University at Buffalo, State University of New York

IBM Netezza powers the study of exponentially large gene interaction associated with multiple sclerosis

The State University of New York (SUNY) at Buffalo is home to one of the leading multiple sclerosis (MS) research centers in the world. MS is a devastating, chronic neurological disease that affects nearly one million people worldwide. The disease causes physical and cognitive disabilities in individuals and is characterized by inflammation and neurodegeneration of the brain and spinal cord. From the beginning, the genetics of MS were known to be complex and it was apparent that no single gene was the likely cause for the disease.

Since 2007, the SUNY team has been looking at data obtained from scanned genomes of MS patients to identify genes whose variations could contribute to the risk of developing MS. New technologies now enable hundreds of thousands of genetic variations, called single nucleotide polymorphisms (SNPs), to be obtained from single samples. According to research lead, Dr. Murali Ramanathan, a critical fact in the study of MS is that “gene products work by interacting with both other gene products and environmental factors.”

Because of this, researchers have postulated that multiple SNPs — combined with environmental variables — would better explain the risk of developing MS. Finding these combinations is analogous to the proverbial search for a needle in a haystack says Dr. Ramanathan. Identifying a candidate's environmental factors that could be used to prevent the disease from progressing in patients was of great importance. Examples of this include sun exposure and vitamin D levels, Epstein-Barr virus infections and smoking.

The researchers needed to create algorithms that would efficiently identify interactions and attempt ‘parsimony’ — an extreme efficiency. The team was looking for the fewest number of SNPs, environmental and phenotypic variable combinations that would help explain the presence of MS.
“Developing knowledge of pertinent environmental factors opens the way for new therapeutic and prevention strategies and can enable individualized treatment and management of MS patients.”

—Dr. Murali Ramanathan, Research Lead, The State University of New York (SUNY) at Buffalo

The researchers at the State University of New York developed an approach they called AMBIENCE that is distinctive in its use of new information theoretic methods along with its versatility and scalability. The theoretic underpinnings of AMBIENCE enable the detection of both linear and non-linear dependencies in the data. The AMBIENCE algorithm is capable of conducting an efficient search of the large combinatorial space because of the unique nature of the information theoretic metrics, which allows for greedy search identification of the most promising combinations.

The need
The computational challenge in gene-environmental interaction analyses is due to a phenomenon called ‘combinatorial explosion.’ Considering that there are thousands of SNPs, the number of combinations of SNPs that have to be assessed for uncovering potential interaction becomes astronomical.

Figure 1: A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide—A, C, G or T—in the shared genome differs in an individual. If many individuals with a disease (or ‘phenotype’) have the same polymorphisms that healthy individuals do not, this may be a clue to finding out genetic clues to what causes a disease (Wikipedia).
The sheer number of SNPs combined with environmental variables and phenotype values mean that the amount of computations necessary for data mining could number in the quintillions (18 zeros).

To identify how many interactions could exist, based on the number of SNPs included, the SUNY researchers graphed a mathematical function. If they were to mine through one million SNPs to look for the number of combinations containing four SNPs, they'd be looking for a possible sextillion interactions, or 10 to the 21st power.

The researchers attempted to run the algorithms against commodity hardware. In this method they found that a simple run of the algorithm took almost a week. Extrapolating this performance, they realized it would take many weeks to run the algorithm against the volume of data they wanted. Meanwhile, they knew that the algorithm results would lead to additional questions, algorithm adjustments, data changes and more, which would be untenable. They needed processing power that would speed this up by 100 times in order to be able to make meaningful discoveries and publish them.

The solution

The researchers knew they needed the kind of processing only available in a high-performance computer (HPC), previously known as a supercomputer. In addition, the research team needed capabilities found in relational databases — not often included in HPC platforms. HPCs typically ran their processing only in parallel, breaking up calculations to run simultaneously across many processors. They also often used field-programmable gate array (FPGA) architectures for additional speed.
Unfortunately HPCs can cost tens of millions of dollars and do not include storage for large volumes of data. They are designed for data to be read ‘into memory’ making their use for the ‘big data’ computations needed by the SUNY team implausible.

That is when the researchers discovered IBM® Netezza®. Testing the IBM Netezza data warehouse appliance confirmed it delivered all the performance they wanted and much more. IBM Netezza’s architecture is designed to accommodate ‘big data’ storage and processing, combined with the ‘in memory’ performance of an HPC.

IBM Netezza, a pioneer of data-intensive HPC that are easy to administer at a fraction of the cost of today’s HPC platforms, easily emerged as the comprehensive solution the researchers at SUNY were seeking.

**The benefit**

Once deploying IBM Netezza data warehousing appliance as their research analytics infrastructure, the SUNY researchers were able to:

- Use the new algorithms and commence analyses that were previously nearly impossible to conduct.
- Reduce the time required to conduct analysis from 27.2 hours without Netezza to 11.7 minutes with it. “We went from days to minutes,” said Dr. Ramanathan.
- Carry out their research with little to no database administration. (Unlike other HPC platforms or databases available, Netezza was designed to require an absolute minimum of maintenance.)
- Publish multiple articles in scientific journals, with more in-process.
- Proceed with studies based on ‘vector phenotypes’ — a more complex variable that will further push the IBM Netezza platform.

---

"With the IBM Netezza research analytics infrastructure in place running an algorithm became a matter of minutes instead of days"

— Dr. Murali Ramanathan
A call to action
One of the most debilitating hurdles to productive scientific research is the latency between being ready to start — having a solid question, the right data, and a robust calculation — and getting an answer. With the aggressive price/performance ratio of IBM® Netezza®, researchers can take advantage of modern, secure compute and storage power. And one of the chief benefits is that the researchers can ask and answer many more questions – in the same time period – than ever before.

For more information
To learn more about IBM Netezza data warehouse appliances, please contact your IBM or Netezza sales representative or IBM Business Partner, or visit the following website: www.netezza.com.

For more information on SUNY Buffalo, visit: www.buffalo.edu/, or SUNY’s research, consult the following sample publications:

**AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes.**
Chanda P, Sucheston L, Zhang A, Brazeau D, Freudenheim JL, Ambrosone C, Ramanathan M.

**Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits.**
Chanda P, Sucheston L, Liu S, Zhang A, Ramanathan M.
PMID: 19889230.

**Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity.**
Sucheston L, Chanda P, Zhang A, Tritchler D, Ramanathan M.
BMC Genomics. 2010 Sep 3; 11:487. PMID: 20815886.

**The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors.**
Chanda P, Sucheston L, Zhang A, Ramanathan M.
Epub 2009 Mar 18. PMID: 19293841.