

Software

Open Access

## SitesIdentify: a protein functional site prediction tool

Tracey Bray<sup>1</sup>, Pedro Chan<sup>1</sup>, Salim Bougouffa<sup>1</sup>, Richard Greaves<sup>1</sup>,  
Andrew J Doig<sup>2</sup> and Jim Warwicker\*<sup>1</sup>

Address: <sup>1</sup>Faculty of Life Sciences, The University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK and <sup>2</sup>Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

Email: Tracey Bray - t.bray@postgrad.manchester.ac.uk; Pedro Chan - pedro.chan-5@postgrad.manchester.ac.uk; Salim Bougouffa - s.bougouffa-2@student.manchester.ac.uk; Richard Greaves - richard.b.greaves@gmail.com; Andrew J Doig - andrew.doig@manchester.ac.uk; Jim Warwicker\* - jim.warwicker@manchester.ac.uk

\* Corresponding author

Published: 18 November 2009

Received: 5 June 2009

BMC Bioinformatics 2009, 10:379 doi:10.1186/1471-2105-10-379

Accepted: 18 November 2009

This article is available from: <http://www.biomedcentral.com/1471-2105/10/379>

© 2009 Bray et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** The rate of protein structures being deposited in the Protein Data Bank surpasses the capacity to experimentally characterise them and therefore computational methods to analyse these structures have become increasingly important. Identifying the region of the protein most likely to be involved in function is useful in order to gain information about its potential role. There are many available approaches to predict functional site, but many are not made available via a publicly-accessible application.

**Results:** Here we present a functional site prediction tool (SitesIdentify), based on combining sequence conservation information with geometry-based cleft identification, that is freely available via a web-server. We have shown that SitesIdentify compares favourably to other functional site prediction tools in a comparison of seven methods on a non-redundant set of 237 enzymes with annotated active sites.

**Conclusion:** SitesIdentify is able to produce comparable accuracy in predicting functional sites to its closest available counterpart, but in addition achieves improved accuracy for proteins with few characterised homologues. SitesIdentify is available via a webserver at <http://www.manchester.ac.uk/bioinformatics/sitesidentify/>

### Background

Efforts, primarily by structural genomics groups, have provided a rapidly growing number of protein structures with little or no functional annotation. This has caused new interest in the relationship between structure and function and has increased focus on ways to elucidate a protein's function from its structure rather than solely from sequence. In order to investigate the role of a protein using its structure, it is useful to be able to identify the portion of the protein that is most closely involved with its

function. In the case of enzymes this is its active site, whilst non-enzymes have functionally important regions that are involved in ligand-binding or protein-protein interactions.

There are currently several computational approaches that predict functional sites which use either structural or sequence information. The most widely used methods rely on sequence information in order to predict functionally important residues, due to the greater availability of

sequence data as opposed to structural data for uncharacterised proteins. Sequence based methods mainly centre around the concept of functionally important residues being more highly conserved through evolution and identify the most conserved residues by comparing positions in a multiple sequence alignment with homologous proteins. Some methods use only sequence conservation information in making predictions [1,2], whilst others also include additional computed sequence features [3], or structural properties predicted from sequence such as predicted secondary structure and solvent accessible surface area [4,5], particularly in order to distinguish between residues conserved for function and those conserved for structure [6,7]. Many methods focus on predicting catalytic residues in enzyme active sites, but measures of sequence conservation have also been successfully used to predict residues in contact with a ligand [1,5,8] or in contact with other proteins, although sequence conservation has been shown to perform less well as a predictive feature in the latter cases [1,9].

Whilst there are a large number of sequence-based methods available, there are also a growing number of methods that predict functional sites based on structural information. These methods fall into two main categories: those that identify structural similarities and transfer annotation from a protein with a known functional site and those that predict functional sites by non-homology related structural features such as geometrical or electrostatic properties [5,10,11].

There are many resources that store structural and sequence information about proteins with known active sites, such as PdbFun [12], CSA [13], PDBSite [14] and ProSite [15]. A protein of unknown active site location can be compared to these resources (CSS [16] scans the CSA and PDBSiteScan [17] scans PDBSite), or to databases derived specifically for the prediction method, to identify any structural similarities with known active sites [18-25]. While these methods often produce accurate results, they assume the existence of a functionally annotated homologue of similar active site structure in their respective databases. As one of the aims of structural genomics initiatives is to obtain structures for proteins that occupy remote fold space, these methods may be of limited use for such proteins.

In this situation, *ab initio* methods that do not rely on the existence of a functionally characterised homologue may be of more value. A wide range of structural properties have been used, showing that the relationship between a protein's structure and its function is affected by many structural characteristics. A study of catalytic residues and their properties [26] showed that they are likely to exist in regions of the protein that are not in helix or sheet second-

ary structure, have a higher propensity to be a charged residue and exhibit lower B-values than non-catalytic residues. A number of methods have used these characteristics to predict residues involved in catalysis [27,28]. Bartlett et al. noted that catalytic residues tend to line the surface of large surface clefts, yet remain relatively buried within the protein geometry. It was also observed in a study of 67 single-chain enzymes that 83% of enzyme active sites are found in the largest surface cleft [29], resulting in methods to predict active sites by finding surface clefts [30,31].

Previous work by this group [32] attempted to identify functional sites by locating peak electrostatic potentials near to the surface of a protein resulting from the interaction of charged residues that are under electrostatic strain. The greatest functional site prediction accuracy, however, was obtained by applying a uniform charge weighting across the protein rather than using actual charges. This uniform charge weighting essentially acts as a cleft-finding algorithm and will predict the most buried surface cleft. This gave a prediction accuracy of 77%, where a successful prediction is when the peak potential was within 5% of the protein surface from the real active site centre.

Other studies have successfully used electrostatics calculations to predict active site and ligand-binding site residues [33-37]. Elcock identified residues that had destabilizing effects on the stability of the protein using continuum electrostatics methods and found that these correlated with residues involved in protein functionality [33]. This method, however, was not tested on a large experimentally annotated dataset and so it is hard to interpret the degree of accuracy it achieved. Another approach predicts enzyme active sites by identifying residues with unusually-shaped titration curves [35,38] as well as predicting enzyme function [39]. Other chemistry-based approaches, such as identifying residues that are unusually hydrophobic for their position in a structure have also been successful [40].

Other *ab initio* methods use the degree of connectivity of residues to predict those involved in function. A number of methods assess the closeness centrality of residues [41-43], whilst one study found that catalytic residues are more likely to exist in close proximity to the molecular centroid [44].

Perhaps the best accuracies can be achieved by combining structural approaches and sequence conservation. Residues may be evolutionarily conserved due to structural as well as functional constraints and a number of studies have attempted to distinguish these two factors by considering the degree of conservation and the residue's structural environment [6,45]. Mapping the degree of

evolutionary conservation onto the structure is useful in identifying clusters of conserved residues in the structure that may indicate a functional site [46,47]. Combining the types of structural information used in *ab initio* structural methods with sequence conservation can be effective [10,11,34,48,49].

Despite the success of the large number of varied approaches, only a relatively small subset of these methods are currently available either via a software package or a web-server. Tools report various levels of accuracy that are difficult for a user to compare due to their separate test datasets, outputs and reporting methods. Here we present a user-friendly functional site prediction tool, SitesIdentify, based on previously published work by this group [10,32]. This is made publicly available via a web-server [50], and is compared to other accessible tools in a comparison of performance on a common dataset.

## Implementation

### Functional Site Prediction Methods

SitesIdentify can predict functional site location by two separate approaches, which have been described in more detail in previous publications [10,32]. In brief, the first method [32] places a 2Å grid over the protein structure and applies a uniform charge to each non-hydrogen atom. The electrostatic potential is calculated using Finite Difference Poisson-Boltzmann calculations with no dielectric boundary. The peak potential is predicted as the centroid of the functional site.

The second method [10] combines the electrostatics method used above with sequence conservation information. Close homologues are found by running the sequence through PSI-BLAST with an E value cut-off of 1e-20. A normalised conservation score is calculated for each residue based on the amino acid and stereochemical diversity and the gap occurrence at that position,  $C(x) = (1-t(x))^\alpha(1-r(x))^\beta(1-g(x))^\gamma$ , where  $t$  is the normalised symbol diversity,  $r$  is the normalised stereochemical diversity (based on the BLOSUM-62 matrix) and  $g$  the gap cost. Each of these terms are weighted by integral values ranging between 0 and 5 ( $\alpha$ ,  $\beta$  and  $\gamma$ ), the values for which are defined as those giving the best predictive performance in the original publication [10]. The peak potential is then calculated in the same way as the first method, but now with a single central atom in each amino acid weighted with the conservation scores.

### SitesIdentify Workflow

Upon submission of a job, SitesIdentify starts a number of programs depending on which method the user requested. If the conservation approach is selected, the in-house Conserved Residue Colouring program(CRC) is run first, which identifies homologues by running the

sequence contained in the SEQRES records in the PDB file through PSI-BLAST [51]. PSI-BLAST is run for one iteration (in default settings) on the non-redundant database with an E-value cut-off for inclusion of sequences of 1e-20. A profile file containing the conservation scores for each residue is produced. SitesIdentify uses the conservation scores as charge weightings on a single atom for each amino acid ( $C_\beta$  or  $C_\alpha$  for glycine), and calculates the location of the peak potential as described above [10]. If no homologue can be identified for a protein using CRC then the method automatically switches to only charge-based calculations. If the conservation method is not selected then the CRC program is omitted and the location of the peak potential is calculated using the uniform charge-weighting method [32]. A sphere of user-supplied radius is drawn around the predicted centroid coordinates and residues are selected that have at least one atom within that sphere and also exhibit more than 5Å<sup>2</sup> of solvent-accessible surface area (SASA) as calculated using the Lee and Richards method [52]. This list of residues represents the predicted functional site, which is given in the results as a text list and also highlighted on the PDB structure using Jmol [53].

### SitesIdentify Usage

SitesIdentify is available for use via a web browser and is freely accessible without license or an account registration. The main web page allows a user to enter either a pre-existing PDB structure ID (and whether to use the biological unit or the asymmetric unit) or upload a structure file, the radius around the predicted site to use, the method to use and an email address so that a user can be notified and emailed the results link upon job completion.

If a user has submitted their own structure file then this is validated to ensure that contains an acceptable PDB-format structure, the rules for which are given in the user guide available from the website. The file must be less than 2 MB in size and contain only text. It also must contain at least SEQRES and ATOM records and be spaced exactly as the standard PDB format. If the user-supplied information is invalid (non-existent PDB ID or invalid email address) then the job is not initialized and the user informed of the incorrect information via the browser. Upon successful completion of a job the web-server directs the user to the results page and also sends an email to the user at the address specified with a link to the results page.

## Results and Discussion

### SitesIdentify Web-server

SitesIdentify is available to run for single protein entries at <http://www.manchester.ac.uk/bioinformatics/sitesidentify/> or can be downloaded to run offline for multiple pro-

teins (Additional File 1). It requires some basic user-input via a web-browser (see Figure 1). Once this information is validated a new job is initiated. The average calculation time per protein is approximately 6 minutes when using the method including conservation information and approximately 2 minutes if only using charge-based calculations. If the protein takes longer than 45 minutes to produce results, which may occur for very large proteins, the job is terminated and the user is notified by email.

Upon completion of a job an email is sent to the user at the address specified which provides a link to the results page. The results page displays a Jmol applet illustrating the protein structure with the predicted site residues highlighted, a text list of the predicted residues and a link to a text file containing the predicted residue information (see Figure 2 for an example).

The methods used in SitesIdentify can distinguish between enzyme and non-enzyme with a high degree of accuracy [32] and so an enzyme/non-enzyme prediction is also given along with the functional site prediction. Cleft size has also been used as a discriminator between

enzyme and non-enzyme with enzymes more likely to exhibit large surface accessible clefts than non-enzymes [54]. Since the charge-based method essentially identifies buried clefts it is likely to perform better for enzymes than non-enzymes, although it still may be able to detect small ligand-binding pocket clefts in non-enzymes. In addition, the second SitesIdentify method incorporates sequence conservation information which has also been shown to be useful in predicting other biologically important regions such as non-enzyme ligand binding sites [49], protein-protein interaction sites [55-57] and DNA-binding sites [58]. It is worth noting however, that a study of four non-enzyme families by Magliery et al. found that rather than binding sites being conserved, they showed a higher degree of variation than the rest of the protein [8]. This may explain why some conservation approaches report better accuracies in predicting functional sites of enzymes than non-enzymes [49,59].

It is unsurprising therefore that SitesIdentify performs better for enzymes than non-enzymes although it is still able to identify non-enzyme ligand-binding sites with comparable accuracies to other non-enzyme specific functional

### ENTER THE PDB IDENTIFIER

The PDB ID:	<input type="text"/>	<b>Example: 12as (lower-case)</b>
Asymmetric unit or biological unit?:	<input type="text" value="Asymmetric Unit"/>	<b>Default: Asymmetric Unit</b>
Radius (in Angstroms):	<input type="text"/>	<b>Default: 10 Angstroms.</b>
Choose which Method to use	Charge-based calculations only: <input type="radio"/>	<b>Default: With sequence profiles</b>
	With sequence profiles: <input checked="" type="radio"/>	
Your Email Address:	<input type="text"/>	<input type="button" value="Submit"/>

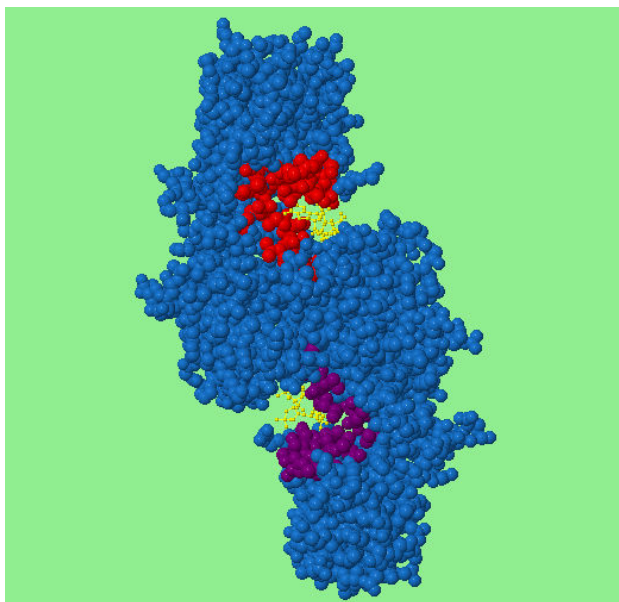
### OR UPLOAD YOUR OWN PDB FILE

Your PDB file:	<input type="text"/>	<input type="button" value="Browse..."/>	<b>Example: 12as.pdb</b>
Radius (in Angstroms):	<input type="text"/>		<b>Default: 10 Angstroms.</b>
Choose which Method to use	Charge-based calculations only: <input type="radio"/>		<b>Default: With sequence profiles</b>
	With sequence profiles: <input checked="" type="radio"/>		
Your Email Address:	<input type="text"/>	<input type="button" value="Submit"/>	

**Figure 1**

**Screenshot showing the required user input fields.** A user can either input a pre-existing PDB code and whether to use the asymmetric or biological unit structure or upload their own PDB-style structure file. All fields are compulsory.





**Figure 3**  
**An example of highlighted residues in an alternative predicted site.** The biological unit structure for 2af4 (phosphotransacetylase) is a homodimer and identical active sites are present on both chains. SitesIdentify identifies only one site (in red), but the annotation is transformed onto the other chain in order to identify the other active site (shown in purple).

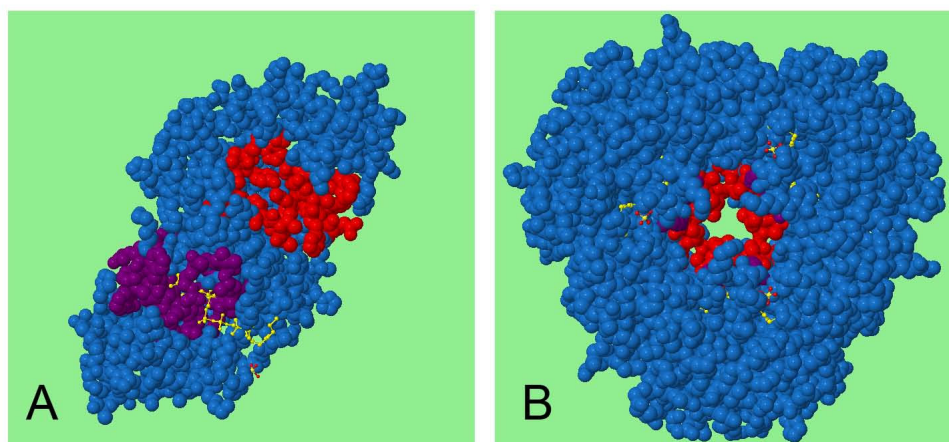
unit, running SitesIdentify on the asymmetric unit may fail to give the correct prediction.

Some biological units, however, may give a false prediction particularly where there is an internal void formed by a cyclical arrangement of subunits. Such voids tend to be well-buried, more so than the real surface clefts, and the residues on the edges of these voids may be evolutionarily conserved in order to retain the quaternary structure. These voids are therefore sometimes incorrectly selected as predicted functional sites, and so where a biological unit has an internal void it would be useful to also run SitesIdentify on the asymmetric unit. For example, running the asymmetric unit for 1B6T through the SitesIdentify server locates the functional site in the correct location, however the site is predicted incorrectly for the biological unit as the void formed in the centre of the molecule (see Figure 4).

#### Comparison to Other Applications

It is difficult to objectively compare the accuracy levels reported by the authors of the various existing functional site prediction tools as they use different datasets and report predictions differently. Some tools predict only the residues specifically involved in the protein function (e.g. catalysis) whilst others predict all residues in and around the functional site. Here, we have attempted to compare suitable methods on a common dataset of non-redundant proteins with known functional sites.

Some methods only predict enzyme active sites whilst others can identify functional residues in other types of proteins (for example PDBSiteScan and Q-SiteFinder).



**Figure 4**  
**An example of differential site prediction between asymmetric and biological unit structures.** The active site predicted for the asymmetric unit of 1b6t (phosphopantetheine adenylyltransferase) is reasonably close to the bound ligand shown in part A. The biological unit is formed by a cyclical arrangement of the asymmetric unit and when SitesIdentify is run on this structure it incorrectly identifies the central void as the enzyme active site (part B).

Enzyme active sites are the most easily defined functional sites in proteins and are the most common prediction targets for these tools; therefore the dataset we have used here contains enzymes with known catalytic residues (results for an analysis on a small set of non-enzymes is available in Additional File 2). The Catalytic Sites Atlas (CSA) [26] is a valuable resource for storing information about catalytic residues that are annotated from literature and at the time of creation of this dataset (November 2008) it contained 880 enzymes with literature-annotated catalytic residues (version 2.2.1). These were then culled for redundancy to ensure that no two structures contained an active-site domain from the same SCOP [60] superfamily (detail of this culling procedure has been reported [61]). This produced a non-redundant set of 237 enzymes for which there are annotated catalytic residues (see Additional File 3 for the list of PDB codes).

In order to be included in this analysis, a method had to adhere to the following criteria:

- The method must require no prior knowledge about the active site.
- It produces output that identifies the active site either by a coordinate location, the identities of catalytic residues or identities of residues found in the binding site.
- It produces results within a reasonable time scale. The method should return results for a test protein with 330 residues in 10 minutes or less.
- It does not simply access known annotation about the test protein.

The applications that met these criteria are listed in Table 1. Other applications that were considered but were not included in this study, along with the reason for not including them, are listed in Additional File 4. Where a method only accepts one chain from a PDB structure, the first chain is used. All predictions are run on the asymmetric unit structure.

In order to put predictions into the same context as those given by SitesIdentify, a central PDB coordinate point is calculated for each prediction given by each method. For example, if a method only predicts catalytic residues, the central coordinate point (centroid) is defined as the geometric average of the  $C_{\beta}$  atom ( $C_{\alpha}$  for glycine) coordinates of the catalytic residues. Similarly to the SitesIdentify output, a sphere with a  $10\text{\AA}$  radius is drawn around this centroid and residues are selected if they have at least one atom within this radius and also have a SASA of  $5\text{\AA}^2$  or

more. These residues are termed the standardised predicted residues.

There are three measures of accuracy used in this analysis. The first is the average percentage of annotated catalytic residues for each protein that are included in the standardised predicted residues (average absolute recall rate). Second is the average absolute recall rate for the method divided by the absolute recall rate of catalytic residues returned by the real centroid (the average relative recall rate). Third is simply the Cartesian distance from the real centroid and the predicted centroid.

It is more representative to consider the relative recall rate for each method as opposed to the absolute recall rate as for some proteins less than 100% of the annotated catalytic residues are recalled by selecting residues that have at least  $5\text{\AA}^2$  SASA within a  $10\text{\AA}$  radius. It is therefore unlikely for these proteins that even a very accurate prediction would give an absolute recall rate of 100%.

The prediction accuracies achieved for each method are shown in Table 2 and comparison of the distances between the predicted centroid and real centroid for each method are shown in Figure 5.

The conservation-based method of SitesIdentify achieved an average relative recall rate of 74.7%, which is comparable to that of the method with the highest accuracy, Consurf (78.1%). In order to extract site predictions for Consurf, all residues with a conservation score of 9 were assumed to be functional residues. For structures with more than one chain residue predictions were taken from the first chain only in order to avoid calculating the incorrect active site centroid from separate sites on multiple chains. Consurf was therefore effectively run on monomer structures rather than the true asymmetric unit. It is worth noting that when SitesIdentify is also run on monomer structures formed from only the first chain in the file it achieves a very similar performance to Consurf (see Figure 6).

Both Consurf and SitesIdentify are based around predicting conserved residues as functional site residues but whilst Consurf appears to perform slightly better overall, it could not produce predictions for three of the proteins in the set (1C3I, 1DMU and 1PGS) as it was unable to identify enough homologues. SitesIdentify uses both a combination of residue conservation information with an electrostatics-based cleft-finding algorithm and so it still gives predictions where there is little or no conservation information available. SitesIdentify was able to recall 100% of the annotated catalytic residues for the three proteins in this set for which Consurf did not make any prediction. SitesIdentify, therefore, is likely to give better

**Table 1: Functional site prediction tools included in the comparison analysis.**

Application	Method Category	Description	Reference
<b>SitesIdentify</b>			
Uniform charge method	CF	A uniform charge weighting is applied to each C $\alpha$ atom on the protein and the electrostatic potential (Finite Difference-Poisson-Boltzmann calculation with no dielectric boundary) is sampled at points on a 2 grid across the protein volume. The peak potential indicates the position of the predicted active site.	Bate and Warwicker, (2004)
Conservation method	SC, CF	As for the above method, except that the charge weightings applied across the protein are replaced with conservation weights derived from normalised sequence profile scores reflecting the amino acid diversity, the stereochemical diversity and the gap occurrence.	Greaves and Warwicker, (2005)
<b>Consurf</b>	SC	Consurf calculates the degree of evolutionary conservation for each residue in a structure and gives them an integer score from 1 to 9, with 9 being the most conserved residues. A graphical representation of the structure is then coloured according to these residue conservation scores, which allows visual identification of highly conserved patches that are predicted to be functional sites.	Landau et al. (2005)
<b>Crescendo</b>	SC	Predicts active sites by identifying clusters of residues that have higher than usual evolutionary restraint. Evolutionary constraint was identified by three measures: 1) whether there was a higher degree of evolutionary conservation than expected at a position, 2) whether environment specific substitution tables made weak predictions of the amino acid substitution patterns, and 3) residues that have spatially conserved positions when structures of proteins within the same family are superimposed.	Chelliah et al. (2004)
<b>FOD</b>	HP	The active site residues are predicted to be those with the highest hydrophobic deficiency score. This is the difference between the expected hydrophobicity and the observed hydrophobicity value for each residue. The expected hydrophobicity of a residue is determined by a residues relative position to the theoretically most hydrophobic point in the protein. The observed hydrophobicity is a combination of the hydrophobicity value of that residue and the effect on the residues position of other sidechains around it.	Brylinski et al. (2007)
<b>Q-SiteFinder</b>	CF	Non-bonded interaction energies are calculated by placing a 3D grid over the whole protein and then evaluating the interaction energy between the protein and a methyl group at each point on the grid. The positions of the probes on the grid that gave the best interaction energies were then spatially clustered to identify groups of close probes. These clusters are then assigned a single interaction energy based on the energies of their member probes. The clusters are then ranked by their representative interaction energy and the highest ranked cluster is predicted as the active site.	Laurie and Jackson (2005)
<b>PDBSiteScan</b>	TM	PDBSiteScan takes 3D fragments of a protein structure and compares them to 3D structure fragments of known active sites. The known active sites structures are held in a collection called PDBSite that is formed from annotation in the PDB SITE field and also REMARK 800 fields. Results were discounted if they compared to annotation held for the test protein.	Ivanisenko et al. (2004)
<b>PASS</b>	CF	PASS (Putative Active Site Spheres) is essentially a geometric cleft-finding method. The shape, volume and depth of the cleft determine which clefts are predicted as active site clefts.	Brady and Stouten (2000)
<b>Thematics</b>	CP	Thematics identifies ionisable residues with unusually perturbed titrations curves. Active sites are predicted where two or more of these ionisable residues form a cluster in 3D space.	Wei et al. (2007)

A description of the seven tools used in this analysis along with a brief description of each method. Method categories are as follows: CF = cleft-finding, SC = sequence conservation, HP = hydrophobicity, TM = structural template matching, CP = chemical properties.

predictions for structures from uncharacterised families, such as those being generated by structural genomics initiatives.

As discussed previously, residue conservation is known to be less indicative of functionality for non-enzymes than for enzymes[8,49,59], and here purely conservation-

based approaches, such as Consurf and Crescendo, achieved a lower average recall rate compared to both SitesIdentify methods on a small set of non-enzymes (see Additional File 2).

PDBSiteScan achieved the lowest absolute and relative recall rates (28.1% and 38.4%, respectively) and also the

**Table 2: Prediction accuracies achieved for each functional site prediction method.**

Method	Absolute Recall Rate	Relative Recall Rate	Average Distance between Predicted and Real Centroid (Å)
SitesIdentify			
Uniform charge method	47.6%	63.0%	11.2
Conservation method	56.9%	74.7%	9.4
Consurf	58.6%	78.2%	8.2
Crescendo	46.9%	63.8%	10.3
FOD	39.7%	56.1%	10.6
QSiteFinder	40.1%	53.0%	13.0
PDBSiteScan	28.1%	38.4%	15.5
PASS	36.6%	49.3%	14.8
Thematics	35.8%	48.9%	13.5

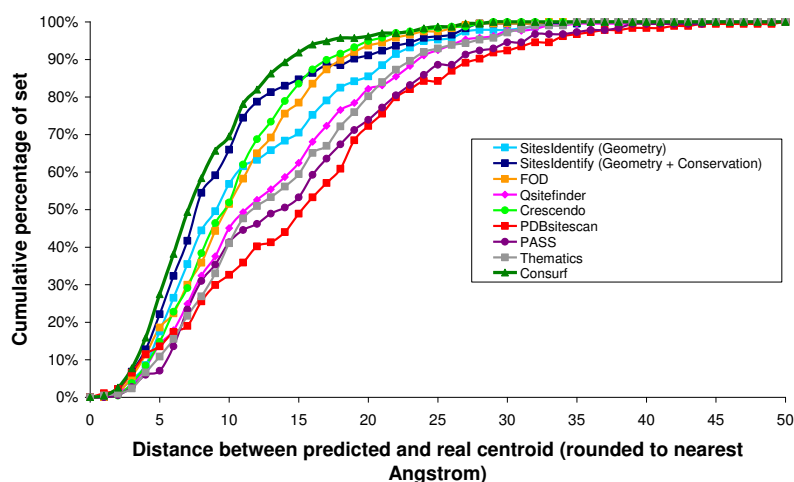
The absolute and relative recall rates achieved along with the average distance between real and active site centroids for each method.

largest average distance between predicted and real active-site centroids (15.5Å). PDBSiteScan scans the query protein against proteins of known annotation. In this analysis the test set consists of enzymes with known annotation and therefore it was necessary to reject predictions that simply accessed the annotation of any of these test proteins. As the number of proteins with well-characterised active site information is limited, removing these proteins from the set that PDBSiteScan compares to will obviously reduce the prediction power of the method. If tested on proteins outside of this set (i.e. proteins with uncharacterised functional sites) the prediction accuracy may increase.

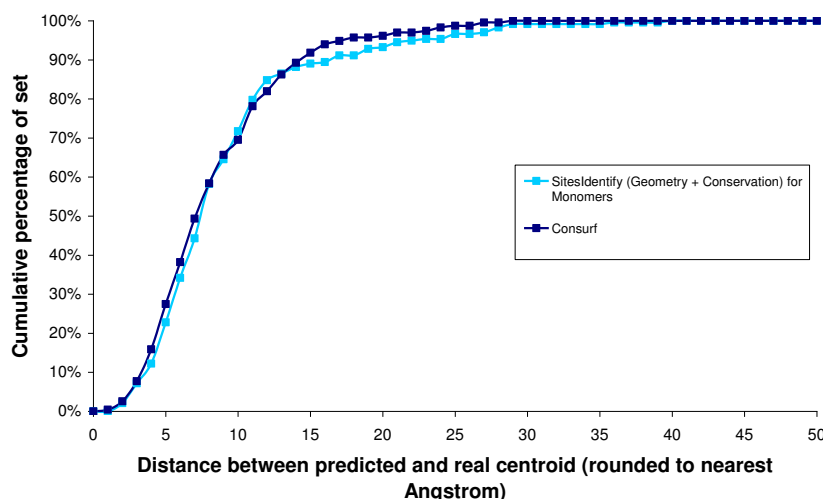
Q-SiteFinder identifies energetically favourable methyl binding sites by calculating the interaction energy between the protein and a methyl probe and then ranking

clusters of probes by their total interaction energy. Similar to the electrostatics-based method of SitesIdentify, Q-SiteFinder is essentially a cleft-finding algorithm. Despite similar approaches the uniform charge method of SitesIdentify achieves a 10% higher relative recall rate than Q-SiteFinder. Both Q-SiteFinder and SitesIdentify performed better than the other cleft-finding method, PASS, which also selects for cleft depth. Since SitesIdentify implicitly detects the atom density around a cleft rather than the cleft geometry itself, it suggests that this may be a contributing factor to the increased accuracy over PASS.

It is interesting that whilst SitesIdentify (charge-based) and Crescendo use very different approaches they give very similar accuracies on this dataset, suggesting that both conservation and geometrical information are equally useful in identifying functional sites. The combi-

**Figure 5**

**Comparison of distances between the real centroid and the predicted centroid for each method.** The cumulative percentage of the set that have differences between the real and predicted active site centroids at each distance are shown for each method.



**Figure 6**  
**Comparison of distances between the real centroid and the predicted centroid for Consurf and SitesIdentify run on monomer structures.** The cumulative percentage of the set that have differences between the real and predicted active site centroids at each distance are shown for both methods.

nation of both of these approaches in the conservation-based method of SitesIdentify further improves the accuracy achieved by either one alone.

### Conclusion

Here we present a functional site prediction tool, SitesIdentify. We have shown that this tool compares favourably to other available functional site prediction tools in a comparison of methods on a non-redundant set of 237 enzymes with annotated active sites. The combination of structure-based and conservation-based approach in this tool produces accurate results, whilst a non-conservation based approach is also available for proteins that perhaps occupy remote fold-space and have no closely related homologues. Such methods are useful for identifying functional sites, and therefore informing about potential protein function, for structures of uncharacterised proteins.

### Availability and Requirements

**Project name:** SitesIdentify

**Project home page:** <http://www.manchester.ac.uk/bioinformatics/sitesidentify/>

**Operating system(s):** Platform independent

**Programming language:** PHP, Perl, Fortran, Jmol, Javascript.

**Other requirements:** e.g. Javascript enabled web browser

**License:** Free for all users

**Any restrictions to use by non-academics:** None

### Authors' contributions

TB carried out the comparison analysis, created the web-server application and wrote the manuscript, whilst JW supplied electrostatics code, PC and RG supplied conservation calculation code and PC and SB provided some website code. JW and AJD directed the design of the application and critically revised the manuscript. All authors read and approved the final version.

### Additional material

#### Additional file 1

*SitesIdentify source code.* Compressed file containing the source code for SitesIdentify.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-379-S1.gz>]

#### Additional file 2

*Non-enzyme ligand binding comparison.* A table showing the prediction accuracies achieved for each functional site prediction method on 13 non-redundant non-enzyme structures with bound ligands from the Q-Site-Finder test set (Laurie and Jackson, 2005).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-379-S2.doc>]

**Additional file 3**

*PDB ID codes for the test dataset. A list of all PDB ID codes for the structures used in the comparison test.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-379-S3.doc>]

**Additional file 4**

*Functional site prediction tools not included in the comparison analysis. A list of the functional site prediction tools not used in the comparison analysis and the reason for their non-inclusion.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-379-S4.doc>]

**Acknowledgements**

Thank you to Nidhi Tyagi, Andrew Cawley, James Kitchen and Simon Lovell for helpful discussion. We would also like to thank members of Dr Mary Jo Ondrechen and Dr Simon Lovell's groups for help in obtaining output from their respective tools. We are grateful to the BBSRC for providing funds for a PhD studentship (TB) and to the Algerian Ministry of Higher Education for the award of PhD funding to SB.

**References**

- Capra JA, Singh M: **Predicting functionally important residues from sequence conservation.** *Bioinformatics* 2007, **23(15)**:1875-1882.
- Manning JR, Jefferson ER, Barton GJ: **The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction.** *BMC Bioinformatics* 2008, **9**:51.
- Zhang T, Zhang H, Chen K, Shen S, Ruan J, Kurgan L: **Accurate sequence-based prediction of catalytic residues.** *Bioinformatics* 2008, **24(20)**:2329-2338.
- Fischer JD, Mayer CE, Soding J: **Prediction of protein functional residues from sequence by probability density estimation.** *Bioinformatics* 2008, **24(5)**:613-620.
- Liang S, Zhang C, Liu S, Zhou Y: **Protein binding site prediction using an empirical scoring function.** *Nucleic Acids Res* 2006, **34(13)**:3698-3707.
- Chelliah V, Chen L, Blundell TL, Lovell SC: **Distinguishing structural and functional restraints in evolution in order to identify interaction sites.** *J Mol Biol* 2004, **342(5)**:1487-1504.
- Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N: **ConSeq: the identification of functionally and structurally important residues in protein sequences.** *Bioinformatics* 2004, **20(8)**:1322-1324.
- Magliery TJ, Regan L: **Sequence variation in ligand binding sites in proteins.** *BMC Bioinformatics* 2005, **6**:240.
- Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES: **Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Sci* 2004, **13(1)**:190-202.
- Greaves R, Warwicker J: **Active site identification through geometry-based and sequence profile-based calculations: burial of catalytic clefts.** *J Mol Biol* 2005, **349(3)**:547-557.
- Wang K, Horst JA, Cheng G, Nickle DC, Samudrala R: **Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information.** *PLoS Comput Biol* 2008, **4(9)**:e1000181.
- Ausiello G, Zanzoni A, Peluso D, Via A, Helmer-Citterich M: **pdb-Fun: mass selection and fast comparison of annotated PDB residues.** *Nucleic Acids Res* 2005:W133-137.
- Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004:D129-133.
- Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA: **PDBSite: a database of the 3D structure of protein functional sites.** *Nucleic Acids Res* 2005:D183-187.
- Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuče BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ: **The 20 years of PROSITE.** *Nucleic Acids Res* 2008:D245-249.
- Torrance JW, Bartlett GJ, Porter CT, Thornton JM: **Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families.** *J Mol Biol* 2005, **347(3)**:565-581.
- Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA: **PDB-SiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins.** *Nucleic Acids Res* 2004:W549-554.
- Binkowski TA, Freeman P, Liang J: **pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins.** *Nucleic Acids Res* 2008:D245-249.
- Chang DT, Weng YZ, Lin JH, Hwang MJ, Oyang YJ: **Protomot: prediction of protein binding sites with automatically extracted geometrical templates.** *Nucleic Acids Res* 2006:W303-309.
- Jambon M, Andrieu O, Combet C, Deleage G, Delfaud F, Geourjon C: **The SuMo server: 3D search for protein functional sites.** *Bioinformatics* 2005, **21(20)**:3929-3930.
- Kleywegt GJ: **Recognition of spatial motifs in protein structures.** *J Mol Biol* 1999, **285(4)**:1887-1897.
- Shulman-Peleg A, Nussinov R, Wolfson HJ: **SiteEngines: recognition and comparison of binding sites and protein-protein interfaces.** *Nucleic Acids Res* 2005:W337-341.
- Stark A, Russell RB: **Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures.** *Nucleic Acids Res* 2003, **31(13)**:3341-3344.
- Kristensen DM, Chen BY, Fofanov VY, Ward RM, Lisewski AM, Kimmel M, Kaviraki LE, Lichtarge O: **Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity.** *Protein Sci* 2006, **15(6)**:1530-1536.
- Goyal K, Mohanty D, Mande SC: **PAR-3D: a server to predict protein active site residues.** *Nucleic Acids Res* 2007:W503-505.
- Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: **Analysis of catalytic residues in enzyme active sites.** *J Mol Biol* 2002, **324(1)**:105-121.
- Tseng YY, Liang J: **Predicting enzyme functional surfaces and locating key residues automatically from structures.** *Ann Biomed Eng* 2007, **35(6)**:1037-1042.
- Tang YR, Sheng ZY, Chen YZ, Zhang Z: **An improved prediction of catalytic residues in enzyme structures.** *Protein Eng Des Sel* 2008, **21(5)**:295-302.
- Laskowski RA, Luscombe NM, Swindells MB, Thornton JM: **Protein clefts in molecular recognition and function.** *Protein Sci* 1996, **5(12)**:2438-2452.
- Gutteridge A, Bartlett GJ, Thornton JM: **Using a neural network and spatial clustering to predict the location of active sites in enzymes.** *J Mol Biol* 2003, **330(4)**:719-734.
- Brady GP Jr, Stouten PF: **Fast prediction and visualization of protein binding pockets with PASS.** *J Comput Aided Mol Des* 2000, **14(4)**:383-401.
- Bate P, Warwicker J: **Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods.** *J Mol Biol* 2004, **340(2)**:263-276.
- Elcock AH: **Prediction of functionally important residues based solely on the computed energetics of protein structure.** *J Mol Biol* 2001, **312(4)**:885-896.
- Ota M, Kinoshita K, Nishikawa K: **Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation.** *J Mol Biol* 2003, **327(5)**:1053-1064.
- Tong W, Williams RJ, Wei Y, Murga LF, Ko J, Ondrechen MJ: **Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines.** *Protein Sci* 2008, **17(2)**:333-341.
- Dessailly BH, Lensink MF, Wodak SJ: **Relating destabilizing regions to known functional sites in proteins.** *BMC Bioinformatics* 2007, **8**:141.
- Laurie AT, Jackson RM: **Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.** *Bioinformatics* 2005, **21(9)**:1908-1916.
- Wei Y, Ko J, Murga LF, Ondrechen MJ: **Selective prediction of interaction sites in protein structures with THEMATICS.** *BMC Bioinformatics* 2007, **8**:119.

39. Ondrechen MJ, Clifton JG, Ringe D: **THEMATICS: a simple computational predictor of enzyme function from structure.** *Proc Natl Acad Sci USA* 2001, **98(22)**:12473-12478.
40. Brylinski M, Prymula K, Jurkowski W, Kochanczyk M, Stawowczyk E, Konieczny L, Roterman I: **Prediction of functional sites based on the fuzzy oil drop model.** *PLoS Comput Biol* 2007, **3(5)**:e94.
41. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanel D, Venger I, Pietrovski S: **Network analysis of protein structures identifies functional residues.** *J Mol Biol* 2004, **344(4)**:1135-1146.
42. del Sol A, Fujihashi H, Amoros D, Nussinov R: **Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families.** *Protein Sci* 2006, **15(9)**:2120-2128.
43. Chea E, Livesay DR: **How accurate and statistically robust are catalytic site predictions based on closeness centrality?** *BMC Bioinformatics* 2007, **8**:153.
44. Ben-Shimon A, Eisenstein M: **Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces.** *J Mol Biol* 2005, **351(2)**:309-326.
45. Cheng G, Qian B, Samudrala R, Baker D: **Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design.** *Nucleic Acids Res* 2005, **33(18)**:5861-5867.
46. Landgraf R, Xenarios I, Eisenberg D: **Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins.** *J Mol Biol* 2001, **307(5)**:1487-1502.
47. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N: **ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.** *Nucleic Acids Res* 2005:W299-302.
48. Thibert B, Bredesen DE, del Rio G: **Improved prediction of critical residues for protein function based on network and phylogenetic analyses.** *BMC Bioinformatics* 2005, **6**:213.
49. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM: **A method for localizing ligand binding pockets in protein structures.** *Proteins* 2006, **62(2)**:479-488.
50. **SitesIdentify** [<http://www.manchester.ac.uk/bioinformatics/sitesidentify>]
51. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
52. Lee B, Richards FM: **The interpretation of protein structures: estimation of static accessibility.** *J Mol Biol* 1971, **55(3)**:379-400.
53. **Jmol: an open-source Java viewer for chemical structures in 3D** [<http://www.jmol.org/>]
54. Dobson PD, Doig AJ: **Distinguishing enzyme structures from non-enzymes without alignments.** *J Mol Biol* 2003, **330(4)**:771-783.
55. Aytuna AS, Gursoy A, Keskin O: **Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces.** *Bioinformatics* 2005, **21(12)**:2850-2855.
56. Hu Z, Ma B, Wolfson H, Nussinov R: **Conservation of polar residues as hot spots at protein interfaces.** *Proteins* 2000, **39(4)**:331-342.
57. Zhou HX, Shan Y: **Prediction of protein interaction sites from sequence profile and residue neighbor list.** *Proteins* 2001, **44(3)**:336-343.
58. Hwang S, Gou Z, Kuznetsov IB: **DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins.** *Bioinformatics* 2007, **23(5)**:634-636.
59. Burgoyne NJ, Jackson RM: **Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces.** *Bioinformatics* 2006, **22(11)**:1335-1342.
60. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247(4)**:536-540.
61. Bray T, Doig AJ, Warwicker J: **Sequence and Structural Features of Enzymes and their Active Sites by EC Class.** *J Mol Biol* 2009, **386(5)**:1423-1436.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

