



National Institute of Allergy and Infectious Diseases
www.niaid.nih.gov

NIAID/DMID BIOINFORMATICS FOR INFECTIOUS DISEASES— TRANSLATING DATA TO KNOWLEDGE

APRIL 26, 2012 | WASHINGTON, DC



Report of the 2012 NIAID/DMID Bioinformatics for Infectious Diseases Workshop

Translating Data to Knowledge

Division of Microbiology and Infectious Diseases, Office of Genomics and Advanced Technologies

National Institute of Allergy and Infectious Diseases

National Institutes of Health

Table of Contents

Introduction.....	3
Background.....	3
The Workshop.....	3
Guiding Principles and Recommendations	4
Appendix 1. Participation List	7
Appendix 2. Agenda	11

Introduction

In April 2012 the Division of Microbiology and Infectious Diseases (DMID), National Institute of Allergy and Infectious Diseases (NIAID), NIH convened a workshop on Bioinformatics for Infectious Diseases. The goals included discussing current NIAID/DMID bioinformatics activities and providing guidance on future undertakings to better assist the infectious diseases research community with the management, accessibility, analysis, and interpretation of research data sets that are increasingly large and complex.

The discussions and important themes that emerged from the workshop are summarized below. The workshop recommendations will be critical to assist NIAID/DMID in charting and planning future bioinformatics activities and resources for the infectious diseases scientific community, particularly in light of the unprecedented volume of data requiring increasing bioinformatics capacity to more easily translate data and information into knowledge.

Background

Over the past 10 years DMID has made a significant investment in supporting bioinformatics activities that provide resources to the infectious disease research community. These resources, many of which are freely available, include databases, open source analysis tools, and bioinformatics services.

NIAID/DMID-supported programs established by the Office of Genomics and Advanced Technologies (OGAT) that focus on bioinformatics or have significant bioinformatics components include the [Genomic Sequencing Centers for Infectious Diseases](#) (GSCIDs); [Bioinformatics Resource Centers](#) (BRCs); [Structural Genomics Centers for Infectious Diseases](#); [Clinical Proteomics Centers for Infectious Diseases and Biodefense](#), and the program for [Systems Biology for Infectious Disease Research](#). General information on these programs can be found on the [NIAID website](#). NIAID/DMID has also been encouraging the sharing and public dissemination of research datasets by establishing and administering [Data Sharing and Release Guidelines](#) for a number of DMID-funded programs that generate data and metadata that are expected to be valuable research resources for the broad scientific community.

The Workshop

Approximately 25 workshop participants including bioinformaticians and infectious disease researchers ensured that both research fields and their needs and priorities were represented. Dr. Rick Stevens from the Department of Computer Science at the University of Chicago and Dr. Adolfo Garcia-Sastre from the Department of Microbiology at the Mount Sinai School of Medicine served as the workshop co-chairs. Participants assembled for a one-day meeting on April 26, 2012. (See Appendix 1 for the workshop agenda and Appendix 2 for the list of participants). Presentations by Dr. Ralph Baric from the University of North Carolina and Dr. Ewan Birney from the European Bioinformatics Institute [EBI] provided examples of how the integration of computational techniques and bioinformatics approaches is generating remarkable progress for infectious disease research. The speakers also highlighted opportunities and challenges for these activities in the upcoming years.

The diversity, value, accessibility, and overall services offered by the bioinformatics components of the OGAT programs were discussed by workshop participants organized into breakout groups. In addition they were asked to identify what worked well and what could be improved and changed. The groups provided recommendations for programs and other activities to be implemented in the short (1-3 years), medium (4-6 years) and long term (more than 6 years) focusing on the challenges and growing opportunities for translating data into knowledge. These include data management and integration, infrastructure and emerging information technologies, and data analysis and modeling approaches.

In reviewing the last few years of DMID OGAT's bioinformatics activities, the workshop participants acknowledged the achievements and efforts dedicated to supporting infectious disease research. In particular, participants recognized the progress in making data and other resources accessible to the scientific community. This includes making microbial genome analysis services and data visualization interfaces freely available and easily accessible, and enabling systematic sharing of data and reagents generated by the programs. Participants also noted that the collaborative efforts between OGAT and the scientific community to establish metadata standards for clinical samples are to be commended as a model of collaboration. For example, with support from OGAT, investigators from the (BRCs) and the (GSCIDs) have collaborated to develop processes for standardized capture of metadata and clinical data from microbial sequencing projects. These data are valuable for supporting epidemiologic and genotype-phenotype association studies. The BRCs have also collaborated with the Systems Biology Centers to establish standards for collecting omics data and to provide such data in a consistent and integrated fashion for unrestricted access by the infectious disease research community.

Guiding Principles and Recommendations

Workshop discussions identified the following as priorities for bioinformatics activities in order to maximize the extraction of knowledge from data over the next decade and advance infectious disease research.

1. Further expand the establishment and adoption of metadata standards

With the considerable amount of data generated by high-throughput omics technologies from basic, translational, and clinical studies and the increasing dissemination of such data to the scientific community, it is critical to systematically collect the associated metadata and clinical data and to take into account data quality for long "shelf life." The availability of experimental metadata in standardized set and format is essential to increase data integration and uptake by the scientific community.

- Develop metadata standards for a wide spectrum of scientific investigations. Examples include metadata about experimental design, geographic location of sample collection, environmental information about sample collection, and phenotype.
- Develop analysis tools and graphical user interfaces that utilize standardized metadata.
- Ensure data quality, extensive data and metadata curation, and long-term access to sustainable databases for key reference datasets (e.g. genome sequences of reference strains, and/or of samples that are unique, expensive, or hard to duplicate or regenerate).

2. Establish an infrastructure and bioinformatics support for infectious disease clinical research data and metadata

It is expected that an increasing number of clinical isolates will become available for sequencing, omics profiling and other types of experimental analyses. Issues related to privacy and risk of re-identification of human subjects from which the clinical isolates are collected need to be addressed while still allowing access to the phenotypic information associated to the sample.

- Establish infrastructure and utilize standards for clinical data collection/management and expand standards when needed.
- Provide computational tools and query interfaces to facilitate comparison of results from clinical studies in multi-center collaborations by both in- and out-of-network investigators.
- Standardize consent forms of DMID-funded clinical research to facilitate omics research.

- Facilitate clinical data and metadata access for computational biology and epidemiology research.

3. *Expand analysis platforms and computational resources*

DMID has made a considerable investment in providing investigators with data repositories and analysis resources, especially for genome sequence data. However, more efforts are needed to facilitate interactive access, processing, visualization, modeling, and synthesis of the large amount of omics and other research data expected in the next few years.

- Improve computational approaches to functional annotation of microbial genomes.
- Empower the widespread distribution of omics technologies through the development of analysis tools, standards, data management, and computational pipelines.
- Continue to investigate and, as appropriate, encourage adoption of new solutions for network and computational architecture infrastructures (e.g., cloud).
- Consider centralization of research data storage to facilitate accessibility and integration of multiple computational analysis pipelines.
- Support collaborative computational environments (i.e. workbenches) for scientists to deposit, integrate, and analyze data in a controlled access setting prior to public release.
- Facilitate the establishment of computational projects ('*spokes*') to develop analysis and visualization tools to expand and improve the bioinformatics offerings of the large Centers ('*hubs*').
- Ensure that budgets of funded research projects address not only data generation costs, but also storage and network costs for bioinformatics activities, including data analysis and public data dissemination are supported in research projects.

4. *Community engagement & training*

Successful bioinformatics resources are those that actively engage the scientific community they are serving. Some infectious diseases researchers are not yet familiar enough with bioinformatics approaches and analysis tools and do not have the capacity to use multiple omics technologies to process and analyze complex data sets, such as those derived from systems biology studies.

- Increase the number of researchers with expertise in computation who can provide bioinformatics support and services to outside laboratories with bioinformatics/analysis needs.
- Create opportunities for inter-disciplinary teams for hypothesis- and mission-driven research to better identify the bioinformatics needs of the community.
- Increase bioinformatics expertise in the developing world.
- Establish opportunities for cross-disciplinary training, such as between computational biologists and bench scientists.
- Facilitate the expansion of microbiology and infectious diseases academic curricula to include more mathematics, computer science, and programming skills.

5. *Establish bioinformatics resources for host/pathogen interaction data and analysis*

Workshop participants noted that while considerable bioinformatics resources have been established to support pathogen research data and information including generating different omics data sets as transcriptomics and proteomics, there has been limited conceptual and technical integration with host-related datasets. For example, integration is needed with databases that support studies of pathogenesis and host transmission, as well as development of therapeutics, diagnostics, and vaccine development.

- Support integration of omics data between host, pathogen, and vector interaction studies.
- Support collection of host and pathogen genetics data and the associated disease phenotype information, including host polymorphisms and susceptibility to disease, as well as epigenetic variability and disease processes.

- Improve integration of omics and epidemiologic data, as well as information derived from clinical research, including the analysis of electronic health records.
- Develop tools to monitor temporal and spatial shifts in host and pathogen populations in response to disease and to predict epidemics.

Appendix 1. Participation List

Adam Arkin
University of California Berkeley
Berkeley, CA
aparkin@lbl.gov

Ralph Baric
Systems Biology Program
University of North Carolina
Chapel Hill, NC
rbaric@email.unc.edu

Clifton Barry
Tuberculosis Research Section
NIAID/NIH
Bethesda, MD
cbarry@niaid.nih.gov

Lara Bethke
Wellcome Trust
London, United Kingdom
l.bethke@wellcome.ac.uk

Ewan Birney
European Bioinformatics Institute
Cambridge, United Kingdom
birney@ebi.ac.uk

Bruce Birren
Genome Sequencing Center
Broad Institute
Cambridge, MA
bwb@broadinstitute.org

Frank Collins
VectorBase Bioinformatics Resource Center
University of Notre Dame
Notre Dame, IN
frank@nd.edu

Greg Deye
Office of the Director, DMID
NIAID/NIH
Bethesda, MD
deyega@niaid.nih.gov

Valentina Di Francesco
Office of Genomics and Advanced Technology
NIAID/NIH
Bethesda, MD
vdifrancesco@niaid.nih.gov

Dennis Dixon
Bacteriology and Mycology Branch
NIAID/NIH
Bethesda, MD
dmdixon@niaid.nih.gov

Peter Dudley
Office of Genomics and Advanced Technology
NIAID/NIH
Bethesda, MD
dudleype@niaid.nih.gov

Patrick Duffy
Laboratory of Malaria Immunology and
Vaccinology
NIAID/NIH
Bethesda, MD
patrick.duffy@nih.gov

Najib El-Sayed
University of Maryland
College Park, MD
elsayed@umd.edu

Claire M. Fraser
Genome Sequencing Center
University of Maryland
Baltimore, MD
cmfraser@som.umaryland.edu

Adolfo Garcia-Sastre
Centers of Excellence for Influenza Research
and Surveillance
Mt. Sinai Medical School
New York, NY
adolfo.garcia-sastre@exchange.mssm.edu

William Gelbart
VectorBase Bioinformatics Resource Center
Harvard University
Cambridge, MA
gelbart@morgan.harvard.edu

Mark Gerstein
Yale University
New Haven, CT
mark.gerstein@yale.edu

Maria Giovanni
Office of Genomics and Advanced Technology
NIAID/NIH
Bethesda, MD
mgiovanni@niaid.nih.gov

Irene Glowinski
Office of the Director, DMID
NIAID/NIH
Bethesda, MD
iglowinski@niaid.nih.gov

Peter Good
NHGRI/NIH
Bethesda, MD
goodp@mail.nih.gov

Chuck Hackett
Office of the Director, DAIT
NIAID/NIH
Bethesda, MD
chackett@niaid.nih.gov

Lee Hall
Parasitology and International Programs Branch
NIAID/NIH
Bethesda, MD
llhall@niaid.nih.gov

Carole Heilman
Office of the Director, DMID
NIAID/NIH
Bethesda, MD
cheilman@niaid.nih.gov

Eddie Holmes
Pennsylvania State University
State College, PA
ech15@psu.edu

Lei Huang
Office of Clinical Research Affairs
NIAID/NIH
Bethesda, MD
huangle@niaid.nih.gov

Yentram Huyen
Office of Cyber Infrastructure and
Computational Biology
NIAID/NIH
Bethesda, MD
huyeny@niaid.nih.gov

Deirdre Joy
Parasitology and International Programs Branch
NIAID/NIH
Bethesda, MD
djoy@niaid.nih.gov

Tom Kepler
Boston University School of Medicine
Boston, MA
tbkepler@bu.edu

Jessie Kissinger
EuPathDB Bioinformatics Resource Center
University of Georgia
Athens, GA
jkissing@uga.edu

David Knipe
Harvard University
Cambridge, MA
david_knipe@hms.harvard.edu

Michael Kurilla
Office of Biodefense Research Affairs
NIAID/NIH
Bethesda, MD
mkurilla@niaid.nih.gov

Linda Lambert
Respiratory Diseases Branch
NIAID/NIH
Bethesda, MD
llambert@niaid.nih.gov

Daniel Lawson
VectorBase Bioinformatics Resource Center
European Bioinformatics Institute
Cambridge, United Kingdom
lawson@ebi.ac.uk

Eun Mi Lee
Office of Genomics and Advanced Technology
NIAID/NIH
Bethesda, MD
eunmi.lee@nih.gov

David Lipman
NCBI/NLM/NIH
Bethesda, MD
dlipman@ncbi.nlm.nih.gov

Punam Mathur
Office of Genomics and Advanced Technology
NIAID/NIH
Bethesda, MD
mathurpu@niaid.nih.gov

Victoria McGovern
Burroughs Wellcome Fund
Durham, NC
vmcgovern@bwfund.org

John J. McGowan
Office of the Director, NIAID
NIAID/NIH
Bethesda, MD
jmcgowan@niaid.nih.gov

Barbara Mulach
Office of the Director, DMID
NIAID/NIH
Bethesda, MD
bmulach@niaid.nih.gov

Karen Nelson
Genome Sequencing Center
J. Craig Venter Institute
Rockville, MD
knelson@jcv.org

William Nierman
Genome Sequencing Center
J. Craig Venter Institute
Rockville, MD
wnierman@jcv.org

Francis Ouellette
Ontario Institute for Cancer Research
Toronto, Ontario
francis@oicr.on.ca

Malu Polanski
Office of Genomics and Advanced Technology
NIAID/NIH
Bethesda, MD
polanskim@niaid.nih.gov

Mihai Pop
University of Maryland
College Park, MD
mpop@umiacs.umd.edu

Diane Post
Respiratory Diseases Branch
NIAID/NIH
Bethesda, MD
postd@niaid.nih.gov

Julia Puzak
Office of Genomics and Advanced Technology
NIAID/NIH
Bethesda, MD
puzakjp@niaid.nih.gov

Raul Rabadan
Columbia University
New York, NY
rabadan@dbmi.columbia.edu

David Roos
EuPathDB Bioinformatics Resource Center
University of Pennsylvania
Philadelphia, PA
droos@sas.upenn.edu

Richard Scheuermann
IRD and ViPR Bioinformatics Resource Centers
University of Texas Southwestern
Dallas, TX
richard.scheuermann@utsouthwestern.edu

Clare Schmitt
Office of BioDefense Research Affairs
NIAID/NIH
Bethesda, MD
cschmitt@niaid.nih.gov

Gary Schoolnik
Systems Biology Program
Stanford University
Stanford, CA
gary.schoolnik@stanford.edu

Bruno Sobral
PATRIC Bioinformatics Resource Center
Virginia Bioinformatics Institute
Blacksburg, VA
sobral@vt.edu

Rick Stevens
PATRIC Bioinformatics Resource Center
Argonne National Laboratory
Argonne, IL
stevens@anl.gov

Granger Sutton
Genome Sequencing Center
J. Craig Venter Institute
Rockville, MD
gsutton@jcv.org

Mike Tartakovsky
Office of Cyber Infrastructure and
Computational Biology
NIAID/NIH
Bethesda, MD
mtartakovs@niaid.nih.gov

Diane Wax
Office of the Director, NIAID
NIAID/NIH
Bethesda, MD
waxd@niaid.nih.gov

Owen White
University of Maryland
Baltimore, MD
owhite@som.umaryland.edu

Jennifer Wortman
Genome Sequencing Center
Broad Institute
Cambridge, MA
jwortman@broadinstitute.org

Alison Yao
Office of Genomics and Advanced Technology
NIAID/NIH
Bethesda, MD
yaoal@niaid.nih.gov

Appendix 2. Agenda

8:00-8:30am	Registration
8:30-8:40am	Welcome Carole Heilman Director, Division of Microbiology and Infectious Diseases, NIAID, NIH
8:40-9:00am	Welcome and Expected Goals Co-Chairs Rick Stevens, Argonne National Laboratory Adolfo Garcia-Sastre, Mt. Sinai Medical School
9:00-9:30am	DMID Bioinformatics Activities Overview Valentina Di Francesco, DMID/NIAID Alison Yao, DMID/NIAID
9:30– 10:30am	Keynote Presentations
9:30-10:00am	Bioinformatics for Infectious Diseases: Advancing the Infectious Disease Research Agenda Claire M. Fraser, University of Maryland
10:00-10:30am	High Throughput Biology and Computational Techniques for Infectious Biology Ewan Birney, European Bioinformatics Institute
10:30-11:00am	BREAK
11:00-12:00pm	Breakout Group Session AM - Current NIAID/DMID Bioinformatics Activities for Infectious Diseases Breakout #1 (Room 2C13, 10401 Fernwood) Chairperson: Patrick Duffy, NIAID/NIH NIAID/DMID Person: Michael Kurilla Breakout #2 (Room 1202, 6700B Rockledge)

Chairperson: Owen White, University of Maryland

NIAID/DMID Person: Lee Hall

12:00 – 1:00pm

LUNCH

1:00-1:30pm

Current Program-Breakout Group Report

Group reports – presented by Chairpersons of each breakout group

1:30-3:30pm

Breakout Group Session PM - Future NIAID/DMID Bioinformatics Activities for Infectious Diseases

Breakout #1 (Room 2C13, 10401 Fernwood)

Chairperson: Francis Ouellette, Ontario Institute for Cancer Research

NIAID/DMID Person: Barbara Mulach

Breakout #2 (Room 1202, 6700B Rockledge)

Chairperson: William Gelbart, Harvard University

NIAID/DMID Person: Irene Glowinski

3:30-4:00pm

BREAK

4:00-4:30pm

Future Program - Breakout Group Reports

Group reports – presented by Chairperson of each breakout group

4:30-5:00pm

Summary

Rick Stevens, Argonne National Laboratory

Adolfo Garcia-Sastre, Mt. Sinai Medical School

5:00pm

Adjournment