Refinement with the program - REFMAC

Garib N Murshudov

MRC-LMB

Cambridge

Contents

1) Introduction
2) Problems of low-resolution refinement
3) TWIN refinement in REFMAC: some warnings about twin refinement
4) Some tools for low resolution: ncs, external restraints, B value restraints and “jelly” body
5) Some approaches to low resolution refinement
6) Map sharpening: General approach and some applications

Available refinement programs

- SHELXL
- CNS
- REFMAC5
- TNT
- BUSTER/TNT
- Phenix.refine
- RESTRAINT
- MOPRO
- XD
- MAIN

What can REFMAC do?

- Simple maximum likelihood restrained refinement
- Twin refinement: Warning
- Phased refinement (with Hendrickson-Lattmann coefficients)
- SAD/SIRAS refinement
- Structure idealisation
- Library for more than 10000 ligands (from the next version)
- Covalent links between ligands and ligand-protein
- Rigid body refinement
- NCS local, restraints to external structures, jelly body
- TLS refinement
- Map sharpening: Inverse problem, Bvalues etc
- etc

Considerations in refinement

- Function to optimise (link between data and model)
  - Should use experimental data
  - Should be able to handle chemical (e.g bonds) and other (e.g. NCS, structural) information
- Parameters
  - Depends on the stage of analysis
  - Depends on the amount and quality of the experimental data
- Methods to optimise
  - Depends on stage of analysis: simulated annealing, conjugate gradient, second order (normal matrix, information matrix, second derivatives)
  - Some methods can give error estimate as a by-product. E.g second order.

Two components of target function

Crystallographic target functions have two components: one of them describes the fit of the model parameters into the experimental data and the second describes chemical integrity (restraints).

Currently used restraints are: bond lengths, angles, chirals, planes, ncs if available, some torsion angles, reference structures

Crystallographic refinement

The function in crystallographic refinement has a form:

$$L(p)=wL_X(p)+L_G(p)$$

Where $L_X(p)$ is -loglikelihood and $L_G(p)$ is -log of prior probability distribution - restraints.

It is one of many possible formulations. It uses Bayesian statistics. Other formulation is also possible. For example: stat physical energy, constrained optimisaton, inverse problem etc.

Various forms of X-ray component

- SAD functions uses observed F+ and F- directly without any preprocessing by a phasing program (It is not available in the current version but will be available soon)
- SIRAS – uses native and derivative anomalous data directly
- MLHL - explicit use of phases with Hendrickson Lattman coefficients
- Rice - Maximum likelihood refinement without phase information

-loglikelihood

-loglikelihood depends on assumptions about the experimental data, crystal contents and parameters. For example with assumptions that all observations are independent (e.g. no twinning), there is no anomolous scatterers and no phase information available, for acentric reflections it becomes:

$$L_X(p) = \sum \frac{|F_o|^2 + |F_c|^2}{\Sigma} - \log(I_0(2|F_o||F_c|/\Sigma)) + \log(\Sigma) + const$$

And for centric reflections:

$$L_X(p) = \sum \frac{|F_o|^2 + |F_c|^2}{2\Sigma} - \log(\cosh(|F_o||F_c|/\Sigma)) + 0.5\log(\Sigma) + const$$

All parameters (scale, other overall and atomic) are inside $|F_c|$ and Σ

Note that these are loglikelihood of multiples of chi-squared distribution with degree of freedom 2 and 1

Map calculation

- After refinement programs usually give coefficients for two type of maps: 1) 2Fo-Fc type maps. They represent the content of the crystal. 2) Fo-Fc type of maps. They represent difference between contents of the crystal and current atomic model. Both maps should be inspected and model should be corrected if necessary.
- Refmac gives coefficients:

  $2\ m\ F_o - D\ F_c$ – to represent contents of the crystal

  $m\ F_o - D\ F_c$ - to represent differences

m is the figure of merit (reliability) of the phase of the current reflection and D is related to model error. m depends on each reflection and D depends on resolution. Unobserved reflections are replaced by DFc.

If phase information is available then map coefficients correspond to the combined phases.

Problems of low resolution refinement

1) Function to describe fit of the model into experiment: likelihood or similar
    1) Data may come from very peculiar “crystals”: Twin, OD, multiple cell
    2) Radiation damage
    3) Converting I-s to |F| may not be valid: weak reflections, modulated crystals
2) Limited and noisy data: use of available knowledge
    1) High B value and spread of B values (!!!)
    2) Severe incompleteness of models
3) Smeared electron density with vanishing side chains, secondary structures, domains: High B values and series termination

TWIN

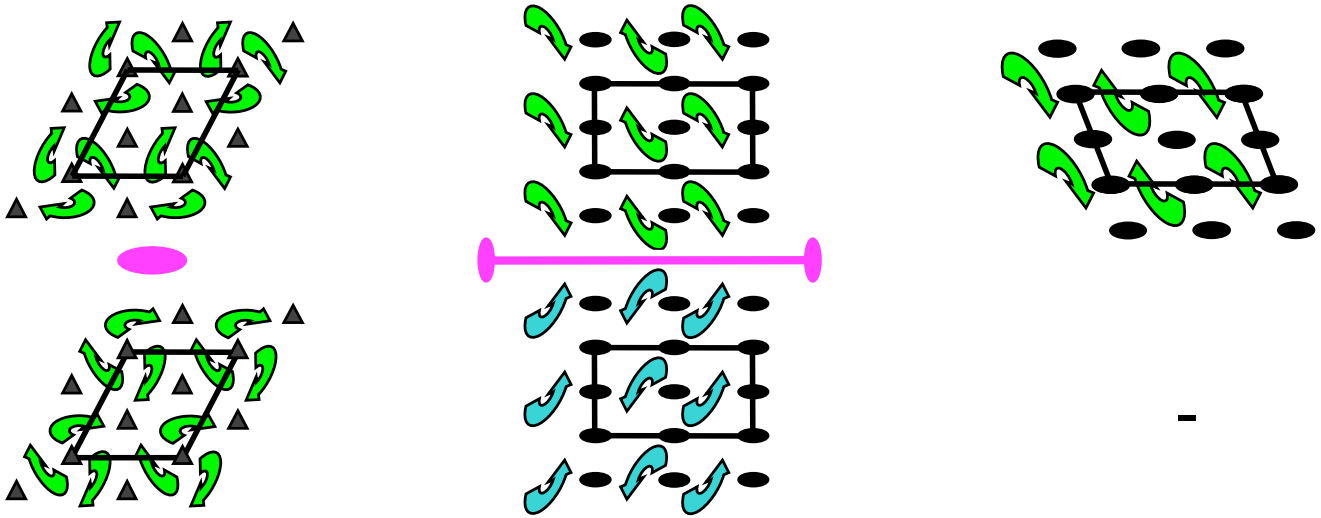
merohedral and pseudo-merohedral twinning

Crystal symmetry:	**P3**	**P2**	**P2**
Constrain:	**-**	**β = 90º**	**-**
Lattice symmetry *:	**P622**	**P222**	**P2**
(rotations only)			
Possible twinning:	**merohedral**	**pseudo-merohedral**	**-**

Domain 1

Twinning operator

Domain 2

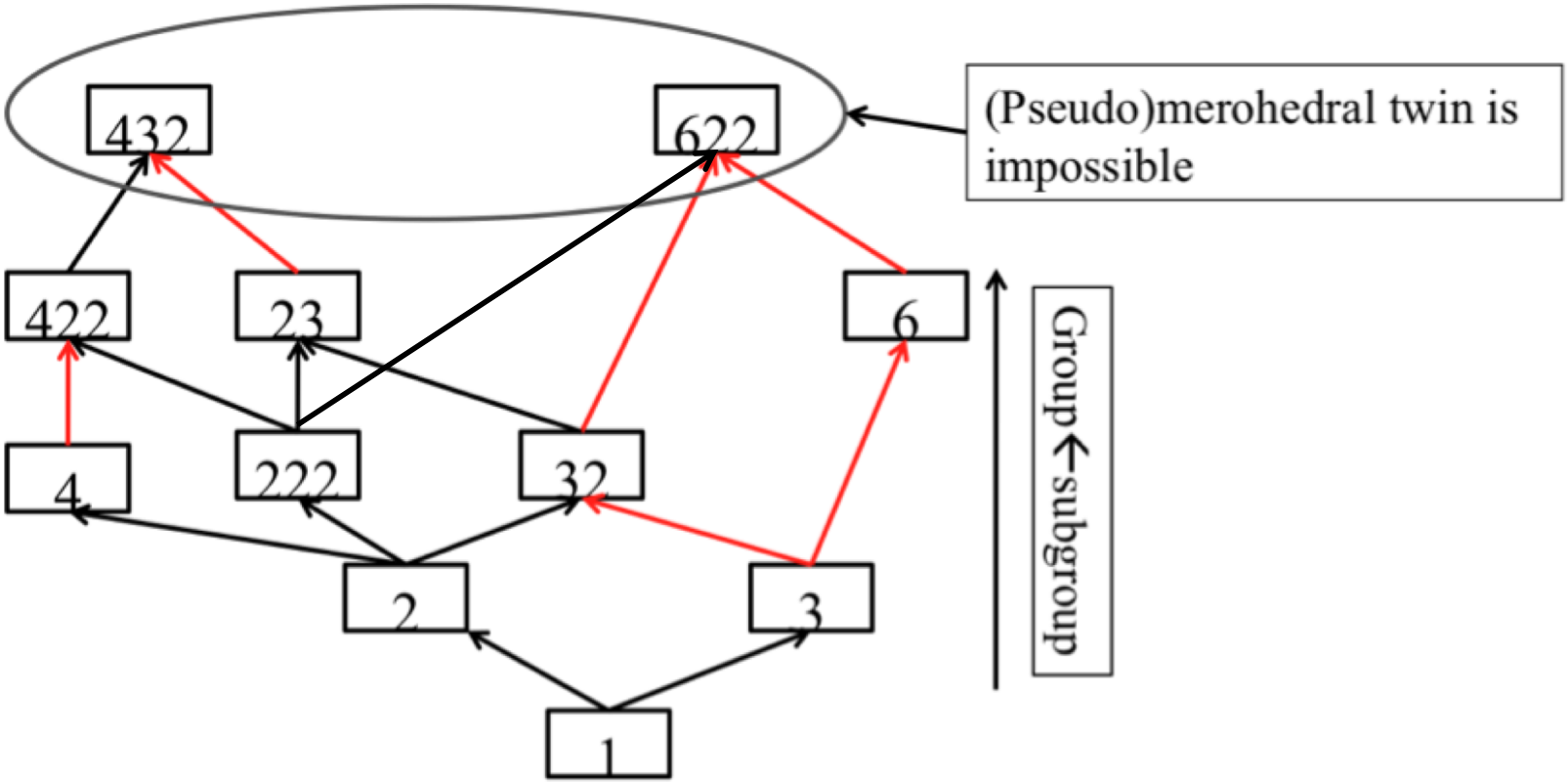


-

Crystal lattice is invariant with respect to twinning operator.

The crystal is NOT invariant with respect to twinning operator.

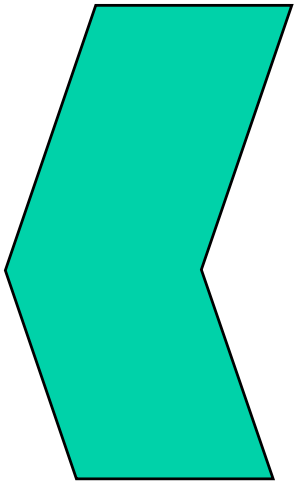
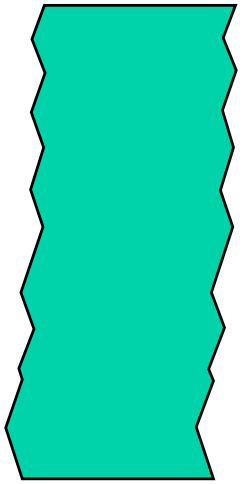
Twin refinement: Group/subgroup



Red arrows: No constraints are needed, merohedral twin could happen
Black arrow: Additional constraints on cell parameters are needed, pseudo merohedral twinning can happen

31

The whole crystal: twin or polysynthetic twin?

	twin	polysynthetic twin
		
A single crystal can be cut out of the twin:	yes	no

The shape of the crystal suggested that we dealt with polysynthetic OD-twin

Twinning

If we have only two domains related with twin operator then observed intensities will be

$$I_{T1} = (1-\alpha)I_1 + \alpha I_2$$
$$I_{T2} = (1-\alpha)I_2 + \alpha I_1$$

I_{T1} and I_{T2} are observed intensities, I_1 and I_2 are intensities from single crystals, α is proportion of the second domain. α is between 0 and 0.5. When it is 0.5 then twin called perfect twin.

In principle these equations can be solved and I_1 and I_2 (intensities for single crystal) can be calculated. It is called detwinning. It turns out that detwinning increases errors in intensities. Also, completeness after detwinning can decrease substantially. Moreover when α=0.5 it is impossible to detwin.

For some purposes (e.g. for phasing, sometimes for molecular replacement) detwinning may give reasonable results. For refinement general rule is to avoid detwinning and use the data directly. Almost all known refinement programs can handle twinning to a certain degree.

Twin refinement in REFMAC

Twin refinement in refmac (5.5 or later) is automatic.

- Identify “twin” operators
- Calculate “Rmerge” ($\Sigma|I_h-\langle I\rangle_{twin}|\ /\Sigma I_h$) for each operator. If Rmerge < 0.44 keep it: Twin plus crystal symmetry operators should form a group
- Refine twin fractions. Keep only “significant” domains (default threshold is 5%): Twin plus symmetry operators should form a group

Intensities can be used

If phases are available they can be used

Maximum likelihood refinement is used

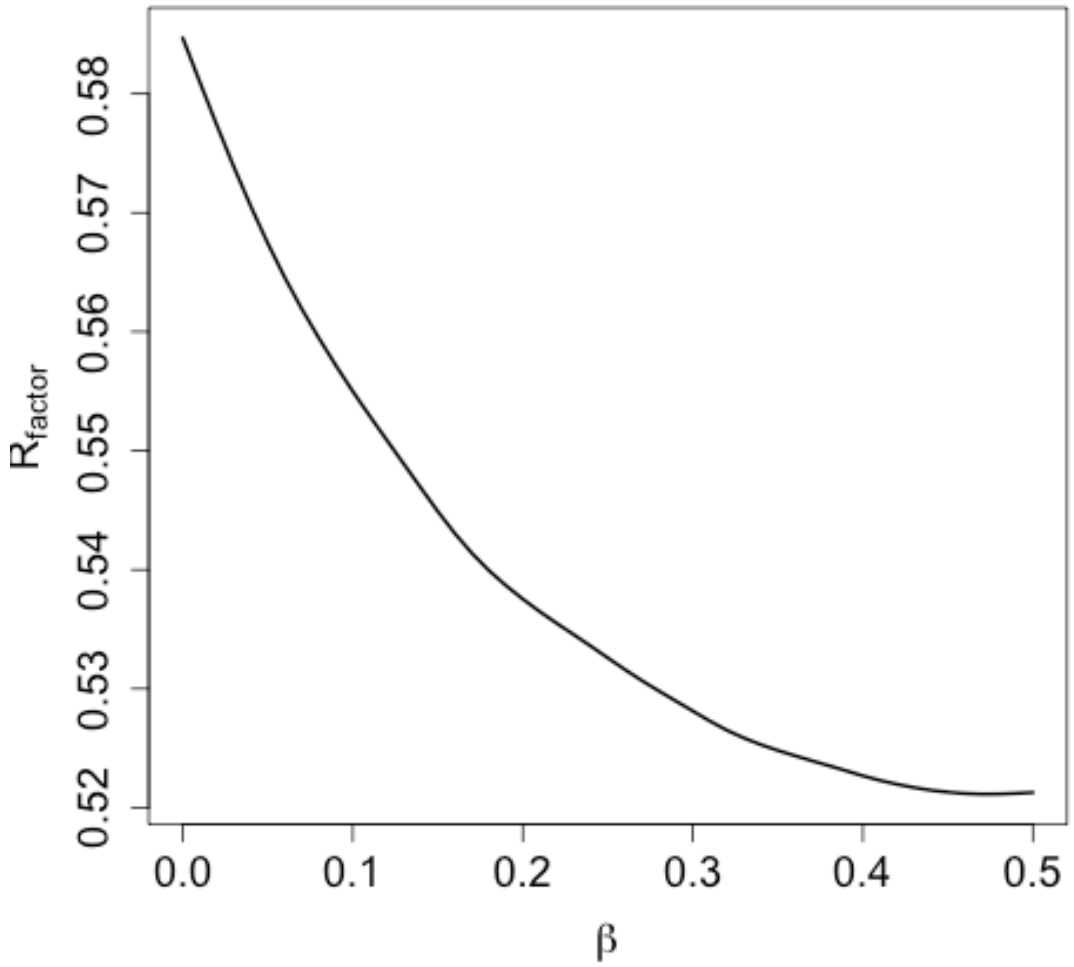
Twin: Few warnings about R values

Rvalues for random structures (no other peculiarities)

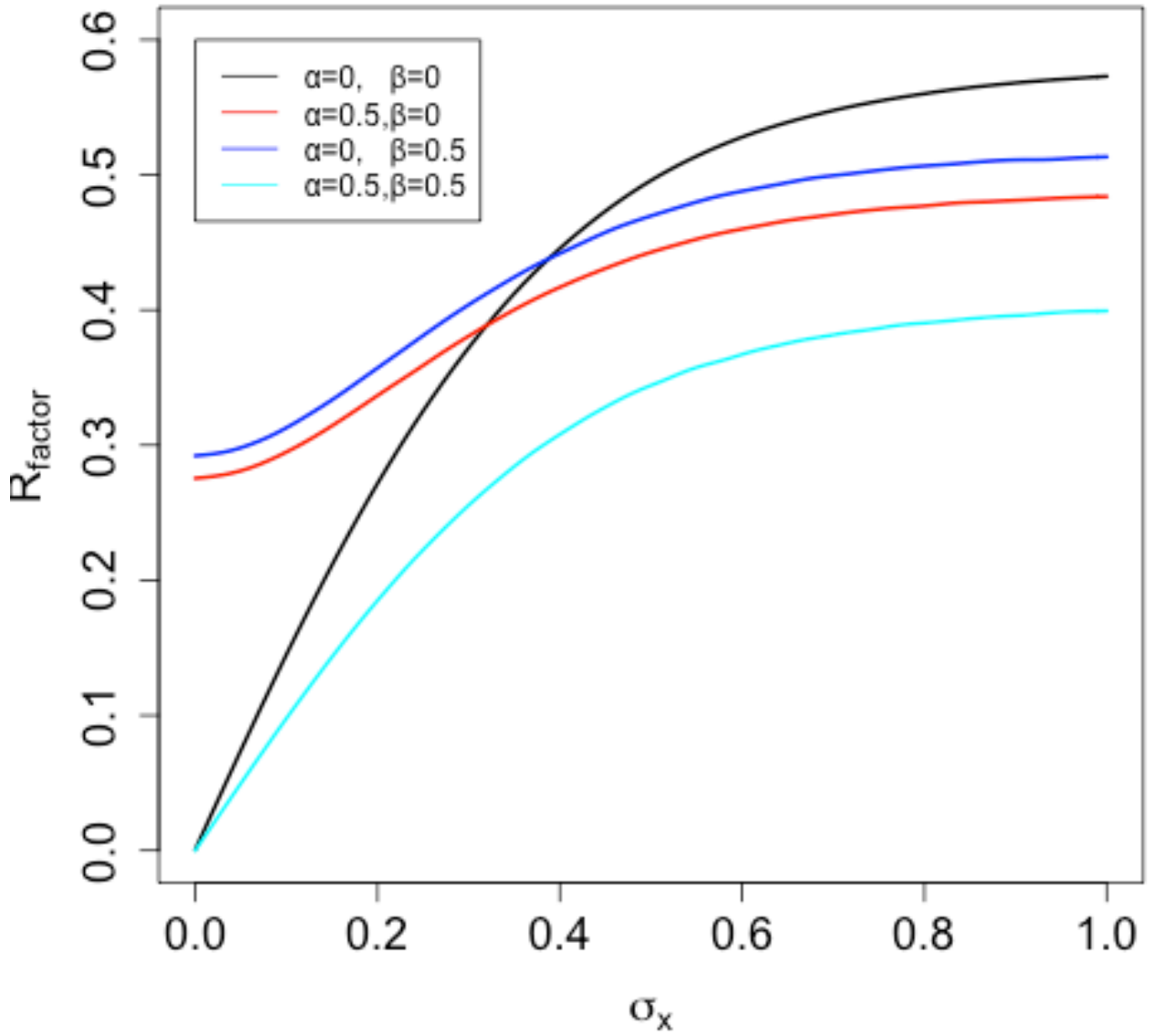
Twin	Modeled	Not modeled
Yes	0.41	0.49
No	0.52	0.58

Murshudov GN “Some properties of Crystallographic Reliability index – Rfactor: Effect of Twinning” Applied and Computational Mathematics”, 2011:10;250-261

Twin is not present, random structure: Rvalues vs “twin fraction”



Rvalue for structures with different model errors:
Combination of real and modeled perfect twin fractions



Low resolutions refinement tools

Use of available knowledge

1) NCS local
2) Restraints to known structure(s)
3) Restraints to current inter-atomic distances (implicit normal modes or “jelly” body)
4) Better restraints on B values

These are available from the version 5.6

Note
Buster/TNT has local NCS and restraints to known structures
CNS has restraints to known structures (they call it deformable elastic network)
Phenix has B-value restraints on non-bonded atom pairs and automatic global NCS
Local NCS (only for torsion angle related atom pairs) was available in SHELXL since the beginning of time

Auto NCS: local and global

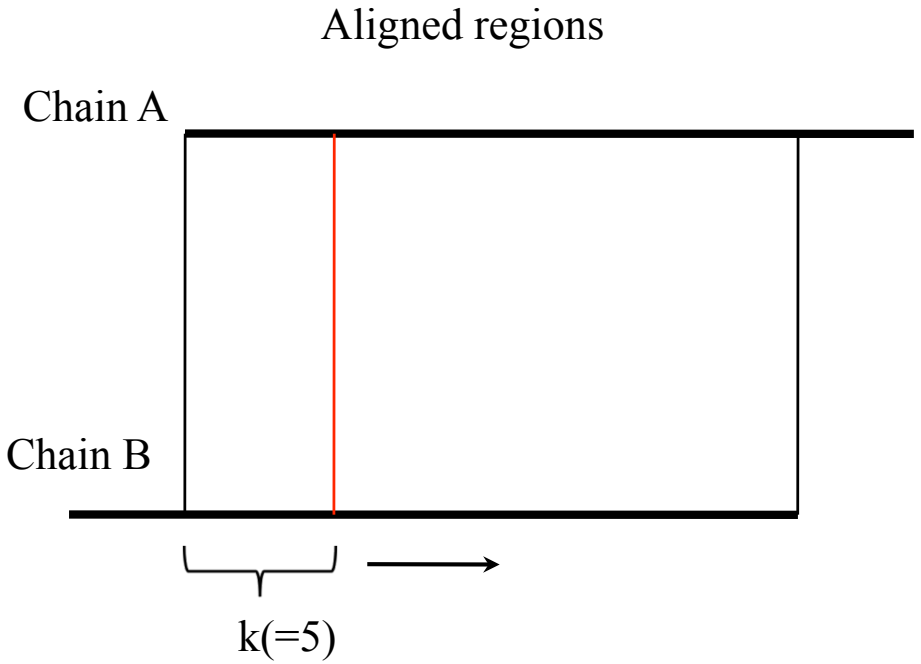
1. Align all chains with all chains using Needleman-Wunsh method
2. If alignment score is higher than predefined (e.g.80%) value then consider them as similar
3.Find local RMS and if average local RMS is less than predefined value then consider them aligned
4. Find correspondence between atoms
5. If global restraints (i.e. restraints based on RMS between atoms of aligned chains) then identify domains
6.For local NCS make the list of corresponding interatomic distances (remove bond and angle related atom pairs)
7.Design weights

The list of interatomic distance pairs is calculated at every cycle

Auto NCS

Global RMS is calculated using all aligned atoms.

Local RMS is calculated using k (default is 5) residue sliding windows and then averaging of the results



$$Ave(RmsLoc)_k = \frac{1}{N-k+1}\sum_{i=1}^{N-k+1} RmsLoc_i$$

$$RMS = Ave(RmsLoc)_N$$

Auto NCS: Iterative alignment

Example of alignment: 2vtu.
There are two chains similar to each other. There appears to be gene duplication

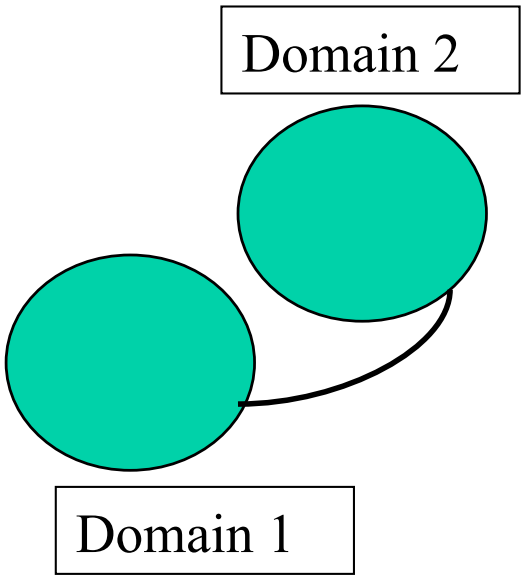
RMS – all aligned atoms
Ave(RmsLoc) – local RMS

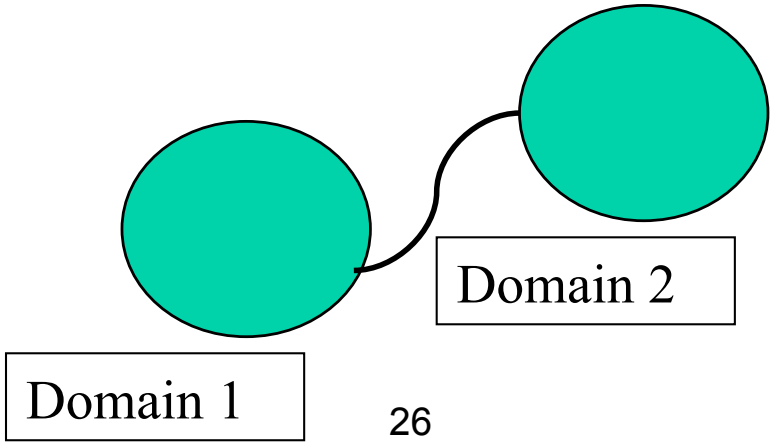
********* Alignment results *********

N:	Chain 1 :	Chain 2 :	No of aligned :	Score :	RMS :	Ave(RmsLoc):
1	J(131 - 256)	J(3 - 128)	126	1.0000	5.2409	1.6608
2	J(1 - 257)	L(1 - 257)	257	1.0000	4.8200	1.6694
3	J(131 - 256)	L(3 - 128)	126	1.0000	5.2092	1.6820
4	J(3 - 128)	L(131 - 256)	126	1.0000	3.0316	1.5414
5	L(131 - 256)	L(3 - 128)	126	1.0000	0.4515	0.0464

Auto NCS: Conformational changes

In many cases it could be expected that two or more copies of the same molecule will have (slightly) different conformation. For example if there is a domain movement then internal structures of domains will be same but between domains distances will be different in two copies of a molecule





External (reference) structure restraints

Restraints to external structures are generated by the program ProSmart:
1) Aligns structure in the presence of conformational changes. Sequence is not used
2) Generates restraints for aligned atoms
3) Identifies secondary structures (at the moment helix and strand, but the approach is general and can be extended to any motif).
4) Generates restraints for secondary structures

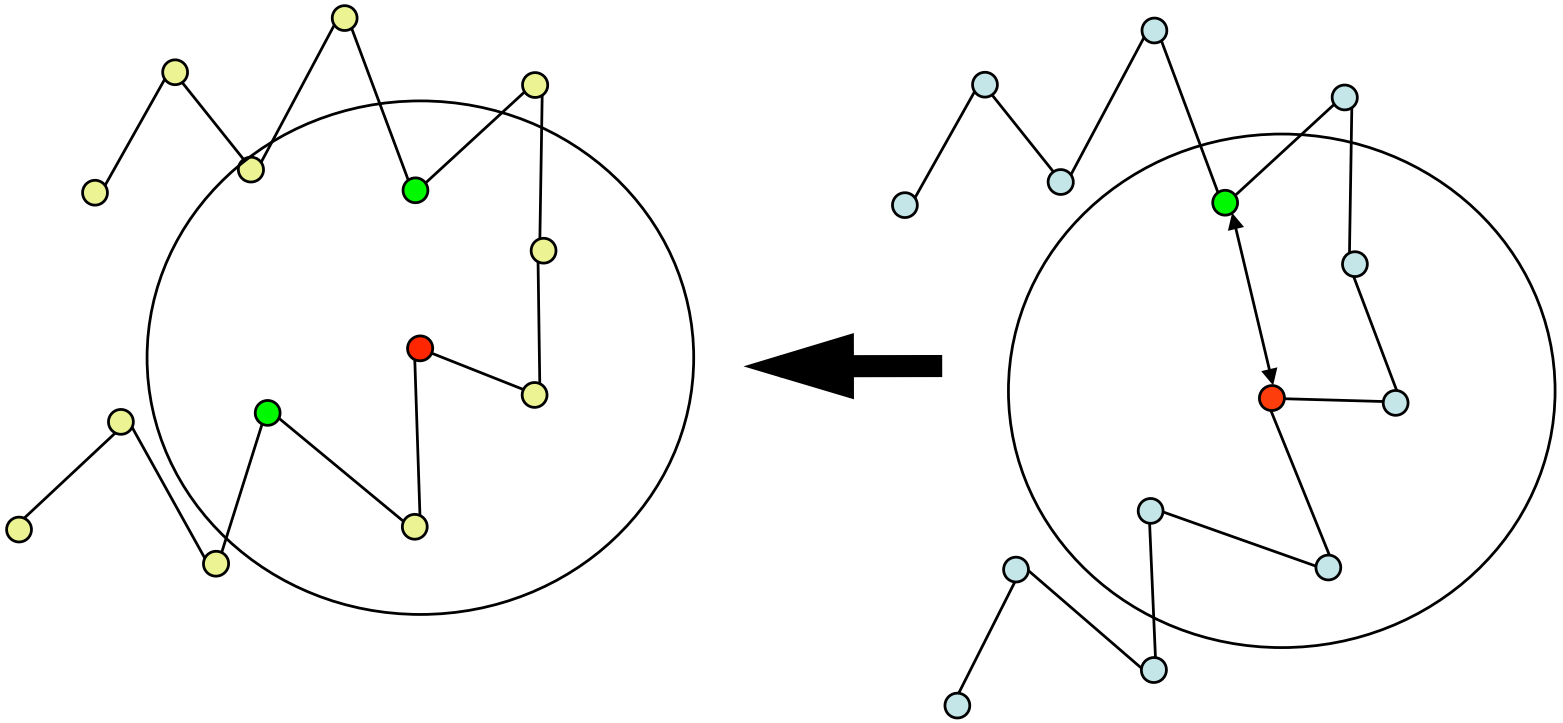
Note 1: ProSmart has been written by Rob Nicholls and available from him (now). It will be distributed by ccp4 (hopefully from the next release)

Note 2: Robust estimator functions are used for restraints. I.e. if differences between target and model is very large then their contributions are downweighted

Reference Structure Restrain

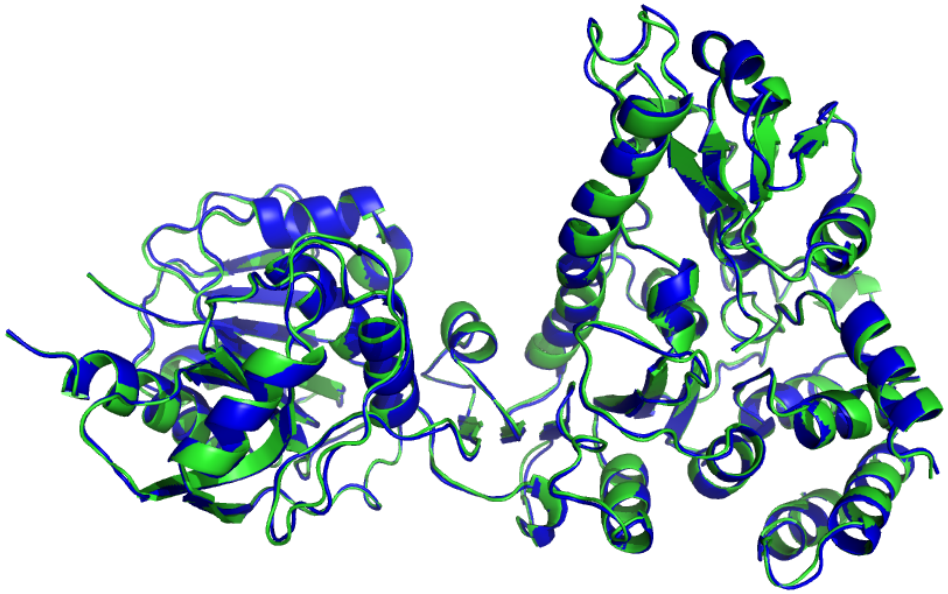
structure to be refined

known similar structure (prior)

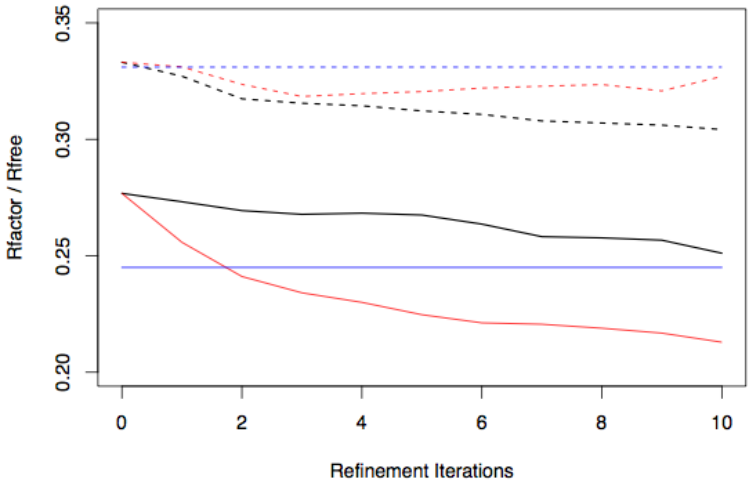


Remove bond and angle related pairs

Automated re-refinement of 3.4Å
External structure 2.5Å.
When using external restraints, both R and R_{free} decrease



Green- reference structure
Blue - target structure



Blue – original statistics
Red - without external structure
Black – with external structure

Restraints to current distances

The term is added to the target function:

$$\sum_{pairs} w(|d| - |d_{current}|)^2$$

Summation is over all pairs in the same chain and within given distance (default 4.2A). $d_{current}$ is recalculated at every cycle. This function does not contribute to gradients. It only contributes to the second derivative matrix.

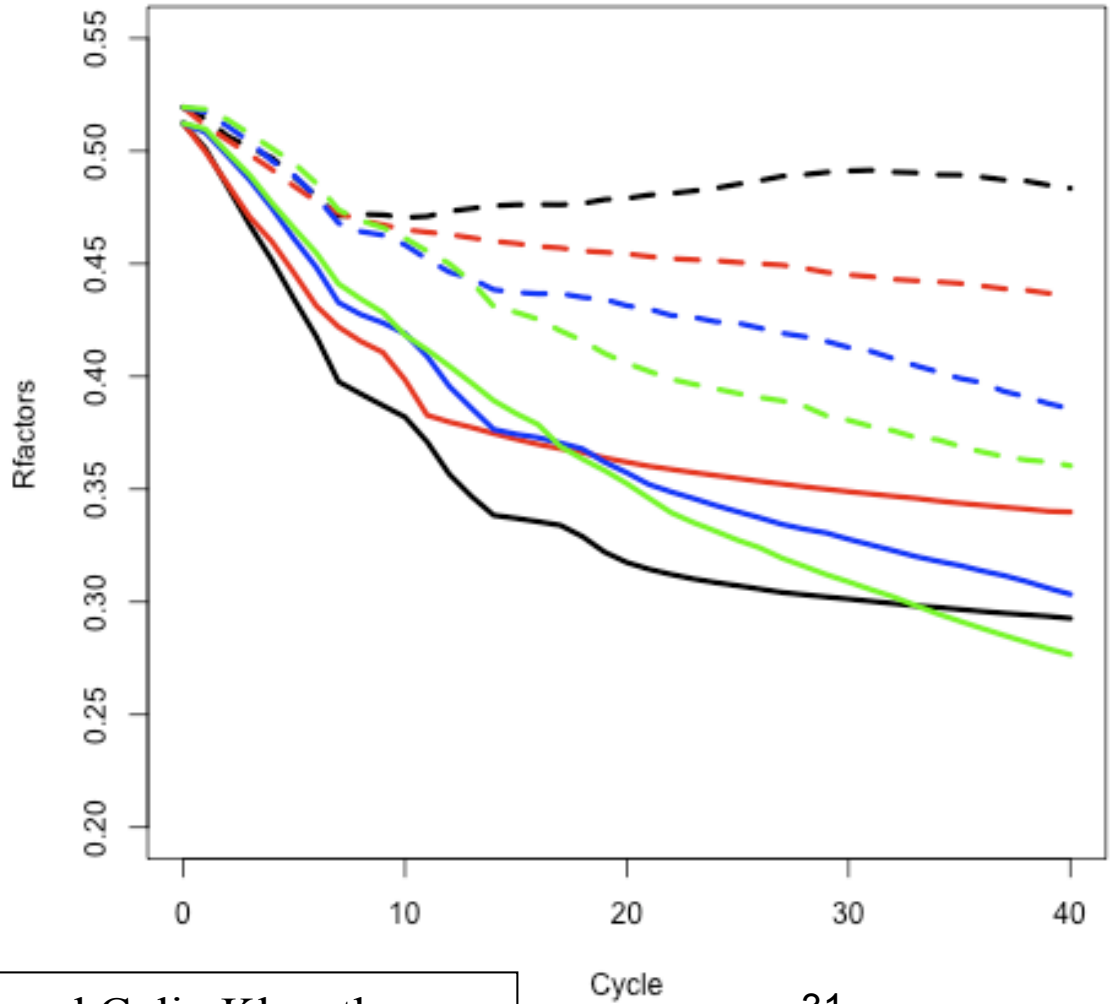
It is equivalent to adding springs between atom pairs. During refinement inter-atomic distances are not changed very much. If all pairs would be used and weights would be very large then it would be equivalent to rigid body refinement.

It could be called “implicit normal modes”, “soft” body or “jelly” body refinement.

Example, after molecular replacement
3A resolution, data completeness 71%

Rfactors vs cycle
Black – simple refinement
Red – Global NCS
Blue – Local NCS
Green – "Jelly" body

Solid lines – Rfactor
Dashed lines - Rfree



Data provided by: Marek Brzozovski and Colin Kleanthous

Example: 4A resolution, data from pdb 2r6c

Starting R/Rfree = 36.0/35.6

R/Rfree after 40 cycles of refinement

	None	Ncs local	Jelly body	External structure
R	20.80	21.44	23.72	23.38
Rfree	32.69	31.79	29.82	28.69

Map Sharpening

MAP SHARPENING: INVERSE PROBLEM

Newer refmac has an option to sharpen map after refinement. Usual map sharpening applies negative B value to the structure factors thus increasing effect of high resolution terms. But it may cause problem: effects of noise and series termination may become larger.

We have implemented a regularized map sharpening that increases effect of high resolution terms while reducing effects of noises

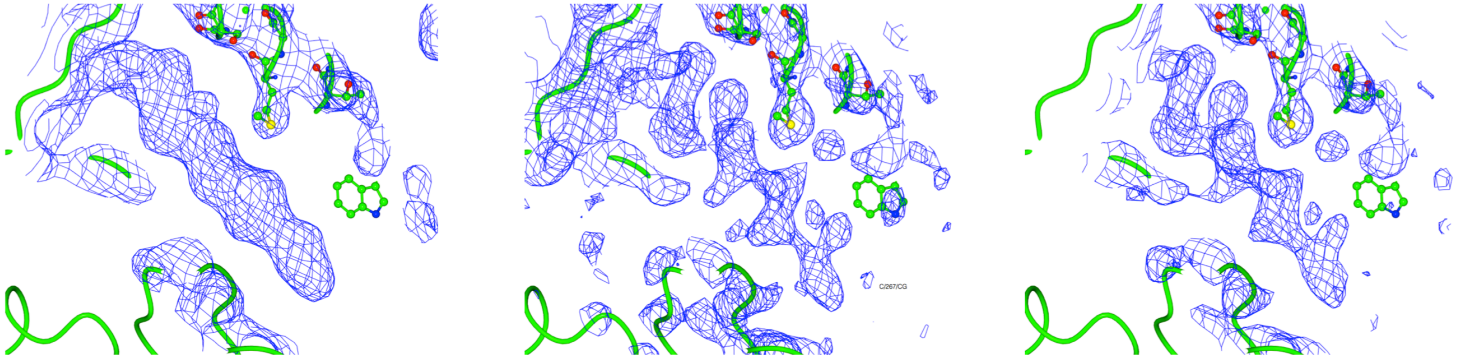
Vogel,CR, Computational Methods for Inverse problems

Example, 2r6c, Electron density

No sharpening

Sharpening: no regularisation

Sharpening: with regularisation

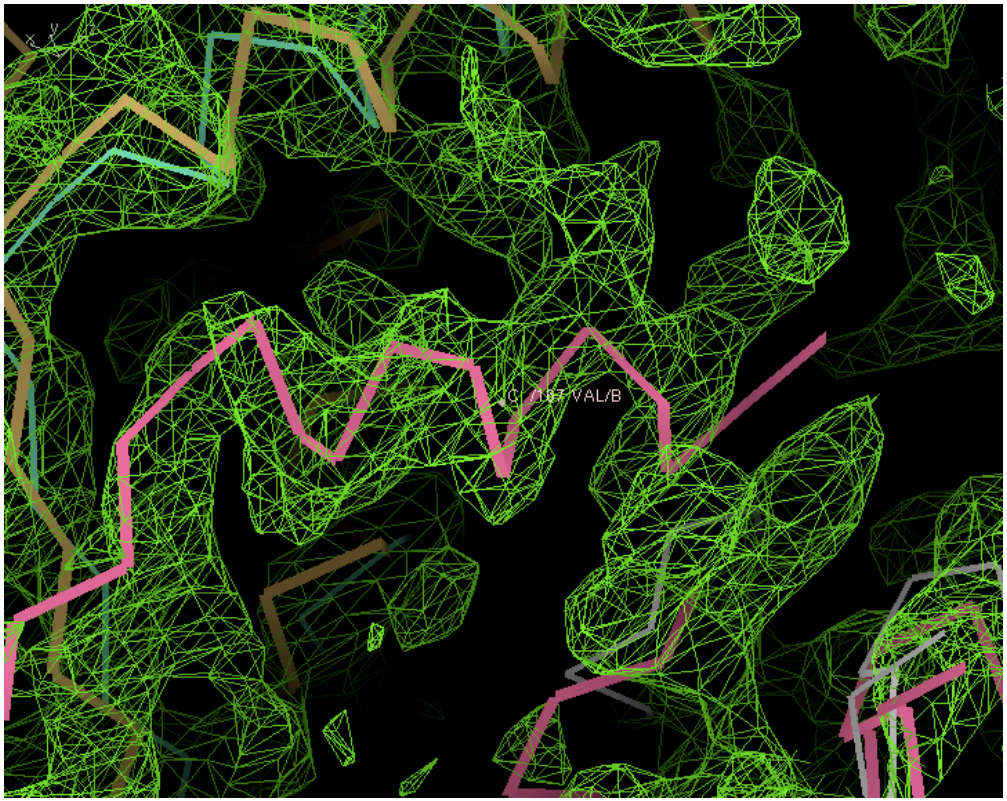


a) b) c)

For sharpening B value median of atomic B values is used.
Anisotropy is also used

Example, 2r6c, Electron density

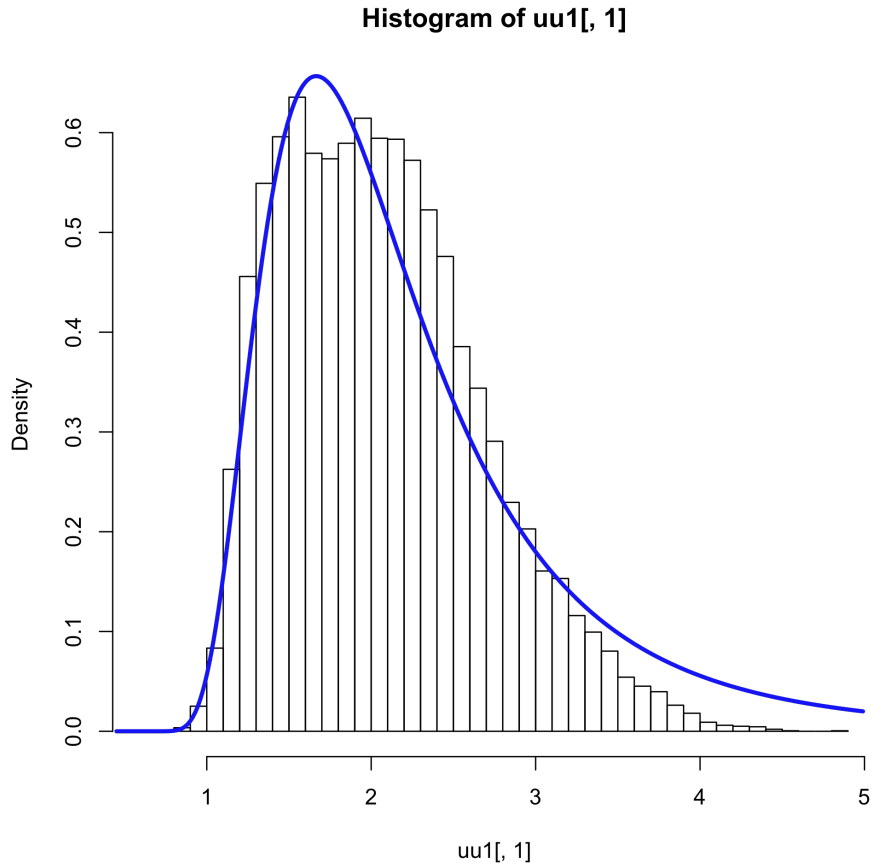
Known structure (2r6a) superimposed to 2r6c structure. There is a helix. Side chains are visible to some degree



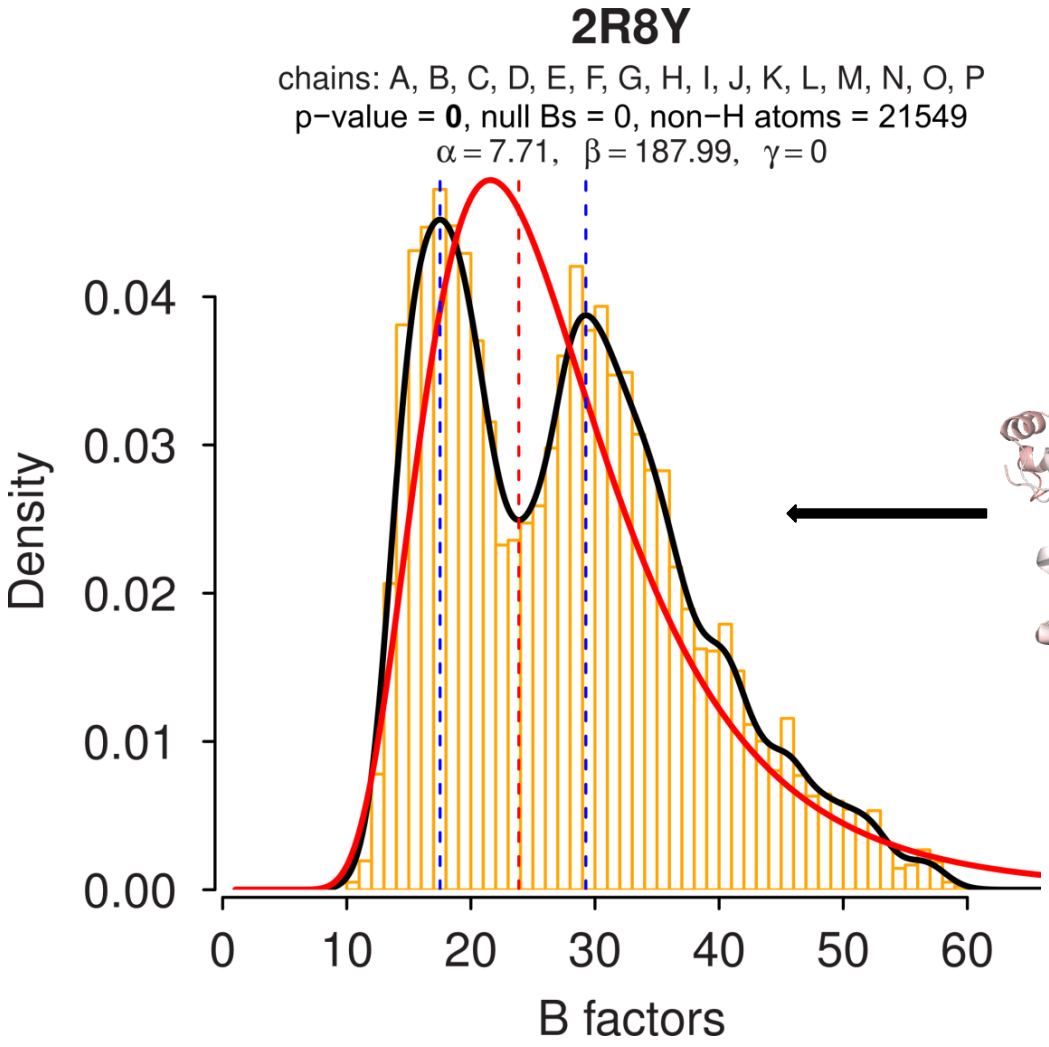
Effect of B value distribution

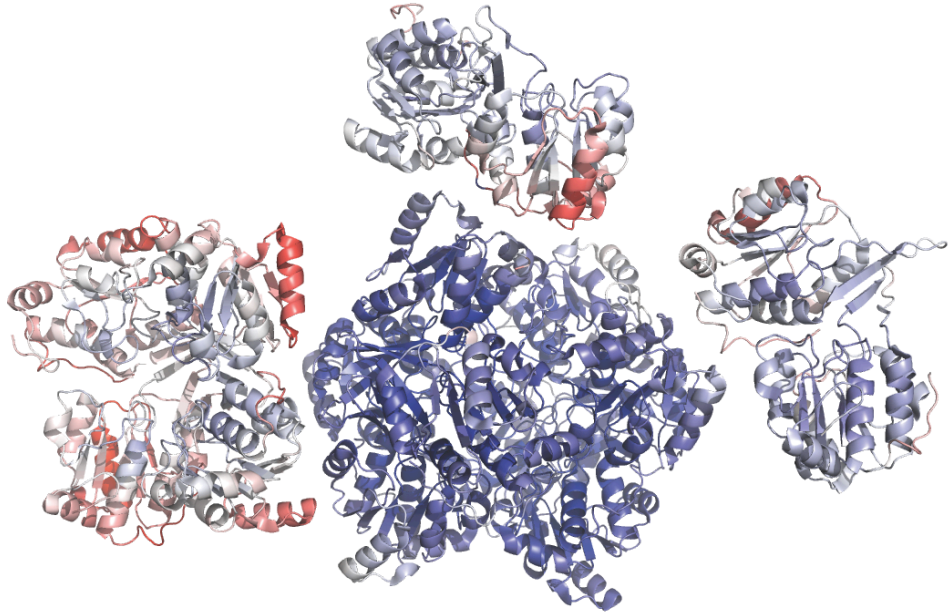
4Å resolution data.
Histogram: empirical distribution of B values
Blue line: Shifted inverse gamma distribution

$$P(B) = \frac{\beta^{\alpha}(B - B_0)^{-\alpha-1}}{\Gamma(\alpha)} e^{-\beta/(B-B_0)}$$



Multimodality at chain level (1)





PDB id 2R8Y; haloacid dehalogenase superfamily phosphatase structure at 1.85Å; R_{work}=0.215, R_{free}=0.244

Conclusion

- Twin refinement improves statistics and occasionally electron density: Rfactors may be misleading
- Use of known structures improves reliability of the derived model: Especially at low resolution
- NCS restraints must be done automatically: but conformational flexibility must be accounted for
- “Jelly” body works better than I thought it should
- Regularised map sharpening looks promising. More work should be done on series termination and general sharpening operators

Future work

- Release reticular twinning, multiple cells, modulations
- Refinement in the presence of radiation damage
- Local TLS
- Bayesian sharpening and denoising for map calculation
- Multicrystal refinement
- More restraints for RNA/DNA and carbohydrates
- Etc

Acknowledgment

Cambridge/York

Alexei Vagin

Andrey Lebedev

Rob Nicholls

Fei Long

Others

Pavol Skubak

Raj Pannu

Jacopo Negroni

CCP4, YSBL people

REFMAC is available from CCP4 or from York's ftp site:

www.ysbl.york.ac.uk/refmac/latest_refmac.html

This and other presentations can be found on:

www.ysbl.york.ac.uk/refmac/Presentations/