Bias and Variance in Cryo-Electron Microscopy

Carlos Oscar S. Sorzano, PhD[1]; Jose Maria Carazo, PhD[1]
[1]Natl. Center of Biotechnology (CSIC), Madrid, Spain

Abstract

In this poster we discuss about possible caveats along the image processing path in CryoEM and on how to avoid them in order to have a reliable 3D structure. Some of these problems are very well known in the community and we may refer to them as sample related (like specimen denaturation at interfaces or non-uniform projection geometry leading to underrepresented projection directions). The rest are algorithmic related, and while some of them have been discussed in depth in the literature, like using an incorrect choice of initial volume, there are others that have received much less attention but, however, they are fundamental in any data analysis approach. Chiefly among them we refer to instabilities in the estimation of many of the key parameters required for a correct three-dimensional reconstruction that happen all along the processing workflow and that may affect significantly the reliability of the whole process.

In the field, the term overfitting has been coined to refer to some particular kind of artifacts. We argue that overfitting is actually statistical bias in key steps of particle estimation in the 3D reconstruction process, including intrinsic algorithmic bias. We also show that common tools (FSC) and strategies (gold standard), that we normally use to detect or prevent overfitting, do not fully protect us against it. Alternatively, we propose that detecting the biases that lead to overfitting is much easier when addressed at the level of parameter estimation, rather than detecting it once we have combined the particle images into a 3D map. Parameter bias can be detected by comparing the results from multiple algorithms (or at least, independent executions of the same algorithm). Then, these multiple executions could be averaged in order to have a lower variance estimate of the underlying parameters.

Introduction

In this poster we focus on the structural bias, that is, the difference between our estimated structure, $V(\boldsymbol{r})$, and the true underlying structure, $\hat{V}(\boldsymbol{r})$. Obviously, in a single experiment we will never have access to the underlying true structure, if only because the measurement noise will cause some random fluctuation around it. We will model our observation as:

$$\hat{V}(\mathbf{r}) = V(\mathbf{r}) + \Delta V(\mathbf{r}) + \epsilon(\mathbf{r})$$

Where $\Delta V(\boldsymbol{r})$ is the structural bias, and $\varepsilon(\boldsymbol{r})$ is a random fluctuation with zero mean. The random noise, $\varepsilon(\boldsymbol{r})$, normally decreases with the number of measurements, suggesting that it does not pose a major problem in the current era of automatic acquisition of thousands of micrographs. The problem is with the bias, $\Delta V(\boldsymbol{r})$, that systematically distorts our structure preventing us from visualizing the true structure. This bias may be related to missing information, violations of the assumptions of the 3D reconstruction process, incorrect prior about the underlying structure, local minima in the search of the parameters of each image, incorrect use of the programs, software bugs, or even the 3D reconstruction workflow itself.

Methods and Materials

Experimental causes of bias:

a. Use of incorrect particles: damaged particles, different conformations, particle superposition. Particle picking algorithms have a false positive rate between 10-30%.
b. Use of incorrect symmetry: especially important for helices.
c. Missing information: missing projection directions, uneven angular distributions

Algorithmic causes of bias:

d. Initial volume: well-known cause, many algorithms try to avoid this
e. Incorrect estimation of particle parameters: The parameters to estimate are whether the particle belongs to the structure or not (class) and its orientation parameters (projection direction and in-plane shift and rotation).
   Let us assume that N_1 particles are correctly estimated giving the true structure V_1, and N_2 parameters are misestimated giving the wrong structure V_2. In a linear reconstruction algorithm, we would have

$$\hat{V}_1(\mathbf{r}) = \frac{N_1}{N_1+N_2}V_1(\mathbf{r}) + \frac{N_2}{N_1+N_2}V_2(\mathbf{r}) = V_1(\mathbf{r}) + \frac{N_2}{N_1+N_2}(V_2(\mathbf{r}) - V_1(\mathbf{r}))$$

f. The 3D reconstruction algorithm: it uses weights in real or Fourier space to combine the experimental images into a 3D structure.

Causes a-d are well known and they are avoided as much as they can by experimentalists. **Causes e and f are less known** and are the ones highlighted in this work.

Results I

Cause e. Classification

Experiment 1. The following table shows the classification results of EMPIAR 10028 by Relion on 3 classes, executed three independent times.

Run	Class 1	Class 2	Class 3
Run 1	46.3%	46.0%	7.7%
Run 2	37.2%	31.8%	31.0%
Run 3	40.1%	31.1%	28.8%

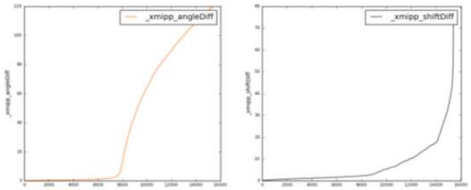
However, the sets of particles that were always classified together represent: 8.0%, 6.8%, 6.0%, 5.8%, 5.8%, 5.6%, 5.6%, 5.2%, 5.2%, 5.1%, 4.9%, 4.7%, 4.3%, 4.3%, 4.0%, 3.9%, 3.9%, 3.7%, 3.6%, and 8 groups with less than 1%.

Experiment 2. We classified five times the images of EMPIAR 10333 in two classes. The largest class contained: 58%, 71%, 87%, 92%, and 93%. The largest set of particles consistently put together was only 43%.

Results II

Cause e. Image alignment

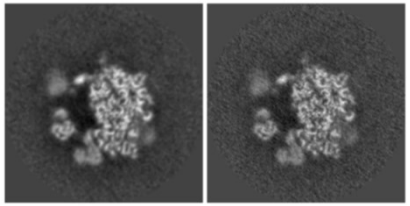
Experiment 3. The following figure shows the differences between the alignment parameters for the images in EMPIAR 10013 estimated by Relion autorefine and Xmipp highres.



Experiment 4. If we compare two independent executions of Relion autorefine on two different datasets, the angular assignment coincides between 62% (EMPIAR 10025) and 88% (EMPIAR 10028).

Cause f. Reconstruction weights

Experiment 5. We reconstructed the images in EMPIAR 10028 with Relion autorefine (left) and Relion reconstruct Fourier (right) using exactly the same angular assignment. The differences in the reconstruction correspond to the differences in weights when constructing the 3D volume in each one of the algorithms.



Conclusions

In this work we have shown that there might be significant differences between the particle parameter estimates of different algorithms, or even different runs of the same algorithm. Whichever the strategy we choose to deal with incorrect estimates at this level, identifying possible bias and reducing the variance, invariably requires multiple independent estimations of the alignment and class belonging parameters. Although not entirely protected against bias (two coincident estimates could be simultaneously biased), this approach could help to produce better, more robust and reliable 3D reconstructions of CryoEM data.

Contact

Carlos Oscar S. Sorzano
Natl. Center of Biotechnology (CSIC)
c/Darwin, 3. 28049 Madrid Spain
coss@cnb.csic.es
+34 91 585 4510