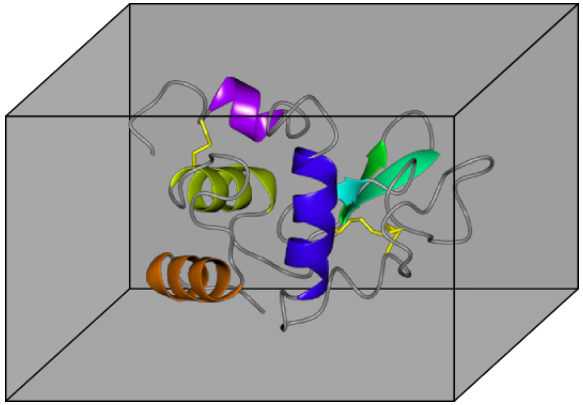
Molecular Replacement

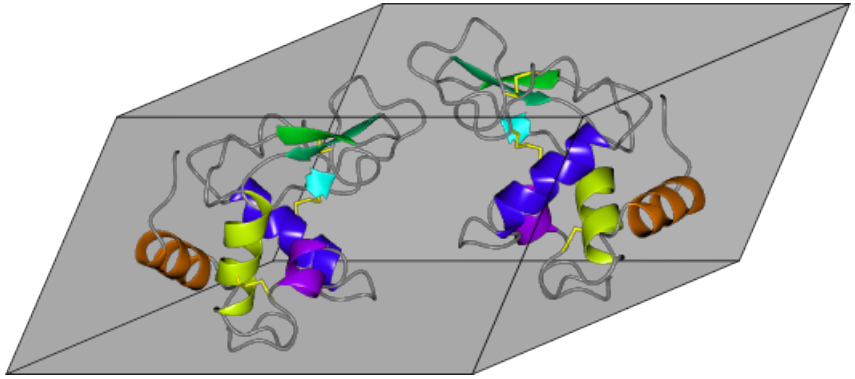
Andrey Lebedev
CCP4

MR Problem

Known crystal structure



New crystal structure

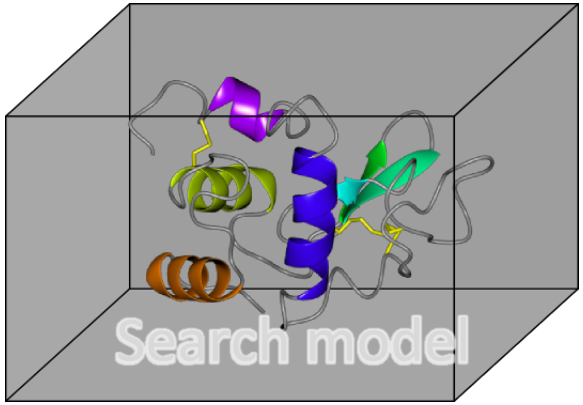


Given:
- Crystal structure of a homologue
- New X-ray data

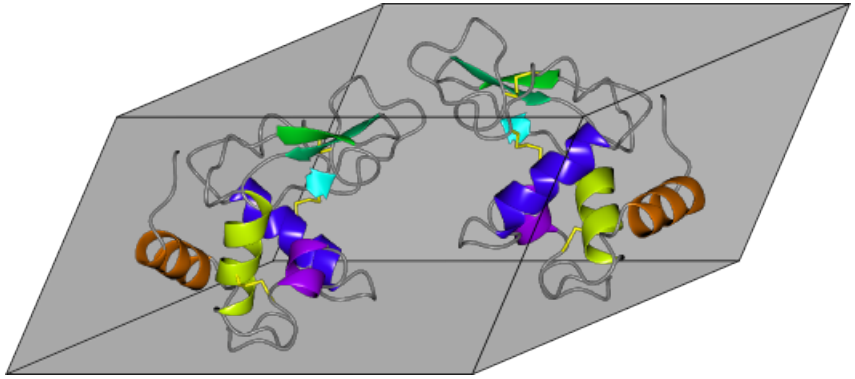
Find:
- The new crystal structure

MR Technique

Known crystal structure



New crystal structure



Method:

- 6×N - dimensional global optimisation
  - one 6-d search for each molecule in the AU
    - >> split further to orientation + translation searches = 3 + 3
    - >> fast search step using FFT

Required:

- Scoring
  - the match between the data and an (incomplete) crystal model
  - ideally: the highest score = correct solution

Real and Reciprocal spaces

- Terms may refer real space but actual calculations may be performed in the reciprocal space:
  - "Search in the electron density"
  - "Patterson search"

- The concepts formulated in real space are more intuitive

Functions in Real and Reciprocal spaces

Real space ← $\mathscr{F}$ → Reciprocal space

Map $\rho(r)$ ← $\mathscr{F}$ → $F(s)$ structure factors

↓ ↓

convolution ← $\mathscr{F}$ → product

↓ ↓

Patterson $P(r)$ ← $\mathscr{F}$ → $F(s)\,F^{*}(s) = I(s)$ intensities

Structure factors and Electron density map

Structure factors $F(h,k,l)$

– A discrete complex function in the reciprocal space

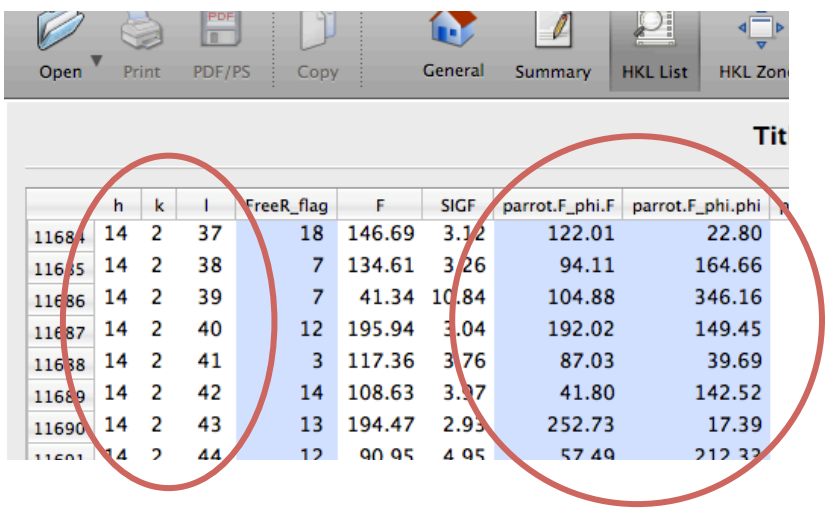
At given *h, k, l*

– Complex number:

$$F = A + iB$$

– Can be expressed via structure amplitude and phase

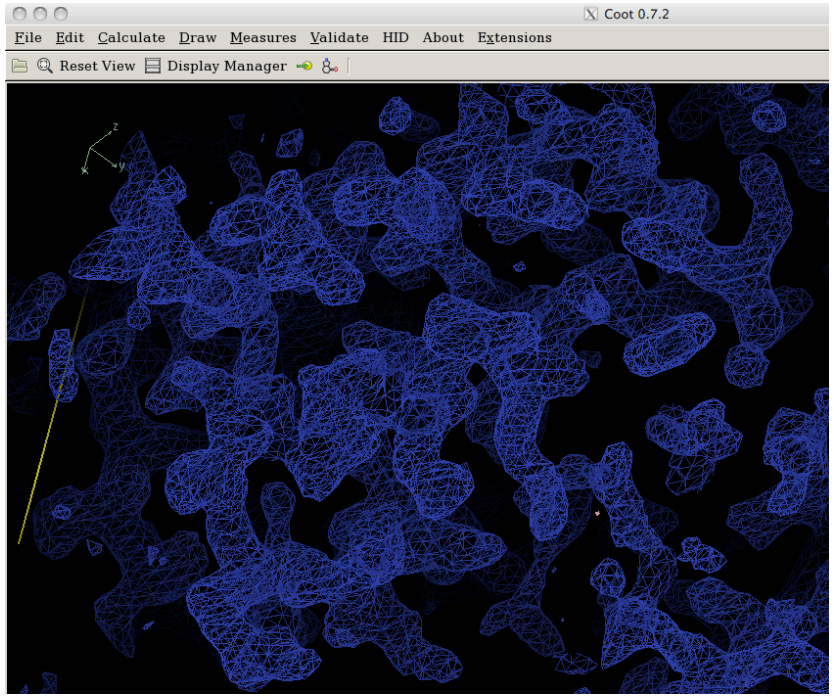
$$F = |F| \exp(i\phi)$$



	h	k	l	FreeR_flag	F	SIGF	parrot.F_phi.F	parrot.F_phi.phi
11684	14	2	37	18	146.69	3.12	122.01	22.80
11685	14	2	38	7	134.61	3.26	94.11	164.66
11686	14	2	39	7	41.34	10.84	104.88	346.16
11687	14	2	40	12	195.94	3.04	192.02	149.45
11688	14	2	41	3	117.36	3.76	87.03	39.69
11689	14	2	42	14	108.63	3.97	41.80	142.52
11690	14	2	43	13	194.47	2.93	252.73	17.39
11691	14	2	44	12	90.95	4.95	57.49	212.33

Electron density map

– periodic 3-d function in real space



is directly interpretable

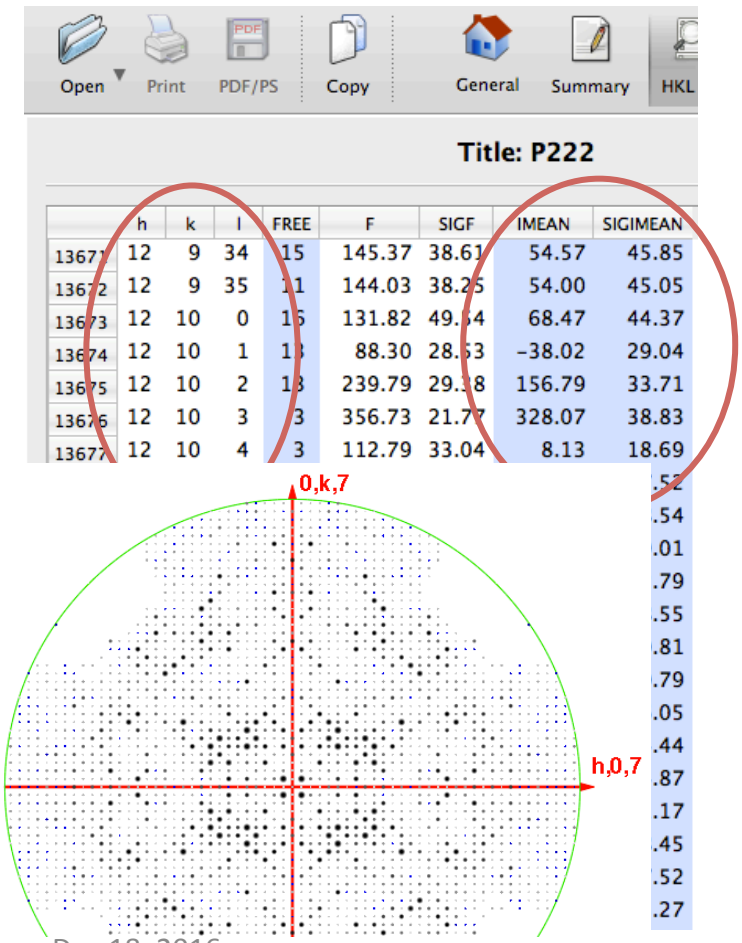
– model building

– real-space fitting of fragments

Intensities and Patterson map

Intensities $I(h,k,l)$

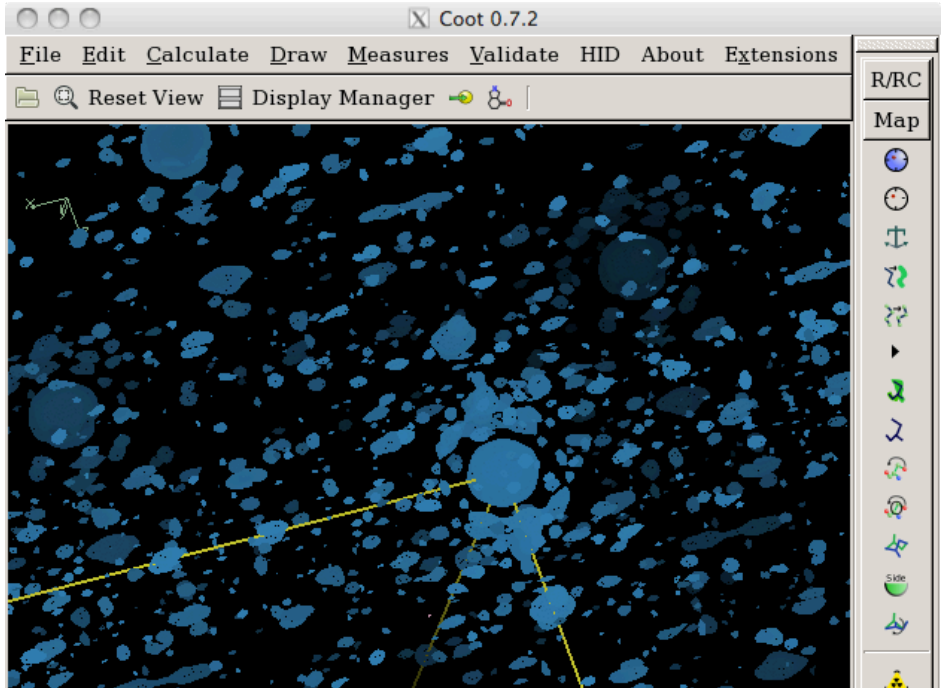
– 3-d discrete real function in the reciprocal space



	h	k	l	FREE	F	SIGF	IMEAN	SIGIMEAN
13671	12	9	34	15	145.37	38.61	54.57	45.85
13672	12	9	35	11	144.03	38.25	54.00	45.05
13673	12	10	0	16	131.82	49.54	68.47	44.37
13674	12	10	1	13	88.30	28.53	-38.02	29.04
13675	12	10	2	13	239.79	29.18	156.79	33.71
13676	12	10	3	3	356.73	21.77	328.07	38.83
13677	12	10	4	3	112.79	33.04	8.13	18.69

Patterson map:

– 3-d function in real(*) space



– Nothing reminds a protein molecule

– **Model building**, residue by residue, is **impossible**

Data and MR

MR: Two distinct cases dependent on availability of phases

- Data = structure factors (include phases)
  - "Search in the electron density"
    - Electron density maps are compared: calculated *vs*. observed
    - Model building is a more straightforward approach
  - » Useful in special cases

- Data = observed intensities (no phases)
  - "Patterson search"
    - Patterson maps are compared: calculated *vs*. observed
    - Direct model building is impossible in the absence of phases
  - » The most common case of MR

As a rule, all computations are in the reciprocal space

Self and cross vectors

Electron density map = peaks from all atoms

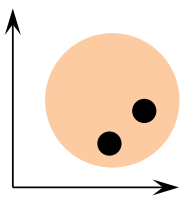
Patterson map = peaks from all interatomic vectors

- self-vectors: vectors between atoms belonging to the same molecule
- cross-vectors: vectors between atoms belonging to different molecules

One molecule

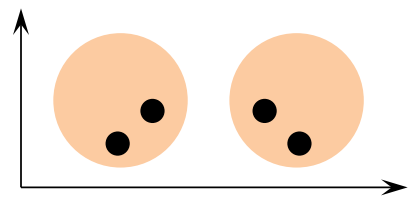
Two molecules

Electron density map

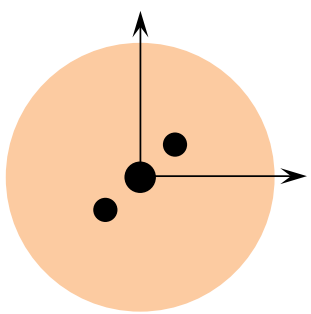


r_1

r_2



Patterson map

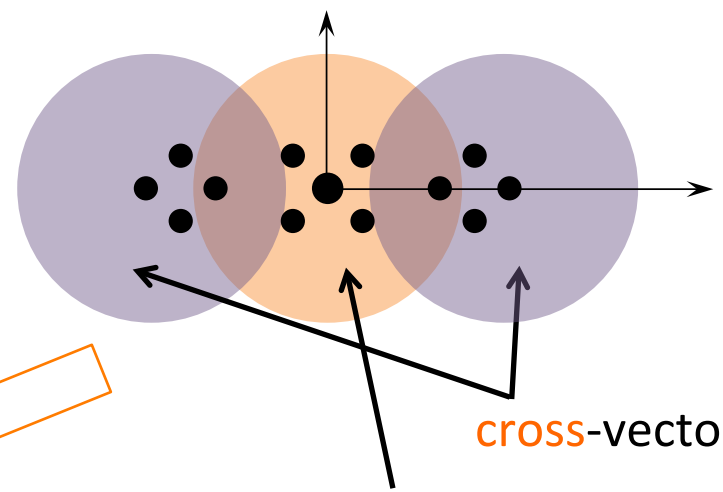


$r_1 - r_1$

$r_2 - r_2$

$r_1 - r_2$

$r_2 - r_1$



Contribution from self-vectors
– is centred at the origin
– dominates in a vicinity of the origin

What can be seen to be separated?

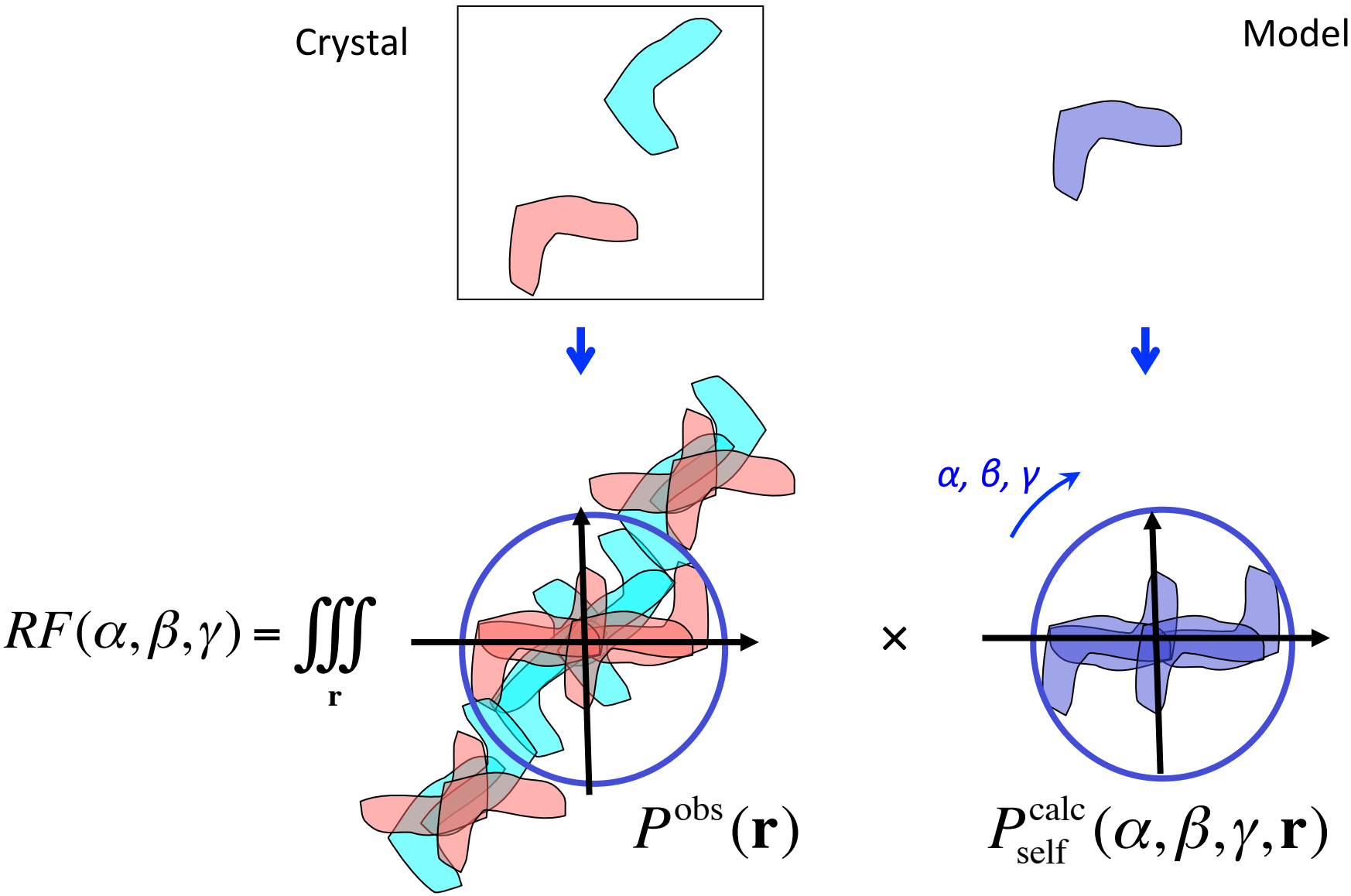
Electron density maps:

- Peaks from atoms or larger fragments are separated in space
  - Model building is possible

Patterson map:

- Contribution from self-vectors is centred at the origin
- Self-vectors are, in average, shorter than cross-vectors
  - Peaks from self-vectors dominates in a vicinity of the origin
  - Peaks from cross-vector dominates away from the origin
- One 6-dimensional search splits into
  - Rotation Function: 3-dimensional search (using self-vectors)
  - Translation Function: 3-dimensional search (using cross-vectors)

Rotation Function



Rotation Function

$$RF(\alpha,\beta,\gamma) = \iiint P^{\mathrm{obs}}(\mathbf{r}) \times P^{\mathrm{calc}}_{\mathrm{self}}(\alpha,\beta,\gamma,\mathbf{r})\,\mathrm{d}\mathbf{r}^3$$

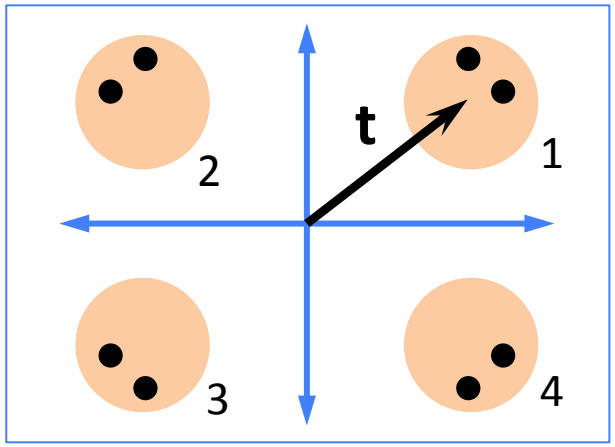
$P^{\mathrm{calc}}_{\mathrm{self}}(\alpha,\beta,\gamma,\mathbf{r})$ contains only

- self-vectors

$P^{\mathrm{obs}}(\mathbf{r})$ contains

- self-vectors (signal from one of the orientations in the crystal)
- cross-vectors (noise)

Translation Function



Analytical steps

- The centre of molecule 1
  - Parameter t
- Centres of molecules 2, 3 and 4
  - form symmetry operations
- $F_{\mathbf{h}}^{\text{calc}}(\mathbf{t})$, $I_{\mathbf{h}}^{\text{calc}}(\mathbf{t})$
- $TF(\mathbf{t}) = \sum_{\mathbf{h}} I_{\mathbf{h}}^{\text{obs}} \times I_{\mathbf{h}}^{\text{calc}}(\mathbf{t})$
- Can be converted to a form:

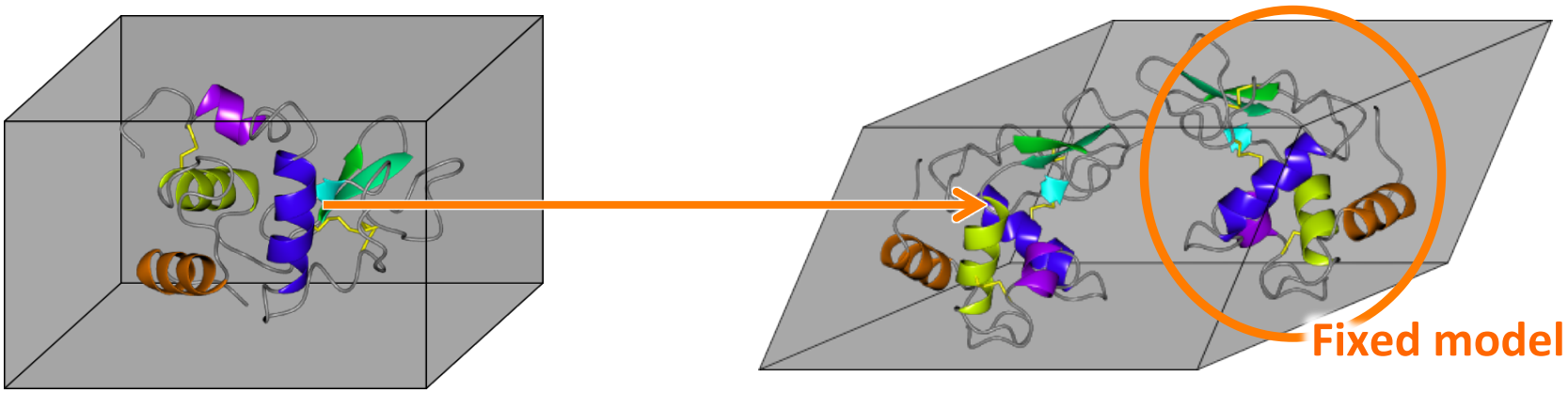
$$\boxed{TF(\mathbf{t}) = \sum_{\mathbf{h}} G_{\mathbf{h}} \exp(2\pi i\, \mathbf{h}\cdot\mathbf{t})}$$

  - FFT techniques can be applied

Numerical calculations

- $G_{\mathbf{h}}$
- FFT: $G_{\mathbf{h}} \rightarrow TF(\mathbf{t})$
- Peak search in $TF(\mathbf{t})$: best **t**

Fixed partial model



Almost the same equation as for a single molecule search,

$$TF(\mathbf{t}) = \sum_{\mathbf{h}} I_{\mathbf{h}}^{\text{obs}} \times \left| F_{\mathbf{h}}^{\text{fixed}} + F_{\mathbf{h}}^{\text{calc}}(\mathbf{t}) \right|^2$$

Again, can be converted to a form:

$$\boxed{TF(\mathbf{t}) = \sum_{\mathbf{h}} G_{\mathbf{h}} \exp(2\pi i\, \mathbf{h} \cdot \mathbf{t})}$$

and FFT technique can be used

Translation Function

$$TF(\mathbf{t}) = \iiint P^{\mathrm{obs}}(\mathbf{r}) \times P^{\mathrm{calc}}_{\mathrm{cross}}(\mathbf{t},\mathbf{r})\,\mathrm{d}\mathbf{r}^3$$

$P^{\mathrm{calc}}_{\mathrm{cross}}(\mathbf{t},\mathbf{r})$ contains

- cross-vectors

$P^{\mathrm{obs}}(\mathbf{r})$ contains

- self-vectors (background or noise)
- cross-vectors (relevant vectors: signal, others: noise)

Packing considerations

Molecules in the crystal do not overlap

How can we use this information?

» Patterson map does not explicitly reveal molecular packing

Reject MR solutions

- Restrict distance between centres of molecules
- Count close interatomic contacts

Modify TF

- Divide by Overlap Function
- Multiply by Packing function

Fast Packing Function

Estimation of overlap

- Using mask from search model
- FFT

Packing Function

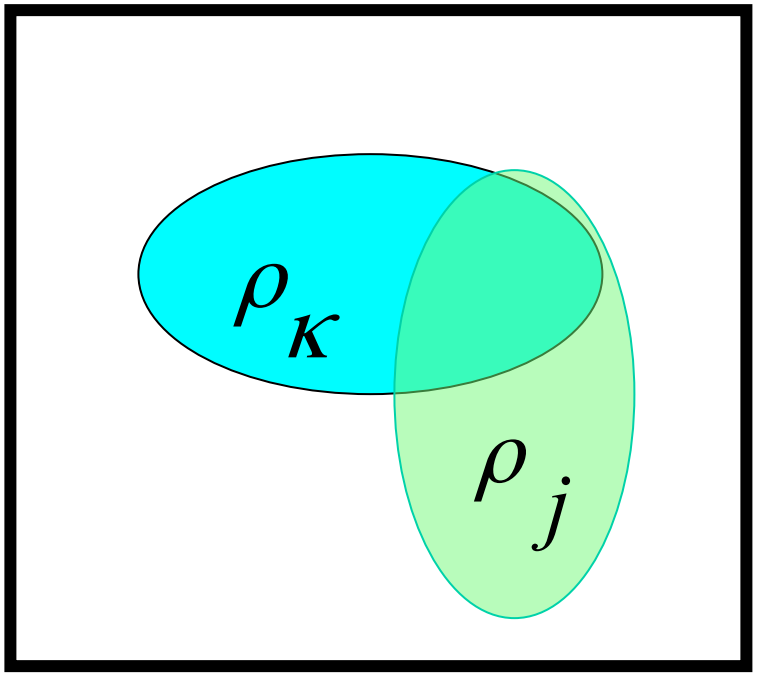
$PF = 1 - \text{overlap}$

Modified Translation Function

$TF_{mod} = TF \times PF$

Peak search

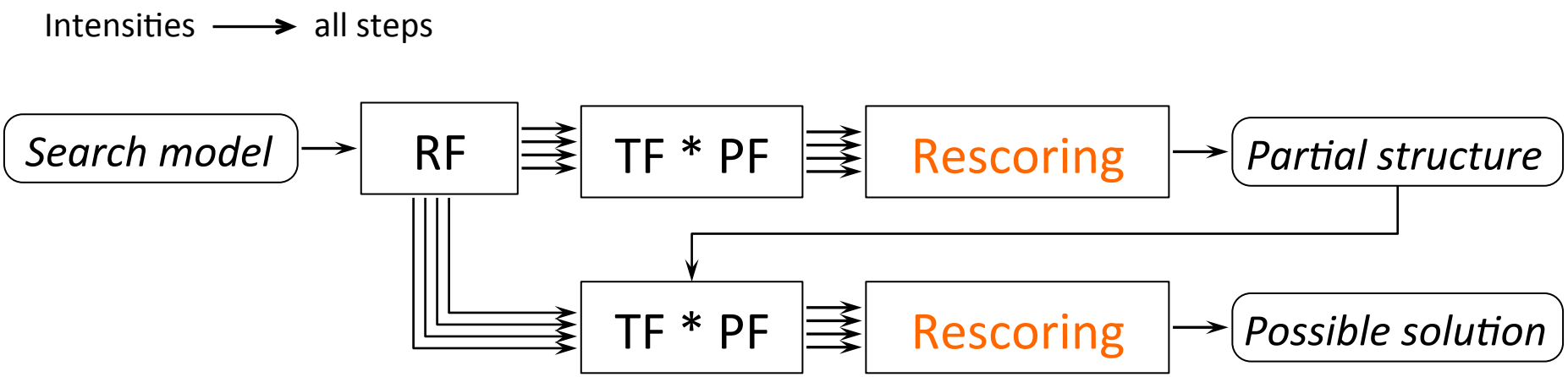
- Using TF_{mod}
- No irrelevant peaks are passed to rescoring step



Implemented in *MOLREP*

A conservative MR protocols for two copies of a model

As implemented in *MOLREP*



$RF = \sum_{hkl} w * I_O * I_C(\alpha\beta\gamma)$

$TF = \sum_{hkl} w * I_O * I_C(xyz)$

Rescoring: Correlation Coefficient* PF

Specialised MR techniques

- Search in the density (phased MR)
- Handling Translational Non-Crystallographic symmetry
  - Non-origin peaks in the Patterson map indicate the presence of TNCS
  - Requires special handling of model errors (Phaser)
  - Molecules related by TNCS can be found in one go as they have nearly the same orientation
- Self Rotation Function
- Locked RF and TF
  - Using point symmetry of oligomers
- Exhaustive searches
- Stochastic searches

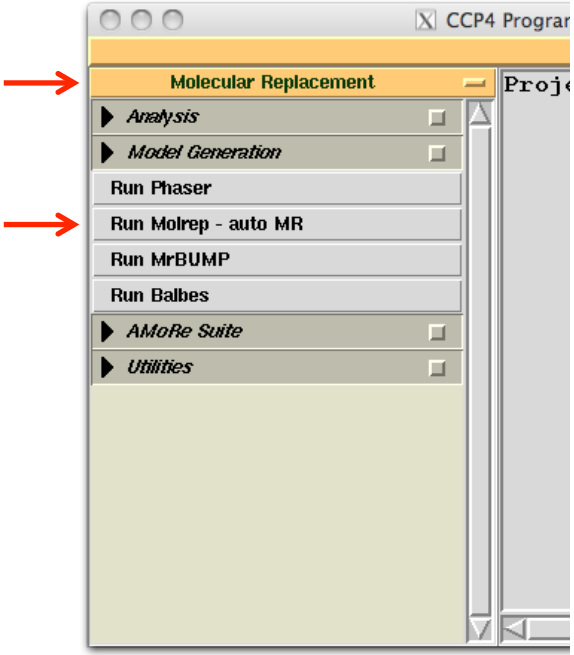
Molrep

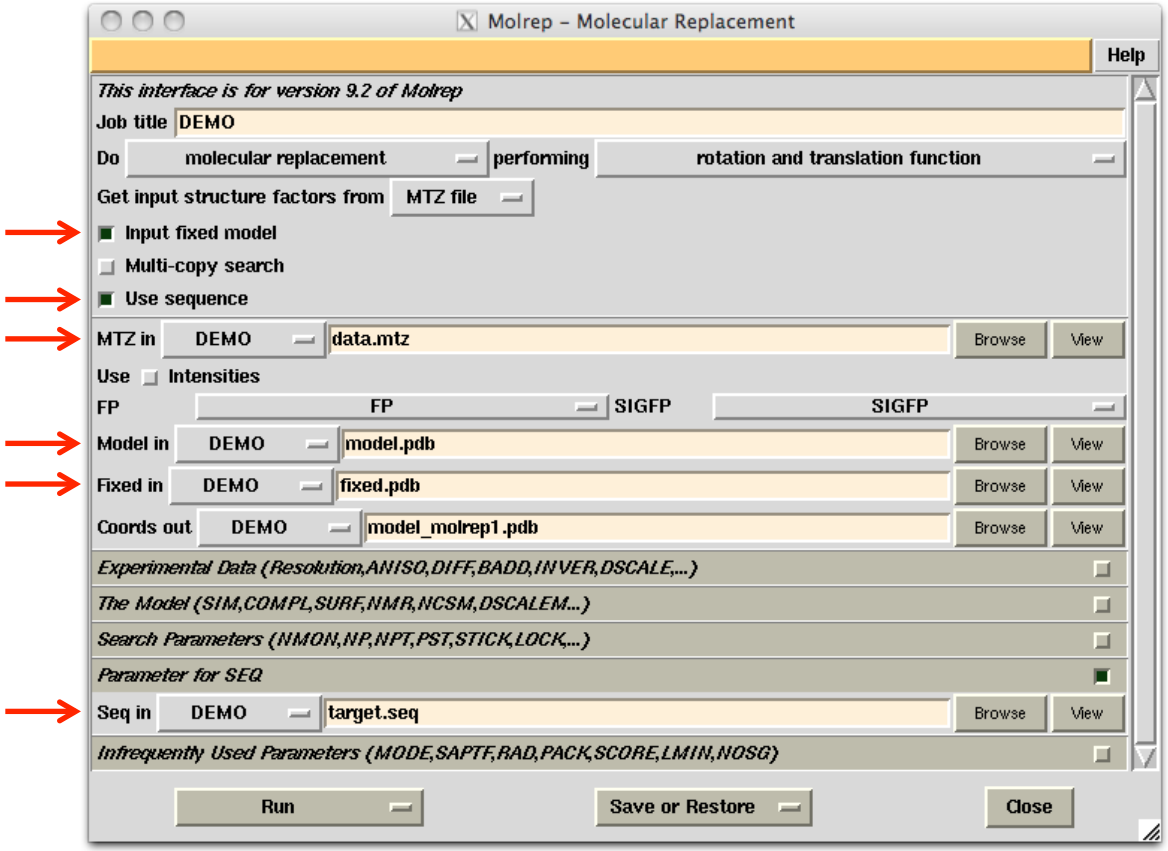
Alexey Vagin

YSBL University of York

Molrep

```
molrep -f data.mtz -m model.pdb -mx fixed.pdb -s target.seq
```





Default protocol

```
molrep -f data.mtz -m model.pdb -mx fixed.pdb -s target.seq
```

- model correction if sequence provided
- defines the number of molecules per AU
- modification of the model surface
- anisotropic correction of the data
- weighting the data according to model completeness and similarity
- check for pseudotranslation and use it if present
- 30+ peaks in Cross RF for use in TF (accounts for close peaks)
- applied packing function, TF × PF
- rescoring using CC × PF
- make use of partial structure (fixed model)

Default protocol: log-file

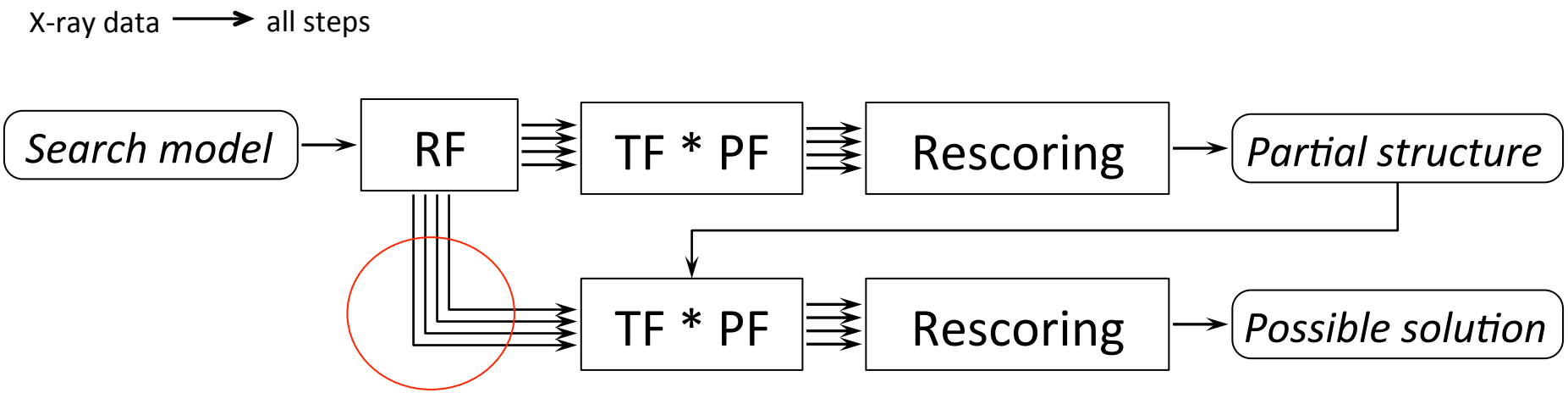
CCP4I fileviewer 1_molrep.log

Help

```
   14   4      8.444       2.920    1.00   1.00 -21.06   0.599   0.110    7.33 ( 0.242)
INFO: contrast is good enough. Stop this run

                                --- Summary ---
+-----------------------------------------------------------------------------------+
|      RF   TF    theta     phi      chi     tx      ty      tz     TFcnt   wRfac   Score |
+-----------------------------------------------------------------------------------+
|   1    1    1    72.59    38.64   179.42  0.825   0.649   0.480   10.06   0.560   0.242 |
|   2    2    1    72.41    38.93   177.39  0.820   0.650   0.480   10.91   0.565   0.217 |
|   3    4    1    72.18    38.99   176.40  0.819   0.652   0.480    9.61   0.573   0.195 |
|   4    7    2    77.85    58.68   142.53  0.445   0.292   0.483    5.03   0.602   0.121 |
|   5    3    2   107.48  -166.00   160.39  0.637   0.790   0.175    4.51   0.599   0.120 |
|   6    6   10    52.26    91.15    50.93  0.416   0.376   0.163    1.95   0.603   0.111 |
|   7   13   12    82.51   133.98   129.34  0.542   0.566   0.253    2.80   0.601   0.110 |
|   8   14    4    81.86    91.66   108.52  0.780   0.260   0.469    2.92   0.599   0.110 |
|   9    9    4   113.57   167.71   124.63  0.757   0.436   0.021    3.39   0.603   0.110 |
|  10    8   13    87.47   114.84   104.62  0.644   0.955   0.369    1.21   0.605   0.109 |
|  11    5    3   108.24  -136.26   176.12  0.816   0.651   0.479    2.79   0.602   0.109 |
|  12   12    1    97.58   104.76    90.32  0.585   0.049   0.166    2.33   0.607   0.107 |
|  13   10    2    98.10   104.76    89.79  0.586   0.049   0.166    1.93   0.607   0.107 |
|  14   11    9    36.40    73.27   110.10  0.394   0.165   0.289    1.16   0.610   0.097 |
+-----------------------------------------------------------------------------------+

Contrast =      7.33

After stick correction:
 Move closer to origin
 I_sym_operator     :    11
 new position(frac):  -0.176   -0.351    0.020


 Nmon RF   TF    theta     phi      chi    tx      ty      tz      TFcnt   wRfac   Score
   1   1    1    21.87  -179.03   106.86 -0.176  -0.351   0.020   10.06   0.560   0.242

--- convert "molrep.crd" to "molrep.pdb" ---
 Time:      1h 27m 58s   Elapsed:      0h   0m 57s
MOLREP(ccp4):  Normal termination
Times: User:       53.8s System:      2.4s Elapsed:      0:57
```

Find | Show Log Graphs | Show Summary | Quit

Molrep default protocols for two copies of a model

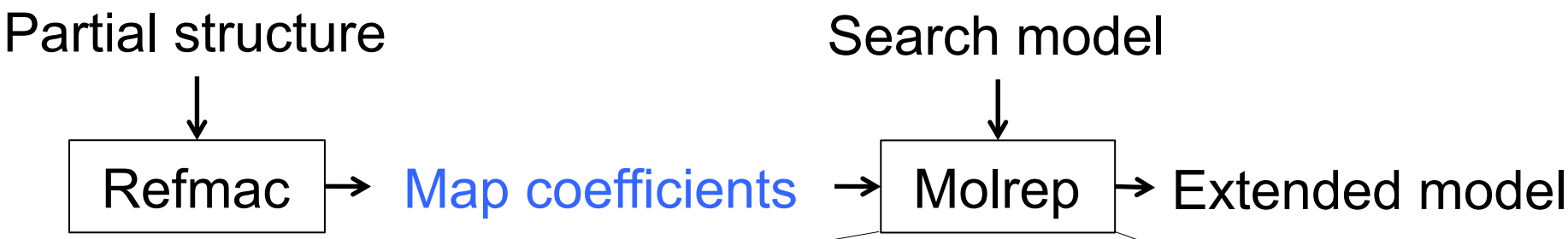


$RF = \sum_{hkl} w * I_O * I_C(\alpha\beta\gamma)$

$TF = \sum_{hkl} w * I_O * I_C(xyz)$

Rescoring: Correlation Coefficient* PF

Search in the electron density map



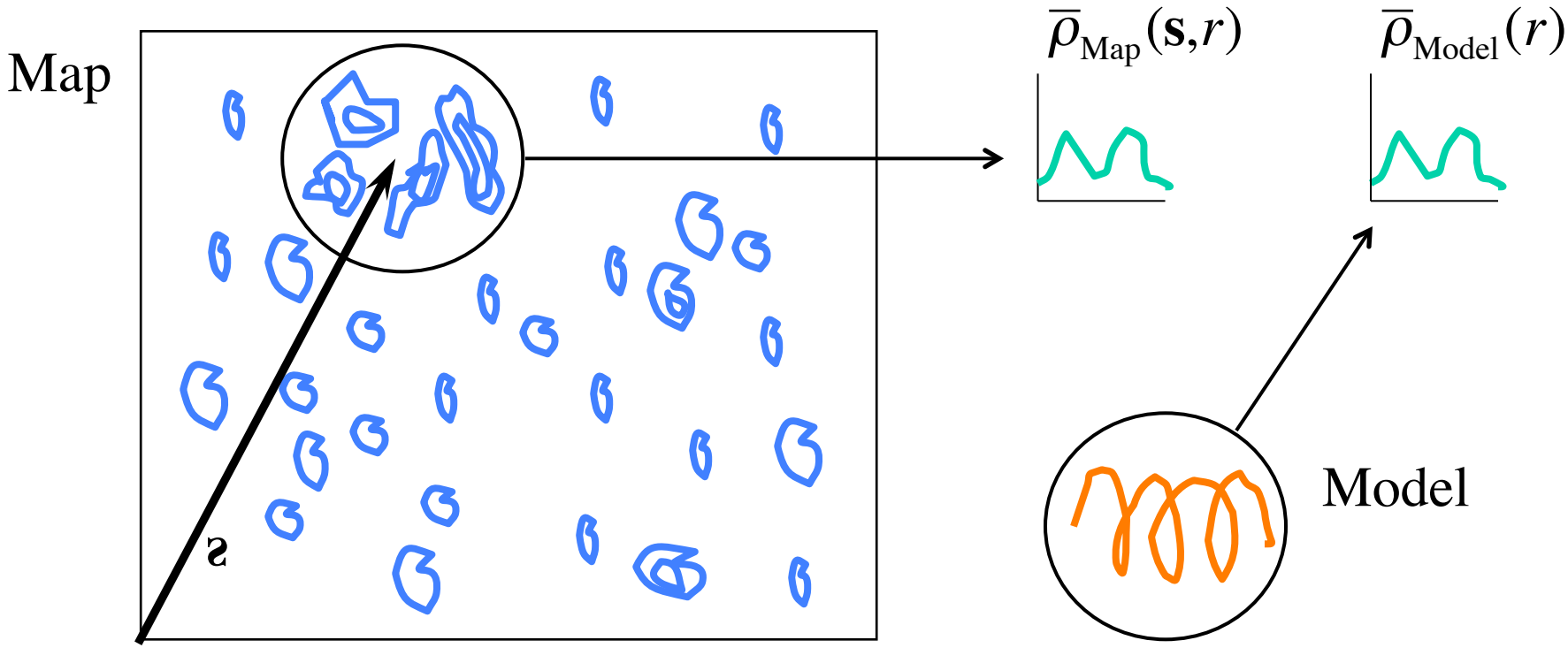
Search in the map

- Calculate 2-1 or 1-1 maps after restrained refinement of partial structure
- Flatten the map corresponding to the known substructure
- Calculate structure amplitudes from the modified map
- Use these modified amplitudes in Rotation Function
- And finally – Phased TF

Molrep: SAPTF

Spherically Averaged Phased Translation Function
(FFT based algorithm)

$$\mathrm{SAPTF}(\mathbf{s}) = \int \bar{\rho}_{\mathrm{Map}}(\mathbf{s},r)\, \bar{\rho}_{\mathrm{Model}}(r)\, r^2\, \mathrm{d}r$$



Molrep: Search in the map with SAPTF

1. Find approximate position:
   Spherically Averaged Phased Translation Function

2. Find orientation:
   Local Phased Rotation Function
   - Local search of the orientation in the density

3. Verify and adjust position:
   Phased Translation Function

Molrep: Search in the map with SAPTF

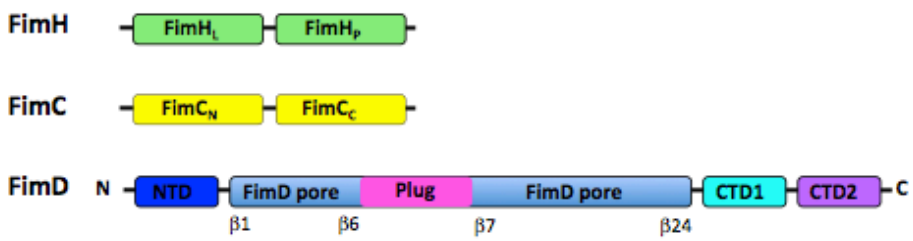
1. Find approximate position:
   Spherically Averaged Phased Translation Function

2. Find orientation:
   Local Rotation Function
   - Structure amplitudes from the density within the SAPTF sphere

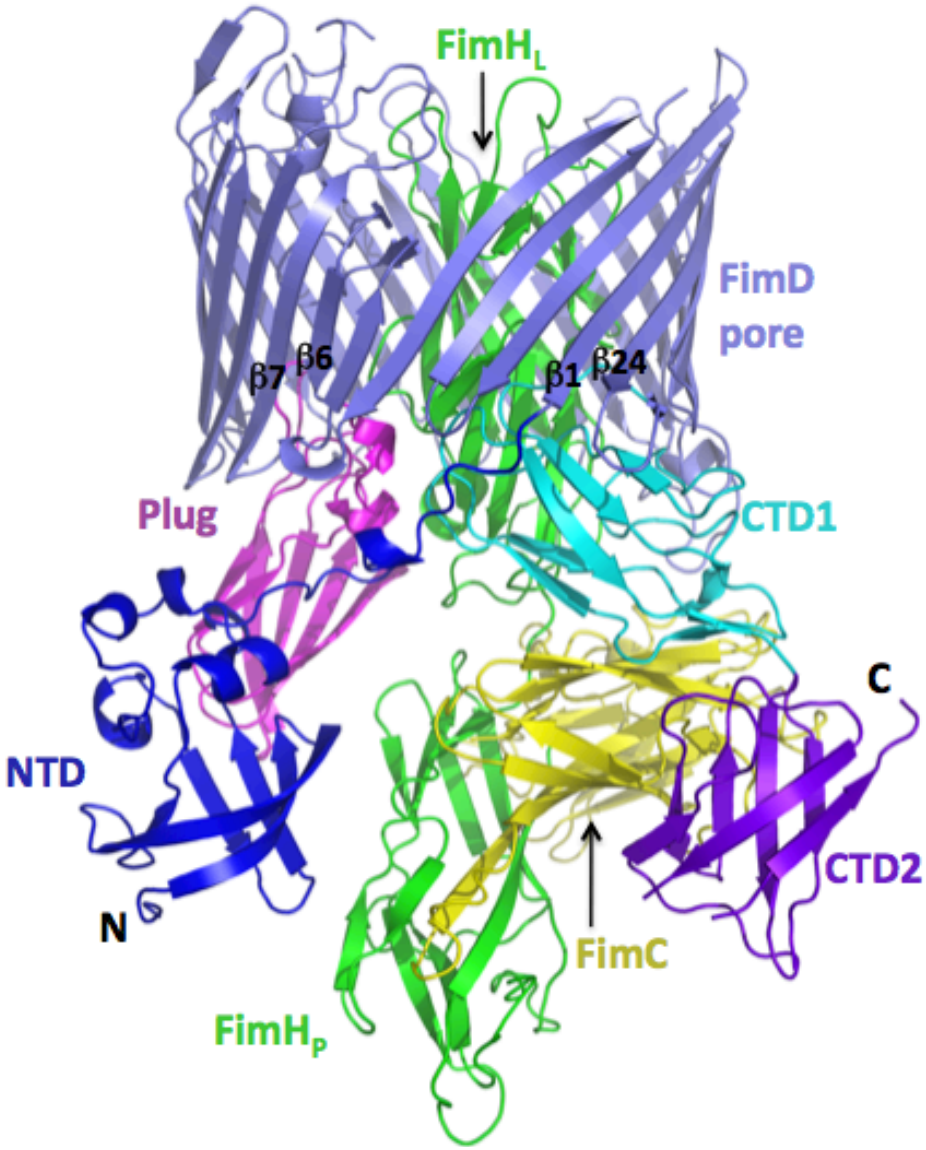
3. Verify and adjust position:
   Phased Translation Function

- Local RF is less sensitive than Phased RF to inaccuracy of the model position

Example



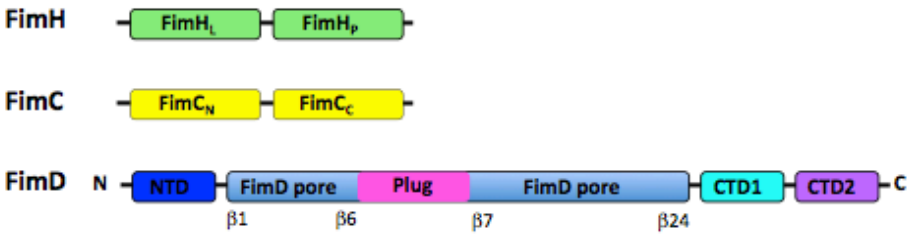
- Asymmetric unit two copies
- Resolution 2.8 Å



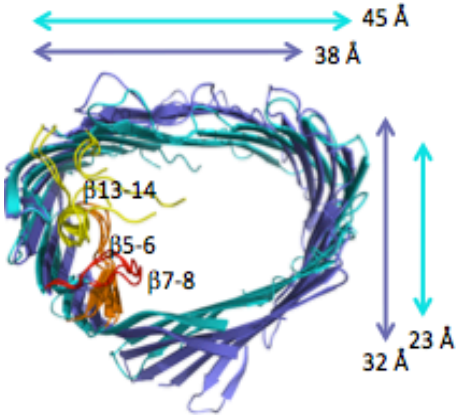
Phane et. al (2011) Nature, 474, 50-53

Usher complex structure solution

1. Conventional MR
   - FimC-N + FimC-C
   - FimH-L + FimH-P
   - FimD-Pore



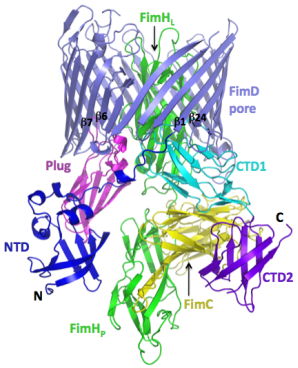
2. Jelly body refinement (Refmac)
   - FimD-Pore



3. <u>Fitting into the electron density</u>
   - FimD-Plug
   - FimD-NTD
   - FimD-CTD-2

4. Manual building
   - FimD-CTD-1

Performance of fitting methods



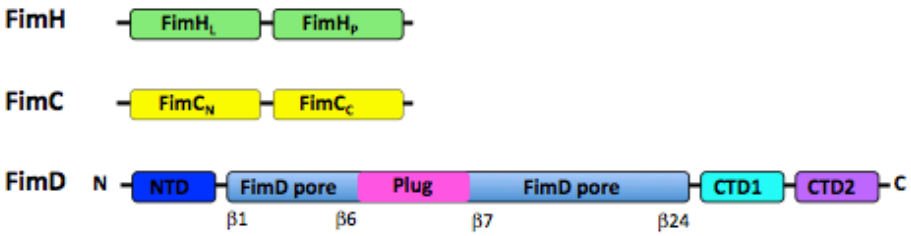


Table: the number of copies found by different methods

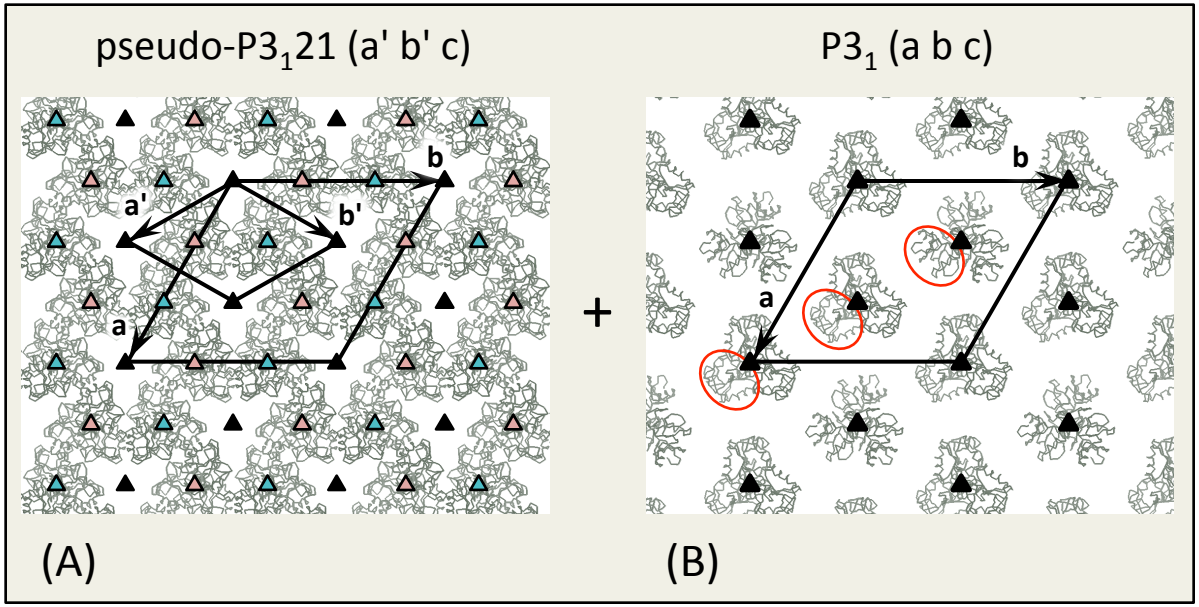
- Trying several methods is a good idea (also because of cross-validation)

	search model	sequence identity	"Masked" RF PTF	SAPTF PRF PTF	SAPTF Local RF PTF
FimD-Plug	3fip_A	38.5%	2	–	1
FimD-NTD	1ze3_D	100%	2	1	2
FimD-CTD-2	3l48_A	33.3%	–	2	–

Twinned crystal with pseudo-symmetric substructure

Human macrophage receptor CLEC5A for dengue virus

Watson, A. A. et al. (2011). J Biol Chem 286, 24208-18.



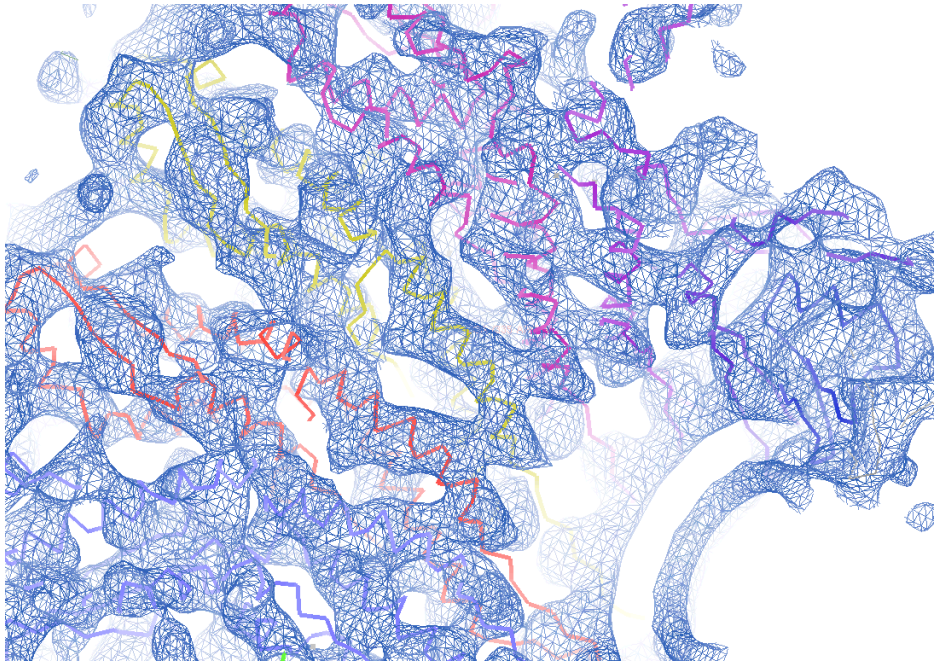
3-fold axes with respect to the true structure:

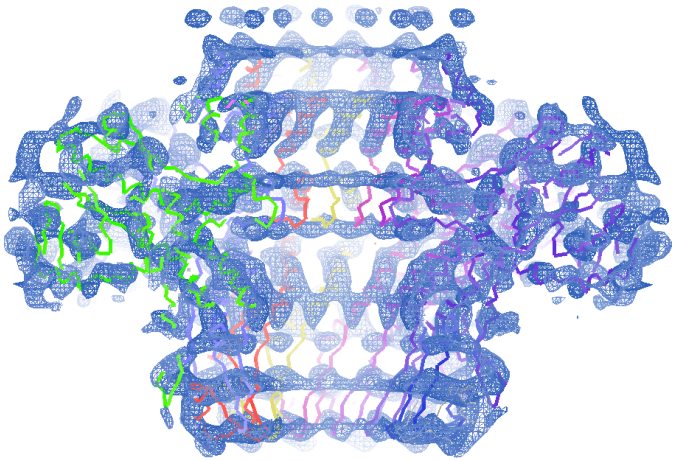
▲ crystallographic

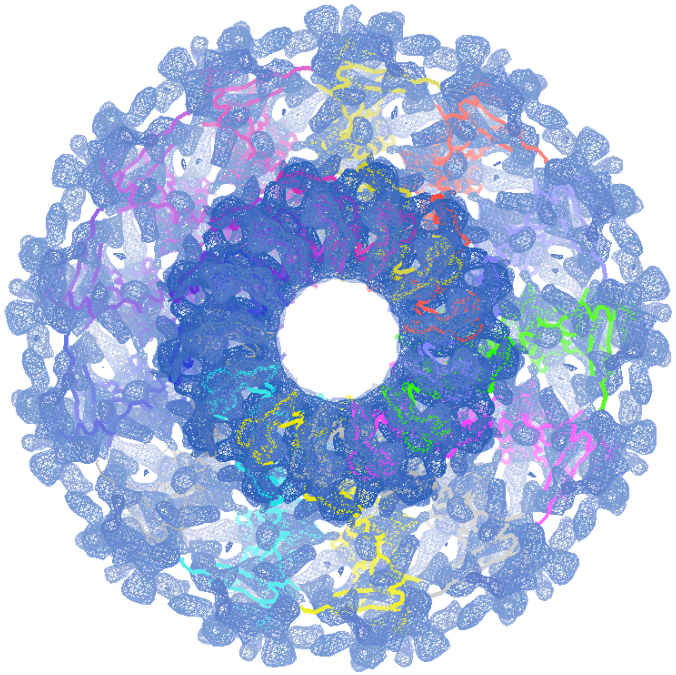
△ △ pseudosymmetry for (A)

- Substructure (A): Patterson search (4 copies), search in density (2 copies)
- Substructure (B): Search in the density (3 copies)

Fitting into EM maps

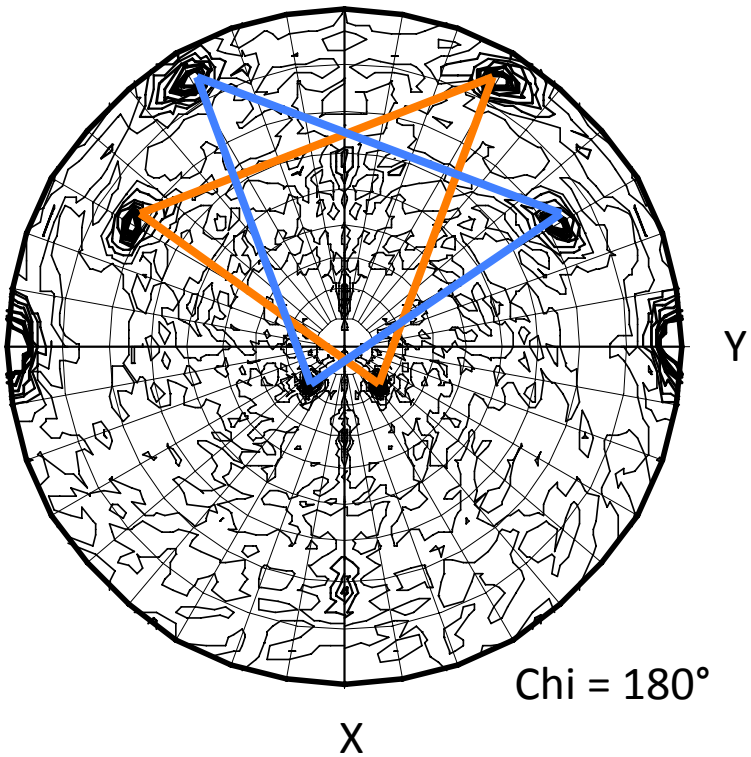






SPP1 portal protein

Self Rotation Function (SRF)



Example of SRF

- Space group P21
- One 222-tetramer in the AU

Preliminary analysis of X-ray data

- Oligomeric state of the protein in crystal
- Selection of oligomeric search model

Limited use

- No clear interpretation or even artifact peaks in high symmetry point groups (e.g. 622)
- different oligomers with the same symmetry

Locked Rotation Function

- Uses SRF to derive NCS operations
- Averages RF over NCS operations
- In favorable cases Improves signal to noise ratio in RF

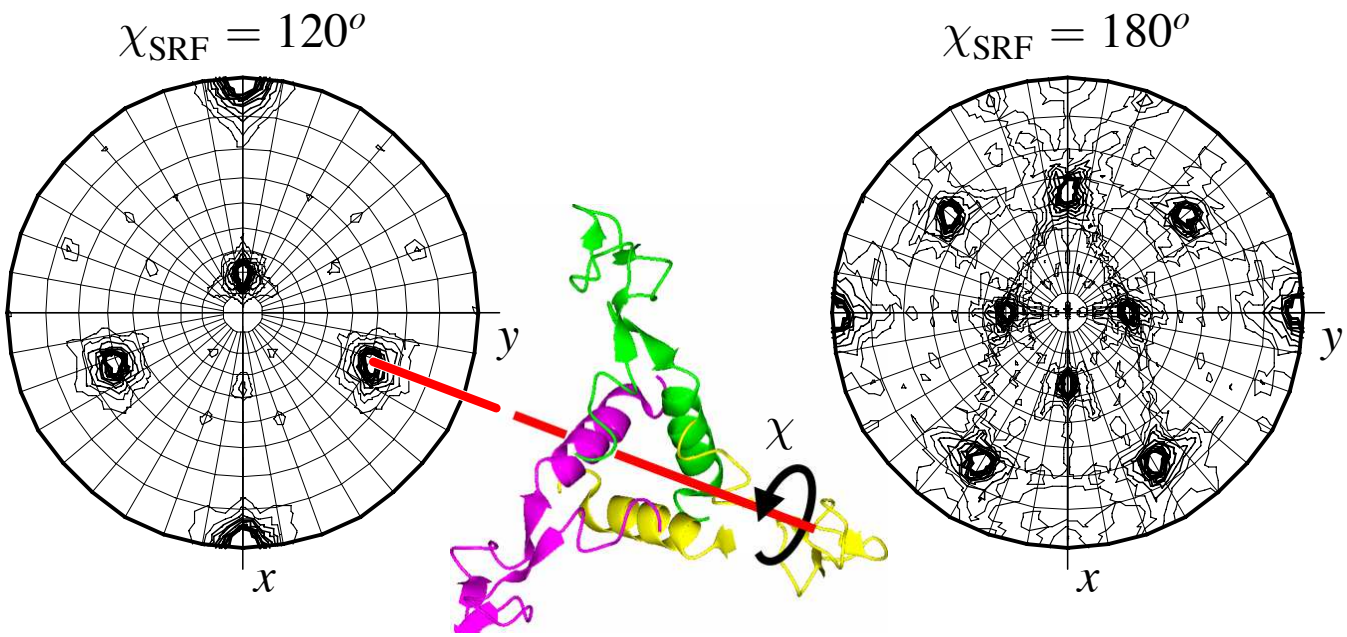
Automatic mode:

```
molrep -f s100.mtz -m monomer.pdb -s s100.seq -i <<+
lock y
+
```

There is an option of selecting specific SRF peaks

Also implemented in CCP4I

One-dimensional exhaustive search (exotic case)

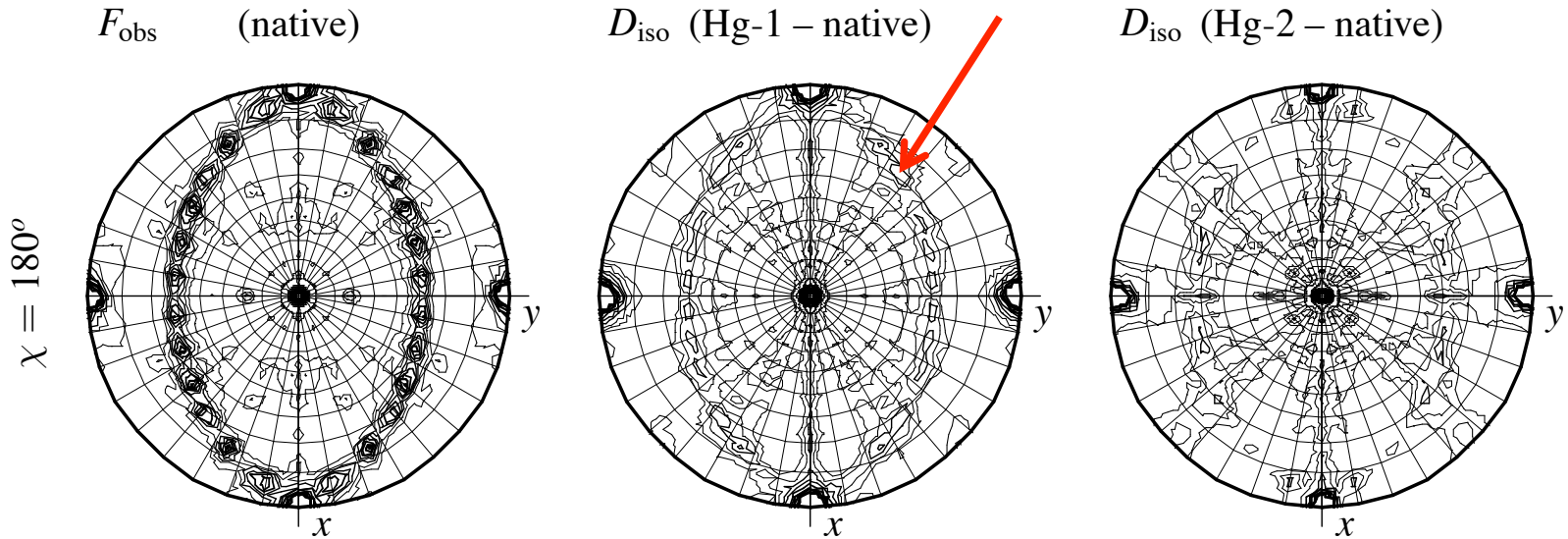


SRF helps restrict dimensionality in an exhaustive search

- Orientation of the trimer is known from the analysis of SRF
- Unknown parameter: rotation about 3-fold axis
- One-parametric exhaustive search using TF as score function

MR substructure solution (exotic case)

- Select isomorphous derivative
  - by comparing native SRF and SRF from D-iso



- Hg-substructure is a 13-atom ring (from native SRF analysis)
  - Orientation of the ring is known from the analysis of SRF
  - Unknown parameters: radius of the ring, rotation about 13-fold axis
- Two-parametric exhaustive search

Which direction does MR go?

Automation:

✖ Collection of tricks

✔ Improvement of "standard" methods
✔ Better scoring system

✔✔ Models

Model improvement

✓ Structural information in MR models

✓ Extra information in sequence alignment

✓ Many homologous structures

- Remove residues that do not align
- Remove "excessive" atoms from aligned residues
- Ensemble models (several superposed models)

CCCP4 programs for model preparation

Single model correction:

- Chainsaw
- Molrep
- Sculptor

Preparation of ensemble models – fitting models:

- Lsqkab
- ProSMART
- SSM (also in Coot)
- Gesamt

Automated preparation of ensemble models:

- Ensembler

Molecular Replacement in CCP4

MR Programs

- AMoRe
- Molrep
- Phaser

MR Pipelines

- MrBUMP
- Balbes
- AMPLE
- Arcimboldo