Xray data collection and processing: a (short?) introduction

Laurent Maveyraud, Oléron 2018

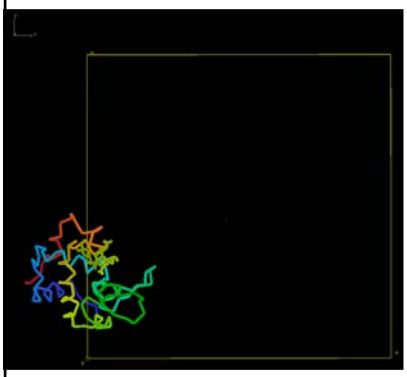
overview

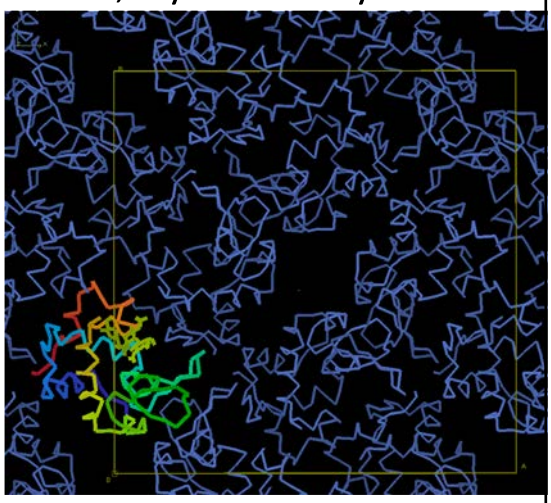
- Reminders about diffraction: structure factors, reciprocal lattice, Ewald's sphere
- Data collection: practical aspects
- Data processing: XDS, mosflm, assessing data quality
- "Claudine, Jean-Luc, how do we solve a structure with these data?"

overview

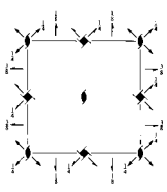
- **Reminders about diffraction: structure factors, reciprocal lattice, Ewald's sphere**
- Data collection: practical aspects
- Data processing: XDS, mosflm, assessing data quality
- "Claudine, Jean-Luc, how do we solve a structure with these data?"

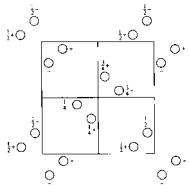
Protein crystals, symmetry





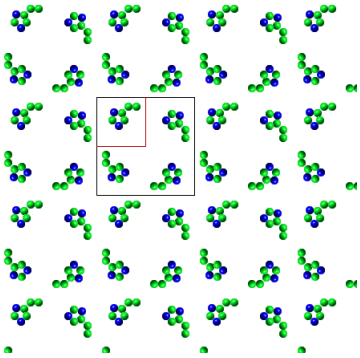
$P4_32_12$



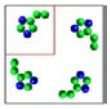


Further reading : Dauter, 2010, J. Appl. Cryst., 43:1150

Protein crystals and symmetry



crystal

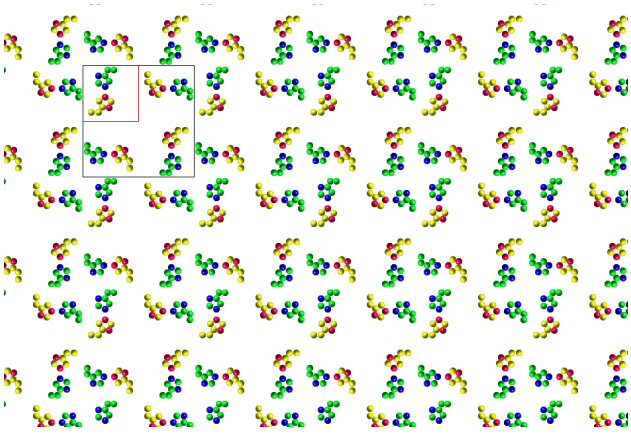


unit cell

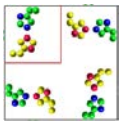
asymmetric unit

As proteins are chiral, only rotation and translations allowed in protein crystals: 65 possible space groups.
Symmetry results in equivalent positions.

Protein crystals and symmetry



crystal

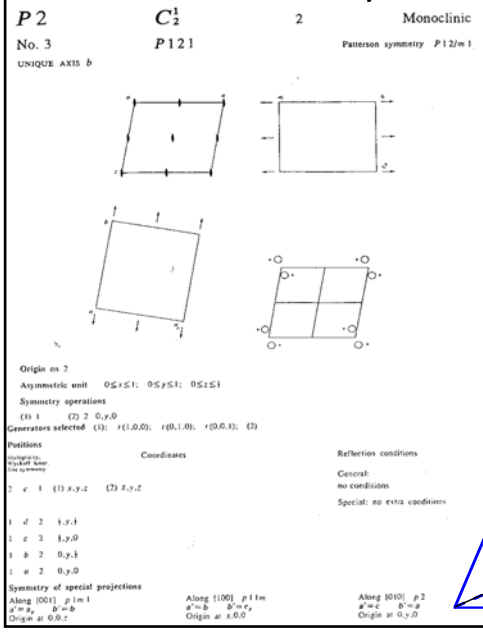


unit cell

asymmetric unit

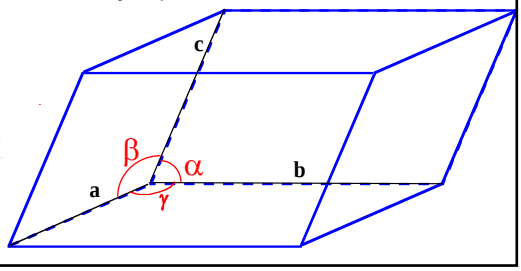
You can have more than one copy of the protein in the asymmetric unit (Non Crystallographic Symetry)

Protein crystals and symmetry

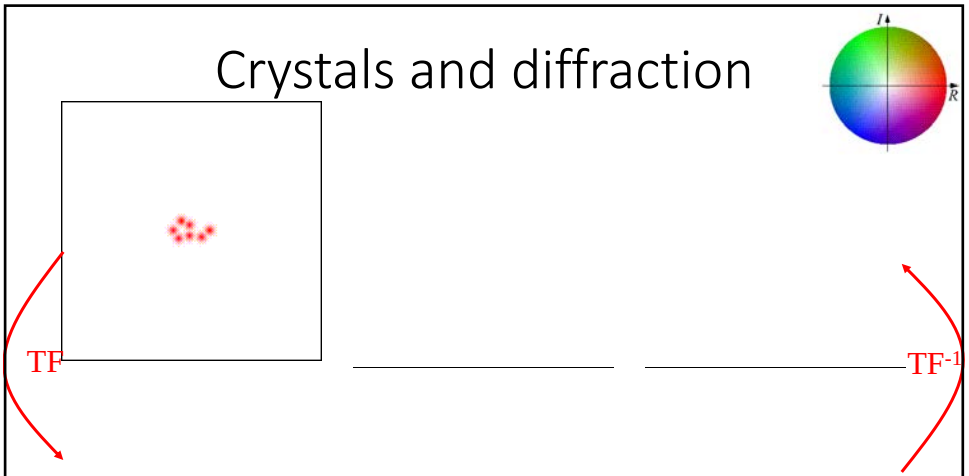


Symmetry operations in the crystals **impose** constraints on **cell parameters**.

cubic symmetry imposes a=b=c and $\alpha=\beta=\gamma=90°$



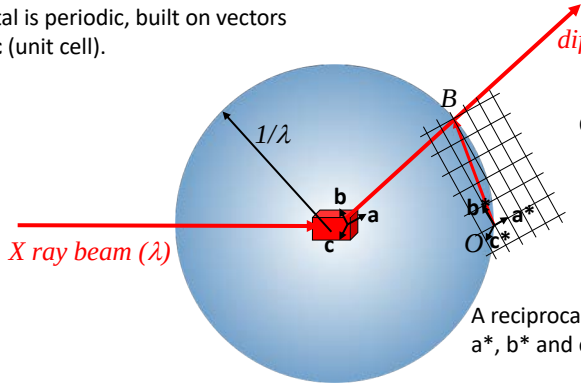
Crystals and diffraction



Spot position depends on cell parameters (**a**, **b** and **c**)
Spot intensity depends on the structure of the molecule

Crystals and diffraction: Ewald's sphere

The crystal is periodic, built on vectors **a**,**b** and **c** (unit cell).



OB = h**a*** + k**b*** + l**c***

A reciprocal lattice is built with vectors a*, b* and c* (reciprocal unit cell).

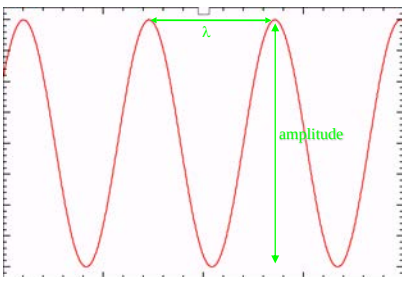
A wave is scattered when a node of the reciprocal lattice (indices h k l) touches the Ewald's sphere. The structure factor (amplitude F and phase ϕ) of the diffracted wave is :

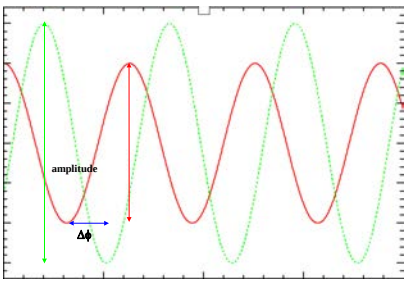
$$\mathbf{F}(hkl) = N_{cell} \cdot \Sigma f_j . \exp(-2\pi(hxj + kyj + lzj))$$

Collecting data: measure of the diffracted beams characteristics

Reminder: our goal is to determine a molecular structure, that is to determine the electron density in a unit cell in the crystal

$$\rho(x,y,z) = \frac{1}{V_{maille}} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} m_{hkl} |F_{hkl}| e^{i\phi_{hkl}} e^{-2i\pi(hx+ky+lz)}$$





Amplitudes are **easy**… phases are a **problem**…

overview

- Reminders about diffraction: structure factors, reciprocal lattice, Ewald's sphere
- **Data collection: practical aspects**
- Data processing: XDS, mosflm, assessing data quality
- "Claudine, Jean-Luc, how do we solve a structure with these data?"

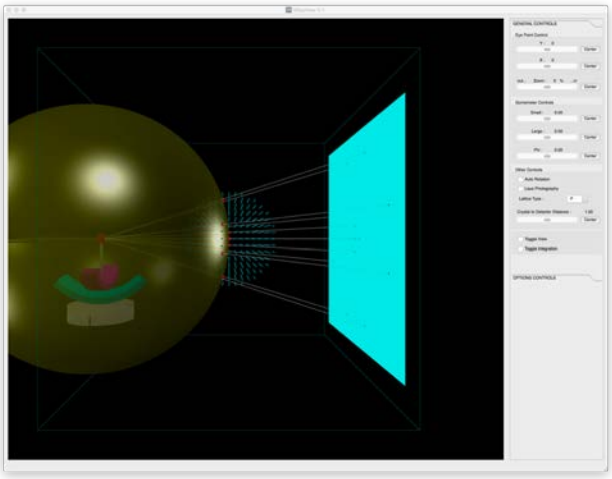
Data Collection

- One crystal – one structure: using the "standard" data collection strategy - the oscillation method
- A few crystals – one structure: the "*in situ*" situation
- A serie of crystals – one structure: serial crystallography

Data Collection

- One crystal – one structure: using the "standard" data collection strategy - the oscillation method
- A few crystals – one structure: the "*in situ*" situation
- A serie of crystals – one structure: serial crystallography

Collecting data



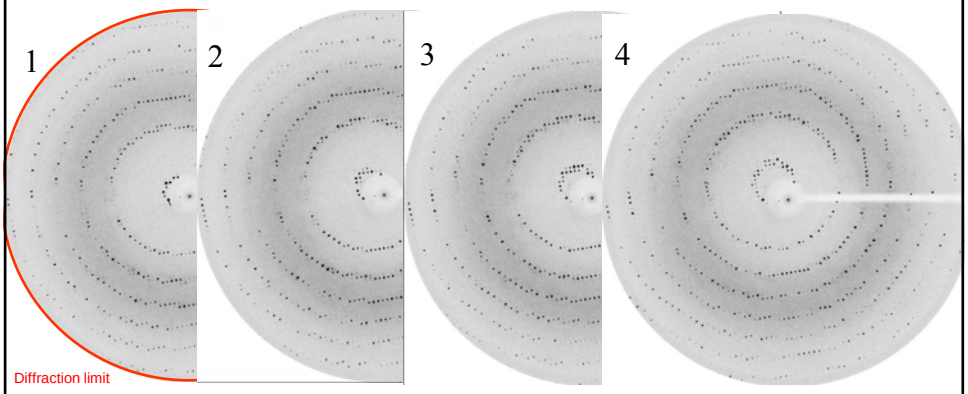
You want to be sure to collect **every** diffracted beam! That is, all nodes of the reciprocal lattice should hit the Ewald's sphere: rotate the crystal while exposing it to Xray… this is the **oscillation method**

Getting ready for data collection

- Xrays can fry your crystals: better cool them !

Further readings : Pfulgrath, 2015, Acta Cryst F, 71:622

Collecting data



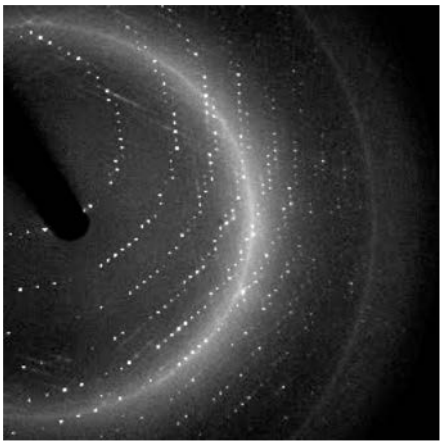
Collecting data: the oscillation method

- How many images to collect ?
  - Crystal symmetry, phasing method
- Which oscillation angle ?
  - Cell parameters, type of detector, type of processing
- Which crystal to detector distance ?
  - Resolution limit of the crystal, cell parameters
- Which exposure time ?
  - Type of detector, no saturated spots

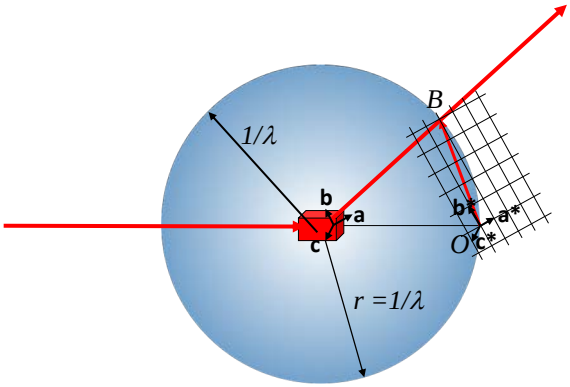
Further readings : Evans, 1999, Acta Cryst, D55:1771
Dauter, 1999, Acta Cryst, D55:1703

Collecting data: the oscillation method

With recent detectors (Pilatus) the crystal is rotated continuously (shutterless data collection).



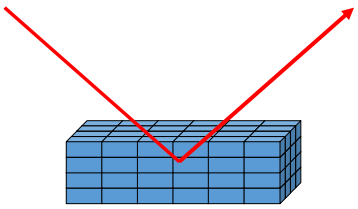
Collecting data: the oscillation method



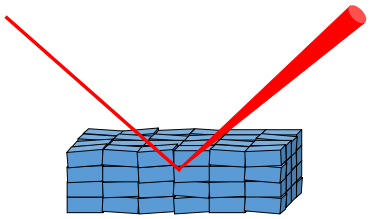
In theory :

- perfect crystal: reciprocal lattice is built of points
- perfect beam (no wavelength dispersion, no divergence…)

Collecting data: let's face reality



- perfect crystal: reciprocal lattice is built of points
- perfect beam (no wavelength dispersion, no divergence…)

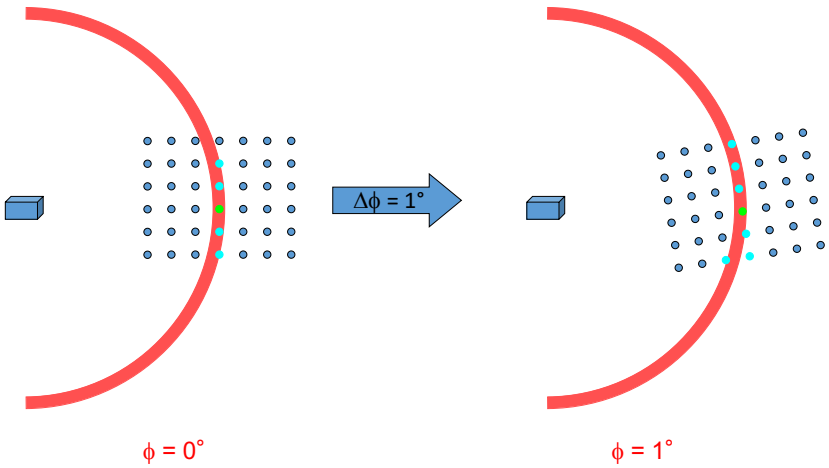


Real life:

- mosaic crystal
- real beam (wavelength dispersion, divergence…)

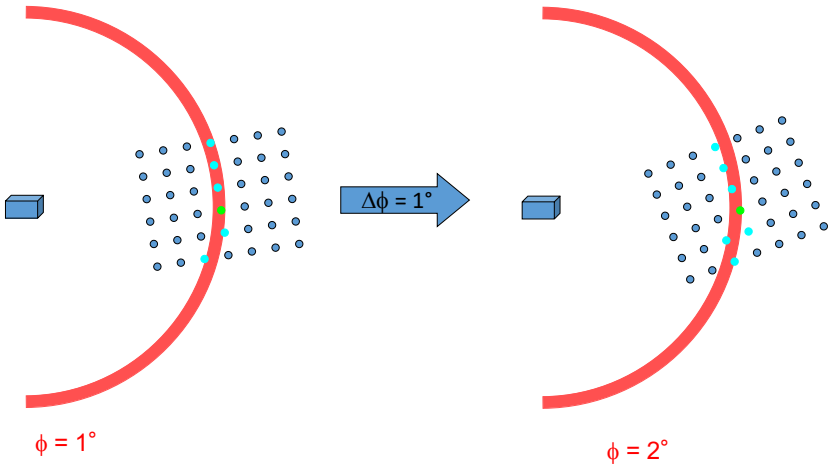
Collecting data: let's face reality

Consequences for the Ewald's construction



Collecting data: let's face reality

Consequences for the Ewald's constuction



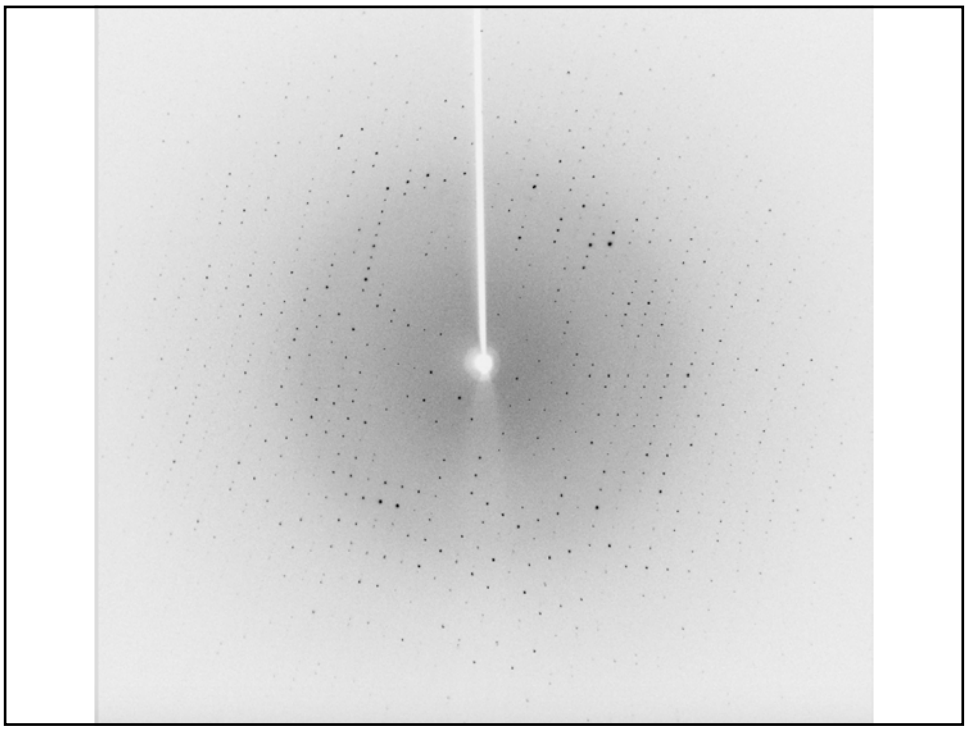
overview

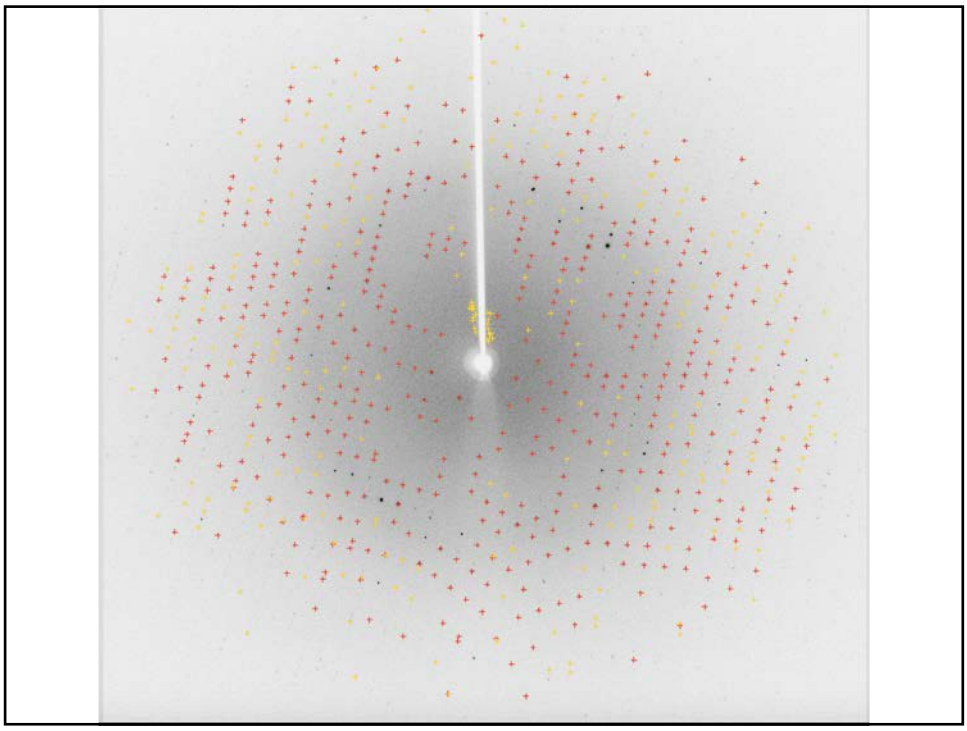
- Reminders about diffraction: structure factors, reciprocal lattice, Ewald's sphere
- **Data collection: practical aspects**
- Data processing: XDS, mosflm, assessing data quality
- "Claudine, Jean-Luc, how do we solve a structure with these data?"

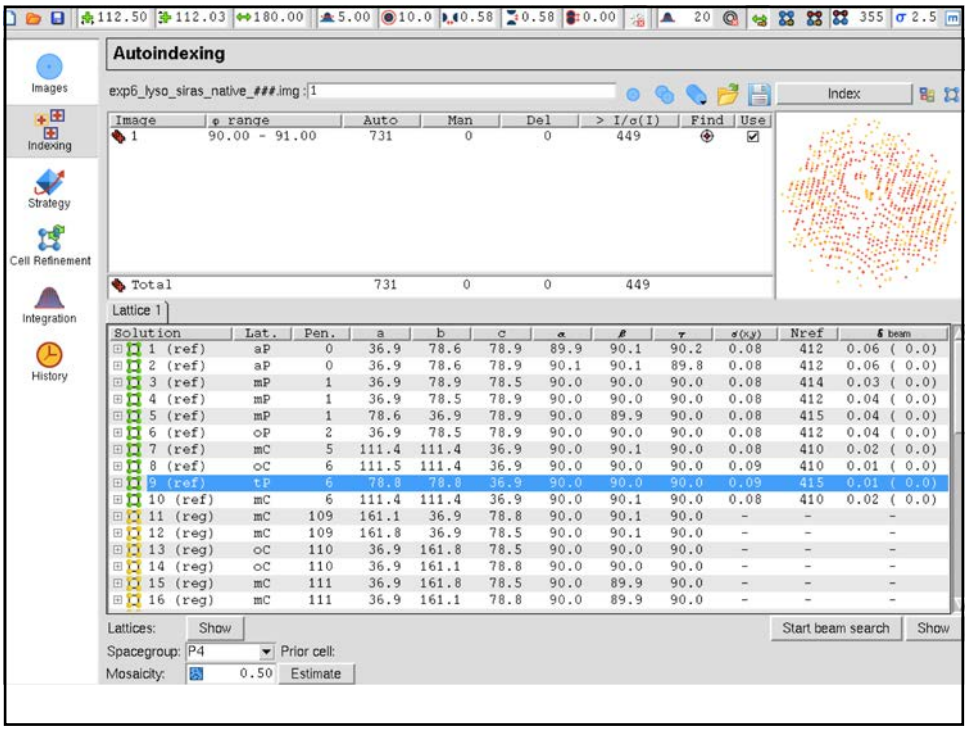
Processing data: XDS, iMosflm

Three steps for data processing :

- Indexing data: find possible cell parameters, crystal orientation, guestimate symmetry
  - For each diffraction spot, you know Miller indices
  - Symmetry derived from cell parameters: it's only a hypothesis !!!!



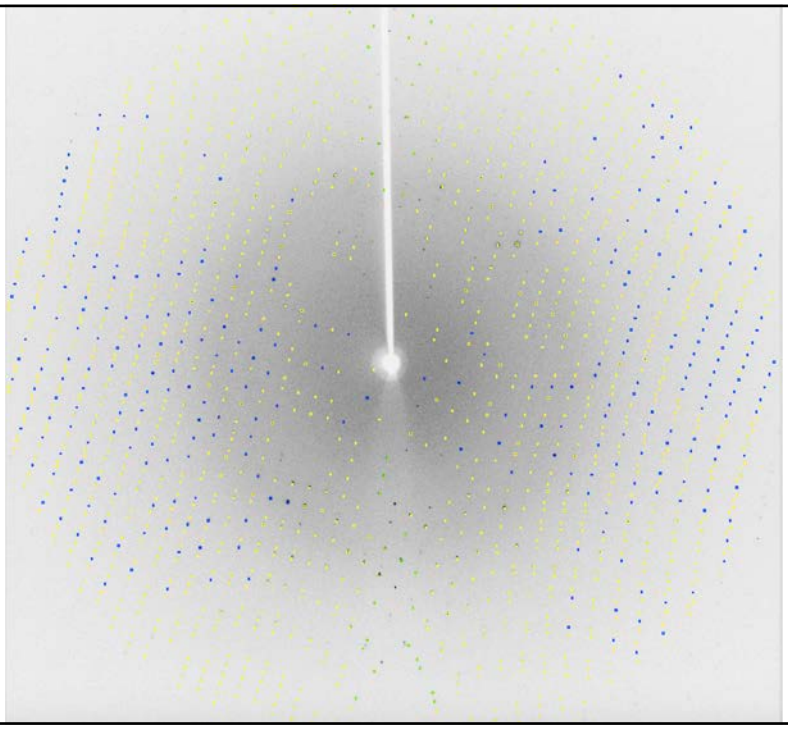




Processing data: XDS, iMosflm

Three steps for data processing :

- Indexing data: find possible cell parameters, crystal orientation, **guestimate** symmetry
  - For each diffraction spot, you know Miller indices
  - Symmetry derived from cell parameters: it's only a hypothesis !!!!
    If the cells seems to obey to some symmetry constraints, it's likely because this symmetry is present in the crystal.
- Now that we have a unit cell and an orientation, we can predict spot position on any frames



Processing data: XDS, iMosflm

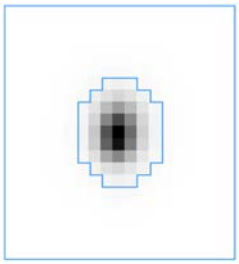
Three steps for data processing :

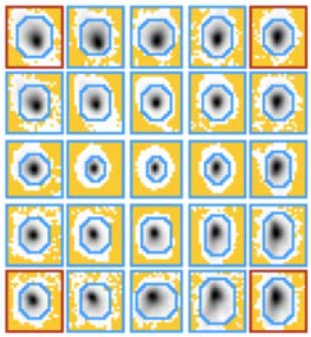
- Indexing data: find possible cell parameters, crystal orientation, guestimate symmetry
  - For each diffraction spot, you know Miller indices
  - Symmetry derived from cell parameters: it's only a hypothesis !!!!
- Integration: for each spot on each frames, measure the intensity
  - Locate spot, assign pixel to « background » or to « spot »
  - Sum the intensity for « spot » pixels
  - Profile fitting (2D iMosflm, 3D XDS)

Further readings: Rossman 1999, Aca Cryst, D55:1631
Leslie, 1999, Acta Crsyt D55:1696

Processing data: iMosflm

2D profile fitting: wide slicing

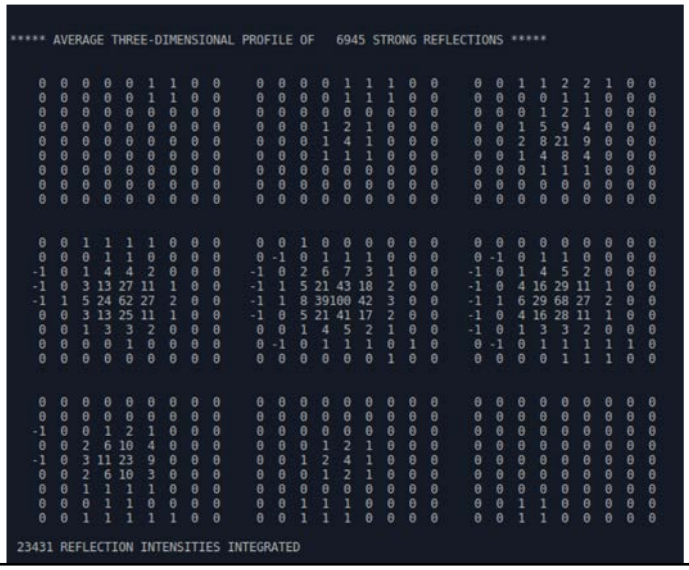




Detector surface splitted in 9 or 25 regions.
Profiles are learned for intense well defined spots.

Processing data: XDS

3D profile fitting: fine slicing

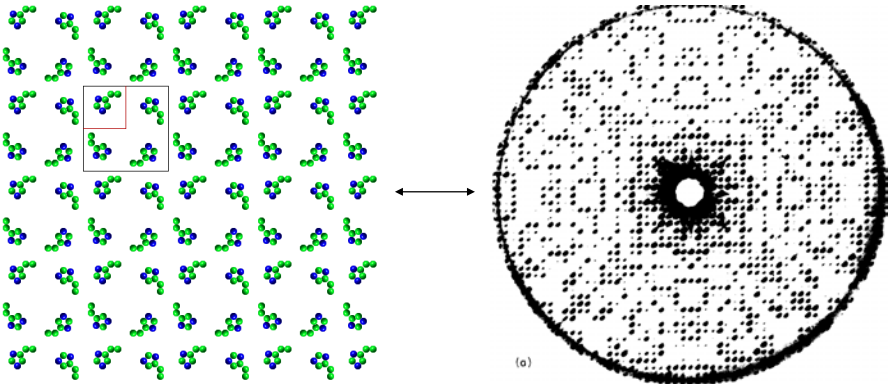


Processing data: XDS, iMosflm

Three steps for data processing :

- Indexing data: find possible cell parameters, crystal orientation, guestimate symmetry
  - For each diffraction spot, you know Miller indices
  - Symmetry derived from cell parameters: it's only a hypothesis !!!!
- Integration: for each spot on each frames, measure the intensity
  - Locate spot, assign pixel to « background » or to « spot »
  - Sum the intensity for « spot » pixels
  - Profile fitting (2D iMosflm, 3D XDS)
- Scaling of data: correct for variation in diffracting volume, beam intensity variations,…
  - Use equivalent reflections to place all images: uses the symmetry of the crystal!

Symmetry of reciprocal space



Crystal symmetry: equivalent positions

x, y, z
y, -x, z
-x, -y, z
-y, x, z

Friedel's law

Symmetry of diffracted intensities: equivalent reflections

h, k ,l	-h, -k, -l
k, -h, l	-k, h, -l
-h, -k, l	h, k, -l
-k, h l	k, -h, -l

Processing data: XDS, iMosflm

Three steps for data processing :

- Indexing data: find possible cell parameters, crystal orientation, guestimate symmetry
  - For each diffraction spot, you know Miller indices
  - Symmetry derived from cell parameters: it's only a hypothesis !!!!
- Integration: for each spot on each frames, measure the intensity
  - Locate spot, assign pixel to « background » or to « spot »
  - Sum the intensity for « spot » pixels
  - Profile fitting (2D iMosflm, 3D XDS)
- Scaling/merging of data:
  - Scaling: correct for variation in diffracting volume, beam intensity variation,. Use the symmetry of the crystal (validate, or not, the symmetry hypothesis from the indexing step)
  - Merging: average different observations of equivalent reflections, compute data processing statistics

Checking the quality of your data

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)	Anomal Corr
5.35	6059	778	779	99.9%	2.1%	2.7%	6059	67.08	2.3%	100.0*	54*
3.80	10814	1395	1395	100.0%	2.7%	2.7%	10814	67.86	2.9%	99.9*	26*
3.11	13860	1797	1797	100.0%	2.9%	2.8%	13860	63.55	3.1%	99.9*	16*
2.69	16578	2139	2139	100.0%	3.4%	3.4%	16578	49.96	3.7%	99.9*	6
2.41	18603	2406	2406	100.0%	4.2%	4.1%	18603	42.43	4.5%	99.9*	5
2.20	20632	2675	2675	100.0%	4.9%	4.9%	20632	35.82	5.2%	99.9*	8
2.04	22300	2899	2899	100.0%	6.0%	6.1%	22300	29.20	6.4%	99.8*	2
1.91	23848	3113	3113	100.0%	8.4%	8.7%	23848	21.33	9.0%	99.7*	5
1.80	24479	3304	3312	99.8%	12.2%	13.0%	24467	14.55	13.1%	99.4*	1
total	157173	20506	20515	100.0%	3.9%	3.9%	157161	37.30	4.2%	99.9*	7

Completeness: which proportion of the possible diffracted beams did we collect?
Rsym, Rmerge: disagreement between all observations of a reflection (and equivalent)
I/sigma: signal to noise ratio
Rmeas: multiplicity corrected Rsym
CC(1/2): half datasets correlation coefficient
Anomal Corr

Checking the quality of your data

Table 1

	55.70 – 1.80 Å	1.84 – 1.80 Å
N observations	156,728	8,565
N unique	11,204	646
Multiplicity	14.0	13.3
Completeness (%)	100.0	100.0
Rsym or Rmerge	0.053	0.145
I/σ	34.8	15.2

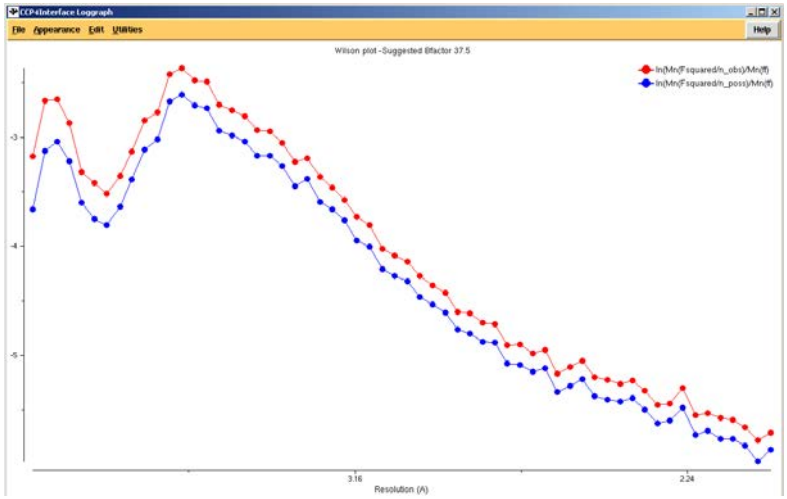
Is Rsym/Rmerge a good indicator of data quality?

Checking the quality of your data

	55.70 – 1.80 Å	1.84 – 1.80 Å
N observations	156,728	8,565
N unique	11,204	646
Multiplicity	14.0	13.3
Completeness (%)	100.0	100.0
Rsym or Rmerge	0.053	0.145
Rmeas	0.057	0.155
CC1/2	0.999	0.995
I/σ	34.8	15.2

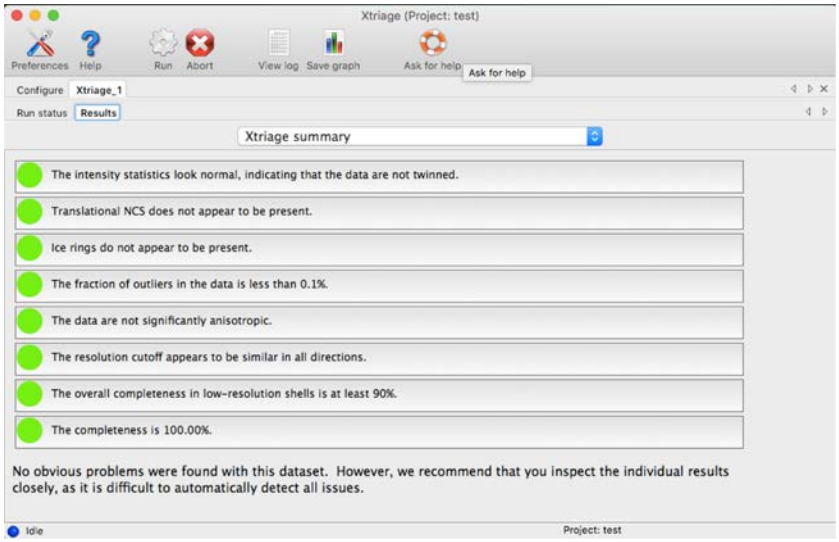
Checking the quality of your data

Wilson Plot



Crystal/dataset pathologies

XTRIAGE analysis (Phenix)



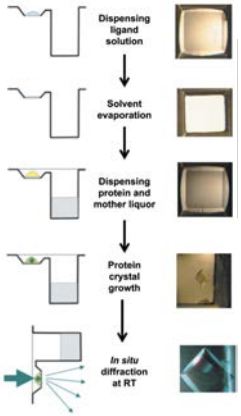
Data Collection

- One crystal – one structure: using the "standard" data collection strategy - the oscillation method
- **A few crystals – one structure: the "*in situ*" situation**
- A serie of crystals – one structure: serial crystallography

- If crystals are mechanically too fragile to be cooled, mounted
- For ligand screening

In situ crystallography : application to ligand screening

- Diffraction data are collected directly from a crystal in its crysallisation drop
- Protein is crystallized in the presence of dry fragment (Gelin, 2015, Acta Cryst D)



McPherson, 2000, J. Appl. Cryst.
Gelin, 2015, Acta Cryst D

In situ crystallography : application to ligand screening

377 datasets collected (75 to 80° oscillation range)

	1	2	3	4	5	6	7	8	9	10	11	12
A	3	1	2	2	2	3		2	3	3	4	
	3	3		3	3	1	2	3	2	3	2	1
B	2	3	2	2	3	3	3	3	3		3	2
	3	2	3		3	3	3	2	2	4	1	2
C	1	3	3	3		2	2	3	4	2	1	3
	4	3	3	2			2	3		1	1	1
D	3	3	2	2		2	3	6	2	3	3	
	2	3	1	3		2	1	4	2			
E	3	3	2		2	1	1		2	2	2	3
	1	1	3		1	1	2	1	2	1		
F	4		3	2	2	4	4	4	3			3
	2		3	3	1		2	3	2	3	2	1
G	4		3	3	3	4	2	3			1	3
	1	2	1		2	2		1				3
H	3	3	1	4		5			4	1	2	3
		4	2	1	2	4	4	2	2	3	3	1

Plateforme Intégré de Criblage de Toulouse

In situ crystallography : application to ligand screening

Different datasets for a given conditions should be **merged** !!

And this is done **manually**… !

Statistics for 285 unique datasets

High resolution limit: 2.33 Å
(53 datasets between 2.0 and 2.1 Å)
Rsym : 0.32
Completeness : 86 %
CC(1/2) : 87 %
Multiplicity : 3.5

Statistics for 146 merged datasets

High resolution limit: 2.13 Å

Rsym : 0.213
Completeness : 92.3 %
CC(1/2) : 97.3 %
Multiplicity : 5.25

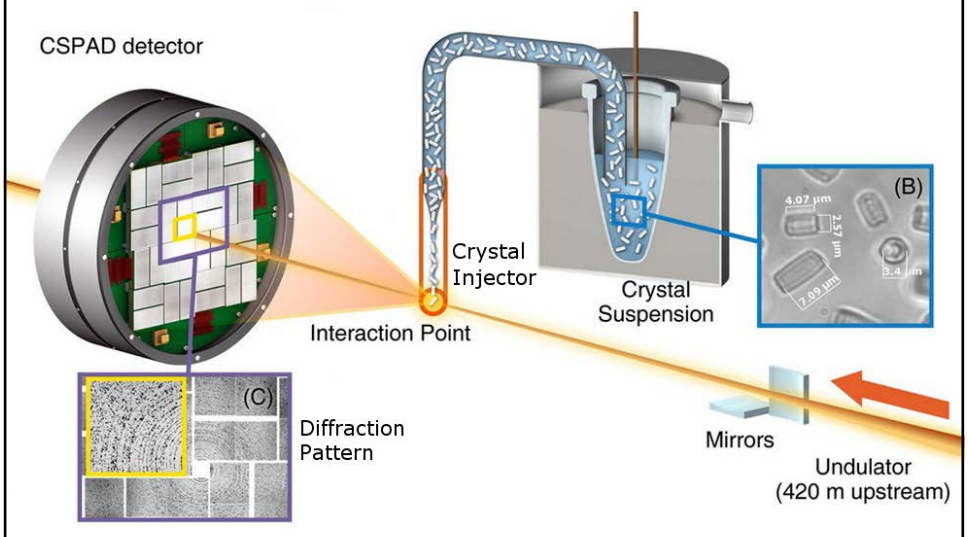
Plateforme Intégré de Criblage de Toulouse

Data Collection

- One crystal – one structure: using the "standard" data collection strategy - the oscillation method
- A few crystals – one structure: the "*in situ*" situation
- **A serie of crystals – one structure: serial crystallography**

- Many, many, many tiny crystals: just flow them!
- Data collected at room temperature

Serial crystallography: one crystal = one image



Serial crystallography: one crystal = one image

Data set	Lysozyme	$A_{2A}R$ SMX	$A_{2A}R$ Cryo	$A_{2A}R$ 19 h 6 keV	$A_{2A}R$ SFX	$TD1_{Apo}$	$TD1_{Col}$	MOSTO
X-ray energy (keV)	12.4	12.4	12.4	6.0	9.5	12.4	12.4	12.4
Measurement time (h)	0.3	6.6	~4	19.4	0.36	7.7	8.6	5.9
Nozzle size	50	50	–	50	50	50	75	75/100
Beam size (µm)	5×5	20×5	20×5	15×5	1×1	20×5/10×5	10×5	20×5/40×5
Flux (ph/s)	3.9×10^{11}	1.5×10^{12}	1.5×10^{11}	4×10^{11}	2.4×10^{11} [a]	$1.5\times10^{12}/0.7\times10^{12}$	0.7×10^{12}	1.5×10^{12}
Frame rate	50 Hz	50 Hz	10 Hz	50 Hz	120 Hz[b]	50 Hz	50 Hz	50 Hz
Crystal size (µm³)	15×10×10	30×30×5	30×30×5	30×30×5	30×30×5	15×10×10	15×10×10	50×20×20
Collected images	58,000	1,180,705	3500	3,496,230	155,241	1,388,078	1,544,487	1,054,366
Crystals used	–	–	6	–	–	–	–	–
Indexed patterns	27,000	128,086	3500	186,688	3563	6,6271	6,2245	68,788
Patterns indexed (%)	46.5	10.8	100	5.3	2.3	4.8	4.0	6.5
Resolution	24.84–1.50	25.2–2.14	50.0–1.95	34.0–2.67	20.2–1.70	36.1–2.13	35.0–2.05	24.2–1.70
Number of reflections	9,140,532	31,065,416	306,759	58,802,388	1,325,959	26,000,036	12,155,884	84,306,711
Number of unique reflections	20,181	30,837	32,392	30,230	56,793	65,679	62,424	110,141
Multiplicity	452.9 (4.6)	1007.4 (8.1)	9.5 (2.0)	1945.2 (684.7)	23.3 (3.0)	395.9 (21.8)	194.5 (6.2)	765.4 (113.8)
Completeness	94.7 (49.23)	99.4 (93.83)	87.6 (61.7)	100 (100)	93.7 (53.6)	100 (100)	91.9 (49.5)	100 (100)
I/ sigma	8.35 (0.72)	13.17 (0.70)	10.36 (1.13)	24.76 (3.57)	2.93 (0.44)	6.56 (0.74)	5.19 (0.52)	5.29 (0.32)
CC*	0.99 (0.53)	0.99 (0.47)	0.99 (0.79)	0.99 (0.35)	0.99 (0.45)	0.99 (0.56)	0.99 (0.82)	0.99 (0.83)
CC1/2	0.996 (0.17)	0.99 (0.12)	0.99 (0.46)	0.99 (0.07)	0.97 (0.11)	0.99 (0.19)	0.99 (0.51)	0.99 (0.53)

h	k	l	F	SIGF	DANO	SIGDANO	F(+)	SIGF(+)	F(-)	SIGF(-)
0	0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0	2	-1.00	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00
0	0	3	-1.00	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00
0	0	4	101.12	6.29	0.00	0.00	100.92	9.00	100.05	9.11
0	0	5	5087.18	868.91	5087.18	868.91	5087.18	868.91	5004.75	871.44
0	0	6	-1.00	868.91	-1.00	868.91	-1.00	868.91	5004.75	871.44
0	0	7	-1.00	868.91	-1.00	868.91	-1.00	868.91	5004.75	871.44
0	0	8	712.77	26.26	0.00	0.00	713.90	35.18	706.38	40.04
0	0	9	251303.12	24365.59	251303.12	24365.59	251303.12	24365.59	246856.75	27390.66
0	0	10	-1.00	24365.59	-1.00	24365.59	-1.00	24365.59	246856.75	27390.66
0	0	11	-1.00	24365.59	-1.00	24365.59	-1.00	24365.59	246856.75	27390.66
0	0	12	374.42	11.63	0.00	0.00	377.39	14.45	367.19	19.85
.....										
.....										
36	20	1	239.06	4.01	-32.37	8.15	221.41	6.19	253.78	5.30

resolution limit

What can we do with these data ?

Stéphane… tell us about phases