Machine learning may boost protein production for better pharmaceuticals

December 22 2017

A machine learning program developed by an international team of researchers may help pharmaceutical companies produce higher quantities of cutting-edge drugs needed for medical treatments.

In a study, the team developed a computer algorithm using gene expression data of Chinese hamster ovary cells—a cell line often used by biopharmaceutical researchers for medical research—to optimize the production of proteins in those cells.

"The pharmaceutical industry typically relies on ovary cells of a Chinese hamster—CHO cells—for research to create effective drugs, but, because the cells do not produce much protein per cell, it requires large-scale production," said Claudio Angione, senior lecturer in computer science, Teesside University. "What we show is that, compared to other methods, combining this metabolic modelling with data-driven methods could be a vast improvement to the automation of cultures design, by accurately identifying optimal growth conditions for producing target therapeutic compounds."

The researchers, who reported their findings at the Second International Electronic Conference on Metabolomics, combined machine learning and a computational model that reconstructs the metabolism of the Chinese hamster ovary cells—CHO—to maximize the cell's efficiency.

"This is a novel step because, for the first time, we are combining two methodologies usually used individually in bioprocessing studies," said

Angione.

The researchers were able to predict the production of lactate—a toxic waste product—inside the cells, in terms of both their genetic and metabolic states.

"Production of lactate is generally undesired as it hinders cell growth and consequently limits the yield of desired products," said Macauley Coggins, research assistant, Teesside University. "By predicting the cellular conditions where lactate accumulation is minimized it is possible to reduce—or possibly avoid—long series of experimental trials."

Therapeutic proteins, like the ones produced in CHO cells, have a wide range of applications in medicine.

"Some of them are used in vaccines and protect against infectious agents such as viruses," added Guido Zampieri, a doctoral student in genomics and bioinformatics, CRIBI Biotechnology Center, University of Padua. "Other proteins with special targeting activity can be used to treat patients that lack those proteins due to genetic conditions. Anticancer drugs are another example."

Machine learning is a field that explores how computers can learn how to solve problems and undertake specific tasks without being programmed, according to Coggins. To do this, researchers usually develop an algorithm to train a computer to recognize patterns, a machine learning technique often referred to as supervised learning.

"It's a lot like how you teach a child to recognize different shapes by showing them what each shape is and what it looks like"

In the future, this method could be used to optimize other metabolites or proteins, the researchers suggest. Producing higher quantities of drugs

could also lead to less expensive treatments.

"We see several interesting research directions," said Angione. "Primarily, we aim at pushing forward the integration of different computational methodologies such as machine learning and biological modelling. This is important as they possess different strong points, which if combined could allow adopting more precise bioengineering interventions.

Particularly, machine learning can extract useful knowledge from experimental data, while metabolic modelling provides insights about local and global mechanisms in biochemical networks.

"We also want to explore other bioengineering steps that could benefit from this integrated optimization. The final goal is to obtain a set of computational tools that can guide industrial processes across multiple levels."

The researchers used data from a publicly available large-scale gene expression dataset from two different CHO cell lines with 295 microarray profiles with expression values for 3592 genes from 121 CHO cell cultures. For genome reconstruction, the researchers used a recently developed genome-scale metabolic model—GSMM—used to accurately predict growth phenotypes. The model is currently the largest reconstruction of CHO metabolism.

They then combined the model of CHO cell metabolism with the gene expression data to create condition and cell line-specific polyomics models.

Provided by Teesside University