Learning circadian times of gene expressions

Eduardo García-Portugués • Department of Statistics, UC3M
Yolanda Larriba • Department of Statistics, Universidad de Valladolid

Inferring circadian time from gene expression is key to uncovering the temporal organization of biological processes across the 24 h cycle and to enabling time-aware applications (e.g., optimal drug dosing). The task is to estimate an internal biological time for samples without explicit time-of-day labels, a common limitation in large transcriptomic datasets, especially human, where ethical and clinical risks of biopsies often necessitate reliance on post-mortem samples.

Two methodological families have emerged toward this goal: dimensionality-reduction methods and autoencoder-based approaches. Early work with Circular Principal Component Analysis (CPCA) (Scholz, 2007) learned a circular latent representation using a multilayer perceptron with a circular bottleneck, i.e., a component layer constrained to a closed curve to capture circular structure. Within this line, CYCLOPS (Anafi et al., 2017) is a CPCA-inspired circular autoencoder to recover latent circadian phase and reveal circular/elliptical structure in high-dimensional data. Its successor, CYCLOPS 2.0 (Hammarlund, 2025), further incorporates covariates (e.g., batch, tissue, disease status) and adds regularization to stabilize inference under sparse or heterogeneous sampling, improving accuracy and robustness in human expression data. As an alternative formulation, Larriba et al. (2020) proposed ORI-TSP, which estimates temporal ordering via an order-restricted optimization problem solved as a Traveling Salesman Problem. From a dimensionality-reduction perspective, Larriba et al. (2023) introduced CIRCUST, whose core methodology uses the top two eigengenes and projects their scores onto the circle as a simple and efficient implementation of classical CPCA (hereafter CPCA*). Other nonlinear embeddings, such as circular variants of the widely used t-SNE algorithm (Maaten & Hinton, 2008; Rodríguez García, 2022), offer additional promising directions. Despite this diversity, a unified, simulation-based comparison of these techniques under realistic conditions remains lacking.

This thesis aims to review, implement, and compare state-of-the-art methods for learning circadian phases from gene-expression data. The methods to be studied include CPCA, CPCA*, CYCLOPS 2.0, ORI-TSP, and circular t-SNE. The project will involve implementing in R the most recent Julia version of CYCLOPS 2.0 (https://github.com/ranafi/CYCLOPS-2.0) and developing reproducible interfaces across methods. A systematic benchmarking study will evaluate accuracy, robustness, and computational efficiency using simulated and real transcriptomic datasets. Real data will include both active (time-labelled) and passive (post-mortem) expression profiles, enabling validation of inferred circadian times and assessment of each method's capacity to recover biologically meaningful rhythmic structure.

References

Anafi, R. C., Francey, L. J., Hogenesch, J. B., & Kim, J. (2017). CYCLOPS reveals human transcriptional rhythms in health and disease. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(20), 5312–5317. https://doi.org/10.1073/pnas.1619320114

Hammarlund, J. A. (2025). *CYCLOPS 2.0: Improving and applying circadian phase reconstruction in confounded datasets* [PhD thesis, Drexel University]. https://doi.org/10.17918/00011113

Larriba, Y., Mason, I. C., Saxena, R., Scheer, F. A. J. L., & Rueda, C. (2023). CIRCUST: A novel methodology for temporal order reconstruction of molecular rhythms; validation and application towards a daily rhythm gene expression atlas in humans. *PLOS Computational Biology*, *19*(9), e1011510. https://doi.org/10.1371/journal.pcbi.1011510

Larriba, Y., Rueda, C., Fernández, M. A., & Peddada, S. D. (2020). Order restricted inference in chronobiology. *Statistics in Medicine*, *39*(2), 265–278. https://doi.org/10.1002/sim.8397

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

Rodríguez García, L. Á. (2022). *t-Stochastic neighbor embedding on the polysphere [Master's thesis]*. Universidad Carlos III de Madrid.

Scholz, M. (2007). Analysing periodic phenomena by circular PCA. *Bioinformatics Research and Development (BIRD 2007)*, *4414*, 38–47. https://doi.org/10.1007/978-3-540-71233-6_4