Refinement of Macromolecular Crystal Structures

Garib N Murshudov

MRC-LMB

Cambridge

Contents

1) Purpose of refinement
2) Refinement and map calculation
3) Model and Data
4) Reference structure restraints
5) Parameterisation

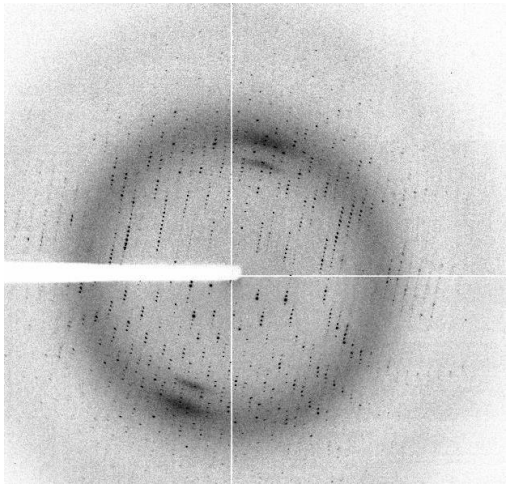
Purpose of refinement

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) map so that atomic model can be rebuild

Model Refinement



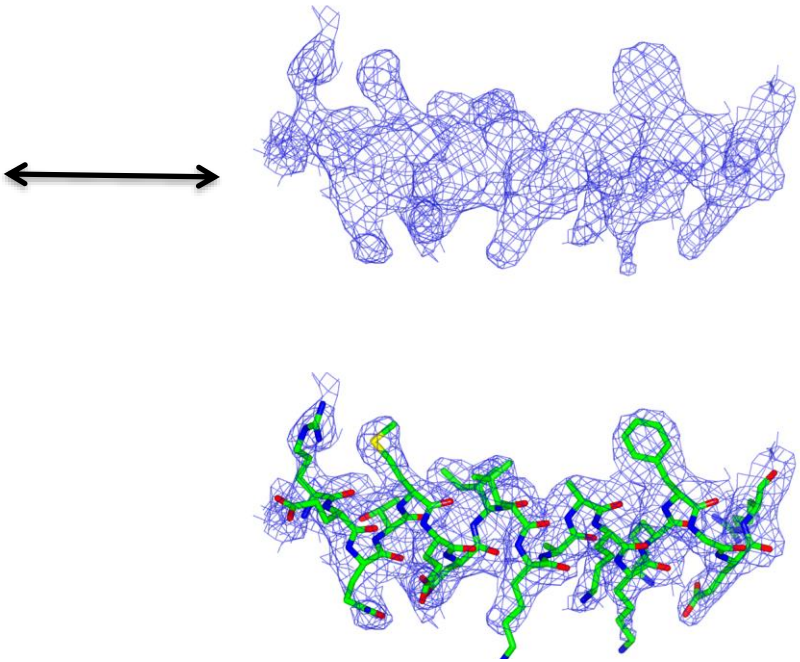
Idea:

Iteratively improve the model, optimising the agreement between $|F_{obs}|$ and $|F_{calc}|$

Purpose: improve phase estimates: φ_{calc}

H, K, L	$\lvert F_{obs} \rvert$	φ
…		
5, 5, 5	348	-
5, 5, 6	392	-
5, 5, 7	157	-
5, 5, 8	312	-
…		

H, K, L	$\lvert F_{calc} \rvert$	φ_{calc}
…		
5, 5, 5	355	27°
5, 5, 6	387	8°
5, 5, 7	146	75°
5, 5, 8	340	31°
…		



Likelihood and posterior distribution

We need:

Probability distribution of observations given parameters – likelihood:

parameter refinement

Probability distribution of ”ideal” maps given observations and parameters:

map calculation

Bayesian statistics is perfect for this type of problems. The problem can be recast as regularization of ill-posed problems also or projection to parameter space or many different ways

Refinement and map calculation

For refinement we need (log likelihood function – we need to integrate out unknowns)

$$L_X(p) = \log(P(obs; model)) = \log(\int_{true} P(obs; true); P(true; model))$$

Candidates are:

$\sum_{obs} \left\| F_o(h) \right| - \left| F_c(h) \right\|^2$	least-squares
$-\sum_{obs} \log(R(\left| F_o(h) \right|, \left| F_c(h) \right|))$	maximum likelihood using Rice distribution
$-\sum_{obs} \log(P(I_o \left| F_c \right|)$	Intensity based likelihood (e.g. used in twin)

Many others. For ideal refinement target we need functions that account for model errors, observational errors and variations due to changes in space (within crystal) and time (during data collection). If such a function would be implemented then we would not need to cut data.

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.

The problem is we do not have enough data to derive the model dependent only on data. We need additional information.

Map calculation

For parameter refinement we need the likelihood function

For map calculation we need the probability distribution of ”True” Fourier coefficients given observations and model parameters. General formula should be:

$$< F_T > = \int_{F_T, ModelError} F_T \, P(F_T; obs, model) \mathrm{P(ModelError; obs) dF_T d\ error}$$

We need the probability distribution of “True” Fourier coefficients as well as the distribution of model errors.

Similar calculations should be done for difference map calculations. With the same distributions we need to find:

$$< F_T - F_c >$$

Expected value of differences between “True” and model Fourier coefficients.

Probability distribution is derived using some basic assumptions like independence of model and observations if “true” map is known.

Sources of prior information

Knowledge about macromolecules used in refinement
As regularisers or prior knowledge

1. Macromolecules consist of atoms bonded to each other in a specific way
   ↳ **Standard restraints: bonds, angles etc**
2. Oscillation of atoms close to each other in 3D cannot be dramatically different
   ↳ **B-factor restraints, TLS restraints**
3. If there are two copies of the same molecule present then they will likely be similar to each other
   ↳ **NCS/local symmetry restraints**
4. If there are two molecules with sufficiently high sequence identity then it is likely that they will be structurally similar
   ↳ **External restraints to homologous structures - ProSMART**
5. Proteins tend to form secondary structures
   ↳ **Generic H-bonding restraints - ProSMART**
6. DNA/RNA tend to form base-pairs, stacked bases tend to be parallel
   ↳ **Generic base-pair and stacking restraints - LibG**

Refmac

About REFMAC

Refmac is a program for refinement of atomic models into experimental data

It was originally designed for Macromolecular Crystallography

It is based on some elements of Bayesian statistics: it tries to fit chemically and structurally consistent atomic models into the data.

It also can fit atomic models into cryo-EM maps.

It can do some manipulation of maps, e.g. sharpening/blurring

It is available from CCP4 and CCPEM

Restraints

Standard restraints (used by default) include:

- Bond lengths
- Angles
- Chirals
- Planes
- Some torsion angles
- B-values
- VDW repulsions

These help to ensure that the model is chemically sensible

Note – we generally deal with restraints, not constraints

Note – In ccp4 there is a program – AceDRG to generate dictionary for new compounds

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

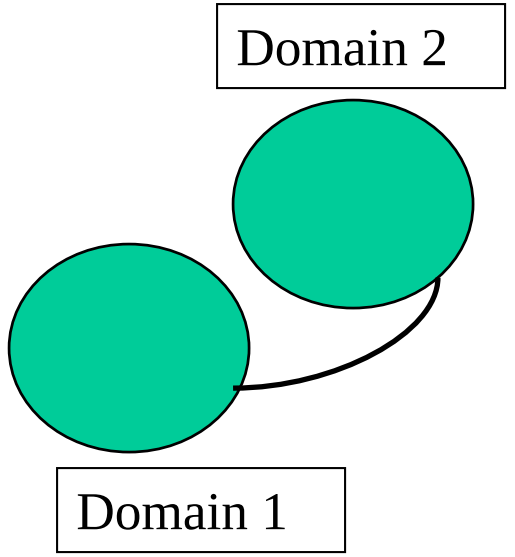
Auto NCS: local and global

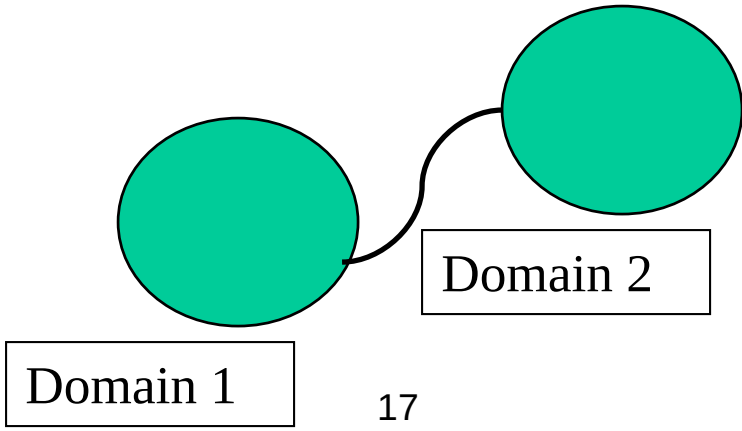
1. Align all chains with all chains using Needleman-Wunsh method
2. If alignment score is higher than predefined (e.g.80%) value then consider them as similar
3.Find local RMS and if average local RMS is less than predefined value then consider them aligned
4. Find correspondence between atoms
5. If 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: 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 CCP4.

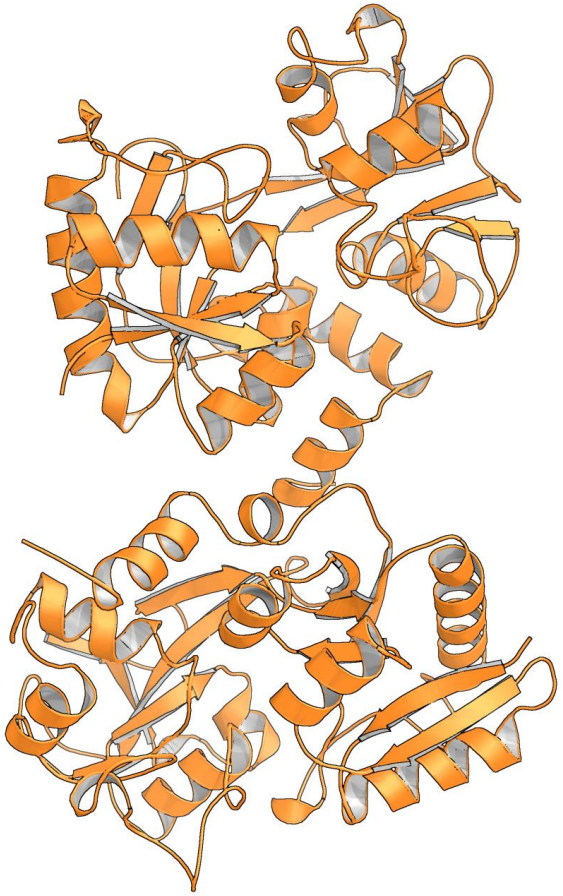
Note 2: Robust estimator functions are used for restraints. I.e. if differences between target and model is very large then their contributions are down-weighted

An Example

Ovotransferrin

1ryx – 3.5Å

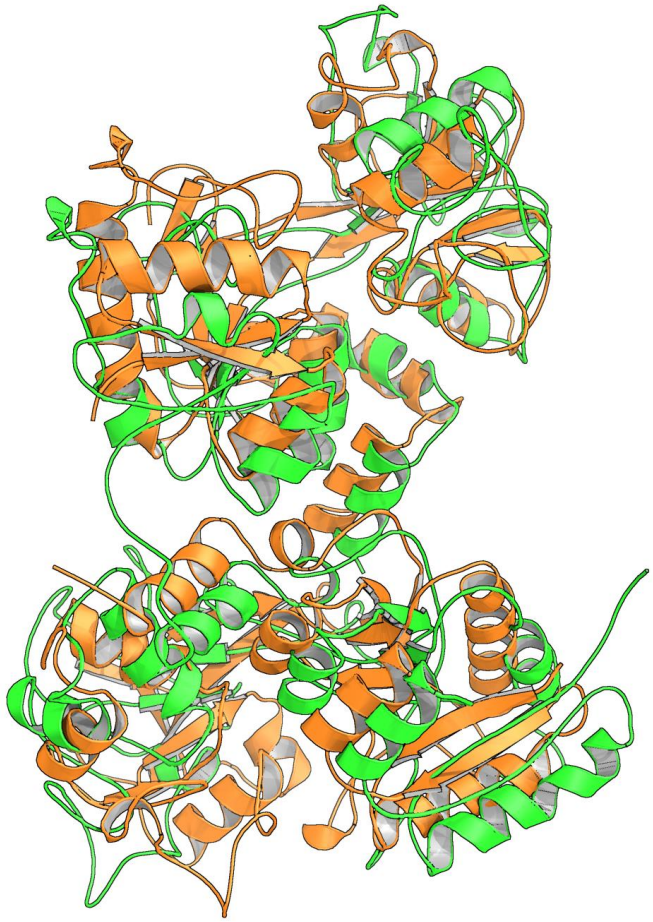
High-resolution homologue



2d3i – 2.15Å

An Example

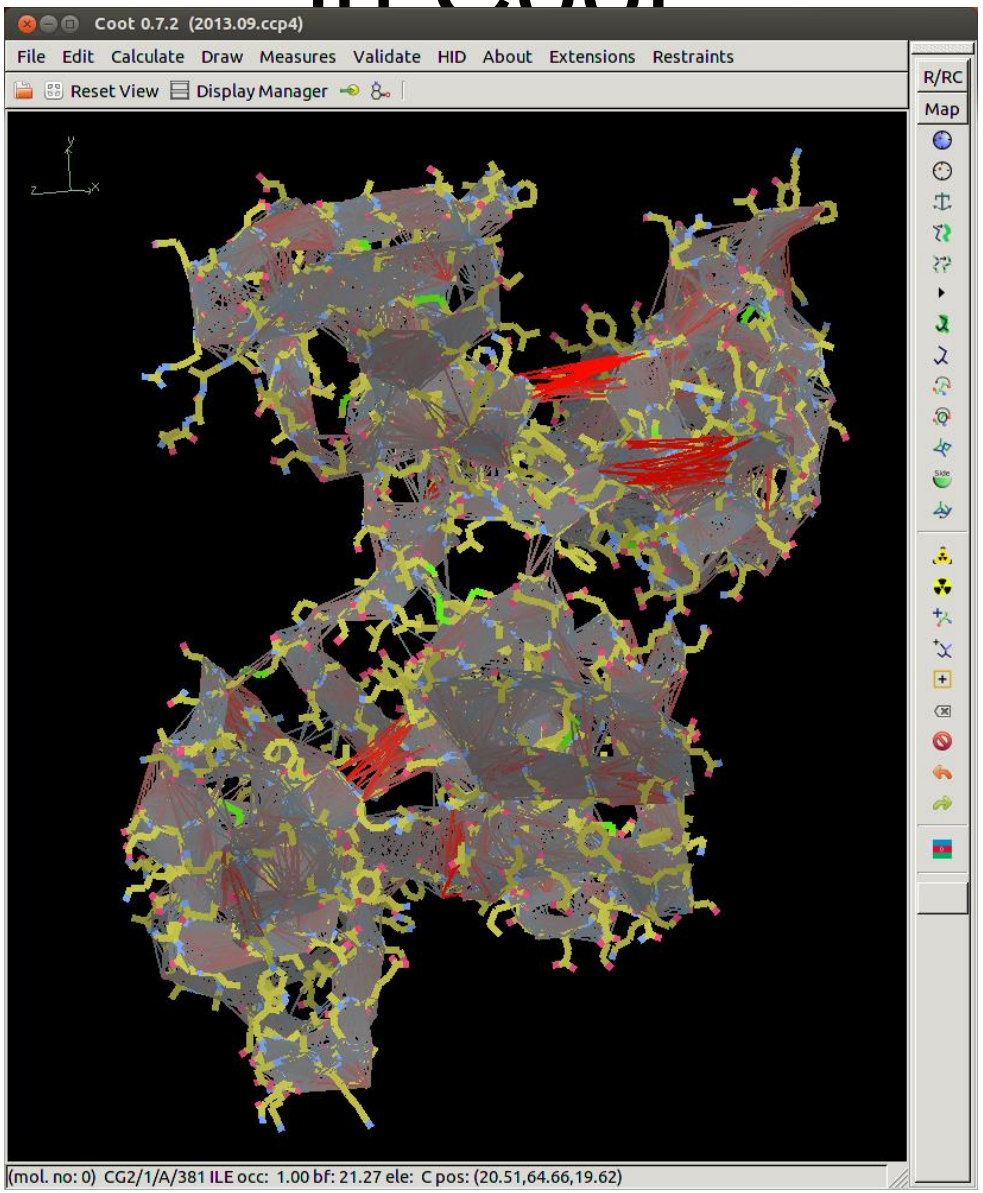
Ovotransferrin



Models don’t superpose well

ProSMART Restraint Visualisation in Coot

Backbone Restraints



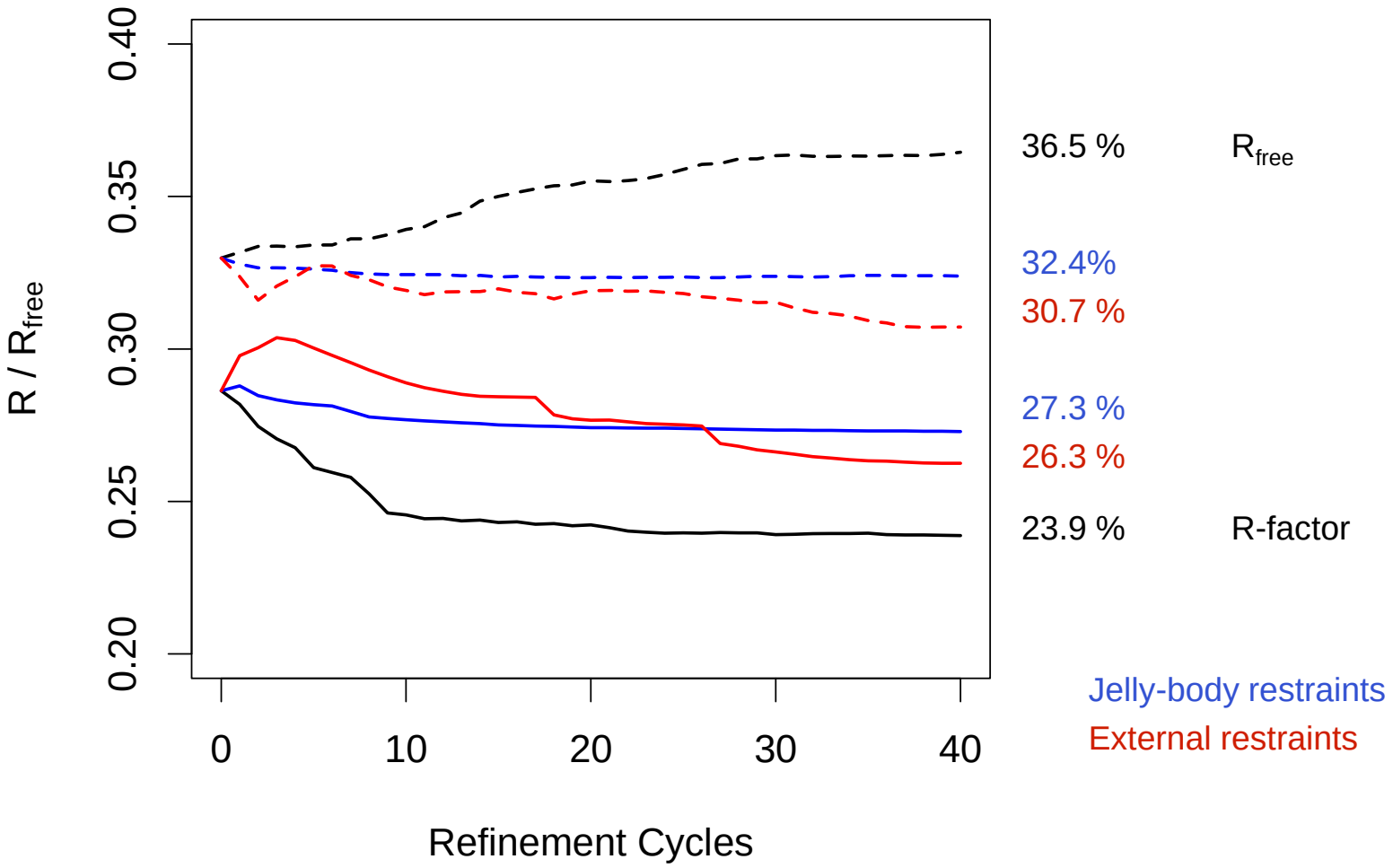
Ovotransferrin

1ryx (3.5Å)
restrained to
2d3i (2.15Å)

Thanks to Paul Emsley

External Restraints

Ovotransferrin

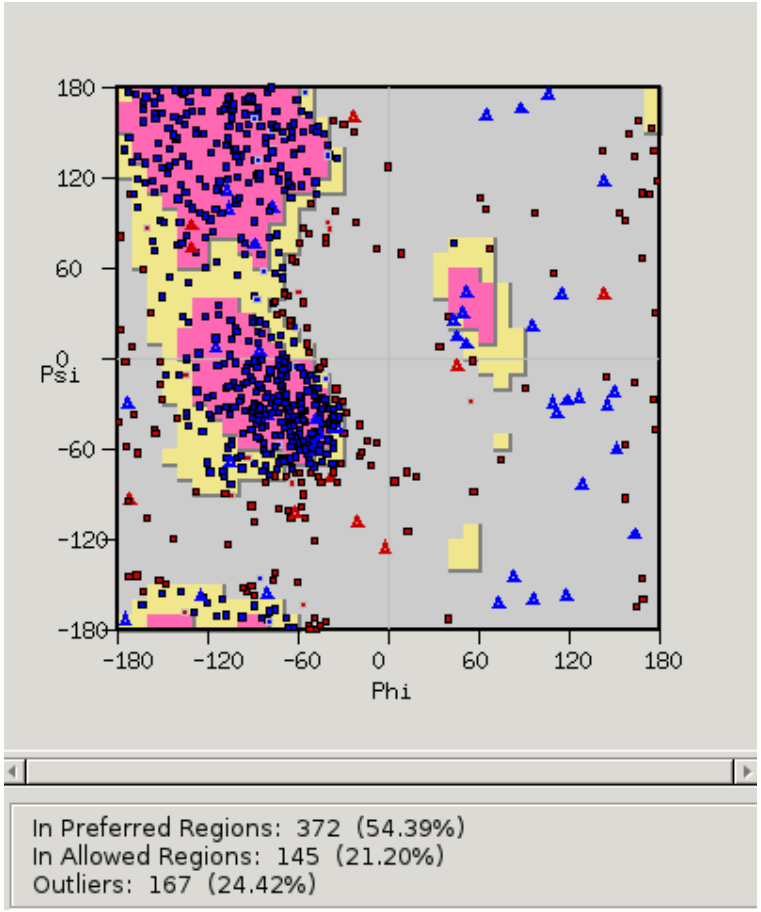


External Restraints

Ovotransferrin

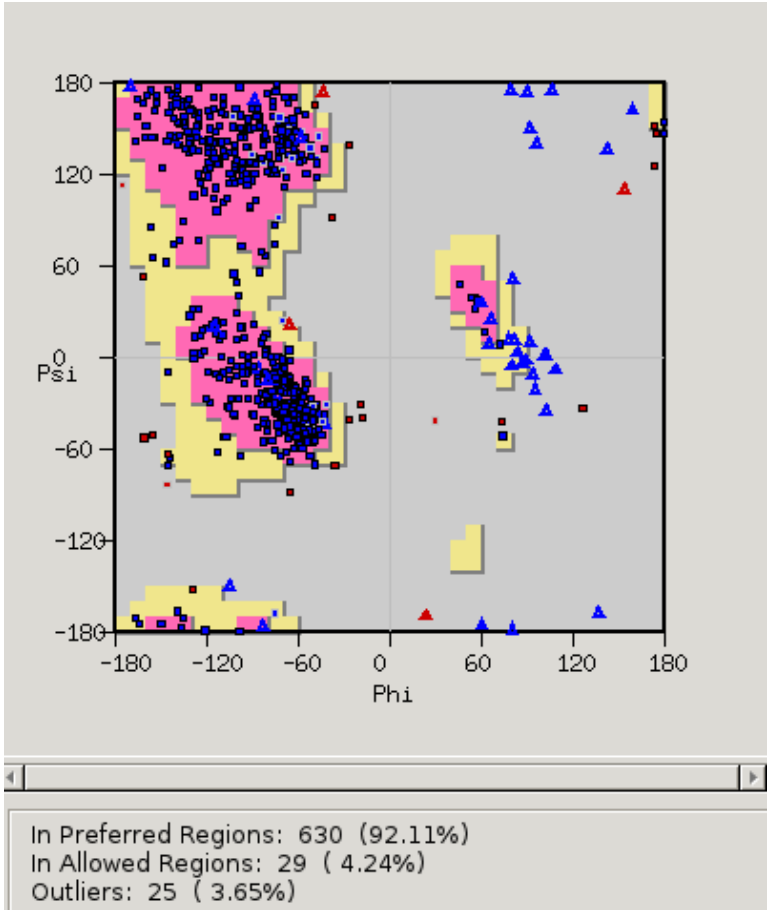
Original Structure

R/R_{free} : 0.286/0.330

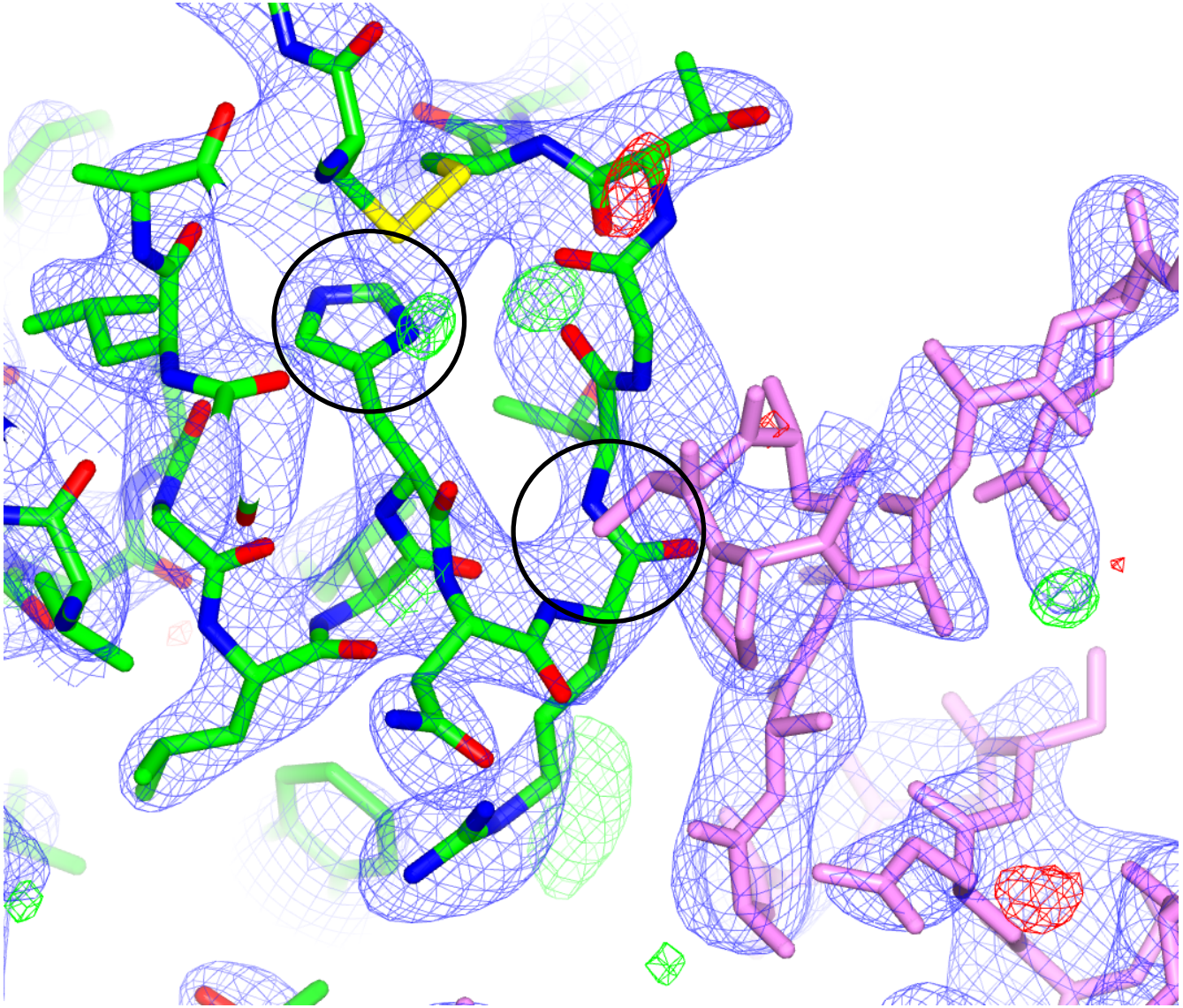


Re-refined with External Restraints

R/R_{free} : 0.263/0.307



External Restraints



1.3σ

Original Structure

R/R_{free} : 0.286/0.330

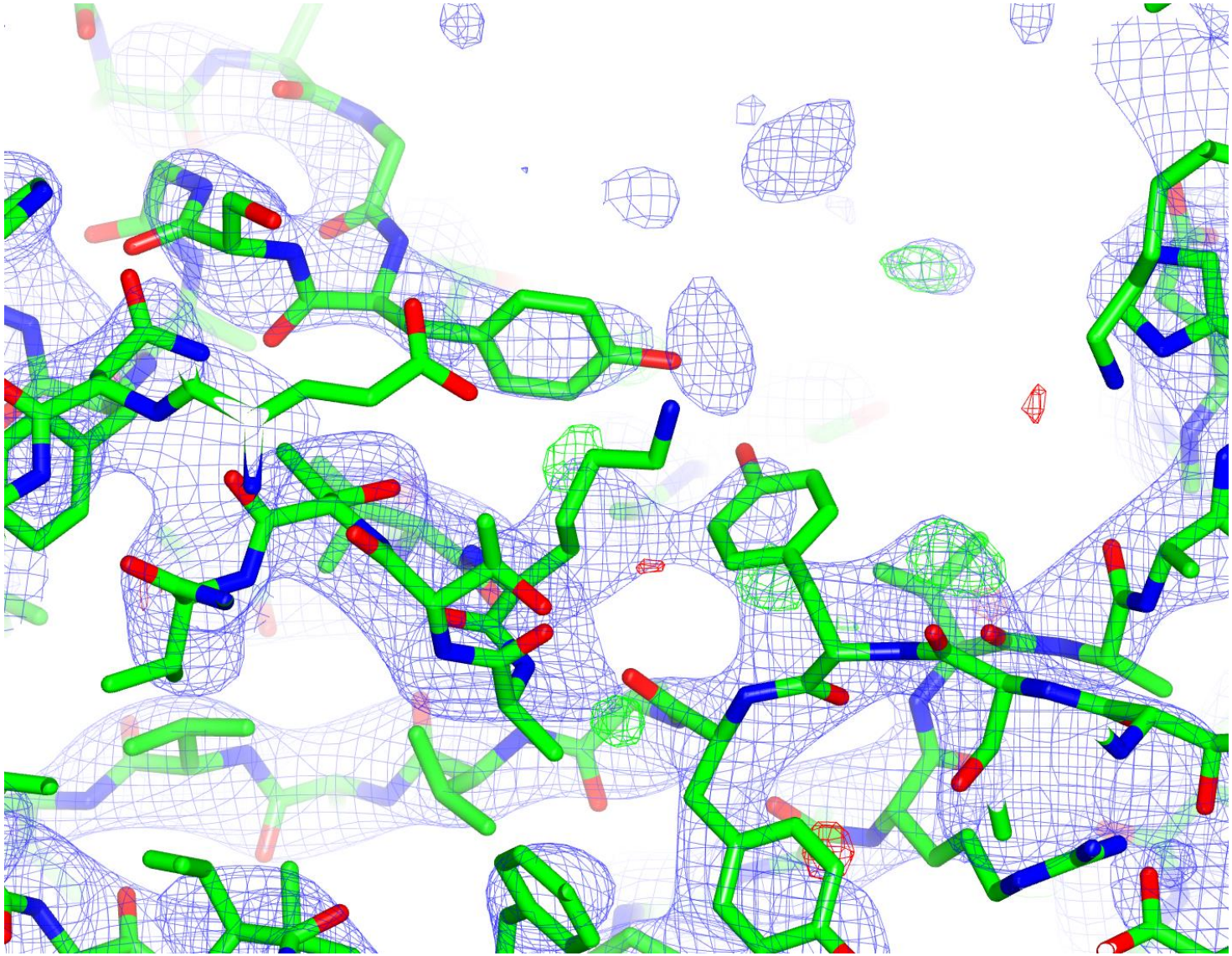
↓

External restraints

(40 cycles)

R/R_{free} : 0.263/0.307

External Restraints



1.3σ

Original Structure

R/R_{free} : 0.286/0.330

⬇

External restraints

(40 cycles)

R/R_{free} : 0.263/0.307

⬇

Build TYR92

Modify LYS209

⬇

Jelly body

(40 cycles)

R/R_{free} : 0.252/0.307

External Restraints

When refining at low resolution, check:

- Refinement statistics - ***Not always conclusive***
- Geometry - ***Not always conclusive***
- Electron density - ***Not always reliable***

Conclusion: At low resolution, everything has to add up!

Quality of prior information is important – consider manual re-refinement

Re-refinement can be done using PDB_REDO

What if there are no high-resolution homologues?

We still need to stabilise refinement…

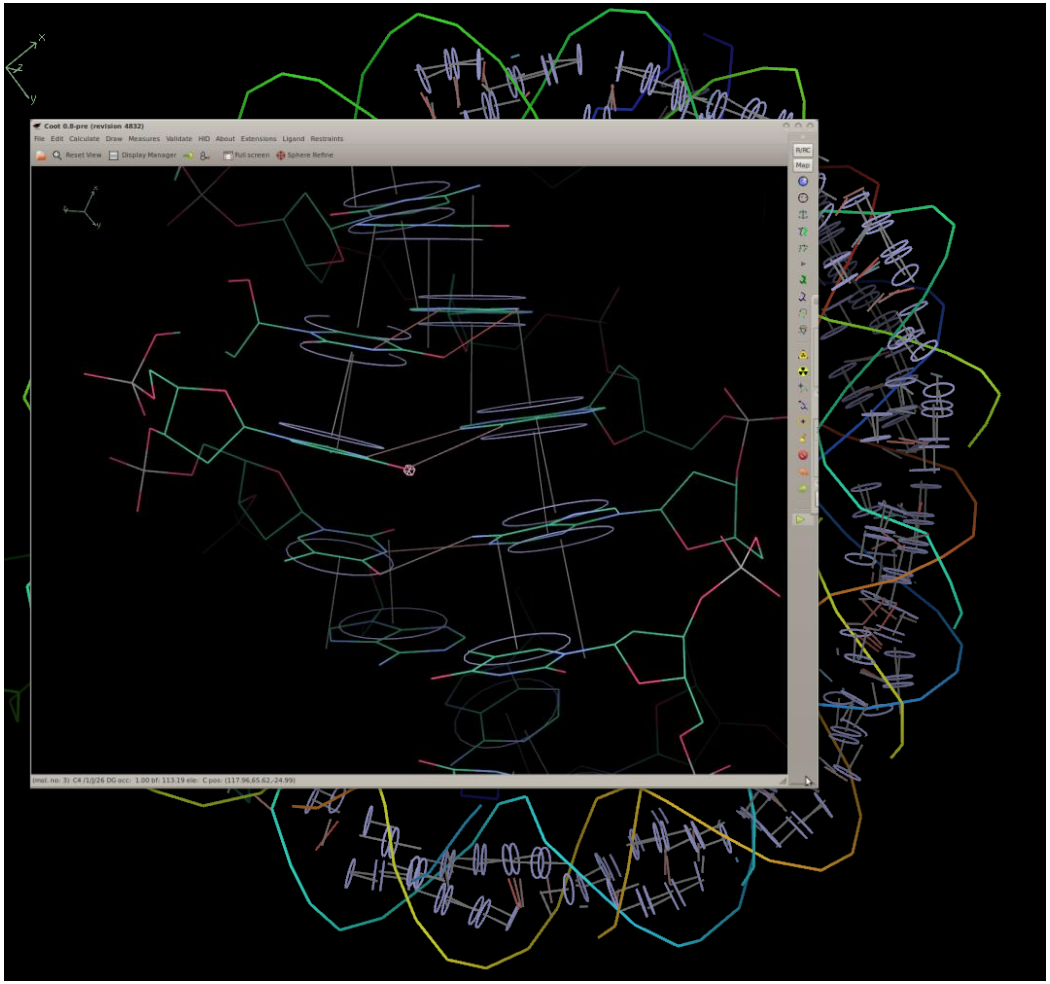
- Jelly-body restraints

- Generic external restraints:
  - ProSMART - protein (secondary-structure)
  - LIBG - DNA/RNA (base-pair, base-stacking)

LIBG Restraints for DNA/RNA

LIBG – for the generation of nucleic acid restraints

Base-stacking restraints:
(parallel plane restraints)



Restraints to current distances (jelly-body)

The term is added to the target function:

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

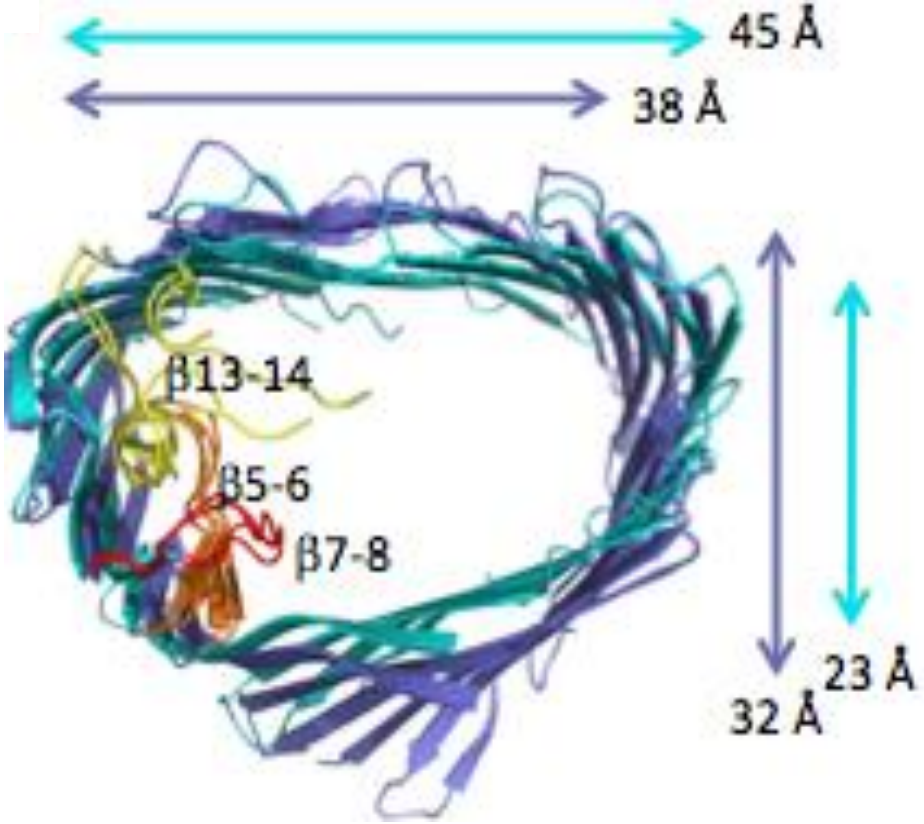
Summation is over all pairs in the same chain and within given distance (default 4.2A). $d_{current}$ is recalculated at every cycle. This function does not contribute to gradients. It only contributes to the second derivative matrix.

It is equivalent to adding plastic 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.

Usher complex structure solution

Jelly body refinement (Refmac)



Crystallographic Data

Different types of data:

1. Amplitudes of structure factors from single crystals:

   Observed amplitudes and sigmas: $|F_{obs}|$, σ_{obs}
2. Intensities/amplitudes from “twinned” crystals
3. SAD – amplitudes available for $|F_+|$ and $|F_-|$
4. Amplitudes available from multiple crystal forms

Note 1: Multiple crystal refinement is not available yet
Note 2: In all cases maximum likelihood refinement is used

TWIN

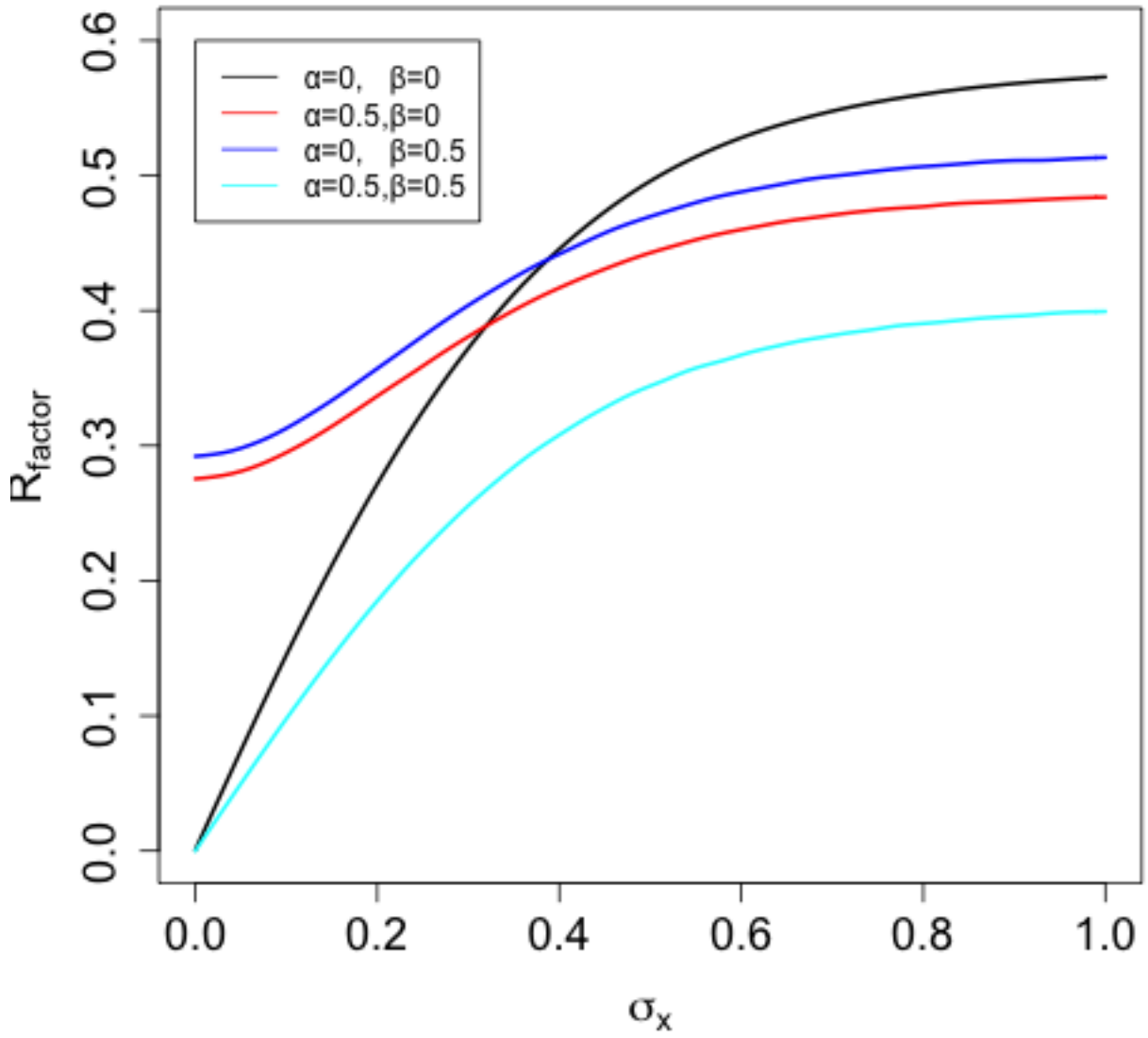
Twin: Few warnings about R values

Rvalues for random structures (no other peculiarities)

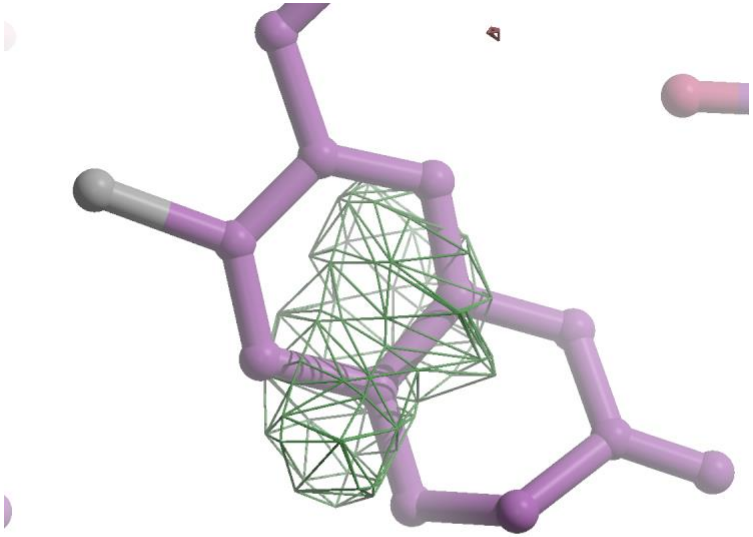
Twin	Modeled	Not modeled
Yes	0.41	0.49
No	0.52	0.58

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

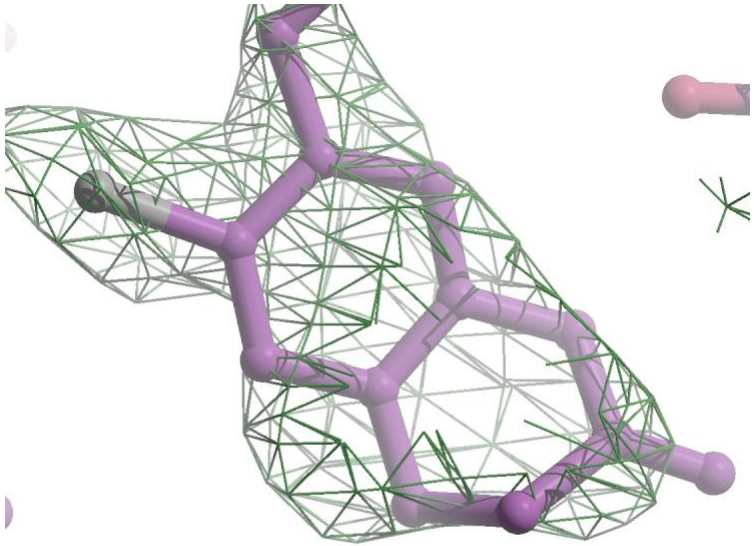
Rvalue for structures with different model errors:
Combination of real and modeled perfect twin fractions



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

Borrowed from B. Bax, GSK, Stevenage, UK

Problem: refinement statistics

Statistical tool: cross validation

1) Divide data into *k* roughly equal groups.

2) Refine against data excluding those from group *i*

3) Calculate statistics using the group *i*

4) Repeat 2 and 3 *k* times

This technique should be done from the beginning. It could take long time

Brunger used very simplified version. You take only one of the sets and leave them from refinement and model building. Usually 5% of the data excluded from refinement. It is the essence of freeR calculation.

Note 1: selection of the subset should be random

Note 2: different subset of the data will give different freeR statistics. Drop of freeR more important than its actual value

Brunger, Nature, 1992, 472-475
Luebden and Gruene, PNAS, 2015, 8999-9003

Problem: refinement statistics

R factors are most commonly used refinement statistics.

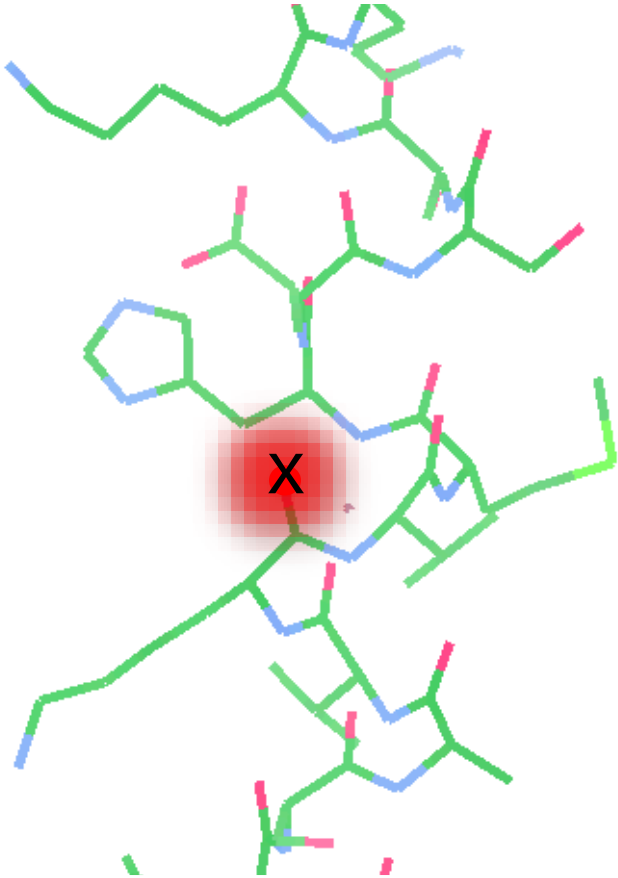
They depend on the distribution of the data: narrower distribution results in lower R factors

We need better statistics to monitor refinement statistics: LLG or information gain are candidates.

Another option is correlation. It seems to have better properties, especially when model is too far from perfet

Model Parameterisation

Standard refienable parameters



Atomic model:

- Position – (x,y,z) coordinates
- **Uncertainty – B-factors**
- (Occupancies)

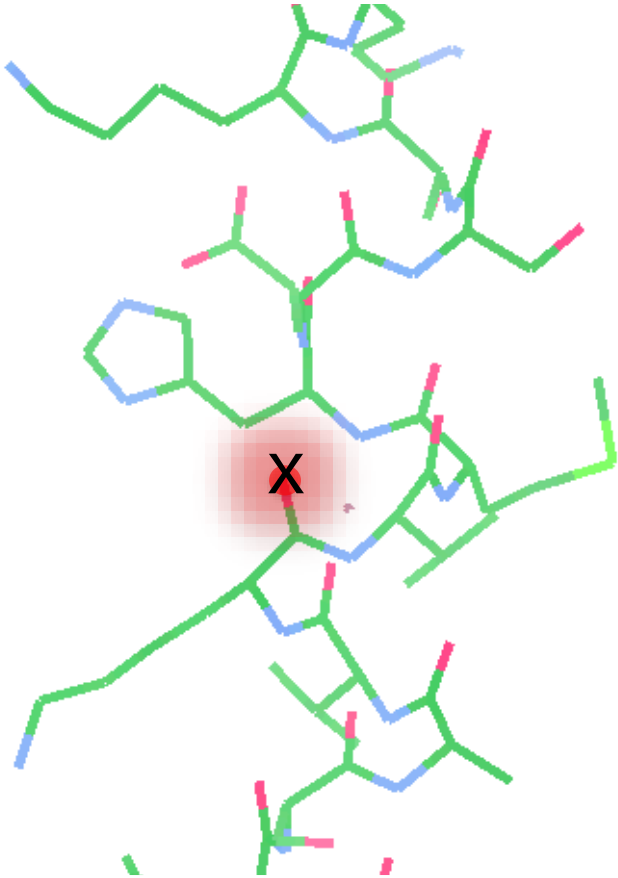
Overall parameters (scaling)

- Overall B-factor (and anisotropic U)
- Solvent treatment

```
ATOM      5  CB  ASP A   8     -30.909   9.723  18.264  1.00 33.70           C
ATOM      6  CG  ASP A   8     -31.252   9.345  16.825  1.00 41.96           C
ATOM      7  OD1 ASP A   8     -31.072  10.248  15.981  1.00 46.18           O
```

Model Parameterisation

Standard refineable parameters



Atomic model:

- Position – (x,y,z) coordinates
- Uncertainty – B-factors
- **(Occupancies)**

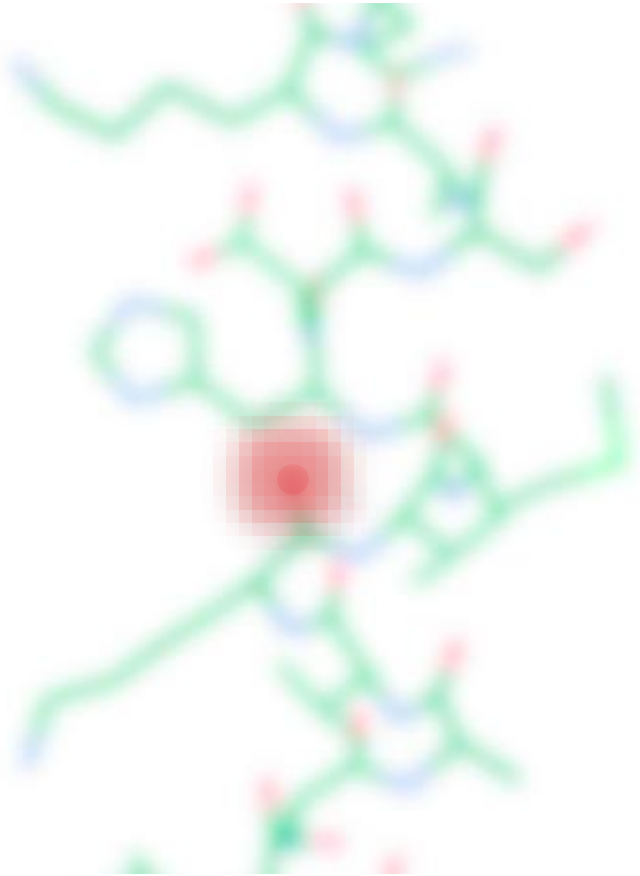
Overall parameters (scaling)

- Overall B-factor (and anisotropic U)
- Solvent treatment

```
ATOM      5  CB  ASP A   8     -30.909   9.723  18.264  1.00 38.70           C
ATOM      6  CG  ASP A   8     -31.252   9.345  16.825  1.00 41.96           C
ATOM      7  OD1 ASP A   8     -31.072  10.248  15.981  1.00 46.18           O
```

Model Parameterisation

Standard refineable parameters



Atomic model:

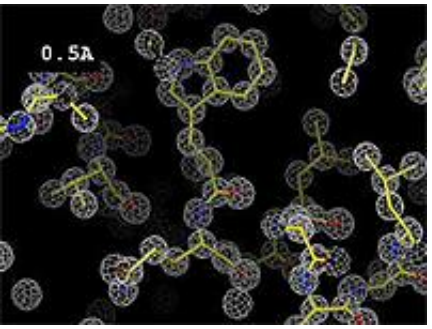
- Position – (x,y,z) coordinates
- Uncertainty – B-factors
- (Occupancies)

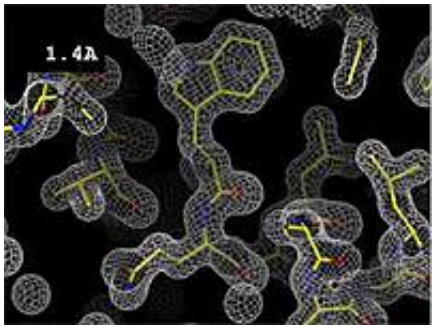
Overall parameters (scaling)

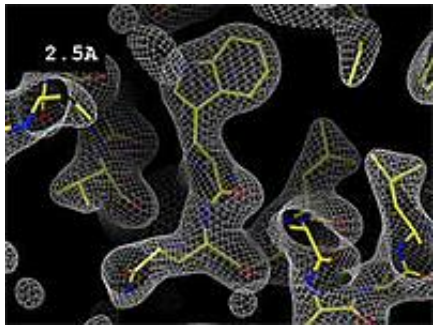
- **<u>Overall B-factor (and anisotropic U)</u>**
- Solvent treatment

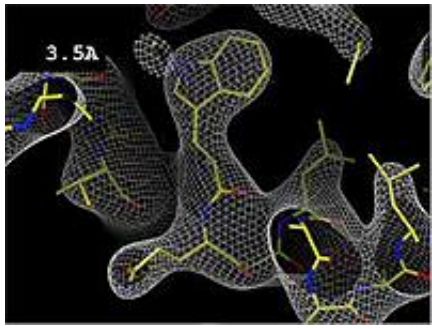
```
ATOM      5  CB  ASP A   8     -30.909   9.723  18.264  1.00 33.70           C
ATOM      6  CG  ASP A   8     -31.252   9.345  16.825  1.00 41.96           C
ATOM      7  OD1 ASP A   8     -31.072  10.248  15.981  1.00 46.18           O
```

Model Parameterisation







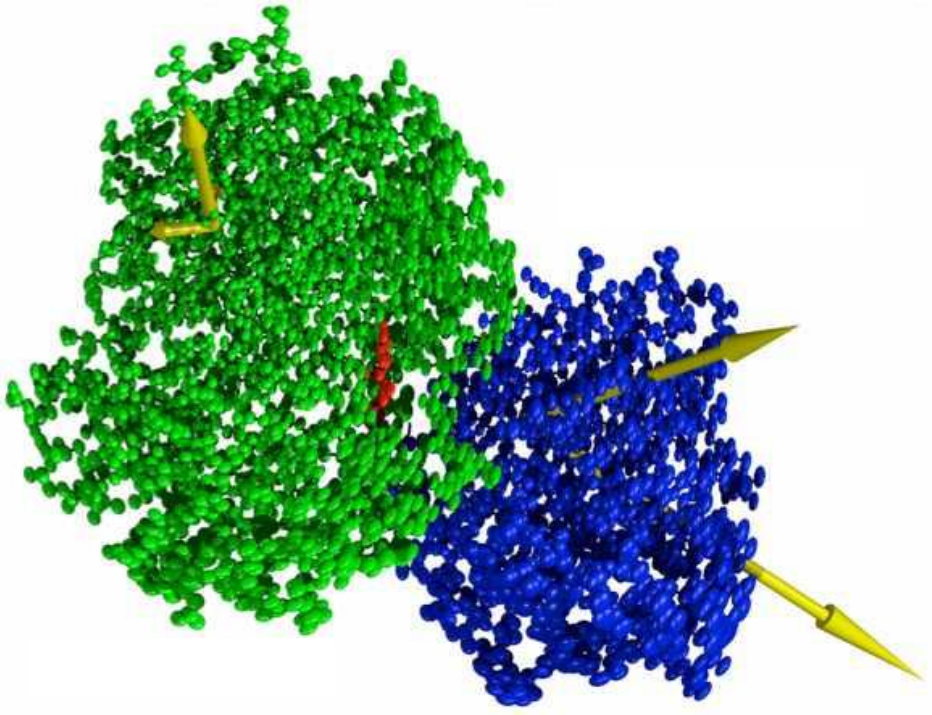


As resolution increases we see more and more details. At higher resolution we can afford to use more parameters (e.g. anisotropic). At lower resolution isotropic B values with restraints would be sufficient.

TLS Groups

Describe rigid body motion – e.g. for chains/domains/subunits

Suitable for medium resolution, when full anisotropy is impossible



Per group (20 parameters):

- **T**ranslation – 6 parameters
- **L**ibration – 6 parameters
- **S**crew rotation – 8 parameters

Define groups using CCP4i

or TLSMD webserver:

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

Overall Parameters: Solvent model

Two methods:

1. Babinet's bulk solvent correction

   Uses the fact that at low resolution solvent and protein contributions anticorrelate

1. Mask-based bulk solvent correction (default)

   It is assumed that solvent molecules are uniformly distributed outside the protein region

Map calculation

For parameter refinement we need the likelihood function

For map calculation we need the probability distribution of ”True” Fourier coefficients given observations and model parameters. General formula should be:

$$< F_T > = \int_{F_T, ModelError} F_T \, P(F_T; obs, model) \mathrm{P(ModelError; obsd F_T d\ error}$$

We need the probability distribution of “True” Fourier coefficients as well as the distribution of model errors.

Similar calculations should be done for difference map calculations. With the same distributions we need to find:

$$< F_T - F_c >$$

Expected value of differences between “True” and model Fourier coefficients.

Map calculation

In practice, currently, suboptmila Fourier coefficients are calculated. These are good when model errors and experimental errors are small.

Refmac calculates 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.

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**
- **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**
- **Helical, point group NCS constraints**
- **TLS refinement**
- **Fit into EM map**
- **Map sharpening**
- **etc**

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 - they could be added always
- Phased refinement - at early and medium stages of refinement
- SAD – at the early srages
- Twin – when you are sure that crystal is twinned
- NCS local – always?
- Ligands - as soon as you see them
- What else?

Summary

Tools to help with model building and refinement:

REFMAC: Jelly body refinement, map sharpening/blurring

ProSMART: External restraints, comparative analysis

LIBG: Nucleic acid restraints

ACEDRG: Ligand description dictionary and conformer generation

Many tools are applicable to cryo-EM as well as MX

Acknowledgements

Computational Crystallography Group:

Rob Nicholls
Paul Emsley
Oleg Kovalevskiy
Fei Long
Michal Tykac
Rangana Warshamanage

Our collaborators
Lamzin group, Hamburg, Germany
Raj Pannu and Pavol Skubak, Leiden Netherlands
Robbie Joosten, Amsterdam, Netherlands

Many Thanks:

Marcus Fischer
Stuart McNicholas
Tom Burnley & Martyn Winn
Alan Brown
Ben Bax
Jake Grimmett & Toby Darling

CCP4 core team
Colleagues from MRC-LMB and CCP4

MRC Laboratory of Molecular Biology

CCP4

wellcome trust

BBSRC bioscience for the future

LMB courses and other videos

MRC LMB youtube

There are number of different videos. Go to courses section and find course you are interested in (e.g. macromolecular crystallography)

Refmac and related tutorials

MRC Murshudov and go to the personal webpage

from there to software and tutorials.