# The Phyre2 web portal for protein modeling, prediction and analysis

Lawrence A Kelley[1], Stefans Mezulis[1], Christopher M Yates[1,2], Mark N Wass[1,2] & Michael J E Sternberg[1]

[1]Structural Bioinformatics Group, Imperial College London, London, UK. [2]Present addresses: University College London (UCL) Cancer Institute, London, UK (C.M.Y.); Centre for Molecular Processing, School of Biosciences, University of Kent, Kent, UK (M.N.W.). Correspondence should be addressed to L.A.K. (l.a.kelley@imperial.ac.uk).

**Phyre2 is a suite of tools available on the web to predict and analyze protein structure, function and mutations. The focus of Phyre2 is to provide biologists with a simple and intuitive interface to state-of-the-art protein bioinformatics tools. Phyre2 replaces Phyre, the original version of the server for which we previously published a paper in *Nature Protocols*. In this updated protocol, we describe Phyre2, which uses advanced remote homology detection methods to build 3D models, predict ligand binding sites and analyze the effect of amino acid variants (e.g., nonsynonymous SNPs (nsSNPs)) for a user's protein sequence. Users are guided through results by a simple interface at a level of detail they determine. This protocol will guide users from submitting a protein sequence to interpreting the secondary and tertiary structure of their models, their domain composition and model quality. A range of additional available tools is described to find a protein structure in a genome, to submit large number of sequences at once and to automatically run weekly searches for proteins that are difficult to model. The server is available at http://www.sbg. bio.ic.ac.uk/phyre2. A typical structure prediction will be returned between 30 min and 2 h after submission.**

## INTRODUCTION

In September 2014, The UniProtKB/TrEMBL protein database contained over 80 million protein sequences. The Protein Data Bank contains just over 100,000 experimentally determined 3D structures. This ever-widening gap between our knowledge of sequence space and structure space poses serious challenges for researchers who seek the structure and function of a protein sequence of interest.

Fortunately, advances in computational techniques to predict protein structure and function can substantially shrink this gap. On average, 50–70% of a typical genome can be structurally modeled using such techniques[1]. The key principles on which such techniques work are: (i) that protein structure is more conserved in evolution than protein sequence, and (ii) that there is evidence of a finite and relatively small (1,000–10,000) number of unique protein folds in nature[2]. These principles permit the protein structure prediction problem to be considered as a problem of matching a sequence of interest to a library of known structures, rather than the more complex and error-prone approach of simulated folding.

For over 30 years researchers have developed and refined computational methods for protein structure prediction. Such methods include simulated folding using physics-based or empirically derived energy functions, construction of models from small fragments of known structure, threading where the compatibility of a sequence with an experimentally derived fold is determined using similar energy functions and template-based modeling (TBM), in which a sequence is aligned to a sequence of known structure on the basis of patterns of evolutionary variation. TBM encompasses the strategies that have been called homology modeling, comparative modeling and fold recognition. It is this technique that has become the most universally reliable and widely used technique by both the modeling and wider bioscience communities. The success of TBM over other methods is due to three main factors: (i) the development of powerful statistical techniques to extract evolutionary relationships from homologous sequences; (ii) the enormous growth in sequencing projects, which provides the raw information; and (iii) the power of computing to process large databases with a fast turn-around.

Today, the most widely used and reliable methods for protein structure prediction rely on some method to compare a protein sequence of interest with a large database of sequences, to construct an evolutionary or statistical profile of that sequence and to subsequently scan this profile against a database of profiles for known structures. This results in an alignment between two sequences, one of unknown structure and one of known structure. One can then use this alignment, or set of equivalences, to construct a model of one sequence on the basis of the structure of another. When the sequence similarity between the protein of interest and the database protein(s) is low, then detection of the relationship and the subsequent alignment can be enhanced if structural information is included to augment the sequence analysis.

Phyre2 is the successor of Phyre, which we previously published in *Nature Protocols*[3]. Although the original Phyre[3] and the new Phyre2 share the common aim of protein modeling, the new Phyre2 system described in this updated protocol has been designed from scratch, and it shares no components with its predecessor. Phyre2 was launched in January 2011, but to date users have needed to reference the original Phyre protocol[3], which has been cited over 2,400 times. Phyre2 is one of the most widely used protein structure prediction servers, and it serves ~40,000 unique users per year, processing ~700–1,000 user-submitted proteins per day. In collaboration with other groups, we have applied Phyre2 to the annotation of a wide range of genomes[4–6].

### Comparison with other methods

There exist a number of other powerful structure prediction servers on the web. However, for the majority of modeling tasks, the differences in accuracy between such tools are minor[7]. The key differentiating factor for Phyre2 is ease of use. One of the primary objectives of the Phyre2 server is to provide a user-friendly interface to cutting-edge bioinformatics methods.

This enables biologists inexpert in bioinformatics to use state-of-the-art techniques without the very steep learning curve typical of many online modeling tools.

Some of the most widely used web servers for protein modeling are Phyre2, i-TASSER[8], Swiss-Model[9], HHpred[10], PSI-BLAST–based secondary structure prediction (PSIPRED)[11], Robetta[12] and Raptor[13]. In international blind trials of protein structure prediction methods (critical assessment of protein structure prediction or CASP)[7], it is observed that for the majority of modeling tasks there is no major difference in the accuracy of these methods. In extremely difficult modeling tasks, in which remote homology is uncertain and in which key regions of a sequence cannot be matched to a known structure, i-TASSER[8] has shown a small but consistent performance improvement over other methods. Phyre2 has been tested in the CASP9, 10 and 11 experiments (results can be seen at http://predictioncenter.org/index.cgi). To compare the performance with other systems, we consider fully automated systems for TBM and the average model quality (known as GDT_TS in CASP) that they produce over the course of the CASP experiment (120 protein domain targets in CASP9 and 98 targets in CASP10). Typically, these domains share <30% sequence identity with an identified template. As a single research group may submit multiple servers in CASP, we consider only the single best-performing server from each participating group. In the CASP9 experiment, Phyre2 ranked sixth out of ~55 unique groups. The five superior groups to Phyre2 had an average improvement in model quality of 2.8%, with only i-TASSER showing a 5% improvement. In CASP10, Phyre2 ranked tenth out of ~45 groups. Excluding i-TASSER (8% improvement), the remaining eight superior groups showed an average improvement over Phyre2 of 3.7%.

To understand these improvements in a structural context, one should note that in a typical 200-residue protein, a 1% improvement in model quality roughly corresponds to two extra residues being within 4.5 Å of the native conformation. CASP11 data for average model quality is not yet available from the prediction center website. We consider that the primary difference between these servers and Phyre2 is not in accuracy, but in the ease of use by non-bioinformaticians.

### Limitations
There are two principal limitations to the methods used by Phyre2 and other related servers. First, if homology cannot be detected between a user-supplied sequence and a sequence of known structure, then modeling will either be impossible or extremely unreliable. This reflects the wider ongoing difficulty of the protein-folding problem. There are still no reliable methods to predict a protein structure purely from sequence alone without reference to known structures.

The second limitation, again applicable to all widely used methods, is predicting the structural effects of point mutations. Phyre2 has functionality to predict the phenotypic effect of a point mutation, but it is unable to accurately determine, beyond the estimated position of a side chain, the wider structural effect of a point mutation. This means that a user attempting to model several single-position variants using Phyre2 will receive essentially identical models with a different side chain at the position of the variant. Generally, no alterations of the backbone of the protein will be observed.

It is often the case that users do not want only a single-chain model of their protein but a model of a multimer. This is not yet possible in Phyre2, but work is currently under way to add this functionality by using known multimeric structures as templates for complex building.

Finally, it is important to understand the potential limitations of modeling multidomain proteins using the 'intensive' mode of Phyre2 (described in stage 3b of the 'Modeling a single sequence' section, below). If homology models of separate domains without any mutual overlap are combined using the *ab initio* techniques described in stage 3b, the relative orientation of the domains in the resulting multidomain homology model is very likely to be incorrect. Such cases can be discerned by examining the table discussed in Step 12B. This can also apply to transmembrane proteins in which a homologous crystal structure of the globular or hydrophilic domain may be found and then grafted onto a transmembrane domain from another protein. This limitation will not apply if homology can be detected to a structure that spans the entirety of the user sequence. Future versions of Phyre2 will automatically detect these cases and provide a warning to the user.

### The Phyre2 server
The Phyre2 system is a combination of a large number of disparate software components created by our own group and others written in multiple languages. The system runs on a shared Linux farm of ~300 CPU cores. The Phyre2 server may be used in several different ways depending on the focus of the user's research. The most commonly used facility is the prediction of the 3D structure of a single submitted protein sequence. Advanced facilities include: (i) Backphyre, to search a structure against a range of genomes; (ii) batch submission of a large number of protein sequences for modeling; (iii) one-to-one threading of a user sequence onto a user structure; (iv) Phyrealarm, for automatic weekly scans for proteins that are difficult to model; and (v) Phyre Investigator for in-depth analysis of model quality, function and the effects of mutations. First, modeling of a single sequence will be discussed, followed by brief explanations of these tools. The procedure will deal mainly with a single query submission to Phyre2. These advanced facilities will not be detailed in the PROCEDURE, with the exception of the use of Phyre Investigator (optional; PROCEDURE Steps 35–39), as the results that they produce and their interpretation largely follow what we describe for a single sequence.

### Modeling a single sequence
The core method of Phyre2 for generating a 3D model of a protein sequence is composed of four underlying technical stages, described below and illustrated in **Figure 1**. There is also an optional intensive mode that attempts to create a complete full-length model of a sequence through a combination of multiple template modeling and simplified *ab initio* folding simulation. This is described in stage 3b and illustrated in **Figure 2**. These stages and their corresponding figures refer to the underlying algorithm being used for structure prediction. In contrast, the steps in the PROCEDURE are a guide to user navigation and analysis of the results of this algorithm. Throughout, the term 'query' refers to the user-submitted protein sequence.

**Stage 1: gathering homologous sequences.** The first stage is to determine an evolutionary profile for the query that captures the

**Figure 1** | Normal mode Phyre2 pipeline showing algorithmic stages. Stage numbers are shown in circles, and elements within a stage are surrounded by a dashed box. Stage 1 (gathering homologous sequences): a query sequence is scanned against the specially curated nr20 (no sequences with >20% mutual sequence identity) protein sequence database with HHblits. The resulting multiple-sequence alignment is used to predict secondary structure with PSIPRED and both the alignment and secondary structure prediction combined into a query hidden Markov model. Stage 2 (fold library scanning): this is scanned against a database of HMMs of proteins of known structure. The top-scoring alignments from this search are used to construct crude backbone-only models. Stage 3 (loop modeling): indels in these models are corrected by loop modeling. Stage 4 (side-chain placement): amino acid side chains are added to generate the final Phyre2 model.
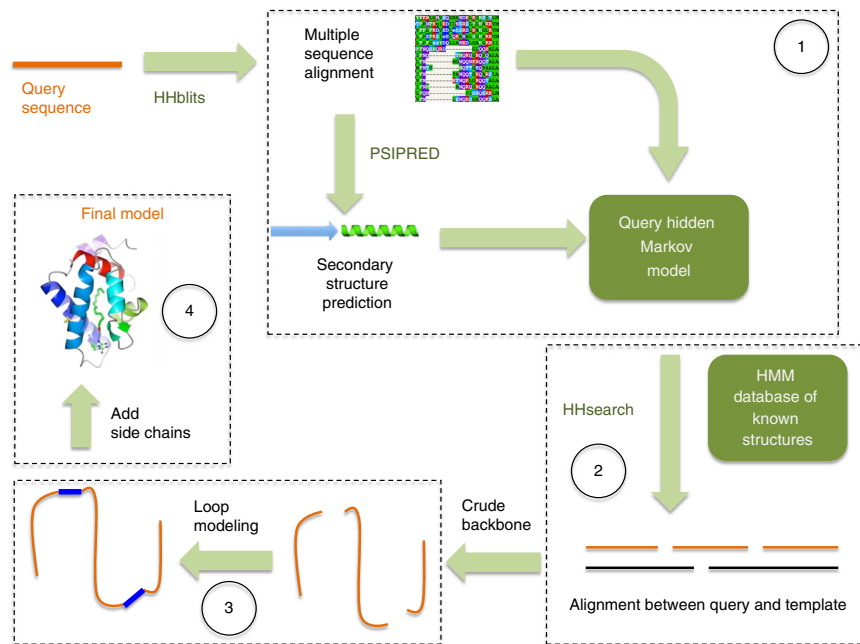
residue preferences at each position along its length. To construct an evolutionary profile, one needs to gather a large number of diverse yet true homologs. Diversity is key to the creation of a statistically representative distribution of amino acid preferences at each position in the protein, whereas avoiding false positives is vital so as not to pollute this distribution. Diversity may be achieved by searching the ever-growing protein sequence databases. In the past, the sequence database was mined using programs such as Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST)[14] that iteratively evolve a profile through multiple BLAST[14] scans of the sequence database— so-called sequence-profile matching. However, the most powerful approach to specific and sensitive collection of homologs is through profile-profile matching. Unfortunately, applying such a technique to large sequence databases is computationally prohibitive. Fortunately, recent powerful heuristics have been developed that overcome much of this computational burden. These heuristics effectively reduce profile-profile matching to sequence-profile matching by discretizing the vectors of 20 amino acid probabilities at each position into a restricted alphabet. This method, known as HHblits[15], demonstrates 50–100% increase in sensitivity (percentage of all true homologs detected) over PSI-BLAST and more accurate alignments without sacrificing computational speed. HHblits is used to scan the query against a sequence database in which no pair of sequences shares >20% identity, resulting in a sequence profile. In addition, the secondary structure of the query is predicted using PSIPRED[16]. PSIPRED is one of the most widely used methods for secondary structure prediction, and it uses neural networks
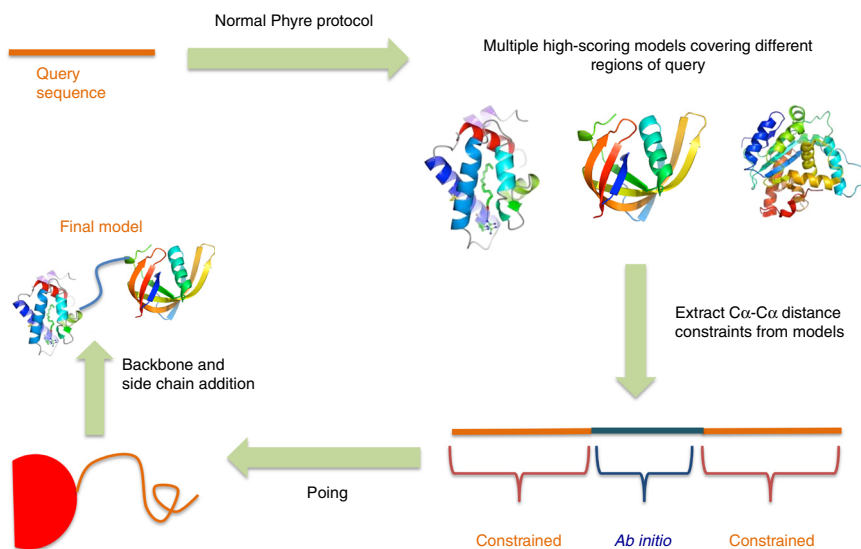
**Figure 2** | Intensive mode Phyre2 pipeline. Once a set of models has been generated, as shown in stages 1–3 of **Figure 1**, models are chosen by heuristics to maximize both confidence and coverage of the query sequence. Pairwise Cα-Cα distances are extracted from these models and treated as linear inelastic springs in Poing. Regions not covered by templates are handled by the *ab initio* components of the Poing algorithm: preferentially, bombardment of hydrophobic residues by notional solvent molecules to encourage burial, predicted secondary structure springs to maintain α-helix or β-strand conformations, and prevention of steric clash. The new protein is 'synthesized' from a virtual ribosome in the context of these forces, and the final Cα structure is used to construct a full backbone using Pulchra followed by side chain addition using R3.

trained on protein sequence profiles to predict the presence of α-helices, β-strands and coils with an average three-state accuracy of 75–80%.

**Stage 2: fold library scanning.** The profile calculated in stage 1, together with the predicted secondary structure, is converted to a hidden Markov model (HMM). This HMM is then scanned using HMM-HMM matching against a precompiled database of HMMs of known structure known as the fold library. The fold library is composed of a representative set of experimentally determined protein structures whose profiles have been calculated using the same approach as stage 1. The alignment algorithm used in Phyre2 is HHsearch[10], which is one of the leading homology detection methods, as demonstrated in international blind trials of protein structure prediction (CASP)[7]. The end result of the fold library scan is a list of query-template alignments ranked by their posterior probabilities, as produced by HHsearch. These alignments are used to generate crude backbone models that often contain insertions and deletions (indels) and that do not contain side chains.

**Stage 3: loop modeling.** Indels are handled using a library of fragments of known protein structures from lengths of 2–15 amino acids. This library is constructed by a complete fragmentation of the structure database, followed by structural clustering. A given gap in a model is characterized by its sequence, geometry of flanking regions and distance between endpoints. For insertions, a sequence-profile search is performed using the missing inserted sequence to detect fragments with similar sequence composition and endpoint distances, which creates a short list (typically 100 members) of potentially useful fragments. Similarly, for deletions, the sequence encompassing a window on either side of the deletion is used. These fragments are fitted to the crude model using cyclic coordinate descent[17], which is a robotic arm algorithm that attempts to fit the ends of the fragment to the crude model while minimizing changes in the dihedral angles of the fragment. Finally, fitted fragments are ranked using a combination of empirical energy terms and the top scoring model selected. In some cases, it is not possible to fit a fragment to an indel and such gaps remain in the backbone. This is often an indication of errors in the original alignment. See Steps 25–28 of the PROCEDURE for details on alignment interpretation.

**Stage 3b: multiple template modeling with Poing.** This stage is only performed if intensive mode is used. The aim of this stage is to create a complete model of the query protein even when different regions or domains are modeled by separate templates, or when there are no templates at all (*ab initio* modeling). To do this, we use Poing[18], a simplified protein-folding simulator. First, heuristics are used to select a subset of models produced in stages 2 and 3 that increase coverage of the query protein while maintaining high confidence, as reported by HHsearch. These input models provide distance constraints between different pairs of residues. These restraints are modeled as linear inelastic springs. In Poing, restraints are added as the query protein is slowly synthesized from a virtual ribosome. Residues that are not constrained by input models are modeled *ab initio* by Poing's solvent bombardment model, predicted secondary structure springs and penalization of steric clashes. 5–100 models are synthesized

in this way depending on the protein size (fewer for large proteins owing to computational demand) and clustered to choose a final representative model. As this model is composed only of α-carbons, its backbone is reconstructed using Pulchra[19].

**Stage 4: Side chain placement.** Side chain fitting to the backbone generated in stage 3 or 3b is performed using the R3 protocol[20] that involves a fast graph-based technique and a side chain rotamer library to place side chains in their most probable rotamer while avoiding steric clashes. This technique is ~80% accurate if the backbone is correct.

### Advanced facilities in Phyre2
**Backphyre: detecting a structure across genomes.** Frequently, users have a protein structure of interest and want to determine whether homologous structures exist in other genomes. For this purpose, HMM libraries must be generated for genomes of interest. Phyre2 currently contains such libraries for 30 genomes, and this number is constantly growing on the basis of user requests.

In Backphyre, a user uploads a structure in PDB format. The sequence of this structure is extracted and processed as in stage 1 above, while also including the known secondary structure within the HMM. This HMM is scanned as in stage 2 against one or more user-selected genomes from the 30 available. The final output screen is similarly laid out to that described in Steps 23–30 of the PROCEDURE.

**Batch analysis.** It is possible to run the single-sequence protocol on a large number of sequences uploaded by a user. By default, users are permitted to upload 100 sequences at a time, but this limit can be changed on request. Batch jobs are processed on spare computing power as it becomes available, and thus they are often somewhat slower than individually submitted jobs. Phyre2 processes on average 16,000 individual submissions per month and 7,000 batch sequences a month. Batch jobs can be monitored during processing. Summary pages for batch jobs are made available, as are facilities to download detailed or summary results for the entire batch. Each individual sequence has associated results pages whose interpretation is the same as in Steps 23 and 34 of the PROCEDURE.

**One-to-one threading.** Although the detection of remote homologous structures by Phyre2 has high specificity and sensitivity, it is sometimes the case that a user wishes to use a particular structural template on which to model their protein. Perhaps a user has a newly solved structure that is not yet published or a user has some biological information that indicates that their chosen template would produce a more accurate model than the one(s) automatically chosen by Phyre2. One-to-one threading allows a user to upload both a sequence to be modeled and the template on which to model it. HMMs of both the sequence and uploaded structure are calculated as in stage 1 above and aligned using the HHsearch algorithm. Unlike ordinary Phyre2 results, one-to-one threading does not of course produce a list of hits. Instead, the user is presented with an alignment view and a model of the protein together with information on the confidence of the match. See Steps 25–28 of the PROCEDURE for how to interpret this.

**Phyrealarm.** On the basis of statistics from 30,000 Phyre2 submissions over 2 months, on average more than 50% of all proteins submitted have had over 75% of their length modeled with >90% confidence. Of the remaining 50% of submissions, 25% have had less than a quarter of their sequence modeled and 25% have had between a quarter and three-quarters confidently modeled. A failure to detect confident structural matches for significant regions of a query is typically caused by one of three factors: (i) a lack of a sufficient number and diversity of homologous sequences to the query to create a useful profile/HMM; (ii) the evolutionary distance between the query and any known structure being too great to detect with the HMM-HMM matching method; or (iii) the query adopting a novel fold that is not present in the current structural database.

Fortunately, both the protein sequence database and structure database are growing every week, meaning that a currently undetectable homology may become detectable in the near future. For this reason, the Phyrealarm service was developed. If a user query cannot be modeled confidently, the protein may be added to an automated scan of new structures and new sequence databases each week. Every week ~100 new structures are added to the Phyre2 fold library, and every few months the clustered sequence database used for profile construction is updated. If a confident match is detected to this newly released data, the user query is automatically processed through the full Phyre2 modeling pipeline, and the user is sent the results and links by e-mail.

**Phyre Investigator.** Given a confident model produced by Phyre2, it is often desirable to perform more in-depth analyses of model quality, potential function and the effects of mutations (see optional PROCEDURE Steps 35–39). For these purposes, we developed Phyre Investigator. Any model produced by Phyre2 can be submitted to Phyre Investigator with one click from the results page for a range of analyses, including the following (see also refs. 10, 21–28):

- Model quality assessment by ProQ2 (ref. 21)
- Alignment confidence from HHsearch[10]
- Clashes, rotamers and Ramachandran analysis by Molprobity[22]
- Pocket detection by fpocket2 (ref. 23)
- Catalytic site detection from the CSA[24]
- Mutational analysis by SuSPect[25]
- Conservation analysis using Jensen-Shannon Divergence[26]
- Interface detection using the ProtinDb (http://protindb.cs.iastate.edu/) and PI-site[27] databases
- Detection of sequence features using the Conserved Domain Database (CDD)[28]

The Phyre Investigator interface (**Fig. 3**) has been designed to make this large amount of data easy to navigate and interpret simultaneously in a sequence and structural context. The screen is divided into three main sections from top to bottom: the information box, the structure view and analyses buttons, and the sequence view.

The structure view and analyses section is itself divided into three regions, from left to right: the JSmol interactive viewer (http://www.jmol.org/), the analyses buttons and two graphs showing sequence profile and mutational predictions. Clicking on an analysis button will display, in the information box, a brief summary of whichever analysis is currently active and links to downloadable raw data. It will also color the structure in the JSMol view in accordance with the analysis chosen and display a color-coded key to the left of the structure. Finally, it will add an extra row to the sequence view, illustrating the same information but in a sequence context.
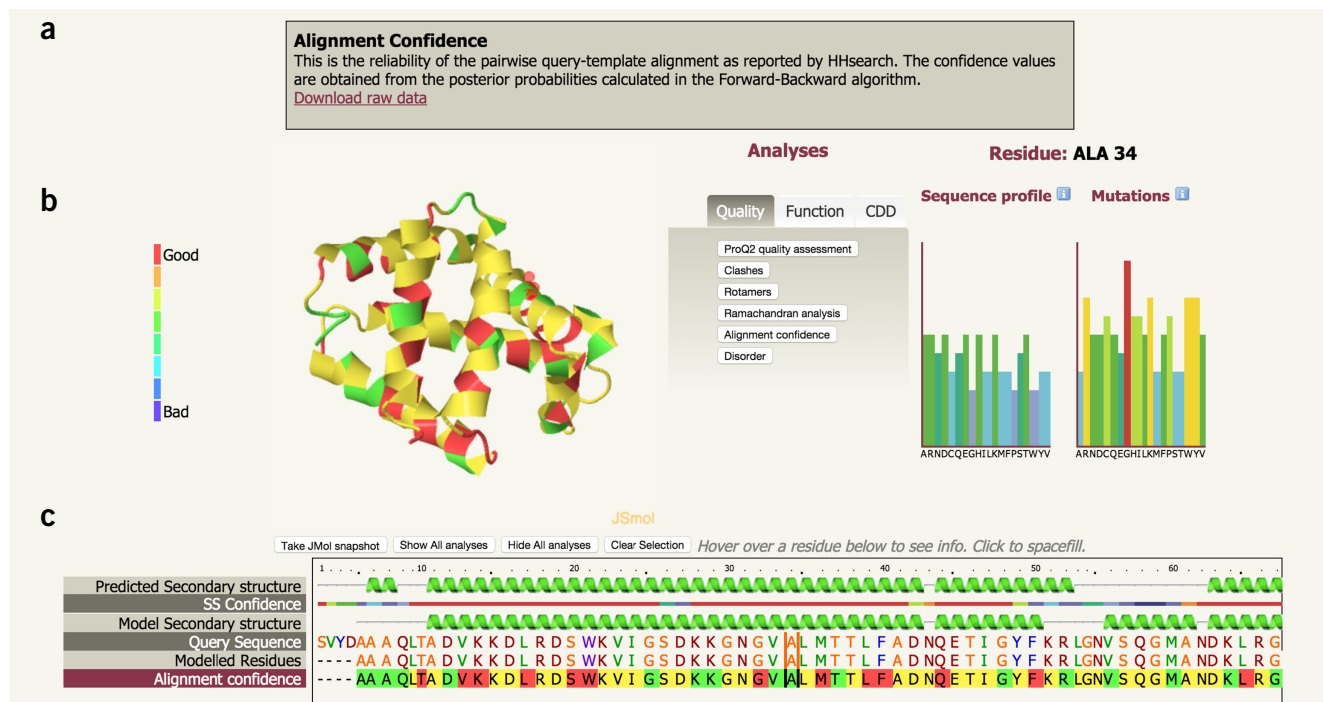


**Figure 3** | Phyre Investigator user interface. (**a**) Information box. (**b**) Structure and analyses view. (**c**) Sequence view. The structure and analyses view shows an interactive 3D JSmol viewer, buttons to toggle different analyses and two bar graphs (in this case for residue A34) showing the sequence profile preferences and predicted likelihood of a phenotypic effect from each of the 20 possible mutations at this position.

The sequence view displays the predicted secondary structure of your sequence, the confidence in this prediction, the secondary structure of the model, the amino acid sequence and which regions have been modeled. In addition, clicking on an analysis button will reveal an extra row showing the corresponding information from the analysis in a sequence context.

Hovering over a sequence position will highlight that position with vertical bars to either side of the residue in question. It will also highlight that residue in the JSmol 3D viewer as a red halo around the atoms of that residue. Finally, it will show the appropriate sequence profile and mutation graphs for that position described later. Clicking on a residue will cause that residue to be spacefilled in the JSmol viewer. You may select multiple residues by repeated clicking. At any time, you can clear your selection by clicking the 'Clear selection' button above the sequence view. You may also take a snapshot of the structure at any time using the 'Take JMol snapshot' button.

The sequence profile graph represents residue preferences in your protein at a particular sequence position. Residue preference for each amino acid type is displayed as a vertical colored bar, with tall, red bars being more favorable than shorter blue bars. These values are taken from the position-specific scoring matrix (PSSM) calculated by a PSI-BLAST scan of the query against a large sequence database (Uniref50).

The mutational analysis graph represents the predicted effect of mutations at a particular position in your sequence. Tall, red bars above a residue type indicate that a mutation to this residue is strongly predicted to have a phenotypic effect. These predictions are made using the SuSPect[25] method. SuSPect uses sequence conservation, solvent accessibility and protein-protein interaction network information to predict how likely a variant is to lead to disease in humans, demonstrating superior benchmark performance over other available methods, such as PolyPhen-2 (ref. 29), SIFT[30] and Condel[31]. The SuSPect method is available as a stand-alone web server (http://www.sbg.bio.ic.ac.uk/suspect/), with more options for uploading sets of sequences, viewing precalculated results for the entire human proteome and more.

When you are using SuSPect through Phyre Investigator, it is important that your sequence is wild type. Submitting a mutant protein to Phyre2 and then to Investigator will lead to misleading predictions from SuSPect. If the protein is a human protein, precalculated scores will be returned. For nonhuman proteins, scores will be calculated using a version of SuSPect incorporating protein structure but no network information. By incorporating network information, SuSPect performs best on mutations in human proteins.

**Phyre2 job manager.** If users register with the Phyre2 server (which is free), they gain access to various other tools including the Phyre2 job manager. This is accessed via the 'View past jobs' link at the top of the home page when logged in. Clicking the job manager takes the user to a page that allows them to see a summary and links to all of their past and running jobs. Every completed job has a link to results, which, when hovered over with the mouse, displays an image of the top scoring model with summary confidence and coverage information. Completed jobs remain by default on the server for 30 d. The job manager permits a user to select past jobs and renew them to prevent expiry, or delete them. This is also possible within the results page, as described in Step 8 of the PROCEDURE.

## MATERIALS

### EQUIPMENT
• A personal computer with an Internet connection and a web browser

**Data: the data required depend on the facilities used (as follows)**
• Standard protocol, single-sequence modeling. Amino acid sequence of the protein of interest 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
• Advanced facilities, BackPhyre. A protein structure file in PDB format
• Advanced facilities, one-to-one threading. Both an amino acid sequence and PDB structure
• Advanced facilities, batch processing. A file containing multiple amino acid sequences with no gaps in FASTA format

## PROCEDURE
### Sequence submission
**1|** Go to the Phyre2 home page (http://www.sbg.bio.ic.ac.uk/phyre2).

**2|** Enter your e-mail address. Results will be mailed to this address on completion.

**3|** Enter an optional job description.

**4|** Copy and paste your amino acid sequence (see Equipment section for data format) into the form provided.
**? TROUBLESHOOTING**

**5|** Choose 'normal' or 'intensive' mode (see description of stages 3 and 3b in the INTRODUCTION for further information) by clicking on the appropriate radio button in the form.
▲ **CRITICAL STEP** Normal is the default. It is recommended to use normal mode first and decide, on the basis of the critical point after Step 12 and Steps 20–22, whether it is worthwhile resubmitting your protein in intensive mode.

**6|** To submit the sequence, click the 'Phyre Search' button. On clicking the button, the user will be taken to a job monitoring page that is automatically updated every 30 s. This page shows a progress bar for the job, information on the job and an estimate of the time it will take. In addition, various user tips are periodically displayed at the bottom of the page to give users more information about the server. The user may choose to bookmark the page, or simply wait for completion. On completion, an e-mail will be sent to the user and the page will be re-directed to a generated results page.

**Obtaining results**

**7|** Upon job completion, an e-mail is sent to the user containing summary information, a unique job identifier and a link to the main results page and an attachment containing the top scoring model in PDB format. Note that job results are only visible to those with the unique job identifier (useful for sharing results) or the user's e-mail and password if they have registered. Click on the link in the e-mail. This will open a web page of results.

**Main results page navigation**

**8|** At the top of the results page is a box with information about the submission date, description, link to the input sequence and expiry date of the job. Check the expiry date. A new submission will say '30 d'. If the job is approaching expiry, the text will be orange or red. If so, click on the 'Renew for 30 d' link to reset the expiry date for the results.

**9|** A button labeled 'Download zip of all results' is present, which allows users to keep an off-line copy of the entire directory structure of the results page. Click this link to save a copy of results locally.

**10|** The page is divided into several main sections: summary (Steps 11 and 12; **Fig. 4**), sequence analysis (Steps 13–16), secondary structure and disorder prediction (Steps 17–19; **Fig. 5a**), domain analysis (Steps 20–22; **Fig. 5b**) and detailed template information (Steps 23 and 34; **Figs. 5c** and **6**). In some cases, depending on the protein submitted, two further sections will appear: binding site prediction and/or transmembrane helix prediction. Most sections can be dynamically shown or hidden using the show/hide links next to the section titles to reduce screen clutter. Below each section title are links to in-depth help on that section and to PDF versions of that section. Video tutorials for each section are under construction. Scroll to the top of the web page to view the 'Summary section'.

**Summary section**

**11|** The summary section (**Fig. 4**) displays an image of the top-scoring model (and its dimensions in Å) produced by Phyre2. Click the image to download the model in PDB format. This file is the same as the attachment in the e-mail received by the user.

**12|** The information to the right of the image is dependent on whether the user selected normal (option A) or intensive mode (option B).

  **(A) Normal mode results**
    (i) Normal results will show information about the known structural template on which this top model is based, the confidence and coverage of the model, a link to display and interact with the model using JSmol within the browser and a link to a FAQ regarding other available 3D molecular visualization tools. Click on the 'Interactive 3D view in JSmol' link to view and rotate the model. Click on 'Close JSmol' to return to the static image.



**Figure 4** | Example Phyre2 summary results page. On the left is an image of a large all-β structure. Clicking on the image will download a PDB-formatted file containing this structure. On the right are various data regarding the model, including the following: PDB code of the template used, information about the protein template extracted from the PDB file, confidence in the model and coverage of the query sequence (100% and 28%, respectively). In this case, there is additional text informing the user that although only 28% of the query could be modeled by a single template, other high-confidence templates were also detected that could increase this coverage to 55% by using Phyre's intensive mode. Finally, there is a link to launch the JSmol 3D viewer in the browser and a link to a FAQ describing popular external molecular viewing software.

**a**



**b**



**c**
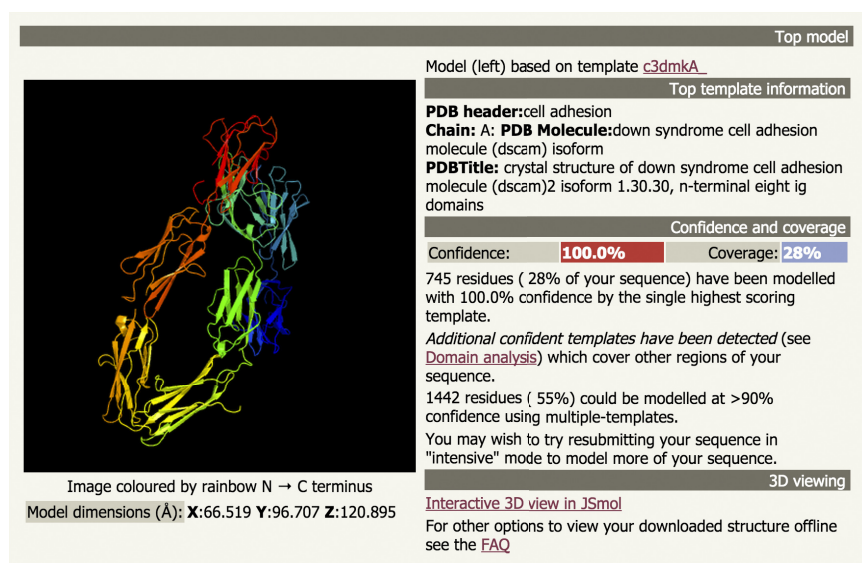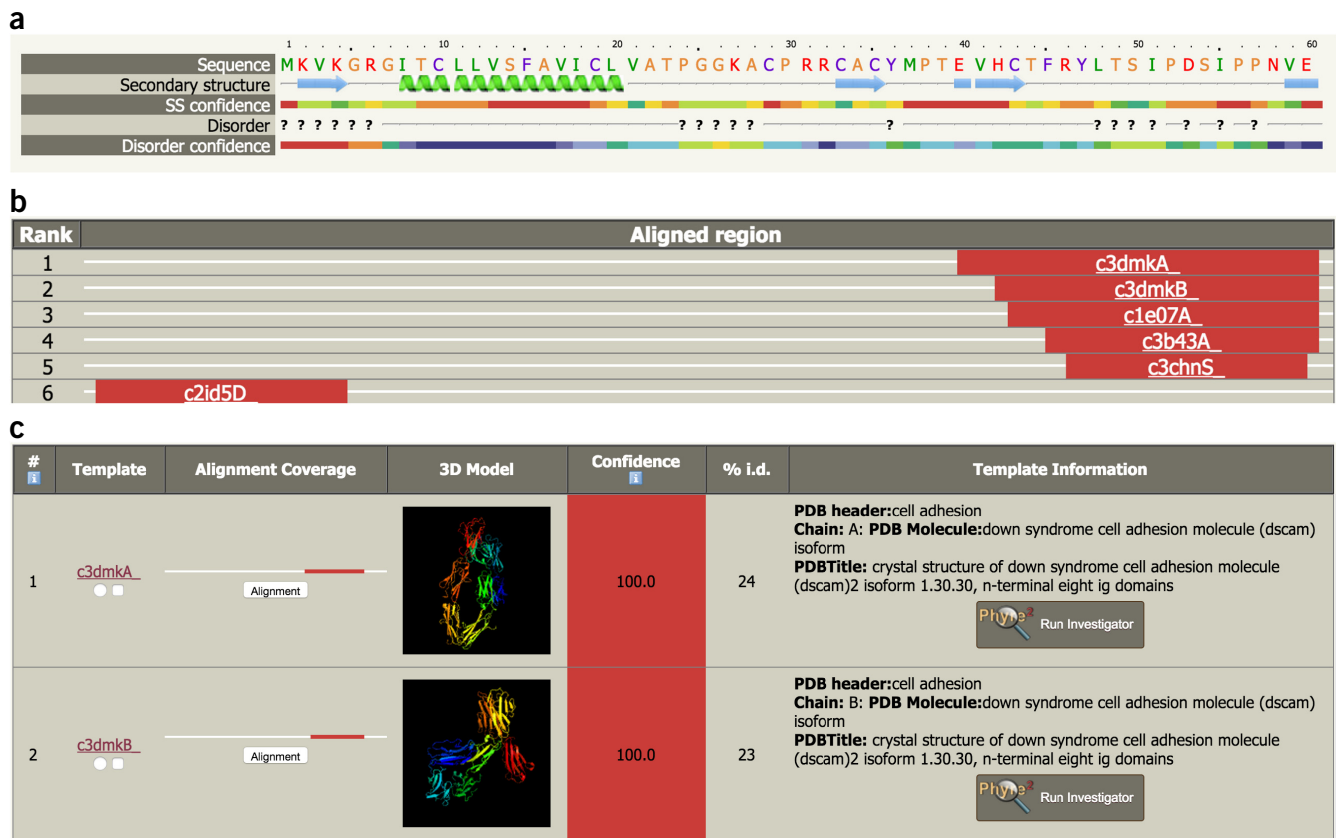


**Figure 5** | Examples of the three main sections of a typical Phyre2 results page. (**a**) Example secondary structure and disorder prediction. The query sequence is colored as described in Step 17. Question marks indicate predicted disordered regions. Each type of prediction is associated with a rainbow color-coded confidence (red, highest confidence; blue, lowest confidence). (**b**) Example of the domain analysis results section described in Steps 20–22. The width of the box indicates the length of the query sequence. In this example, confident (red) matches have been found at the N terminus (rank 6) and the C terminus (ranks 1–5), but no confident matches have been found to the intervening segment. (**c**) Example of the detailed table of results described in Steps 23 and 24 and 29–32. In this example, the rank 1 and rank 2 matches have confidence of 100% and sequence identities of 23% and 24%, respectively.

**(B) Intensive mode results**

(i) Intensive results will show a horizontal (possibly multiline) colored bar that represents the user sequence together with a color-coded confidence key. The colors indicate the predicted confidence of the model along the sequence. Regions modeled by the *ab initio* approach of Phyre2 are always colored blue to indicate minimum confidence. Other colors are inherited from the confidence in the template(s) used to model that region. Click the 'Details' link to be taken lower down the page to the 'Multi-template and *ab initio* information' table. This table indicates which structural templates were used for which regions of the user sequence and their associated color-coded confidence. Click the browser's 'back' button to return to the top of the page of results.

▲ **CRITICAL STEP** Sometimes the confidence of the top model is too low to be useful. It is not recommended to consider models with a confidence value of <90%. Similarly, it may be that the top model does not cover a substantial fraction of the user protein. Sometimes this is because there are multiple domains in the protein covered by separate templates. See Steps 20–22 and the associated TROUBLESHOOTING section to see whether intensive mode may be valuable here. Phyre2 attempts to automatically determine whether other templates covering additional portions of your protein are available and whether they will provide a message to that effect and a recommendation to try intensive mode. However, if the confidence is poor (<90%) and there are no extra templates, the user is alerted to use Phyrealarm. In this case, clicking on the Phyrealarm icon or link will take the user to a prefilled web form wherein one click will add the sequence to the Phyrealarm system. As discussed in the INTRODUCTION, once it has been added to Phyrealarm, the sequence will automatically be scanned against new structures as they become available in the fold library each week. If a confident hit is detected, a full Phyre2 modeling job is automatically run and the user is e-mailed the results. If this happens, the user can resume the protocol at Step 7.

**Figure 6 |** Example alignment between user query sequence and known structure, as described in Steps 25–28. Sequence coloring is as described in Step 17. Identical residues between query and template have a gray background. Secondary structures (predicted and known) are displayed: in this case α-helices. Color-coded per-residue confidence in both the alignment (from HHsearch) and in secondary structure prediction is displayed. The level of residue conservation for both the query and template sequences is also shown, where thicker horizontal bars indicate greater degrees of conservation.
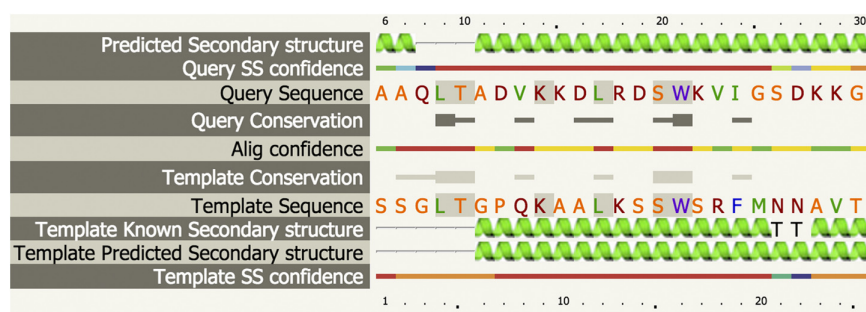
## Sequence analysis

**13|** In the 'Sequence analysis' section of the main result page, click the button entitled 'View PSI-Blast pseudo-multiple sequence alignment' to open a new window. This contains the results of scanning the query sequence against an up-to-date nonredundant protein sequence library using PSI-BLAST.

**14|** To determine how many homologous sequences were found, assess the number of rows in the newly opened window by looking at the left-hand side that indexes the number of homologs detected. 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 user sequence, the E-value reported by PSI-BLAST, the percentage sequence identity to the query and a sequence identifier for the homolog.

**15|** To ascertain whether highly informative alignments that are most likely to generate an accurate secondary structure prediction were obtained, assess 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 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 error-prone secondary structure prediction, a weak sequence profile and consequently poor overall structure prediction accuracy.

**16|** Close the window. Next to the 'View PSI-Blast pseudo-multiple sequence alignment' button is a link to a zipped FASTA-formatted version of the multiple sequence alignment. Click this to download the alignment for importing into any standard multiple-sequence viewer.

## Secondary structure and disorder prediction

**17|** In the secondary structure and disorder prediction section of the main results page (**Fig. 5a**), the position in the sequence is indicated in the top line. The sequence is represented on the next line with residues colored according to a simple property-based scheme: (A,S,T,G,P: small/polar) are yellow, (M,I,L,V: hydrophobic) are green, (K,R,E,N,D,H,Q: charged) are red and (W,Y,F,C: aromatic + cysteine) are purple. The secondary structure prediction comprises three states: α-helix, β-strand or coil. Green helices represent α-helices, blue arrows indicate β-strands and faint lines indicate coils. The 'SS confidence' line indicates the confidence in the prediction from PSIPRED, with red indicating high confidence and blue showing low confidence. Assess which regions are predicted with high and low confidence. A large amount of blue or green in the confidence line is indicative of few homologous sequences detected and a consequent low probability of modeling success.

**18|** The 'Disorder' line contains the prediction of disordered regions in your protein by DisoPred[32,] and such regions are indicated by question marks (?). Assess whether a large proportion of your sequence is predicted to be disordered both visually and by looking at the statistics shown below the prediction.
! CAUTION Secondary structure and disorder prediction is on average 78–80% accurate (i.e., 78–80% of the residues are predicted to be in their correct state). However, this accuracy is only reached if there is a substantial number of diverse sequence homologs detectable in the sequence database (Steps 13–16). If your sequence has very few homologs (something you can check by looking at the PSI-BLAST results via the button near the top of the results page), then the accuracy falls to ~65%. In addition, there are no predictions of β-turns, β-bends, π-helices or 310-helices. These classes are merged such that β-turns and β-bends are treated as coil, and π-helices and 310-helices are considered α-helices.
! CAUTION If a large (>50%) proportion of the query is predicted to be disordered, Phyre2 presents a warning that attempting to model the protein may not be meaningful as the protein is unlikely to adopt a globular structure.

**19|** The user sequence is also scanned against the CDD[28] for features of interest. When detected, these are also highlighted as colored dots at the appropriate positions in the sequence with a color key at the bottom of this section (disorder confidence line). Click on a feature to be taken to more detail at the CDD website.

**Domain analysis**

**20|** Click 'Show' next to the heading 'Domain analysis' on the main results page. This will open a scrollable table (**Fig. 5b**) whose width represents the length of the user protein. Matches by Phyre2 to known structures are shown as colored rows in which the colors represent the confidence in the homology (red is high confidence, whereas blue is low confidence).

**21|** At most, the top 20 high-scoring matches are model built. Lower-ranked hits are not modeled to conserve computing resources. (To model lower-ranked hits, see TROUBLESHOOTING for Step 29.) These have links in the centers of their aligned regions that are named after the template used. Hover over these to see a pop-up summary picture of the model and further information. Click this link to go to the equivalent entry in the detailed table of results (Step 23).

**22|** Scroll the table to determine which regions of your protein have been modeled. This enables you to see whether there are regions that cannot be modeled at all, or whether there are multiple templates covering different regions of your protein. This in turn is an indicator of the probable domain structure of your protein. Return to the top of the table and click 'Hide' to collapse the table.
**? TROUBLESHOOTING**

**Detailed template information**

**23|** Scroll down to the 'Detailed template information' table on the main results page (**Fig. 5c**). This displays information on the template code, alignment coverage, 3D model, confidence, percentage sequence identity and text description of the protein template. Look for red in the 'confidence' column for reliable models.

**24|** The matches are ranked by a raw alignment score (not shown) that is based on the number of aligned residues and the quality of alignment. This in turn is based on the similarity of residue probability distributions for each position, secondary structure similarity and the presence or absence of insertions and deletions. Each row provides information on the template used for the model and a small graphic indicating where along your sequence the match color-coded by confidence occurs. Beneath that line graphic is an alignment button. Hover over that button to view statistics about the start, end and percentage coverage of the alignment. Each model built by Phyre2 is based on an alignment generated by HMM-HMM matching. Both the predicted secondary structure of your sequence and the known and predicted secondary structure of the template are used in conjunction with the sequence information in generating the alignment.

**25|** Click the alignment button to go to a new page containing detailed alignment information (**Fig. 6**) and the ability to interactively inspect the model using JSmol. See Step17 for basic interpretation. One of the extra rows present here is entitled 'Template known secondary structure.' In the 'Template known secondary structure' row you will sometimes see 'S', 'T', 'G', 'I' and 'B' characters. These are assigned secondary structure types by the program DSSP. They represent the following: G = 3-turn helix (310 helix), I = 5-turn helix (pi helix), T = hydrogen bonded turn, B = residue in isolated β-bridge and S = bend. Identical residues in the alignment are highlighted with a gray background.

**26|** Click the links below the alignment to download a text version of the alignment, a simple pairwise representation of the alignment in FASTA format and the coordinates of the model in PDB format. Beneath these links is an image of the model. Click on the image to launch the JSmol applet inside the browser to interactively inspect the model. Click the 'Close JSmol' link when finished.

**27|** At the top of the screen are two buttons that can add or remove detail from the alignment. Click both buttons to display all extra information, as shown in **Figure 6**. The 'Conservation' rows contain information on residue conservation across the detected sequence homologs classified into three states. No symbol indicates unconserved, a thin gray bar indicates moderate conservation and a large block indicates a high degree of conservation.

**28|** Click on 'Return to main results' in the top left corner of the page and return to the detailed table of results.

**29|** The '3D Model' column contains a picture of the model constructed for your sequence based on that template. Click on the picture to download the coordinates of the model in PDB format for input to any other viewing or analysis programs you may have.
**? TROUBLESHOOTING**

**30|** The next two columns are 'Confidence' and '% i.d.'. Confidence represents the probability (from 0 to 100%) that the match between your sequence and this template is a true homology. Sequence identity is the proportion of user protein residues equiv-alenced to identical template residues in the generated alignment. Check the value of these two columns. If both values are colored red, you are dealing with a high-confidence close homology model. If only the confidence column is red, you are dealing with a remote homology that has still been modeled well but with greater expected deviation from native than a close homolog. **! CAUTION** Confidence does not represent the expected accuracy of the model—although the two are intimately related. If you have a match with confidence >90%, one can generally be very confident that your protein adopts the overall fold shown and that the core of the protein is modeled at high accuracy (2–4 Å r.m.s.d. from the native, true structure). However, surface loops will probably deviate from the native.

**31|** The 'Template Information' column provides either the fold, superfamily and family of the template, as determined from Structural Classification of Proteins (SCOP; http://scop.mrc-lmb.cam.ac.uk/scop/), or description fields taken from the 'PDB Header' and 'Title' fields if the structure in question is not present in the current version of SCOP. Check to see whether the functional information in this column matches any previous biological expectation you may have (ideally from experiment) about the function of your protein. Agreement in general functional class lends greater support to the prediction.

**32|** Also in this column is a button called 'Phyre Investigator'. When a model or match is particularly interesting, it is possible to perform a range of more in-depth analyses by clicking this button and submitting the model to Phyre Investigator. Find a model that interests you, and then click this button.

**33|** You will be taken to a page that explains Phyre Investigator and a 'submit' button. Click this button. This will show you a progress bar for the Phyre Investigator job. In the top left is a link 'Return to main results'. Click this link.

**34|** On the main results page, scroll down to the entry for the model that you submitted. You will see a message saying 'Investigator running'. Processing typically takes 5–10 min. When the job is complete, the message will change to a link saying 'View investigator results'. If you have been waiting longer than 10 min, refresh the page in your browser. Click this link and proceed with the optional Steps 35–39.

**Phyre Investigator (optional)**
**35|** The Phyre Investigator interface is composed of several sections (**Fig. 3**). Click the 'Quality' tab in the 'Analyses' section to see a list of buttons including ProQ2 (a model quality prediction system), clashes, rotamers and Ramachandran analysis. Click each of these in turn looking for regions of poor model quality. If these are far from residues in which you are inter-ested or in loops that are unlikely to be functionally important, they are not a concern. However, if problematic residues lie near functional sites, you should exercise caution and investigate other alternative models that may avoid these problems.

**36|** Click the 'Function' tab in the 'Analyses' section to display options such as conservation, interfaces, pockets and mutational sensitivity. Conservation can give clues as to likely functional residues. Highly conserved residues that are also present in a pocket are an even stronger indication of likely functional importance.

**37|** Click one of the interface buttons if available. Are predicted interface residues also conserved? If so, this increases confidence in transferring the known interface of the template to your protein.

**38|** If a residue appears problematic from the 'Quality' measures, or if it is likely to be functionally important from the 'Function' measures, hover over it in the sequence view and look at its profile and mutational graphs. Are there strong preferences for some types of amino acid? Are some mutations strongly predicted to have a phenotypic effect? This information can guide mutagenesis experiments or aid in interpreting SNPs.

**39|** When finished, click on 'Return to main results' in the top left corner of the screen.

**Superposition of models**
**40|** At the bottom of the main table (**Fig. 5c**) is a button entitled 'Generate superposition of selected models'. Beneath each template name (column two) of the main table are two buttons: first, a radio button that allows you to select one single master model on which other models will be superposed, and second, a tick box to select models (slaves) to be superposed on the master. Superposition is performed using the MaxSub algorithm[33]. This algorithm attempts to find the maximum subset of atoms between two structures that can be superposed within 4.5 Å. Typically, one would choose as the master either the top-ranked model or a model judged on some previous background biological knowledge to be most interesting. Click on the radio button to select the model that you wish to be the master model.

**41|** After selecting a master model, you may tick as many slave models as you wish. Slave models would typically be chosen as alternative structures with comparable confidence values to the top rank model or master model.

**42|** Click on the 'Generate superposition' button below the table; the ticked templates will be superposed on the master model. This will then take you to a page with a large JSMol window that displays the superposition for interactive viewing, together with descriptive help text.

**43|** Rotate the superposition, looking for the regions of the models that are in close agreement in 3D space. Often one will observe a conserved core with variable surface loops that can indicate where there is likely to be a modeling error or structural flexibility. This can be helpful to establish which regions of the models agree and disagree, which in turn can give you a sense for which regions of the model are trustworthy and which regions you should be cautious about. Take note of the structural similarity between models as indicated by the TM score, which is explained in the text on the web page.

**44|** Use the back button in your browser to return to the main results page.

**Binding site prediction**
**45|** If the top-ranked model is assigned a probability of >90%, the model and sequence are submitted automatically to the 3DLigandSite[34] server for ligand binding site prediction. Scroll down the main results page beyond the detailed template information table to the 'Binding site prediction' section. If your top model is >90% confident, a message to this effect and a link to results is presented. If the top rank model is either below 90% confidence or is predicted to contain substantial (>50%) disorder, it is not sent to 3DLigandSite and a message to this effect is displayed. A link is available to submit the model regardless of confidence, but this requires user intervention. Click this link to go to the 3DLigandSite page and explore the results. See the 3DLigandSite[34] paper and the site's FAQ for how to interpret these results. Return to the main results page.

**Transmembrane helix prediction**
**46|** Your sequence and the set of homologs detected by PSI-BLAST are processed by a support vector machine (a powerful machine-learning tool) to determine whether your sequence is likely to contain transmembrane helices, as well as to predict their topology in the membrane. For this, Phyre2 uses memsat-svm[35], which has demonstrated an average accuracy of 89% on a large test set. Scroll to the bottom of the main results page just below the 'Binding site prediction' section. If transmembrane helices are predicted, an image will be seen showing the extracellular and cytoplasmic sides of the membrane and the beginning and end of each transmembrane helix, illustrated with a number indicating the residue index. This information is not explicitly used in model generation but is presented as additional useful information for the user.

**? TROUBLESHOOTING**
**Step 4: how should long sequences be handled?**
There is currently a sequence length limit of 1,200 amino acids. Work is under way to extend this limit. If the query exceeds this limit, it is advised that the query be submitted to the CDD[29] to determine probable domain boundaries. The query may then be chopped at these boundaries to ensure that the length is below the limit and resubmitted to Phyre2. Future versions of Phyre2 will automate this step and display optional cut points to the user.

**Step 4: what if I only have an identifier and no sequence?**
If the user has only an identifier or descriptor of the protein of interest as opposed to the sequence itself, they can click the 'sequence finder' on the main submission page. This performs a rapid keyword search of a number of sequence databases to retrieve probable matches to the user query. Matches are returned as a table of sequences, species and Uniprot descriptors. One click inserts the chosen sequence into the main form.

**Step 22: should I resubmit my protein in intensive mode?**
This step gives you vital information on whether you should consider using the intensive mode of Phyre2. If you see multiple, high-confidence, largely nonoverlapping hits, this indicates that your protein contains multiple domains, each of which can be modeled confidently. In this case, you should consider trying intensive mode, as it will attempt to connect these individual domains together using *ab initio*–modeled connecting segments where required. Note that if you observe long (>100-residue) unmodeled segments, you can try intensive mode, but such regions are extremely unlikely to be well modeled owing to the limitations of *ab initio* protein modeling.

## Step 29: what if a template is found but not modeled?

If a structural template of interest is present lower down the list and thus has not been automatically modeled, you can generate a model using this template by using the one-to-one threading method. Clicking on the identifier in the 'Template' column of the detailed results table takes the user to the Phyre2 fold library, where the user can download the PDB coordinates of the template. Users may then upload their sequence and this template to the one-to-one threading method. Simply return to the Phyre2 home page, switch to 'expert mode' (in the top left of the home page once you are logged in to Phyre2) and then navigate to one-to-one threading.

● **TIMING**

In normal mode, job completion typically takes from 20 min to several hours depending on sequence length, number of detected homologous sequences and server load. Intensive mode jobs can take considerably longer (2–6 h) if there is a substantial portion of the sequence that cannot be modeled by known homologous structures, or if the protein is large (>700 amino acids).

## ANTICIPATED RESULTS

Once the job is completed, the user is notified by an e-mail that contains information on the confidence of the modeling, a link to a web page of results and an attachment containing the top scoring model in PDB format (Step 7). The web page of results contains the following:

- Facilities to interactively view all models in the browser using JSmol (Steps 12 and 26)
- Secondary structure, disorder and functional site predictions (Steps 17–19; **Fig. 5a**)
- Graphical summary table showing locations of matched homologs giving information on potential domain boundaries (Steps 20–22; **Fig. 5b**)
- The top 20 all-atom 3D models and their associated alignments and estimated confidence values (Steps 23–31; **Fig. 5c**)
- Ligand binding site predictions (Step 45) and transmembrane predictions (Step 46) where applicable

**COMPETING FINANCIAL INTERESTS**
The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Mukherjee, S., Szilagyi, A., Roy, A. & Zhang, Y. Genome-wide protein structure prediction. in *Multiscale Approaches to Protein Modeling* (ed. Kolinski, A.) Ch. 11, 255–279 (Springer, 2010).
2. Koonin, E.V., Wolf, Y.I. & Karev, G.P. The structure of the protein universe and genome evolution. *Nature* **420**, 218–223 (2002).
3. Kelley, L.A. & Sternberg, M.J.E. Protein structure prediction on the web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–371 (2009).
4. Mao, C. *et al.* Functional assignment of *Mycobacterium tuberculosis* proteome by genome-scale fold-recognition. *Tuberculosis* **1**, 93 (2013).
5. Lewis, T.E. *et al.* Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucl. Acids Res.* **41**, D499–D507 (2013).
6. Fucile, G. *et al.* ePlant and the 3D data display initiative: integrative systems biology on the world wide web. *PLoS ONE* **6**, e15237 (2010).
7. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—round X. *Proteins* **82** S2: 1–6 (2014).
8. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
9. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL Workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195–201 (2006).
10. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
11. Lobley, A., Sadowski, M.I. & Jones, D.T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*. **25**, 1761–1767 (2009).
12. Raman, S. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77** (suppl. 9), 89–99 (2009).
13. Källberg, M. *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **7**, 1511–1522 (2012).
14. 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).
15. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
16. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* **292**, 195–202 (1999).
17. Canutescu, A.A. & Dunbrack, R.L. Cyclic coordinate descent: a robotics algorithm for protein loop closure. *Protein Sci.* **12**, 963–972 (2003).
18. Jefferys, B.R., Kelley, L.A. & Sternberg, M.J. Protein folding requires crowd control in a simulated cell. *J. Mol. Biol.* **397**, 1329–1338 (2010).
19. Rotkiewicz, P. & Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29**, 1460–1465 (2008).
20. Wei, X. & Sahinidis, N.V. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics* **22**, 188–194 (2006).

21. Arjun, R., Lindahl, E. & Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinformatics* **13**, 224 (2012).
22. Davis, I.W. *et al*. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids Res* **35** (suppl. 2), W375–W383 (2007).
23. Schmidtke, P., Le Guilloux, V., Maupetit, J. & Tufféry, P. Fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic acids Res* **38** (suppl. 2), W582–W589 (2010).
24. Porter, C.T., Bartlett, G.J. & Thornton, J.M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic acids Res* **32** (suppl. 1), D129–D133 (2004).
25. Yates, C.M., Filippis, I., Kelley, L.A. & Sternberg, M.J. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol*. **426**, 2692–2701 (2014).
26. Capra, J.A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).
27. Higurashi, M., Ishida, T. & Kinoshita, K. PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res*. **37** (Database issue): D360–D364 (2009).
28. Marchler-Bauer, A. *et al*. CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* **41** (D1): D348–D352 (2013).
29. Adzhubei, I.A. *et al*. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
30. Sim, N. *et al*. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids Res*. **40** W1: W452–W457 (2012).
31. González-Pérez, A. & López-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet*. **88**, 440–449 (2011).
32. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. & Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol*. **337**, 635–645 (2004).
33. Siew, N., Elofsson, A., Rychlewski, L. & Fischer, D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*. **16**, 776–785 (2000).
34. Wass, M.N., Kelley, L.A. & Sternberg, M.J. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res*. **38**, W469–W473 (2010).
35. Jones, D.T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **3**, 538–544 (2007).