Introduction to cryoSPARC

cryoSPARC is an algorithm for rapid unsupervised cryo-EM structure determination. It consists of two algorithms - The first of these algorithms makes it possible to perform unsupervised *ab initio* 3D classification where multiple 3D states of a molecule can be determined without prior structural knowledge and without the assumption that the 3D states resemble each other. The second algorithm performs high resolution refinement rapidly based on efficiently computable heuristics.

Like the **RELION** algorithm, even cryoSPARC frames the optimization problem in the Bayesian likelihood framework.

$$\arg\max_{V_{1,..,K}} = logp(V_{1,..,K}|X_{1,..,N}) = \arg\max_{V_{1,..,K}} \sum_{i=1}^{N} log \sum_{j=1}^{K} \frac{1}{K} \int p(X_i, \phi_i|V_j) d\phi_i + logp(V_{1,..,K}) \tag{1}$$

The aim of this optimization is to find the 3D structures ($V_{1,..,K}$) that best explain the observed projections. In cryoSPARC an SGD optimization scheme is used to quickly identify several low-resolution 3D structures that are consistent with the set of observed images. Unfortunately the optimization problem above is a non-convex algorithm and starting with any random model may quickly converge to a local optima. Sensitivity to local optima is seen in most optimization algorithms, and as a result, refinement programs typically require a reasonably accurate initial model. To circumvent these problems **SGD - Stochastic Gradient descent** is proposed as a key tool for the optimization of nonconvex functions.

0.1 Stochastic Gradient Descent

SGD iteratively optimizes an objective function by computing approximate gradients and taking steps in the parameter space according to those gradients.

Objective Function - The optimization function is the log posterior probability distribution over K 3-D densities, given N particle projections, i.e,

$$\begin{aligned}
&= \arg\max_{V_1,V_2,..,V_K} logp(V_1, V_2, .., V_K|X_1, X_2, .., X_N) \\
&= \arg\max_{V_1,V_2,..,V_K} logp(X_1, X_2, .., X_N|V_1, V_2, .., V_K) + logp(V_1, V_2, .., V_K) \\
&= \arg\max_{V_1,V_2,..,V_K} \sum_{i=1}^{N} logp(X_i|\mathbf{V}) + \sum_{j=1}^{K} logp(V_j) \\
&= \arg\max_{\mathbf{V}} f(\mathbf{V})
\end{aligned} \tag{2}$$

In Equation 2, the second term is a joint prior over the 3-D structures. This prior can be set, for example, to restrict density to be strictly positive or to penalize high-frequency noise in structures. In cryoSPARC, the prior is assumed to be independent over the structures, meaning that it can factor over the structures as we did in Equation 2.

We know that,

$$\begin{aligned}
p(X_i|\mathbf{V}) &= \sum_{j=1}^{K} \pi_j p(X_i|V_j) \equiv U_i \\
p(X_i|V_j) &= \int p(X_i|\phi, V_j) p(\phi) d\phi
\end{aligned} \tag{3}$$

Therefore,

$$f(\mathbf{V}) = \sum_{i=1}^{N} log\left(\sum_{j=1}^{K} \pi_j \int p(X_i|\phi, V_j) p(\phi) d\phi \right) + \sum_{j=1}^{K} logp(V_j) \tag{4}$$

The mixing probabilities between the different 3-D structures are given by π_j and in cryoSPARC, the mixing probabilities are assumed to be uniform over all classes. i.e., $\pi_j = K^{-1}$. Also a prior over poses p_ϕ can be specified, and in this work an uniform distribution is again used.

Gradient - SGD optimizes the objective function in Equation 4 by iteratively updating the parameters $V_1, .., V_K$. The gradient of Equation 4 with respect to each structure is computed in order to take steps. The gradient is

$$\begin{aligned}\frac{\partial f}{\partial V_k} &= \sum_{i=1}^{N} \frac{1}{U_i}\frac{\partial U_i}{\partial V_k} + \frac{\partial}{\partial V_k} logp(V_k) \\ &= \sum_{i=1}^{N} \frac{1}{U_i}\pi_k \int \frac{\partial}{\partial V_k} p(X_i|\phi, V_k)p(\phi)d\phi + \frac{\partial}{\partial V_k} logp(V_k)\end{aligned} \tag{5}$$

Approximate gradient - The sum giving the gradient in Equation 5 is over all projections in the dataset. In SGD, the sum is approximated using subsampling. At each iteration, SGD selects a subset of the projections and uses only those projections to approximate Equation 5. The size of each batch $\mathbf{M}$ can vary over iterations and in cryoSPARC, it is set automatically based on the current resolution and the number of classes K. The approximate gradients are given by,

$$\frac{\partial f}{\partial V_k} \approx G_k \equiv \frac{N}{M} \sum_{i \in M} \frac{1}{U_i}\pi_k \int \frac{\partial}{\partial V_k} p(X_i|\phi, V_k)p(\phi)d\phi + \frac{\partial}{\partial V_k} logp(V_k) \tag{6}$$

SGD update rule with momentum - The approximate gradient in Equation 6 points in a direction within the space of 3-D structure that will, in expectation over random selections of minibatches, improve the objective functions in Equation 4. It is well known that in the general case, optimization of non convex functions like Equation 4 is difficult and SGD only provides guarantees of local convergence. Nevertheless, in practice, SGD performs well and finds the correct 3D structures. SGD computes the update at the current iteration, $dV_k^{(n-1)}$, by scaling the current gradient $G_k^{(n)}$ with a step size η_k and combining this linearly with the previous update in a ration given by μ. This linear averaging is known as momentum and serves to smooth the noisy approximate gradient directions in SGD.

$$\begin{aligned}dV_k^{(n)} &= (\mu)dV_k^{(n-1)} + (1-\mu)(\eta_k)G_k^{(n)} \\ V_k^{(n+1)} &= V_k^{(n)} + dV_k^{(n)}\end{aligned} \tag{7}$$

0.2 Results

On applying cryoSPARC for *ab initio* structure determination and 3D classification, we saw convergence to correct low-resolution structures from arbitrary random initialization. When applied to a data set of conformationally heterogeneous Thermus thermophilus, the algorithm discerned three different conformational states. This finding is notable as previous analysis with with reference based classification revealed only two out of the three states. This observation illustrates the importance of reference-free ab initio classification for unbiased identification of states.

0.3 Rapid refinement of maps to high resolution

The primary computational burden in map refinement is the search for orientation parameters that best align with each projection. In cryoSPARC, a branch and bound algorithm is used which accelerates this search by quickly and inexpensively ruling out large regions of the search space that cannot contain the optimum of the objective function. To find the optimal pose, an inexpensive lower bound on the error is first computed across the entire space of poses. At the pose which minimizes the lower bound, the computationally expensive true function is evaluated.

Although conceptually straightforward, application of the branch-and-bound strategy requires an informative and inexpensive lower bound for the objective function. We now proceed to derive this lower bound. In this work, the pose variable ϕ is broken down into two parts with r denoting the 3-D orientation and t denoting the 2-D translation of the projection. As is common in the literature, the probability of observing a projection from a particular pose is given in the Fourier domain as follows:

$$p(X|\phi, V) = p(X|r, t, V) = \frac{1}{Z} exp\left(\sum_l \frac{-1}{2\sigma_l^2} |C_l Y_l(r) - S_l(t) X_l|^2 \right) \tag{8}$$

where

$$Y_l(r) = \Theta_l(r) V$$

This is a log-likelihood sum over all Fourier coefficients l . $Y_l(r)$ denotes the projection of model V according to pose r, at frequency l. C denotes the contrast transfer function of the (CTF) of the microscope, $\Theta_l(r)$ is a linear projection operator, corresponding to the slice operator in Fourier space, with pose r, for wavevector l. S denotes the 2-D phase shift corresponding to a 2-D translation of t pixels. The noise parameter σ_l represents the level of Gaussian noise expected at each frequency, with a possibly different variance for each Fourier coefficient. For notational clarity in deriving the lower bound, we assume that $\sigma_l = \sigma = 1$.

Taking the negative log of Equation 8 gives the image alignment error, which is the squared error in Fourier coefficients:

$$E(r, t) = \sum_l \frac{1}{2} |C_l Y_l(r) - S_l(t) X_l|^2 \tag{9}$$

Our aim is to find the r and t that minimize this function for the given projections X and model V.

Deriving the lower bound - The derivation starts with a well-known intuition: If an image aligns poorly to a structure at low resolution, it will not align well at high resolution. This means that, if we evaluate the likelihood of an image across poses at low resolution it should give us an idea about which pose is worth pursuing at high resolution. The intuition above indicates that, if the Fourier coefficients of the model at higher resolutions have limited power, there is a limit to how much they can impact the squared error E. In cryoSPARC, inexpensive evaluations of squared error are first used to bound true values of E to eliminate search regions without evaluating Equation 9 entirely.

So we can split Equation 9 into two parts,

$$E(r, t) = \sum_{||l|| \leq L} \frac{1}{2} |C_l Y_l(r) - S_l(t) X_l|^2 + \sum_{||l|| \geq L} \frac{1}{2} |C_l Y_l(r) - S_l(t) X_l|^2 \tag{10}$$

In the branch and bound algorithm, L is initialized to a small value, and then increased with each iteration until reaching Nyquist frequency. In order to bound E we compute the first term directly which is inexpensive and bound the second term from below. To derive that bound it is convenient to bound the second term (B) into three parts:

$$\begin{aligned} B(r, t) &= \sum_{||l|| \geq L} \frac{1}{2} |C_l Y_l(r) - S_l(t) X_l|^2 \\ &= \sum_{||l|| \geq L} \frac{1}{2} |X_l|^2 + \sum_{||l|| \geq L} \frac{1}{2} C_l^2 |Y_l(r)|^2 - \sum_{||l|| \geq L} C_l R(Y_l(r) * S_l(t) X_l) \end{aligned} \tag{11}$$

First lets consider the third term - B_3. An upper bound on this term corresponds to a lower bound on B. According to the Cryo-EM image formation model,

$$X_l = C_l \bar{X}_l + \epsilon_l \tag{12}$$

Inserting Equation 12 in B_3 gives us,

$$\begin{aligned} B_3 &= \sum_{||l||\geq L} C_l^2 R(Y_l(r) * S_l(t)\bar{X}_l) + \sum_{||l||\geq L} C_l R(Y_l(r) * S_l(t)\epsilon_l) \\ &= \sum_{||l||\geq L} C_l^2 R(Y_l(r) * S_l(t)\bar{X}_l) + H \\ &\leq \sum_{||l||\geq L} C_l^2 |Y_l(r)||\bar{X}_l| + H \end{aligned} \tag{13}$$

Returning to B_2 and B_3 ans substituting Equation 13, we get:

$$\begin{aligned} B_2 - B_3 &= \sum_{||l||\geq L} \frac{1}{2} C_l^2 |Y_l(r)|^2 - \sum_{||l||\geq L} C_l R(Y_l(r) * S_l(t) X_l) \\ &\geq \sum_{||l||\geq L} \frac{1}{2} C_l^2 |Y_l(r)|^2 - \sum_{||l||\geq L} C_l^2 |Y_l(r)||\bar{X}_l| - H \\ &= \underbrace{\sum_{||l||\geq L} \frac{1}{2} (C_l^2 |Y_l(r)|^2 - 2C_l^2 |Y_l(r)||\bar{X}_l|)}_{\mathrm{Q}} - H \end{aligned} \tag{14}$$

Each term in Q is a positive-definite quadratic of $Y_l(r)$. Therefore the above can be bounded from below as:

$$Q \geq \min_{Y_l(r)} \sum_{||l||\geq L} \frac{1}{2} (C_l^2 |Y_l(r)|^2 - 2C_l^2 |Y_l(r)||\bar{X}_l|) \tag{15}$$

attained at $Y_l(r) = \bar{X}_l$

$$= \sum_{||l||\geq L} -\frac{1}{2} C_l^2 |\bar{X}_l|^2 \tag{16}$$

We know that there would be at least one corresponding pose r which would correspond to $Y_l(r^*) = \bar{X}_l$. Therefore,

$$\begin{aligned} Q &\geq \sum_{||l||\geq L} -\frac{1}{2} C_l^2 |Y_l(r^*)|^2 \\ Q &\geq \min_r \sum_{||l||\geq L} -\frac{1}{2} C_l^2 |Y_l(r)|^2 \\ &= -\max_r \sum_{||l||\geq L} \frac{1}{2} C_l^2 |Y_l(r)|^2 \end{aligned} \tag{17}$$

which is attained at $r = \hat{r}$, with $\hat{Y}_l = Y_l(\hat{r})$. As a consequence it follows that

$$Q \geq -\sum_{||l||\geq L} \frac{1}{2} C_l^2 |\hat{Y}_l|^2 \equiv \hat{Q} \tag{18}$$

Here, $\hat{Y}$ is the slice of model V that has the maximum CTF-modulated total power, as given by Equation 17. Finding this slice is simple, because it does not depend on the image X or shift t. Once $\hat{Y}$ (and the corresponding pose $\hat{r}$) is identified, the bound $\hat{Q}$ on Q is fixed. Inserting all these bounds in Equation 11 we get the final lower bound on $B(r,t)$ as:

$$B(r,t) \geq \sum_{||l||\geq L} \frac{1}{2} |X_l|^2 - \sum_{||l||\geq L} \frac{1}{2} C_l^2 |\hat{Y}_l|^2 - H \tag{19}$$

Inserting the above bound on $B(r,t)$ into Equation 10 yields a lower bound on $E(r,t)$:

$$E(r,t) \geq \sum_{||l||\leq L} \frac{1}{2}|C_l Y_l(r) - S_l(t)X_l|^2 + \sum_{||l||\geq L} \frac{1}{2}|X_l|^2 - \sum_{||l||\geq L} \frac{1}{2}C_l^2|\hat{Y}_l|^2 - H \tag{20}$$

With some derivation we can show that,

$$\begin{aligned} H &= \sum_{||l||\geq L} C_l R(Y_l(r) * S_l(t)\epsilon_l) \\ &= \mathcal{N}\left(0, \sum_{||l||\geq L} \frac{1}{2}C_l^2|Y_l(r)|^2\right) \end{aligned} \tag{21}$$

Therefore, due to the presence of H in Equation 20, the expression is a probabilistic bound on E, giving the probability of E being greater than the value of the expression. In practice, a probability of 0.999936, corresponding to four standard deviations of H, provides a threshold that serves as an upper bound on H, and hence a deterministic lower bound for E above. That is,

$$\begin{aligned} H &\leq 4\sigma_H(with probability 0.999936) \\ &= 4\sqrt{\sum_{||l||\geq L} \frac{1}{2}C_l^2|Y_l(r)|^2} \\ &\leq 4\max_r \sqrt{\sum_{||l||\geq L} \frac{1}{2}C_l^2|Y_l(r)|^2} \end{aligned} \tag{22}$$

from which it follows that,

$$H \leq 4\sqrt{\sum_{||l||\geq L} \frac{1}{2}C_l^2|\hat{Y}_l|^2} \tag{23}$$

Incorporating this, finally yields a complete lower bound on $E(r,t)$,

$$\begin{aligned} E(r,t) &\geq \sum_{||l||\leq L} \frac{1}{2}|C_l Y_l(r) - S_l(t)X_l|^2 + \sum_{||l||\geq L} \frac{1}{2}|X_l|^2 - \sum_{||l||\geq L} \frac{1}{2}C_l^2|\hat{Y}_l|^2 - 4\sqrt{\sum_{||l||\geq L} \frac{1}{2}C_l^2|\hat{Y}_l|^2} \\ &\equiv \beta_L(r,t) \end{aligned} \tag{24}$$

Equation 24, with very high probability, bounds $E(r,t)$ from below. The bound is inexpensive to compute for a particular r,t. To compute the bound, the slice of the model that has the most power is first found. Then the expression for $\beta_L(r,t)$ is used to compute values of the lower bound. As L is increased, the bound becomes more expensive but tighter. In this way, at each iteration, we reject the poses which do not come close to minima of the lower bound error and focus on the pose which minimizes this error. We then make the bound tighter by increasing L and search more extensively around the pose which gave the minima and in this way we keep refining the model at each iteration.