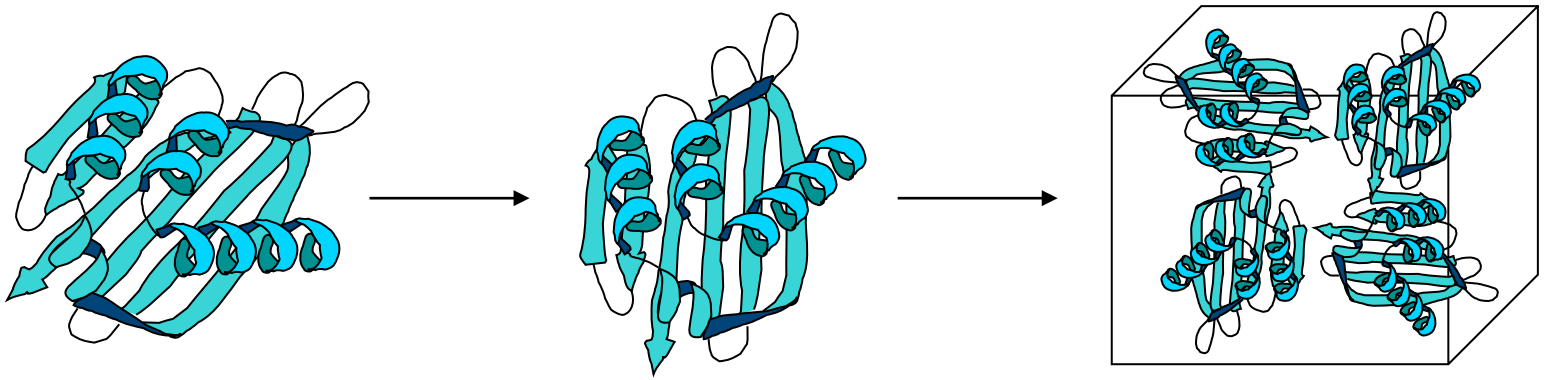
Likelihood-based molecular replacement in *Phaser*

UNIVERSITY OF CAMBRIDGE

R J Read, Department of Haematology
Cambridge Institute for Medical Research

Phasing by molecular replacement

- Phases can be calculated from atomic model
- Rotate and translate related structure to match the observed diffraction intensities



Models for molecular replacement

- Ease of molecular replacement depends on quality and completeness of model
- Quality depends on:
  - sequence identity
  - experiment (resolution, NMR/X-ray)
  - flexibility
  - sophistication of homology modelling
- Maximum likelihood methods can cope with poorer models

Choosing the right model

- The best model may not be the top hit
  - correlation between sequence identity and quality is approximate
  - conformational change
- Test multiple choices of model
  - easier in a pipeline: phenix.MRage, Balbes, MrBUMP
  - multiple templates,
    multiple manipulations of template

Model manipulation

- Ensembler
  - multiple structure superposition to make ensemble of possible models
  - optionally trim non-conserved surface loops
- Sculptor (see also Chainsaw)
  - use sequence alignment to:
    - trim parts of template not in target
    - adjust B-factors of poorly-conserved regions
  - use surface accessibility to:
    - adjust B-factors of surface regions

Homology modeling and MR

- *Rosetta*: sophisticated modeling program from David Baker's group
  - computationally intensive (Rosetta@home)
  - combination of physics, database knowledge and conformational search algorithms
- Templates from NMR structures and distant homologues can be improved for MR
  - Bin Qian, Rhiju Das *et al.* (2007)
- Complete (possibly ambiguous) solution from poor model: phenix.mr_rosetta
  - Frank diMaio, Tom Terwilliger *et al.* (2011)

Known knowns…

- Donald Rumsfeld, 12 February 2002:
  - “There are known knowns; there are things we know we know. We also know there are known unknowns, that is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know.”

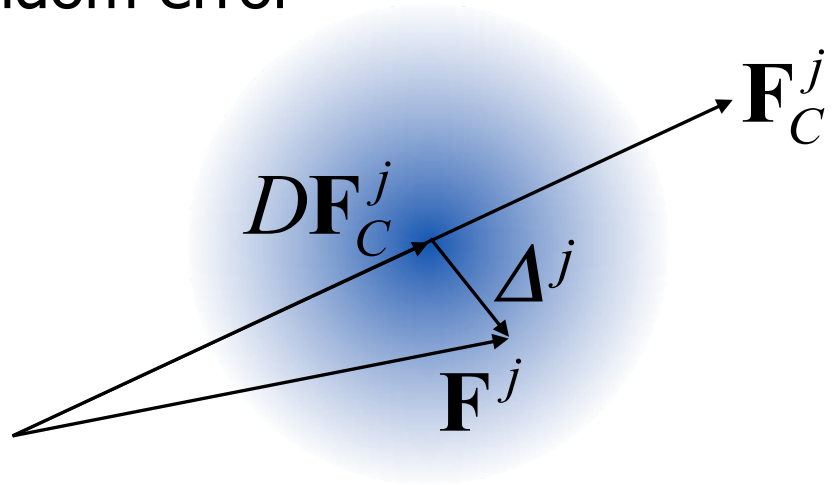
Known knowns	Known unknowns
Unknown knowns	Unknown unknowns

Likelihood-based molecular replacement

- Likelihood target:
  - probability of observed amplitude given (set of) model structure factor contributions
    - account for effect of unknown relative phases
- Benefits of likelihood
  - account for expected size of errors in model
  - account for lack of completeness of model
  - exploit knowledge from partial solutions
  - allow ensemble of possible models
    - also useful for MR with NMR

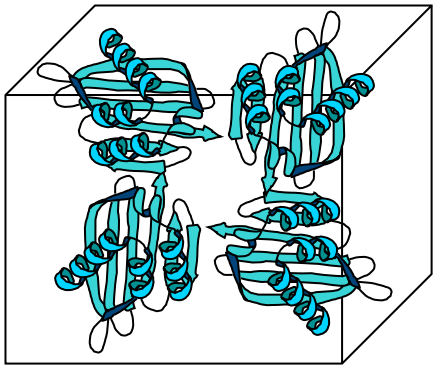
Effect of errors on structure factor distribution

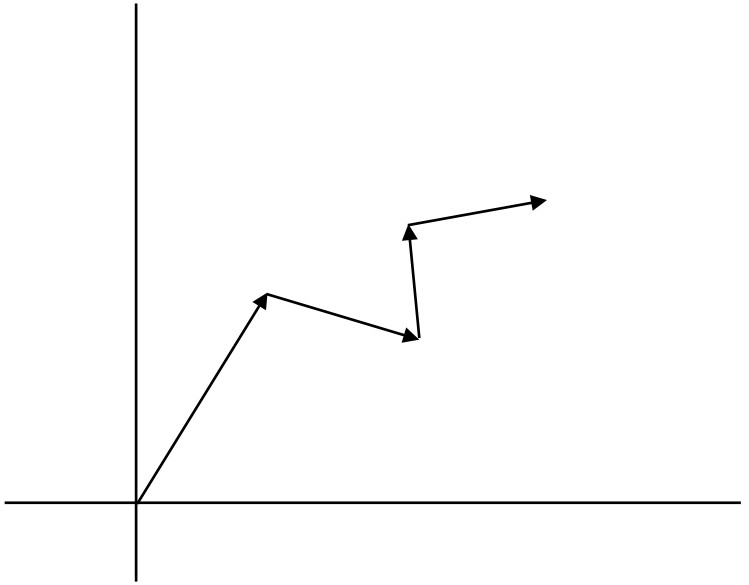
- Errors and incompleteness in model lead to errors in calculated structure factor
  - probability is complex Gaussian
- Only part of model structure factor is correct
  - plus random error

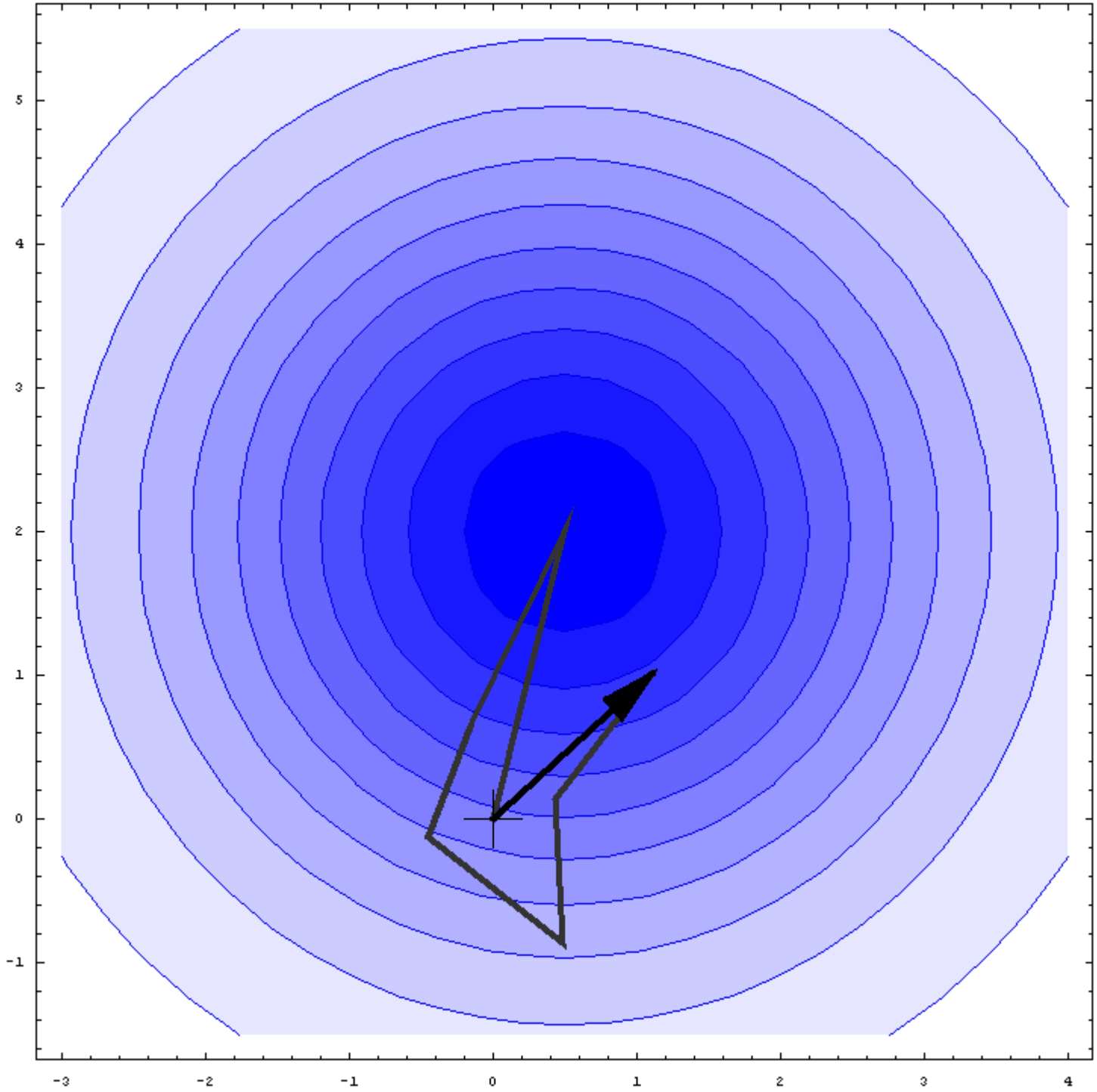


Rotation likelihood function

- What structure factors could be obtained from an oriented model?
  - add up contributions from molecules in unit cell, but unknown relative phase

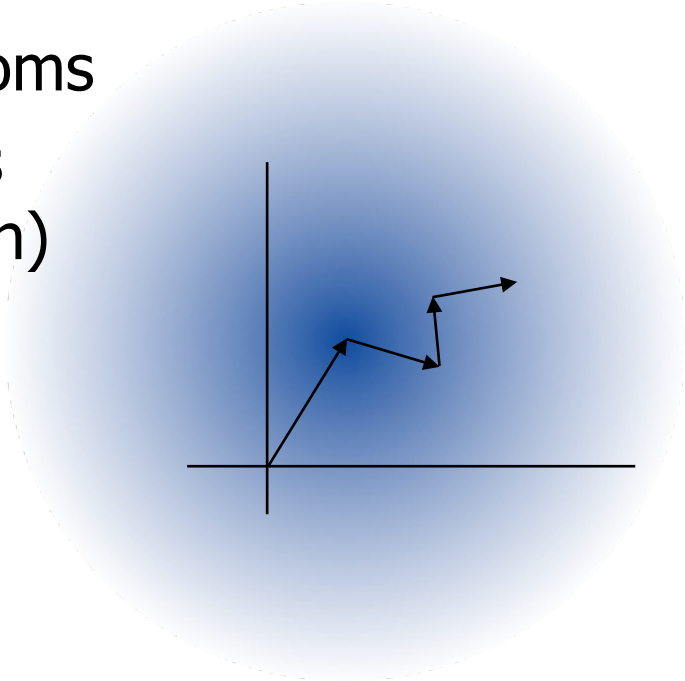





5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
4

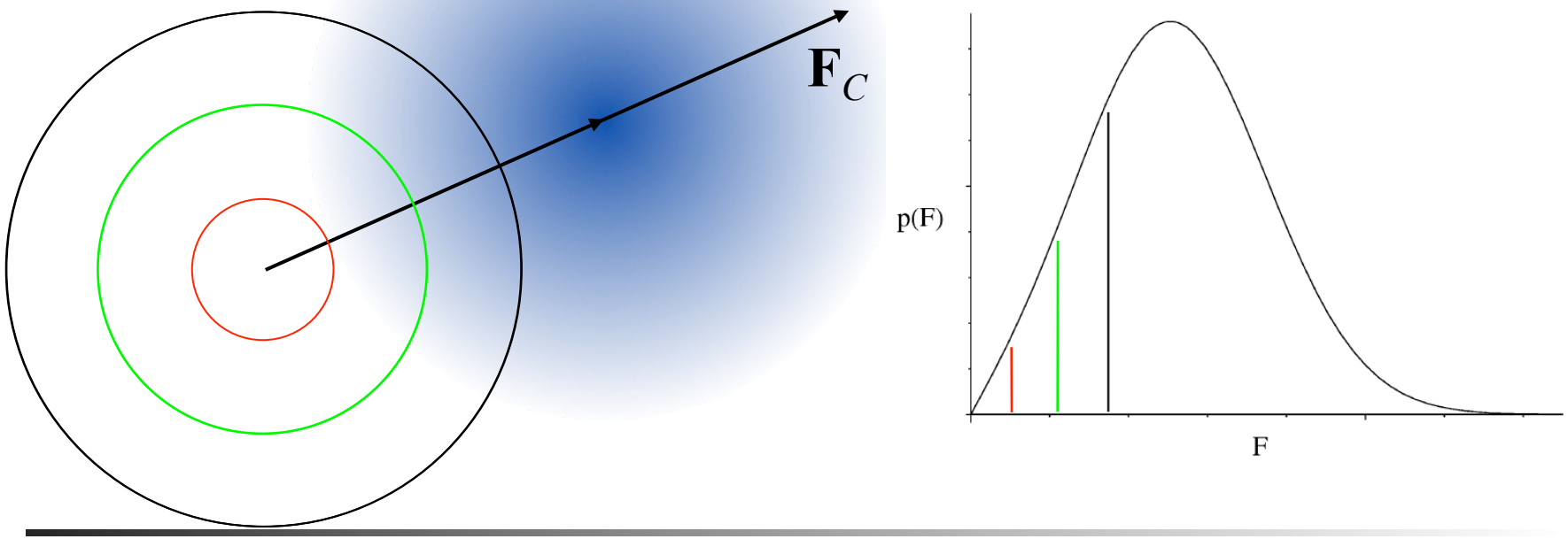
Molecular replacement likelihood function in *Phaser*

- Take biggest single contribution as $\mathbf{F}_C$
  - multiply by D
- Gaussian noise includes:
  - effects of model error, missing atoms
  - other symmetry-related molecules in cell (but not if translation search)



Amplitude probability distribution

- Have p($\mathbf{F}$), but data are |$\mathbf{F}$| so need p(|$\mathbf{F}$|)
- Integrate over unknown phase angle to get Rice (Luzzati, Sim, Srinivasan) distribution



Fast rotation and translation functions

- Full likelihood functions are expensive to evaluate
- Search orientations with likelihood-based fast rotation function
  - rescore plausible solutions with full rotation likelihood
- Search translations with likelihood-based fast translation function
  - rescore with full likelihood target
  - refine against full likelihood target

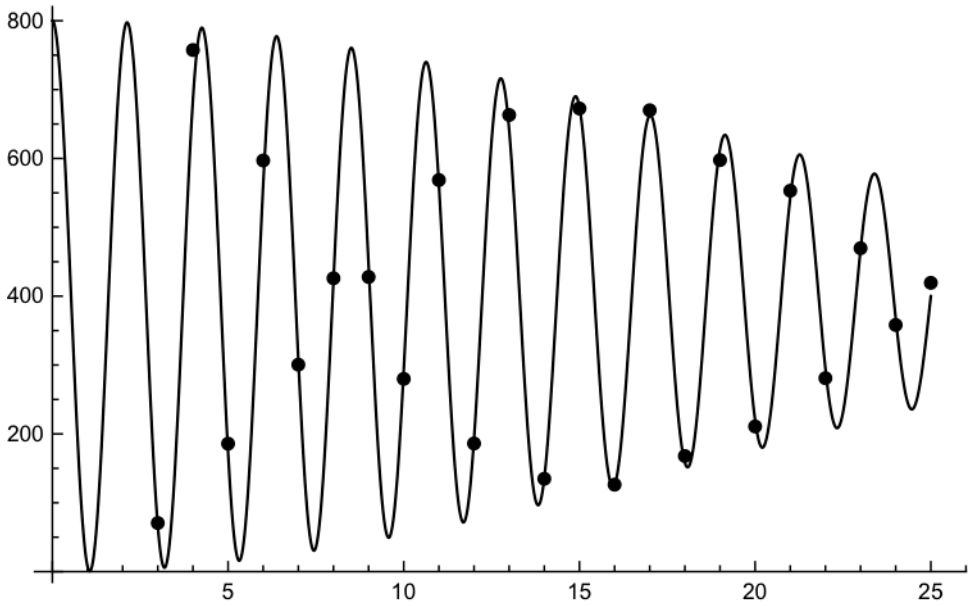
Translational non-crystallographic symmetry (tNCS)

- Found in about 8% of PDB entries

Photo courtesy of Laurie Betts

Accounting for translational NCS

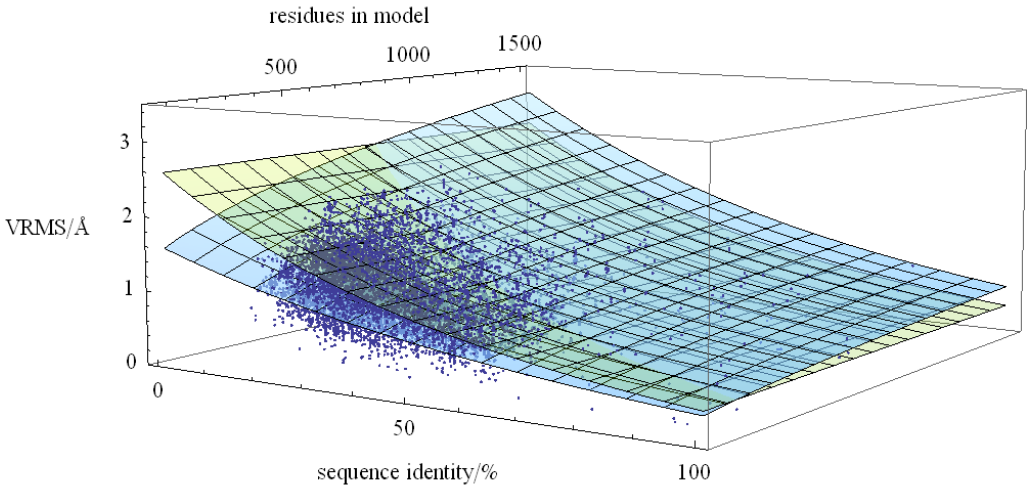
- Model effect of translation combined with small rotation and random differences between copies

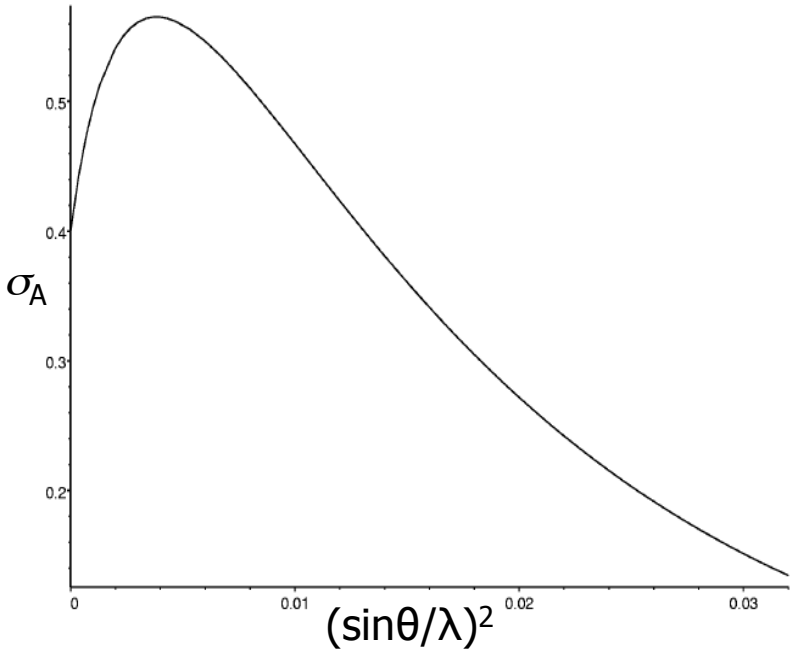


Hyp-1:
Sliwiak, Jaskolski, Dauter, McCoy, Read (2014)

Defining an *a priori* σ_A curve

- Need initial estimates of completeness, disordered solvent, coordinate errors
  - function of sequence identity and size of structure





Can I solve it?

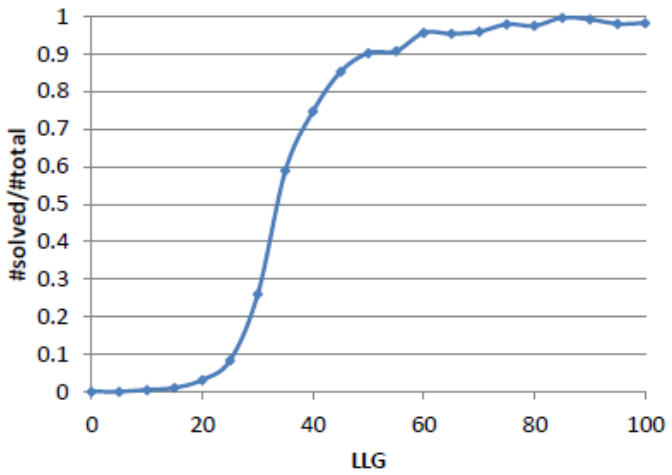
- Is my sequence identity too low?
- Is my model too incomplete?
- Is the resolution too low?

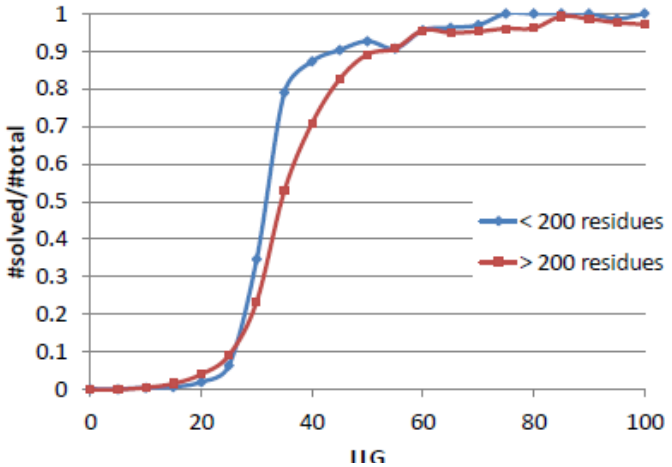
- Have I solved it?

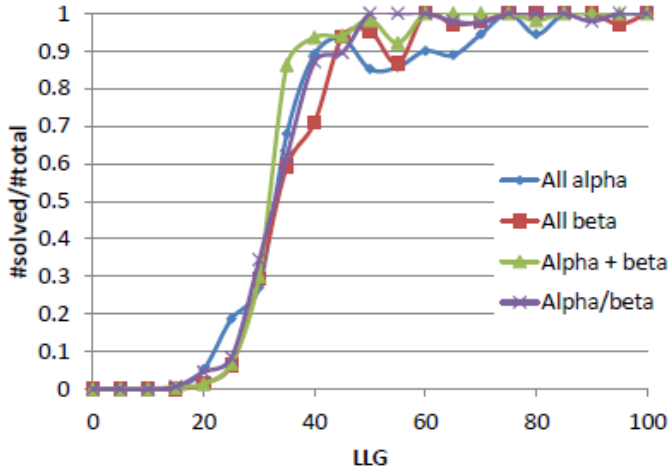
Understanding the LLG score

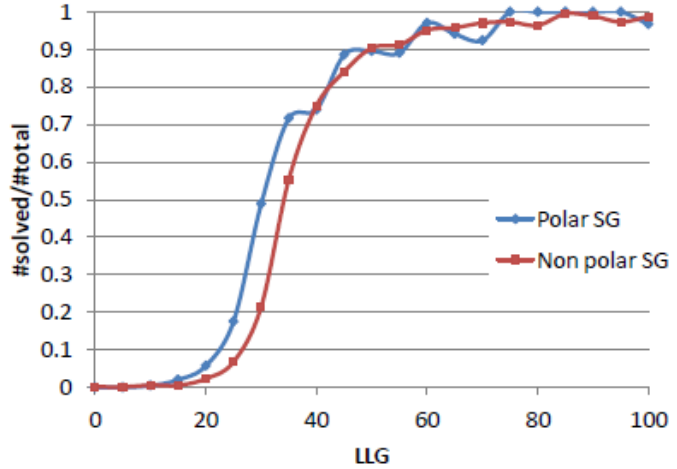
- LLG (log-likelihood-gain) value depends on
  - quality and completeness of model
    - what fraction of the scattering it explains (σ_A^2)
    - σ_A^2 is generally a function of resolution
  - number of reflections

LLG: measure of confidence in solution (Rob Oeffner)



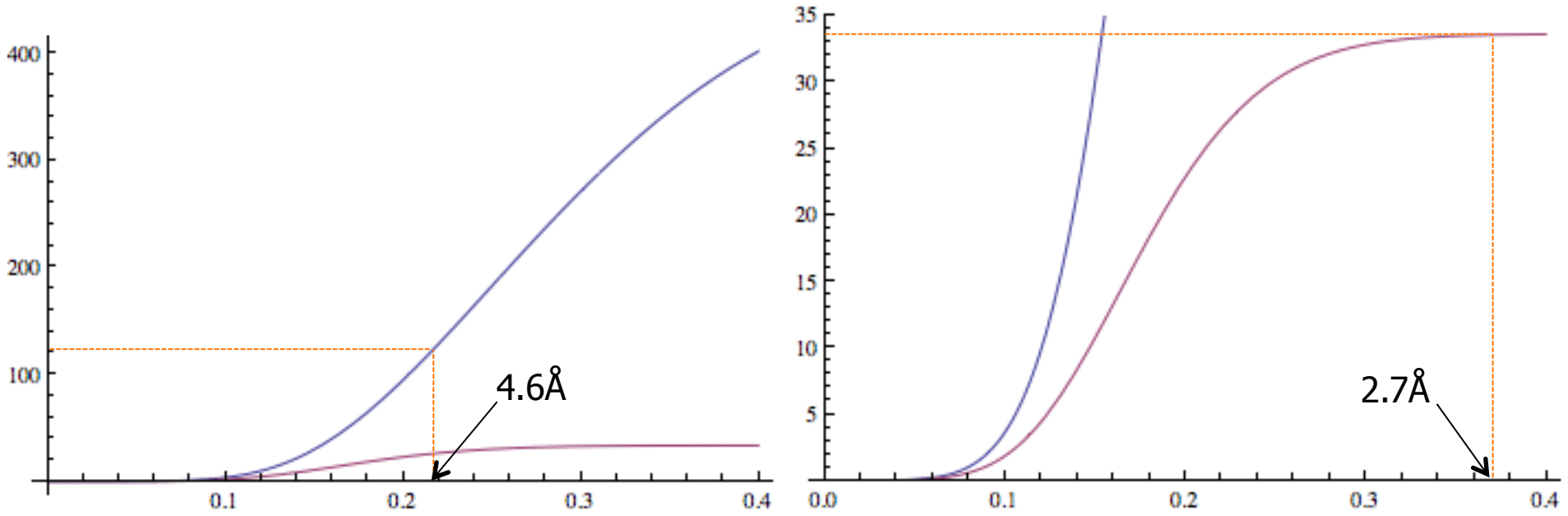






Adaptive strategies based on <LLG>

- Consider case of good model (0.8Å rms) *vs* bad model (1.5Å rms), data to 2.5Å resolution



Solving the structure of the ribosome

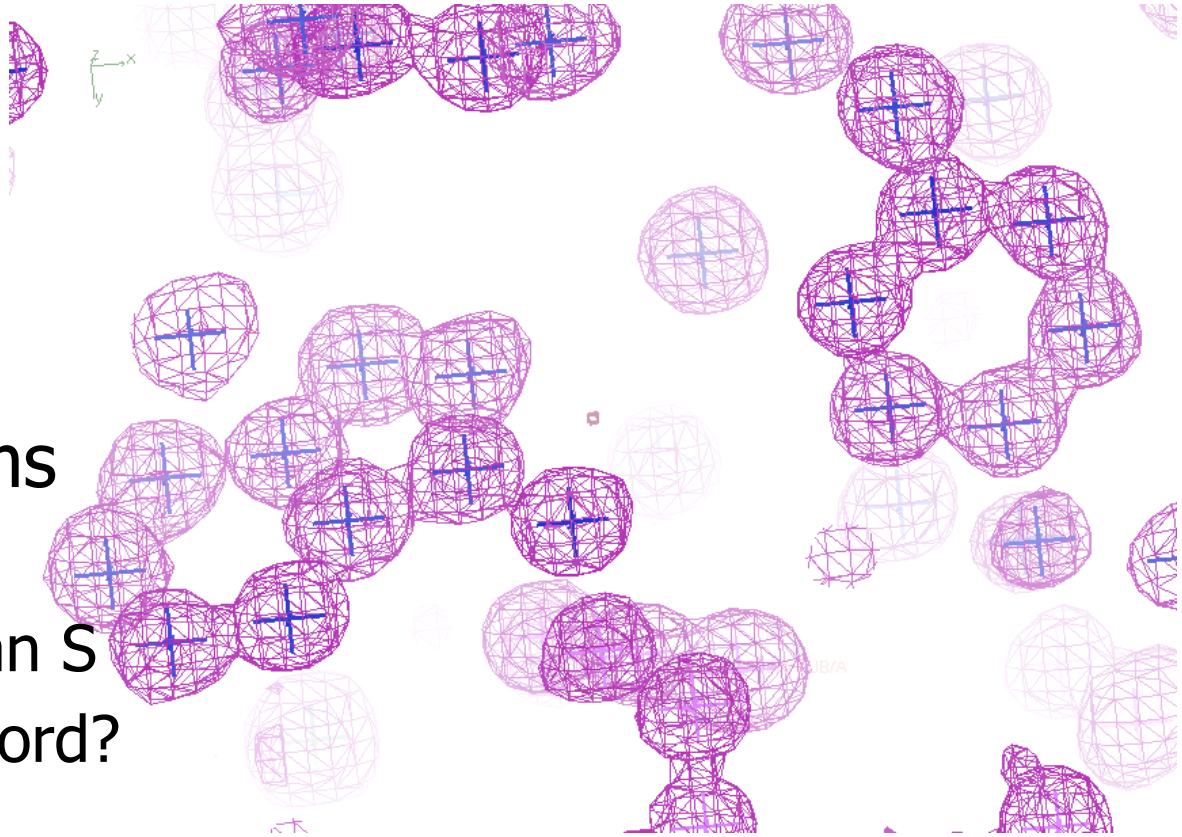
- 2j00: *Thermus thermophilus* 70S ribosome
  - Ramakrishnan group, 2006
  - two copies in a.u
  - data to 2.8Å resolution (1.3M reflections)
- Models:
  - 1j5e: *Thermus thermophilus* 30S small subunit
  - 1ffk: *Haloarcula marismortui* 50S large subunit
- *Phaser* chooses limit of 7.5Å (79K reflections)
  - sufficient to use data to 12Å (19K reflections)

Arcimboldo

- Rodríguez, Grosse, Himmel, González, de Ilarduya, Becker, Sheldrick & Usón, *Nature Methods* **6**: 651-653 (2009)
- *Phaser* and *SHELXE*

Aldose reductase (Antoni Wrobel)

- 36 kDa, 0.78Å resolution (3bcj)
- Find 4 S (<3h)
- Complete with N atoms (3h)

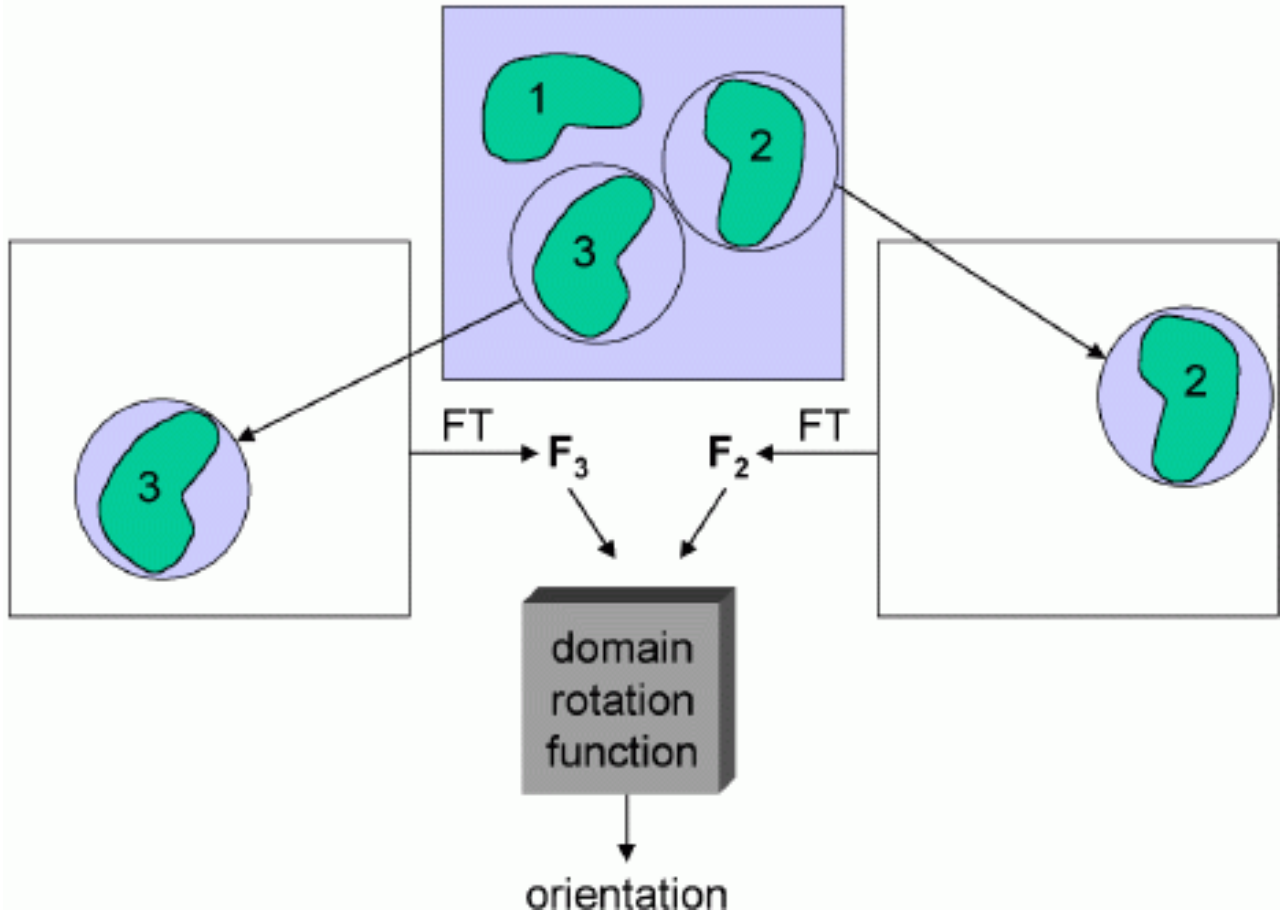


- 2525 non-H atoms in structure
  - none heavier than S
  - new *ab initio* record?

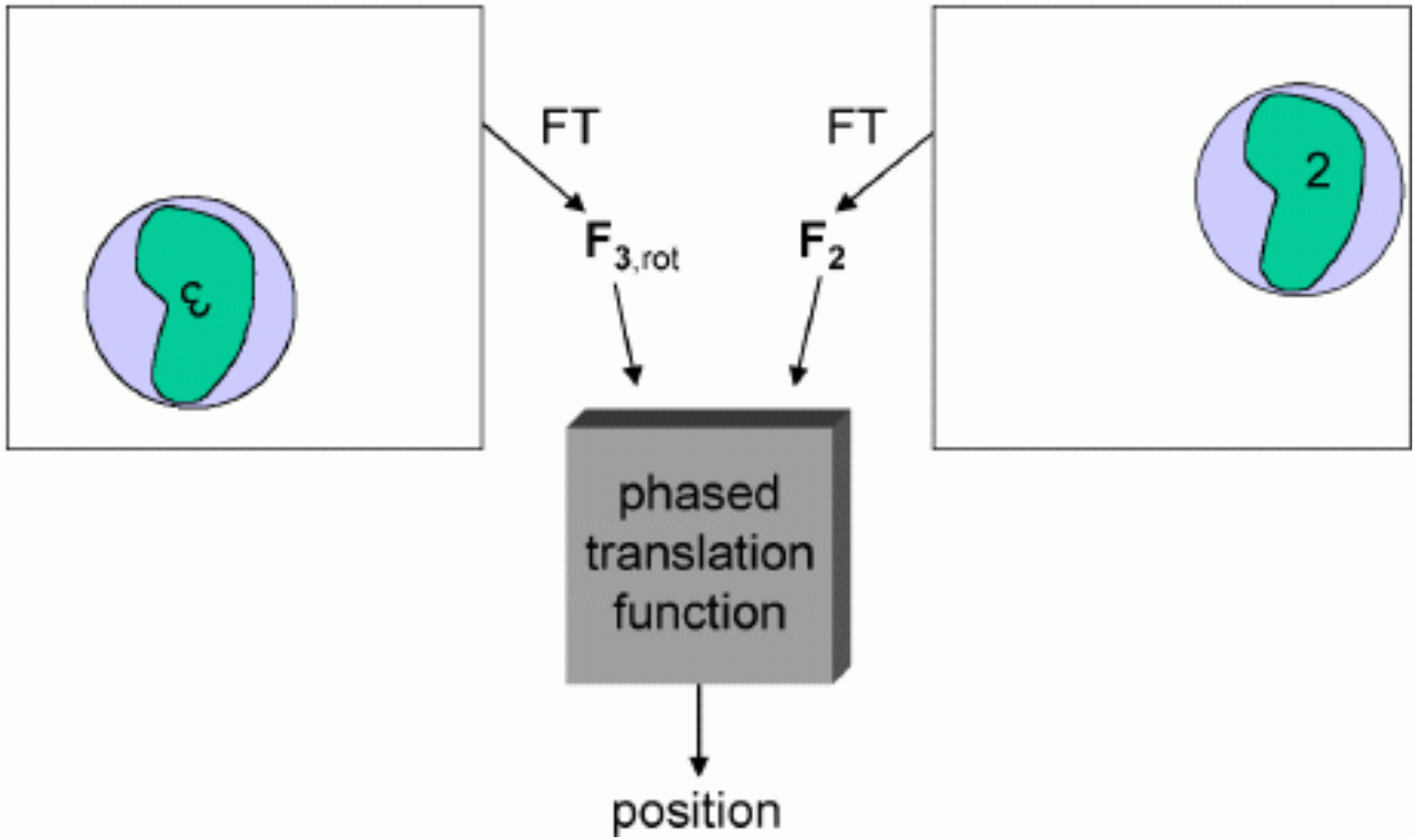
Real-space molecular replacement

- Use phase information in two ways:
  - use electron density as model
    - calculate structure factors from isolated density, then proceed as with atomic model
    - possible in *Phaser*
  - fit model into electron density
    - "domain rotation function"
    - "phased translation function"

Domain rotation function



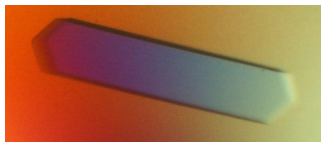
Phased translation function



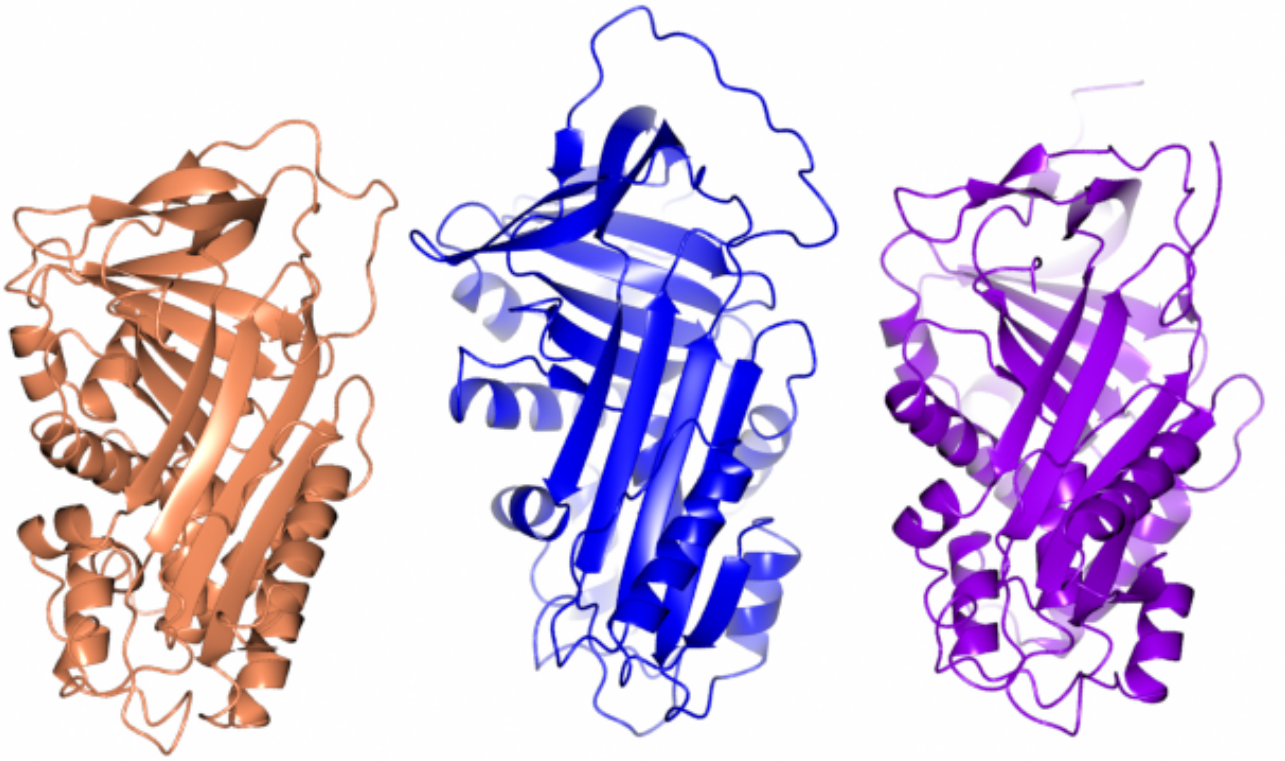
Angiotensinogen crystals

- Human: 1 crystal form
  - 3.3Å, 1 copy
- Rat: 2 crystal forms
  - 2.8Å, 2 copies
  - 3.15Å, 2 copies
- Mouse: 2 crystal forms
  - 2.1Å, 1 copy
  - 2.95Å, 4 copies

Human angiotensinogen: molecular replacement



human

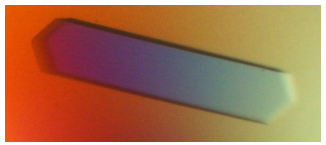


heparin cofactor II
(20% identical)

α_1-antitrypsin
(21% identical)

thyroxine-binding globulin
(20% identical)

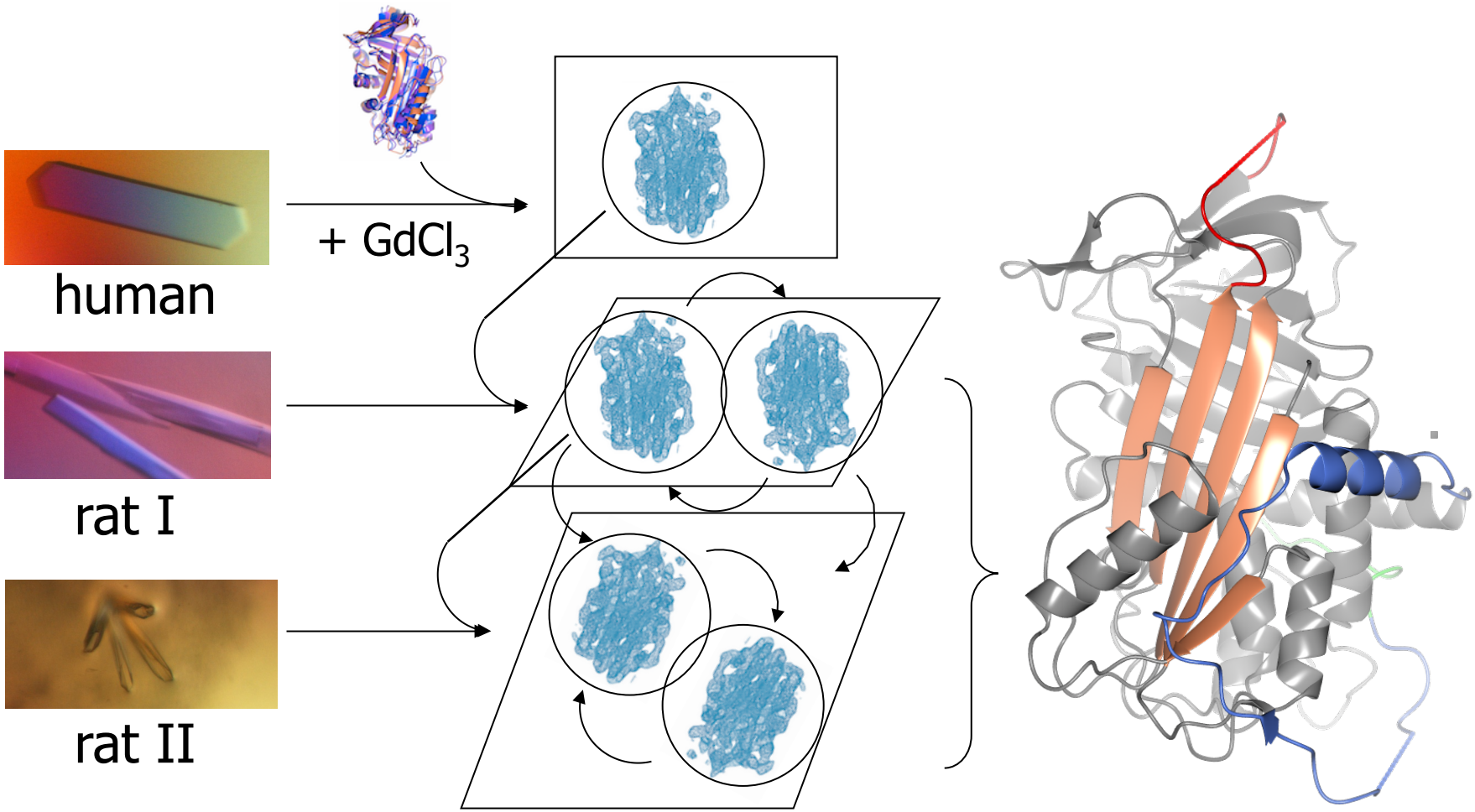
Human angiotensinogen: molecular replacement



human

Trimmed ensemble

Solving angiotensinogen structures



Solving *Drosophila* GST2 (1M0U)

- Difficult structure from Bogos Agianian (Piet Gros)
- Find one of two copies with ensemble of 3 structures (28-30% identity)
  - search for second copy fails
- Find second copy as density from first
  - this succeeds

Automation in the fast search mode

- Adaptive strategies based on judging whether component is placed correctly
- Ambiguous solution?
  - back up and try a different choice of model
  - back up and try placing a different component
  - back up and try higher resolution data
- Unambiguous solution?
  - if searching for multiple copies, try placing several copies at once (amalgamation)

Practical aspects of MR in *Phaser*

- Provide information about model quality
  - estimated RMS error to calibrate σ_A curve
- Provide information about cell content
  - sequence, molecular weight, percent solvent…
  - used to determine model completeness
- Consider possibility of conformational change
  - alternative models
  - search with isolated domains
- Let *Phaser* try first to place everything!

Background information

- "*Phaser* crystallographic software", McCoy, Grosse-Kunstleve, Adams, Winn, Storoni & Read (2007), *J. Appl. Cryst.* **40**, 658-674.
  - plus papers cited here
- "Liking likelihood", Airlie J. McCoy (2004), *Acta Cryst. D***60**, 2169-2183.

- http://www.phaser.cimr.cam.ac.uk/index.php
- http://www.phaser.cimr.cam.ac.uk/index.php/Tutorials
- http://www-structmed.cimr.cam.ac.uk/Course

The Phenix Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Nat Echols, Richard Gildea, Ralf Grosse-Kunstleve Jeff Headd, Nigel Moriarty, Nicholas Sauter, Peter Zwart

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Cambridge University

Duke University

Jane & David Richardson, Vincent Chen, Lindsay Deis, Bradley Hintze, Swati Jain, Gary Kapral, Chris Williams

An NIH/NIGMS funded Program Project

Funding


wellcome trust


NATIONAL INSTITUTES OF HEALTH


British Heart Foundation


CCP4