The Potential for Exascale to Transform Healthcare

Patricia Kovatch, Anthony Costa, Rong Chen, Robert Klein, Kevin D. Costa, Roger J. Hajjar, Zahi Fayad, Laurie Margolies, David Mendelson, David Yankelevitz
Icahn School of Medicine at Mount Sinai

J. Jeremy Rice, Viatcheslav Gurev
IBM T. J. Watson Research Center

Abstract
High performance computing has already been enlisted in the quest to better understand, diagnose and treat human disease. Through the expert guidance of computational scientists and advanced computing and data analytic infrastructures, advances have been made in such areas as drug discovery and genomic sequencing. However, enormous scientific challenges lie ahead to realize the promise of personalized medicine. To achieve personalized medicine's full potential, commensurate advances to reach exascale also need to be made. This paper will outline the specific scientific challenges and impacts for three areas of medicine: personalized cardiac therapy, precision medicine and real-time accurate imaging diagnosis. Then it will discuss the limitations of existing HPC along with the expected computational and data parameters and new capabilities needed for each of these areas in 2025.

Patient-specific Heart Modeling for Personalized Cardiac Therapy

The Specific Scientific Challenge
Heart disease persists as a leading human health issue worldwide reflecting our inability to predict adverse cardiac events for individual patients, as well as longstanding challenges in evaluating efficacy and risks of novel therapies on a clinically relevant population. We propose a new generation of patient-specific, multi-scale computational models that simulate cardiac performance and response to therapeutic intervention. These models would feed patient-specific data into a whole heart model that simulates the interaction of the chemical, cellular, whole organ and vascular system. These models could be correlated with large medical image data sets to obtain the necessary geometric and structural data to construct patient specific heart models and help select treatments based on the outcomes of other similar patients. With this capability, we could predict, through patient-specific simulation, how much time would pass before a specific person might have a "heart attack," which would represent a disruptive breakthrough in the fight to conquer cardiovascular disease.

Such models would require the following capabilities:

1) Patient-specific cardiac anatomy and muscle fiber direction,
2) Functionally accurate finite element models of whole-heart contractile performance incorporating electro-mechanical coupling and solid-fluid interaction,
3) Multi-scale modeling from sub-cellular signaling to whole organ function in health and disease
4) Systems-level coupling of cardiovascular physiology and metabolic function, and
5) Functional outputs to simulate clinically measureable indices of cardiovascular performance and correlate with other high dimensional data sets such as medical records and genomics data.

Early versions of (1)-(5) already exist, including: magnetic resonance imaging and diffusion tensor imaging (from Mount Sinai) that provide patient specific cardiac anatomy and fiber architecture; validated finite element models of the mechanical function of a healthy whole heart (developed at IBM and dubbed Cardioid). Disease models are currently under development in collaboration between Mount Sinai and IBM. Electromechanical coupling and solid-fluid interaction models are under active development by colleagues at other leading research institutions including The Johns Hopkins University and Stanford University. Our approach is synergistic with the multi-scale modeling and systems biology coupling at other partner institutions like UC San Diego. The Mount Sinai Health System is the largest medical network in New York and provides unparalleled access to cardiologists and cardiac patients. IBM's Cardioid model has been used to simulate the introduction of an anti-arrhythmic drug into the bloodstream and watch its absorption.

The Potential Impact
The ability to predict the timing and severity of a heart attack brings the possibly of prevention through the application of specific therapies prior to the expected attack. This information may also help individuals to

choose healthier lifestyles. For those patients unable to avoid the impact of heart disease, the ability to optimize patient-specific therapies in silico would maximize therapeutic outcomes and reduce the disease burden for a substantial segment of society. It could impact the national economy by reducing health care costs associated with the greatest health problem in the developed world. It would also streamline the discovery and development of novel drug-, gene-, and cell-based treatments that could generate revenue as the next generation of cardiac therapies. These models will advance the field tremendously by both increased predictive ability and being highly customized based on individual patient data.

The Specific Limitations of Existing HPC

Exascale-sized data repositories with secure encryption capabilities will also be required for machine learning to help correlate data from simulations with patient medical records and genomic data to help develop a specific precision therapy for a specific patient. Expert systems such as Watson could help correlate disparate data sets to suggest effective treatment plans.

Related Research Areas

Related research areas that may benefit include modeling of other debilitating diseases that require patient-specific detail and could be minimized by advanced warning, such as aortic aneurysm or intervertebral disc herniation. Solving multi-scale cell-to-organ biology problems would impact research areas in diseases such as diabetes, neurodegeneration, and various age-related illnesses. Similar approaches could help inform other organs models such as the brain. Successful multi-physics integration of solid mechanics, fluid dynamics, and electrical conduction systems could impact the study of human physiology in extreme conditions such as high altitude, deep sea, or outer space exploration, and could directly translate to a wide range of industrial applications including future energy, transportation, and manufacturing systems.

Computational Parameters Expected in 2025

The simulation for the finite element models for the contractile performance in Cardioid ((2) above) already require petascale computing capabilities. This model has already been run on 1.6 million compute cores on Sequoia BlueGene/Q at Livermore National Laboratory, simulating an hour of heart activity in seven hours of wall clock time at 0.1mm resolution. The combination of higher resolution, additional multi-physics and other multi-scale models correlated with image, genomic and medical records will clearly constitute an exascale simulation. In addition, ideally the simulations would run in hours or less for practical utility in a healthcare setting.

Other Capabilities Needed

In addition to increased computational horsepower to run these multi-scale, multi-physics, patient-specific cardiac models, other capabilities needed by the end-to-end system will include enhanced 3D visualization tools for interacting with the medical image data and for finite element modeling assembly and inspection for multi-physics compatibility; analytical tools for machine-learning interpretation, and accessible presentation (including to physicians and patients) of the complex multi-dimensional data sets generated by the model simulations; data encryption capabilities to ensure confidentiality of patient medical records; and data storage and sharing capabilities to transfer and integrate image data, medical records, and modeling software across multiple sites and platforms. Tools for automated image registration and segmentation and feature extraction will need to be developed and validated as current tools are only semi-automated. Simulation software that accurately mimics the chemicals and their coordinated efforts with cells, organs and the entire vascular system also need to be developed and validated. The resolution of Cardioid needs to be improved to 0.05 mm for greater accuracy.

Delivering on the Promise of Precision Medicine

The Specific Scientific Challenge

The insight gained from the integration of multi-scale data from new genomics technologies, national databases (e.g. dbGAP, TCGA), advanced electronic medical records, imaging, and personal biometric devices (e.g. FitBits) will transform healthcare. Precision medicine requires the dynamic correlation of multi-dimensional data of a specific patient with extremely large BioBanks with multi-modal, high dimensional data on tens of millions of patients. Currently, the ability to perform such analyses is severely limited by the scale and performance of available computational and storage resources. There is an urgent need for computational, storage and database systems to serve massive amounts of geographically-distributed databases data to many

researchers at the same time. Taking full advantage of this scientific opportunity will require I/O performance and storage infrastructure several orders of magnitude greater than what's available now.

The Potential Impact

While modern medicine is based on treating the average patient, we now recognize that the molecular and physiological details of disease in any one patient can vary greatly from the norm. Through the integration of genomic and biometric technologies with electronic medical records, it will be possible to develop sophisticated predictors of disease onset and response to treatment, enabling more precise medicine. A reduction of unnecessary tests and ineffective treatments will save significant amounts of healthcare dollars for patients as well as for local, state and federal governments. Leveraging these vast datasets to identify at-risk patients and optimize treatment strategies will reduce morbidity and mortality from myriad diseases.

The Specific Limitations of Existing HPC

Genomic data processing workflows are I/O-intensive, read and write an enormous number of tiny files and require an extremely large number of Input/Output Operations Per Second (IOPS). This is very different from traditional HPC workflows that write and read from a few, large sequential files. Although a few technologies and techniques exist today that improve the efficiency of these workflows (e.g. flash), it is unclear if they will be affordable or scale sufficiently to handle exascale-sized datasets [1]. In addition, the exascale-sized genomics data will need to be correlated with geographically-distributed national data sets and from medical records and personal biometric devices. There is no current HPC capability available to stream datasets of this size from disparate locations, let alone with any sort of performance nor is there robust infrastructure and community-adopted best-practice methodology for integrating even single modality data sets at this scale.

Computational Parameters Expected in 2025

As a typical study in this field, we imagine a study following 1,000,000 people across the country: this number was chosen to be the same order of magnitude as the Million Veteran's Study from the VA or President Obama's Precision Medicine Initiative. For each participant, there will be a single genome sequenced, five tissues sampled and analyzed with RNA-seq and three epigenomic assays every six months for five years, and the use of portable biometric devices such as FitBit to record five physiological measurements every minute. While this study is reasonable in scale and the data could easily be generated over the next decade, the storage space required is immense. For each participant:

1 Human Genome 130 GB (Estimate from recent whole genome sequencing we performed on Illumina X10 machines)
1 RNA-seq profile 5 GB x 5 tissues x 10 timepoints = 250 GB
3 Epigenomic assays 1 GB x 5 tissues x 10 timepoints x 3 = 150 GB
5 64-bit biometric measures 8 bytes x (60 minutes/hour x 24 hours/day x 365 days/year x 5 years) = 0.02 GB

This is 395 GB/person x 1,000,000 people = 377 petabytes of raw input data for this single study. Analysis of the data through community-driven genomic sequencing pipelines (e.g., alignment to references, refinement, and variant calling) is an iterative process that at each step requires significant (though non-exascale) compute, doubling of storage, and massive I/O throughput for communication and storage of intermediate files. As the data produced scales linearly with number of tissues and timepoints, the data storage required for such a study would scale similarly as the scope of the study increases. Together the total data footprint will grow to over an exabyte for a single study of this magnitude, especially as the computational methodology is refined and analyses are rerun. On a typical machine, there will be many studies conducted with similar computational and storage requirements, necessitating the need for truly exascale storage and I/O performance. Current HPC I/O is in the terabytes per seconds range for block data. At this rate, it will take a million seconds or 11.5 days to read an exabyte of data, without the random reads and writes typical of the iterative "best-practices" genomics pipeline. It should be noted that the sample study described here only touches the surface of the range of tissues and timepoints that can be sampled using epigenomic and RNA-seq studies.

Other Capabilities Needed

New applications with the capability to stream I/O from local storage and remote, national databases in real-time will be needed to avoid the latency from reading and writing from hard drives. Flash can assist but it is currently cost-prohibitive and would need to be of a size large enough to store locally copies of geographically dispersed databases. Community-driven, open best-practices data management, security, curation, and

distributed access are key requirements to the successful implementation of large-scale longitudinal studies like the one proposed as an example above.

Potential developments and pitfalls

Significant efforts have been devoted in recent years towards alleviating some of the pain experienced by the heavy I/O load leveraged on HPC resources by genomics-style workflows. These include efforts in envisioning computational genomics as a big data problem well suited for Hadoop-style resources, which is still in its infancy and has significant hurdles to overcome for efficient compute through a distributed data-driven style. Another style alleviates the random I/O bottlenecks (though not the streaming and data footprint limitations) by implementing pipelines using in-memory only models. These too have not been widely deployed. The research, results-driven and fast-paced style of development in this area have precluded the successful implementation of these kinds of approaches. I/O-heavy on-disk communication has continued to dominate the genomics space, and this will likely continue as computational methodology is refined over the next decade or more.

Real-time Accurate Imaging Diagnosis

The Specific Scientific Challenge

Healthcare providers need to determine the right test at the right frequency for optimal early detection of breast cancer, with limited false positives. Today, breast radiologists synthesize data from imaging informatics, the radiology subspecialty that includes Picture Archiving and Communication System (PACS), electronic medical records, structured reporting, computer assisted diagnosis, natural language processing (NLP), archiving, radiation dosimetry, data mining products, peer review, the exchange of health information and real-time education, to make the best diagnosis possible. As the breast imaging profession positions itself for the next twenty years of innovation in the era of big data, structured reporting and the ability to create and populate large databases that allow for data mining are critical to improve diagnosis and treatment. The ability for computer-aided diagnosis software to more quickly analyze and refine data from a much larger dataset would be transformative for patient care. For instance, if we could identify tumors with a high level of accuracy in real-time, it would revolutionize healthcare.

Developing highly accurate machine learning algorithms to perform feature extraction will require exabyte sized image repositories and exascale computing for data analysis. Correlations of specific image features would be matched with similar images from others along with genetic and other medical record information. We use all modalities, even for a simple study, asking questions such, “Does ultrasound or tomosynthesis add benefit to high risk breast screening if an MRI was done? Who needs a mammogram and at what frequency? Who needs supplemental ultrasound or MRI? More tests? Fewer tests?”

High dimensional data analysis could help answer these questions tailored to each patient. By performing some of this analysis in advance, we can work towards a real-time diagnosis suggested by software. To ensure that this computer-aided diagnosis software is highly accurate, we would need image, genetic and medical record data from tens to hundreds of millions of patients, thus requiring exabyte sized data warehouses and analysis engines. Ultimately, we want to mine previously uncollectable data to determine new, true risk factors beyond the traditional ones such as how much you walk from your cell phone, radiation exposure from flying, second hand or primary smoke exposure, food, medication, genes beyond BRCA1 and 2 and pTen and others. This would enable us to perform meaningful data mining and to improve accuracy and specificity need to amalgamate images from many institutions along with genetic and lifestyle and other risk factor data. When have all this data combined then it can be mined for the ultimate goal of preventing cancer or employing much more effective screening algorithms.

The Potential Impact

More accurate, real-time identification and diagnosis of lesions will enable faster treatment for afflicted patients and reduce patient stress for healthy patients. Coupled with genetic, medical record and other data from the digital universe, doctors will be empowered to provide more precise treatments for specific patients. The reduction of unnecessary biopsies and related tests will save significant amounts of healthcare dollars for patients as well as for local, state and federal governments.

The Specific Limitations of Existing HPC

This capability is currently limited by (1) the accuracy and complexity of the algorithms learning to automatically identify tumors, (2) the scale of the available images the algorithms need to learn from, and (3) the lack of correlation with other high dimensional data sets such as medical records and genomics data. Existing centers do not have the exabyte storage capacity to host the images or the computing power to perform the necessary exascale data analytics to correlate them genetic and medical record data. Scalable software stacks and sophisticated machine learning algorithms for imaging are also missing.

Computational Parameters Expected in 2025

At ISMMS, we already store over 10 million images in all modalities (x-ray, ultrasound, MRI, mammography, CT Scan, MRI, PET, etc.) in over 4 petabytes of storage for over seven million patients since 2003. As imaging modalities become more complex and accurate, the file sizes for each modality will grow. We estimate that an average patient has 20 images taken over their lifetime representing 1 TB/patient. If we wanted to correlate images with genomic and medical record data from 1 million patients through President Obama's Precision Medicine Initiative, then we would need an exabyte of storage to hold all of this data.

Other Capabilities Needed

To achieve our stated goal of improved diagnosis, we will need access to an order of magnitude more images in order to train algorithms to successfully and accurately identify tumors. To do this, there will need to be a national repository for images, genetic information and medical records and/or a mechanism for sharing them in a decentralized way so that all researchers will be able access the data for additional studies. We need software that facilitates communication and integration of patient data from numerous, geographically distributed sources. The repository and software could be linked to President Obama's million person precision medicine initiative for additional scientific insight and analysis.

References

[1] Kovatch, P., Costa, A., Giles, Z., Fluder, E., Cho, H., Mazurkova, S. (2015). *Big Omics Data Experience*. Supercomputing 2015. DOI: 10.1145/2807591.2807595.