Unbiased Measurement of Population Size in Cryo-EM Micrographs

Nadav Gasner[1], Marcus Brubaker[1,2]

Abstract— **The molecular environment of a cell is highly chaotic, as a given particle can have many possible configurations depending on its energy state and underlying dynamics. Determination of the probability of a biological molecule being in a specific configuration has yet to be achieved in an unbiased manner. Electron cryomicroscopy (Cryo-EM) is a recent Nobel-prize winning experimental technique which can produce detailed 3D structures of particles from 2D micrographs generated by a transmission electron microscope. Enabled by computational processes, Cryo-EM can determine structures of particles in multiple configuration or states. Our objective is to develop an algorithm which use cryo-EM generated micrographs and a particles various 3D structures to accurately measure population sizes of configurations in an unbiased manner. The initial problem of particle detection was tackled by implementing common template-matching techniques involving cross-correlations and the convolution theorem. As Cryo-EM uses low-intensity electron beams, the resultant micrographs are subject to heavy noise, which, in addition to the microscope-specific contrast transfer function, need to be addressed and cause detection to be a formidable task. For each of the located particles, the probabilities of it being from a specific configuration are calculated and compared using Bayesian inference. By counting the occurrence of different configurations over many micrographs, the result is an estimate of the relative likelihoods of the different molecular configurations in solution. This data can provide a deeper knowledge of the energetics of molecular configurations and provide greater understanding to the working environment of a cell.**

I. INTRODUCTION

The methods outlined in this paper describe the process of taking a Cryo-EM micrograph and particle stack projects in order to determine the proportions of different particle populations present. The pipeline includes a template-matching function which produces a variation of the cross-correlation between the micrograph and template projection. Next, using Bayesian Inference, the relative negative log probabilities for each particle stack are calculated. Lastly, these are passed into a sorting algorithm which determines which particle is most likely for each detection in the micrograph.

II. MATH

A. *Cross-Correlation*

Convolution is the integral of the point-wise multiplication of two functions as one is reversed and shifted:

$$P * Q = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} P(\tau)Q(t-\tau)d\tau \tag{1}$$

where τ is the sliding window of the convolution.

The convolution theorem states that a convolution in the time domain is equivalent to a multiplication in the Fourier domain:

$$\mathcal{F}[P * Q] = \mathcal{F}(P)\mathcal{F}(Q) \tag{2}$$

The Cross-correlation is similar to the convolution, with the difference that one of the functions is not reversed:

$$P \circledast Q = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^*(\tau)g(t+\tau)d\tau \tag{3}$$

where τ is the displacement. This function assesses the similarity between two functions, comparing the probe image, f, position-by-position to the test image, g. The result is a 2D matrix that has a maximal value where the probe most closely matched the test image. The easiest way to calculate this is by the convolution theorem, using the test image and flipping the probe image. The normalized cross-correlation is computed by dividing the cross-correlation by both a template normalization term, and an image normalization term:

$$ncc(I,T) = \frac{\sum_i I_i T_i}{\sqrt{\sum_i I_i^2 \sum_i T_i^2}} \tag{4}$$

But, as we have multiple structures with many projections, we do not care what specific projection the detected particle is, we calculate the error for each projection, which will be combined probabilistically to determine the likelihood of the detected particle being a specific sturcute, regardless of the projection. The error calculation for rotation j or particle k for a micrograph with noise variance σ is:

$$\mathcal{E}_{jk} = \frac{1}{2\sigma^2} \sum_i (I_i - T_i^{(j,k)})^2 \tag{5}$$

which can be expanded as follows:

$$\mathcal{E}_{jk} = \frac{1}{2\sigma^2} (\sum_i I_i^2 + \sum_i T_i^{2(j,k)} - 2\sum_i I_i T_i^{(j,k)}) \tag{6}$$

In order to calculate the error term for an entire particle over all of its projections, we combine the $\mathcal{E}_{jk}$ terms as follows:

$$\mathcal{E}_k = -\log \sum_j e^{-\mathcal{E}_{jk}} \tag{7}$$

and this is done using the LogSumExp algorithm explained below.

B. *LogSumExp*

LogSumExp is a function used in the calculation of log likelihood of a given event by means of the cross entropy

[1] York University, Toronto, Canada.
[2] Borealis AI, Toronto, Canada.

loss due to the Softmax function. A loss function represents the difference between an actual and predicted output.

$$\begin{aligned} Softmax\{x_1,...x_i...x_n\} &= \frac{e^{x_j}}{\sum_i^n e^{x_i}} \qquad (8)\\ CrossEntropyLoss &= \log(Softmax)\\ &= x_j - \log\sum_i^n e^{x_i} \qquad (9)\end{aligned}$$

The calculation of the second term quickly leads to overflow and underflow errors due to the presence of the exponential function. To counter this, we can use a trick which shifts the centre of the exponential sum, reducing the x_i terms to ones which can be handled properly:

$$\log\sum_i^n e^{x_i} = a + \log\sum_i^n e^{x_i - a} \qquad (10)$$

for an arbitrary value of a. Typically a is selected to be the maximum, which forces the other terms to 0. This cross entorpy loss calculation is a good way to produce a probability distribution.

C. Probability Theory

Basics: The probability of two events, a and b, occurring is the probability of event a occurring multiplied by the conditional probability of event b occurring given a has occurred:

$$\mathcal{P}(a,b) = \mathcal{P}(a)\mathcal{P}(b|a) \qquad (11)$$

Marginalization: We are able to take a distribution over two sets and combine one axis of the distribution, producing a marginalized distribution over one variable:

$$\mathcal{P}(a) = \int \mathcal{P}(a,b)db = \sum_b \mathcal{P}(a,b) \qquad (12)$$

Bayes's Theorem: Based on the equivalence of:

$$\mathcal{P}(a,b) = \mathcal{P}(a)\mathcal{P}(b|a) = \mathcal{P}(b)\mathcal{P}(a|b)$$

we can derive an expression for the conditional probability:

$$\mathcal{P}(a|b) = \frac{\mathcal{P}(a)\mathcal{P}(b|a)}{\mathcal{P}(b)} \qquad (13)$$

Particle Probability: Assuming a Gaussian noise model, we can determine the negative log probability of the micrograph, I, given a specific template projection, T^{jk} from the equation for a Gaussian distribution where X is the image, μ is the projection and σ is the noise variance:

$$\begin{aligned} \mathcal{P}(X|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}\\ \log\mathcal{P}(X|\mu,\sigma^2) &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(X-\mu)^2\\ \log\mathcal{P}(X|\mu,\sigma^2) &= C - \frac{1}{2\sigma^2}(X-\mu)^2\\ -\log\mathcal{P}(X|\mu,\sigma^2) &= \frac{1}{2\sigma^2}(X-\mu)^2\\ -\log\mathcal{P}(I|T^{(j,k)}) &= \mathcal{E}_{jk} = \frac{1}{2\sigma^2}(I-T^{(j,k)})^2 \qquad (14)\\ \mathcal{P}(I|T^{(j,k)}) &= Ce^{-\frac{(I-T^{(j,k)})^2}{2\sigma^2}} \qquad (15)\end{aligned}$$

Through a quick rearrangement of, it can be shown:

$$\mathcal{P}(I|T^{(j,k)}) = e^{-\mathcal{E}_{jk}} \qquad (16)$$

In order to combine all the projections of a certain template, T_k, we need to perform marginalization over the probabilities for each projection to yield $\mathcal{P}(I|T^k)$.

$$\begin{aligned} \mathcal{P}(I|T^k) &= \int \mathcal{P}(I,V|T^k)dV\\ &= \int \mathcal{P}(I|T^k,V)\mathcal{P}(V|T^k)\\ &= C\int \mathcal{P}(I|T^k,V)dV\\ &= \int \mathcal{P}(I|T^{jk})dj\\ \mathcal{P}(I|T^k) &= \sum_j e^{-\mathcal{E}_{jk}}\end{aligned}$$

This gives the folloeing calculation for the negative log probability of one template stack being:

$$-\log\mathcal{P}(I|T^k) = -\log\sum_j e^{-\mathcal{E}_{jk}} = \mathcal{E}_k \qquad (17)$$

That was the calculation of the micrograph occuring given the particle stack, however we want the opposite, the probability of the particle stack given the micrograph, in essence, the probability of the detected particle belonging to a given stack. From Bayes's Theorem we have:

$$\mathcal{P}(T^k|I) = \frac{\mathcal{P}(T^k)\mathcal{P}(I|T^k)}{\mathcal{P}(I)} \qquad (18)$$

$\mathcal{P}(T^k)$ is the prior, which we are going to assume is constant, although the derivation for known, varying priors will also be done. We know what $\mathcal{P}(I|T^k)$ is from above, and we can solve for $\mathcal{P}(I)$ via marginalization as follows:

$$\begin{aligned} \mathcal{P}(I) &= \sum_{k=1}^N \mathcal{P}(I,T^k)\\ &= \sum_{k=1}^N \mathcal{P}(T^k)\mathcal{P}(I|T^k)\\ &= \sum_{k=1}^N \mathcal{P}(T^k)e^{-\mathcal{E}_k}\end{aligned}$$

Now we have two opions:

1) $\mathcal{P}(T^k)$ is constant, which gives:

$$\begin{aligned}
\mathcal{P}(I) &= \mathcal{P}(T^k)\sum_{k=1}^{N} e^{-\mathcal{E}_k} \\
\mathcal{P}(T^k|I) &= \frac{\mathcal{P}(T^k)\mathcal{P}(I|T^k)}{\mathcal{P}(T^k)\sum_{k=1}^{N} e^{-\mathcal{E}_k}} \\
\mathcal{P}(T^k|I) &= \frac{\mathcal{P}(T^k)\sum_j e^{-\mathcal{E}_{jk}}}{\sum_{k'} \mathcal{P}(T^{k'})e^{-\mathcal{E}_{k'}}} \\
\mathcal{P}(T^k|I) &= \frac{\sum_j e^{-\mathcal{E}_{jk}}}{\sum_k e^{-\mathcal{E}_k}} \qquad (19) \\
-\log\mathcal{P}(T^k|I) &= -\log\sum_j e^{-\mathcal{E}_{jk}} + log\sum_k e^{-\mathcal{E}_k} \qquad (20)
\end{aligned}$$

2) $\mathcal{P}(T^k)$ is varied, in which case it is combined with the exponential:

$$\begin{aligned}
\sum_k \mathcal{P}(T^k)e^{-\mathcal{E}_k} &= \sum_k e^{\log\mathcal{P}(T^k)-\mathcal{E}_k} \\
\mathcal{P}(T^k|I) &= \frac{\mathcal{P}(T^k)\sum_j e^{-\mathcal{E}_{jk}}}{\sum_k e^{\log\mathcal{P}(T^k)-\mathcal{E}_k}} \qquad (21) \\
-\log\mathcal{P}(T^k|I) &= -\log\mathcal{P}(T^k) - \log\sum_j e^{-\mathcal{E}_{jk}} + \\
&\quad log\sum_k e^{\log\mathcal{P}(T^k)-\mathcal{E}_k} \qquad (22)
\end{aligned}$$

D. Contrast Transfer Function

Each microscope has a corresponding contrast-transfer function (CTF), which can be modeled as a dense matrix representing the CTF multiplying the projection of the particle, akin to a convolution. It represents the information transferred as a function of frequency. In the Fourier domain this becomes a multiplication of the template with the CTF. We go from

$$\sum(I_i - T_i)^2 \to \sum(I_i - CT_i)^2$$

By Parseval's Theorem, we can state that the magnitude of the Fourier transform is the same as in the time domain, as the Fourier transform is just a vector rotation.

$$\sum_i (I_i - CT_i)^2 = \sum_\omega ||(I_\omega - \tilde{C}_\omega\tilde{T}_\omega)||^2 \qquad (23)$$

As the squared magnitude of a complex value isn't simply the square, rather itself times its conjugate, $||c||^2 = c^*c$, we need to work out how to solve this:

$$\begin{aligned}
\sum_\omega ||(\tilde{I}_\omega - \tilde{C}_\omega\tilde{T}_\omega)||^2 &= (\tilde{I}_\omega - \tilde{C}_\omega\tilde{T}_\omega)^*(\tilde{I}_\omega - \tilde{C}_\omega\tilde{T}_\omega) \\
&= (\tilde{I}^*_\omega - \tilde{C}_\omega\tilde{T}^*_\omega)(\tilde{I}_\omega - \tilde{C}_\omega\tilde{T}_\omega) \\
&= \tilde{I}^*\tilde{I} - \tilde{I}^*\tilde{C}\tilde{T} - \tilde{I}\tilde{C}\tilde{T}^* + \tilde{C}^2\tilde{T}\tilde{T}^* \\
&= \tilde{I}^*\tilde{I} + \tilde{C}^2\tilde{T}\tilde{T}^* - 2\tilde{C}\mathbb{R}[\tilde{T}^*\tilde{I}]
\end{aligned}$$

If the transform is Unitary, it has the proper scaling to be combined with the sigma factor.

E. Equations

$$\mathcal{E}_{jk} = \frac{1}{2\sigma^2}(\sum_i I_i^2 + \sum_i T_i^{2(j,k)} - 2\sum_1 I_i T_i^{(j,k)}) \qquad (24)$$

$$\tilde{\mathcal{E}}_{jk} = \frac{1}{2\sigma^2}(\tilde{I}^*\tilde{I} + \tilde{C}^2\tilde{T}\tilde{T}^* - 2\tilde{C}\mathbb{R}[\tilde{T}^*\tilde{I}]) \qquad (25)$$

$$\mathcal{E}_k = -\log\sum_j e^{-\mathcal{E}_{jk}} \qquad (26)$$

$$-\log\mathcal{P}(T^k|I) = -\log\sum_j e^{-E_{jk}} + \log\sum_k e^{-E_k} \qquad (27)$$

$$\begin{aligned}
&-\log\mathcal{P}(T^k|I) = \\
-\log\mathcal{P}(T^k) - \log\sum_j e^{-\mathcal{E}_{jk}} &+ log\sum_k e^{\log\mathcal{P}(T^k)-\mathcal{E}_k} \qquad (28)
\end{aligned}$$

III. METHOD

A. Step-By-Step, no CTF

The process starts by a function call on a micrograph and a folder containing particle stacks of various particles in many different projections. For my testing, I've primarily used 3 different particles under the file names 26001, 27001 and 28001. 26001 is a small, rod shaped protein, 27001 is larger and spherical with a diffuse center, while 28001 is larger and spherical with a dense center. Each particle has a stack of 100 different projections, each one 256*256 padded to 1024*1024. The micrograph used was 4096*4096 with a noisy background and a random selection of the three particles with projections not from the stack.

1) The calculation of σ is performed if none is provided. This is done by the random selection of N points, the calculation of the mean of the points and the subsequent calculation of the variance of the points, by the formula:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(V_i - \hat{V})^2 \qquad (29)$$

This calculation was repeated M times and the median of the calculated variances was used as the best estimate.

2) The 2D Real-Fourier transform of the image is computed at this point. A padded image the size of the micrograph (4096*4096), consisting of a white square the size of the template (1024*1024) at its center is created, and the 2D Real-Fourier transform of it is subsequently computed.

3) Each of the stacks of particle projections, k, are dealt with separately. Depending on the memory of the computer and the size of the stack, each stack is broken down into R sub-stacks of S projections. For each of the S projections, a padded image is created being the same size as the micrograph, with the flipped projection placed in the center. The error of the given projection j of particle k is calculated as follows, with I being the micrograph, T being the given projection,

padding being the padded white square as in (1), and mask being the padded inverted projection as above:

$$\begin{aligned}
\mathcal{E}_{j,k} &= \sum_i (I_i - T_i^{(j,k)})^2 \\
&= \sum_i I_i^2 + \sum_i T_i^{2(j,k)} - 2\sum_1 I_i T_i^{(j,k)} \\
&= \frac{1}{2\sigma^2}((I^2 \circledast P)(x) + \sum_i T_i^{2(j,k)} - 2(I \circledast T^{(j,k)})(x)) \\
\text{image\_norm} &= \mathcal{IF}(\mathcal{F}(\text{micrograph}^2)\mathcal{F}(\text{padding})) \\
\text{temp\_norm} &= \sum_i T^{2(j,k)} \\
\text{cross\_corr} &= -2\mathcal{IF}(\mathcal{F}(\text{micrograph})\mathcal{F}(\text{mask})) \\
\mathcal{E}_{j,k} &= \frac{1}{2\sigma^2}(\text{image\_norm} + \text{temp\_norm} + \text{cross\_corr})
\end{aligned}$$

4) Each of these $\mathcal{E}_{j,k}$ for a particle k in sub-stack S are added by taking the negative LogSumExp of the negative errors, $\mathcal{E}_S = -LSE(-\mathcal{E}_{j,k})$.The sub-stacks are then combined by taking the negative of the above calculation, and performing LogSumExp over all the sub-stacks. This circumvents any computational restraints and results in the LogSumExp of all the errors for a give particle stack, $\mathcal{E}_k = -LSE(-\mathcal{E}_S)$.
5) For each of the particles, the calculated $\mathcal{E}_j$ are appended onto an error stack. A term representing the above calculation done with a stack of zeros, instead of particle projections, is also appended onto the error stack. The calculation for the zero stack is:

$$\mathcal{E}_0 = \frac{1}{2\sigma^2}\mathcal{IF}(\mathcal{F}(\text{micrograph}^2)\mathcal{F}(padding)) - \log(z)$$

where z is the number of projections in a stack.
6) The error stack sum is calculated by taking the LogSumExp of the negative error stack, in essence this is the LogSumExp over the errors of all projection of all particle and the zero stack. $\mathcal{E}_{sum} = LSE(-\mathcal{E}_k)$.
7) For each of the particle stacks, the corresponding negative log probability is calculated by adding the $\mathcal{E}_k$ of each particle to the $\mathcal{E}_{sum}$.

B. With CTF Functionality

The process is overall very similar to without the CTF functionality, but I will point out the differences:

1) Sigma is estimated as outlined in 1. All matrices created have complex data types.
2) There is no padded image produced, rather the CTF is calculated with the input parameters, and reshaped to be the same size as the template, in this case 1024*1024. This CTF is then padded to being the same size as the image, 4096*4096, and the 2D Fourier transform is taken with unitary normalization. The Fourier-transform of the image is also unitarily normalized.
3) Each stack is taken separately as described above, and each of the projections in the stack has its error computed. To do that, once the inverse of the template projection is padded to the size of the micrograph, its 2D Fourier transform is taken with unitary normalization. We now have the Fourier Transforms of the micrograph (I), the padded CTF (C), and the padded inverse projection (mask). We also calculate the conjugate of (I) and (mask), giving (I^*) and $(mask^*)$ We now calculate the error of the projection as follows:

$$\begin{aligned}
\tilde{\mathcal{E}}_{j,k} &= \sum_\omega ||(I_\omega - \tilde{C}_\omega \tilde{T}_\omega)||^2 \\
&= \frac{1}{2\sigma^2}(\tilde{I}^* \tilde{I} + \tilde{C}^2 \tilde{T} \tilde{T}^* - 2\tilde{C}\mathbb{R}[\tilde{T}^* \tilde{I}])
\end{aligned}$$

4) The $\tilde{\mathcal{E}}_{j,k}$ are used to calculate $\tilde{\mathcal{E}}_k$ using LogSumExp as described above.
5) The matrix appended to the error stack representing the calculation done with a zero stack is:

$$\mathcal{E}_0 = \frac{1}{2\sigma^2}(\tilde{I}^* \tilde{I}) - \log(z)$$

where z is the number of projections in a stack.
6) The error sum and probabilities are calculated for each template stack as outlined above.

IV. CONCLUSIONS