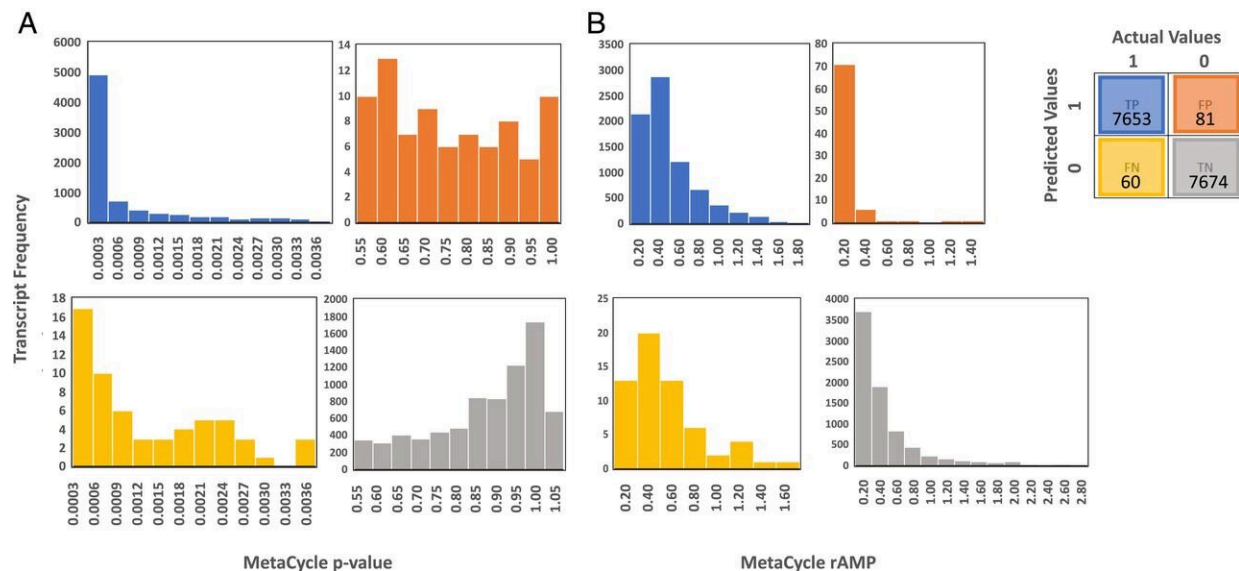# New machine learning technology explores circadian rhythms

August 10 2021

Arabidopsis circadian/noncircadian comparative ML binary classification with 12 transcriptomic timepoints. Class 0 = Noncircadian and Class 1 = Circadian. Histograms in A and B relate to the best model from SI Appendix, Fig. S1A generated using LightGBM; the histograms are color coded as per the confusion matrix shown to the right (i.e., showing where our model assigned TP labels, FP labels, FN labels, and TN labels). The histograms show bins representing the frequency of transcripts that had various P values (A) or rAMPs (B) assigned to them by Metacycle. Credit: DOI: 10.1073/pnas.2103070118

We all have an internal clock but what makes us tick? Scientists at the Earlham Institute and IBM Research have developed new artificial

intelligence (AI) and machine learning (ML) technology to understand how gene expression regulates an organism's circadian clock.

The inner 24-hour cycles—or circadian rhythms—are key to maintaining human, plant and animal health, which could provide valuable insight into how broken clocks impact health.

Circadian rhythms, such as the sleep-wake cycle, are innate to most living organisms and critical to life on Earth. The word circadian originates from the Latin phrase "circa diem" which means "around a day."

Biologically, the circadian clock temporally orchestrates physiology, biochemistry, and metabolism across the 24-hour day-night cycle. This is why being out of kilter can affect our fitness levels, our health, or our ability to survive. For example, experiencing jet lag is a chronobiological problem—our body clocks are out of sync because the normal external cues such as light or temperature have changed.

The circadian clock isn't unique to humans. In plants, an accurate clock helps to regulate flowering and is crucial to synchronizing metabolism and physiology with the rising and setting sun. Understanding circadian rhythms can help to improve plant growth and yields, not to mention revealing new avenues for tackling human diseases.

Beyond plants

For this latest research, the team applied ML to predict complex temporal circadian gene expression patterns in model plant Arabidopsis thaliana. Taking newly generated datasets, published temporal datasets, and Arabidopsis genomes, the team of scientists trained ML models to make predictions about circadian gene regulation and expression patterns.

Featured in the journal *PNAS*, the work demonstrates the power of AI and ML-based approaches to enable more cost-effective analysis and deeper insight into the function of the circadian clock and its regulation. These approaches are redefining how scientists use public data and generate testable hypotheses to understand gene expression control in plants and humans.

Lead author Dr. Laura-Jayne Gardiner from IBM Research Europe (formerly at the Earlham Institute where the research was carried out), said: "Essentially, our inner rhythm is driven by a circadian clock, which is a biochemical oscillator synchronized with solar time or the position of the sun in the sky. In most living things, including animals, plants, fungi and even cyanobacteria, internally synchronized circadian clocks make it possible for an organism to anticipate daily environmental changes corresponding with the day-night cycle and adjust its biology and behavior accordingly."

Detecting circadian rhythms

Prof Anthony Hall, Group Leader at the Earlham Institute, said: "Genes involved in the circadian clock typically show an oscillation between off-on state rhythmic patterns throughout a 24-hour period. This pattern is called circadian rhythmicity.

"Detecting circadian rhythmicity with existing methods is challenging as it requires using sequencing technologies to generate long, high-resolution, time-series transcriptome datasets to measure gene expression throughout the day. Not only is this expensive, it is also time-consuming for laboratory scientists. Consequently, our knowledge to date of how genes are controlled and regulated in a circadian clock is limited."

The development of AI and ML based technology was initially applied to the model plant Arabidopsis, progressing to testing other complex or

temporal gene expression patterns as well as other species across Arabidopsis ecotypes. Furthermore, the team have adapted the ML approach for wheat to show that the methods used allow accurate analysis of key food crops.

Arabidopsis thaliana is a popular scientific model organism used by plant biology and genetics. The first plant to have its genome sequenced, it has been used to understand the molecular biology and genetics of many plant traits, including circadian regulation.

"Our ML models classify circadian expression patterns using iteratively lower numbers of transcriptomic timepoints, which is an improvement in accuracy compared to the existing state-of-the-art models," explained Prof Hall.

"We developed a ML model which generates a proxy gene set to predict the circadian time (phase) from a single transcriptomic sampling time point in the day. There are thousands of public transcriptomic datasets and by comparing this predicted time with the experimental time, we can identify specific genes or conditions that alter the clock function. Therefore increasing our understanding of the mechanism and function of the clock."

"We re-defined the field by developing ML models to distinguish circadian transcripts that don't use transcriptomic timepoint information, but instead DNA sequence features generated from public genomic resources. Therefore, allowing us to predict the circadian regulation of genes simply by analyzing the genome DNA sequence."

The researchers based their study on the theory that a major mechanism of gene expression control, be it circadian or other mechanisms, is through transcription factors (and other factors) that bind to a regulatory DNA sequence.

Transcription factors are vital molecules that can control gene expression—directing when, where and to what degree genes are expressed. They bind to specific sequences of DNA and control the transcription of DNA into mRNA.

Explainable AI

Dr. Gardiner adds: "Our ML models and their application in crops, where circadian rhythms are critical to maintaining healthy growth and development, could lead to increased yields as agricultural scientists and farmers begin to use the model to understand the inner rhythms of the plants they grow and harvest.

"However, the technology we developed goes beyond the scope of plants. We are now looking at different species to investigate the circadian clock and its link to disease in humans, for example, where the dysregulation of the circadian clock has been associated with a range of diseases from depression to cancer."

Dr. Gardiner is clear about the value of ML and AI in gaining deeper insights into circadian regulation: "What makes our models more informative is our usage of explainable AI algorithms," she explains. "We wanted to use the interpretation of our ML models to illuminate what's inside the 'black box," so that we can better understand the predictions they make.

"We used local model explanations that are transcript specific to rank DNA sequence features, which provide a detailed profile of the potential circadian regulatory mechanisms for each transcript. Using the local explanation derived from ranked DNA sequence features allows us to distinguish the temporal phase of transcript expression and, in doing so, reveal hidden sub-classes within the circadian class. E.g., whether a transcript is likely to show its peak expression in the morning, afternoon,

evening or night."

"Interpreting machine learning models to investigate circadian regulation and facilitate exploration of clock function" is published in *Proceedings of the National Academy of Sciences* (*PNAS*).