ISSN 2161-3974

JeSLIB 2018; 7(2): e1129
doi:10.7191/jeslib.2018.1129

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Developing a Bioinformatics Program and Supporting Infrastructure in a Biomedical Library

Nathan Hosburgh

National Institutes of Health, Bethesda, MD, USA

Abstract

Background: Over the last couple decades, the field of bioinformatics has helped spur medical discoveries that offer a better understanding of the genetic basis of disease, which in turn improve public health and save lives. Concomitantly, support requirements for molecular biology researchers have grown in scope and complexity, incorporating specialized resources, technologies, and techniques.

Case Presentation: To address this specific need among National Institutes of Health (NIH) intramural researchers, the NIH Library hired an expert bioinformatics trainer and consultant with a PhD in biochemistry to implement a bioinformatics support program. This study traces the program from its inception in 2009 to its present form. Discussion involves the particular skills of program staff, development of content, collection of resources, associated technology, assessment, and the impact of the program on the NIH community.

Conclusion: Based on quantitative and qualitative data, the bioinformatics support program has been heavily used and appreciated by researchers. Continued success will depend on filling key staff positions, building on the existing program infrastructure, and keeping abreast of developments within the field to remain relevant and in touch with the medical research community utilizing bioinformatics services.

Correspondence: Nathan Hosburgh: Nathan.Hosburgh@nih.gov
Keywords: bioinformatics, bioinformatics support program, biomedical library

Introduction and Background

In the context of an ever-expanding information landscape, those involved in biomedical research have become increasingly reliant on the use of bioinformatics to analyze large amounts of complex data. Bioinformatics is an interdisciplinary field involving molecular biology and genetics, computer science, mathematics, and statistics. Large-scale biological problems, such as modeling biological processes, are addressed from a computational point of view so that inferences can be made from aggregate data (Can 2014). As stated by Rein (2006), "Bioinformatics research advances in such areas as gene therapy, personalized medicine, drug discovery, the inherited basis of complex diseases influenced by multiple gene/ environmental interactions, and the identification of the molecular targets for environmental mutagens and carcinogens have wide ranging implications for the medical and consumer health sectors." (Rein 2006). The field of bioinformatics has seen explosive growth since the mid-1990s, spurred by the Human Genome Project and rapid advances in DNA sequencing technology.

Despite the importance of bioinformatics in advancing scientific research, it has been observed that most researchers in the life sciences do not have the necessary training to take advantage of the array of bioinformatics tools and resources available to them due to the rapidly evolving, interdisciplinary nature of the field (Schneider et al. 2010). Extensive technological changes, new databases and software, and changes in the types and quantity of data combine to pose formidable challenges to the uninitiated. Likewise, few biomedical librarians have the training, experience, or subject expertise required to provide robust bioinformatics services such as interpretation of molecular sequence database search results, pathway analysis, and data analysis from the latest biotechnology advances. Therefore, some institutions have recruited individuals with advanced degrees in biology or biochemistry and a strong background in bioinformatics to assess molecular biological information needs of researchers and design strategies to enhance library resources and services in the areas of consultation, education, and resource development (Li, Chen, and Clintworth 2013; Yarfitz 2000; Rein 2006).

As library involvement in bioinformatics has grown, particularly across research and clinical settings, the role of the health information professional as "informationist" has become more prominent. Specifically, in the "bioinformaticist" role, the information professional possesses advanced subject knowledge in information science as well as applied technical and biological skills (Davidoff et al. 2000; Helms et al. 2004). Those responsible for building library bioinformatics programs must discern user needs and skills, identify existing services, develop plans for new services, recruit and train specialized staff, establish collaborations with other centers at their institutions, and assess the success of such programs (Geer 2006; Lyon, Tennant, and Messner 2006). If executed effectively, library involvement in bioinformatics support services has the potential to contribute to the process of scientific discovery and save the research community valuable time and money.

Study Purpose

The purpose of this case study is to outline the process of creating, developing, and assessing a bioinformatics support program at the National Institutes of Health in Bethesda, Maryland.

Case Presentation

The National Institutes of Health (NIH), a part of the U.S. Department of Health and Human Services, is the nation's medical research agency. Located in the Clinical Research Center at the heart of campus, the NIH Library supports the clinical care and research of the intramural community, which leads to discoveries that improve public health and save lives. In addition to bioinformatics, the NIH Library provides services in bibliometrics, custom information solutions, data management and analysis, document delivery, editing, literature searching, research assistance, systematic reviews, training, and translations (National Institutes of Health Library 2018b).

In 2008, the National Center for Biotechnology Information (NCBI) scaled back its bioinformatics training program, creating a need for other groups to offer the training previously provided by NCBI. The NIH Library, in keeping with its objective to support intramural research in genetics and bioinformatics more comprehensively, stepped in to fill that void by offering training specifically geared towards NIH investigators.

In February 2009, the NIH Library hired an expert bioinformatics trainer and consultant, Dr. Medha Bhagwat, to support bioinformatics research at NIH. Up to this point, the Library did not offer bioinformatics support services. Dr. Bhagwat arrived from NCBI with 11 years of bioinformatics experience as well as diverse expertise in biochemistry and structural biology.

During her tenure at NCBI, Dr. Bhagwat developed and taught several two-hour mini-courses dealing with the effective use of specialized bioinformatics tools. These included "quick start" courses on analyzing microbial genomes, structural analysis, identification of disease genes, correlating disease genes and phenotypes, understanding DNA and protein sequences, and utilizing tools such as BLAST, Entrez Gene, MapViewer, and GenBank. Leveraging the courses and training she had previously developed at NCBI, Dr. Bhagwat was able to create classes tailored to the specific bioinformatics needs of the NIH intramural research community (Bhagwat 2006). Previous work as a bench scientist endowed her with an understanding of the needs and terminology particular to biomedical researchers. The fact that Dr. Bhagwat had been employed on the NIH campus since 1994 meant that she had also generated a strong internal network and was able to feel the pulse of the research community. These qualities combined to immediately make Dr. Bhagwat a valuable resource in her new role at the NIH Library.

Although Dr. Bhagwat had the expertise, experience, and training as a bioinformaticist, preliminary work was necessary to build a comprehensive bioinformatics support program. She began by researching bioinformatics support programs at prominent medical libraries and found that such programs include one or more of the following: instruction, licensing, computing software, collections, resource development such as online tutorials, and setting up collaborations among researchers. She then sought to identify the requirements of the NIH research community via a three-pronged approach: interviews with bioinformatics specialists at several NIH institutes, direct interaction with researchers during early training and consultation sessions, and a formal survey of NIH scientists. An initial bioinformatics support program was established, consisting of classroom training, one-on-one tutorials and consultation, online tutorials, software and database licenses, high-performance computers, and a collection of books, journals, and other literature.

Classroom training is taught by NIH Library staff as well as outside speakers, including subject and product experts supplied by bioinformatics software vendors. Most of the classroom instruction is provided in the library training room with additional live streaming over WebEx in some cases. Dr. Bhagwat formed strategic partnerships with several institutes to teach on-site training programs offered to extramural scientists, medical professionals, educators, and students at other facilities. These partnerships have helped expand the reach of the NIH Library's bioinformatics support program and fostered a network of bioinformatics experts across campus. Examples include the National Institute of Nursing Research (NINR) Precision Health Boot Camp (National Institute of Nursing Research 2018a) and the Summer Genetics Institute for nurses (National Institute of Nursing Research 2018b), the National Human Genome Research Institute (NHGRI) Short Course in Genomics (National Human Genome Research Institute 2018) for middle- and high-school teachers, community college, and tribal-college faculty, and the National Library of Medicine's (NLM) remote hands-on classes hosted by university libraries for academic researchers (Charles R. Drew University of Medicine and Science 2016; University of Maryland Health Sciences and Human Services Library 2013). Dr. Bhagwat taught a 2-credit course "Practical Bioinformatics" at the Foundation for Advanced Education Sciences (FAES) at NIH annually during the fall semester (Foundation for Advanced Education in the Sciences 2015). She gave lectures at Georgetown University as adjunct faculty and provided continuing education courses at both the Medical Library Association (Bhagwat 2012) and Special Library Association conferences (Bhagwat 2010). The annual NIH Library Bioinformatics Research Symposium serves as a great example of a collaborative endeavor in which the Library organizes a two-day event featuring a series of scientific presentations highlighting practical applications of the analysis tools and databases licensed by the NIH Library for NIH researchers. The presenters are all scientists from NIH or relevant companies offering such bioinformatics tools (National Institutes of Health Library 2018a).

Examples of bioinformatics classes led by Dr. Bhagwat at NIH include: Making Sense of DNA and Protein Sequences, Gene Resources: From Transcription Factor Binding Sites to Function, Sequence Similarity Search: BLAST, Sequence Similarity Search: BLAT, Protein Structural Analysis: Binding Sites to Distant Homologs, Genome Browsers, Identification of Disease Genes, Correlation of Disease Genes to Phenotypes, Microbial Genome Analysis, Gene Expression Microarray Data Analysis, Next Gen Sequence Analysis, Gene Expression Omnibus, and Introduction to Clinical Genomics.

In addition, specific training is done by vendor-provided experts on the following proprietary bioinformatics software: CLC Biomedical Genomics Workbench, DNASTAR Lasergene, ArrayStar Qseq, and SeqMan NGen, Metacore and MetaGeneMark, GeneIndexer, GeneSpring, Genomatix Genome Analyzer, Golden Helix SVS and VarSeq, Human Gene Mutation Database Professional, Ingenuity Pathways Analysis, Partek Genomics Suite, Pathway Studio, and ProteinLounge.

Depending on the software, the library provides online access via floating licenses or directly on three specialized bioinformatics workstations, two of which have identical specifications for typical high-throughput analysis: Windows 7 64-bit, 8 cores, 48 GB RAM, and 2 TB disk space. The third workstation is designed specifically to run CLC Genomics Workbench, an application for analyzing and visualizing next generation sequencing (NGS) data. The specifications of this computer are more robust due to the demanding requirements of this sort of data analysis: Red Hat Enterprise Linux 6 64-bit, 28 cores, 512 GB RAM, and 24 TB disk space.

Even with these computing capabilities, the workstations often run overnight in order to complete such analyses.

In order to bolster support for the burgeoning bioinformatics program, a second staff member was hired in August 2010. Dr. Lynn Young has a PhD in physics with computer programming experience as well as expertise in microarray and next-generation sequencing data analysis. Employing years of teaching experience, Dr. Bhagwat provides classroom instruction and organizes vendor-led instruction, while Dr. Young devotes more time to individual and small group consultations, either on the bioinformatics workstations or in her office. Due to her background in computer science and bioinformatics, Young is uniquely positioned to collaborate with NIH researchers by assisting with software, writing scripts, and interpreting the results of complex analyses. When a researcher needs a tutorial before Dr. Young is available, she is able to refer them to a short video tutorial outlining the analysis of next-generation sequencing data using specific software and follow up later with an in-person meeting. Examples of tutorial and consultation topics include: download upstream gene sequence and identify transcription factor binding sites, gene set enrichment/pathway analysis from microarray experiments, and next-gen sequence analysis: RNA-Seq, ChIP-Seq, and miRNA-Seq.

In response to the heavy demands of instruction and consultation, the Bioinformatics Workgroup was formed to handle some of the administrative functions of the program. This workgroup consists of library staff members who are not bioinformaticists but support the program in various ways; these support roles were realized by reallocating resources among existing NIH Library staff. Support activities include communicating with vendors, scheduling and keeping an up-to-date training calendar, organizing qualitative and quantitative data from testimonials and evaluation forms, and compiling statistics on classes, tutorials, off-site presentations, workstation reservations, software usage, and other metrics that feed into assessment of the program.

The most comprehensive formal assessment covers the 2016 calendar year in which 50 training sessions were provided to a total of 1,475 participants. The Bioinformatics Workgroup adjusts strategies for advertising and works with the library Communication Workgroup to make such training available to the most attendees possible. For example, the group decided to raise the cap on registrants for each class and to publicize to people on the waiting list that, if they arrive early to class and sign in, they would be given any empty seats once the class began.

Figure 1 is a list of vendor-led training during 2016. This training for fee-based resources is typically provided as part of the Library's subscription. It gives vendors an opportunity to promote their resources and enables the user community to gain targeted experience with specialized tools.

Partnerships have been formed with other NIH institutes to provide training in library facilities, while library program staff also provide training for them at their own centers (see Figure 2). For example, the National Cancer Institute (NCI) and the National Institute of Allergy and Infectious Diseases (NIAID) offered an Exome Sequencing Analysis class in the library during 2016.

JeSLIB 2018; 7(2): e1129
doi:10.7191/jeslib.2018.1129

Figure 1: Vendor-Led Training

Class	Attendees
GeneSpring 13.1	16
Pathway Studio	9
GeneSpring 13.1	16
Partek Flow	17
Partek Genomics Suite	14
Ingenuity Pathway Analysis	27
Genomatix	19
Pathway Studio	16
GeneSpring 13.1	16
QIAGEN Ingenuity Variant Analysis	12
Partek Flow	20
QIAGEN CLC Genomics Workbench	24
Pathway Studio	15
Partek Genomics	14
MetaCore	12
GeneSpring 14.5	12
GeneSpring 14.5	15

Figure 2: Strategic Partner-Led Training

Class	Attendees
RNA-Seq	34
OmicCircos	14
ChiP-Seq Analysis	27
Exome Sequencing Analysis	33
Pathway Analysis	16

JeSLIB 2018; 7(2): e1129
doi:10.7191/jeslib.2018.1129

The list in Figure 3 represents classes led by NIH Library bioinformatics staff during 2016.

Figure 3: Staff-Led Training

Class	Location	Attendees	Instructor
Genome Browsers	NIA, Baltimore	24	Bhagwat
TCGA	NIHL	20	Bhagwat
Introduction to Clinical Genomics	NIHL	24	Bhagwat
Introduction to Clinical Genomics	NLM and remote	40	Bhagwat
Gene Expression Omnibus	NIHL	19	Bhagwat
Gene Resources	Georgetown University	20	Bhagwat
Genome Browsers	NIHL	14	Bhagwat
Pathway Analysis	NICHD	25	Bhagwat
Gene Resources	NIHL	27	Bhagwat
Sequence Analysis	NIHL	26	Bhagwat
Bioinformatics Introduction to SQL	NIHL	11	Young
NINR SGI Program	NIHL	36	Bhagwat
Bioinformatics Symposium	NIHL	320	Bhagwat
BLAST	NIHL	20	Bhagwat
NINR Boot Camp	NIHL	170	Bhagwat
NHGRI Bio 101	NIHL	25	Bhagwat
BLAT	NIHL	13	Bhagwat
Making Sense	NIHL	20	Bhagwat
Gene Resources	NIHL	20	Bhagwat
Genome Browser	NIHL	20	Bhagwat
GEO	NIHL	20	Bhagwat
Clinical Genomics	NIHL	20	Bhagwat
BLAST	NIHL	20	Bhagwat
BLAT	NIHL	20	Bhagwat
Next Gen	NIHL	20	Bhagwat
TCGA	NIHL	20	Bhagwat

In order to use networked bioinformatics resources, NIH affiliates are required to register for access to a particular resource so that an individual account is created. The highest number of new registrations in 2016 were recorded for Ingenuity software; the National Cancer Institute had the most new registrants overall. A total of 524 reservations were made for the bioinformatics workstations. Workstation 2, the only workstation with CLC Genomics Workbench software used to align short sequence reads to a genome sequence (big data analysis), had the most reservations. Partek Genomics Suite was used the most on workstation 1. Genomatix, Golden Helix SVS, and Pathway Studio represent the software reserved most frequently on workstation 3. The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) booked the most reservations of any institute in 2016.

Although quantitative data is useful in evaluating the program, researchers have often indicated the value of instruction and consultation by providing qualitative feedback. This is most often received by bioinformatics staff via email and surveys. Below is a one of many positive comments from a course participant.

> *Dear Medha,*
>
> *Year after year, we put together an outstanding selection of speakers for Short course. Year after year — against that background of excellence — your bioinformatics workshop literally blows up the brains of our teachers. One super experienced teacher in the area of bioinformatics summed it all for me by simply saying "the best bioinformatics workshop I ever attended". Thank you for your commitment to our educational projects. You are truly a cornerstone of our course.*

Discussion

The NIH Library bioinformatics program has served more than 10,000 participants directly through classroom training and individual consultations since 2009. Drawing from the quantitative and qualitative data, it is clear that the NIH Library Bioinformatics Support Program is well-used and appreciated by researchers. However, in order to remain relevant, it is important to understand the evolving needs of the NIH research community. Based on the experience of bioinformatics staff, it is necessary to be in regular contact with NIH researchers as well as the larger bioinformatics and library communities. Focused conferences, seminars, and individual consultations with investigators offer excellent opportunities for keeping track of current trends. Within the NIH research community, consultations in particular afforded staff the best opportunity for understanding the topics, effective modes of training, and resources required by researchers for bioinformatics analysis. These interactions indicated that many users benefit from individualized training in ways that large group training and webinars cannot address. In this setting, bench scientists and clinicians are able to engage in substantive conversation with informationists, discuss ideas, and directly apply knowledge to real-world problems in real-time. Forming a network of bioinformatics experts throughout the NIH community has also been a key factor in the growth and success of the program.

In the coming years, as new biotechnologies emerge, staff must identify cutting edge trends and emerging needs and make modifications — both qualitatively and quantitatively — in certain aspects of the program. For example, more in-person classes may be needed to accommodate demand for this format as evidenced by feedback on evaluation forms. And more online offerings tailored towards the library community at large might be provided to reach a broader audience and enable learned application of general bioinformatics concepts using practical techniques. In regard to data infrastructure, storage, and analysis, staff will need to work closely with the NIH Library Information Architecture Branch to investigate the merits of cloud computing versus high performance workstations and associated servers supported by NIH, although reliable network speed is a potential limiting factor for moving in this direction. Government security in a networked environment is also a perennial concern and the Library must find comprehensive solutions for data backup and storage.

In July 2017, Dr. Bhagwat retired from the NIH Library. She was instrumental in creating the bioinformatics support program in 2009 and has been a cornerstone since that time. It remains to be seen whether her role can be filled by someone with the necessary experience, enthusiasm, and vision, not only to keep the program running, but to foster innovation and build on past successes. As with the NIH, institutions that have recruited individuals with advanced degrees in the biosciences into such roles have been able to create and sustain successful bioinformatics support programs (Rein 2006; Li, Chen, and Clintworth 2013; Yarfitz 2000). While it takes a leader to spearhead such an endeavor, a dedicated support team is necessary to handle some of the administrative aspects such as scheduling, promotion, and data collection. In this way, subject experts can devote more of their time to directly assisting researchers.

The NIH Library Bioinformatics Support Program has grown to encompass staff and vendor-led classes, in-person consultations, online tutorials, high-performance workstations, analysis tools and databases, and other curated bioinformatics resources (National Institutes of Health Library 2018a). As this program evolves, the NIH Library strives to provide a dynamic and valuable suite of bioinformatics services to NIH and the larger medical research community well into the future.

Acknowledgements

The author would like to thank Dr. Medha Bhagwat for supplying much of the information regarding the details of the bioinformatics program, Dr. Lynn Young for leading the initiative to evaluate the NIH Library Bioinformatics Program, and Lisa Federer for proposing that such a case study be written.

Disclosure

The author reports no conflict of interest. Products named are for informational purposes only. The NIH Library does not endorse specific software or databases.

References

Bhagwat, Medha. 2006. “Mini Courses - NCBI Resources.” *NCBI Website*. https://www.ncbi.nlm.nih.gov/Class/minicourses/#bioinformatics

———. 2010. “Genetic Resources: From Chromosomal Location to 3-D Structure.” In *2010 Special Libraries Association Annual Conference*. New Orleans, LA: Special Libraries Association. http://dbiosla.org/publications/pubs/biofeedback/Spring2010.pdf

———. 2012. “Microbial Genome Analysis and Comparisons: Web-Based Protocols and Resources.” In *2012 MLA Annual Meeting.* Seattle, WA: Medical Library Association. http://www.mlanet.org/d/do/1854

Can, Tolga. 2014. “Introduction to Bioinformatics.” In *miRNomics: MicroRNA Biology and Computational Analysis, Methods in Molecular Biology*, edited by Malik Yousef and Jens Allmer, 1107: 51-71. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-748-8_4

Charles R. Drew University of Medicine and Science. 2016. “Bioinformatics: Clinical Genomics Subject of Mini Course.” *Newsletter CDU*. Published April 1. https://www.cdrewu.edu/CDUNewsletters/activenews_view.asp?articleID=719

Davidoff, Frank and Valerie Florance. 2000. “The Informationist: A New Health Profession?” Annals of Internal Medicine 132(12): 996-998. https://doi.org/10.7326/0003-4819-132-12-200006200-00012

Foundation for Advanced Education in the Sciences. 2015. “Practical Bioinformatics.” *2015-2016 Catalog of Courses and Student Handbook*. https://faes.org/sites/default/files/files/FAES Catalog 2015-16 FINAL.pdf

Geer, Renata C. 2006. “Broad Issues to Consider for Library Involvement in Bioinformatics.” Journal of the Medical Library Association 94(3): 286-298, E152-E155. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525323

Helms, Alison J, Kevin D. Bradford, Nancy J. Warren, and Diane G. Schwartz. 2004. “Bioinformatics Opportunities for Health Sciences Librarians and Information Professionals.” *Journal of the Medical Library Association* 92(4): 489-493. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC521520

Li, Meng, Yi-Bu Chen, and William A. Clintworth. 2013. “Expanding Roles in a Library-Based Bioinformatics Service Program: A Case Study.” *Journal of the Medical Library Association* 101(4): 303-309. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3794686

Lyon, Jennnifer A., Michele R. Tennant, and Kevin R. Messner. 2006. “Carving a Niche: Establishing Bioinformatics Collaborations.” Journal of the Medical Library Association 94(3): 330-335. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525329

National Human Genome Research Institute. 2018. “NHGRI Short Course in Genomics.” Last Modified February 7. https://www.genome.gov/10000217/nhgri-short-course-in-genomics

National Institute of Nursing Research. 2018a. “NINR "Precision Health: Smart Technologies, Smart Health” Boot Camp.” Accessed March 9. https://www.ninr.nih.gov/training/trainingopportunitiesintramural/bootcamp

———. 2018b. “Summer Genetics Institute (SGI).” Accessed March 9. https://www.ninr.nih.gov/training/trainingopportunitiesintramural/summergeneticsinstitute

National Institutes of Health Library. 2018a. “Bioinformatics Support Program.” Accessed March 9. https://nihlibrary.nih.gov/services/bioinformatics-support

———. 2018b. “NIH Library - About Us.” Accessed March 9. https://nihlibrary.nih.gov/about-us

Rein. 2006. “Developing Library Bioinformatics Services in Context: The Purdue University Libraries Bioinformationist Program.” *Journal of the Medical Library Association* 94(3): 314-320, E193-E197. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525331

Schneider, Maria Victoria, James Watson, Teresa Attwood, Kristian Rother, Aidan Budd, Jennifer McDowall, Allegra Via, Pedro Fernandes, Tommy Nyronen, Thomas Blicher, Phil Jones, Marie-Claude Blatter, Javier De Las Rivas, David Phillip Judge, Wouter van der Gool, and Cath Brooksbank. 2010. “Bioinformatics Training: A Review of Challenges, Actions and Support Requirements.” *Briefings in Bioinformatics* 11(6): 544-551. https://doi.org/10.1093/bib/bbq021

University of Maryland Health Sciences and Human Services Library. 2013. “March 2013 – Volume 7 – Number 4.” *Connective Issues*. http://www2.hshsl.umaryland.edu/newsletter/?p=1434

Yarfitz. 2000. “A Library-Based Bioinformatics Services Program.” *Bulletin of the Medical Library Association* 88(1): 36-48. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC35196