

AMPLE

Ab Initio Modelling of Proteins for Molecular Replacement

Jens Thomas

Institute of Integrative Biology

December 2015



UNIVERSITY OF
LIVERPOOL

Outline

- What is AMPLE?
- How it works
- When it works
- Why it works
- How to use it
- What we're doing now

AMPLE

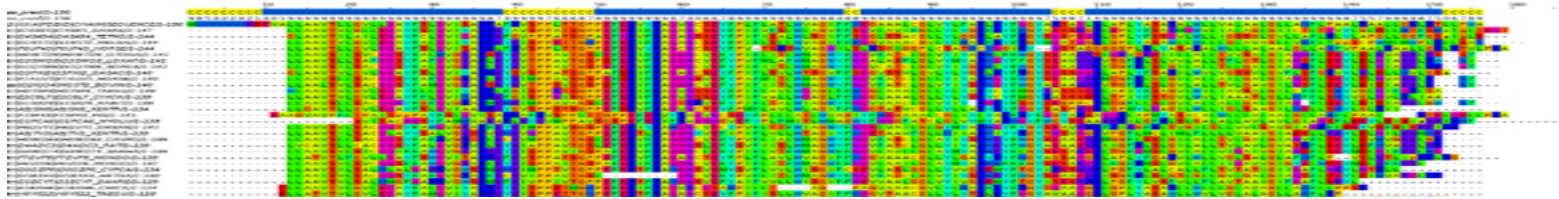
Ab initio **M**odelling of **P**roteins for mo**L**Ecular replacement

- Joint development by the University of Liverpool and CCP4

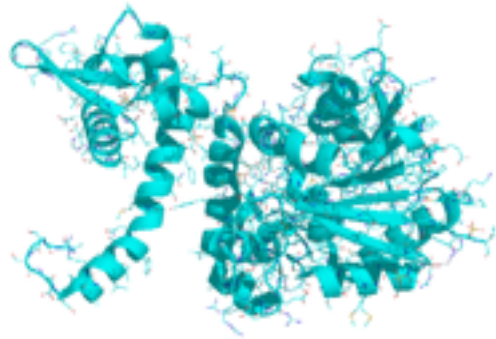


- AMPLE is a comprehensive project to assess the suitability of unconventional models in molecular replacement
- An additional goal of the project is to make AMPLE into an automated software tool that can be made generally available to potential users through CCP4

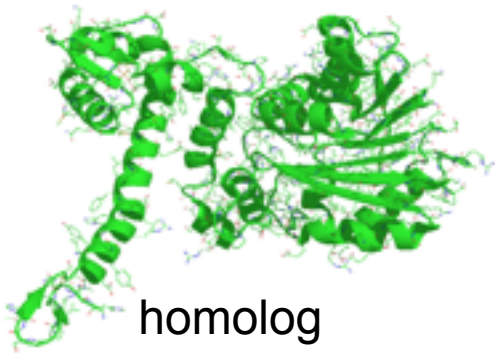
Molecular Replacement



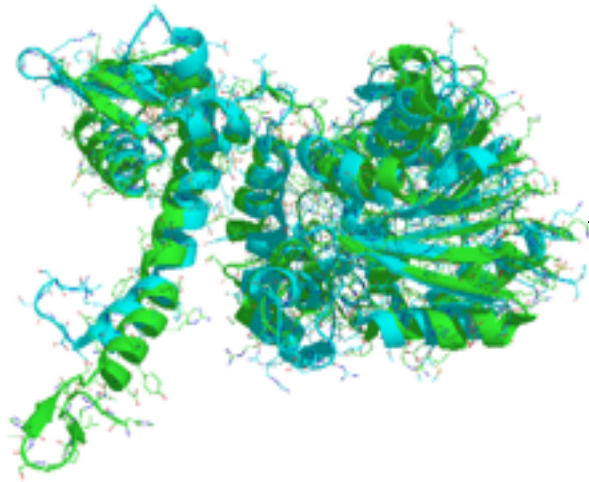
Sequence search for a suitable homolog



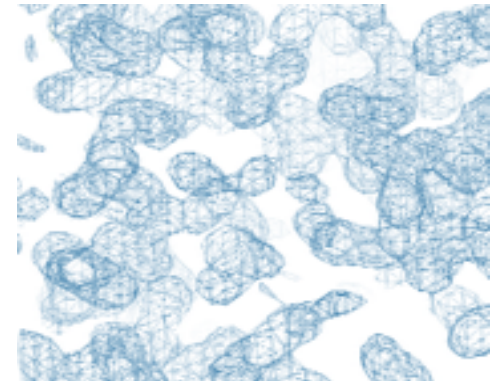
target



homolog



- Phases from homolog
- Intensities/positions from target



electron density map

Replacing Molecular Replacement

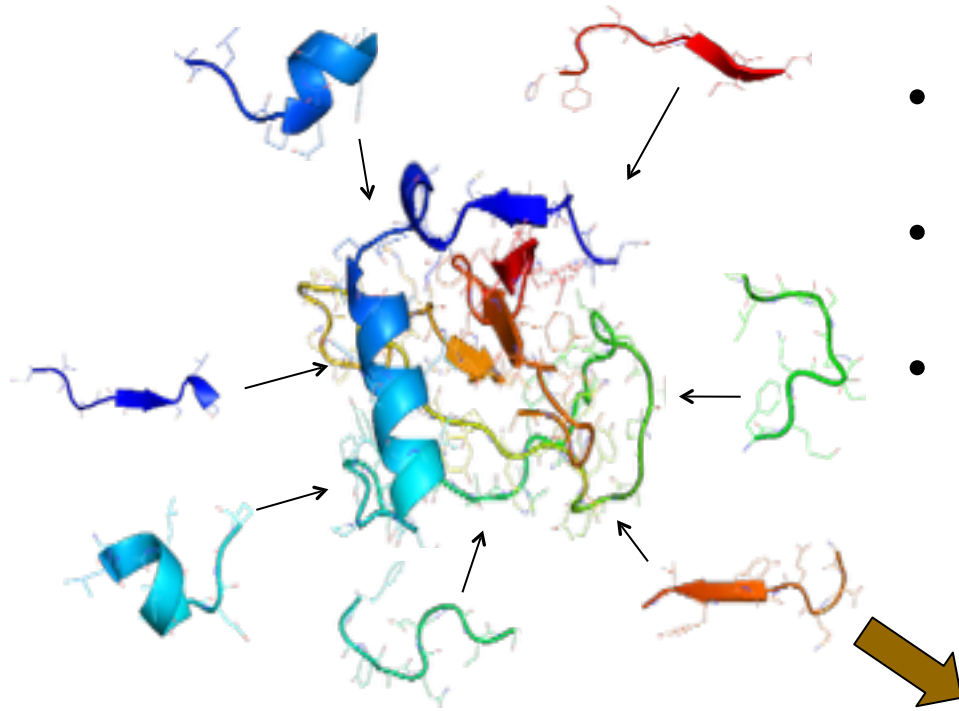
- Globular proteins with new folds or distant homologies
- Coiled-coils (small changes in sequence or crystallisation conditions can lead to different polymorphs crystallising)
- Helical membrane proteins
- NMR structures as search models
- Not sure exactly what is in the crystal



What to do?

- **AMPLE**
- **ARCIMBOLDO**

Ab Initio Modelling



- Create database of 3- and 9-residue fragments.
- Start from linear chain and splice in fragment geometry.
- Randomly move the fragments under simplified forcefield
- Score the resulting structures.

- 1000's of “decoys” created.
- Decoys are clustered and centroid representatives of largest cluster are considered candidate fold predictions
- Refinement under a more realistic physics-based force field



Decoys vs. final models

- for normal usage > 20,000 decoys are recommended
- refinement stage can take weeks on a supercomputer

This is beyond the capabilities of most crystallography groups.

But:

- creating 1000 decoys ~ 1/2 day on a single CPU with ROSETTA
- QUARK modelling is available online and takes ~ 1 day

Ab Initio Models and Molecular Replacement

Simple *ab initio* modelling

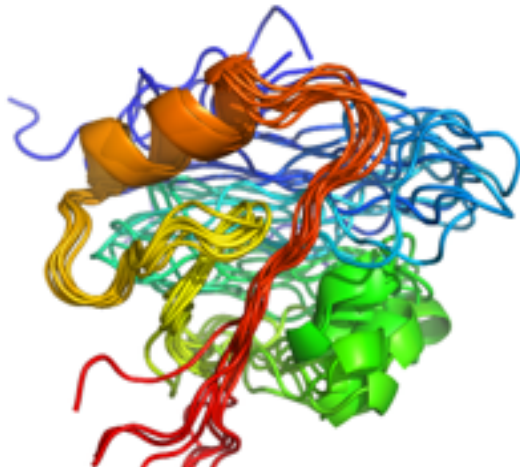
Clusters of similar structures

Within/between clusters, similarity indicates accuracy $\therefore$ trim inaccurate regions leaving more reliable **core**

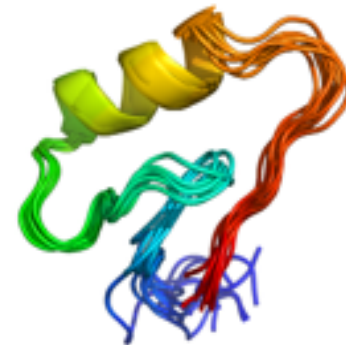
Molecular Replacement

Works well with superposed **ensembles** approximating the target

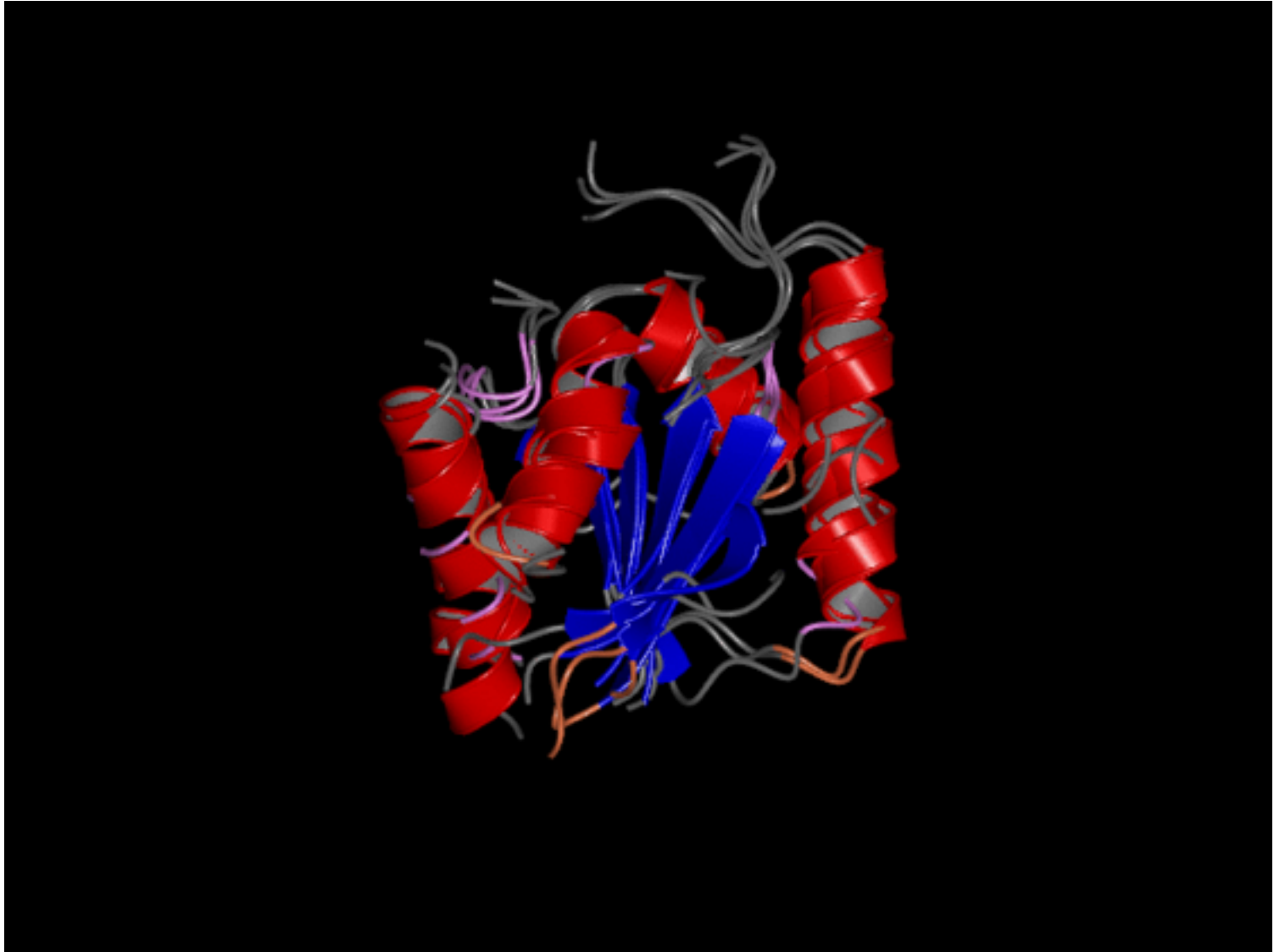
May only require a **partial** model



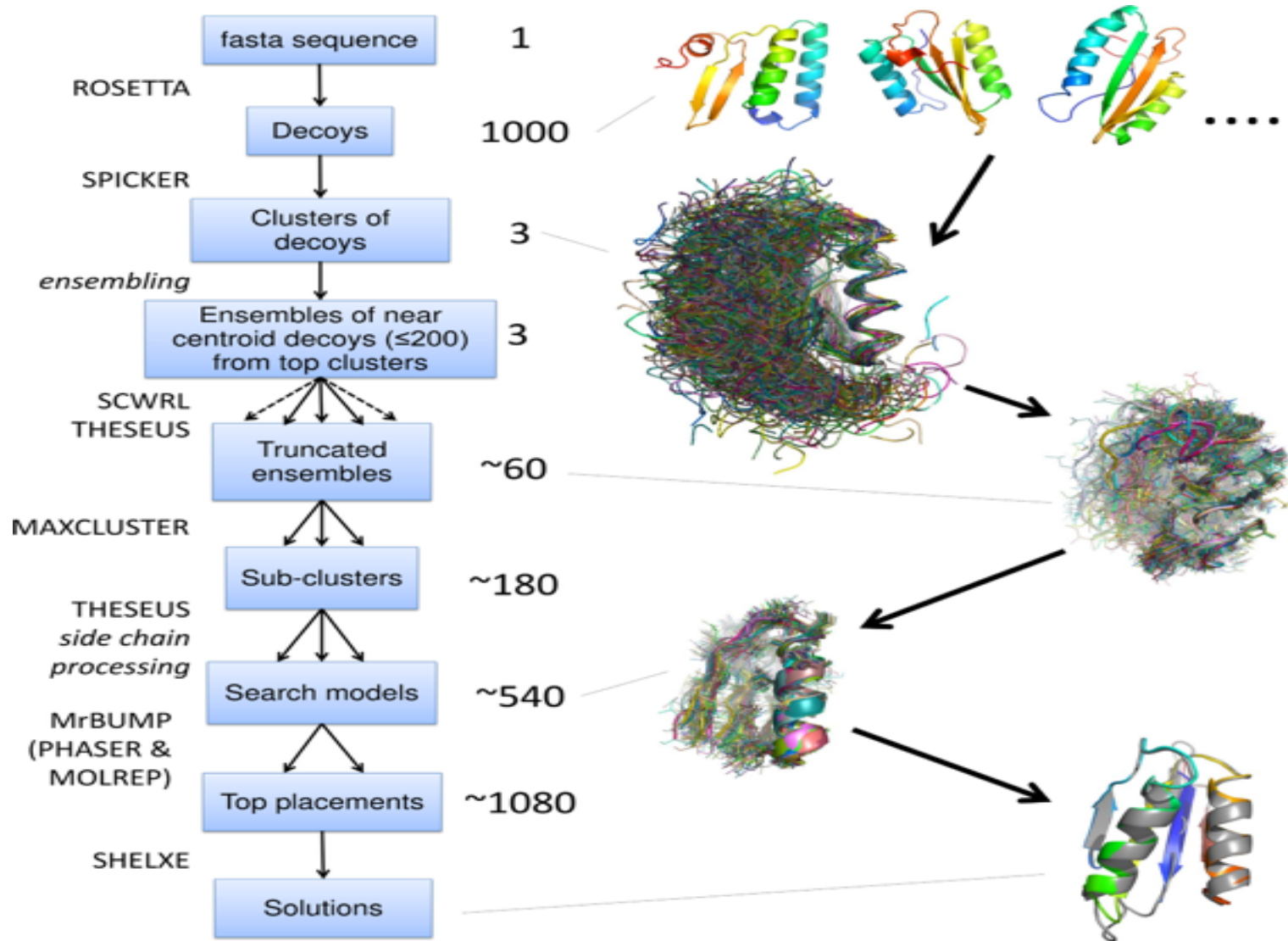
→
trim to accurate core



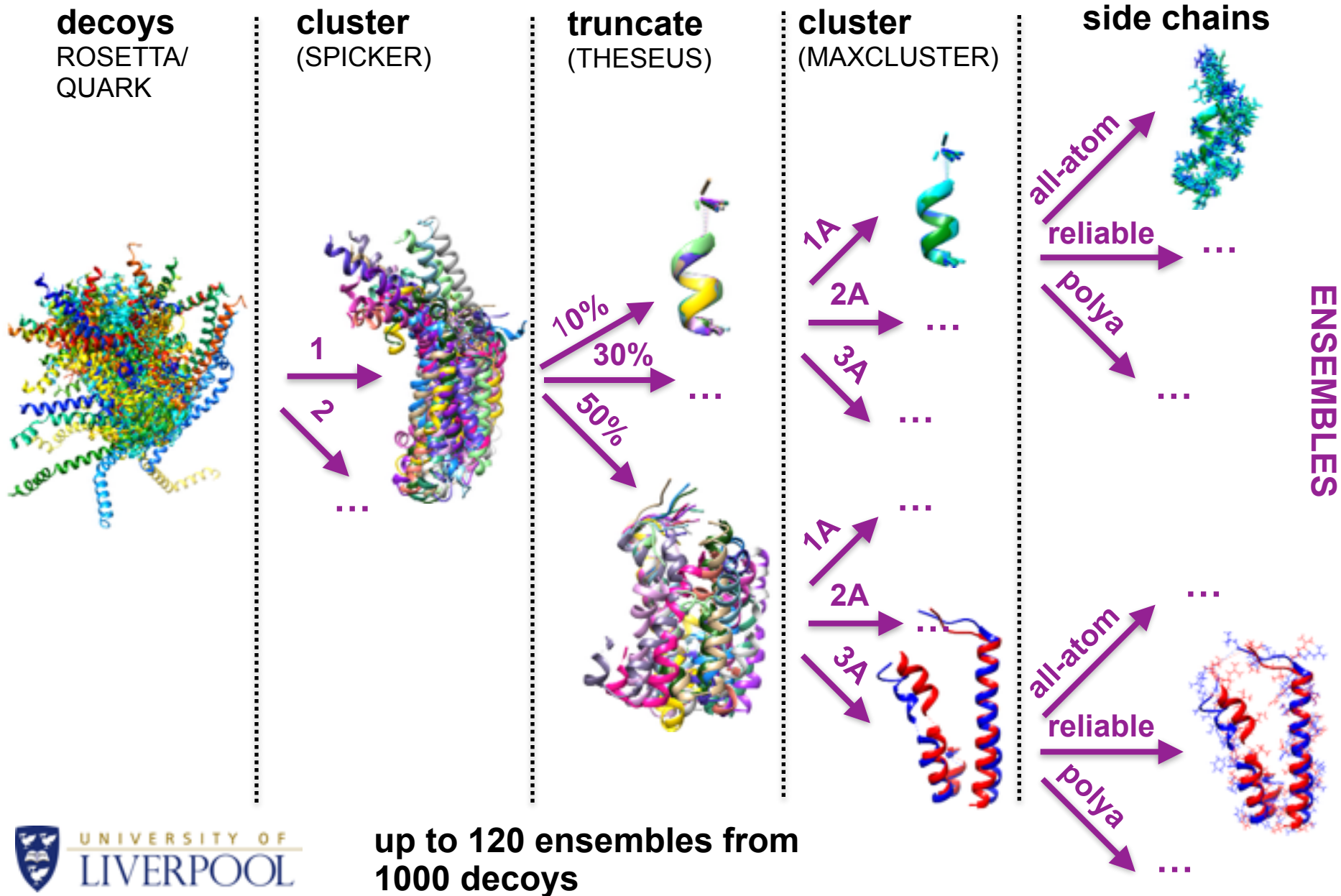
Truncation in action



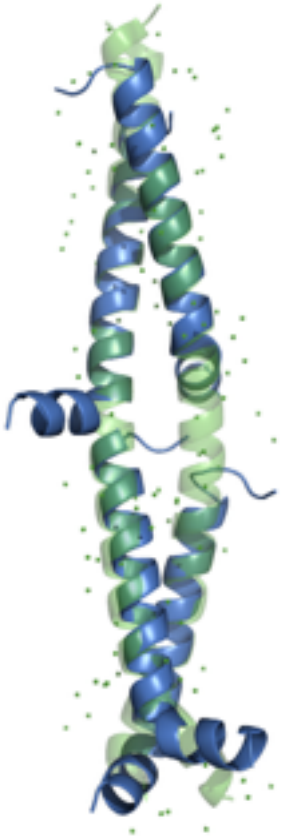
The AMPLE pipeline



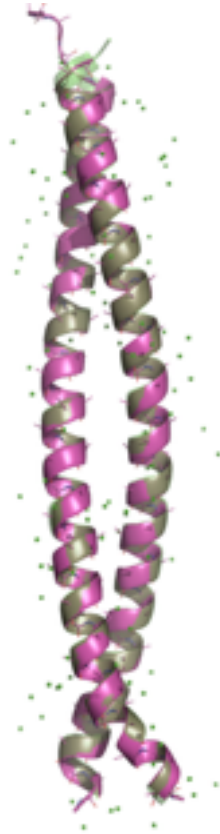
Pathways to ensembles



MRBUMP MR pipeline



PHASER: places search model



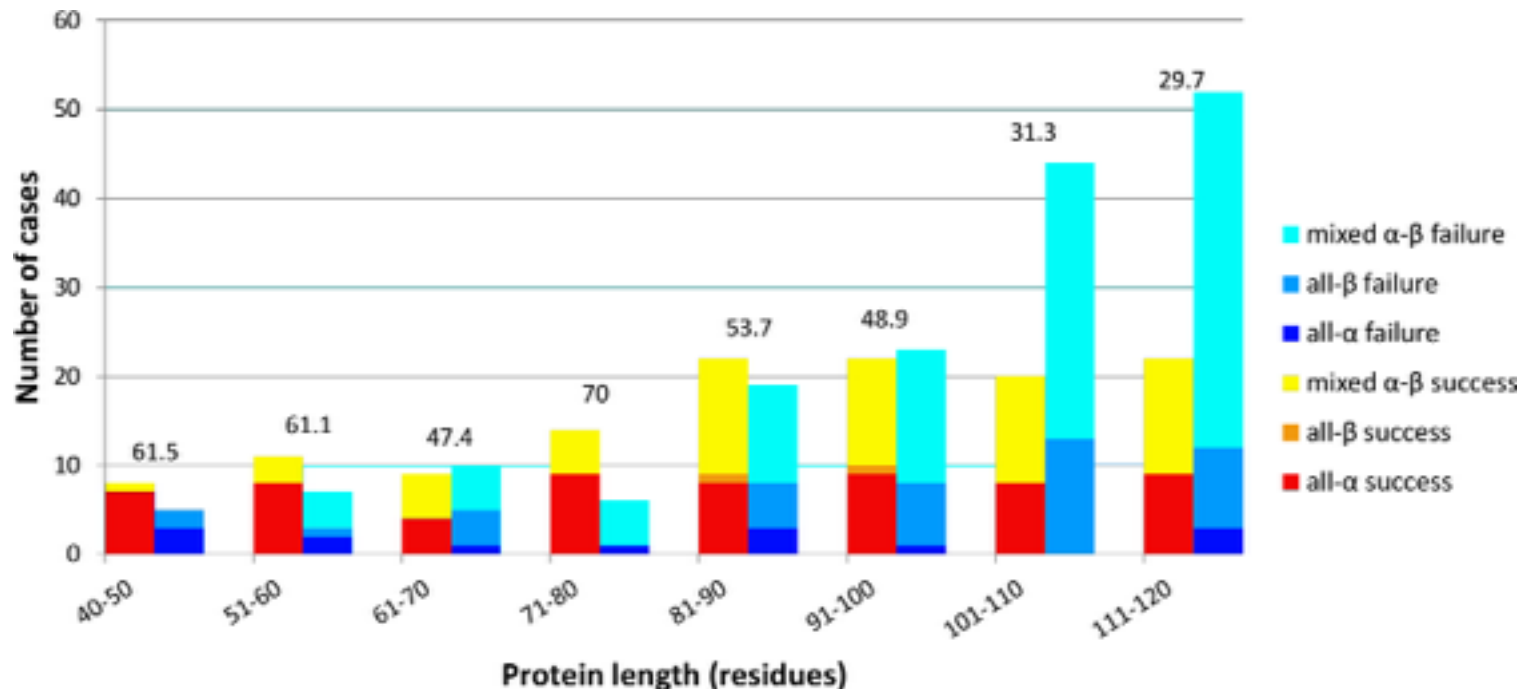
SHELXE: density modification and tracing of the main chain



ArpWARP/BUCCANEER completes the rebuild

Initial AMPLE success

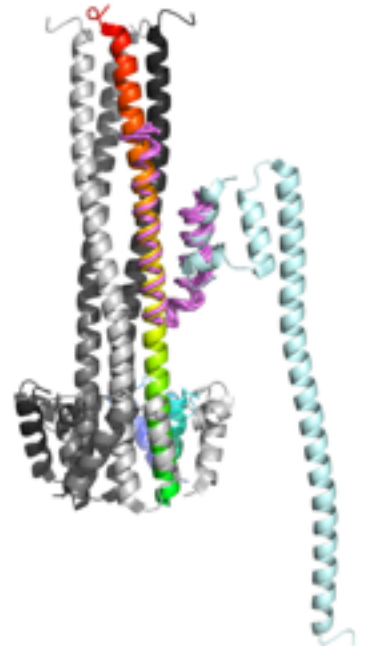
- Test set of 295 small globular proteins (40-120 residues) from the PDB, resolution of 2.2 Å or better
- Single molecule in the asymmetric unit
- Mixture of all- α , all- β and mixed α - β secondary structure



Initial success rate: all- α : 80%, all- β : 2%, mixed α - β : 37%

Coiled-coils

- Among the most ubiquitous and functionally important class of proteins
- 2-5 amphipathic helices wound around each other in a supercoil with sequence repeats of 7 or 11 residues
- Notoriously difficult to solve with MR techniques
- But, are exclusively α -helices
- Selected 94 cases from the PDB:
 - $12 < \text{residues} < 253$
 - $0.9 < \text{resolution } \text{\AA} < 2.91$

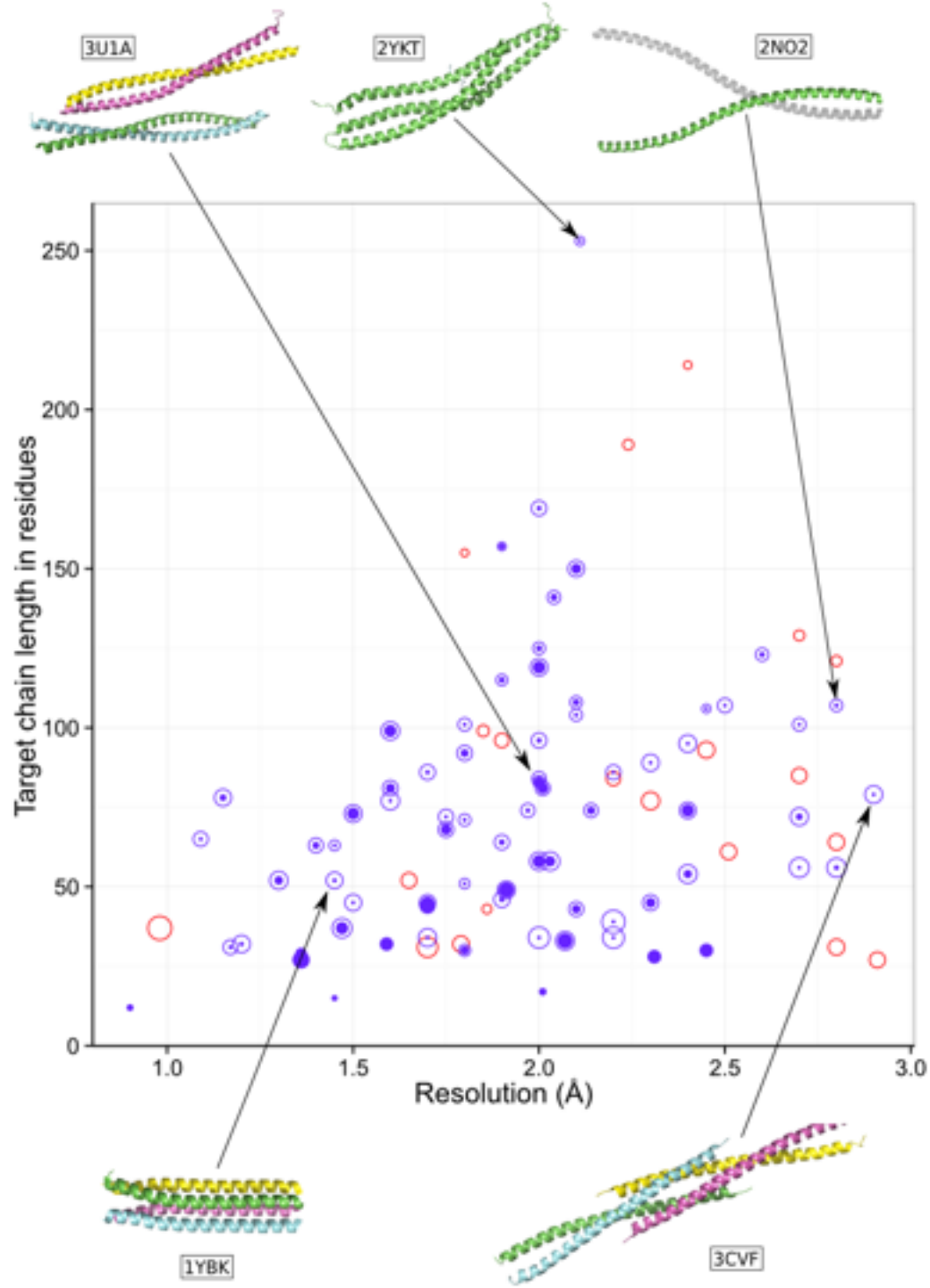


Coiled-coils

Solved more than **78%** with an average CPU time of less than 24 hours.

Successes included:

- **3U1A**: 334 residues
- **3CVF**: resolution of 2.8Å
- **2NO2**: a domain of Huntingtin-interacting protein 1, that contains a long, unconventional coiled-coil-like assembly originally phased experimentally using MAD
- **1YBK** right-handed coiled coil

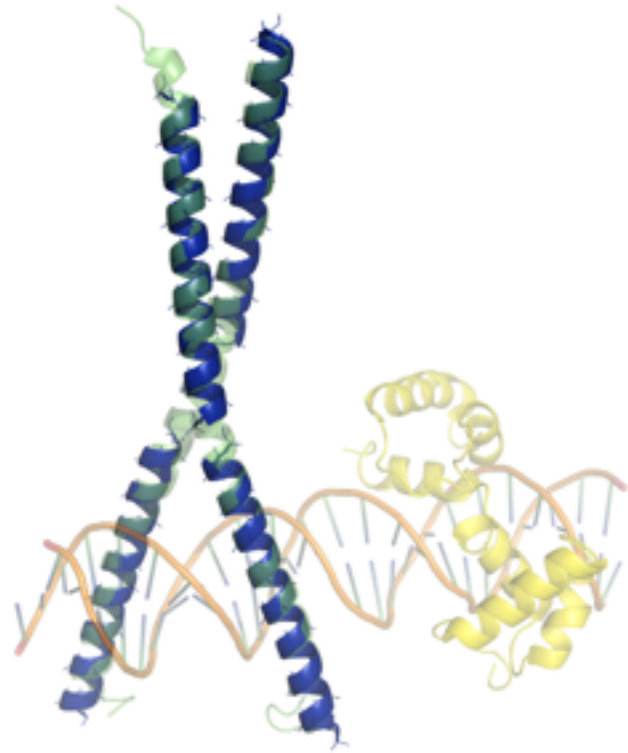


Complex Ambitions

Coiled-coils are often involved in complexes - could AMPLE's ability to solve coiled-coils extend to complexes?



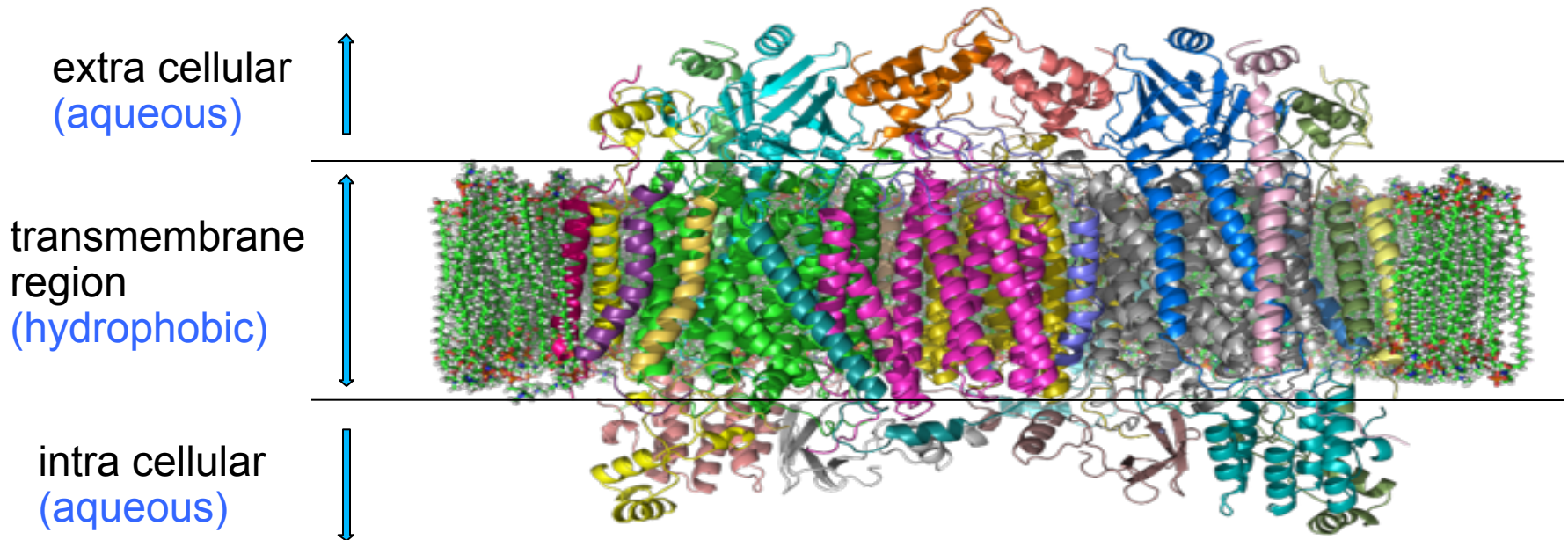
- **1X79**: three chains, 322 residues, ~70% coiled-coil @ 2.41Å



- **1H8A**: MyB/DNA complex of 278 residues ~ 60% coiled-coil @ 2.23Å.

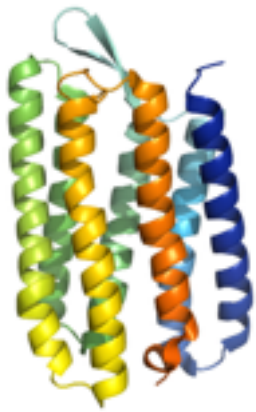
Transmembrane Proteins

- Experimentally very difficult to work with/crystallise
- Represent ~30% of all proteins
- Make up < 3% of structures in the Protein Data Bank
- Presence in the membrane constrains their shape, so they can be easier to model
- MR with *ab initio* transmembrane models hasn't been tried yet

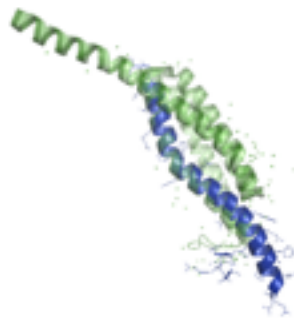


Transmembrane Results

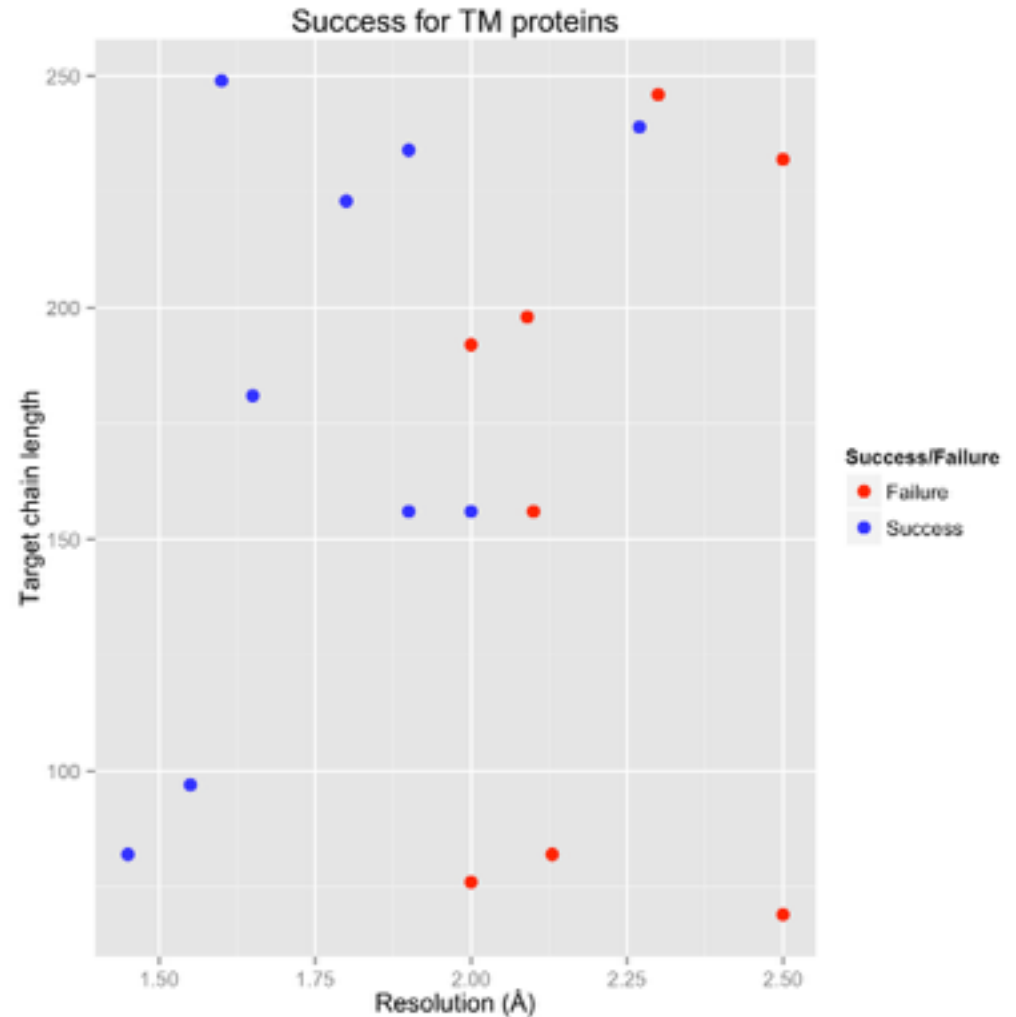
- Selected 18 alpha-helical transmembrane proteins:
 - 23 – 249 residues
 - 1.45 – 2.5Å resolution
- 9 clear successes
- Successes extended to 249 residues (3HAP).



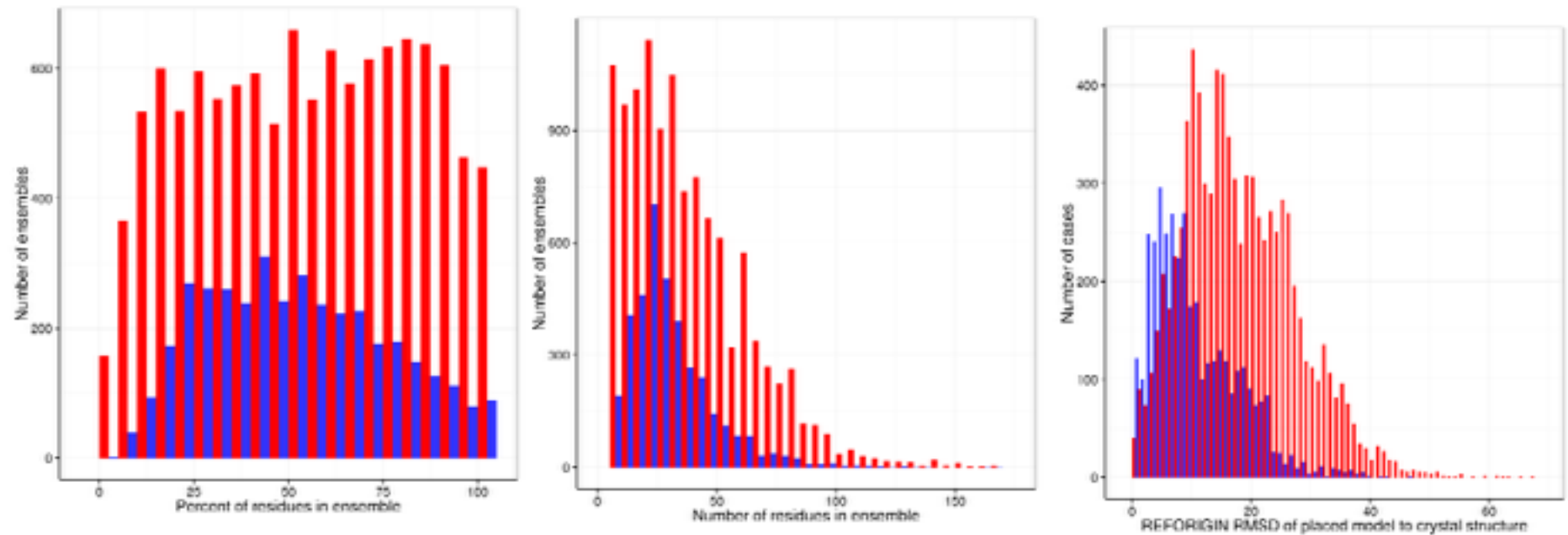
3HAP



3PCV

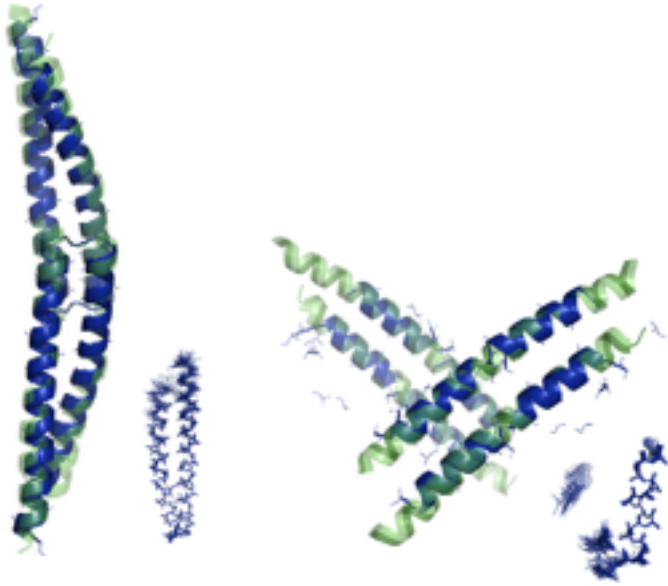


What is solving these structures?



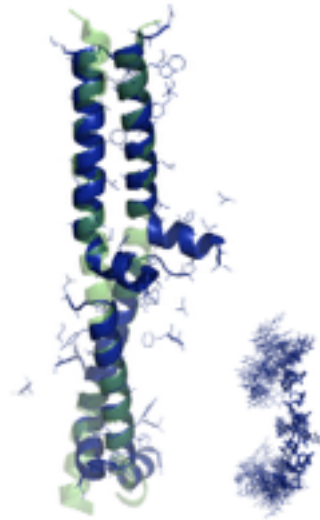
- Sample all proportions of the model but usually less than half
- Short (~25-30 residues) search models
- Search model seems to be in the wrong place!

What is solving these structures?

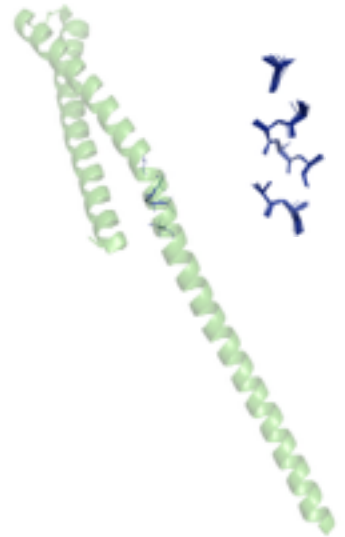


2IC9: 2 copies
RIO=125
RIO_in=125

3H00: 8 copies
RIO=110
RIO_out=110



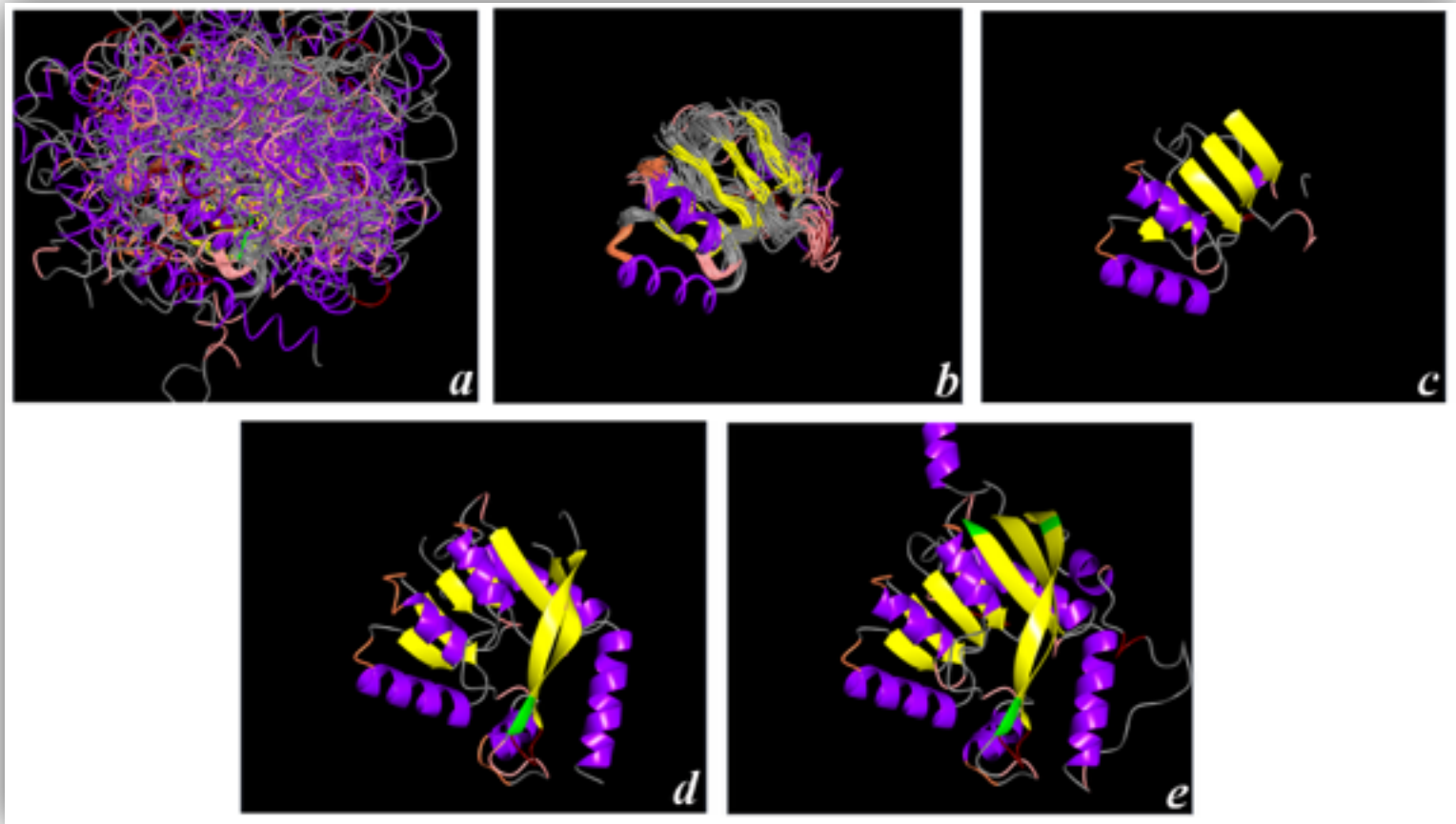
1UII: 4 copies
RIO=91
RIO_in=23
RIO_out=68



3OKQ: 1 copy
RIO=4
RIO_out=4

A mixture of in-register and out-of-register placements

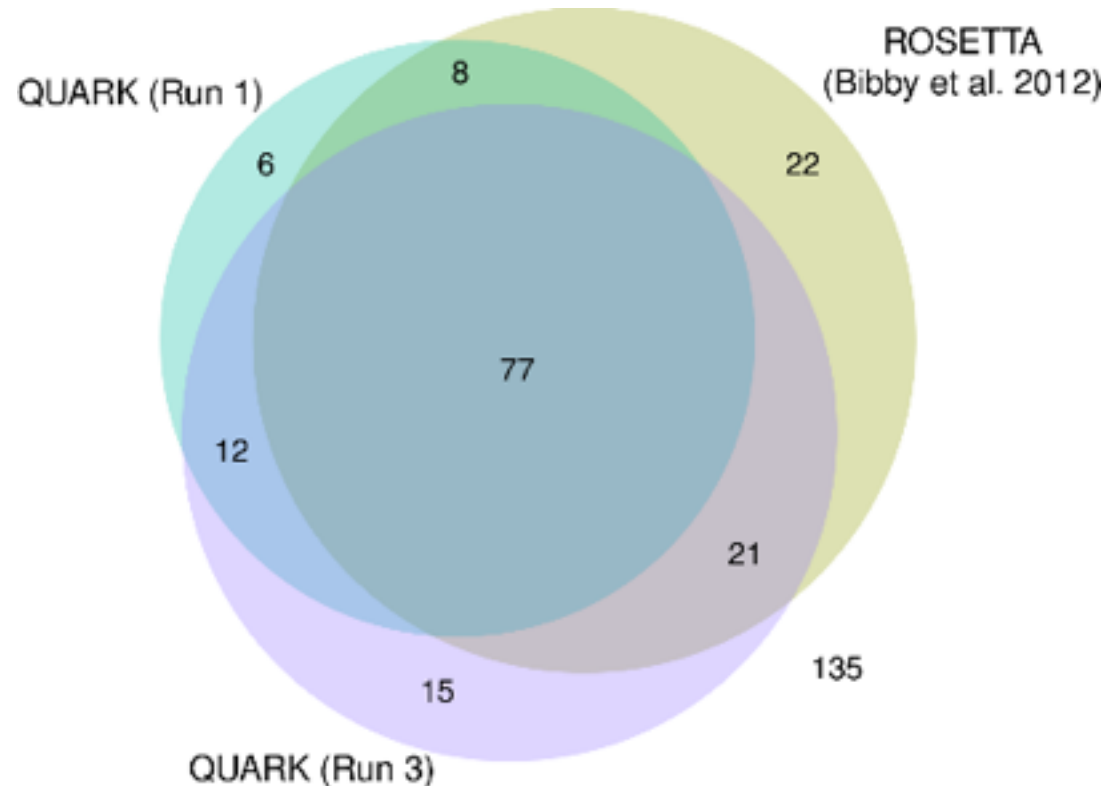
Exploiting distant homologues



a. Clustered decoy models, *b*. Truncated ensemble, *c*. Positioned MR solution, *d*. Shelxe c-alpha trace, *e*. Completed structure

QUARK vs ROSETTA

- Re-did testing using QUARK-generated decoy models
- Results are similar to ROSETTA but there are some targets that could only be solved by one or the other

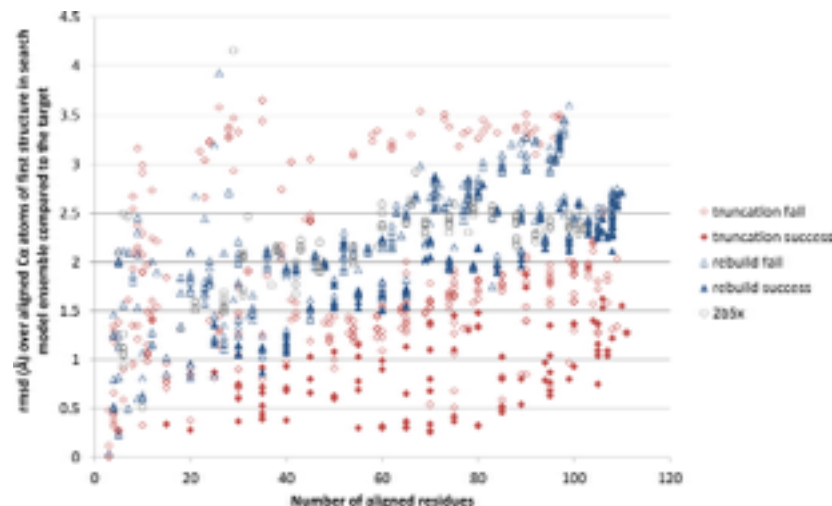


Remodelling related NMR structures or distant homologues

- **Truncation**, where the default AMPLE truncation strategy is applied to the models in the NMR ensemble.
- **Rebuilding**, where the NMR ensemble models are subjected to idealisation, comparative modelling and relaxation with ROSETTA, before being subjected to the standard truncation strategy.

25 test cases (FINDCORE)

- 19 by truncation alone
- 3 with remodelling

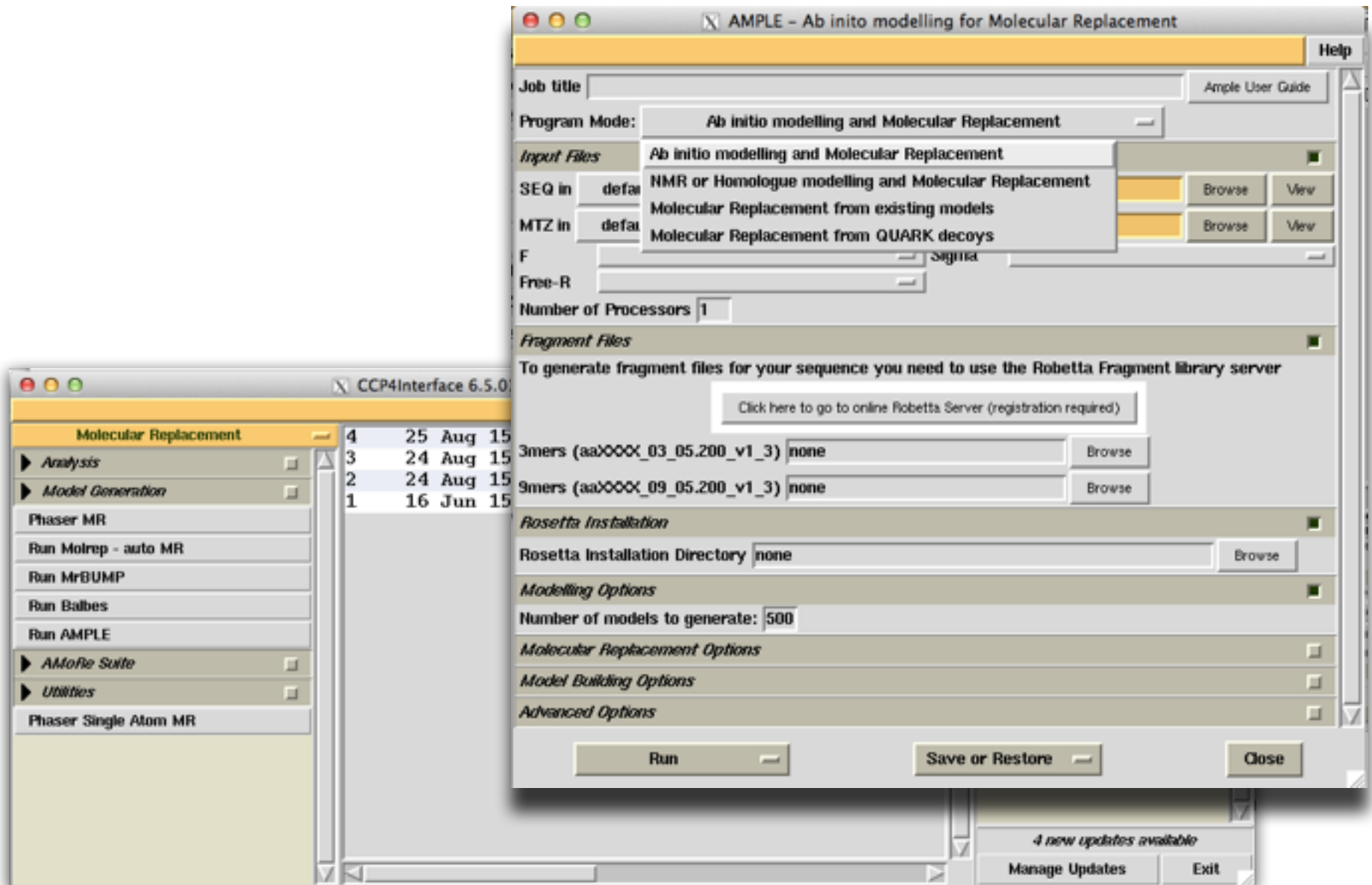


Practicalities



- Available since CCP4 6.3.0.
- Requires installation of several non-CCP4 packages (all freely available to academics):
 - **SHELXE** (now in CCP4 7.0)
 - **Rosetta** (optional)
 - **SCWRL** (optional)
- QUARK models now available so can do modelling online.
- Models can come from anywhere – just need a directory of PDB files
- Can exploit multiple processors on a local desktop, but also runs out-of-the-box on a cluster/supercomputer (SGE/LSF).

Available in CCP4i



Create models with Quark

← → ↻ zhanglab.ccmb.med.umich.edu/QUARK/ ☆ ⌵ 🔴 ☰

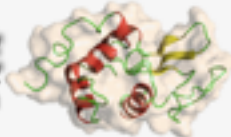
 **Zhang Lab**

UNIVERSITY OF MICHIGAN

Home Research Services Publications People Teaching Job Opening Laborly

Online Services

- I-TASSER
- QUARK
- LOMETS
- COACH
- COFACTOR
- MUSTER
- SEGMENT
- FO-MD
- ModRefiner
- REMO
- SPRING
- COTH
- BSPred
- SVMSEQ
- ANGLO
- BSP-SUM
- SAXSTER
- ThreeDom
- EvoDesign
- gPCR-I-TASSER
- TM-score
- TM-align
- MM-align
- NW-align

 **QUARK ONLINE**
Ab Initio Protein Structure Prediction 

QUARK is a computer algorithm for ab initio protein folding and protein structure prediction, which aims to construct the correct protein 3D model from amino acid sequence only. QUARK models are built from small fragments (1-20 residues long) by replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field. QUARK was ranked as the No 1 server in Free-modeling (FM) in [CASP9](#) and [CASP10](#) experiments. Since no global template information is used in QUARK simulation, the server is suitable for proteins which are considered without homologous templates.

Go to [Job Q24927](#) to view an example of QUARK output. The description of predicted feature files can be seen in [readme.txt](#). Questions about the QUARK server can be posted at the [Service System Discussion Board](#).

Cut and paste your sequence (in [FASTA format](#), less than 200 AA. Please submit bigger proteins to [I-TASSER Server](#)):

```
>1X79:AIPO8ID(CHAIN)SEQUENCE
GPLCSKSKRVNAIEEVNNVKKLTENVMSHSQCGAAAGSSDLMKELYQRCERMPTLFRLASOTEDNDEALADIQAN
DNLTQVINLYKQVRCEE
```

Or upload the sequence from your local computer:

No file chosen

Email: (mandatory, where results will be sent to. Only academic email accounts are acceptable.)

ID: (optional, your given name of the protein)

Optional: You can assign additional distance restraints for modeling (example is in [distrestraint.txt](#); format is described in [readme.txt](#)).

No file chosen



Quark Decoys

QUARK Ab Initio Results for Job Q20

Submitted Protein Sequence

Predicted Secondary Structure

Predicted Tertiary Structure

Predicted Solvent Accessibility

Download Model 1

Download Model 2

Download Model 3

Download Model 4

Download Model 5

Download Model 6

Download Model 7

Download Model 8

Download Model 9

Download Model 10

Estimated TM-score of Model 1: 0.3668 ±0.0833
Estimated TM-score of the Best of Top 10 Model: 0.4183 ±0.0764
[Download the Decoy Set Used for Clustering](#)

Estimated TM-score of Model 1: 0.3668 ±0.0833

Estimated TM-score of the Best of Top 10 Model: 0.4183 ±0.0764

[Download the Decoy Set Used for Clustering](#)

decoys.tar.gz

Quark Decoys

AMPLE - Ab initio modelling for Molecular Replacement

Help

Job title

Program Mode: **Molecular Replacement from QUARK decoys**

Input Files

SEQ in **PROJECT** Browse View

MTZ in **PROJECT** Browse View

F Sigma

Free-R

Number of Processors **1**

QUARK decoys

QUARK decoys can be generated using the QUARK server. Download the the decoy set used for clustering (decoys.tar.gz). Note: decoys can take up to 24 hours to generate.

[Click here to open QUARK Server in a browser](#)

QUARK decoys.tar.gz file: Browse

Molecular Replacement Options

Model Building Options

Advanced Options

Run Save or Restore Close

Running Rosetta

CCP4 Program Suite 6.3.0 CCP4Interface 2.2.0 running on ccp4i1 Project: aps5-raj

Molecular Replacement

- Analysis
- Model Generation
 - Phaser MR
 - Run Molrep - auto MR
 - Run MrBUMP
 - Run Balbes
 - Run AMPLE (beta)
- AMoRe Suite
- Utilities

Job	Date	Status	Description
15	11 Jul 12	FINISHED	arp_warp_classic automated model building starting
14	11 Jul 12	FINISHED	buccaneer_pipeline [No title given]
13	11 Jul 12	FINISHED	refmac5 refine jao soln
12	10 Jul 12	FINISHED	buccaneer_pipeline [No title given]
11	10 Jul 12	FINISHED	arp_warp_classic automated model building starting
10	10 Jul 12	FINISHED	refmac5 refine jao soln
9	10 Jul 12	FINISHED	refmac5 refine jao soln
8	10 Jul 12	FINISHED	refmac5 refine jao soln
7	10 Jul 12	FINISHED	refmac5 refine jao soln
6	10 Jul 12	FINISHED	arp_warp_classic automated model building starting
5	10 Jul 12	FINISHED	buccaneer_pipeline [No title given]
4	10 Jul 12	FINISHED	
3	10 Jul 12	FINISHED	
2	10 Jul 12	FINISHED	
1	10 Jul 12	FINISHED	

AMPLE (beta) - Ab initio modelling for Molecular Replacement initial parameters from /home/rolo/

Job title: 2cmp ab initio model MR

Input Files

SEQ In: ample 2cmp.fasta Browse View

MTZ In: ample 2cmp.mtz Browse View

F: FP Sigma SIGFP

Free-R: FREE

Number of models to generate: 500

Number of Processors: 1

Fragment Files

To generate fragment files for your sequence you need to use the Rosetta Fragment library server

Click here to go to online Rosetta Server (registration required)

3mers: aXXXX_03_05.200_v1_3 /home/rmk65/CCP4-Projects/ample Browse

9mers: aXXXX_09_05.200_v1_3 /home/rmk65/CCP4-Projects/ample Browse

Rosetta Installation

Rosetta Installation Directory: /home/rmk65/opt/src/rosetta3.4 Browse

Run Save or Restore Close

ROBETTA BETA
Full-chain Protein Structure Prediction Server
www.bakerlab.org

REGISTRATION
[Register / Update] [Login]

DOCUMENTATION
[Docs / FAQs]

SERVICES
Domain Partitioning & 3-D Modeling
[Queue] [Submit]

Interface Amino Acid Scanning
[Queue] [Submit]

Fragment Libraries
[Queue] [Submit]

DNA Interface Protein Scanning
[Queue] [Submit]

RELATED SITES
RosettaBackrub Server
RosettaAntibody Server
RosettaDesign Server
RosettaDock Server
Rosetta Commons
Folp
RosettaOnline
RosettaProtein Folding Project

Model 1 Target - 70513

2.66 Å over 42 residues

0.84 Å over 29 residues

68 www.bakerlab.org by Rosetta in CCP4

Results: Ensembles

CCP4 on-line

AMPLE Results

Print Refresh

Configure Exit

Summary Results Log file

Generated 105 ensembles

☒ Ensembling Results

Name	Truncation Level	Variance Threshold (σ^2)	No. Residues	Radius Threshold (δ -)	No. Decays	Number of Atoms	Sidechain Treatment
c1_t18_r1_polyAla	8	0.043932	5	1	13	25	polyAla
c1_t18_r1_reliable						78	reliable
c1_t18_r1_allatom						87	allatom
c1_t14_r1_polyAla	14	0.053448	8	1	13	40	polyAla
c1_t14_r1_reliable						114	reliable
c1_t14_r1_allatom						133	allatom
c1_t19_r1_polyAla	19	0.072878	11	1	13	55	polyAla
c1_t19_r1_reliable						145	reliable
c1_t19_r1_allatom						190	allatom
c1_t24_r1_polyAla	24	0.090143	14	1	13	70	polyAla
c1_t24_r1_reliable						185	reliable
c1_t24_r1_allatom						242	allatom
c1_t29_r1_polyAla	29	0.102773	17	1	13	85	polyAla
c1_t29_r1_reliable						200	reliable
c1_t29_r1_allatom						297	allatom
c1_t34_r1_polyAla	34	0.156564	20	1	13	100	polyAla
c1_t34_r1_reliable						215	reliable
c1_t34_r1_allatom						348	allatom
c1_t39_r1_polyAla	39	0.257561	23	1	13	114	polyAla
c1_t39_r1_reliable						241	reliable
c1_t39_r1_allatom						391	allatom
c1_t44_r1_polyAla	44	0.494188	26	1	11	128	polyAla



Results: MrBUMP

CCP4 on-line

AMPLE Results

Print Refresh

Configure Exit

Summary Results Log file

Ensembles

MRBUMP

MRBUMP Results

ensemble_name	MR_program	Solution_Type	PHASER_ILG	PHASER_TFZ	REFMAC_Rfact	REFMAC_Rfree	BUCC_final_Rfact	BUCC_final_Rfree	SHELXE_CC	SHELXE_ACL
c1_t100_r2_polyAla	PHASER	MARGINAL	52.0	5.3	0.4755	0.497	0.3175	0.3867	42.01	17
c1_t169_r2_reliable	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t134_r1_allatom	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t114_r1_allatom	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t159_r2_allatom	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t154_r2_polyAla	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t149_r2_polyAla	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t124_r1_allatom	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t139_r1_reliable	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t149_r1_polyAla	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t190_r1_reliable	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t164_r2_polyAla	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t159_r1_polyAla	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t139_r1_allatom	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t185_r1_reliable	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t175_r2_reliable	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t195_r3_allatom	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t159_r1_allatom	None	no_job_directory	None	None	None	None	None	None	None	None
c1_t190_r3_polyAla	None	no_job_directory	None	None	None	None	None	None	None	None

Results: Files








[Summary](#)
[Results](#)
[Log file](#)

Top 3 SHELXE Results

[c1_0100_r3_polyAa](#)
[c1_0100_r3_reliable](#)
[c1_0100_r3_reliable](#)

Results for ensemble: c1_0100_r3_polyAa

- Summary**

ensemble_name	MR program	Solution Type	PHASER LLG	PHASER TFZ	REFMAC R _{factor}	REFMAC R _{free}	BUCC final R _{factor}	BUCC final R _{free}	ARP final R _{factor}	ARP final R _{free}	SHELXE CC	SHELXE AC1
c1_0100_r3_polyAa	PHASER	MRASYSOL	45.0	6.7	0.4929	0.4725	0.3636	0.4115	0.2708	0.347	48.75	27
- Ensemble Search Model**

☒ Structure
- PHASER Outputs
- REFMAC Outputs
- SHELXE Outputs

Top 3 PHASER Results

[c1_0100_r3_polyAa](#)
[c1_0100_r3_reliable](#)
[c1_0100_r3_reliable](#)

Results for ensemble: c1_0100_r3_reliable

- Summary**

ensemble_name	MR program	Solution Type	PHASER LLG	PHASER TFZ	REFMAC R _{factor}	REFMAC R _{free}	BUCC final R _{factor}	BUCC final R _{free}	ARP final R _{factor}	ARP final R _{free}	SHELXE CC	SHELXE AC1
c1_0100_r3_reliable	PHASER	MRASYSOL	45.0	6.5	0.497	0.4755	0.3635	0.4178	0.2602	0.303	47.43	55
- Ensemble Search Model**
- PHASER Outputs
- REFMAC Outputs
- SHELXE Outputs**

☒ Structure and electron density
- SHELXE Logfile**

AMPLE server - coming soon!

CCP4 online

ccp4serv6.rc.harwell.ac.uk:8080/ccp4online/servlet/controller/RunnableProgramsTable

Apps New Tab

 **Collaborative Computational Project No. 4**
Software for Macromolecular X-Ray Crystallography 

Home (Logout) > Login > Programs > AMPLE > New AMPLE Run Username: rokeegan2

New AMPLE Run

To generate QUARK models for use with AMPLE, please use the **QUARK server**. This will give you a file decoys.tar.gz which you can input below using the "PDB models" option.

The file formats accepted for input are **mtz** and **clf** (structure factors) and **FASTA** (sequence target). The PDB models file can be a zip file or a tar file compressed with gzip or bzip2 (file extensions .zip, .tgz, .tar.gz or .tar.bz2).

Structure Factors: 1dtx.mtz

Sequence Target: toxid.fasta

Instead of entering a Sequence Target file you can paste your **FASTA** sequence below:
(Note that a comment line beginning with a ';' character must precede each sequence)

PDB Models: models.tar.gz

(after clicking submit, **PLEASE WAIT** for your files to upload - this may take some time)

CCP4 STFC Rutherford Appleton Laboratory, Harwell Campus
Didcot Oxon OX11 0FA
Tel: (+44) 1235 567726 | Fax: +44 (+44) 1235 567720 | Email: ccp4@ccp4.ac.uk

[About](#) | [Legal Statements](#)

 **UNIVERSITY OF LIVERPOOL**

Ongoing developments

Residue-residue contact prediction

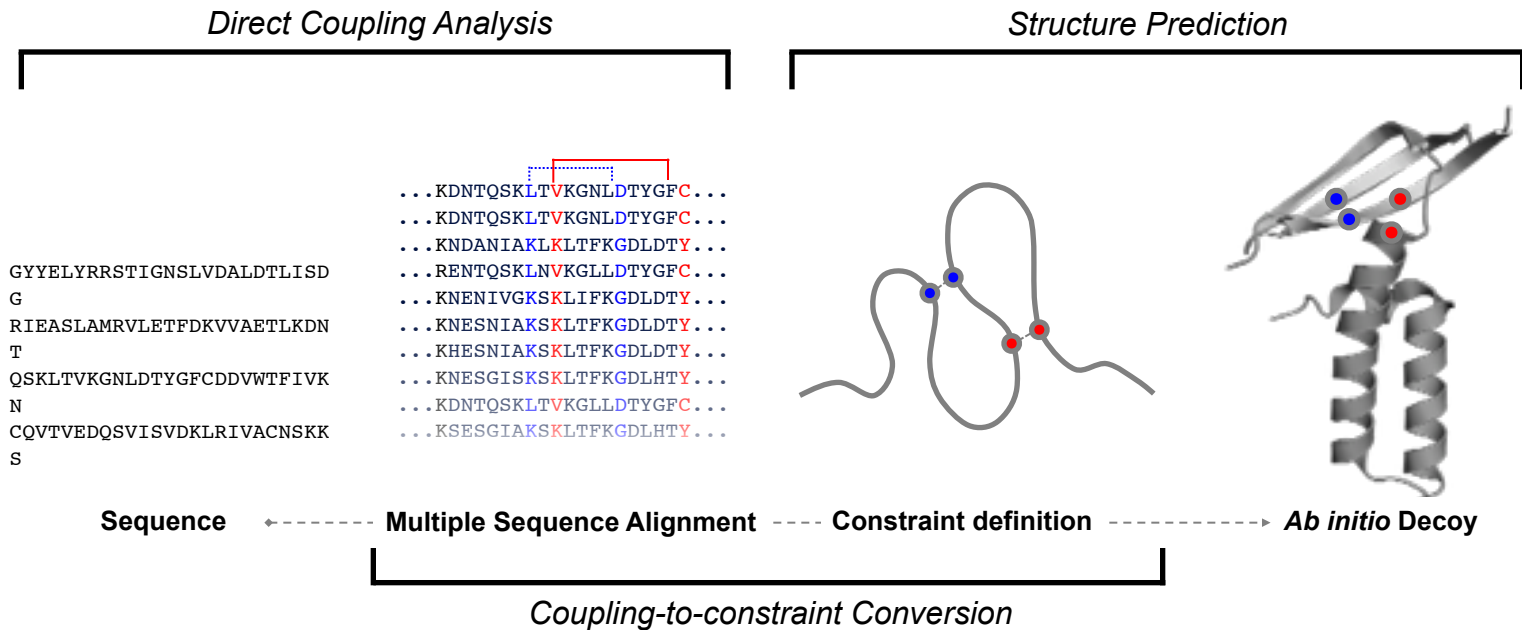


Figure 10: Average TM-score for Largest Cluster vs. Target Chain Length in Residues

The scatter plot displays the Average TM-score for the Largest Cluster (Y-axis, 0.20 to 0.80) against the Target Chain Length in Residues (X-axis, 50 to 210). A vertical dashed line is drawn at 120 residues. Data points are categorized by AMPLE success (Failure, Rosetta, PconsC2, PconsC2 + bbcontacts) and Protein Fold (all- α , mixed α/β , all- β). Two protein structures are shown: 1EAZ (top left) and 1E0S (bottom right).

Protein ID	Target Chain Length (Residues)	Average TM-score	AMPLE success	Protein Fold
1E0S	~175	~0.32	PconsC2 + bbcontacts	mixed α/β
4W97	~195	~0.56	Failure	all- α
2WY4	~140	~0.77	Failure	all- α
1A6M	~150	~0.75	Failure	all- α
4CL9	~125	~0.60	Failure	mixed α/β
1C44	~125	~0.55	Failure	mixed α/β
3W56	~130	~0.50	Failure	all- β
1VMB	~140	~0.46	Failure	mixed α/β
4WTX	~95	~0.48	Failure	all- β
4U3H	~100	~0.45	Failure	all- β
1KW4	~90	~0.45	Failure	all- α
2NUZ	~70	~0.64	Rosetta	all- β
1ABA	~90	~0.63	Rosetta	mixed α/β
3EGU	~65	~0.27	Failure	all- β
1E0S	~175	~0.32	PconsC2 + bbcontacts	mixed α/β
1E0S	~175	~0.32	PconsC2 + bbcontacts	all- α
1E0S	~175	~0.32	PconsC2 + bbcontacts	all- β

Summary

- AMPLE – a pipeline for Molecular Replacement with unconventional models.
- works particularly well with α -helical proteins.
- solves structures with fragments or more complete models.
- usually best for small proteins with reasonable resolution.
- favourable cases can work > 330 residues and > 2.8Å resolution.
- very easy to use.

Credits

- Dan Rigden, Olga Mayans – IIB, Liverpool University
- Felix Simkovic, Adam Simpkin and Jaclyn Bibby – Liverpool University
- Ronan Keegan, Martin Wynn – CCP4 (STFC)
- BBSRC
- Hartree Center
- **Phaser**: Randy Read, Airlie McCoy & Gabor Bunkozci
- **SHELX**: Andrea Thorn, Tim Gruene & George Sheldrick
- **Refmac**: Garib Mushudov, LMB-MRC Cambridge
- Developers of **Rosetta** and **Quark**
- **Molrep**: Alexei Vagin & Andrey Lebedev
- Thanks to authors of all underlying programs (**Maxcluster**, **Theseus**, **SPICKER**, **SCWRL**)

