BIOINFORMATICS & BIODEFENSE
KEYS TO UNDERSTANDING NATURAL & ALTERED PATHOGENS

This report is based on a colloquium, sponsored by the American Academy of Microbiology, convened November 2008 in Baltimore, Maryland.

The American Academy of Microbiology is the honorific leadership group of the American Society for Microbiology. The mission of the American Academy of Microbiology is to recognize scientific excellence and foster knowledge and understanding in the microbiological sciences. The Academy strives to include underrepresented scientists in all its activities.

The American Academy of Microbiology is grateful for the generosity of the National Interagency Genome Sciences Coordinating Committee for support of this project.

The opinions expressed in this report are those solely of the colloquium participants and do not necessarily reflect the official positions of our sponsors or the American Society for Microbiology.
A REPORT FROM THE AMERICAN ACADEMY OF MICROBIOLOGY

BIOINFORMATICS & BIODEFENSE
KEYS TO UNDERSTANDING NATURAL & ALTERED PATHOGENS

BOARD OF GOVERNORS, AMERICAN ACADEMY OF MICROBIOLOGY

R. John Collier, Chair
Harvard Medical School

Kenneth I. Berns
University of Florida Genetics Institute

E. Peter Greenberg
University of Washington

Carol A. Gross
University of California, San Francisco

Lonnie O. Ingram
University of Florida

J. Michael Miller
Centers for Disease Control and Prevention

Stephen A. Morse
Centers for Disease Control and Prevention

Edward G. Ruby
University of Wisconsin, Madison

Patricia Spear
Northwestern University

COLLOQUIUM PARTICIPANTS

Michael Ashburner
European Bioinformatics Institute, Cambridge, England

Thomas Brettin
Los Alamos National Laboratory

Charles R. Cantor
Sequenom, San Diego, California

Thomas Cebula
Food and Drug Administration

Rita R. Colwell
University of Maryland

Valentina Di Francesco
National Institute of General Medical Sciences, National Institutes of Health

Dawn Field
Centre for Ecology and Hydrology, Molecular Evolution, and Bioinformatics, Oxford, England

Yuriy Fofanov
University of Houston

James E. Galagan
Broad Institute, Massachusetts Institute of Technology

Robert Jones
Craic Computing, LLC, Seattle, Washington

Cheryl Kraft
National Institute of Allergy and Infectious Diseases, National Institutes of Health

Eliot Lefkowitz
University of Alabama, Birmingham

Baochuan Lin
Naval Research Laboratory
COLLOQUIUM PARTICIPANTS
(CONTINUED)

Anthony Malanoski
Naval Research Laboratory

Barbara Mann
University of Virginia

Michael Pear
IBIS Biosciences, Inc., Carlsbad, California

Wilfred Pinfold
Intel Corporation, Hillsboro, Oregon

Tim Read
Naval Research Laboratory

Margaret Riley
University of Massachusetts, Amherst

Daniel L. Rock
University of Illinois

Forest Rohwer
San Diego State University

Bruno Sobal
Virginia Bioinformatics Institute, Virginia Tech

George F. Sprague
Institute of Molecular Biology, University of Oregon

Granger Sutton
J. Craig Venter Institute, Rockville, Maryland

Doreen Ware
Cold Spring Harbor Laboratory, Cold
Spring Harbor, New York

George Weinstock
Washington University at St. Louis

Owen R. White
University of Maryland School of Medicine

Jagjit S. Yadav
University of Cincinnati College of Medicine

PARTICIPANTS: NATIONAL INTERAGENCY GENOME SCIENCES COORDINATING COMMITTEE

Juan Arroyo
Mitre Corporation

Kim Bishop-Lily
Naval Research Laboratory

James Burrans
Battelle National Bioforensics Institute

Daniel Drell
U.S. Department of Energy

Maria Giovanni
NIAID, NIH

Elaine Mullen
Mitre Corporation

Evan Skowronski
Edgewood Chemical Biological Center

Ronald Walters
Intelligence Community

Mark Wolcott
USAMRIID
EXECUTIVE SUMMARY

Biodefense research is a high national priority, not only for the sake of improving security in the face of threats from infectious disease, but also for the insights that biodefense and its related bioinformatics offer to the basic science of understanding pathogens. A colloquium was convened to bring together experts in biodefense, bioinformatics, molecular biology, microbiology, information technology, and other fields to discuss ways in which bioinformatics practice and support can be changed to improve the speed of discovery for the benefit of national security and of scientific progress. Participants made specific recommendations for improving Data Description and Repositories, Algorithm and Software Resources, and Infrastructure Resources.

Many features of databases, including their funding, consistency, and usability, pose problems for bioinformaticists. It is critical that bioinformatics database resources remain in the government-funded domain. Moreover, a plan for data sharing should be defined for any new bioinformatics database resources that are developed, and researchers who make use of such a database should be held to a contract of collaboration with the database.

Efforts to develop improved ontologies (constructs for annotating data in a semantic and computationally usable form) to meet the special needs of biodefense should be nurtured and intensified to fill gaps in the breadth of pathogen-specific ontologies. While additional work is needed on ontologies in many aspects of biological science, ontologies relevant to biodefense are needed to accelerate the construction of integrated databases relevant to both biodefense and human health programs.

Researchers working in biodefense and bioinformatics need more and better software tools for making phylogenetic assignments, strain typing, and for detecting engineered organisms. In the future, as sequencing technology improves and the rate by which data are produced continues to climb, all of our current big, centralized, monolithic systems for handling and storing data will, inevitably, break. We should implement scalable new approaches and move to more distributed systems now.

There is a deficit in statistical training for bioinformaticists that represents a significant, discreditable weakness in the field. Educators must begin to provide appropriate statistical training for all experimentalists.

The submission tools of the National Center for Biotechnology Information (NCBI) are not adequate for the size and complexity of the data sets users want to submit, making the submission of sequence data to NCBI’s GenBank a bottleneck in the process of publishing bioinformatics results. Although it is clear that sequence management is in need of a new paradigm, the characteristics of a new paradigm are not apparent. There may be a role for Google and its various tools and resources to enable information to be shared in a more open and scalable manner. Scaling NCBI to meet the needs of biological data for the next decade should be a high priority.
INTRODUCTION

In the face of threats from bioterrorism and emerging infectious diseases confronted by the United States and other nations, supporting a rigorous research agenda in biodefense is imperative. For the purpose of this report, biodefense is defined to mean research and operations designed to enhance national security by supporting rapid detection and subsequent determination of whether pathogen incidents affecting humans or plants/animals of economic importance are an act of nature, or an act of biocrime or bioterrorism. Clearly, there is a large and imprecisely-defined overlap between biodefense and infectious disease research for human, animals, and plants. Additionally, whether some particular project is considered basic research or biodefense may be in the eye of the beholder. Bioinformatics, the application of computer analysis to molecular biology, is a fundamental corollary to biodefense research.

Advancements in biodefense research and bioinformatics are meant to lead to improvements in national security, but the impacts of these scientific discoveries will also inevitably echo across almost all disciplines in biology. In short, the pursuit of research in biodefense and bioinformatics can be expected to increase our understanding of life. These dual positive impacts in security and science—today’s need and tomorrow’s ambition—thrust biodefense and bioinformatics toward the top of the long list of national priorities.

Specific fields of scientific inquiry that benefit from biodefense and bioinformatics research include, but are not limited to:

- **Genetic Diversity.**
  In research ostensibly directed toward biodefense, many microbial genomes are being sequenced to identify the presence or absence of genes, genes under unusual selective pressure, and global genetic rearrangements. This effort is expected to lead not only to the ability to predict pathogenesis from genomic sequence and to improve our ability to assess the threat level of a given organism, but also to insights into population genetics. For example, sequencing the genomes of various species of *Yersinia pestis*, the bacterium responsible for plague, which is a biosecurity threat, will enable population geneticists to identify the factors associated with the host-specificity and explain the underlying mechanics of how different strains infect different hosts and how those strains are interrelated and distributed geographically.

- **Epidemiology.**
  Determining the presence or absence of biothreat organisms among the population, the environment, and the factors that influence the abundance of infectious agents in the environment are endeavors that have direct relevance to epidemiology. Understanding the rapid evolution of many pathogens, even during the course of a single outbreak, is also highly relevant.
**Vaccinology.**
High throughput sequencing has been applied to speed identification of vaccine targets for biodefense. Signatures discovered in this pursuit could also be used to identify organisms with resistance to vaccines. In addition, Next Generation Sequencing Technologies (NGST) are being applied to examine the portions of the human genome which encode t-cell and b-cell receptors, allowing us to monitor the immune systems within the host and to track changes over time for biosecurity monitoring. This also allows researchers to make predictions of susceptibility and to aid in rational design of vaccines (e.g., influenza vaccine).

**Global Health.**
NGST is expected to help with large-scale monitoring of species in the environment for the sake of biodefense. This will also aid in revealing the “big picture” of bacterial populations in the environment, which is preferable to the current approach in which single strains are tracked, a technique that provides an incomplete picture of the species that interact with humans. Global microbe tracking could be tailored for specific organisms (e.g., flu) or designed to examine the complete metagenomic profile of a geographical region. This would permit (a) measurement of the “background” population of species in the environment and monitoring outbreaks of virulent and antibiotic resistant organisms, and (b) development of epidemiological models that predict outbreaks and identify the environmental factors that are relevant to the incidence of disease by tracking metagenomic profiles in the context of environmental factors, such as air and water movement, temperature, and rainfall.

**Metabolic Reconstruction.**
Improvements in bioinformatics methods will probably lead to an improved understanding of the metabolic capability of the cell. By coupling this understanding with advancements in metagenomic sequencing methods, scientists expect to be able to evaluate and model the metabolic capabilities of collections of bacteria.

**Systems Biology.**
Post-sequencing bioinformatics techniques, such as *in silico* pathway reconstruction, proteomics, and expression analysis, are expected to advance our understanding of regulatory networks, gene-to-gene interactions, and signaling pathways. This will enable improved prediction of metabolite use in the cell and help identify other factors involved in regulation. It is also anticipated that this work, when used in combination with knowledge of human systems, will lead to models for host-pathogen interactions and ultimately development of intervention strategies to reduce our susceptibility to pathogens.

**Personalized Medicine.**
High-throughput analysis of infectious organisms of concern to biodefense, in combination with genotyping of individuals to identify human factors associated with susceptibility to infection, will lead to insights into an individual's susceptibility to disease, as well as intervention strategies specific to each individual. This research area is expected to have broad impact in the health field, but it is also anticipated to be relevant to generating intervention strategies in response to biothreat agents.

This report summarizes the proceedings of a colloquium convened to make recommendations for ways in which bioinformatics and biodefense practice and support can be changed to improve the speed of discovery for the benefit of national security.
and of scientific progress. Colloquium participants discussed the goals of biodefense research and barriers to moving forward; they made specific recommendations for overcoming current and foreseen difficulties with data description and repositories, algorithm and software resources, and infrastructure resources.

**DATA DESCRIPTION AND REPOSITORIES**

**NIAID’s Bioinformatics Resource Centers: One-Stop Shopping.** In 2004 the National Institute of Allergy and Infectious Diseases (NIAID) funded eight Bioinformatics Resource Centers (BRCs) to establish databases of genomics and associated data of pathogens of biodefense interest. The BRCs are required to create user friendly web and software interfaces to access/query/visualize the genomics data, as well as to develop tools and algorithms to facilitate the analysis of such data. One of the goals of the BRCs is to reach out to scientists working in basic research on infectious diseases to provide bioinformatics training and support and to establish collaborations to ultimately bridge the gap between these labs and the BRCs. Overall, the BRCs meet these goals and provide superb “one stop shopping” for researchers involved in pathogen-focused research. There are certain weaknesses in the BRCs, however, and areas for improvement have been identified.

On the positive side, the BRCs are largely focused on and driven by the infectious disease community, but their services are also utilized by those working on biodefense. The expert annotation and active curation at the BRCs go far beyond the data curation and maintenance offered by the National Center for Biotechnology Information (NCBI), and the organism-focused nature of the BRCs has encouraged effective integration across multiple datasets generated for individual organism by a number of data sources. The BRCs are creating useful pathogen sequence databases along with functional links (or preparing the framework for functional data entries), a model that will be useful for similar initiatives to organize general or specialized databases for other fields like agriculture or biodefense.

Future efforts at the BRCs will deal with data interoperability; in fact, the renewal of the NIAID’s BRCs program will include a single portal to link and integrate the data from all the funded BRCs. The currently existing eight BRCs will also be condensed into four centers, a move that will require integration of the existing data from decommissioned centers into the remaining databases.

Due to the requirements of the contracts on which they were founded, some of the particular weaknesses of the BRCs include:

- The various BRCs are currently focused on arbitrary sets of different types of pathogens and vectors and there is no easy way to search across them.
- Some BRCs are easier to work with than others (this pertains both to ease of use of their web interfaces, as well as to other types of interactions).
- Although all the BRCs have established and adopted a sequence and annotation data exchange format that follows the GFF3 exchange requirements, BRCs should augment the adoption of standards for storing and exchanging data.
- The BRCs have focused their development efforts on providing infectious diseases researchers with access to genomics data and analysis tools through web interfaces. The BRCs were also expected to build web services or APIs, although this was not considered to be a high-priority activity since those services are typically used by “hard core” bioinformaticists, who are not the BRCs primary targeted users. Hence, for the most part,
the BRCs lack these services, and this adds a layer of complexity to those attempting usage beyond the standard interactive web interfaces.

- Colloquium participants who are members of various BRCs noted that BRCs receive frequent changes in priority and focus from NIAID program staff.
- Owing to the focus on infectious disease, the fragmented organization of the BRCs is sub-optimal for researchers focused on biodefense, many of whom require integrated access across the entire range of pathogens, including animal and plant pathogens with large economic consequences.

Researchers receiving investigator-initiated (e.g., RO1) funding from NIAID are not typically required to work with a BRC, but these bench researchers need access to BRC resources, hardware, people, etc., and extra funding should be made available for small-scale research labs to interact with BRCs and other such centers. Other outreach activities should be pursued as well.

The service of the BRCs could be greatly improved by maintaining all relevant data in a common format as is done for some large initiatives. For example, the NIH’s Human Microbiome Project (HMP) has its own database generation and maintenance plan implemented by the NIH-funded HMP’s Database and Analysis Coordination Center, which is tasked with collecting relevant information on the HMP data. Treating the BRCs as a distributed virtual center, rather than a loose federation of independent centers, and scaling them to be capable of integrating all relevant “-omics” data of the future would render them more useful to both infectious disease and biodefense research and development.

Resources that offer information integration methods that would be useful for biodefense include:

- **Models of Infectious Disease Agent Study (MIDAS).**
  This NIGMS-funded MIDAS project seems to be a good model for some epidemiological aspects of biodefense. MIDAS uses ecosystem level data (including census data and others) to model public health outbreaks and help inform policy.

- **The Antibiotic Resistance Database (ARDB).**
  This new database provides a centralized compendium of information on antibiotic resistance to facilitate the consistent annotation of resistance information in newly-sequenced organisms and to facilitate the identification and characterization of new genes.

**RECOMMENDATIONS FOR DATABASES**

Much of the data related to biodefense need to be published publicly, and the related federal agencies need to enforce a requirement for researchers to do so. Data publication in traditional print or online journal articles is not good enough. Questions related to who should host these data and for how long, and who will fund a database, remain to be answered.

**Granularity**

The question of what does and does not belong in a database—its granularity—can only be answered in the specific context of users and of the questions they ask of the database. To determine the appropriate granularity for a database to serve biodefense needs, database designers must determine who the prospective users are and
what the users want to know. It may be possible to make portions of the information available at different levels of detail to different classes of users. The following issues need particular attention in this context:

- Data entries need to be correct.
- Use of a controlled vocabulary to identify function.
- Standards for nomenclature should be defined.
- Introduction of ways to incorporate metadata.

Rather than structure data, it may be preferable to tag data so they can be found and used. Data must be persistent (preserving the previous versions when it is modified) in order to use a tagged model, and funding agencies need to budget accordingly. Web crawlers (automated tools that search the World Wide Web) would also be needed to find the data. It may be overly naïve to assume that we will be able to define a single common way to structure all the data pertinent to infectious disease and biodefense in a traditional “database” schema. The increasing complexity of biological information and uses of it may drive us towards other models of annotating and organizing knowledge. We cannot structure information we do not yet understand.

Data that exist only in one place and on one drive can be impossible to access. The future of data storage lies in moving data to a “cloud,” and funding agencies need to drive the field in this direction. Examples of this kind of data storage can be found in the Biomedical Informatics Research Network (BIRN).

**Funding Models**

Bioinformatics database resources should remain in the government-funded domain. It is critical that federal agencies consider funding (and otherwise supporting) databases that are relevant to their mission, in collaboration with other federal agencies. Bioinformatics databases are common national infrastructure, and they cross the mission space of several federal agencies. It is likely that all federal agencies are underfunding essential bioinformatics.

In this context, there is a need to develop and sustain an umbrella database with specialized segments on biodefense, agriculture, etc. Interagency contributions to the content development of this umbrella database could provide a funding model for the database. For example, the Department of Homeland Security (DHS) will contribute for biodefense database needs and the U.S. Department of Agriculture (USDA) for agricultural database needs. This is not as straightforward as it may first seem. What portion of agricultural biodefense pathogens should DHS fund versus USDA? A similar question could be asked about food safety and FDA versus DHS. Recent problems with food safety clearly indicate there are major gaps that urgently need to be filled, regardless of the funding source(s). NCBI comes closest to an umbrella database model that can be expanded with satellite branches of specialized databases (such as small databases for biodefense or agriculture). However, an umbrella database model does not guarantee that there will be adequate integration and not merely a collection of information “silos.” The goal to be sought is an adequate integration of information, supported by all relevant agencies, that allows the range of missions (human health, biodefense, food and agriculture safety, force protection, etc.) to be adequately served without major gaps or gratuitous duplication.

It will be necessary to establish metrics to measure the utility of federally-funded database projects. The National Institutes of Health (NIH) has supported many useful model organism and broad-purpose databases, including fly base, yeast database, *E. coli*, BRCs, PDB (Protein Data Bank), and others. Other agencies, including the Nation-
Bioinformatics and Biodefense: Keys to Understanding Natural and Altered Pathogens

al Science Foundation, the DHS, and USDA have not been as good at establishing and supporting such database resources.

One important aspect of funding in particular often gets overlooked during planning: funding for curation and maintenance into the future. Money—on the order of 10% or more of the total database budget—should be specifically earmarked for these activities. Most agencies do not deal well with these issues, but the National Human Genome Research Institute (NHGRI) has developed a good funding model, allotting money for funding databases over the long term.

Effective database funding by federal agencies can be thwarted by difficulties with imposing the processes of hypothesis-driven science and peer review on agency infrastructures. Barriers to interagency cooperation can also hinder funding; interagency funding (as opposed to ownership by an individual agency) can fail in the face of pressure to hoard funding for agency issues (called “stovepiping”). The DHS has yet to contribute significant and sustained funding for bioinformatics databases that are relevant to their mission—perhaps because the culture of the agency emphasizes short-term deliverables, not research.

Collaborations and data sharing via data resources

Timely data sharing is essential, but it cannot be accomplished satisfactorily by simply publishing the results of a given project. A plan for data sharing should be defined for any new bioinformatics database resources that are developed, and researchers who make use of such a database should be held to an agreement of data sharing with the database. Funding agencies need to insist on data sharing and open access to data and follow through with penalties (such as lack of renewal) for those who do not share in a timely fashion. One model for this are the terms set forth in the Fort Lauderdale Agreement (http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf). Although this example is relevant to genomic sequence data, similar data release policies and active enforcement are needed for all other federally-funded data types. Data sharing can be facilitated by adding supplementary money for database usage monitoring. Upfront assessment of potential users should also be encouraged.

To encourage basic researchers and bioinformaticists to come together, it may be advisable to introduce a pilot funding mechanism for bench researchers in the BRC initiative. We are encouraged to learn that this is a part of the BRC renewal. There is also a need for uniform identifiers and standards to enable interaction between experimentalists and bioinformaticists. Data need to be tagged so they can be found and used in a distributed fashion and web crawlers need to be in place to find the data.

Curation: by the user or a curator?

If the proper steps are taken, it is possible to supplant the work of a database curator with user annotation, but it may be preferable to enable users to assist an expert curator with annotation efforts. User annotation can meet the standards of expert curation if due attention is paid to the following areas:

- Imparting proper training to the user.
- Continuing classroom teaching using annotation tools and methods for next generation users (this is currently being pursued in a number of BRCs, although interactive web-based training may be a feasible alternative).
- Automated systems of annotation such as the RAST (Rapid Annotation using Subsystem Technology) Server can help.
- Focus expert curation on reference genomes for a species or genus and utilize an automated system to extend it to all genomes and genes in that species or genus.
A scientist needs a profile on file when data is submitted so that scientist can become an expert curator in the future. Details of data sharing and collaboration on annotation need to be worked out up front.

Some argue that the goal is not to replace the curator, but rather to assist the curator in his or her annotation efforts. A hierarchical approach that combines human and automated techniques is best when building annotation data. There is also an opportunity here to make community-based annotation part of the hierarchy, but there are issues related to attribution and evidence code that remain to be resolved. The advent of sequencing many thousands of genomes and metagenomes each year means that human annotation will be limited to certain “flagship” projects. This will put increasing pressure on automated annotation and possibly also community annotation.

Data archiving policies that work
In research involving microorganisms, archives are needed at the level of sample, strain, and sequence. Clearly, there are different challenges in archiving physical specimens than in archiving sequence data. At this stage, it may be premature to think about not archiving certain types of data. There are no “throw-away data.” In an ideal world, all data would be archived. It is a difficult balance, but researchers worry about discarding items like trace files and mass spectrometry data in haste and repenting that decision at leisure. In the end, technology advances so rapidly that, in some cases, it may be better to simply recreate missing data—including tissue matching, metagenome rRNA, etc.—than to archive it indefinitely. Current sequencing technologies may create 1 Tb of raw image data per sequencing run, which software will turn into several hundreds of millions of bases of raw sequence reads. The sheer cost of archiving “all” data needs to be weighed against the cost and ability to recreate it. Raw data from a rare fossil clearly would have different storage value than that of a circulating isolate of *Yersinia pestis*.

If the right usage-based infrastructure is put in place, archival decisions can be easy to make. Distributed systems, for example, permit one to monitor use and archive data as appropriate. In such a system, the community helps to inform decisions about whether the information is important to leave in open format.

HOW ADVANCEMENTS IN SEQUENCING TECHNOLOGY WILL IMPACT BIOINFORMATICS

In the near future, there will be two major sources of sequence information: large pre-existing centers and independent laboratories, both utilizing Next Generation Sequencing Technologies (NGST). Unfortunately, it is also possible that much of the data generated by independent laboratories will never become publicly available. Also, many standards for metadata have been developed, and there should be a broader adherence to these standards across the relevant scientific communities. There is a concern that without collection of metadata associated with the sequence generated by all producers, this information will have decreasing utility.

For small sequencers with NGST, there will inevitably be severe problems with data uniformity at the level of DNA preparation source identification, sequence quality, annotation and semantic consistency, and metadata information attached to those data. There will also be a shortage of tools required for assembly, annotation of draft/complete genome quality, metagenomic analysis, and resequencing studies. Many of these problems, especially with respect to uniformity and semantic consistency of
annotation and metadata, will also be relevant for large centers generating sequence information. It is likely the large centers will participate in more collaborations for experimental design and for obtaining strains.

It is generally accepted that the existing sequence databases contain many errors. Many users of these data have to cope with this, perhaps by only using finished genomes from reputable genome centers. However, even these genomes can contain errors, and all users of sequence data need to be aware that even complete genomes are a statistical approximation of a potentially complex population.

The NGST permit resequencing to be used not only to search for polymorphisms but also to cross-check existing sequence data. Resequencing presents special problems in bioinformatics. Current repositories are not set up to manage these data or to allow users to view resequencing information in the context of complete genomes. We also lack services for (1) uploading resequencing information and comparing it to complete genomes, and for (2) annotation and management of this kind of data. Researchers need evaluation systems for draft quality data.

**ONTOLOGIES FOR BIODEFENSE**

Ontologies are constructs for annotating data in a semantic and computationally usable form. Although a biodefense ontology, as such, does not exist, “biodefense” research can more precisely be termed as a super-set of “infectious disease” research. There are many ontologies for infectious disease that are relevant to biodefense and can support the regularization of data relevant to biodefense, but they do not cover all aspects of this pursuit. These include gene ontologies, environment, pathogen, host, disease, symptom, and mode of transmission ontologies.

The usefulness of the various ontologies depends on the purpose for which they are used. In the future, ontologies should be developed in the context of existing ontologies as appropriate. The Ontology for Biomedical Investigations (OBI, http://obi-ontology.org/page/Main_Page) and the Open Biological Ontologies (OBO, http://www.obofoundry.org/) provide some community management to maintain sets of standards for ontologies.

Although the ontologies that have been developed and are being developed encompass some aspects of biodefense and cover them well, ontology efforts need to be nurtured and intensified for the following reasons:

- To fill gaps in the breadth of identification ontologies.
- To facilitate functional annotation (expand the focus on Gene Ontology and increase incorporation of both bacterial and viral component terms).
- To accelerate construction of integrated databases (with sequence and function information).
- To address key needs in biodefense efforts (pathogenicity and virulence prediction, to identify non-naturals, etc.). Use cases from the biodefense community would be useful in determining the specifics of these needs.

Biodefense and identification ontologies will not be developed without agency and community involvement; community involvement will be a key component of the success of these projects. The Gene Ontology and NCI Thesaurus are both good examples of ontologies that work.
While an abundant number of ontologies seem to exist, there are several obstacles that reduce general utilization of these systems. For example, some ontologies suffer from being too large, making them difficult for novice users to use. There are multiple ontologies for the same research domain, creating the potential for redundant and incompatible systems. Effort should be directed toward:

- “Slim” ontology systems which would be more useful to biodefense users.
- Training systems for the use of ontologies by novice users.
- Improved interfaces that promote rapid ontological assignment in production environments.
- Improved software systems that are able to rapidly apply ontologies to unstructured data.
- Improved incorporation of ontology information into public repositories (e.g., GenBank).

If ontologies were applied to the data on infectious agents, investigators will eventually be able to perform computational reasoning on that information. This could lead to systems with predictive powers in the areas of pathogenicity, disease-causing factors, and threat agents. These systems might also be capable of performing complex tasks, such as diagnosing disease based on patient symptoms or predicting the impact of changes in the environment on disease outbreaks. They might also be used in biodefense command and control systems. Development in this area would be useful.
CURRENT NEEDS: INADEQUACIES AND SERVICE GAPS IN TOOLS FOR BIODEFENSE AND SEQUENCE ANALYSIS

Like ontologies, software designed for biodefense is not necessarily limited in its utility to biodefense alone; they are generally useful outside of the biodefense area. In biodefense applications, common parameters include sample information, location, geo-temporal data, and others—variable shared in common with infectious disease investigations, for example.

Many useful tools exist for analyzing biological sequences; examples include ENTREZ, BLAST, and CLUSTAL. Multiple sequence analysis tools are needed in an outbreak since no one tool could provide all the answers in a complex real-world situation. With respect to biodefense, which has certain specific needs, there is a particular need to develop and validate tools and software, including:

■ **Phylogenetic assignment tools.**
   Improved tools are needed for multiple sequence alignments. Currently, phylogeny work requires the investigator to start from a curated set of alignments. Manual curation of multiple sequence alignments is still needed. It may also be advisable to optimize the existing phylogeny tools to improve their capability to identify gene conversion and or transfer, non-naturals, microbial diversity, etc. Tools that can scale to aligning many thousands of genomes are needed.

■ **Strain typing tools.**
   Highly specific genomic markers and genetic signatures and non-genetic signatures are needed in order to more easily identify strains quickly and accurately. Genome-wide analysis assays are also needed.

■ **Tools for detecting engineered organisms.**
   Researchers need to improve the existing recombination detection tools (such as those available for viral recombination). It is equally important to develop compatible databases to support these tools.

   In addition to the specific tools mentioned above, researchers require certain generic functions in order to more efficiently pursue sequence analysis. Tools with the following functions should be developed and widely disseminated so that individual researchers do not carry the burden of developing very specific tools for their purposes that are of limited utility to others:

   ■ Operations that can automatically find the recombination hot spots in the genome.
   ■ Functions to answer higher level and more complex questions (e.g., predict expression changes from genomic changes, etc.).
Functions that apply to whole genomes (most tools today work on gene by gene basis, but this approach needs to be adapted to more complex data) and across multiple genomes.

Visual ways of defining and using workflows/pipelines.

Application programming interfaces for both databases and tools so that automated pipelines can be more readily constructed.

Configurable, documented, and scalable attributes.

Layers of use so that the program can be adapted to both sophisticated and less advanced or occasional users.

Interfaces that are comprehensible for any of the various potential users, including analysts, policy experts, microbiologists, and others. Informaticists and users should engage in long and detailed conversations to create these interfaces, with significant help from funding agencies. However, it is almost certain that no single interface will satisfy all possible classes of users.

**BLAST**

BLAST is, arguably, the most commonly used tool in initial sequence analysis, and although it is powerful, colloquium participants noted several suggestions to improve BLAST. Many of these relate to the desire to link to specialized databases and/or to have better ways to deal with the flood of search results:

- Users need BLAST to interface with a standard trusted database of 16S rRNA sequences (or other marker sequences) to improve initial ID analysis for certain classes of organism identification.
- Users need BLAST to access a non-redundant viral sequence database.
- Users need BLAST to provide a set of tiered databases for query searching. For example, users could organize their search results based on similarity, based on experimental evidence, based on taxonomic classification, and other parameters.
- Users need the ability to search against an “experimental subset” after the global “similarity search.”
- BLAST does not scale well to modern uses and database sizes. Users need an easier approach to filtering search results. It is not easy to get rid of the many irrelevant results in a BLAST report, nor is it easy to define what is irrelevant in the general case.
- BLAST output parsing remains problematic for automated pipeline use. Homology tools scaled and designed for today’s problems are required, particularly with respect to the increasing need to perform metagenomic data analysis.

**SCALING: OFF-THE-CHARTS DATA GROWTH MEETS OUTMODED TOOLS**

Due to advances in sequencing technology, the rate at which scientists create biological sequence data is expected to continue to accelerate over the coming years, eventually achieving a rate that surpasses the growth of computational abilities to store and analyze those data. Existing databases, modes of curation and annotation, software, etc. will not be able to handle this much data or the newer types of complex data that are being produced (for instance, metagenomics, array, and metadata analysis). The current doubling rate of data production ensures that all current big, centralized, monolithic systems for handling and storing will, inevitably, break. We should implement new approaches and move to more distributed systems now.
Standardizing tools is not recommended at this time; biological sequence analysis is still in the learning phase, and attempting to force standardizing the available tools would stifle innovation and flexibility. However, it would be good to have a set of standard software engineering criteria to recommend for tool development.

Current computer architectures are geared toward the computing needs of today's business environment. Few biological problems are similar to e-commerce transactions, and, thus, it is reasonable to ask whether the different algorithms, data usage patterns, and memory needs for biological problems (including biodefense) could benefit from different computer architectures. Such research should probably be led by the Defense Advanced Research Projects Agency (DARPA) and the Department of Homeland Security (DHS) because industry will not likely see sufficient profit motive to pursue this.

**KEY DATA TO INTEGRATE**

With respect to biological sequence analysis, different needs should inspire different ways of visualizing the data. Visualization is critical, but browsing large volumes of sequence data is difficult at best. Researchers need visualization software for the analysis of large data sets, entry of data into ontological systems, and to improve general uniformity of data.

The technology of visualization has changed in recent years; AJAX, semantic queries, and Google Maps represent some of the new technologies and paradigms that can be exploited for visualizing sequences. Software designers need to work effectively with the research community to figure out the real problems and model useful workflows. Feedback and lists of requirements from the community will be critical to the success of any new software.

Bringing visualization and usability experts together with computer scientists who have cognitive training is also recommended for designing new systems and for training people to use them. This will be expensive, but if you are building expensive platforms, there is no way around it. There are existing models for how to do this; perhaps the best model involves getting the educators into the labs and teaching students and postdocs with their own data.

For integrating biodefense information, the optimal balance of the various means of representing data—including databases, graphs, and text-indexes—depends on the users and the questions being asked; standards should not be dictated. Infrastructure should be built to suit the demand. There has to be decision logic built into the query system. It needs to be forward thinking; any queries need to be able to handle any kind of data storage type. Also, queries need to anticipate the volume of data and provide appropriate output format choices.

There are currently gaps in the “query-ability” of datasets related to biodefense. In some cases, large data sets may be stored using systems such as relational databases or text retrieval systems, which may not readily support optimal querying across semantic networks or other “graph” based data. Researchers should develop systems that support optimized operations across semantic graphs, text retrieval systems, and relational databases by exploiting the information in ontology concept graphs. The key point here is that no single data organization methodology is optimal for all biological data, and how to optimize a query across multiple data organization methodologies is currently an unsolved problem.
VALIDATING ALGORITHMS AND SOFTWARE

There is little quality control on the algorithms used in bioinformatics. Some software packages endure peer-review, but most do not. As a result, debate rages over reliability (does it always give you the correct answer?) versus reproducibility (does it always produce the same results?) in these packages. Moreover, the line between commercial- and research-grade software is unclear.

It is recommended that all new algorithms and softwares be subjected to a validation process. Unfortunately, there is no ISO9000-like set of tests to consult for validating software. The field needs a “gold standard” for validation. Sponsors would also need to understand that they must provide adequate funding to support any validation process or it will not be done.

Open source software is not industry standard and never will be. Clearly, software validation is needed from a regulatory and legal perspective, but from an academic perspective, this is not necessarily true. The funding agency that requires the software to fulfill its research goals should take some responsibility for validating it. (In one arrangement that follows this logic, the Department of Defense takes responsibility for developing vaccines for diseases that could be deployed as bioweapons, since there is little profit motive for commercial entities to develop these resources as compared with vaccines for other, more commonly-encountered infectious diseases.)

A feedback mechanism in open access software may help to validate and improve it (this would require a Wikipedia-like interface to collect community feedback). For instance, we could develop a repository of certified software and associated user training courses with certificates. Criteria to certify the software could include: repeatable regression tests, compliance with standard operating systems, and the presence of sufficient documentation for the user to get started. The life cycle of software informs the level of software engineering rigor that goes into creating it.

GAPS IN STATISTICAL ANALYSIS OF DATA

There are a number of significant problems with the statistical analysis of biological data that may lead to false or unsupported inferences. A recognized, standardized set of statistical tools and techniques does not exist for bioinformatics; users seem to select suitable approaches to suit their individual needs. The problems and needs with statistics in bioinformatics generally fall into two categories: problems originating with software tools and problems originating with the user.

STATISTICAL TOOLS

Judging from the quality of the software tools available for bioinformatics, there is not enough communication between statisticians and bioinformaticists to design the proper tools. Moreover, there is apparently variation between the tools. In a recent test of statistical tools used in bioinformatics, the six most common array analysis packages give different results in the analysis of a single data set.

Software developers need to incorporate false positive and false negative analysis into the existing tools. There is also a need for more accurate statistical tools for phylogenetic assignment in order to meet identification and forensic needs. Researchers also need tools that incorporate mixture-genome elements in the analysis.
USER PROBLEMS

The lack of statistical training for bioinformaticists is a significant, discreditable weakness in the field. This deficit in training is not limited to bioinformatics: most biologists do not have even the minimum statistical skills set for analysis and interpretation of their data. Moreover, biologists do not know how to design experiments so that the analysis can be done with these mega datasets. Educators must begin to provide appropriate statistical training for all experimentalists. This is a crucially important issue.
Currently, two rate-limiting factors with respect to networking and computation in bioinformatics are submission of sequence data to GenBank and the disconnected arrangement of databases and informatics tools, which creates walls between users and prevents resource sharing.

**GenBank Sequence Submission**

NCBI's submission tools are not adequate for the size and complexity of the data sets users want to submit, making the deposition of genomes into GenBank the rate-limiting step in publishing sequence data and sequence analysis results. Submission of genome sequences to GenBank is extremely cost-intensive and requires an inordinate amount of researchers’ time. Unfortunately, management at NCBI is reactive rather than proactive on this issue, and progress has been unsatisfactory, leaving the research community to wonder about how GenBank will scale in response to ever-increasing data volumes.

NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease (http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html). GenBank serves as a repository of virtually all biological sequence data, but there is a concern that an overwhelming volume of data incorporated into that system may become less and less valuable over time. GenBank submission processes are not transparent, and, given the heterogeneity in data quality in this resource, researchers and funding agencies may want to consider putting other repositories with more focused missions on-line. The NIAID-funded BRCs and NHGRI-funded FlyBase are examples of data resources that provide highly integrated “model” databases for biodefense pathogens and Drosophila researchers. To avoid misunderstandings about data quality, it may be advisable to develop a lower quality database that accepts any sequence regardless of the metadata.

NCBI should lower its submission hurdles by publishing their standards and ensuring that the format of submissions are consistent and stable. The bar could also be lowered by, for example, creating a “video drop box” for sequence data. An ongoing dialogue between the research community and NCBI is also recommended. A meeting to address the backlog at genome centers and in the research community at-large and to reduce the hurdles would also smooth the path going forward. There should be an active discussion of the GenBank submission bottleneck at the national level.

Although it is clear that sequence management is in need of a new paradigm, the characteristics of a new paradigm are not apparent. NCBI may choose to employ a sub-contractor, such as Google, Amazon, or Microsoft, to help expand their capacity.
GenBank needs to take a “Google-like” approach to distributing its storage around the country. Outsourcing this function to industry should be considered.

**Resource Sharing**

The current architecture of bioinformatics creates walls between users. A distributed computing model for databases and informatics tools is recommended to overcome the barriers to sharing these resources. A distributed infrastructure that relies on the power of cloud computing (scalable internet based services and resources) is the wave of the future, and the details of any arrangement will inevitably change over time. Outreach and training are integral to the shift to distributed “cloud” systems that needs to take place. This model allows for needed compute resources to be made available on demand. The Biomedical Research Integrated Domain Group (BRIDG) model from the National Cancer Institute (among other sponsors) serves as a good example of the hardware, software, and terminology that would be involved in the recommended model.

**BENEFITS OF OPEN-SOURCE PARADIGM**

“Open-source” is a broad term, but in this context it refers to an approach in which the user has access to a product’s source data. “Open-source” resources do not necessarily mean “open-access”; these two approaches come with different licensing models. An open-source policy is mutually beneficial to the data generator (or submitter) and the user, and by allowing users to add to and extend databases, types of data, etc., it provides much-needed flexibility. The ability to make use of software infrastructures developed by sequencing centers, large-scale annotation systems, and data storage repositories is essential. Open-source strategies should be mandated by the funding agencies involved in the development of these resources.

Despite the benefits of the open-source paradigm, university rewards structures do not support credit for some open source and data sharing endeavors of this nature. Open-source projects need support and encouragement from universities and funding agencies alike.

An open-source policy for biodefense is most useful on the “user” side of the relationship because secrecy requirements for some types of data (such as active cases involving biocrimes or national security) prevent sharing. Additionally, there may be problems in getting open-source software approved to be run on high-security systems. Nevertheless, opportunities for authorized access to the biodefense sample-linked data for purposes such as academic pursuit may be an option to consider.

**THE POTENTIAL ROLES OF GOOGLE (OR GOOGLE-LIKE SYSTEMS)**

Google and its various tools and resources are potentially very useful in creating a scalable scientific infrastructure for bioinformatics. Potential roles include:

- Replacement to Entrez/Pubmed.
- Linkage into Google Earth for displays of locations of outbreaks or pathogen isolates.
- Monitoring outbreaks.
- Google could assist in fostering a type of volunteerism that circumvents government agencies by allowing users to report the incidence of disease.
- Users performing DNA and protein searches on Google.
Google’s tools would be particularly useful for visualizing metagenomic data, individual genomes and resequencing data, and links between genomes and the literature (including genome reports).

Google could also serve as a distributed annotation system for sequence storage. In such an arrangement, users would leave their data on indexable ftp sites, and web crawlers would traverse these sites to support searches. Other forms of data could be indexed in this way as well. For example, all data could be marked up in Resource Description Framework (RDF) and indexed in an improved manner, including metadata, annotation, and incidence of disease. This model could serve as an initial step in addressing the problem of rate-limited submission of data into GenBank.

Google and its resources provide a good model for some of the improvements needed in bioinformatics, and NCBI needs to adopt some of Google’s adaptability to become more facile at growing with the rapidly expanding and higher complexity database, curation, and maintenance needs. That said, Google should not and could not replace NCBI; NCBI has scientific knowledge and expertise in sequence analysis, and Google does not. NCBI will surely learn a lot from Google and should probably evolve to look more like Google in the future.

APPLYING THE SOCIAL NETWORKING MODEL TO BIOINFORMATICS

The popularity of social networking sites could be exploited to assist scientists in the identification of other users with common interests (e.g., curators of gene families, automated annotation systems) and to assist linkages between scientists of different disciplines (e.g., finding biostatisticians). Social networking sites can also aid in developers advertising software, analytical tools, and data resources.

There is also the potential to connect social networking sites to workflow systems, permitting monitoring of workflows, as well as making contact with other users, and training. In such an arrangement, users would also be able to rate software.

Social networking could also be used in training efforts, for example:

- Training members outside of the domain (e.g., computer scientists).
- Novice users for annotation.
- Novice users for ontologies.
- Instruction on lab-based protocols.
- Interviews with scientists post-publication.
RECOMMENDATIONS

DATA DESCRIPTION AND REPOSITORIES

1. The first recommendation is to focus on increasing the general awareness and perceived value of collaboration between investigator-initiated research and federally funded bioinformatics based data repositories. This report does not imply that these collaborations do not exist, but does mean to draw attention to the value of these collaborations and that there needs to be new and creative ways to increase these collaborations. Researchers receiving investigator-initiated (e.g., RO1) funding often need access to bioinformatics resources, (hardware, software, people, etc.) but find it difficult to figure out how to productively collaborate with labs or centers with these resources. Extra funding should be made available for small-scale research labs to interact with BRCs and similar centers. Other outreach activities should be pursued as well.

2. The very essence of the first recommendation presumes federally funded bioinformatics based data repositories. Hence, the second recommendation is to ensure that bioinformatics database resources remain in the government-funded domain. It is critical that federal agencies consider funding (and otherwise supporting) databases that are relevant to their mission space in collaboration with other federal agencies. Federal agencies should review whether bioinformatics budgets are consistent with the current state of technology and user needs. Better coordination between federal agencies involved with bioinformatics data resources is required.

3. There are multiple ontologies for the same research domains, creating the potential for redundant and incompatible systems. Effort should be directed toward:
   - “Slim” ontology systems which would be more useful to production-level annotators.
   - Training systems for the use of ontologies by novice users.
   - Improved interfaces that promote rapid ontological assignment in production environments.
   - Improved software systems that are able to rapidly apply ontologies to unstructured data.
   - Improved incorporation of ontology information into public repositories (e.g., GenBank).
   - Encouraging funded researchers to participate with appropriate ontology standards efforts.
INFRASTRUCTURE RESOURCES

4. We recommend moving more towards scalable distributed systems. Existing databases, software, and modes of curation and annotation will be unlikely to scale at the same rate as sequence data production in light of advances in sequencing technology. The current doubling rate of data production ensures that all current big, centralized, monolithic systems for handling and storing will, inevitably, break. Funding for infrastructure projects that focus on distributed storage and computing resources is rapidly becoming a critical impediment to the biodefense efforts. Advances in high speed wide area networks are sorely needed to support sharing large amounts of sequence data. Coordination between federal agencies involved in biodefense is far from optimal when it comes to providing for adequate bioinformatics infrastructure.

ALGORITHM AND SOFTWARE RESOURCES

5. The lack of statistical rigor for bioinformatic algorithms and analyses is a significant, discreditable weakness in the field and must be addressed. This deficit may be due to the immature nature of the field, but none the less, needs to be addressed if the field of bioinformatics is to be taken seriously in the future. We recommend continued support for interdisciplinary training focusing on the statistical skills set for analysis and interpretation of data and algorithms.