Bioinformatics, 37(3), 2021, 351–359
doi: 10.1093/bioinformatics/btaa712
Advance Access Publication Date: 11 August 2020
Original Paper

Structural bioinformatics

ResiRole: residue-level functional site predictions to gauge the accuracies of protein structure prediction techniques

Joshua M. Toth[1], Paul J. DePietro[1], Juergen Haas[2] and William A. McLaughlin [1,*]

[1]Department of Medical Education, Geisinger Commonwealth School of Medicine, Scranton, PA 18510, USA and [2]Biozentrum, University of Basel and SIB Swiss Institute of Bioinformatics, CH-4056 Basel, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on December 12, 2019; revised on July 31, 2020; editorial decision on August 3, 2020; accepted on August 5, 2020

Abstract

Motivation: Methods to assess the quality of protein structure models are needed for user applications. To aid with the selection of structure models and further inform the development of structure prediction techniques, we describe the ResiRole method for the assessment of the quality of structure models.

Results: Structure prediction techniques are ranked according to the results of round-robin, head-to-head comparisons using difference scores. Each difference score was defined as the absolute value of the cumulative probability for a functional site prediction made with the FEATURE program for the reference structure minus that for the structure model. Overall, the difference scores correlate well with other model quality metrics; and based on benchmarking studies with NaïveBLAST, they are found to detect additional local structural similarities between the structure models and reference structures.

Availabilityand implementation: Automated analyses of models addressed in CAMEO are available via the ResiRole server, URL http://protein.som.geisinger.edu/ResiRole/. Interactive analyses with user-provided models and reference structures are also enabled. Code is available at github.com/wamclaughlin/ResiRole.

Contact: wmclaughlin@som.geisinger.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As three-dimensional protein structures are currently available for only a fraction of all known protein sequences, protein structure prediction techniques are utilized to expand the coverage of protein structure space (Baker and Sali, 2001). Once generated, protein structure models may be examined to gain insights into the relationships between three-dimensional structure and biological function (Grabowski *et al.*, 2007). One question is whether structure models are accurate enough to enable the identification of functional sites that are localized around specific residues. Such identification was shown to be possible, as described for the functional assessments of the CASP12 experiment (Liu *et al.*, 2018). Also, methods like COFACTOR (Zhang *et al.*, 2017), 3DLigandSite (Wass *et al.*, 2010) and FunFold (Roche and McGuffin, 2016) have independently shown that functional sites in experimental structures can be reconstituted in structure models.

Here, we address the somewhat reversed question of whether protein structure prediction techniques can be effectively ranked according to their capacities to generate structure models with functional site predictions like those of experimental reference structures. To address this question and to provide an additional means to assess model quality, we developed the residue role in assessment method (ResiRole) to examine the matching between the functional site predictions made for reference structures versus those made at the corresponding sites in the structure models. The premise is that if a functional site is predicted to be centered on a specific residue within a reference structure and similarly predicted for a structure model, then the structure model has reconstituted the structural and physiochemical environment necessary for the functional site prediction.

To perform the study, we retrieved collections of structure models available through the Continuous Automated Model EvaluatiOn (CAMEO) server (Haas *et al.*, 2018). CAMEO retrieves the amino acid sequences of structures to be released in the Protein Data Bank (PDB), a few days prior to the release of the three-dimensional coordinates (Berman, 2000). CAMEO submits a selected set of these sequences to the enrolled structure prediction servers; and timestamped structure prediction results are generated. The structure models produced are subsequently evaluated using model quality assessment (MQA) programs, such as lDDT (Mariani *et al.*, 2013) and GDT-TS (Zemla *et al.*, 1999). The entire protocol is run with a weekly update cycle in coordination with the release of structures from the worldwide PDB (Berman *et al.*, 2007).

The ResiRole method considers functional site predictions centered on specific amino acid residues, as analyzed with the FEATURE program (Wu *et al.*, 2008). FEATURE was selected due to the breadth of the functional sites addressed and the availability of benchmarking data (Buturovic *et al.*, 2014). The FEATURE program enumerates the physiochemical properties of the environment surrounding an anchor atom of a given residue into a computational vector representation via sampling of multiple localized concentric spherical volumes. The physiochemical properties addressed include descriptions such as atom and residue types, partial charges, secondary structure assignments and van der Waals radii. FEATURE can test this representation against the representation of a known or potential functional site to predict the likelihood that the environment surrounding the anchor atom constitutes a functional site. The types of functional sites analyzed include small molecule and ion binding sites which have corresponding sequence motifs described in PROSITE (Hulo, 2006). Here, we measure the average degree of matching between the functional site predictions of the reference experimental structures versus those for the structure models to compare structure prediction techniques.

2 Materials and methods

2.1 Overview of the analysis pipeline

A flow diagram that describes the overall stages involved in comparing the structure prediction techniques is provided inFigure 1. In stage A, structure models produced by structure prediction techniques hosted in CAMEO (Haas *et al.*, 2018) were retrieved. CAMEO also provided a source to retrieve the coordinates of the corresponding PDB reference structures. In stage B, functional site predictions were made with the SeqFEATURE models using the FEATURE program (Wu *et al.*, 2008).

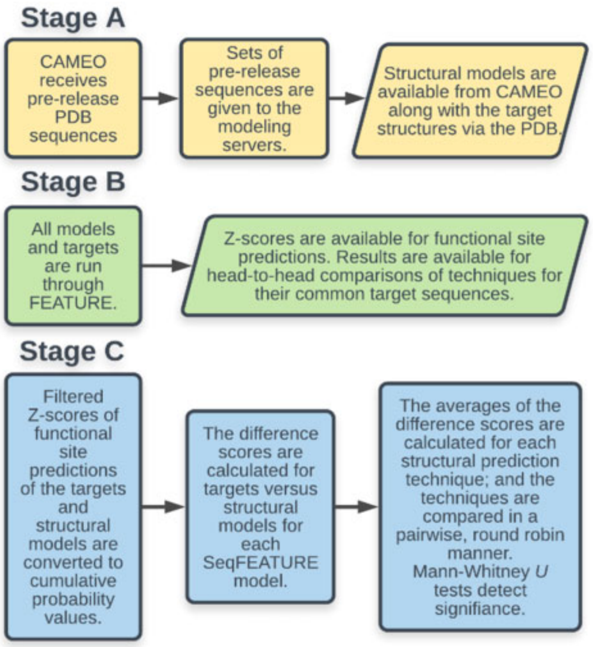
Stage A
CAMEO receives pre-release PDB sequences
Sets of pre-release sequences are given to the modeling servers.
Structural models are available from CAMEO along with the target structures via the PDB.
Stage B
All models and targets are run through FEATURE.
Z-scores are available for functional site predictions. Results are available for head-to-head comparisons of techniques for their common target sequences.
Stage C
Filtered Z-scores of functional site predictions of the targets and structural models are converted to cumulative probability values.
The difference scores are calculated for targets versus structural models for each SeqFEATURE model.
The averages of the difference scores are calculated for each structural prediction technique; and the techniques are compared in a pairwise, round robin manner. Mann-Whitney U tests detect signifiance.

Fig. 1. Flow diagram of the data analysis stages for comparing structure prediction techniques regarding their capacities to have functional site predictions like those of the reference structures

In stage C, each difference score was obtained as the absolute value of the functional site cumulative prediction probability for a site in the reference structure minus that for the corresponding site in the structure model. For each structure prediction technique, the difference scores were averaged across all the analyses done separately with each of the functional site prediction models (SeqFEATURE models).

Structure prediction techniques were compared in a pairwise, head-to-head, round-robin manner and ranked according to their overall average difference scores. Mann–Whitney U tests were used to estimate the significance level for each pairwise comparison of structure prediction techniques.

2.2 Acquisition of the reference structures and structure models

Protein structure models were generated from the sets of prereleased target sequences from the PDB (Berman, 2000) using the structure prediction techniques represented in CAMEO. The techniques included HHPredB (Söding *et al.*, 2005), IntFOLD2-TS (Buenavista *et al.*, 2012), IntFOLD3-TS (McGuffin *et al.*, 2015), IntFOLD4-TS (McGuffin *et al.*, 2018), M4T (Fernandez-Fuentes *et al.*, 2007), NaïveBLAST (Cozzetto *et al.*, 2007), Phyre2 (Kelley *et al.*, 2015), PRIMO (Hatherley *et al.*, 2016), Princeton-TEMPLATE (Khoury *et al.*, 2014), RaptorX (Källberg *et al.*, 2012), RBO Aleph (Mabrouk *et al.*, 2015), Robetta (Kim *et al.*, 2004), SPARKS-X (Yang *et al.*, 2011) and SWISS-MODEL (Schwede, 2003). The NaïveBLAST method selects the first template returned by searching the target protein sequence against the full PDB archive and employs MODELLER (Šali and Blundell, 1993) with default parameters to produce a baseline structure model.

Although CAMEO offers its benchmarking data for multiple public modeling techniques, only the structure prediction techniques that modeled at least 100 target sequences in common with each other during the time frame considered, August 8, 2014 to February 24, 2018, were included for the study. For a complete list of the reference structures used for each pairwise comparison, see the data file provided in Supplementary Material.

2.3 Predictions of the functional sites

All coordinate data for the reference structures and the protein structure models were analyzed using the FEATURE program (version 3.0), URL https://simtk.org/projects/feature (Halperin *et al.*, 2008). FEATURE is trained to associate certain physiochemical environments with known functional sites utilizing supervised machine learning. The results are a collection of classification models, called SeqFEATURE models (Wu *et al.*, 2008), that can be used to predict the likelihood that a given environment has structural features like the classification model against which it is scored. FEATURE uses a Bayesian scoring algorithm, which treats each property as an independent event. We converted the raw prediction score produced by FEATURE for each SeqFEATURE model to a Z-score based on the mean and standard deviation of the dataset that consisted of all predictions for that SeqFEATURE model made for the reference structures. As experimentally determined structures can sometimes lack coordinate data for regions that are available in the structure models and vice versa, only residues for which coordinate data existed in both were used in the analyses.

2.4 Comparisons of the structure prediction techniques using functional site predictions

The difference score was defined as the absolute value of the cumulative probability obtained for a functional site prediction within the reference structure minus the cumulative probability obtained for the function site prediction at the corresponding site within the structure model. The cumulative probabilities were obtained by converting the Z-scores to cumulative probabilities using the cumulative density function in SciPy 1.1.0 (Jones *et al.*, 2014).

A supporting analysis stage was performed to define the range of probability values to be utilized for the study. For that purpose, a *matching* specificity for each SeqFEATURE model was calculated as the number of instances predicted as negative for both the reference structures and structure models divided by the total number of negatives predicted for the reference structures. We found that a 90% matching specificity was, within the errors of the measurements, identical to a functional site prediction specificity of 90%. Based on this finding and to focus on predictions which have relatively high *Z*-scores, only predictions in the reference structures that had *Z*-score values corresponding to functional site prediction specificities within the range of 90–100 were utilized. For that purpose, the *Z*-scores that corresponded to specificity thresholds of 90% for each SeqFEATURE model were extracted from the results of a benchmarking study by Buturovic *et al.* (2014). These thresholds were applied such that only those functional site predictions that had a *Z*-score greater than the *Z*-scores corresponding to specificity levels of 90% for the SeqFEATURE models for the reference structures were included. All other functional site predictions were removed or 'filtered' from the subsequent analyses. Further details regarding the supporting analysis stage are described in Supplementary Material.

Using the remaining set of 'filtered' functional site predictions, an average difference score was calculated separately for each SeqFEATURE model for the set of structure models produced by each structure prediction technique. The protein structure prediction techniques were then compared in a pairwise head-to-head, round-robin manner using Mann–Whitney *U* tests that were performed on the lists of average difference scores from the SeqFEATURE models. For each pairwise comparison of structure prediction techniques, only target sequences modeled in common between the two techniques were considered.

All comparisons between structure prediction techniques were repeated after the targets were categorized according to lDDT score ranges to produce easy, medium and hard subsets, as described by Haas *et al.* (2018). These categories corresponded, respectively, to lDDT score ranges of greater than or equal to 75, between 50 and 75 and <50.

2.5 Comparisons of the difference score to other assessment metrics

The difference scores were compared with other metrics for assessing model quality. These metrics included lDDT (Mariani *et al.*, 2013), which provides a measure of local model quality. lDDT-BS was used as a local measure of model quality around ligands found in complex within the reference structures (Haas *et al.*, 2018). TM-score (Zhang and Skolnick, 2005), GDT-TS (Zemla *et al.*, 1999), GDT-HA (Read and Chavali, 2007) and GDC (Keedy *et al.*, 2009; Mariani *et al.*, 2013) provided measurements of global model quality. For the comparisons to the difference score, averages of the metrics were found separately for the entire set of structure models produced by each structure prediction technique. Each other metric was averaged across all targets that received at least one score according to a SeqFEATURE model.

The average difference scores obtained for the structure predictions techniques were plotted against the averages of the other metrics for assessing model quality. Best fit linear regression lines were obtained for each of the plots to interpret the degree to which the difference score provides a comparable estimate of structure model quality. Standardized residuals obtained for deviations of individual structure prediction techniques away from their predicted values based on the best fit lines were used to interpret potential outliers.

3 Results

3.1 Example illustration of functional site predictions

Consider an example of a functional site prediction in a reference structure and the corresponding predictions within structure models. Figure 2 provides a representation of the crystal structure of the group I dockerin domain of hydrolase GDSL protein from the bacteria *Ruminococcus flavefaciens* (PDB ID: 5M2O, chain B) (Bule *et al.*, 2017). The experimental structure is shown to be aligned with the structure models produced by RaptorX and Robetta using PyMOL (DeLano, 2002). The type of functional site is a calcium binding site centered on the oxygen of asparagine's side chain. The functional site prediction model, or SeqFEATURE model, is abbreviated as EF_HAND_1.5.ASN.OD1. For the asparagine at position 32, a positive prediction was made in the reference structure; but negative predictions were made for the Robetta and RaptorX models. Notice that the asparagines in the RaptorX and Robetta models adopt different rotamer states relative to that found for the reference structure. These changes in conformations contribute to alterations of the physiochemical properties assigned to the feature vectors used for the function site predictions such that the RaptorX model received a 3.49% functional site prediction cumulative probability and the Robetta model received a 26.27% probability. The target, in which the anchor asparagine is known to participate in a calcium ion binding site (green), received a 99.96% functional site prediction probability. The examples in Figure 1 illustrate that the prediction probabilities can be used to measure the degree to which the physicochemical and structural environment of a functional site has been accurately reconstituted within a structure model.

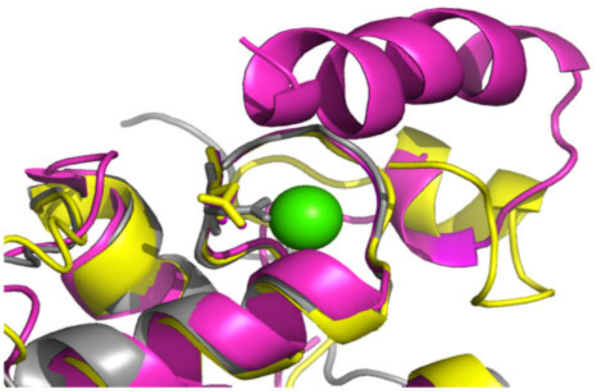

Fig. 2. Comparison of the crystal structure of the group I dockerin domain of hydrolase GDSL along with structure models at the location of a calcium binding site. The function is predicted to be centered on an asparagine residue, which is shown in stick representations. The experimental structure is in white, whereas the Robetta model and RaptorX models are in yellow and magenta, respectively. The calcium ion from the reference structure is shown in green

3.2 Example calculations of the difference score

To further illustrate how the difference scores were calculated, consider examples of functional site predictions made with the EF_HAND_1.5.ASN.OD1 SeqFEATURE model. Shown in Figure 3 is a plot of the cumulative probabilities of functional site predictions for the reference structures versus those made at the corresponding sites in the Robetta models, which had targets in common with RaptorX models.

Included in the plot are functional site predictions that had *Z*-scores in the reference structures above a functional prediction specificity of 90%.

See that there is a wide distribution along the ordinate of the scatter plot. But a significant correlation between the probabilities found for the reference structures versus the models is apparent. The correlation coefficient, Pearson's *r* is 0.1358, with an associated *P* value of 7.108×10^{-9}. The point indicated by the arrow corresponds to the probabilities of the predictions in the reference structure and the Robetta model that are described in Figure 2. The difference score for that point is 0.9996–0.2627, or 0.74.

3.3 Comparisons of structure prediction techniques

The lists of average difference scores associated with the SeqFEATURE functional site models that were applied to the sets of structure models generated by each the structure prediction technique were used to compare the structure prediction techniques in a head-to-head, round-robin manner using Mann–Whitney *U* tests.

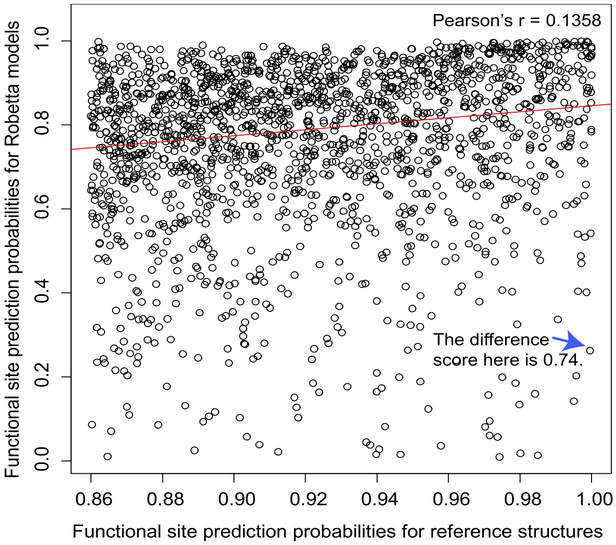

Fig. 3. Scatter plot of all the cumulative probabilities for functional site predictions made in the reference structures versus those for the corresponding sites in the Robetta models based on the SeqFEATURE model EF_HAND_1.5.ASN.OD1. Pearson's r is 0.1358 for the least-squares regression line. An example of a difference score calculation of 0.74 is shown for the point indicated by the arrow

The results are presented in Table 1. A total of 3382 targets were utilized for the study based on the available structure models from CAMEO for the study's time frame. We see that the difference scores provide a way to obtain a relatively high statistical significance for most the pairwise comparisons of the structure prediction techniques. The *P* values in yellow are those below a threshold value of 6.54×10^{-7}, the required threshold to meet the Bonferroni correction with a base *P* value of 0.0001 for a total of 153 different pairwise comparisons conducted across the 18 different structure prediction techniques. The top five structure prediction techniques, ranked by their global average difference scores were Robetta, IntFOLD4-TS, SWISS-MODEL, M4T and NaïveBLAST.

Our submitted question was whether structure prediction techniques can be effectively ranked according their capacities to enable residue-level functional site predictions. Based on the head-to-head pairwise comparisons, we infer that the average difference score provides an effective means to do so. We infer that if a given a pair of structure prediction techniques were found to differ significantly by their average difference scores, the one with lower difference score can, on average, produce structure models which reconstitute more of the local structural features of the reference structures at the predicted functional sites.

3.4 Results for the hard targets

Results of the assessments of the structure prediction techniques can vary according the difficulty of modeling the target sequences. To address that matter, CAMEO has devised categories of target sequences based on lDDT score ranges (Haas *et al.*, 2018); these designations are easy, medium and hard. The average difference scores for structure prediction techniques for structure models produced for hard target sequences are provided in Table 2. Here, Robetta, IntFOLD4-TS, SWISS-MODEL, M4T and RaptorX were found to be the top five structure prediction techniques regarding their abilities to produce structure models with functional site predictions like those of the reference structures. Otherwise, the order would be different if going by lDDT. The results of all pairwise, head-to-head comparisons for the easy and medium targets are provided in Supplementary Material.

3.5 Consideration of NaïveBLAST as a benchmark technique

The NaïveBLAST technique provides a baseline method to evaluate the other structure prediction techniques because it identifies the sequence of an available three-dimensional structure which is similar to the target sequence based on a BLAST search (Altschul, 1997). The carbon alpha backbone of the closest template then provides the structure on which to perform template-based structure prediction with MODELLER (Šali and Blundell, 1993). The resulting model is energy minimized. By comparing the model produced from another structure prediction technique with the structure model produced by NaïveBLAST, the degree to which the former model yields additional information found in the reference structure, but cannot be obtained from the nearest template structure, can be estimated.

As shown in Table 1, we see that Robetta, IntFOLD4-TS, SWISS-MODEL and M4T outperform NaïveBLAST according to the difference scores regarding the analyses done for all targets. That is, these structure prediction techniques produced structure models that had, on average, difference scores significantly better when compared to the structure models produced by NaïveBLAST. For the hard target sequence category, RaptorX, Phyre2 and HHPredB, and RBO Aleph additionally outperformed NaïveBLAST. Our interpretation is that the models produced by these structure prediction techniques have, on average, reconstituted more of the structural features associated with the functional site predictions than the structure models generated by NaïveBLAST for the hard targets. The results provide evidence that these structure models contain additional information regarding the structural features required for functional site predictions and corroborate previous findings that describe the utility of structure models for enabling functional site predictions (Liu *et al.*, 2018), especially for the target sequences in the hard category.

3.6 Comparison the difference score to other assessment metrics

Consider the correlations between the difference score and other metrics which assess the quality of structure models in relative to the reference structures. In Figure 4, we present scatter plots of the average difference scores versus the averages for the other metrics, as obtained for each of the 18 structure prediction techniques. Each plot has a significant correlation coefficient, as indicated by the associated *P* value.

See the linear regression fit for the plot of the average difference scores versus the average GDC values (Fig. 4D). We see that the Pearson *r* value is -0.856, which is higher than the Pearson *r* value for the plot of the difference scores versus GDT-TS (Fig. 4B), which was -0.7785. The difference between these two correlation coefficients is significant at a *P* value of .05, that is, when considering that Pearson *r* for the correlation between GDC and GDT-TS is 0.9606, as found using the paired.r module in R. For the same comparison based on the hard targets only, the *P* value was 0.02.

GDC examines the distances between all superposed atoms, whereas GDT-TS examines only the interatomic distances of the superposed alpha carbon atoms. We infer that the higher absolute value of Pearson *r* was obtained for GDC because functional site predictions are more accurately represented by the relative distances between all atoms of the side chain residues rather than just the atoms of the alpha carbon backbone. Since GDT-TS only examines the distances between the aligned and superposed alpha carbon atoms, the lower correlation may mean that GDT-TS does not capture as many of the structural features required for the functional site predictions described by the SeqFEATURE models, as compared to GDC when using the difference score as the benchmark metric.

See that the average difference score versus the average lDDT-BS metric is shown in Figure 4F. The lDDT-BS metric describes the average lDDT score for atoms surrounding a ligand in complex with the reference structure (Haas *et al.*, 2018). The difference score correlates well with the lDDT-BS metric with a Pearson *r* value of -0.877. These metrics complement each other in the sense that difference score calculation does not require the reference structure to be

Table 1. Results of round-robin, head-to-head comparisons using difference scores

Server	DS	RB	I4	SM	M4T	NB	HB	I3	P2	RX	I2	RA	PBCL	PR	SX	PB3D	PH3D	PHCL
RB	0.1528																	
I4	0.1718	3.E-01 213																
SM	0.1721	*** 1509	*** 397															
M4T	0.1736	*** 659	*** 166	*** 1135														
NB	0.177	*** 1107	*** 326	*** 2247	*** 777													
HB	0.1786	*** 993	8.E-05 281	*** 1834	*** 580	*** 1386												
I3	0.1795	*** 436	*** 349	3.E-01 699	*** 308	*** 545	*** 491											
P2	0.1799	*** 1230	*** 268	*** 2532	*** 997	*** 1690	*** 1304	*** 548										
RX	0.1802	*** 1489	*** 402	6.E-04 2621	3.E-03 978	*** 1919	*** 1665	4.E-01 707	*** 2040									
I2	0.1862	*** 948	*** 197	6.E-04 1230	* 510	*** 867	*** 744	1.E-01 410	*** 1021	*** 1217								
RA	0.1909	*** 642	*** 229	*** 1284	*** 674	*** 804	*** 642	*** 327	*** 1156	*** 1028	*** 552							
PBCL	0.192	*** 307	*** 302	*** 875	5.E-01 178	*** 791	*** 572	*** 361	* 494	*** 817	*** 219	*** 277						
PR	0.1924	*** 308	*** 303	*** 877	5.E-01 178	*** 791	*** 576	*** 362	* 492	*** 818	*** 219	*** 275	4.E-01 885					
SX	0.1954	*** 815	*** 360	*** 2161	*** 847	2.E-01 1454	*** 979	*** 666	*** 1770	*** 1720	*** 741	*** 955	* 750	* 752				
PB3D	0.1968	*** 305	*** 302	*** 874	*** 178	3.E-01 790	*** 575	*** 357	*** 491	*** 812	*** 218	*** 274	*** 879	*** 880	*** 748			
PH3D	0.1998	*** 284	*** 263	*** 818	*** 158	*** 733	*** 556	*** 333	*** 445	*** 772	*** 200	*** 244	*** 799	*** 803	*** 707	* 796		
PHCL	0.2014	*** 289	*** 272	*** 836	*** 158	*** 743	*** 570	*** 341	*** 457	*** 786	*** 206	*** 248	*** 814	*** 817	*** 720	*** 811	* 825	
PT	0.2075	*** 973	*** 357	*** 2285	*** 1007	*** 1523	*** 1065	*** 631	*** 2077	*** 1790	*** 862	*** 1181	*** 560	*** 559	*** 1839	*** 557	*** 516	*** 527

Note: The overall average difference scores are displayed in the left column next to the technique identification. *P* values that are associated with each pairwise comparison are based on the Mann–Whitney *U* tests. Yellow ***, ** and * indicate comparisons in which the techniques were statistically different after applying the Bonferroni-corrected *P* value threshold of 6.54×10^{-7}, as calculated based on initial *P* values of .0001, .001 and .01, respectively. The number of targets in common for each pair of techniques are given for each comparison.

DS, average difference score; RB, Robetta; HB, HHPredB; I4, IntFOLD4-TS; I3, IntFOLD3-TS; I2, IntFOLD2-TS; P2, Phyre2; NB, NaïveBLAST; RA, RBO Aleph; RX, RaptorX; PB3D, PRIMO-BST-3D; PBCL, PRIMO-BST-CL; PH3D, PRIMO-HHS-3D; PHCL, PRIMO-HHS-CL; PR, PRIMO; PT, Princeton-TEMPLATE; SM, SWISS-MODEL; SX, SPARKS-X.

in complex with a ligand. Also, the difference score examines predicted rather than verified functional sites.

3.7 The ResiRole server

The ResiRole server, http://protein.som.geisinger.edu/ResiRole/, is established to provide routine updates to the analyses of structure models addressed in the CAMEO project. The accuracies of the structure prediction techniques according to their average difference scores are provided. Calculations are done for defined release intervals in CAMEO. For example, results are provided for weekly, monthly and yearly updates.

The results on the ResiRole server are further categorized according to target difficulty. The categories are all, easy, medium and hard, according to the lDDT score ranges described in CAMEO (Haas *et al.*, 2018). The ResiRole server also provides the average difference score for each structure model versus its reference structure, thereby enabling structure prediction techniques to be compared at the per target granularity.

A means for users to analyze their own structure model versus a reference structure through an interactive web submission page is also enabled. Here, the user uploads the coordinates of their structure model along with the coordinates of the reference structure. The overall average difference score for the functional site probabilities is calculated and sent to the user via email. The interactive page thereby enables analyses of structure models generated by additional structure prediction techniques and targets outside of the scope of CAMEO. A potential application is for the Critical Assessment of Structure Prediction (CASP) experiment (Kryshtafovych *et al.*, 2019).

4 Discussion

4.1 Overall assessments of structure prediction techniques

ResiRole may be used to assess the accuracies of structure prediction techniques for comparisons between different techniques and between different versions of the same technique. Consider for

Table 2. Results of round-robin, head-to-head comparisons using difference scores for the hard targets

Server	DS	RB	SM	M4T	RX	P2	I4	HB	RA	NB	I3	SX	PT	I2	PH3D	PHCL	PR	PBCL
RB	0.2229																	
SM	0.2501	*** 254																
M4T	0.2539	*** 51	6.E-04 143															
RX	0.2545	*** 262	*** 686	*** 119														
P2	0.2597	*** 214	*** 686	3.E-02 119	*** 539													
I4	0.2604	*** 37	*** 107	*** 15	*** 111	*** 72												
HB	0.2666	*** 176	*** 493	*** 84	*** 434	*** 350	4.E-01 72											
RA	0.2669	*** 113	8.E-03 280	8.E-03 70	*** 224	*** 257	2.E-01 42	*** 143										
NB	0.2775	*** 145	*** 536	*** 102	*** 445	*** 377	*** 79	*** 329	*** 138									
I3	0.278	*** 80	5.E-01 188	5.E-04 27	*** 198	*** 157	*** 100	*** 131	*** 82	*** 125								
SX	0.2795	*** 127	*** 606	9.E-03 107	*** 483	*** 505	*** 93	*** 283	5.E-01 203	*** 355	3.E-02 183							
PT	0.2816	*** 166	*** 640	*** 129	*** 496	9.E-05 586	*** 96	*** 313	*** 270	*** 361	* 174	*** 512						
I2	0.2827	*** 188	3.E-03 327	3.E-01 60	*** 322	*** 289	*** 55	** 193	*** 142	*** 184	4.E-02 119	4.E-03 218	2.E-02 252					
PH3D	0.3136	*** 38	*** 239	*** 23	*** 228	*** 120	*** 73	*** 163	*** 46	*** 197	*** 95	*** 202	*** 145	*** 56				
PHCL	0.3151	*** 40	*** 247	*** 22	*** 237	*** 125	*** 81	*** 170	*** 47	*** 203	*** 100	*** 210	*** 152	*** 58	*** 245			
PR	0.323	*** 39	*** 250	*** 23	*** 236	*** 128	*** 83	*** 165	*** 47	3.E-01 212	*** 97	*** 208	*** 157	*** 57	*** 233	*** 240		
PBCL	0.3233	*** 38	*** 247	*** 23	*** 233	*** 128	*** 82	*** 161	*** 47	3.E-01 210	*** 96	*** 205	*** 156	*** 57	*** 229	*** 236	5.E-01 256	
PB3D	0.3423	*** 39	*** 247	*** 23	*** 231	*** 127	*** 82	*** 162	*** 46	*** 210	*** 95	*** 204	*** 155	*** 56	*** 229	*** 236	*** 255	*** 253

example the significant improvement of the IntFOLD4-TS over IntFOLD3-TS, as accessed with the difference scores described in Table 1. We find that the results confirm the significant improvement previously reported based the lDDT metric (Mariani *et al.*, 2013; McGuffin *et al.*, 2018). The advancements in model quality between the IntFOLD4-TS over IntFOLD3-TS server versions are described as being due in part to the use of ModFOLD6 for the selection of the final model. ModFOLD6 evaluates the agreement between the contacts found in the model versus the contacts predicted using MetaPSICOV (Maghrabi and McGuffin,2017). MetaPSICOV utilizes covariation to identify residue contact and predict long-range hydrogen bonds (Jones *et al.*, 2015).

4.2 Relative performances of Phyre2 and NaiveBLAST

When the results presented in Figure 4 are further considered, we see instances in which the structure prediction techniques had either less than or more than their expected values relative to what was predicted by the line formulas. To identify potential outliers, we calculated standardized residuals for the structure prediction techniques using the standard method in R (Team, 2017).

See the plot of average difference scores versus the average lDDT values and notice the point for Phyre2, which is indicated by the arrow in Panel E of Figure 4. The standardized residual for this point is -2.611 Further, the standardized residual for the plot of difference score versus lDDT for Phyre2 for the hard targets was -2.20 (see Supplementary Fig. S2). Our interpretation of the relatively large residuals is that as Phyre2 explicitly includes measures of functional information for structure model generation in the form of local sequence conservation, template binding site information, cleft detection and consensus binding site conservation (Kelley *et al.*, 2015; Kelley and Sternberg, 2009). The lDDT score may underestimate the ability of Phyre2 to produce models that accurately reconstitute functional site predictions using the difference score as the benchmark metric. But the linear fit of average lDDT-BS value versus the difference score is relatively high; and Phyre2 did not have a residual indicative of a possible outlier. Since the lDDT-BS metric focuses only on the accuracy of ligand binding sites, the result may provide more evidence to the above claim.

We also find that the standardized residuals for the NaïveBLAST method are relatively pronounced for all the plots shown in Figure 4. Consider the regression fits for the difference score versus TM-score, GDT-TS, GDT-HA, GDC, lDDT and lDDT-BS in which the standardized residuals for NaïveBLAST were, respectively, -1.70, -1.46, -1.26, -1.73, -1.63 and -2.77. Also, the corresponding values for analyses regarding the hard targets are interpreted as outliers at -3.37, -3.00, -2.78, -3.21, -2.51 and -3.08. (See Supplementary Tables S4 and S5.) NaïveBLAST is thereby ranked inaccurately with the other assessment metrics when using the difference score as the benchmark. Its ability to produce structure models with detailed structural features of accurately predicted functional sites is underestimated by the other metrics. We interpret the result by considering that NaïveBLAST produces a baseline model using the nearest template structure. If the nearest template structure already has the functional site to be predicted, the precise local described as being due in part to the use of ModFOLD6 for the selection

formulas. To identify potential outliers, we calculated standardized resid-

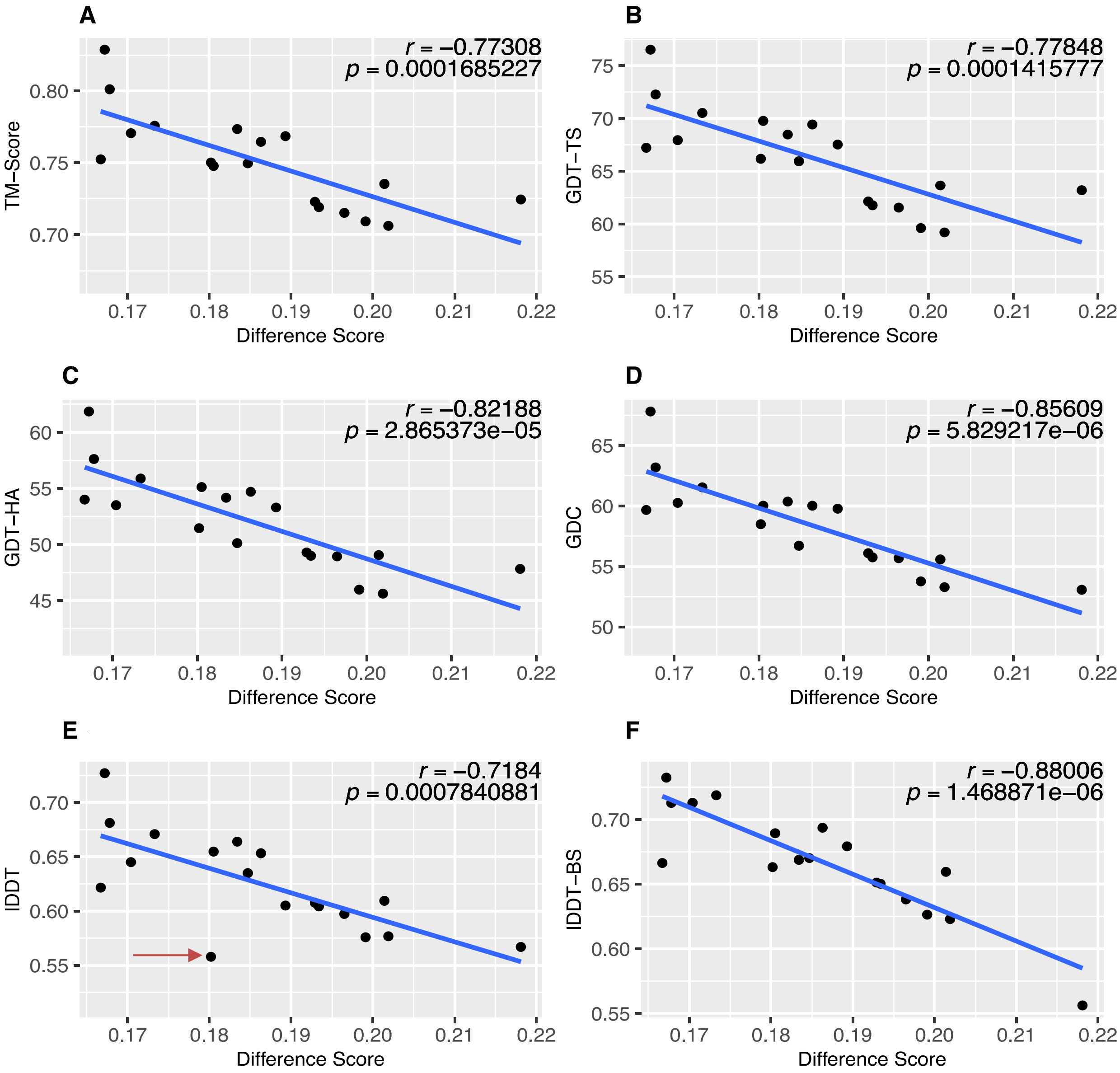

Fig. 4. Correlations between the difference scores and other metrics for assessing structure model quality. The regression lines for the scatter plots off the average difference scores for the 18 different structure prediction techniques versus their corresponding average values of the other quality assessment metrics are provided. The other metrics are TM-Score (**A**), GDT-TS (**B**), GDT-HA (**C**), GDC (**D**), lDDT (**E**) and lDDT-BS (**F**)

structural features required for that functional site may be represented in the corresponding structure model *a priori*, thereby accounting for the relatively low difference score measurements. We infer the other metrics did not capture the relevant structural features of the functional site predictions as well as the difference score.

4.3 Utilities of the ResiRole server

Although from a theoretical standpoint, it may be obvious that structure model quality is the cause of similar functional site predictions between the structure models and the reference structures, an advance that made the overall assessments possible here was the benchmarking studies by Buturovic et al. (*et al*2014). Using these benchmarking results, the same functional specificity threshold range could be applied for each of the functional site prediction models (SeqFEATURE models). See supporting analyses in Supplementary Material for a description of the selection of the specificity ranges used for the study. That enabled the calculation of the difference score as an overall, normalized metric that could be averaged across different types of functional site predictions.

The ResiRole framework does not necessitate that each of the predicted functional sites be present within the reference structures. seqThe analysis thereby provides an objective means to measure the ability of the structure prediction technique to reconstitute the structural features found at local sites within the reference structures.

There will be advantages and disadvantages for each selected structure quality assessment metric which a user may select based on an intended application. If the goal is to identify whether an overall predicted fold is accurate, global measurements such as TM-Score or GDC may be preferred. If the application entails evaluation of documented or potential local functional sites, then an estimate of structure model quality using the difference score may provide a more accurate method to estimate the likelihood that the structure model reconstitutes the local structural features required to observe similar functional site predictions.

As future development, it would be useful to extend the ResiRole method to include analyses of functional site predictions made with multiple functional site prediction methods which evaluate the features of the three-dimensional coordinates. We anticipate that, in addition to PROSITE mappings to SeqFEATURE models, other primary sequence motifs, such as those available from BioSeq-Analysis2.0 (Liu *et al.*, 2019), may be utilized to generate 3D functional site prediction models. For their subsequent use in structure MQA, we see a need to obtain benchmarking results for each of the 3D functional site prediction models that would include correspondences between the *Z*-scores of the predicted functional sites and the functional site specificity thresholds. Via such benchmarking results, the *Z*-scores may be confidently mapped to the corresponding cumulative probabilities. These cumulative probabilities would then be comparable across different types of functional site predictions. The average difference score obtained using the different types of functional site prediction models may provide additional means to assess the accuracies of the structure prediction techniques more accurately.

The ResiRole method currently does not serve the purpose of benchmarking how accurately experimentally verified functional sites are identified via the functional site predictions. That would require the collection of experimentally verified functional sites available for the CAMEO targets and curating these about their correspondences to the SeqFEATURE models. We therefore see that future developments are needed in that area. It would be worthwhile to estimate the capacity of the structure models to reconstitute experimentally verified functional sites that are found in the reference structures. These assessments would enable ResiRole to additionally serve as a tool for evaluating whether functional site predictions in the structure models are likely to correspond to actual functional sites within the target structures. But we find implementing that additional aspect is not a requirement for estimating the relative accuracies of the structure models based on their capacities to have functional site predictions like those of the reference structures.

5 Conclusion

Here, we describe the ResiRole method as means to assess the average quality of structure models produced by each structure prediction technique based on comparing the functional site predictions in the reference structures to predictions at the corresponding sites within structure models. The method provides an objective means to assess the relative accuracies of structure prediction techniques since it uses parameters not directly linked to the generation of the structure models. Consider for example, that physics-based or empirically derived energy functions that are used in structure prediction (Kelley *et al.*, 2015), are not used in calculating ResiRole's assessment metric. A disadvantage of the method is that the reference structures need to be available.

The top four structure prediction techniques based on difference scores for all targets considered were Robetta, IntFOLD4-TS, SWISS-MODEL and M4T. These techniques had a statistically higher performance relative to NaïveBLAST, which indicates that, on average, their structure models provide more information regarding the predicted functional sites of the reference structures than can be obtained from the nearest template structures. For targets in the hard category, as defined by a low lDDT score, several other techniques such as SWISS-MODEL, M4T, RaptorX and Phyre2 were found also to on average to outperform NaïveBLAST.

The difference score metric provides a direct measure of the accuracy of each structure prediction technique to generate structure models that reconstitute functional site predictions of the reference structures. As shown with outlier analysis with NaïveBLAST as the benchmark technique and the difference as the benchmark metric, we infer the other metrics do not capture the relevant structural features of the functional site predictions as well as the difference score.

The ResiRole server provides routine updates for the analyses of structure prediction techniques represented in CAMEO with the goal of providing a complementary means for structure MQA. An interactive web submission site enables the evaluation of user-provided models. We expect that the results will further inform the development of more accurate structure prediction techniques and aid with the selection of models for user applications.

Acknowledgements

The authors thank Mike Wong for compiling the results of the benchmarking studies that describe the correspondences between the Z-scores and the specificity thresholds for functional site predictions made with the FEATURE program. Karissa Barbarevech and Emily Holzman helped to organize the data. James Basting provided expertise to help setup the server.

Funding

This work was support in part by a grant from the National Institute of General Medical Sciences [grant number 5U01 GM093324-02].

Conflict of Interest: none declared.

References

Altschul,S.F. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–303.

Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Buenavista,M.T. *et al.* (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics*, **28**, 1851–1857.

Bule,P. *et al.* (2017) Assembly of *Ruminococcus flavefaciens* cellulosome revealed by structures of two cohesin–dockerin complexes. *Sci. Rep.*, **7**, 759.

Buturovic,L. *et al.* (2014) High precision prediction of functional sites in protein structures. *PLoS One*, **9**, e91240.

Cozzetto,D. *et al.* (2007) Assessment of predictions in the model quality assessment category. *Proteins Struct. Funct. Bioinf.*, **69**, 175–183.

DeLano,W.L. (2002) Pymol: An open-source molecular graphics tool. CCP4 Newsletter on protein crystallography, **40**, 82–92.

Fernandez-Fuentes,N. *et al.* (2007) M4T: a comparative protein structure modeling server. *Nucleic Acids Res.*, **35**, W363–W368.

Grabowski,M. *et al.* (2007) Structural genomics: keeping up with expanding knowledge of the protein universe. *Curr. Opin. Struct. Biol.*, **17**, 347–353.

Haas,J. *et al.* (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins Struct. Funct. Bioinf.*, **86**, 387–398.

Halperin,I. *et al.* (2008) The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics*, **9**, S2.

Hatherley,R. *et al.* (2016) PRIMO: an interactive homology modeling pipeline. *PLoS One*, **11**, e0166698.

Hulo,N. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.

Jones,D.T. *et al.* (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**, 999–1006.

Jones,E. *et al.* (2014) {SciPy}: open source scientific tools for {Python}.

Källberg,M. *et al.* (2012) Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.*, **7**, 1511–1522.

Keedy,D.A. *et al.* (2009) The other 90% of the protein: assessment beyond the Cαs for CASP8 template-based and high-accuracy models. *Proteins Struct. Funct. Bioinf.*, **77**, 29–49.

Kelley,L.A. *et al.* (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.*, **10**, 845–858.

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.

Khoury,G.A. *et al.* (2014) Princeton_TIGRESS: protein geometry refinement using simulations and support vector machines. *Proteins Struct. Funct. Bioinf.*, **82**, 794–814.

Kim,D.E. *et al.* (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.*, **32**, W526–W531.

Kryshtafovych,A. *et al.* (2019) Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins Struct. Funct. Bioinf.*, **87**, 1011–1020.

Liu,B. *et al.* (2019) BioSeq-Analysis2. 0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.*, **47**, e127–e127.

Liu,T. *et al.* (2018) Biological and functional relevance of CASP predictions. *Proteins Struct. Funct. Bioinf.*, **86**, 374–386.

Mabrouk,M. *et al.* (2015) RBO Aleph: leveraging novel information sources for protein structure prediction. *Nucleic Acids Res.*, **43**, W343–W348. gkv357.

Maghrabi,A.H. and McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. *Nucleic Acids Res.*, **45**, W416–W421.

Mariani,V. *et al.* (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**, 2722–2728.

McGuffin,L.J. *et al.* (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Res.*, **43**, W169–W173.

McGuffin,L.J. *et al.* (2018) Accurate template-based modeling in CASP12 using the IntFOLD4-TS, ModFOLD6, and ReFOLD methods. *Proteins Struct. Funct. Bioinf.*, **86**, 335–344.

Read,R.J. and Chavali,G. (2007) Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins Struct. Funct. Bioinf.*, **69**, 27–37.

Roche,D.B. and McGuffin,L.J. (2016) In silico identification and characterization of protein-ligand binding sites. In: *Computational Design of Ligand Binding Proteins*. Humana Press, New York, NY, pp. 1–21.

Šali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Schwede,T. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.

Söding,J. *et al.* (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.

Team,R.C. (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Wass,M.N. *et al.* (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.

Wu,S. *et al.* (2008) The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. *Genome Biol.*, **9**, R8.

Yang,Y. *et al.* (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, **27**, 2076–2082.

Zemla,A. *et al.* (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins Struct. Funct. Bioinf.*, **37**, 22–29.

Zhang,C. *et al.* (2017) COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.*, **45**, W291–W299.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.