TEMPORAL STRUCTURE IN THE YEAST TRANSCRIPTOME

JACKSON POND, ISAAC SMITH, HENRY TULLIS, WYATT WIMMER

Abstract. *Saccharomyces cerevisiae*, or budding yeast, can undergo spontaneous, synchronized metabolic oscillation when grown in a continuous culture. Here we use time series data derived from sampling ambient mRNA in a *s. cerecisiae* culture to determine that the period of budding yeast's metabolic cycle is 5 hours, and we distinguish components of the cycle. We also investigate which transcripts drive each component and interpret our findings with the help of gene annotation data. Code available here: `https://github.com/jackdpond/hmm_gene_model`

1. Problem Statement and Motivation

Transcriptomic tools measure the transcriptome, the subset of the genome (the set of a cell's DNA) being transcribed into RNA at any given time [CGG+21]. While individual cells' genomes are generally static, the transcriptome is dynamic, illuminating internal mechanisms that drive the dance of life. Even in simple eukaryotes like yeast, transcriptomic data are high dimensional, stochastic, and variable between individual cells [NRIW+19]. Interpreting these data in yeast and other organisms is ambitious but important for understanding and building comprehensive models of their biology.

In this paper, we use a time series transcriptomic dataset [TKRM05a] to investigate a yeast population undergoing spontaneous, synchronized metabolic oscillations. Our two guiding questions are: 1. What is the transcriptomic structure of these oscillations, and 2. How should we interpret this structure biologically? We independently determine the period of the metabolic cycle and identify distinct transcriptional components involved in the cycle to address the first question. We interpret these components using text annotations on each distinct transcript measured to address the second question. Our investigation validates the conclusions of [TKRM05a] using distinct analysis methods. It also provides new biological insights, and highlights advantages and drawbacks of various data analysis techniques on this high-dimensional, periodic dataset.

2. Data

2.1. **Description.** Our dataset consists of a sequence of yeast microarray assays. Each microarray was designed to measure the abundance of the

Date: April 16, 2026.

same set of 6,400 yeast gene transcripts along with other RNA transcripts of interest, bringing the total to 9725 transcriptomic features per assay. These assays were performed separately on 36 samples taken at regular 25-minute intervals from a metabolically oscillating yeast population. Our dataset is publicly available at the Gene Expression Omnibus [TKRM05b] and comes from a peer-reviewed publication in *Science* [TKRM05a]. Included with this dataset are natural language annotations of various categories describing the biological context of each transcript.

2.2. **Preprocessing.** Individual microarrays are subject to variability due to variations in staining and other experimental parameters. We median-normalized intensity values for each microarray as indicated by [TKRM05a] to reduce noise-based differences between the microarrays' distributions of expression intensities. This preprocessing step was a standard practice at the time of this publication. However, it restricts our use of the normalized data to relative comparisons, not absolute comparisons, about intensities of the each gene across timesteps.

The means and variances of each of the transcripts occupy many orders of magnitude, with higher mean intensity corresponding to higher variance in expression. To attempt to 'stabilize' the variance of our time series across varying mean intensities, we took an *arcsinh* transformation of all intensity values. Using *arcsinh* functions is standard practice in transcriptomics, similar to log-transforming data but with better handling of zero, which is a common value in transcriptomic data. It should be noted that applying or even learning an affine transformation to the data first can enhance the variance stabilization effect of *arcsinh* [HvHS$^+$02]. In our case, visual inspection of the data confirmed that the time series' variances were more comparable than before and we decided to proceed without that additional step.

The *arcsinh*-normalized data was used for our analyses that require comparing relative expression intensities of different transcripts, but for our analyses that compare the temporal expression patterns, we decided to also standardize the *arcsinh*-transformed relative intensities of each gene via a Z-score across each individual time series. Normalizing by Z-score allows us to identify transcripts with similarly timed peaks and troughs in relative expression, even when those transcripts have different means and variances in *arcsinh* relative intensity.

2.3. **Train-test Split.** Our data involves a relatively small number of time points. For some analyses such as clustering, we used the entire dataset because evaluation was accomplished visual inspection, etc. In one case, we applied a train test split useing the first two periods (24 time steps) for training and the third period (the last 12 time steps) for the testing set. The forecasting we attempted would have also used this train test split.

3. Methods

Our anlysis seeks to answer two basic questions: (1) What is the transcriptomic structure of the observed metabolic oscillations (i.e., what is the period and what are the distinct phases)? and (2) What is the biological interpretation of this structure?

3.1. **Period Length Discovery.**

We expected that the metabolic cylce would have distinct *phases* that repeat in order, identifiable by the groups of co-expressed transcripts that characterize them. However, the large number of transcripts made individual time series hard to study. We fit Gaussian Mixture Models (GMMs) to the full set of *arcsinh*-transformed gene expressions to cluster the 36 timesteps by their expression values and thereby recover the phases and underlying period of the yeast metabolic cycle.

Approach 1: High-variance feature selection. A naive GMM fit on all 9,275 transcripts as features suffers from the curse of dimensionality ($n \ll d$), producing overconfident, nearly deterministic cluster assignments. To mitigate this, we discarded all but the n highest variance transcripts (rows) and fit a 5-component GMM on all 36 time points. The intuition is that high-variance time points are the most discriminative for separating phases of the metabolic cycle. We swept n from 200 to 1100 in steps of 100, observing that cluster assignments were stable regardless of n and reflected a neatly cyclic sequence consistent with a 300 s period, or three total periods in the experiment.

However, even this large reduction did not produce real soft-clustering, as the assignments were still essentially deterministic. We decided to reduce to even less than 200 dimensions.

Approach 2: PCA + GMM on time steps. As a more principled dimensionality reduction, we projected the 9,275-gene feature space down to 5 principal components and fit a GMM with tied covariance (also to prevent overfitting) on the reduced representation.

In order to ascertain the best number k of clusters, we split the data temporally — the first 24 time points for training and the remaining 12 for testing — to avoid leakage. We selected the number of components $k \in 2, \ldots, 7$ by evaluating held-out log-likelihood on the test set over 100 random restarts per k, finding $k = 5$ to be consistently well-supported.

Predictions over the full 36-time-point sequence with $k = 5$ show a pattern in cluster sequence, confirming a 300s (12 timestep) oscillation period (Figure 1A). Interestingly enough, even with an observation space of just 5 dimensions and a hidden state space of just 5, the assignments are still very high probability, indicating high separability.

Additionally, as a sanity check, we fit a GMM on the 200 highest variance transcripts and then projected to 2 principal components, and graphed the 36 time steps colored by their GMM assignment. There are 5 distinct clusters in 2 dimensions which are correctly discovered by the GMM in 200

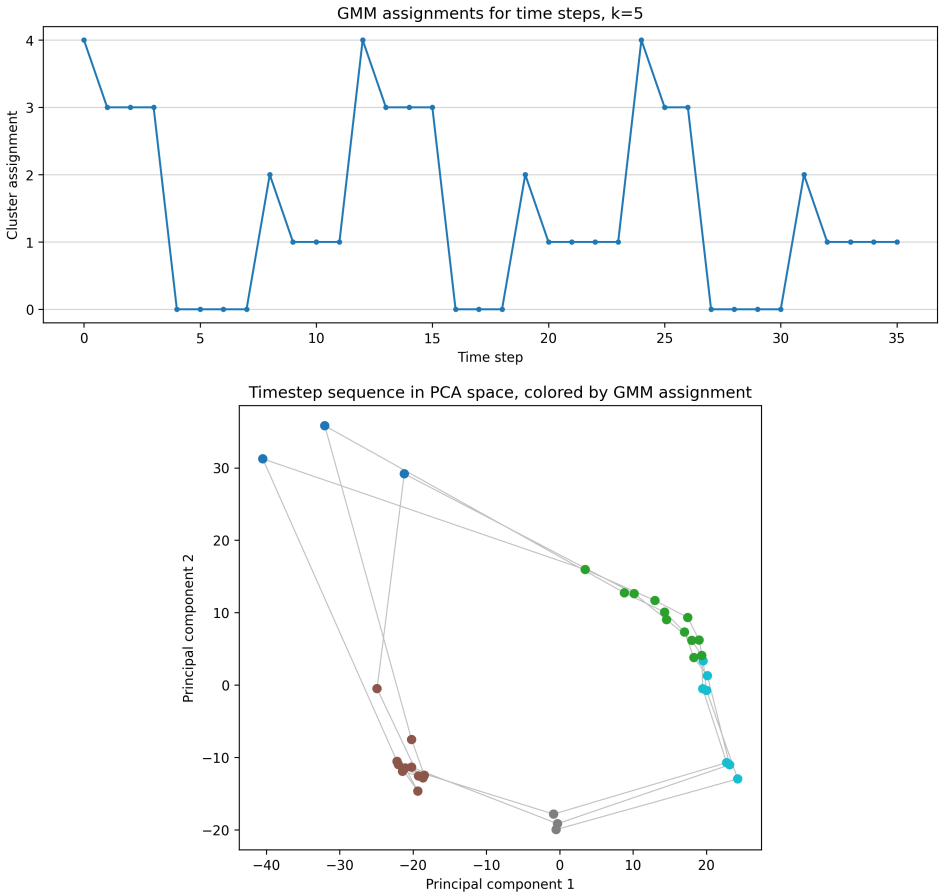

FIGURE 1. A (top): GMM cluster assignments over time–the 12-time-step period is clear. B (bottom): Projection to two principal components of 200-dimensional time-steps, colored by GMM assignment and connected in order. The clusters and cycle are clear.

dimensional space. Plotting a line connecting the time-points in order reveals the looping, periodic nature of the data (Figure 1B).

Approach 3. Continuous Density Hidden Markov Model. We also attempted to discover hidden states using a Continuous Density Hidden Markov Model (CDHMM) approach. We using PCA on our *arcsinh* normalized dataset to capture the directions of highest variance, and then used *GaussianHMM* method from the *hiddenhmm* package on the transformed data to compute the hidden states. A CDHMM approach using 5 states shows robust periodicity (Figure 2).

3.2. **Periodic Gene Identification.** Once an overall 12-timestep period length was identified, we used permutation testing to identify which of the

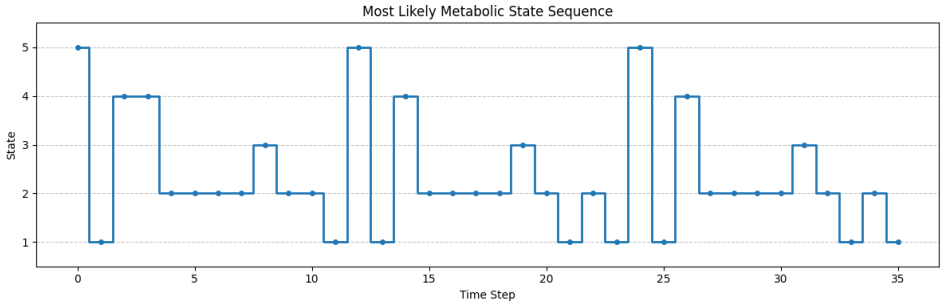

FIGURE 2. The hidden states predicted by a CDHMM approach over the course of the metabolic cycle also support an overall 12-timestep periodicity.

9725 transcripts were oscillatory on that period. We computed the autocorrelation of each transcript's time series with delay equal to the period length (12 timesteps). We called these values the *true series autocorrelation* for each gene. We then randomized the order of the intensities in each time series 100 times and computed the lag-12 autocorrelations of these shuffled time series. If the absolute value of the true series autocorrelation was at the 95th percentile or greater relative to the absolute values of the randomized series, we accepted the gene as likely periodic and labeled it accordingly for downstream analyses. The other transcripts were rejected as non-periodic on the 12-timestep period, since significant numbers of randomized permutations of their time series had similar lag-12 autocorrelations.

We used PCA to visualize the distributions of expression patterns between the periodic and non-periodic transcripts. Our visualizations on the *arcsinh*-transformed relative intensities (Figure 3A) appears consistent with the hypothesis that that both periodic and non-periodic transcripts can be found with wide ranges of average expression levels (appears consistent with PC1), while periodic transcripts otherwise vary more in the temporal modlulation of expression profiles than the non-periodic transcripts (appears consistent with PC2). To remove the variation in average expression between transcripts from the visualization, we also visualized the time-relativized (Z-scored) expressions with PCA (Figure 3B), finding that in two dimensions, the time-relativized periodic time series tend to a wider distribution than the non-periodic transcripts, surrounding them with a disc or annulus-like shape.

3.3. **Classical Time Series Decomposition and Analysis.** Once a suitable overall period and individual periodic transcripts were identified, we used classical time series analysis to investigate the periodic transcripts specifically in more depth. We obtained the classical time series decomposition the *arcsinh*-transformed data for each individual gene via centered

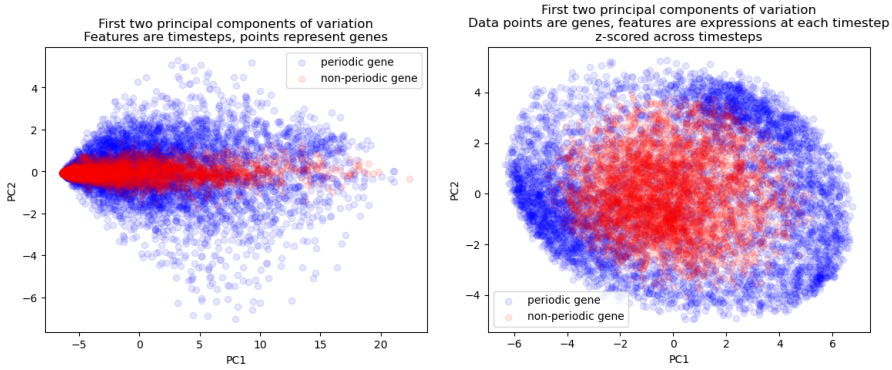

FIGURE 3. Distributional differences in the time series of periodic and non-periodic transcripts. A (left): A PCA plot of individual transcripts by their *arcsinh*-transformed relative intensities across timesteps. B (right): A similar plot, using the time-relativized (Z-scored) versions of those values.

moving averages over the 24 central time steps (as the moving average requires timesteps at the beginning and end of the time series). We then Z-scored the seasonal components for each transcript across time points in that transcript's seasonal component. We clustered these Z-scored seasonal components used GMMs of varying cluster counts, finding 8 clusters to minimize the BIC values. Note that here we cluster *transcripts*, whereas earlier we clustered *time steps*. The time-step clustering recovered period-length, and the transcript clustering recovered distinct temporal modules of co-expressed transcripts.

We then inspected the temporal patterns of the transcripts in each in more depth through visualization of the time series in each cluster. We removed one cluster which did not appear to have a consistent temporal pattern, but rather seemed like a catchall for transcripts that were not contained in any other cluster. The other clusters appeared to represent synchronized temporal expression modules, with different phases and shapes visible among the modules (Figure 4B). For the transcripts in these clusters, a UMAP plot of Z-scored expression patterns confirmed our suspicion that the topology of the transcription modules is annular. The phases of each modules' peak expression appears to be correlated with the angular position of the gene on the UMAP plot (Figure 4A). A video animation confirmed this observation.

3.4. **ARIMA.** We attempted to fit the data to an ARMA model. Largely, these efforts are unsuccessful. While we can fit individual transcripts as ARMA models, efforts to include interactions between transcripts failed. To try to combat the curse of dimensionality, we attempted to fit multivariate ARMA models to reduced data from PCA, or to the mean trends of clustered

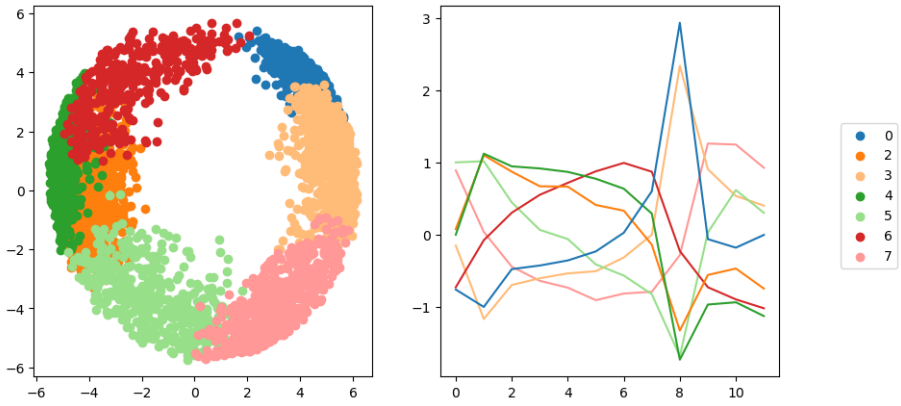

FIGURE 4. Temporal expression modules identified by GMM clustering (a catchall cluster with no apparent periodic synchronization was removed from the analysis). **A** (left): UMAP of transcripts according to the seasonal component of their standardized periodic expressions, colored by temporal module. **B** (right): Periodic expression trends of each synchronized module. Z-scored seasonal component values were averaged by module across one period, the other period of the seasonal component looks similar.

data. However, these attempts also failed, either due to lack of convergence or to nonsensical results.

3.5. **Semantic Cluster Interpretation.** In order to aid our biological interpretation, we performed Latent Dirichlet Allocation (LDA) using the natural language descriptions each of the transcripts' biological context. The natural language descriptions for each gene were constructed by concatenating several text fields in the metadata CSV: "Sequence Source", "Target Description", "Gene Title", "Gene Ontology Biological Process", "Gene Ontology Cellular Component", and "Gene Ontology Molecular Function". For transcript targets not associated with a gene, the gene-related fields were often blank.

We considered each gene as a document, and performed LDA on the documents to identify conserved 10 topics. We iteratively added to the list of stopwords until the words appearing in the topics were composed entirely of meaningful biological terms and identifiers. We then extracted the top 20 keywords used in each of the 10 topics, and gave these words to Claude Sonnet 4.6 to classify into themes.

We then analyzed the contributions of each topic to each gene's contextual label, identified the topic with the highest contribution for each gene, and investigated the concordnace of toipcs and transcriptional modules.

#	LABEL	TOP 10 KEYWORDS
T1	Cytoplasmic translation & ribosome structure	translation ribosomal subunit ribosome structural constituent large 0003735 complex factor
T2	Cell cycle & DNA damage checkpoint kinases	kinase binding regulation cellular DNA genetic bud phosphorylation repair Database
T3	Proteasome & ATP-dependent protein degradation	binding nucleus cytoplasm catabolic 0005737 ATP complex 0005634 hydrolase ATPase
T4	RNA Pol II transcription & chromatin regulation	transcription polymerase binding RNA DNA regulation II promoter factor DNAtemplated
T5	ER membrane & uncharacterized transmembrane proteins	membrane Database reticulum endoplasmic component function integral unknown 0016021 wall
T6	Mitotic/meiotic chromosome segregation & nuclear RNA processing	complex chromosome binding nuclear mRNA spindle nucleus RNA mitotic I
T7	Mitochondrial metabolism & oxidative phosphorylation	mitochondrial mitochondrion 0005739 membrane inner binding Database complex oxidoreductase dehydrogenase
T8	Vesicular transport & endomembrane system	transport membrane transmembrane transporter component integral 0016021 Golgi vacuole 0006810
T9	Retrotransposon biology (Ty elements) & nucleic acid metabolism	binding biosynthetic TYB helicase DNA is acid Gag structural tRNA
T10	Ribosome biogenesis & rRNA processing (nucleolus)	rRNA processing nucleolus SSUrRNA 0005730 biogenesis 58S nucleus LSUrRNA binding

FIGURE 5. Some of the keywords used by Claude Sonnet 4.6 to summarize each toipc.

4. Results and Analysis

We first performed GMM clustering and fit a CDHMM across the time steps. The key takeaways were 1. the consistent recovery of the period of the yeast's metabolic cycle (300 minutes), which independently verifies the $\approx$ 300 minute period recovered by Tu et al. [TKRM05a] via distinct methods; and 2. the consistent identification of the optimal number (5) of distinct yeast transcription states/clusters, which was not identified in [TKRM05a].

Permutation testing identified 5243 of the transcripts–over half of all transcripts–as periodic with the 12 timestep period. This number roughly agrees with [TKRM05a], while our methods (use of permutation testing) are unique. Classical time series decomposition of the periodic transcripts enabled identification of transcriptional modules–clusters of transcripts that peaked and trough in sync. These clusters peaked with different phases (Figure 4), which were correlated with transcript biological function as identified via LDA (Figure 6 A).

The temporal ordering of biological functions we identified in the metabolic cycle (Figure 6 A) generally agreed with the ordering of oxidative, reductive building, and reductive charging phases (Figure 6 B) described in [TKRM05a]. Our biological module identification was more granular than

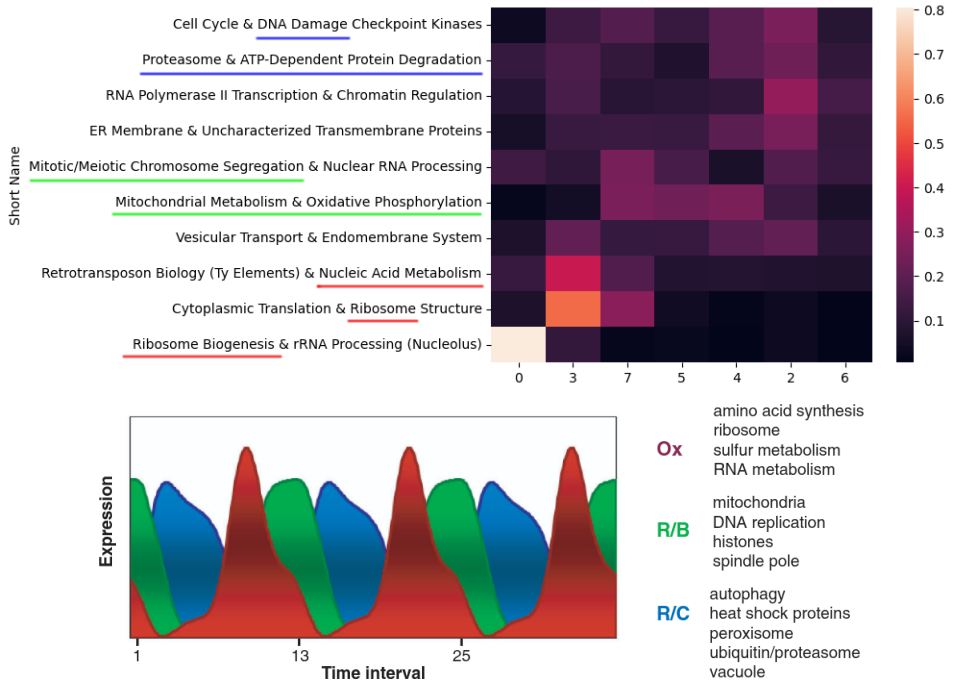

FIGURE 6. **A** (top) Row-normalized heatmap counting transcripts with each combination of module (column) and topic (row) membership. The columns are sorted temporally in terms of increasing phase as seen in Figure 4, and a correlation between module phase and the distribution of its transcripts' topics is observed. **B** (bottom) The three metabolic modules identified in [TKRM05a]. Note the correlation (red, green, blue) of module ordering with topics in **A**.

[TKRM05a], with more temporal modules identified (7 vs 3), and interpreting these extra modules would be best done with a domain expert, but the general agreement of the overall trends between our analyses is encouraging, especially given that we used unsupervised methods.

ARMA models were explored but failed to converge meaningfully (see Methods 3.4). We attribute this result to the inherent periodicity in our data. This violates ARMA's assumption of a covariance stationary time series (the expected value of the time series is different at different time steps). In our case, first and second differencing (ARIMA) did not eliminate the problems, which makes sense because the trend was not linear or quadratic but oscillatory. Some literature search showed that seasonal differencing could be an appropriate solution for future work [RG26].

5. Ethical Implications and Conclusions

Our data are derived exclusively from yeast cells, which avoids many of the ethical dilemmas associated with using a dataset derived from the human transcriptome. However, we recognize that important scientific and medical insights are obtained by thorough analysis of the human transcriptome and genome, and that investigations of those datasets require significant discression, prudence, and strict privacy regulations to ensure the safety of all participants.

In addition to considerations about extensions to human-derived data, we also consider the impact our work has on scientific replicability. The ability to produce consistent conclusions when reexamining and validating the analysis of a dataset is critical.When multiple teams find consistent results using different analysis methods, the scientific interpretations become more reliable and stronger [BNea20].The analysis of large genomic and transcriptomic data sets is involving more complicated analysis pipelines, so practicing replication is a positive ethical practice that we attempt to perform here. When several techniques converge to the same conclusion, results obtain an additional robustness.

Obviously, mathematical evaluation of mathematical data does *not* address the other side of the coin - experimental replication, where a experiment is repeated to collect additional data to validate previous results. However, robust strategies of reproducibility can lead to new questions about the same data, spurring more interesting future experiments that not only verify but build upon the findings of previous experiments. Thus, deployment of various strategies to analyze complicated, high-dimensional data is important both to validate previous conclusions about a dataset, and encourage novel research that helps qualify and extend those results honestly.

6. Conclusion

Analyzing the data provided an independent confirmation of some of the conclusions of Tu et al. [TKRM05a]. For example, the period they identified from the data (300 minutes) for the yeast metabolic cycle oscillations via a maximum likelihood analysis was rediscovered by our independent use of GMMs to cluster the timesteps, and we likewise identified slightly more than half of the dataset's transcripts as periodic with this period. This concordance reinforces the strength of those conclusions about their data. We also were able to provide novel analyses and interpretations of the data that Tu et al. did not reach. For example, our extraction of the seasonal components of the standardized time series and our use of GMMs to cluster them revealed 5-7 distinct periodic modules of activation, which improves upon the granularity of the original analysis [TKRM05a], which identified only three clusters using k-means. This validation combined with increased granularity represents an effort to replicate analyses, a validation of our computational toolset, and an opportunity to identify interesting new biological questions.

References

[BNea20] Rotem Botvinik-Nezer et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, June 2020.

[CGG+21] Claudia Caudai, Antonella Galizia, Filippo Geraci, Loredana Le Pera, Veronica Morea, Emanuele Salerno, Allegra Via, and Teresa Colombo. AI applications in functional genomics. *Computational and Structural Biotechnology Journal*, 19:5762–5790, January 2021.

[HvHS+02] Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl_1):S96–S104, 2002.

[NRIW+19] Mariona Nadal-Ribelles, Saiful Islam, Wu Wei, Pablo Latorre, Michelle Nguyen, Eulàlia de Nadal, Francesc Posas, and Lars M. Steinmetz. Sensitive high-throughput single-cell RNA-Seq reveals within-clonal transcript-correlations in yeast populations. *Nature microbiology*, 4(4):683–692, April 2019.

[RG26] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice (3rd ed)*. April 2026.

[TKRM05a] Benjamin P. Tu, Andrzej Kudlicki, Maga Rowicka, and Steven L. McKnight. Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, 310(5751):1152–1158, 2005.

[TKRM05b] Benjamin P. Tu, Andrzej Kudlicki, Maga Rowicka, and Steven L. McKnight. Yeast metabolic cycle microarray data. NCBI Gene Expression Omnibus, accession GSE3431, 2005.