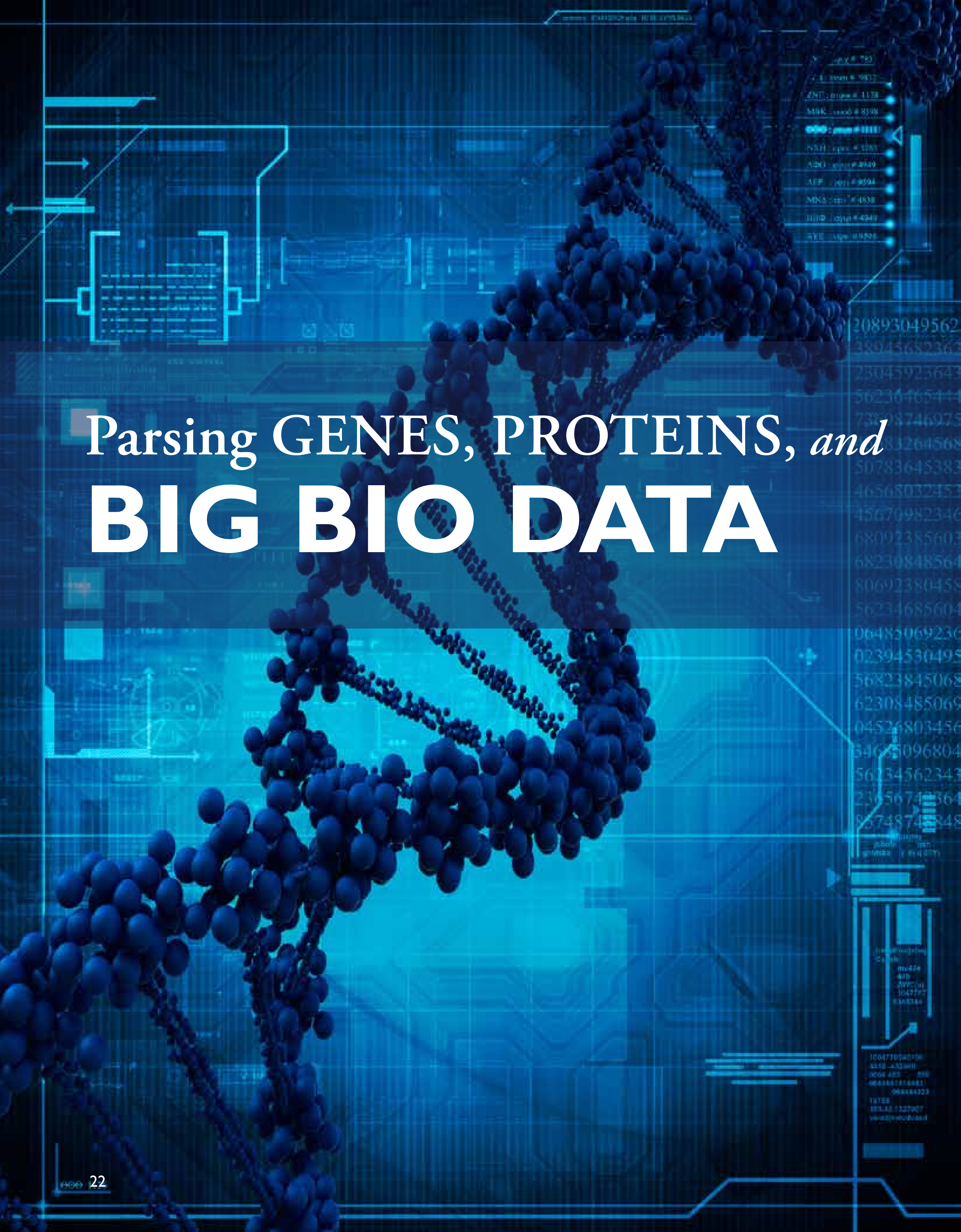


Parsing GENES, PROTEINS, *and* BIG BIO DATA

Wayne Pfeiffer is an SDSC Distinguished Scientist. His current research focuses on computationally challenging problems in bioinformatics, including genomic variation analyses.

SDSC's data-intensive computing resources have proven to be a boon to biologists interested in rapidly sifting through ever-expanding amounts of data or trying to tame the tidal wave of genomic data used to sequence the DNA of an organism, whether human, plant, or jellyfish.

"Next-generation sequencing has profoundly transformed biology and medicine, providing insight into our origins and diseases," according to Wayne Pfeiffer, a Distinguished Scientist at SDSC. "However, obtaining that insight from the data deluge requires complex software and increasingly powerful computers."

Available for use by industry and government agencies, SDSC's *Gordon* and *Trestles* are part of the NSF's (National Science Foundation) XSEDE (eXtreme Science and Engineering Discovery Environment) program, a nationwide partnership comprising 16 supercomputers as well as high-end visualization and data analysis resources.

Following are examples of how *Gordon* and *Trestles*, as well as the development of "science gateways" that give researchers Web-based access to these and other HPC systems, are improving the scope and accessibility of essential biological research databases, while creating faster and more effective ways to assemble genomic information.

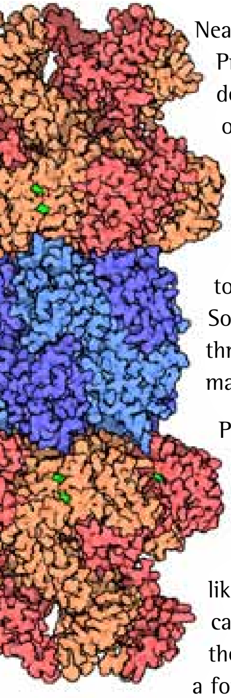
One Search to Bind Them: IntegromeDB

The diversity of biological fields has spawned thousands of databases and millions of public biomedical, biochemical, and drug and disease-related resources. For researchers interested in collecting information from those resources, search engines such as Google are of limited use since they are unable to comprehend the language of biology. They return results on the basis of keywords rather than in terms of scientific importance.

Michael Baitaluk and Julia Ponomarenko, principal investigators at SDSC, created a "smart" search for biologists, one able to return gene- and protein-centered knowledge in a biologically meaningful way—for example, by pathways, binding partners, structures, mutations, associated diseases, splice variants, or experiments. Called IntegromeDB for its ability to integrate biomedical data, the resource includes more than 16 million experimental findings. Receptor binding data for drugs and bioactive compounds, kinetic information for drug-metabolizing enzymes, and relevant signaling proteins are semantically linked to nearly 120 ontologies with a controlled vocabulary of approximately 70 million synonyms. Since its launch in January 2012, more than 4,000 users have visited the resource, and it has become an official science gateway for the NSF.

Stored in a PostgreSQL database, IntegromeDB contains more than 5,000 tables, 500 billion rows, and 50 terabytes of data. Baitaluk predicts IntegromeDB will eventually require more than a single petabyte of storage. The resource utilizes 16 compute and four I/O nodes of *Gordon* and 150 terabytes on SDSC's *Data Oasis* storage system.

Unclogging a Bottleneck for the Protein Data Bank



Nearly 250,000 scientists take advantage of the RCSB Protein Data Bank each month, every one of them depending on the resource to quickly provide details on more than 90,000 proteins, nucleic acids, and complex assemblies. While the PDB's current resources can easily handle research requests such as pairwise protein comparisons, the calculation of large numbers of protein structure alignments is too computationally intensive to be done in real time. So the PDB pre-calculates a large number of pairwise three-dimensional protein structure alignments and makes them available via its website.

Periodically, those alignments are recalculated as new protein structures and deposited into the database. However, the process of updating slows at the server that stands between the PDB and the nodes used to perform the alignment calculations. Much like an overwhelmed traffic interchange, the system cannot keep up with the data going and coming from the database, creating a data traffic jam. Phil Bourne, a former professor of pharmacology at UC San Diego, and Andreas Prlić, a senior scientist with the university, investigated whether this process could be improved using *Gordon*.

They found that the calculations were sped up 3.8 times; previous calculations that required 24 hours now took 6.3 hours. To gauge whether I/O performance might improve even further, Bourne and his colleagues tried the same calculations using Intel's new Taylorsville flash drives. The drives delivered twice as much bandwidth and read IOPS, and 13 times more write IOPS than Intel's Lyndonville flash drives. This drove the time down to 4.1 hours.

"With its excellent communications capabilities, *Gordon* can be used to greatly reduce the time to solution over the systems we currently use," said Bourne.

Taming the tidal wave of genomic data

Knowing the whole genome of various species underlies biological and medical research, such as understanding evolution pathways or identifying the causes of diseases. However, existing sequencing techniques produce huge amounts—billions for a high organism such as a human—of overlapping short sequences randomly sampled from the genome. A major challenge in genome research is to assemble these short reads, which vary from ten to several hundred bases, back into the whole genome, a task that requires vast amounts of memory. It would be similar to gluing together an encyclopedia from a haystack of words and sentence fragments.

Using *Trestles*, Xifeng Yan, the Venkatesh Narayanamurti Chair of Computer Science at the University of California, Santa Barbara, and his colleagues demonstrated that a new algorithm called MSP reduces one of the steps required so that it uses significantly less memory, a mere 10 gigabytes, than widely used algorithms. The results promise to remove one of the bottlenecks to processing whole genomes, thus making it possible to assemble large genomes using smaller, less expensive, commodity clusters.

"High-quality genome sequencing is foundational to many critical biological and medical problems," said Yan. "With the advent of massively parallel DNA sequencing technologies, how to manage and process the big sequence data has become an important issue. Experimental results showed that MSP can not only successfully complete the tasks on very large datasets within a small amount of memory, but also achieve better performance than existing state-of-the-art algorithms."

AAA+ Protease, image courtesy of the RCSB Protein Data Bank (above)

SDSC's *Trestles* supercomputer (right)

Mark Miller leads the Next Generation Tools for Biology group at SDSC and is founder of the CIPRES science gateway.

Reducing Research Barriers through Science Gateways

Making supercomputers more accessible to researchers is another area of focus at SDSC. One solution is the development of science gateways, or virtual environments that provide researchers with web-based access to tools, applications, computing resources, and data archives to further their scientific studies. Researchers can access top-tier resources, such as applications running on a supercomputer, remote instruments such as telescopes or electron microscopes, or curated data collections.

SDSC last year received a $1.5 million NSF award to make access to supercomputing resources simpler and more flexible for phylogenetic researchers. The award, which follows an earlier NSF grant that ran from 2003 to 2008, was for the CIPRES Science Gateway, a web site that allows researchers to explore evolutionary relationships between species using SDSC supercomputers as well as systems in XSEDE's repertoire. CIPRES stands for CyberInfrastructure for Phylogenetic Research.

"The CIPRES Gateway allows scientists to conduct their research in significantly shorter times without having to understand how to operate supercomputers," said Mark Miller, principal investigator for the gateway and an SDSC researcher. At of the end of 2013, the CIPRES Science Gateway supported more than 8,600 users and led to more than 700 publications of phylogenetic studies involving species in every branch of the Tree of Life.

Beyond phylogenetics, SDSC is a partner under a $5 million NSF grant to help build a science gateway service platform that will give researchers improved access to a variety of hosted or cloud services. Called SciGaP, the project is a collaboration among researchers at Indiana University and the University of Texas aimed at significantly lowering the development overhead for communities that wish to create new science gateways, allowing gateway creators to focus on developing new capabilities that are unique to an individual gateway's scientific community.

Science Gateway group leaders Miller and Amit Majumdar are leading SDSC's participation in the project. "With the SciGaP project we hope to enable a large number of existing and new science gateways from various domain sciences," said Majumdar, interim director of SDSC's Data Enabled Scientific Computing division.

Tree of Life image courtesy of Nick Kurzenko, Greg Rouse, and the U. S. Fish and Wildlife Service.