The Phyre2 web portal for protein modelling, prediction and analysis

Lawrence A Kelley[1*], Stefans Mezulis[1], Christopher M Yates[1,2], Mark N Wass [1,3] and Michael JE Sternberg[1]

[1] Structural Bioinformatics Group, Imperial College London, London SW7 2AZ, UK
[2] Present address:UCL Cancer Institute, 72 Huntley Street, London WC1E 6DD
[3] Present address: Centre for Molecular Processing, School of Biosciences, University of Kent, Kent CT2 7NH, UK.

* Corresponding author. Email l.a.kelley@imperial.ac.uk

Keywords: protein modelling, protein structure prediction, homology modelling, phyre2, poing, structural bioinformatics, nsSNPs, disease variants, protein modelling server. CASP.

EDITORIAL SUMMARY: Phyre2 is a web-based tool for predicting and analysing protein structure and function. Phyre2 uses advanced remote homology detection methods to build 3D models, predict ligand binding sites, and analyse amino-acid variants in a protein sequence.

TWEET: Using Phyre2 for protein structure and function prediction.

Abstract/Summary

Phyre2 is a suite of tools available on the web to predict and analyse 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 protocol. In this updated protocol, we describe Phyre2, which uses advanced remote homology detection methods to build 3D models, predict ligand binding sites, and analyse the effect of amino-acid variants (e.g. nsSNPs) for a user's protein sequence. Users are guided through results by a simple interface at a level of detail determined by them. This protocol will guide a user 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 difficult to model. The server is available at http://www.sbg.bio.ic.ac.uk/phyre2. A typical structure prediction will be returned between 30mins and 2 hours 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 seeking 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 modelled using such techniques[1]. The key principles on which such techniques work are a) that protein structure is more conserved in evolution than protein sequence and b) 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 the model 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 modelling, where a sequence is aligned to a sequence of known structure based on patterns of evolutionary variation. Template-based modelling encompasses the strategies that have been called homology modelling, comparative modelling and fold recognition. It is this technique that has become the most universally reliable and widely used by both the modelling and wider bioscience communities. The success of template-based modelling over other methods is due to three main factors: 1, the development of powerful statistical techniques to extract evolutionary relationships from homologous sequences; 2, the enormous growth in sequencing projects which provides the raw information; and 3, 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 to a large database of sequences, construct an evolutionary/statistical profile of that sequence, and 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 based on 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, for which we previously published a protocol[3] Although the original Phyre[3] and the new Phyre2 share the common aim of

protein modelling, the new Phyre2 system described in this updated protocol has been designed from scratch and shares no components with its predecessor. Phyre2 was launched in January 2011, but users have needed **[LK: We prefer this phrasing]** 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 serves approximately 40,000 unique users per year, processing approximately 700-1000 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 to other methods

There exist a number of other powerful structure prediction servers on the web. However, for the majority of modelling 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 on-line modelling tools.

Some of the most widely used web servers for protein modelling are Phyre2, i-TASSER[8], Swiss-Model[9], HHpred[10], PSI-Pred[11], Robetta[12] and Raptor[13]. In international blind trials of protein structure prediction methods (CASP)[7], it is observed that for the majority of modelling tasks, there is no significant difference in the accuracy of these methods. In extremely difficult modelling tasks, where remote homology is uncertain and where significant regions of a sequence cannot be matched to a known structure, i-TASSER[8] has shown a small but significant performance improvement over other methods. Phyre2 has been tested in CASP9, 10 and 11 experiments (results can be seen at: http://predictioncenter.org/index.cgi) **[AU: Is it possible to add a brief comment on accuracy to address Reviewer #2's comments in the manuscript, as detailed in your response to reviewers letter?]** To compare performance to other systems we consider fully automated systems for template-based modelling (TBM) and the average model quality (known as GDT_TS in CASP) they produce over the course of the CASP experiment (120 protein domain targets in CASP9, and 98 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 CASP9 Phyre2 ranked 6th out of approximately 55 unique groups. The 5 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 10th out of approximately 45 groups. Excluding i-TASSER (8% improvement) the remaining 8 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 2 extra residues being within 4.5Å of the native. CASP11 data for average model quality is not yet available from the prediction centre website. We

consider that the primary difference between these servers and Phyre2 is not in accuracy, but in ease-of-use by non-bioinformaticians.

Limitations

There are two principle 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 modelling will either be impossible or extremely unreliable. This reflects the wider on-going 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 is unable to accurately determine, beyond the estimated position of a sidechain, the wider structural impact of a point mutation. This means a user attempting to model several single position variants using Phyre2 will receive essentially identical models with a different sidechain at the position of the variant. No alterations of the backbone of the protein will generally be observed.

It is often the case that a user does 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 underway to add this functionality by using known multimeric structures as templates for complex building.

Finally it is important to understand the potential limitations of modelling multi-domain proteins using the 'intensive' mode of Phyre2 (described in stage 3b of the 'Modelling 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 multi-domain 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 where a homologous crystal structure of the globular/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 approximately 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 a) Backphyre to search a structure against a range of genomes, b) batch submission of a large number of protein sequences for modelling, c) one-to-one threading of

a user sequence onto a user structure, d) Phyrealarm for automatic weekly scans for proteins difficult to model and e) Phyre Investigator for in-depth analysis of model quality, function and the effects of mutations. First modelling of a single sequence will be discussed, followed by brief explanations of tools a) to e). The Procedure will deal mainly with a single query submission to Phyre2. The advanced facilities will not be detailed in the Procedure, with the exception of the use of Phyre Investigaor (optional Procedure Steps 35-39) as the results they produce and their interpretation largely follow that described for a single sequence.

Modelling a single sequence

The core method of Phyre2 for generating a 3D model of a protein sequence is composed of 4 underlying technical stages, described below and illustrated in Figure 1. There is also an optional 'intensive mode' which attempts to create a complete full-length model of a sequence through a combination of multiple template modelling and simplified *ab initio* folding simulation. This is described in stage 3b and illustrated in Figure 2. These stages and the 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' will refer 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 residue preferences at each position along its length. In order to construct an evolutionary profile, one needs to gather a large number of diverse yet true homologues. Diversity is key in order to create a statistically representative distribution of amino acid preferences at each position in the protein, whilst 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 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 homologues 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 (% of all true homologues detected) over PSI-Blast and more accurate alignments without sacrificing computational speed. HHblits is used to scan the query against a sequence database where no pair of sequences shares more than 20% identity, resulting in a sequence profile. In addition, the secondary structure of the query is predicted using PSI-Pred[16]. PSI-Pred is one of the most widely-used methods for secondary structure prediction and uses neural networks trained on protein sequence profiles to predict the presence of alpha helices, beta strands and coils with an average 3-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 pre-compiled 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 often containing insertions and deletions (indels) and without sidechains.

Stage 3: Loop modelling. 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 characterised by its sequence, geometry of flanking regions and distance between end points. For insertions, a sequence-profile search is performed using the missing inserted sequence to detect fragments with similar sequence composition and end-point distances, creating a short list (typically 100 members) of potentially useful fragments. Similarly for deletions, the sequence encompassing a window either side of the deletion is used. These fragments are fitted to the crude model using cyclic coordinate descent (CCD)[17], a robotic arm algorithm that attempts to fit the ends of the fragment to the crude model whilst minimising 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 on alignment interpretation in the Procedure.

Stage 3b: Multiple template modelling 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/domains are modelled by separate templates, or when there are no templates at all (*ab initio* modelling). 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 whilst maintaining high confidence as reported by HHsearch. These input models provide distance constraints between different pairs of residues. These restraints are modelled as linear inelastic springs. In Poing, restraints are added as the query protein is slowly synthesised from a virtual ribosome. Residues not constrained by input models are modelled *ab initio* by Poing's solvent bombardment model, predicted secondary structure springs and penalisation of steric clashes. 5-100 models are synthesised in this way depending on protein size (fewer for large proteins due to computational demand) and clustered to choose a final representative model. As this model is composed only of alpha carbons, its backbone is reconstructed using Pulchra[19].

Stage 4: Sidechain placement. Sidechain 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 sidechain rotamer library to place sidechains in their most probable rotamer whilst avoiding steric clashes. This technique is approximately 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 if 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 based on 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, whilst 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 **[AU: Edits correct?] [LK: Adjusted from steps 23-34 to steps 23-30 as the further steps do not apply to Backphyre output]**.

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 so 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 23-34 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 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 modelled **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. Based on statistics from 30,000 Phyre2 submissions over two months, on average more than 50% of all proteins submitted have had over 75%

of their length modelled with >90% confidence. Of the remaining 50% of submissions, 25% have had less than a quarter of their sequence modelled and 25% have between a quarter and three quarters confidently modelled. A failure to detect confident structural matches for significant regions of a query is typically caused by one of three factors: 1. A lack of a sufficient number and diversity of homologous sequences to the query to create a useful profile/HMM, 2. The evolutionary distance between the query and any known structure being too great to detect with the HMM-HMM matching method, or 3. The query adopting a novel fold not present in the current structural database.

Fortunately, both the protein sequence database and structure database are growing every week, meaning a currently undetectable homology may likely become detectable in the near future. For this reason the Phyrealarm service was developed. If a user query cannot be modelled confidently, the protein may be added to an automated scan of new structures and new sequence databases each week. Every week approximately 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 modelling pipeline and the user sent the results and links by email.

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 Phyre Investigator was developed. Any model produced by Phyre2 can be submitted to Phyre Investigator with one click from the results page for a range of analyses including:

- Model quality assessment by ProQ2[21]
- Alignment confidence from HHsearch[10]
- Clashes, Rotamers, and Ramachandran analysis by Molprobity[22]
- Pocket detection by fpocket2[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[28]

The Phyre Investigator interface (Figure 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 3 main sections from top to bottom: 1. the information box, 2. the structure view and analyses buttons and 3. the sequence view.

The structure view and analyses section is itself divided into 3 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 colour the structure in the JSMol view in accordance with the analysis chosen and display a colour-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.

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 modelled. 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 snaphot" 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 coloured bar, with tall, red bars being more favourable 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[29], SIFT[30] and Condel[31]. The SuSPect method is available as a standalone web server (http://www.sbg.bio.ic.ac.uk/suspect/), with more options for uploading sets of sequences, viewing pre-calculated results for the entire human proteome and more.

When using SuSPect through Phyre Investigator, it is important that your sequence is the wild-type. Submitting a mutant protein to Phyre2 and then Investigator will lead to misleading predictions from SuSPect. If the protein is human, pre-calculated scores will be returned. For non-human 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 a user registers 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 allowing 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 days. 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 depends on the facilities used, as follows:
  - *Standard protocol, single sequence modelling*: 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 email 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 for data format) into the form provided.

? TROUBLESHOOTING

5. Choose 'Normal' or 'Intensive' mode (see description of Stages 3 and 3b above 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, based on the critical point after Step 12 and Steps 20-22 whether it is worthwhile re-submitting 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 seconds. 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 email will be sent to the user and the page will be redirected to a generated results page.

Obtaining results

7. Upon job completion, an email 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 email and password if they have registered. Click on the link in the email. 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 days'. If the job is approaching expiry the text will be orange or red. If so, click on the 'Renew for 30 days' link to reset the expiry date for the results.
9. A button labelled '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 (see Steps 11-12 and Figure 4), sequence analysis (see Steps 13-16), secondary structure and disorder prediction (see Steps 17-19 and Figure 5a), domain analysis (see Steps 20-22 and Figure 5b) and detailed template information (see Steps 23-34 and Figure 5c, Figure 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 (Figure 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 email 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 visualisation 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.

B. Intensive mode results

i) 'Intensive' results will show a horizontal (possibly multi-line) coloured bar representing the user sequence together with a colour-coded confidence key. The colours indicate the predicted confidence of the model along the sequence. Regions modelled by the *ab initio* approach of Phyre2 are always coloured blue to indicate minimum confidence. Other colours 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 colour-coded confidence. Click the browser '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 <90%. Similarly it may be that the top model does not cover a significant 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 to see if 'intensive mode' may be valuable here. Phyre2 attempts to automatically determine whether other templates covering additional portions of your protein are available and 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 pre-filled web form where one click will add their sequence to the Phyrealarm system. As discussed in the Introduction, once 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 modelling job is automatically run and the user emailed the results. If this happens, the user would resume the protocol at Step 11.

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 non-redundant 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 homologues 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 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 potentially 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 (see Figure 5a), the position in the sequence is indicated in the top line. The sequence is represented on the next line with residues coloured 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 is 3-state: α-helix, β-strand or coil. Green helices represent α-helices, blue arrows indicate β-strands and faint lines indicate coil. The 'SS confidence' line indicates the confidence in the prediction from PSI-Pred, with red being high confidence and blue 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 modelling 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 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 homologues detectable in the sequence database (see Steps 13-16. If your sequence has very few homologues (something you can check by looking at the PSI-Blast results via the button near the top of the results page) then accuracy falls to approximately 65%. Also 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 Conserved Domain Database[28] (CDD) for features of interest. When detected, these are also highlighted as coloured dots at the appropriate positions in the sequence with a colour key at the bottom of this section (disorder confidence line). Click on a feature to be taken to more detail at the CDD web site.

Domain analysis

20. Click 'Show' next to the heading 'Domain analysis' on the main results page. This will open a scrollable table (Figure 5b) whose width represents the length of the user protein. Matches by Phyre2 to known structures are shown as coloured rows where the colours represent the confidence in the homology (red is high confidence while blue is low confidence).
21. At most, the top 20 high scoring matches are model built. Lower ranked hits are not modelled to conserve computing resources. (To model lower ranked hits see Troubleshooting for Step 29.) These have links in the centre of their aligned regions 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 take you 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 modelled. This enables you to see whether there are regions that cannot be modelled at all, or whether there are multiple templates covering different regions of your protein. This in turn is an indicator of the likely 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 (Figure 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 colour-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 take you to a new page containing detailed alignment information (Figure 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. S = bend. Identical residues in the alignment are highlighted with a grey 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.Beneath the image of the model is a link that launches the JSmol applet inside the browser to interactively inspect the model **[AU: there isn't an image of the model on Figure 6? Are you referring to the image shown on the main results page here, i.e. Figure 5c? If so we should move this bit earlier?] [LK: The image of a model is not shown in Figure 6, however on the page there will be an image of the model. So this does not refer to the earlier results. I have updated the text to reflect this.]**
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 homologues classed into 3 states. No symbol indicates unconserved, a thin grey 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 equivalenced to identical template residues in the generated alignment. Check the value of these two columns. If both values are coloured 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 modelled well but with greater expected deviation from native than a close homologue.

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 modelled at high accuracy (2-4Å rmsd 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 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 if 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’. In cases where 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 click this button.
33. You will be taken to a page explaining Phyre Investigator and a submit button. Click the submit 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 you submitted. You will see a message saying ‘Investigator running’. Processing typically takes 5-10 minutes. When the job is complete the message will change to a link saying ‘View investigator results’. If you have been waiting longer than 10 mins, refresh the page in your browser. Click this link and proceed with the optional steps 35-39.

Phyre Investigator

35. The Phyre Investigator interface is composed of several sections (Figure 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 interested or in loops 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 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 the 'Return to main results' in the top left corner of the screen.

Superposition of Models

40. At the bottom of the main table (Figure 5c) is a button entitled "Generate superposition of selected models". Beneath each template name (column 2) of the main table are two buttons. 1. A radio button allowing you to select one single master model on which other models will be superposed. 2. 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 4.5Å. Typically one would choose as the master either the top rank model, or a model judged on some previous background biological knowledge to be most interesting. Click on the radio button for a model 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 and the ticked templates will be superposed on the master model. This will then take you to a page with a large JSMol window displaying the superposition for interactive viewing, together with a range of descriptive help.
43. Rotate the superposition, looking for which regions of the models 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 modelling 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, explained in the text on the web page.
44. Press the back button in your browser to return to the main results page

Binding site prediction

45. If the top rank model is assigned a probability >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 was >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 displayed. A link is available to submit the model regardless, but this requires user intervention to do so. 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 homologues 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 and 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. **[AU: Please clarify which web page we are on here? From where is the transmembrane helix prediction accessed?] I've added text in Steps 45 and 46 to clarify that the user is on the 'main results page'.** 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 to handle long sequences?

There is currently a sequence length limit of 1200 amino acids. Work is underway to extend this limit. If the query exceeds this limit, it is advised that the query be submitted to the Conserved Domain Database[29] to determine likely domain boundaries. The query may then be chopped at these boundaries to ensure 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 likely 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 the 'intensive' mode of Phyre2. If you see multiple, high confidence, largely non-overlapping hits, this indicates that your protein contains multiple domains each of which can be modelled confidently. In this case, you should consider trying 'intensive' mode as it will attempt to connect these individual domains together using *ab initio* modelled connecting segments where required.

CAUTION

If you observe long (>100 residue) unmodelled segments, you can try 'intensive' mode, but such regions are extremely unlikely to be well modelled due to the limitations of *ab initio* protein modelling.

Step 29: What if a template is found but not modelled?

If a structural template of interest is present lower down the list and thus has not been automatically modelled, 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. The user 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 logged in to Phyre2) and navigate to One-to-one threading.

TIMING

In 'Normal' mode, job completion typically takes between 20 minutes to several hours depending on sequence length, number of detected homologous sequences and server load. 'Intensive' mode jobs can take considerably longer (2-6 hours) if there is a significant amount of the sequence that cannot be modelled by known homologous structures or the protein is large (>700 amino acids).

Anticipated Results

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

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

Acknowledgments

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) (LA Kelley: BB/J019240/1, M Wass: BB/F020481/1), the Medical Research Council (MRC) (C Yates: MRC Standard Research Student (DTA) G1000390-1/1) and the Engineering and Physical Sciences Research Council (EPSRC) (S Mezulis: EPSRC Standard Research Student (DTG) EP/K502856/1).

Author contribution

L.A.K. designed the Phyre2 system and wrote the paper. M.J.E.S. supervised the project. S.M. developed the multiple template modelling protocol. C.M.Y.

developed the SuSPect method. M.N.W. developed the 3D-Ligandsite web resource.

Competing financial interests

MJES is a Director and shareholder in Equinox Pharma Ltd which uses bioinformatics and chemoinformatics in drug discovery research and services.

REFERENCES

1. Srayanta Mukherjee, Andras Szilagyi, Ambrish Roy, Yang Zhang. Genome-wide protein structure prediction. Multiscale approaches to protein modeling: structure prediction, dynamics, thermodynamics and macromolecular assemblies, Chapter 11, Edited by Andrzej Kolinski, (Springer-London, 2010), P. 255-280.
2. Koonin, E.V. et al. The structure of the protein universe and genome evolution. *Nature* **420**, 218-223 (2002).
3. Kelley, L.A. and 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 TE, 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** (D1), 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(1)**: e15237 (2010).
7. Moult, J. et al. Critical assessment of methods of protein structure prediction (CASP)—round X. *Proteins* **82**.S2, 1-6 (2014).
8. Roy A. et al. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725-738 (2010)
9. Arnold K. et al. 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. et al. 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. *Nature protocols* **7,** 1511-1522 (2012).
14. Altschul, SF., 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, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **9**, 173-175 (2012).
16. Jones, DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* **292**, 195-202 (1999).
17. Canutescu, AA., and Dunbrack, RL. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Prot. Sci.* **12,** 963-972 (2003).

18. Jefferys, BR., et al. Protein folding requires crowd control in a simulated cell. *J. Mol. Biol.* **397,** 1329-1338 (2010).
19. Rotkiewicz, P., and Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29,** 1460-1465 (2008).
20. Wei, X. and Sahinidis, NV. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics* **22,** 188-194 (2006).
21. Arjun, R. et al. Improved model quality assessment using ProQ2. *BMC bioinformatics* **13** 224 (2012).
22. Davis, IW. 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. et al. Fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic acids Res* **38**.suppl 2, W582-W589 (2010).
24. Porter, CT. et al. 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, CM. et al. SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *J Mol Biol.* **426,** 2692-2701 (2014).
26. Capra, JA. and Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23,** 1875-1882 (2007).
27. Higurashi, M. et al. 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), D348-52 (2013).
29. Adzhubei, IA. et al. A method and server for predicting damaging missense mutations. *Nature 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. and 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 JJ, et al. 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, et al. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*. **16(9)**, 776-85 (2000).
34. Wass MN, et al. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res* **38**, W469-73 (2010).
35. Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **23(5)** 538-44 (2007).

Figure legends

Figure 1. Normal mode Phyre2 pipeline showing algorithmic stages. **[AU: Please add outline boxes to this figure to show the grouping of elements into stages 1, 2, 3 and 4 as described in Introduction]**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 PSI-pred 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 modelling): Insertions and deletions in these models are corrected by loop modelling. Stage 4 (Side chain placement): Finally amino acid side chains are added to generate the final Phyre2 model.

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 maximise 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 alpha helix or beta strand conformations, and prevention of steric clash. The new protein is 'synthesised' 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 sidechain addition using R3 **[AU: Could these stages be illustrated/labelled in the figure?]. [LK: There are really just two stages shown. I have added 'Backbone and sidechain addition' to the final flow chart arrow in the figure to differentiate it from the 'Poing' stage. I hope this is satisfactory.]**

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.

Figure 4. Example Phyre2 summary results page.
On the left is an image of a large all-beta structure. Clicking on the image will download a PDB formatted file containing this structure. On the right are various data regarding the model including: 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 modelled 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.

Figure 5. [AU: This figure needs a title, a sentence to describe the figure as a whole before describing individual panels]
Samples of the three main sections of a typical Phyre2 results page. The sections are labelled a-c and discussed below.

a. Example secondary structure and disorder prediction.
The query sequence is coloured as described in Step 17. Question marks indicate predicted disordered regions. Each type of prediction is associated with a rainbow colour-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 23-24, and 29-32. In this example, the rank 1 and 2 matches have confidence of 100% and sequence identities of 23 and 24% respectively.

Figure 6. Example alignment between user query sequence and known structure, as described in Steps 25-28. Sequence colouring is as described in Step 17. Identical residues between query and template have a grey background. Secondary structures (predicted and known) are displayed; in this case alpha helices. Colour-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.

Figure 1

Query sequence

HHblits

multiple sequence alignment

1

PSI-Pred

secondary structure prediction

Query hidden Markov model

FINAL MODEL

4

Add sidechains

HHsearch

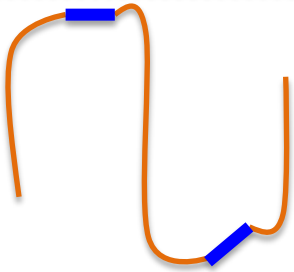
HMM database of known structures

2

Alignment between Query and Template

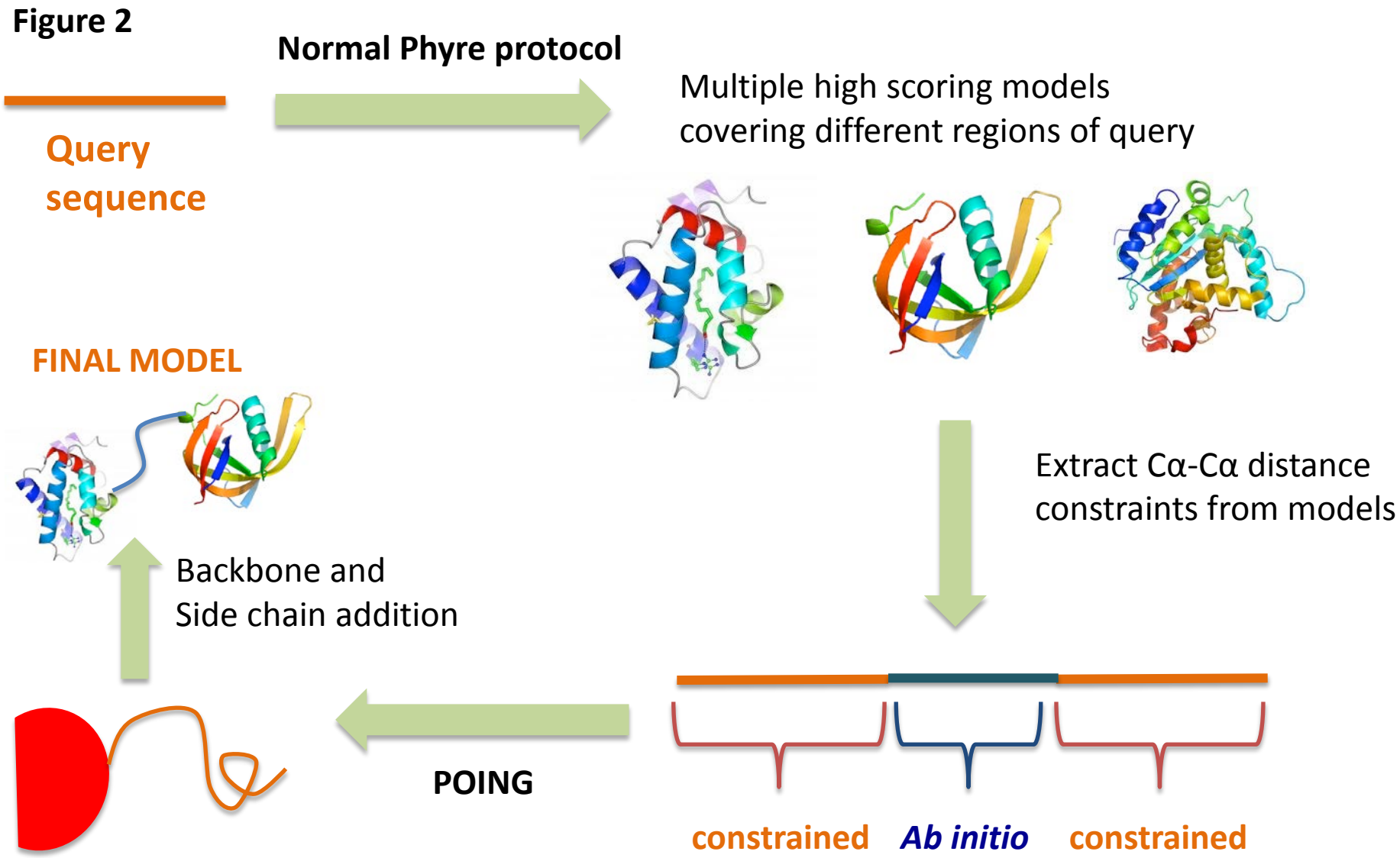
Crude backbone

Loop modelling



3

Figure 2



POING: Synthesise from virtual ribosome. Springs for constraints. *Ab initio* modelling of missing regions. Backbone and side chain reconstruction.

Figure 3

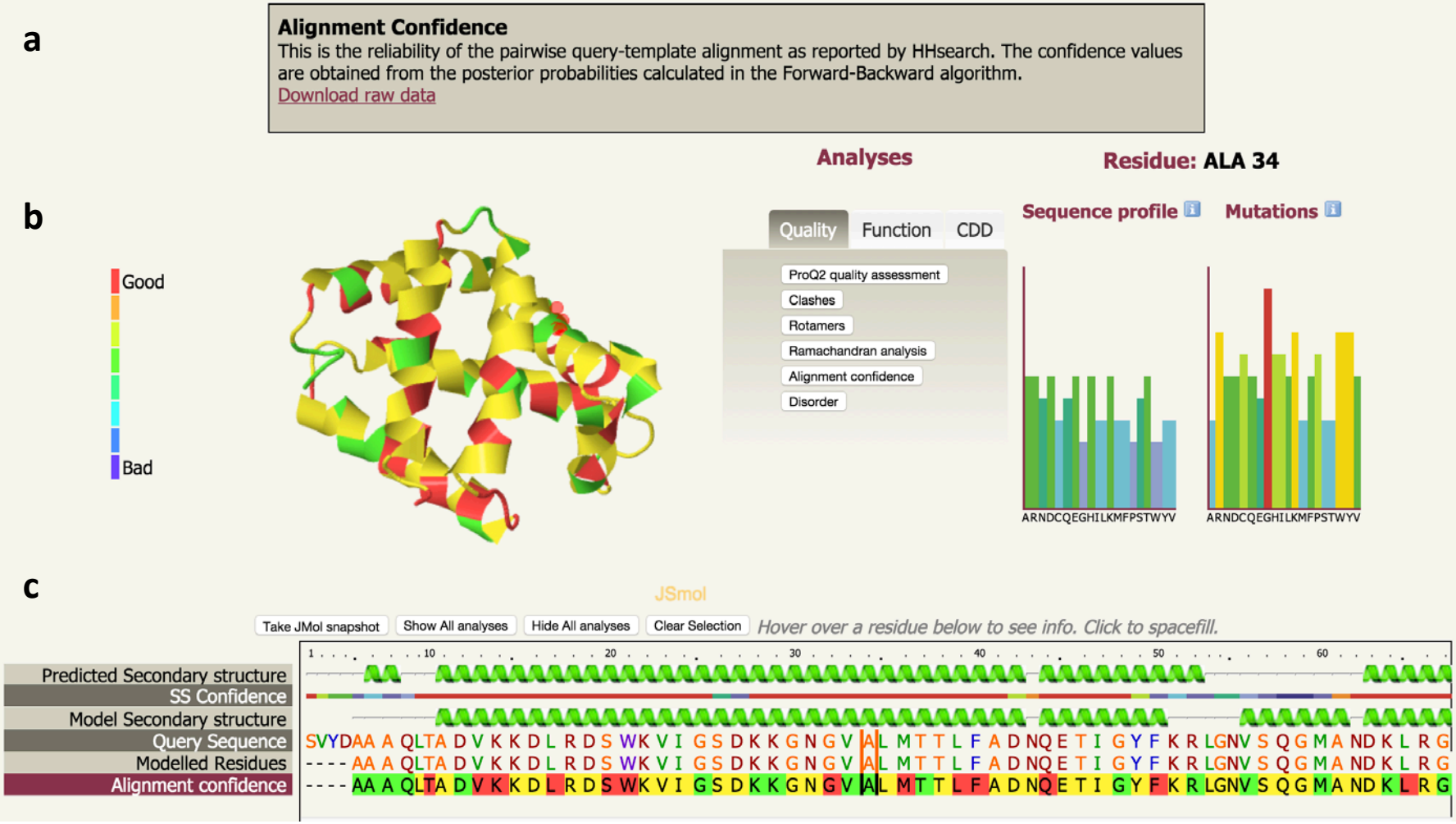

a
Alignment Confidence
This is the reliability of the pairwise query-template alignment as reported by HHsearch. The confidence values are obtained from the posterior probabilities calculated in the Forward-Backward algorithm.
Download raw data
b
Analyses
Residue: ALA 34
Good
Bad
Quality
Function
CDD
ProQ2 quality assessment
Clashes
Rotamers
Ramachandran analysis
Alignment confidence
Disorder
Sequence profile
Mutations
ARNDCQEGHILKMFPSTWYV
ARNDCQEGHILKMFPSTWYV
c
JSmol
Take JMol snapshot
Show All analyses
Hide All analyses
Clear Selection
Hover over a residue below to see info. Click to spacefill.
Predicted Secondary structure
SS Confidence
Model Secondary structure
Query Sequence
Modelled Residues
Alignment confidence

Figure 4

Top model

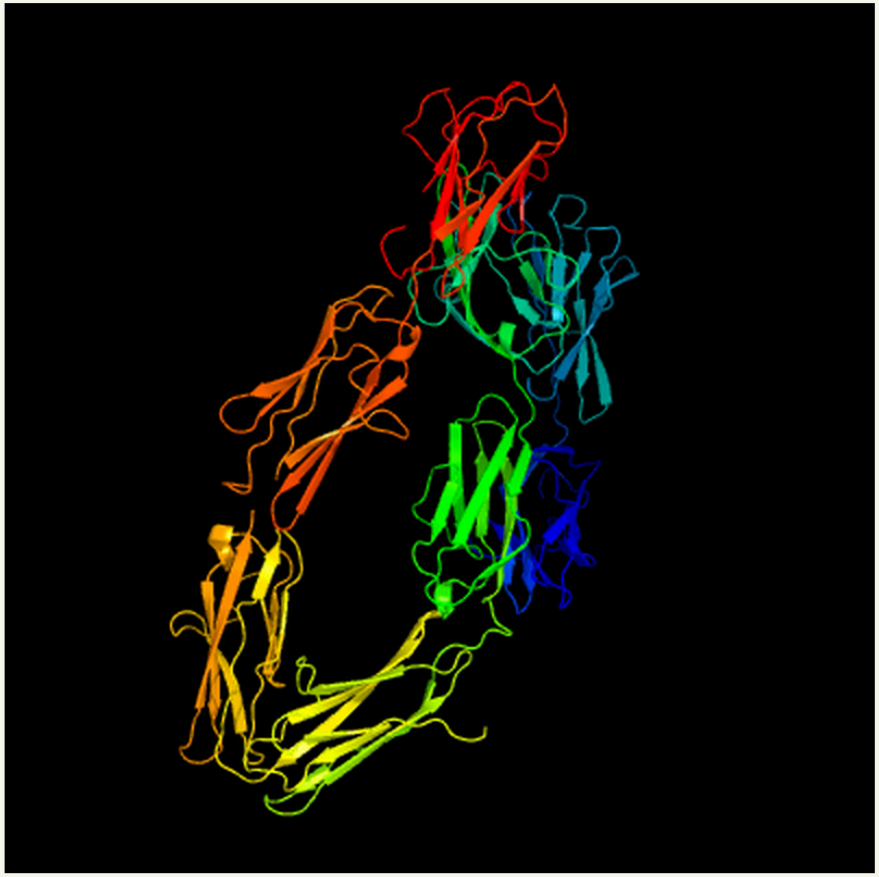


Image coloured by rainbow N → C terminus

Model dimensions (Å): **X**:66.519 **Y**:96.707 **Z**:120.895

Model (left) based on template c3dmkA_

Top template information

PDB header:cell adhesion
Chain: A: **PDB Molecule:**down syndrome cell adhesion molecule (dscam) isoform
PDBTitle: crystal structure of down syndrome cell adhesion molecule (dscam)2 isoform 1.30.30, n-terminal eight ig domains

Confidence and coverage

Confidence:	**100.0%**	Coverage:	**28%**

745 residues (28% of your sequence) have been modelled with 100.0% confidence by the single highest scoring template.

Additional confident templates have been detected (see Domain analysis) which cover other regions of your sequence.

1442 residues (55%) could be modelled at >90% confidence using multiple-templates.

You may wish to try resubmitting your sequence in "intensive" mode to model more of your sequence.

3D viewing

Interactive 3D view in JSmol

For other options to view your downloaded structure offline see the FAQ

Figure 5

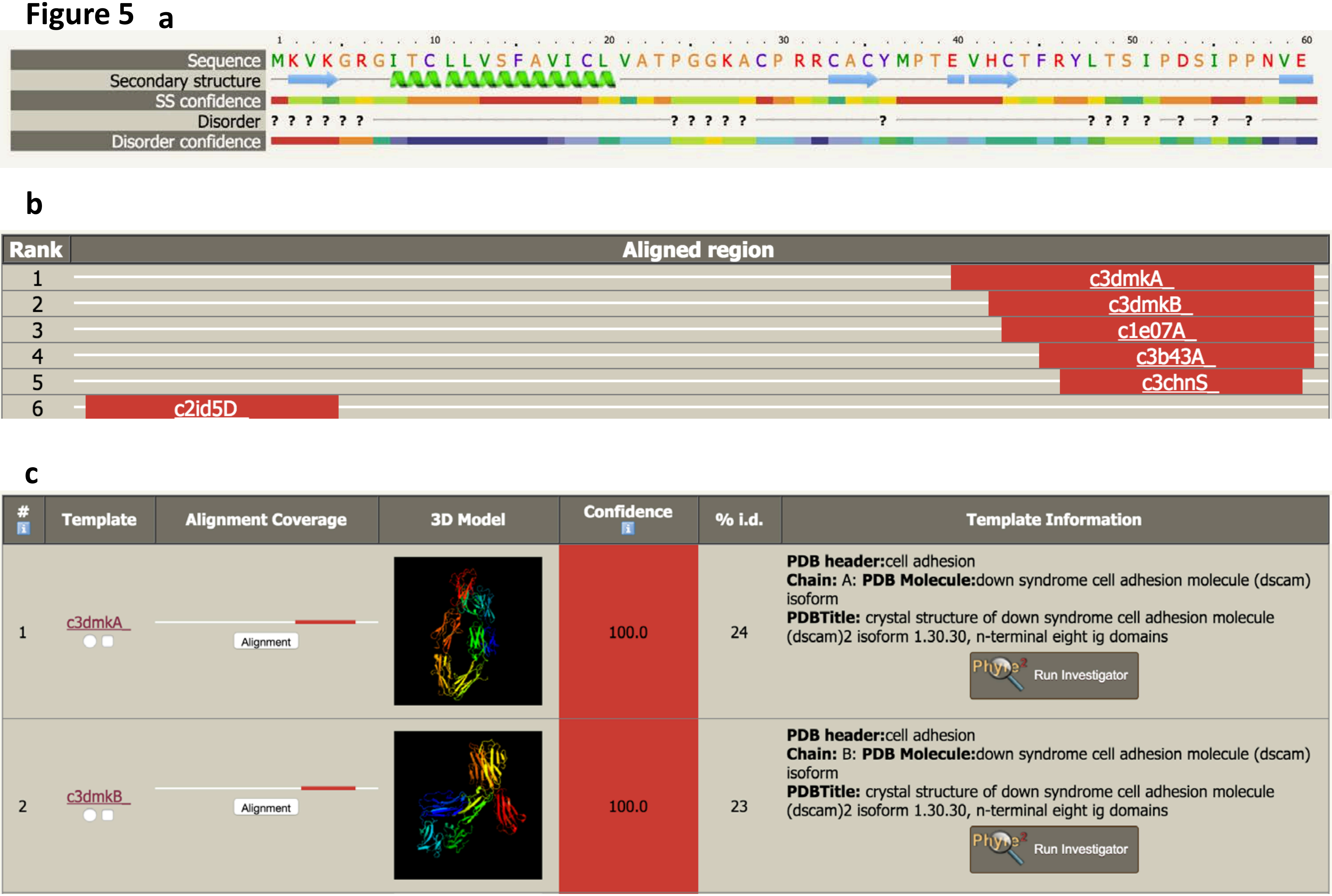
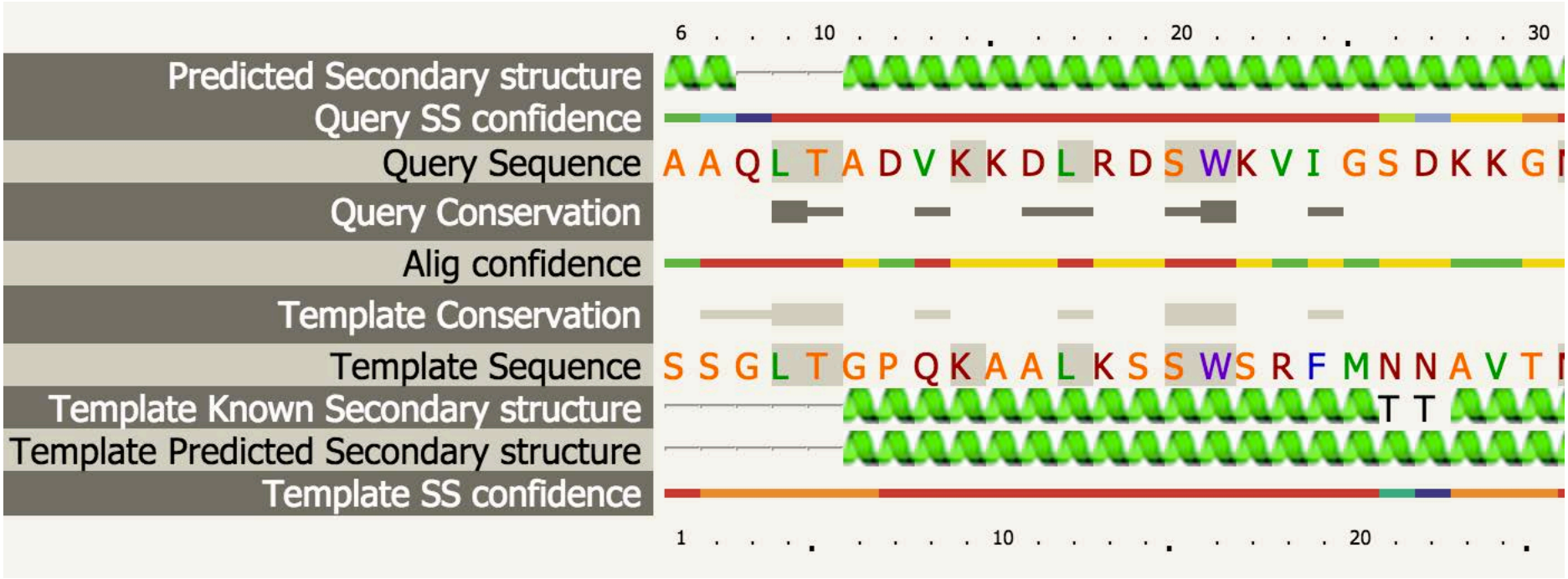

a
Sequence
Secondary structure
SS confidence
Disorder
Disorder confidence
M K V K G R G I T C L L V S F A V I C L V A T P G G K A C P R R C A C Y M P T E V H C T F R Y L T S I P D S I P P N V E
b
Rank
Aligned region
1 c3dmkA_
2 c3dmkB_
3 c1e07A_
4 c3b43A_
5 c3chnS_
6 c2id5D_
c
#
Template
Alignment Coverage
3D Model
Confidence
% i.d.
Template Information
1
c3dmkA_
Alignment
100.0
24
PDB header:cell adhesion
Chain: A: PDB Molecule:down syndrome cell adhesion molecule (dscam) isoform
PDBTitle: crystal structure of down syndrome cell adhesion molecule (dscam)2 isoform 1.30.30, n-terminal eight ig domains
Run Investigator
2
c3dmkB_
Alignment
100.0
23
PDB header:cell adhesion
Chain: B: PDB Molecule:down syndrome cell adhesion molecule (dscam) isoform
PDBTitle: crystal structure of down syndrome cell adhesion molecule (dscam)2 isoform 1.30.30, n-terminal eight ig domains
Run Investigator

Figure 6


Predicted Secondary structure
Query SS confidence
Query Sequence
Query Conservation
Alig confidence
Template Conservation
Template Sequence
Template Known Secondary structure
Template Predicted Secondary structure
Template SS confidence
6
10
20
30
A A Q L T A D V K K D L R D S W K V I G S D K K G
S S G L T G P Q K A A L K S S W S R F M N N A V T
T T
1
10
20