Advanced MR with AMPLE and SIMBAD

Adam Simpkin

UNIVERSITY OF LIVERPOOL

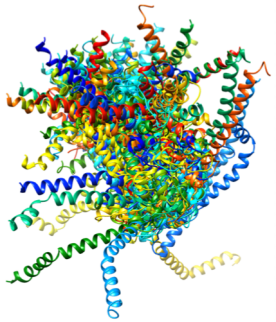
Conventional MR techniques

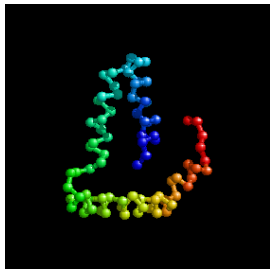
1. Homologues are identified by sequence
2. Crystal structures or homology models are used as search models
3. Mainly use single structures
4. A small number of search models are used
5. Any editing to the search model is gentle

What is AMPLE?

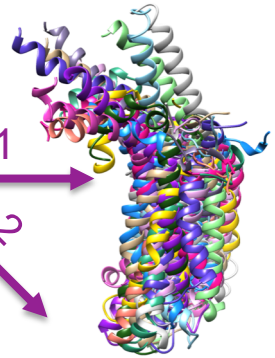
- **AMPLE** (***A****b initio* **M**odelling of **P**roteins for mo**LE**cular replacement)
- Pipeline that uses *ab initio* models to generate suitable search models in scenarios where no homologues can be identified.
- A sophisticated method to process search models in order to get the most out of them

ab initio models (ROSETTA/ QUARK)

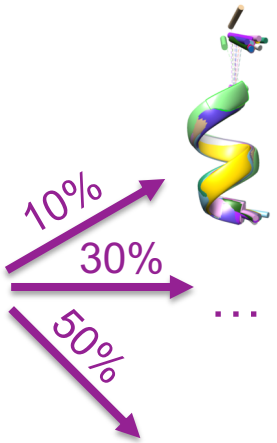




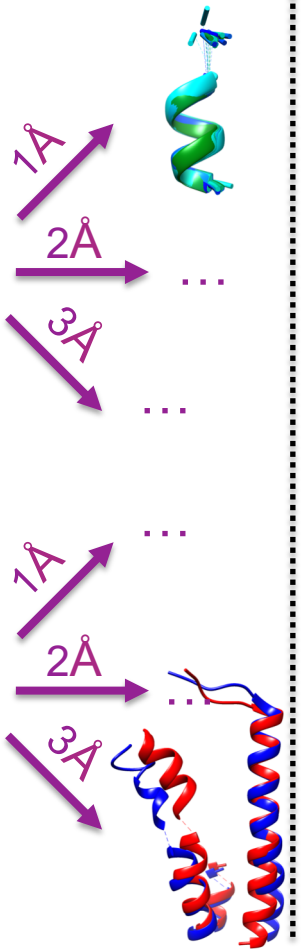
cluster (SPICKER)



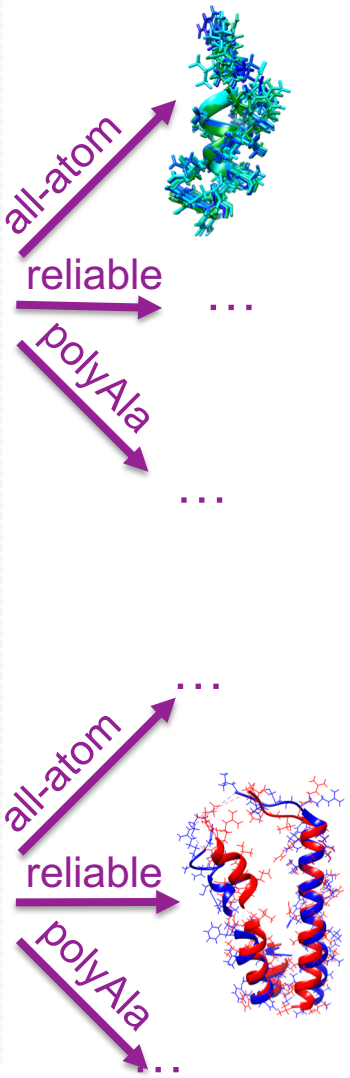
truncate (THESEUS)

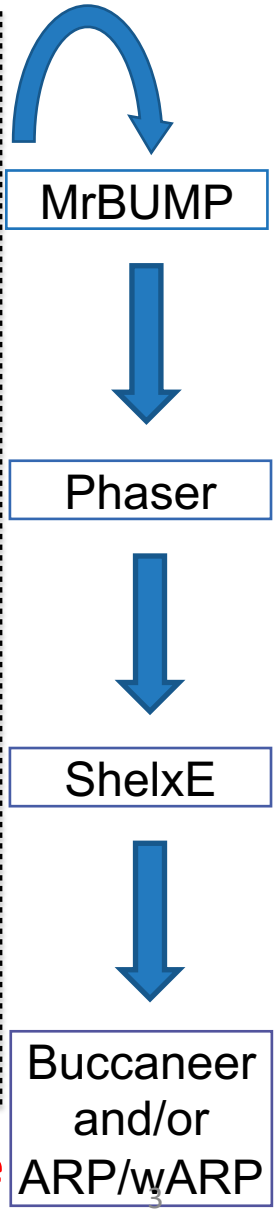


cluster (MAXCLUSTER)



side chains

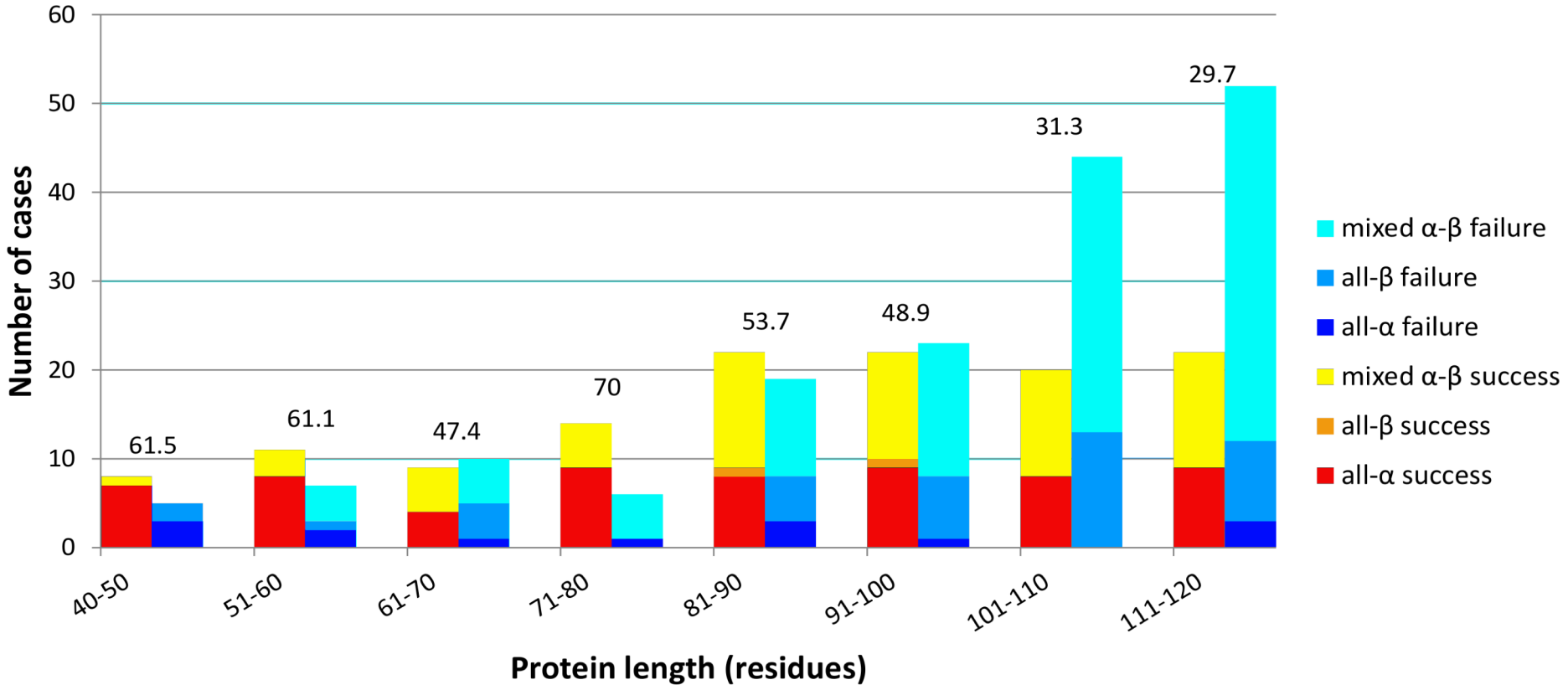




http://www.cs.ucl.ac.uk/staff/D.Jones/t42morph.html

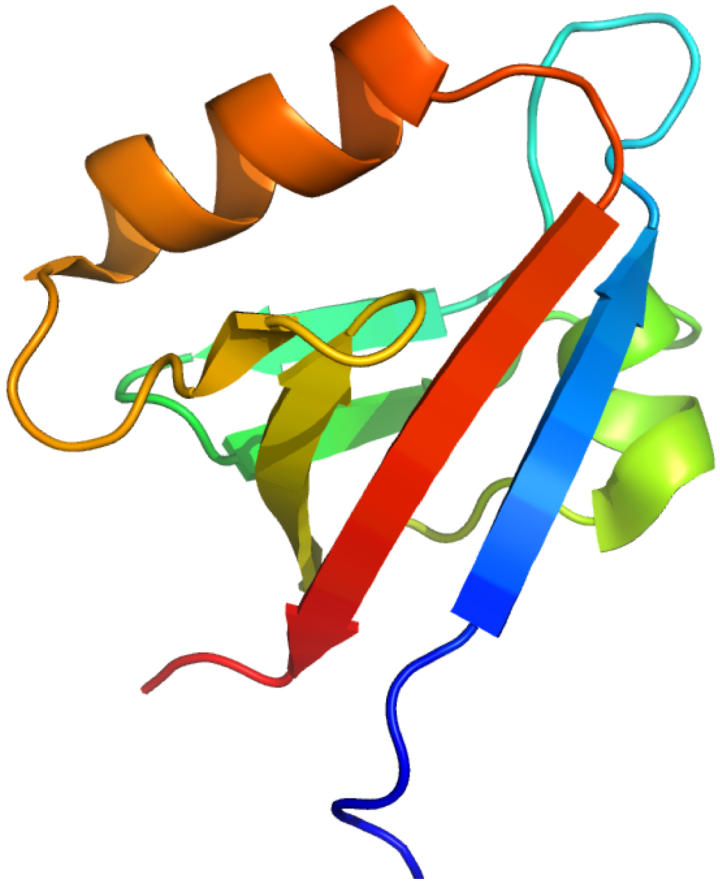
up to 180 ensemble search models

Testing on a large set of small proteins

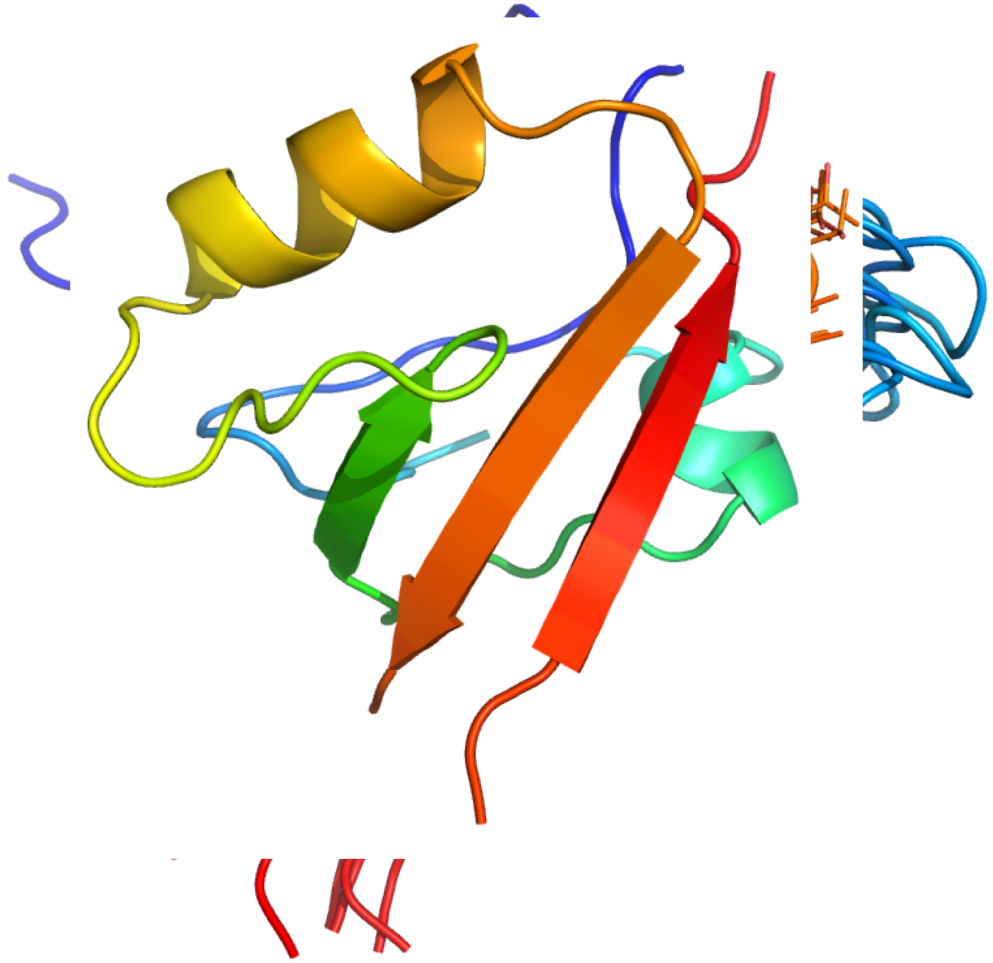


- 126 out of 295 (43%) solved. Treated as novel folds.
- all-α, 80%; all-β, 2%; mixed α_β, 37%

Example success (1R6J, a PDZ domain)

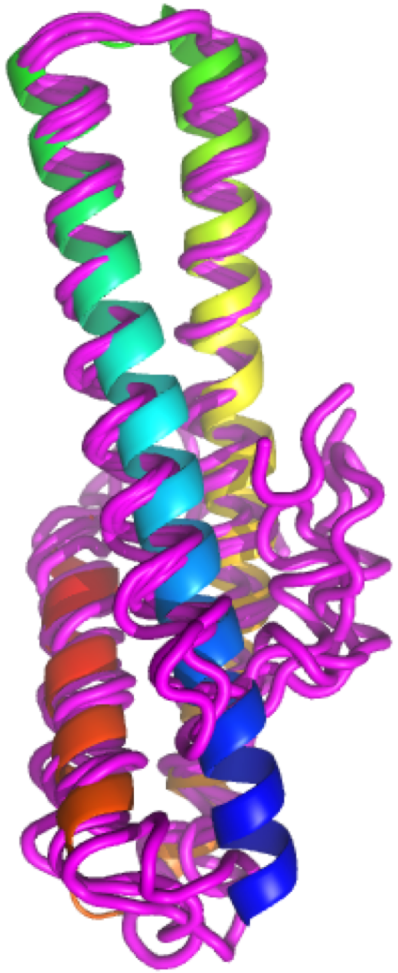


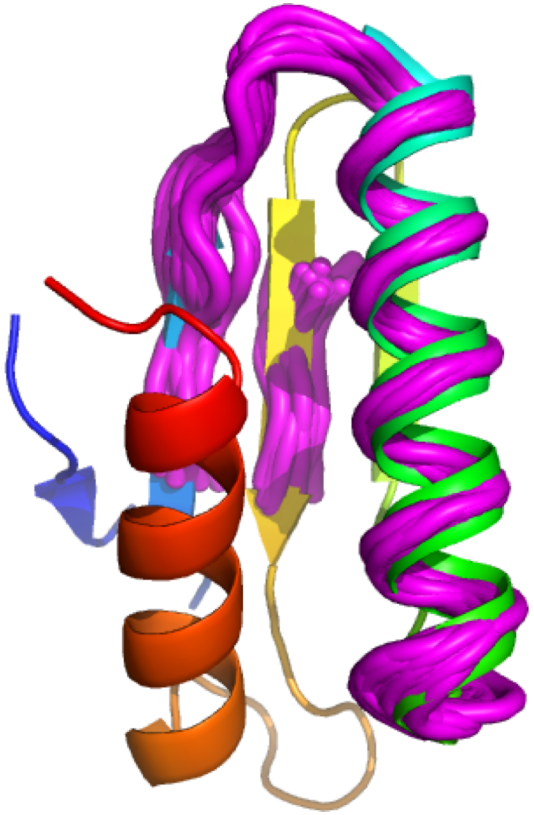
Crystal structure

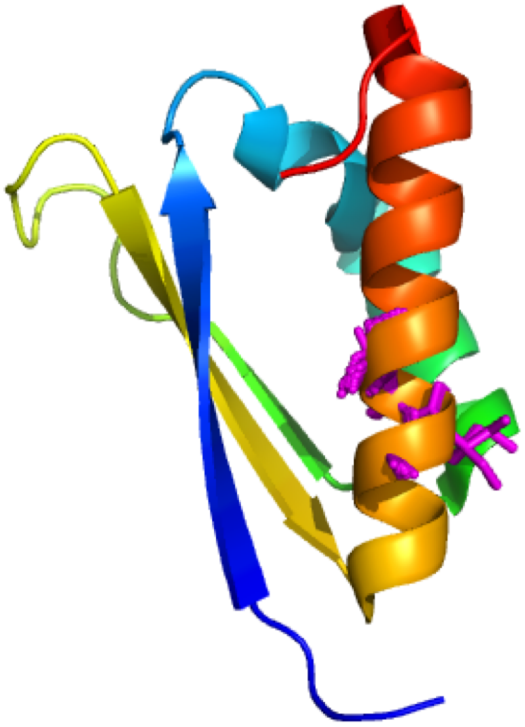


Truncated Rosetta cluster

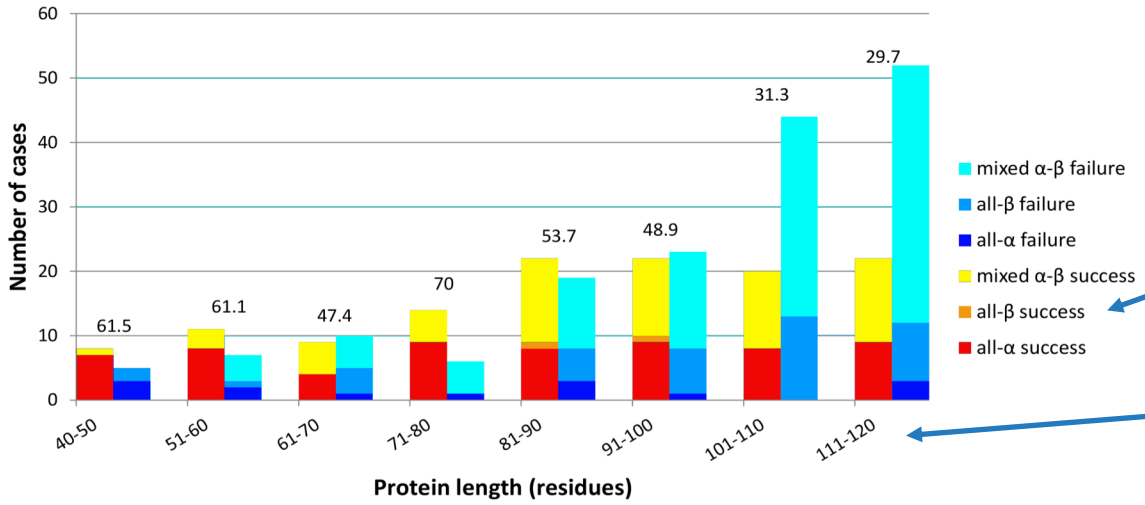
Success with various sized search models







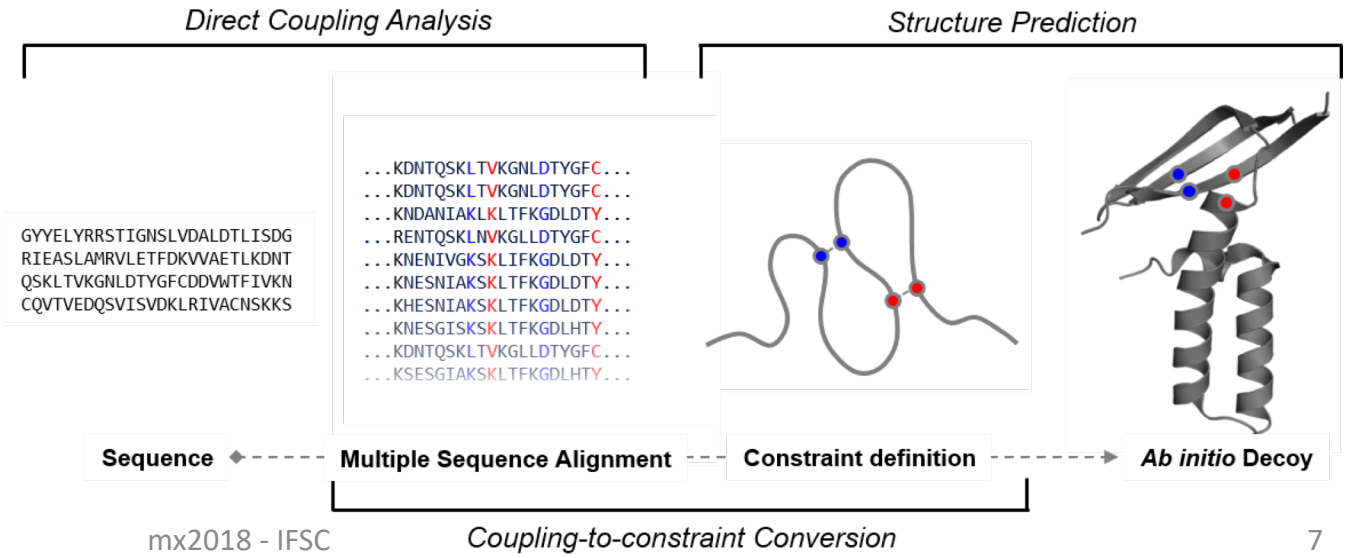
Contact assisted *ab intio* modelling



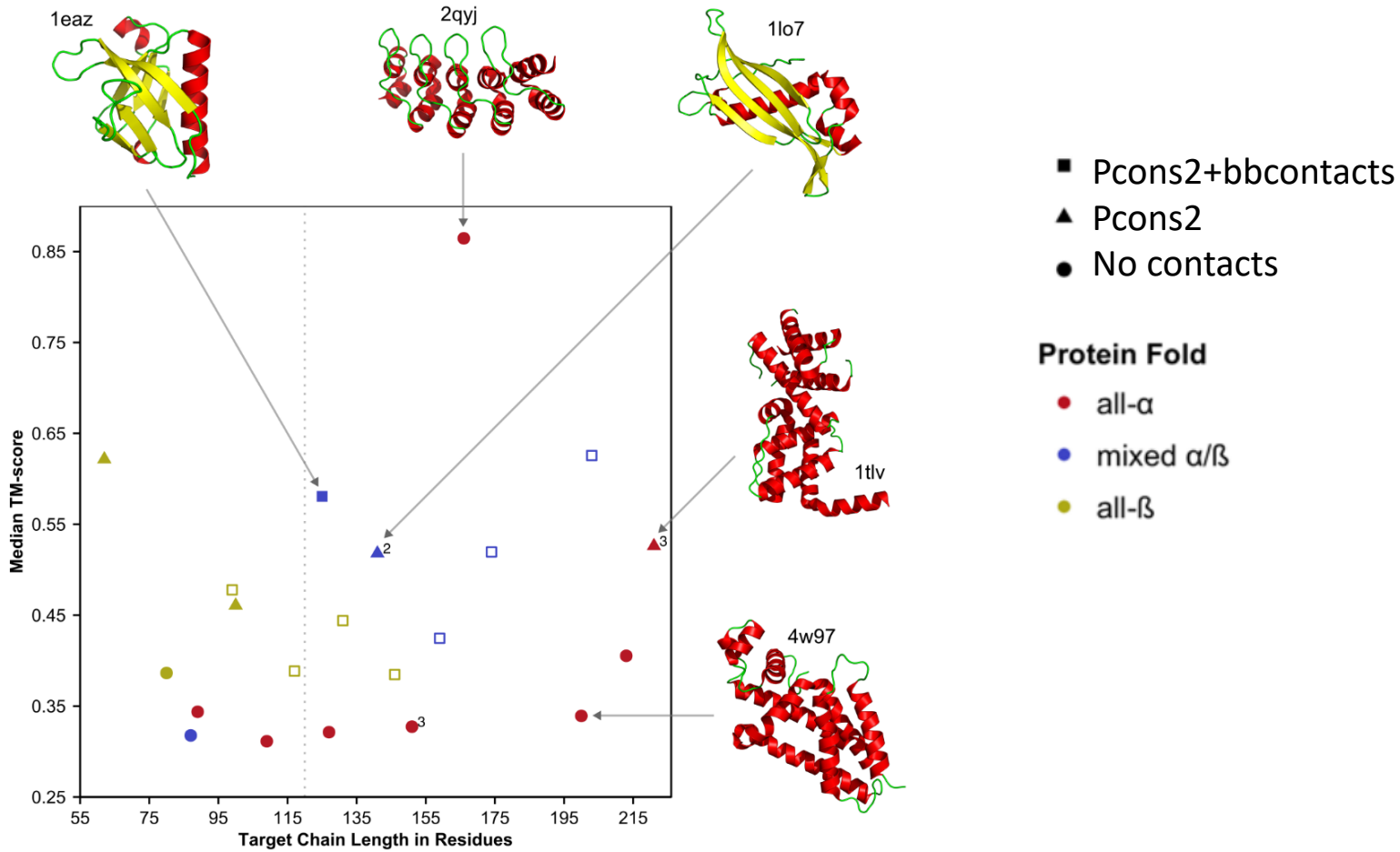
Very low (2%) success rate with all-β proteins

Lower success at upper size limit (120 residues)

Can the extra info from contact predictions helps successfully address larger proteins and harder β-rich structures?



With contact predictions AMPLE succeeds with larger and more β-rich proteins



Simkovic *et al.* (2016) IUCr J. **3**, 259-270

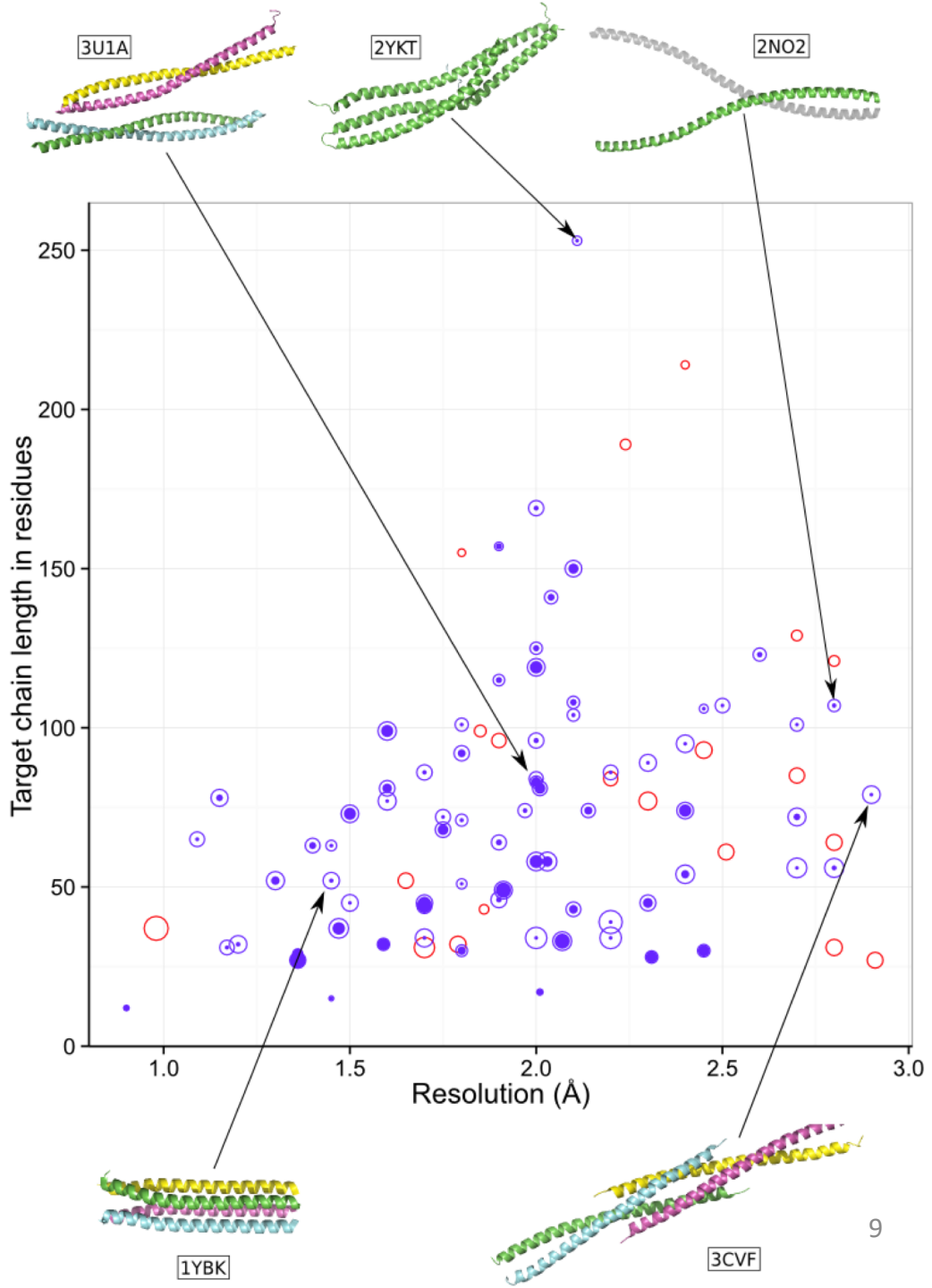
Coiled-coils

Coiled-coils generally considered awkward for MR

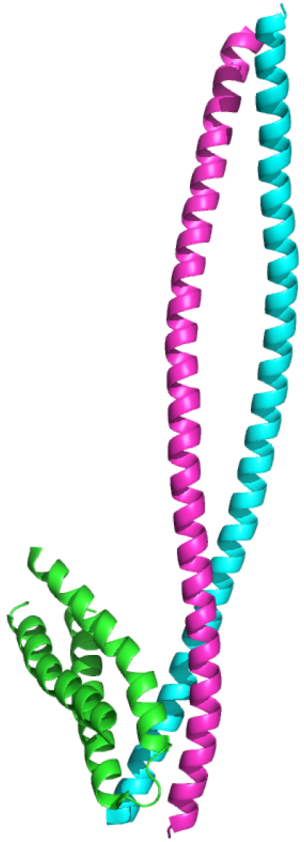
Nevertheless, AMPLE solved **~80%.**
No knowledge of oligomeric state required

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
- **1YBK** right-handed coiled coil

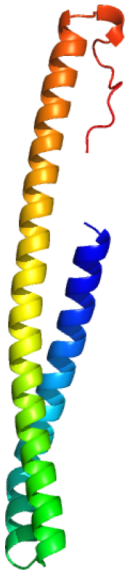
Small ideal helix library solved half



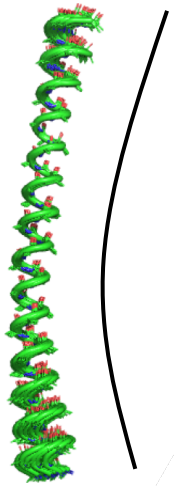
Coiled-coil example: 1X79



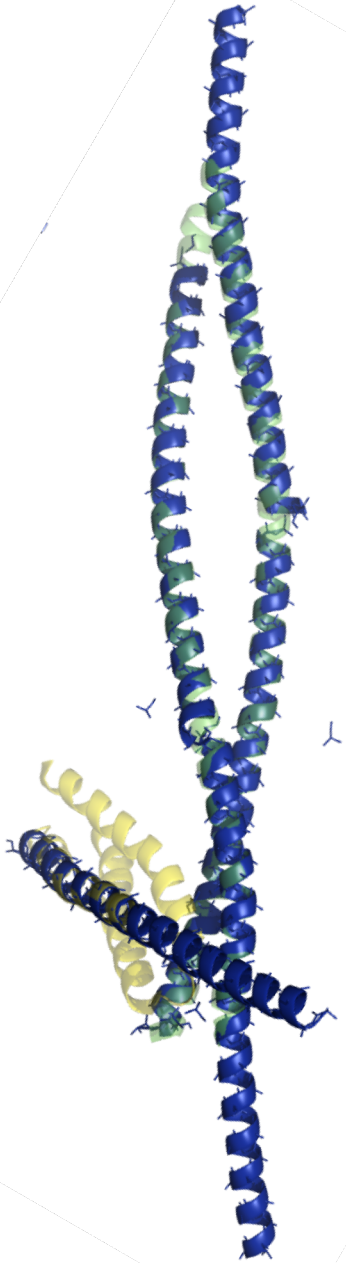
Crystal structure



Rosetta model



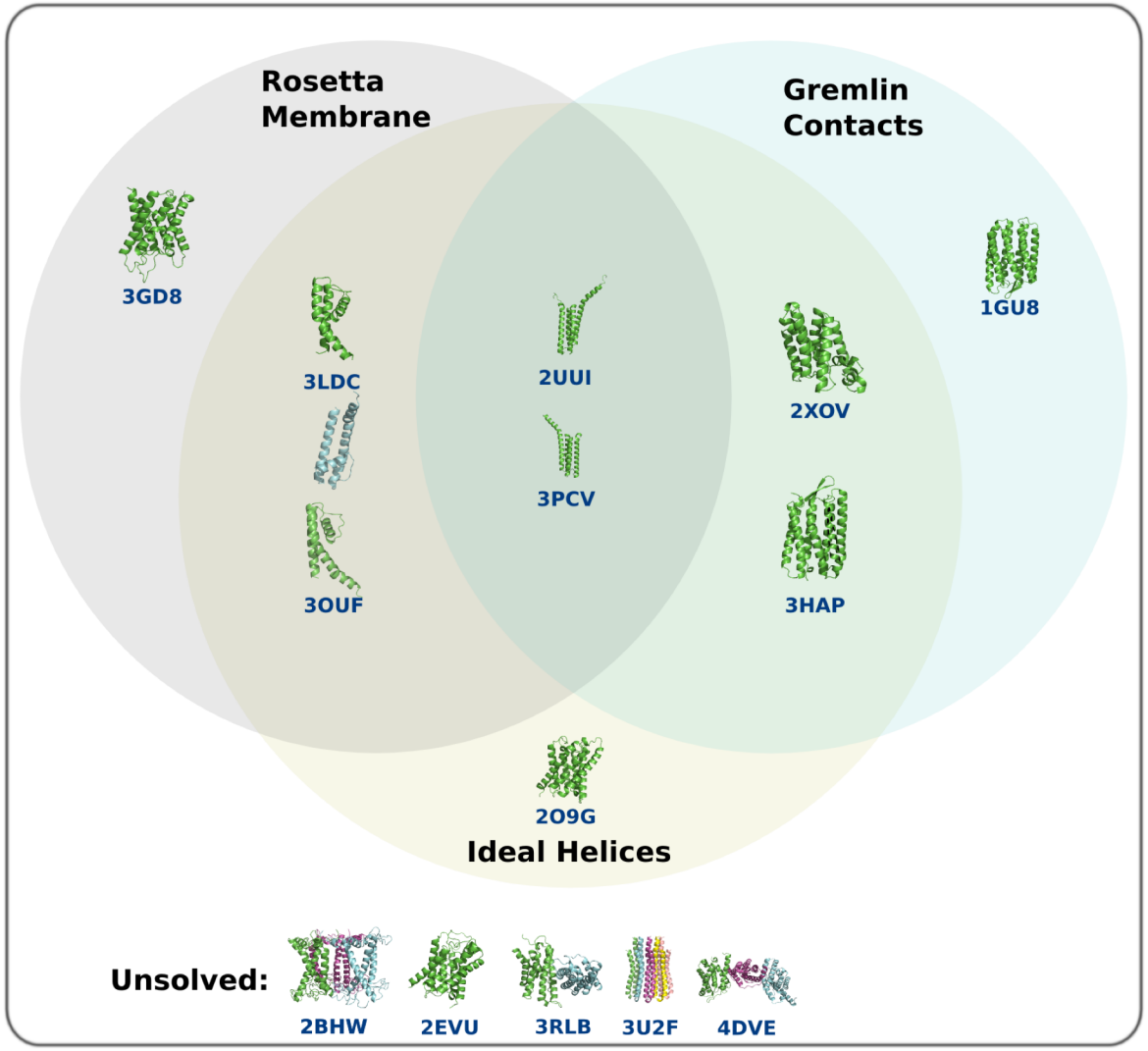
AMPLE search model



Three chains,
322 residues,
~70% coiled-coil,
2.41Å

Thomas *et al.* (2015) IUCr J. **2**, 198

Transmembrane helical proteins



- 15 transmembrane helical proteins
  - 23 – 249 residues
  - 1.45 – 2.5Å resolution
- 10 clear successes
- Largest has 223 residues (3GD8).

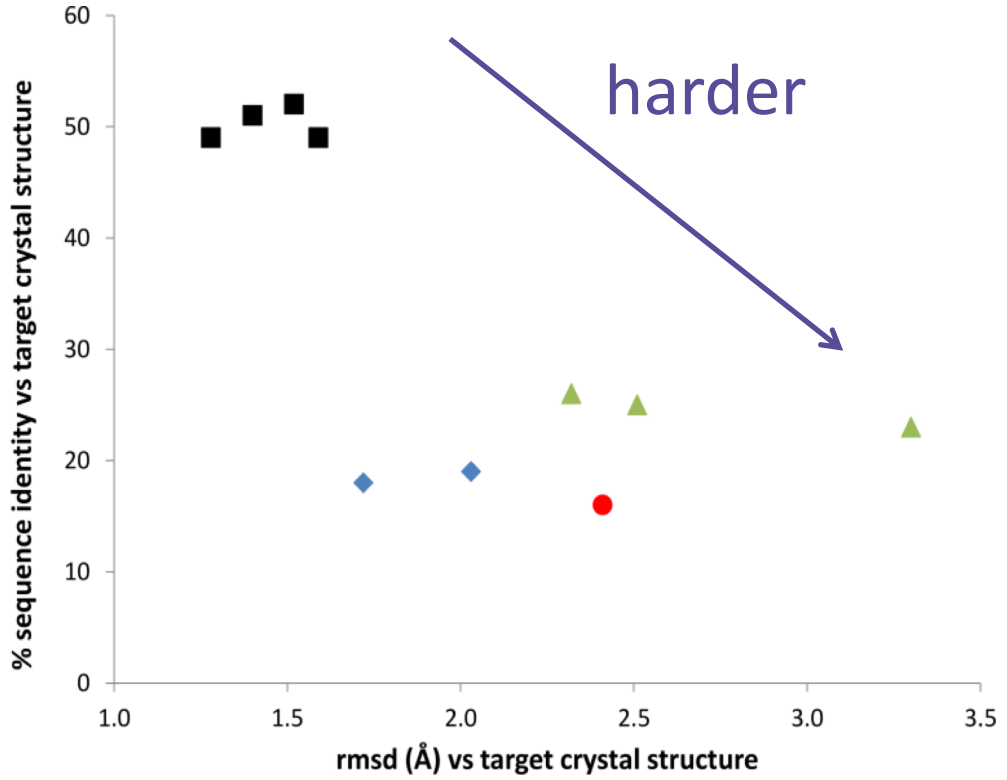
Thomas *et al.* (2017) *Acta Cryst.* D**73**, 985-996

Repurposing AMPLE's truncation protocol

What is truncated?	Rationale
Ab initio models	Variability predicts inaccuracy
NMR models	Variable regions are less well-determined/more flexible
Set of distant homologues	Eliminate variable regions from a superposition to find the conserved core likely shared with the target
CONCOORD ensembles from a single distant homologue	Reveal the conserved core in a single structure by exploiting the correlation between close packing and evolutionary conservation

AMPLE and NMR

- For MR with NMR structures variable regions are least well-defined, provide least phasing information and can erroneously trigger packing problems
- NMR structures are traditionally tricky for MR
- AMPLE works better than FindCore in harder cases, down to 18% sequence identity.

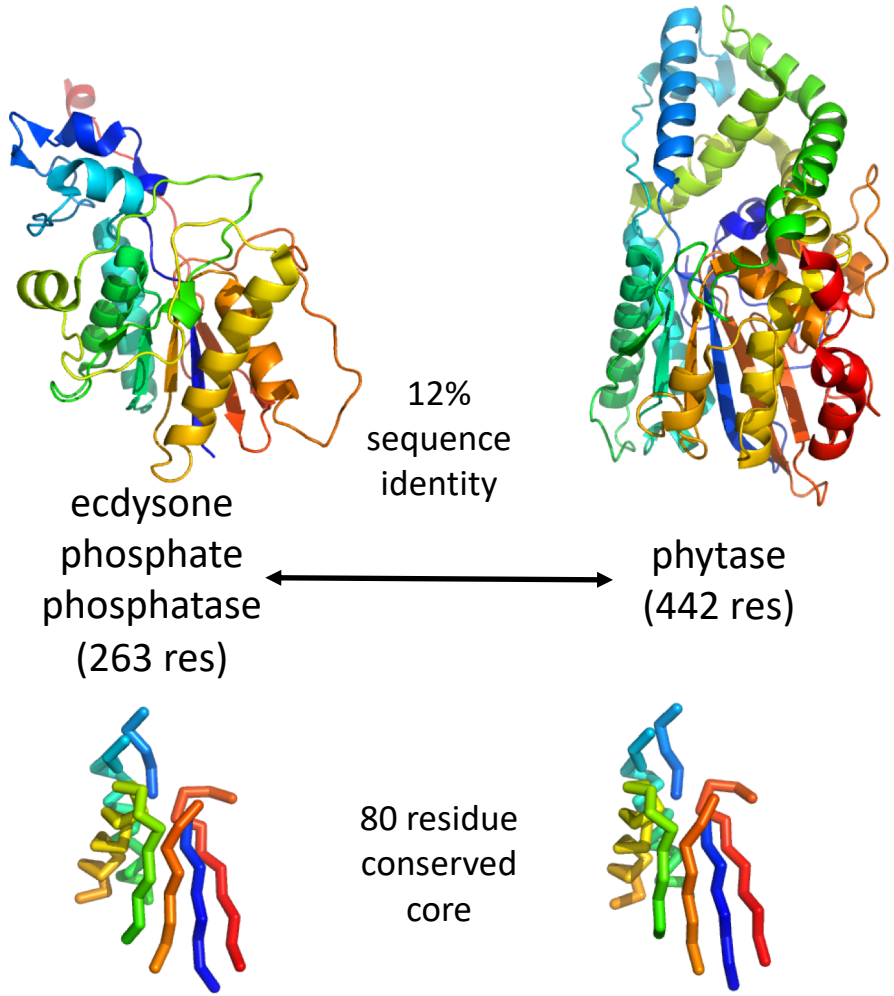


- FindCore and AMPLE succeed
- FindCore fails, AMPLE succeeds (truncation only protocol)
- FindCore fails, AMPLE succeeds (Rosetta rebuild protocol)
- Both fail

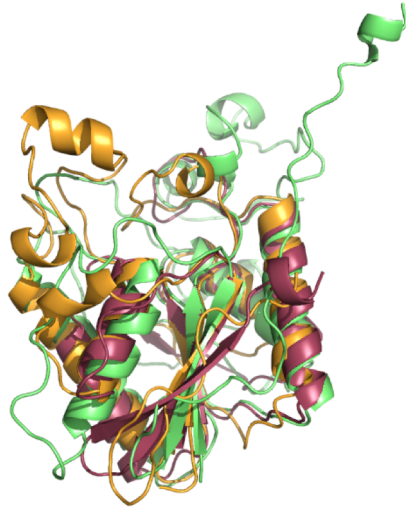
Bibby *et al.* (2013) Acta Cryst D**69**, 2194

AMPLE for processing crystal structures

- Often have homologous structures but they are too divergent to solve the target
- AMPLE can help find small, better conserved core for MR
- With **multiple** structures compare them directly
- With a **single** structure exploit correlation between rigidity and evolutionary conservation

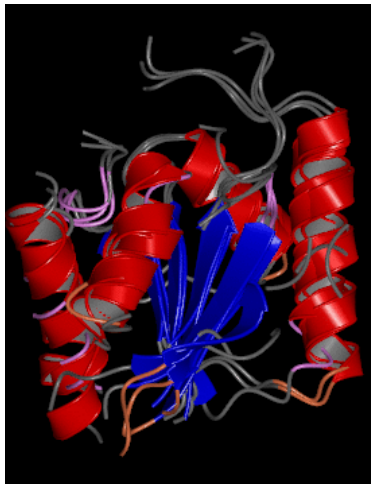


Multiple homologue truncation

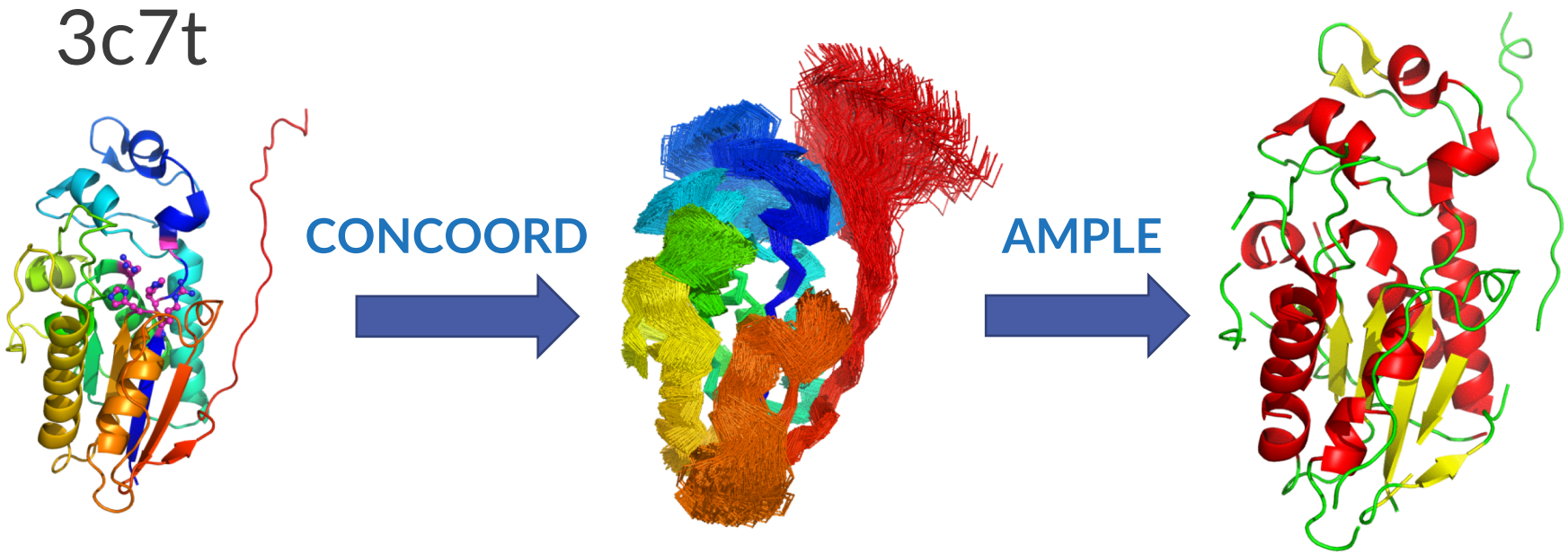


- 7 distantly homologous superfamily members (7-28% identity)
- Can 2-3 in superposition solve other 4?

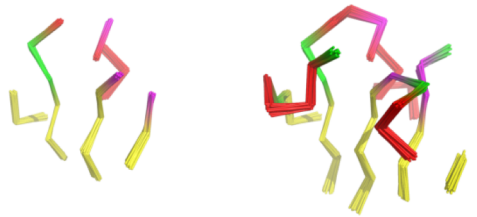
	AMPLE				MrBUMP	
Target PDB	1UJB/2A6P/3C7T	1UJB/2A6P	2A6P/3C7T	1UJB/3C7T	All 3	no 1UJB
1E59	0/57	-	-	-	0/10	-
1EBB	11/57	0/57	1/57	14/57	3/10	0/7
2QNI	34/57	14/57	19/57	27/57	0/10	-
3DCY	45/57	41/57	40/57	45/57	2/10	0/7



Single homologue truncation based on flexibility



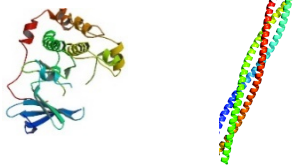
	%id vs 3c7t	length	res (Å)	AMPLE solved with CONCOORD results	MrBUMP with manual edits
1ujb	22	156	2.1	Yes (51-103 res)	No
2qni	13	194	1.8	Yes (25-142 res)	Yes
1e59	19	239	1.8	No	No
4e09	18	240	2.45	Yes (90 res)	No
1ebb	23	202	2.3	Yes (77 res)	No
3dcy	20	269	1.75	Yes (38-64 res)	No



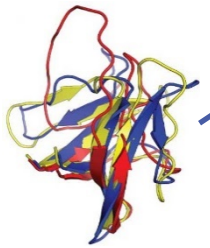
Brutally truncated search model ensembles capture best conserved catalytic core

Modelling

```
>seq
FASDGITF
DRSLFFGH
```



NMR models



ab initio modelling

Contact-assisted *ab initio*

Truncated search model ***ensembles***

MrBUMP (Phaser, Molrep)

SHELXE

BUCCANEER

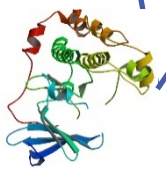
ARP/wARP

CONCOORD pseudodynamics

Truncated search models

AMPLE

Transforming and processing experimental structures



When to use AMPLE

- If your target is a novel or divergent globular fold and not too large
- If your target contains a coiled-coil protein (or maybe is a TM helical protein)
- If you have one or more distant homologues available, but they cannot solve your target by conventional means
- If you have an NMR structure for a homologue of your target

What is SIMBAD?

SIMBAD (**S**equence **I**ndependent **M**olecular replacement **B**ased on **A**vailable **D**atabase)

Pipeline to screen everything against your experimental data

Lattice Parameter Search against every known crystal structure

Exhaustive but fast contaminant search

Brute-force search with every non-redundant known fold

!!! No sequence required !!!

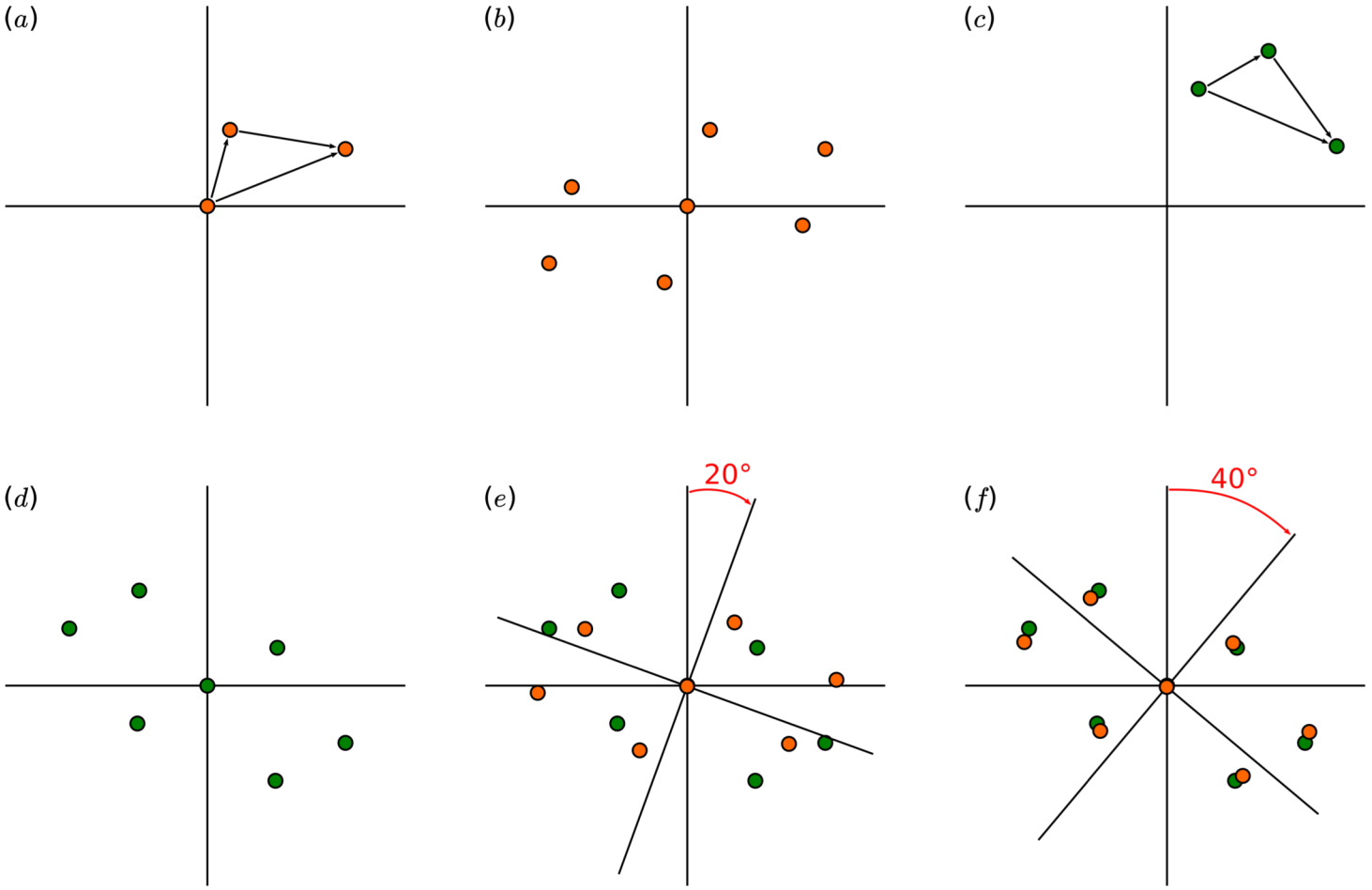
Scenarios SIMBAD can help

- A Contaminant protein has been crystallised
- The sequence of the protein is unknown
- Unexpected fragmentation (Unit cell is smaller than the protein!)
- Homologues are not obvious from sequence
- High % sequence ID homologues that undergo large conformational changes

How does SIMBAD work?

- Molecular replacement is split up into a rotation search and a translation search
  - The rotation search identifies the orientation of the structure
  - The translation search identifies the position of the structure
- SIMBAD uses the rotation search in order to screen a database of search models in a brute force manner

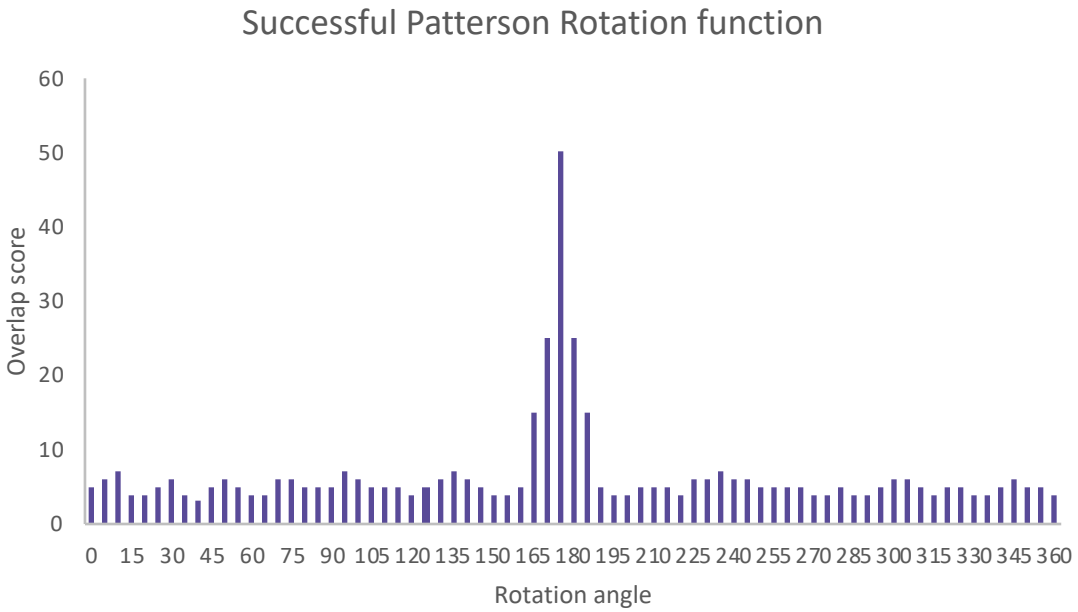
The rotation search – Patterson based

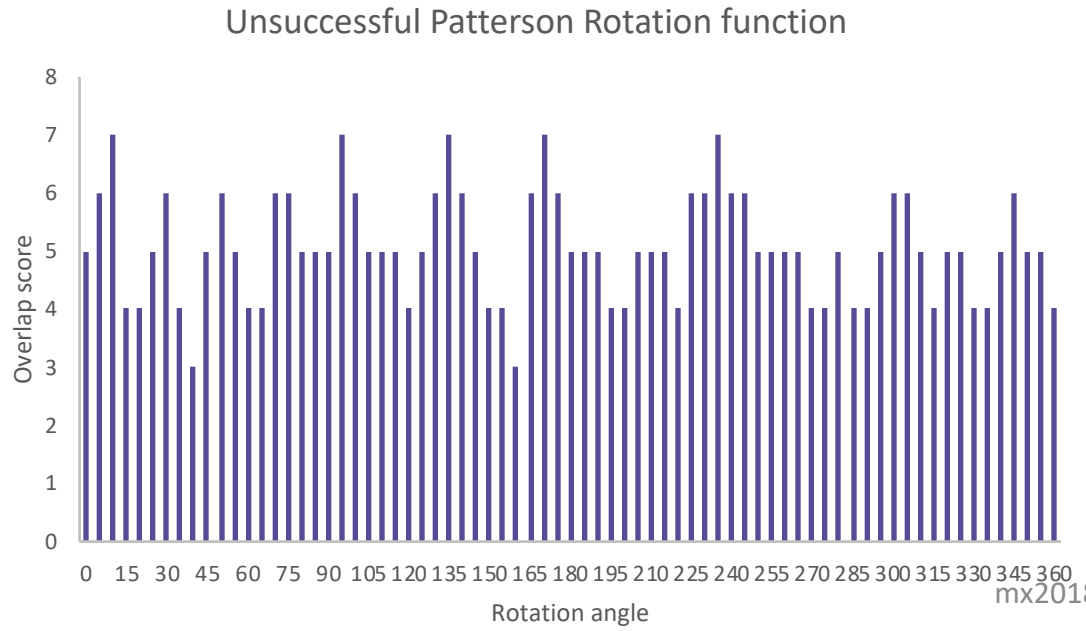


- Consider the Patterson map for our crystal structure (b) and our search model, (d).
- Using an overlap function we can measure the agreement between the two maps at different orientations and thus determine the orientation of the crystal structure

Exploiting the rotation search

- Successful molecular replacement first requires a successful rotation search
- Therefore we can exploit the characteristics of a successful rotation search in order to identify potential search models





- A successful rotation function will give a large overlap score at a specific rotation angle.
- Meanwhile an unsuccessful rotation search will by unable to distinguish the correct orientation from noise.

Taking a Z-score of the peaks from the rotation function allows us to identify this feature and compare it for difference search models.

The SIMBAD method

1. Run the rotation search against every structure in a given database and compare the output Z-scores.
2. Rank the models by Z-score
3. Select the top scoring models to be used as search models in full MR

The MoRDa database

1. The MoRDa database was far less redundant than the PDB.
2. The MoRDa database is made up of domains

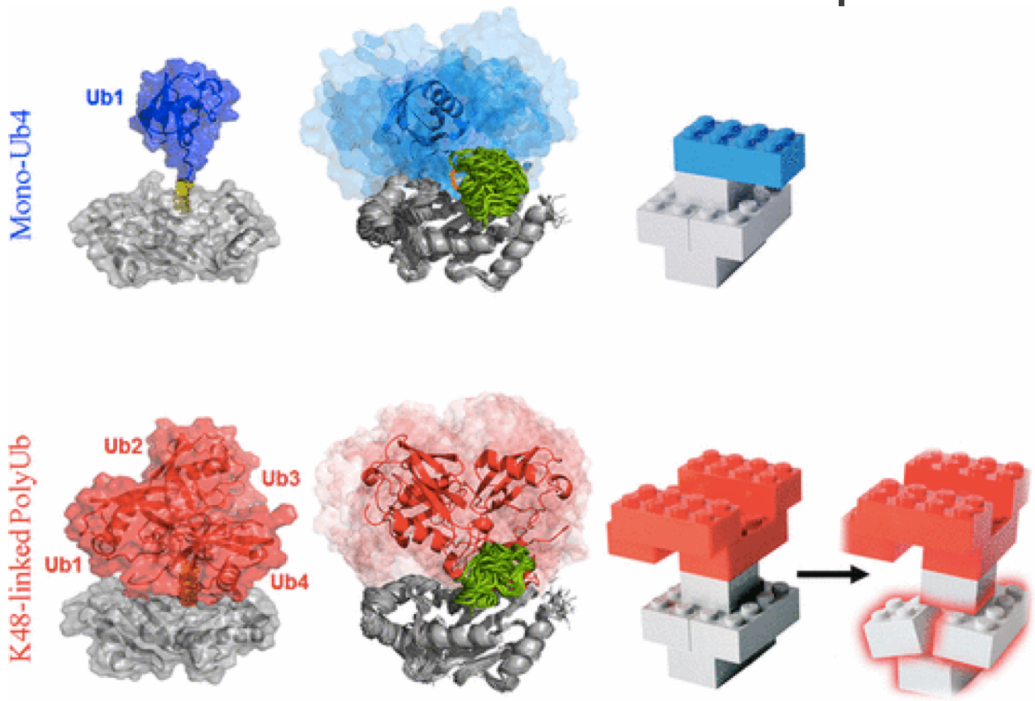
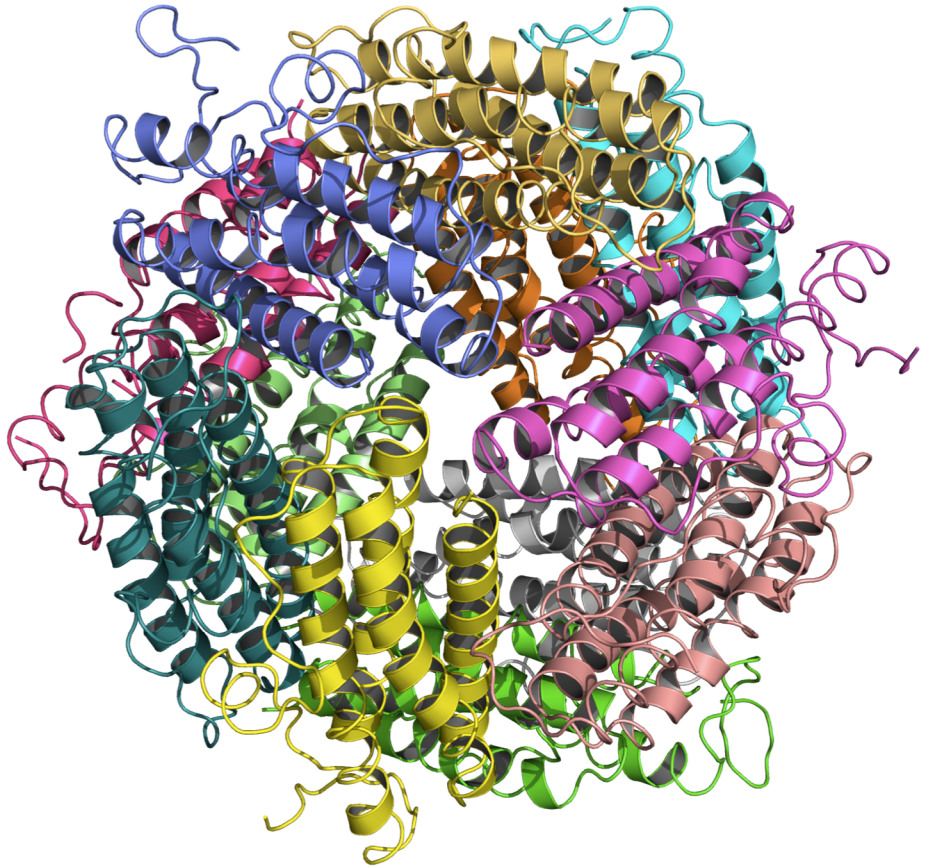


Diagram from Levy (2017) Biochemistry **56**(38), 5040–5048

Example solution: DPS

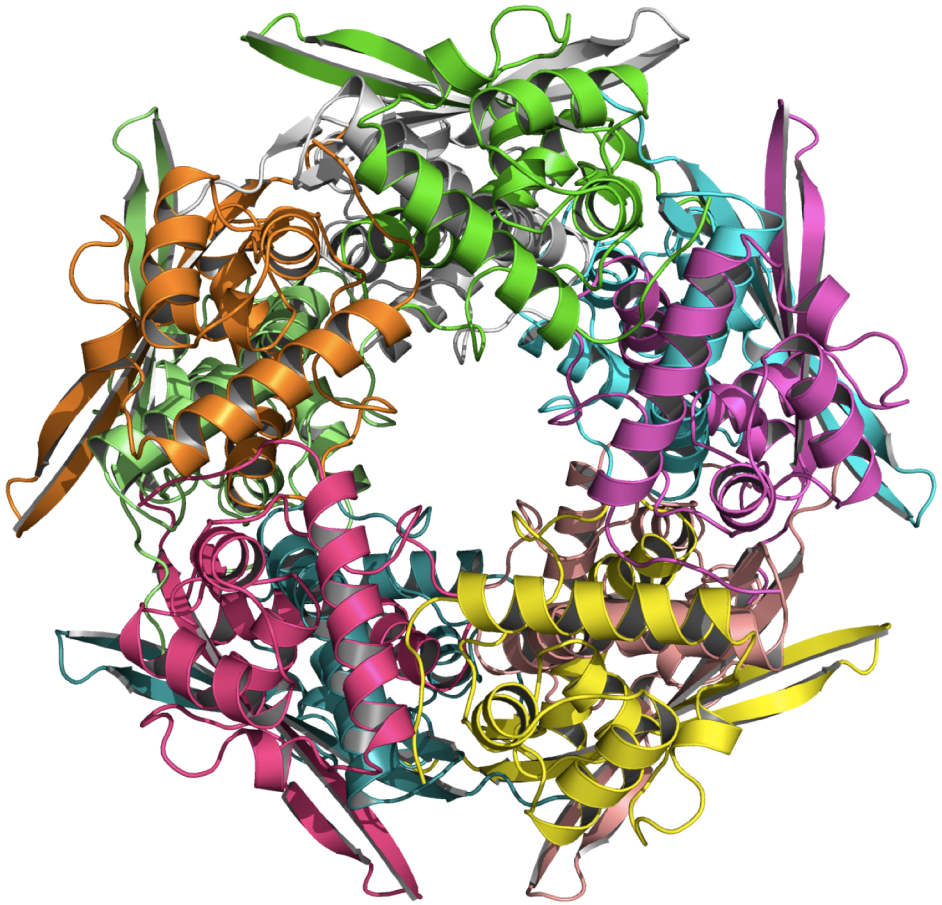


- The target protein was expressed in *E. coli* cells
- Crystals were obtained after several months of incubation.
- MR attempts with homologues failed.
- SIMBAD readily identified that the crystals in fact contained DPS

CONTAMINANT

Simpkin *et al.* (2018) Acta Cryst D**74**, 595-605

Example solution: Cyanase

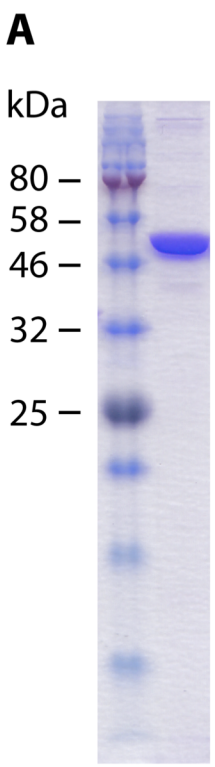


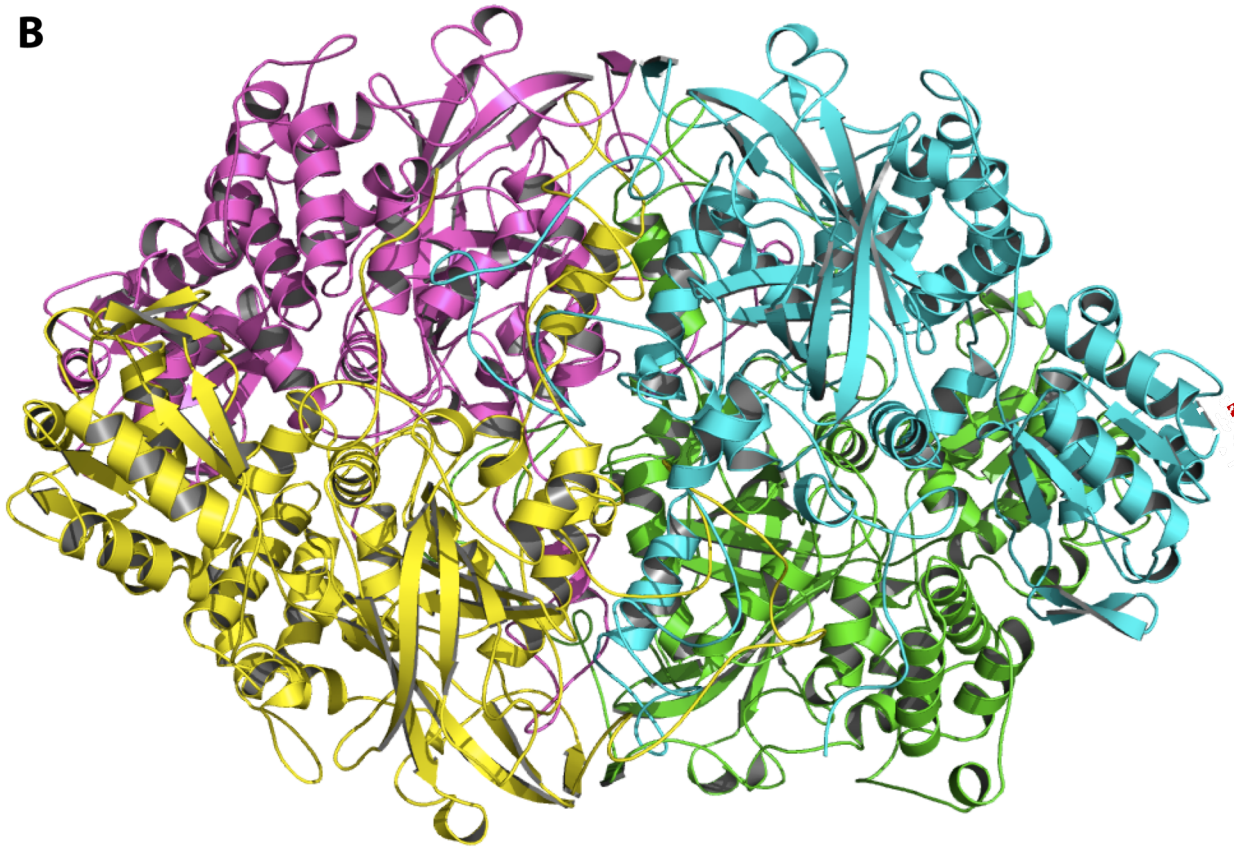
- The target protein was expressed in an insect cell line.
- Crystals were obtained after 6 months of incubation.
- MR attempts with homologues failed
- SIMBAD readily identified that the crystals in fact contained Cyanase

CONTAMINANT

Simpkin *et al.* (2018) Acta Cryst D**74**, 595-605

Example solution: Catalase





- The target protein was expressed in *E. coli* cells
- Mass spectrometry (MS) confirmed the identity of the purified target protein.
- Crystals were obtained after 3 months of incubation.
- MR attempts with homologues failed as well as methods such as AMPLE/ARCIMBOLDO
- SIMBAD readily identified that the crystals in fact contained Catalase

CONTAMINANT

Simpkin *et al.* (2018) Acta Cryst D**74**, 595-605

Contaminants: a trend?

- Contamination is the most common scenario where SIMBAD works
  - This has led to Ronan's alternative name for the pipeline: SIMBAD: **SIM**pkin's **BAD** news pipeline

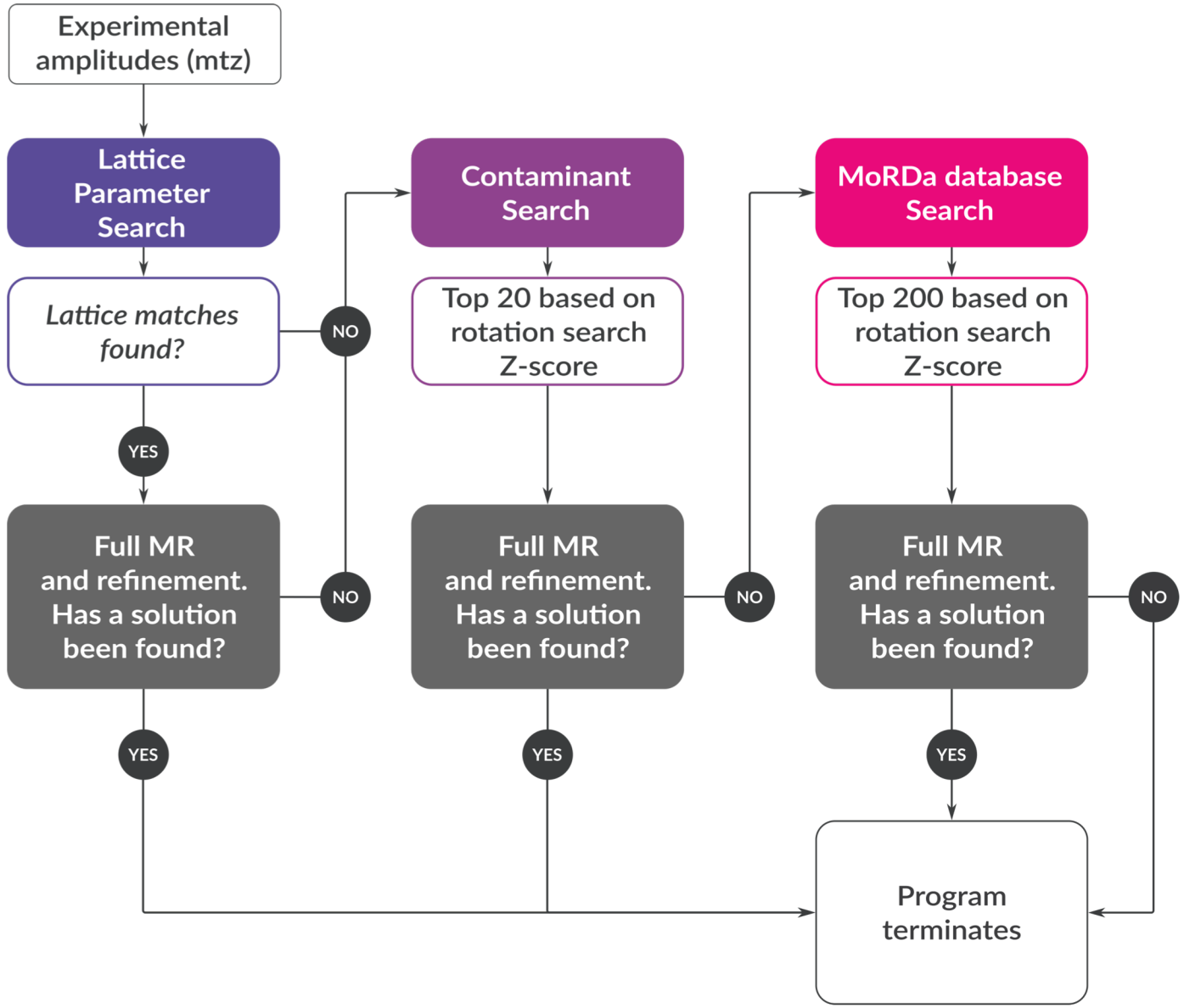
- However, it is worth mentioning that SIMBAD has been able to solve novel structures, for instance unsequenced proteins crystallised from snake venom.
  - Unfortunately I have no pretty pictures from that case ☹

Contaminant search

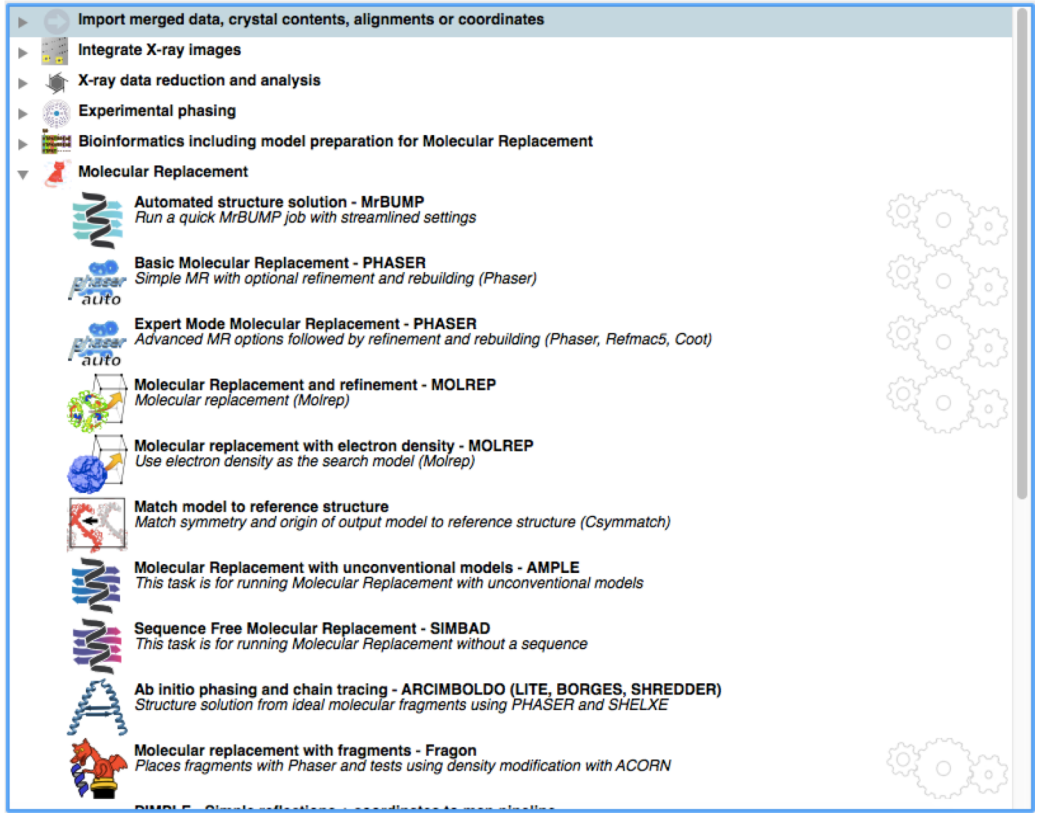
- Around 350 commonly known contaminants [Marcin Wojdyr @ CCP4/GlobalPhasing & Contabase]
  - Cover different expression organisms
- High screening performance achieved by AMoRe fast rotation cross function
  - ~15 min on desktop with 4 cores
- Allows us to find known contaminants in novel crystal forms.

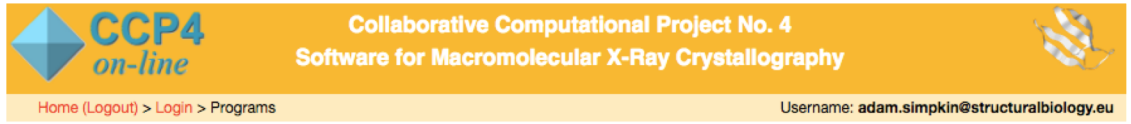
Lattice parameter search

- The lattice parameters from the experimental data are compared to a database of all the lattice parameters in the PDB.
  - This search is extremely quick
- Allows us to find novel contaminants in known crystal forms.


Experimental amplitudes (mtz)
Lattice Parameter Search
Lattice matches found?
NO
YES
Full MR and refinement. Has a solution been found?
NO
Contaminant Search
Top 20 based on rotation search Z-score
Full MR and refinement. Has a solution been found?
NO
YES
MoRDa database Search
Top 200 based on rotation search Z-score
Full MR and refinement. Has a solution been found?
NO
YES
Program terminates

Availability of the programs

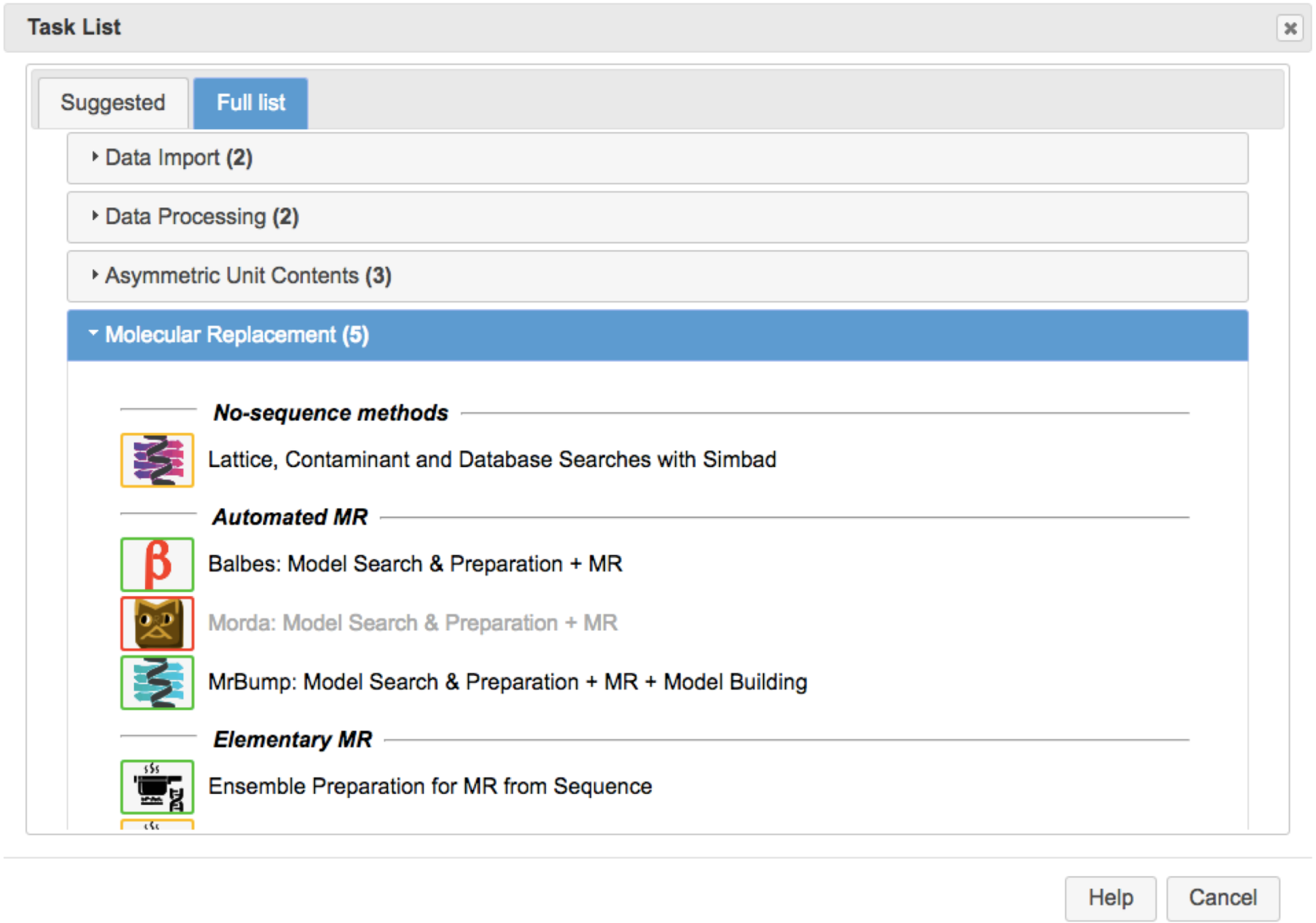




Programs

Note: *You must have a CCP4 licence to run these programs.*

Balbes	An automated Molecular Replacement (MR) pipeline - Balbes integrates into one system all the components necessary for solving a crystal structure by Molecular Replacement
MrBUMP	An automated Molecular Replacement (MR) pipeline - Given a target sequence and experimental structure factors, it will search for homologous structures, create a set of suitable search models from the template structures, do molecular replacement, and test the solutions with some rounds of restrained refinement.
Zanuda	Space group and crystallographic origin validation
AMPLE	Automated ab initio search model generation for molecular replacement.
SHELX	Automated SHELXC/D/E structure solution pipeline for fast routine experimental phasing. Accepts data in XDS, Scalepack, SHELX hkl or mtz formats and outputs phases and a poly-Ala trace. If a protein sequence is provided, BUCCANEER and REFMAC complete the structure.
CRANK2	Automated structure solution pipeline for experimental phasing using maximum likelihood methods.
MoRDa	MoRDa is a pipeline for molecular replacement protein structure solution based on its own domain database. Models relevant to the target sequence are further adjusted before molecular replacement search.
SIMBAD	Sequence-independent molecular replacement, good for identifying if your crystal contains a contaminant protein. SIMBAD can also do full search of homologous structures in difficult-to-solve novel target cases, but this functionality is not yet available through CCP4-Online.


Task List
Suggested
Full list
Data Import (2)
Data Processing (2)
Asymmetric Unit Contents (3)
Molecular Replacement (5)
No-sequence methods
Lattice, Contaminant and Database Searches with Simbad
Automated MR
Balbes: Model Search & Preparation + MR
Morda: Model Search & Preparation + MR
MrBump: Model Search & Preparation + MR + Model Building
Elementary MR
Ensemble Preparation for MR from Sequence
Help
Cancel

Acknowledgements

Daniel Rigden

Ronan Keegan

Charles Ballard

Villi Uski

Andrey Lebedev

William Shepard

Martin Savko

Felix Simkovic

Jens Thomas

Jaclyn Bibby

ample.readthedocs.io

Bibby *et al.* (2012). Acta Cryst. D**68**, 1622-1631.

simbad.readthedocs.io

Simpkin *et al.* (2018). Acta Cryst. D**74**, 595-605.