MedicalXpress

Study shows promise in automated reasoning, hypothesis generation over complete medical literature

August 25 2014

With approximately 50 million scientific papers available in public databases– and a new one publishing nearly every 30 seconds – scientists cannot know about every relevant study when they are deciding where to take their research next.

A new tool in development by computational biologists at Baylor College of Medicine and analytics experts at IBM research and tested as a "proof-of-principle" may one day help researchers mine all public medical literature and formulate hypotheses that promise the greatest reward when pursuing new scientific studies.

Knowledge Integration Toolkit or KnIT

In a retrospective case study involving published data on p53, an important tumor suppressor protein, the team showed that this new resource called the Knowledge Integration Toolkit (KnIT) is an important first step in that direction, accurately predicting the existence of proteins that modify p53 – proteins that were subsequently found to do just that.

Details from the study published online today in the Association for Computing Machinery's digital library. Dr. Olivier Lichtarge, director of the Center of Computational and Integrative Biomedical Research at Baylor and the principle investigator on the study, will discuss details of

the study in a presentation Aug. 27 at the 20th annual Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining conference in New York City, the premier data mining conference.

"On average, a scientist might read between one and five research papers on a good day," said Lichtarge, also a professor of molecular and human genetics, biochemistry and molecular biology at Baylor. "But, to put this in perspective with p53, there are over 70,000 papers published on this protein.

Even if a scientist reads five papers a day, it could take nearly 38 years to completely understand all of the research already available today on this protein."

Scientists formulate hypotheses based on what they read and know, but because there is so little that they can actually read, hypotheses can be biased, Lichtarge said. "A computer certainly may not reason as well as a scientist but the little it can, logically and objectively, may contribute greatly when applied to our entire body of knowledge."

Collaboration with IBM

Together with colleagues at IBM led by Scott Spangler, principal data scientist at IBM, the team initiated a research project to develop a knowledge integration tool that took advantage of existing text mining capabilities, such as those used by IBM's Watson technology (cognitive technology that processes information more like a human than a computer.)

"Our hope is that scientists and researchers will be able to use Watson's cognitive capabilities to accelerate the understanding of biology underlying diseases," said Spangler. "Better understanding the biology of

diseases can eventually lead to better treatments for some of the most complex and challenging diseases, like cancer."

They came up with KnIT, a system that aims to mine the information contained in the scientific literature, represents it explicitly in a network that can be queried, and then further attempts to use these data to generate new reasonable and testable hypotheses that can be used to help direct laboratory studies.

P53 kinases

In the first test using KnIT, the team sought to identify new protein kinases that phosphorylate (or turn on) the protein tumor suppressor p53. There are over 500 known human kinases and 10s of thousands of possible proteins they can target. Thirty-three are currently known to modify p53.

In the study, the team used KnIT to mine the medical literature up to 2003 when only half of the 33 phosphorylating protein kinases had been discovered.

Using KnIT, 74 kinases were extracted as potential modifiers. Of these, prior to 2003, 10 were known to phosphorylate p53, nine were discovered at a later date. Of the 10 already known, KnIT accounted for them in reasoning as well as ranking the likelihood that the other 64 kinases targeted p53. Of the nine found nearly a decade later, KnIT accurately predicted seven.

"This study showed that in a very narrow field of study regarding p53, we can, in fact, suggest new relationships and new functions associated with p53, which can later be directly validated in the laboratory," said Lichtarge, who holds The Cullen Foundation Endowed Chair at Baylor.

The remaining kinases identified in the case study, but not previously identified in real time, may be further studied in the laboratory, he said.

Long-term goals

"Our long-term hope is to systematically extract knowledge directly from the totality of the public medical literature. For this we need technological advances to read text, extract facts from every sentence and to integrate this information into a network that describes the relationship between all of the objects and entities discussed in the literature," said Lichtarge. "This first study is promising, because it suggests a proof of principle for a small step towards this type of knowledge discovery. With more research, we hope to get closer to clinical and therapeutic applications."

Provided by Baylor College of Medicine