

THE COMPUTERWORLD HONORS PROGRAM

CASE STUDY

LOCATION:
*Cambridge, Massachusetts,
United States*

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Science

NOMINATING COMPANY:
EMC

ORGANIZATION:

Whitehead Institute for Biomedical Research

PROJECT NAME:

Genome-wide Location Analysis

Summary

Genome sequencing gives biologists a list of the genes that run the life of an organism. Intricate regulatory networks control the expression of all of these genes. One key component of these networks is transcription factors, which are proteins that act as control switches, turning the genes on and off. Transcription factors give the orders; genes follow them.

Until now, unraveling the complex relationships between genes and transcription factor proteins was a painstakingly long and tedious process, involving years of individual biochemical and molecular experiments. Advanced high-throughput biological methods and computing technologies developed by Richard Young at Whitehead Institute for Biomedical Research, David Gifford at Massachusetts Institute of Technology, and their colleagues provide a way to describe and understand these relationships in a global fashion and allow researchers to do in weeks what would have taken years to achieve in the past. This is a major step in unraveling the regulatory networks that contribute to normal human development, cellular differentiation and environmental response and opens new avenues for developing targeted pharmaceutical treatments for human diseases.

Introductory Overview

Whitehead Institute for Biomedical Research is a nonprofit, independent research and educational institution with pioneering programs in cancer research, developmental biology, genetics, infectious disease research, and transgenic science. As computational capabilities and resources continue to develop, the use of computer science and technology by the researchers at Whitehead Institute is increasing dramatically. The fusion of biomedical research, highly advanced information technologies (including flexible, highly available computing and storage capacity), and a group of people highly skilled in bioinformatics and information technologies, offers substantial benefits in support of the mission of the Institute. One example of the vital programs being supported by Information Technology (IT) and Bioinformatics and Research Computing (BaRC) is the breakthrough research being done by Richard Young at Whitehead, David Gifford at Massachusetts Institute of Technology, and their colleagues in genome-wide location analysis.



THE COMPUTERWORLD HONORS PROGRAM

CASE STUDY

ORGANIZATION:
*Whitehead Institute for
Biomedical Research*

PROJECT NAME:
Genome-wide Location Analysis

LOCATION:
*Cambridge, Massachusetts,
United States*

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Science

NOMINATING COMPANY:
EMC

One of the next frontiers of the post-genome sequencing era is understanding how the expression of genes is controlled. Complex processes, such as the growth of an organism from a single cell, are the result of many cells precisely executing many gene expression programs in an ordered fashion. Intricate regulatory networks are needed to control and coordinate these expression programs. One key component of those networks is transcription factors, proteins that act as control switches, flipping the genes on and off. Transcription factors give the orders; genes follow them.

Until now unraveling the complex relationships between genes and transcription factors was a painstakingly long and tedious process, involving years of biochemical and molecular experiments. Recent advancements in high-throughput biological methods and computing technologies have provided a way to describe and understand these relationships in a global fashion and allow researchers to do in weeks what would have taken years to achieve in the past. This methodology, called location analysis, enabled rapid detection of DNA sequences bound by proteins, by using microarrays, a highly parallelized experimental platform where microscopic spots of DNA are printed on glass slides,

Recently, a key bottleneck in extending location analysis to the entire human genome on a highly reliable, easily obtainable platform was solved at the Whitehead. Due to the huge size of the genome, most previous efforts with location analysis could only focus on the regions of genes closest to the transcription start site. While the majority of known regulatory events seem to occur in this region, these regions only represent about one percent of the entire genome. In contrast, the new arrays used at Whitehead are designed to cover almost all of the human genome, greatly expanding the potential for understanding the full complexity of the regulatory network.

This involved first developing, optimizing and then running a software package which systematically combed the 3 billion basepairs of the human genome and evaluated over 1 billion candidate DNA fragments for inclusion on micorarrays. Additional software and algorithms were developed to select the subset of these DNA fragments that balanced the highest quality fragments with complete coverage of the genome. The sequences of the selected fragments were then sent to a commercial array manufacturer. The resulting set of slides, containing approximately 5 million spots of DNA, represents the whole human genome.

This achievement was made possible by an infrastructure of high throughput, high capacity and high availability data storage and computing resources. The supporting infrastructure also included people with the expertise necessary to develop, optimize and implement innovative software and implement and manage complex hardware resources. This infrastructure gave Whitehead researchers fast, on-demand access to the tools and information they needed.

This technique has already been used to create tools which can be used to identify the location of transcription regulators in human, yeast, mouse, zebrafish and fruit fly genomes. Combining this genome-wide location data with additional sources such as gene expression and sequence data enables researchers to construct detailed models of genome regulatory networks and create a far better picture of cellular function.

It is now possible to envision genome expression maps that explain how cell growth, cell cycle, nutrient utilization, stress responses and development are controlled and coordinated. These maps will be important for understanding human health and disease and for developing new, specifically targeted pharmaceutical approaches. They offer a new approach to researching hu-



THE COMPUTERWORLD HONORS PROGRAM

CASE STUDY

ORGANIZATION:
*Whitehead Institute for
Biomedical Research*

PROJECT NAME:
Genome-wide Location Analysis

LOCATION:
*Cambridge, Massachusetts,
United States*

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Science

NOMINATING COMPANY:
EMC

man diseases associated with mutations in transcription factors, including cancer, hypertension and immunological and neurological disorders.

Benefits

Research:

By using these advanced biological and computing technologies, Whitehead is in the forefront of determining how biological processes are regulated on a whole-genome scale. These techniques allow scientists to accomplish in a few weeks what would have taken years to achieve in the past. This is a major step in developing targeted pharmaceutical approaches to treating a wide range of diseases. For instance, identifying the control mechanism underlying cell division is key to finding out what goes wrong in cells causing cancer, where cells divide uncontrollably.

“In the whole history of biology to date, we’ve come to understand gene regulation for only a few dozen genes,” says Dr. Young. “In the next few years, this new technology will allow us to unravel gene regulation across the entire genome of any organism, and thus the script underlying fundamental life processes.”

Researchers at the Whitehead have used these biological and computational techniques and the related processing and data storage technologies to focus on a key transcriptional repressing complex in human embryonic stem cells. This complex, PRC2, was known to play an essential role in maintaining stem cell identity. If disabled, the stem cell immediately begins to differentiate and is thus no longer a stem cell. But it was not known how these proteins instruct stem cells to be pluripotent.

To answer this question, the entire genome of a human embryonic stem cell was analyzed and genes regulated by PRC2 were identified. The research team discovered that this complex binds and represses a special group of other transcriptional regulators that are required for development.

This new work provides the first genome-wide evidence that embryonic cells shut down gene expression programs that would lead to loss of stem cell identity by targeting transcriptional regulators.

Enabling New Methodology:

The investment in infrastructure and the experience gained from the project has reaped great benefits as there is now a pool of resources and expertise available on-demand for novel experimental techniques throughout the Institute. Examples include the development of microarrays for new organisms; the development of new arrays to examine the role of small, non-coding RNAs, protein-RNA interactions, and nuclear organization; and the ability to handle the assembly of new genomes for new experimental systems upon request

Online Dissemination of Genomic Breakthroughs:

Beneficiaries of Whitehead’s genomic research go far beyond the Institute itself. Thanks to the Institute’s data storage and archiving capabilities, all Whitehead research reports are made public by posting on the Whitehead Web site. As a result, biomedical researchers worldwide can access the latest, cutting-edge research. Whitehead’s goals aren’t just to achieve new, crucial breakthroughs but to disseminate them as widely as possible for the ultimate medical benefits they



THE COMPUTERWORLD HONORS PROGRAM

CASE STUDY

ORGANIZATION:
*Whitehead Institute for
Biomedical Research*

PROJECT NAME:
Genome-wide Location Analysis

LOCATION:
*Cambridge, Massachusetts,
United States*

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Science

NOMINATING COMPANY:
EMC

can bring to all people.

The Importance of Technology

Whitehead's Information Technology department provides a high performance computing, network and storage infrastructure in support of the scientific/bioinformatics requirements at the Institute. These scientific computing resources are made available to all researchers on an as needed basis and are considered an important part of the scientist's toolset.

The Institute's research cluster consists of 60 HP ProLiant DL360 servers running SUSE Linux, 10 of which were reserved for exclusive use by the Young lab in this project. Much of the research data is managed using Oracle databases on Sun Solaris servers. Providing data storage for these critical applications are two EMC CLARiiON CX600 systems with both ATA and Fibre Channel disk storage. Two EMC Celerra NS704G systems serve as the storage front end, speeding the deployment of storage resources in what would otherwise be a time-consuming process.

Because this is such data-intensive research, flexible data storage and on-demand access to information are critical. EMC storage technology enables Whitehead Institute's IT group to quickly reallocate storage capacity when researchers need it. Thirty-two terabytes of EMC storage, including EMC CLARiiON networked storage and EMC Celerra network-attached storage (NAS), as well as EMC data management software, support all high-end research efforts at the Institute. In this project processing a single species genome produced approximately a terabyte of data. Using both CLARiiON Fibre Channel and CLARiiON ATA disks, one terabyte of storage was always available to them; when that terabyte was used another was immediately allocated while the first was archived to tape.

In addition to these hardware resources an important part of the support of this and other research at the Institute is the expertise provided by BaRC. This group brings expertise in genomic and sequence database analysis, and the automation and optimization of computational analysis.

To accomplish this project, nearly 5,000,000 segments of the human genome needed to be selected using several unique characteristics. When Young envisioned and initiated this study, a benchmark processing of a single segment took four minutes. BaRC worked with the lab scientists and the systems administrators in the IT department to reduce this time to less than a second per segment by optimizing both the algorithms and the processing.

The whole-genome location analysis project was the product of a highly successful synergy created by the collaboration of the scientists in the various laboratories, BaRC, and IT staff as well as the availability of advanced, biological and computational technologies. From the design of the microarrays to the processing of the data generated by the assays, modern information technology is the key to this new biological tool and all that it will produce.

Originality

Whitehead Institute's genome-wide location analysis is a new, highly-focused approach for studying gene function.

Studying the factors that control gene expression rather than products of that expression – the



THE COMPUTERWORLD HONORS PROGRAM

CASE STUDY

ORGANIZATION:
*Whitehead Institute for
Biomedical Research*

PROJECT NAME:
Genome-wide Location Analysis

LOCATION:
*Cambridge, Massachusetts,
United States*

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Science

NOMINATING COMPANY:
EMC

more traditional methodology – brings improved insight to the processes controlling cell functions.

Success

Initially, the Whitehead research team expected it would take many months to create the first genome-wide location analysis. But, thanks to the availability of both scientific and IT resources and expertise, they accomplished this in about three weeks. This has enabled new species to be quickly added to the list for study. To date, they have created the tools to enable genome-wide location analyses of yeast, human, mouse, fruit fly, and zebrafish genomes. And this is only the beginning. The project researching embryonic stem cells described above is a prime example of the use of these biological and computational techniques and the related processing and data storage technologies.

In keeping with its core values, the Whitehead Institute has made these findings available to researchers around the world, enabling them to obtain important profiles of DNA-protein binding. In addition, the international scientific community has access to more complete and reliable information for networks of gene expression in specific cells, tissues, or entire organisms in a matter of weeks – information that previously would have taken years to develop.

Information Technology is one of the keys to Whitehead being able to deliver on its mission of serving the greater good – not generating profit – by making information openly available to the entire international research community.

This innovation has already been used and produced published results including the embryonic stem cell research previously cited.

Difficulty

The primary difficulty was the magnitude of the data to be analyzed and therefore the apparent time that would be necessary to complete the processing of a single genome. As described above the team of scientists, Bioinformatics specialists and IT staff working together reduced the processing time from the initially predicted year to a mere three weeks.

Additionally an infrastructure of sufficient storage and computational resources was in place when it was needed. With this high-performance computing and data storage infrastructure, Whitehead can easily and reliably satisfy the rigorous, computing and data-intensive requirements of its research environment. It is also possible to store and manage the growing amount of biological data and ensure that it is well protected and easily accessible to researchers.

There were no objections to overcome. Innovative research is one of the founding principles of the Whitehead Institute and the human and infrastructure resources necessary for this and other research are provided as part of the Institute's core scientific support.