



<http://www.scientificcomputing.com/articles-HPC-Genomics-on-the-Petascale-052311.aspx>

Genomics on the Petascale

How HPC is transforming biological research

Mark Borodovsky and David A. Bader

As a new era centered on human health dawns around the world, the life sciences — accelerated by the tremendous bloom of genomics — are poised to open new horizons. And, in this relatively new, interdisciplinary branch of biology called genomics, computing plays a critical role, particularly in such areas as genome assembly, analysis and interpretation.

Genome sequences are available for many organisms, but making biological sense of the genomic data requires high performance computing methods and an evolutionary perspective, whether one is trying to understand how genes of new functions arise, why genes are organized as they are in chromosomes, or why these arrangements are subject to change.

In the last five years, next-generation sequencing technologies, pioneered by 454 Life Sciences, Solexa/Illumina, SOLiD/ABI and other industry beacons, have produced an acceleration of sequencing speed three orders faster than the best Sanger machines renowned for their use in the Human Genome Project. Today, with the genomes of more than 5,000 biological species sequenced or well in progress, the complexity of algorithms applied for genome decoding, gene identification, comparison and inference of biological function and evolution grows fast.

In any normally equipped bioinformatics lab, one will find computers whose running times of bioinformatics applications are simply too slow for effective workflow. Hence, genomics becomes a research area where high performance computing applications are perhaps in the highest demand.

Decoding *Fragaria vesca*

Consider as one example the success of an international consortium working on the genome of the woodland strawberry, an endeavor that depended heavily on the accuracy and speed of two new gene-prediction algorithms developed in the lab of Mark Borodovsky at Georgia Tech, one of the authors of this article.

The first algorithm, GeneMark.hmm-ES, takes as input an anonymous genomic sequence, then works in iterations to converge on algorithm parameters that deliver probabilistic (Hidden Markov) models of genomic regions that carry and do not carry the genetic code. Afterward, the thus-defined models are used in Viterbi algorithms that parse the genomic sequence into coding and non-coding regions. The second algorithm, GeneMark-ES+, "pre-processes" the transcriptome sequence data that provide additional evidence for protein coding genes and integrates this evidence as a restriction into the Viterbi optimal parse.

Working on a strawberry genome of some 240 million nucleotides, and 11 million long sequences from the expressed transcripts, it would take two weeks for each full run of these algorithms on conventional computers. And, to repeat these runs on several genome assemblies and transcriptome versions would add another four or five months to the project.

Fortunately, the Borodovsky lab had access to a teraflop cluster sponsored by NIH with the ability to perform parallel computations on 300 processors. This computer reduced the time of a full run of the two algorithms to just four hours, and the results of all necessary runs were obtained within a total of 48 hours. The algorithms confidently — and quickly! — identified 34,809 strawberry genes encoding strawberry proteins.

A paper on the strawberry genome was published in the journal *Nature Genetics* in December 2010. From a genetic standpoint, the woodland strawberry, formally known as *Fragaria vesca*, is similar to the cultivated strawberry, but less complex, making it easier to study, breed and improve.

Supercomputing toward phylogenies

Still, decoding of larger and more complex genomes, such as the 20 GB genome of Norway spruce (otherwise known as a Christmas tree), is posing even larger challenges. Borodovsky and his fellow Georgia Tech researchers are hoping to address these open problems by developing new algorithms that will delineate large swaths of surely non-coding regions, prior to pinpointing exact locations of protein-coding genes.

Given the ready availability of sequenced genomes, a natural question arises: How are these sequences related in evolutionary terms? And how can we use those similarities to reconstruct their phylogenetic history? A phylogeny is an evolutionary tree reconstructed from its leaves (each of which represents a different species) by comparing DNA sequences or gene data with a plausible model of evolution. Because phylogenies are crucial to answering many

fundamental open questions in biomolecular evolution, biologists have a strong interest in algorithms that enable resolution of such ancient relationships.

A considerable body of applied research depends on these algorithms as well. Pharmaceutical companies use phylogenetic analysis in drug discovery — for instance, in discovering biochemical pathways unique to target organisms. Health organizations study the phylogenies of such organisms as HIV to understand their epidemiologies and to aid in predicting the course of disease over time within an individual. Government laboratories work to develop improved strains of basic foodstuffs, such as rice, wheat and potatoes, using an understanding of the phylogenetic distribution of variation in wild populations. Finally, the reconstruction of large phylogenies could yield fundamental new insights into the process of evolution itself.

Technological advances in high-throughput DNA sequencing have opened up the possibility of determining how living things are related by analyzing the ways in which their genes have been rearranged on chromosomes. However, inferring such evolutionary relationships from rearrangement events is computationally intensive even on the most advanced computing systems available today.

GRAPPA-ling with genomes

Georgia Tech computational scientist David A. Bader, one of the authors of this article, leads a research group that has developed the software package GRAPPA for reconstructing evolutionary histories using gene-order data. GRAPPA was first implemented to use breakpoint distance between genomes. Bader's research took a new approach: using the *inversion* distance between genomes, which is a more biologically accurate measure. His team designed new techniques for reconstructing large-scale phylogenies with hundreds to thousands of taxa. For example, on a dataset of a dozen bellflower genomes, the latest version of GRAPPA determined the flowers' evolutionary relatedness a billion times faster than the original implementation, which did not utilize parallel processing or optimization.

Research recently funded by the American Recovery and Reinvestment Act of 2009 aims to develop computational tools that will utilize next-generation petascale computers to understand genomic evolution. The four-year \$1 million project, supported by the National Science Foundation's PetaApps program, was awarded to a team of universities that includes Georgia Tech, University of South Carolina (USC) and Pennsylvania State University.

Even on today's fastest parallel computers, it could take centuries to analyze genome rearrangements for large, complex organisms. That is why the research team — led by Bader at Georgia Tech and also including Jijun Tang, an associate professor of computer science and engineering at USC; and Stephen Schaeffer, an associate professor of biology at Penn State — is focusing on future generations of petascale machines, which will be able to process more than a thousand trillion (or 10^{15}) calculations per second. Today, most personal computers can only process a few hundred thousand calculations per second.

The researchers are developing a new high performance software package called COGNAC (Comparing Orders of Genes using Novel Algorithms and high-performance Computers) and will test the performance of their new algorithms by analyzing a collection of fruit fly (*Drosophila*) genomes. The analysis of genome rearrangements in *Drosophila* will provide a relatively simple system to understand the mechanisms that underlie gene order diversity, which can later be extended to more complex mammalian genomes, such as primates.

These new algorithms, the researchers believe, will make genome rearrangement analysis more reliable and efficient, while potentially revealing new evolutionary patterns. Armed with this understanding, scientists in fields from drug discovery to food production and dozens of others will be better equipped to address many of the grand challenges facing humanity today.

References

1. Strawberry genome paper: http://opal.biology.gatech.edu/GeneMark/PAPERS/Strawberry_NG.pdf < br /> 2. GeneMark-ES: <http://nar.oxfordjournals.org/content/33/20/6494.full>
3. GRAPPA: "Industrial Applications of High-Performance Computing for Phylogeny Reconstruction," D.A. Bader, B. M.E. Moret, and L. Vawter, SPIE ITCOM: Commercial Applications for High-Performance Computing (SPIE ITCOM2001), Denver, CO, SPIE Vol. 4528, pp. 159-168, August 21-22, 2001.
4. COGNAC: "Rec-DCM-Eigen: Reconstructing a Less Parsimoniousbut More Accurate Tree in Shorter Time," Seunghwa Kang, Jijun Tang, Stephen W. Schaeffer, and David A. Bader. Technical Report, Georgia Institute of Technology, February 2011.
5. COGNAC: "On the Design of Architecture-Aware Algorithms for Emerging Applications," Seunghwa Kang, Ph.D. Dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, January 2011.

Mark Borodovsky is Regents' Professor in the Wallace H. Coulter Department of Biomedical Engineering and the School of Computational Science & Engineering, as well as Director of the Center of Bioinformatics and Computational Genomics at the Georgia Institute of Technology. David A. Bader is Professor in the School of Computational Science & Engineering and Executive Director of High Performance Computing at the Georgia Institute of Technology. They may be reached at editor@ScientificComputing.com.

Copyright 2011 Scientific Computing, Advantage Business Media