Cell Systems
Peer Review

Evaluation of Wu et al.: Comprehending Global and Local Structure of Single-Cell Datasets

Wenjun Kong[1,2,3] and Samantha A. Morris[1,2,3,*]
[1]Department of Developmental Biology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA
[2]Department of Genetics, Washington University School of Medicine in St. Louis, St. Louis, MO, USA
[3]Center of Regenerative Medicine, Washington University School of Medicine in St. Louis, St. Louis, MO, USA
*Correspondence: s.morris@wustl.edu
https://doi.org/10.1016/j.cels.2018.12.003

One snapshot of the peer review process for "Visualizing and interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding" (Wu et al., 2018).

Editor's Note: This is a first-round review of "Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding" by Yan Wu, Pablo Tamayo, and Kun Zhang; it was written for Cell Systems *as part of the peer review process. We chose to feature it because in addition to being remarkably fair-minded, conscientious, and incisive, this review also demonstrates powerful ways of thinking about two fundamentally important topics that slip under the radar too often—data visualization and making head-to-head comparisons properly. It also improved Wu et al. (2018) in part by giving its authors additional perspective on the core of their work. After the first round of review, Wu et al. (2018) was revised to take the reviewers' comments into account, resubmitted, re-reviewed, accepted for publication, and then published in this issue of* Cell Systems*. For comparison, an earlier version of Wu et al. was deposited on bioRxiv ahead of review and can be found here: http://biorxiv.org/lookup/doi/10.1101/276261. Wenjun Kong and Samantha Morris blinded their identities during the peer review process but have chosen to reveal them here. Wu et al. support the publication of this Peer Review; their permission to use it was obtained after their paper was officially accepted. This Peer Review was not itself peer reviewed. It has been lightly edited for stylistic polish and clarity. No scientific content has been substantively altered.*

In their manuscript, Wu et al. describe a new visualization technique for single-cell RNA-seq datasets, "similarity weighted nonnegative embedding" (SWNE). Prior to this method, single-cell datasets have commonly been visualized and interpreted via t-distributed stochastic neighbor embedding (t-SNE). However, although t-SNE captures the local structure of the data, global structures are lost. This study uses an alternative approach to t-SNE, supporting improved capture of both the global and local structure. This new visualization technique adapts and combines various different methods, such as matrix factorization with imputation, shared nearest neighbor (SNN) network construction, and Sammon mapping. This integration enables the projection of high-dimensional data into two dimensions, with accurate relative positioning of the data points.

Overall, the approach described in this manuscript is elegant and potentially impactful, although the power of this technique could be better demonstrated. The major comparison that has been performed to evaluate SWNE is t-SNE. Although t-SNE could not capture the global structure, it succeeds in capturing the clustering structure within the single-cell dataset. If this new method, SWNE, has been developed primarily for the purpose of clustering, comparisons with t-SNE and other clustering methods would be sufficient. However, since the clustering power of t-SNE is robust, the necessity for further improvement isn't immediately obvious. Therefore, if the authors could identify and showcase other applications of this method (for example trajectory inference), and then compare it with a broader range of existing methods for those applications (such as Monocle), the paper would more powerfully demonstrate why this method would be beneficial for the analysis of single-cell gene expression datasets.

In addition to these broad concerns, we have some specific comments regarding the methodology.

(1) For the optimal construction of factorization matrices, approximately 20% to 25% of the gene expression matrix is randomly selected by the authors and set to "unknown." Our concern is that random selection and imputation of the matrix may change the biological interpretation of the dataset. Specifically, although the mean squared error (MSE) is minimized, the factorization could result in a perturbed gene expression matrix. For instance, a biologically true zero indicating lack of expression of a gene could be imputed to have non-zero values. If such improper imputation is indeed a problem, the deviation or error comparing imputed to original values is likely to be affected. It would be helpful if the authors would provide the error or percent error for each dropped value after factorization to see if it is within an acceptable range.

(2) For weighted sample embedding, based on the equations provided, it seems probable that different cells have the same coordinates as the factors. For example, if there are three factors with two cells possessing the coordinates (4.2, 0, 0) and (3.1, 0, 0) in the H matrix, then both cells would have the coordinates of factor 1, although they have somewhat different embeddings. Would these cells be considered the same in space or they could be separated via subsequent smoothing?

(3) The SNN network appears to play a critical role in SWNE. Without SNN, it seems that SWNE could be

outperformed by principal-component analysis (PCA). In other methods based on SNN, t-SNE results can be sensitive to the point at which graph-based clustering is run. For example, in Seurat, graph-based clustering (also based on SNN) is usually conducted before executing t-SNE to provide information for more effective t-SNE clustering. However, for the script previously provided for testing the efficacy of this SWNE-based approach, it appears that this graph-based clustering step with SNN is not executed for the datasets used for testing prior to t-SNE. As explained above, graph-based clustering can largely facilitate the performance of t-SNE. Would the results from t-SNE together with SNN be more comparable with SWNE? Providing some clarification about cluster identification when t-SNE is executed would be helpful.

(4) Based on Figure S2, both PCA and Multidimensional Scaling (MDS) appear to perform well, relative to SWNE, both with respect to separate different groups with relatively correct distances in different datasets and for inferring the major trajectory. Although the authors refer to this in the paper, it would be better if the authors consider another example that demonstrates the distinct power and superiority of the SWNE method.

Finally, we have some minor suggestions.

(1) For Figure 1B, Figure S1A could be more explanatory for the overview of the method, since it does mark the "NAs." It doesn't quite make sense to us when there is an arrow pointing to Figure 1B.

(2) For gene expression and pseudotime overlaid on the SWNE plot in Figure 3, what is the range of values for the coloring scheme? It only had the color bar within the figure and the scale is not described in either the figure or legend currently.

(3) For the runtime limitation of this application, would it be helpful to implement parallelization for nonnegative matrix factorization (NMF)?

REFERENCES

Wu, Y., Tamayo, P., and Zhang, K. (2018). Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding. Cell Syst. *7*, this issue, 656–666.