doi:10.1017/S1431927620019480
Microsc. Microanal. 26 (Suppl 2), 2020

Embedding Heterogeneous Cryo-EM Data with 3D Principal Component Analysis and Variational Autoencoders

Dimitry Tegunov

Max Planck Institute for Biophysical Chemistry, Goettingen, Niedersachsen, Germany

As function often follows form in proteins, macromolecular machines undergo compositional and conformational changes in order to perform a multitude of complex functions [1]. The ability to resolve an entire population of molecular states in a sample is one of the biggest strengths of cryo-electron microscopy coupled with single-particle analysis (SPA). Currently popular maximum likelihood methods [2] divide the data into a pre-defined number of discrete 3D classes during refinement. They can be a powerful tool for disentangling multiple states on a conformational landscape with well-defined energetic minima [3] that lead to a clearly multimodal distribution. However, in case of a unimodal or sufficiently smooth multimodal distribution, this approach will miss many states and won't be able to model smooth transitions between the captured ones. More advanced approaches for continuous motion analysis [4, 5] can only model coarse conformational changes while ignoring the composition entirely. Thus, algorithms that can calculate a lower-dimensional embedding for the entire compositional and conformational space of a macromolecular complex are needed.

Voxel intensity-based 3D principal component analysis (PCA) is a powerful tool for embedding heterogeneity [6]. However, because motion has to be represented as co-variance between voxel intensities, 3D PCA's modeling power for long-range motion is limited and many principal components are required for its accurate representation, leading to potential overfitting. Variational autoencoders (VAE) [7] for SPA data can be modeled after 3D PCA, but introduce additional non-linearity in their hidden layers to vastly improve their modeling power for conformational changes using fewer dimensions. Additional constraints can be introduced during their training to regularize the solution.

Tilde is a new tool that implements 3D PCA and VAE training for conventional 2D particle images and tilt series data. The ability to analyze and grasp the results is equally important to an accurate embedding. Here, Tilde offers an interactive graphical interface that helps the user to quickly understand the biological meaning of individual latent space dimensions and inter-dimensional relations, make sense of highly heterogeneous data sets, and communicate the analysis results efficiently. Tilde's workflow and features are demonstrated using *in vitro* data of a highly flexible transcription initiation complex, as well as *in situ* data of a 70S ribosome.

References

1. Rodnina, M.V. and W. Wintermeyer, *The ribosome as a molecular machine: the mechanism of tRNA-mRNA movement in translocation.* Biochem Soc Trans, 2011. **39**(2): p. 658-62.
2. Scheres, S.H., *A Bayesian View on Cryo-EM Structure Determination*, in *J Mol Biol*. 2012. p. 406-18.
3. Hofmann, S., et al., *Conformation space of a heterodimeric ABC exporter under turnover conditions.* Nature, 2019. **571**(7766): p. 580-583.
4. Schilbach, S., et al., *Structures of transcription pre-initiation complex with TFIIH and Mediator.* Nature, 2017. **551**(7679): p. 204-209.
5. Nakane, T., et al., *Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION.* Elife, 2018. **7**.

6. Tagare, H.D., et al., *Directly reconstructing principal components of heterogeneous particles from cryo-EM images.* J Struct Biol, 2015. **191**(2): p. 245-62.
7. Kingma, D.P. and M. Welling, *Auto-Encoding Variational Bayes.* 2013.

https://doi.org/10.1017/S1431927620019480 Published online by Cambridge University Press