Supplementary Methods Accompanying Reboul *et. al* "Single-particle Cryo-EM—Improved *Ab Initio* 3D Reconstruction with SIMPLE/PRIME"

July 26, 2017

Contributors:
cyril.reboul@monash.edu
michael.eager@monash.edu
dominika.elmlund@monash.edu
hans.elmlund@monash.edu
Adress:
Biomedicine Discovery Institute
Monash University
Clayton, VIC, Australia, 3800
Webpage:
www.simplecryoem.com
Contact:
`http://simplecryoem.com/contact.html`

1 Equation describing a cluster centre

Previously, we relied on the "phase flipping" method that simply involves changing the sign of the resolution bands subjected to sign inversion by the CTF. This results in a cluster centre $\overline{\mathbf{C}}_j$ defined as

$$\overline{\mathbf{C}_j} = \frac{\sum_{i=1}^{N_p} \delta_{ij} \mathbf{X}_i sgn(CTF_i)}{\sum_{i=1}^{N_p} \delta_{ij}} \tag{1}$$

where i is the particle image index, j is the cluster index, N_p is the number of particle images, δ_{ij} is the Kronecker delta function that is one if particle image i belongs to cluster j and zero else, $\mathbf{X}_i$ is the particle image Fourier transform, and $sgn(CTF_i)$ denotes the sign of the CTF. Assuming a simplified image formation model

$$\mathbf{X}_i = \mathbf{F}_i CTF_i + \mathbf{N}_i \tag{2}$$

where $\mathbf{F}_i$ is the 2D structure factor and $\mathbf{N}_i$ is a Gaussian noise term, we can write the equation for the cluster centre as

$$\overline{\mathbf{C}_j} =< \mathbf{F}_i |CTF_i| > + < sgn(CTF_i) \mathbf{N}_i > \tag{3}$$

where $< . >$ denotes expectation value. If we let the number of images in cluster j, $p(C_j)$, where

$$p(C_j) = \sum_{i=1}^{N_p} \delta_{ij} \tag{4}$$

approach infinity

$$\lim_{p(C_j)\to\infty} \overline{\mathbf{C_j}} =< \mathbf{F}_i |CTF_i| > \tag{5}$$

the cluster centre approaches the expectation value of the structure factor multiplied with the absolute value of the CTF. If the number of observations are large and the defocus values vary considerably, phase-flipping will provide a good estimate of the structure factor. However, in 2D analysis the number of images per cluster are often few, which results in the signal being incorrectly weighted in the Fourier domain when using phase-flipped images. Furthermore, the weighting of the information around the zero crossings of the CTF, where the signal component is absent and only noise is present, is incorrect for phase-flipped images. Re-defining the cluster centre as

$$\overline{\mathbf{C}_j} = \frac{\sum_{i=1}^{N_p} \delta_{ij} \mathbf{X}_i CTF_i}{\sum_{i=1}^{N_p} \delta_{ij} CTF_i^2} =< \mathbf{F}_i > + < \mathbf{N}_i > \tag{6}$$

results in a better behaved average that in the limit of infinite population

$$\lim_{p(C_j)\to\infty} \overline{\mathbf{C}_j} =< \mathbf{F}_i > \tag{7}$$

approaches the expectation value of the 2D structure factor that we seek to reconstruct. This theory is identical to that originally put forward by Frank and Penczek, reviewed in [Penczek(2010)].

2 Hadamard matrix theory for accelerated single-particle orientation search

In PRIME2D, the global correlation function G subject to maximisation is the sum of all individual particle-cluster centre correlations

$$G = \sum_{i=1}^{N_p} \lambda_{ij}^{(\phi_i)} \tag{8}$$

where $\lambda_{ij}^{(\phi_i)}$ denotes the band-pass limited cross correlation between the i:th polar particle Fourier transform $\mathbf{X}_i^{(\phi_i)}$ in rotational state ϕ_i and the j:th polar reference Fourier transform $\mathbf{Y}_j$.

$$\lambda_{ij}^{(\phi_i)} = \frac{\mathrm{Re}\{\sum_\kappa \mathbf{Y}_\mathrm{j} \mathbf{X}_\mathrm{i}^{(\phi_\mathrm{i})*}\}}{\sqrt{\sum_\kappa |\mathbf{X}_i^{(\phi_i)}|^2 \sum_\kappa |\mathbf{Y}_j|^2}} \tag{9}$$

where κ defines a band-pass limited area in 2D polar Fourier space. Polar Fourier transforms are generated by convolution interpolation [Yang and Penczek(2008)]. We assume that no CTF-dependent modifications have been applied to the experimental images. Therefore, the reference needs to be multiplied with the CTF

$$\mathbf{Y}_j = \overline{\mathbf{C}_j} CTF_i \tag{10}$$

to accomplish correct weighting of the correlation, which is composed of two terms: a signal-dependent component that depends on the linear association between the 2D structure factor and the signal in the cluster centre and a noise component that embodies many sources of noise-dependent errors that are difficult to express in mathematical formulae. Therefore, we low-pass limit the projection matching to avoid including frequencies with too weak SSNR.

Calculating correlations parameterised over in-plane rotations and origin shifts uptake most of the computations of our stochastic SAC solver. It is therefore important that these operations are efficiently implemented. Let the $m \times n$ dense complex matrices $\mathbf{Y} := \{y(\phi, k) \in \mathbb{C} \,|\, \phi \in \mathbb{R}, k \in$

$\mathbb{Z}\}$ and $\mathbf{X} := \{x(\phi, k) \in \mathbb{C} \mid \phi \in \mathbb{R}, k \in \mathbb{Z}\}$ denote polar Fourier transforms (PFTs) with the first dimension $\phi \in [0, 2\pi]$ spanning in-plane rotation and the second dimension $k \in [k_H, k_L]$ spatial frequency. The entry-wise multiplication between $\mathbf{Y}$ and $\mathbf{X}$ is known as the Hadamard product, denoted by $\circ$ and defined as

$$[\mathbf{Y}]_{\phi k}[\mathbf{X}]_{\phi k} = [\mathbf{Y} \circ \mathbf{X}]_{\phi k}. \tag{11}$$

The PFTs are represented by a cyclically ordered set of n one-dimensional arrays of complex numbers (the columns of $\mathbf{Y}$ or $\mathbf{X}$). Each individual array contains components from low to high Fourier index $k \in [k_H, k_L]$ where k_H denotes the high-pass limit and k_L denotes the low-pass limit. The complex matrix $\mathbf{Y}$ represents the fixed reference PFT and $\mathbf{X}$ represents the rotating PFT that we seek to register with $\mathbf{Y}$. In-plane rotations $\phi \in [0, 2\pi]$ are represented by cyclic permutations, where the successor $\gamma(\phi)$ of a given in-plane rotation ϕ is defined

$$\gamma(\phi) = \begin{cases} 0 & \text{if } \phi + \Delta\phi > 2\pi \\ \phi + \Delta\phi & \text{otherwise} \end{cases} \tag{12}$$

where $\Delta\phi = 1/r_{\text{mask}}$ is the angular resolution in the plane and r_{mask} is the circular mask radius in pixels. Solving SAC involves many billions of correlation function evaluations per iteration. We begin noting that for any contiguous PFT segment in the range $[\phi, \phi + \pi]$ where $\phi \in [0, 2\pi]$ the PFT has constant power.

$$\Pi(\mathbf{Y}) = \sum_{\phi=0}^{\pi} \sum_{k=k_H}^{k_L} |y_{\phi k}|^2 \quad \text{and} \quad \Pi(\mathbf{X}) = \sum_{\phi=0}^{\pi} \sum_{k=k_H}^{k_L} |x_{\phi k}|^2. \tag{13}$$

These terms are pre-calculated because during one iteration (one pass over all particle images), the references and the resolution range remain unchanged. Organising the data structure for correlation calculations to take advantage of how the CPU uses the cache to access memory can lead to substantial performance gains. Consider doubling the space complexity for the rotating PFT $\mathbf{X}$ by concatenating two copies of $\mathbf{X}$ to create a new matrix $\mathbf{X}'$ with elements $\mathbf{X}'_{\phi k}$ for $\phi \in [0, 4\pi]$ and $k \in [k_H, k_L]$. In this setting, the lower ϕ_l and upper ϕ_u rotational index bounds for rotation ϕ are calculated as follows

$$\phi_l\,(\phi) = \phi \quad \text{and} \quad \phi_u\,(\phi) = \phi + \pi. \tag{14}$$

This allows reformulation of the normalised cross correlation coefficient between the reference PFT $\mathbf{Y}$ and the particle PFT $\mathbf{X}'$ using the Hadamard product

$$\lambda_{ij}^{(\phi)} = \frac{\sum_{\phi=\phi_l(\phi)}^{\phi_u(\phi)} \sum_{k=k_H}^{k_L} \operatorname{Re}\left\{[\mathbf{Y} \circ \mathbf{X}'^{*}]_{\phi \mathrm{k}}\right\}}{\sqrt{\Pi(\mathbf{Y})\,\Pi(\mathbf{X}'^{*})}}. \tag{15}$$

This restructuring of the calculation allows us to express discrete rotations using pointer arithmetic to contiguous registers in random access memory, removing unnecessary loops and index modifications, leading to more efficient cache utilisation and improved performance. In the SHC update step, the optimal in-plane angle $\tilde{\phi}_i$ is identified for the j^{th} reference $\mathbf{Y}_j$ by solving the stochastic optimisation problem

$$\begin{aligned} &\underset{\forall\phi}{\text{identify}} && \lambda_{ij}(\phi) \\ &\text{subject to} && \lambda_{ij}(\phi_i^{(t)}) \geqslant \lambda_{ij}(\phi_i^{(t-1)}) \end{aligned} \tag{16}$$

where t denotes iteration number and $\forall\phi$ means any in-plane rotation. In words, we seek to stochastically identify any cluster assignment and in-plane rotation that improves the correlation obtained with the previous best parameters. This approach is called first-improvement heuristic or stochastic hill climbing (SHC).

We separate the rotation search from the origin shift search. Once a candidate reference PFT and an in-plane angle has been identified, origin shifts are explored for this reference/in-plane rotation pair using continuous optimisation in polar Fourier space. A shift in real-space corresponds to a linear phase change in the Fourier domain. Using 2D Cartesian coordinates and a shift of (x_0, y_0) pixels, the expression is

$$\mathcal{F}\left[f\left(x - x_0, y - y_0\right)\right] = F\left(h, k\right)\, exp\left[-2\pi i\left(\frac{h\, x_0}{N_x} + \frac{k\, y_0}{N_y}\right)\right] \tag{17}$$

where (N_x, N_y) is the even dimensions of the real image and (h, k) are the Fourier indices. The shift can equivalently be expressed in polar coordinates via

$$F_{\text{shifted}}\left(\phi, k\right) = F\left(\phi, k\right)\, exp\left[-i\left(t^{(x_0)}(\phi, k)\, x_0 + t^{(y_0)}(\phi, k)\, y_0\right)\right] \tag{18}$$

where

$$t^{(x_0)}(\phi, k) = 2\pi\left(\frac{k\, \cos\left(\phi\right)}{N_x}\right) \quad \text{and} \quad t^{(y_0)}(\phi, k) = 2\pi\left(\frac{k\, \sin\left(\phi\right)}{N_y}\right) \tag{19}$$

denote the elements of the transfer matrices, $\mathbf{T}^{(x_0)}(\phi, k)$ and $\mathbf{T}^{(y_0)}(\phi, k)$, respectively. These matrices are constant throughout the orientation search and can be pre-calculated and stored in memory. The complex elements $o_{\phi k}^{(x_0,y_0)}$ of the shift transformation matrix $\mathbf{O}(x_0, y_0)$ for a given rotational origin shift of (x_0, y_0) are obtained from

$$\begin{aligned} \mathrm{Re}\left\{\mathrm{o}_{\phi\mathrm{k}}^{(\mathrm{x_0,y_0})}\right\} &= \cos(\, t^{(x_0)}(\phi, k)\, x_0) & (20) \\ \mathrm{Im}\left\{\mathrm{o}_{\phi\mathrm{k}}^{(\mathrm{x_0,y_0})}\right\} &= \sin(\, t^{(y_0)}(\phi, k)\, y_0) & (21) \end{aligned}$$

and a continuously shifted PFT is obtained with a simple Hadamard product

$$\mathbf{Y}_{\text{shifted}} = \left[\mathbf{Y} \circ \mathbf{O}(x_0, y_0)\right]_{\phi k} \tag{22}$$

Assuming that an optimal in-plane angle $\tilde{\phi_i}$ has been identified for the i^{th} reference $\mathbf{Y}_i$, the below optimisation problem is solved to identify the optimal origin shift $(\tilde{x}_0, \tilde{y}_0)$

$$\lambda_{ij}^{(\tilde{\phi_i})}(\tilde{x}_0, \tilde{y}_0) = \underset{x_0, y_0 \in [x_l, x_u]}{\operatorname{argmax}} \sum_{\phi=\phi_l(\tilde{\phi_i})}^{\phi_u(\tilde{\phi_i})} \sum_{k=k_h}^{k_L} \mathrm{Re}\left\{\left[\left[\mathbf{Y}_\mathrm{i} \circ \mathbf{O}(\mathrm{x_0}, \mathrm{y_0})\right]_{\tilde{\phi}_\mathrm{i},\mathrm{k}} \circ \mathbf{X}'^{*}\right]_{\tilde{\phi}_\mathrm{i},\mathrm{k}}\right\} \tag{23}$$

The Nelder–Mead method [Nelder and Mead(1965)] is used to solve the problem. The reference PFT $\mathbf{Y}$ is shifted because, as explained above, the particle PFT $\mathbf{X}'$ spans $[0, 4\pi]$ rotations whereas the reference only spans $[0, \pi]$. Hence, shifting the reference is more efficient. A shift vector independent of the in-plane rotation is obtained in Cartesian coordinates according to the mapping

$$\begin{pmatrix} \tilde{x}_0 \\ \tilde{y}_0 \end{pmatrix} \longmapsto \begin{pmatrix} \tilde{x}'_0 \\ \tilde{y}'_0 \end{pmatrix} = \begin{pmatrix} \tilde{x}_0 \\ \tilde{y}_0 \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \tag{24}$$

References

[Nelder and Mead(1965)] Nelder, J. A., Mead, R., 1965. The Computer Journal 7 (4), 308–313.

[Penczek(2010)] Penczek, P. A., 2010. Image restoration in cryo-electron microscopy. Methods Enzymol 482, 35–72.

[Yang and Penczek(2008)] Yang, Z., Penczek, P. A., Aug 2008. Cryo-em image alignment based on nonuniform fast fourier transform. Ultramicroscopy 108 (9), 959–69.