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

DOI: 10.1089/omi.2011.0006

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

Arnold Smith,[1,2] Magdalena Balazinska,[2] Chaitan Baru,[3] Mark Gomelsky,[4] Michael McLennan,[5] Lynn Rose,[1] Burton Smith,[6] Elizabeth Stewart,[1] and Eugene Kolker[1,2]

Abstract

The life sciences are poised at the beginning of a paradigm-changing evolution in the way scientific questions are answered. Data-Intensive Science (DIS) promise to provide new ways of approaching scientific challenges and answering questions. This article is a summary of the life sciences issues and challenges as discussed in the DIS workshop in Seattle, September 19–20, 2010.

Introduction

More than 3 years have passed since Jim Gray, pioneering computer scientist, was lost at sea, but his words captured in his January 11, 2007 talk still ring true: "We have to do better at producing tools to support the whole research cycle—from data capture and data curation to data analysis and data visualization." Although this statement is just a very small part of what he said that day, it captures some of the challenges we are still facing with research in the 21st century. We have yet to fully realize the potential of the fourth paradigm of science, data-intensive science, because of complex social and technological challenges that must be met. The goal of the Data-Intensive Science (DIS) Workshop's Part One was to identify the gaps in our abilities and resources that keep us from realizing that potential.

Current State

In the life sciences today, researchers recognize the need and potential benefits of sharing select data sets between scientists, integrating data sets, and performing various types of vertical studies such as how a drug will affect an individual given his/her genome. Although the need is recognized, this type of analysis is difficult to perform today. With the life sciences we have many and varied data sources, which lead to fragmented, redundant, and incomplete data sets that are difficult to coalesce. Even the databases that hold large data sets are often specialized and fragmented, which hinders integration.

Although many researchers may wish to utilize the data-sets of other scientists, not all researchers wish to share their own work. Sharing requires a great deal more effort than not sharing. Considerable effort is spent cleaning and curating the data sets, adding appropriate meta-data to the files, and re-running samples due to ambiguous results. This substantial effort that goes into sharing data is generally insufficiently recognized and rewarded.

Data sets are being underutilized because the right people do not know about them or cannot get to them. For example, http://clinicaltrials.gov/ provides clinical trial details but not the data. Many resources have data that have gone stale or should have been archived to ease data access, but without robust storage systems this does not happen. There are significant technological challenges with data sharing. People wish they could get data easily, rather than shipping disks, but for the most part there is insufficient bandwidth to move large files. Most sharing is ad hoc/word of mouth rather than through an organized clearinghouse.

In addition, a lack of validation/quality control of the data (NCBI takes virtually all submissions of sequence data) as well as a lack of standardization persists. This lack of terminology standardization leads to confusion because the same terms are used with different meanings. In addition, there are too many one-off data analysis tools, with diverse outputs that make data integration and tool integration difficult. When confronted with a new data set, all too often the only recourse is to open yet another Excel workbook for analysis. This approach neither scales well nor lends itself to data sharing.

[1]Seattle Children's Research Institute, Seattle, Washington.
[2]University of Washington, Seattle, Washington.
[3]San Diego Supercomputer Center, UC San Diego, San Diego, California.
[4]University of Wyoming, Laramie, Wyoming.
[5]Purdue University, West Lafayette, Indiana.
[6]Microsoft Research, Redmond, Washington.

Most life science researchers are not adequately cross-trained in the computer sciences or in crossdisciplinary team science approaches to be able to effectively solve these DIS challenges.

Barriers

To our knowledge, the most significant barrier to reaching our fourth paradigm potential is a lack of a community-wide mindset within life sciences. Absence of such mindset results in absence of a strong and smart push for transformative policies and mechanisms that increase the efficacy of biology. Experiments must be done with a clear understanding that the data, without question, will be shared, and therefore certain standards must be followed. Typical promotion/tenure policies and departmental structures are based on PI-centric research paradigms. Incentives are needed because adopting this new mindset, as it can be onerous, until the overall benefit begins to appear. Education at the undergraduate and graduate levels targeted at developing this data-sharing paradigm is lacking. Funding for data integration groups, which must be cross-disciplinary, is insufficient and may be best served by several funding agencies cofunding a project.

Although some journals require data-publishing for federally funded studies, no legal requirement exists for phase I data publication. Additionally, negative results rarely get published, leading to repetition of experiments, wasting time and money. Finally, no economic models exist for data sharing, and the competition inherent in the grant mechanism and profit-motives is at odds with data sharing.

Fear is also associated with data sharing: fear of being scooped, fear of a mistake being found, fear of legal repercussions, and fear of data security (i.e., HIPAA and informed consent regulations place a huge burden of data security on the data producers).

The technological barriers to effective DIS are multifaceted. First, we have a data infrastructure that cannot handle huge data sets for storage, transfer, or analysis. Although it would be optimal in many ways to have data generation colocalized with data storage and analysis, that is not realistic. As we push for a "data for all" reality, we need to be able to give access across the globe, not require the researcher to come to the data. Second, we have data sets that are fragmented and not easily integrated due to lack of standardization and quality control. All too often, a given term's definition is not clear to the researchers in the field, much less researchers in another discipline. As a result, data integration become a data fusion challenge that is part science, part art, and very difficult to automate accurately. Third, the tools for analysis are not standardized and, thus, do not work well together in pipelines. Transforming the output of one tool to a form that can serve as input for the next tool is a task that continues to take inordinate amounts of time (usually graduate students' time), hampering productivity considerably. Fourth, the life sciences have datasets of data types not easily handled by current technologies. For example, *in situ* hybridization images, electrocardiograms, and other highly visual data types are difficult to crossreference and search in standard systems. As these issues are not unique to the life sciences, the lessons learned in this field could be applied elsewhere.

Future/Outlook

We cannot afford to continue on the same trajectory. Too much time, effort, and money is being wasted on inefficient data generation, management, and analysis. Resources are being wasted at a time when funding is not keeping pace with our need to compete in an increasingly competitive global science arena. Our best approach is more effective utilization of current resources.

Change begins with education. We need community-wide education on data sharing and analysis. We need to transition from the mindset that dominates the life sciences, focusing on a single process or organism, toward that which can be seen in the field of physics. Integrated training in biology and computer science is needed to broaden knowledge of experts in both disciplines. Mandatory computer science/data management classes for undergraduate and graduate biologists and natural science courses for computer science students would encourage collaboration. Because no one person will be able to learn everything, there must be education on team interactions and team management. Although science used to be seen as a solitary pursuit, it is transforming into a team effort. Much of the training for an MBA is through team projects, in which the student learns how to get the best from his/her teammates. This mindset needs to be added to research training.

Education must also occur in other venues. Funding agencies and career advancement panels must see metrics beyond citations, such as analytics, usage stats of applications and the number of useful datasets generated and shared. We could have a Lifesciences Appstore, and make it worthwhile to produce tools and datasets for others to use. We could pose DIS Challenge questions to allow researchers to win awards for solving challenge problems that carry enough prestige to influence career advancement.

Funding agencies need to recognize and reward the impact of multifaceted data-sharing projects. Similar problems emerged at the start of the Genome Project as individual scientists wondered how working with a team of 100 scientists would possibly help their own career. As was done with the Genome Project, there could be numerous funded small data integration/data management pilot projects. The best would rise to the top and continue their development. Merit due to effectiveness needs to be awarded. Many efforts have been made to manage and integrate large data sets in the life sciences, efforts such as Entrez (http://www.ncbi.nlm.nih.gov/sites/gquery) and Ensembl (http://www.ensembl.org/index.html). The Genome Project provided impetus for large data set management tools, but data generation has gone far beyond the Genome Project and much more is needed (Fig. 1).

National Science Foundation (NSF) is one funding agency that has made major steps to support data management. NSF grant applications require a "data management" section in the proposal because data management can no longer be left to chance. As was noted by Cora Marrett, NSF Acting Deputy Director: "This is the first step in what will be a more comprehensive approach to data policy. It will address the need for data from publicly-funded research to be made public." This requirement will, hopefully, spur thoughtful design, reflecting back on the overall proposal for more efficient data generation, analysis, and propagation.

Quality control, validation, and tests of data are very much needed. Similarly, data producers must know how to best

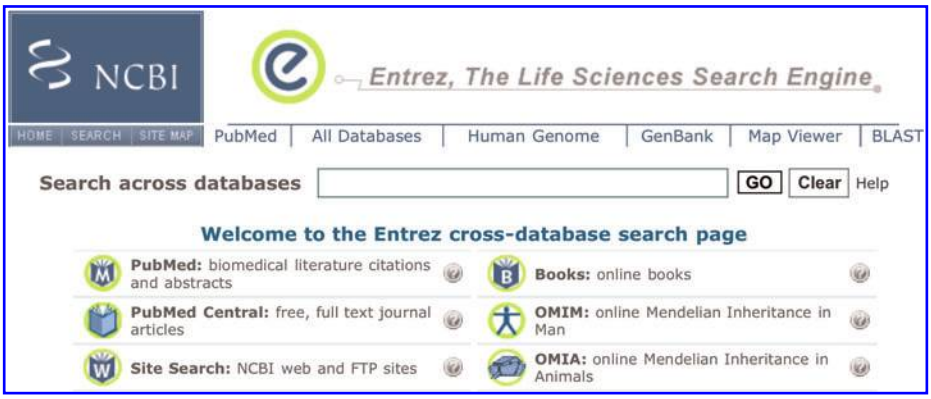

FIG. 1. The top of the Entrez home page. Entrez integrates and enables searching of over 30 NCBI databases.

utilize and navigate through the copious data sets. A review process akin to Amazon's 5-star system along with the ability to select only data sets that meet the given standards have to become normal practice. Standard, customizable tools are needed. Therefore, a number of tools should be developed and lessons learned from genome sequence analysis tool evolution (BLAST won by popular usage) should be applied. We need to let data filter out. If it is not accessed, perhaps it does not need to be easily accessed and could be archived. Indexing, as implemented by Google, can help the "best" data come back most quickly. We need semantic data as well as better data management, versioning, and archiving. As a result, we will be able to get the needed data. The NSF Office of Cyberinfrastructure Task Force on Data and Visualization recommends that each key research domain triage its own essential data, and then have an open call with storage providers to allow competition to provide favorable price-points (NSF_OCI_TFDV, in press).

We need database integration to allow crossreferencing of different layers of data. For example, the challenge of referencing from the genome to the transcriptome to the proteome to the metabolome remains unsolved despite numerous attempts. Cloud computing can become part of our future but will need increased security and more seamless interaction. Although highly specialized data, such as clinical data, have exceptional security needs, hackers must be stopped for all data types. Different approaches need to be examined to determine the best use of technologies such as clouds. Will the government host all federally funded data sets? Will smaller entities, such as universities, host data sets for all to access? With wireless sensor data, Dartmouth provides a good example. When interesting publications came by, they asked for, hosted, and made publicly available the data sets (this effort is still ongoing at http://crawdad.cs.dartmouth.edu/index.php). This successful grassroots approach can be extended and improved.

Representation and Engagement with Society

We need a representative society or alliance but, if a good fit already exists, we should find and collaborate with existing groups. The second workshop should invite representatives from existing societies (XLDB (Extremely Large Databases), HUPO (Human Proteome Organization, etc.). We also need representatives from e-science groups.

Our representation is also needed to influence the policies of the future. We want to influence the scientific, methodological, technical, and funding solutions. We would encourage joint funding by different NSF Directorates and joint NSF and by National Institutes of Health (NIH) to promote collaborative projects between biologists with IT challenges and computer scientists with resources looking for problems. Funding should not be restricted by agency boundaries. The OSTP/NITRD (Office of Science and Technology Policy/Networking and Information Technology Research and Development Program) could be a useful resource. The Department of Defense (DOD), Department of Energy (DOE), Centers for Disease Control and Prevention (CDC), and Food and Drug Administration (FDA), under similar data-intensive pressures, can also be involved. The policies of the future will need to include funding for managing on-going hosted datasets. This suggestion, echoed by the NSF_OCI_TFDV, may require a new funding model to manage the data-intensive challenges (NSF_OCI_TFDV, in press).

We recommend reexamining the basic tenets behind the Bayh-Dole act. Should data be protected for and owned by the public or protected for and owned by entities to allow for privatization? Adopted in 1980, the act permits a university, small business, or nonprofit institution to elect to pursue ownership of an invention in preference to the government. This may be at odds with the goal of data-sharing, as public disclosure can negate an entity's competitive edge.

Representation in society is particularly important with regard to society's view of our work. Scientists, as a whole, face continuing challenges to make science accessible, understandable, and nonthreatening. This is not a new issue, and is not particular to DIS. However, certain efforts within the life sciences that involve laypersons (i.e., volunteer clinical trials) can contribute to data intensive science. We need to ensure that the layperson feels safe and worthwhile to contribute to data-gathering efforts. Many volunteers want to track their data and receive results from studies. We need to enable that involvement without compromising privacy. Citizen science may be an important value for societal involvement and social network venues.

Facebook, Wikipedia, and podcasts are social media tools that can help to put a face on science. Facebook has a "Layperson Science" page (http://www.facebook.com/pages/Layperson-Science/375516482719), but there is currently no content. Wikipedia, perhaps one of the most astounding examples of societal involvement, is heavily used by the public.

However, a search of Wikipedia does not bring up any direct page to Data-Intensive Sciences or to Fourth Paradigm.

Conclusion

A cultural shift is needed for data-intensive life sciences to flourish. This shift will require an integration of top-down approaches promoting data management and grass roots initiatives. No single entity, whether from government, private industry or academia, can make this happen. No one discipline can foresee all necessary changes. Untapped knowledge is waiting to be mined from extant data sets with more data being generated every day. We cannot wait for the perfect plan, the all-encompassing blueprint to the future of data-intensive science. We must take thoughtful steps now so the journey can commence.

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