Data mining algorithm explains complex temporal interactions among genes

June 22 2010

Researchers at Virginia Tech, New York University (NYU), and the University of Milan, Italy, have created a data mining algorithm they call GOALIE that can automatically reveal how biological processes are coordinated in time.

Biological processes such as cell division, metabolism, and development must be carefully synchronized for proper cell function. How such events are coordinated in time is a complex problem in the field of systems biology. While researchers can gather temporal data about the activity of thousands of genes simultaneously, interpreting these datasets in order to understand higher order phenomena such as cell division requires the development of new analysis tools. The mathematically rigorous data mining algorithm GOALIE (Gene Ontology based Algorithmic Logic and Invariant Extractor) reconstructs temporal models of cellular processes from gene expression data. The researchers describe this algorithm in the early on-line Proceedings of the National Academy of Sciences (*PNAS*) to be released the week of June 21, 2010.

The researchers developed and applied their algorithm to time-course gene expression datasets from the well-studied organism Saccharomyces cerevisiae, a budding yeast that is also used for raising bread dough and the manufacture of beer, wine, and distilled spirits. They applied their novel temporal logic-based algorithm to a range of yeast data sets involving cell division, metabolism, and various stresses. "A key goal of GOALIE is to be able to computationally integrate data from distinct stress experiments even when the experiments had been conducted

independently," said Naren Ramakrishnan, professor of computer science at Virginia Tech, and lead author.

"GOALIE is part of a broader effort to combine data mining with modeling tools", said Bud Mishra, professor of computer science and mathematics with the Courant Institute of Mathematical Sciences at NYU, and corresponding author. Mishra, also a professor of cell biology with the NYU School of Medicine, is investigator on a $10 million National Science Foundation (NSF) Expeditions grant to develop novel computational reasoning tools for complex systems, focusing on biological organs to complex diseases as well as engineered systems. "GOALIE can not just mine patterns but also extract entire formal models that can then be used for posing biological questions and reasoning about hypotheses," said Mishra.

A hypothesis in the yeast example is how genes organize into groups to perform a specific concerted behavior. "However, these gene groupings are not permanent, but shift as the cell begins orchestrating its next step. These transitions correspond to significant 'regrouping' of genes, which is indicative of a change in cellular state," said Richard Helm, associate professor of biochemistry at Virginia Tech, and co-author. Tracking down these transitions in time-based experiments is difficult, especially with thousands of genes changing in levels simultaneously. "When confronted with datasets this large we tend to focus on our 'favorite' genes or processes, leading potentially to a biased viewpoint," said Helm.

"GOALIE blends techniques from mathematical optimization, computer science data mining, and computational biology," said Layne Watson, professor of computer science and mathematics at Virginia Tech, and co-author. "It automatically mines the data in an unsupervised manner, identifying temporal relationships between groups of genes in order to gain a more unbiased and holistic understanding of time-based cellular behavior."

Specific strains of S. cerevisiae have been shown to have two robust biological cycles occurring simultaneously, namely the metabolic and cell division cycles. While the yeast cell division cycle has been well studied, its relationship to and coordination with metabolism are only now being worked out. GOALIE was able to recover the underlying temporal metabolic and cell cycle relationships in the datasets studied. "Through our temporal models, we have shown that S. cerevisiae reacts in a somewhat unified fashion, with cellular fate depending on core metabolism and cell division," the authors write in their paper.

"The metaphor that emerges from this analysis is that the metabolic state of the cell is essentially a fuel gauge, and there must be enough 'fuel in the tank' before permitting another key biological process, such as reproduction, to commence," said Helm. "The availability of energy controls whether a yeast cell divides or not."

"Our tools bring out the nature of temporal 'hardwiring' manifest in biological processes," said Ramakrishnan.

Helm adds: "In particular, they open up questions related to whether it would be possible to manipulate the system to adopt an aberrant cell state or make it proceed along a desired temporal order. The identification of well-defined states, such as found in hydrogen peroxide treatments, suggests that at this stage it may be possible to force the organism to adopt aberrant states."

For instance, the biotechnology industry currently employs microbes for a number of important commodity and specialty compounds, ranging from biofuels to pharmaceutical products. If cell division could be unlinked from metabolism, the microbial system would only need nutrients for maintaining metabolism, with fewer resources diverted to cell division. "This scenario would reduce overall bioproduction costs for the chemical of interest," said Helm.

Ramakrishnan, Mishra, and co-author Marco Antoniotti, associate professor of computer science with the University of Milan, are also inventors on a US patent application about GOALIE for which a notice of allowance has been issued. "We hope in the future our work can become key to understanding other important phenomena, like disease progression, aging, host-pathogen interactions, stress responses, and cell-to-cell communication," said Mishra.

Provided by Virginia Tech