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

Lawrence A Kelley* and Michael JE Sternberg

*corresponding author

Abstract

Determining the structure and function of a novel protein sequence is a cornerstone of many aspects of modern biology. Over the last three decades a number of state-of-the-art computational tools for structure prediction have been developed. It is critical that the broader biological community are aware of such tools and, more importantly, are capable of using them and interpreting their results in an informed way. This protocol provides a guide to interpreting the output of structure prediction servers in general and details one such tool in particular, the Phyre server. Phyre is widely used by the biological community with over 150 submissions per day and provides a simple interface to what can often seem an overwhelming wealth of data.

Introduction

Currently over 7 million protein *sequences* have been deposited in the public databases and this number is growing rapidly. 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 sizes of the sequence and structure databases has driven research towards computational methods of predicting protein structure from sequence. Computational methods grounded in simulation of the folding process using only the sequence itself as input (so-called *ab initio* or *de novo* approaches) have been pursued for decades and are showing some progress. However, in general, these methods are either computationally intractable or demonstrate poor performance on everything but the smallest proteins (<100 amino acids).

The most successful general approach for predicting the structure of proteins involves the detection of homologues of known 3D structure – so-called template-based homology modelling 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. Thus, given a protein sequence of interest, one may compare this sequence to the sequences of proteins with experimentally determined structures. If a homologue can be found, an alignment of the two sequences can be generated and used directly to build a three-dimensional model of the sequence of interest.

Every two 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 homologues and in the accuracy of the resulting homology models. With the advent of large sequence databases, and powerful programs to mine that data such as PSI-Blast[2], Hidden Markov Models[3] and recently, profile-profile matching algorithms[4] 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 homologues. This process can be repeated 5 to 10 times, as the profile is iteratively modified. 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.

However, a solution to the protein-folding problem, the 'holy grail' of structural bioinformatics, remains out of reach. Thus the techniques that have been developed to tackle structure prediction, though 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 analysing 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 focussing on a step-by-step procedure for one system in particular, we hope the principles described can be more clearly understood in a practical context.

In brief, to use the Phyre system a user simply pastes their amino acid sequence into a webpage, together with their email address and clicks a button. Approximately 30 minutes later the user will receive an email containing, amongst other things, a link to a web page of results, including full downloadable three-dimensional models of their protein and associated confidence estimates (Figure 1).

A detailed description of the methods used by the Phyre server may be found in Bennett-Lovsey et al.[5]. 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[6] and augmented with newer depositions in the Protein Data Bank (PDB)[7]. The sequence of each of these structures is scanned against a non-redundant 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 non-redundant *sequence* database, a profile constructed and its secondary structure predicted. This profile and secondary structure is then scanned against the fold library using a profile-profile alignment algorithm detailed in Bennett-Lovsey et al.[5] This alignment process returns a score on which the alignments are ranked. The top 10 highest scoring alignments are then used to construct full three-dimensional 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.

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.

PROCEDURE

Domain parsing for long sequences

1 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[8] search service at the NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) or PFAM[9] (http://pfam.sanger.ac.uk/). Domains clearly identified by these programs should be extracted from the query sequence and processed individually through the remainder of this protocol. For optimal performance, sequences less than 1000 residues are preferred for use in Phyre. When presented with long sequences with no clear domain boundaries one should submit 1000 residue segments and consult step 4 below.

Sequence submission

2 Sequence submission consists of visiting the Phyre home page (http://www.sbg.bio.ic.ac.uk/phyre/) supplying your e-mail address, pasting your sequence into the provided form and supplying an optional description of your job. The sequence can be in FASTA format or simply a string containing the amino sequence. You are encouraged to supply a meaningful name for your job. Upon clicking on the 'Search' button near the bottom of the web form, the server will take you to a page confirming your submission (or returning a message detailing any problem with the submission) together with a link to permit you to follow the progress of your job in real time. You are free to monitor progress or await an e-mail confirming that your job has been completed.

The e-mail you receive contains job information, the 3-dimensional 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. The rest of this protocol concerns the interpretation of the detailed results presented when following the link supplied in the e-mail. A typical results page is shown in Figure 1.

CAUTION: 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 off-line results.

Sequence homologue detection

3 Near the top of the results page is a button entitled “View PSI-Blast pseudo-multiple sequence alignment”. Clicking this button opens a new window containing the results of scanning the query sequence against an up-to-date non-redundant protein sequence library. Five iterations of PSI-Blast are used to gather both close and remote sequence homologues. 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 below.

(A) Assessing the number and diversity of homologues

Up to 1000 homologous sequences may be presented in this alignment. Each row of the table contains the region of the homologue matched to the query, the E-value reported by PSI-Blast, the percentage sequence identity to the query and a clickable link to the NCBI for complete information on the homologue.

A large number of high confidence (low E-value) homologues with extensive sequence diversity is indicative of a highly informative alignment which is likely to generate an accurate secondary structure prediction and powerful sequence profile. Conversely, a very small number of homologues or a large number of highly similar homologues (>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.

(B) Assessing alignment coverage

Another valuable source of information about the query sequence is the pattern and density of aligned sequences across the length of the query. Regions of dense alignment (i.e. columns containing many homologues) often correspond to domains, whereas poorly populated columns may correspond to domain linkers or independent domains with few homologues in the sequence database. Thus regions where alignment density changes rapidly may indicate potential domain boundaries where, depending on the focus of the researcher, one may wish to chop the sequence and resubmit separate regions to the Phyre server.

(C) Alignment interpretation

Amino acids in the alignment are coloured according to biophysical properties to aid in the visual assessment of strongly conserved motifs. Lower case characters are used to indicate regions where the homologue contains an inserted (not shown) sequence relative to the query. A link is also present to download the alignment in FASTA format for use in any common external program for viewing, printing or manipulating the alignment. Finally, some regions of the homologues may contain ‘X’ characters indicating that these regions have been identified as low complexity (possibly disordered) regions. This low-complexity information can be used in conjunction with the explicit disorder prediction described below in step 5 in making a general assessment of what regions of the query may be accurately modelled.

Secondary structure prediction

4 Further down the results page is the secondary structure section. Three independent secondary structure prediction programs are used in Phyre: Psi-Pred[10], SSPro[11] and JNet[12]. The output of each program is displayed beneath the query sequence as a 3-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. This consensus prediction is used in all subsequent processing by the system.

In addition, the program Disopred[13] is run to calculate a 2-state prediction of which regions of the query are likely to be structurally ordered (o) and which disordered (d). Again 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 analysing predicted functional sites (step 8). Finally, beneath the disorder prediction are the results of a ProSite[14] search, and any ProSite motif detected in the query is highlighted beneath the sequence with gold dots.

Structural homology detection and fold recognition

5 The bulk of the results page is occupied by a table containing a ranked list of the top 10 highest scoring matches to known template structures in the Phyre fold library and their respective models. The table consists of several columns which will be discussed in turn below. The first column, ‘View Alignments’ is discussed in step 6.

(A) SCOP code

This column indicates the unique identifier for the template structure matched by Phyre, the length of the template and the percentage sequence identity between the query and template calculated relative to the shortest sequence. The identifier itself is of the form [d/c][PDB code][chain identifier][domain number]. The initial ‘d’ or ‘c’ character indicates the structure is a SCOP **d**omain or a whole **c**hain 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 multi-domain, yet single chain, of a protein.

Matches with high percentage sequence identity (>40%) are highlighted in red. 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. It is now commonplace, using protocols such as Phyre, to achieve high accuracy models at very low sequence identities. 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 modelled with an rmsd to the native structure of 2-4Å even at such a low sequence identity. This shortcoming of sequence identity as a measure of predictive accuracy is why the ‘Estimated precision’ described in part C below has been developed as a more useful guide.

(B) View model

This column contains an image of the 3D model of the query protein. Clicking on the image permits the user to download the 3D coordinates in standard PDB format for use in any external application. In addition, there are several icons beneath the image, the most immediately useful being the ‘JMol’ icon. Clicking this icon opens the model in the web browser using the powerful JMol molecular viewer (www.jmol.org). 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.

(C) Estimated Precision

This column indicates the confidence that the query sequence is homologous to the template in question. 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 homologues according to the SCOP database. The confidence values are colour coded from red to blue indicating high and low confidence respectively.

CAUTION: 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 rmsd 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])

(D) Fold, Superfamily and Family annotation

These three columns contain information about the template extracted from the SCOP database or, in the absence of such information, from the header information supplied in the coordinate file deposited in the Protein Data Bank. This information can be particularly helpful in providing hints at the possible function of the query and in assessing the level of consensus of the matches in terms of fold and function. The presence of four or five templates with similar folds or functions lends more weight to a prediction than a singleton.

Alignment assessment

6 The first column of the main fold recognition results table is a link to a page containing extensive information on the alignment, patterns of conservation and predicted functional sites. 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 homologues, and in determining the final accuracy of the three-dimensional model. As with model accuracy assessment described above, 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. Below we highlight some generic features of an alignment that one should examine. Figure 2 shows a screenshot of a typical alignment view in Phyre.

(A) Secondary structure matching

Secondary structure elements tend to be conserved in remote homologues despite extensive changes in their amino acid sequence. For this reason, a correctly aligned pair of homologues is expected to display considerable agreement between the secondary structures at aligned positions. 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.

For the query sequence we only have available its predicted secondary structure. However, for the template we possess both the predicted and known secondary structure. Disparities between predicted and known secondary structure may indicate certain (perhaps unusual) sequence features of the template that tend to be mispredicted by secondary structure prediction schemes. Thus the user should also consult the confidence values in the secondary structure prediction of the query (see Step 4) to determine whether mismatches in aligned secondary structure elements are misleading due to low confidence values in these regions of the secondary structure prediction.

(B) Domain boundaries

The beginning and end of the alignment with respect to either template or query can indicate domain boundaries where the user may consider chopping the query sequence and resubmitting to the Phyre system. The domain boundaries elucidated in this way may be more informative than those using the initial PSI-

Blast search (Step 3) as the Phyre profile-profile alignment strategy is more powerful in determining remote homology than the simpler sequence-profile matching performed by PSI-Blast.

(C) Alignment accuracy and match quality

A method for automating the assessment of alignment accuracy has been implemented similar to that of Tress et al.[15]. 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. These positions are colour-coded to indicate high and low scoring matches. 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 likely to be poorly aligned and are highlighted with a blue bar.

Functional site prediction

8 A common requirement for many users of protein structure prediction tools is to predict the residues likely to be involved in the function of the protein. To this end we have implemented a variety of tools to analyse and combine information from the alignment and the three-dimensional 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[16] (www.sbg.bio.ic.ac.uk/confunc/).

(A) Conservation

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 homologues at each position in the pseudo-multiple sequence alignment described in Step 3. The calculation is performed at different sequence identity thresholds to compensate for possible redundancy or paucity of homologues at each position in the query. Large numbers of highly similar sequences in the alignment may skew conservation scores, whilst 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 to straightforward conservation, the sequence (Shannon) entropy is calculated at each position. Finally, we apply the Evolutionary Trace algorithm of Yao et al.[17] to predict a score of 'functional importance' for each position based on the correlations of variations in the sequence homologues with their phylogenetic tree.

(B) Template binding site information

In some cases the template structure detected by Phyre may contain binding site information. We use the eF Site database[18] as a source of such information. Any residues in the template known to be proximal to a biologically relevant ligand in the crystal structure are highlighted in the alignment view with the letter '*f*'. If that 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 grey. Thus, at a glance one can assess whether the query and template are likely to share a common functional site.

(C) Cleft detection

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[19]. To locate such clefts we use the Pocket program[20]. 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. The result is usually a small number (between 1 and 5) of pockets ranked by volume from largest to smallest. Those residues found within the five largest pockets are labelled according to the index of their pocket. For example a residue labelled '1' belongs to the first and largest pocket.

(D) Consensus functional site prediction

Integrating the above sources of data into a final prediction is non-trivial. The Phyre server creates a weighted average of the information from steps 8A-8C for each position in the query sequence normalised between 0 and 9, 9 being the highest confidence prediction of a functionally relevant residue. These confidence scores are mapped to a colour scale and are use to colour a space-filling model of the query which can be interactively viewed using the JMol browser application. A clickable image of this structure is shown below the main alignment.

When examining this rendering in JMol, one is most interested in tightly clustered, red-coloured residues. This is a strong indicator of a potential functional site. An example can be found on the Phyre web site illustrating the detection of the heme binding site in a Globin fold by clicking on the 'example' link on the Phyre home page. This is also shown in Figure 3.

Building a broad consensus across methods

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 familiarise themselves with each of them. 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.

Troubleshooting

Indels and missing coordinates

Although the Phyre system uses powerful loop modelling techniques to model insertions and repair deletions in the alignment to the template, in some cases this system will fail. 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.

Modelling 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 three-dimensional model. In certain cases, e.g. if the point mutation lies in a loop that is processed by our loop modelling software, a change will be observed. However, the accuracy of loop modelling is in general insufficient to permit any firm conclusions to be drawn from such differences.

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: 1) The query adopts a known fold but is so remote from any solved structure that this homology or analogy cannot be detected, or 2) The query constitutes a novel fold. As mentioned above it is always wise to use a broad spectrum of available tools and search for a consensus. Even when none of the available tools is capable of providing a confident assignment, where there is a commonly occurring fold or superfamily returned by several diverse predictive systems, this can be provide important clues to guide further research.

Novel folds and small proteins

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[21] and refined and improved by the lab of David Baker[22]. Such methods fragment the query sequence into fully overlapping short stretches of amino acids (usually 9 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[23] 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.

As mentioned in the introduction, *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 which otherwise would be completely unavailable in the absence of an experimentally derived structure.

References

1. CASP 7 special issue. *Proteins* **69** (S8) 1-207 (2007).

2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation ofprotein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

3. Karplus, K., Barrett, C. & Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856 (1998).

4. 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).

5. 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**(3) 611-625 (2008).

6. 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).

7. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. The protein data bank. *Nucleic Acids Res* **28** 235–242 (2000).

8. Marchler-Bauer, A. *et al.* CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res*. **35** (Database Issue) D237-40 (2007).

9. Finn, R.D., *et al*. The Pfam protein families database. *Nucleic Acids Res.* **36** (Database Issue) D281-D288 (2008).

10. McGuffin, L.J., Bryson, K. & Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics* **16** 404–405 (2000).

11. 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**(2) 228-35 (2002).

12. Cole, C., Barber, J.D., & Barton, G.J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36** (Web Server issue) W197-201 (2008).

13. 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**(13) 2138-9 (2004).

14. Hulo, N. *et al.* The 20 years of PROSITE. *Nucleic Acids Res.* **36** (Database issue) D245-9 (2008)

15. Tress, M.L., Jones, D.T. & Valenica, A. Predicting Reliable Regions in Protein Alignments from Sequence Profiles. *J Mol Biol.* **330**(4) 705-718 (2003).

16. Wass, M.N. & Sternberg, M.J.E. ConFunc - functional annotation in the twilight zone. *Bioinformatics* **24**(6) 798-806 (2008).

17. Yao, H., *et al.* An Accurate, Sensitive, and Scalable Method to Identify Functional Sites in Protein Structures. *J. Mol. Biol.* **326** 255-261 (2003).

18. 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).

19. Laskowski, R.A. *et al.* Protein clefts in molecular recognition and function. *Prot. Sci.* **5**(12) 2438-2452 (1996).

20. Liang, J, Edelsbrunner, H, Fu, P, Sudhakar, PV and Subramaniam, S. Analytical shape computation of macromolecules I and II. *Proteins* **33** 1-17 and 18-29 (1998).

21. Jones, D.T. Predicting novel protein folds by using FRAGFOLD. *Proteins* **45** (S5) 127-32 (2001).

22. 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).

23. Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69** (S8) 108-17 (2007).

Table 1. Popular web servers for remote homology/fold recognition. 'Consensus' indicates the server collates results from multiple independent servers to form a final prediction, whereas 'single' indicates 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*, etc). 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'*).

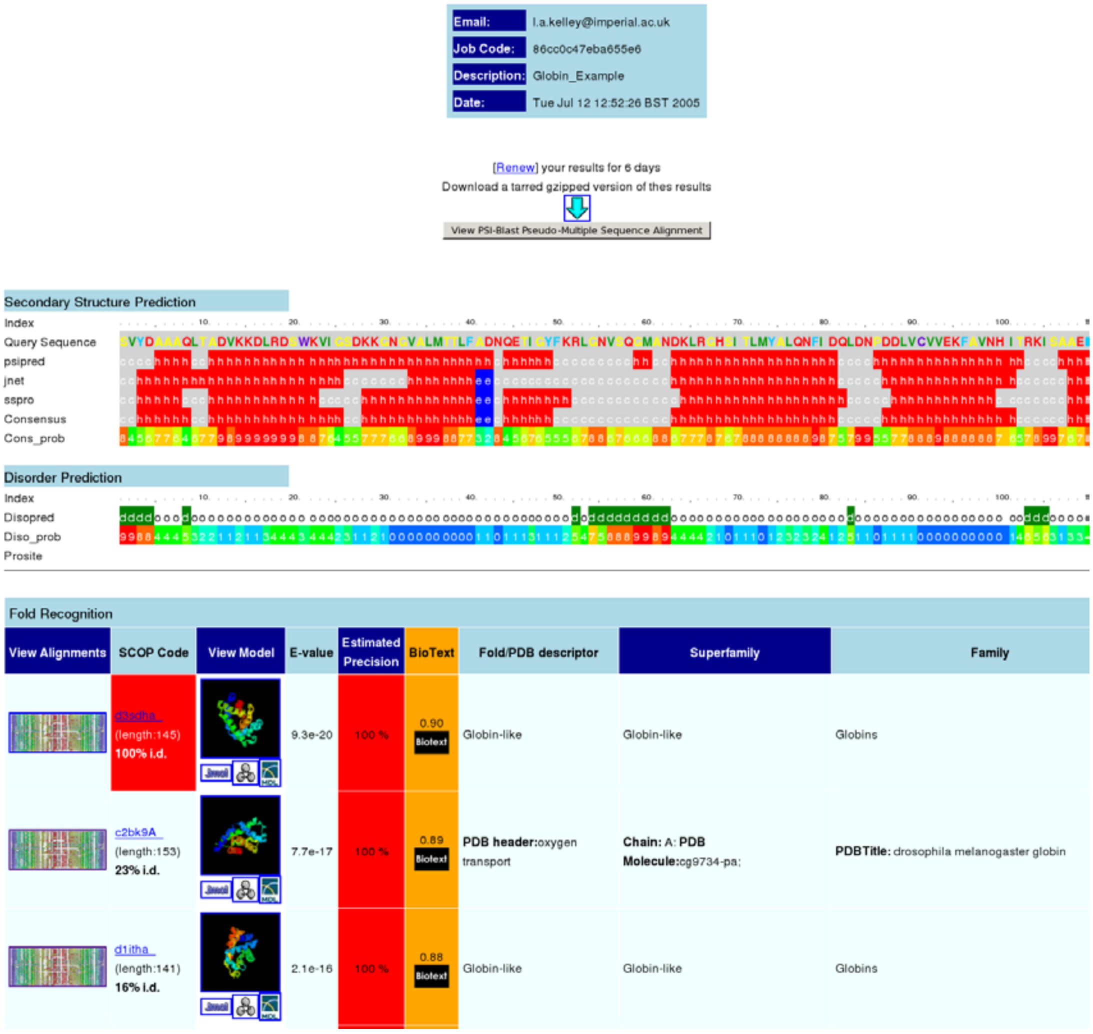
Server Name	*Web address*	*Consensus/ single*	*Model Building/confidence measure?*	*FR/ abinitio*
Phyre	http://www.imperial.ac.uk/phyre/	Single	Model + confidence	FR
I-TASSER	http://zhang.bioinformatics.ku.edu/I-TASSER/	Single	Model + confidence	FR+ ab initio
SAM-T06	http://www.soe.ucsc.edu/compbio/SAM_T06/T06-query.html	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 +Pcons 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

Figure Legends

Figure 1. Example of a typical Phyre results page The page is roughly divided into three sections entitled Secondary Structure Prediction, Disorder Prediction and Fold Recognition from top to bottom respectively. Details regarding each section can be found in the text.

Figure 2. Example of a typical Phyre alignment view For each of the ten modelling 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 coloured according to the confidence of the prediction (See Figure 3 for more details).

Figure 3. Example of predicted functional sites coloured 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 colouring 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 favourably with the known functional sites in the template.



Email:	l.a.kelley@imperial.ac.uk
Job Code:	86cc0c47eba655e6
Description:	Globin_Example
Date:	Tue Jul 12 12:52:26 BST 2005

[Renew] your results for 6 days

Download a tarred gzipped version of thes results

View PSI-Blast Pseudo-Multiple Sequence Alignment

Secondary Structure Prediction

```
Index           . . . . . . . . . 10. . . . . . . . . 20. . . . . . . . . 30. . . . . . . . . 40. . . . . . . . . 50. . . . . . . . . 60. . . . . . . . . 70. . . . . . . . . 80. . . . . . . . . 90. . . . . . . . . 100. . . . . . . .
Query Sequence  SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTLFADNQETIGYFKRLGNVSQGMANDKLRGHSITLMYALQNFIDQLDNPDDLVCVVEKFAVNHITRKISAAE
psipred         cccchhhhcchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhchhhhhhccccccccchhcchhhhhhhhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhcccchhh
jnet            cchhhhhhhhhhhhhhhhhhhhhhhcccccccchhhhhhhhheecccccccccccccccccccchhhhhhhhhhhhhhhhhhcccccchhhhhhhhhhhhhhhcccccchh
sspro           cchhhhhcccchhhhhhhhhhhhhhcccccchhhhhhhhhhhhheechhhhhhhhhcccccccccccchhhhhhhhhhhhhhhhhhhcccccchhhhhhhhhhhhcccccccchh
Consensus       cchhhhhhcchhhhhhhhhhhhhhhhhcchhhhhhhhhhhhheechhhhhhcccccccccccccchhhhhhhhhhhhhhhhhhhcccccchhhhhhhhhhhhhhcccccchh
Cons_prob       84567778467798999999988784557778689998877328456765556788676668867778767888888889875799557788898888887657899767
```

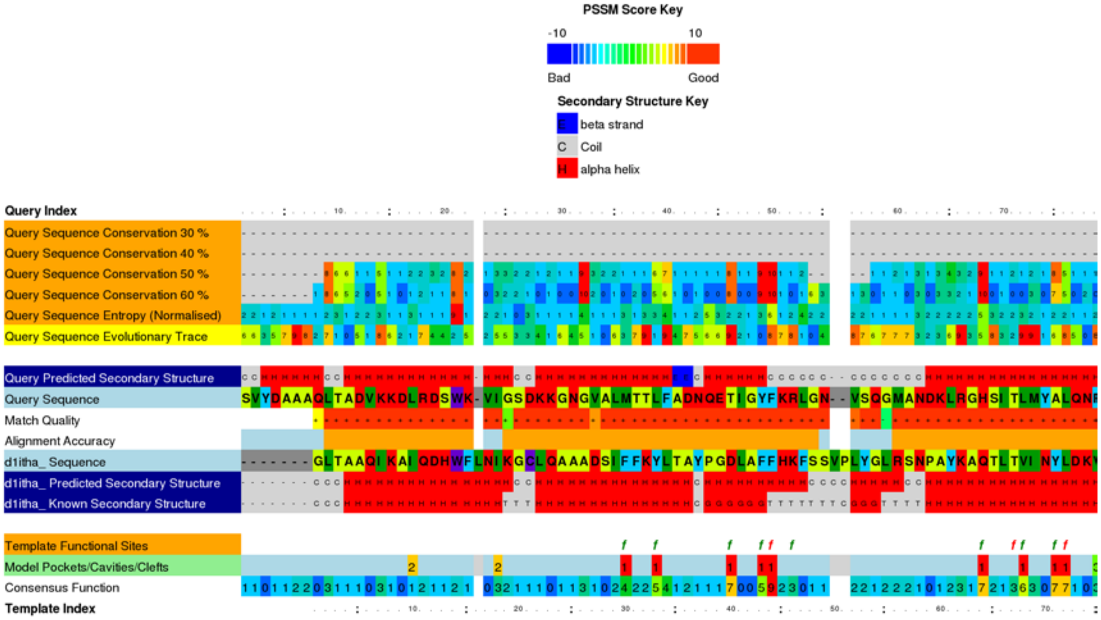
Disorder Prediction

```
Index           . . . . . . . . . 10. . . . . . . . . 20. . . . . . . . . 30. . . . . . . . . 40. . . . . . . . . 50. . . . . . . . . 60. . . . . . . . . 70. . . . . . . . . 80. . . . . . . . . 90. . . . . . . . . 100. . . . . . . .
Disopred        ddddooooodoooooooooooooooooooooooooooooooooooooooooooododdddddddddooooooooooooooooooooodoooooooooooooooooooooodddooooo
Diso_prob       99884445322112113444344423112100000000001101113111254758889989444421011101232324125110111100000000001465631334
Prosite
```

Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily	Family
	d3sdha_ (length:145) **100% i.d.**	Jmol	9.3e-20	100 %	0.90 Biotext	Globin-like	Globin-like	Globins
	c2bk9A_ (length:153) **23% i.d.**	Jmol	7.7e-17	100 %	0.89 Biotext	**PDB header:**oxygen transport	**Chain:** A: **PDB Molecule:**cg9734-pa;	**PDBTitle:** drosophila melanogaster globin
	d1itha_ (length:141) **16% i.d.**	Jmol	2.1e-16	100 %	0.88 Biotext	Globin-like	Globin-like	Globins
]

FASTA Text version of alignment


PSSM Score Key
-10
10
Bad
Good
Secondary Structure Key
E beta strand
C Coil
H alpha helix
Query Index
Query Sequence Conservation 30 %
Query Sequence Conservation 40 %
Query Sequence Conservation 50 %
Query Sequence Conservation 60 %
Query Sequence Entropy (Normalised)
Query Sequence Evolutionary Trace
Query Predicted Secondary Structure
Query Sequence
SVYDAAAQLTADVKKDLRDSWK-VIGSDKKGNGVALMTTLFADNQETIGYFKRLGN--VSQGMANDKLRGHSITLMYALQNF
Match Quality
Alignment Accuracy
d1itha_ Sequence
-------GLTAAQIKAIQDHWFLNIKGCLQAAADSIFFKYLTAYPGDLAFFHKFSSVPLYGLRSNPAYKAQTLTVINYLDKV
d1itha_ Predicted Secondary Structure
d1itha_ Known Secondary Structure
Template Functional Sites
Model Pockets/Cavities/Clefts
Consensus Function
Template Index
10
20
30
40
50
60
70

Consensus Functional Sites Mapped onto Model