Molecular Replacement

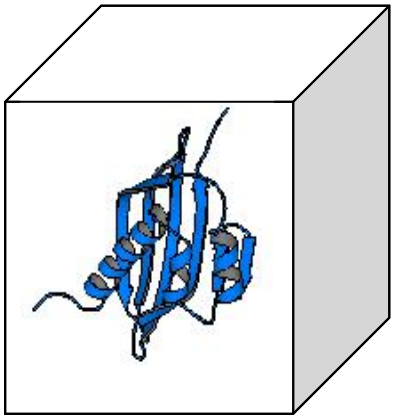
Airlie McCoy

UNIVERSITY OF CAMBRIDGE

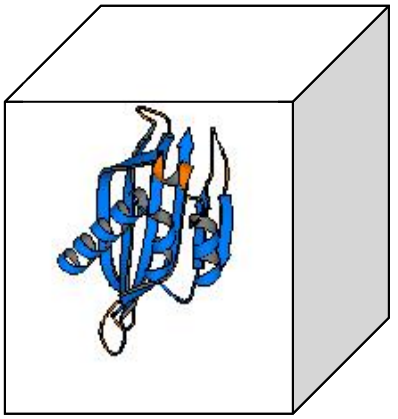
Molecular Replacement

- Find orientation and position where model overlies the target structure
- **Borrow** the phases
- Then it becomes a refinement problem – the phases change

known structure



unknown structure



H K L	F	ϕ
0 0 1	12.6	120
0 0 2	2.1	10
0 0 3	69.9	280

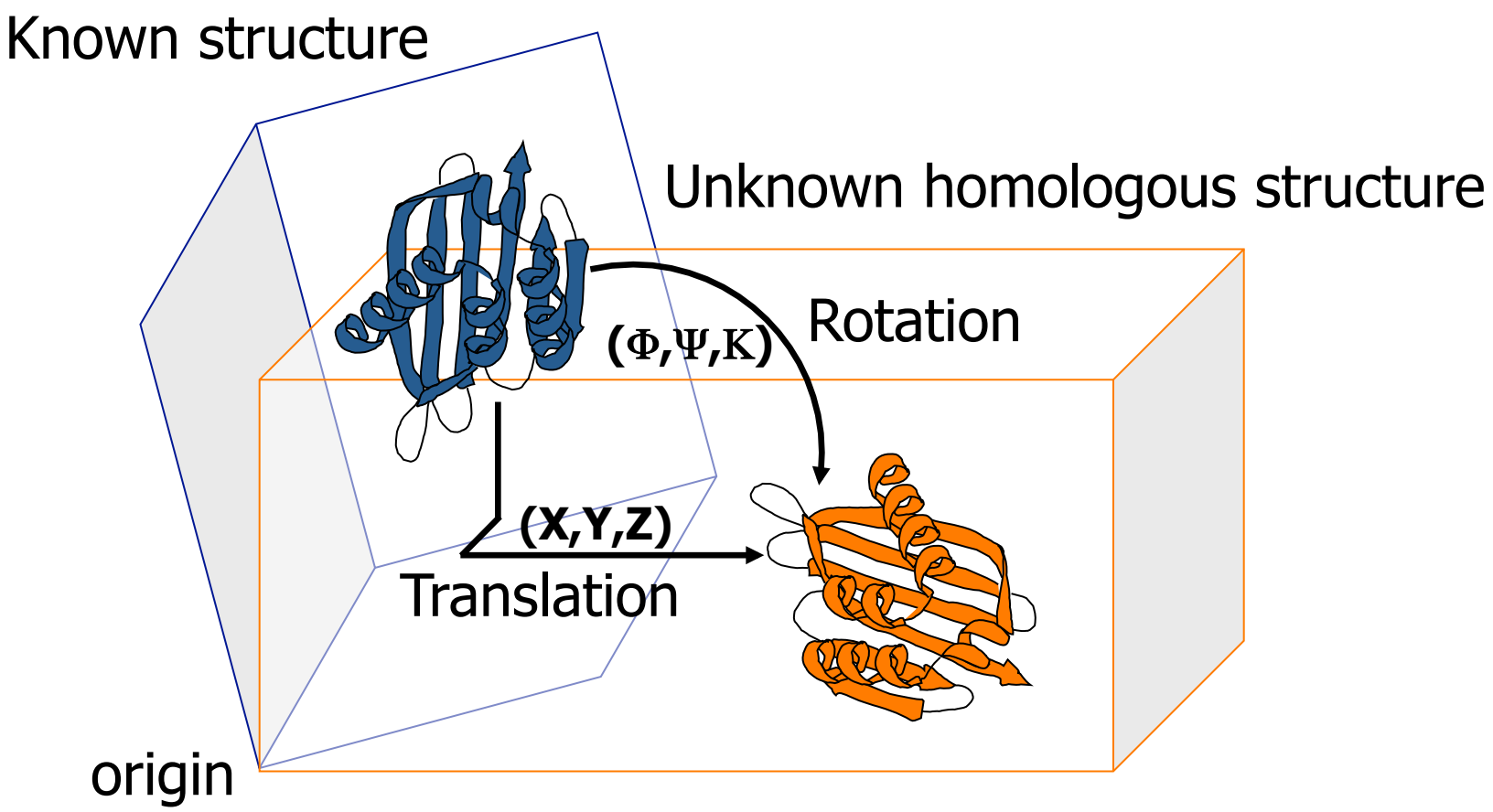
etc...

copy →

H K L	F	ϕ
0 0 1	10.4	120
0 0 2	3.1	10
0 0 3	52.2	280

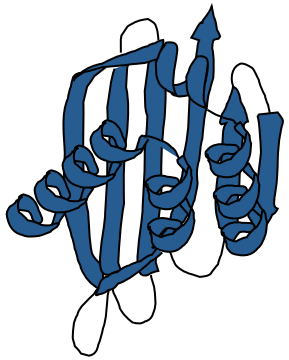
etc...

Molecular Replacement



1. Collect data

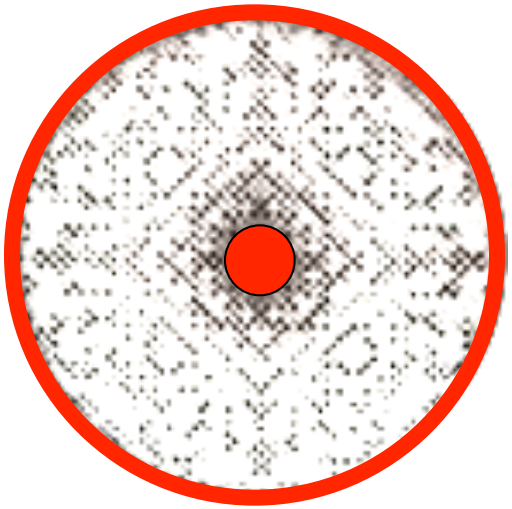
2. Find model

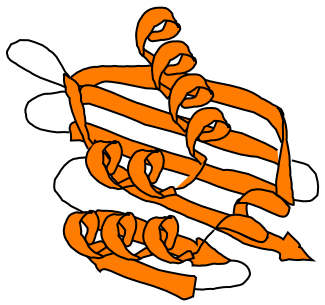


3. Rotate & translate model

4. Calculate "diffraction" from rotated & translated model

5. Compare and find best match

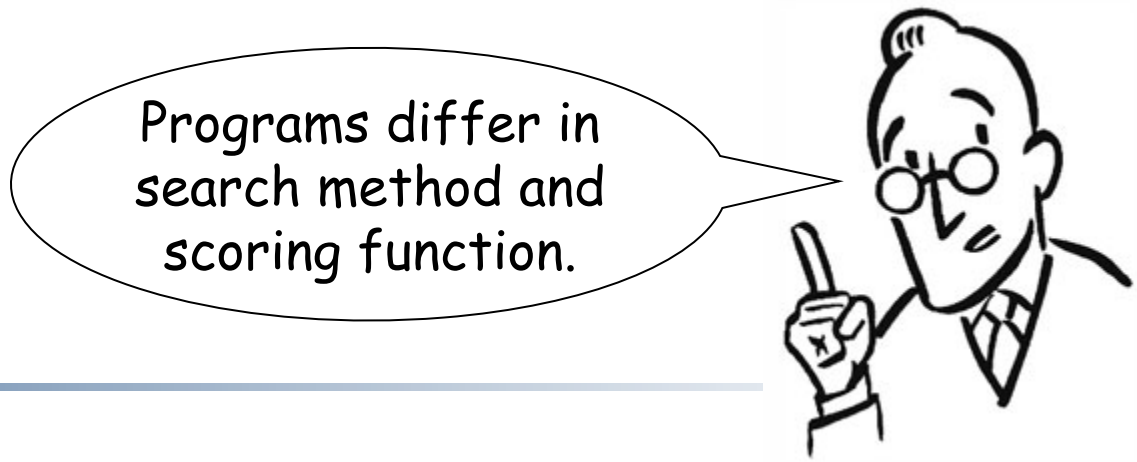




6. Phase observed data

Molecular Replacement

- Issues
  1. How to score each orientation and position so as to find when the model best fits the target structure
  2. How to search for solutions: strategies for exploring rotations and translations
- MR can fail due to suboptimal choices in either
- These issues are not independent



Software: Scoring

- MolRep
- AMoRe
- EPMR
- COMO
- CNS
- XPLOR
- SOMoRe
- Queen of Spades

Patterson

- Phaser

Maximum Likelihood

Software: Search

- MolRep
- AMoRe
- CNS
- XPLOR

} 2x3D, independent molecules

- EPMR
- COMO
- SOMoRe
- Queen of Spades

} 6D (but not exhaustive)

- Phaser

} 2x(3D,3D), cumulative, amalgamation

Some Important Features of Phaser

- The likelihood scoring functions can take account of errors in the model and the data
  - Intensity based targest
- The scoring functions allow solutions to be built up by addition
- Expected LLG allows artificial intelligence to be built into resolution selection, search order, and termination criteria
  - Bias free occupancy refinement
- Translational NCS correction allows new classes of structures to be solved by MR

Year	Phaser	PDB	X-ray	Phaser % X-ray PDB	Phaser % total PDB
2004	13	5181	4424	0.3	0.3
2005	214	5361	4452	4.8	4.0
2006	518	6474	5567	9.3	8.0
2007	876	7199	6206	14.1	12.2
2008	1190	6959	6248	19.0	17.1
2009	1648	7385	6746	24.4	22.3
2010	2122	7901	7296	29.1	26.9
2011	2712	8087	7496	36.2	33.5
2012	3055	8903	8268	36.9	34.3
2013	3662	9603	8852	41.3	38.1

Phaser has been used to solve 20% of the X-ray PDB

General search strategy

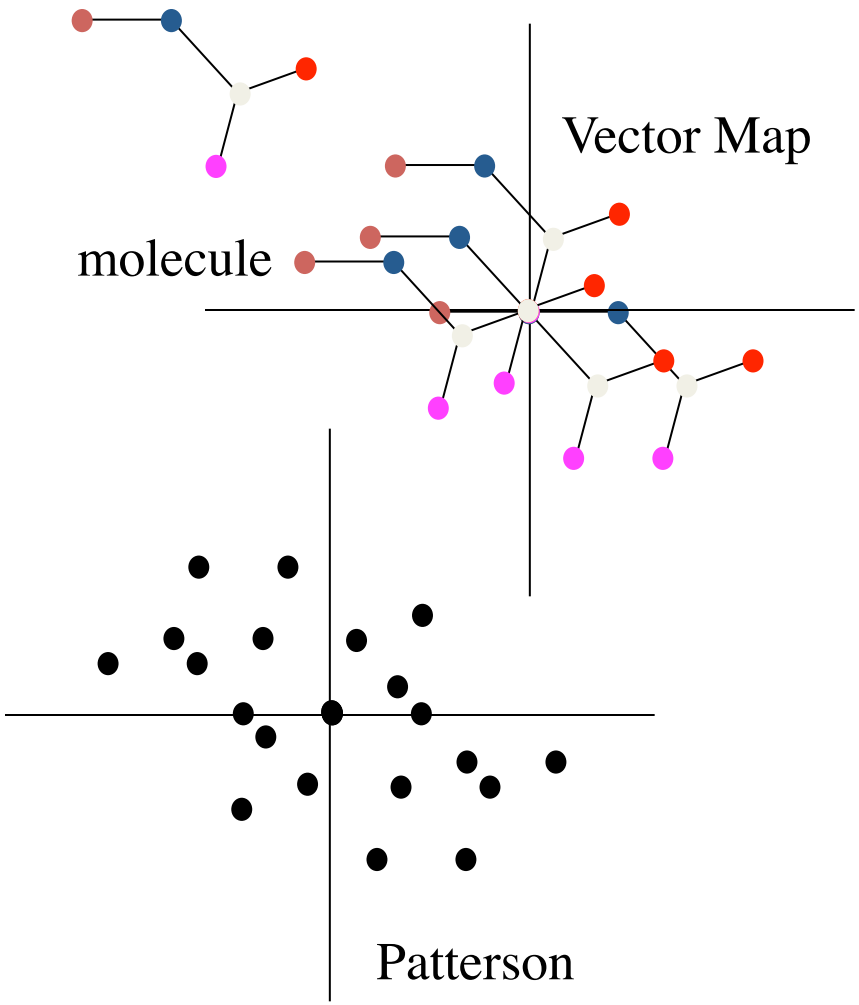
- Each molecule needs 6 parameters (6D)
- An exhaustive search is too big
  - All angles sampled at 2.5°; $N_{rot} = 1.5\times10^{6}$ points
  - All positions sampled at 1Å in a 100Å cubic cell; $N_{tra} = 1.0\times10^{6}$ points
  - 6 dimensional search is $N_{rot}\times N_{tra} = 1.5\times10^{12}$ points
  - Programs that do "6 dimensional" searches use genetic, random or limited sampling
- MR search strategies can be divided into rotation and translation separately (2x3D)
  - $N_{rot}+N_{tra} = 2.5\times10^{6}$ points

Scoring

- What is the "best match" between the observed and calculated structure factors for the rotation function and translation function
  - 2x3D = rotation + translation function targets
  - 6D search = translation function target
- Can use
  - Patterson methods
  - Maximum Likelihood methods

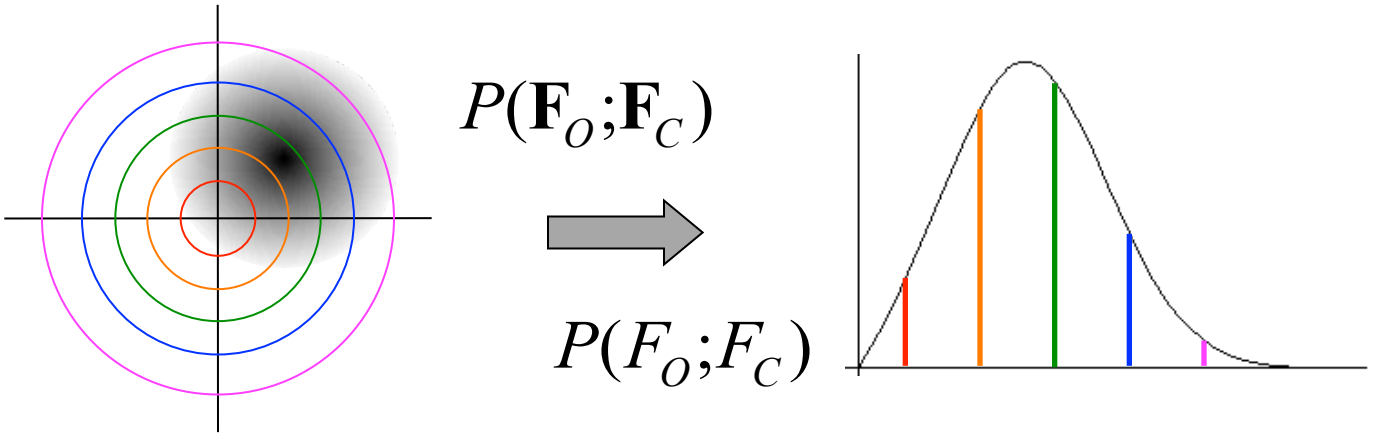
Patterson

- Patterson is the FT of the amplitude^2 and the phases set to zero
  - Can be calculated from the intensities
  - This is the vector map of the atoms
- So we are lucky – we have a function that we can calculate AND that is useful



Maximum Likelihood Scoring

- Use **probability**
- Probabilities account for errors
  - Patterson methods cannot do this!

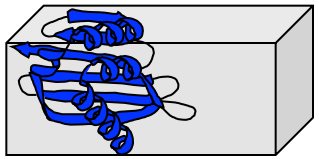


$$MLTF = \min\left[-\sum_{\mathbf{h}} \log\left(\frac{2F_O}{\sigma_\Delta^2} e^{-\frac{F_O^2 + D^2 F_C^2}{\sigma_\Delta^2}} I_0\left(\frac{2F_O D F_C}{\sigma_\Delta^2}\right)\right)\right]$$

Brute translation function search

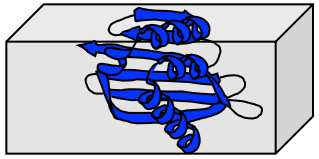
- Place model at points in unit call and calculate probability that it is in each position

Search orientation and position #1



Etc…

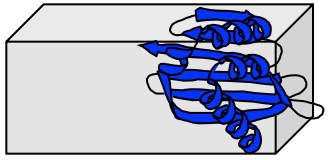
Search orientation and position #7211



Best "match" of structure factor amplitudes

Etc…

Search orientation and position #9999

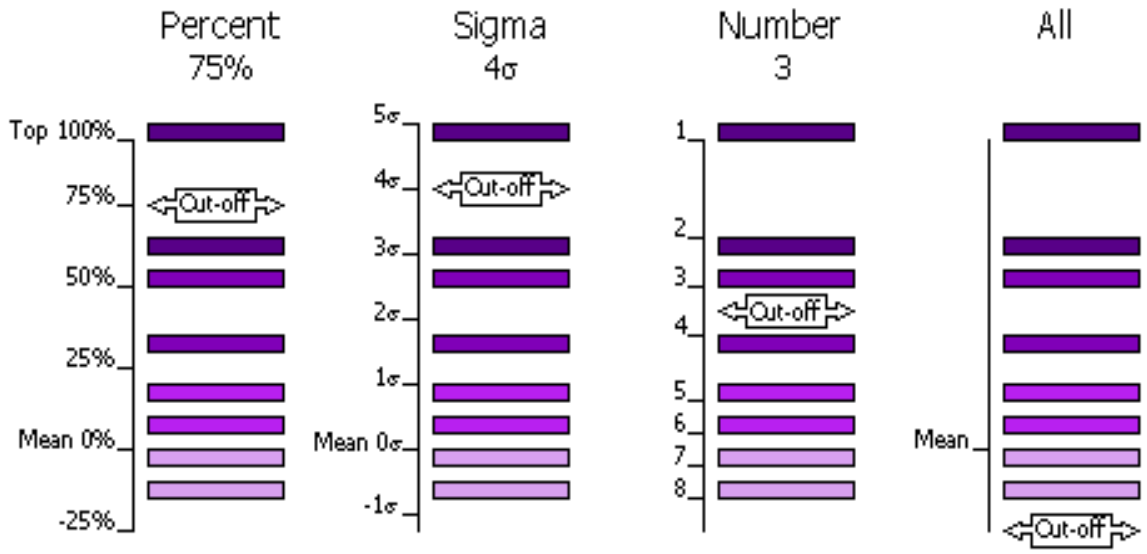


Fast Searches

- ML searches are very slow
- An approximation can be calculated by fast Fourier transform (FFT)
  - Values for all orientations/positions obtained at once
  - FFT function doesn't discriminate the "correct" solution as well as the full ML functions
- Rescore the top solutions with the slower but more accurate ML functions
- As long as the "correct" orientation/position was in the list from the fast search, it should rise to the top of the rescored list
  - If not, then more fast search points must be rescored

Peak selection

- Must chose a selection criteria to carry potential solutions through to the next step

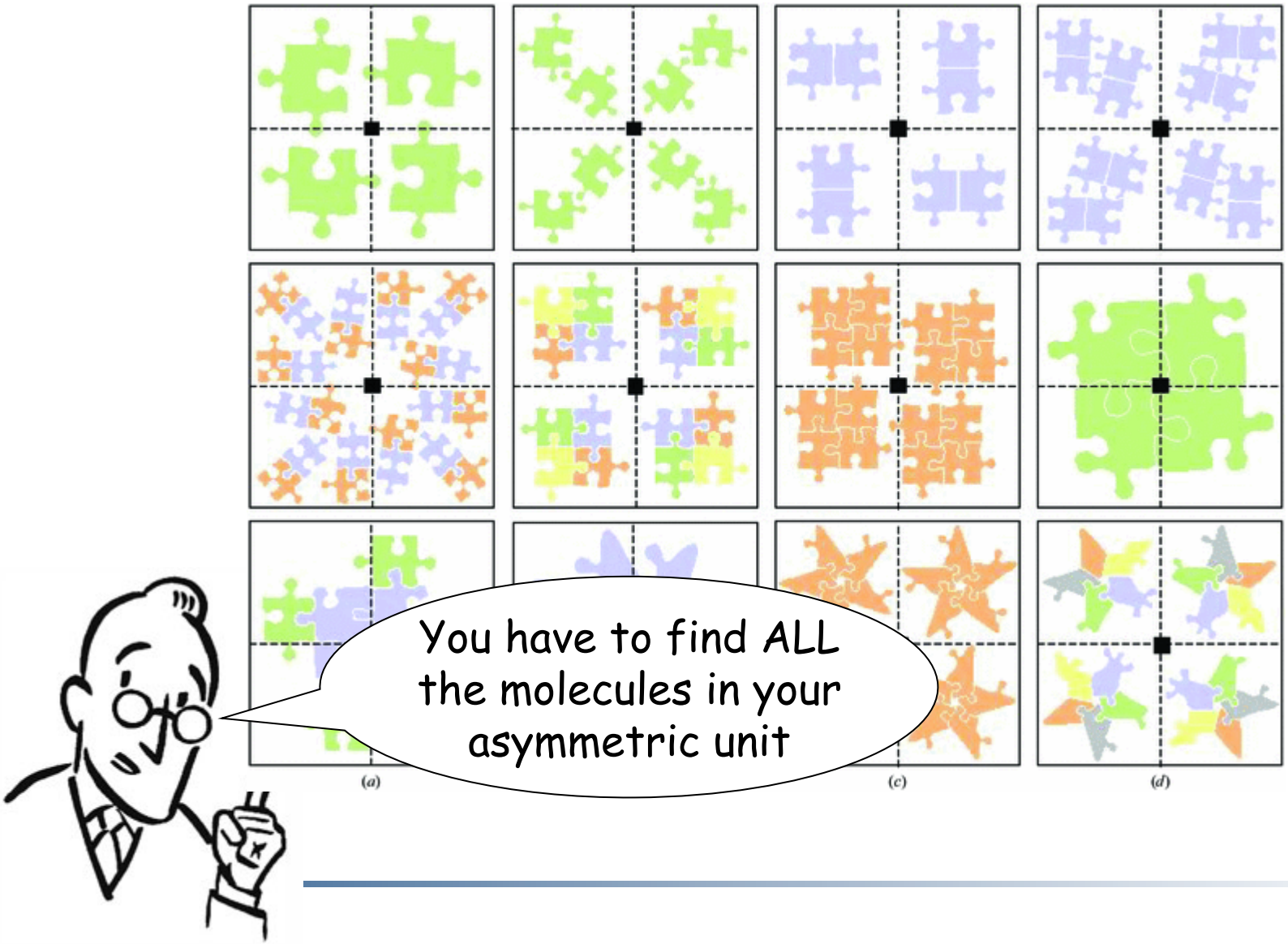


- By default, solutions over 75% of the difference between the top peak and the mean are selected

Using Partial Structure

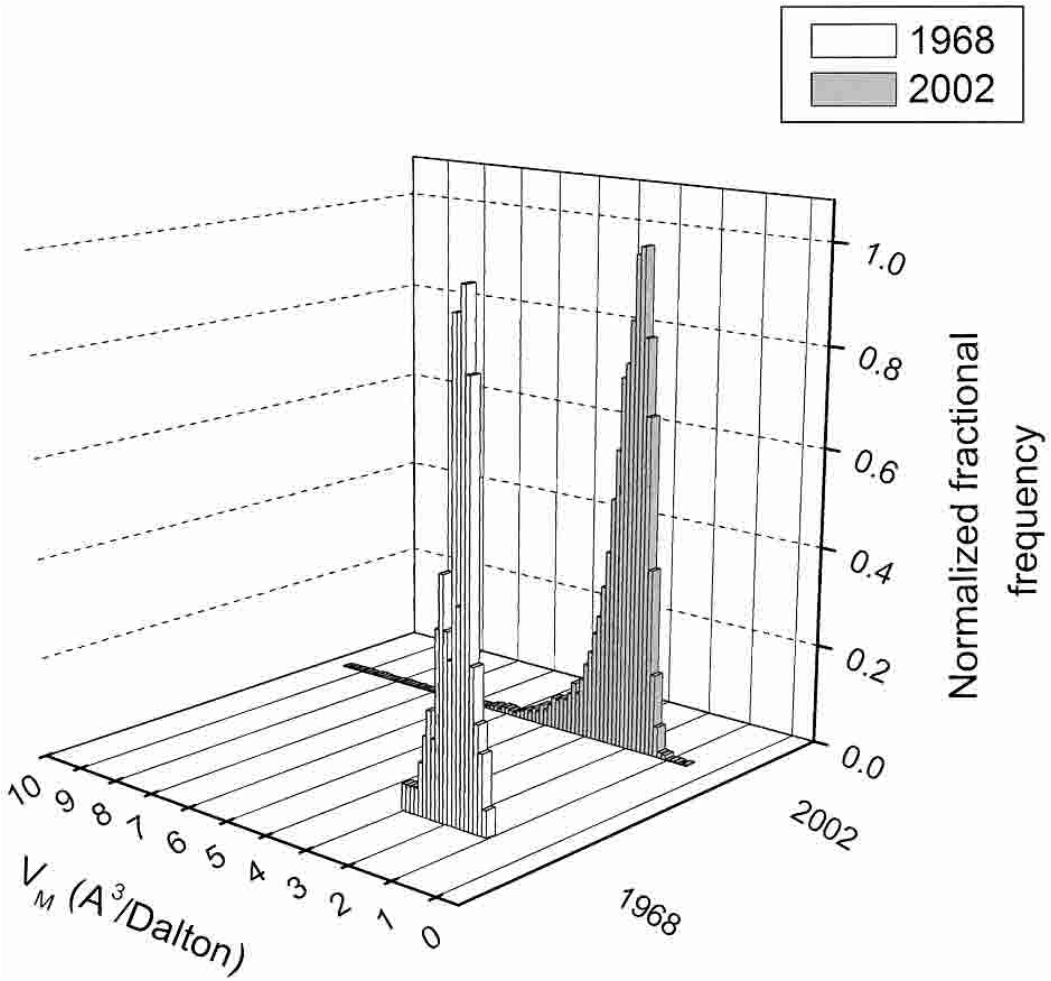
- The MLRF and MLTF can use models that have already been placed in the asymmetric unit
- Patterson RF cannot account for placed models
  - The translation Correlation Coefficient can account for known positions…
  - But most of the difficulty in MR is the rotation function, because the signal to noise ratio is much lower

Contents of the Asymmetric Unit

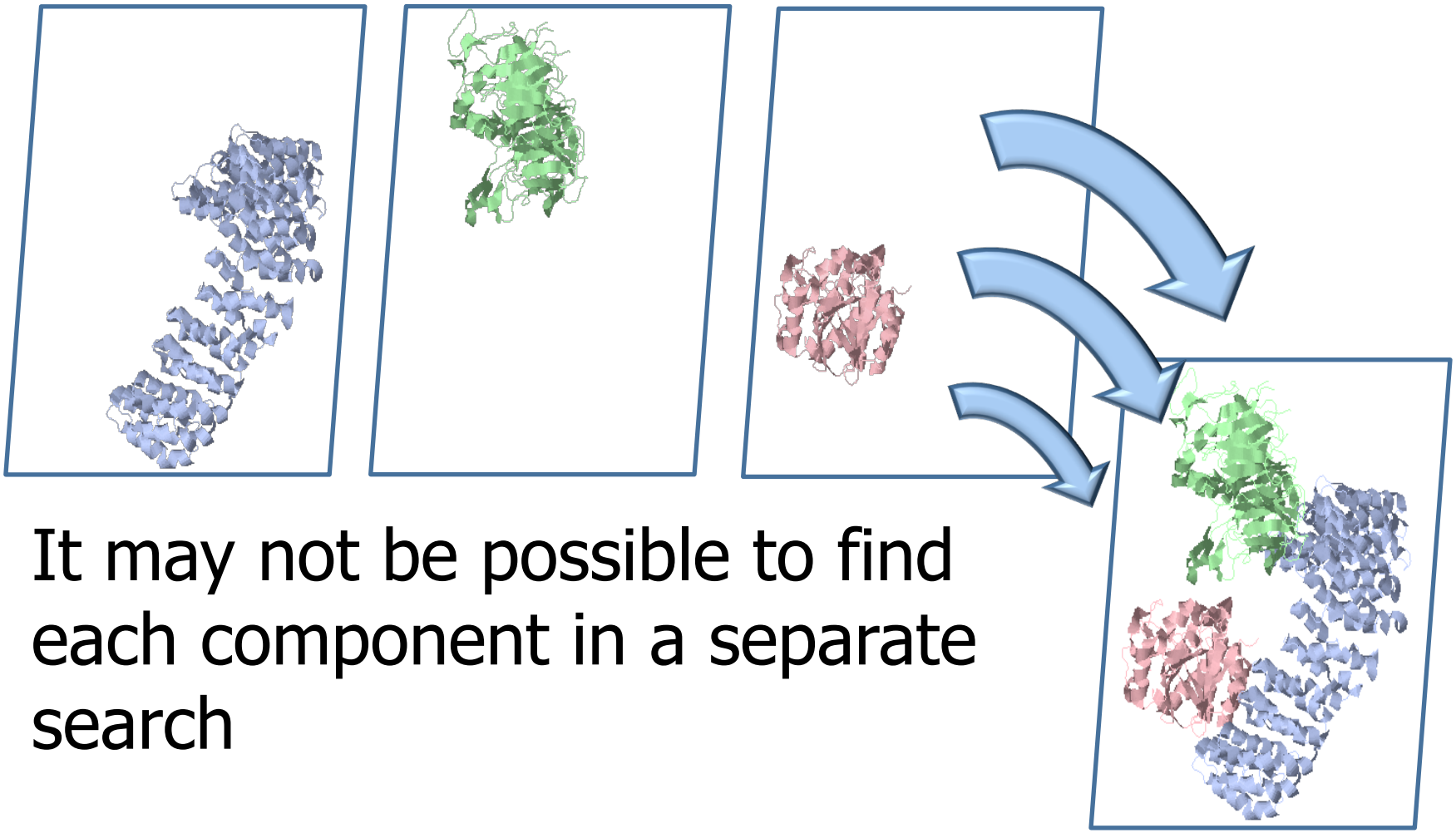


Matthew's coefficient

- First calculated by Brian Matthews in 1968 (over 3500 citations)
- Most crystals are 50% protein by volume
- Can be used to estimate the contents of the asymmetric unit



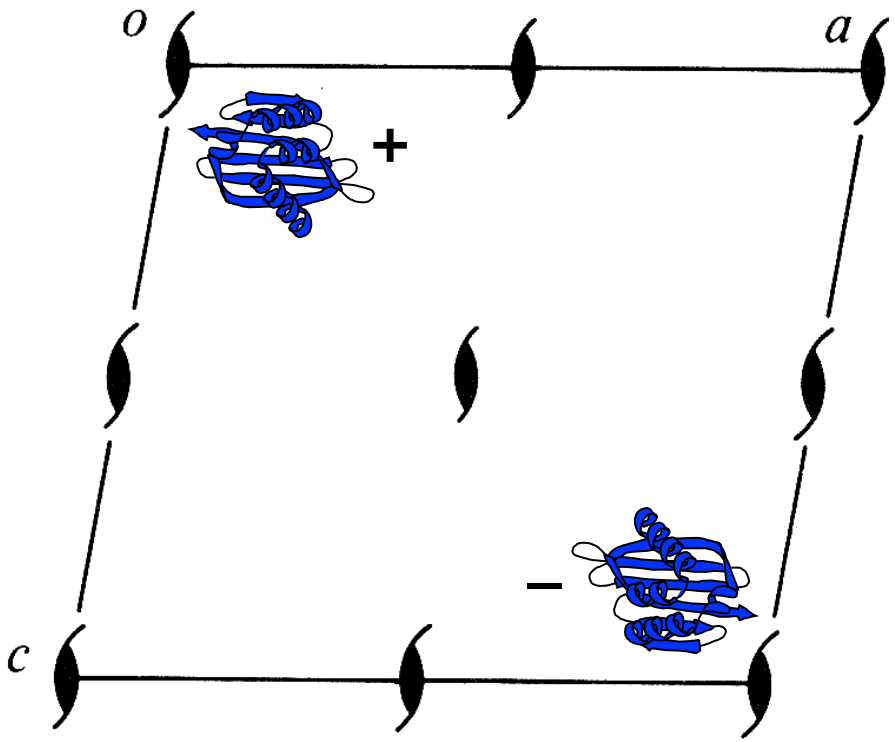
Searches for multiple components



It may not be possible to find each component in a separate search

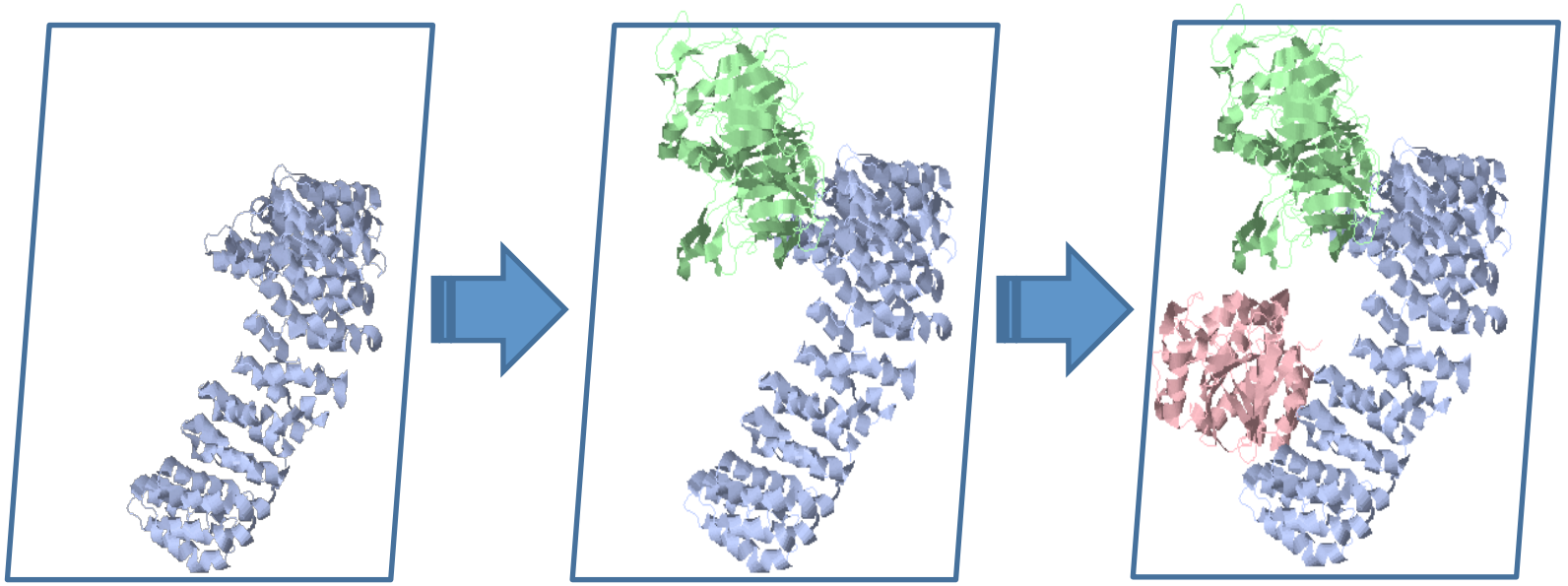
Origins

- Can only find the translation perpendicular to a rotation axis
  - If an axis has no rotation symmetry, there is no translation to find!
- If there are multiple symmetry axes of the same type (i.e. same order of rotation) in the same plane then the translation can be defined with respect to any one of these
  - These are equivalent to different choices of origin
  - Different MR solutions may be on "different origins" and look different when they are really the same


o
a
+
–
c

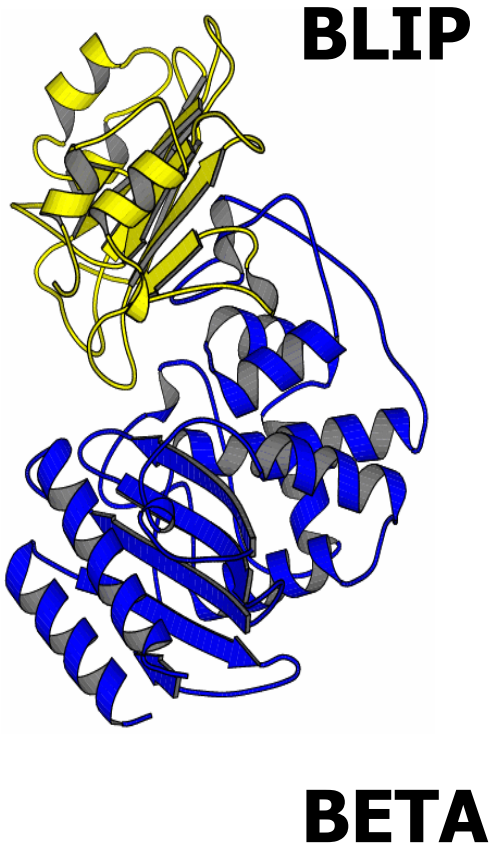
Searches for multiple components

- Maximum Likelihood (ML) functions are able to include partial structure information from previous placements
- Structure builds up by addition



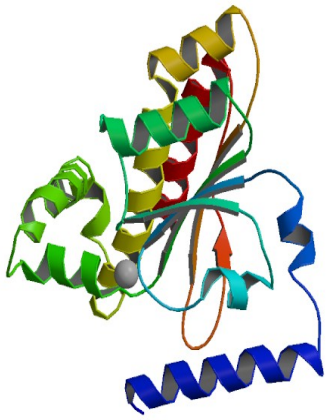
β-lactamase/β-lactamase inhibitor

- BETA and BLIP previously solved separately
- Structure of complex solved with Amore
  - BETA 62% residues - easy
  - BLIP 38% residues - difficult
  - BLIP found by selecting 1000 RF peaks and running a TF for each
- Data are very anisotropic
- Trivial with Phaser



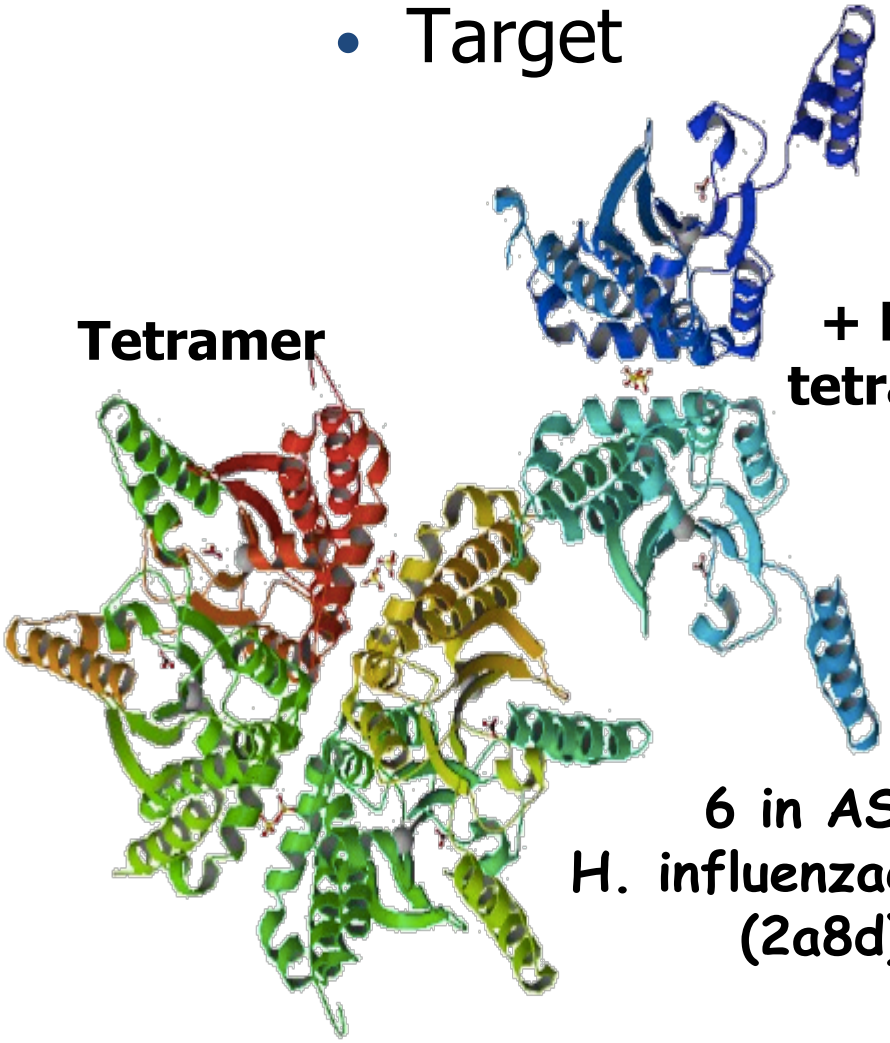
β-Carbonic Anhydrase

- Model



E.coli β-CA
(1i6p)
61% identity

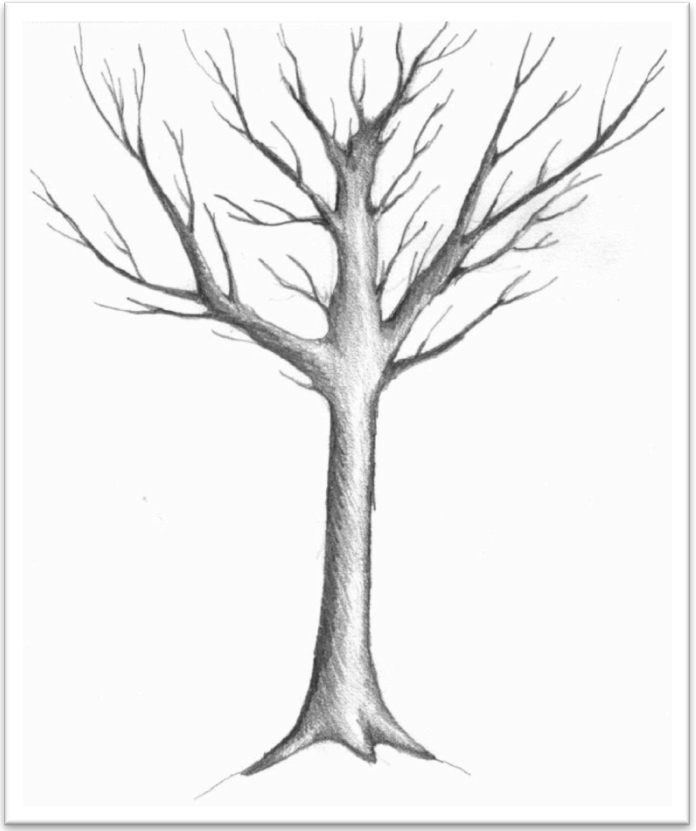
- Target



6 in ASU
H. influenzae β-CA
(2a8d)

Multi-copy searches

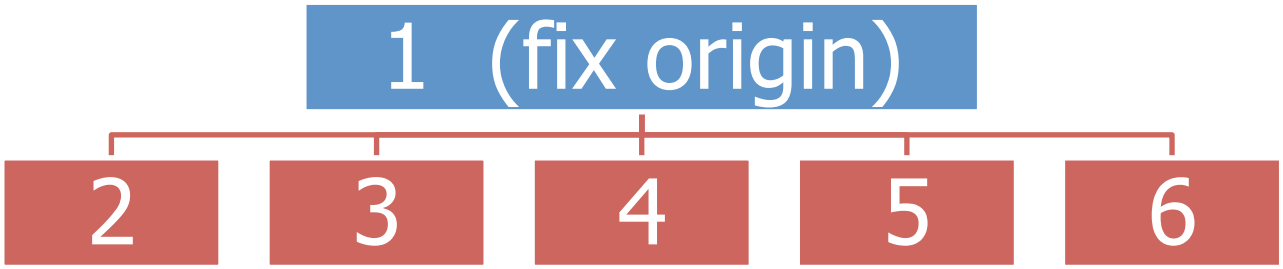
- One RF finds all orientations
- One TF for each orientation finds component
- Tree search generates a heavily branched search
- All solutions equivalent after 6 sequential RF/TF searches



Naive search does more work than necessary

Fast Search Algorithm

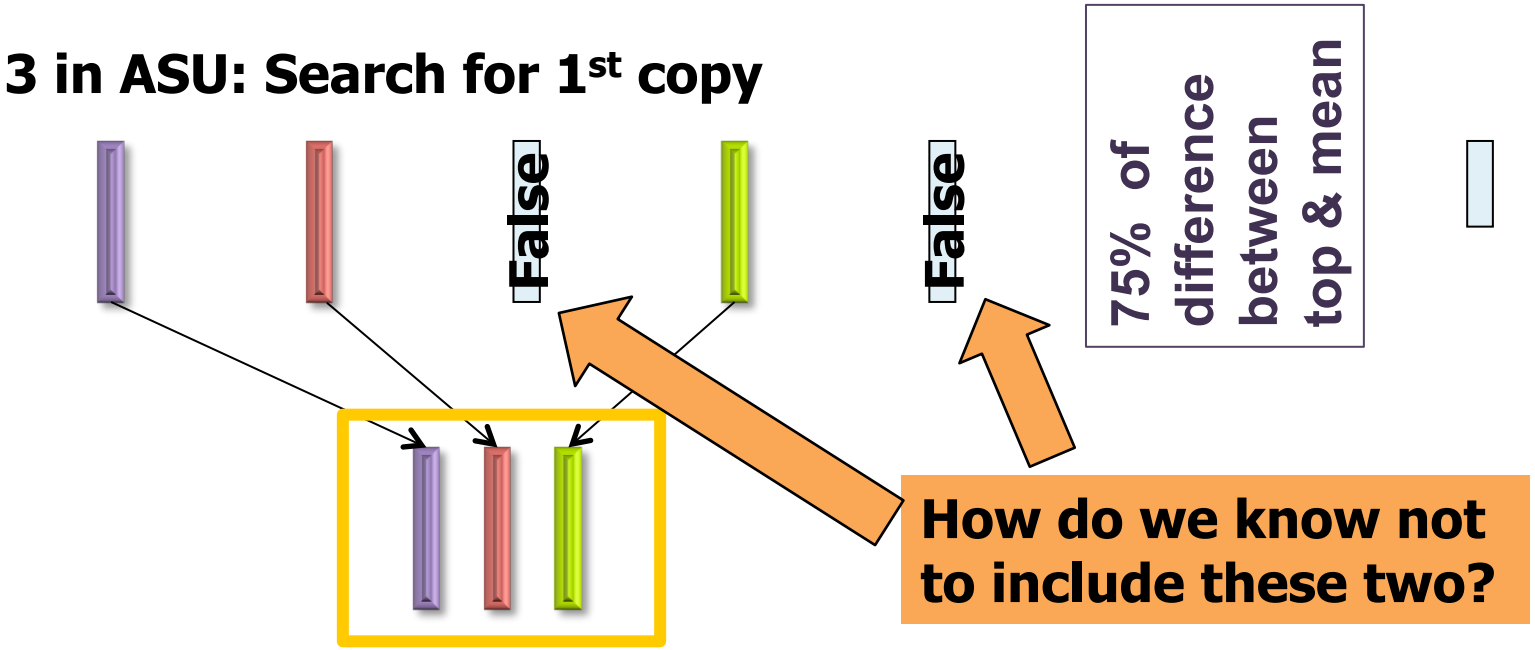
- Phaser has a search algorithm that amalgamates more than one solution per RF/TF pair
- Fixes origin with first, then reuses RF peaks to find other placements



When is a model correctly placed?

TF Z-score	Solved?
< 5	no
5 - 6	unlikely
6 - 7	possibly
7 - 8	probably
> 8	definitely

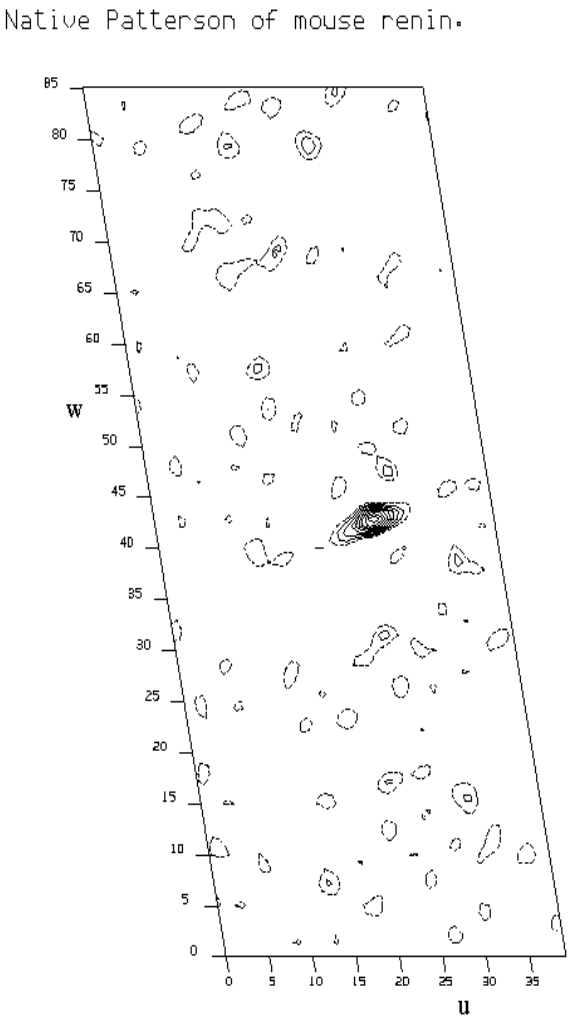
Fast Search: Amalgamation



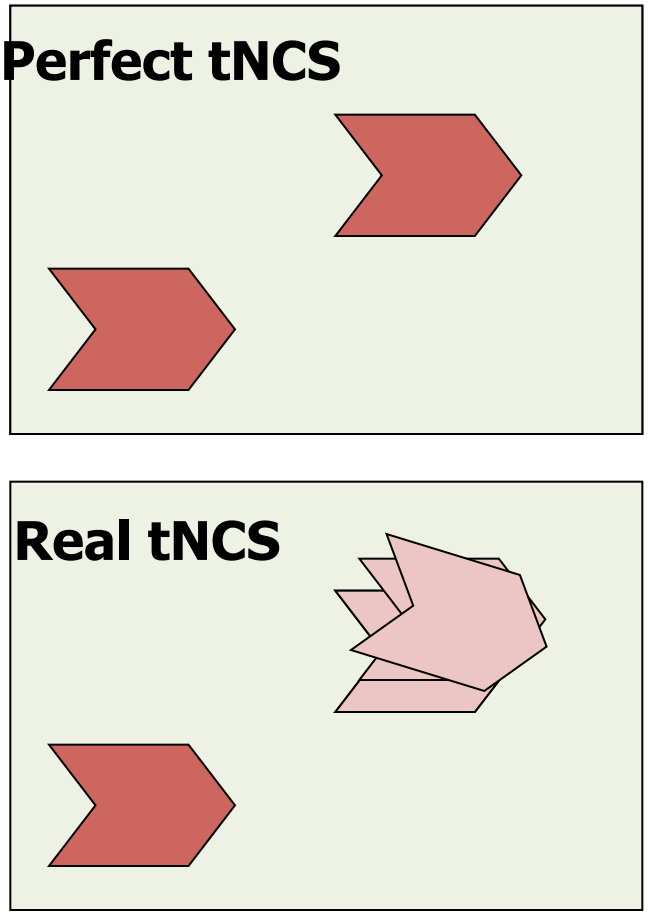
1. Amalgamated placements must have TFZ > 8
2. Amalgamated placements must pack
3. Amalgamated placements must increase LLG

Translational NCS

- If tNCS is not accounted for then TFZ > 8 does not indicate a correct placement
  - TFZ values are always higher
  - TFZ > 12 can be wrong
- When TFZ is accounted for the TFZ values are those expected of data without tNCS



Translational NCS



- Three tNCS parameters refined from data alone
  1. Rotation
  2. Translation
  3. RMSD between copies
- tNCS correction factors used in MR and SAD
- Two classes of tNCS cases accounted for
  - These cover the majority of cases

Translational NCS

1. Pairs of molecules related by one vector
   - One peak in Patterson
   - Molecules in pairs
   - There can be any number of pairs of molecules related by the same tNCS vector
2. Molecules related by multiples of one vector
   - Peaks in Patterson are multiples of same vector
   - Molecules in sets related by same vector
   - There can be any number of sets of molecules related by the same tNCS vector

Twin Detection

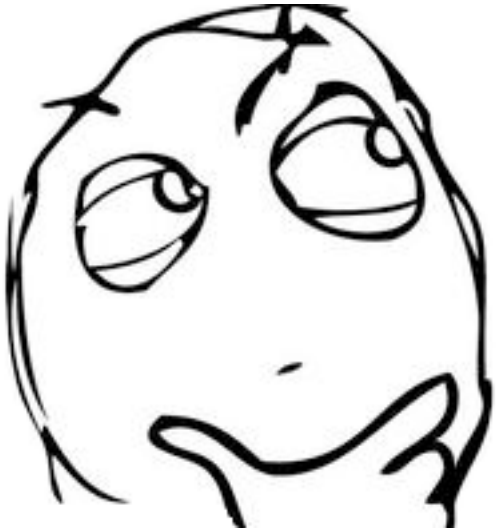
- Translational NCS masks twinning
- Correcting the data for tNCS unmasks twinning
- Phaser generates cumulative intensity plots for centric and acentric reflections after correction for tNCS and anisotropy
- Phaser gives a P-value for the probability of twinning

Models

Devolution

Will MR work for me?

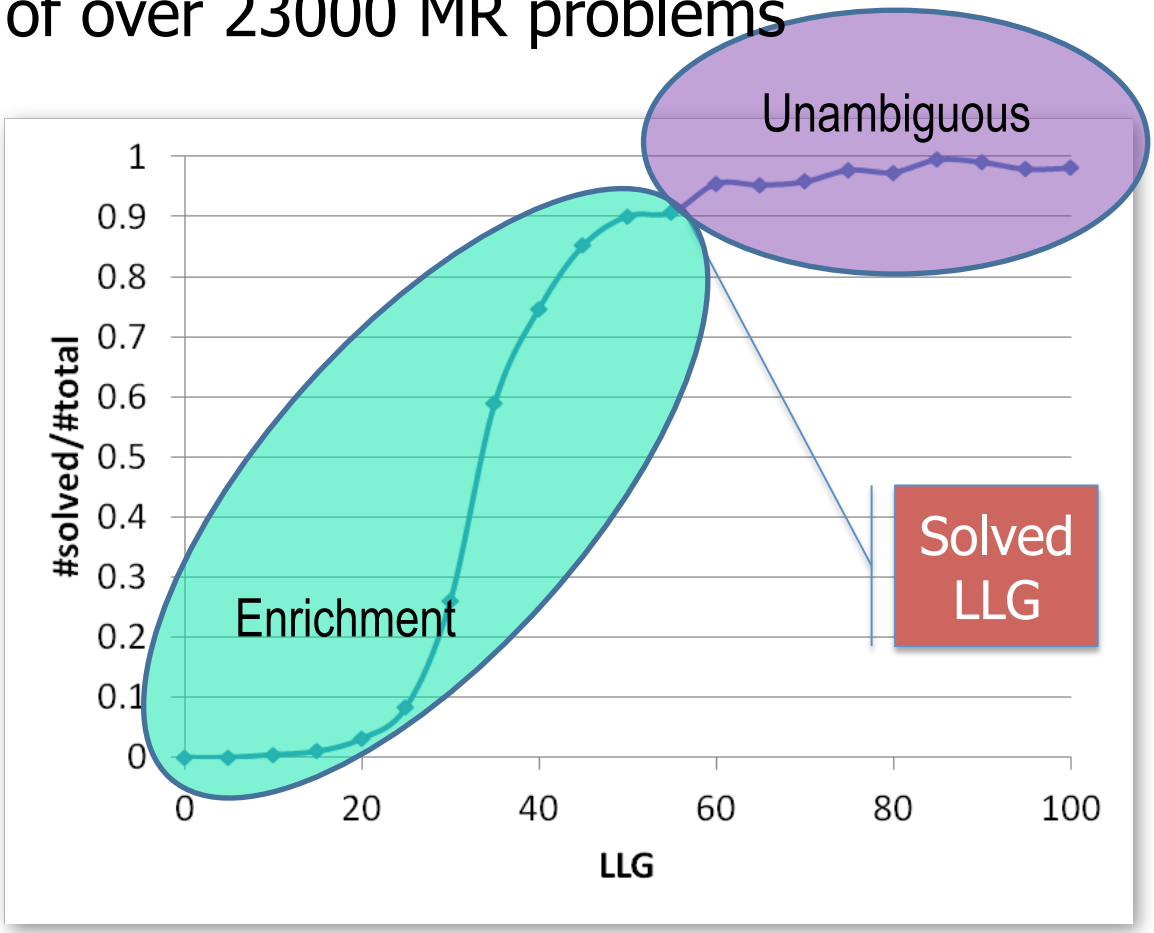
- Is my model good enough?
- Is my data good enough?
- Do I need to place multiple models simultaneously to get a signal?
- Will fragment-based MR work?
- Will α-helices work?
- How big does my helix/fragment have to be?
- Will single-atom MR work?



First, find the criteria for deciding that MR has worked!

Final LLG for MR solutions

Database of over 23000 MR problems



Plot of LLG versus success in structure solution

When is a model correctly placed?

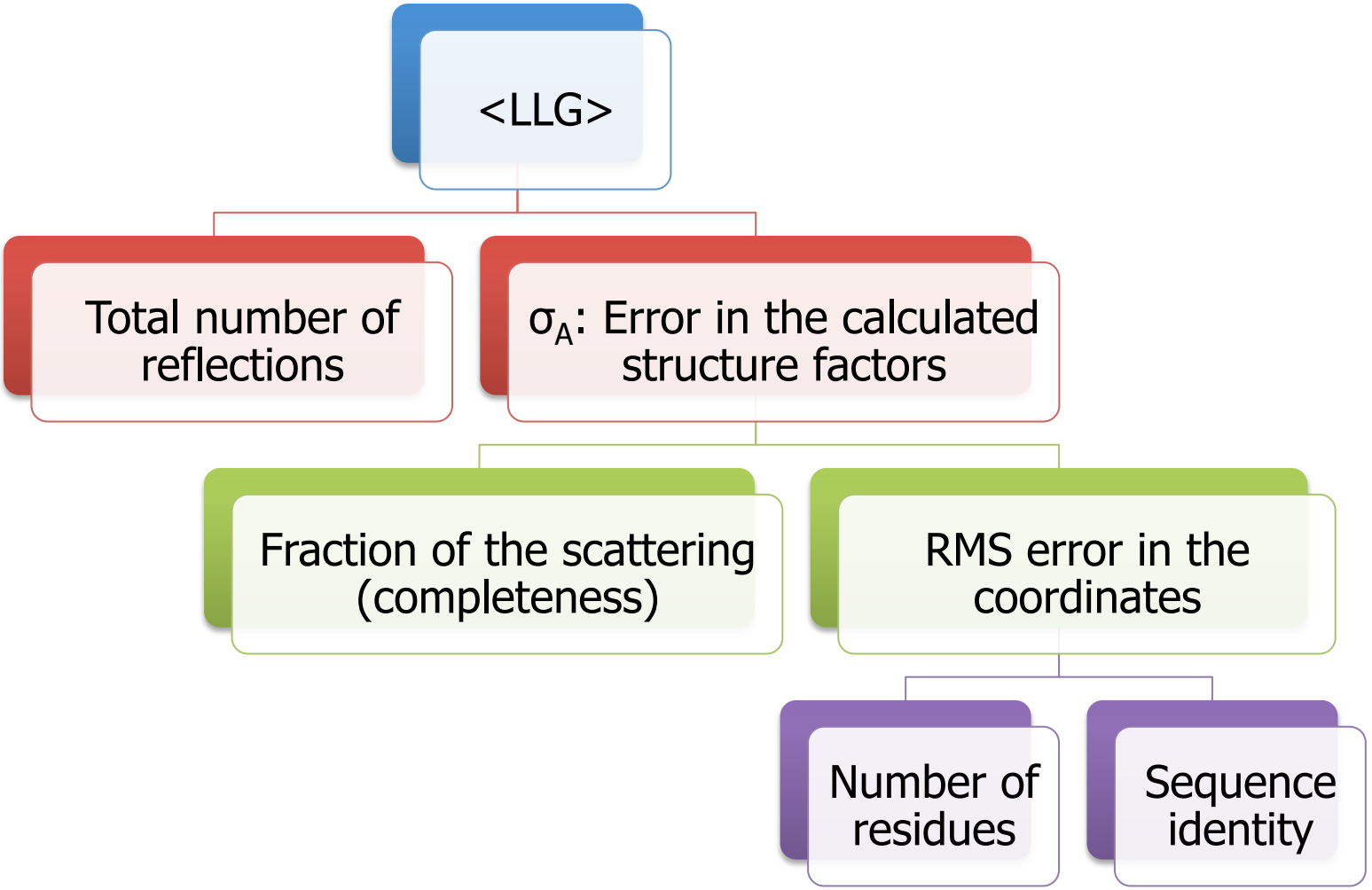
TF Z-score	LLG score	Solved?
< 5	< 25	no
5 - 6	25 - 36	unlikely
6 - 7	36 - 49	possibly
7 - 8	49 - 64	probably
> 8	> 64	definitely

Predicting LLG of solution

- So if you can predict the LLG…
  - You know how easy/difficult will be MR
  - You can prioritize structure solution strategies
- Removes uncertainty in MR
  - Knowing when to start/stop has always been the problem with MR

You can minimize the time to structure solution

Expected LLG

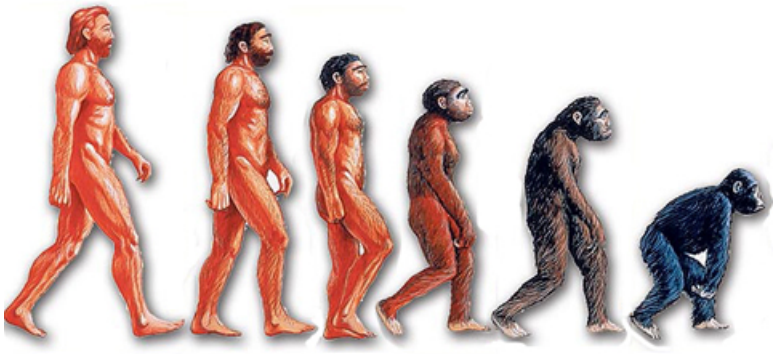


MR with Fragment/Atom

- Fragments or even atoms are just models with low completeness
  - Low σ_A
- Success of fragment/atom based MR relies on other contributions to the <LLG> being favourable
  - Lots of reflections
  - Low RMS

Success does NOT depend on
The resolution (strictly)
The type of fragment

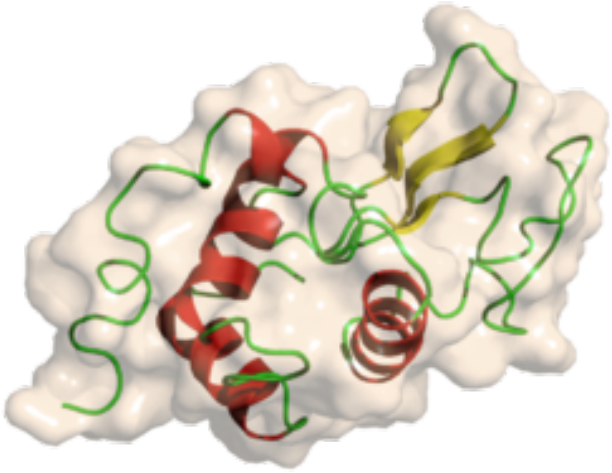
Devolution

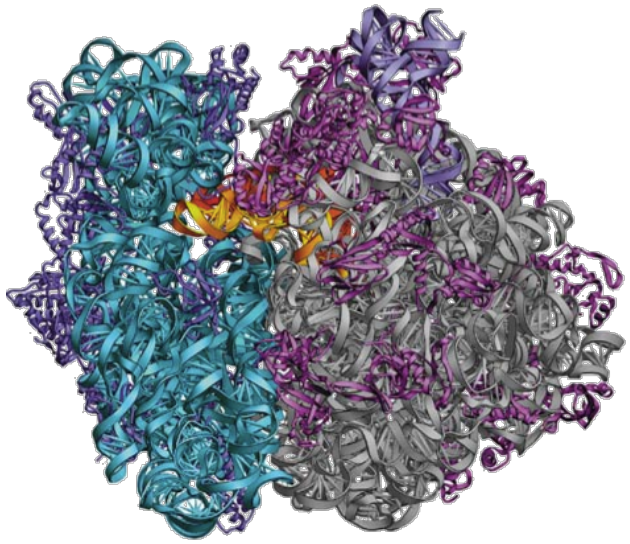


- ↓ Homologous structures
- ↓ Domains
- ↓ Ab Initio models
- ↓ Fragments
- ↓ Helices
- Atoms

<LLG> and Resolution

Example	Data	<LLG> target	resolution
HEWL	1.9 Å	120	5.6 Å
Ribosome	3.6 Å	120	10.8 Å





2hhr with CASP T0283

2hhr

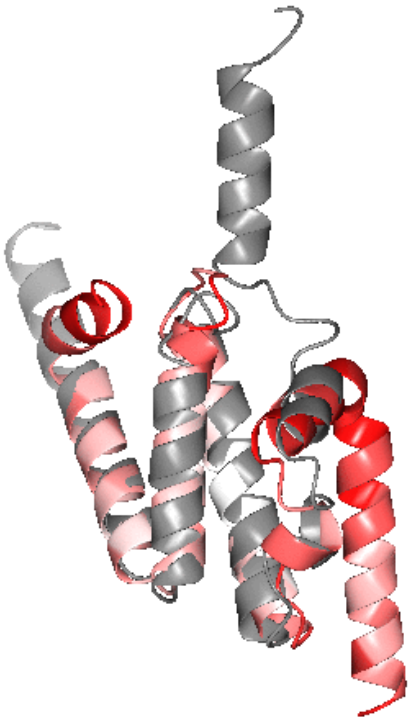
T0283

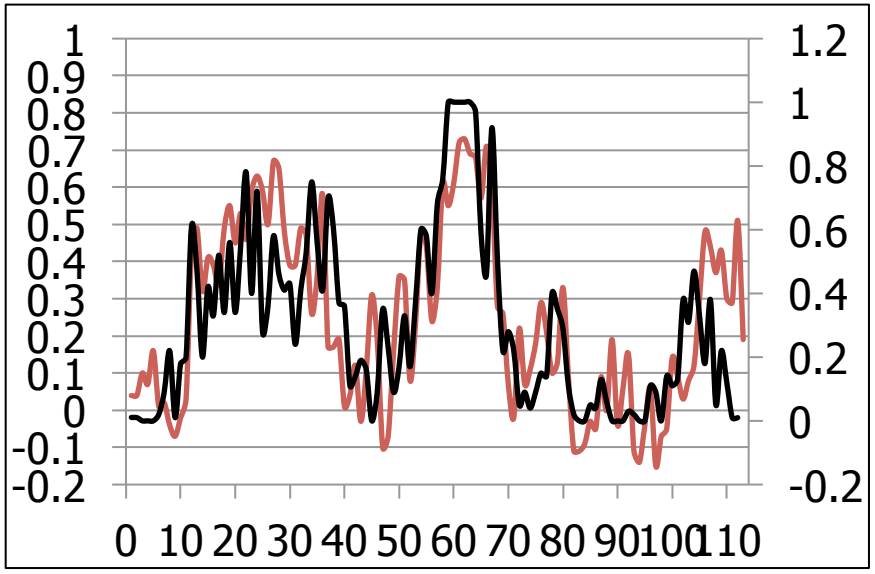
Predict expected change in LLG

- Total of 112 residues in 2hhr structure
- <LLG> of 55.6 for full model

	Fraction omitted	ΔLLG predicted
1 residue	0.009	1.0
Window of 3 residues	0.027	2.9
Window of 5 residues	0.045	4.9
Window of 11 residues	0.098	10.4

Occupancy refinement



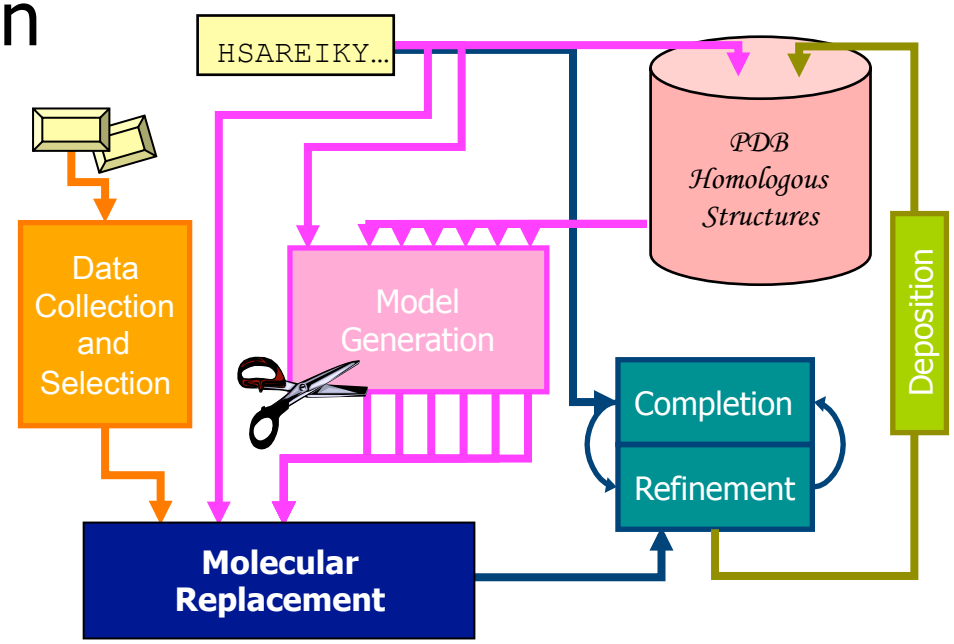


Pipelines

Evolution

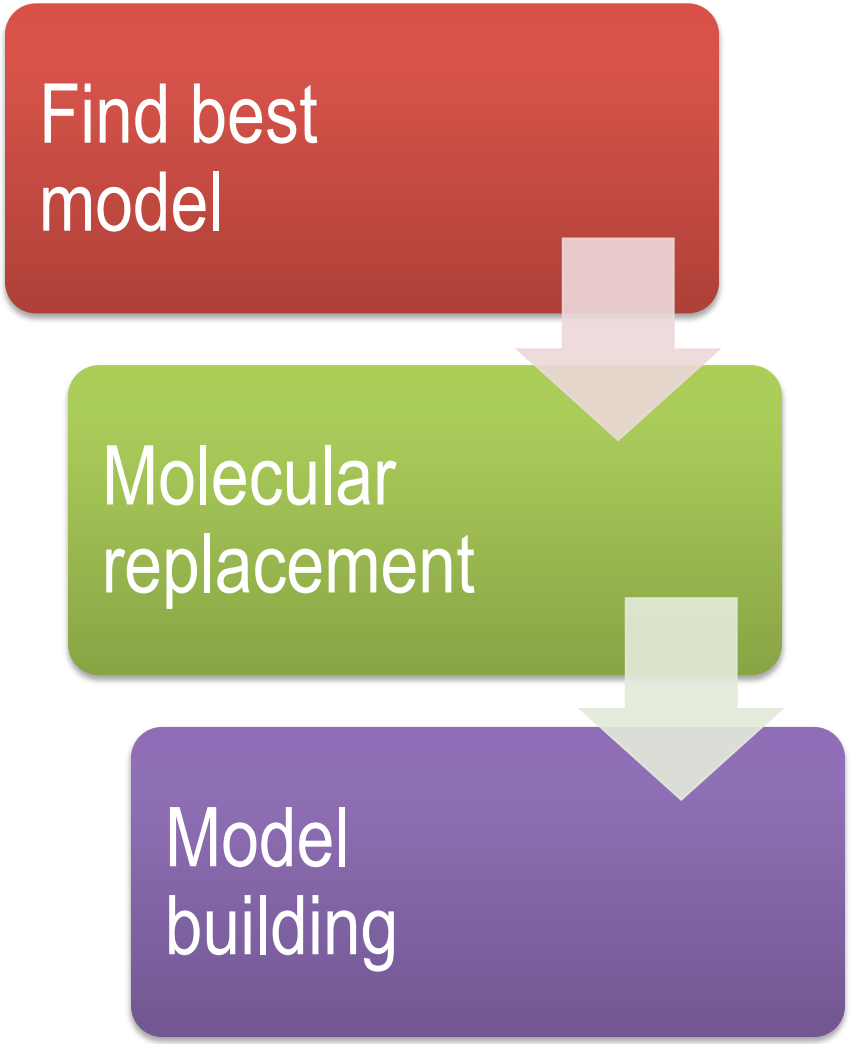
Pipelines for Molecular Replacement

- Phaser performs MR in
  - MrBUMP
  - Ample
  - phenix.automr
  - phenix.mr_rosetta
  - phaser.mrage
  - MRGrid
  - Arcimboldo
- Phaser does not perform MR in *Balbes*



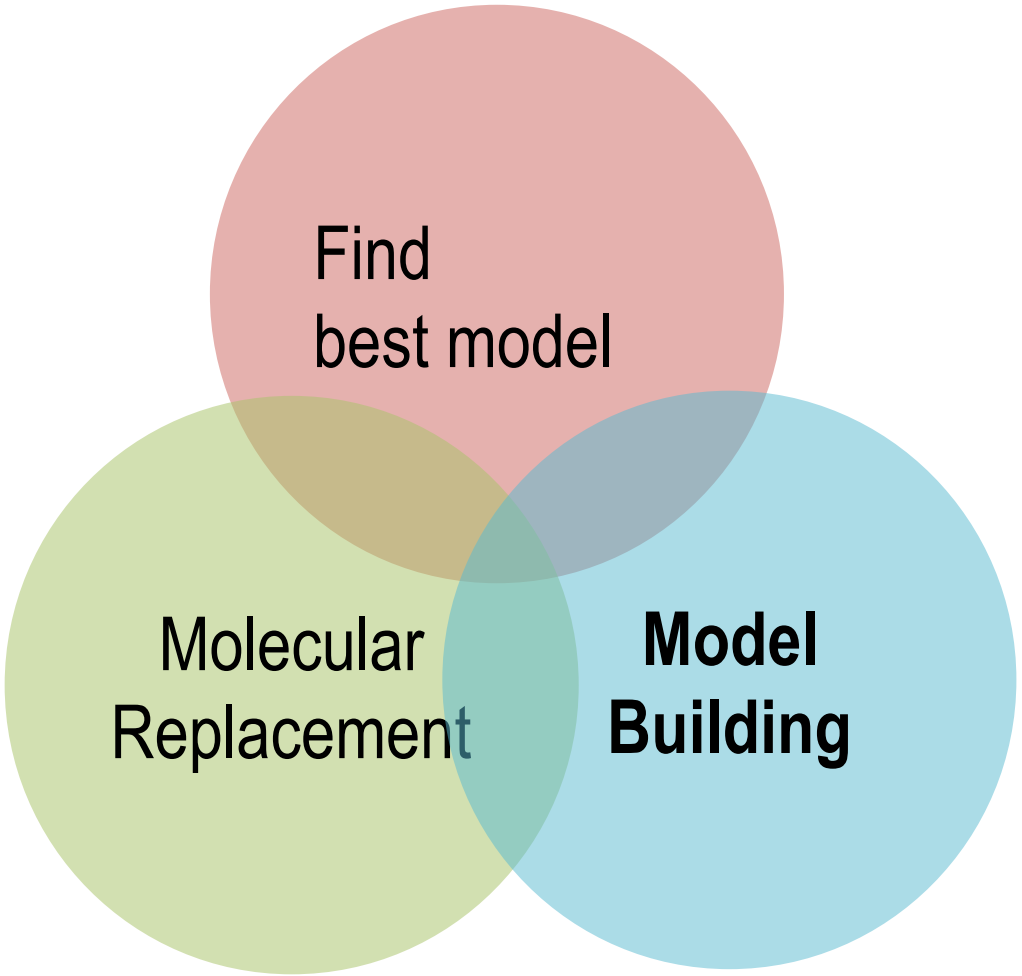
The pathway of structure solution

- Historically, there has been a linear progression through structure solution
- You had to be sure each step is correct before progressing to the next
- When signal is low you cannot be sure (of anything)



New approaches

- Take multiple possibilities for each step and uses subsequent steps to distinguish correct from incorrect solutions
- Enables structure solution when signal is low



phaser.mrage

- Fetches models from PDB and processes using sculptor
- For each partial structure model MR is farmed out to a cluster in a highly parallel manner
  - Calculations are performed in the order of sequence identity or LLG score at each stage
- Exploration continues until a solution is found
- All alternative models are superposed onto the solution and refined. This allows the quick evaluation of model quality for a potentially large number of alternative models.

Phaser.MRage: automated molecular replacement
Bunkoczi G, Echols N, McCoy AJ, Oeffner RD, Adams PD, Read RJ
Acta Cryst. (2013). D69, 2276-2286

phenix.mr_rosetta

- Find MR solutions with Phaser, rebuild them with ROSETTA using techniques from ab initio modelling (ROSETTA energy term) to bring the structures within the radius of convergence of standard rebuilding/refinement in phenix.autobuild
- Correct solution must be in list passed to ROSETTA
  - phenix.mr_rosetta takes top 5 by default, regardless
  - relies on enrichment
- Rebuilding in ROSETTA includes map information through a density term in the ROSETTA energy

Increasing the Radius of Convergence of Molecular Replacement by Density and Energy Guided Protein Structure Optimization
DiMaio et al (2011) Nature, 473, 540-543.

The Phenix Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Youval Dar, Nat Echols, Nigel Moriarty, Nader Morshed, Ian Rees, Oleg Sobolev

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner, Richard Mifsud

Cambridge University

Duke University

Jane & David Richardson, Chris Williams, Bryan Arendall, Bradley Hintze

An NIH/NIGMS funded Program Project