Clinical Chemistry 59:9
1301–1305 (2013)

Bioinformatics: What the Clinical Laboratorian Needs to Know and Prepare For

Moderator: Gregory J. Tsongalis[1*]
Experts: Elizabeth Chao,[2] Jill M. Hagenkord,[3] Tina Hambuch,[4] and Jason H. Moore[5]

The introduction into the clinical laboratory of novel molecular diagnostics technologies that cover a wide variety of applications has occurred at a record-setting pace and has led to revolutionary changes in the field. Microarrays, for example, are routinely used in the clinical laboratory as the first line of testing for constitutional abnormalities associated with developmental delay and autism. In addition, microarrays employing millions of probes for each analysis are used for large-scale genotyping and for gene expression profiling in association with specific clinical algorithms. Next generation or massively parallel sequencing (NGS)[6] is also becoming routine in the clinical laboratory for targeted whole-gene, exome, and genome sequencing. The amount of data generated by these analyses is unprecedented and requires a sophisticated knowledge of bioinformatics for the proper storage, analysis, and mining of these data sets. While clinical laboratories have experience with informatics and in handling large numbers of results, the systems used for those tasks are inadequate for handling the data from omics studies. In this Q&A, several leading investigators from academia and industry, who routinely use bioinformatics for omics studies, were invited to discuss the importance of bioinformatics and how clinical laboratorians can best prepare themselves for handling the increasing amount and complexity of data generated by their laboratories in conducting these studies.

Bioinformatics and biostatistics are often used interchangeably despite substantial differences. Can you define these terms and indicate how bioinformatics may impact the clinical laboratory?

Elizabeth Chao: Biostatistics applies statistics to biological topics. The emphasis is placed on using statistical analysis to design experiments involving large populations. Biostatistics highlights significant results through statistical inference, thus sorting out signal from noise. Bioinformatics, on the other hand, is a modern interdisciplinary science that unites biology, computer science, applied mathematics, and statistics into one discipline. The primary aim of bioinformatics is to develop and use computer programs to study biological processes.

Both bioinformatics and biostatistics, despite their differences, are integral to the analysis of large data sets. In clinical laboratories, bioinformatics is indispensable to accommodate the entry of high-throughput instruments and genomic data sets.

Jill M. Hagenkord: In my opinion, bioinformatics is the utilization of computational tools to study large, complex biological data sets. Bioinformatics is quite interdisciplinary and includes elements of biology, chemistry, mathematics, computer science, and statistics. Examples include algorithms for determining the DNA sequence of the exome from millions of overlapping fragmented reads or

[1] Director of Molecular Pathology and Co-Director of the Translational Research Program, Geisel School of Medicine at Dartmouth, Dartmouth-Hitchcock Medical Center, Lebanon, NH; [2] Director of Translational Medicine, Ambry Genetics, Aliso Viejo, CA; [3] Chief Medical Officer and Senior Vice President, Complete Genomics, Inc., Mountain View, CA; [4] Illumina Clinical Services Laboratory, Illumina, Mountain View, CA; [5] Professor of Genetics and Community and Family Medicine, Director of the Institute for Quantitative Biomedical Sciences, Associate Director for Bioinformatics, Norris-Cotton Cancer Center, and Editor-in-Chief of BioData Mining, Geisel School of Medicine at Dartmouth, Dartmouth-Hitchcock Medical Center, Lebanon, NH.

* Address correspondence to this author at: Geisel School of Medicine at Dartmouth and the Dartmouth-Hitchcock Medical Center, One Medical Center Drive, Lebanon, NH, 03756. Fax 603-650-6120; e-mail gregory.j.tsongalis@hitchcock.org.

Received May 1, 2013; accepted May 9, 2013.

[6] Nonstandard abbreviations: NGS, next generation sequencing; HIPAA, Health Insurance Portability and Accountability Act of 1996; IOM, Institute of Medicine; ACMG, American College of Medical Genetics.

predicting the biological impact of a DNA sequence variant on protein function. I think of biostatistics as the application of statistics to clinical trials—for example, how to sufficiently power a study or how to represent statistical significance of results.

We are seeing the migration of more and more "massively parallel" testing methods from research laboratories into clinical laboratories. Cytogenomic arrays, which essentially rebuild the chromosomes inside the computer from disrupted DNA hybridized onto a microarray, are already the first-line testing for many constitutional genetic disorders and are gaining use for cancer applications. NGS is being used to sequence all or parts of entire genomes—germline, cancer, and microbes. There is virtually no area of pathology that won't be impacted by NGS assays in the near future. Designing, validating, implementing, and interpreting clinical genomic assays will fall under the purview of pathology and each step requires a great deal of bioinformatics. All the data from NGS are digital—we are essentially turning bases into bytes and then visualizing them as bases again inside the computer. Compared to static data, this digital information readily interfaces with additional algorithms or databases to help the clinical laboratorian interpret the data in the context of the clinical question for each patient. It is critical that the responsible pathologist or laboratory scientist understands the assumptions behind each algorithm and the relative strengths and limitations of external databases. It is likely that each laboratory will need to employ one or more clinical bioinformaticians.

Tina Hambuch: Bioinformatics is the processing of biological datasets, and it encompasses logistical as well as analytical processes. Biostatistics is specifically a subset of the analytical processes focused on identifying specific types of patterns/trends through the use of statistical tools. As the scope of data becomes larger, bioinformatics becomes more critical. It also allows for clinical laboratory science to become more standardized and quantitative, in terms of both interoperator and interlaboratory performance. However, as few clinical laboratories have a strong bioinformatics contingent, this will be a significant challenge for them in the short term, and the danger of misapplied bioinformatics is substantial. For example, different programs are optimized for detecting different types of genetic variants and their performance may vary considerably; thus, understanding how the data are analyzed and the assumptions and optimizations of the various tools used for their analysis is critical to their appropriate application.

Jason H. Moore: Biostatistics is a formal discipline that employs mathematics to carry out point estimation and hypothesis testing with the goal of answering biological and biomedical problems. Point estimation focuses on the accurate estimation of population parameters such as mean and variance or other measures such as the slope of a regression line. Hypothesis testing focuses on the formal process of inference from testing a null hypothesis about one or more parameters. Understanding the basic concepts and methods in biostatistics is critical to clinical laboratorians and all others in the biomedical sciences.

Bioinformatics is a relatively new discipline that combines biostatistics with computer science to tackle complex biomedical problems. Bioinformatics had its origins in the 1970s when the need to store, manage, and analyze DNA sequence data arose. It then took off in the 1990s with the spread of the internet and affordable computers. Much of bioinformatics is focused on the development, evaluation, and application of new databases and computational algorithms for the analysis of high-dimensional data from technologies such as DNA sequencing and mass spectrometry. This focus on computer science and its many subdisciplines, such as machine learning and visualization, are what set bioinformatics apart from biostatistics. However, good bioinformaticists are able to integrate sound biostatistical methods into their analytical strategies. Both biostatistics and bioinformatics are critical for the clinical laboratory setting. For example, sequencing clinical DNA samples may require formal biostatistical methods for QC while bioinformatics would be needed to store the raw data followed by integration with research or clinical databases.

In clinical laboratories, microarray and NGS technologies represent diagnostic tools known to generate large quantities of data. What are some of the options for data analysis, storage, and mining?

Elizabeth Chao: Clinical laboratories house internal IT computing infrastructures to interpret and transmit large amounts of data. Vendors can provide solutions that offer a wide variety of high-performance comput-

ing and storage solutions to service unique laboratory needs.

Jill M. Hagenkord: Cytogenomic microarrays generate around a gigabyte of data and whole genome sequencing generates around a terabyte. However, it is rare that we ever need to go back to the raw data and it is sufficient, in my opinion, to store processed data. The processed data for a whole human genome is only a few gigabytes, which becomes financially reasonable to store. Besides, Mother Nature has already figured out the cheapest way to store genomes—it's called DNA. So, it is actually cheaper to store the processed data files for clinical samples rather than the raw data, and then reprocess the DNA on the rare occasions that it is necessary to do so. It's a different model than what we are used to in the clinical lab, but it maintains the spirit of the requirements to retain data from clinical samples.

Another consideration is to store genomic data in "The Cloud"—raw or processed—and access it remotely to write reports as needed. The final report goes into the patient's medical record, but the genomic data reside in a secure, Health Insurance Portability and Accountability Act of 1996 (HIPAA)-compliant database. If one wants to minimize costs, store the DNA in your clinical freezer and put the processed data into cold storage in The Cloud.

Tina Hambuch: There are many options available and the challenge is to identify which of these are the most appropriate. This will depend on how the particular laboratory is trying to use the data and what questions are being asked. That said, standardization around certain formats, such as BAM (binary alignment/map) or VCF (variant call format), will improve things greatly. For data storage, we use Isilon (EMC) and for the data analysis, a Sun grid engine (compute cluster). While some commercial software packages are becoming available, we use a set of custom programs and scripts.

Jason M. Moore: Two of the most widely used tools for DNA sequence analysis are Galaxy and CLC Genomics Workbench. Galaxy is a free web-based software package that includes dozens of analysis tools for DNA sequence manipulation, QC, alignment, variant detection, statistical analysis, etc. It has quickly emerged as one of the primary resources for DNA sequence analysis. CLC Genomics Workbench is a commercial software package that has much of the same functionality with a much more intuitive graphic–user interface. Although expensive, CLC Genomics Workbench is very accessible to nonbioinformaticists.

What is the "Cloud"? Is this data storage option suitable for laboratories governed by financial constraints, HIPAA, and other regulatory issues?

Elizabeth Chao: The Cloud provides easy access and distribution to data and processing capabilities. It is unrestricted geographically, allowing its services to be available at all times. Furthermore, it is a cost-effective business solution aimed to scale storage and processing capabilities in response to company needs. This idea is further explored when comparing private and public cloud network capabilities. The primary difference is the host infrastructure that predetermines the transmission and processing of data. Public infrastructure uses the World Wide Web to communicate and is inherently a risk from an information technology perspective. An in-house information technology support team responsible for safeguarding the data, on the other hand, supports private infrastructure. The latter requires stringent information security programs to be put in place.

Jill M. Hagenkord: An easy way to conceptualize The Cloud is to think of Yahoo Mail or Gmail. These are web-based email services that run on remote servers using remote software. As the end user, you do not need to install or maintain either the hardware or the software. The Cloud lets the end users utilize (and pay for) only the bandwidth that they need while enjoying the economies of scale provided by the "shared-user" model. Software, IT, security, and data center experts take care of the technology in The Cloud so that, in the case of hospital laboratories, end users can focus on using genomic information for patient care. HIPAA regulations introduce some interesting hurdles to Cloud computing, but they are not insurmountable. For example, to maintain HIPAA compliance, data should not leave US soil; it's not always transparent to the end user *where* The Cloud is at any given time. This is just one of many ways that omics-based testing sets yesterday's compliance policies on their heads. Try interpreting CLIA '88 in the context of operating a compliant clinical genome factory! However, there are many options that can be leveraged when implementing Cloud solutions. Services like Virtual Private Clouds, encryption, and access control can be leveraged to build HIPAA and other regulatory-compliant solutions. There are currently many services running in production on The Cloud that fulfill those requirements.

Genomics and Cloud computing will continue to mature and gain acceptance and are going to be part of the future landscape of genomic medicine. As laboratory professionals, we need to be aware of the current regulatory policies and ensure that our omics solutions

are adhering to those guidelines, even as we educate policy makers on what may be more appropriate for genomic testing.

Tina Hambuch: We are exploring this, but there are significant concerns around HIPAA. Ultimately, I think it will enable the exploration and optimized utility of the information that is generated, but in the short term, there are many significant challenges regarding data security and the ability to share them appropriately.

Jason H. Moore: Cloud computing is a data storage and analysis service that is offered over the internet. Public or private entities sell access to their high-performance computing and data storage hardware to users that are geographically distributed across the internet. The advantage of this model is that you don't need to maintain any computing resources yourself and pay only for those services that you use. For a clinical laboratory, this could mean using The Cloud for all data storage needs. You would pay by the gigabyte and length of time you want the data stored. What you get is cheap data storage with some redundancy that protects against loss of data due to hardware failures. Of course, the downside is that data that leaves your facility may not meet your privacy and security standards. Once the data are on someone else's server, you no longer have control over their protection. Many in clinical laboratories are not willing to take that risk at this time.

In March 2012, the Institute of Medicine (IOM) issued a report on validation of datasets generated by high-complexity testing and the subsequent analysis (Evolution of Translational Omics). What are your thoughts on the reproducibility of bioinformatic analyses and how clinical laboratories can avoid the problematic issues presented in the report?

Elizabeth Chao: Reproducibility of bioinformatics analyses has been a looming issue with the rapid introduction of genomics into clinical work. It is reassuring that both IOM and the American College of Medical Genetics (ACMG) chose to address these practices so quickly. ACMG put forth policy statements and are actively working on specific guidelines for whole-exome and whole-genome sequencing.

As leaders and early adopters in this field, our company relied on internal benchmarks for reliability and reproducibility, which are more rigorous than those in the current IOM report. Having said that, we are pleased that these minimum standards will be in place to ensure more standardization in the future. The investment required to get to this stage should not be overlooked, and is more than worthwhile.

Jill M. Hagenkord: There are different kinds of omics assays, some more transparent than others. Expression patterns can be particularly opaque to the end user, and variables must be tightly controlled to get consistent results. In addition, these types of assays require careful clinical validity and utility studies as well as technical validation. Other types of omics assays have more familiar output. For example, a cytogenomic array that produces a karyogram, showing a deletion of chromosome 13q14 in the DNA from tumor cells of a patient with chronic lymphocytic leukemia. Although there are complex algorithms that convert the disrupted DNA into the in silico karyogram, the representation of the data is familiar (a chromosome) and matches our understanding of the tumor biology of chronic lymphocytic leukemia, which provides a bit of a sanity check. Because cytogenomic arrays used in this way are merely an alternative method to detect changes with established clinical significance, technical validation and/or diagnostic yield studies may suffice.

Clinical pathologists and laboratory scientists undergo years of training to validate tests for clinical use. We hold each other to very high standards in this regard. Bioinformatics will be part of the design, validation, implementation, and interpretation of laboratory-developed tests going forward. Clinical laboratories offering genomic tests need to have clinical bioinformaticians intimately involved in the process and in close and constant communication with the laboratory-testing personnel and medical directors. The algorithms and outputs need to be validated and version controlled as part of the laboratory-developed test.

Tina Hambuch: Reproducibility is critical to the accuracy of the data, and we certainly evaluate reproducibility in our technical validations of our bioinformatic software. It is possible to achieve, but cannot be assumed.

Jason H. Moore: Clinical laboratories are by design extremely careful to avoid errors in their measurements. Many safeguards and checks and balances have been put in place to make sure that clinical data are reliable and accurate because patient care depends on it. Omics data are inherently less reliable owing to the nature of high-throughput technology such as DNA sequencing. Further, large-scale omics data often require many rounds of processing to produce useful information. Useful information is then converted to knowledge through the application of different bioinformatics and biostatistics analysis methods. Each step in the analysis pipeline from QC to final analysis and interpretation can generate unintentional errors. For example, many machine-learning methods have numerous settings

that each can have huge effects on the results. It is easy to mis-specify and misreport a setting such that the results are invalid and/or subsequently not replicable.

There is a movement in the field of bioinformatics, and computer science more generally, to provide the software and exact settings that were used with a particular data set to generate published results. Anyone should be able to download both the data and the software and easily reproduce a finding. The culture of sharing data and methodology has been slow to change but is necessary if we are to believe published results. This is another good reason why it is critical for clinical laboratorians to have a working knowledge of both bioinformatics and biostatistics.

If you had only one resource available to educate yourself about bioinformatics, what would that be and why?

Elizabeth Chao: The field of bioinformatics evolves faster than current educational resources. Therefore, global input from researchers and leaders in the field referencing new methods for data analysis remains a primary educational resource. This directly translates to attending local and national conferences and joining speakers in various universities, as well as having personal meetings with experts in the field.

Jill M. Hagenkord: That's a hard question. I don't think one could learn what they need to know for clinical genomics from one source. But, that said, I think the best way to learn is by doing. Professional societies should conduct hands-on workshops for interested members and trainees should present publically available genomic cases in case conferences on a regular basis. Many genomic companies and software vendors have publically available data sets for educational purposes, if your institution is not yet offering genomic assays. I have been fortunate to keep myself in close proximity to many very smart bioinformaticians to help clarify questions about algorithms and apparent inconsistencies in the representation of the data. Clinical laboratorians need to make an effort to get themselves educated and connected to bioinformaticians—and to be patient and keep their sense of humor while medical professionals and bioinformaticians learn to communicate with each other. I have countless funny stories about how hard it can be for a doctor and a bioinformatician to communicate.

Tina Hambuch: There really isn't a single good resource—the field is perhaps too diverse and also too nascent. It is also complex enough that people need comprehensive training that goes beyond a single book or a website. Also, resources and computation infrastructure are necessary for learning this field.

Jason H. Moore: The R statistical programming software package is an ideal focus for learning bioinformatics. R is open source and freely available and will run on Linux, Mac, and Windows operating systems. The advantage of R is that is has quickly become the primary bioinformatics analysis software package. This is partly because it is free but also because it is extensible. Many bioinformaticists release their new methods as packages within R that anyone can download and immediately use. There is extensive online documentation and a collection of bioinformatics tools in a package called bioconductor. The initial learning curve for R is a bit steep. However, the time invested is well worth the doors that knowing R opens. R now includes packages for just about anything you would want from bioinformatics and biostatistics.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: E. Chao, Ambry Genetics; J.M. Hagenkord, InVitae Corporation; T. Hambuch, Illumina.
Consultant or Advisory Role: None declared.
Stock Ownership: T. Hambuch, Illumina.
Honoraria: None declared.
Research Funding: None declared.
Expert Testimony: None declared.
Patents: None declared.

Previously published online at DOI: 10.1373/clinchem.2012.198226