What if one would have many noisy 2D projections of assumedly identical 3D objects in unknown orientations, and one would want to know that 3D structure?

And how strictly adhering to theory helps...
Questions?

Conventional and ML 3D (projection matching) refinement (and the differences between them), Bayesian extension, how to avoid overfitting, how to avoid model bias, multi-reference refinement (classification).
Cryo-EM inconveniences

• Electrons damage biological material
  - Low dose: large amounts of noise

• We need to defocus to get contrast
  - Strong artefacts (CTF)

• We can't control how the particles fall on the grid
  - Unknown orientations & classes

Incomplete, ill-posed inverse problem
Inverse problems
The forward model

\[ X_i = \text{CTF}_i \mathbf{P}_\phi V_k + N_i \]

Given V and CTF, we can simulate X very well.

But the other way around is more difficult!
Incompleteness
Incomplete data problems

• Part of the data was not observed experimentally
  – Orientations
  – Class assignments

• Difficult to solve!
  – Iterative methods?

• Complete data problem would be very easy to solve

• (Another famous one: the phase problem in XRD)
Incomplete data problems

Observed data ($X$): images

Missing data ($Y$): orientations

Not easy
Complete data problems

white Gaussian noise

\[ L(\Theta) = P(X \mid \Theta) \]

\[ A^{MLE} = \frac{1}{N} \sum_{i=1}^{N} X_i \]

*Observed data* (X): images
Incomplete data problems

Observed data ($X$): images

Missing data ($Y$): orientations

Not easy
Incomplete data problems

• Option 1: add $Y$ to the model

$$L(Y, \Theta) = P(X \mid Y, \Theta)$$

Maximum cross-correlation / least-squares

• Option 2: marginalize over $Y$

$$L(\Theta) = P(X \mid \Theta) = \int_Y P(X \mid Y, \Theta)P(Y \mid \Theta)d\phi$$

Maximum Likelihood

Probability of $X$, regardless $Y$
The maxCC approach
Statistical data model

\[ X_i = P_{\varphi} V_k \]
Reference-based alignment

- Starts from some initial guess about the structure

$A^{(n)}$

Cross-correlation

Compare initial guess with each experimental image
Align and average

Iterate!
Align and average

Iterate!
The ML approach
Statistical data model

\[ X_i = P_{\varphi} V_k \]
Statistical data model

\[ X_i = P \varphi V_k + N_i \]

white / coloured Gaussian noise

Statistical description of the noise
Maximum likelihood

\[ A^{(n)} \quad X_i \quad \text{Statistical model} \quad \rightarrow \quad P(X_i | \phi, \Theta) \]
Maximum likelihood

\[ A^{(n)} \]

\[ X_i \]

Do not assign discrete orientations if the noise in the data does not allow this...
Incomplete data problems

• **Option 1**: add $Y$ to the model

$$L(Y, \Theta) = P(X | Y, \Theta)$$

In the limit of **noiseless data** the Two techniques are equivalent!

• **Option 2**: marginalize over $Y$

$$L(\Theta) = P(X | \Theta) = \int Y P(X | Y, \Theta)P(Y | \Theta) d\phi$$

Probability of $X$, regardless $Y$

maxCC projection matching

• Compare $X_i$ with $\text{CTF}_i \Phi V$ for all $\phi$, and select optimal $\phi^*$ based on some similarity measure (e.g. CC)

• Reconstruction:

$$V^{(n+1)} = \frac{\sum_{i=1}^{N} P_{\phi^*}^T \text{CTF}_i X_i}{\sum_{i=1}^{N} P_{\phi^*}^T \text{CTF}_i^2}$$

• Least-squares solution to $V (?)$
Maximum likelihood refinement

• Calculate a probability $P(X_i|\phi, \Theta)$ for all $\phi$, based on an explicit noise model (e.g., Gaussian)

• Probability-weighted angular assignment:

  $$\sum_{\phi} \int \cdots = \Gamma \cdot \cdots$$

  $$\sum_{\phi} \int \cdots = \Gamma \cdot \cdots$$

  $$\sum_{\phi} \int \cdots = \Gamma \cdot \cdots$$

  $$\sum_{\phi} \int \cdots = \Gamma \cdot \cdots$$

  $$\sum_{\phi} \int \cdots = \Gamma \cdot \cdots$$

  $$\sum_{\phi} \int \cdots = \Gamma \cdot \cdots$$

The theory says this is the best one can do

*(in the limit of infinitely large data sets)*

• Maximum likelihood estimate of $V$
Remaining issues

1. What to use as initial guess?
   - Local optimizer:
   - Wrong initial model -> wrong answer!
   - Model bias!
Model bias

- common-lines models are difficult
  - 2D projections are OK
  - Their combination in 3D is not

- Better (?)
  - RCT, sub-tomogram averaging, homologous structure

- EMAN(2) better than projection matching
  - But also not guaranteed...
Remaining issues

1. What to use as initial guess?
   - Wrong initial guess may lead to wrong answers!
   - **Model bias**!

2. What if multiple structures are present?
   - Cannot align against 1 reference
   - Alignment + **classification** problem
Prelim. ribosome reconstruction

91,114 particles; 9.9 Å resolution

In collaboration with Haixiao Gao & Joachim Frank
Seed generation

80 Å filter

4 random subsets; 1 iter ML
ML-derived classes

no ratcheting; no EF-G; 3 tRNAs

(raticheting, EF-G, 1 tRNA)

(Results coincided with a supervised classification)
BUT....

- 3D-classification is not a cure for bad data....
- Works best for few well-defined states
- Not all variability can be resolved
  - Continuous heterogeneity -> compromises
  - Many states may be tricky (expensive at least)
    - Supervised classification may be an alternative:
      - Fischer et al, Nature, 2010 (>20 states, 2M particles)
    - Ultimately a signal-to-noise ratio issue
Remaining issues

1. What to use as initial guess?
   1. Wrong initial guess may lead to wrong answers!
   2. Model bias!

2. What if multiple structures are present?
   1. Cannot align against 1 reference
   2. Alignment + classification problem

3. What if I do not infinite amounts of data...
Ill-posedness
The bad news

• The experimental data alone is not enough to determine a unique solution! (*ill-posed*)
  – Noise tends to accumulate in the reconstruction
The bad news

• The experimental data alone is not enough to determine a unique solution!
  – Noise tends to accumulate in the reconstruction
  – Overfitting
  – Over-estimation of resolution
  – Incorrect interpretations
The good news

• By incorporating external information, a different problem may be solved for which a unique solution does exist!

• Regularization

• Conventional approaches
  – Wiener filtering
  – Low-pass filtering
2D Wiener filter

• Assume **noise is independent**
  – with spectral power $\sigma^2(\nu)$

• Assume **signal is independent**
  – with spectral power $\tau^2(\nu)$

• **Minimise noise in 2D average:** (optimal filter)

\[
A_j^{(n+1)} = \frac{\sum_{i=1}^{N} R_{\phi^*}^T \frac{\tau^2(\nu)}{\sigma^2(\nu)} CTF_{ij} X_{ij}}{\sum_{i=1}^{N} R_{\phi^*}^T \frac{\tau^2(\nu)}{\sigma^2(\nu)} CTF_{ij}^2 + 1}
\]

Damp $A$ for those Fourier components where all CTFs are zero or $\tau^2/\sigma^2$ is small

Correct CTF AND low-pass filter!
Reconstruction methods based on backprojection and direct Fourier inversion methods require the implementation of a form of Wiener filter, which schematically is written as (see Chapter 2):

\[
D = \sum_n \frac{\text{CTF}_n^2 \text{SSNR}_n G_n}{\sum_n \text{CTF}_n^2 \text{SSNR}_n + 1}.
\]

The summation in the numerator can be realized as a backprojection of the Fourier transforms of \((n - 1)D\) projections multiplied by their respective CTFs and SSNRs, so the result is \(nD\). However, it is far from obvious how the summation in the denominator can be realized such that the result would have the intended meaning after the division is performed.
3D Wiener filter

• Same assumptions
• Plus (often):
  \[
  \frac{\tau^2(v)}{\sigma^2(v)} = \text{SSNR}(v) = 1/C \quad \text{BUT THIS IS NOT TRUE!!!!}
  \]
  Low-pass filtering effect is lost!

\[
V^{(n+1)} = \frac{\sum_{i=1}^{N} P^T \phi^* CTF_i X_i}{\sum_{i=1}^{N} P^T \phi^* CTF_i^2 + C}
\]

“Wiener constant”
“Arbitrary” low-pass filters

• Many different ones exist
  – choose shapes, effective resolution, width, etc.

• User expertise is required!
A Bayesian view on regularization

\[ P(\Theta | X) = \frac{P(X | \Theta)P(\Theta)}{P(X)} \]

Posterior = Likelihood * Prior

Evidence

Maximum A Posteriori estimation
Likelihood

- Assume noise is Gaussian and independent
  - in Fourier space
  - with spectral power $\sigma^2(\nu)$: coloured noise

\[
P(X_i \mid k, \phi, \Theta) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}} \exp \left( \frac{\|X_{ij} - CTF_{ij}(P_{\phi}V_k)_j\|^2}{-2\sigma_{ij}^2} \right)
\]
Prior

- Assume signal is Gaussian and independent
  - in Fourier space
  - Limit power $\tau^2(\nu)$: \textit{smoothness in real space!}

\[ P(\Theta) = \prod_l \frac{1}{2\pi\tau_{kl}} \exp \left\{ \frac{\|V_{kl}\|^2}{-2\tau_{kl}^2} \right\} \]
Expectation maximization

\[ V^{(n+1)} = \frac{\sum_{i=1}^{N} \int \Gamma_{i \phi}^{(n)} P_{\phi}^{T} \frac{CTF_i}{\sigma_i^2(n)} X_i d\phi}{\sum_{i=1}^{N} \int \Gamma_{i \phi}^{(n)} P_{\phi}^{T} \frac{CTF_i^2}{\sigma_i^2(n)} d\phi + \frac{1}{\tau^2(n)}} \]

Wiener filter for 3D reconstruction

\[ \sigma_i^{2(n+1)} = \frac{1}{2} \int_{\phi} \Gamma_{i \phi}^{(n)} \left\| X_i - CTF_i P_{\phi} V^{(n)} \right\|^2 d\phi \]

Estimate resolution-dependent power of noise from the data

\[ \tau^{2(n+1)} = \frac{1}{2} \left\| V^{(n)} \right\|^2 \]

Estimate resolution-dependent power of signal from the data

\[ \Gamma_{i \phi}^{(n)} = \frac{P(X_i | \phi, \Theta^{(n)}) P(\phi | \Theta^{(n)})}{\int_{\phi'} P(X_i | \phi', \Theta^{(n)}) P(\phi' | \Theta^{(n)}) d\phi'} \]
3D Wiener filter

\[ V^{(n+1)} = \frac{\sum_{i=1}^{N} \int \Gamma_i(n) P_i^T \frac{CTF_i}{\sigma_i^2(n)} X_i d\phi}{\sum_{i=1}^{N} \int \Gamma_i(n) P_i^T \frac{CTF_i^2}{\sigma_i^2(n)} d\phi + \frac{1}{\tau^2(n)}} \]

- Calculates SSNR(\(\nu\)) (as a 3D function)
- Handles uneven orientational distribution
- Handles anisotropic CTFs & CTF envelopes
- Corrects CTF & low-pass filters
- **Optimal linear filter**

WITHOUT ARBITRARINESS!
Recapitulating...

• **Inverse** problem: needs **iterating**
• **Incomplete** problem: needs **marginalizing**
• **Ill-posed** problem: needs **regularizing**

• **Bayesian approach:**
  – Does all 3 things in optimizing a single function!
  – “Learns” optimal parameters from the data
  – No *ad-hoc* parameters to tune by the user
Preventing overfitting

A little detour...

Scheres & Chen (2012) Nature Methods
The pitfalls of undetected overfitting

- 20k simulated GroEL particles
- Conventional projection matching

![Graph showing FSC vs resolution with simulated and experimental images compared.](image)
Overfitting-free refinement
Only lower resolution data drive alignment

Orientations from half-reconstructions are AS GOOD AS those from whole-reconstructions!
Experimental data

- 5,053 GroEL particles*
- 50,330 β-galactosidase particles
- 5,403 hepatitis B capsid particles**
- High-resolution crystal structures!

* kindly provided by NCMI/Steven Ludtke
** kindly provided by Tony Crowther
GroEL
Hepatitis B capsid
$\beta$-galactosidase
Conclusions

• Overfitting may be avoided without loss of resolution
  – Gold-standard FSCs between 2 independent models

• In the absence of overfitting
  – Higher-resolutions may be obtained
  – Maps are clean and easy to interpret, fit, etc.
  – FSC=0.143 is a reliable resolution estimate
Back to the statistical approach
Gold-standard FSC in the Bayesian approach

- Refine two models independently
- At each iteration: calculate $\tau^2(\nu)$ based on $\text{FSC}_{\text{gold}}$
Running RELION

Using the GUI

The graphical user interface (GUI) has been designed to provide an intuitive interaction with the user. **It is strongly recommended to always run the GUI from the same directory for a given data set.** The GUI may be launched from the command line by typing the command `relion`. A screenshot is given below.

RELION may be used to perform different tasks (run types). The following run-types may be selected from the drop-down menu at the top of the GUI:

- 2D averaging: calculate reference-free 2D class averages
- 3D reconstruction: perform 3D (multi/single-reference) refinements
Some results

Tom Walz: test new programs on old data!
Classify structural variability

- Standard data set *(i.e. used by many groups...)*
  - 10,000 70S ribosomes *(50% +EFG; 50% -EFG)*
  - MAP-refinement K=4

8 hrs on 64 CPUs
# 3D auto-refine results

<table>
<thead>
<tr>
<th>Sample characteristics</th>
<th>$\beta$-galactosidase</th>
<th>groEL</th>
<th>hepatitis B</th>
<th>rotavirus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (MDa)</td>
<td>0.45</td>
<td>0.8</td>
<td>4</td>
<td>60</td>
</tr>
<tr>
<td>Symmetry</td>
<td>D2</td>
<td>D7</td>
<td>I</td>
<td>I</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Microscopy settings</th>
<th>FEI Polara G2</th>
<th>Jeol 3000SFF</th>
<th>Hitachi HF2000</th>
<th>FEI Tecnai F30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microscope</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voltage (kV)</td>
<td>80</td>
<td>300</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>Defocus range ($\mu$m)</td>
<td>1.2-2.7</td>
<td>1.9-3.2</td>
<td>1.0-2.0</td>
<td>1.2-2.9</td>
</tr>
<tr>
<td>Detector</td>
<td>Kodak SO163</td>
<td>Kodak SO163</td>
<td>Kodak SO163</td>
<td>Kodak SO163</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data characteristics</th>
<th>$100 \times 100$</th>
<th>$128 \times 128$</th>
<th>$220 \times 220$</th>
<th>$400 \times 400^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image size (pixel$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pixel size (Å)</td>
<td>2.93</td>
<td>2.12</td>
<td>2.00</td>
<td>2.40</td>
</tr>
<tr>
<td>Nr. particles</td>
<td>50,330</td>
<td>5,053</td>
<td>5,403</td>
<td>3,700</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RELION parameters</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Particle mask diameter (Å)</td>
<td>200</td>
<td>205</td>
<td>400</td>
<td>785</td>
</tr>
<tr>
<td>Initial low-pass filter (Å)</td>
<td>60</td>
<td>60</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Initial angular sampling (°)</td>
<td>7.5</td>
<td>7.5</td>
<td>3.7</td>
<td>3.7</td>
</tr>
<tr>
<td>Local searches from (°)</td>
<td>1.8</td>
<td>1.8</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Initial offset range (pixel)</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Initial offset step (pixel)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RELION results</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall-clock time (hr)</td>
<td>13.6</td>
<td>2.0</td>
<td>8.2</td>
<td>41.5</td>
</tr>
<tr>
<td>Reported resolution (Å)</td>
<td>9.8</td>
<td>8.2</td>
<td>7.3</td>
<td>5.6</td>
</tr>
</tbody>
</table>

| Resolution vs X-ray (Å) | 10.1 | 8.4 | 7.3 | 4.4$^b$ |

<table>
<thead>
<tr>
<th>Previous results</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Refinement program</td>
<td>XMIPP$^c$</td>
<td>EMAN2$^d$</td>
<td>MRC</td>
<td>FREALIGN$^e$</td>
</tr>
<tr>
<td>Reported resolution (Å)</td>
<td>13.9</td>
<td>8.4</td>
<td>7.4</td>
<td>$\approx$6</td>
</tr>
<tr>
<td>Resolution vs X-ray (Å)</td>
<td>12.7</td>
<td>8.7</td>
<td>7.5</td>
<td>4.4$^b$</td>
</tr>
</tbody>
</table>
3D auto-refine results

Non-overfitted maps are clean!
More exciting RELION results

• DNA-origami object @ 11.5 Å resolution
  – See poster (Xiao-chen Bai)
Conclusions

• 3D-EM reconstruction is ill-posed, incomplete inverse problem
  – Needs: regularization, marginalization and iteration

• Initial model generation & classification remain problematic in some projects

• Overfitting may be avoided w/o loss of reconstruction quality
  – Use gold-standard FSCs, or high-res limited refinement!

• Bayesian framework provides a firm theoretical basis for 3D-EM
  – Learns optimal parameters from the data
  – Very little user input -> objective and easy-to-use
  – Excellent quality reconstructions
Acknowledgements

• HepB data
  – Tony Crowther
  – Greg McMullan

• GroEL data
  – Steven Ludtke

• 70S Ribosome data
  – Haixiao Gao
  – Joachim Frank

• β-galactosidase data
  – Shaoxia Chen
  – Richard Henderson

• Rotavirus data
  – James Chen
  – Niko Grigorieff

• 80S ribosome data
  – Xiaochen Bai
  – Israel Sanchez
  – Venki Ramakrishnan

• Computing
  – Jake Grimmett
  – Toby Darling

• Some code in RELION
  – Xmipp (Carazo et al.)
  – Bsoft (Heymann et al.)

• Discussions
  – LMB colleagues

• Funding

MRC Laboratory of Molecular Biology