

New Search Engine Can be Used for Creative Discovery



Virginia Tech's supercomputer, System X, is an 1100 Apple Xserve G5 cluster. It is part of the university's research computing resource, available for faculty research.

When you ask a supercomputer to tell a story, you might not expect a creative outcome – or any. But a group of Virginia Tech researchers are using System X, the university's supercomputer, to test a new search program that can tell the stories of life – the connections between gene sets, for instance, or the connections between discoveries reported in biomedical articles on the U.S. National Library of Medicine PubMed database.

We are all familiar with search engines that provide a list of hits on the terms we enter. Researchers in computer science and biochemistry at Virginia Tech have created a search capability that they call Storytelling that will discover connections between information that appears dissimilar. It discovers a sequence of events or relationships to create a chain of concepts between specified start and end points. Imagine, for instance, asking for a connection from the concept “traveling in London,” to the concept “places popes are interred.” The Storyteller might postulate “the history of codes” as an intermediary and find the Da Vinci Code – if it existed.

“The stories are pieced together by analyzing large volumes of text or other data” said Naren Ramakrishnan, associate professor of computer science at Virginia Tech who works with life scientists to create software for data mining and information analysis tasks arising in biology. The aim is to help scientists make connections in the complex, burgeoning world of scientific discovery. “Everyday, there are new research results reported in the literature and there are discoveries waiting to be made by exploring connections,” said Ramakrishnan.

“Our minds cannot correlate all available datasets efficiently and with any high degree of confidence without the aid of computational biology,” said Richard Helm, associate professor of biochemistry. “Attempting to find significant correlations within the ocean of online datasets is daunting. However, there may be experiments that have been published in the literature that look at particular subsets of a biological process. The storytelling algorithm links ‘distant’ objects by finding these closer connections and drawing them together in a storyline. Evaluation of these stories can provide hypotheses that can be tested at the bench, potentially resulting in new insights into the role of a particular molecular event in the process you are interested in.”

The design of the storytelling algorithm is modeled after large scale search engines such as Google. Each “node” in System X, an 1100 Apple Xserve G5 cluster supercomputer, is responsible for indexing a portion of the biological literature and the nodes exchange information among each other to help define links and make connections. “Some of our larger storytelling runs process hundreds of thousands of papers and work with up to 200 nodes simultaneously,” said Ramakrishnan.

Helm and his colleague Malcolm Potts, professor of biochemistry, are studying the processes and strategies used by organisms to enter into and exit from a state of reduced metabolic activity, such as dormancy or suspended animation. Application of such processes to mammalian cells could lead to the development of robust cell-based biosensors, long-term storage of cell components, and vaccines that do not require refrigeration.

So they decided to use Storytelling to study such processes on the budding yeast (*Saccharomyces cerevisiae*) by exploring connections between yeast papers. They used Storytelling to explore article abstracts -- not agreed upon code or nomenclature, but sentences and paragraphs that present thoughts from different people using different phrases and jargon and not thinking about the same problems.

The researchers used Storytelling to discover the relationship between two PubMed (PM) articles – “Early expression of yeast genes affected by chemical stress” (PMID: 15713640) and “Heat stress transcription factors from tomato can functionally replace HSF1 in the yeast *Saccharomyces cerevisiae*” (PMID: 9268023). They asked for connections using abstracts of 140,000 publications about yeast. Keywords were developed from 3,756 abstracts containing the keywords “yeast” and “stress.”

“The [discovered] story begins with a high throughput experiment that links chemical stress to gene expression in *Saccharomyces cerevisiae*, and ends with heat stress transcription factors in tomato,” they reported at an international meeting. “The ‘story line’ was initiated through comparisons between oxidative and heavy metal stresses. This led to a paper identifying a gene from *Candida* sp. that was expressed when the cells are exposed to cadmium but not copper, mercury, lead, or manganese.”

It turned out to be a novel protein. “The link between tomato heat stress transcription factors and a cadmium-specific gene with no known match in the current databases was through work with a different species of yeast (the fission yeast *Schizosaccharomyces pombe*) where a study looked specifically at heat and cadmium stress responses. This story identifies key players in the systems biology of related chemical stresses,” Helm said.

“The holy grail of applying computing to biology is to understand a particular organism or process at a higher level than we are used to considering,” said Helm.

The study aging in humans is an example of such “systems biology” – the study of integrated systems. And it is an example of an area where the Storytelling approach can make a difference. “There are a lot of molecular components to the aging process, yielding physical and emotional “phenotypes.” All these interactions generate an aged human, the results of which are slightly different for everybody,” said Helm. “Can we make connections between seemingly dissimilar molecular events to piece together the complex aging process? That is a tall question, but there may be experiments performed and published in the literature that look at smaller subsets of the problem.”

Source: Virginia Tech

This document is subject to copyright. Apart from any fair dealing for the purpose of private study, research, no part may be reproduced without the written permission. The content is provided for information purposes only.