Update

Perspectives on High-Performance Data Analysis: The Life Sciences

Steve Conway
Chirag Dekate, Ph.D.
Earl C. Joseph, Ph.D.

IN THIS UPDATE

This IDC update presents near-verbatim notes from talks on high-performance data analysis (HPDA) in the life sciences given at IDC-organized HPC User Forum meetings. Together, these talks provide a good feeling for the wide variety of life science applications that have begun to exploit high-performance computing (HPC) resources for HPDA. IDC coined the term *high-performance data analysis* to refer to the use of HPC resources to run big data workloads, especially advanced analytics and data-intensive modeling and simulation. IDC believes that the life sciences will become one of the most economically and socially important markets for HPDA.

"Drug Discovery at the Petascale: Opportunities and Challenges": Jerome Baudry, University of Tennessee

I'm a computational biophysicist by training. We're new to the HPC world and went to HPC to answer some equations that couldn't be answered without HPC. We run on top machines at Oak Ridge National Laboratory (ORNL) and the University of Tennessee (UT). The pharmaceutical industry very much needs HPC. A drug is a small molecule that binds to a much larger molecule that is linked to disease. The drug discovery process is linking a disease with a protein. It starts with a screening of many small molecules with an interesting protein target to test their capacity to bind with the protein target. This process costs $10–20 per test and there are 80 million molecules in chemical databases out there.

This is too many to screen in this way, so you have to find ways to reduce this number. Computationally, you want to simulate this process of *in vitro* binding. The docking calculations involve a lot of physics and chemistry. You have to translate the chemical information into physical properties and then calculate the interaction's potential/free energy. Then you rank the compounds and energies and structures to see which ones are likely to bind. This is the reduced set of potential small molecules you work with. Many companies, including Schrödinger and Accelrys, are in this business.

Beyond Single-Protein Screening

We aim to screen massive databases of chemicals (ligands) against a large number of proteins. The pharma industry suffers from a dismal failure rate. Why do drugs fail after 10 years and $1 billion of investment? Half of small molecules that work *in vitro* move on to preclinical trials, but fewer than 10%

go on to survive late-stage clinical testing. So 99% of initial chemicals fail. It's like a car plant where 99% of the cars fail to qualify.

Why? Many molecules bind to more than one protein. So this is a combinatorial problem. Millions of protein/drug candidate pairs need to be processed, and for this, we need supercomputers. We have access to Titan at ORNL (20PFLOPS, 300,000 cores, and 19,000 GPUs) and Kraken at UT (1PFLOPS).

Results

On Titan, we reduced the time for docking 100,000 compounds with 100 protein snapshots from 38 years to 1 hour (theoretically). But piling up cores is not enough. Throwing docking jobs at a big computer is very inefficient. This process plateaus very quickly. At best we can push it to maybe 512 cores. There is no way we can render on 300,000 cores. The bottlenecks are all the calculations to prepare for screening/docking and the data traffic jams this creates. The solution: parallelize MPI and you can exploit more than 16,000 cores. To get beyond this, we sort the tasks by complexity (ligand complexity and multiple proteins) and make sure every "worker" has an equal number of tasks to carry out. Our results scaled to 85% of Kraken (84,672 cores) and to about half of Titan. We expect to be able to scale to use all of Titan. If we do this, we could run 250 million docking events per day. We could do even better by learning how to exploit the GPUs better.

Big Data Issues

The existing technology has not been very concerned with data so far, but multiple-protein docking gets insane quickly. This will generate exabytes of data, so we have to store only what we really need. And visualization is important for interacting with collaborators. We don't know how we can visualize all this data. Another issue is moving the data. We need faster, bigger communications pipes.

Conclusions

Virtual assays move from simulating a test to simulating a patient. We have the hardware and software, but we are facing a major quantitative and qualitative data issue.

The drug industry handles well what happens in the test tube but still has this miserable failure rate getting to the point where it works well in humans. Their motto seems to be: fail early, fail cheap.

Comment: Industry partners: What role do you and they play?

Speaker: We are publicly funded and this affects how we can use our technology, but we partner with pharmaceutical companies and others. The pharma industry is moving away from early-stage discovery and focusing more on the clinical trials, and they are subcontracting a lot of the work.

"Public Health Applications of HPC": Shawn Brown, Pittsburgh Supercomputing Center

I'm a quantum chemist by training. Computational modeling in public health is a relatively new field. Subdisciplines include social science, epidemiology, systems science, big data analytics/data analysis, economics, ecology, and logistics. Computational modeling and data analytics are essential for making sense of this complexity. We tell people in this industry that they're modeling every day, so moving to computational modeling should not be too foreign. Our Public Health Applications group was formed in late 2012. We have partnerships with Johns Hopkins University, the University of Pittsburgh, RTI, UNICEF, WHO, and the University of Notre Dame.

[He showed a visualization of influenza spread through Pennsylvania.] We drop influenza into Pennsylvania in this simulation and see how it spreads. About 12 million people are in this model. It's supported by a very multidisciplinary team. Our agent-based models are based on synthetic, statistically accurate populations; households; schools; workplaces; and communities. We move people throughout their day and watch how the disease progresses. Our base community is Allegheny County, with 1.2 million people, taking about 7GB of memory. Pennsylvania, with 11.9 million people, takes about 20GB. The entire United States takes 74–300GB of RAM to run. We have to run these simulations multiple times stochastically to gain statistical significance. There's a lot of stochasticity in the models. This is really a big data analytics problem. The software we use is FRED (Framework for Reconstructing Epidemiological Dynamics). We developed this under an NSF grant and it's now available to anyone. The Pennsylvania model runs in 220 seconds and the United States in about 2 hours, thanks to parallelizing the threads in the algorithms.

We published a paper on whether a school closure for the 2009 H1N1 influenza epidemic would have been worth the cost. The model shows the effect of school closure on how many get infected, and we showed you'd have to keep the school closed for eight weeks to make a real difference, which is enormously costly to society. So our study told policymakers that extended school closings are not a policy to be taken lightly.

Another project is GAIA, to visualize disease spreads and case diagnoses. It's a visualization Web service, and we're trying to build interactivity into it. Yet another model is CLARA, which looks at dengue fever in developing countries. 2.5 billion people per year are at risk for this. It needs distributed memory parallelism, and we need to go to this. The model models the full mosquito life cycle and includes humans, too.

We're also providing cyberinfrastructure for public health applications to put powerful HPC tools into the hands of decision makers and create user-facing tools. VECNet is a consortium providing Web-based platforms for the eradication of malaria. This is funded by the Bill & Melinda Gates Foundation. PSC is part of this consortium.

My favorite project is modeling vaccine supply chains in developing countries. We are in the Decade of Vaccines, according to the world public health community. The problem is, once you get the vaccines into the country, many of them are wasted because the supply chains can't get them to the people who need them. To address this, we developed a software package named Hermes, which simulates the supply chain in great detail. This lets countries explore their supply chains and what it would take and cost to optimize them.

"From Personalized Medicine to Clean Energy Production — Accelerating Multicellular Biological System Simulation Using BioCellion": Seunghwa Kang, Pacific Northwest National Laboratory

My main training is in HPC, but I'm now 70% HPC and 30% a computational biologist and am aiming for 50/50. I'll talk about simulating biological systems with many, many cells. Bio systems are very important in many domains, from the natural environment to infectious diseases. In biofuels, one important thing is finding the right biological strain for the conversion process. In bioremediation, the challenge is creating genetically engineered bacteria. In breast cancer, one challenge is establishing guidelines for when secondary surgery is needed.

Our simulations included individual cell behavior, cell-cell interactions, and cell-environment interactions. There are two main modeling approaches: population-based approaches, which are computationally inexpensive, and discrete agent-based modeling, which works better but is computationally demanding (>10^6 cells).

Major Challenge and BioCellion Approach

There are no established, stable mathematical models. Models vary and are changed all the time. They get outdated even before you finish coding, many times. With BioCellion, users provide the model specifics, and BioCellion addresses the high-performance parallel computing challenges. This makes modeling with several million cells feasible.

Evolution is a slow process, so biological systems don't change that fast. But what we don't know are the details. BioCellion is being used for many things, including micro tumor growth in yeast colonies. I'm working with others on modeling bacteria systems using thermodynamic principles. The plan is to move from single cell to multicell simulations using BioCellion. Ilya Shmulevich of the Institute for Systems Biology and others are building prototype skin models in partnership with Procter & Gamble.

Our future road map includes mapping a single cell to multiple discrete agents and also modeling water flow in soil aggregate and modeling microvascular blood flow.

Conclusions

With help from the HPC community, the computational modeling of discrete agents should not be a bottleneck. We plan to release the first version of BioCellion soon.

"Accelerating Individual Treatments for Pediatric Cancer": Glen Otero, Dell

Research on Adults' Cancers Versus Children's Cancers

Precision medicine: with adult cancers, we know how to do this very well. With the genome, we're looking for specific vulnerabilities that we can target a drug to. Everyone's cancer behaves differently, so we need customized treatments and this really helps survival rates. Not many studies have been done with children. In adult oncology, we've seen great results with many types of cancers using personalized treatments based on sequencing the individuals' genomes. Pharma companies have

funded very little in the way of research on children. We're targeting children's blastoma, which has very poor outcomes even after remission.

The Collaboration with TGen

Dell and TGen (Translational Genomics Research Institute) are collaborating to cure more children and get the treatment model out there to the whole world. I started this project about two years ago. I'm a bioinformatics architect at Dell, and my job is to take this whole pipeline and figure out how we can make this precision medicine pipeline work better and faster. It goes from tumor sample to molecular characterization of tumor to mapping against a database and then mapping the individual result to Dell cluster systems. When we started, it took 9–10 weeks from start to finding a treatment for a child. At Arizona State University, these jobs had to wait in a queue and it took days to get the data there, and these delays were unnecessary. Now we use faster pipes and there's no wait in the queue, so we save 4–5 days.

The next level is to look at the genome processing pipeline. This involves RNA-sequence data analysis. This first took 7 days, with a goal to reduce it to 5 days. We worked with TGen to parallelize its RNA-sequence pipeline. We accomplished this. Our M420s had the best energy efficiency, it turned out. They were 64% more energy efficient than alternatives. Many smaller sites focus more on this than on GFLOPS per dollar. One 10U chassis of 32 x M420s can complete the RNA-sequence pipeline in 4 hours, versus the goal of 5 days. This is akin to the time it takes to administer and get results from a blood test.

The Dell Scalable Unit for Life Sciences: SANGER

SANGER (Sequence Analysis 'N' Genomics Research) is basically a genomics platform in a rack. It's been built and tested to run the pipelines I discussed. 12 human genomes (30x) are processed per day. It consumes 20.23kWh/genome. We have this on a HIPAA-compliant cloud that allows all the collaborators to work together. In 2014, we did five trials with a total of 600 children. We expect these methods to become standard at hospitals in three to five years.

Comment: What are the challenges going forward?

Speaker: The biggest hurdle is the software, much of which is open source. There is a reference human genome but not a standard genome that we could go into the clinic with. That's one thing that's out of my control. The second thing is that there's no easy way to package and deploy the software. Things change so fast, and the software is coming in from all over the earth.

"HPC Software Applications for Biotechnology": Jaques Reifman, U.S. Army Medical Research and Materiel Command

I made a career change a number of years ago and moved to computational biology. I am with DoD's Biotechnology HPC Software Applications Institute. This was started in 2004 on a grant to develop software applications using HPC to support the Armed Forces. There was little use of HPC then in military medical research. We later decided to apply our applications to the sponsor's medical challenges. Our chain of command starts with the Medical Command (MEDCOM) through USAMRMC

and down to Headquarters Lab. Our work cuts across all the missions of the Army. We now have 41 staff with diverse backgrounds in DOE labs, big pharma, academia, and NIH.

Our three research programs are human physiology, cells and biological systems, and molecules, proteins, and DNA sequences. 40% of our computation involves HPC and the rest is desktop. We do a lot of interdisciplinary research. The Medical Chemical and Biological Defense program is under DTRA (Defense Threat Reduction Agency). One project involves finding the biomarkers of Burkholderia (BH) mallei, which affects horses and other animals. What causes virulence and what are the host proteins that make the transmission possible? To find this out, we put together a multidisciplinary team and took our usual systems biology approach that included *in vitro* assays and other steps. We scan each of the 5,500 proteins of BH mallei using HPC, and this lets us generate a hypothesis. This lets us efficiently identify biomarkers through systems biology.

Another study underway is trying to understand drug resistance in the influenza virus. The high mutation rate leads to drug resistance. We did six different simulations to predict changes in binding affinity. This has benefits for combat casualty care. The goal here is to reduce morbidity and mortality. Most deaths are from bleeding to death. We're studying clot formation and just started this a few months ago. With military operational medicine, the stressors are heat, cold, anxiety, and others. The mechanisms causing heat stress are not really well known. We use animal experimentation to predict organ damage as a result of exposure to extreme temperatures.

"Translational Bioinformatics at GE Global Research": Chinnappa Kodira, General Electric

This refers to translating complex medicine into precision medicine. Leroy Hood said, "Disease is a consequence of one or usually multiple perturbed networks and will require a systems biology approach." Healthcare analytics should span the whole continuum of businesses and services, from R&D to the delivery of services. It should span from early diagnosis to selecting optimal therapy to predicting and reducing adverse reactions, gene mapping, and more. Our focus is on cardiovascular oncology and related topics. Cancer is a complex disease. It is not just a result of single mutations but of complex interactions between molecules in biological systems. The same gene can mutate across multiple diseases/cancers. But what's important are the pathways. We are focusing on the impact of mutations on pathways. There's a widespread adoption of next-generation genomics. We need to develop analytical frameworks to correlate the mutations with the diseases.

Cancer Therapy for Recurrent Mutations

A man developed multiple melanomas and was given an experimental drug that completely cured him, but 23 weeks later the cancer returned and the patient died. So there was a secondary mutation in another gene that bypassed the pathways addressed by the drug, and so the patient died. We have to interrogate variants that often escape detection in the first phase of the disease. In the early days, cancer was diagnosed based on cell morphology. Then came basic molecular tests and expanded molecular analysis to find out what type of the disease the person has, so we can decide what treatment will produce the best outcome. NGS-based personalized medicine is about delivering high-quality and cost-effective healthcare. If we used this today, we estimate there would be a 34% reduction in ineffective chemo for breast cancer and other benefits. Big data challenges include scalable solutions where we can bring data from many sources and build analytics on top of that.

General Electric (GE) is an engineering powerhouse. We make turbines, engines, and more. GE's industrial Internet is pushing the boundaries of minds and machines. This is connecting the world's machines with intelligent software, advanced sensors, and more. Then comes advanced analytics. We now have 350 people on this initiative, and it will double in coming months. Tomorrow's medicine will combine many elements:

- We recently bought a sequencing provider and Clarient for cancer diagnostics.
- GE is working with the NFL on a $40 million project to study brain concussions. NGS bioinformatics enables hypotheses and data-driven research, using data interpretation, analytics, and outcome prediction. We sequence, annotate, analyze, and visualize.
- MultiOmyx is a technology for probing tumor heterogeneity. Through repeated staining and bleaching, we can see the expressions of more than 100 proteins. This helps to show how a host is responding to a cancer. It has led to many collaborations with pharmas and academia.
- Multimodal analysis of cancer combines medical imaging, pathology, Omics data, and clinical, EMR, and bio databases. The goal is to make data sources talk to each others.

"ORNL NIH Collaboration Using Social Media for Epidemiological Research — The Application and Value of Infodemiology": Georgia Tourassi, Oak Ridge National Laboratory

This is part of ORNL's Health Data Sciences Institute. Today, I'll talk about one of our projects (we do health informatics, health economics, and more). Infodemiology is using social media information to know more about epidemics. It includes detecting disparities in information availability, monitoring sentiment analysis, tracking the effectiveness of health marketing campaigns, and investigating unknown drug side effects and complications. Social media use among Internet users really tapers off after age 55. We have to be careful not to bias conclusions by placing too much stress on certain age groups or other factors.

NCI studies environmental cancer risk and migration pattern. Typical studies would take people with similar migration patterns and look at patterns of disease (e.g., the link between breast cancer and diet among Asian women who stayed in Asia and Asian women who moved to California). Today, people are far more mobile than before, so we apply a cyberinformatics approach to the problem. We proposed to find profiles of people on the Internet (that they volunteered) to establish lifelines for these people and their locations and create environmental, spatiotemporal profiles so we can compare incidences of cancer or other diseases based on commonalities. This is a knowledge discovery approach. This works well because cancer patients are common and represent one out of five Internet users and a growing number of cancer patients share experiences online. The only way to evaluate the reliability of the findings is to try to reproduce the findings via very costly immunological studies.

Breast Cancer and Childbirth

To assess breast cancer risk, you take a group diagnosed with breast cancer and a group without breast cancer, then look at the proportion that gave birth. It turns out women who have children have a lower risk of breast cancer in their lifetimes. We decided to proceed by looking at online obituaries, which often have a lot of information. What we need is females, age, children or not, and cause of death. But there are limits, such as that obits don't always distinguish between biological children and

stepchildren and adopted children. We identified 20,000 with breast cancer and 16,000 without, after cleaning the obituary data. We looked at the age distribution and as expected the "without" group lived on average longer. The difference between those who had given birth and those who hadn't was statistically significant. We discovered that the data from our methods were very similar to those reported by the NCI. We think our approach can be used as a hypothesis generator and does not replace established approaches. Our approach lets us monitor situations dynamically. Our work is supported by the NCI.

"Modeling Biological Systems and Analyzing Large-Scale Data Sets": Ilya Shmulevich, Institute for Systems Biology

Institute for Systems Biology (ISB) is a nonprofit research institute. I'm involved in cancer research, a large project called The Cancer Genome Atlas (TCGA). We are HPC users. This is a large project funded by NCI and another government agency to identify 25 types of cancer using multiple data types (e.g., clinical analysis, protein expression, treatment history, and pathologic reports/images). We have about 10,000 cancer samples in our collection so far. More than a dozen centers are involved across the United States. The data collected are very heterogeneous: numerical, binary, discrete, visual, and more. We want to integrate the data to learn more about the molecular basis of cancer and its progression in order to help develop more effective treatments. In colorectal cancer, there are multiple measures of aggressiveness, and we can find correlations of these with molecular features such as gene expressions. We found one molecule, FBXW7, that is strongly associated with aggressive colorectal cancer. We can then map all of these features onto the genome and see what elements on the genome are strongly associated with aggressive disease. Regulome Explorer is an interactive Web application that let users explore multivariate relationships in data.

Last year, we collaborated with Google to implement our Random Forest application on its Google Compute Engine. It spun up 600,000 cores within several minutes to support real-time processing of our application. Our collaborators include MD Anderson Cancer Center in Houston. Now that we have the same types of data on 25 different cancers, we can ask questions such as, Do some of these cancers behave similarly and might they therefore respond to the same treatments? We compute a mutational investment score to show how much of a cancer's efficacy is related to certain pathways and hallmarks. This type of analysis produces billions of associations to identify the ones that are most useful in research and to answer questions such as whether a drug developed for a certain type of cancer might be useful for a different type of cancer.

Our Approach Needs Big Computers

We create a very large graph of associations. Then we throw a lot of other things into this graph, including semantic associations from the literature, TCGA data, databases, and more. We've been using the Urika appliance from Cray YarcData because it has no locality of reference and other useful features. We load the graph as RDF triples and do SPARQL queries. We crawl the graph to find associations related to a subtype of cancer, and we go on from there.

"Accelerating Life Sciences and Personalized Medicine": Paolo Narvaez, Intel

The players in computing for personalized medicine are the payer, provider, patient, and life sciences.

Key Life Science Challenges

- Many apps are single threaded and benefit from a single address space.
- Some algorithms scale quadratically with problem size. Large data sets exceed available memory and storage.
- International collaboration is an imperative, and bioinformatics expertise is scarce.
- Databases are distributed. Data is siloed and will likely stay that way.

There's a need for a balanced computing infrastructure. Life sciences are at the intersection of HPC, cloud, and open source, all of them transformative forces. Genomics is a big data problem. It takes 313EB to sequence the genes of everyone in the United States, and 495EB to sequence the genes of every U.S. cancer patient every two weeks.

"The Anton Project at the National Resource for Biomedical Supercomputing": Jim Kasdorf, Pittsburgh Supercomputing Center

The National Resource for Biomedical Supercomputing

The National Resource for Biomedical Supercomputing (NRBSC) pursues leading-edge research in high-performance computing and the life sciences and fosters exchange between Pittsburgh Supercomputing Center (PSC) expertise in computational science and biomedical researchers nationwide.

NRBSC's focus is twofold: computational biomedical research and outreach to the national biomedical research community through education and publications. Research at NRBSC is centered in three areas: microphysiology, volumetric visualization and analysis, and computational structural biology.

NRBSC's education arm includes not only user training but also software distribution, publications, and other outreach activities such as online courses and workshop Webcasts.

The National Resource for Biomedical Supercomputing, formerly the Biomedical Initiative, was established at the Pittsburgh Supercomputing Center in 1987 as the first extramural biomedical supercomputing program in the country funded by the National Institutes of Health (NIH). Funding is subject to renewal every five years. In 2006, the NRBSC acquired its current name.

The "Anton" Supercomputer

"Anton" is a massively parallel supercomputer designed and built by D. E. Shaw Research (DESRES) for molecular modeling. It runs simulations fully in hardware, using custom ASICs and novel simulation algorithms. Anton uses customized numerical precision ranging from less than 8 bits to 72 bits, rather than fixed 32-bit or 64-bit precision.

The molecular dynamics challenge includes:

- Protein folding

- How drug molecules interact with proteins and nucleic acids
- Mode of operation of many important cellular proteins, such as ion channels

Most relevant biological events occur on a timescale of milliseconds or longer.

On a bio benchmark, Anton was thousands of times faster than a parallel supercomputer. On a 512-processor Anton, systems with up to 120,000 atoms can be simulated effectively. DESRES made available an Anton machine without cost for noncommercial research use by not-for-profits. Operations are funded by a two-year, $2.7 million NIH grant. It has four cabinets, each with 128 special-purpose ASICs.

In spring 2010, 15 proposals were awarded at a total of 100,000 node hours. The phase 1 investigators are now wrapping up their work. Anton has 162 nanosecond end-to-end latency.

"Public Cloud Computing in Healthcare": Pavan Pant, CloudSwitch

I'll talk about how healthcare, pharmaceutical, and other life science companies leverage the cloud. Reasons for going to the cloud:

- Need massive computing power for sequencing
- Used to scaling resources
- Huge cost and delays for internal provisioning (reduce capex)
- Shared resource environment that offers economies of scale

Which apps are appropriate:

- Next-generation DNA sequencing (pattern recognition, data mining)
- Molecular modeling and simulation
- Protein docking

Companies have a need for bursting/peak scale out, also for improving the application life cycles and for collaboration.

Eli Lilly, Novartis, Genentech, and Pfizer are leading the charge.

Eli Lilly

- 64-machine clusters using Amazon EC2
- Cost: $6.40
- Plan to have up to 10 HPC applications "cloud enabled" by end of 2011

Pfizer

Pfizer used Amazon EC2 to develop and refine models in antibody-antigen docking runs, shortening the process from days to hours.

What's needed to make the cloud work:

- Security
- Flexibility
- Orchestration layer
- More high-compute resources
- More streamlined procurement
- Ease of deployment/transparency with the datacenter
- Cloud resource provisioning

CloudSwitch Product Architecture

- You install (in 30 minutes) a downloadable virtual appliance in your datacenter.
- We then provide an interface that lets you virtualize your machines or create virtual machines in the cloud.
- All the data is encrypted.

Bioinformatics Use Case in the Cloud

The task from the pharma was to set up a cluster for them in the cloud. We created one and then quickly cloned it.

- 1,000 cores in Amazon EC2 to create 500 compute node clones in approximately 30 minutes. These were not the Amazon cluster instances (CCI), which are quite expensive. The total cost was less than $10,000. We started the task with a 48-hour test. The goal is to establish a more permanent footprint in the cloud.

Common use cases are capacity for burst/peak demand and development/test environments.

Comment: Is there a performance penalty for the virtualization?

Speaker: Yes, 10–20%, depending which cloud provider you use.

Comment: What about charges for uploading and downloading data? I've heard those can be very high.

Speaker: Many companies already are paying telcos for large pipes, so that's covered, but Amazon and others do charge otherwise for this.

"The Role of HPC in Defining the Phenomic Landscape for Genes, Chemicals, and Diseases": Keith Cheng, Penn State Hershey Medical Center

Why imaging? Morphology plays a key role in biomedical sciences but has often been downplayed. We need to look at the whole animal and know when one gene is on and another is off.

Grand Challenges

- **Genomics:** Know the functions of the 24,000 human genes
- **Toxicology:** Human and environmental chemical exposure
- **New drug output:** Flatlined despite increased investment because of low throughput of toxicological assessment

In the history of science and biology, we try to identify genes involved in specific functions. These are called phenotypes. You find those genes and related genes. Today, we sequence first, but we need to start with the genes and find all their functions.

Why high throughput? Scans now take 20 minutes each and we need to do multiple per animal. For 10,000, it would take 200 years, so we need to increase the speed 100 times to 12 seconds or so. The same goes for chemicals.

We need cell-level resolution to diagnose cancer, detect infectious parasites, do autopsies, and so forth.

How do we most quickly gain insight into all animal genes? What voxel size is needed for cell resolution (tomographic imaging)? The zebrafish is used as a model system. Histological analysis provides more information than gross analysis. MRI takes too long and doesn't get to cellular (micro) resolution. We've found that genes whose deficiencies cause similar loss of function have similar functions. Most gene deficiencies affect multiple organs.

The Zebrafish Genome Project

We will look at all mutants and chemicals to measure differences and identify the functions of chemicals. HPC is needed to characterize what "normal" is and to visualize. The challenges require interdisciplinary collaborations.

Comment: Are these problems computational or data processing and visualization problems?

Speaker: My understanding of HPC is that it involves faster, larger throughput.

LEARN MORE

Related Research

Additional research from IDC in the high-performance data analysis (HPDA) program includes the following documents:

- *Micron Demonstrates Technologies to Address Emerging Challenges in Big Data Applications* (IDC #244843, December 2013)
- *HPDA Pulse: 2013 Software and Consulting Market Analysis* (IDC #244513, November 2013)
- *HPDA Pulse Results: 2013 Hardware and Storage Market Analysis* (IDC #244493, November 2013)
- *Catalyst Supercomputer Heralds Shift to More Balanced Architectures* (IDC #lcUS24437513, November 2013)
- *China Eyes 10,000-Fold Data Reduction for Internet of Things* (IDC #lcUS24392513, October 2013)
- *High-Performance Data Analysis in the Life Sciences: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243774, October 2013)
- *High-Performance Data Analysis – PayPal Breaks New Ground: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243772, October 2013)
- *European Conference Underscores Movement of Enterprise Big Data Analytics to HPC Capabilities* (IDC #243618, October 2013)
- *HPDA: Importance of Hardware in Effective Big Data Solutions* (IDC #243495, September 2013)
- *eBay Deploys Innovative Green Power-Efficient Datacenter Technologies* (IDC #lcUS24350013, September 2013)
- *Worldwide High-Performance Data Analysis 2013-2017 Forecast* (IDC #241315, June 2013)
- *Changing Market Dynamics: HPC Meeting Big Data and IDC's Projected Evolution of the Market* (IDC #240365, March 2013)
- *High-Performance Data Analysis: HPC Meets Big Data* (IDC #DR2013_LSIS1_SC_CD, March 2013)
- *High-Performance Data Analysis: The Visible Human Project* (IDC #238253, December 2012)
- *High-Performance Data Analysis at NASA JPL* (IDC #238254, December 2012)

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-insights-community.com
www.idc.com