YALE JOURNAL OF BIOLOGY AND MEDICINE 85 (2012), pp.305-308.

FOCUS: EDUCATING YOURSELF IN BIOINFORMATICS

Introduction

Jonathan Rothberg[a], Barry Merriman[b], and Gadareth Higgs[c*]

[a]CEO of Ion Torrent, a division of Life Technologies, Inc., Director of The Rothberg Institute for Childhood Diseases, Guilford, Connecticut; [b]Lead Architect for Advanced DNA Sequencing Technology at Life Technologies, Inc., Visiting Professor of Human Genetics at UCLA, Los Angeles, California; [c]Computational Biology & Bioinformatics PhD Candidate, Yale University and YJBM September Issue Editor, New Haven, Connecticut

We believe the field of bioinformatics for genetic analysis will be one of the biggest areas of disruptive innovation in life science tools over the next few years.

— Isaac Ro, Goldman Sachs analyst [1]

While the term "bioinformatics" originated from Paulien Hogeweg, a Dutch theoretical biologist, in 1970 [2], the groundwork was laid by the founding fathers of information theory and computer science, Claude Shannon and Alan Turing.

Shannon, widely known for his Master's thesis work in the application of Boolean logic to digital design, outlined the basis for bioinformatics in his 1940 PhD thesis entitled "An Algebra for Theoretical Genetics" [3,4]. However, as Shannon's main interest was in communication theory and cryptanalysis, he never pursued this work any further [4].

In 1943, Turing visited Shannon at Bell Labs to share British code-breaking methods, and the two exchanged ideas. Later on, near the end of his life, Turing applied his computing acumen to biology, penning his famous 1952 work in spatial modeling of morphogenesis [5]. Building on work by D'Arcy Thompson, Turing, essentially employing bioinformatics, modeled biological growth in fir cones, using the Manchester University Mark I computing machine [4].

From the beginning, bioinformatics has been a field driven by big data. It is now defined as the interdisciplinary toolset for applying computer science, mathematics, and statistics to the classification and analysis of biological information. At the highest level, bioinformatics is used to analyze large datasets to help answer biological questions, both in fundamental biology and in the biology that underlies disease.

As such, developments in this field have been driven by our ever-increasing ability to accrue large amounts of biologi-

To whom all correspondence should be addressed: Gadareth Higgs, 300 George St., YCMI, Suite 501, P.O. Box 208009, New Haven, CT 06520-8009; Tele: 203-737-6029; Fax: 203-737-5708; Email: gadareth.higgs@yale.edu.

†Abbreviations: DNA, deoxyribonucleic acid; HGP, Human Genome Project; YJBM, Yale Journal of Biology and Medicine; NIH, National Institutes of Health; NHGRI, National Center for Human Genome Research.

Keywords: bioinformatics, microarrays, sequencing, next-generation sequencing, 454, Ion Torrent, medical records, electronic health records

cal data, and they ultimately trace back to the underlying data generation technology. The first major effort that propelled this technology was the International Human Genome Project (HGP†) [6,7], which was enabled by the introduction of automated DNA sequencing instruments in the mid-1980s. This resulted in the unprecedented development of sequence analysis techniques for assembling our genome.

This project turned into a 10-year, multinational, multi-billion dollar effort to generate the sequence of a single reference human genome. While this project was a necessary impetus for the development of the field, such large-scale projects are not easily sustained. Nevertheless, it led to the democratization of data generation.

The first democratizing technology was the DNA microarray, which was introduced in the mid-1990s [8]. Microarrays enabled researchers to perform a single-day, low-cost experiment that surveyed the gene activity of all genes in a cell sample — an effort that would have otherwise taken several man-years to produce. This thousand-fold advance resulted in many labs producing enormous datasets of gene activity, and it stimulated a great deal of bioinformatics development in the areas of deducing gene function and classifying disease conditions. Microarray technology was further extended to provide sequence information at a large number of variable marker sites in the genome. This enabled large-scale statistical studies of the association between gene variants and the risk for disease [9]. In these gene expression and genotyping roles, and in the hands of many researchers and labs, DNA microarray data was a major driving force for bioinformatics development between 1998 and 2008.

The democratization of data generation was further propelled by the introduction of massively parallel DNA sequencers in 2005 [10]. These provided a major performance advance over the automated sequencers that powered the HGP and ushered in a new wave of diverse DNA sequence data. As a display of the power of this new technology, it was used to sequence the first genome of a specific individual, that of Dr. James Watson, co-discoverer of the double helical structure of DNA and Director of the original HGP [11]. Taking merely a few months, this process cost less than $1 million and reproduced the entire HGP on a scale that could be performed by a single lab. Such sequencers facilitated developments in personal genome and next-generation sequencing technology [12]. Eventually, genome-scale sequencing became accessible to local research centers and larger labs, as costs dropped to $5,000-$10,000 genomes, deliverable in 1 to 2 weeks. Ultimately, sequencing-based approaches began to supplant DNA microarrays for many research applications.

Currently, and in the foreseeable future, the explosion of data emanating from human genome sequencing will drive bioinformatics. However, the field will undergo a critical inflection, as more of this data will be generated in translational and clinical settings. This change was recently reflected at the highest levels of government planning, when, earlier this year, the NIH created the National Center for Advancing Translational Sciences with a first-year budget of $575 million and a mandate to develop this new field of genomic medicine. This landmark initiative is the first national funding shift in genomics since the inception of the original HGP and parallels the events of 1989, when the NIH created the National Center for Human Genome Research (now NHGRI) as the organizing and funding agency that would initiate and drive the HGP. While thousands of research subject genomes are currently being sequenced per year, it is anticipated that millions of patient genomes will be sequenced annually, in combination with medical records. This, in turn, will result in unprecedented power and need to use bioinformatics to resolve the genetic components of human diseases.

In order for this goal to be realized, continued advancements in data creation must be made. Such advances are needed to deliver the $1,000 same-day human genome, a performance landmark that has long been considered the threshold for when genome sequencing can enter the realm of routine medical diagnostics [13,14]. While such advanced tech-

nology will most likely propel the creation of an entirely new industry, the first platform capable of this was introduced this year [15,16]. The technological advance employed was to perform the sequencing entirely using a standard semiconductor sensor chip, thereby placing sequencing on the same technological foundation as computers, digital cameras, and smartphones. This enables sequencing to directly leverage the trillion-dollar investment that the semiconductor industry has made in chip manufacturing capability and to directly benefit from Moore's Law — the doubling of chip performance every two years. To highlight the importance of this historical scaling, the first human genome sequenced with this new chip-based technology was that of Gordon Moore [15], founder of Intel Corporation and originator of Moore's Law [17]. Such sequencing technology will prove vital in driving the next wave of clinical genome sequencing.

Yale has played an important role in this technological progression, as alumnus Jonathan Rothberg (PhD Yale, '91, Biology) introduced the first next-generation sequencing platform, the 454 system [10]. Most recently, through Ion Torrent [15], Rothberg introduced the first platform that could provide the $1,000 genome.

As a university with one of the leading research medical centers and departments of Human Genetics, Yale is ideally positioned at the forefront of this next wave of clinical genome sequencing. Yale's Center for Genomic Medicine, under the direction of Richard Lifton, Chair of Human Genetics and Professor of Medicine, is pioneering the use of clinical genome sequencing with the latest Ion Torrent technology.

Because bioinformatics is data-driven biology, it is highly interdisciplinary and requires the application of mathematical knowledge, computer science, and statistical skills for working with biological data. Constructing a sound education in this field is challenging, as there are no time-honored curricula to follow.

In the first article, Bagga describes the challenges of developing a well-structured bioinformatics undergraduate degree program. This piece is an invaluable summary of a decade of experience in the creation and evolution of such a curriculum. Auerbach then addresses these issues from the perspective of a bioinformatics graduate student, with a helpful guide to navigating complex educational choices.

As the future of bioinformatics will increasingly be driven by the clinical use of genome sequencing and patient data handling, bioinformaticians who work in clinical environments and physicians familiar with bioinformatics techniques will be in high demand. This raises further educational issues, and Rubinstein outlines the ideal educational path for those who will be working at this critical nexus between clinical medicine and large-scale data analysis.

The ultimate bioinformatics challenge in medicine is to elucidate the genetic basis and treatment of cancer. The diversity and complexity of tumor genetics and biology, combined with the severity and ubiquity of cancer, create unmatched potential for bioinformatics applications. Two articles illustrate the status and challenge of cancer to bioinformatics.

The review article by Fendler et al. summarizes what we have learned from years of DNA microarray research and early efforts at sequencing cancer genomes. It also highlights many of the remaining challenges and opportunities.

The original research article by Parisi et al. provides an excellent illustration of the complexity of cancer biology, particularly the complexity arising from heterogeneous cell populations within a single tumor. The paper then shows how this complexity can be teased apart by integrative bioinformatics analysis drawing on technologies such as targeted genome sequencing via DNA microarray assessment of SNP markers and DNA microarray assessment of genome-wide gene expression.

The major target for bioinformatics in the clinical arena will be massive databases of electronic health care records, combined with genome sequence data. The creation of such databases is essential to discovering genetic risk factors for disease causation and

treatment response. A pair of articles by Kerr et al. and Ronquillo describes the factors that will support the creation of such databases, the barriers to their creation, and the ultimate utility they can provide in personalized medicine.

We hope that this special issue of *YJBM* provides both inspiration and guidance to those interested in becoming bioinformaticians or merely learning more about this exciting field. The new era of clinical genome sequencing is just beginning, and there has never been a better time for bioinformaticians to make a real and lasting impact on human health.

Acknowledgments: JR and BM compiled the original manuscript.GH reordered, edited, formatted, and added to the piece.

REFERENCES

1. Pollack A. DNA Sequencing Caught in Deluge of Data [Internet]. The New York Times. 2011 Nov 30. Available from: http://www.nytimes.com/2011/12/01/business/dna-sequencing-caught-in-deluge-of-data.html?pagewanted=all.
2. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol. 2011;7(3):e1002021.
3. Shannon C. An algebra for theoretical genetics [PhD thesis]. [Cambridge, MA]: MIT; 1940.
4. Searls DB. The Roots of Bioinformatics. PLoS Comput Biol. 2010;6(6):e1000809.
5. Turing A. The chemical basis of morphogenesis. Phil Trans R Soc London Ser B. 1952;237(641):37-72.
6. Watson JD. The human genome project: past, present, and future. Science. 1990;248(4951):44-9.
7. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
8. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467-70.
9. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78.
10. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. Nat Biotechnol. 2008;26(10):1117-24.
11. Rothberg JM, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452(7189):872-6.
12. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010;11(1):31-46.
13. Robertson JA. The $1000 genome: ethical and legal issues in whole genome sequencing of individuals. Am J Bioeth. 2003;3(3):W-IF1.
14. Service RF. Gene sequencing. The race for the $1000 genome. Science. 2006;311(5767):1544-6.
15. Rothberg JM. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348-52.
16. Westly E. Device brings $1000 Genome within reach [Internet]. MIT Technology Review. 2012 Jan 12. Available from: http://www.technologyreview.com/news/426603/device-brings-1000-genome-within-reach/
17. Moore GE. Cramming more components onto integrated circuits. Electronics. 1965;38(8):114-7.