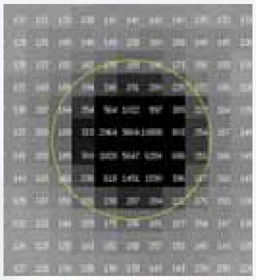
HKL2000

Dominika Borek
UT Southwestern Medical Center at Dallas

Data processing

X-ray data processing = changing detector output to estimate of square of structure factors amplitudes

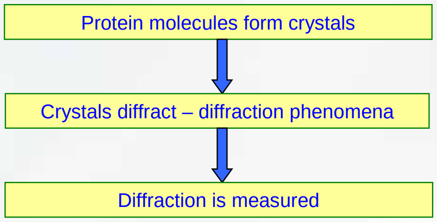


$$\rightarrow \left|\mathbf{F}\right|^2$$

Data processing in HKL2000

- Detector description (e.g. site file)
- Autoindexing (Denzo) and visual assessment (XDisplayF)
- Refinement of experimental parameters and optimization of integration parameters (Denzo)
- Integration (Denzo)
- Scaling (Scalepack)
- Merging and statistical assessment (Scalepack and HKL2000)

Data model



Protein molecules form crystals

Crystals diffract – diffraction phenomena

Diffraction is measured

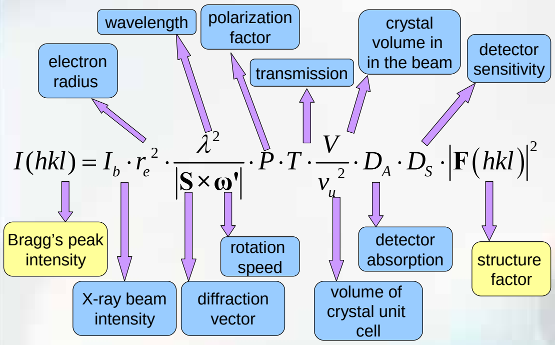
Data Model of a Crystal

Crystal ≡ ideal space group symmetry in a perfectly ordered infinite crystal lattice

Deviations:

- Finite crystal size
- Ideally imperfect crystal (no double scattering and no extinction)
- Observable mosaicity
- Multiple lattices due to phase transition
- Twinning
- Pseudosymmetry
- Modulated structures (Wang, J. (2001) J. Struct. Biol. 134, 1524; Bochtler *et al.* (2001) J. Struct. Biol. 135, 281)

Data model - diffraction



$$I(hkl) = I_b \cdot r_e^2 \cdot \frac{\lambda^2}{\left|\mathbf{S}\times\boldsymbol{\omega}'\right|} \cdot P \cdot T \cdot \frac{V}{v_u^2} \cdot D_A \cdot D_S \cdot \left|\mathbf{F}(hkl)\right|^2$$

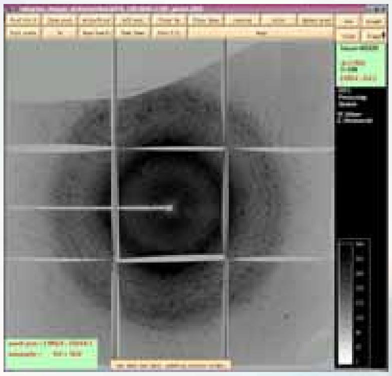
Diffraction - Deviations

- Radiation damage
- Double scattering
- Uneven exposure
- Uneven rotation
- Contaminating wavelength
- Absorption

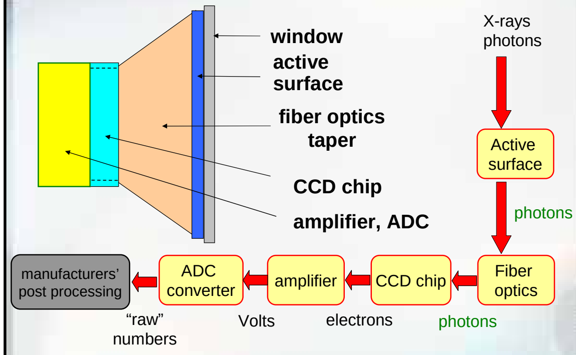
Data Model of Measurements - I

Obscuration:

1. Beam stop
2. Cryo-colling
3. Goniostat



Data Model of Measurements - II



window

active surface

fiber optics taper

CCD chip

amplifier, ADC

X-rays photons

Active surface

photons

manufacturers' post processing

ADC converter

amplifier

CCD chip

Fiber optics

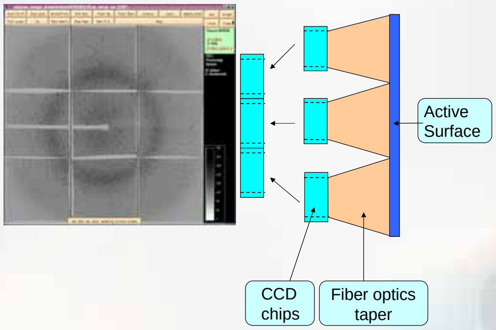
"raw" numbers

Volts

electrons

photons

Assembly of diffraction pattern

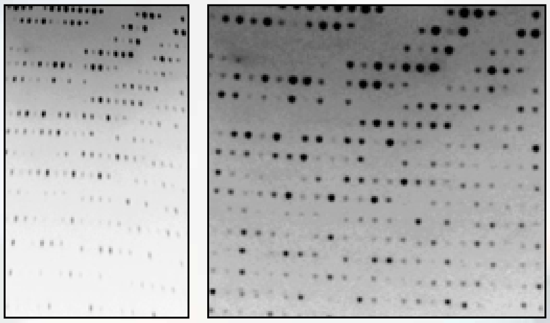


Active Surface

CCD chips

Fiber optics taper

Image distortion - magnified



Detector geometric corrections - method I

Correction of images

- allows for the use of integration software that does not apply distortion corrections
- looks nice from a distance
- closer view:
  - non-uniformly broadens spots – creates overlaps, makes profile fitting less accurate
  - flattens variations – affects error model, creates moiré pattern
- allows to simulate spherical detectors

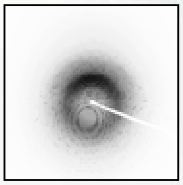
Detector geometric corrections - method II

Correction of diffraction pattern

- requires programs to understand detector specifics
- optimizes overlaps and profile fitting
- produces a better error model
- looks a bit strange (fake gaps)

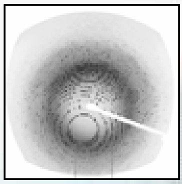
Detector description – Site file

- The site file contains numerical parameters describing how reciprocal space is distorted on diffraction image. These parameters belong to two groups: one describing geometry of distortion and optional second describing sensitivity of each pixel on the detector.



- Wrong site file results in:

- misindexation, misprediction of spots' positions, wrong refinement of processing parameters
- wrong correction of intensities due to wrong values of pixels' sensitivity



Indexing

Assigning hkl index to diffraction maxima (spots)

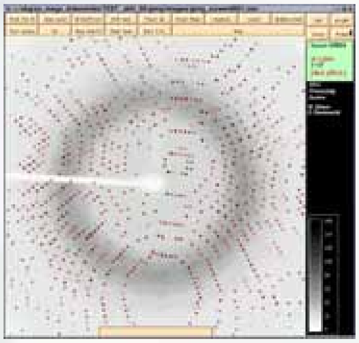
Requirements:
- approximate description of detector geometry (x beam, y beam !!!, distance, detector orientation)
- free of artifacts list of peaks (peak search)
    - * twins, ice, zingers, satellites
        - - manual editing, resolution limits
- proper procedure
    - * spots separation
        - longest vector = distance*λ/(spot size)
    - * oscillation range

viruses	0.25º
proteins	1º
small molecules	2.5º

Autoindexing

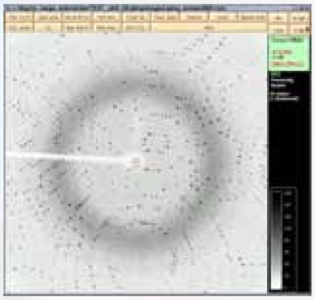
- peak search
- autoindexing in primitive lattice
- choice of Bravais lattice (lattice symmetry)
- reindexing to standard symmetry
- if more than one crystal involved – checking the consistency of indexing between crystals
    - needed only for some space groups
    - after scaling of data from crystals separately

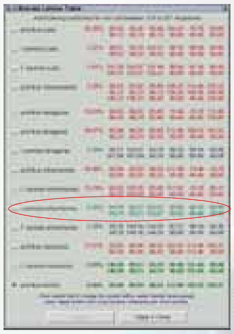
Peak Search



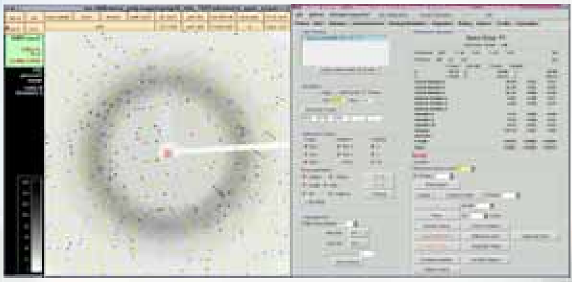
It finds the strongest intensity peaks

Autoindexing in primitive lattice and choice of higher symmetry Bravais lattice (if possible)

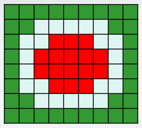




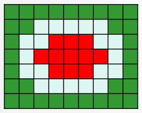
Choice of integration parameters- Spot size too small



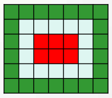
Spot size and integration parameters



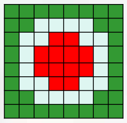
ibox 8 9
background 6 x 7
spot 4 x 5



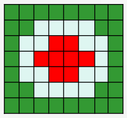
ibox 7 9
background 5 x 7
spot 3 x 5



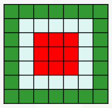
ibox 6 7
background 4 x 5
spot 2 x 3



ibox 8 8
background 6 x 6
spot 4 x 4



ibox 7 8
background 5 x 6
spot 3 x 4



ibox 7 7
background 5 x 5
spot 3 x 3

Why autoindexing may fail?

Peak search list problems:

- less than 10 diffraction peaks
- twin crystals, ice, satellites crystals, zingers

→ Resolution limits
Manual editing
I/σ(I) selection

Detector description:

x beam, y beam → Refined values from previous experiment
Direct beam exposure
Center of powder diffraction (ice)

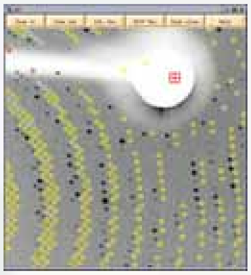
Unknown system

- spindle direction
- correction files

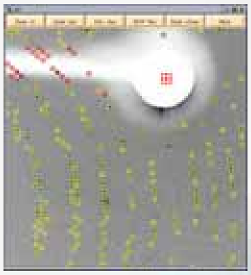
→ Site documentation
Four possibilities (only one works when indexing in 3D-mode)

Xbeam, Ybeam

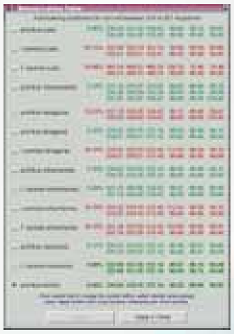
150.6 151.2



150.0 150.0



Xbeam, Ybeam

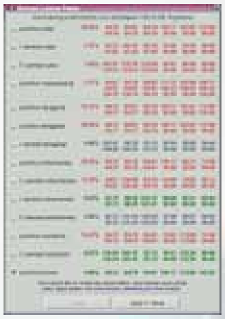


Xbeam, Ybeam - Importance of precision depends on unit cell parameters

150.6 151.2

For glucose isomerase

With shorter unit cell parameters refinement converges and gives correct Bravais lattice



Why autoindexing may fail?

- Procedure problems:

$$longest\ vector = \frac{distance \cdot \lambda}{spot\ size}$$

- spot size reduce spot radius
- distance re-collect image at longer distance
- mosaicity too large reorient the crystal if only one axis is affected
- rotation range too large decrease for large unit cells, but even if indexing works there may be too many overlaps

Refinement of instrument and crystal parameters

Crystal: - orientation (rotx, roty, rotz)
- unit cell
- mosaicity

Beam: - focus parameters (crossfire x y xy - 0 values for beam focused on the detector)

Detector: - distance (distance)
- orientation (rotx, roty, rotz)
- position (x beam, y beam)
- internal geometry (radial offset, angular offset, y scale, skew, distortions)

The parameters could be the same or different for consecutive images

Refinement - target

Minimization of target function

$$\chi^2_{total} = \chi^2_x + \chi^2_y + \chi^2_p$$

$$\chi^2_x = \sum_{spots} \frac{\left(x_{predicted} - x_{observed}\right)^2}{\sigma^2_x}; \quad \chi^2_y = \sum_{spots} \frac{\left(y_{predicted} - y_{observed}\right)^2}{\sigma^2_y}$$

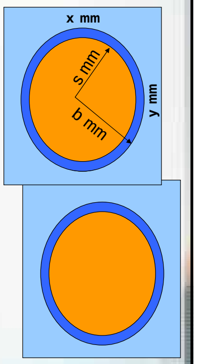
$$\chi^2_p = \sum_{hkl} \frac{\left(p_{predicted} - p_{observed}\right)^2}{\sigma^2_p}; \quad p = \frac{I_{hkl,\ frame}}{I_{hkl,\ total}}$$

The displayed values of χ^2 are divided by number of observations

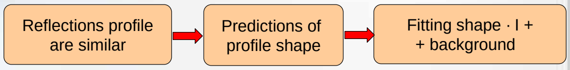
Integration of diffraction peaks - I

- based on analysis of local environment of peaks - “box” (box x_mm y_mm or ibox x_pixels y_pixels)
- definition of spot area (spot radius s_mm)
- background is outside of spot area (including other reflections) and outside of background circle (background radius b_mm)
- background is analyzed for slope (linear variations with positions) and artifacts

Spot and background are symmetric with respect to the center of the box.



Integration of diffraction peaks - II



Reflections profile are similar → Predictions of profile shape → Fitting shape · I + + background

close in x,y,φ
same relations to pixel boundary

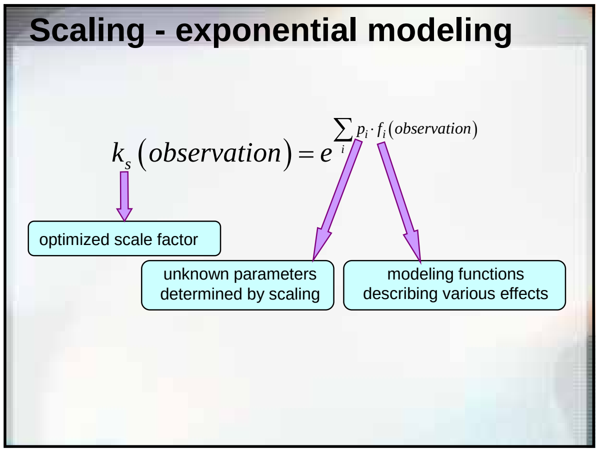
normalized average of neighbors
profile fitting radius p mm

optimal weighted fitting proper for strong and weak reflections

Scaling - definition

$$I(hkl) = \underbrace{I_b \cdot r_e^{\,2} \cdot \frac{\lambda^2}{\left|\mathbf{S} \times \boldsymbol{\omega}'\right|} \cdot P \cdot T \cdot \frac{V}{v_u^{\,2}} \cdot D_A \cdot D_S}_{\text{scale factor } K} \cdot \left|\mathbf{F}(hkl)\right|^2$$

$$K = \underbrace{k_{overall}}_{\text{From comparison of data to atomic model}} \cdot \Big(\underbrace{k_{Lorentz} \cdot k_{polarization} \cdot k_{detector}}_{\text{From calibration and diffraction geometry}} \cdot \underbrace{k_{absorption} \cdot \ldots}_{\text{From comparison of symmetry related reflections SCALING}}\Big)$$



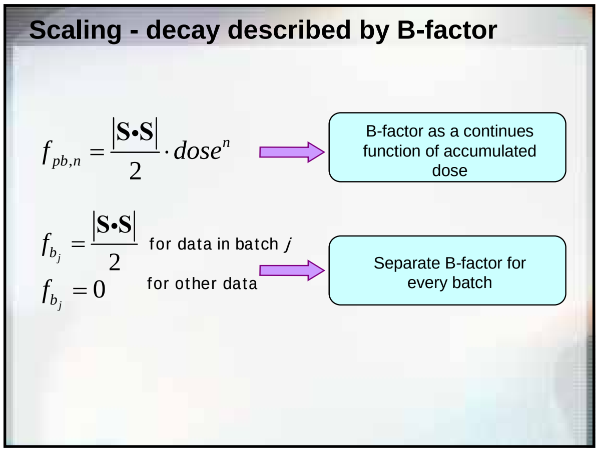
Scaling - exponential modeling

$$k_s\left(observation\right) = e^{\sum_i p_i \cdot f_i\left(observation\right)}$$

optimized scale factor

unknown parameters determined by scaling

modeling functions describing various effects



Scaling - decay described by B-factor

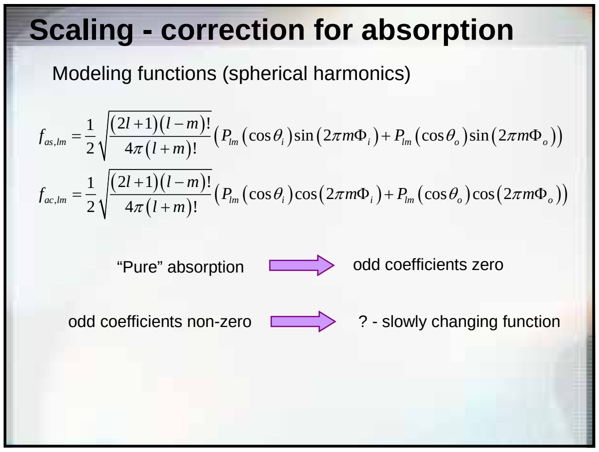
$$f_{pb,n} = \frac{\left|\mathbf{S}\bullet\mathbf{S}\right|}{2} \cdot dose^n$$

B-factor as a continues function of accumulated dose

$$f_{b_j} = \frac{\left|\mathbf{S}\bullet\mathbf{S}\right|}{2}$$ for data in batch *j*

$$f_{b_j} = 0$$ for other data

Separate B-factor for every batch



Scaling - correction for absorption

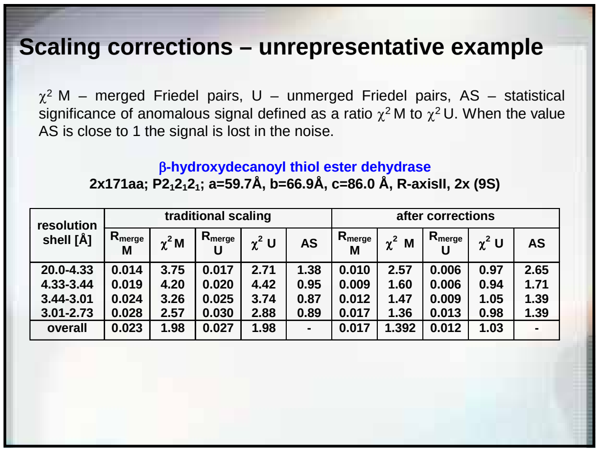
Modeling functions (spherical harmonics)

$$f_{as,lm} = \frac{1}{2}\sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}\left(P_{lm}\left(\cos\theta_i\right)\sin\left(2\pi m\Phi_i\right) + P_{lm}\left(\cos\theta_o\right)\sin\left(2\pi m\Phi_o\right)\right)$$

$$f_{ac,lm} = \frac{1}{2}\sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}\left(P_{lm}\left(\cos\theta_i\right)\cos\left(2\pi m\Phi_i\right) + P_{lm}\left(\cos\theta_o\right)\cos\left(2\pi m\Phi_o\right)\right)$$

"Pure" absorption ⇨ odd coefficients zero

odd coefficients non-zero ⇨ ? - slowly changing function

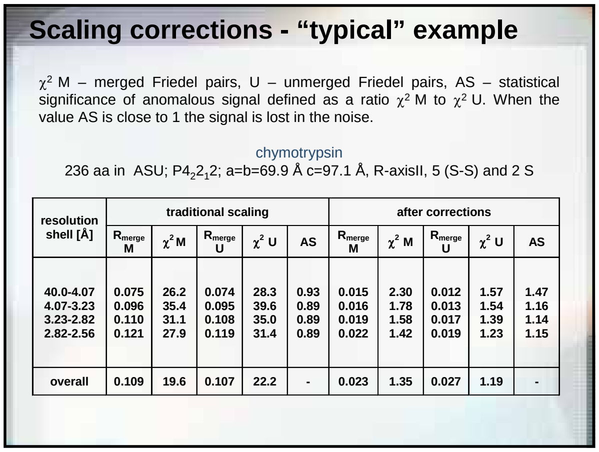


Scaling corrections – unrepresentative example

χ^2 M – merged Friedel pairs, U – unmerged Friedel pairs, AS – statistical significance of anomalous signal defined as a ratio χ^2 M to χ^2 U. When the value AS is close to 1 the signal is lost in the noise.

β-hydroxydecanoyl thiol ester dehydrase
2x171aa; P$2_12_12_1$; a=59.7Å, b=66.9Å, c=86.0 Å, R-axisII, 2x (9S)

resolution shell [Å]	traditional scaling					after corrections				
	R_{merge} M	χ^2 M	R_{merge} U	χ^2 U	AS	R_{merge} M	χ^2 M	R_{merge} U	χ^2 U	AS
20.0-4.33	0.014	3.75	0.017	2.71	1.38	0.010	2.57	0.006	0.97	2.65
4.33-3.44	0.019	4.20	0.020	4.42	0.95	0.009	1.60	0.006	0.94	1.71
3.44-3.01	0.024	3.26	0.025	3.74	0.87	0.012	1.47	0.009	1.05	1.39
3.01-2.73	0.028	2.57	0.030	2.88	0.89	0.017	1.36	0.013	0.98	1.39
overall	0.023	1.98	0.027	1.98	-	0.017	1.392	0.012	1.03	-

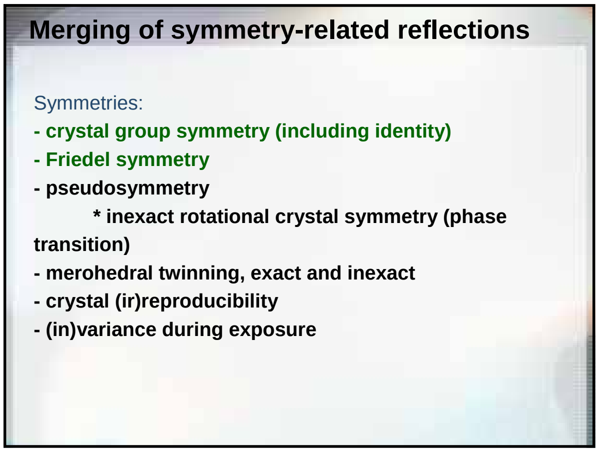


Scaling corrections - "typical" example

χ^2 M – merged Friedel pairs, U – unmerged Friedel pairs, AS – statistical significance of anomalous signal defined as a ratio χ^2 M to χ^2 U. When the value AS is close to 1 the signal is lost in the noise.

chymotrypsin
236 aa in ASU; P4_22_12; a=b=69.9 Å c=97.1 Å, R-axisII, 5 (S-S) and 2 S

resolution shell [Å]	traditional scaling					after corrections				
	R_{merge} M	χ^2 M	R_{merge} U	χ^2 U	AS	R_{merge} M	χ^2 M	R_{merge} U	χ^2 U	AS
40.0-4.07	0.075	26.2	0.074	28.3	0.93	0.015	2.30	0.012	1.57	1.47
4.07-3.23	0.096	35.4	0.095	39.6	0.89	0.016	1.78	0.013	1.54	1.16
3.23-2.82	0.110	31.1	0.108	35.0	0.89	0.019	1.58	0.017	1.39	1.14
2.82-2.56	0.121	27.9	0.119	31.4	0.89	0.022	1.42	0.019	1.23	1.15
overall	0.109	19.6	0.107	22.2	-	0.023	1.35	0.027	1.19	-



Merging of symmetry-related reflections

Symmetries:

- **crystal group symmetry (including identity)**
- **Friedel symmetry**
- **pseudosymmetry**
  - *** inexact rotational crystal symmetry (phase transition)**
- **merohedral twinning, exact and inexact**
- **crystal (ir)reproducibility**
- **(in)variance during exposure**

Merging - analysis

1. Determination of point group symmetry

- metric pseudosymmetries and relative indexing of different crystals

2. Parameters of error model (error scale factor, error systematic, rejection probability)

3. Assessment of data quality

- random events (signal-to-noise ratio)
- non-random events (outliers, ice-rings, bad frames etc.)
- non-isomorphism (radiation damage, pseudosymmetry)

4. Assessment of data content (significance of anomalous signal, systematic absences, translational pseudosymmetry - pseudosystematic absences)

Symmetry determination

Scaling (1 cycle) in all subgroups

Statistical analysis:

- **Increase of χ^2**
- **How many reflections for higher symmetry contributed to the χ^2**

Data collection – where to look?

- I/σ(I)
- R-merge
- χ^2 statistic
- Error model
- Detector area
- Phasing signal

I/σ(I), R-merge, % of reflections measured with I/σ(I) > 3

I/σ(I) is weighted statistics

- 2 quite reasonable limit

R-merge is unweighted statistics

- make no sense to calculate it for whole data set

- in resolution shells it gives valuable information, particularly at low resolution

χ^2 statistics

Squared ratio of differences between equivalent measurements divided by expected errors

$$\frac{\left(I^{1}_{hkl} - \langle I_{hkl} \rangle\right)^2}{\sigma^2_{I_{hkl}} + \sigma^2_{\langle I_{hkl} \rangle}}$$

Expected value is around 1 for reasonable model of errors, however some departures are acceptable

0.9	-	5% overestimated errors
1.05	-	2.5% underestimated errors
1.1	-	5% underestimated errors
1.5	-	22% underestimated errors
2.0	-	40% underestimated errors

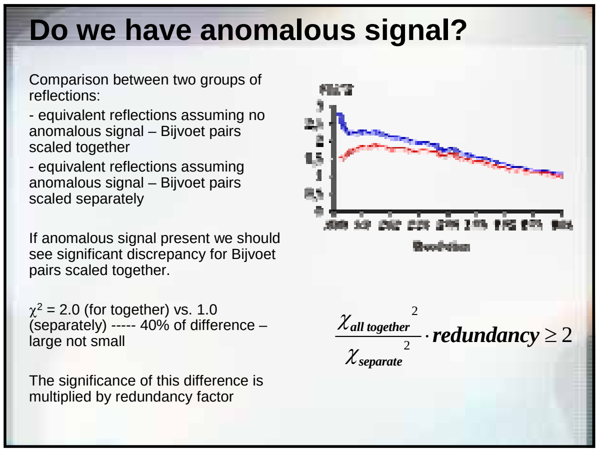
Error model

Based on the χ^2 test we can change the error model:
In HKL2000:

- error model (default value = 0.03)
  - change in resolution shells – be careful
  - if you have to go over 0.10 – something bad happened in experiment
- scale factor
  - more impact at higher resolution
  - default value 1.3
  - if you have to go over 2.0:
    - increase error density value in Denzo
    - non-isomorphism – accept χ^2

Non-optimal error model can kill phasing

The problem of error estimates – consequences of non-optimality of error estimates grow as a square of this non-optimality.


3
2.5
2
1.5
1
0.5
0
Resolution

Do we have anomalous signal?

Comparison between two groups of reflections:

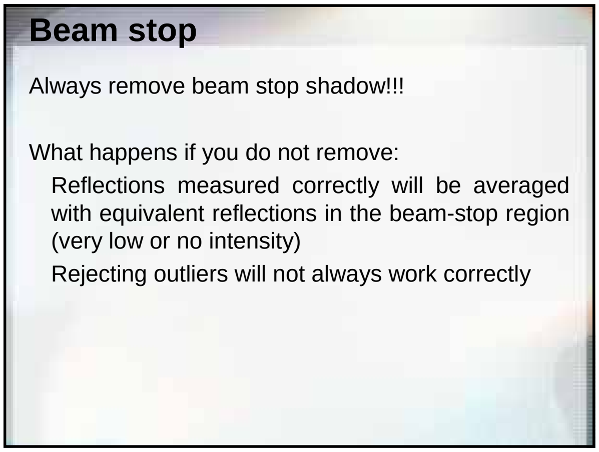
- equivalent reflections assuming no anomalous signal – Bijvoet pairs scaled together
- equivalent reflections assuming anomalous signal – Bijvoet pairs scaled separately

If anomalous signal present we should see significant discrepancy for Bijvoet pairs scaled together.

χ^2 = 2.0 (for together) vs. 1.0 (separately) ----- 40% of difference – large not small

$$\frac{\chi_{all\ together}{}^{2}}{\chi_{separate}{}^{2}} \cdot \boldsymbol{redundancy} \geq 2$$

The significance of this difference is multiplied by redundancy factor



Beam stop

Always remove beam stop shadow!!!

What happens if you do not remove:

Reflections measured correctly will be averaged with equivalent reflections in the beam-stop region (very low or no intensity)

Rejecting outliers will not always work correctly