Published in final edited form as:
Proteins. 2009 ; 77(Suppl 9): 147–151. doi:10.1002/prot.22513.

Prediction of ligand binding sites using homologous structures and conservation at CASP8

Mark N. Wass and **Michael J.E. Sternberg**
Structural Bioinformatics Group, Centre for Bioinformatics, Imperial College London, London, SW7 2AZ.

Abstract

The Critical Assessment of protein Structure Prediction experiment (CASP) is a blind assessment of the prediction of protein structure and related topics including function prediction. We present our results in the function/binding site prediction category. Our approach to identify binding sites combined the use of the predicted structure of the targets with both residue conservation and the location of ligands bound to homologous structures. We obtained average coverage of 83% and 56% accuracy. Analysis of our predictions suggests that overprediction reduces the accuracy obtained due to large areas of conservation around the binding site that do not bind the ligand. In some proteins such conserved residues may have a functional role. A server version of our method will soon be available.

Keywords

bioinformatics; binding site; CASP; function prediction; structural biology

Introduction

We present our results for the prediction of ligand binding sites during the eighth Critical Assessment of protein Structure Prediction experiment (CASP8). CASP is a biennial blind assessment primarily of the prediction of protein structure, which also includes categories such as function prediction. The growth of uncharacterized protein sequence data has made the prediction of functional features such as ligand binding and functional sites a vital area of research. Numerous approaches have been developed for the prediction of functional sites, many use sequence conservation1, which may be mapped onto protein structure2-4. Other methods identify clefts on the protein surface5,6 and recent methods have begun to utilize the ligands present in homologous structures7,8.

Methods

Central to our human predictions was the use of the CASP8 server structure predictions, which provided models of the target structure. 3D-JURY9 was used to filter the models, with the top three selected for modeling. Binding site predictions were made using two different sources of information. Our primary approach was the identification of structures homologous to the target with bound ligands. This is similar to approaches recently developed in firestar7 and FINDSITE8. Our approach also builds upon the concept of supersites10, which demonstrated the conservation of binding sites between even non-

Corresponding Author Address: Sternberg Michael J.E. Biochemistry Building Division of Molecular Biosciences Faculty of Natural Sciences Imperial College London, London SW7 2AZ,UK. m.sternberg@imperial.ac.uk Tel +44 (0)20 7594 5212 Fax +44 (0)20 7594 5264.

Europe PMC Funders Author Manuscripts

homologous proteins (i.e. conservation within a fold). Alignment of the homologous structures with the modeled target structures superimposed the ligands bound to the homologues on to the target structure, indicating possible binding sites. In addition we considered residue conservation. Conserved residues were identified using ConFunc11, our function prediction server, and from Pfam12 alignments. Conserved residues were mapped onto the modeled protein structure, with the aim of identifying clusters of conservation that may be indicative of functional sites. Functional data from UniProt13, Pfam, ConFunc, the catalytic site atlas14 and other sources were used to gain insight into the likely function of the target and thus the type of ligand that may be bound to the structure.

Results and Discussion

We made human predictions for 79 CASP8 targets. Many structures are solved without bound ligands, so it is not possible to assess all of the predictions made in this category15. We made predictions for 23 of the 27 targets assessed by the CASP8 assessors15. We assessed our results using the Matthews correlation coefficient16 (MCC; see Table 1) and measures of coverage and accuracy as used in CASP717, where coverage indicates the percentage of the binding site that was predicted (TP/(TP+FN) where TP are true positives and FN are false negatives) and accuracy (TP/(TP+FP) where FP are false positives) the percentage of predicted residues that were correct. In this assessment we have defined the binding site as all residues within 0.5Å + van der Waals radii from the ligand15. Our predictions obtained an average MCC of 0.63, 83% coverage and 56% accuracy. The results for each individual target show that in all but three of the predictions, more than 70% of the binding site was correctly identified (Fig. 1), demonstrating that we generally predicted most of the binding site. Much greater variation in accuracy is observed, which suggests a trend of overprediction. However, accuracy is also sensitive to the small number of binding site residues compared to the total number of residues in the target proteins and MCC (Table 1) is a more reliable measure of our predictions.

We assessed the effect of the quality of the target models used on our predictions. However, this assessment is complicated because we used multiple models for most targets and information from them was combined manually. It is also difficult to isolate these factors from other sources of error, such as the use of conservation (see below). Plotting the MCC against the GDT_TS of the models used, shows that there is little correlation between the quality of the models and the MCC obtained (Supplementary Figure 1). This is largely because the assessed targets generally have good predicted models. For example only three of our predictions used structures with GDT_TS less than 60. Target T0476 was a borderline free modeling target for which the GDT_TS of the models used are 25 or less. We obtained an MCC of 0.19 for this target. Our prediction did not use homologous structures and relied upon other functional data. Despite the lack of information available for this target, the poor quality models used will have affected our prediction.

We also investigated the relationship between the sequence identity between the target and template structures and our predictions. For most targets we used multiple homologous structures. We do not observe any correlation between the sequence identity and predictive performance (Supplementary Figure 2). Indeed the few targets with homologues with greater than 40% sequence identity obtain some of the lowest MCC scores. This observation may be caused by the use of multiple structures and the manual nature of our predictions.

We visualized our predictions on the target structures to further assess our predictions. A single site was predicted for each target, so false positive predictions are due to the prediction of an area larger than the actual binding site. There is considerable overprediction for 15 of the targets, of which nine have metal ligands. Interestingly some of the lowest

accuracies are obtained for targets that bind a single metal ion. This may be indicative of our approach of using functional data to identify potential ligands that may be bound to the structure. Where functional data suggested that a target was an enzyme we considered where its substrate would be likely to bind. Three targets (T0426,T0444,T0461) that are enzymes, are only bound by metal ions in the absence of substrate. This is a source of overprediction because for these targets we considered the larger functional site that would bind the substrate. Two of these targets (T0444 and T0461) bind RNA or DNA and the assessment was limited to small molecule ligands. However, target T0426 is likely to bind bicarbonate and in our analysis the prediction of an area that may bind this substrate would result in lower performance. We are aware that the assessors have addressed the issue of missing ligands in their final assessment15.

The use of conservation to make predictions has also resulted in overprediction. Residue conservation often occurs in more than one area on the protein and may not be indicative of a binding site as residues are conserved for functional and structural reasons other than ligand binding. Conservation was predominantly used where there was agreement with the location of ligands from homologous structures.

Target T0422 (pdb code 3d8b) provides an example of how overprediction was caused by large area of conservation around the binding site. Twelve of the fifteen residues contacting the ligand (coverage=0.8) were correctly identified, along with a further 9 residues that were not classified as part of the binding site (accuracy=0.57, MCC=0.66). T0422 is an ATPase domain from human fidgetin-like protein 1. ADP is bound in the active site in the solved structure (Fig. 2). It is possible that with ATP bound the binding site would be larger and the extra phosphate group may extend further towards the entrance of the cleft, contacting part of the larger binding site in our prediction. However it is difficult to know how ATP would bind to this protein. Superimposing ligands from homologous structures onto the target structure (See Supplementary Figure 3) suggests that the ATP may extend further out of the cleft but it could also have a different conformation in the binding site, with the adenosine contacting different residues. This highlights a difficulty of assessing such predictions; the target is solved with a ligand bound but different ligands may bind the same site and the residues in direct contact may differ. Similarly some of the residues predicted for the binding site of target T0477 (pdb code 3dkp), which is described as a probable ATP dependent DNA helicase, may bind ATP rather than ADP. We note that the assessors have considered this in the their final assessment15.

Our approach of mapping conservation onto the modeled target structures in addition to the ligands bound to the homologous ligands is the main difference between our method and that of the other top performing group (LEE)18. Our use of conservation often resulted in the prediction of larger binding sites than bound by the ligand in the assessment. In contrast the LEE method solely maps ligands from homologous structures onto their model of the target and using this approach they appear to predict smaller binding sites, which agree more closely with the assessed binding site. This is demonstrated by the assessors' analysis of the predictions using different distance cutoffs, which show that the performance of our predictions increases with the distance cutoff, whereas LEE performance reduces and the difference in MCC is negligible at 2.0Å15.

We made a good prediction for target T0453 (pdb code 3ded). Multiple homologous structures (2oai, 2pls and 2p4p) were identified in the pdb that bind either calcium, magnesium or both ions. Aligning these structures with the target model gave a strong indication of a likely binding site. Residue conservation was not very useful for this target, so the prediction was largely based on the location of the superimposed ligands on the modeled structure. This resulted in a good prediction, correctly identifying the four residues

that bind calcium, while only incorrectly predicting two further residues (Fig. 2). These erroneous predictions appear to be caused by differences between the modeled structures used for the prediction and the solved structure.

Another good prediction was made for target T0440 (pdb code 3dcy). Numerous homologous structures existed with bound ligands. We based our prediction on structures 2yxo, 1pb0 and 1m65. This target binds two iron ions and one zinc ion. The homologous structures contained a combination of different ligands; 2yxo has a virtually identical binding site to the target, 1pbo binds two zinc ions and 1m65 binds a zinc and two sodium ions. Our resulting prediction correctly identified all nine residues that contact the metal ions and only two further residues (Figure 2). Other structures (pdb codes 1av8, 1jk0 1mrr, 1pfr), identified with bound ligands suggested that metal ions may be bound in a different area of the same pocket. However they were disregarded because there was less agreement between their structure and the predicted server structures used.

Concluding Remarks

We have demonstrated the ability to use homologous structures and conservation to make blind predictions of ligand binding sites in CASP8. Our results demonstrate that binding sites were successfully identified in all but one of our assessed predictions. Homologous structures were useful even when they bound ligands different to those bound to the solved structure. Conservation supported the sites identified using homologous structures. However, the use of conservation was also a large source of error, resulting in regular overprediction of the binding site area. We have shown that in some proteins these larger conserved areas may form part of a binding site for a different ligand and for some targets these include the binding site for other proteins or DNA/RNA, which is beyond the scope of the CASP assessment. Errors in the modeled structures used also led to errors in our predictions, however it was their use that was essential to our approach. It enabled us to superimpose ligands from homologous structures onto the target and visualize areas of sequence conservation. We would not have been able to obtain similar predictions using only sequence, which demonstrates how useful predicted structures can be for the prediction of binding sites.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Lawrence Kelley for comments and suggestions on the prediction process and Suhail Islam for technical support. MNW is supported by BBSRC grant BB/F020481/1.

References

1. 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. [PubMed: 14871869]
2. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 1996; 257(2):342–358. [PubMed: 8609628]
3. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol. 2001; 311(2):395–408. [PubMed: 11478868]

4. 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; 33(Web Server issue):W299–302. [PubMed: 15980475]
5. 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. [PubMed: 16304646]
6. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995; 13(5):323–330. 307–328. [PubMed: 8603061]
7. Lopez G, Valencia A, Tress ML. firestar--prediction of functionally important residues using structural templates and alignment reliability. Nucleic Acids Res. 2007
8. Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A. 2008; 105(1):129–134. [PubMed: 18165317]
9. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 2003; 19(8):1015–1018. [PubMed: 12761065]
10. Russell RB, Sasieni PD, Sternberg MJ. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol. 1998; 282(4):903–918. [PubMed: 9743635]
11. Wass MN, Sternberg MJ. ConFunc--functional annotation in the twilight zone. Bioinformatics. 2008; 24(6):798–806. [PubMed: 18263643]
12. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res. 2006; 34(Database issue):D247–251. [PubMed: 16381856]
13. The_Uniprot_Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2007; 35(Database issue):D193–197. [PubMed: 17142230]
14. 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; 32(Database issue):D129–133. [PubMed: 14681376]
15. Tress M. CASP8 Function Assessment Paper in this special issue of proteins. CASP8 Function Assessment Paper in this special issue of proteins. 2009
16. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405(2):442–451. [PubMed: 1180967]
17. Lopez G, Rojas A, Tress M, Valencia A. Assessment of predictions submitted for the CASP7 function prediction category. Proteins. 2007; 69(Suppl 8):165–174. [PubMed: 17654548]
18. LEE. Binding site prediction paper In this issue.

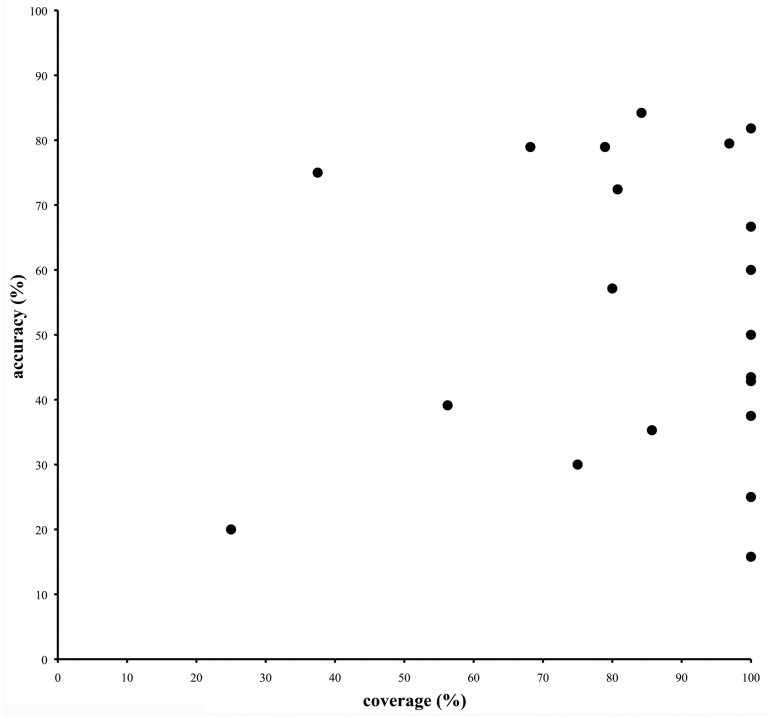

Figure 1.
Coverage and accuracy assessment of CASP8 predictions.

Europe PMC Funders Author Manuscripts

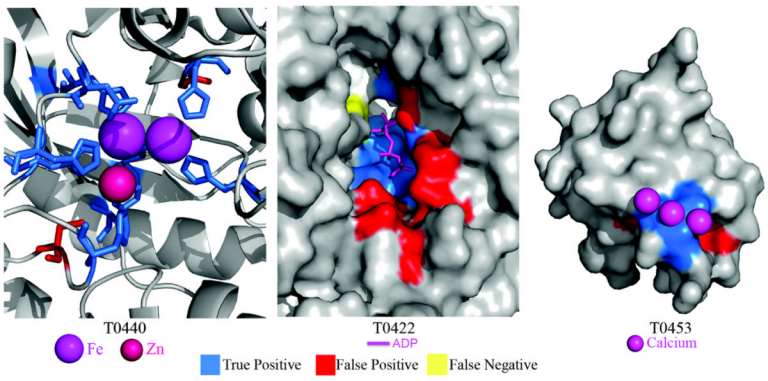

Figure 2.
Example predictions. The agreement of predictions with the binding site is shown for targets T0440, T0422 and T0453. The prediction is mapped onto the target structure and each of the ligands shown (magenta). Residues are colored blue, if they were correctly predicted, red if they were predicted as part of the binding site but are not (i.e. false positives) and yellow for residues in the binding site that were not predicted (i.e. false negatives).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Matthew's Correlation Coefficient (MCC) for binding site predictions.

Target	Ligand	MCC
T0391	FeS	0.51
T0394	PO4	0.53
T0396	FAD	0.67
T0406	Ni	0.60
T0407	Zn	0.90
T0410	Fe	0.65
T0422	ADP	0.66
T0425	Zn	0.70
T0426	Zn	0.39
T0440	Fe	0.90
T0444	Fe	0.46
T0450	FAD	0.87
T0453	Ca	0.81
T0457	Mg	0.49
T0461	Zn	0.65
T0470	Zn	0.49
T0476	Zn	0.19
T0477	ADP, MG	0.64
T0480	Zn	0.68
T0483	ADP	0.44
T0485	SAM8	0.77
T0490	FAD	0.75
T0508	SAM	0.83