

October 26, 2016 | By Liezel Labios

## New Analysis of Big Data Sheds Light on Cell Functions



*By integrating multiple omic data types into genome-scale models, researchers discovered new biological patterns inside cells.*

Researchers have developed a new way of obtaining useful information from big data in biology to better understand—and predict—what goes on inside a cell. Using genome-scale models, researchers were able to integrate multiple different data sets and discovered new biological patterns among different cellular processes. The research, led by bioengineers at the University of California San Diego, was published online Oct. 26 in *Nature Communications*.

Scientists have been relying more on big data to make new quantitative discoveries in biology with respect to the genome, the microbiome, personalized medicine and disease modeling, for example. With today’s technology, scientists are able to generate data about a cell’s or organism’s complete set of genes, proteins, RNA profiles, metabolites and much more—known as omic data. Using omic data, scientists can model complex biological interactions and gain a more holistic view of different cellular processes. But a challenge is analyzing and making sense of these large data sets.

“When doing big data analysis, it is important to know how all these different data types are related. Now we have a way of connecting multiple different data types to generate fundamental answers to biological questions,” said Bernhard Palsson, Galetti Professor of Bioengineering at the Jacobs School of Engineering at UC San Diego and senior author of the study.

“While all these data types are derived from the same cell, they represent processes occurring at very different scales. Our work is about getting multiple different data types synchronized so that we can understand the coordination of these processes and derive meaning from them,” said Elizabeth Brunk, a postdoctoral researcher in Palsson’s lab and a co-first author of the study.

This study is part of a larger effort to address a grand challenge posed by the National Institutes of Health called “Big Data to Knowledge”—translating large, complex biological data sets into information that can be understood based on fundamentals.

In this study, researchers collected multiple omic data types (RNA sequences, ribosome profiles, protein data, metabolic data) from *E. coli* grown in different growth environments. The team then integrated these different data types into next-generation genome-scale models of metabolism, which were developed in Palsson’s lab.

They examined the relationships between omic data types and discovered new regularities, which are biological consistencies throughout a change in environment. Among the regularities they found were that during protein translation, ribosomes consistently pause at particular sites along a messenger RNA transcript, and that these pause sites dictate the protein’s three-dimensional structure.

Pause sites exist so that a protein has time to fold and form its overall shape, which is important for the protein to function correctly, Palsson explained. This knowledge is useful for studying cancer biology. If a tumor has a genetic mutation that eliminates a pause site, translation will yield a protein that’s not folded correctly and malfunctions.

“Now we have a fundamental explanation for these pause sites that we didn’t have before. It’s as if we’re witnessing an intricate dance with a certain rhythm to make sure that a protein is formed the right way,” Palsson said.

The team also developed what’s called a parameterized model that can be used to predict which genes are expressed when a cell experiences a change in environment.

“Thanks to the high-quality topological information provided in the genome-scale models developed by Dr. Palsson’s lab, we can obtain a better understanding of the connection between genes, proteins and metabolites and place multi-omic data into the context of these biochemical networks,” Brunk said.

Full paper: “Multi-omic data integration enables discovery of hidden biological regularities.”  
Authors of the study are: Ali Ebrahim,\* Elizabeth Brunk,\* Justin Tan,\* Edward J. O’Brien,  
Donghyuk Kim, Richard Szubin, Joshua A. Lerman, Anand Sastry, Aarash Bordbar, Adam M.  
Feist and Bernhard O. Palsson of UC San Diego; and Anna Lechner of UC Berkeley.

\*These authors contributed equally to this work.

This work was supported by the Novo Nordisk Foundation (grant NNF16CC0021858), the US  
Department of Energy (DE-FOA-000014) and the National Institutes of Health (NIH R01  
GM057089).

---

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>.