Published online 31 May 2010 Nucleic Acids Research, 2010, Vol. 38, Web Server issue W469–W473
doi:10.1093/nar/gkq406

3DLigandSite: predicting ligand-binding sites using similar structures

Mark N. Wass, Lawrence A. Kelley and Michael J. E. Sternberg*

Structural Bioinformatics Group, Centre for Bioinformatics, Imperial College London, London, SW7 2AZ, UK

Received February 11, 2010; Revised April 15, 2010; Accepted May 2, 2010

ABSTRACT

3DLigandSite is a web server for the prediction of ligand-binding sites. It is based upon successful manual methods used in the eighth round of the Critical Assessment of techniques for protein Structure Prediction (CASP8). 3DLigandSite utilizes protein-structure prediction to provide structural models for proteins that have not been solved. Ligands bound to structures similar to the query are superimposed onto the model and used to predict the binding site. In benchmarking against the CASP8 targets 3DLigandSite obtains a Matthew's correlation co-efficient (MCC) of 0.64, and coverage and accuracy of 71 and 60%, respectively, similar results to our manual performance in CASP8. In further benchmarking using a large set of protein structures, 3DLigandSite obtains an MCC of 0.68. The web server enables users to submit either a query sequence or structure. Predictions are visually displayed via an interactive Jmol applet. 3DLigandSite is available for use at http://www.sbg.bio.ic.ac.uk/3dligandsite.

INTRODUCTION

Proteins often perform their function on ligands (e.g. enzyme substrates) or are regulated by them. Therefore the identification of ligand-binding sites is important. The explosion of protein sequences from genome sequencing projects makes it essential for automated methods to predict ligand-binding sites. Further, protein structures are often solved in the absence of ligands, making it important that we are able to identify binding sites for such proteins.

Many methods have been developed for the prediction of ligand-binding sites (reviewed in 1). Sequence conservation is commonly used to predict binding sites (2,3). Many methods combine sequence conservation with structural data (4–6). Evolutionary trace was the first approach to do this by mapping and clustering conserved residues onto protein structure (4,5). Other approaches use probes to identify pockets on the protein surface that are likely to be binding sites (7–9). More recent approaches have focused on using ligand-binding data from similar structures (10–13). Firestar (10) generates sequence alignments of a query with ligand-bound proteins present in the Protein Data Bank (PDB) and combines these with residue conservation to make predictions. FINDSITE (11) superimposes and clusters ligands from similar structures onto a query structure and makes consensus predictions of residues that contact ligands in a cluster. FINDSITE is available as part of the PSiFR web server (12).

Here we present 3DLigandSite, a web server for the prediction of ligand-binding sites, which automates the manual process we used for ligand-binding site prediction in the eighth round of the Critical Assessment of techniques for protein Structure Prediction (CASP8) (13), where we were one of the top-performing groups (14). In CASP8, we superimposed ligands from similar structures onto a structural model of the target protein, in an approach similar to FINDSITE. We also mapped residue conservation onto the protein surface and made predictions combining data from both approaches. 3DLigandSite uses ligands from similar structures to make predictions. It also provides details of conservation as a further guide for the user, but residue conservation information is not currently used in the predictive process.

3DLigandSite performance has been assessed on two sets of proteins; the CASP8 targets, and a set of 617 proteins from the FINDSITE test set (11). On the CASP8, set of 28 protein targets 3DLigandSite obtained a Matthew's Correlation co-efficient (MCC) (15) of 0.64 and coverage and accuracy of 71 and 60%, respectively.

METHODS

The 3DLigandSite algorithm

Figure 1 shows an overview of the 3DLigandSite algorithm. Users may either submit a protein structure or sequence. For sequence submission, Phyre (16), our

*To whom correspondence should be addressed. Tel: +44 20 7594 5212; Fax: +44 20 7594 5264; Email: m.sternberg@imperial.ac.uk

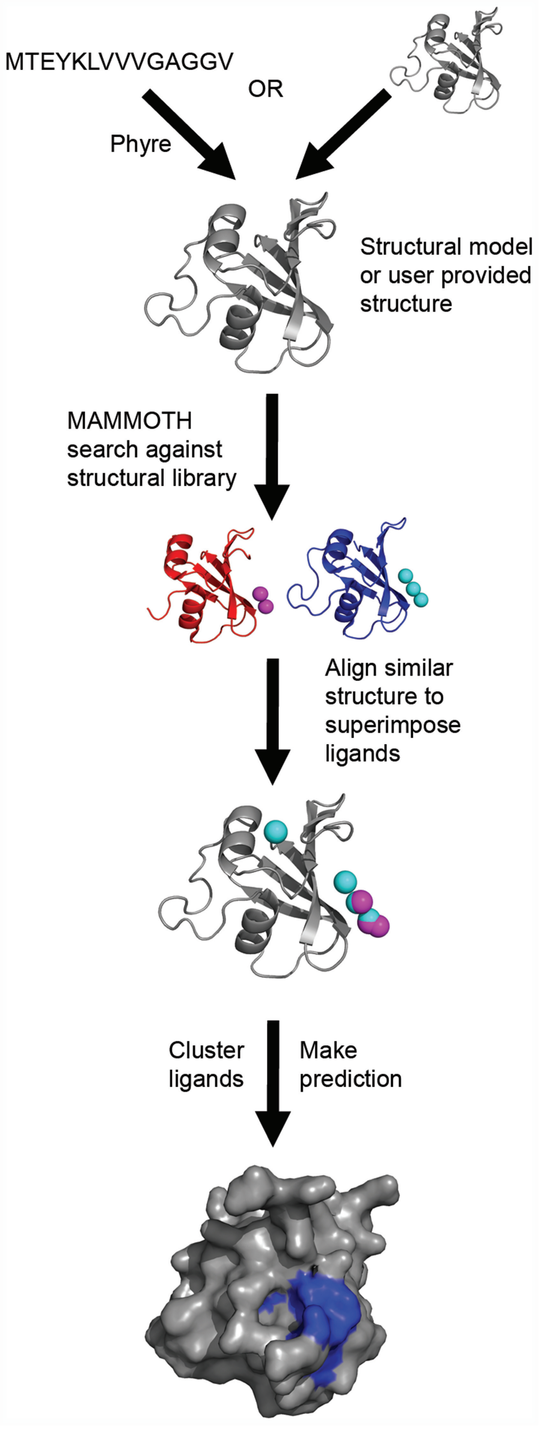

MTEYKLVVVGAGGV
OR
Phyre
Structural model or user provided structure
MAMMOTH search against structural library
Align similar structure to superimpose ligands
Cluster ligands
Make prediction

Figure 1. Overview of 3DLigandSite process.

in-house structure prediction server, is run to generate a model of the protein. Structures similar to the query are identified using MAMMOTH (17) to perform a structural scan of the user-provided structure or Phyre model (referred to as the query structure) against a library of protein structures that have bound ligands. The alignment of the similar structures with the query superimposes their bound ligands onto the query structure. The ligands from the top 25 MAMMOTH hits are retained for analysis. To remove weak structural matches only MAMMOTH hits with a –LnE score >7 are retained. Single linkage clustering is performed on the ligands with a maximum separation of 0.5Å + van der Waals radii between ligands in a cluster. The cluster with the highest number of ligands is selected as the general area of the binding site. The number of ligands within a distance threshold of each residue is used to predict the residues that are part of the binding site. Residue conservation is also calculated and mapped onto the target structure (see below). More details of these steps are described in the sections below.

Generating the structural library

The identification of ligands present in the PDB (18) that are biologically relevant and not present as solvent molecules can be difficult to perform automatically. We used a list of heterogens provided by Uniprot (19) that are unlikely to be present in protein structures as solvent and manually supplemented it with further heterogens that we considered to be likely to be biologically relevant.

Selecting residues

The chosen cluster is used to predict the binding site in the query protein. The number of ligands within a fixed distance (distance cut off) of a residue is used to determine if it is predicted to form part of the binding site. The threshold number of ligands is a proportion of the total number of ligands in the cluster and is set using Equation (1) (where m is a constant that determines the proportion of the ligands that need to be within the distance cut off to be predicted as part of the binding site). The threshold needs to account for variation between the modeled and real structure and between the ligands in the cluster, so a range of distances from 0.2Å to 2.0Å ('Evaluating 3DLigandSite performance' section and Figure 2) and a range of m values in the equation between 0.10 and 0.35 were considered. The server uses a distance setting of 0.8 Å and Equation (1) with m set to 0.24.

$$\text{Threshold} = \text{m} \times \text{cluster size} + 1 \qquad (1)$$

Calculating residue conservation

Residue conservation is calculated using the Jensen Shannon divergence (JSD) score (20). PSI-BLAST (21) is run for the query sequence. The full length sequences of PSI-BLAST hits with *E*-values below 1*e*-03 are aligned with MUSCLE (22) to generate a multiple sequence alignment which is used to calculate conservation. JSD is calculated using default settings as described by Capra and Singh (20), which uses BLOSUM-62 as a background distribution. Residue conservation is not used in the 3DLigandSite prediction. It is provided as a feature of the server which the user can use in conjunction with the 3DLigandSite prediction.

EVALUATING 3DLIGANDSITE PERFORMANCE

3DLigandSite has been benchmarked on the set of structures that were used for the assessment of FINDSITE (11) and on the targets assessed for the ligand-binding category

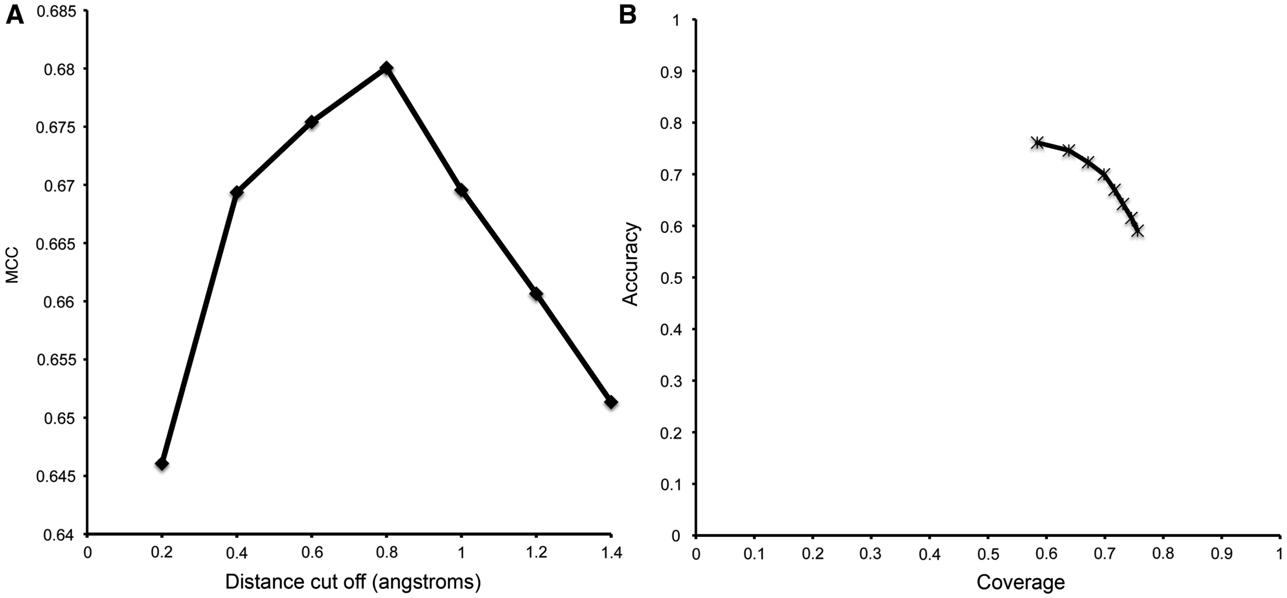


Figure 2. 3DLigandSite performance. (**A**) MCC obtained at different distance cutt offs. (**B**) A graph of accuracy versus coverage for the distance thresholds from (A).

in CASP8 (14). The FINDSITE data set was filtered using our list of accepted ligands, which resulted in a set of 617 test structures. 3DLigandSite made predictions for all but three of these structures. 3DLigandSite performance was assessed using a range of distance cut offs between 0.2Å and 2.0Å at 0.2Å intervals and with m in Equation (1) set to values in the range 0.10–0.35. We assessed the predictions using the MCC (15), and coverage and accuracy, all of which have been used for assessment in recent CASP experiments (14,23). The results for the m setting of 0.24 are displayed in Figure 2 (the full set of results is shown in Supplementary Figure S1). At low-distance cut offs, high accuracy and lower coverage are obtained and as the distance cut off increases the accuracy lowers while the coverage increases. The maximum MCC of 0.68 is obtained at a 0.8Å distance cut off. The MCC decreases at lower and higher cut offs (Figure 2). At this distance cut off 70% coverage and accuracy are obtained. This setting was selected for use in the 3DLigandSite server and for the analysis of the CASP8 targets.

To make our predictions of the 28 CASP8 targets comparable with the predictions made during CASP8, the structural library was restricted to structures present in the PDB before May 2008 (the start of CASP8). Using the settings described earlier, 3DLigandSite obtained a MCC of 0.64 and coverage and accuracy of 71 and 60%, respectively. These results are comparable to our human performance in CASP8, where we obtained MCC of 0.63, 83% coverage and 56% accuracy (13).

THE 3DLIGANDSITE WEB SERVER

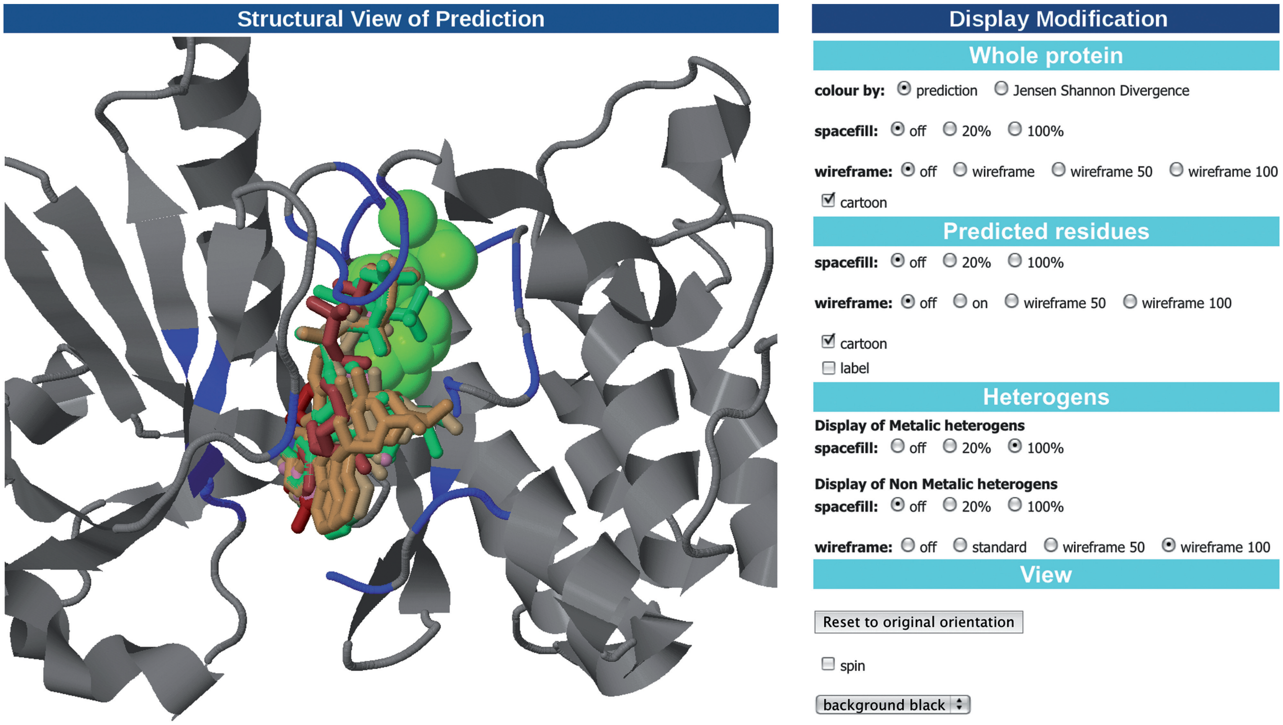
The 3DLigandSite server is available at http://www.sbg.bio.ic.ac.uk/3dligandsite. Users can submit either a protein sequence or a structure. For sequence submission, the first step of the prediction process is to model the structure of the protein using Phyre (16).

Results output

3DLigandSite output is split into four main sections. The first provides details of the phyre model used (only where a sequence has been submitted) and details of the search against the structural library. This information provides details of confidence in two individual steps of the prediction process, which can aid the user in deciding their confidence in the prediction.

The second section shows a table of the ligand clusters identified. The cluster containing the greatest number of ligands is automatically selected for prediction by 3DLigandSite. This table provides details of the other clusters and allows the user to view the potential sites associated with these clusters. Links are provided to Jmol applets for each cluster, similar to that of the main prediction (see below and Figure 3).

The final two sections display the 3DLigandSite prediction. A table lists all of the predicted binding-site residues with details of the number of ligands that they contact, the average distance between the residue and the residue conservation score (JSD). A table of the heterogens present in the cluster is provided together with details of the source structures from the structural library. A Jmol (www.jmol.org) applet enables visualization of the modeled protein, the ligand cluster and the predicted binding site (Figure 3). Jmol is java based and only requires users to have a java runtime environment installed on their machine. By default, the protein is displayed in cartoon format with metallic ligands in spacefill and non-metallic ligands in wireframe representation. A table to the right of the applet provides controls for users to modify the display in the applet. Options are available to modify the display of the whole protein, predicted residues and ligands.



colour by: ● prediction ○ Jensen Shannon Divergence

spacefill: ● off ○ 20% ○ 100%

wireframe: ● off ○ wireframe ○ wireframe 50 ○ wireframe 100

☑ cartoon

spacefill: ● off ○ 20% ○ 100%

wireframe: ● off ○ on ○ wireframe 50 ○ wireframe 100

☑ cartoon

□ label

Display of Metalic heterogens

spacefill: ○ off ○ 20% ● 100%

Display of Non Metalic heterogens

spacefill: ● off ○ 20% ○ 100%

wireframe: ○ off ○ standard ○ wireframe 50 ● wireframe 100

Reset to original orientation

□ spin

background black

Figure 3. 3DLigandSite visualization of prediction for target T0483 in CASP8. The Jmol applet displays the protein structure with predicted binding site colored blue. The ligands in the cluster used to make the prediction are displayed with ions in spacefill and organic molecules in wireframe formats. In this example, Magnesium, AMP, ADP and ATP form the ligand cluster. The panel to the right enables the user to modify the display of the protein and the ligand cluster.

The protein and predicted binding site can be shown in cartoon, spacefill or wireframe formats. The protein can be colored to show the predicted binding site or residue conservation. The user can also label the predicted residues making it easier for them to investigate the predicted binding site. Further, spacefill and wireframe options are also available for displaying the ligand cluster. These multiple viewing options provide a powerful way for the user to interrogate the prediction displayed in the applet.

CONCLUDING REMARKS

3DLigandSite was developed to automate our manual approach for predicting ligand-binding sites used in CASP8 (13). We have demonstrated that 3DLigandSite is able to obtain performance comparable to ours in CASP8 and that this performance is also retained for a much larger test set. In CASP8 we found that extensive use of residue conservation reduced performance of our approach and as a result the use of residue conservation is limited in 3DLigandSite, so future work will try to incorporate conservation in a way that improves predictive performance. We also intend to develop more sophisticated thresholds over the simple distance measure that is currently used by 3DLigandSite.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Suhail Islam for technical support.

FUNDING

Biotechnology and Biological Sciences Research Council (grants BB/F020481/1 to M.N.W. and BB/E000940/1 to L.A.K., respectively). Funding for open access charge: Biotechnology and Biological Sciences Research council.

Conflict of interest statement. M.J.E.S. is a founder director of Equinox Pharma Ltd, holds shares in the company, and has obtained remuneration from the company. Equinox Pharma Ltd is exploiting computational methods for drug discovery and markets software.

REFERENCES

1. Gherardini,P.F. and Helmer-Citterich,M. (2008) Structure-based function prediction: approaches and applications. *Brief. Funct. Genomic Proteomic*, **7**, 291–302.
2. Berezin,C., Glaser,F., Rosenberg,J., Paz,I., Pupko,T., Fariselli,P., Casadio,R. and Ben-Tal,N. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, **20**, 1322–1324.
3. Fischer,J.D., Mayer,C.E. and Soding,J. (2008) Prediction of Protein Functional Residues from Sequence by Probability Density Estimation. *Bioinformatics*, **24**, 613–620.
4. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
5. Aloy,P., Querol,E., Aviles,F.X. and Sternberg,M.J. (2001) 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.*, **311**, 395–408.

6. Capra,J.A., Laskowski,R.A., Thornton,J.M., Singh,M. and Funkhouser,T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.
7. Glaser,F., Morris,R.J., Najmanovich,R.J., Laskowski,R.A. and Thornton,J.M. (2006) A method for localizing ligand binding pockets in protein structures. *Proteins*, **62**, 479–488.
8. Huang,B. and Schroeder,M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Bio.*, **6**, 19.
9. Hernandez,M., Ghersi,D. and Sanchez,R. (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.*, **17**, W413–W416.
10. Lopez,G., Valencia,A. and Tress,M.L. (2007) firestar–prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, **35**, W573–W577.
11. Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA*, **105**, 129–134.
12. Pandit,S.B., Brylinski,M., Zhou,H., Gao,M., Arakaki,A.K. and Skolnick,J. (2010) PSiFR: an integrated resource for prediction of protein structure and function. *Bioinformatics*, **26**, 687–688.
13. Wass,M.N. and Sternberg,M.J. (2009) Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins*, **77(Suppl. 9)**, 147–151.
14. Lopez,G., Ezkurdia,I. and Tress,M.L. (2009) Assessment of ligand binding residue predictions in CASP8. *Proteins*, **77(Suppl. 9)**, 138–146.
15. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.*, **405**, 442–451.
16. Kelley,L.A. and Sternberg,M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.*, **4**, 363–371.
17. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
18. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
19. The UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
20. Capra,J.A. and Singh,M. (2008) Characterization and prediction of residues determining protein functional. *Bioinformatics*, **24**, 1473–1480.
21. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
22. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
23. Lopez,G., Rojas,A., Tress,M. and Valencia,A. (2007) Assessment of predictions submitted for the CASP7 function prediction category. *Proteins*, **69(Suppl. 8)**, 165–174.