Powerful machine-learning technique uncovers unknown features of important bacterial pathogen

January 21 2016, by Karen Kreeger

A powerful new machine-learning technique can be applied to large datasets in the biological sciences to uncover previously unknown features of organisms and their genes, according to a team led by researchers from the Perelman School of Medicine at the University of Pennsylvania. For example, the technique learned the characteristic gene-expression patterns that appear when the bacterium is exposed to low-oxygen conditions and robustly identified changes that occur in response to antibiotics.

The technique employs a recently developed algorithm called a "denoising autoencoder," which learns to identify recurrent features or patterns in large datasets without being told what specific features to look for. In 2012, for instance, when Google-sponsored researchers applied a similar method to randomly selected YouTube images, their system successfully learned to recognize major recurrent features of those images—including cats.

In the new study, published in the online journal *mSystems* this week, Casey Greene, PhD, an assistant professor of Systems Pharmacology and Translational Therapeutics, in collaboration with Deborah Hogan, PhD at Dartmouth College, used a system of denoising autoencoders to analyze many large datasets that measure how genes in the bacteria are expressed in different conditions.

"The system learned fundamental principles of bacterial genomics just from these data," Greene said. "We expect that this approach will be particularly useful to microbiologists researching bacterial species that lack a decades-long history of study in the lab. Microbiologists can use these models to identify where the data agree with their own knowledge and where the data seem to be pointing in a different direction." Greene thinks that these are cases where the data may suggest new biological mechanisms.

Last year, Greene and his team published the first demonstration of the new method in a biological context: an analysis of two gene-expression datasets of breast cancers. The new study was considerably more ambitious—it covered all 950 gene-expression arrays publicly available at the time for the bacterium Pseudomonas aeruginosa, from 109 distinct datasets. This bacterium is a notorious pathogen in the hospital and in individuals with cystic fibrosis and other chronic lung conditions and is often difficult to treat due to its high resistance to standard antibiotic therapies.

First author Jie Tan, a graduate student at Dartmouth, where Greene, until recently, had his laboratory, developed ADAGE (Analysis using Denoising Autoencoders of Gene Expression) and applied it to the P. aeruginosa datasets. The data included only the identities of the roughly 5,000 P. aeruginosa genes, their measured expression levels in each published experiment. The goal was to show that this "unsupervised" learning system could uncover important patterns in P. aeruginosa gene expression and clarify how those patterns change when the bacterium's environment changes, for example when in the presence of an antibiotic.

Even though the model built with ADAGE was relatively simple—roughly equivalent to a brain with only a few dozen neurons—it had no trouble learning which sets of P. aeruginosa genes tend to work together or in opposition. To the researchers' surprise, the ADAGE system also detected differences between the main laboratory strain of P. aeruginosa and strains isolated from infected patients. "That turned out to be one of the strongest features of the data," Greene said.

"We were struck by the similarities between P. aeruginosa grown in association with cultured lung epithelial cells and these bacteriataken directly from the lungs of individuals with cystic fibrosis," said John H. Hammond, a graduate student in the Hogan Lab who collaborated on this project. "We are excited to continue to use ADAGE in combination with data from patient samples and experiments using laboratory models to discover better ways to find therapies to treat cystic fibrosis lung infections."

"We think that the proliferation of 'big data' provides an opportunity, through the use of unsupervised machine-learning, to find completely new things in biology that we didn't even know to look for," Greene said.

More information: Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders, www.ncbi.nlm.nih.gov/pubmed/25592575

Provided by University of Pennsylvania School of Medicine

Citation: Powerful machine-learning technique uncovers unknown features of important bacterial pathogen (2016, January 21) retrieved 24 April 2024 from https://phys.org/news/2016-01-powerful-machine-learning-technique-uncovers-unknown.html