ECHO v4.0: Methods and Comparisons

Hannah De los Santos[1,2], Kristin P. Bennett[1,2,3], and Jennifer M. Hurley[4,5]

May 2020

[1]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, U.S.A
[2]Institute of Data Exploration and Applications, Rensselaer Polytechnic Institute, Troy, NY, U.S.A
[3]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, U.S.A
[4]Department of Biological Sciences, Rensselaer Polytechnic Institute, Troy, NY, U.S.A
[5]Center for Biotechnology and Interdisciplinary Sciences, Rensselaer Polytechnic Institute, Troy, NY, U.S.A

1 Summary

In our work for MOSAIC[1][1], we have developed improved starting points for ECHO models. As such, we have integrated these starting points for the ECHO model for the newest version, v4.0. Methods for these improved starting appear in subsequent sections, as well as comparisons with datasets used in the *Bioinformatics* publication. Due to these improvements, we highly recommend that new users use ECHO v4.0+. However, the journal version will still be available for those who prefer it[2].

2 Improved ECHO Starting Points

The ECHO model is as follows:

$$x(t) = Ae^{\frac{-\gamma t}{2}}\cos(\omega t + \phi) + y \tag{1}$$

Improved starting points for the ECHO model (1) are similar to those specified in [2], with some enhancements to both the starting points and the data. First, the averaged data is smoothed with a 1-2-1 weighting scheme described in [2]. Then starting points are calculated based on the smoothed averaged data as follows:

$$A_0 = \begin{cases} |\text{peaks}(1)|, & \text{if \# of peaks} > 0 \\ \max(\overline{x_s(t)}) - y_0, & \text{otherwise} \end{cases} \tag{2}$$

$$\gamma_0 = \begin{cases} \frac{1}{\sqrt{1+(\frac{2\pi}{\delta})^2}}, & \text{if \# of peaks } \geq 2 \text{ \& peak(1) } > \text{ peak (2)} \\ \frac{-1}{\sqrt{1+(\frac{2\pi}{\delta})^2}}, & \text{if \# of peaks } \geq 2 \text{ \& peak(1) } < \text{ peak (2)} \\ 0.01, & \text{if peaks } = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\omega_0 = 2\pi / \begin{cases} (\text{\# of time points})(\text{resolution})(\text{\# of peaks}), & \text{if \# of peaks} > 1 \\ (\text{\# of time points})(\text{resolution})(\text{\# of peaks} + 1), & \text{if \# of peaks} = 1 \\ (\text{\# of time points})(\text{resolution})(H + L)/2, & \text{if \# of peaks} = 0 \end{cases} \tag{4}$$

$$y_0 = \overline{\overline{x_s(t)}} \tag{5}$$

where # is number, $\overline{x_s(t)}$ is the smoothed averaged data, and resolution is the difference, in hours, between time points. Similarly to [2], we calculate ϕ_0 by partitioning the possible phase

[1]https://github.com/delosh653/MOSAIC
[2]https://github.com/delosh653/ECHO/releases/tag/v3.22

range into 12 parts such that $\phi_0 = \frac{i\pi}{6}, i = 0, ..., 11$. Then, using the other previously calculated initial values, we choose the ϕ_0 that minimizes the absolute value of the the difference between $\overline{x_s(t)}$ and proposed initial fit at certain time points. The compared values are calculated at time points from the beginning, middle, and end of the time course: time points 2 and 3, $\lfloor \frac{(\#\text{ of time points})}{2} \rfloor$ and $\lfloor \frac{(\#\text{ of time points})}{2} \rfloor + 1$, and (# of time points) − 2 and (# of time points) − 1.

Equations (2, 3, 4) depend on the peaks vector, which is calculated from $\overline{x_s(t)}$. A value is determined to be a "peak" by being either the maximum of minimum of a set consisting of the value and its surrounding points; i.e., by being a peak or a trough. These adjacent surrounding points are the same as those stated in [2].

The final vector of "peaks" is calculated by comparing the vector of all peaks and all troughs found across the time course for consistency, defined by the following:

$$C = -|\tau_{est} - \tau_{eff}| \tag{6}$$

$$\tau_{est} = \frac{\text{time course length(hours)}}{\#\text{of peaks}} \tag{7}$$

$$\tau_{eff} = \text{mean}_{t_j > t_i}(\text{peak}_j - \text{peak}_i) \tag{8}$$

where C is consistency, τ_{est} is estimated period, τ_{eff} is effective period, and a "peak", in this notation, corresponds to a peak or a trough. Consistency is calculated for both the set of peaks and the set of troughs.

If troughs have a higher consistency than peaks, or there is one peak or less and the number of troughs is greater than the number of peaks, then the troughs are selected for the "peaks" vector. Otherwise, the peaks are selected. By changing this "peaks" vector calculation, we more accurately account for changes in phase and mitigate noise effects.

The calculation of the logarithmic decrement, as included in (3), has been amended to the following:

$$\delta = \begin{cases} \delta = \ln(\frac{x_s(t)}{x_s(t+T)}) & \text{if peak(1) } > \text{ peak (2)} \\ \delta = \ln(\frac{x_s(t+nT)}{x_s(t)}) & \text{otherwise} \end{cases} \tag{9}$$

where t is the time of the first peak and $t + T$ is the time of the next peak. The previous iteration of the logarithmic decrement underestimated the effects of damping or forcing, respectively.

After all initial parameters are calculated, ECHO's starting points are then fit using nonlinear least squares as in [2], with an update to using the `nls.lm` function from the `minpack.lm` package [3]. This function removes the possibility of "nonstarter" results, which previously had too much noise, as this function always converges.

3 Comparisons to Previous Datasets

In our work, we compared ECHO to 6 publicly available 48-hour datasets: transcriptomic and proteomic datasets gathered from *Neurospora crassa* [5, 6], a transcriptomic dataset of pooled liver samples taken *in vivo* from mice [4], an *in vitro* transcriptomic dataset with NIH3T3 mouse fibroblast cells [4], and two transcriptomic datasets of *Anopheles gambiae* heads with differences based on lighting scheme (complete darkness [DD] and 12 hours of light/12 hours of darkness [LD]) [7]. We used the same processing as described in [2] for all comparisons.

3.1 Comparisons of Total Recovered Circadian CCEs

Tables 1 to 6 contain counts of circadian CCEs (clock-controlled elements), as well as the AC categories for all CCEs, for the journal ECHO version and ECHO v4.0. A BH-adjusted p-value cutoff of 0.05 was used for all counts.

ECHO v4.0 was able to recover more circadian CCEs in 4 out of 6 datasets, and for those where it lost circadian genes, most were close to the AC coefficient cutoff for overexpressed/repressed CCEs. It should also be noted that, in all cases, the amount of forced rhythms increased. This was due to the fact that the whole rhythm is considered for various starting points; previous starting points were based on the beginning of the rhythm, which created more bias towards damped rhythms.

Count	Journal (Before)	ECHO v4.0 (After)	Change (After - Before)
Circadian	2405	2443	38
Damped	811	781	-30
Forced	623	722	99
Harmonic	971	940	-31

Table 1: **Comparisons of *Bioinformatics* ECHO version and ECHO v4.0 counts for *Anopheles* DD data.** Counts of circadian, damped, forced, and harmonic CCEs for *Anopheles* DD data [7].

Count	Journal (Before)	ECHO v4.0 (After)	Change (After - Before)
Circadian	3277	3282	5
Damped	687	657	-30
Forced	777	816	39
Harmonic	1813	1809	-4

Table 2: **Comparisons of *Bioinformatics* ECHO version and ECHO v4.0 counts for *Anopheles* LD data.** Counts of circadian, damped, forced, and harmonic CCEs for *Anopheles* LD data [7].

Count	Journal (Before)	ECHO v4.0 (After)	Change (After - Before)
Circadian	5036	4978	-58
Damped	1372	1311	-61
Forced	1944	2005	61
Harmonic	1720	1662	-58

Table 3: **Comparisons of *Bioinformatics* ECHO version and ECHO v4.0 counts for mouse liver data.** Counts of circadian, damped, forced, and harmonic CCEs for mouse liver data [4].

Count	Journal (Before)	ECHO v4.0 (After)	Change (After - Before)
Circadian	3121	2877	-244
Damped	2088	1920	-168
Forced	490	533	43
Harmonic	543	524	-19

Table 4: **Comparisons of *Bioinformatics* ECHO version and ECHO v4.0 counts for NIH3T3 mouse fibroblast data.** Counts of circadian, damped, forced, and harmonic CCEs for NIH3T3 mouse fibroblast data [4].

Count	Journal (Before)	ECHO v4.0 (After)	Change (After - Before)
Circadian	6685	6765	80
Damped	3428	3309	-119
Forced	1069	1318	249
Harmonic	2188	2138	-50

Table 5: **Comparisons of *Bioinformatics* ECHO version and ECHO v4.0 counts for *Neurospora crassa* transcript data.** Counts of circadian, damped, forced, and harmonic CCEs for *Neurospora crassa* transcript data [5].

Count	Journal (Before)	ECHO v4.0 (After)	Change (After - Before)
Circadian	2146	2295	149
Damped	935	907	-28
Forced	594	715	121
Harmonic	617	673	56

Table 6: **Comparisons of *Bioinformatics* ECHO version and ECHO v4.0 counts for *Neurospora crassa* protein data.** Counts of circadian, damped, forced, and harmonic CCEs for *Neurospora crassa* protein data [6] .

3.2 Heat Maps and AC Coefficient Density Plots

Heat maps and AC coefficient density graphs for *Neuospora crassa* and mouse datasets, as in Figure 3 of [2], appear in Figure 1. Overall distributions remain largely unchanged, with a slight increase in forced CCEs throughout all datasets.

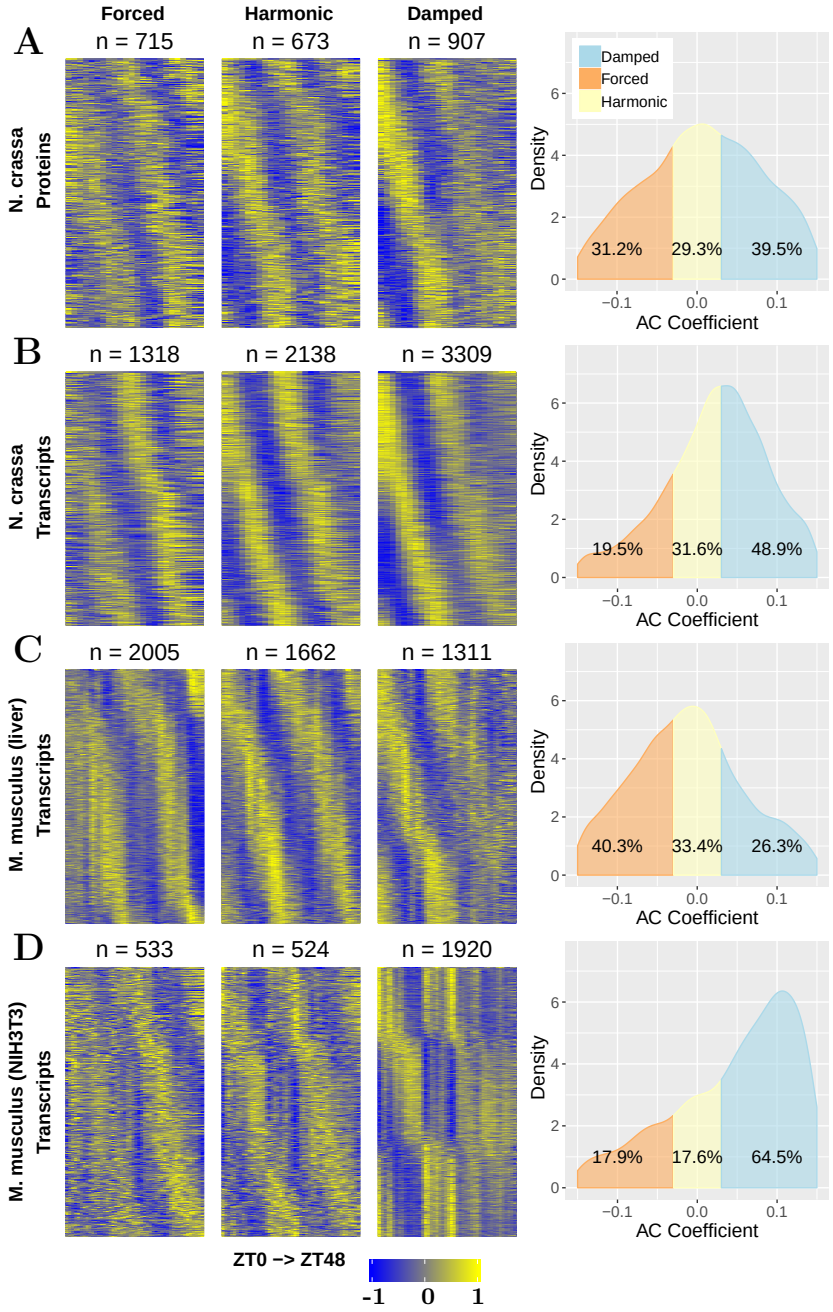

Figure 1: **The ratio of harmonic to damped to forced CCEs depends on sampling conditions, as in Figure 3 of De los Santos, et al. (2020) [2]** Heatmaps and AC coefficient density graphs of the CCEs determined by ECHO to be circadian in A. the *N. crassa* proteome [6], B. the *N. crassa* transcriptome [5], C. the *M. musculus* liver transcriptome [4], and D. the *M. musculus* NIH3T3 transcriptome [4]. For each dataset, the heat maps show mean-centered normalized expression values at a given time point for the transcripts that fall into the AC coefficient categories damped, forced, or harmonic, and are sorted vertically by phase.

3.3 Anopheles Comparisons

Heat maps, AC coefficient density graphs, a Venn Diagram, and confusion matrix for the *Anopheles gambiae* datasets, as in Figure 4 of [2], appear in Figure 2. Overall distributions remain largely unchanged, with a slight increase in forced CCEs throughout both datasets.

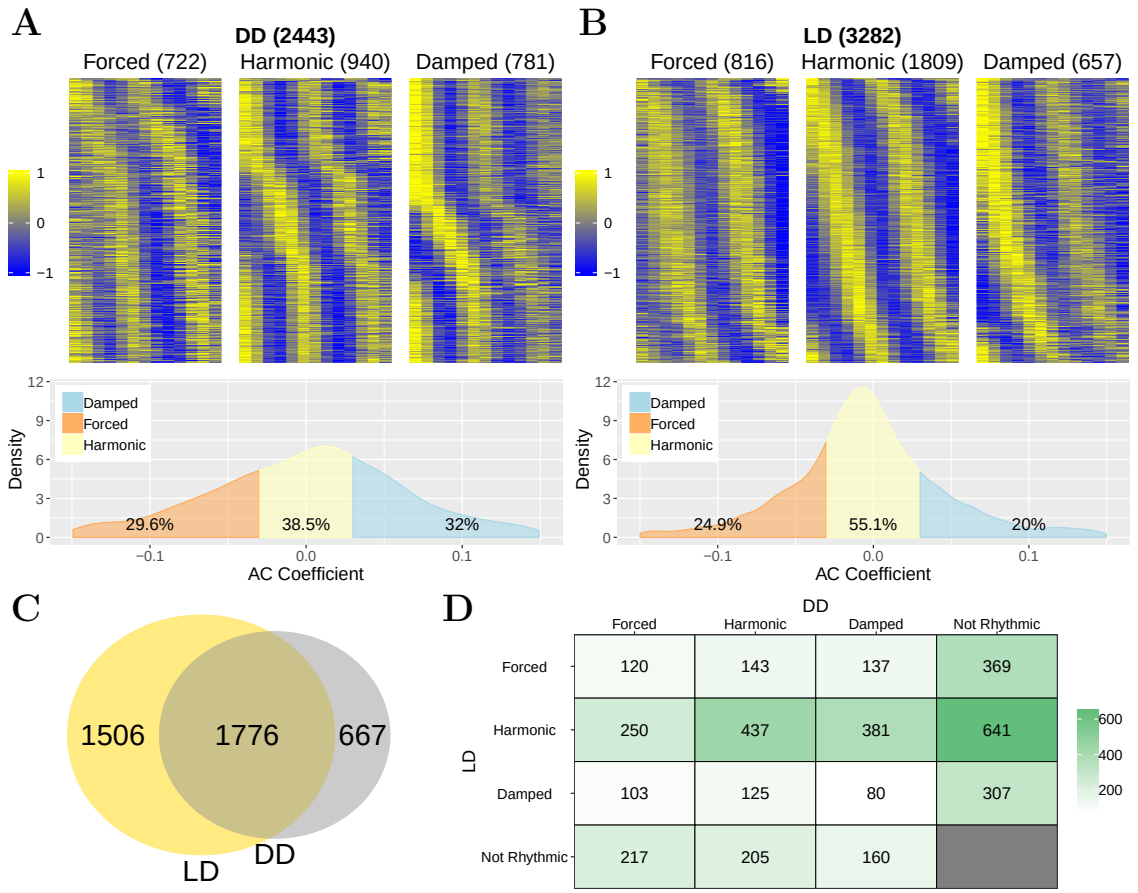

Figure 2: **The ratio of damped, forced, and harmonic CCEs varies depending on lighting schemes in *Anopheles gambiae*, as in Figure 4 of De los Santos, et al. (2020) [2]** A. and B. Heat maps and AC coefficient density graphs of the transcripts defined as damped, forced, or harmonic by ECHO in *Anopheles gambiae* (mosquito) heads gathered in either complete darkness (DD) (A) or 12:12 Light/Dark (LD) (B) [7]. C. A Venn diagram describing the overlap of transcripts found to be rhythmic in Anopheles in DD and LD. D. A confusion matrix comparing the best fit models (damped, forced, and harmonic) for CCEs in Anopheles identified as circadian by ECHO in either DD or LD conditions.

References

[1] Hannah De los Santos, Kristin P. Bennett, and Jennifer M. Hurley. MOSAIC: a joint modeling methodology for combined circadian and non-circadian analysis of multi-omics data. *bioRxiv*, 2020.

[2] Hannah De los Santos, Emily J Collins, Catherine Mann, April W Sagan, Meaghan S Jankowski, Kristin P Bennett, and Jennifer M Hurley. ECHO: an application for detection and analysis of oscillators identifies metabolic regulation on genome-wide circadian output. *Bioinf.*, 36(3):773–781, February 2020.

[3] Timur V. Elzhov, Katharine M. Mullen, Andrej-Nikolai Spiess, and Ben Bolker. *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds*, 2016. R package version 1.2-1.

[4] Michael E. Hughes, Luciano DiTacchio, Kevin R. Hayes, Christopher Vollmers, S. Pulivarthy, Julie E. Baggs, Satchidananda Panda, and John B. Hogenesch. Harmonics of circadian gene transcription in mammals. *PLoS Genet.*, 5(4):e1000442, April 2009.

[5] Jennifer M. Hurley, Arko Dasgupta, Jillian M. Emerson, Xiaoying Zhou, Carol S. Ringelberg, Nicole Knabe, Anna M. Lipzen, Erika A. Lindquist, Christopher G. Daum, Kerrie W. Barry, Igor V. Grigoriev, Kristina M. Smith, James E. Galagan, Deborah Bell-Pedersen, Michael Freitag, Chao Cheng, Jennifer J. Loros, and Jay C. Dunlap. Analysis of clock-regulated genes in Neurospora reveals widespread posttranscriptional control of metabolic potential. *Proc. Natl. Acad. Sci.*, 111(48):16995–17002, October 2014.

[6] Jennifer M. Hurley, Meaghan S. Jankowski, Hannah De los Santos, Alexander M. Crowell, Samuel B. Fordyce, Jeremy D. Zucker, Neeraj Kumar, Samuel O. Purvine, Errol W. Robinson, Anil Shukla, Erika Zink, William R. Cannon, Scott E. Baker, Jennifer J. Loros, and Jay C. Dunlap. Circadian proteomic analysis uncovers mechanisms of post-transcriptional regulation in metabolic pathways. *Cell Syst.*, 7(6):613–626.e5, December 2018.

[7] Samuel S. C. Rund, Tim Y. Hou, Sarah M. Ward, Frank H. Collins, and Giles E. Duffield. Genome-wide profiling of diel and circadian gene expression in the malaria vector Anopheles gambiae. *Proc. Natl. Acad. Sci.*, 108(32):E421–E430, June 2011.