Refinement

Garib N Murshudov

MRC-LMB

Cambridge

Contents

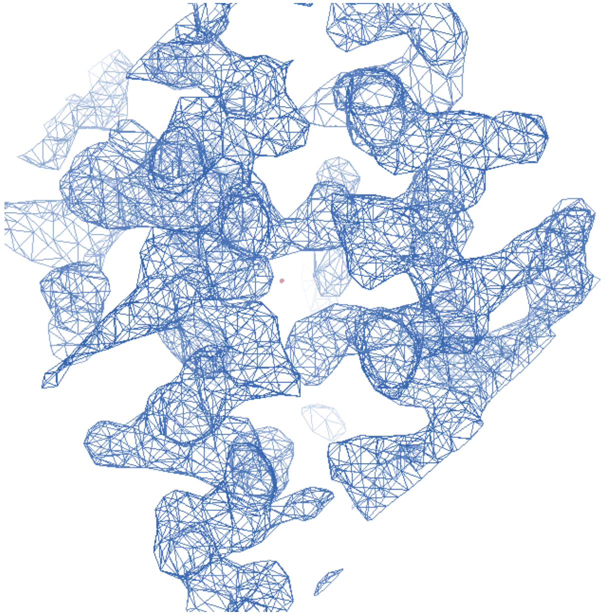
1) Purpose of refinement
2) Data and model
3) Model: what do we know about macromolecules
4) Bulk solvent modelling
5) Crystallographic data and their property
6) Model parameters

Purpose of refinement

Crystallographic refinement has two purposes:

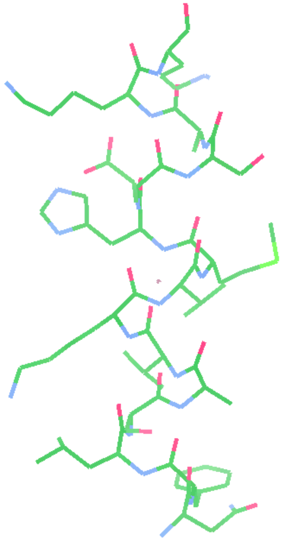
1) **Fit chemically and structurally sensible atomic model into observed– X-ray crystallographic data**

2) **To calculate best possible electron density so that atom model can criticised**



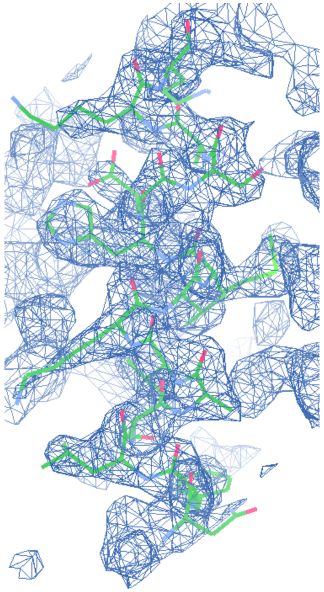
Data

+



Atomic model

=



Fit and refine

We want to fit currently available model to the data and calculate differences between them.

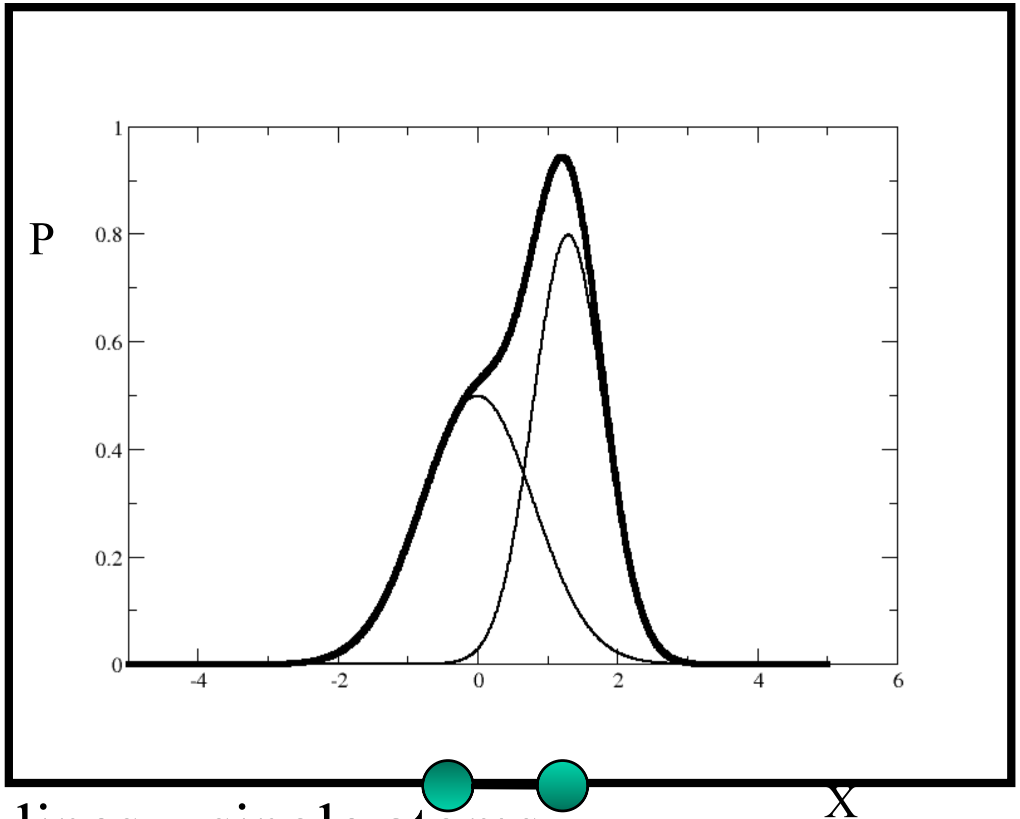
To do this fit properly we must use as much as possible information about model and data.

What do we know about macromolecules?

1) Macromolecular consists of atoms that are bonded to each other in a specific way
2) If there are two molecules with sufficiently high sequence identity then it is likely they will be similar to each other in 3D
3) It is highly likely that if there are two copies of a the same molecule they will be similar to each other (at least locally)
4) Oscillation of atoms close to each other in 3D cannot be dramatically different
5) Proteins tend to form secondary structures
6) DNA/RNA tend to form basepairs

Why restraints:
Two atoms ideal case

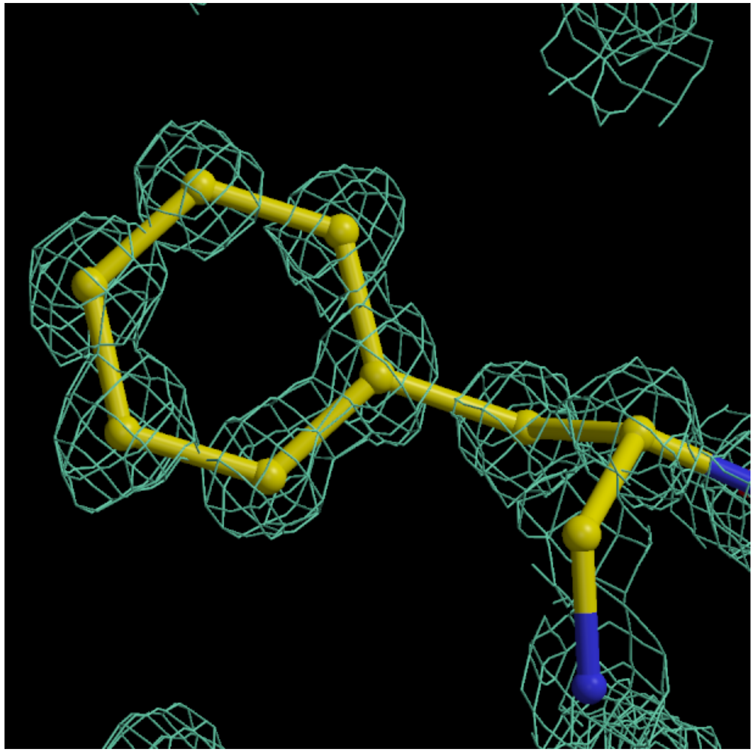
- Distance between atoms 1.3Å. B values 20 and 50



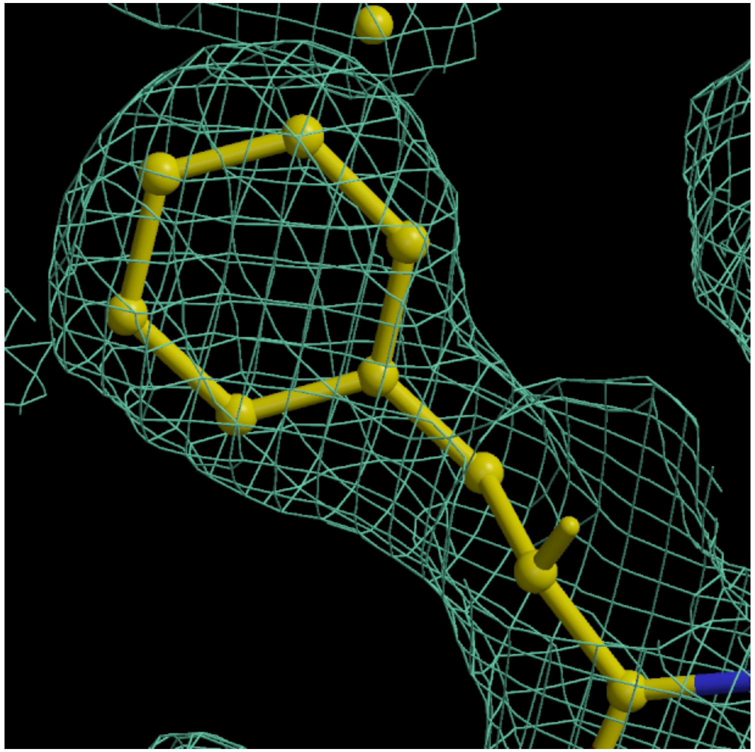
- Thin lines – single atoms
- Bold line - sum of the two atoms

Chemical information: Phe at two different resolutions

- 0.88 Å

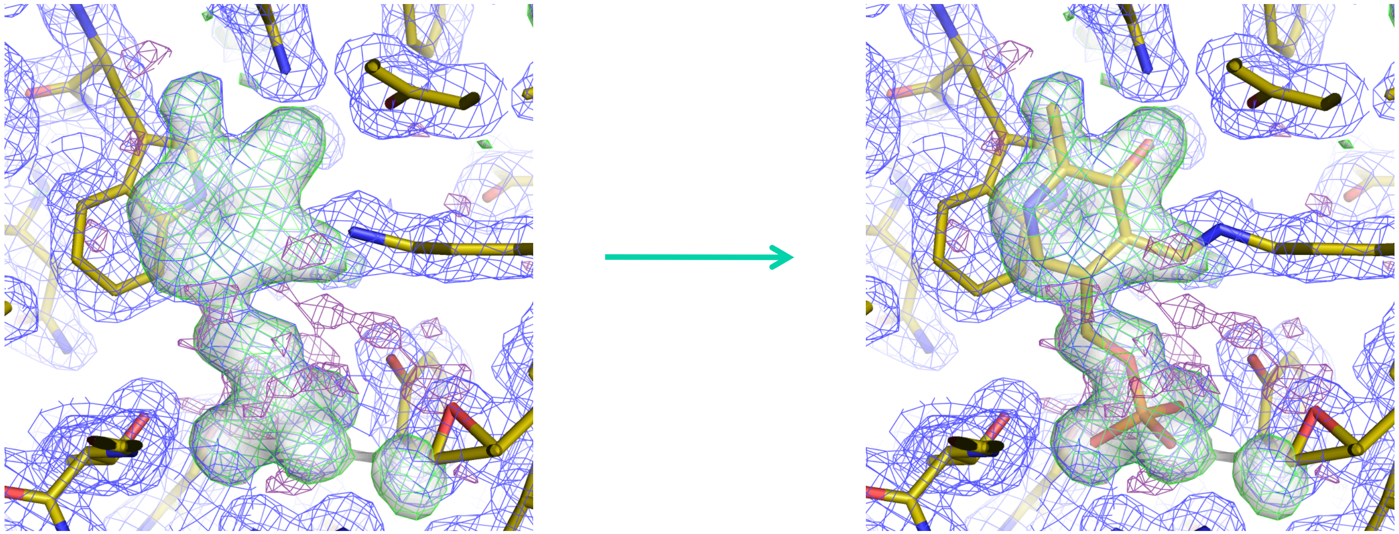


2 Å and High mobility



Basic chemical information

We know that our crystal contains pyridoxal phosphate and it makes covalent bond to Lys



There is a tutorial how to deal with these cases on:
http://www.ysbl.york.ac.uk/mxstat/JLigand/

NCS

Three ways of dealing with NCS

1) NCS constraints: copies of molecules are considered to be exactly same. Only one set of atomic parameters per molecule is refined, other copies are kept to be exactly same
2) NCS restraints: Molecules are superimposed and difference between corresponding atoms after superposition minimised.
3) NCS local restraints: Molecules are assumed to be locally similar, globally they may be different

Auto NCS: local and global

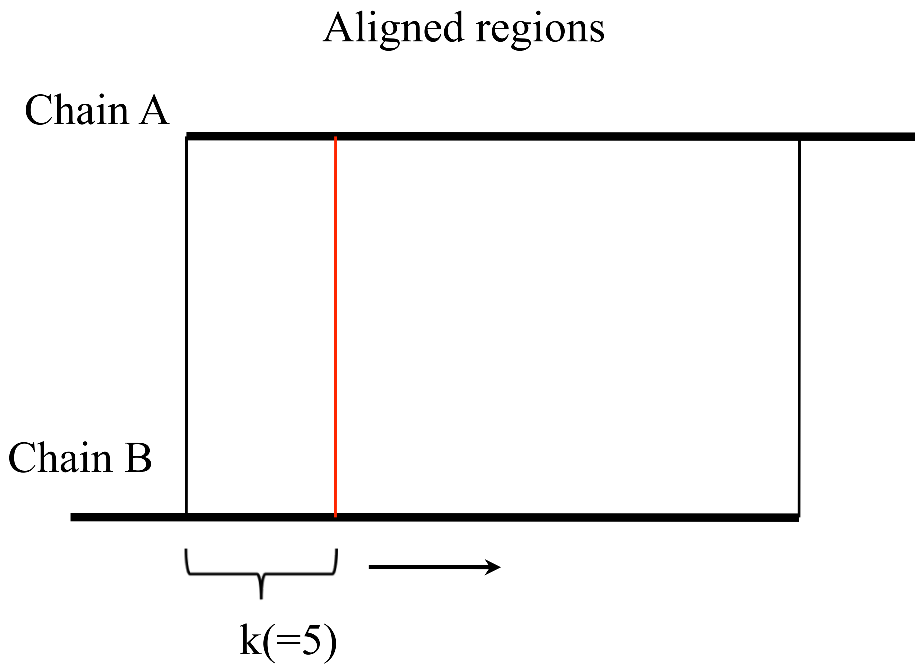
1. Align all chains with all chains using Needleman-Wunsh method
2. If alignment score is higher than predefined (e.g.80%) value then consider them as similar
3.Find local RMS and if average local RMS is less than predefined value then consider them aligned
4. Find correspondence between atoms
5. If 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

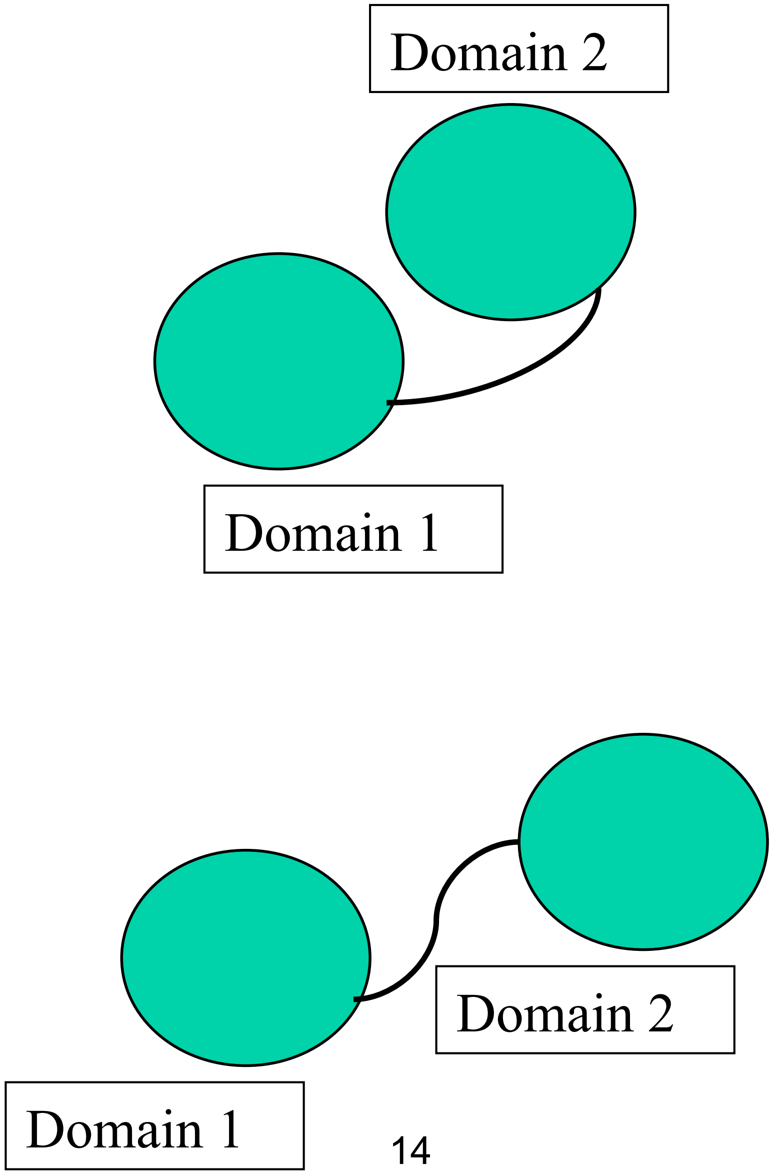
Example of alignment: 2vtu.
There are two chains similar to each other. There appears to be gene duplication

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

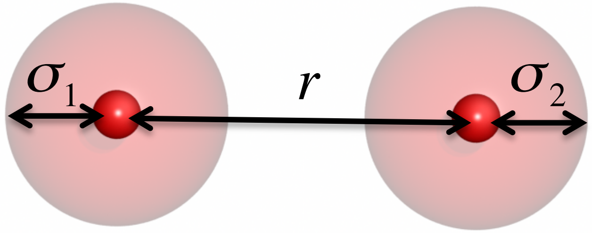
Note 1: ProSmart has been written by Rob Nicholls and available from him and CCP4.

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

External Restraints

Restraint Parameter Estimation:

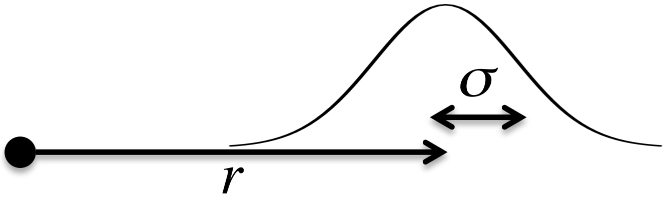
Atoms in homologous structure:



σ_i estimated as by Murshudov & Dodson (CCP4 Newsletter 1997)

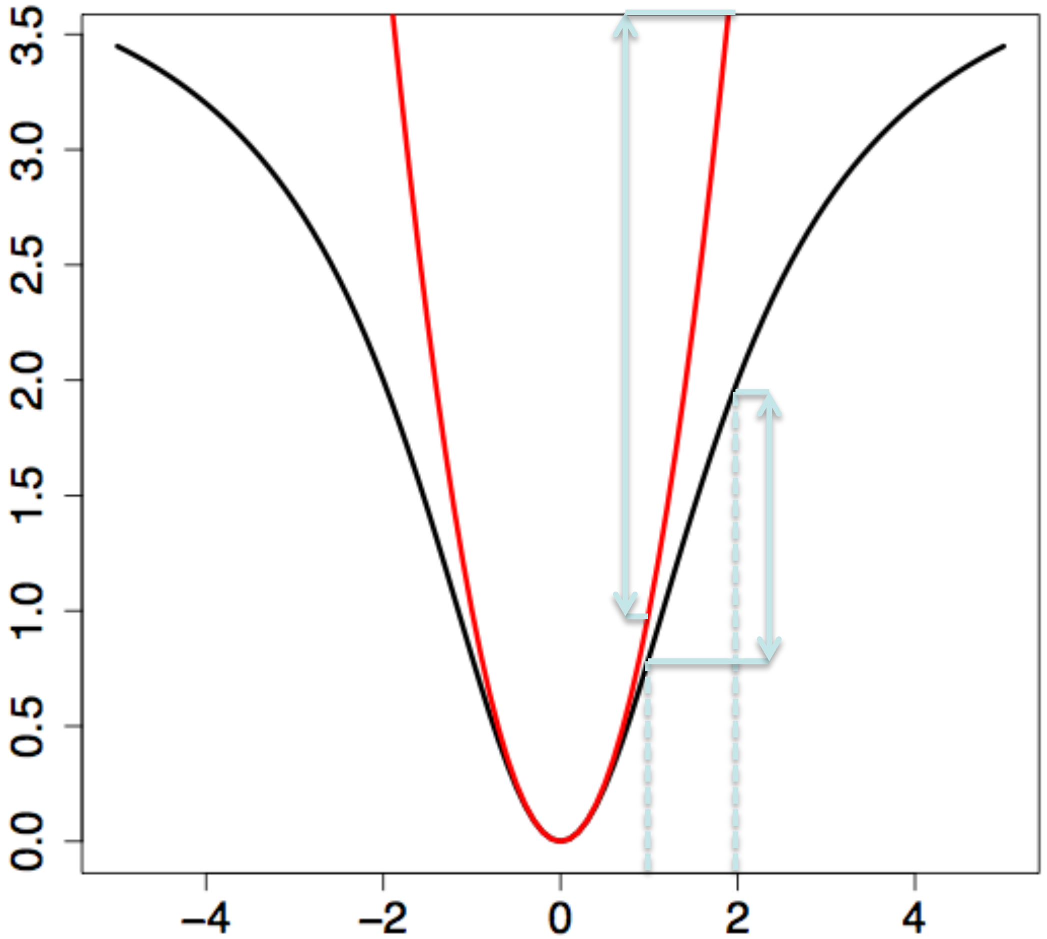
Dependencies:

- B-value
- Resolution
- Model/data completeness
- Data quality



Plus a modification that increases robustness to outliers (use of Geman-McClure function)

External Restraints



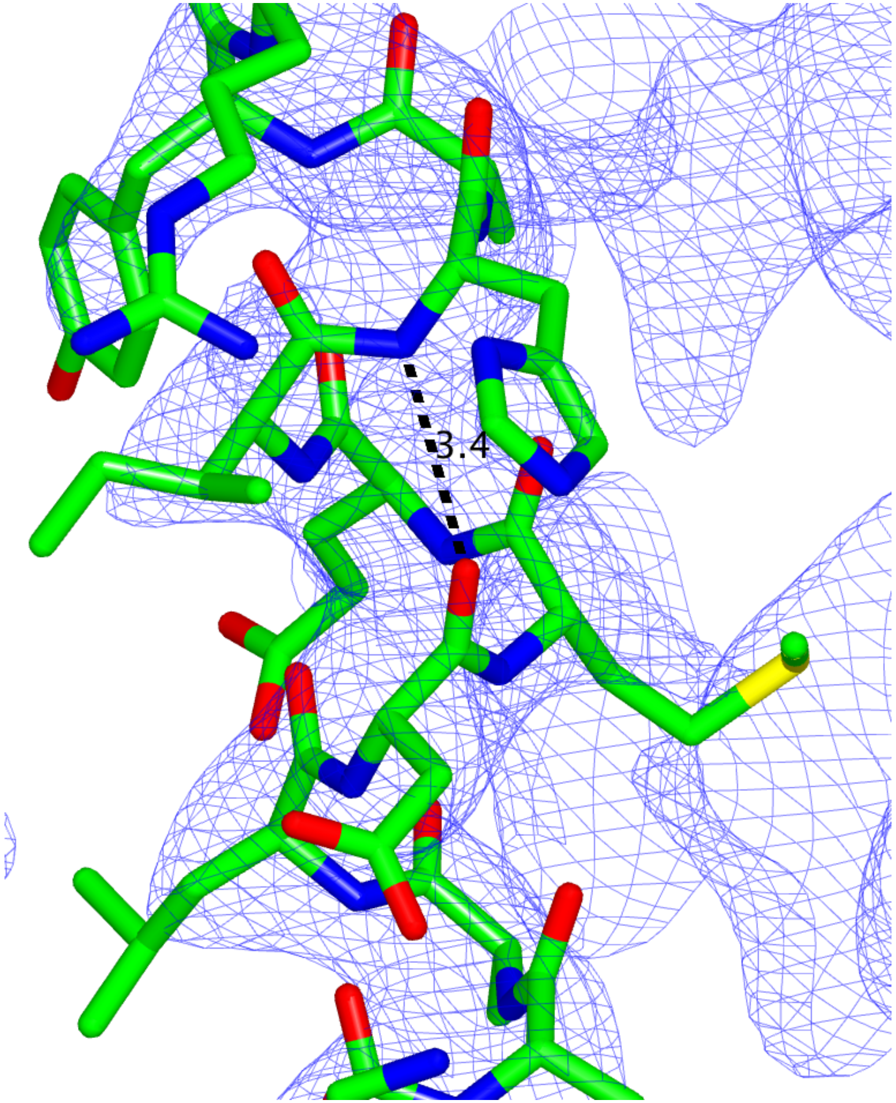
$$x = \frac{d - r}{\sigma}$$

Least squares: x^2

Geman–McClure: $\dfrac{x^2}{1 + wx^2}$

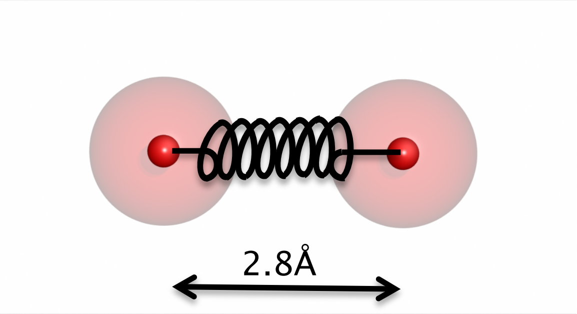
Also use external B-value restraints and Van-der-Waals repulsions in recent versions of REFMAC5.

External Restraints



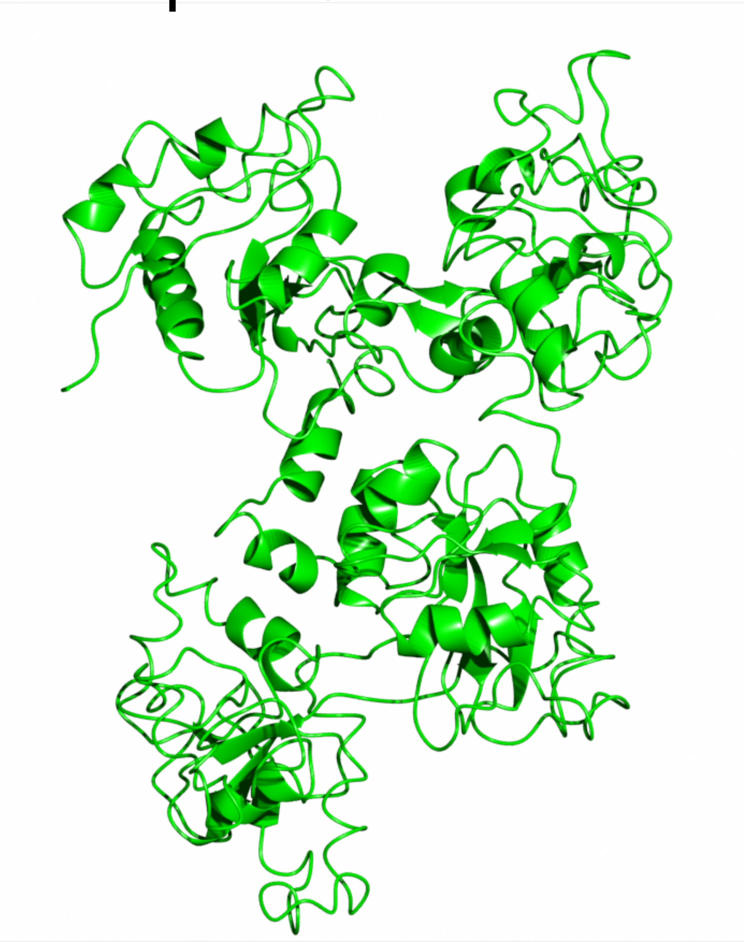
3g4w – 3.7Å

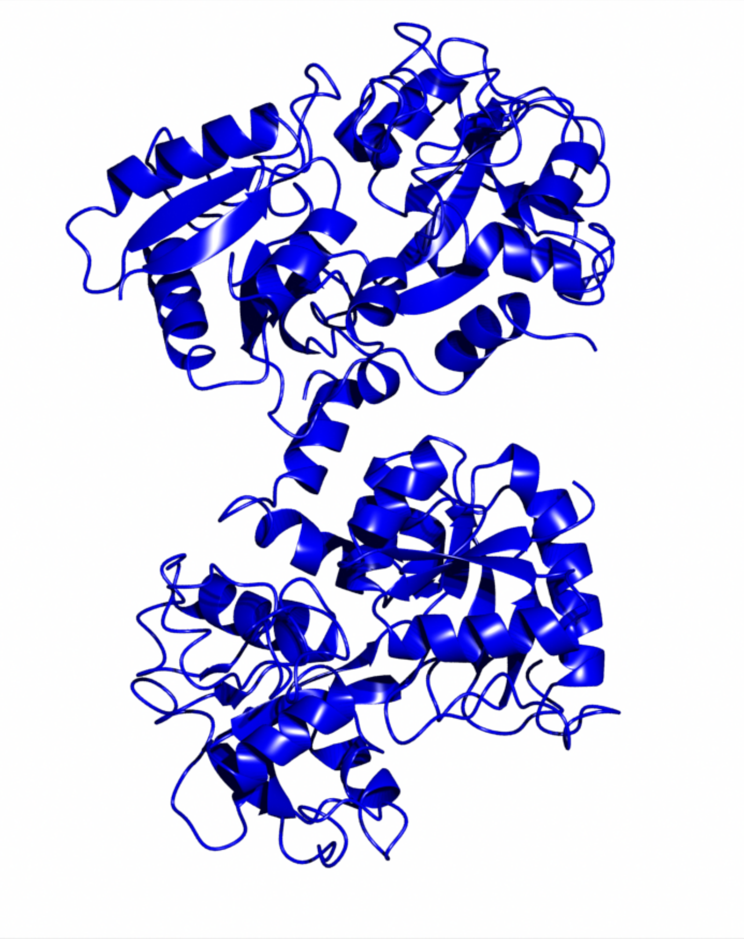
Prior information:



External Restraints

Example: Ovotransferrin



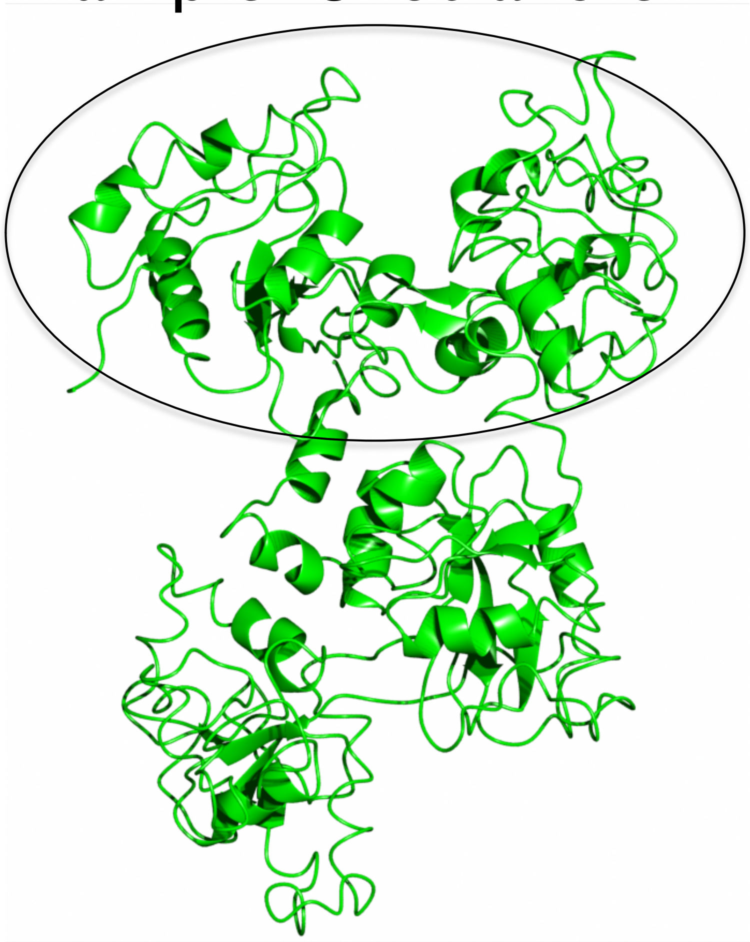


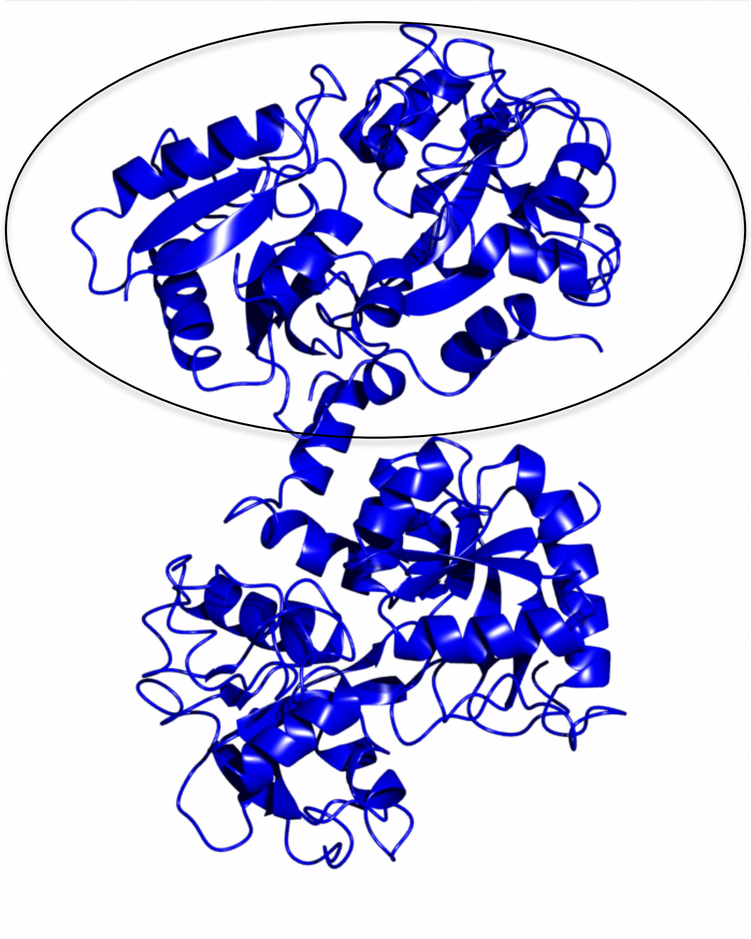
1ryx – 3.5Å

2d3i – 2.15Å

External Restraints

Example: Ovotransferrin



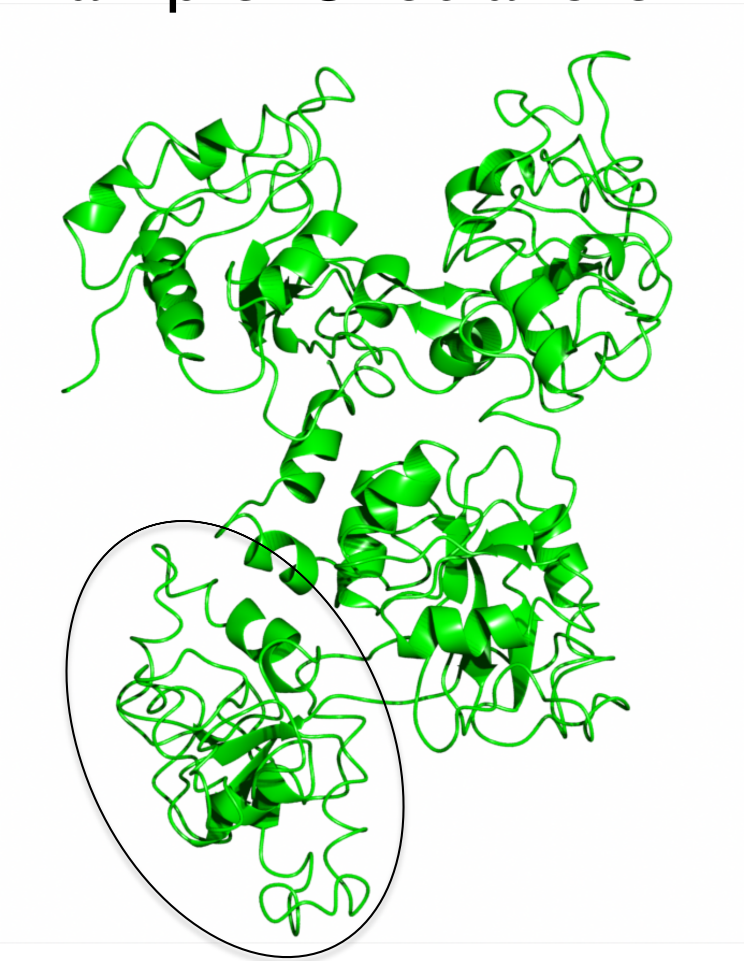


1ryx – 3.5Å

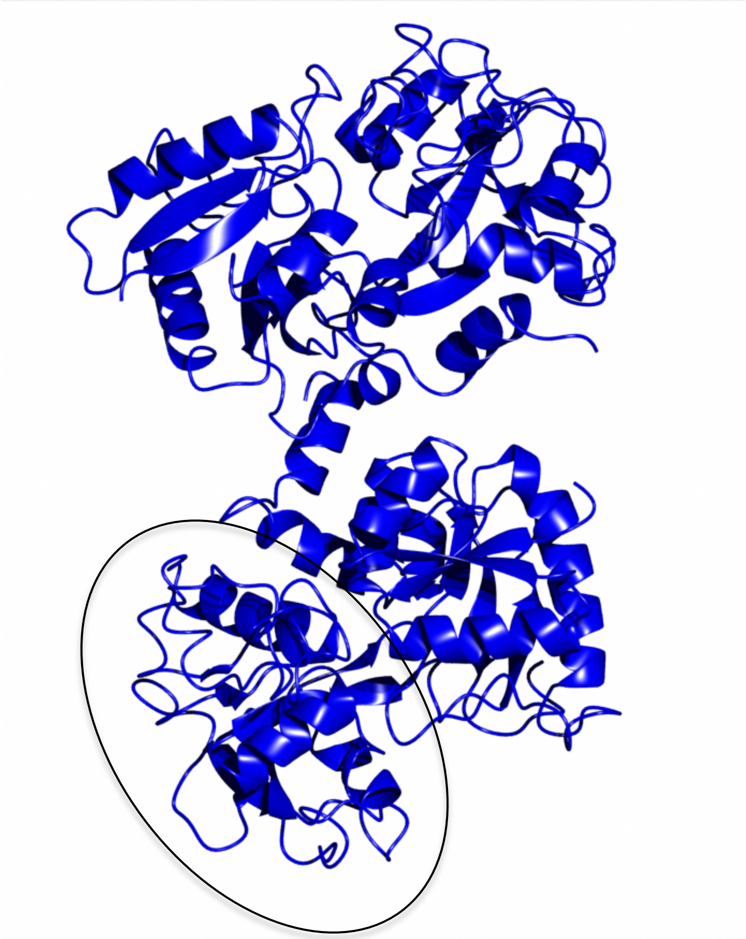
2d3i – 2.15Å

External Restraints

Example: Ovotransferrin



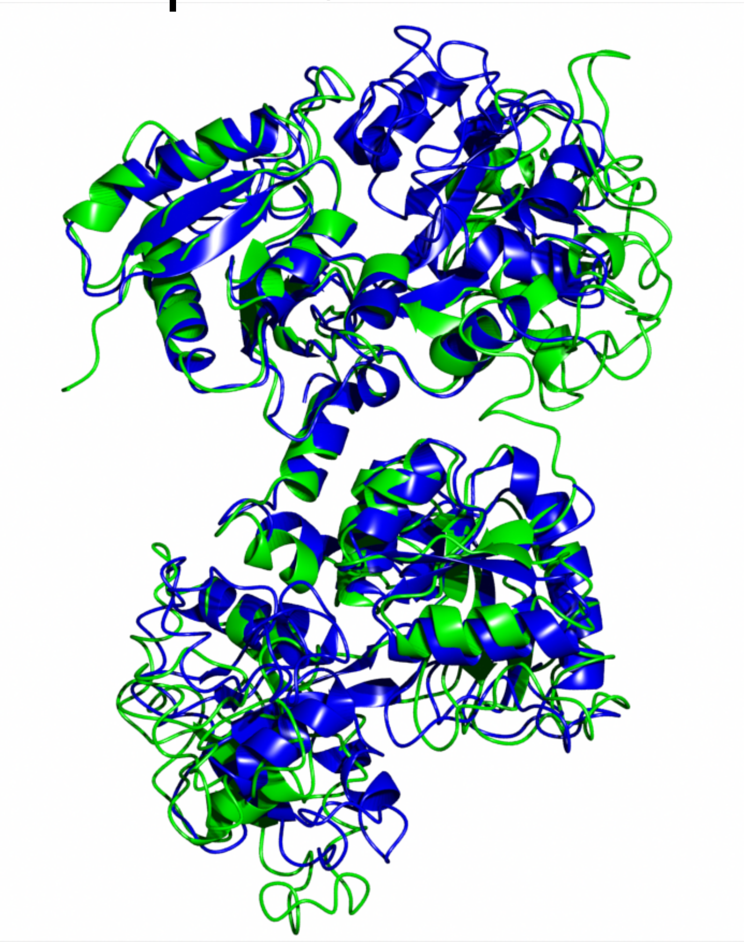
1ryx – 3.5Å



2d3i – 2.15Å

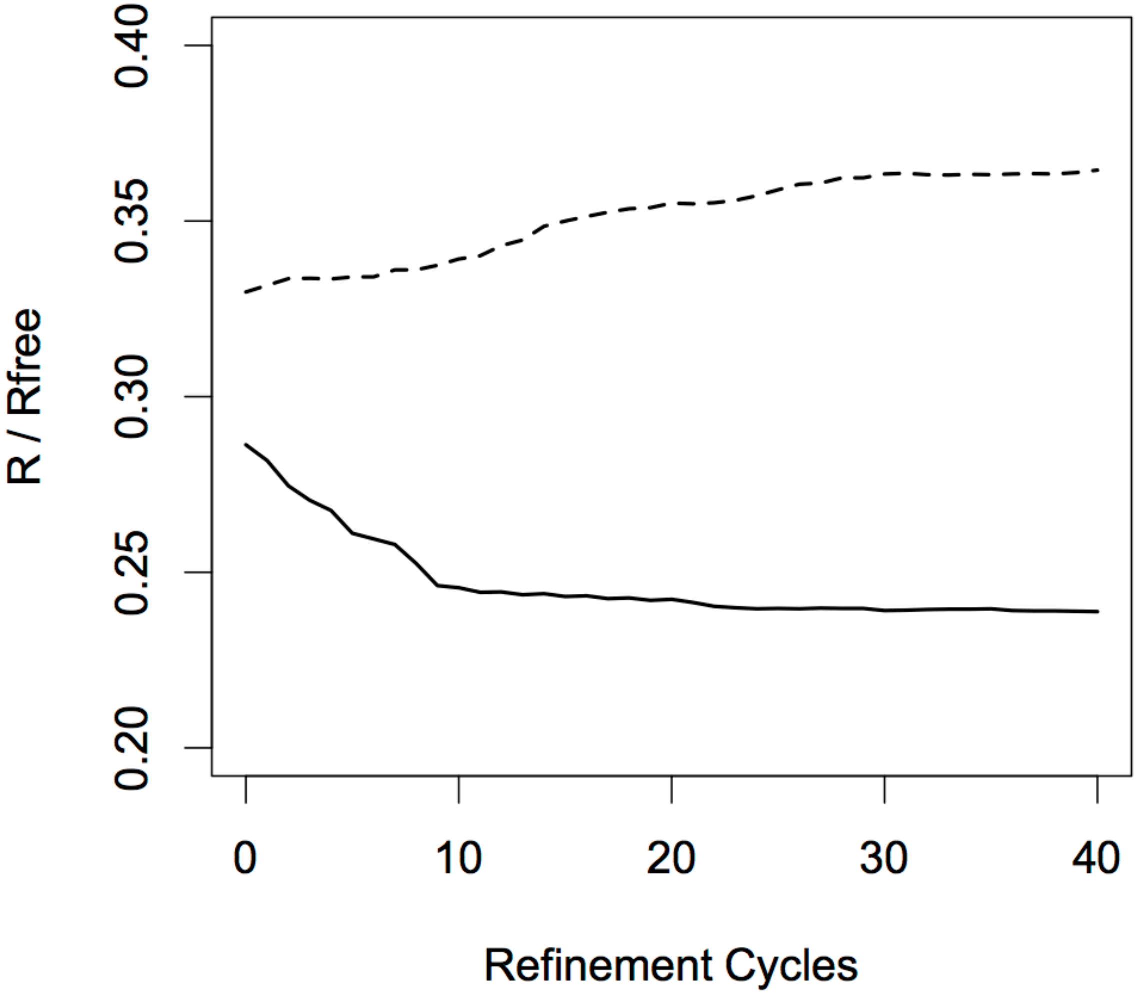
External Restraints

Example: Ovotransferrin



External Restraints

Example: Ovotransferrin



0.365

0.322

0.307

0.265

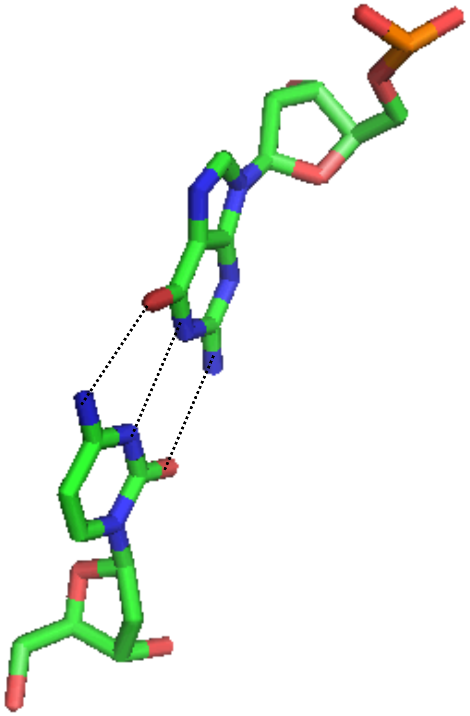
0.263

0.239

Basepair restraints

```
monomers C G   label C:G
 bond atom N4 C atom O6 G value 2.91 sigma 0.15
 bond atom N3 C atom N1 G value 2.95 sigma 0.1
 bond atom O2 C atom N2 G value 2.86 sigma 0.15
 chiral atom N1 G atom C6 G atom C2 G atom N3 C value 0.0 sigma 0.5
 chiral atom N3 C atom C4 C atom C2 C atom N1 G value 0.0 sigma 0.5
 torsion atom C4 C atom N3 C atom N1 G atom C2 G value 180.0 sigma 10.0
```

Basepair restraints should allow twists, buckles etc to occur



Restraints to current distances (jelly-body)

The term is added to the target function:

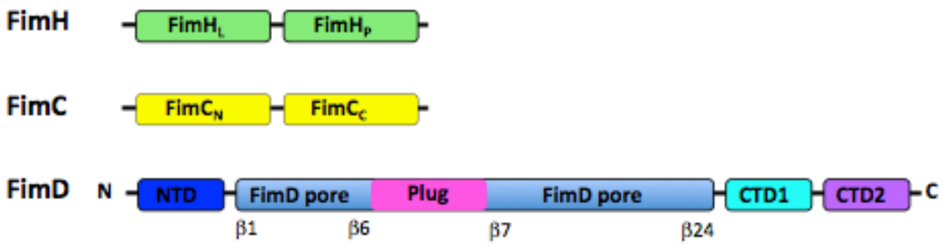
$$\sum_{pairs} w(|d| - |d_{current}|)^2$$

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.

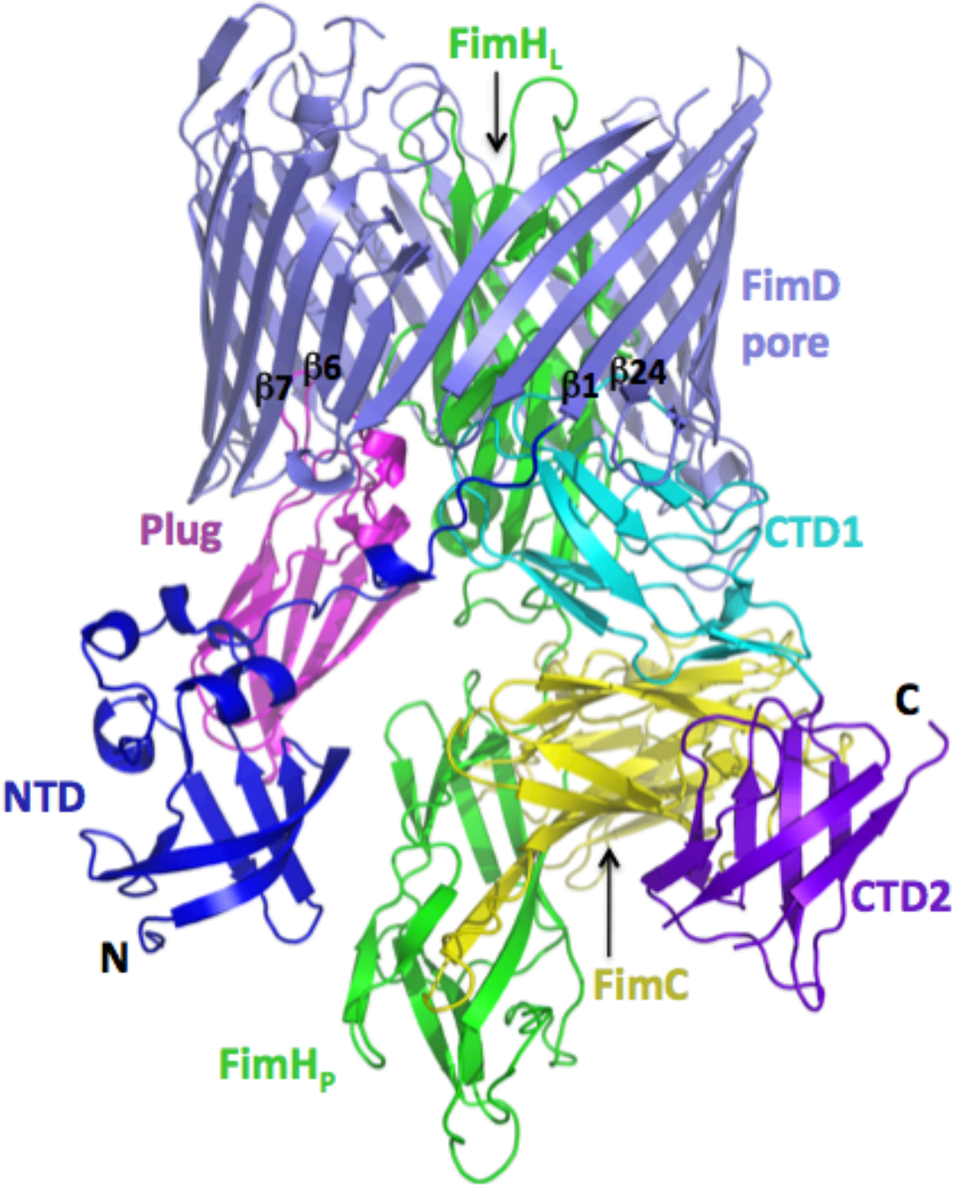
It could be called “implicit normal modes”, “soft” body or “jelly” body refinement.

Effect of "jelly" body refinement: Example is provided by A.Lebedev



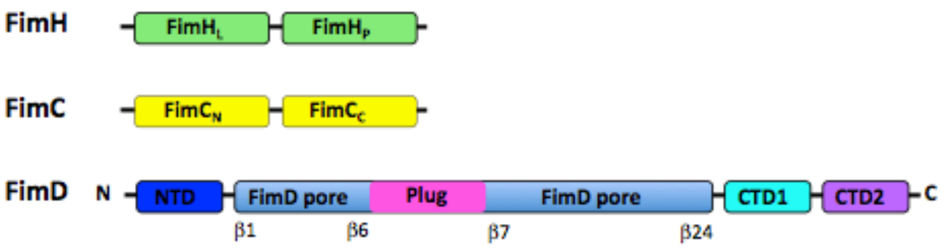
– Asymmetric unit copies	two
– Resolution	2.8 Å

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

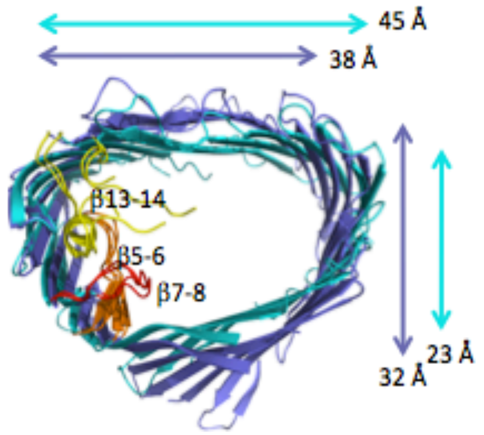


Usher complex structure solution

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



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



- 3．Fitting into the electron density
  - FimD-Plug
  - FimD-NTD
  - FimD-CTD-2

4. Manual building
   - FimD-CTD-1

Data and their property

1) Amplitudes of structure factors from single crystals - $|F_o|$, σ_o. It is the most common case. Usually structure factors are reliable, uncertainties are not so.
2) Intensities or amplitudes are from "twinned" crystals
3) Amplitudes of structure factors are available for |F+| and |F-| - SAD case
4) Amplitudes of structure factors are available from multiple crystal forms

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 one 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: bond lengths, angles etc.

It is one of many possible formulations. It uses Bayesian formulation. Other formulation is also possible.

-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:

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

And for centric reflections:

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

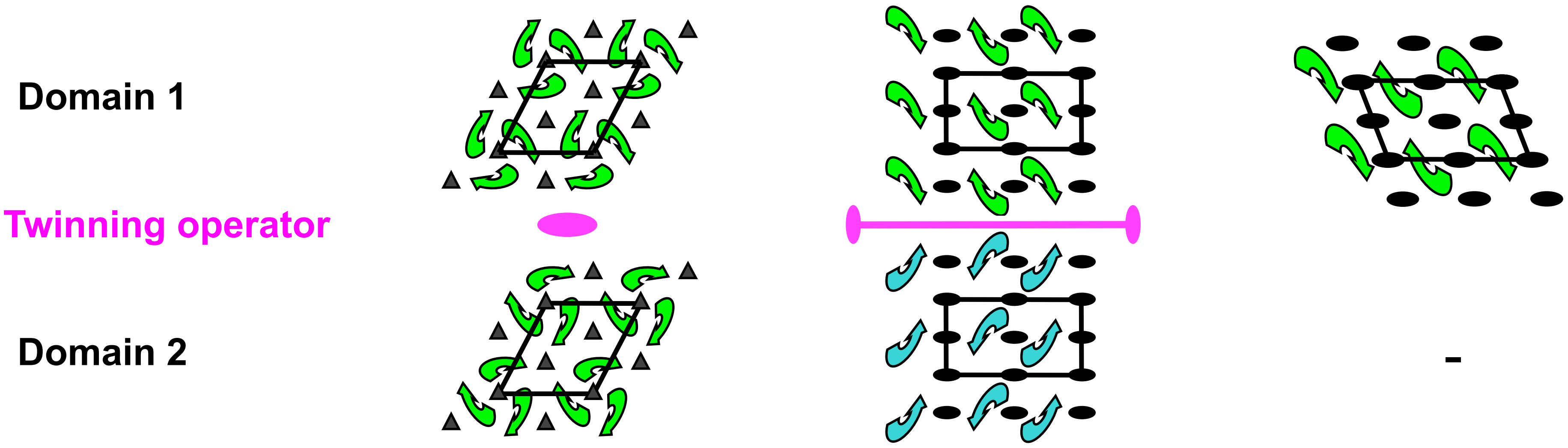
All parameters (scale, other overall and atomic) are inside $|F_c|$ and Σ

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

TWIN

merohedral and pseudo-merohedral twinning

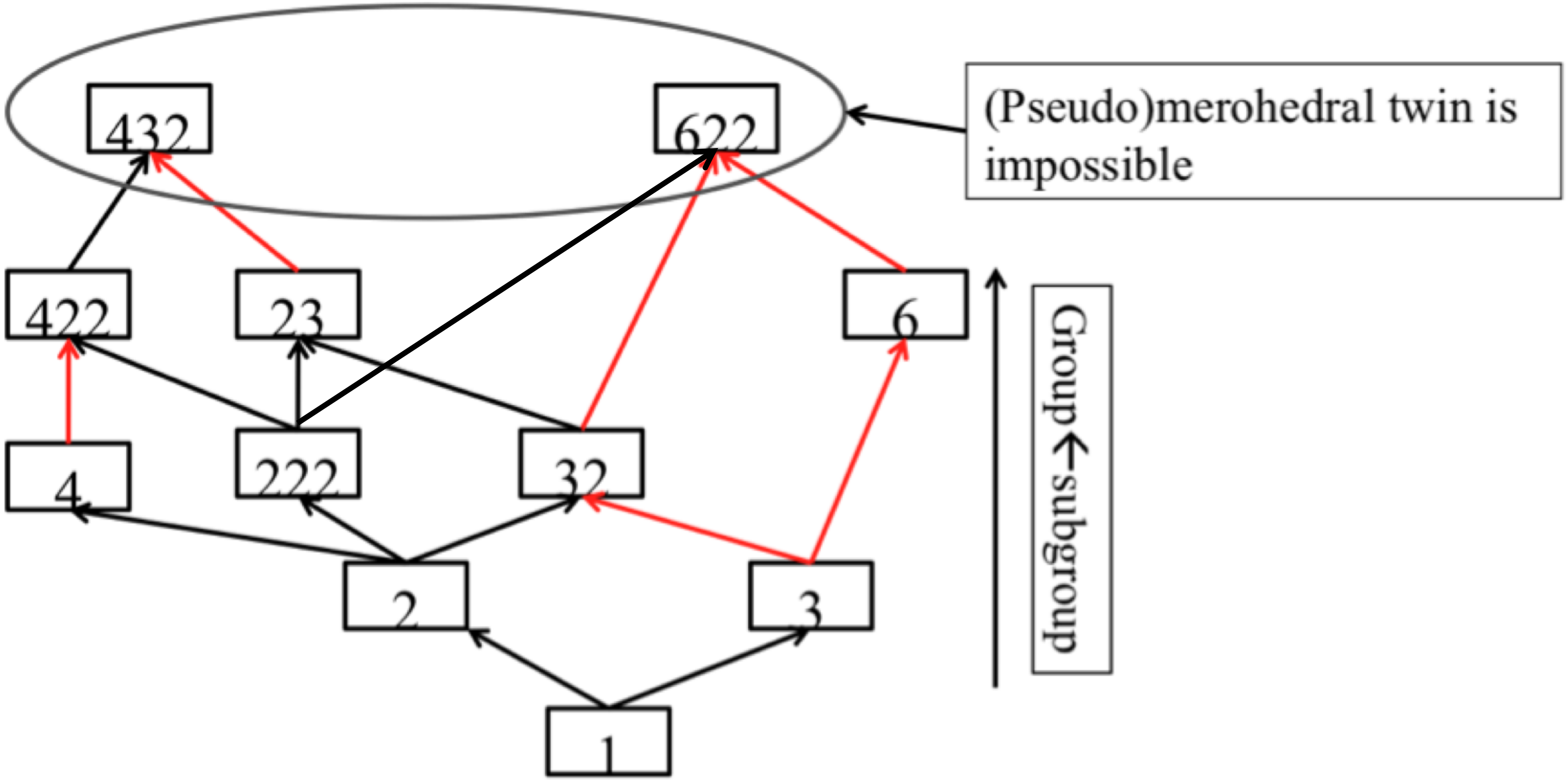
Crystal symmetry:	P3	P2	P2
Constrain:	-	β = 90º	-
Lattice symmetry *: (rotations only)	P622	P222	P2
Possible twinning:	merohedral	pseudo-merohedral	-



Crystal lattice is invariant with respect to twinning operator.

The crystal is NOT invariant with respect to twinning operator.

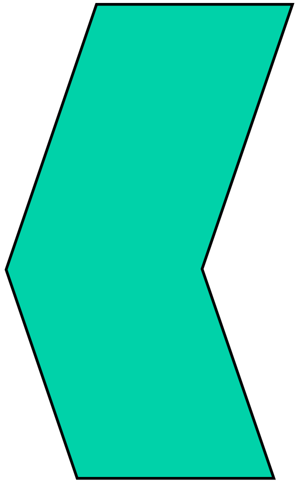
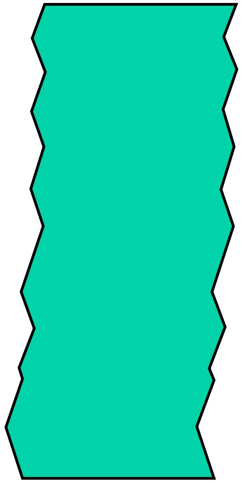
Twin refinement: Group/subgroup



(Pseudo)merohedral twin is impossible

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

Effect on intensity statistics

Take a simple case. We have two intensities: weak and strong. When we sum them we will have four options w+w, w+s, s+w, s+s. So we will have one weak, two medium and one strong reflection.

As a results of twinning, proportion of weak and strong reflections becomes small and the number of medium reflections increases. It has effect on intensity statistics

In probabilistic terms: without twinning distribution of intensities is χ^2 with degree of freedom 2 and after perfect twinning degree of freedom increases and becomes 4. χ^2 distributions with higher degree of freedom behave like normal distribution

Twin refinement in REFMAC

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

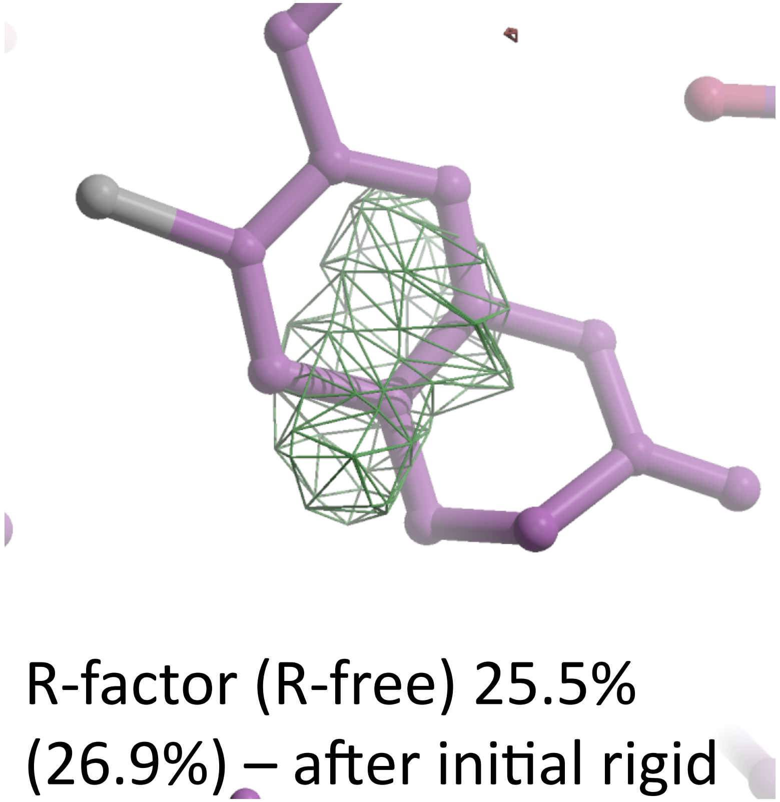
- Identify “twin” operators
- Calculate “Rmerge” ($\Sigma|I_h$-$<I>_{twin}|$ /ΣI_h) for each operator. If Rmerge>0.44 remove this operator: Twin plus crystal symmetry operators should form a group
- Refine twin fractions. Keep only “significant” domains (default threshold is 7%): Twin plus symmetry operators should form a group

Intensities can be used

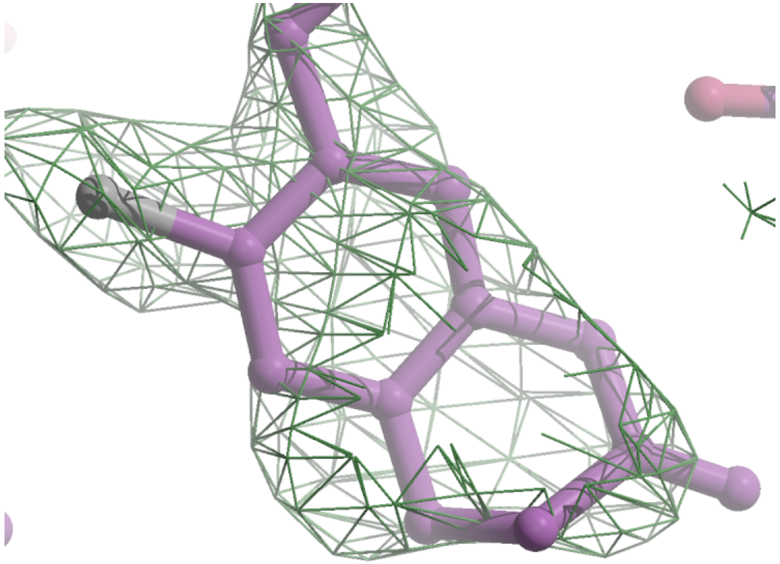
If phases are available they can be used

Maximum likelihood refinement is used

Where's the density for my ligand (2.15A)?



R-factor (R-free) 25.5% (26.9%) – after initial rigid body and restrained refinement.
Fo-Fc – 3 sigma



R-factor (R-free) 15.9% (16.3%) – re-run restrained ref. with twin on (refined twin fractions 0.6043/0.3957).
Fo-Fc – 3 sigma

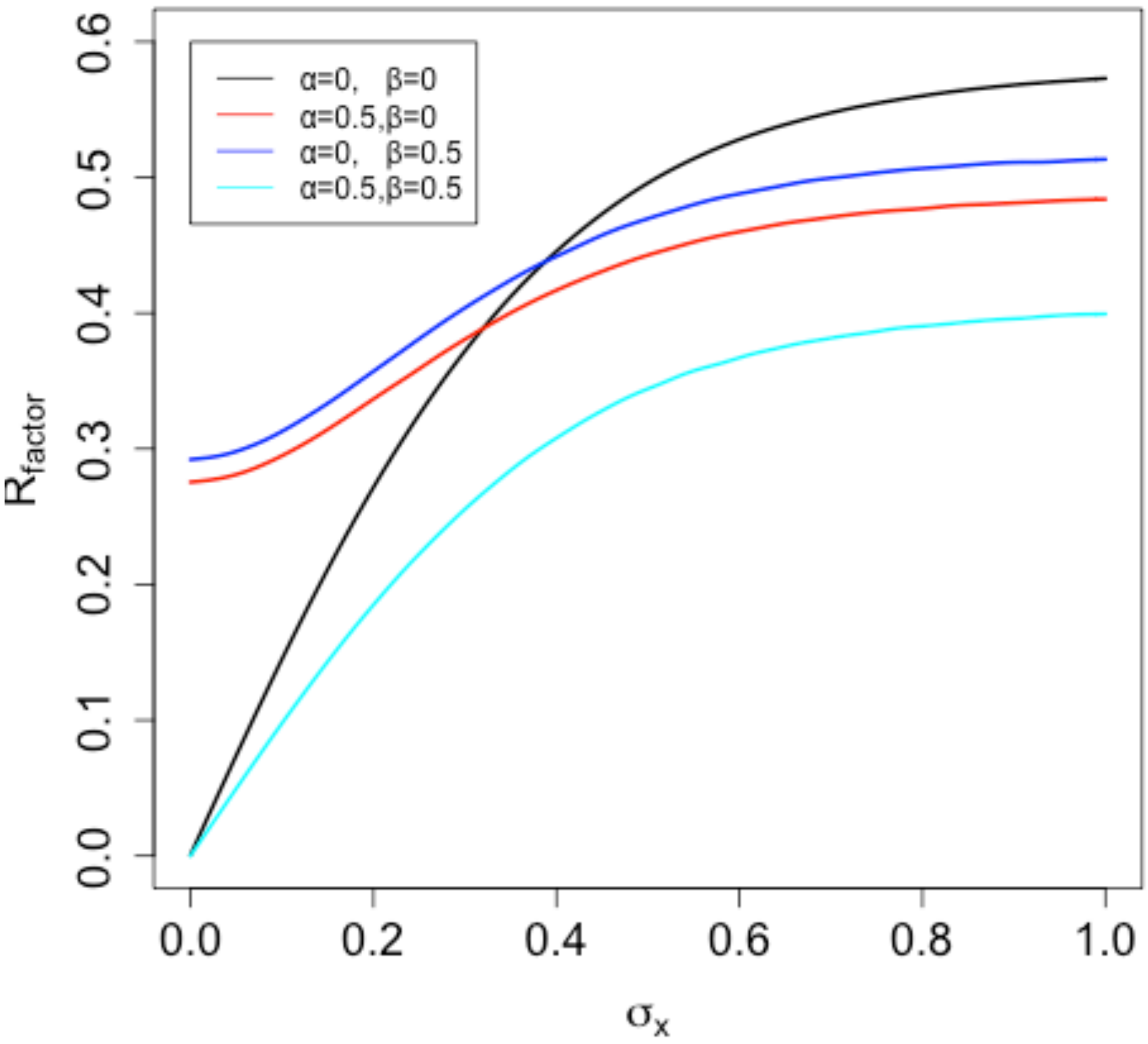
Twin: Few warnings about R values

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

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



Parameters

Usual parameters (if programs allow it)

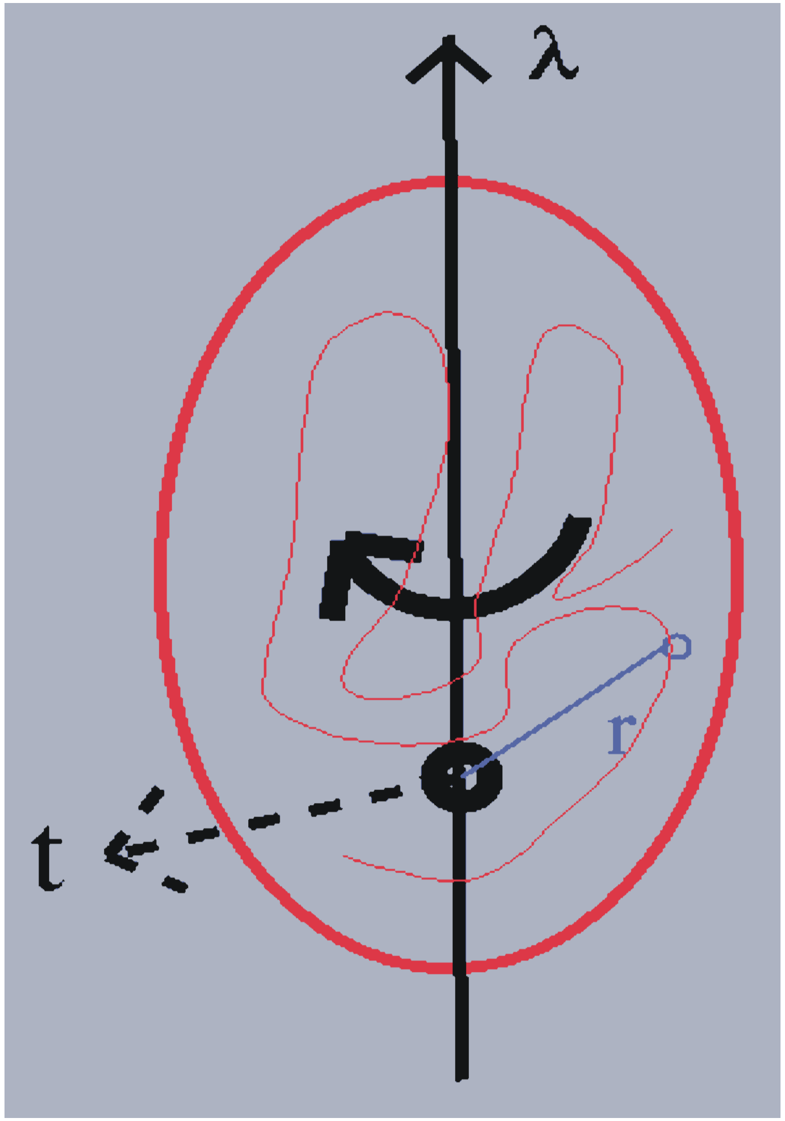
1) Positions x,y,z
2) B values – isotropic or anisotropic
3) Occupancy

Derived parameters

4) Rigid body positional
   - After molecular replacement
   - Isomorphous crystal (liganded, unliganded, different data)
5) Rigid body of B values – TLS
   - Useful at the medium and final stages
   - At low resolution when full anisotropy is impossible
6) Torsion angles

TLS

Rigid-body motion



General displacement of a rigid-body point can be described as a rotation along an axis passing through a fixed point together with a translation of that fixed point.

$$\underline{\mathbf{u}} = \underline{\mathbf{t}} + \mathbf{D}\underline{\mathbf{r}}$$

for small librations

$$\underline{\mathbf{u}} \approx \underline{\mathbf{t}} + \underline{\boldsymbol{\lambda}}\times\underline{\mathbf{r}}$$

D = rotation matrix

λ = vector along the rotation axis of magnitude equal to the angle of rotation

TLS parameters

Dyad product:

$$\underline{uu}^T = \underline{tt}^T + \underline{t\lambda}^T\times\underline{r}^T - \underline{r}\times\underline{\lambda t}^T - \underline{r}\times\underline{\lambda\lambda}^T\times\underline{r}^T$$

ADPs are the time and space average

$$U_{TLS} = \langle\underline{uu}^T\rangle = T + S^T\times\underline{r}^T - \underline{r}\times S - \underline{r}\times L\times\underline{r}^T$$

$T = \langle\underline{tt}^T\rangle$ 6 parameters, TRANSLATION

$L = \langle\underline{\lambda\lambda}^T\rangle$ 6 parameters, LIBRATION

$S = \langle\underline{\lambda t}^T\rangle$ 8 parameters, SCREW-ROTATION

TLS groups

Rigid groups should be defined as TLS groups. As starting point they could be: subunits or domains.

If you use script then default rigid groups are subunits or segments if defined.

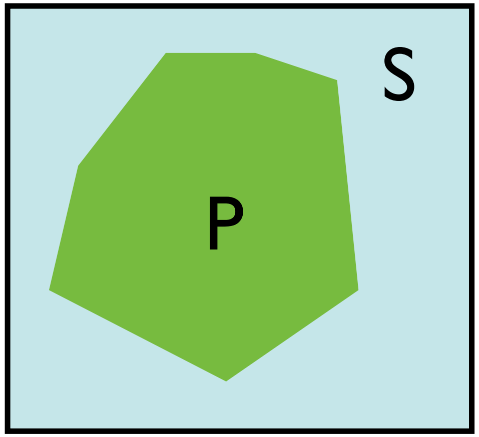
In ccp4i you should define rigid groups (in the next version default will be subunits).

Rigid group could be defined using TLSMD webserver:

http://skuld.bmsc.washington.edu/~tlsmd/

Bulk solvent
Method 1: Babinet's bulk solvent correction

At low resolution electron density is flat. Only difference between solvent and protein regions is that solvent has lower density than protein. If we would increase solvent just enough to make its density equal to that of protein then we would have flat density (constant). Fourier transformation of constant is zero (apart from F000). So contribution from solvent can be calculated using that of protein. And it means that total structure factor can calculated using contribution from protein only

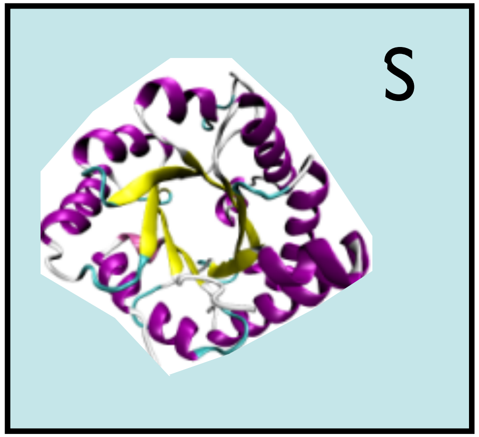


$$\rho_s+\rho_p=\rho_T \quad \Longleftrightarrow \quad F_s+F_p=F_T$$
$$\rho_s+k\rho_p=c \quad \Longleftrightarrow \quad F_s+kF_p=0$$
$$F_s=-kF_p \quad \Longrightarrow \quad F_T=F_p-kF_p=(1-k)F_p$$

k is usually taken as $k_b \exp(-B_b s^2)$. k_b must be less than 1. k_b and B_b are adjustable parameters

Bulk solvent
Method 2: Mask based bulk solvent correction

Total structure factor is the sum of protein contribution and solvent contribution. Solvent region is flat. Protein contribution is calculated as usual. The region occupied by protein atoms is masked out. The remaining part of the cell is filled with constant values and corresponding structure factors are calculated. Finally total structure factor is calculated using



$$\boxed{F_T = F_p + k_s F_s}$$

k_s is adjustable parameter.

Mask based bulk solvent is a standard in all refinement programs. In refmac it is default.

Overall parameters: Scaling

There are several options for scaling:

1) Babinet's bulk solvent assumes that at low resolution solvent and protein contributors are very similar and only difference is overall density and B value. It has the form:

$$k_b= 1-k_b\ e(-B_b\ s^2/4)$$

1) Mask bulk solvent: Part of the asymmetric unit not occupied by atoms are asigned constant value and Fourier transformation from this part is calculated. Then this contribution is added with scale value to "protein" structure factors. Total structure factor has a form:

$$F_{tot} = F_p+s_s\exp(-B_s\ s^2/4)F_s.$$

1) The final total structure factor that is scaled has a form:

$$s_{aniso}s_{protein}\ k_bF_{tot}$$

Map calculation

- After refinement programs give coefficients for two type of maps: 1) 2Fo-Fc type maps. 2) Fo-Fc type of maps. 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.

What and when

- Rigid body: At early stages - after molecular replacement or when refining against data from isomorphous crystals
- “Jelly” body – At early stages and may be at low resolution
- TLS - at medium and end stages of refinement at resolutions up to 1.7-1.6A (roughly)
- Anisotropic - At higher resolution towards the end of refinement
- Adding hydrogens - Higher than 2A but they could be added always
- Phased refinement - at early and medium stages of refinement
- SAD – at the early srages
- Twin - always (?). Be careful at early stages
- NCS local – always?
- Ligands - as soon as you see them
- What else?

Conclusion

- Information about ligands and their chemistry should be used in refinement
- NCS restraints are useful tool
- External restraints can be a powerful tool for reliable atomic model derivation.
- “Jelly” body can be very powerful at the early stages of refineement
- Twin refinement improves statistics and occasionally electron density: Rfactors may be misleading
- Refinement is just one step in X-ray structure analysis – it is often used as part of model building

Acknowledgment

Alexei Vagin
Andrey Lebedev
Rob Nicholls
Fei Long

Pavol Skubak
Raj Pannu

CCP4, LMB people

REFMAC is available from CCP4 and from our website:

http://www2.mrc-lmb.cam.ac.uk/groups/murshudov/