Protein structure prediction using Phyre² and understanding genetic variants

Prof Michael Sternberg
Dr Lawrence Kelley
Mr Stefans Mezulis
Dr Chris Yates

Imperial College
London
Timetable

- 10.00 – 11.00 Lecture
- 11.00 – 11.30 Tea/Coffee
- Courtyard, West Medical Building
- 11.30 – 1.00 Hands on workshop using Phyre:\(^2\)
- Computer Cluster 515, West Medical Building

Many thanks to Glasgow Polyomics and Amy Cattanach
Overview

- Methods
- Interpretation of results
- Extended functionality
- Proposed developments

- Publications:
  The Phyre2 web portal for protein modeling, prediction and analysis
  Kelley LA, Mezulis S, Yates CM, Wass MN & Sternberg MJES

Predict the 3D structure adopted by a user-supplied protein sequence
http://www.sbg.bio.ic.ac.uk/phyre2
How does Phyre2 work?

- “Normal” Mode
- “Intensive” Mode
- Advanced functions
Search the 30 million known sequences for homologues using PSI-Blast.
Capture the mutational propensities at each position in the protein

An evolutionary fingerprint
Phyre2

~ 100,000 known 3D structures

Extract sequence

HAPTLVRDC......
Phyre2

~ 100,000 known 3D structures

Extract sequence

HAPTLVRDC.......

PSI-Blast

Hidden Markov model
for sequence of KNOWN structure
Phyre2

~ 100,000 known 3D structures

~ 100,000 hidden Markov models
Phyre2

~ 100,000 known 3D structures

Hidden Markov Model Database of KNOWN STRUCTURES
Capture the mutational propensities at each position in the protein

An evolutionary fingerprint
Alignments of user sequence to known structures ranked by confidence.

**Phyre2**

**ARDLVIPMIYCGHGY** → **PSI-Blast** → **HMM**

**Hidden Markov Model DB of KNOWN STRUCTURES**

**HMM-HMM Matching (HHsearch, Soeding)**

**ARDL--VIPMIYCGHGY**

**AFDLCDLVIPV--CGMAY**

Sequence of known structure
Phyre2

**Sequence of known structure**

ARDLVIPMICYCGHGY

AFDLCDLIPV--CGMAY

**Hidden Markov Model DB of KNOWN STRUCTURES**

HMM-HMM Matching (HHsearch, Soeding)

3D-Model

ARDLVIPMICYCGHGY

AFDLCDLIPV--CGMAY

Sequence of known structure

PSI-Blast

HMM
Phyre2

ARDLVIPMIYCGHGY → PSI-Blast → HMM

Very powerful – able to reliably detect extremely remote homology

Routinely creates accurate models even when sequence identity is <15%

Hidden Markov Model DB of KNOWN STRUCTURES

HMM-HMM Matching (HHsearch, Soeding)

3D-Model

ARDL--VIPMIYCGHGY
AFDLCDLIPV--CGMAY

Sequence of known structure
From alignment to crude model

**Query (your sequence):**

```
ARDL--VIPMIYCGHGY
AFDLCDLIPV--CGMAY
```

**Known Structure**

**Known 3D Structure coordinates**
From alignment to crude model

Re-label the known structure according to the mapping from the alignment.

Insertion (handled by loop modelling)

Homology model
Loop modelling

ARDAKQH

d
Loop modelling
• Insertions and deletions relative to template modelled by a loop library up to 15 aa’s in length

• Short loops (<=5) good. Longer loops less trustworthy

• Be wary of basing any interpretation of the structural effects of point mutations
Sidechain modelling
Sidechain modelling
Sidechain modelling

Optimisation problem

• Fit most probable rotamer at each position
• According to given backbone angles
• Whilst avoiding clashes
• Sidechains will be modelled with ~80% accuracy IF.....the backbone is correct.
• Clashes *will* sometimes occur and if frequent, indicate probably a wrong alignment or poor template
• Analyse with Phyre Investigator
Example results

- Top model info
- Secondary structure/disorder
- Domain analysis
- Detailed template information
Example results

Model (left) based on template d1pmx

**Fold:** Insulin-like  
**Superfamily:** Insulin-like  
**Family:** Insulin-like

**Confidence and coverage**

Confidence: **100.0%**  
Coverage: **46%**

70 residues (46% of your sequence) have been modelled with 100.0% confidence by the single highest scoring template.

---

You may wish to submit your sequence to Phyrealarm. This will automatically scan your sequence every week for new potential templates as they appear in the Phyre2 library.

**Warning:** 54% of your sequence is predicted disordered. Disordered regions cannot be meaningfully predicted.

Interactive 3D view in Jmol
Example results

- Top model info
- Secondary structure/disorder
- Domain analysis
- Detailed template information
Example SS/disorder prediction
Secondary structure and disorder

- Based on neural networks trained on known structures.
- *Given a diverse set of homologous sequences*, expect ~75-80% accuracy.
- Few or no homologous sequences? Only 60-62% accuracy
Example results

- Top model info
- Secondary structure/disorder
- Domain analysis
- Detailed template information
Example domain analysis

<table>
<thead>
<tr>
<th>Rank</th>
<th>Aligned region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>c2vv5D_</td>
</tr>
<tr>
<td>2</td>
<td>d2vv5a1</td>
</tr>
<tr>
<td>3</td>
<td>d2vv5a3</td>
</tr>
<tr>
<td>5</td>
<td>d1nv9a_</td>
</tr>
<tr>
<td>6</td>
<td>cze6zA_</td>
</tr>
<tr>
<td>7</td>
<td>d1npna2</td>
</tr>
<tr>
<td>8</td>
<td>c2kvqG_</td>
</tr>
<tr>
<td>9</td>
<td>c2lvV_</td>
</tr>
<tr>
<td>10</td>
<td>d2hqa1</td>
</tr>
<tr>
<td>11</td>
<td>c3p8bB</td>
</tr>
<tr>
<td>12</td>
<td>c2zkr</td>
</tr>
</tbody>
</table>
Domain analysis

• Local hits to different templates indicate domain structure of your protein

• Multiple domains can be linked using ‘Intensive mode’
Example results

- Top model info
- Secondary structure/disorder
- Domain analysis
- Detailed template information
### Main results table

<table>
<thead>
<tr>
<th>#</th>
<th>Template</th>
<th>Alignment Coverage</th>
<th>3D Model</th>
<th>Confidence (% I.d.)</th>
<th>Template Information</th>
</tr>
</thead>
</table>
| 1 | d1b97a  |                    | ![3D Model](image1.png) | 99.7 | 17 | Fold: Globin-like  
Superfamily: Globin-like  
Family: Globins  
[View investigator results](#) |
| 2 | d2qka1  |                    | ![3D Model](image2.png) | 99.7 | 12 | Fold: Globin-like  
Superfamily: Globin-like  
Family: Globins  
[Run investigator](#) |
| 3 | c2wyoA  |                    | ![3D Model](image3.png) | 99.7 | 14 | PDB header: Oxygen transport  
Chain: A: PDB Molecule: globin-like protein;  
PDBTitle: high resolution 3d structure of c. elegans globin-like 2 protein glob-1  
[Run investigator](#) |

**Actual Model!**  
Not just a picture of the template – click to download model
Interpreting results

How accurate is my model?

• Simple question with a complicated answer!
• RMSD very commonly used, but often misleading
• Modelling community uses **TM score** for benchmarking: essentially the percentage of alpha carbons superposable on the answer within 3.5Å. Prediction of TM-score coming soon.
• Focused on the protein core, rather than loops and sidechains.
**MAIN POINT:** The confidence estimate provided by Phyre2 is NOT a direct indication of model quality – though it is related...

- It is a measure of the **likelihood of homology**
- Model quality can now be assessed using the new Phyre Investigator (more later)
- New measure of model quality coming soon..
Sequence identity and model accuracy

- High confidence (>90%) and High seq. id. (>35%): almost always very accurate: TM score > 0.7, RMSD 1-3Å
- High confidence (>90%) and low seq. id. (<30%) almost certainly the correct fold, accurate in the core (2-4Å) but may show substantial deviations in loops and non-core regions.
Interpreting results

100% confidence,
56% sequence identity, TM-score 0.9
Interpreting results

100% confidence,
24% sequence identity, TM-score 0.8
Interpreting results

Checklist

• Look at confidence
• Given multiple high confidence hits, look at % sequence identity
• Biological knowledge relating function of template to sequence of interest
• Structural superpositions to compare models – many similar models increase confidence
• Examine sequence alignment
<table>
<thead>
<tr>
<th>#</th>
<th>PDB ID</th>
<th>Chain</th>
<th>% Identity</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>d1pya</td>
<td>A</td>
<td>100.0</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>c124xA</td>
<td>A</td>
<td>100.0</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>d2rbp</td>
<td>A</td>
<td>99.9</td>
<td>34</td>
</tr>
</tbody>
</table>
Alignment view
Alignment view

<table>
<thead>
<tr>
<th>Predicted Secondary structure</th>
<th>Query SS confidence</th>
<th>Query Sequence</th>
<th>Template Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Conservation</td>
<td>Align confidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Template Conservation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Template Known Secondary structure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Template Predicted Secondary structure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Template SS confidence</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Alignment view:**

- **Predicted Secondary structure**
- **Query SS confidence**
- **Query Sequence**
- **Template Conservation**
- **Align confidence**
- **Template Sequence**
- **Template Known Secondary structure**
- **Template Predicted Secondary structure**
- **Template SS confidence**
Checklist

• Secondary structure matches
• Gaps in SS elements indicate potentially wrong alignment
• Active sites present in the Catalytic Site Atlas (CSA) for the template highlighted – look for identity or conservative mutations when transferring function
• Alignment confidence per residue
Mutations

• The STRUCTURAL effects of point mutations on structure will NOT be modelled accurately

Checklist

• Is it near the active site?
• Is it a change in the hydrophobic core?
• Is it near a known binding site? (can predict with e.g. 3DLigandSite)
• Phyre Investigator can help (see later)
Is my model good enough?

All depends on your purpose.

• Good enough for drug design? – probably if the sequence identity is very high (>50%)
• Sometimes good enough if far lower seq id but accurate around site of interest.
• High confidence but low seq i.d. still very likely correct fold, useful for a range of tasks.
How does Phyre2 work?

• “Normal” Mode
• “Intensive” Mode
• Advanced functions
• Individual domains in multi-dom proteins often modelled separately
• Regions with no detectable homology to known structure unmodelled
• Does not use multiple templates which, when combined could result in better coverage

Thus need a system to fold a protein without templates and combine templates when we have them
Poing – simplified folding model

structure simplification

Backbone
C-alpha

Protein backbone

Small hydrophilic sidechain

Large hydrophobic sidechain
**Phyre + Poing**

**ARNDLSLDLVCS........**

**PSI-Blast**

**HMM**

**FINAL MODEL**

**POING:** Synthesise from virtual ribosome. Springs for constraints. Ab initio modelling of missing regions.

**Hidden Markov Model DB of KNOWN STRUCTURES**

**HMM-HMM matching**

**Extract pairwise distance constraints**
Intensive mode

Image coloured by rainbow N → C terminus

Confidence Summary:
- 1: 200
- 201: 400
- 401: 600

Confidence Key:
- High (9)
- Low (0)

48% of residues modelled at >90% confidence (Details)

Publication-ready images:
- Hi-Res image (black background)
- Hi-Res image (white background)

Interactive 3D view in Jmol

Jmol Viewer
• Designed to handle multiple domains or proteins with substantial stretches of sequence without detectable homologous structures.
• POOR at ab initio regions
• GOOD at combining multiple templates covering different regions
• Relative domain orientation will NOT generally be correct if those domains come from different PDB’s with little structural overlap.
• Relative domain orientation will NOT generally be correct if those domains come from different PDB’s with little structural overlap.
“Intensive” does not always equal “Better”!

Checklist

- Always use normal mode first to understand what regions can be well modelled
- Multiple overlapping high confidence domains? Good, try intensive. Otherwise skip it.
- Danger of “spaghettification”
- Active development, new version ‘soon’
How does Phyre2 work?

- “Normal” Mode
- “Intensive” Mode
- Advanced functions
  - Phyre Investigator on web page including mutational analysis by SuSPect
  - Log in to use expert mode
• What parts of a model are reliable?
• What parts may be functionally important? (guide mutagenesis, understand mutants/SNPs)
• What residues are involved in interactions with other proteins?
Phyre Investigator

- Clashes
- Rotamer outliers
- Ramachandran outliers
- ProQ2 model quality assessment
- Alignment confidence (HHsearch)
- Conservation/evolutionary trace (Jenson-Shannon divergence – far faster and just as accurate as ET)
- Catalytic Site Atlas
- Disorder
- Pocket detection (Fpocket)
- Protein interface residues (PI-Site, ProtinDB)
- Conserved Domain Database ‘conserved features’ for NCBI-curated domains
Effect of Mutations?

- Will a SNP effect my protein’s function?
- New method: SuSPect by Chris Yates
- Integrated into Phyre Investigator
- Also standalone server

<table>
<thead>
<tr>
<th>#</th>
<th>Template</th>
<th>Alignment Coverage</th>
<th>3D Model</th>
<th>Confidence</th>
<th>% i.d.</th>
<th>Template Information</th>
</tr>
</thead>
</table>
| 1 | d1lm5a   | Alignment          | ![3D Model](image1) | 99.1       | 19     | Fold: beta-hairpin-alpha-hairpin-repeat  
Superfamily: Plakin repeat  
Family: Plakin repeat |
| 2 | d1lm7a   | Alignment          | ![3D Model](image2) | 98.8       | 16     | Fold: beta-hairpin-alpha-hairpin-repeat  
Superfamily: Plakin repeat  
Family: Plakin repeat |
Phyre Investigator

Beta Testing (What is this?)

Email Lawrence Kelley with problems or suggestions

Phyre Investigator allows you to interactively examine many features of your protein model including:

- Model quality predictions (clashes, rotamers etc)
- Conservation analysis
- Pocket detection and interface prediction
- Predicted effects of mutations using SuSPect

Processing typically takes 5-10 minutes

Warning: You should be using an up-to-date HTML5 compliant browser to view Phyre Investigator results

Confirm Phyre Investigator submission

Rank 3 model (template c3pt8B_)  
Job Description: globin_example
Phyre Investigator

**ProQ2 quality assessment**
ProQ2 is a model quality assessment algorithm that uses support vector machines to predict local as well as global quality of protein models. If you use this information, please cite: Improved model quality assessment using ProQ2. Arjun Ray, Erik Lindahl and Björn Wallner. BMC Bioinformatics 2012, 13:224.

Download raw data
SuSPect – Phenotypic effect of amino acid variants

Sequence conservation
- PSSM
- Pfam domain
- Jensen-Shannon entropy

Structural features
- Predicted solvent accessibility

Network features
- Protein-protein interaction (PP as domain centrality)
SuSpect – Results on non-training data (VariBench)

Specificity = \( \frac{TP}{TP + TN} \)

Sensitivity = \( \frac{TP}{TP + FP} \)

Benchmark consists of 20k SNPs (15k Neutral, 5k pathogenic)
Neonatal diabetes

- Arg 201 His in ATP-sensitive inward rectifier potassium channel 11 (Kir6.6)
- SuSPect gives score of 87/100 – high probability of disease associated
Phyre2 yields model which suggest structural basis for disease

Most variants predict to be disease associated

Arg 201 forms H-bond with main chain O

His in variant could not form similar interaction
Register and Log in to access Expert Mode
Advanced functions

- **PhyreAlarm** — automatically re-run tricky sequences every week
- **BackPhyre** — compare a structure to up to 30 genomes
- **One-To-One Threading** — use specific PDB for model building
- **Batch Jobs** — run many sequences at once
- **Job Manager** — keep track of your jobs and history
Advanced functions

- **PhyreAlarm** — automatically re-run tricky sequences every week
- **BackPhyre** — compare a structure to up to 30 genomes
- **One-To-One Threading** — use specific PDB for model building
- **Batch Jobs** — run many sequences at once
- **Job Manager** — keep track of your jobs and history
• Sometimes no confident homology detected
• Automatically try every week as new structures are deposited in the PDB
• Receive an email if hit found
• PhyreAlarm auto-suggested in cases where sequence has low coverage by confident hits
• Two clicks adds your sequence to the alarm queue
PhyreAlarm

User sequence
SVYDAAAQLTADVKKD......

Newly solved PDB
Structures added WEEKLY

Email results
New 3D model

Perform full Phyre modelling

Newly added structure HMMs

HMM-HMM matching

Confident hit?
Yes
No
Try again next week
Advanced functions

- **PhyreAlarm** — automatically re-run tricky sequences every week
- **BackPhyre** — compare a structure to up to 30 genomes
- **One-To-One Threading** — use specific PDB for model building
- **Batch Jobs** — run many sequences at once
- **Job Manager** — keep track of your jobs and history
• Does a structure I’m interested in exist in an organism?
• 30 searchable genomes to-date.
• Scan multiple genomes at a time. Quite fast.
• New version will allow users to upload their own genomes of interest.
**BackPhyre**

**User structure**

![User structure image]

<table>
<thead>
<tr>
<th>Rank</th>
<th>Hit</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gi...</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Gi..</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Gi..</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

SVYDAAAQLTADVKKDLRD SW
KVIGSDKKGNGVALMTTLF AD
NQETIGYFKRLGNV SQGM AN D
KLRGHSITL MYALQNFIDQLD
NPDSL DLVCS.......
Advanced functions

- **PhyreAlarm** — automatically re-run tricky sequences every week
- **BackPhyre** — compare a structure to up to 30 genomes
- **One-To-One Threading** — use specific PDB for model building
- **Batch Jobs** — run many sequences at once
- **Job Manager** — keep track of your jobs and history
• Useful if you:
  a) Know a better template than found by Phyre2
  b) Have your own structure not yet in the PDB
  c) Model a a lower-ranked (>20) template
  d) Want more expert control over alignment options: local/global, secondary structure weight etc.
One to one threading

User structure

User sequence
KLRGHSITLMYALQN NPDSLDLVCS........

HMM of User structure

HMM-HMM matching

Final model

SVYDAAAQLTADVKK DLRDSWDLVCS........

HMM of user sequence
Future
PhyreStorm

- **Searching Topology with Rapid Matching**
- Structural search and alignment of the **entire** PDB in under 1 minute.
- Go directly from a Phyre2 model and find all other similar structures rapidly.
- Beta released
PhaserPhyre

Molecular replacement tool: PhaserPhyre

Model d3sda.h.1

Chopper

Info

Ensemble of 19 models (>0.5 TM to master)

Fraction of model displayed: 95%
Fraction of Query modelled: 95%
Number of residues: (138/146)

Download model trimmed by Ensemble at 100 cut-off
PhyreRisk (with Prof R Houlston ICR)

Integrate disease networks, SNPs, GWAS, protein structure and complexes
PhyreRisk

PhyreRisk is a tool designed to analyze and predict the impact of single amino acid variants (SAVs) in proteins. It offers various features including:

- Structural explanation of a SAV affecting the interface
- Network with structures & interfaces highlighted
- Phenotype predictions (from SuSPect)
- Structure of human proteome - experimental and Phyre2
- Analysis of SAV susceptibility
- Prioritisation of disease genes via network clustering

Examples of PhyreRisk information
Protein structure prediction using Phyre² and understanding genetic variants.

Prof. Michael Sternberg
Dr. Lawrence Kelley
Mr. Stefans Mezulis
Dr Chris Yates

Imperial College
London

BBSRC
20 Years of Pioneering
Great British Bioscience

MRC
Medical Research Council

Wellcome Trust
• 10.00 – 11.00 Lecture
• 11.00 – 11.30 Tea/Coffee
• Courtyard, West Medical Building
• 11.30 – 1.00 Hands on workshop using Phyre²
• Computer Cluster 515, West Medical Building

Many thanks to Glasgow Polyomics and Amy Cattanach