phaser and phasing

Airlie McCoy

UNIVERSITY OF CAMBRIDGE

1

software for phasing

Score: find the pose that best fits the data

$$LLGI = \sum_{hkl_{info}} log\left[\frac{2E_e}{1-D_{obs}^2\sigma_A^2}exp\left(-\frac{E_e^2+D_{obs}^2\sigma_A^2E_C^2}{1-D_{obs}^2\sigma_A^2}\right)I_0\left(\frac{2E_eD_{obs}\sigma_AE_C}{1-D_{obs}^2\sigma_A^2}\right)\right]$$

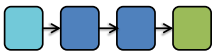
Search: strategies for exploring poses

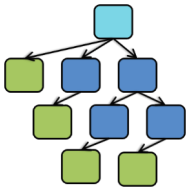
2

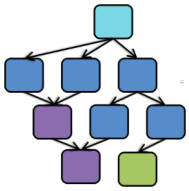
2

process flow

- path
- tree
- directed acyclic graph



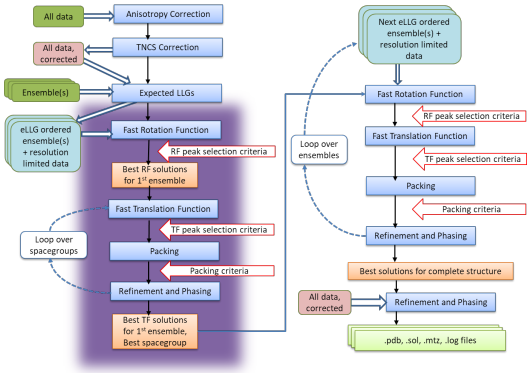




- termination condition

3

3

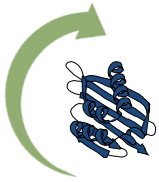


4

4

brute rotation function search

- Place model at orientations and calculate probability of each being correct



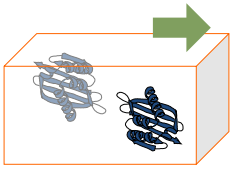
$$LLGI = \sum_{\mathbf{h}} \log\left(\frac{2E_e}{1-D_{obs}^2\sigma_A^2}\exp\left(-\frac{E_e^2+D_{obs}^2\sigma_A^2E_C^2}{1-D_{obs}^2\sigma_A^2}\right)I_0\left(\frac{2E_eD_{obs}\sigma_AE_C}{1-D_{obs}^2\sigma_A^2}\right)\right)$$

5

5

brute translation function search

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

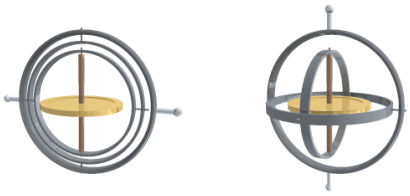


$$LLGI = \sum_{\mathbf{h}} \log\left(\frac{2E_e}{1-D_{obs}^2\sigma_A^2}\exp\left(-\frac{E_e^2+D_{obs}^2\sigma_A^2E_C^2}{1-D_{obs}^2\sigma_A^2}\right)I_0\left(\frac{2E_eD_{obs}\sigma_AE_C}{1-D_{obs}^2\sigma_A^2}\right)\right)$$

6

6

euler angles

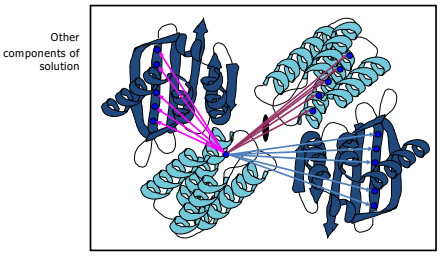


7

7

packing

- clash test

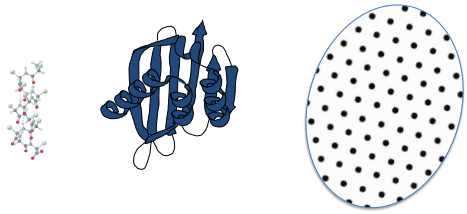


8

8

packing

- Small – Medium – Large
- all atoms – Cα atoms – Hexagonal Grid

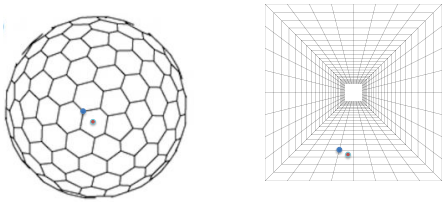


9

9

refinement

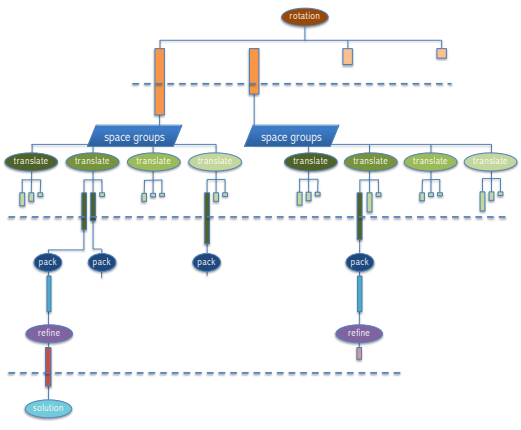
- Rotation and Translation searches are scored on a grid



$$LLGI = \sum_{\mathbf{h}} \log\left(\frac{2E_e}{1-D_{obs}^2\sigma_A^2}\exp\left(-\frac{E_e^2+D_{obs}^2\sigma_A^2E_C^2}{1-D_{obs}^2\sigma_A^2}\right)I_0\left(\frac{2E_eD_{obs}\sigma_AE_C}{1-D_{obs}^2\sigma_A^2}\right)\right)$$

10

10

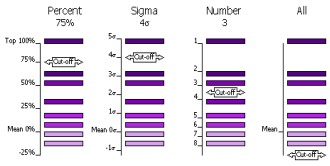


11

11

peak selection

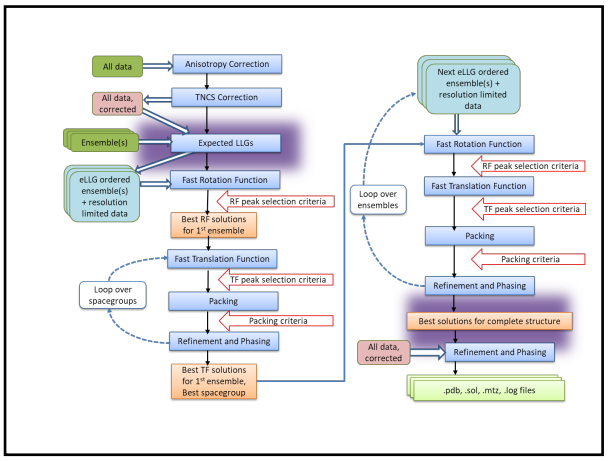
- Must chose a selection criteria to carry possible solutions through to the next step



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

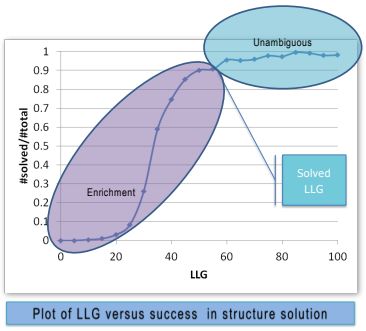
12

12



13

log-likelihood gain for solutions



Database of over 23000 MR problems

Plot of LLG versus success in structure solution

R.D. Oeffner

14

14

log-likelihood gain for solutions

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

15

15

predicting log-likelihood gain

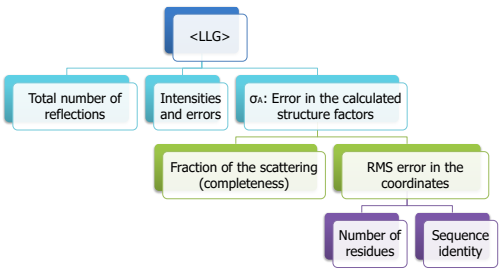
- If you can predict the log-likelihood gain you know how easy/difficult will be molecular replacement
- The expected log-likelihood gain for a model is calculated by "integrating out" the calculated structure factors from the equation
  - The error in the calculated structure factors (σ_A) is included
- We have a similar analysis for SAD phasing*
- Used to guide decision making at every step of the search strategy
  - Prioritize resources

*unpublished

16

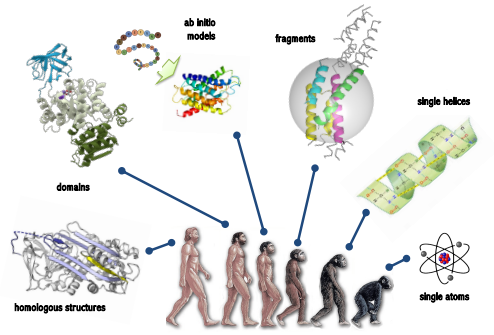
16

expected log-likelihood gain



17

17



18

fragments

- Fragments are just models with low completeness
  - Low σ_A
- Success of fragment 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

19

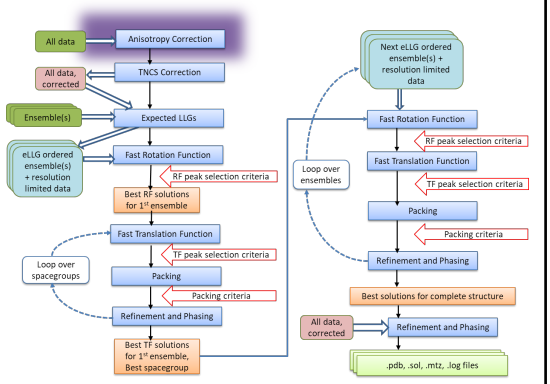
19

What chance of success?

- *Better than 40% identity: usually easy, unless large conformational changes are involved.*
- *30-40%: MR possible, but sometimes more difficult.*
- *20-30%: MR difficult, careful model preparation required.*
- *Less than 20%: MR unlikely to work in most cases.*
- *In terms of RMSD, above 2.5A is very unlikely to work, while 1.5A or less is preferable.*

20

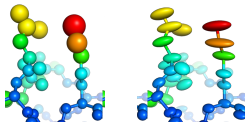
20



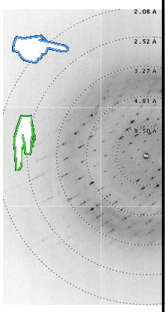
21

Anisotropy

- Systematically weak and strong diffraction directions
- Caused by underlying anisotropic movement of atoms



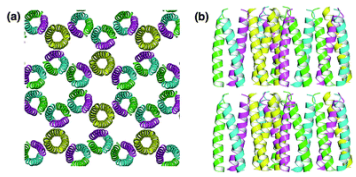
Visualization of individually refined isotropic (left) versus anisotropic (right) B-factors for a high-resolution structure



22

22

Anisotropy in coiled coils



(a) Honeycomb-like P6 crystal lattice of a computationally-designed coiled-coil protein building block. (b) The layer structures were stacked by hydrogen bond interactions between layers. 4DAC

23

23

Anisotropy

Before

- Easy phasing cases and/or mild anisotropy, can ignore
- In difficult cases, must corrected before structure solution
- Global Anisotropic B-factors calculated with reference to the Wilson distribution

After

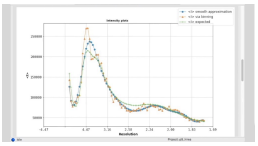
- Corrections will be better after structure solution when you have a model
- Global Anisotropic B-factor calculated with reference to calculated structure factors
  - TLS
  - Atomic anisotropic B-factors

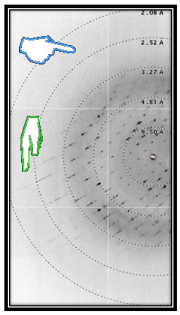
24

24

Wilson Plot

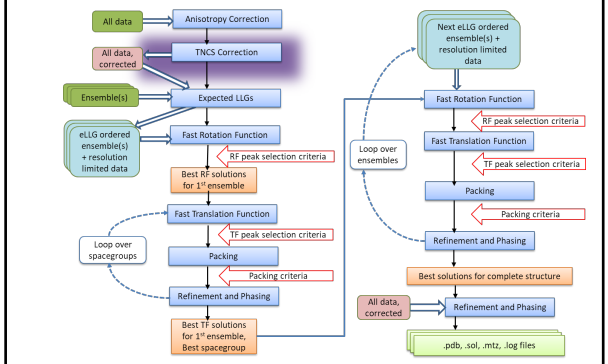
- Correct data with reference to the Wilson plot
- Make intensity distribution the same in all directions by refining parameters of the anisotropy tensor





25

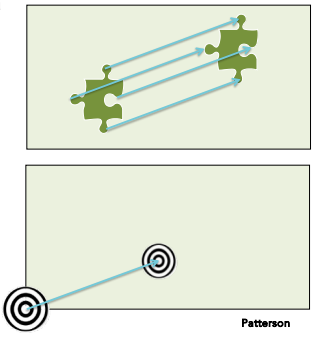
25



26

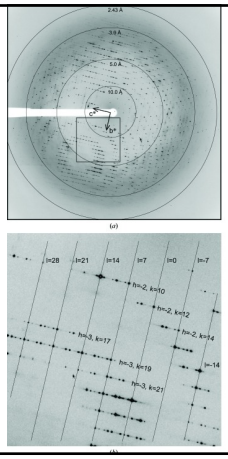
Translational NCS

- Molecules related by a vector translation
- Patterson shows a single vector
  - and symmetry copies



27

27

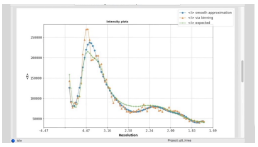


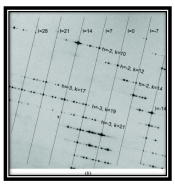
Hyp-1 (PR-10)
(a) A sample full image, with resolution rings and reciprocal-lattice directions indicated,
(b) Enlarged fragment on which layer lines of constant l = 7n are marked and annotated
(c) A histogram of intensity distribution in layers of l.

28

Wilson Plot

- Correct data with reference to the Wilson plot
- Make intensity distribution in tncs modulation directions the same by refining parameters of the tncs





29

29

TNCS

Before

- In difficult cases, must corrected before structure solution
- TNCS parameters calculated with reference to the Wilson distribution
- TNCS parameters generate epsilon factors for each reflection, to make them closer to the Wilson distribution

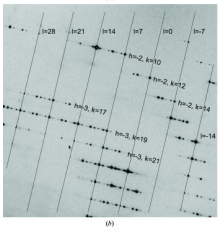
After

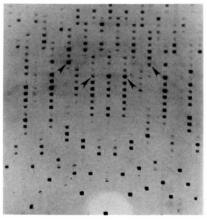
- Corrections will be better after structure solution when you have a model
- TNCS parameters calculated with reference to the oriented and placed models in the asymmetric unit

30

TNCS and Twinning

- TNCS – broadens intensity distribution
- Twinning – narrows intensity distribution





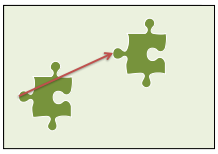
31

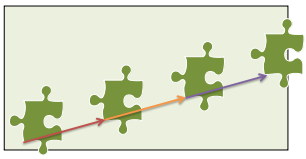
31

PHASER: Two classes of tNCS

Pairs of molecules related by one vector

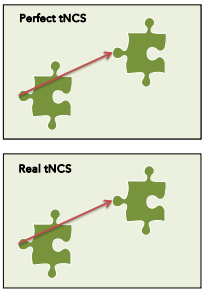
Molecules related by multiples of one vector





32

Translational NCS: Pairs

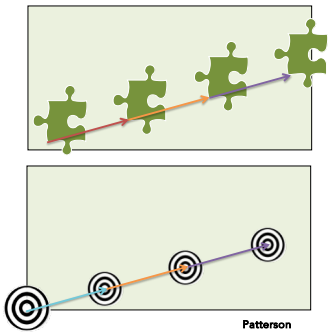


- Three TNCS parameters refined from data alone
  1. Rotation
  2. Translation
  3. RMSD between copies

33

33

Translational NCS: Multiples



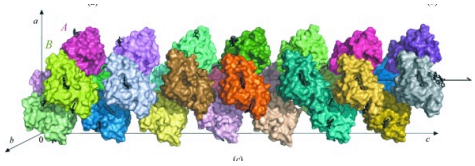
- Two TNCS parameters refined from data alone
  1. Translation
  2. RMSD between copies

34

34

Hyp-1 (PR-10)

- Repeats 7/2 along c* (sevenfold translational NCS)

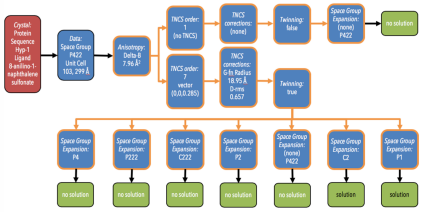


Acta Cryst D STRUCTURAL BIOLOGY

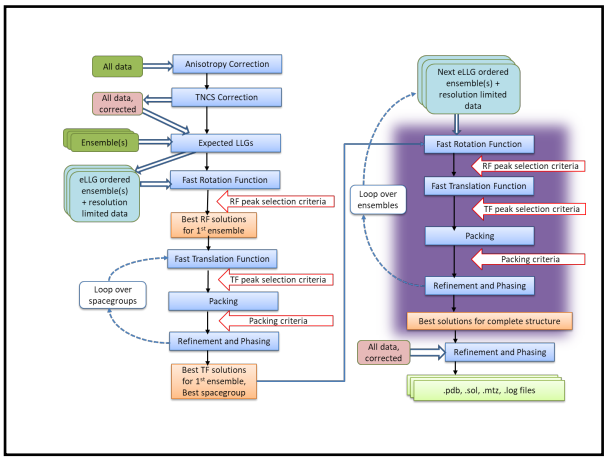
Likelihood-based molecular-replacement solution for a highly pathological crystal with tetartohedral twinning and sevenfold tNCS Sliwiak J, Jaskolski M, Dauter Z, McCoy AJ, Read RJ. Acta Cryst D70, 471 (2014)

35

35



36



37

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
- *Self Rotation Function*
- *TNCS Order*

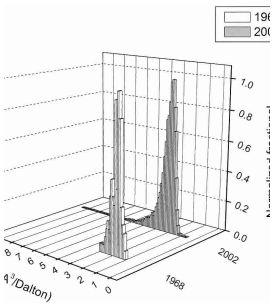
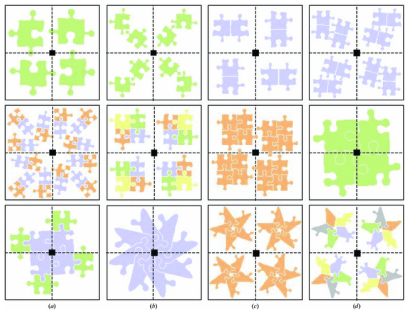


Figure 1: Kantardjieff and Rupp (2003)

38

38

contents of the asymmetric unit



39

39

using partial structure

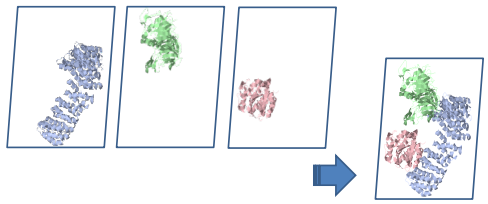
- The maximum likelihood rotation and translation functions can use models that have already been placed in the asymmetric unit
- Patterson Rotation Function cannot account for placed models
  - The Patterson Correlation Coefficient Translation Function can account for known positions (but not their errors)
  - Most of the difficulty in molecular replacement is the rotation function, because the signal to noise ratio is much lower

40

40

searches for multiple components

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

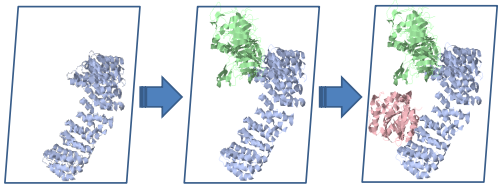


41

41

searches for multiple components

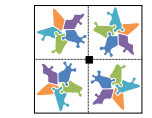
- maximum likelihood functions include partial structure information from previous placements
- structure builds up by addition



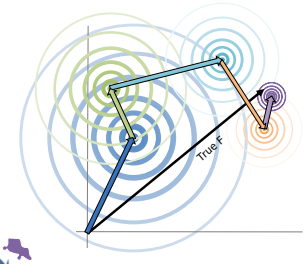
42

42

σ_A and fragments



Structure builds by addition



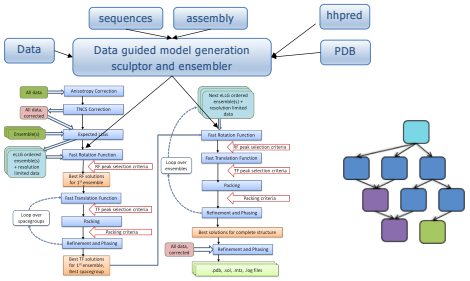
Whenever a component is placed, the variances are reduced, thus increasing the signal-to-noise of the search for the next component

43

43

phaser.voyager

- Data guided, statistical model generation
- DAG tracking and control



44

44

Acknowledgements

UNIVERSITY OF CAMBRIDGE

Randy Read
Rob Oeffner
Massimo Sammito
Kaushik Hatti
Duncan Stockwell
Tristan Croll
Claudia Millán

ICREA

Isabel Usón
Iracema Caballero

BERKELEY LAB

Billy Poon
Pavel Afonine

Phenix
CCP4

45

45