



De novo structure inference of fibrillar proteins using X-ray fiber diffraction

Wojtek Potrzebowski

Ingemar André Lab

28.07.2013



Fiber diffraction origins

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

Performance

Rfree

Conclusions





LUND
UNIVERSITY

Fiber diffraction origins

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

Performance

*R*free

Conclusions



1952:
Watson and Crick
propose B-DNA
structure

Fiber diffraction origins

Fiber
Diffraction
basics

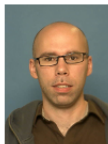
FD data and
Rosetta

Benchmark

Performance

Rfree

Conclusions



2011:
Ingemar
started
working on it



1952:
Watson and Crick
propose B-DNA
structure

Fiber diffraction origins

Fiber
Diffraction
basics

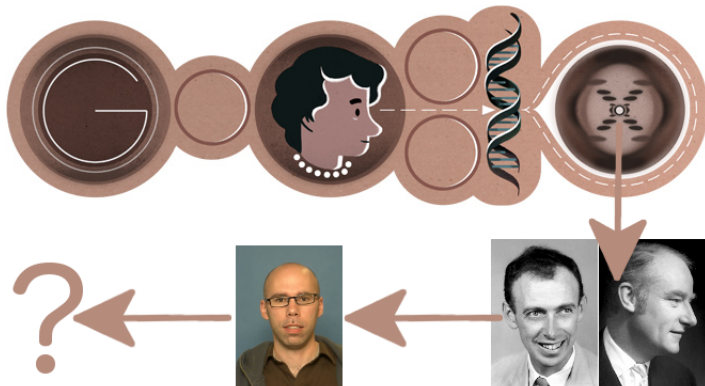
FD data and
Rosetta

Benchmark

Performance

*R*free

Conclusions



2011:
Ingemar
started
working on it

1952:
Watson and Crick
propose B-DNA
structure

Fiber Diffraction experiment setup

Fiber
Diffraction
basics

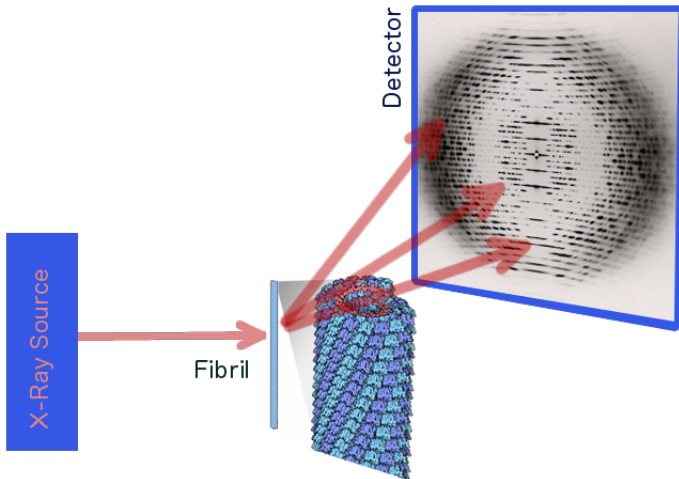
FD data and
Rosetta

Benchmark

Performance

*R*free

Conclusions



Continuous helix

Layer lines arise from repeats along the fiber axis

Fiber
Diffraction
basics

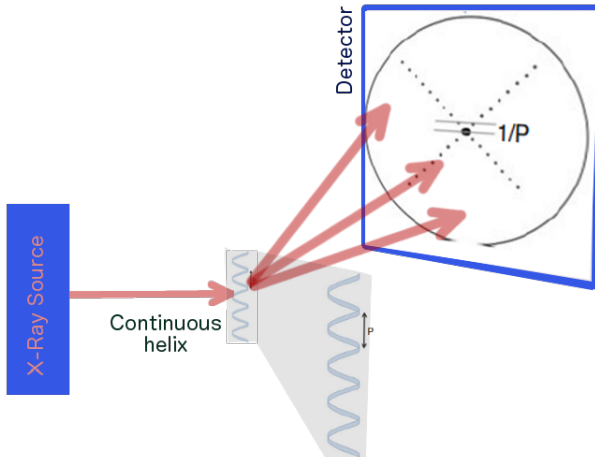
FD data and
Rosetta

Benchmark

Performance

*R*free

Conclusions



Discontinuous helix

Diffraction in vertical and horizontal directions

Fiber
Diffraction
basics

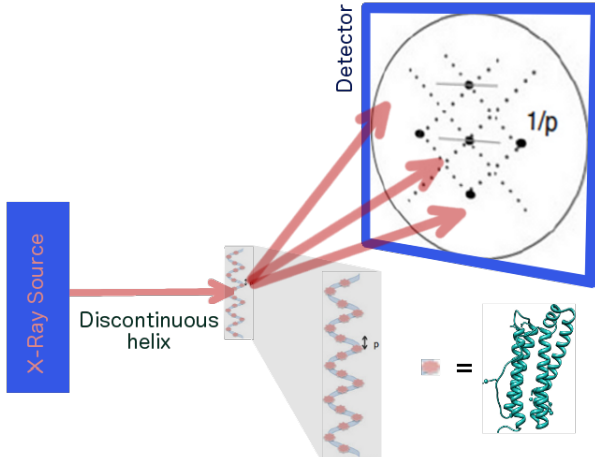
FD data and
Rosetta

Benchmark

Performance

*R*_{free}

Conclusions

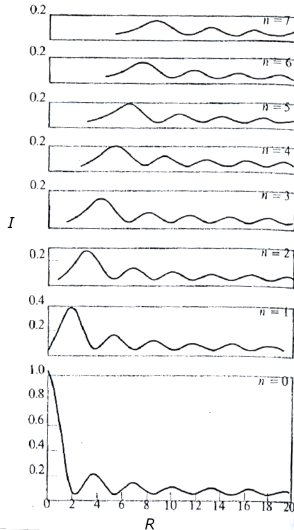


We can infer helical parameters from diffraction pattern.



Discontinuous helix

Intensity along the layer lines



Intensity along the layer line:

- is a continuous function
- reflects regularly repeating molecules on the helix



LUND
UNIVERSITY

Real-life fiber diffraction experiment

Bundle of aligned fibrils

Fiber
Diffraction
basics

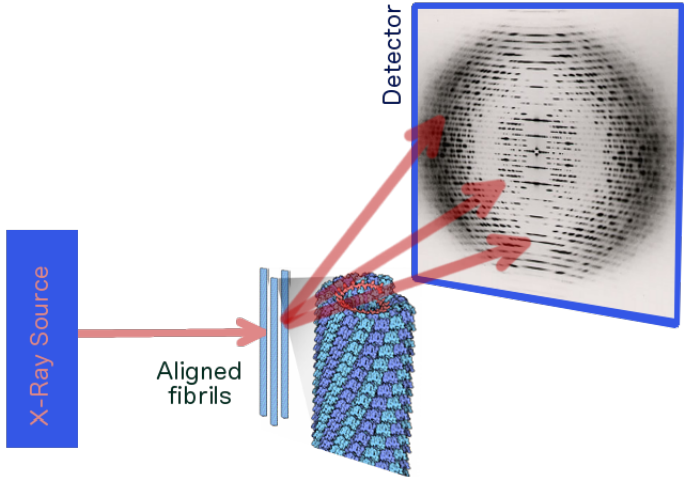
FD data and
Rosetta

Benchmark

Performance

*R*free

Conclusions



Real-life fiber diffraction experiment

Bundle of aligned fibrils - top view

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

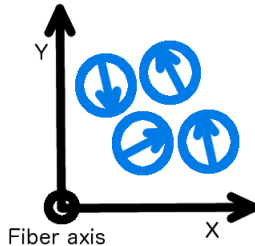
Performance

Rfree

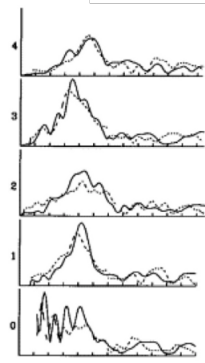
Conclusions



Side view



Top view



Layer lines

Randomly oriented fibrils in XY plane lower resolution!

Real-life fiber diffraction experiment

Misaligned fibrils

Fiber
Diffraction
basics

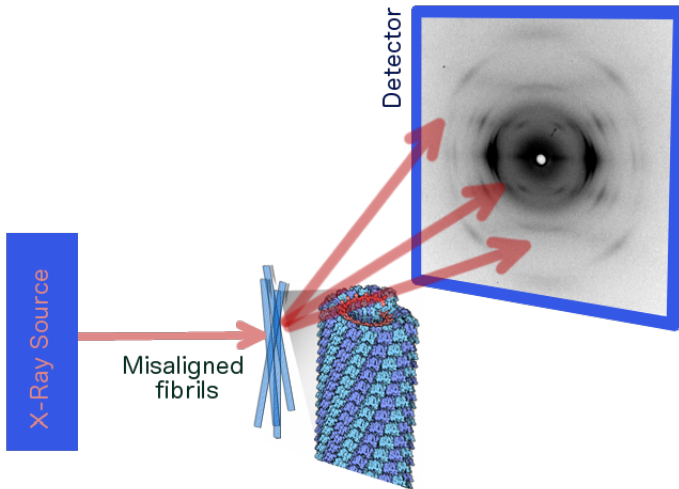
FD data and
Rosetta

Benchmark

Performance

*R*_{free}

Conclusions



Fiber Diffraction provides 2D information

Fiber
Diffraction
basics

FD data and
Rosetta

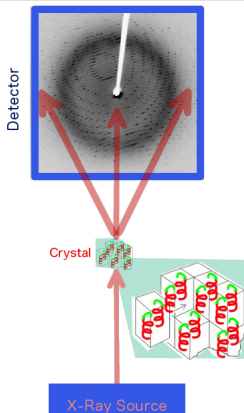
Benchmark

Performance

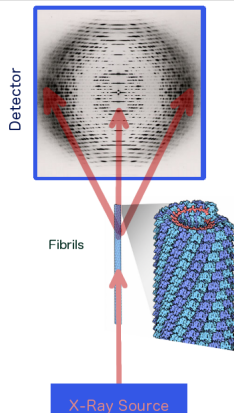
Rfree

Conclusions

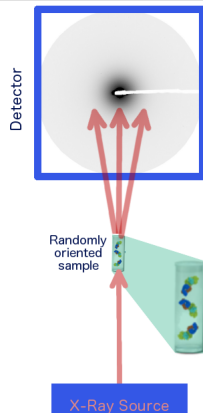
Xtallography (3D)



Fiber Diff. (2D)



SAXS (1D)





Limitations

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

Performance

Rfree

Conclusions

Fiber diffraction limitations:

- Provides less information than X-Ray Crystallography
- Crystallographic methods don't work for fiber diffraction data
- More than one model can explain experimental data
- Alignment of fibrils is difficult to obtain
- There is no method to process data from misaligned fibrils

Motivation

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

Performance

Rfree

Conclusions

Major goals:

- Combine fiber diffraction data with **modeling**
- Develop a fully automated structure solution method
- Determine structures *de novo*
- Obtain high-resolution structure for misaligned fibrils
- ...and potentially from single molecule X-FEL experiment

Rosetta with experimental restraints

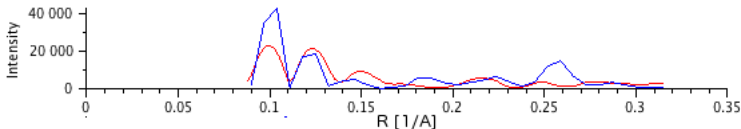
Total energy calculation:

$$E_{total} = E_{structure} + weight * E_{experimental}$$

$$E_{structure} = E_{Rosetta}$$

$$E_{experimental} = \frac{\sum (I_{calc} - I_{exp})^2}{\sum I_{exp}^2} \Leftrightarrow Rfactor$$

Intensity on a layer line: **Red** - experimental, **Blue** - calculated:



Incorporating Fiber Diffraction data into Rosetta

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

Performance

Rfree

Conclusions

Intensity calculations:

$$I_l(R) = \sum_n |G_{n,l}|^2$$

$G_{n,l}$ calculation - reciprocal space

$$G_{n,l} = \sum_n \sum_i f_i J_n(2\pi r_i R) \exp(i[-n\phi_i + (2\pi l z_i/c)])$$

$$I_l(R) = \sum_n \sum_{i,j} f_i f_j J_n(2\pi r_i R) J_n(2\pi r_j R) \cos(\text{phase})$$

$$\text{where } \text{phase} = (\phi_i - \phi_j) - 2\pi l(z_i - z_j)/c$$

Computationally costly: for 46aa proteins and 27 layer lines
 10^8 iterations...

Incorporating Fiber Diffraction data into Rosetta

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

Performance

Rfree

Conclusions

Intensity calculations :

$$I_l(R) = \sum_n |G_{n,l}|^2$$

$G_{n,l}$ calculation - real space:

$$G_{n,l} = \int_0^\infty g_{n,l}(R) J_n(2\pi r R) 2\pi r \delta r$$

$$\text{where } g_{n,l} = (c/2\pi) \int_0^c \int_0^{2\pi} \rho(r, \phi, z) e^{i(\phi - 2\pi l z/c)} \delta\phi \delta z$$

Computationally less expensive but less accurate.

Intensity calculation - methods comparison

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

Performance

Rfree

Conclusions

reciprocal space

Pros:

- Accurate
- Derivatives can be calculated

Cons:

- Computationally expensive (scales with atoms²)
- Calculated in reciprocal space

real space

Pros:

- Weak dependence on number of atoms
- Calculated in cartesian coordinates

Cons:

- Less accurate
- Derivatives cannot be calculated

De novo modeling flowchart

Fiber
Diffraction
basics

FD data and
Rosetta

Benchmark

Performance

*R*_{free}

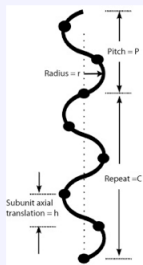
Conclusions

Protein sequence

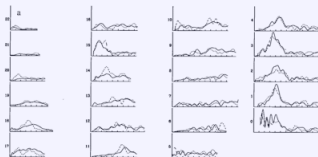


Inputs

Helical symmetry (from data)

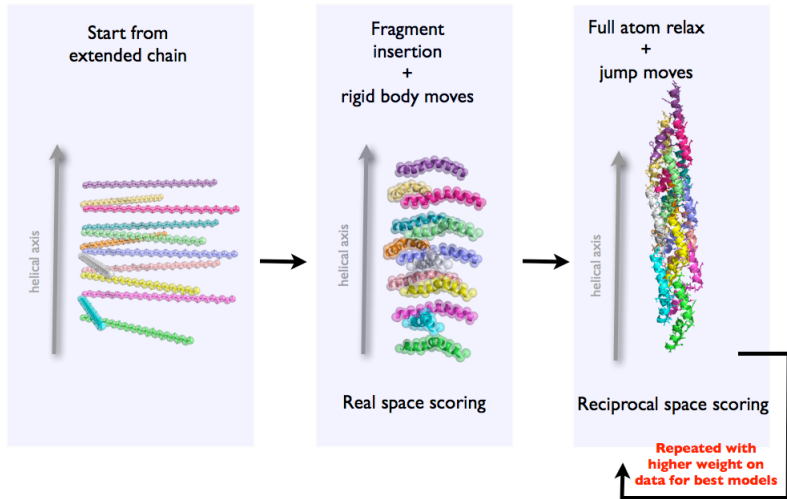


Experimental layer lines



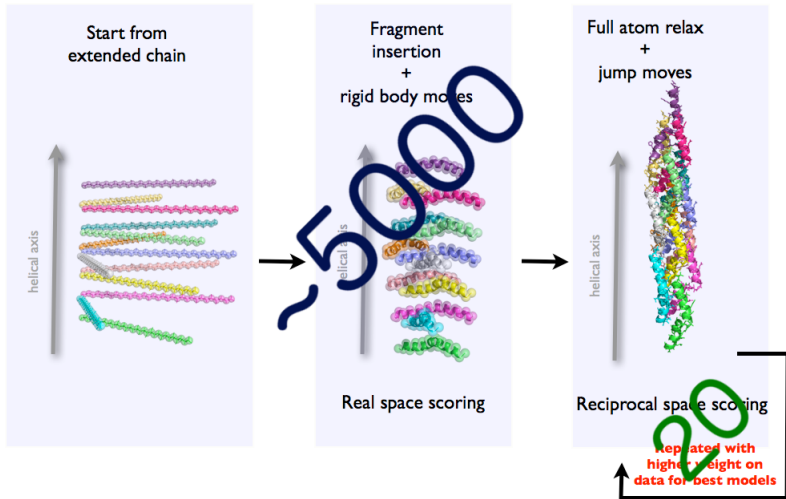
De novo modeling flowchart

Fold-And-Dock simulations:



De novo modeling flowchart

Fold-And-Dock simulations:

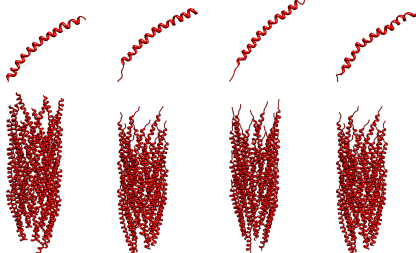


Test set

Inoviruses - bacteriophage viruses

PDB:	1lfp	1ql1	4ifm	1hgv
Phage:	Pf3	Pf1	Pf1	PH75
Number of residues:	44	46	46	46
Helix units/turns:	27/5	27/5	71/13	27/5

Monomer:



Assembly:

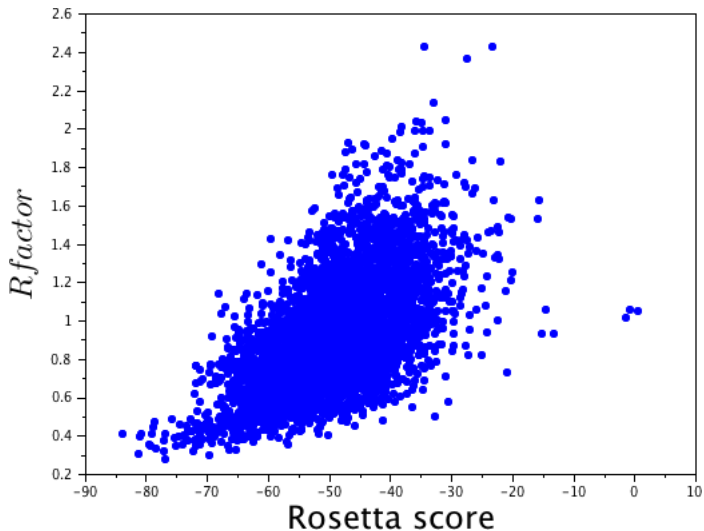


LUND
UNIVERSITY

Rfactor vs. Rosetta score

PF3 filamentous bacteriophage (1lfp)

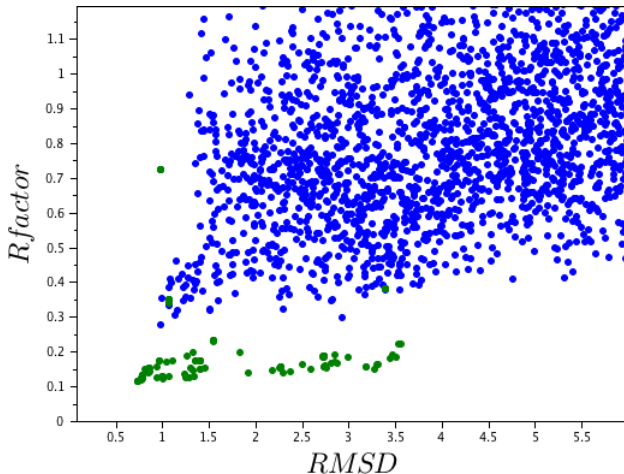
Fiber
Diffraction
basics
FD data and
Rosetta
Benchmark
Performance
Rfree
Conclusions





Rfactor vs. RMSD(monomer)

PF3 filamentous bacteriophage (1lfp)



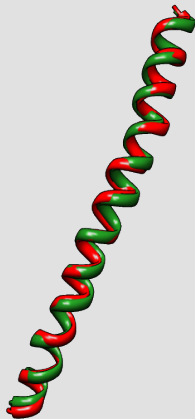


LUND
UNIVERSITY

Comparison of lowest *Rfactor* model with native PF3 filamentous bacteriophage (1lfp)

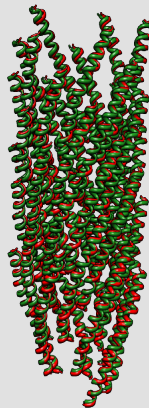
Fiber
Diffraction
basics
FD data and
Rosetta
Benchmark
Performance
Rfree
Conclusions

Monomer



RMSD: 0.7Å

Assembly



RMSD: 0.8Å, *Rfactor*: 0.11



LUND
UNIVERSITY

Comparison of lowest *Rfactor* model with native PF1 bacteriophage (1ql1)

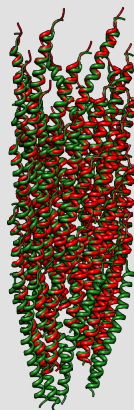
Fiber
Diffraction
basics
FD data and
Rosetta
Benchmark
Performance
R_{free}
Conclusions

Monomer



RMSD: 1.6Å

Assembly



RMSD: 1.7Å, *Rfactor*: 0.12



LUND
UNIVERSITY

Comparison of lowest *Rfactor* model with native PF1 bacteriophage (4ifm)

Fiber
Diffraction
basics

FD data and
Rosetta

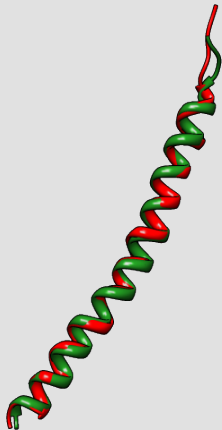
Benchmark

Performance

R_{free}

Conclusions

Monomer



RMSD: 1.6Å

Assembly



RMSD: 1.7Å, *Rfactor*: 0.07



LUND
UNIVERSITY

Comparison of lowest *Rfactor* model with native PH75 bacteriophage (1hgv)

Fiber
Diffraction
basics

FD data and
Rosetta

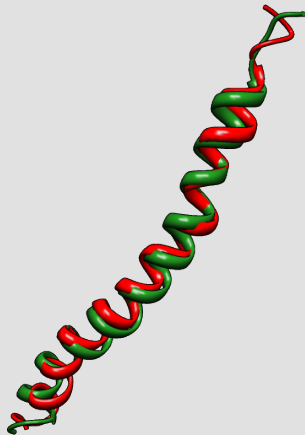
Benchmark

Performance

Rfree

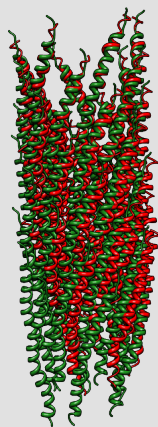
Conclusions

Monomer







RMSD: 2.0Å

Assembly



RMSD : 2.6Å, *Rfactor*: 0.25

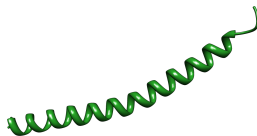
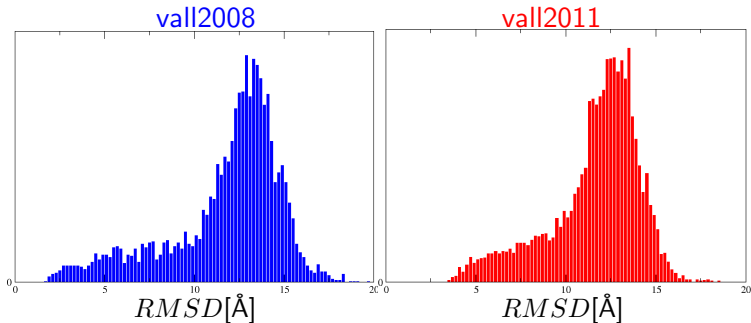
Benchmark on Inoviruses - summary

PDB:	1ifp	1ql1	4ifm	1hgv
Phage:	Pf3	Pf1	Pf1	PH75
Number of residues:	44	46	46	46
Helix units/turns:	27/5	27/5	71/13	27/5
				
Monomers (cmp.):				
Rfactor:	0.11	0.12	0.07	0.25
RMSD (monomer):	0.7Å	1.6Å	1.6Å	2.0Å
RMSD (assembly):	0.8Å	1.7Å	1.7Å	2.5Å

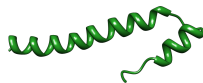
It works!

Still fragments are crucial...

RMSD distribution for Pf1 bacteriophage (1q11)



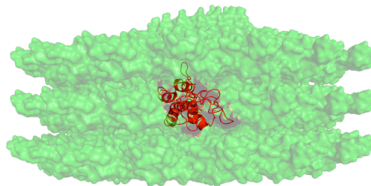
$RMSD = 1.6\text{\AA}$



$RMSD = 11.8\text{\AA}$

Hibiscus Latent Singapore virus

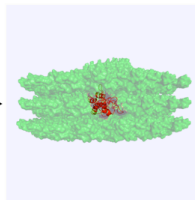
Each subunit consist of 162 amino acid residues.



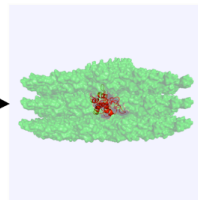
Homology models



Data guided docking of
models into a helix



All atom refinement



Real space scoring function

Reciprocal space scoring
function



Reciprocal scoring on CPU

```
for each layer_line l
  for each bessell order n
    for each reciprocal R
      for each atom_i
        for each atom_j
```

... gives 10^8 iterations for 46aa and 27 layer lines and takes 2-3s



LUND
UNIVERSITY

Execution Times

Scoring in reciprocal space

Fiber
Diffraction
basics

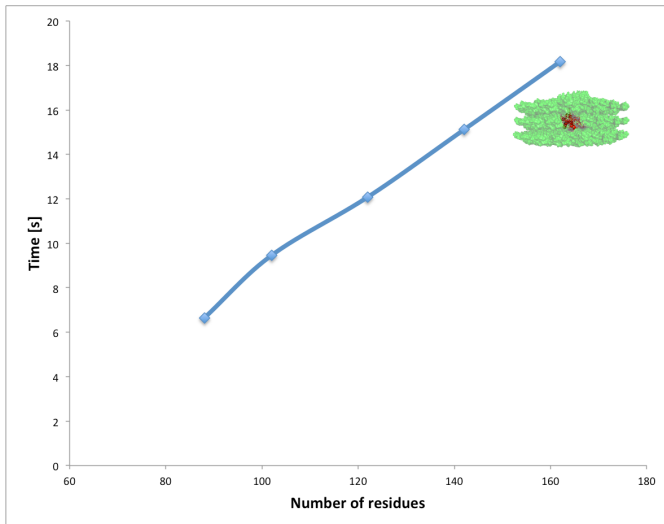
FD data and
Rosetta

Benchmark

Performance

R_{free}

Conclusions





LUND
UNIVERSITY

Execution Times

Derivatives calculation in reciprocal space

Fiber
Diffraction
basics

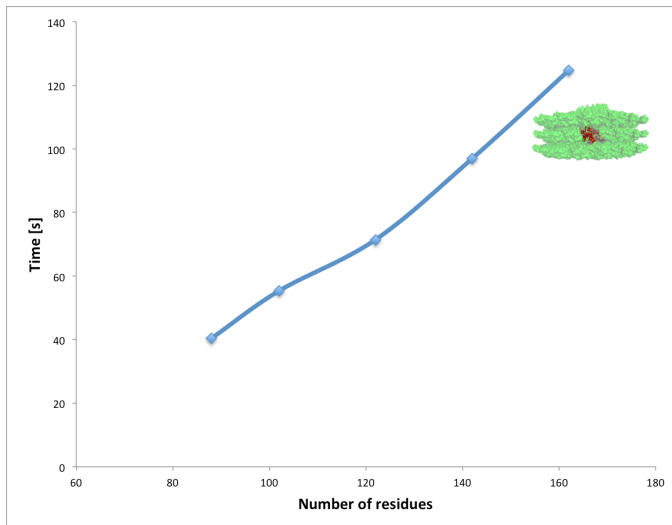
FD data and
Rosetta

Benchmark

Performance

R_{free}

Conclusions



Software and hardware optimizations





LUND
UNIVERSITY

Scoring times comparison

Optimization of trigonometric functions

Fiber
Diffraction
basics

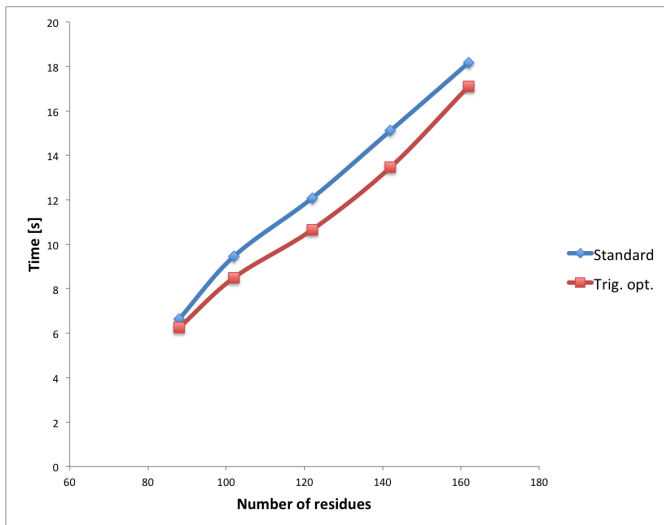
FD data and
Rosetta

Benchmark

Performance

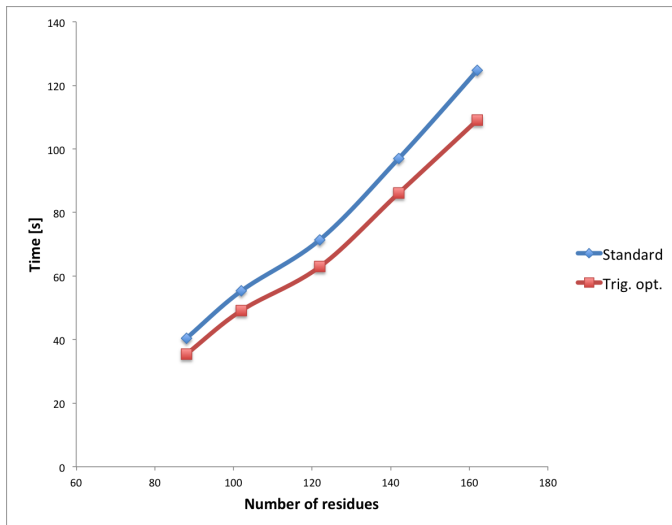
R_{free}

Conclusions



Derivatives calculation time comparison

Optimization of trigonometric functions

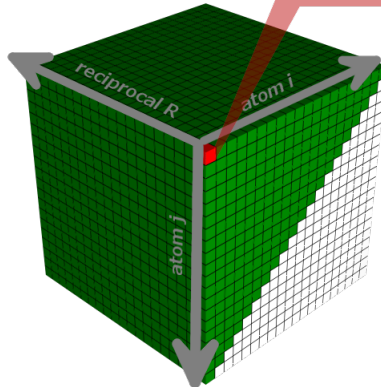


Scoring calculation on GPU

```
for each layer_line l
  for each bessel order n
```

$$f_i f_j \cdot J_n(2\pi R r_i) \cdot J_n(2\pi R r_j) \cdot \cos(\text{phase})$$

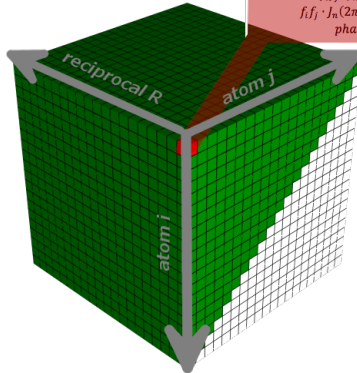
$$\text{phase} = (\varphi_i - \varphi_j) - 2\pi l(z_i - z_j)/c$$



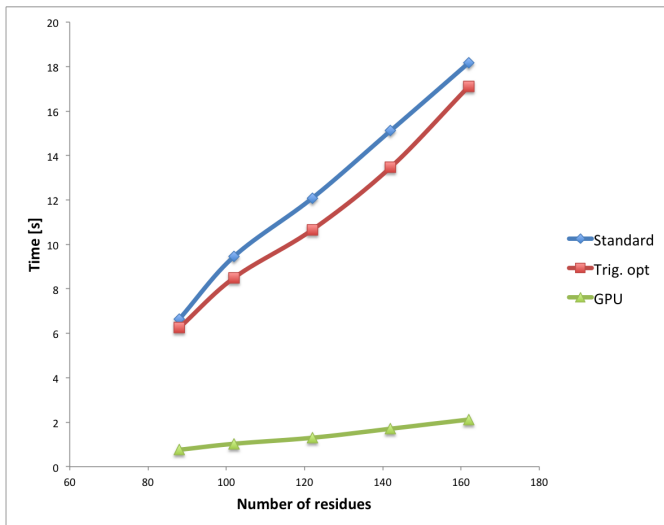
Derivatives calculation on GPU

```
for each layer_line l
  for each bessell order n
```

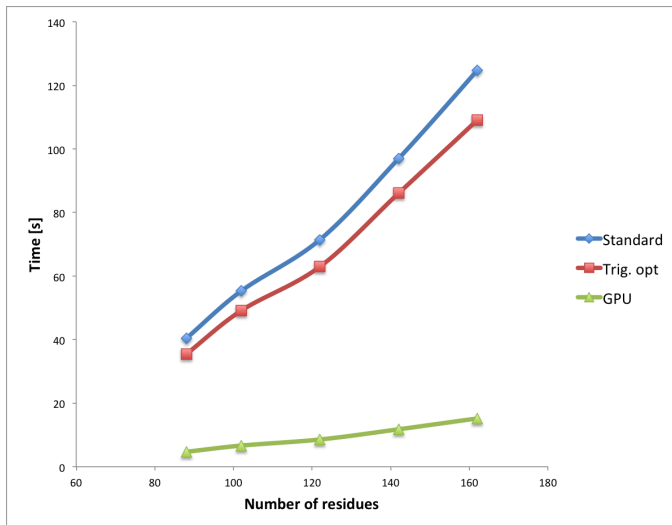
$$\begin{aligned} & f_i f_j \cdot \cos(\text{phase}) J_n(2\pi R r_i) [-2\pi R r_i J_{n+1}(2\pi R r_i) + n/r_j J_{n+1}(2\pi R r_i)] \\ & f_i f_j \cdot J_n(2\pi R r_i) \cdot J_n(2\pi R r_j) [-n \sin(\text{phase})] \\ & f_i f_j \cdot J_n(2\pi R r_i) \cdot J_n(2\pi R r_j) \cdot 2\pi/z_i [-n \sin(\text{phase})] \\ & \text{phase} = n(\varphi_i - \varphi_j) - 2\pi l(z_i - z_j)/c \end{aligned}$$



Reciprocal space scoring times comparison

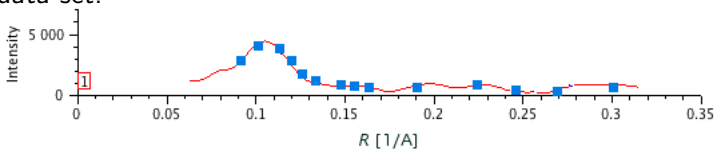


Derivatives calculation time comparison



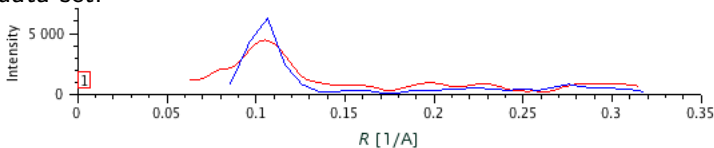
*R*_{free} - a cross-validation method

- Modeling based on *R*_{factor} is prone to over-fitting.
- Because of low redundancy of data we cannot directly use crystallographic *R*_{free}.
- We can, however optimally choose points from processed data set.



*R*_{free} - a cross-validation method

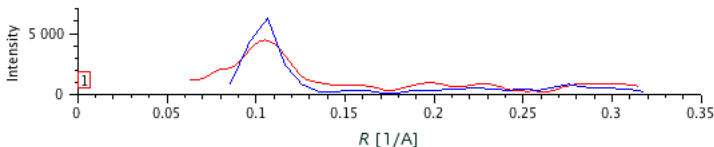
- Modeling based on *R*_{factor} is prone to over-fitting.
- Because of low redundancy of data we cannot directly use crystallographic *R*_{free}.
- We can, however optimally choose points from processed data set:



*R*_{free} - a cross-validation method

- Modeling based on *R*_{factor} is prone to over-fitting.
- Because of low redundancy of data we cannot directly use crystallographic *R*_{free}.

for each set_of_optimal_points



$$R_{\text{free}} = \text{average}(R_{\text{factor}})$$

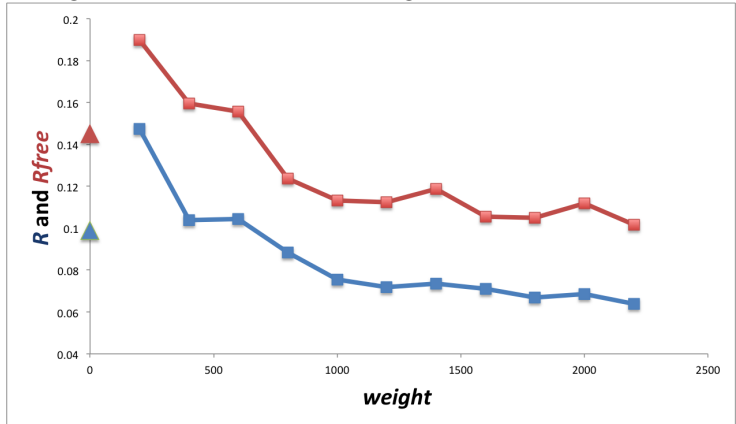


*R*_{factor} and *R*_{free}

a structure of Pf3 phage's capsid (1fp)

$$E_{total} = E_{structure} + weight * E_{experimental}$$

Triangles - native structure, Rectangles - relaxed native structure





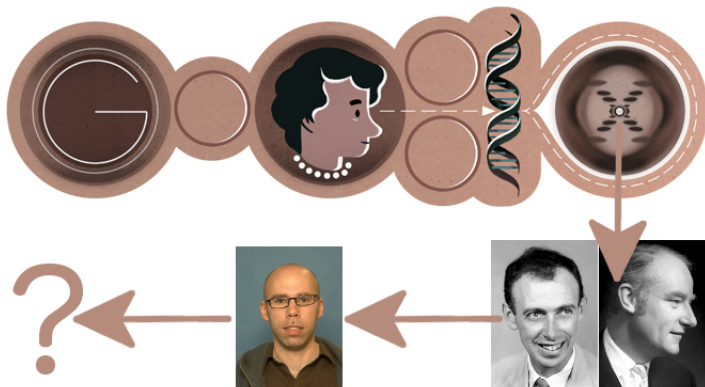
Conclusions

LUND
UNIVERSITY

Fiber
Diffraction
basics
FD data and
Rosetta
Benchmark
Performance
Rfree
Conclusions

- We have successfully developed fiber diffraction modules for Rosetta
- We can *de novo* solve structures directly from fiber diffraction data!
- Larger systems can be approached with GPU based computing.
- Our approach presents an alternative to state-of-the-art programs: CLEARER and X-PLOR

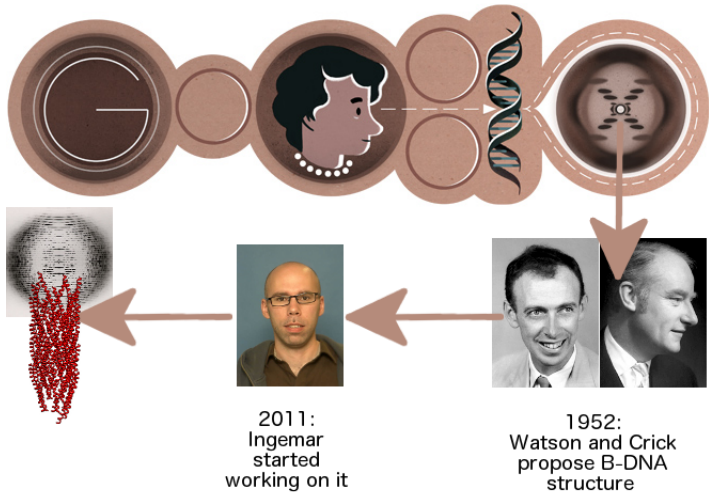
Conclusions



2011:
Ingemar
started
working on it

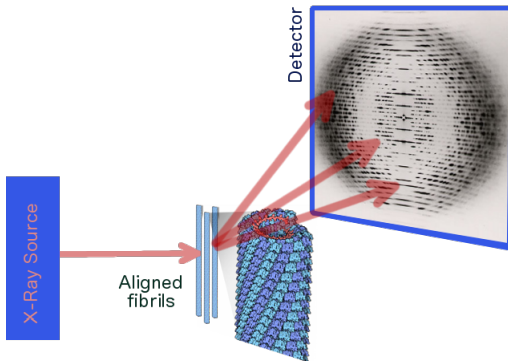
1952:
Watson and Crick
propose B-DNA
structure

Conclusions



Future overview

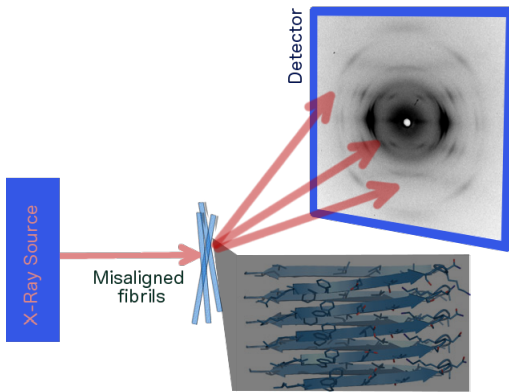
Bundle of aligned fibrils - we can solve it!



But fibrils are not always willing to align...

Future overview

Misaligned fibrils - we hope we can solve it!



A lot of data available a no method to interpret them at the moment!



Acknowledgements

Acknowledgements:

- Ingemar André
- Robert Lizatovic
- Sebastian Ramisch
- Sabine Kaltofen

- Frank DiMaio

Funding:

- Crafoord Foundation



LUND
UNIVERSITY

Thank you!

Fiber
Diffraction
basics
FD data and
Rosetta
Benchmark
Performance
Rfree
Conclusions

Thank you for your attention!