percentage of analyzed structures having at least one correctly predicted PbS. (**B**) Average number of FP predictions produced per structure. (**C**) Matthews Correlation Coefficient (MCC). (**D**) Final score. The score is the fraction of identified proteins divided by the average number of FP predictions per structure.

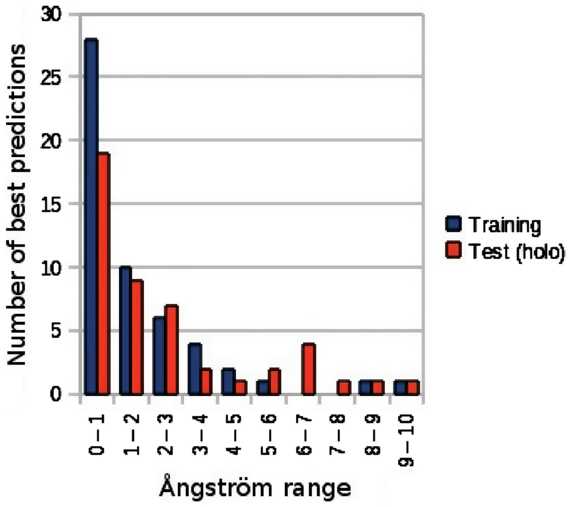

Figure 3. Distribution of the distances between the crystallographic position of the bound phosphate and the best prediction obtained for each protein structure in the training and test sets.

phosphate group, and the lowest average value of FP predictions. Figure 2D shows the optimized score for all the parameters combinations tested. The best performance was obtained with a RMSD of 0.7 Å and a residue substitution threshold of 1. Although this set of parameters has the highest score, the number of proteins without at least one correct prediction is too high (44%) to make this parameters combination usable. The second best performance value was obtained with both the parameters 0.7 (RMSD)/0 (substitution) and 0.9/1. However the latter parameters resulted in a higher MCC (0.39 versus 0.35). The 0.9/1 combination allows the method to identify at least one correct prediction in 69% of the proteins, with a conservation threshold of 66, an average FP predictions number of 3.7 ± 0.4 and a high conservation threshold area under curve (AUC) value (0.81). The TP predictions (i.e. the ones closest to a crystallized phosphate group) made on the 59 proteins of the training set are evenly spread in a 5 Å radius from the crystallized phosphate group. The distribution of the distances between the crystallographic positions and the best predictions in the training set is reported in Figure 3.

Discrimination between phosphate-binding and non-binding proteins

Pfinder is mainly aimed at identifying the location of PbSs on protein structures. We decided to also measure its ability to discriminate which proteins can bind phosphate at all. To this end we built a negative data set of 59 protein structures that do not bind phosphate-containing ligands. These 59 proteins were chosen from different PISCES (27) groups (30% redundancy level) entirely composed of protein structures that do not bind phosphorylated ligands. We compared the results of Pfinder on these two sets, using the most stringent parameters combination (0.7 and 1). Considering as positives all the proteins with at least one predicted PbS, the method obtained an MCC value of 0.26, in discriminating the structures belonging to the two sets. This result shows that, even though a signal is present, it is too weak to discriminate binding from non-binding structures.

Alternative representation of the amino acids

The Query3D structural comparison algorithm, which was used in this work, represents each residue with the Cα and the geometric centroid of the side-chain (‘Materials and Methods’ section). This is the same representation that was used to build the data set of PbM (18). This representation takes into account both the backbone and the spatial orientation of the side-chain.

We also tested the performance of Pfinder with a novel comparison program, Superpose3D (23), which allows for user-defined residue representations. We used a detailed representation that is focused on the physicochemical properties of specific side-chain groups. Each residue is therefore described by the Cα plus points centered on the most important chemical groups of the side-chain. This representation was adapted from the one used by Schmitt *et al.* (35). Each chemical group is labelled as either hydrogen-bond donor, acceptor, donor/acceptor, hydrophobic aliphatic or aromatic. During the structural comparison only groups of the same type are matched.

We obtained an average of 504.2 predicted PbMs per structure on the training set, using an RMSD threshold of 0.7. We sorted all the PbMs first by the length of the match and then by RMSD. We then identified the threshold that better discriminates TP from FP PbSs. The best MCC value we obtained was 0.04, with an RMSD threshold of 0.655 Å and a minimum match size of five amino acids. With these parameters the method was able to identify 67.7% of the PbS of the Training set with an average of 45.1 ± 13.2 FP predictions. We conclude that this amino acid representation introduces a large amount of FP predictions without helping the method in predicting PbS missed during the training.

Surface cavities filtering

The majority of the protein ligands are bound in one of the top four largest pockets of the interacting protein (36). We decided to take advantage of this property to further improve the PbSs predictions made by Pfinder. Indeed many predictions are located in proximity of convex surface regions that are unlikely to be able to interact with any ligand. We decided to evaluate if a filter discarding all the predictions falling outside one of the top four largest cavities could further reduce the number of FPs without affecting the number of TPs. We used the Surfnet (37) program with default parameters to identify the four biggest protein clefts using the whole PDB complexes. A total of 64.7% of the 326 predictions placed outside the solvent accessible surface are located inside one of the four biggest clefts. When predictions falling outside one of the four largest pockets are discarded, the average number of FP predictions decreases from 3.7 ± 0.4 to 2.1 ± 0.3, but five structures lose their TP predictions because they do not reside in any of the four biggest pockets. (see Supplementary Table S4 for complete results).

Alternative phosphate-binding positions

The distribution of the distances between the best prediction made by Pfinder and the phosphate crystallographic position in the training set (shown in Figure 3) shows that 43.4% of the TP predictions almost perfectly overlap with the real phosphate group of the ligand (distance <2.0 Å). The remaining TP predictions (56.6%) could represent alternative binding positions for the phosphate groups of the ligand. To investigate this possibility we selected an example from the results on the training set with three TP predictions on the γ-phosphate of an ATP molecule, one at 1.4 Å, one at 3.7 Å and one at 4.7 Å.

The *Oryctolagus cuniculus* phosphorylase kinase, PDB code 1PHK (38), binds a molecule of adenosine-5′-triphosphate in the kinase active site. The binding site also contains two Mn ions in coordination with the γ-phosphate of the ATP molecule. Pfinder predicts three PbSs close to the γ-phosphate, in a position that makes the coordination of the Mn ions possible. The first two PbSs almost exactly overlap with the crystallized γ-phosphate group, but the third PbS is located 4.7 Å away from the γ-phosphate representing a putative alternative binding site. If this is true, the γ-phosphate group could be shifted from the crystallized position to the alternative one without moving the rest of the ligand. Figure 4A shows the two predictions close to the γ-phosphate located in the ATP-binding site of the crystallized kinase. The two TP predictions, that overlap with the crystallized γ-phosphate, come from structural matches involving highly conserved residues (conservation values of 97 and 98, respectively) while the alternative binding site involves poorly conserved residues (conservation value of 17).

In order to evaluate if this PbS can act as an alternative binding site we docked the ATP molecule inside the kinase-binding pocket using the Autodock4 (39) software. We constrained the ligand in a 60-Å edge grid (grid point spacing of 0.375 Å) using Autogrid4 and generated 100 solutions with a Lamarckian Genetic Algorithm. The two solutions with the lower RMSD value (Figure 4B and C) with respect to the crystallized ligand (0.97 Å and 1.71 Å, respectively) resulted in the docked γ-phosphate being placed respectively at 0.96 Å from the ligand γ-phosphate and at 1.4 Å from the predicted alternative position. Moreover these solutions have low calculated free energies of −4.44 kcal/mol and −5.67 kcal/mol respectively.

Test of Pfinder on bound and unbound protein structures

To test the method we created a high-quality non-redundant test set, described in detail in the 'Materials and Methods' section. This data set was derived from the LigASite database (34) and consists of 52 proteins that bind ligands containing at least one phosphate group. Each protein was crystallized both in its apo and holo form.

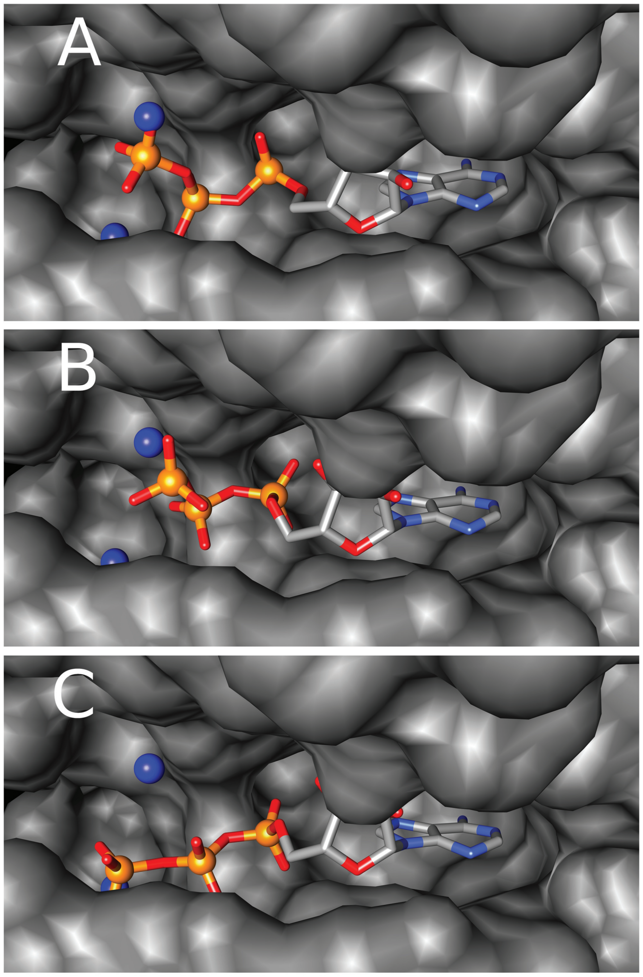

Figure 4. Adenosine-5′-triphosphate molecule bound by the *Oryctolagus cuniculus* phosphorylase kinase (PDB code 1PHK) in the active site. The ligand atoms are colored by atom type (carbon in gray, phosphorus in orange, oxygen in red, nitrogen in blue). The positions of the correctly predicted PbS phosphate in a 5.0 Å radius from the ligand phosphates are represented as blue spheres. (**A**) The ATP molecule crystallized in the active site of the kinase structure. (**B**) The docking solution, with the lowest RMSD value with the crystallized ligand. (**C**) The docking solution with the second lowest RMSD value with the crystallized ligand.

We applied Pfinder to the 52 holo structures using our data set of 215 PbMs and the parameters optimized on the training set. After the surface and conservation filtering we obtained 345 total predictions with an average of 6.6 predictions per structure. A total of 33 of the 52 proteins in the holo set (63%) get at least one correct PbS prediction, with an average of 4.8 ± 0.7 FP per structure and a conservation threshold AUC value of 0.83. As before the predictions that passed the surface filtering were analyzed for their location with respect to protein cavities. The apo-holo structures pair of the *Escherichia coli* pyruvate dehydrogenase E1 component (PDB code 2G67 and 2IEA respectively) is not considered for this analysis because Surfnet could not process these structures. Discarding all the predictions located outside the top four largest pockets eliminated two further structures, and reduced the average FP predictions to 2.8 ± 0.4 per structure.

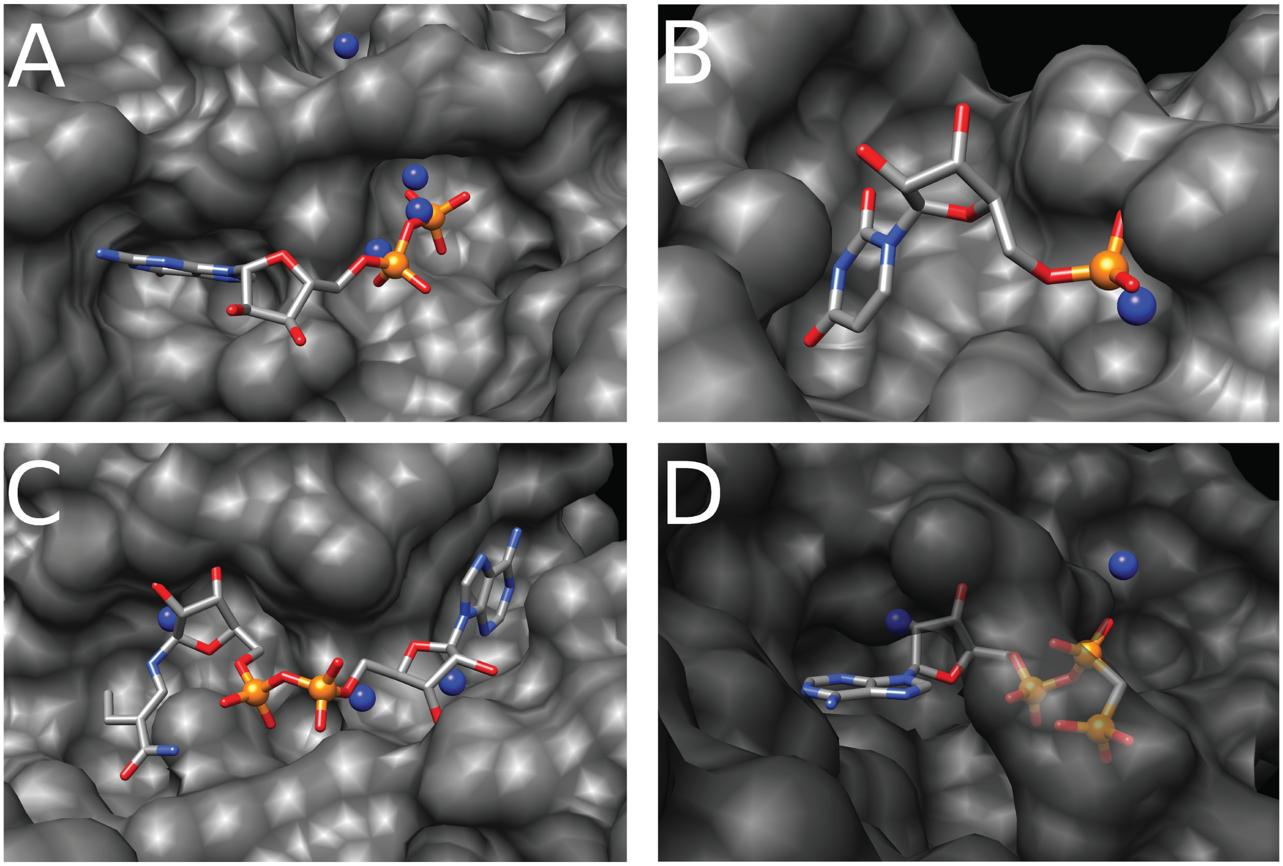

Figure 5. PbS predictions in four apo structures. The protein surface is colored in gray, the protein ligand is colored by atom type (carbon in gray, phosphorus in orange, oxygen in red, nitrogen in light blue). The PbSs are colored in blue and are displayed as spheres. (**A**) Binding site for the guanosine-5′-diphosphate molecule of the GTPase from *Pyrococcus abyssi* (PDB code 1YR6). (**B**) Binding site for uridine-5′-monophosphate molecule of the ribonuclease MC1 from seeds of *Momordica charantia* (PDB code 1BK7). (**C**) Binding site for the 1,4-dihydrodicotinamide adenine dinucleotide of the rabbit L-gulonate 3-dehydrogenase (PDB code 2DPO). (**D**) Binding site for the phosphomethylphosphonic acid adenylate ester of the *T. themophilus* HB8 probable ATP-binding protein (PDB code 1WJG).

The method was tested on the corresponding apo forms of the proteins in the Test Set. The comparison of the results obtained with the apo and holo forms should reveal how much the performance of Pfinder is affected by the conformational changes that occur upon ligand binding. After the surface and conservation filtering, Pfinder produced 351 predictions on the apo Set (6.8 per structure). A total of 32 structures out of 52 obtained at least one correct prediction. The method generated an average of 4.8 ± 0.7 FP per chain, while maintaining a good average value of TP (1.9 ± 0.3 per chain) and a high conservation threshold AUC value of 0.84. If the predictions outside the protein cavities are discarded, the average number of FP decreases to 3.2 ± 0.5 and the number of proteins with at least one correct prediction becomes 28. The 2G67 protein structure, which was excluded due to problems with the pocket detection algorithm (see above), does not have TP predictions. The detailed results for each structure are reported in the Supplementary Table S5. The distribution of the distances between the best predictions and the crystallized phosphate groups for each protein in the holo set, before the surface cavities filtering, is similar to that of the training set (Figure 3). In conclusion the conformational changes occurring upon ligand binding do not affect incisively the performance of Pfinder. Furthermore the results with the test set confirmed that the structural filtering, using surface and pockets information, greatly improves the accuracy of the method.

An apo protein test case

The apo-holo couple of the GTPase protein from *Pyrococcus abyssi* corresponds to the 1YR6-1YR9 PDB codes. The ligand bound by this GTPase is a guanosine-5′-phosphate molecule. Six phosphate-binding sites were predicted on the protein surface. Two of these overlap with the α- and β-phosphate groups of the ligand while a third clearly suggests, according to the enzyme biological function, a binding site for a γ-phosphate (Figure 5A). Only one FP-binding site was predicted outside the ligand-binding site yet still close to it. Other cases depicting predicted binding sites of phosphate belonging to different ligands are shown in Figure 5B–D.

Annotation of protein structures with unknown function

We ran Pfinder on 31 proteins of unknown function whose structure was solved between 1 January 2009 and 1 March

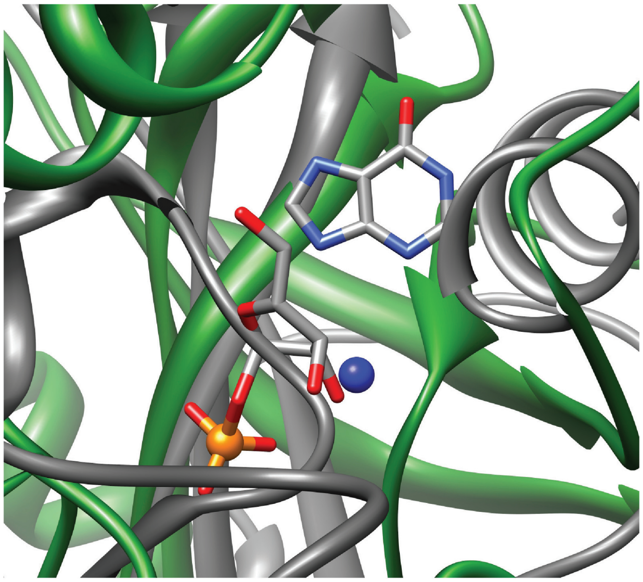

Figure 6. Superimposed ligand-binding sites from protein structures, represented in ribbon style, of Rv2714 protein from *M. tubercolosis* (PDB code 2WAM), colored in gray, and bovine purine nucleoside phosphorylase (PDB code 1A9T), colored in green. The predicted phosphate-binding site on the Rv2714 protein is represented as a blue sphere. The ligands of bovine purine nucleoside phosphorylase, a hypoxantine molecule and a ribose-1-phosphate molecule, are colored by atom type (carbon in gray, phosphorus in orange, oxygen in red, nitrogen in light blue).

2010 and whose PFAM records were available. The method predicted 72 PbSs on the 31 proteins, with an average of 2.3 predictions per structure. The results are available in the Supplementary Table S6. In one case, the Rv2714 protein from *Mycobacterium tubercolosis*, a paper describing the structure (PDB code 2WAM) has been published (40). The crystal structure is a trimer whose protomer is structurally similar to several enzymes spanning a range of diverse functions, from carboxypeptidases, to nucleosidases, to purine nucleoside phosphorylases (PNP). Two out of the nine PNP/nucleosidases detailed in the paper bound a ligand containing a phosphate group. One of these enzymes is the bovine purine nucleoside phosphorylase complexed with a hypoxantine molecule and a ribose-1-phosphate molecule (PDB code 1A9T). We used UCSF Chimera (29) to align the two protomers, resulting in a RMSD of 1.1 Å. After the superimposition one of the phosphate-binding sites we predicted in the Rv2714 protein is located close (5.8 Å) to the phosphate group of the 1A9T ligand (Figure 6). Therefore in this case Pfinder correctly predicted the approximate location of the phosphate-binding site in a protein of unknown function which was crystallized without any bound ligand. The protomer of the *E. coli* uridine phosphorylase structure (PDB code 1RXC) is also very similar to that of the Rv2714 protein. In this case too, after the structural superimposition (0.9 Å), the same predicted phosphate-binding site is located 6.4 Å away from the phosphate group of the ribose-1-phosphate molecule bound by the protein.

Fold distribution in the protein sets

Pfinder has been developed to predict binding sites for the phosphate group independently of the fold of the protein, to which it binds. We analyzed the fold distribution among all the proteins studied and among the structures comprising the PbM data set (Supplementary Table S7). A total of 34 out of 59 training proteins and 35 out of 52 test proteins have a fold that does not belong to the set of widespread nucleotide-binding folds such as the Rossmann-type folds and the P-loop containing nucleotide hydrolases (15, 41).

However, since our Training and Test sets are enriched in nucleotide-binding proteins (and so is the PDB), we wanted to exclude the possibility that PbMs are identified on those proteins due to an obvious global structural similarity. To demonstrate that the method could work on these proteins also in the absence of an overall fold similarity we determined the SCOP folds corresponding to the PbMs that produced TP predictions on the common nucleotide-binding fold (CNBF) structures. Supplementary Tables S8 and S9 show for each of these training and test set structures the number of TP matches that comes from PbM obtained from a non-CNBF. We found that 105 out of 111 structures have at least one TP prediction due to a PbM from a non-CNBF. This means that these binding sites would have been correctly identified even in the absence of an overall fold similarity.

DISCUSSION

In this work we presented Pfinder, the first method, to the best of our knowledge, that predicts PbS on a protein structure. The method works by comparing a protein structure with a data set of known PbMs. Subsequently geometric criteria and sequence conservation are used to filter the predictions, greatly improving the performance of the method. We have trained the method on a set of 59 high-quality structures of proteins complexed with phosphate-containing ligands. Pfinder correctly predicted at least one PbS in 41 of the analyzed chains. The amino acids conservation helps the method in discarding 80% of the FP predictions. After this filtering Pfinder produced 3.7 ± 0.4 FP on average per analyzed protein chain. Pfinder has been tested on a different set composed of apo-holo structures of the same proteins in order to determine how the ligand-induced conformational changes affect the performance of the method. We obtained comparable results between the apo and the holo Sets (62 and 63% of proteins, with at least one correctly identified PbS, respectively) showing that the method can be used to predict PbSs in functionally uncharacterized proteins with approximately the same accuracy obtained when using ligand-bound structures. The method produces the same number (4.8 ± 0.7 for both holo and apo sets) of FP predictions. Therefore more time-consuming methods such as flexible docking can be used to discriminate among the few proposed PbS.

We demonstrated that considering the position of protein clefts lowers the number of FP predictions to nearly a half, to the detriment of the few proteins in the apo test set for which the binding site is not part of any of the four largest pockets. We also showed that the predictions that do not overlap closely (2.5–5.0 Å distance) with the crystallized phosphate groups could represent alternative PbSs. Finally we applied the method to predict phosphate-binding sites on 31 protein structures that do not have an assigned function in order to make those predictions useful for their functional annotation.

To the best of our knowledge Pfinder is the only method for the prediction of phosphate-binding sites, since the existing predictive methods are more focused on binding sites for metals or other types of ligands. Therefore Pfinder constitutes a reference point for future methods and the results of this work can be easily extended to other types of chemical groups. The performance of the method could be improved by decreasing the number of structures that do not have any correct prediction. Therefore the main limitation of Pfinder seems to be the incomplete nature of the PbM templates set that does not contain all the existing PbMs because of its quality criteria. Indeed a PbM is included in the template data set if it is shared by at least two different SCOP folds. To this end we plan to extend the data set of templates and therefore increase the likelihood of finding binding motifs in a specific structure.

The comparison with other ligand-specific binding sites prediction methods is complicated by the very different nature of the ligands both in terms of size and physicochemical properties. Moreover differences in the protein structure data sets used and the criteria for determining the correct predictions complicate the direct comparison of performance values. Metal-binding sites are easier to predict due to their relatively fixed geometry. Indeed the methods devoted to the identification of metal-binding sites have a sensitivity of 87% or better (7,8). When the ligand is carbohydrate-like the sensitivity reaches 65–72% (9,10), even if the exact position of the ligand is not always precisely identified. Indeed even the most successful method (10) considers as TP predictions with an overlap of only 25% between the predicted and real binding pockets. The prediction of phosphate-binding sites by Pfinder attains values of sensitivity similar to those of carbohydrate prediction methods and also permits a precise positioning of the phosphate group (within 5Å).

Pfinder has also been tested (data not shown) on the same data set of nine phospho–peptide-binding domains used by the Joughin *et al.* method (13). Only one binding site has been correctly identified, suggesting that the phosphate group in phosphorylated peptides is recognized by a different set of PbMs. Indeed our data set of PbMs was derived from small-molecule ligands.

Pfinder was designed as a general method to analyze if PbMs can be recognised independently of the identity of the whole ligand. We believe that the main road to follow for a correct identification of a whole ligand may reside in the identification of further binding motifs, specifically associated to other portions of the ligand such as the ribose and nucleobase in the case of nucleotides (19).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Inter-Department Centre for Advanced Computing in Science and Technology (CAST) of the University of Rome Tor Vergata for the precious gift of computing time. The authors also want to thank Alessio Colantoni and Iolanda Mangone for their useful contribution in the analysis of, respectively, protein clefts and amino acids conservation.

FUNDING

European Commission under FP7, LEISHDRUG Project n° 223414 (to M.H.C.) and FIRB 'Futuro in ricerca' Project number E81J10000030001 (to G.A.). Funding for open access charge: EU FP7 project LEISHDRUG.

Conflict of interest statement. None declared.

REFERENCES

1. Hirsch,K.H., Fischer,F.R. and Diederich,F. (2006) Phosphate recognition in structural biology. Angewandte. *Chemie Int. Edn*, **46**, 338–352.
2. Traxler,P. and Furet,P. (1999) Strategies toward the design of novel and selective tyrosine kinase inhibitors. *Pharmacol. Ther.*, **82**, 195–206.
3. Gitlin,J.D. (2003) Wilson disease. *Gastroenterology*, **125**, 1868–1877.
4. Ji,H.F., Kong,D.X., Shen,L., Chen,L.L., Ma,B.G. and Zhang,H.Y. (2007) Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol.*, **8**, R176.
5. Saraste,M., Sibbald,P.R. and Wittinghofer,A. (1990) The P-loop–a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.*, **15**, 430–434.
6. Kleiger,G. and Eisenberg,D. (2002) GXXXG and GXXXA Motifs Stabilize FAD and NAD(P)-binding Rossmann Folds Through Cα–H . . . O Hydrogen Bonds and van der Waals Interactions. *J. Mol. Biol.*, **323**, 69–76.
7. Schymkowitz,J.W.H., Rousseau,F., Martins,I.C., Ferkinghoff-Borg,J., Stricher,F. and Serrano,L. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl Acad. Sci. USA*, **102**, 10147–10152.
8. Deng,H., Chen,G., Yang,W. and Yang,J.J. (2006) Predicting calcium-binding sites in proteins - a graph theory and geometry approach. *Proteins*, **64**, 34–42.
9. Taroni,C., Jones,S. and Thornton,J.M. (2000) Analysis and prediction of carbohydrates binding sites. *Protein Eng.*, **13**, 89–98.
10. Kulharia,M., Bridgett,S.J., Goody,R.S. and Jackson,R.M. (2009) InCa-SiteFinder: a method for structure-based prediction of inositol and carbohydrate binding sites on proteins. *J. Mol. Graph. Model.*, **28**, 297–303.
11. Ghersi,D. and Sanchez,R. (2009) EASYMIFS and SITEHOUND: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics*, **25**, 3185–3186.
12. Gherardini,P.F. and Helmer-Citterich,M. (2008) Structure-based function prediction: approaches and applications. *Brief. Funct. Genomic. Proteomic.*, **7**, 291–302.
13. Joughin,B.A., Tidor,B. and Yaffe,M.B. (2005) A computational method for the analysis and prediction of protein: phosphopeptide-binding sites. *Protein Sci.*, **14**, 131–139.

14. Swindells,M.B. (1993) Classification of doubly wound nucleotide binding topologies using automated loop searches. *Protein Sci.*, **2**, 2146–2153.
15. Vetter,I.R. and Wittinghofer,A. (1999) Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer. *Q. Rev. Biophys.*, **32**, 1–56.
16. Erlanson,D.A. (2006) Fragment-based lead discovery: a chemical update. *Curr. Opin. Biotechnol.*, **17**, 643–652.
17. Hajduk,P.J. and Greer,J. (2007) A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug. Discov.*, **6**, 211–219.
18. Ausiello,G., Gherardini,P.F., Gatti,E., Incani,O. and Helmer-Citterich,M. (2009) Structural motifs recurring in different folds recognize the same ligand fragments. *BMC Bioinformatics*, **10**, 182.
19. Gherardini,P.F., Ausiello,G., Russell,R.B. and Helmer-Citterich,M. (2010) Modular architecture of nucleotide-binding pockets. *Nucleic Acids Res.*, **38**, 3809–3816.
20. Brakoulias,A. and Jackson,R.M. (2004) Towards a structural classification of phosphate binding sites in protein–nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins*, **56**, 250–260.
21. Kinoshita,K., Sadanami,K., Kidera,A. and Go,N. (1999) Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. *Protein Eng.*, **12**, 11–14.
22. Ausiello,G., Via,A. and Helmer-Citterich,M. (2005) Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics*, **6(Suppl. 4)**, S5.
23. Gherardini,P.F., Ausiello,G. and Helmer-Citterich,M. (2010) Superpose3D: a local structural comparison program that allows for user-defined structure representations. *PLoS One*, **5**, e11988.
24. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
25. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
26. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F. Jr, Brice,M.D., Rodgers,J.R., Kennard,O. and Tasumi,T.S. (1978) The protein data bank: A computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.*, **185**, 584–591.
27. Wang,G. and Dunbrack,R.L. Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
28. Connolly,M.L. (1983) Analytical molecular surface calculation Connolly. *J. Appl. Cryst.*, **16**, 548–558.
29. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera–A visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
30. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunesekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38(Database Issue)**, D211–D222.
31. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
32. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
33. Zhao,S., Morris,G.M., Olson,A.J. and Goodsell,D.S. (2001) Recognition templates for predicting adenylate binding sites in proteins. *J. Mol. Biol.*, **314**, 1245–1255.
34. Dessailly,B.H., Lensink,M.F., Orengo,C.A. and Wodak,S.J. (2008) LigASite: a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.*, **36(Database Issue)**, D667–D673.
35. Schmitt,S., Kuhn,D. and And Klebe,G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.
36. Laskowski,R.A., Luscombe,N.M., Swindells,M.B. and Thornton,J.M. (1996) Protein clefts in molecular recognition and function. *Protein Sci.*, **5**, 2438–2452.
37. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.
38. Owen,D.J., Noble,M.E.M., Garman,E.F., Papageorigiou,A.C. and Johnson,L.N. (1995) Two structures of the catalytic domain of phosphorylase kinase: an active protein kinase complexed with substrate analogue and product. *Structure*, **3**, 467–482.
39. Morris,G.M., Goodsell,D.S., Halliday,R.S., Huey,R., Hart,W.E., Belew,R.K. and Olson,A.J. (1998) Automated docking using a Lamarckian genetic algorithm and and empirical binding free energy function. *J. Comput. Chem.*, **19**, 1639–1662.
40. Grana,M., Bellinzoni,M., Miras,I., Fiez-Vandal,C., Haouz,A., Shepard,W., Buschiazzo,A. and Alzari,P.M. (2009) Structure of Mycobacterium tuberculosis Rv2714, a representative of a duplicated gene family in Actinobacteria. *Acta Cryst.*, **F65**, 972–977.
41. Aravind,L., Mazumder,R., Vasudevan,S. and Koonin,E.V. (2002) Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.*, **12**, 392–399.