Accelerating Disease Research

on NIH's MEDLINE using a "Big Data" Approach

The Challenge

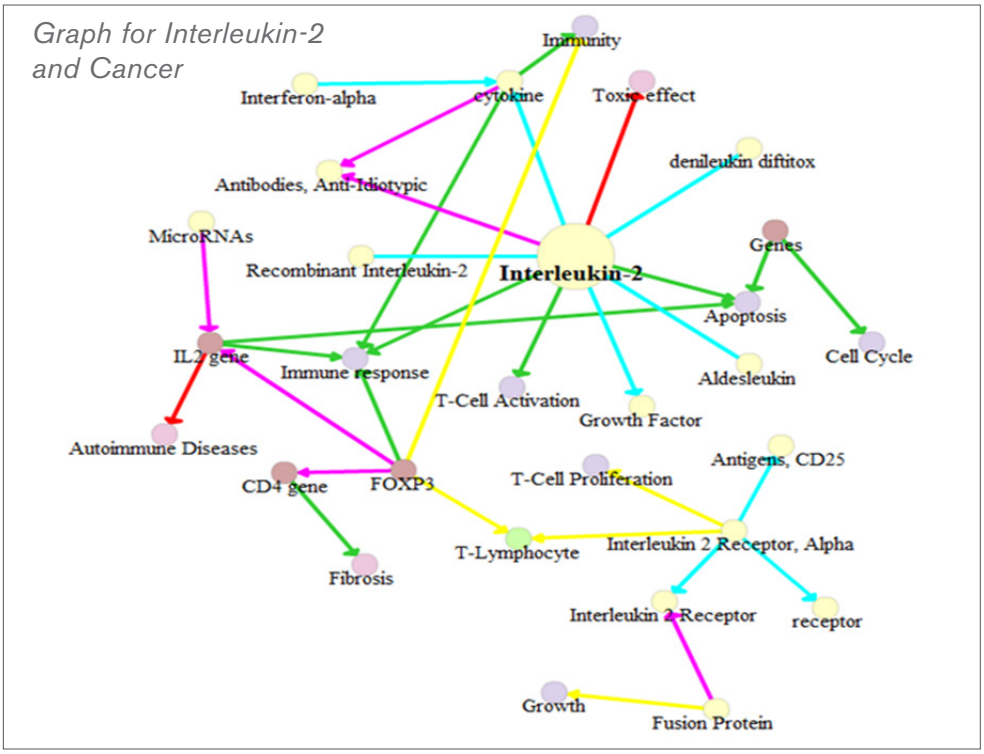
Researchers in the medical field have a wealth of data available to them regarding the characteristics of various types of diseases and the uses of therapeutic drugs and treatments to fight disease, including research from MEDLINE publications as well as publicly available data from other sources of publications, medical research and Web-based content.

This Big Data could be used to elucidate the molecular pathophysiology of diseases for which effective therapies are currently not available. Such enhanced understanding can underpin the development of new drugs, providing insight and guidance to focused lab work. The current MEDLINE database includes 22 million citations. These have been enhanced with 65 million semantic predications, resulting in a total of 2.2 billion RDF triples in the graph database available for this project. With commodity hardware and response times, the NIH team was required to limit their searching to a subset of the available citations and there was no room for expansion.

The Solution

The NIH team worked with YarcData to develop an innovative, real-time approach to disease research discovery using the Urika graph analytics appliance. The team was able to store all of their data resources in the appliance's shared memory (up to 512 terabytes)—eliminating the need to partition data or create time-consuming and complex data models prior to posing a hypothesis. Multiple graph analytics appliances could be linked together, creating a limitless environment for parsing and analyzing vast quantities of scientific data.

The NIH team wanted to identify new disease therapies based on the knowledge represented by roughly 10 million semantic predications extracted from nearly 3 million MEDLINE citations on cancer, far too much information to exploit effectively without sophisticated computational assistance. Working in the context of the literature-based discovery paradigm, which identifies hitherto unnoticed relationships in published research, the team used the system as a tool to augment their thinking process. They focused on the intricate networks of substance interactions that enable cancer hallmarks, such as resisting cell death, inducing angiogenesis, invasion and metathesis. The graph-based processing allowed the team to first focus on the molecular pathophysiology of a particular hallmark and then look at poorly understood aspects of substance interaction phenomena involved. In order to deliver results quickly, the researchers did not want to waste time inspecting large numbers of predications that did not advance their knowledge of the processes under inspection. Instead, the Urika-based application enabled them to develop rich graphs from the data found in the MEDLINE database and the system helped the team clearly visualize connections and associations which could potentially underpin new therapies and instantly rule out those that did not support the current goal.



The impact of using a more powerful analytics solution was immediate—and dramatic: It allowed the team to search the entire collection of citations and quickly gain insight, for example, into the immune system and cancer as the context for promising new cancer therapies. Recent research manipulating the cytokine interleukin-2, in particular, claim, for example, "profound antitumor effects" (PMID: 22660171) and "long-term tumour dormancy" (PMID: 23198850) —validating the effectiveness of the "big data" approach for identifying drug treatment solutions.

In order to rapidly validate scientific hypotheses in real-time and discover new connections between their existing data reserves and add new and unidentified data sources the NIH team needed a powerful appliance-based solution with the capacity to process large amounts of stored data simultaneously.

The Urika Advantage

Urika, with its global shared memory and proprietary Threadstorm massively multithreaded graph processor, allowed the team to immediately access NIH's entire Pub Med data simultaneously, enabling the researchers to formulate connections between disparate elements that could be easily visualized in graph form. Urika can consolidate every data resource available to the NIH team and load it into memory without upfront modeling. No advance knowledge of the relationships is required to identify non-obvious patterns, facilitating true data discovery.

NIH researchers could then investigate multiple graphs to inspect new insights into the molecular pathophysiology of cancer underpinning promising new therapies. Conventional analytics or lab-based approaches would either never discover such relationships or would take months to reveal them. Urika delivers these results quickly—enabling researchers to derive invaluable new insights from their existing data that can lead to new disease treatments and innovations throughout the life sciences industry.

YarcData
*Getting to **Eureka!** faster™*

About the Urika™ YarcData Urika big data appliance for graph analytics helps enterprises gain business insight by discovering relationships in big data. Urika complements an existing data warehouse or Hadoop cluster by offloading graph workloads and interoperating within the existing analytics workflow. Subscription pricing or on-premise deployment of the appliance eases Urika adoption into existing IT environments.

About YarcData YarcData, a Cray company, delivers business-focused real-time graph analytics. Adopters include the Institute of Systems Biology, the Mayo Clinic, Noblis, Sandia National Labs, as well as multiple deployments in the US government. YarcData is based in the San Francisco bay area and more information is available at www.yarcdata.com.