Informatics Working Group Report: Analysis of Existing Strengths, Critical Gaps, and Opportunities for Collaboration
May 2014

Big Data Activities/Strengths at Rutgers University

Protein Data Bank (PDB): The Protein Data Bank was established in 1971 as the single archive for information about biological macromolecules. In 1998 Rutgers University was funded by the National Science Foundation, the National Institutes of Health and the Department of Energy to lead and manage that resource. The resource is used by academic researchers and educators, as well as by pharmaceutical scientists involved in drug discovery. It requires 24/7 operations as well as expert data management, storage and fast networking. The PDB has a very large international user community. For example, in 2012, the site had more than 250,000 unique visitors per month, and more than 350,000,000 downloads of data. Although management of the project and most aspects of the work are done at Rutgers, the current Rutgers networking would not allow us to have a robust data distribution system.

Biomedical Informatics Share Resource at CINJ: The overarching mission of the Biomedical Informatics Shared Resource is to provide the requisite tools, expertise, and training to foster advances in research and discovery while enhancing patient care and initiating and sustaining productive collaborations among investigators at Rutgers Cancer Institute of New Jersey and throughout the clinical and research communities. To meet the needs of the new Precision Medicine and Translational Programs at Rutgers Cancer Institute of New Jersey this Resource has been completely reorganized to include 4 different, but inter-related divisions which includes (1) Computational Imaging; (2) Clinical & Research IT; (3) *In silico*, Chemical Informatics; and (4) Bioinformatics & Systems Biology. The primary objectives of this Resource are (1) to provide support for large-scale, high-throughput analysis of biomedical data; (2) provide data warehousing and data-mining capabilities to facilitate cancer biology studies and precision medicine initiatives; (3) identify, develop and implement cutting-edge computational and imaging tools to facilitate clinical trials and services; (4) provide support for *in silico* chemical informatics analyses and drug discovery; and (5) provide software and applications development to support web-based and other clinical, research and educational activities.

Rutgers University Cell and DNA Repository (RUCDR): RUCDR Infinite Biologics plays a key role in research aimed at understanding the genetic causes of common, complex diseases. RUCDR activities will enable gene discovery leading to diagnoses, treatments and, eventually, cures for these diseases. RUCDR assists researchers throughout the world by providing the highest quality biomaterials, technical consultation, and logistical support. As the world's largest university-based biorepository, RUCDR has been perfecting the science of biobanking, bioprocessing and analytics since 1999. By utilizing a technologically advanced infrastructure and the highest quality biomaterials, RUCDR scientists work to convert precious biosamples into renewable resources thereby extending research capabilities. RUCDR understands that research goals and objectives vary from project to project so we give each client individual and customized attention to ensure "best fit" service

Operational Clinical IT: Rutgers Biomedical and Health Sciences (RBHS) have fully implemented state-of-the-art electronic medical records (EMR) for the faculty practice of the Robert Wood Johnson Medical School (Rutgers Robert Wood Johnson Medical Group, RWJMG), the Rutgers Cancer Institute of New Jersey (CINJ) and the Eric B. Chandler Health Center (Chandler). All 3 practices have achieved 100% adoption of the EMR and all 3 have very high rates of successful attestation for Stage 1 of Meaningful Use. The non-CINJ part of RWJMG and Chandler share the GE Centricity EMR, which greatly facilitates collaborative care of Chandler patients by specialists in the RWJMG. Physicians at Rutgers Cancer Institute have implemented the Aria EMR which features capabilities that are specific to clinical oncology. Aria also has the capability of guiding cancer treatment protocols used in clinical trials. Both EMR's are certified by the Office of the National Coordinator of Health IT (ONC) for Stage 1 of Meaningful Use and certification for Stage 2 of Meaningful Use is expected this year. Both EMR's have advanced features including decision support, electronic prescribing, electronic importation of

laboratory results as discrete data, patient portals to make data available to patients and document imaging systems to manage paper documents from external sources. All 3 practices are closely integrated with the Robert Wood Johnson University Hospital (RWJUH), with a large number of electronic data interfaces to transfer hospital data to the outpatient EMRs in order to facilitate continuity of care. The sources of data include laboratory, radiology, ER notes, admission notes, consultations, operative notes, discharge summaries, and a range of ancillary tests. In addition, an interface is being built to exchange office notes between Aria and GE Centricity in order to improve continuity of care. All 3 practices also have EMR reporting capability. RWJMG and CINJ have successfully made a transition to electronic submission of clinical quality data for the Medicare Physician Quality Reporting System (PQRS)

Use of Health IT to support clinical research: RWJMG, CINJ and Chandler have reporting capability that enables searching the EMR databases for potential candidates for clinical studies. This capability is used to determine whether sufficient patients are available for a particular study, to plan studies (e.g. by adjusting inclusion and exclusion criteria to provide sufficient subjects) and to identify actual candidates for a study. In addition CINJ participates in several disease registries. Rutgers Cancer Institute of New Jersey (RCINJ) has launched several key translational and clinical research and development projects, which will require a standards-based informatics platform which can support both small and large-scale investigations. A cornerstone for these projects is the new clinical data warehouse, which integrates data arising from Electronic Medical Records (EMR); Clinical Trial Management Systems (CTMS), Tumor Registries, Biospecimen Repositories, Radiology and Pathology archives and next generation sequencing devices. Aria has the capability of guiding cancer treatment protocols used in clinical trials.

Academic Clinical Informatics: Both CINJ and RWJMG have researchers who are engaged in NIH-funded clinical informatics research. Dr. Foran at CINJ has established active, successful programs in computational imaging and informatics. He currently leads several, R01-funded, multi-institutional projects involving team-based software development and high-performance computing focused on cancer detection, patient stratification, disease management, and clinical outcomes studies. Dr. Frank Sonnenberg has had a history of NIH funding for informatics research, including multiple RO1's and is currently an active member of the Clinical Decision Support Consortium (CDSC), an AHRQ-sponsored program of research into providing clinical decision support as a service to a variety of EMR's. Robert Wood Johnson Medical School is a designated CDSC demonstration site for the GE Centricity EMR.

Genomics Research Program: The Genomics Research Program at New Jersey Medical School provides investigators with a full complement of expertise for genomics-based research and diagnostics. The program organized into functional units with staff (16 staff members) participating within and across functional groups based on their areas of expertise. This allows us to optimize the full range of services offered by the program, from basic research to clinical diagnostics. The Clinical Genomics Program at Rutgers New Jersey Medical School is CLIA certified- CAP accredited.

Mass Spectrometry-based Proteomics: RBHS also features an active, mass spectrometry based proteomics service, which is an indispensible tool for molecular and cellular biology and for the emerging field of systems biology. Recent advances allow protein mixtures of considerable complexity to be routinely characterized in depth.

Rutgers School of Environmental and Biological Sciences (SEBS): Researchers at SEBS collect real time streaming data such as meteorology, ocean parameters such as temperature, salinity, currents, etc., as well as genomics data for, for example, bacterial populations in soil, sewers, wastewater treatment plants, etc. This data is often posted on department web sites on an ad hoc basis with help from departmental IT support, but with no centralized organization.

Statistics and Biostatistics (SAS): The Department of Statistics and Biostatistics (SAS) has established a strong array of collaborations within and outside the university, including joint faculty appointments and other long term arrangements with CS (Ping Li), genetics (Steve Buysky) and the Institute for Health, Health Care Policy and Aging Research (Donald Hoover) and funded collaborations with Baidu (Tong Zhang), Yahoo (Ping Li), FAA (Regina Liu), Pfizer (Javier Cabrera), and more. A majority of the statistics faculty members are actively engaged in funded research in Big Data and closely related areas. The Department of Statistics and Biostatistics (SAS) has made great effort in promoting Rutgers' profile

in Big Data. It is currently hosting jointly with CS a highly successful "Statistics/CS Big Data Seminar" which is partially funded by Yahoo! Labs. It will host a "Statistics and the Century of Data" symposium in May 2014 featuring several leaders in Big Data research as speakers. It has proposed to host a Big Data symposium beginning in Fall 2014 to bring together speakers and participants among world-class researchers on the fundamental development of Big Data theory, methodology, and practice. The symposium, designed as an integral component of the Rutgers Big Data Initiative, will greatly enhance the research and collaborations on Big Data across many units of the University, including Statistics and Biostatistics, Computer Science, Electric Engineering, Mathematics, Genomics, Information Science, Public Health, Medicine, Astrophysics, and Social Sciences.

Department of Computer Sciences (DCS): The New Brunswick CS department is engaged in research that overlaps with all of the three themes of the working group. Its work focuses on fundamental computational research questions raised by new kinds of data and information, including how data can be analyzed to discover new patterns and knowledge, how data can be efficiently stored, accessed, queried and processed, and how users can manage data while addressing concerns of safety, security and privacy. A major bottleneck for CS research in this area is access to significant data sources, which are often expensive to collect and difficult to share, but invariably bring new challenges on the computational side. Many large funding opportunities require an integrated vision involving both innovative technology and concrete scientific or clinical problems and data sets. So the chance to engage with other researchers, especially at RBHS, who have such data and need to do new things with them, is an exciting and necessary way to raise Rutgers's overall stature.

Rutgers-Newark: While there are a wide range of research activities that are underway on the Newark campus it is important to demonstrate the scope of the areas of expertise. To that end, we have including brief descriptions of a few representative laboratories:

The Ware Lab (Biological Sciences) has sequenced transcriptomes for 1200 species of insects using illumina at Beijing Genomics Institute, BGI. The final dataset has 1200 taxa, with a dataset after orthology prediction and alignment masking of 500,000 amino acid sites.

The Kutska Lab (Earth and Environmental Sciences) recently adapted next generation sequencing methods, using emergent single molecule real time (SMRT) technologies, to evaluate in situ phytoplankton assemblages as well as their responses to Fe and Modified Circumpolar Deep Water in the Ross Sea, Antarctica in incubation experiments.

The Cervantes Lab (Biological Sciences) played a major role in the prediction of RNA secondary and tertiary structure based on sequence and thermodynamic data; data mining of start-codon context sequences in monocots.

The Kirby Lab (Biological Sciences, Rutgers) performs gene expression analysis in plants, focus on poplar; use of Agilent microarrays and RNA seq; design and analysis of gene co-expression networks; GC-MS metabolite analysis

Rutgers University Libraries: Crucial aspects of big data support relate to data management issues while preserving it, making it accessible to researchers today and in the future. This is particularly critical in the biomedical sciences because of the federal mandate, particularly the NIH mandate, to make federally funded data products accessible via PubMed and other open access avenues. The Rutgers University Libraries developed and host RUcore, the Rutgers Community Repository, which supports faculty publications (articles, conference proceedings, presentations), electronic theses and dissertations, special collections, and research data. The repository is built upon the Fedora Commons repository architecture and supports simple and complex data files, including related resources such as multiple trial datasets for an experiment, codebooks, lab notes, images associated with data sets, etc.

Rutgers Discovery Informatics Institute (RDI^2): RDI^2 is a university-wide Computational and Data-Enabled Science and Engineering (CDS&E) resource at Rutgers University that fundamentally integrates of research, education, and infrastructure. Its overarching goal is to stimulate new thinking and new practices in computational and data sciences that are catalyzed by advanced cyberinfrastructure, towards addressing grand challenges in science, engineering, and society. RDI^2 houses the largest openly accessible research-computing platform at Rutgers, which is utilized by a large number of researchers and

students. RDI^2 has also introduced extensive education programs, including a new Master of Business and Science concentration in analytics, new certificate programs, technology boot camps and a distinguished seminar series. It is also coordinating the New Jersey Big Data Alliance, which brings together academic institutions, government organizations and industry across the state to address Big Data challenges and opportunities.

I. Analysis

Relative opportunity to be "best in class" compared to other institutions: RBHS has the opportunity to be "best in class" compared with private institutions and other competing academic institutions in measuring and demonstrating high quality of care. We also have the opportunity to excel in implementation of clinical informatics, collaboration with our partner institutions (e.g. RWJUH and the newly formed ACO) and providing outstanding information services to our patients. We started early, implemented aggressively and have achieved very high rates of Meaningful Use among our providers.

Key Strengths at RBHS: Key current strengths at RBHS/Rutgers include:

- World class investigators, departments in computer science, physics, chemistry, engineering, biology and life sciences, medicine, informatics and other related disciplines whose future success and competitiveness are reliant upon improvements to computational resources and investment in cyber-infrastructure
- Institutes such as CINJ, RDI^2, DIMACS, RUCRDR, PDB and others in which advanced computation is a common thread which will provide insight and improved understanding in applications which transcend the physical, biological and medical sciences
- Several leading departments and schools have already leveraged advanced training and experience in applied scientific computation to establish innovative initiatives and programs which with proper investment and management will lead to the next generation of funded projects and advancements
- Established information technology organization (OIT) with track record and long-term commitment to computing support, network design and data management
- HPC and research support group with vested interest in computational biology, the physical sciences, informatics and medicine
- Planned renovations and recruitment to support improvement s to physical infrastructure, staffing and large-scale computing

Limitations in Existing Capabilities & Critical Gaps: The clinical enterprise at Rutgers presents a unique set of challenges and opportunities. During the course of recent discussions with representatives from the clinical faculty, the following areas were identified as requiring attention:

Administrative:

- Need for institutional mechanisms and policies to allow technology refresh or to respond to unexpected, emerging opportunities, which present challenging computational requirements
- A Rutgers Data Science Center should be established to facilitate collaboration, to solicit and attract external funding, to recruit students interested in Big Data, and to foster closer ties with industry sponsors and partners
- An organizing focus to give clinical informatics research a home and to promote interactions with interested collaborators throughout RBHS

Personnel:

- Technical staff and support with bioinformatics and programming expertise
- Personnel to help support development and analysis of genomic data being generated by the precision medicine initiative
- Internal support staff is insufficient to provide an optimal amount of training, point-of-care application support and customization of clinical content

Technology:

- Limited support for issues surrounding HIPPA Compliance/Data Security available on campus
- Development of additional software tools to enable the clinical faculty to adhere to the emerging guidelines

- Lack of new technologies and interfaces to facilitate access to EMR's
- Lack of interoperability among software systems that support access and interrogation of data originating from EMR's, genomic sequencing analysis, and pre-clinical and clinical imaging is required to enable investigators to collaborate on large-scale, multi-disciplinary studies and projects
- Lack of intelligent repositories that can combine information, which is embedded in EMR with research databases for large-scale data mining
- Limited storage capability to handle the large databases in a highly secure and uniform manner
- Secure HIPAA compliant storage of clinical genomics data; dynamic storage and computational power for translational projects involving genomics data; integration of clinical genomic data into LIS and EMR that is currently used in Rutgers Health
- Need for new imaging and informatics technologies to enable physicians to detect and track response to therapy using objective, reproducible methods that can chronicle changes that occur over the course of longitudinal studies
- Need for an enterprise-wide clinical data warehouse to support precision medicine and drug discovery
- Need to address long-term storage, analysis, archiving, sharing of sequencing, clinical and pre-clinical imaging data is currently
- Network and computational infrastructure support for new precision medicine & translational research programs; Bandwidth to support emerging telemedicine and remote consultation activities; Non-uniform network capabilities within building and across campuses
- New technologies such as computer-based simulations to support interactive physician training and continuing medical education
- Need for bioinformatics infrastructure including clusters, storage and pervasive data access

Anecdotal Evidence of the Impact Capabilities Limitations & Critical Gaps:

Rutgers Cancer Institute of New Jersey (RCINJ): RCINJ is undertaking several key translational and clinical research and development projects, which will require a standards-based informatics platform that can support both small and large-scale investigations. A cornerstone for these projects will be the clinical data warehouse, which will integrate data arising from Electronic Medical Records (EMR); Clinical Trial Management Systems (CTMS), Tumor Registries, Biospecimen Repositories, Radiology and Pathology archives and next generation sequencing devices. Given large and growing volume of data and number of modalities that is actively gathered and archived as part of these investigations, the primary challenges for these undertakings have now become focused on the scarcity of adequate computational and data analytics resources. These applications offer tremendous potential for collaborative projects and research proposals, which transcend basic and clinical research. In order to continue to propel these projects forward, it is essential that we establish the requisite cyber-infrastructure and associated computer and network resource which allow reliable processing of the rich informational content extracted from large patient cohorts. The capabilities provided by these resources will make it feasible to conduct high-throughput screening and mining of large data sets, generate and test hypotheses. These capabilities will enable the community of investigators to stratify patient populations in multi-dimensional space; perform dynamic modeling of the changes in the molecular signatures and morphology; visualize organs, tissues and microstructures in 3D and determine precise localization of biomarkers within the tumor environment throughout the course of disease progression. Together these advances and new technologies will serve to improve prognostic accuracy and therapy planning for subpopulations of patients who have been afflicted with cancer while facilitating investigative cancer research and discovery.

Protein Data Bank (PDB): The Protein Data Bank was established in 1971 as the single archive for information about biological macromolecules. In 1998 Rutgers University was funded by the National Science Foundation, the National Institutes of Health and the Department of Energy to lead and manage that resource. The resource is used by academic researchers and educators, as well as by pharmaceutical scientists involved in drug discovery. It requires 24/7 operations as well as expert data management, storage and fast networking. The PDB has a very large international user community. For example, in 2012, the site had more than 250,000 unique visitors per month, and more than 350,000,000 downloads of data. Although management of the project and most aspects of the work are done at Rutgers, the current

Rutgers networking would not allow us to have a robust data distribution system. A strategic partnership was formed with the University of California San Diego; UCSD has two supercomputer centers and very strong network presence that would allow us to accommodate the very large global community of users. The subcontract to UCSD is for about $2,000,000 per year of which $600,000 is for indirect costs. *Over a ten-year period UCSD administration has received more than $6,000,000 for providing this service, over which, we have had no control.* Clearly it would have been to our advantage from both from the public relations perspective, as well as financially, to be able to run the entire operation at Rutgers.

Rutgers University Cell and DNA Repository (RUCDR): For the past six years, Rutgers University Cell and DNA Repository (RUCDR) has outsourced its advanced computational needs to the Information Sciences Institute (ISI) at the University of Southern California. As part of the most recent five-year, $45 M award for the NIMH Center for Collaborative Genomics Research on Mental Disorders (J. Tischfield, PI), RU subcontracts $850,000 per year ($4.3M over the next five years) to ISI for the development of computational tools accessible through the web. Efforts were made to find a partner at RU but no group was willing or able to provide this service. The main issue is that there are no groups at RU that provide service in response to the relatively circumscribed computational research demands of faculty in diverse fields. To some extent, RUCDR has built limited computational resources using its own IT staff and consultants. However, it won't be possible for RUCDR to compete with large computational groups at institutions such as Harvard, Johns Hopkins and UCSD until it can access advanced computational resources as needed. For the past 15 years RUCDR has been outsourcing its medical informatics needs to Washington University School of Medicine (WU), mainly for its NIMH, NIDA and NIAAA grants and contracts. This subcontract to WU has totaled nearly $10M over this period. Even after integration of the medical schools, RU has no presence in medical informatics as it relates to computational genomics. In particular, neither RWJMS nor NJMS have departments of human genetics or centers for computational services.