Scoring and Stochastic Sampling Methods for Sub Tomogram Averaging

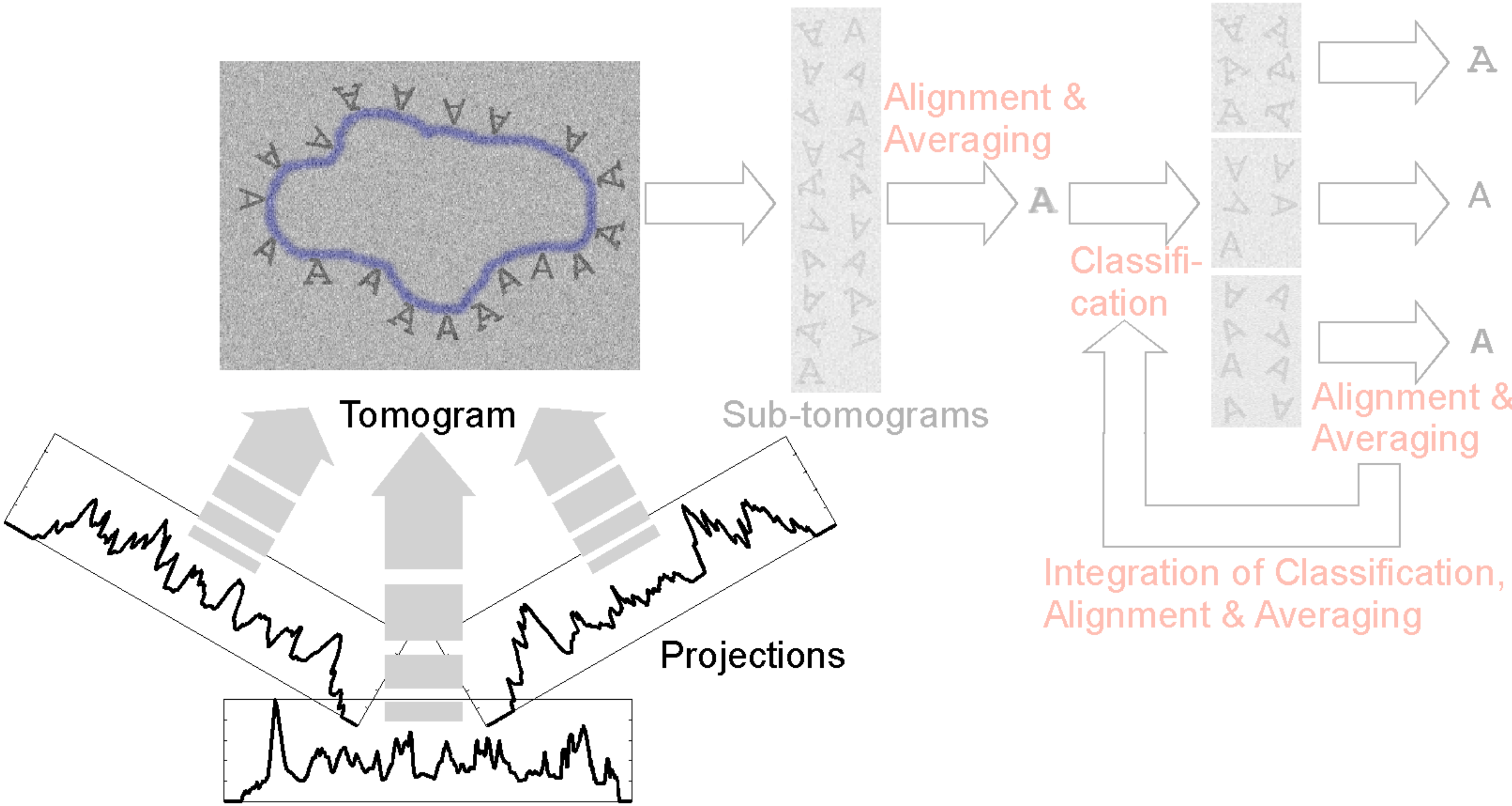
Thomas Hrabe

Modeling of Macromolecular Complexes Group
www.biochem.mpg.de/foerster

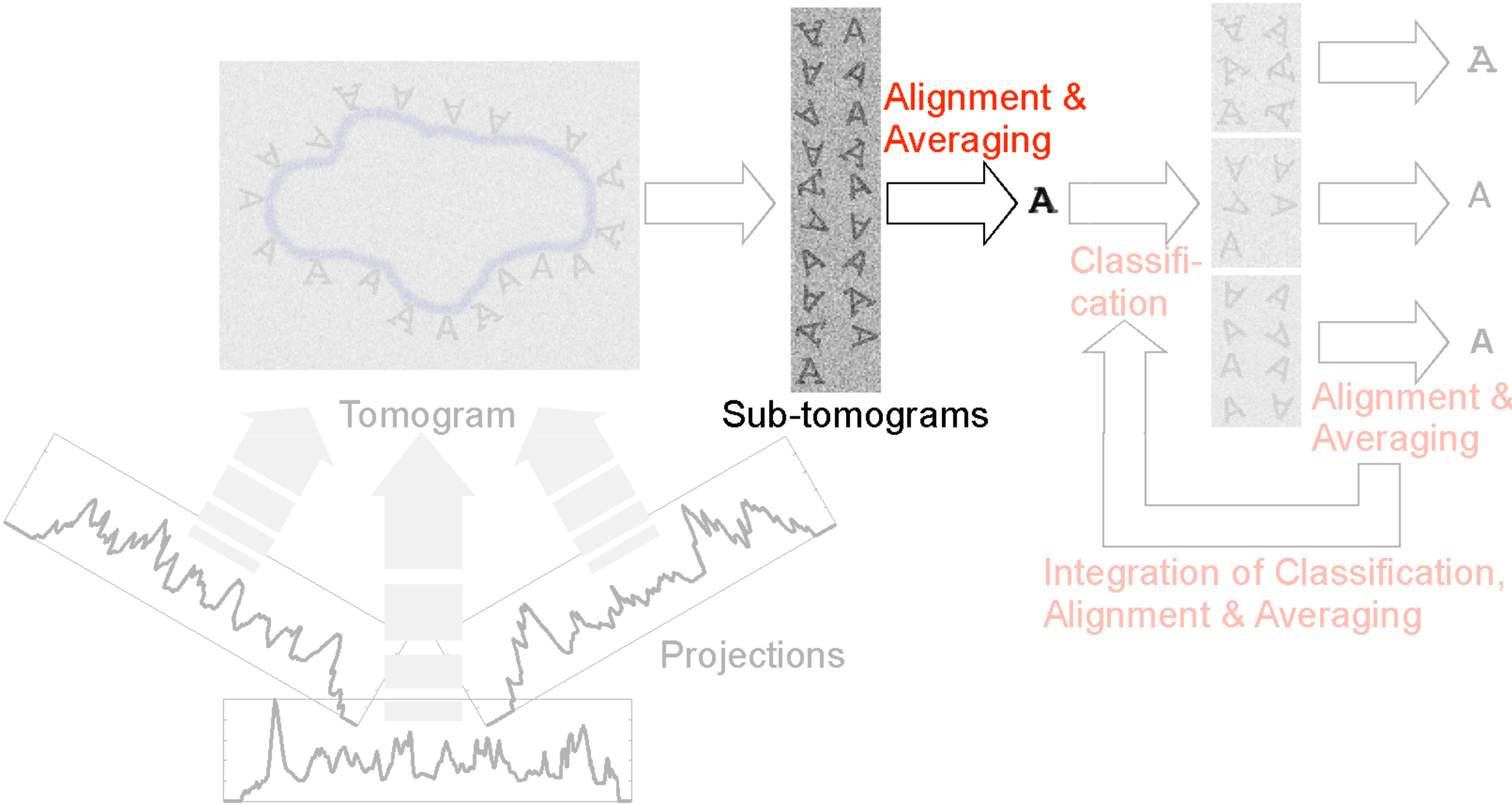
NRAMM Workshop on EM Structure Determination of Challenging Macromolecules - San Diego 2009

Max Planck Institute
of Biochemistry

Sub-Tomogram Alignment



Sub-Tomogram Alignment



A Score measures the similarity of objects

Common approach:

Cross Correlation (XCF)

$$XCF(O_1,O_2) = F^{-1}\left(F((O_1)) * F(O_2)^*\right)$$

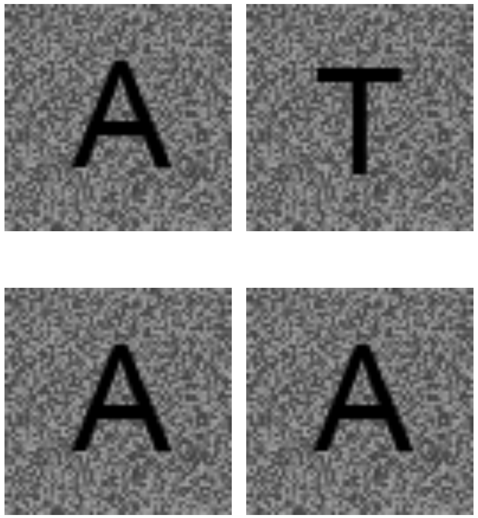
Yields arbitrary values

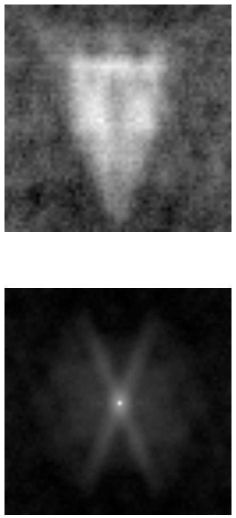
Normalized XCF (NXCF)

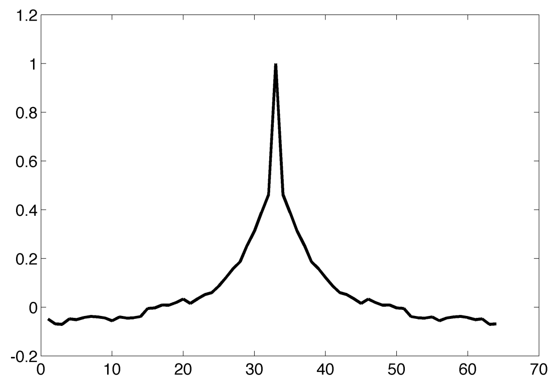
$$NXCF(O_1,O_2) = \frac{XCF(O_1,O_2)}{\|O_1\| \bullet \|O_2\|}$$

Restricts XCF to [-1,1]

O_1 O_2







Max Planck Institute
of Biochemistry

A Score measures the similarity of objects

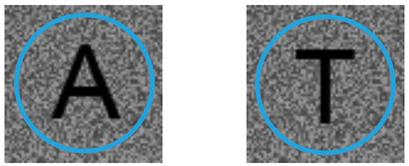
Advanced approach:

Fast Local Correlation (FLCF)

$$FLCF(O_1,O_2) = \frac{1}{P}\sum\frac{\left(O_1 - \overline{O_1}\right)M\left(O_2 - \overline{O_2}\right)}{\|O_1\| \bullet \|MO_2\|}$$

More accurate Normalization
Through use of Mask M

O_1 O_2



Max Planck Institute
of Biochemistry

A Score measures the similarity of objects

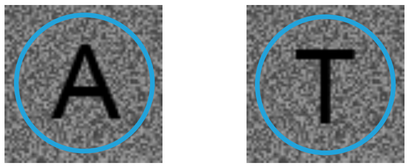
Advanced approach:

Fast Local Correlation (FLCF)

$$FLCF(O_1,O_2) = \frac{1}{P}\sum \frac{(O_1-\overline{O_1})M(O_2-\overline{O_2})}{\|O_1\|\bullet\|MO_2\|}$$

More accurate Normalization
Through use of Mask M

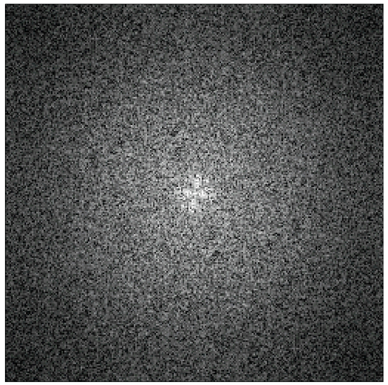
O_1 O_2



weighted Correlation (wXCC)

$$wXCC = \frac{1}{\sum_i N_i}\sum_i N_i\left|XCC(O_1,O_2)^3_{f,i}\right|$$

Noise model weights Correlation



A Score measures the similarity of objects

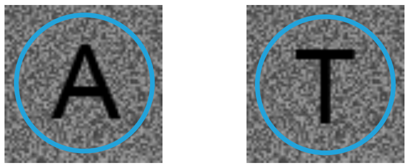
Advanced approach:

Fast Local Correlation (FLCF)

$$FLCF(O_1,O_2) = \frac{1}{P}\sum \frac{\left(O_1 - \overline{O_1}\right)M\left(O_2 - \overline{O_2}\right)}{\|O_1\| \bullet \|MO_2\|}$$

More accurate Normalization
Through use of Mask M

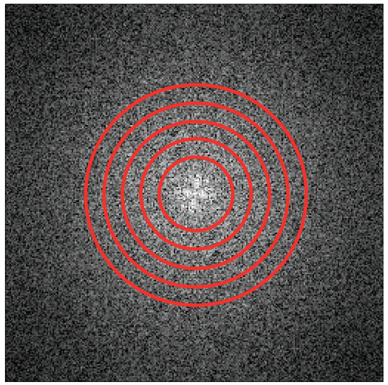
O_1 O_2



weighted Correlation (wXCC)

$$wXCC = \frac{1}{\sum_i N_i}\sum_i N_i \left|XCC\left(O_1,O_2\right)^3_{f,i}\right|$$

Noise model weights Correlation
Weight of Noise in outer bands is higher



A Score measures the similarity of objects

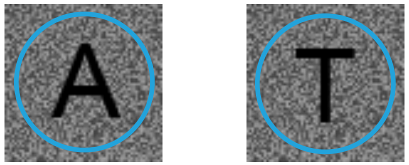
Advanced approach:

Fast Local Correlation (FLCF)

$$FLCF(O_1,O_2) = \frac{1}{P}\sum \frac{(O_1 - \overline{O_1})M(O_2 - \overline{O_2})}{\|O_1\| \bullet \|MO_2\|}$$

More accurate Normalization
Through use of Mask M

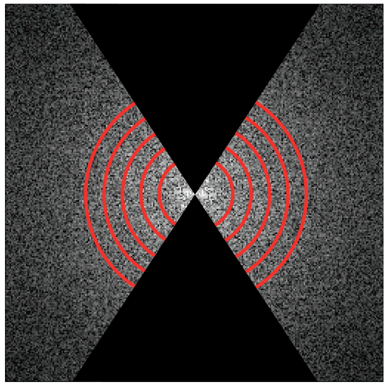
O_1 O_2



weighted Correlation (wXCC)

$$wXCC = \frac{1}{\sum_i N_i} \sum_i N_i \left| XCC(O_1,O_2)^3_{f,i} \right|$$

Noise model weights Correlation
Weight of Noise in outer bands is higher
Missing Wedge is taken into accout



Sampling of parameters space optimizes the scoring function

Common approach :

Expectation Maximization

Expectation Maximization is a rapid optimization method

Assumption

Given configuration is correct for start

What we want

Average of Sub Tomograms (Expected Result *E*)

What we have

Sub Tomograms (Observations *O* )

What we need

Rotation Angles + Shifts for each ST (Hidden Variables θ)

HFSP

Max Planck Institute of Biochemistry

Expectation Maximization is a rapid optimization method

Sub Tomograms

Expectation step

$$E_i = \frac{1}{N}\sum_{O}\tau(O,\theta_i)$$

Maximization step

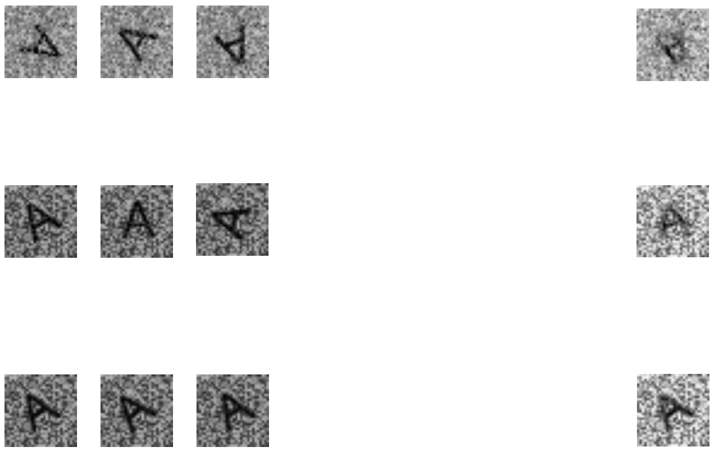
$$\theta_{i+1} = \arg\max_{\theta}\left(SC(O,E_i),\theta\right)$$

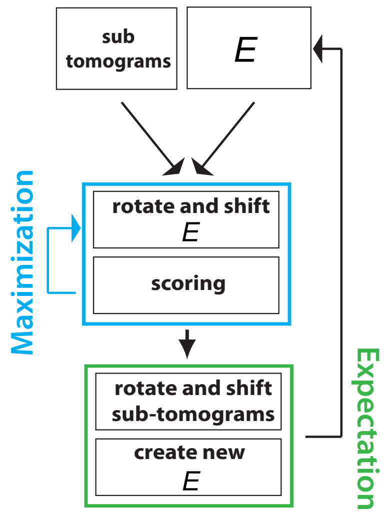
Iteration | *E*

1

2

3





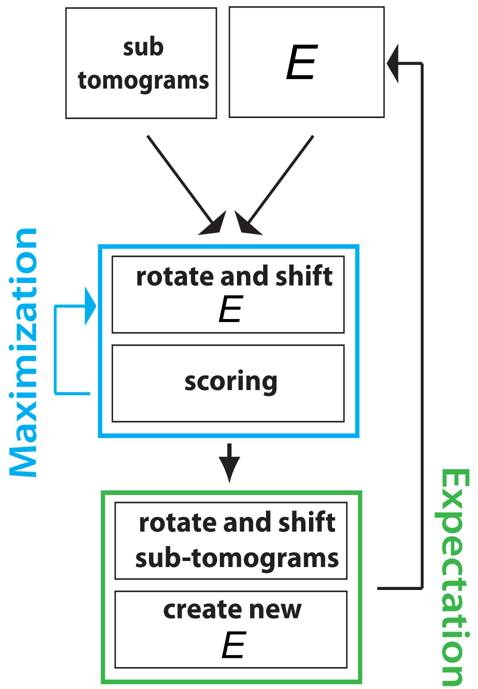
HFSP

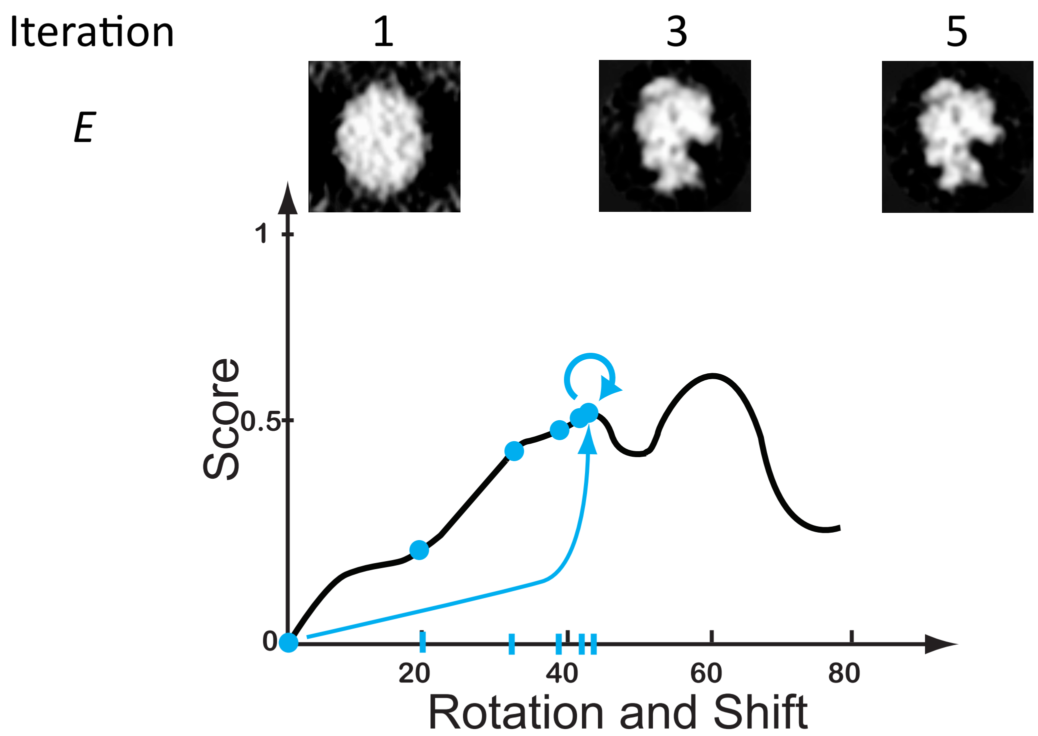
Max Planck Institute of Biochemistry

Sampling of parameters space optimizes the scoring function

Aim: Find best alignment for average

Expectation Maximization



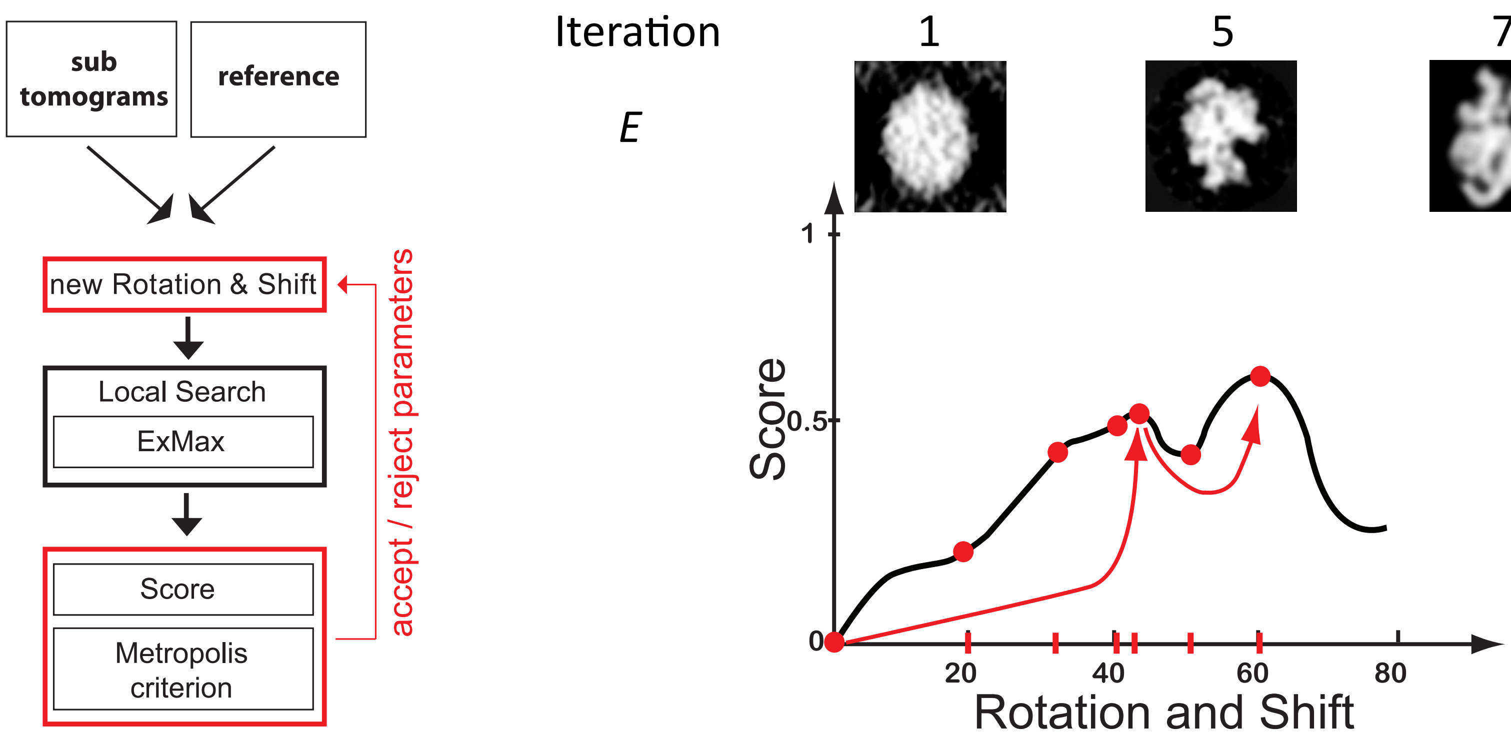


Max Planck Institute of Biochemistry

Sampling of parameters space optimizes the scoring function

Aim: Overcome local optima

Stochastic Sampling

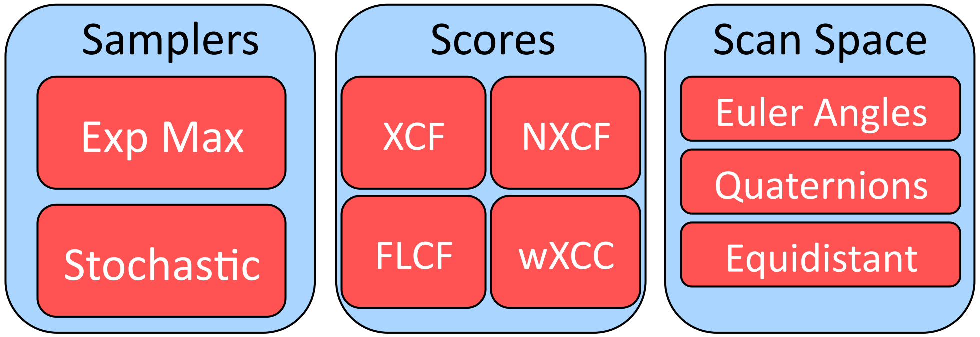


Max Planck Institute of Biochemistry

Modular Software Toolbox : PyTom

Object Oriented Programming

- Choose between components
- Objects serialize to XML



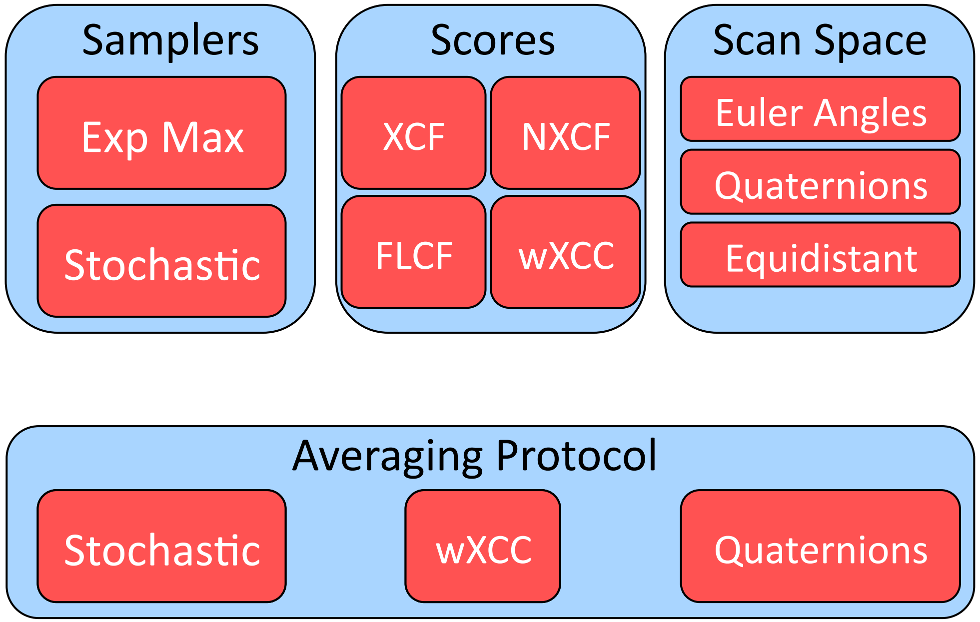
Modular Software Toolbox : PyTom

Object Oriented Programming

- Choose between components
- Objects serialize to XML

‘Mix and Match’ protocols

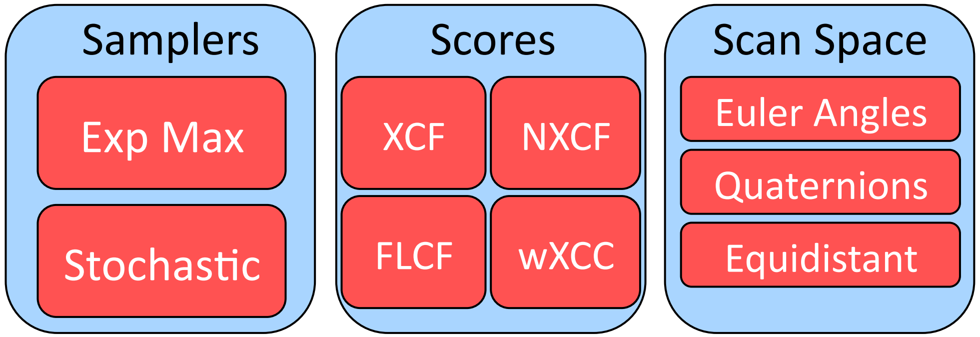
- Easily exchangeable
- Adapt to different data



Modular Software Toolbox : PyTom

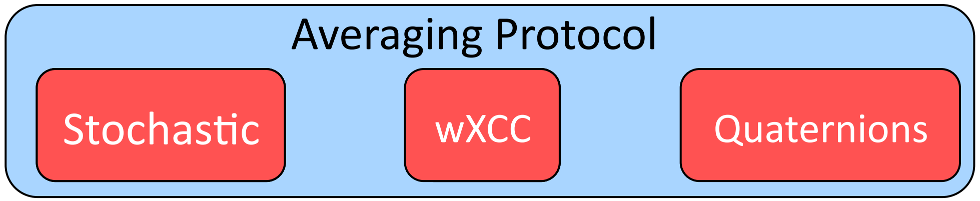
Object Oriented Programming

- Choose between components
- Objects serialize to XML



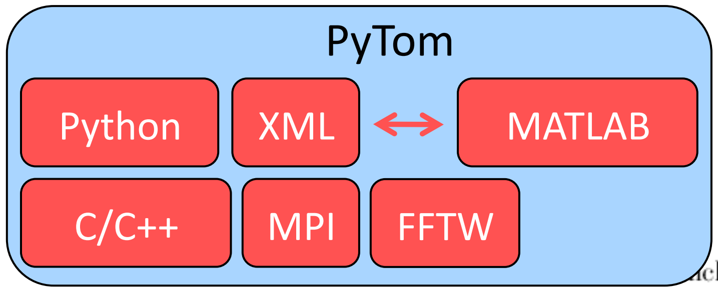
'Mix and Match' protocols

- Easily exchangeable
- Adapt to different data



PyTom

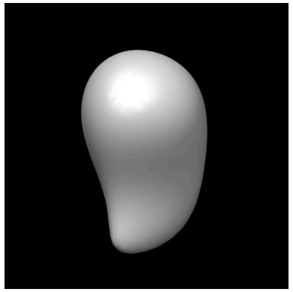
- Understandable Code
- Parallel Number Crunching



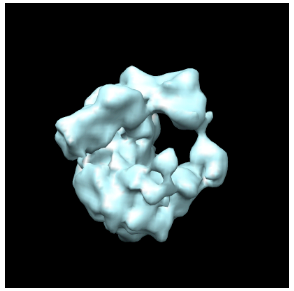
ck Institute
of Biochemistry

Preliminary Results

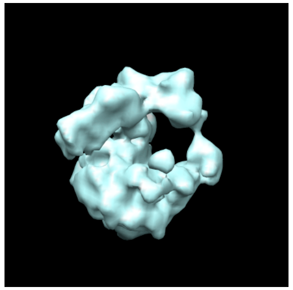
Simulations



Start Model

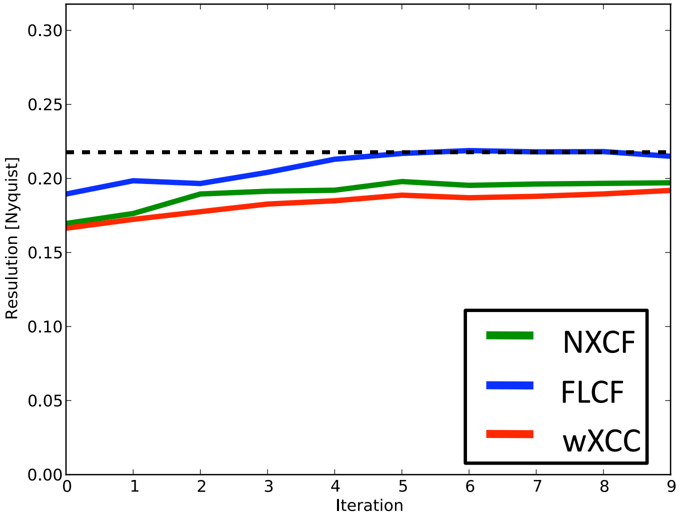


SNR ~ 0.1

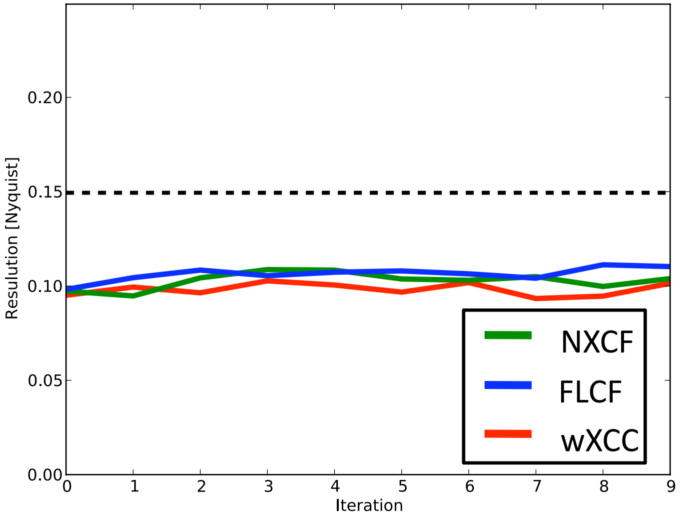


SNR ~ 0.01

SNR ~ 0.1



SNR ~ 0.01



HFSP

Max Planck Institute
of Biochemistry

Summary

PyTom – Modular Software for Sub-Tomogram Processing

easy to use, parallel, stable, free, open-source (once v1.0)

Modularity allows quick comparison of different methods

Mix and Match modules

Different Scoring functions

NXCF, FLCF, wXCC, FRC , …

Sampling

Expectation Maximization

Stochastic Optimizers (to come)