Protein structure prediction on the Web: a case study using the Phyre server

Lawrence A Kelley & Michael J E Sternberg

Structural Bioinformatics Group, Division of Molecular Biosciences, Department of Life Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. Correspondence should be addressed to L.A.K (l.a.kelley@imperial.ac.uk).

Published online 26 February 2009; doi:10.1038/nprot.2009.2

Determining the structure and function of a novel protein is a cornerstone of many aspects of modern biology. Over the past decades, a number of computational tools for structure prediction have been developed. It is critical that the biological community is aware of such tools and is able to interpret their results in an informed way. This protocol provides a guide to interpreting the output of structure prediction servers in general and one such tool in particular, the protein homology/analogy recognition engine (Phyre). New profile–profile matching algorithms have improved structure prediction considerably in recent years. Although the performance of Phyre is typical of many structure prediction systems using such algorithms, all these systems can reliably detect up to twice as many remote homologies as standard sequence-profile searching. Phyre is widely used by the biological community, with >150 submissions per day, and provides a simple interface to results. Phyre takes 30 min to predict the structure of a 250-residue protein.

INTRODUCTION

At present, over six million unique protein sequences have been deposited in the public databases, and this number is growing rapidly (http://www.ncbi.nlm.nih.gov/RefSeq/). Meanwhile, despite the progress of high-throughput structural genomics initiatives, just over 50,000 protein structures have so far been experimentally determined. This enormous disparity between the number of sequences and structures has driven research toward computational methods for predicting protein structure from sequence. Computational methods grounded in simulation of the folding process using only the sequence itself as input (the so-called *ab initio* or *de novo* approaches) have been pursued for decades and are showing some progress[1]. However, in general, these methods are either computationally intractable or show poor performance on everything except the smallest proteins (<100 amino acids)[1].

The most successful general approach for predicting the structure of proteins involves the detection of homologs of known three-dimensional (3D) structure—the so-called template-based homology modeling or fold-recognition. These methods rely on the observation that the number of folds in nature appears to be limited and that many different remotely homologous protein sequences adopt remarkably similar structures[2]. Thus, given a protein sequence of interest, one may compare this sequence with the sequences of proteins with experimentally determined structures. If a homolog can be found, an alignment of the two sequences can be generated and used directly to build a 3D model of the sequence of interest. The practical applications of protein structure prediction are many and varied, including guiding the development of functional hypotheses about hypothetical proteins[3], improving phasing signals in crystallography[4], selecting sites for mutagenesis[5] and the rational design of drugs[6].

Every 2 years an international blind trial of protein structure prediction techniques is held (Critical Assessment of Structure Prediction—CASP)[1]. Over the years, we have observed enormous improvements at CASP in both the detection of ever more remote homologs and in the accuracy of the resulting homology models. With the advent of large sequence databases, and powerful programs to mine the data, such as PSI-Blast[7], Hidden Markov Models[8] and, recently, profile–profile matching algorithms[9], it is now commonplace to accurately detect and model protein sequences with less than 20% sequence identity to a known protein structure. A common feature of all such methods is their use of multiple sequence information. For example, PSI-Blast is a powerful algorithm for iteratively searching a protein sequence database. In each iteration, homologous sequences are collected and used to construct a statistical profile of the mutational propensities at each position in the sequence. This profile is then used in a subsequent round of searching, permitting the detection of further remote homologs. This process can be repeated 5–10 times, as the profile is iteratively modified[7]. Sequence profiles are powerful representations of the evolutionary history of a protein. As such, they form the backbone of many of the most successful structure prediction methods in use today. Newer profile–profile approaches significantly outperform PSI-Blast, producing more accurate alignments, and can detect up to twice as many remote homologs[9].

However, a solution to the protein-folding problem, the 'holy grail' of structural bioinformatics, remains out of reach. The techniques that have been developed to tackle structure prediction, although powerful, are not without their flaws. Although such tools may be used in a fully automated way, gaining the most from them requires human expertise in analyzing the results in the context of biological knowledge. For this reason, this protocol focuses on interpretation, not prescription. Rarely are there certain answers in structure prediction, and what we provide here are guidelines to judgement that can be applied to the output of any structure prediction system. However, by focusing on a step-by-step procedure for one system, in particular, the recently developed Phyre[10] server, we hope the principles described can be more clearly understood in a practical context.

Phyre remote homology modeling server

A detailed description of the methods used by the Phyre server may be found in Bennett-Lovsey *et al.*[10]. However a brief overview will be useful. The Phyre server uses a library of known protein structures taken from the Structural Classification of Proteins

(SCOP) database[11] and augmented with newer depositions in the Protein Data Bank (PDB)[12]. The sequence of each of these structures is scanned against a nonredundant sequence database and a profile constructed and deposited in the 'fold library'. The known and predicted secondary structure of these proteins is also stored in the fold library.

A user-submitted sequence, henceforth known as the 'query', is similarly scanned against the nonredundant sequence database, and a profile is constructed. Five iterations of PSI-Blast are used to gather both close and remote sequence homologs. The (often large number of) pairwise alignments generated by PSI-Blast are combined into a single alignment with the query sequence as the master. This is thus not a true multiple sequence alignment (which would often be computationally too demanding to calculate), yet it provides valuable information, which will be discussed in Steps 8–15 in PROCEDURE.

Following profile construction, the query secondary structure is predicted. Three independent secondary structure prediction programs are used in Phyre: Psi-Pred[13], SSPro[14] and JNet[15]. The output of each program is in the form of a three-state prediction: alpha helix (H), beta strand (E—for extended) and coil (C). Each of these three programs provides a confidence value at each position of the query for each of the three secondary structure states. These confidence values are averaged and a final, consensus prediction is calculated and displayed beneath the individual predictions. In addition, the program Disopred[16] is run to calculate a two-state prediction of which regions of the query are most likely to be structurally ordered (o) and which disordered (d).

This profile and secondary structure is then scanned against the fold library using a profile–profile alignment algorithm detailed in Bennett-Lovsey *et al.*[10]. This alignment process returns a score on which the alignments are ranked. These scores are fitted to an extreme value distribution to generate an *E*-value. The top ten highest scoring alignments are then used to construct full 3D models of the query. Where possible, missing or inserted regions caused by insertions and deletions in the alignment are repaired using a loop library and reconstruction procedure. Finally sidechains are placed on the model using a fast graph-based algorithm and sidechain rotamer library.

During the development of the Phyre protocol, a large benchmark set of protein sequences were processed by the system and the frequency with which different *E*-values were returned for both true- and false-positive matches was recorded. This was used to build a mapping between a reported *E*-value and the empirical frequency of errors. Thus, an estimated precision score of 95% indicates that, on our benchmark, 95% of sequences that received this score or better were true homologs according to the SCOP database. It is important to note that these *E*-values are not those returned by PSI-Blast, but are generated internally from the Phyre profile–profile alignment algorithm itself.

It is now commonplace, using protocols such as Phyre, to achieve high accuracy models at very low sequence identities (15–25%). The term 'high accuracy' has different meanings according to the goals of the user of course, but the core part of a structure can regularly be modeled with a root mean square deviation (r.m.s.d.) to the native structure of 2–4 Å even at such a low sequence identity. This shortcoming of sequence identity as a measure of predictive accuracy is why the 'estimated precision' described above has been developed as a more useful guide.

The Phyre system is typical of many of the freely available structure prediction systems on the Web, and as such, the concepts discussed in this protocol are easily transferable to other systems. In brief, to use the Phyre system, a user simply pastes their amino-acid sequence into a Web page together with their email address and clicks a button. Approximately 30 min later, the user will receive an email containing, among other things, a link to a Web page of results, including full downloadable 3D models of their protein and associated confidence estimates (**Fig. 1**). The present publicly available Phyre server showed performance typical of the majority of other structure prediction servers in CASP7 (ref. 1). Recent major developments to the core algorithms of Phyre have placed it among the best servers in the most recent CASP8 preliminary assessment (http://prodata.swmed.edu/CASP8/evaluation/DomainsAll.First.html), and these developments will be shortly rolled out in the public server.

The accuracy of alignment between a query protein sequence and a known template structure is key in defining the accuracy of the final 3D model. A method for automating the assessment of alignment accuracy has been implemented in Phyre similar to that of Tress *et al.*[17]. Every position along the alignment where a query residue is matched to a template residue is assigned a score from the profile–profile matching algorithm internal to Phyre. Contiguous high-scoring regions are indicative of accurate alignment and are color-coded as described in Step 31.

Domain parsing for long sequences

Long protein sequences often contain multiple domains. Most homology-based structure prediction systems use a library of individual structural domains and are poor at predicting domain–domain orientation. In addition, computing time increases rapidly with increasing length of the query sequence. For these reasons, it is advisable to first establish whether there is any clear domain structure in a long sequence using tools specifically designed for this purpose. We suggest the Conserved Domain Database[18] search service at the NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) or PFAM[19]. Domains clearly identified by these programs should be extracted from the query sequence and processed individually through this protocol. For optimal performance, sequences less than 1,000 residues are preferred for use in Phyre. When presented with long sequences with no clear domain boundaries, one should submit 1,000 residue segments and refer to Step 29 in PROCEDURE.

Functional residue prediction

A common requirement for many users of protein structure prediction tools is to predict the residues most likely to be involved in the function of the protein. To this end, we have implemented a variety of tools to analyze and combine information from the alignment and the 3D model to produce a consensus prediction of functional residues in the query. A complementary approach involving predicting GO functional terms in addition to functional residues is available from the ConFunc server[20] (http://www.sbg.bio.ic.ac.uk/confunc/).

Functionally important residues are expected to be under stronger selective pressure than those involved in maintaining more generic protein structural features. Amino-acid conservation scores are calculated for the query and its sequence homologs at each position in the pseudo-multiple sequence alignment described

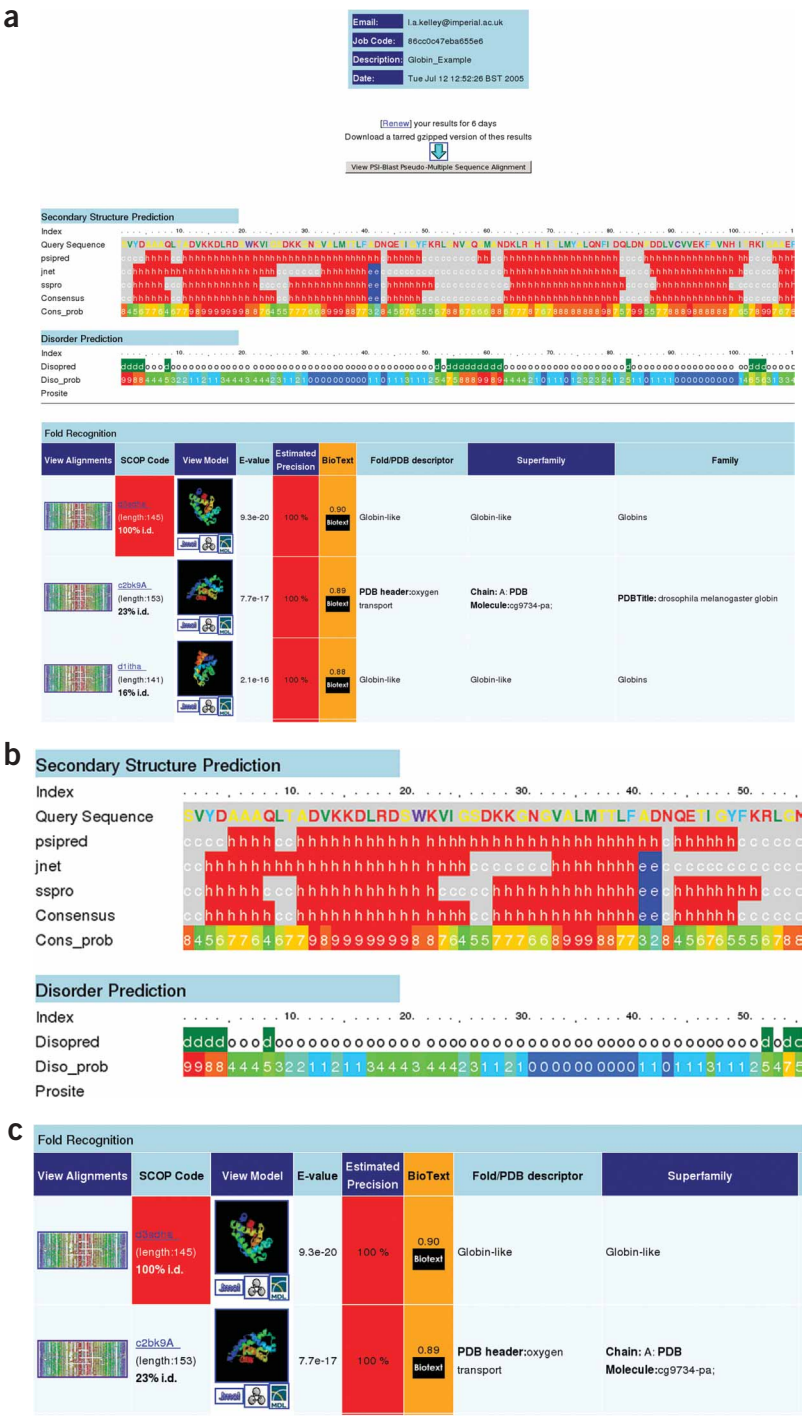

Figure 1 | Example of a typical Phyre results page. (**a**) The page is roughly divided into three sections entitled Secondary Structure Prediction, Disorder Prediction and Fold Recognition from top to bottom, respectively. (**b**) Detailed view of the secondary structure and disorder predictions. The results from three independent secondary structure prediction methods are shown together with a consensus prediction. Red horizontal bars indicate predicted alpha helices, blue bars represent beta strands and gray bars indicate coil. The color-coded numbers in the Cons_prob row indicate the confidence of the prediction at each position from 0 (low confidence) to 9 (high confidence). Similarly, in the Disorder prediction, positions are flagged as 'd' for disordered and 'o' for ordered, with a likelihood of disorder between 0 (low) and 9 (high) reported in the Diso_prob row. (**c**) Cropped view of the top two hits in the primary table of the fold recognition results, including images of the protein models produced and descriptors of the fold and superfamily of the template used. Table cells with a red background highlight particularly confident (>95% probability) predictions.

above. The calculation is performed at different sequence identity thresholds to compensate for possible redundancy or paucity of homologs at each position in the query. Large numbers of highly similar sequences in the alignment may skew conservation scores; however, when sequences are sparse, even highly redundant sequences can provide some useful information. Thus the conservation of each position in the query is calculated by removing all sequences sharing more than threshold sequence identity with any other sequence in the alignment. This threshold takes the values 30%, 40%, 50% and 60%. In addition, Phyre includes an implementation of the potentially more sensitive Evolutionary Trace approach of Yao *et al.*[21].

In some cases, the template structure detected by Phyre may contain binding site information. We use the eF Site database[22] as a source of such information and highlight identities and conservative mutations between query and template residues that form known binding sites. In addition, it has been known for some time that the active site of a protein is frequently found within large clefts or pockets in the protein[23]. To locate such clefts, we use the Pocket program[24]. We apply the cleft searching to a backbone-only model of the query produced by Phyre and use a larger probe radius to compensate for the lack of sidechains. A backbone-only model is used to avoid spurious pocket detection caused by misplaced sidechain rotamers.

Integrating the above sources of data into a final functional residue prediction is nontrivial. The Phyre server creates a weighted average of the information from sequence conservation, evolutionary trace, known template functional sites and pocket detection for each position in the query sequence normalized between 0 and 9, with 9 being the highest confidence prediction of a functionally relevant residue. When one or more residues achieves a consensus score >5, these confidence scores are mapped to a color scale and are used to color a space-filling model of the query that can be interactively viewed using the JMol browser application.

Modeling point mutations

A frequent request from users is to model the effect of point mutations on the structure of their query protein. Unfortunately, an inherent limitation to purely template-based prediction algorithms, such as Phyre, is that such subtle changes to the primary sequence will

not, in general, result in a different 3D model. In certain cases, e.g., if the point mutation lies in a loop that is processed by our loop modeling software, a change will be observed. However, the accuracy of loop modeling is, in general, insufficient to permit any firm conclusions to be drawn from such differences.

Indels and missing coordinates

Although the Phyre system uses powerful loop modeling techniques to model insertions and repair deletions in the alignment to the template, in some cases, this system will fail and this will be evident by missing coordinates in these regions of the 3D model. It is not generally possible to model insertions of more than 15 residues in length. Similarly, if an extensive deletion in the alignment occurs across regions of the template that cannot be bridged by the remaining residues, a gap will remain. Such irreparable deletions are often indicative of a poor alignment or poor choice of template structure. In addition, the template structure itself may not contain coordinates for certain residues due to crystallographic resolution problems that may be indicative of intrinsic disorder in these regions of the template protein.

Low confidence matches

The problems of fold recognition and remote homology detection remain to be solved. Despite advances in the field, users will invariably come across cases where no confident matches to their query can be detected. This may be for two reasons: (i) the query adopts a known fold but is so remote from any solved structure that this homology or analogy cannot be detected, or (ii) the query constitutes a novel fold. Even when Phyre is incapable of providing a confident assignment, where there is a commonly occurring fold or superfamily returned in the top ten predictions, this can provide important clues to guide further research.

No single method of structure prediction is completely trustworthy. Every system has its strengths and weaknesses. This is reflected in the repeatedly demonstrated superior performance of consensus or meta-servers in the international blind trials of structure prediction, CASP[1]. Combining predictions from many sources is the most reliable way of avoiding false-positive fold assignments and of determining the most accurate alignment and model. There are many freely available Web servers for structure prediction on the Internet, and the space limitations of this protocol prohibit a similarly in-depth discussion of their use and interpretation. Nevertheless, we provide a short list of some of the most successful structure prediction systems from recent CASP competitions in **Table 1** and encourage the reader to familiarize themselves with each of them.

In the most difficult cases, the confidence measures for a prediction are extremely low and no consensus can be found across a range of structure prediction tools. If the query protein is sufficiently small (i.e., less than 120 amino acids), then the one remaining avenue is to use one of the few publicly available systems for *ab initio* structure prediction. The most successful of these approaches are based on a principle of fragment assembly, originally pioneered by David Jones[25] and refined and improved by the lab of David Baker[26]. Such methods fragment the query sequence into fully overlapping short stretches of amino acids (usually nine residues in length). Candidate structures for these small fragments are then generated using conventional template-based techniques. These structural fragments are then stochastically sampled, within the context of an empirically derived statistical force field, and assembled to construct a low-energy protein conformation. The I-TASSER server[27] uses a similar approach but includes larger, fixed structural fragments where available. In addition, a fast *ab initio* folding technique not based on fragments, but instead based on a simplified protein representation and Langevin dynamics known as Poing, has been recently developed in our lab and will soon be made available on the Web.

TABLE 1 | Popular Web servers for remote homology/fold recognition.

Server name	Web address	Consensus/ single	Model building/ confidence measure	FR/ *ab initio*
Phyre	http://www.imperial.ac.uk/phyre/	Single	Model + confidence	FR
I-TASSER	http://zhang.bioinformatics.ku.edu/I-TASSER/	Single	Model + confidence	FR + *ab initio*
SAM-T06	http://www.soe.ucsc.edu/compbio/SAM_T06/T06-query.html	Single	Model + confidence	FR
HHpred	http://toolkit.tuebingen.mpg.de/hhpred	Single	Confidence	FR
GenThreader	http://bioinf.cs.ucl.ac.uk/psipred/psiform.html	Single	*P*-value	FR
PCONS	http://pcons.net/	Consensus	Model + P_{cons} score	FR
Bioinfo	http://meta.bioinfo.pl	Consensus	Model + *E*-value	FR
FFAS	http://ffas.ljcrf.edu	Single	FFAS score	FR
Robetta	http://robetta.bakerlab.org/	Single	Model + confidence	FR + *ab initio*
SP[4]	http://sparks.informatics.iupui.edu/SP4/	Single	Model + *Z*-score	FR

'Consensus' indicates that the server collates results from multiple independent servers to form a final prediction, whereas 'single' indicates that a server uses only its own local methods. The Model building/ confidence measure column indicates whether a server provides as output 3D coordinates of a potential model ('*Model*') and a score indicating the confidence in the model (*Z*-score, *P*-value, *E*-value and so on). The 'FR/*ab initio*' column indicates whether the server can produce results based only on remote homology/fold recognition ('*FR*') or can additionally build models in the absence of a template ('*ab initio*').

Ab initio techniques for structure prediction are highly fallible and often do not return any accurate measure of confidence. Nevertheless, such techniques can provide the researcher with valuable clues to structural features that otherwise would be completely unavailable in the absence of an experimentally derived structure.

Finally, although this protocol has been designed as a detailed tutorial on interpreting the results of one specific structure prediction tool, many of the principles discussed are applicable to other similar tools and will hopefully help the user harness cutting-edge bioinformatics for their research.

MATERIALS

EQUIPMENT

• A computer with access to the internet and a web browser

EQUIPMENT SETUP

Data Your input protein sequence should be written in the standard one-letter code. Thus, the allowed characters are ACDEFGHIKLMNPQRSTVWY and also X (unknown). Spaces and line breaks will be ignored and will not affect the predictions.

PROCEDURE

Sequence submission

1| Go to the Phyre home page (http://www.sbg.bio.ic.ac.uk/phyre/).

2| Supply your e-mail address.

3| Paste your protein sequence of interest into the provided form. The sequence can be in FASTA format or simply a string containing the amino sequence.

4| Supply a name for your job.

5| Click on the search button at the bottom of the Web form, and the server will take you to a page confirming your submission.

6| You can follow the progress of your job in real time or wait for an e-mail confirming that your job has been completed. The e-mail you receive contains job information, the 3D coordinates of the top scoring predicted model in PDB format and a link to a Web page of detailed results regarding your job. The PDB-formatted coordinates in the email can be extracted with any text editor and loaded into any of the standard molecular viewers, such as Pymol, Rasmol and others. A typical results Web page is shown in **Figure 1**.

▲ CRITICAL STEP To maintain sufficient free disk space on our servers, the results for your job will be deleted after 7 days. There are two links near the top of the results page. One link permits you to renew your results. Thus, if the results are nearing their expiry time, the user may click this link to keep their results residing on the server for a further 7 days. The second link on the page permits the user to download a zipped file containing all the relevant Web pages. However, certain functionality, such as the in-browser molecular viewer Jmol, will not function with downloaded offline results.

Sequence homolog detection

7| Click the button entitled 'View PSI-Blast pseudo-multiple sequence alignment', found near the top of the results page (**Fig. 1**). The results of scanning the query sequence against an up-to-date nonredundant protein sequence library are displayed in a new window (**Fig. 2**). Steps 8–15 refer to **Figure 2**.

8| To determine how many homologous sequences were found, view the number of rows in the results page. Up to 1,000 homologous sequences may be presented in this alignment. Each row of the table contains the region of the homolog matched to the query, the *E*-value reported by PSI-Blast, the percentage sequence identity to the query and a sequence identifier for the homolog.

? TROUBLESHOOTING

9| To ascertain whether highly informative alignments most likely to generate an accurate secondary structure prediction were obtained, view the number of homologs with low *E*-values (<0.001).

▲ CRITICAL STEP A large number of high-confidence (*E*-value <0.001) homologs with extensive sequence diversity is indicative of a highly informative alignment, which is most likely to generate an accurate secondary structure prediction and powerful sequence

Figure 2 | Example of pseudomultiple sequence alignment from PSI-Blast. A cropped sample of a typical sequence alignment generated by PSI-Blast for a two-domain protein. The leftmost rectangle highlights the PSI-Blast *E*-value for each homologous sequence match. The string of Xs in the central rectangle indicate a region of low complexity sequence. On either side of the alignment are densely packed columns of aligned residues. In the middle is a region largely populated by dashes (-) indicated by a blue curly brace. This is a clear example of a potential domain boundary.

profile. Conversely, a very small number of homologs or a large number of highly similar homologs (>50% sequence identity) are both indicators of a lack of useful evolutionary information, which can lead to potentially error-prone secondary structure prediction, a weak sequence profile and consequently poor overall structure prediction accuracy.

Assessing alignment coverage

10| To determine the pattern and density of aligned sequences across the length of the query, view the regions of dense alignment (i.e., columns containing many homologs). Poorly populated columns may correspond to domain linkers or independent domains with few homologs in the sequence database.

11| To determine where potential domain boundaries are found, view regions where the alignment density changes rapidly. ▲ CRITICAL STEP If potential domain boundaries are defined and significant regions (>20 residues) have not been modeled (consult Step 29), it is possible to chop the sequence and resubmit separate regions to the Phyre server; repeating Steps 1–11. ? TROUBLESHOOTING

Alignment interpretation

12| View the colors of the aligned amino acids to visually assess strongly conserved motifs.

13| View the use of lower-case characters to identify regions where the homolog contains an inserted sequence relative to the query.

14| If required, click on the link near the top of the window entitled 'Click here for Fasta Format Flat File' to download the alignment in FASTA format.

15| To determine regions of low complexity, view the regions of the homologous sequences for 'X' characters. ▲ CRITICAL STEP Low-complexity information can be used in conjunction with the explicit disorder prediction described below in Step 18 in making a general assessment of what regions of the query may be accurately modeled.

Secondary structure and disorder prediction

16| Scroll further down the primary results page to the secondary structure prediction section (**Fig. 1a**).

17| View the consensus prediction score to obtain a confidence value for the secondary structures predicted (0 = low confidence, 9 = high confidence). The consensus prediction is used in all subsequent processing by the system.

18| View the two-state disorder prediction score to ascertain whether regions of the query are structurally ordered (o) or disordered (d). Confidence values are displayed from 0 (low confidence) to 9 (high confidence). Such disordered regions have often been found to be involved in protein function and should be taken into account when analyzing predicted functional sites (Steps 32–37).

19| View the query to determine if any ProSite[28] motifs are detected. These are highlighted beneath the sequence with gold dots.

Structural homology detection and fold recognition

20| View the table in the remainder of the results page (**Fig. 1b**) to determine the top ten highest scoring matches of the query to known template structures in the Phyre fold library and their respective models.

21| View the SCOP code column to determine the percentage sequence identity between the query and template. This is calculated relative to the shortest sequence. Matches with high percentage sequence identity (>40%) are highlighted in red. This column also indicates the unique identifier for the template structure matched by Phyre. The identifier is of the form [d/c][PDB code][chain identifier][domain number]. The initial 'd' or 'c' character indicates that the structure is a SCOP domain or a whole chain taken from the PDB respectively. The PDB code and chain identifier are self-explanatory. The domain number is an index (usually 1–9) supplied by SCOP to identify a particular domain in a multidomain, yet single chain, of a protein. ▲ CRITICAL STEP Usually a high sequence identity will be indicative of a high accuracy model. However, if the template sequence is particularly short relative to the query, percentage identity can be a poor guide to accuracy. Even more importantly, a low sequence identity (~20%) is not necessarily indicative of a poor model.

22| To view the 3D model of the query protein, click on the 'JMol' icon. Launching JMol within the browser permits a quick 3D view of the protein with full rotational and zoom facilities to assess the extent of gaps in the model, overall topology and the presence or absence of protein-like features. For a more detailed analysis, the user is encouraged to download the coordinates and use an in-depth standalone application. To download coordinates in PDB format, click on the image of the model itself.

23| View the 'Estimated Precision' column to determine the confidence that the query sequence is homologous to the template in question. The confidence values are color-coded from red to blue, indicating high and low confidence, respectively.

▲ CRITICAL STEP It is important to be aware that this number reflects the likelihood of homology and not the accuracy of the model. If presented with several high-confidence predictions, it is wise to focus on those involving matches with higher sequence identities and/or functional similarity to the query when known. The prediction of model accuracy (i.e., predicted r.m.s.d. to the true structure) is extremely difficult and is an actively pursued research goal of many groups. (See Model Quality Assessment section of the most recent CASP 7 competition[1].)
? TROUBLESHOOTING

24| View the fold, superfamily and family annotation columns to obtain information on the possible functions of the query. The presence of four or five templates with similar folds or functions lends more weight to a prediction than a singleton.

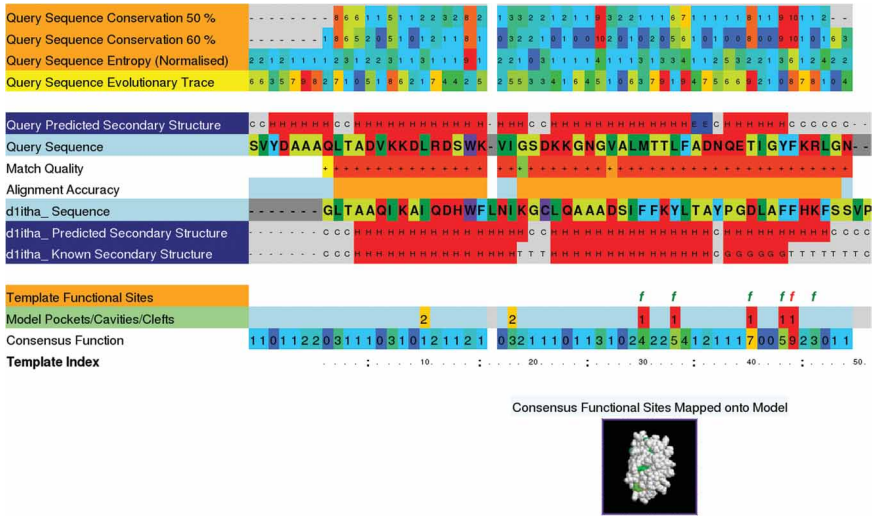

Figure 3 | Example of a typical Phyre alignment view. For each of the ten modeling results shown on the main Phyre results page, there is an accompanying alignment view. This includes alignment accuracy predictions, conservation analysis, cleft detection and functional site prediction as detailed in the protocol text. At the bottom of the figure is a clickable image of a space-filling model of the query protein with predicted functionally important residues colored according to the confidence of the prediction.

25| To obtain information on the alignment, patterns of conservation and predicted functional sites of the query, click on the link in the first column of the main fold-recognition results table. **Figure 3** shows a screenshot of a typical alignment view in Phyre.
▲ CRITICAL STEP The quality of the alignment of the query with the template is the most important feature in determining whether the query and template are true homologs, and in determining the final accuracy of the 3D model. Steps 26–31 outline key features to determine alignment accuracy, and an example is illustrated in **Figure 3**. Determining alignment accuracy with computational methods is still an active research focus of many groups. As such, it is here that the user's expertise and knowledge of their protein of interest comes most directly to bear. The user's knowledge of potentially important residues based on site-directed mutagenesis or other wet-lab and computational studies can be used to discard or reinforce matches made by Phyre, or to help discriminate between a set of similarly scoring models.

26| To determine if the query protein and the homolog show considerable agreement between the secondary structures at aligned positions, visually scan across the alignment looking for mismatching blocks of color between the query and template secondary structure rows.

27| Determine if insertions or deletions are present within the secondary structure elements by looking for dash characters (-) interrupting contiguous blocks of red (alpha helix) or blue (beta strand).
▲ CRITICAL STEP Insertions or deletions within secondary structure elements are usually an indicator of poor alignment. Mismatching elements, such as helices aligned to strands or the complete deletion of elements, is particularly concerning and may be indicative of an incorrect fold. Insertions and deletions are largely to be expected in loop regions.

28| To determine whether mismatches in the aligned secondary structure elements are misleading, view the confidence values for the secondary structure prediction as described in Step 17.

29| View the regions at the beginning and end of the alignments; where distinct boundaries are present, consider dividing the sequence of the original query at these boundaries and repeat Steps 1–28 for the individual segments.

30| View the colors of the cells in the 'Match Quality' row of the alignment to determine the high (red) and low (blue) scoring matches of individual residues.

31| View the highlighted bars in the 'Alignment accuracy' row to determine the predicted accuracy of the alignment at each position. Contiguous high-scoring regions are indicative of accurate alignment and are highlighted by an orange bar. Conversely, low scoring or 'patchy' regions of mixed high and low scoring matches are most likely to be poorly aligned and are highlighted with a blue bar.

Conservation

32| To identify potentially functionally important amino-acid residues, view the conservation score (0 low, 9 high) for the query and its sequence homologs in the 'Query Sequence Conservation *X*%' rows (**Fig. 3**).

33| To more finely discriminate between residues conserved for functional as opposed to structural reasons, view the scores in the 'Query Sequence Evolutionary Trace' row. The scores for each residue are based on the correlations of variations in the sequence homologs with their phylogenetic tree. This is calculated by applying the Evolutionary Trace algorithm of Yao *et al.*[21] (0 = low probability, 9 = high probability of functional importance).

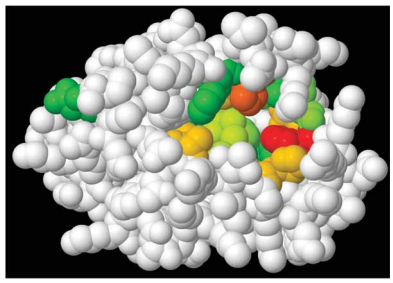

Figure 4 | Example of predicted functional sites colored by prediction confidence. In this example (a model of a globin sequence), one can see a cluster of orange and red residues residing in a deep cleft in the protein. This cleft accommodates the heme prosthetic group in the template structure. The residue coloring indicates that those residues of the query aligned to those in the cleft of the template are highly conserved, provide a strong evolutionary trace signal and match favorably with the known functional sites in the template.

Template-binding site information

34| To determine whether the query and template are most likely to share a common functional site at any residue, view the alignment for amino-acid residues highlighted with an '*f*'. If a position in the alignment matches identical residues, the '*f* ' is red. If the residue type of the match is not identical between query and template but the score for the match is positive, the '*f*' is green. Otherwise the '*f*' is gray.

Cleft detection

35| To determine the number and size of the clefts or pockets that the query contains, view the row labeled as 'Model pockets/cavities/clefts'. Those residues found within the five largest pockets are labeled according to the index of their pocket. For example, a residue labeled '1' belongs to the first and largest pocket.

Consensus functional site prediction

36| Click on the protein structure image below the alignment (if present) to obtain a space-filling model of the query showing the confidence prediction of a functionally relevant residue.

37| Examine the space-filling model for tightly clustered, red-colored residues these are a strong indicator of a potential functional site. To determine potential functional sites, see **Figure 4**.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 2**.

TABLE 2 | Troubleshooting table.

Step	Problem	Possible reason	Solution
8	Few sequence homologs are detected	Query sequence is an orphan sequence or very remote from anything in the genomes sequenced at present	Continue with protocol until Step 23. If no confident matches are found, consider other tools in **Table 1**
11	No domain boundaries are found	Query may be a large single domain sequence or contains domains not yet seen in isolation in genomes sequenced at present	Continue with protocol as normal until Step 23. If no confident hits are detected, consider other tools in **Table 1**. If confident hits are detected, continue to Step 29 and reassess domain coverage
23	No confident hits are found	Query is a new fold or too remote from any known structures to be detected by Phyre	Consider tools in **Table 1**. Also, find remote sequence homologs from Step 7 and submit these to Phyre and other systems

ANTICIPATED RESULTS

The accuracy of protein structure prediction depends critically on sequence similarity between the query and template. If a template is detected with >30% sequence identity to the query, then usually most or all of the alignment will be accurate and the resulting relative positions of structural elements in the model will be reliable. Below this level of sequence identity, confident matches are routinely made by Phyre. Given a high confidence match (>90% confidence), the overall fold of the model will be almost certainly correct and the central core of the model will tend to be accurate, even at sequence identities <20%. However, in such cases, greater deviations from the true structure will be observed in more peripheral regions of the protein and in regions neighboring sequence insertions or deletions in the alignment.

In cases where no confident model can be built by Phyre or other tools, it is important to be aware of the rate at which the structural database is growing. On average, 50 newly solved structures are added to the Phyre fold library every week, any one of which may be a detectable structural homolog of a previously processed query sequence. Thus, it is important to periodically resubmit a sequence to the Phyre server (and others) on a regular basis.

ACKNOWLEDGMENTS L.A.K. is supported by the BBSRC grant number LDAD P06300.

COMPETING INTERESTS STATEMENT The authors declare competing financial interests (see the HTML version of this article for details).

Published online at http://www.natureprotocols.com/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. CASP 7 special issue. *Proteins* **69** (Suppl. 8), 1–207 (2007).
2. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
3. Watson, J.D., Laskowski, R.A. & Thornton, J.M. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* **15**, 275–284 (2005).
4. Qian, B. *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450**, 259–264 (2007).
5. Rava, P. & Hussain, M.M. Acquisition of triacylglycerol transfer activity by microsomal triglyceride transfer protein during evolution. *Biochemistry* **46**, 12263–12274 (2007).
6. Park, H. *et al.* Discovery of novel alpha-glucosidase inhibitors based on the virtual screening with the homology-modeled protein structure. *Bioorg. Med. Chem.* **16**, 284–292 (2008).
7. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
8. Karplus, K., Barrett, C. & Hughey, R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856 (1998).
9. Ohlson, T., Wallner, B. & Elofsson, A. Profile–profile methods provide improved fold-recognition: a study of different profile–profile alignment methods. *Proteins* **57**, 188–197 (2004).
10. Bennett-Lovsey, R.M., Herbert, A.D., Sternberg, M.J.E. & Kelley, L.A. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins* **70**, 611–625 (2008).
11. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
12. Berman, H.M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
13. McGuffin, L.J., Bryson, K. & Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405 (2000).
14. Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228–235 (2002).
15. Cole, C., Barber, J.D. & Barton, G.J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36** (Web server issue): W197–W201 (2008).
16. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. & Jones, D.T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139 (2004).
17. Tress, M.L., Jones, D.T. & Valenica, A. Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. Biol.* **330**, 705–718 (2003).
18. Marchler-Bauer, A. *et al.* CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* **35** (Database issue): D237–D240 (2007).
19. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36** (Database issue): D281–D288 (2008).
20. Wass, M.N. & Sternberg, M.J.E. ConFunc—functional annotation in the twilight zone. *Bioinformatics* **24**, 798–806 (2008).
21. Yao, H. *et al.* An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261 (2003).
22. Kinoshita, K. & Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **12**, 1589–1595 (2003).
23. Laskowski, R.A. *et al.* Protein clefts in molecular recognition and function. *Prot. Sci.* **5**, 2438–2452 (1996).
24. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P.V. & Subramaniam, S. Analytical shape computation of macromolecules I and II. *Proteins* **33**, 1–17 and 18–29 (1998).
25. Jones, D.T. Predicting novel protein folds by using FRAGFOLD. *Proteins* **45** (Suppl. 5): 127–132 (2001).
26. Kim, D.E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32** (Web server issue): W526–W531 (2004).
27. Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69** (Suppl. 8): 108–117 (2007).
28. Hulo, N. *et al.* The 20 years of PROSITE. *Nucleic Acids Res.* **36** (Database issue): D245–D249 (2008).