Maximum-Likelihood Refinement of Electron Microscopy Images

Sjors H.W. Scheres[1], Roberto Marabini[2], Carlos O.S. Sorzano[1,3], Gabor T. Herman[4], Jose-Maria Carazo[1,2]

[1]Biocomputing Unit, Centro Nacional de Biotecnología. 28049. Madrid. Spain
[2]Escuela Politécnica Superior, Universidad Autónoma de Madrid, Cantoblanco, 28049, Madrid, Spain
[3]Escuela Politécnica Superior, Universidad San Pablo-CEU, Boadilla del Monte, 28668, Madrid, Spain
[4]The Graduate Center, City University of New York, New York, NY 10016, USA

scheres@cnb.uam.es

INTRODUCTION

Structural heterogeneity is often a major obstacle in 3D-EM analyses. Maximum-likelihood (ML) refinement of multiple reference volumes may be a promising way to deal with the intertwined problems of orientation assignment and classification of a heterogeneous particle population. The statistical model of the ML approach not only includes the underlying structures in the data, but also a formal description of the experimental noise and the distributions of refinement parameters. For infinitely large data sets, maximizing the likelihood yields less-biased models than those provided by alternative methods[1].

A ML approach to (single-reference) 2D-alignment was introduced by Sigworth[2]. Application of ML to 3D-reconstruction and classification of icosahedral virus particles was presented by Yin *et al.*[3]. Our group first applied ML to classification of projections using self-organizing maps[4]. More recently, we presented a ML approach to 2D multi-reference refinement[5], and a way to speed up the extensive computations[6].

Currently, we are working on ML-refinement of a single 3D-reconstruction, which will subsequently be extended to include multiple reference volumes.

METHODS

The target function: Log-likelihood

We aim to optimize the logarithm of the probability of observing data set $\mathbf{X}$, (containing N images X_i) given a model with parameter set Θ (see below).

$$L(\mathbf{X}|\Theta)=\sum_{i=1}^{N}\log\sum_{k=1}^{K}\int_{\phi}P(X_i|k,\phi,\Theta)P(k,\phi|\Theta)d\phi \qquad (1)$$

The model: Assumptions

1. The images are rotated and translated copies ($X_i(\phi)$, $\phi=\{\psi, x, y\}$) of one of *K* underlying 2D-structures A_k, to which white Gaussian noise with std.dev. σ is added:

$$P(X_i|k,\phi,\Theta)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{M}\exp\left(-\frac{\|A_k-X_i(\phi)\|^2}{2\sigma^2}\right) \qquad (2)$$

2. The in-plane rotations are uniformly distributed; the origin offsets are distributed according to a 2D-Gaussian with std.dev. ξ, centered at the origin; and structures A_k are distributed according to a discrete distribution α_k:

$$P(k,\phi|\Theta)=\alpha_k\frac{1}{2\pi\xi^2}\exp\left(-\frac{x^2+y^2}{2\xi^2}\right)\frac{1}{2\pi} \qquad (3)$$

The optimization: Expectation-Maximization[7]

1. Expectation: Use the current $\Theta^{(n)}$ to calculate a lower bound to $L(\mathbf{X}|\Theta)$.
This involves calculating all:

$$P(k,\phi|X_i,\Theta^{(n)})=\frac{P(X_i|k,\phi,\Theta^{(n)})P(k,\phi|\Theta^{(n)})}{\sum_{k'=1}^{K}\int_{\phi'}P(X_i|k',\phi',\Theta^{(n)})P(k',\phi'|\Theta^{(n)})d\phi'} \qquad (4)$$

2. Maximization: maximize the bound by updating all model parameters:

A. (2D) Multi-reference refinement:

$$A_k^{(n+1)}=\frac{\sum_{i=1}^{N}\int_{\phi}P(k,\phi|X_i,\Theta^{(n)})X_i(\phi)d\phi}{\sum_{i=1}^{N}\int_{\phi}P(k,\phi|X_i,\Theta^{(n)})d\phi} \qquad (5)$$

B. (3D) Volume refinement:

A_k are projections in K different directions (R_k) of a volume V. A better volume $V^{(n+1)}$ is obtained by solving the following weighted least-squares problem:

$$\sum_{k=1}^{K}\sum_{i=1}^{N}\int_{\phi}P(k,\phi|X_i,\Theta^{(n)})d\phi\left\|R_kV^{(n+1)}-\frac{\sum_{i=1}^{N}\int_{\phi}P(k,\phi|X_i,\Theta^{(n)})X_i(\phi)d\phi}{\sum_{i=1}^{N}\int_{\phi}P(k,\phi|X_i,\Theta^{(n)})d\phi}\right\|^2=0 \qquad (6)$$

which is done using **a new type of iterative reconstruction techniques**.

Besides A_k or V, also update the other parameters in Θ: σ, ξ, and α_k. Then proceed using $\Theta^{(n+1)}$ to calculate the lower bound for the next iteration ($n+1$).

REFERENCES

1. J.A. Rice, *Mathematical Statistics and Data Analysis* (1995)
2. F.J. Sigworth, *J. Struc. Biol.* **122**, 328-339 (1998)
3. Z. Yin *et al., J. Struc. Biol.* **144**, 24-50 (2003)
4. A. Pascual-Montano *et al., J. Struc. Biol* **133**, 233-245 (2001)
5. S.H.W. Scheres *et al., J. Mol. Biol.* **348**, 139-149 (2005)
6. S.H.W. Scheres *et al., Bioinformatics*, accepted
7. A.P. Dempster *et al., J. Royal. Statist. Soc. Ser. B.* **39**, 1-38 (1977)
8. C.O.S. Sorzano *et al., J. Struc. Biol.* **148**, 194-204 (2004)

RESULTS

A. (2D) Multi-reference Refinement

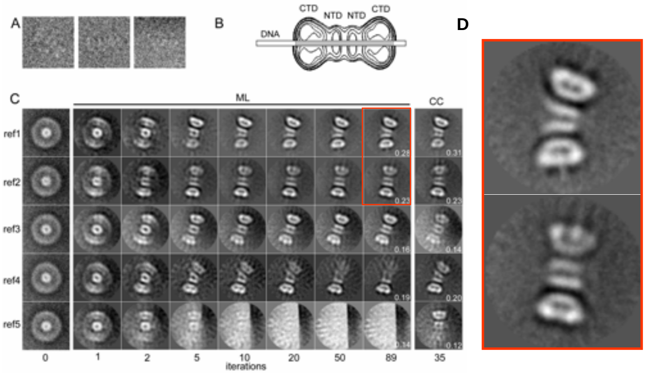


Figure 1: *Five-reference refinement of 7,590 cryo-EM projections (A) of large T-antigen in complex with an symmetric DNA-probe (B).* ***Unbiased starting models*** *were obtained by averaging over five random subsets of the unaligned images (C; column 0). In our experience,* ***ML refinement has been the only method capable of visualizing the complexed DNA*** *(see D, upper image).*

B. A faster approach

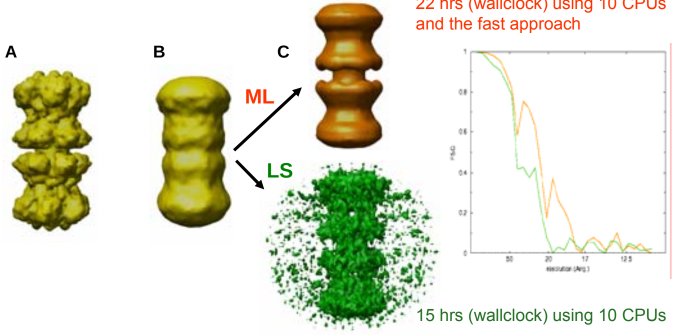
Evaluating Eq. 5 (or 6) for all i, k, and φ is expensive! Alternatively, for all (n) we store those translations ($x_{i,k}^{(n)}, y_{i,k}^{(n)}$) that yield the highest probability of observing image X_i given reference A_k (Eq. 2). Then, for those X_i, A_k and ψ where:

$$P(X_i|k,\psi,x_{i,k}^{(n)},y_{i,k}^{(n)},\Theta^{(n)})<10^{-12}\max_{\psi}\left[P(X_i|k,\psi,x_{i,k}^{(n)},y_{i,k}^{(n)},\Theta^{(n)})\right]$$

we assume that none of the translations will contribute significantly to Eq. 5 (or 6) and in iteration (n+1) the corresponding integration over x and y is skipped.

In many cases, this results in a speed-up of the calculations (**up to 7-fold**), without notably changing the optimization path. The images shown in Fig 1D were obtained using this approach in combination with a relatively fine ψ-sampling of 2°.

C. (3D) Volume Refinement



22 hrs (wallclock) using 10 CPUs and the fast approach

15 hrs (wallclock) using 10 CPUs

Figure 2: *Volume refinement (10 iterations) using 5,000 projections of a 6-fold symmetric phantom representing large T-antigen (A). To all projections white Gaussian noise was added (SNR ~0.02). The initial reference volume (B) was obtained by simulating a RCT experiment. ML optimization (fast protocol) yielded a volume that was less noisy and extended to higher resolution than a conventional protocol of 5D-orientation assignment based on maximum cross-correlation and weighted back-projection with arbitrary tilt geometry (C).*

DISCUSSION

- Our 2D-results indicate that ML is a powerful tool to classify structural differences
- Preliminary results indicate that ML may also be well-suited for volume refinement
- Current efforts focus on optimization of the 3D-refinement algorithm and its extension to include multiple reference volumes
- We further note that:
  - The 2D-program, its fast variant, and an MPI implementation are available through our free program package Xmipp[8] (www.cnb.uam.es/~bioinfo).
  - The assumption of independent noise is incorrect for experimental data. The statistical model in the ML approach may be improved by incorporation of CTF-introduced dependencies between nearby pixels.