OMICS A Journal of Integrative Biology
Volume 15, Number 4, 2011

DOI: 10.1089/omi.2011.0024



# Bioinformatics and Data-Intensive Scientific Discovery in the Beginning of the 21st Century

Roger Barga,[1] Bill Howe,[2] David Beck,[2] Stuart Bowers,[1] William Dobyns,[2,3] Winston Haynes,[3] Roger Higdon,[3] Chris Howard,[3] Christian Roth,[2,3] Elizabeth Stewart,[3] Dean Welch,[3] and Eugene Kolker[2,3]

## Abstract

This article is a summary of the bioinformatics issues and challenges of data-intensive science as discussed in the NSF-funded Data-Intensive Science (DIS) workshop in Seattle, September 19–20, 2010.

## Overview

Bioinformatics is increasingly a data-driven science. Future success for life scientists will depend upon the ability to leverage the large-scale data. By adopting the advances in information technology made by fields that have already faced the type of inflection point bioinformatics face, such as cloud computing, we believe bioinformatics can weave the computational environments that exist today into a solution for our data problems. In our breakout discussions, we examined the current state of the field, current barriers to research, and concluded with an examination on the role of cloud computing to support bioinformatics research.

## Current State

In our discussion on the current state of the field of bioinformatics we addressed a wide variety of issues: data heterogeneity, lack of integration among commonly used tools, lack of data standards, common analysis bottlenecks, and a survey of tools commonly used in labs today.

## The Need for Standards to Facilitate Data Integration and Management

Diverse bioinformatics analyses spawn diverse approaches. Our discussions indicate the need for better: (1) schema-based integration methods (when schemes or ontologies are available), including better mappings to facilitate navigation from one data source to another, (2) complex joins across various databases, (3) support for provenance data, and (4) enriched meta-data catalogues to support resource discovery. The first four necessities are around data integration functionalities. The fifth item reflects the requirement for better meta-data to enable distributed data resource choices. Research activity in bioinformatics is often faced with large, diverse data sets dispersed across many sources. This information should be integrated and crossqueried to support the search for multifaceted questions and answers. Current systems have only a limited ability to cope with the integration challenges and clearly, more effective methods are needed.

A straw poll around the group revealed that researchers spend 10 to 90% of their time wrangling with data (an informal observation was that the percent of time was highly negatively correlated with seniority). Our group summarized this discussion by acknowledging that one of the most fundamental challenges in front of the bioinformatics community is knowledge and data-type integration.

Advances on this path have been twofold: standards development for bioinformatics data types and development of comprehensive repositories of heterogeneous data. Standards for data include, for example, data formats, data exchange protocols, and meta-data controlled vocabularies. A data repository may be separate databases tied together or a collection of diverse data found in one location such as a data warehouse. The National Science Foundation Office of Cyberinfrastructure Task Force on Data and Visualization (NSF_OCI_TFDV, in press) specifically recognized these issues and recommended that the community "identify and share best practices for the critical areas of data management."

## The Need for Tool Integration

Our breakout group discussion turned to the topic of tools and applications. The heterogeneous nature of life sciences data sets has led to a wide range of bioinformatics tools. However, all too often the software is developed without thought toward future interoperability with other software products. As a result, the bioinformatics software landscape is

[1]Microsoft Research, Redmond, Washington.
[2]University of Washington, Seattle, Washington.
[3]Seattle Children's Research Institute, Seattle, Washington.

currently characterized by fragmentation and silos, in which each research group develops and uses only the tools created within their lab. Today, software availability has become influential in bioinformatics research.

Equally important, although only recently given its due, is tool integration. Tools are often created for specialized purposes, yet it has become clear that they must be able to work in ensemble fashion for comprehensive data exploration tasks. Support of common data standards allows tools to communicate data without human intervention.

### Top-down vs. Bottom-up Integration

The development and enforcement of standards represent a top-down approach to data integration. The group advocated that, simultaneously, the community must consider bottom-up integration, where data in non-compliant formats described by incomplete, incorrect, or altogether missing metadata must be tolerated and even embraced. For example, consider that major search engines do not only index HTML-compliant web pages, but rather any resources amenable to keyword search—text files, spreadsheets, presentations, pdf documents, documents created using word processors, and more. Similarly, integration solutions for the bioinformatics community cannot restrict themselves only to clean, standards-compliant data, or we will overlook the majority of important information in existence. In general, the group found that standards, schemas, controlled vocabularies, ontologies, and other prescriptive structures represent a "shared consensus" about the world that is elusive at the frontier of research by definition—if there were global agreement about how the data should be modeled, described, and interpreted, then it would not be research.

### Analysis Bottlenecks

We switched gears and asked an open question to our breakout group: "what is the limiting resource or bottleneck in your research pipelines today?" The number one constrained resource was not computational, but rather time and funding. The time required going from idea or research question to a result or insight was considerable. Several members noted that, although they are funded to do research, much of their time was consumed with data manipulation and writing new tools to carry out the work. The group also noted the lack of agility. Although they may be aware of a new or better algorithm they cannot easily integrate it into their analysis pipelines given the lack of standards across both data formats and tools. It typically requires a complete rewrite of the code in order to take advantage of a new technique or algorithm, requiring time and often funding to hire developers.

The conversation drifted into other pain points, which can be roughly summarized by noting that the reality in bioinformatics is that humans are often the workflow system: chaining codes together, performing manual joins or interpretations of data (humans as sensors). The entire result of which is very labor intensive.

### Commonly Used Tools

We closed our working group session with a survey around the table on tools commonly used and why they were selected. At the top of the list was the common spreadsheet (i.e., Microsoft Excel, OpenOffice) given its ease of use, the tabular storage suits the needs of the researcher, and the analysis capabilities are sufficient for many common tasks. R was commonly used for manipulation of raw data in files. The users of R noted the ease at which they could perform complex analysis on native files without having to reformat the data. On par with R was the community of relational database users, who utilized MySQL, Postgres, Microsoft Access, and SQL Server. The users of relational databases were often maintaining reference data collections, which warrant the additional effort required to define a schema, clean and upload the data into relational tables, and write their analysis routines in SQL. All members identified tools created in their own labs, typically written in C, Java, and Python. Tools and data collections from NCBI, GenBank, and Ensemble were frequently identified. Overall, the group noted a cultural objection and real cost to switching tools, because the investment in code and data collections present a high "switching cost." Further, we noted a reluctance to try new methods without a clear return on investment. Legacy assets are still assets.

### Current Barriers to Research

Whether operating in the cloud or locally, significant resources are spent joining and aggregating data and switching data from one format to another. Additionally, many simple analyses are not automated because data formats are a moving target. In some cases this is made easier through common schemas and ontologies, but more work is needed to provide high quality, broad reaching ontologies, and tools to reference them from experimental datasets. The community has been slow to share tools, partially because tools are not robust against different input formats. Receiving credit for providing a tool to the community is difficult, and this benefit is often outweighed by the cost of maintaining the tool. As a result, new projects frequently start by creating new tools.

Additionally, many forms of analysis require an understanding of metadata, but, because metadata is inherently difficult to schematized, this translates to human involvement throughout analysis. Finally, it is difficult for biologists to keep abreast of changes in hardware and software/algorithms. As a result, it is very difficult to incorporate cost savings that depend on GPGPU, SSD, etc., or new techniques that leverage tools that are difficult to use or host.

### Current Barriers to Cloud-Based Research

Unless one can truly remove the need for a local cluster, the benefits of cloud computing are limited. To really be valuable, the cloud needs to handle each phase of the analysis pipeline in its entirety. The notable exception to this rule is when the ratio of computation to bandwidth is high enough to justify moving data. This represents most of the existing scenarios we see happening in the cloud today: lots of computation but not as much data. The lack of analysis tools available for the cloud requires researchers to either custom code each analysis activity (too expensive and slow) or constantly upload/download data (too expensive and slow). Although some applications can be hosted on existing cloud infrastructure, exposing them through a simple UI is difficult, and scaling them beyond a single machine is not always feasible. Similarly, common orchestration activities

that are handled with scripting for local deployments become difficult to manage when apps and data are hosted in the cloud.

Finally, building a team with expertise in biology and bioinformatics and cloud computing is perceived as prohibitively difficult. By requiring expertise in cloud computing, the communication overhead is increased substantially. To make cloud computing useful in research, bioinformatics providers need to utilize a suite of tools that make the cloud simple and accessible. Additionally, bioinformatics researchers need to provide libraries of tools that run on the cloud but are easy to use and invoke locally by biologists.

## Future/Outlook

New genomics reality, such as next generation sequencing technologies, is producing data at unimaginable rates and volumes. These new technologies will bring terabyte or even petabyte scales of data within reasonable cost. Yet not only small labs, but even large institutions are having difficulty obtaining and maintaining the computational infrastructure necessary to process, store, and integrate these large-scale data sets.

## Conclusion

With current bandwidth limitations it is not realistic to store data on the cloud but compute on it elsewhere. To make the cloud truly useful, data needs to be schematized or linked to high-quality ontologies, a cohesive set of scalable tools needs to be developed, and user-friendly interfaces need to be easy to develop, allowing biologists to use the cloud transparently in daily work. Many of these barriers also apply to local storage and analysis and must be solved locally before mainstream research can migrate to the cloud.

## Acknowledgments

This policy report and DIS workshop were supported by SCRI and NSF Grant DBI- 0969929 to E. Kolker (Principal investigator). The views expressed in this article are entirely personal opinions of the authors and do not necessarily represent positions of their affiliated institutions or NSF.

## Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

## Reference

National Science Foundation, NSF (2011). Office of Cyberinfrastructure, Task Force on Data and Visualization (in press).

Address correspondence to:
*Eugene Kolker, Ph.D.*
*Seattle Children's Research Institute*
*1900 Ninth Avenue*
*C9S-9*
*Seattle, WA 98101*

*E-mail:* eugene.kolker@seattlechildrens.org

**This article has been cited by:**


1. Pedro Martínez-Gómez , Raquel Sánchez-Pérez , Manuel Rubio . 2012. Clarifying Omics Concepts, Challenges, and Opportunities for Prunus Breeding in the Postgenomic Era. *OMICS: A Journal of Integrative Biology* **16**:5, 268-283. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
2. Eugene Kolker , Elizabeth Stewart , Vural Ozdemir . 2012. Opportunities and Challenges for the Life Sciences Community. *OMICS: A Journal of Integrative Biology* **16**:3, 138-147. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
3. Matthew I. Bellgard, Stanley E. Bellgard. 2011. A Bioinformatics Framework for plant pathologists to deliver global food security outcomes. *Australasian Plant Pathology* . [CrossRef]