Exploring Statistical and Computational Workflows in Circadian Metabolomics

Neehal Hussain[1], Dr. Aalim Weljie[2], Dania Malik[2]

[1]University of Pennsylvania (EAS' 23), [2]Department of Pharmacology and Translational Therapeutics

What is Metabolomics?

Metabolomics is:

- The large-scale study of **small molecules** (metabolites) in cells, biofluids, tissues, or organisms
- The study of the **reactants and products** of metabolism (the chemical processes occurring within a biological system)**,** which are influenced by genetic and environmental factors (Figure 1)
- A powerful, direct approach that displays the **underlying biochemical activity** and state of cells and tissues
- Divided into two types: **untargeted** (discovery-based) and **targeted** (validation-based) (Figure 2)

Metabolites are:

- Small molecules ranging in size of 25-5000 atomic mass units (Figure 3)
- The building blocks to the complex series of reactions (**pathways**) that occur in the body, where a product of one reaction becomes the reactant of the next
- Continuously interacting with each other within and between biological systems (Figure 4); the complete set of metabolites at any given time point is the **metabolome**

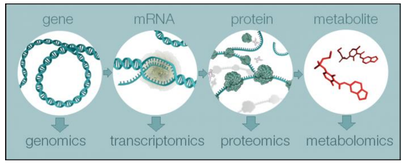

Figure 1: Overview of the four major "omics" fields, from genomics to metabolomics[1]

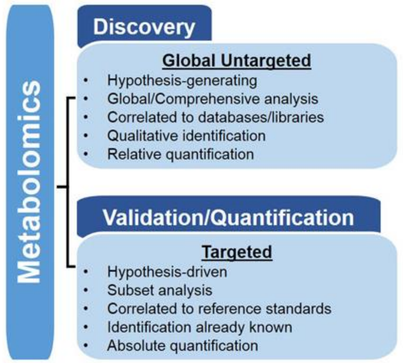

Figure 2: Comparison between untargeted and targeted metabolomics[2]

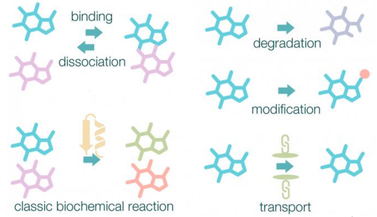

Figure 4: Examples of biochemical reactions that can occur as part of pathways[1]

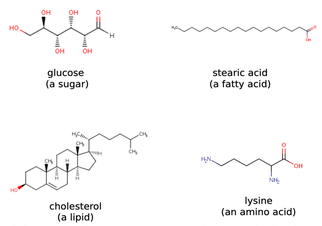

Figure 3: Examples of different classes of metabolites present in biological systems[1]

What is Chronobiology?

Chronobiology is:

- the study of cyclic phenomena (**biological rhythms**) in living organisms and their adaptations to solar- and lunar-related rhythms, including **infradian** (>24 hours), **ultradian** (<24 hours), and – the most highly researched – **circadian** (~24 hours) rhythms

Circadian rhythms are:

- Approximately **24-hour** cycles in the physiological processes of living beings (plants, animals, fungi, etc.)
- **Internally** generated, but can be modulated by environmental cues, especially sunlight
- Affecting brain wave activity, metabolite production, cell regeneration, and many other biological processes

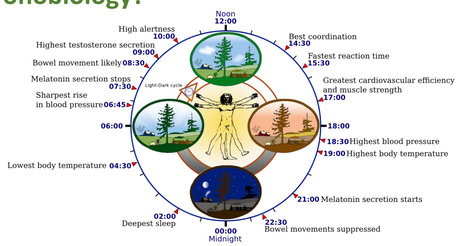

Figure 5: Representation of typical circadian rhythm of high-level bodily functions[3]

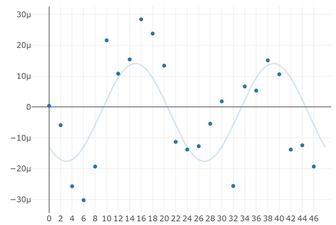

Figure 6: Metabolic time course and corresponding cosine regression curve (x: hrs, y: metabolic expression)[4,6]

Figure 7: Standard metabolomics experiment workflow[5]

Project 1: Determining Circadian Behavior in Metabolite Concentration from Time Series Data

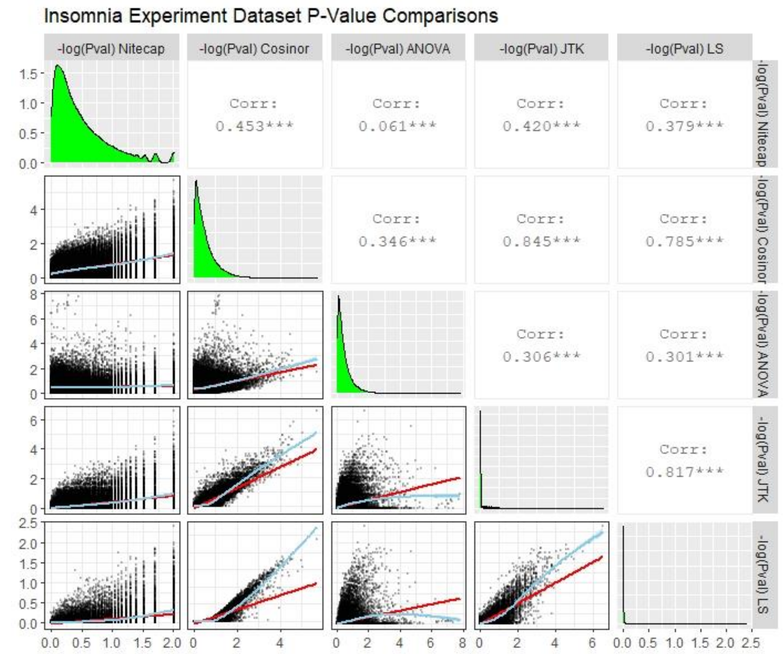

Figure 8: Comparison matrix of the outputted p-values by different circadian detection algorithms[4,7] for approximately *90000 time courses* (each time course with *24 time points sampled evenly across 48 hours*); time courses were sourced from insomnia experiment[6] with *4000 metabolites across 44 subjects* (note some time courses are not considered since they had missing time points and would cause some algorithms to fail as a result). *Upper*: correlation coefficients for relationship between p-values (note they are scaled with base 10 logarithm) across the pairs of algorithms. *Lower:* scatterplots and fitted curves (linear regression, LOESS regression) comparing p-values. *Diagonal*: p-value distributions for each algorithm. Note: $p < 0.05 \rightarrow -\log_{10} p > 1.3$ and $p < 0.01 \rightarrow -\log_{10} p > 2$

Observations:

- P-Value distribution has **similar shape** across the 5 different algorithms (as expected, most metabolites don't exhibit circadian behavior while relatively few do)
- Nitecap algorithm seems to **group p-values together** (as indicated by vertical lines); behaves like q-values
- **LOESS curve** fits the scatterplot data **better than linear** in most cases (outlier: ANOVA vs LS), as the linear regression curve emphasizes the concentration of the non-significant p-values (lower left of each graph)
- One metabolic time course can easily be deemed **significant by one algorithm, but not by another** (as displayed by the presence of points in the upper left or lower right quadrants) in the scatterplot graph
- **JTK, LS, and Cosinor** behave very similarly (r-squared coefficients **above 0.75**) compared to Nitecap and ANOVA (r-squared coefficients **below 0.45**)

Project 2: Consolidating Annotation Workflow

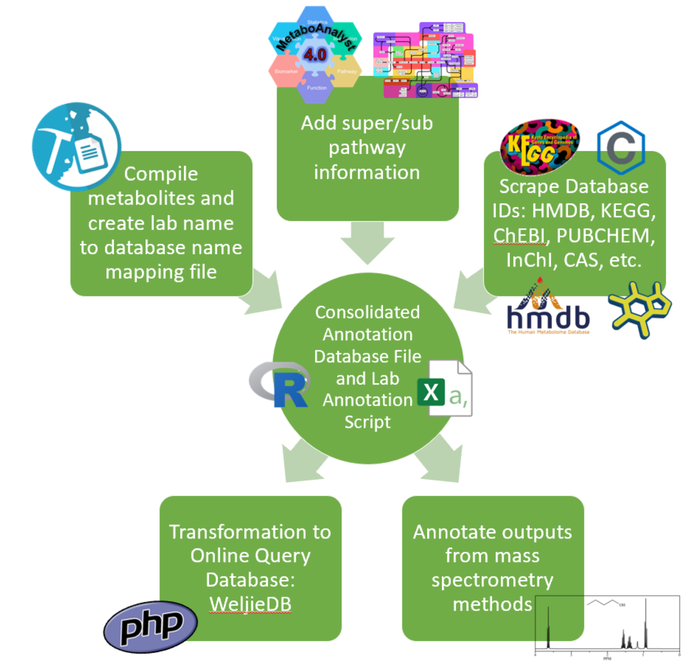

Figure 9: Diagram detailing newly automated annotation workflow

Takeaways

Project 1:

- We still do not have a ground truth about whether a given metabolite is actually circadian or not, particularly due to (1) the issue that different algorithms that attempt to detect cycling differ in their p-value outputs and (2) the nature and definition of p-values themselves.
- Ultimately, all we can conclude is that the circadian rhythm clearly does influence metabolite concentrations, for if there was no influence (null hypothesis) we would expect a uniform, not exponential, p-value distribution for each algorithm.

Project 2:

- There exist many metabolites that have IDs in some databases but do not have IDs in others; hence it was paramount that we scrape multiple databases to gain as much information as possible.
- Many identical metabolites take different names across the lab and across different databases; hence a mapping file was needed to account for these multiple names; in essence, text analysis was needed in a place relatively unexpected, and will continue to be needed with the ongoing development of WeljieDB.

References

[1]"What Is Metabolomics?" What Is Metabolomics | Metabolomics, EMBL-EBI, www.ebi.ac.uk/training-beta/online/courses/metabolomics-introduction/what-is/.
[2]"Metabolomics Research Introduction, Applications, Sample Types and Handling." Center for Innovative Technology, Vanderbilt University, 4 Nov. 2019, www.vanderbilt.edu/cit/introduction-metabolomics-research/.
[3]"Tick Tock: New Clues About Biological Clocks and Health." National Institute of General Medical Sciences, U.S. Department of Health and Human Services, www.nigms.nih.gov/education/Inside-Life-Science/Pages/tick-tock-new-clues-about-biological-clocks-and-health.aspx.
[4]Grant, Gregory, et al. Nitecap, University of Pennsylvania Institute of Translational Medicine and Therapeutics, nitecap.org/.
[5]"Metabolomics for Biomarker Discovery." The Britz-McKibbin Laboratory, britz.mcmaster.ca/research/directed-metabolite-analyses.
[6]Gehrman, Philip, et al. "Altered Diurnal States in Insomnia Reflect Peripheral Hyperarousal and Metabolic Desynchrony: a Preliminary Study." OUP Academic, Oxford University Press, 7 Mar. 2018, academic.oup.com/sleep/article/41/5/zsy043/4924205.
[7]Wu, Gang, et al. Introduction to MetaCycle, 18 Apr. 2019, cran.r-project.org/web/packages/MetaCycle/vignettes/implementation.html.

Acknowledgements

I would like to thank Dr. Aalim Weljie and Penn Undergraduate Research Mentoring for giving me the opportunity to research with the Weljie Lab this summer and beyond. I would like to thank the lab team, especially researchers Dania Malik and Arjun Sengupta for guiding me through understanding the workflows of the lab and giving me the ability to use their past papers' data as baselines for these two projects. I would also like to thank mathematician Dr. Gregory Grant and his team for giving me a deper mathematical perspective into metabolomics and becoming a continued mentor as I continue my studies in mathematics. Finally, I would like to thank Dania (again) for giving me a great poster template for this CURF Research Expo.