University of Ljubljana

Faculty of Computer and Information Science

Večna pot 113, SI-1000 Ljubljana, Slovenia

April 21, 2013

Dear Editors,

We would like to submit our application note on “openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding” for Bioinformatics.

In the paper, we describe an open source library for t-SNE visualization that can handle massive data sets, globally align the resulting clusters, and place new data points into a constructed embedding. The t-SNE algorithm has been extensively used in presentations of multidimensional data sets from biology, but the method has been recently criticized for poor scalability and susceptibility to batch effects. Our proposed implementation addresses all these concerns and includes all the lately-proposed methodological advancements.

We illustrate the utility of our library on visualization of two single-cell datasets. In our example we show that placement of new data to constructed embedding, as implemented in our library, effectively mitigates batch effects. The t-SNE implementation we describe in the paper is open, available through GitHub, and fosters extensibility and experimentation. openTSNE runs orders magnitudes faster than comparable Python-based implementations (e.g., scikit-learn) and can handle data sets containing up to millions of data points.

This is an original manuscript that has not been considered for publication before. Bonnie Berger would be an excellent choice for Editor for this manuscript because of her broad knowledge in the field and her work in single-cell analytics. For reviewers, we kindly propose Dmitry Kobak (U Tuebingen), an author of recently proposed t-SNE tricks that we implemented in the library, Barbara di Camillo (U Padova), single-cell data scientists, or Timo Ropinski (U Ulm), data visualization specialists.

Yours faithfully,

Pavlin G. Poličar, Martin Stražar, Blaž Zupan