

December 04, 2019 | By Liezel Labios

# Machine Learning Provides New Paradigm in Understanding Microbial Gene Regulation

*E. coli* are hardy bacteria, able to live in diverse conditions from the surface of a lettuce leaf to an acidic stomach. To survive and thrive in so many environments, the bacteria must use a network of transcriptional regulators to change their gene expression levels in response to their surroundings. Even in *E. coli*, one of the best characterized bacteria, it is still a significant challenge for scientists to understand how they coordinate the expression of their thousands of genes.

In a paper published in the Dec. 4 issue of *Nature Communications*, bioengineers at the University of California San Diego report a new method to interpret gene expression datasets. By applying a machine learning algorithm designed to separate mixed signals into their original sources, researchers were able to split a large high-quality collection of gene expression data into around 100 signals that represent the targeted effects of transcriptional regulators.

The work was led by Bernhard Palsson, Galletti Professor of Bioengineering at UC San Diego, and Anand V. Sastry, a bioengineering Ph.D. student in the Palsson lab.

When analyzing gene expression datasets, scientists traditionally have had to sift through hundreds of differentially expressed genes, trying to find a cohesive pattern or story that connected them together. However, a problem is that many of these genes may be responding to the same underlying signal, making it difficult to discern the root cause of the organism's response, Sastry explained.

The UC San Diego team pioneered a new framework that automatically extracts the signals for specific transcriptional regulators that cause the measured changes in gene expression. The method also does not require prior knowledge of the transcriptional regulatory network. This makes it easier to apply to less-understood organisms, Sastry said.

The team's analysis was able to characterize two previously-unknown transcription factors, and refined the known targets for many other transcriptional regulators. Additional studies are in progress to validate multiple predictions posed by the study. The team's analysis also identified direct links between mutations in *E. coli* strains and their gene expression states, introducing a new strategy to compare strains across a species.

"Since the transcriptional regulatory network is how bacteria sense their environment, we now have a way to see what they 'see'. We can easily tell if the cell is starved for a nutrient, like iron, or is stressed in any way," Sastry said. "This could be invaluable when studying complex environments, like *in vivo* infections."

Paper title: "The *Escherichia coli* Transcriptome Mostly Consists of Independently Regulated Modules." Co-authors include Ye Gao, Richard Szubin, Ying Hefner, Sibe Xu, Donghyuk Kim, Kumari Sonal Choudhary, Laurence Yang and Zachary A. King.

---

#### MEDIA CONTACT

**Liesel Labios**, 858-246-1124, [llabios@ucsd.edu](mailto:llabios@ucsd.edu)

UC San Diego's [Studio Ten 300](#) offers radio and television connections for media interviews with our faculty, which can be coordinated via [studio@ucsd.edu](mailto:studio@ucsd.edu). To connect with a UC San Diego faculty expert on relevant issues and trending news stories, visit <https://ucsdnews.ucsd.edu/media-resources/faculty-experts>.