In silico identification and characterization of protein-ligand binding sites

Chapter Title: *In silico* identification and characterization of protein-ligand binding sites

Daniel Barry Roche[1,2] and Liam James McGuffin[3]

[1]Institut de Biologie Computationnelle, LIRMM, CNRS, Université de Montpellier, Montpellier, France, [2]Centre de Recherches de Biochimie Macromoléculaire, CNRS- UMR 5237, Montpellier, France and [3]School of Biological Sciences, University of Reading, Reading, UK

Corresponding author: Daniel Barry Roche, email: daniel.roche@lirmm.fr

Running Head: Predicting protein-ligand interactions.

Summary

Protein ligand binding site prediction methods aim to predict, from amino acid sequence, protein-ligand interactions, putative ligands and ligand binding site residues using either sequence information, structural information or a combination of both. *In silico* characterisation of protein-ligand interactions have become extremely important to help determine a protein functionality, as *in vivo* based functional elucidation is unable to keep pace with the current growth of sequence databases. Additionally, *in vitro* biochemical functional elucidation is time consuming, costly and may not be feasible for large scale analysis, such as drug discovery. Thus, *in silico* prediction of protein-ligand interactions need to be utilized to aid in functional elucidation.

Here we briefly discuss protein function prediction, prediction of protein-ligand interactions, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) and the Continuous Automated EvaluatiOn (CAMEO) competitions, along with their role in shaping the field. We also discuss, in detail, our cutting-edge web-server method FunFOLD for the structurally informed prediction of protein-ligand interactions. Furthermore, we provide a step-by-step guide on using the FunFOLD webserver and FunFOLD3 downloadable application, along with some real world examples, where the FunFOLD methods have been used to aid functional elucidation.

Key Words

Protein function prediction; Protein-ligand interactions; Binding-site residue prediction; Biochemical functional elucidation; Critical Assessment of Techniques for Protein Structure Prediction (CASP); Continuous Automated EvaluatiOn (CAMEO); Protein structure

prediction; Structure based function prediction; Quality assessment of protein-ligand binding site predictions; Protein-ligand interactions; Webserver; Downloadable application;

1. Introduction

Proteins play an essential role in all cellular activity, which includes; enzymatic catalysis; maintaining cellular defences; metabolism and catabolism; signalling within and between cells and the maintenance of the cells structural integrity. Hence, the identification and characterization of a protein binding site and associated ligands, is a crucial step in the determination of a proteins functionality [1-3].

1.1 Predicting protein-ligand interactions

Protein-ligand interaction prediction methods can be categorised into two broad groups: sequence based methods and structure based methods [4,1,5]. Sequence based methods utilize evolutionary conservation to determine residues, which may be structurally or functionally important. These methods include firestar [6,7], WSsas [8], INTREPID [9], Multi-RELIEF [10], ConSurf [11], ConFunc [12], DISCERN [13], TargetS [14] and LigandRFs [15]. Structure based methods can additionally be separated into geometric-based methods (FINDSITE [16], Surflex-PSIM [17], LISE [18], Patch-Surfer2.0 [19], CYscore [20], LigDig [21] and EvolutionaryTrace [22,23]), energetic methods (SITEHOUND [24]) and miscellaneous methods that utilize information from homology modelling (FunFOLD [3], FunFOLD2 [2], COACH [25], COFACTOR [26], GalaxySite [27] and GASS [28]), surface accessibility (LigSite^{csc} [29]) and physiochemical properties, utilized by methods including SCREEN [30].

1.2 The role of CASP and CAMEO on the development of protein-ligand interaction methods

In recent years there has been an explosion in the development and availability of protein ligand binding site prediction methods. This is a direct result of the inclusion of a ligand-binding site prediction category in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition [31-33], along with the subsequent inclusion of ligand-binding site prediction in the Continuous Automated EvaluatiOn (CAMEO) competition [34].

Ligand-binding site residue prediction was first introduced in CASP8 [31], with the idea to predict putative binding site residues, in the target protein, which may interact with a bound biologically relevant ligand. The top methods in CASP8 (LEE [4] and 3DLigandSite [35]) utilized homologous structures with bound biologically relevant ligands in their prediction strategies. In both CASP9 [32] and CASP10 [33], protein-ligand interaction methods converged on similar strategies; the structural superposition of models, onto templates bound to biologically relevant ligands [1].

After the CASP10 competition the protein-ligand interaction analysis moved to the CAMEO [34] continuous evaluation competition. This was a direct result of a lack of targets for evaluation, over the 3 month prediction period of the CASP competition. Although, predictions were still accepted for the CASP11 competition. This also resulted in a change of prediction format, where methods not only have to predict potential ligand binding site residues, but also predict the probability that each residue binds to a specific ligand type; I - Ion; O - Organic ligand; N - nucleotide and P - peptide. In addition to the most likely ligand type that the protein may bind [34]. The continuous weekly assessment of CAMEO allows

for a much better picture, of how a method performs, on a large diverse dataset, containing a wide diversity of ligand types [34].

1.3 Metrics to assess protein-ligand interactions

Both CASP and CAMEO utilise a number of different metrics to analyse protein-ligand interaction predictions. The first score utilised in CASP8 [31] was the Matthews Correlation Coefficient (MCC score) [36]. The MCC score is a statistical score for the comparison of predicted ligand binding site residues to observed ligand binding site residues, by analysing the number of residues assigned as true positives, false positives, true negatives and false negatives, resulting in a score between -1 and 1 (1 is a perfect prediction, 0 is a random prediction). The disadvantage of the MCC score is that it is a statistical measure, which does not take into account the 3D nature of a protein. Additionally, it is often a subjective matter to assign observed ligand binding site residues, even in an experimental structure, which is another disadvantage of using a purely statistical metric.

Thus, we proposed a new scoring metric: the Binding-site Distance Test (BDT score) [37], which addresses some of the problems associated with the MCC score. The BDT score takes into account the distance in 3D space a predicted binding site residue is from an observed binding site residue. The BDT score ranges from 0 to 1 (1 is a perfect prediction, 0 is a random prediction). Binding sites which are predicted close to the observed binding site, scores higher that binding sites predicted far from the observed site. The BDT score was used in addition to the MCC score in both the CASP9 [32] and CASP10 [33] assessments and is now a standard assessment metric used in CAMEO [34].

1.4 The FunFOLD2 server for the prediction of protein-ligand interactions

The FunFOLD server has been developed with the user in mind, providing an intuitive interface (Figure 1), which allows users to easily predict protein-ligand interactions for their protein of interest [2]. Additionally, for the more expert user, a PDB file of the top IntFOLD2-TS [38] model containing the biologically relevant ligand cluster can be downloaded for further interrogation, along with predicted ligand-protein interaction quality scores. Additionally, the results are available in CASP FN and CAMEO-LB format. The FunFOLD2 server takes as input a protein sequence, and optionally a short name for the target protein. Also, the user has the option to include an email address, to allow for easy results delivery or the submission page can be bookmarked and returned to later, when results are available. The FunFOLD2 server runs the IntFOLD2-TS structure prediction algorithm to produce a set of models and related templates that can be used to predict protein-ligand interactions. The FunFOLD2 [2] method combines the original FunFOLD method [5] for ligand binding site residue prediction, the FunFOLDQA method [1] for ligand binding site quality assessment and a number of scores to comply with the CAEMO-LB prediction format [34].

The original FunFOLD method [5] was designed based on the following concept: protein structural templates from the PDB containing biologically relevant ligands, and having the same fold (according to TM-align [39]), as the model built for the target under analysis, may contain similar binding sites. Firstly, the FunFOLD algorithm takes as input a model and a set of template PDB IDs (generated by IntFOLD2-TS [38]). Secondly, TM-align [40] is used to superpose each template determined to contain a biologically relevant ligand onto the target model (originally the method used an in-house curated ligand list, now the latest version, FunFOLD3, described below, makes use of the BioLip database [41]). Template-model superpositions having a TM-score ≥ 0.4 are used in the next step.TM-scores ranging from 0.4

to 0.6 has been shown to mark the transition step of significantly related folds [42]. Thirdly, all retained templates are superposed onto the model and ligands are assigned to clusters using an agglomerative hierarchical clustering algorithm, identifying each continuous mass of contacting ligands, thus locating potential binding pockets. Ligands are determined to be in contact within a cluster if the contact distance is less than or equal to the Van der Waal radius of the contacting atoms plus 0.5 Å. The location of the largest ligand cluster is thus determined to be the putative binding site.

Fourthly, putative ligand binding site residues are determined using a novel residue voting method. The distance between all atoms in the ligand cluster and all atoms in the modelled 3D protein are calculated. Again, residues are determined to be in contact with the ligand cluster, if the contact distance between any atom in the residue and any atom in the ligand cluster is less than or equal to the Van der Waal radius of the contacting atoms plus 0.5 Å. Finally, the next step is "residue voting", where all residues determined to be in contact with the ligand cluster are further analysed and included in the final prediction if a residues has at least one contact to 2 ligands within the cluster and at least 25% of the ligands in the cluster [3].

The next tool utilized by the FunFOLD2 server [2] is the FunFOLDQA algorithm [1], which assesses the quality of the FunFOLD prediction [3], outputting a set of quality scores. The FunFOLDQA algorithm produces five feature based scores; BDTalign, Identity; Rescaled BLOSUM62 score; Equivalent Residue Ligand Distance Score and 3D Model Quality (using ModFOLDclust2 [43]), which are subsequently combined using a neural network to produce predicted MCC and BDT scores. The predicted MCC and BDT scores can be used to rank the FunFOLD predictions of the top 10 IntFOLD2-TS models, to find the best prediction. This has been shown to statistically significantly improve protein-ligand prediction quality over

using FunFOLD alone [1]. The BDTalign score basically determines the fit of the model binding site into the binding sites of the templates used in the prediction. The Identity score assesses the relationship between the binding site residues, which are equivalent in 3D space, between the model and the templates, scoring them according to their amino acid identity. The Rescaled BLOSUM62 score utilizes the same concept as the Identity score, but scores equivalent residues in 3D space according to the BLOSUM62 scoring matrix. Furthermore, the Equivalent Residue Ligand Distance score scores equivalent residues in 3D space between the model and each template according to their distance from the bound ligand.

The final component of the FunFOLD2 server [2] is to score the resultant ligand binding site residues, from the top prediction, based on the CAMEO-LB criteria. The first score is a global functional propensity metric, which calculates the probability that the protein will bind to each ligand type (I- Ion; O - Organic; N- Nucleotide; P - Peptide). The second score is the per-residue functional propensity metric, which determines the propensity that each predicted ligand binding site residue is in contact with each ligand type (I, O, N & P) [2].

1.5 The FunFOLD3 algorithm for the prediction of protein-ligand interactions

The FunFOLD3 algorithm, is the latest implementation of FunFOLD. FunFOLD3 was designed to produce predictions to comply with the CAMEO-LB prediction format [34], including the development new metrics to predict per-atom *P-values*. Another major change in FunFOLD3 is the use of the BioLip database [41], for the determination of biologically relevant ligands at multiple binding sites. In addition to the provision of functional annotations, namely EC [44] numbers and GO terms [45]. The FunFOLD3 algorithm along with FunFOLDQA [1] have been integrated into the latest version of the IntFOLD server pipeline [46] and is available as an executable JAR file. The executable version of

FunFOLD3 does not incorporate the FunFOLDQA binding site quality scoring module, which can be downloaded as a separate JAR executable if desired.

The FunFOLD2 method and its previous implementations have been benchmarked at CASP9 and CASP10 and were amongst the top performing methods [33,32]. In addition to CASP the FunFOLD2 and FunFOLD3 methods are now continuously benchmarked by CAMEO [34] (http://www.cameo3d.org). Furthermore, the FunFOLD algorithms have been utilized in numerous studies, including on barley powdery mildew proteins [47,48], calcium binding proteins [49] and olfactory proteins [50], which have resulted in interesting biological findings.

In summary, the use of computational methods for the prediction of protein-ligand interactions is essential in the era of high-throughput next-generation sequencing, as experimental methods are unable to keep pace. The prediction of protein-ligand interactions can lead to the interpretation of a protein's general function. These predictions can be further utilized in subsequent *in silico*, *in vivo* and *in vitro* studies, for the discovery of new functions, as well as in drug discovery, which can impact on issues such as health and disease.

2. Materials and Systems Requirements

2.1 Web server requirements

1. For the FunFOLD2 webserver [2], internet access and a web browser are required. The server is freely accessible at: http://www.reading.ac.uk/bioinf/FunFOLD/ (See Figure 1 and Note 1). The FunFOLD2 server has been extensively tested on Google Chrome and Firefox, which are recommended for proper use. The server also works

on other browsers such as Internet Explorer, Safari and Opera, but these browsers have not been tested as extensively.

2. To run your protein-ligand interaction predictions on the FunFOLD2 server you require an amino acid sequence for your protein of interest, in single letter code format. Additionally, a short name can be given for the target sequence submitted and an email address can be included to inform the user when the prediction is complete. If the length of the target amino acid sequence is longer than 500 amino acids, it is best to divide the target sequence into domains, using PFAM [51] or SMART [52], then submit each domain sequence separately. For a more detailed explanation along with potential problems that can be encountered at the submission stage see Note 1.

2.2 Requirements for the FunFOLD3 downloadable executable

A downloadable version of the FunFOLD3 method is available as an executable JAR file, which can be run locally. The executable has several dependencies and system requirements which are briefly described below. The executable along with a detailed README file and example input and output data can be downloaded from the following location: http://www.reading.ac.uk/bioinf/downloads/. (See Note 3 for potential errors that may be encountered).

2.2.1 FunFOLD3

The system requirements are as follows:

1. A linux based operating system such as Ubuntu.
2. A recent version of Java (www.java.com/getjava/).
3. A recent version of PyMOL (www.pymol.org).
4. The TM-align program [39] (http://zhanglab.ccmb.med.umich.edu/TM-align/) . Please ensure the TM-align program is working on your system before attempting to

run FunFOLD3. Ensure that you have the correct 32bit/64bit version for your hardware and that the TMalign file is made executable: chmod +x TMalign

5. wget and ImageMagick installed system wide.
6. The CIF chemical components database file [53] should be downloaded from here: ftp://ftp.wwpdb.org/pub/pdb/data/monomers/components.cif.

1. The BioLip databases [41] containing ligand and receptor PDB files is also required (up to 30 GB or disc space may be required). The databases need to be downloaded in 2 sections firstly all annotations prior to 6/3/2013 can be downloaded from here for the receptor database: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/receptor_2013-03-6.tar.bz2 (3.6 G) and from here for the ligand database: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/ligand_2013-03-6.tar.bz2 (438 M). The Text File of the BioLip annotations can be downloaded from here: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/BioLiP.tar.bz2. To update the databases to include annotations after 2013-03-6 it is recommended to download and use this perl script which will update the databases: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/download_all_sets.pl. The BioLip text file: http://zhanglab.ccmb.med.umich.edu/BioLiP/download/BioLiP.tar.bz2 and all the weekly update text files should be concatenated to form a large text file containing all of the annotations. Furthermore, it is recommended to regularly update your BioLip and CIF databases. Additionally, a shell script is available downloadBioLipdata.sh, which can be download from here: http://www.reading.ac.uk/bioinf/downloads/, in a compressed directory: FunFOLDPackage.tar.gz. To run the shell script simply edit the file paths for the location of the BioLip databases and the executable directory.

7. Please ensure your system environment is set to English, as utilizing other languages may cause problems with the FunFOLD calculations: export LC_ALL=en_US.utf-8.
8. Note the FunFOLD3 executable does not contain the FunFOLDQA code. The FunFOLDQA code is available to download as a separate executable if desired.

3. Methods

In this section we present a step-by-step guide on utilizing the FunFOLD2 server and the FunFOLD3 downloadable executable, to produce protein-ligand interaction predictions for the user's sequence of interest. We also describe interesting case studies of the FunFOLD3 method and its previous implementations.

3.1 The FunFOLD2 server

3.1.1 The submission process

1. Navigate to the FunFOLD2 submission page: http://www.reading.ac.uk/bioinf/FunFOLD/FunFOLD_form_2_0.html.
2. The next step is to paste the full single-letter format amino acid sequence of your protein of interest into the text box provided on the submission page (see Figure 1).
3. Optionally, the user can provide a short name for their target sequence.
4. The user has the option to supply their email address, which enables an email to be sent to the user once the results of the target sequences become available.
5. Once all of the required information boxes, on the submission page, have been filled, the user then needs to click on the submit button to enable submission of their prediction.
6. Presently, submissions are limited to one per IP address, to enable the maintenance of speed and server capacity. Upon completion of the user's prediction, their IP address

is automatically unlocked and they can then submit their next target sequence. See Note 1 for common problems encountered at the submission step.

3.1.2 How to interpret the results

1. Upon job completion an email is sent to the user, which contains a link to the prediction results for the target sequence. See Figure 2 for an example results page (FunFOLD3 via the IntFOLD server) and Figure 3 for example results from CASP11.
2. The results page contains graphical results for the target sequence, in addition to downloadable machine readable results in CASP format. Firstly, a graphical representation of the ligand binding site, showing putative binding site residues, rendered using PyMOL (www.pymol.org) is shown. The backbone of the protein is shown as a green ribbon, while the putative ligand binding site residues are labelled and shown as blue sticks. Secondly, a link is also available to download a PDB file containing the putative ligand binding site cluster within the top IntFOLD [54] model. Thirdly, the CASP FN format results are shown. This includes a list of putative ligand binding site residues. The most likely ligand, which is the most likely ligand to be bound to the target protein according to the FunFOLD prediction. This is followed by the centroid ligand and a list of all ligands within the putative ligand cluster is also included. The centroid and most likely ligand have an associated residue number that corresponds to their residue number in the downloadable PDB file, the residue number can be easily used to locate the ligand in the PDB file for a more detailed examination of the results.
3. The final section of the results page is a JSmol view of the ligand binding site within the target protein, which can be easily used to examine the prediction in 3D space.

There are a number of options to rotate the protein, show and hide the ligands as well as alter the way the ligands are represented.

4. In addition, predicted quality scores from FunFOLDQA [1] are also provided: BDTalign, Identity; Rescaled BLOSUM62 score; Equivalent Residue Ligand Distance Score and Model Quality along with the predicted MCC and BDT scores (See 1.4 for a description of these scores). Furthermore, the propensity that the target protein binds to each ligand type (I- Ion; O - Organic; N- Nucleotide; P - Peptide) is also provided in CAMEO-LB format [34]. (See Note 2 for potential errors that may be encountered and Note 4 for current method limitations).
5. Moreover, for the version of FunFOLD (FunFOLD3) integrated into the IntFOLD pipeline [46], putative EC [44] and GO [45] codes, derived from templates used in the prediction from the BioLip [41] database are included. (See Note 3 for details on the IntFOLD server [46,55]).

3.2 The FunFOLD3 executable

For large scale analysis or to integrate the FunFOLD3 method into a structure prediction pipeline or web server (See Note 2 and Note 5) a downloadable executable JAR file, which has been developed to run on linux based operating systems is available (http://www.reading.ac.uk/bioinf/downloads/). This version of the program has been tested on recent versions of Ubuntu, but it should work on all linux based systems that have bash installed and meet the system requirements (See Section 2.2.1).

2. To run the program you can simply edit the shell script (FunFOLD3.sh) or you can follow the steps below.
3. The user can optionally set the bash environment variable for Java, TM-align and PyMOL if they have not installed it system wide, along with the location of the

databases and database files e.g.

export LC_ALL=en_US.utf-8

export PYMOL_HOME=/usr/bin/

export TMALIGN_HOME=/home/roche/bin/

export JAVA_HOME=/usr/bin/

export BIOLIP_Directory=/home/roche/bin/BioLip/FunFOLDBioLip/

export BIOLIP_LIGAND=/home/roche/bin/BioLip/FunFOLDBioLip/ligand/

export BIOLIP_RECEPTOR=/home/roche/bin/BioLip/FunFOLDBioLip/receptor/

export BIOLIP_TXT=/home/roche/bin/BioLip/FunFOLDBioLip/BioLiP.txt

export CIF=/home/roche/bin/BioLip/FunFOLDBioLip/components.cif

$BIOLIP_Directory = BioLip directory location

$BIOLIP_TXT = BioLip database text file including the full directory path

$BIOLIP_LIGAND = BioLip ligand directory

$BIOLIP_RECEPTOR = BioLip receptor directory

$CIF = CIF file including the full directory path

4. For example, if the path of your model was "/home/roche/bin/FunFOLD3/MUProt_TS3", your list of templates was "/home/roche/bin/FunFOLD3/T0470_PARENTNew.dat" (all templates should be listed on a single line separated by a space), your FASTA sequence file was "/home/roche/bin/FunFOLD3/T0470.fasta", your output directory was "/home/roche/bin/FunFOLD3/" and your target was called T0470:

   $JAVA_HOME/java -jar FunFOLD3.jar /home/roche/bin/FunFOLD3/MUProt_TS3 T0470 /home/roche/bin/FunFOLD3/ /home/roche/bin/FunFOLD3/T0470_PARENTNew.dat

/home/roche/bin/FunFOLD3/T0470.fasta $BIOLIP_TXT $BIOLIP_LIGAND $BIOLIP_RECEPTOR $CIF

Or, using the shell script provided:

./FunFOLD3.sh /home/roche/bin/FunFOLD3/MUProt_TS3 T0470 /home/roche/bin/FunFOLD3/ /home/roche/bin/FunFOLD3/T0470_PARENTNew.dat /home/roche/bin/FunFOLD3/T0470.fasta

5. Basically, the user requires a model generated for their target protein, this can be achieved using a homology modelling method either in-house or via a webserver such as IntFOLD [38] (see Note 3). Additionally, the user needs a list of structurally similar templates. Again this list of templates can be generated from the list of templates used to generate the target protein model. The program utilises the templates that have the same fold and contain biologically relevant ligands in the prediction process. Furthermore, it is important to download and install the BioLip databases [41] and CIF chemical components library file [53]. Additionally, it is important that the full paths for all input files are used, the output directory should also end with a "/" and must contain the input model, template list and FASTA sequence file.
6. Additionally, a shell script is available called downloadBioLipdata.sh, which can be used to download and update the BioLip and CIF libraries. The shell script and the required perl script can be found on the downloads page, in a compressed directory: downloadBioLip_CIF.tar.gz. To run the shell script simply edit the file paths for the location of the BioLip databases and the executable directory.

A number of output files are produced in the output directory (e.g. "/home/roche/bin/FunFOLD3/") and a log of the prediction process is output to screen as standard output. A description of the output files are as follows:

1. The final ligand binding site prediction file "T0470_FN.txt" is supplied, conforming to CASP FN format. This file contains a list of predicted binding site residues, ligands, along with associated EC and GO terms.
2. The final binding site prediction file "T0470_FN2_CAMEO-LB.txt" is additionally supplied in CAMEO-LB format. This file contains the predicted propensity that each ligand type is in contact with the predicted binding site residues.
3. A PDB file "T0470_lig.pdb", which contains superpositions of all templates, having the same fold and containing biologically relevant ligands, onto the model is produced.
4. A reduced version of the PDB file "T0470_lig2.pdb", which contains only the target model with all possible ligands is also produced.
5. Another reduced version of the PDB file "T0470_lig3.pdb", which contains only the target model with the predicted centroid ligand, is additionally output.
6. A graphical representation of the protein-ligand interaction prediction "T0470_binding_site.png" is automatically generated using PyMOL.
7. Finally, the PyMOL script "pymol.script" that was used to generate the image file is also output.
8. An example of output produced by FunFOLD3 for target T0470 can be found in the compressed directory: "T0470_Results.tar.gz" along with an example of the required input: "T0470_Input.tar.gz". These example directories can be found on the downloads page: http://www.reading.ac.uk/bioinf/downloads/.

3.3 Server fair usage policy

To enable timely throughput and wide use of the server, a fair usage policy is implemented. Users are allowed to submit one prediction per IP address. Once the first job is complete, a notification is sent to the user via email, if an email address has been provided. If a user does not provide an email address, then a link to the results page is provided which users are recommended to bookmark during the submission process. Once the job has been completed, the users IP address is unlocked and the server is ready to receive the next submission. The results for each complete job is saved for 30 days. It is recommended for large scale analysis of a large number of proteins (proteome level) to download the executable version of FunFOLD3 (See section 3.2 and Note 2 and Note 5).

3.4 Case studies

The FunFOLD3 method and its previous implementation have been used in a number of studies [47-50], which have led to interesting biological findings, here we discuss one such study. Furthermore, in-house analysis of the CASP11 FN predictions produced by the FunFOLD3 algorithm, via the IntFOLD server are evaluated (CASP11 group ID: TS133).

3.4.1 Analysis of the barley powdery mildew proteome

The first study combined proteogenomic and *in silico* structural and functional annotations (prediction of protein-ligand interactions), to enable the investigation of the pathogen proteome of barley powdery mildew [47,48]. Basically, genomic scale structure prediction was carried out using IntFOLD [55]. Both the global and per-residue model quality was assessed utilizing ModFOLD3 [54,56] and putative protein-ligand interactions, were additionally predicted using FunFOLD [5]. The results lead to interesting conclusions about the structural and functional diversity of the proteomes. Firstly, only 6 protein could be modelled with a model quality score above 0.4, leading to a conclusion that the genome is very structurally diverse and may have many novel folds. Secondly, for the 6 predicted

structures, FunFOLD [5] was able to predict that the proteins were carbohydrate binding, and using the models and other additional data it was conclude that they were probably glycosyl hydrolases. Furthermore, the putative functionality was experimentally verified. In conclusion the FunFOLD method was crucial in the putative functionality assignment of these enzymes, which were subsequently experimentally verified.

3.4.2 CASP 11 functional prediction

The second case study focuses on the analysis of FunFOLD3 blind predictions from the CASP11 competition. Briefly, all CASP11 targets with associated PDB IDs were analysed. Firstly, targets were analysed using the BioLip [41] database to determine if they contained biologically relevant ligands. Secondly, targets deemed to contain biologically relevant ligands were further investigated to determine ligand binding site residues, using the standard CASP distance cut-off; the Van der Waal radius of the contacting atom of a residue and the contacting ligand atom plus 0.5 Å. This resulted in a set of 11 proteins containing biologically relevant ligands and binding site residues.

In CASP11, the FunFOLD3 method was integrated into the IntFOLD TS predictions (TS133). Protein-ligand interactions were predicted for 8 out of the 11 FN targets (described above), with a mean MCC score of 0.554 and a mean BDT score of 0.478. Four of the top predictions, are subsequently discussed in detail. Figure 3 highlights the four assessed predictions, compared to the observed binding sites, with BDT scores ranging from 0.753 to 0.849. Figure 3A shows the predicted ligand binding site for a HAD-superfamily hydrolase, subfamily IA, variant 1 from *Geobacter sulfurreducens* (CASP ID T0854 and PDB ID 4rn3), with correctly predicted binding site residues in blue (16,18 and 173) and under (177) and over-predictions (19) in red, the MG ligand is coloured by element. The prediction resulted in a BDT score of 0.845 and an MCC score of 0.745. Figure 3B shows the observed binding site

for T0854 (PDB ID 4rn3), with binding site residues coloured in blue and the ligand MG coloured by element. A minority of residues were either under or over-predicted for this target as a result of the centroid ligand and the ligand cluster not being well superposed. The binding sites of the templates were not well superposed onto the model binding site, thus, the ligand cluster was not optimally located in the binding site.

The second CASP11 target is a cGMP Dependent Protein Kinase II from *Rattus norvegicus* (CASP ID T0798 and PDB ID 4ojk). Figure 3C shows the predicted ligand binding site, with correctly predicted binding site residues (14, 15, 16, 17, 18, 19, 29, 30, 31, 117, 118, 120, 121, 147, 148, 149) in blue and under (12, 32) and over-predictions (13, 33, 35, 36, 61, 62) in red, the GDP ligand is coloured by element. This prediction has a BDT score of 0.797 and an MCC score of 0.754. The observed ligand binding site for T0798 (PDB ID 4ojk), with binding site residues coloured in blue and the ligand GDP coloured by element can be seen in Figure 3D. Again, the minority of under and over-predictions are caused by firstly having a very large ligand binding site, which did not have the ligands cluster in the correct location within the large binding site, in part due to a number of templates having larger cofactor ligands and others having an additional MG ion bound with the cofactor.

The third example is of an aldo/keto reductase from *Klebsiella pneumoniae* (CASP ID T0807 and PDB ID 4wgh). Figure 3E shows the predicted ligand binding site, with correctly predicted binding site residues (20, 21, 22, 50, 55, 143, 165, 193, 194, 195, 196, 198, 199, 201, 224, 240, 241, 242, 244, 248, 251) in blue and under (80, 142, 243, 245, 252) and over-predictions (23, 54, 113, 197, 200, 207) in red, the NAP ligand is coloured by element. This prediction resulted in a BDT score of 0.849 and an MCC score of 0.771. In addition, the observed ligand binding site can be seen in Figure 3F, with binding site residues coloured in blue and the ligand NAP coloured by element. Furthermore, the over and under-predictions

seem to be a direct result of a number of templates having an additional ligand bound along with the cofactor, resulting in an extended ligand binding site.

The final CASP11 target that we will analyse is a histidinol-phosphate aminotransferase from *Sinorhizobium meliloti* (CASP ID T0819 and PDB ID 4wbt). Figure 3G shows the predicted ligand binding site, with correctly predicted binding site residues (93, 94, 95, 119, 167, 194, 197, 223, 225, 226, 234) in blue and under (161, 196) and over-predictions (347) in red, the PLP ligand is coloured by element. The prediction results in a BDT score of 0.753 and an MCC score of 0.877. In addition, Figure 3H shows the observed ligand binding site for T0819 (PDB ID 4wbt), with binding site residues coloured in blue and the ligand PLP coloured by element. Here, the under and over-predictions are a result of the incorrect orientation of residues in one case away from the binding site (TYR 161), in the other cases the under-predicted residue (ALA 196) and the over-predicted residue (ARG 347) are located on flexible loops.

These four CASP11 examples and the results [31-33] from previous CASP assessments, along with in-house evaluations [3,1], highlight the usefulness of the FunFOLD methods for the accurate prediction of protein-ligand interactions, for a wide range of proteins and ligand binding sites. See Note 4 for current method limitations.

4. Notes

1. When using the FunFOLD server [1,2,5], several problems may be encountered. These mainly include but are not limited to, providing the incorrect data to the server. It is important to input a sequence in plain text and single letter code format, into the text box labelled “Input sequence of target protein”. Additionally, it is recommended not to submit sequences longer than 500 amino acids. Firstly, these sequences usually contain multiple domains, thus it may not be possible to find a good template to

model multiple domains, resulting in one or more domains not being modelled well. Secondly, if both domains contain ligand binding sites only one will be predicted and displayed in the results page. Hence, it is advisable to partition the sequence into domains and submit each domain sequence as a separate job.

The next place where errors can occur is the next submission box “Short name for protein target”; inputting a short name for your protein sequence is useful to keep track of your prediction by providing a meaningful description. The short descriptor is limited to a set of characters: letters A to Z (either case), the numbers 0 to 9 and the following characters: .~_-. The protein descriptor supplied by the user, is subsequently utilized in the subject line of the email sent to the user, which contains a link to the FunFOLD results for their target protein.

The final text box to be completed is the “E-mail address”. This will enable a link of the graphical and machine readable results to be sent to the user, upon job completion. Here errors can occur if the user incorrectly inputs their email address.

2. For the downloadable Java application FunFOLD3, errors can occur but are not limited to the following reasons: Firstly, errors can occur if the dependencies - Java, TM-align [39], BioLip [41] and PyMOL - are not installed or not installed correctly; Secondly, if the full paths to the input files, BioLip database, CIF database and output directory are not included; Thirdly, if the target model to be analysed is not in the output directory; Fourthly, if the list of templates used in the prediction contains non-existent PDB IDs or the PDB IDs (including chain identifiers) are not all on the same line of the text file, the program will not run; Fifthly, if the input sequence file is not in FASTA format; Finally, it is recommended to limit the template list to 40 template structures, for efficient prediction and this is near the limit of the number of structure

files PyMOL can handle (See Section 3.2 and the README file downloaded with the executable).

Moreover, downloading the BioLip database may be time consuming and is an area where problems may occur if the instructions available on the BioLip website and contained in the README are not followed. Alternatively, if the user has the I-TASSER [57] pipeline installed on their system, the BioLip databases [41] will have been installed as part of the I-TASSER installation process.

3. The IntFOLD server [55,46] is a novel independent server, which gives users easy access to a number of cutting edge methods, for the prediction of structure and function from sequence. The idea behind the IntFOLD server is to provide easy access to our methods from a single location, producing easily understandable integrated output of results, enabling ease of access for the non-expert user. The IntFOLD server provides output in graphical form, enabling users to interpret results at a glance as well as CASP formatted text files, allowing a more in-depth analysis of the prediction results. The IntFOLD pipeline integrates a number of methods, to enable users to simply input a target sequence and produce a set of models (IntFOLD3-TS [38]), with associated global and per-residue model quality (ModFOLD5 [56]), disorder prediction (DISOclust3 [58]), domain partitioning (DomFOLD3) and function prediction results utilizing FunFOLD3 [1,2,5]. The component methods of the IntFOLD server have been ranked amongst the top methods in their respective categories at recent CASP and CAMEO competitions.
4. Predicting protein-ligand interactions is a difficult task, which results in a number of limitations to current prediction methods. The following is a non-exhaustive list of the most common limitations currently encountered in the field; 1. If the server or prediction algorithm is unable to build a model for the target sequence, then no

protein-ligand interactions are predicted. The solution to this problem is to utilize sequence based methods (see Section 1.1 for suggestions of sequence based prediction methods), which are less accurate; 2. If structurally similar templates to the target, which containing biologically relevant ligands cannot be found, then no prediction can be made; 3. The FunFOLD server currently outputs predictions based on the top IntFOLD model, which has the highest global model quality score. This model may not have the best per-residue model quality around the binding site location, resulting in under or over-predicted ligand binding site residues.

5. The user has the option of using the server version of FunFOLD, IntFOLD, or the downloadable java application. The user has to leverage which option they would like to utilize. The server only permits users to submit one job at a time due to server load balancing. If the user would like to carry our large scale analysis, for example predicting protein-ligand interactions for a proteome, it is then recommended to download and use the executable java application for FunFOLD3. This allows the user the freedom in the number of structures they can analyse, provided they have adequate CPU capacity.

   For light use (several predictions a week), server prediction is adequate for the user, whereas for heavy users (greater than 5-10 predictions a week) the downloadable application would be the most useful. Extensive help pages are available for the FunFOLD server. Furthermore, at least 30 GB of disc space is required to download the complete BioLip libraries. In addition, an extensive README file, example input and output files are available to aid the user in the installation and running of the FunFOLD3 downloadable java application.

Acknowledgements

Grant IBC ANR Investissements D'Avenir (To D.B.R).

5. References:

1. Roche DB, Buenavista MT, McGuffin LJ (2012) FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions. PloS one 7 (5):e38219. doi:10.1371/journal.pone.0038219
2. Roche DB, Buenavista MT, McGuffin LJ (2013) The FunFOLD2 server for the prediction of protein-ligand interactions. Nucleic acids research 41 (Web Server issue):W303-307. doi:10.1093/nar/gkt498
3. Roche DB, Tetchner SJ, McGuffin LJ (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC bioinformatics 12 (1):160. doi:1471-2105-12-160 [pii]

10.1186/1471-2105-12-160
4. Oh M, Joo K, Lee J (2009) Protein-binding site prediction based on three-dimensional protein modeling. Proteins 77 Suppl 9:152-156. doi:10.1002/prot.22572
5. Roche DB, Tetchner SJ, McGuffin LJ (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC bioinformatics 12:160. doi:10.1186/1471-2105-12-160
6. Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML (2011) firestar--advances in the prediction of functionally important residues. Nucleic acids research 39 (Web Server issue):W235-241. doi:10.1093/nar/gkr437
7. Lopez G, Valencia A, Tress ML (2007) firestar--prediction of functionally important residues using structural templates and alignment reliability. Nucleic acids research 35 (Web Server issue):W573-577. doi:gkm297 [pii]

10.1093/nar/gkm297
8. Talavera D, Laskowski RA, Thornton JM (2009) WSsas: a web service for the annotation of functional residues through structural homologues. Bioinformatics 25 (9):1192-1194. doi:btp116 [pii]

10.1093/bioinformatics/btp116
9. Sankararaman S, Kolaczkowski B, Sjolander K (2009) INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. Nucleic acids research 37 (Web Server issue):W390-395. doi:gkp339 [pii]

10.1093/nar/gkp339
10. Ye K, Feenstra KA, Heringa J, Ijzerman AP, Marchiori E (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. Bioinformatics 24 (1):18-25. doi:10.1093/bioinformatics/btm537
11. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic acids research 38 Suppl:W529-533. doi:gkq399 [pii]

10.1093/nar/gkq399
12. Wass MN, Sternberg MJ (2008) ConFunc--functional annotation in the twilight zone. Bioinformatics 24 (6):798-806. doi:btn037 [pii]

10.1093/bioinformatics/btn037
13. Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjolander K (2010) Active site prediction using evolutionary and structural information. Bioinformatics 26 (5):617-624. doi:btq008 [pii]

10.1093/bioinformatics/btq008

14. Dong-Jun Y, Jun H, Jing Y, Hong-Bin S, Jinhui T, Jing-Yu Y (2013) Designing Template-Free Predictor for Targeting Protein-Ligand Binding Sites with Classifier Ensemble and Spatial Clustering. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 10 (4):994-1008. doi:10.1109/TCBB.2013.104
15. Chen P, Huang JHZ, Gao X (2014) LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone. BMC bioinformatics 15. doi:Unsp S4

Doi 10.1186/1471-2105-15-S15-S4
16. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A 105 (1):129-134. doi:0707684105 [pii]

10.1073/pnas.0707684105
17. Spitzer R, Cleves AE, Jain AN (2011) Surface-based protein binding pocket similarity. Proteins 79 (9):2746-2763. doi:10.1002/prot.23103
18. Xie ZR, Liu CK, Hsiao FC, Yao A, Hwang MJ (2013) LISE: a server using ligand-interacting and site-enriched protein triangles for prediction of ligand-binding sites. Nucleic acids research 41 (Web Server issue):W292-296. doi:10.1093/nar/gkt300
19. Zhu X, Xiong Y, Kihara D (2015) Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0. Bioinformatics 31 (5):707-713. doi:10.1093/bioinformatics/btu724
20. Cao Y, Li L (2014) Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model. Bioinformatics 30 (12):1674-1680. doi:10.1093/bioinformatics/btu104
21. Fuller JC, Martinez M, Henrich S, Stank A, Richter S, Wade RC (2014) LigDig: a web server for querying ligand-protein interactions. Bioinformatics. doi:10.1093/bioinformatics/btu784
22. Erdin S, Ward RM, Venner E, Lichtarge O (2010) Evolutionary trace annotation of protein function in the structural proteome. Journal of molecular biology 396 (5):1451-1473. doi:10.1016/j.jmb.2009.12.037
23. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. Journal of molecular biology 316 (1):139-154. doi:10.1006/jmbi.2001.5327
24. Hernandez M, Ghersi D, Sanchez R (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic acids research 37 (Web Server issue):W413-416. doi:gkp281 [pii]

10.1093/nar/gkp281
25. Yang J, Roy A, Zhang Y (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29 (20):2588-2595. doi:10.1093/bioinformatics/btt447
26. Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic acids research 40 (Web Server issue):W471-477. doi:10.1093/nar/gks372
27. Heo L, Shin WH, Lee MS, Seok C (2014) GalaxySite: ligand-binding-site prediction by using molecular docking. Nucleic acids research 42 (Web Server issue):W210-214. doi:10.1093/nar/gku321
28. Izidoro SC, de Melo-Minardi RC, Pappa GL (2014) GASS: identifying enzyme active sites with genetic algorithms. Bioinformatics. doi:10.1093/bioinformatics/btu746

29. Huang B, Schroeder M (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 6:19. doi:1472-6807-6-19 [pii]

10.1186/1472-6807-6-19
30. Andersson CD, Chen BY, Linusson A (2010) Mapping of ligand-binding cavities in proteins. Proteins 78 (6):1408-1422. doi:10.1002/prot.22655
31. Lopez G, Ezkurdia I, Tress ML (2009) Assessment of ligand binding residue predictions in CASP8. Proteins 77 Suppl 9:138-146. doi:10.1002/prot.22557
32. Schmidt T, Haas J, Cassarino TG, Schwede T (2011) Assessment of ligand binding residue predictions in CASP9. Proteins: Structure, Function, and Bioinformatics 79 Suppl 10:126-136. doi:10.1002/prot.23174
33. Gallo Cassarino T, Bordoli L, Schwede T (2014) Assessment of ligand binding site predictions in CASP10. Proteins 82 Suppl 2:154-163. doi:10.1002/prot.24495
34. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T (2013) The Protein Model Portal--a comprehensive resource for protein structure and model information. Database : the journal of biological databases and curation 2013:bat031. doi:10.1093/database/bat031
35. Wass MN, Sternberg MJ (2009) Prediction of ligand binding sites using homologous structures and conservation at CASP8. Proteins 77 Suppl 9:147-151. doi:10.1002/prot.22513
36. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405 (2):442-451
37. Roche DB, Tetchner SJ, McGuffin LJ (2010) The binding site distance test score: a robust method for the assessment of predicted protein binding sites. Bioinformatics 26 (22):2920-2921. doi:10.1093/bioinformatics/btq543
38. Buenavista MT, Roche DB, McGuffin LJ (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. Bioinformatics 28 (14):1851-1857. doi:10.1093/bioinformatics/bts292
39. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research 33 (7):2302-2309. doi:33/7/2302 [pii]

10.1093/nar/gki524
40. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research 33 (7):2302-2309. doi:10.1093/nar/gki524
41. Yang J, Roy A, Zhang Y (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic acids research 41 (Database issue):D1096-1103. doi:10.1093/nar/gks966
42. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26 (7):889-895. doi:10.1093/bioinformatics/btq066
43. McGuffin LJ, Roche DB (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics 26 (2):182-188. doi:btp629 [pii]

10.1093/bioinformatics/btp629
44. Webb EC (1989) Nomenclature Committee of the International-Union-of-Biochemistry (Nc-Iub) - Enzyme Nomenclature - Recommendations 1984 - Supplement-2 - Corrections and Additions. Eur J Biochem 179 (3):489-533. doi:DOI 10.1111/j.1432-1033.1989.tb14579.x
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese

JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Consortium GO (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25 (1):25-29
46. McGuffin LJ, Atkins JD, Salehe BR, Shuid AN, Roche DB (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. Nucleic acids research. doi:10.1093/nar/gkv236
47. Bindschedler LV, McGuffin LJ, Burgis TA, Spanu PD, Cramer R (2011) Proteogenomics and in silico structural and functional annotation of the barley powdery mildew Blumeria graminis f. sp. hordei. Methods 54 (4):432-441. doi:10.1016/j.ymeth.2011.03.006
48. Pedersen C, Ver Loren van Themaat E, McGuffin LJ, Abbott JC, Burgis TA, Barton G, Bindschedler LV, Lu X, Maekawa T, Wessling R, Cramer R, Thordal-Christensen H, Panstruga R, Spanu PD (2012) Structure and evolution of barley powdery mildew effector candidates. BMC genomics 13:694. doi:10.1186/1471-2164-13-694
49. Zhou Y, Xue S, Yang JJ (2013) Calciomics: integrative studies of Ca2+-binding proteins and their interactomes in biological systems. Metallomics : integrated biometal science 5 (1):29-42. doi:10.1039/c2mt20009k
50. Don CG, Riniker S (2014) Scents and sense: in silico perspectives on olfactory receptors. Journal of computational chemistry 35 (32):2279-2287. doi:10.1002/jcc.23757
51. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic acids research 42 (Database issue):D222-230. doi:10.1093/nar/gkt1223
52. Letunic I, Doerks T, Bork P (2015) SMART: recent updates, new developments and status in 2015. Nucleic acids research 43 (Database issue):D257-260. doi:10.1093/nar/gku949
53. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. Bioinformatics 20 (13):2153-2155. doi:10.1093/bioinformatics/bth214
54. Roche DB, Buenavista MT, McGuffin LJ (2014) Assessing the quality of modelled 3D protein structures using the ModFOLD server. Methods in molecular biology 1137:83-103. doi:10.1007/978-1-4939-0366-5_7
55. Roche DB, Buenavista MT, Tetchner SJ, McGuffin LJ (2011) The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. Nucleic acids research 39 (Web Server issue):W171-176. doi:10.1093/nar/gkr184
56. McGuffin LJ, Buenavista MT, Roche DB (2013) The ModFOLD4 server for the quality assessment of 3D protein models. Nucleic acids research 41 (Web Server issue):W368-372. doi:10.1093/nar/gkt294
57. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5 (4):725-738. doi:nprot.2010.5 [pii]

10.1038/nprot.2010.5
58. McGuffin LJ (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics 24 (16):1798-1804. doi:10.1093/bioinformatics/btn326

Figure Legends

Figure 1. Submission page for the FunFOLD server.

Figure 2. The IntFOLD3-FN (FunFOLD3) server results page for CASP11 target T0807 (PDB ID - 4wgh).

Figure 3. Comparison of FunFOLD3 ligand binding site predictions (A, C, E, G) for 4 CASP 11 targets, compared to the observed ligand binding sites (B, D, F, H). **A**. Predicted ligand binding site for T0854 (PDB ID 4rn3), with correctly predicted binding site residues in blue and under and over-predictions in red, the MG ligand is coloured by element. BDT score of 0.845 and MCC score of 0.745. **B**. The observed ligand binding site for T0854 (PDB ID 4rn3), with binding site residues coloured in blue and the ligand MG coloured by element. **C**. Predicted ligand binding site for T0798 (PDB ID 4ojk), with correctly predicted binding site residues in blue and under and over-predictions in red, the GDP ligand is coloured by element. BDT score of 0.797 and MCC score of 0.754. **D**. The observed ligand binding site for T0798 (PDB ID 4ojk), with binding site residues coloured in blue and the ligand GDP coloured by element. **E**. Predicted ligand binding site for T0807 (PDB ID 4wgh), with correctly predicted binding site residues in blue and under and over-predictions in red, the NAP ligand is coloured by element. BDT score of 0.849 and MCC score of 0.771. **F**. The observed ligand binding site for T0807 (PDB ID 4wgh), with binding site residues coloured in blue and the ligand NAP coloured by element. **G**. Predicted ligand binding site for T0819 (PDB ID 4wbt), with correctly predicted binding site residues in blue and under and over-predictions in red, the PLP ligand is coloured by element. BDT score of 0.753 and MCC score of 0.877. **H**. The observed ligand binding site for T0819 (PDB ID 4wbt), with binding site residues coloured in blue and the ligand PLP coloured by element. All images were rendered using PyMOL (http://www.pymol.org/).