TRAINING THE NEXT GENERATION OF QUANTITATIVE BIOLOGISTS IN THE ERA OF BIG DATA

KRISTINE A. PATTIN AND ANNA C. GREENE
Institute for Quantitative Biomedical Sciences, Dartmouth College
Hanover, NH 03755, USA
Email: Kristine.A.Pattin@Dartmouth.edu, Anna.C.Greene@Dartmouth.edu

RUSS B. ALTMAN
Department of Genetics, Stanford University
Stanford, CA 94305, USA
Email: russ.altman@stanford.edu

KEVIN B. COHEN, ELIZABETH WETHINGTON, CARSTEN GÖRG
Computational Bioscience Program, University of Colorado School of Medicine
Aurora, CO 80045, USA
Email: kevin.cohen@ucdenver.edu

LAWRENCE E. HUNTER
Computational Bioscience Program, University of Colorado School of Medicine
Aurora, CO 80045, USA
Email: larry.hunter@ucdenver.edu

SPENCER V. MUSE
Department of Statistics, North Carolina State University
Raleigh, NC 27695, USA
Email: muse@ncsu.edu

PREDRAG RADIVOJAC
Department of Computer Science and Informatics, Indiana University
Bloomington, IN 47405, USA
Email: predrag@indiana.edu

JASON H. MOORE
Institute for Quantitative Biomedical Sciences, Dartmouth College
Hanover, NH 03755, USA
Email: Jason.H.Moore@Dartmouth.edu

1. Workshop Focus

Francis Collins recently stated that, "the era of 'Big Data' has arrived, and it is vital that the NIH play a major role in coordinating access to and analysis of many different data types that make up this revolution in biological information."[1] With this, Philip E. Bourne was named as the Associate Director for Data Science at the NIH, the first permanent appointment of this position. Additionally, through the Big Data Initiative started in 2012, the Obama Administration invested $200 million dollars in "big data" research that promises "to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data."[2] The term "big data" extends beyond the research arena into the popular press. CNN Money has named the data scientist job as one of the best jobs in America (#32/100).[3] Harvard Business Review Magazine has named the data scientist as "the sexiest job" of the 21st Century.[4] Yet, what exactly is a data scientist? The focus of this workshop is to discuss key skill sets for biomedical data scientists to determine if they differ from a standard bioinformatics curriculum.

- Is there any substantive difference between a biomedical data scientist and a biomedical informaticist? If so, how do we train one versus the other?
- Are current bioinformatics curricula evolving to encompass the realm of data science?

- Are there obsolete lessons in coursework that could be replaced with more modern technical information?

We will have 6 scientists from various quantitative fields describe how their program's curriculum is structured and what changes they have made or anticipate they could make to strengthen and align the program with current practices in data science and bioinformatics. They will also speak to where future training in bioinformatics should go in the growing era of big data.

2. Workshop Contributions

The speakers were asked to respond to the following question: What is a data scientist? Do the key skill sets for biomedical data scientists differ from a standard bioinformatics curriculum?

Russ B. Altman. I think that the concept of a "Data scientist" has emerged within industries where large amounts of information are collected and managed by individuals with skills in statistics, computer science, information science and related disciplines. For many of these industries, there is no tradition of employees with this skill set—they were used to hiring engineers from the traditional engineering disciplines or (in some cases) natural scientists from biology, chemistry, and physics. Sometimes they may have hired a statistician, but this was usually for study design or analysis of relatively orderly "controlled" data. The phenomenon of an employee with a firm grounding in statistics, but also with ability to write and run programs to handle relatively large amounts of data[Footnote1], and apply the principles of data mining and machine learning is new in many fields. In addition, there are skills from informatics that are also critical including understanding the use and maintenance of controlled terminologies and ontologies.

Within biomedical research, the field of biomedical informatics has existed (arguably) since the early 1960's when Ledley & Lusted outlined in Science[5] some of the major challenges to information sciences in biomedicine. In the early 1980's programs emerged to train professionals in biomedical informatics. The curricula that emerged were, in many cases, very similar to the curricula created today for data scientists; they included a strong background in computer science, statistics, probability, decision theory and (importantly) courses in the domain of application. The final element is quite important so that the individual understands the major questions and challenges in the domain, and knows when certain problems have been solved, and when they are unsolved. The main concern about undifferentiated data scientists who lack domain knowledge is whether they will be as efficient and effective as practitioners with an understanding of the underlying application area. For biomedicine, there is little doubt that the best data scientists will be those who understand the special features and challenges in biology or medicine, and thus make assumptions and approximations that are valid and not fatal.

[Footnote1] In this context, "Big Data" can be defined as any data set that is mission critical to the organization and bigger than what their current infrastructure can handle. As soon as the infrastructure and staff adjust to "Big Data" it becomes regular data.

Kevin B. Cohen, Elizabeth Wethington, Carsten Görg. Examining the advertisements for open positions for data scientists on the popular Monster.com job-seeking website shows that biomedical science is well-represented in the data science job market (at least on the date of search). The search [data scientist] returns 180 job openings, and [data scientist biomedical] returns 8. (The search [data scientist health] returns 30, but many of these simply mention that they provide health insurance to employees.) Examining the advertisements themselves is revealing. The sample returned by the [data scientist biomedical] search is small—8 positions in total—but some trends emerge.

The first thing of note is that the list of required skills for most of these positions is short. This might be somewhat surprising. Data scientists are typically thought of as some sort of engineer or statistician on the one hand, or as a sort of jack-of-all-trades on the other. These advertisements suggest that a jack-of-all-trades is not needed, but rather that a relatively small set of skills will suffice in most (although not all) cases. (This is apparently true of getting the job—whether or not a limited skill set would be sufficient to keep the job is less obvious.) In the small skill sets mentioned in most of these advertisements, two specific skills predominate. Databases are mentioned in three of them, and statistics are mentioned in three of them—not the same three.

How does this compare to a standard bioinformatics curriculum? A recent highly unscientific (and unpublished) survey of bioinformatics doctoral programs showed that databases were not part of any them, and statistics was not covered in an independent class in any of them. It is, however, unlikely that students typically leave a bioinformatics doctoral program without any background in statistics or databases—it is likely that they enter their doctoral program already having a background in these areas. If not, they are likely to pick it up in the course of their education, although our survey suggests that they are not doing so in their coursework. The key skill sets for biomedical data scientists do seem to differ from a standard bioinformatics curriculum.

Lawrence E. Hunter. Taking "Data Science" as defined in Vasant Dhar's CACM article[6] summarized as "the generalizable extraction of knowledge from data" and requiring "an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions"; while there are clear overlaps between bioinformatics and biomedical data science, there are also important differences. Significant aspects of bioinformatics fall outside of this definition. For example, many of the key methods for dealing with protein structural data (e.g. molecular dynamics simulation or structural visualization) are not subsumed by Data Science, but are clearly of fundamental importance in bioinformatics. Likewise, many of the critical techniques for handling massive short read sequencing data would not be included in a reasonable definition of data science. A well-trained bioinformatician knows about computational techniques that are important in contemporary molecular biology, but that are not clearly part of the Data Science toolkit. Furthermore, an effective bioinformatics researcher will have deeper domain knowledge than is typically assumed for a data scientist. Many important innovations in bioinformatics have come from a deep familiarity with the underlying biology or even more frequently with the experimental methodologies that generate the data to be analyzed. Insights into the idiosyncrasies of instruments such as mass spectrometers and hybridization arrays have led to dramatic improvements in informatics methods not available to those who treat data as a "given". Perhaps the most important difference, however, is not about the computational methods or domain knowledge of the practitioners, but about the goals of the scientific work. Philosophers of science Carl Craver and Lindley Darden have eloquently described the central role of elucidation of mechanism in biology.[7] Data science is largely concerned with finding patterns in data. While such patterns have the potential to be extremely helpful to understanding living systems, their identification is the beginning of biology, not the end. Biologists insist on mechanistic understandings of the phenomena they observe, not merely predictive ones. Bioinformatics must always be acutely sensitive to the needs of biologists to hypothesize and test mechanisms, not just to find predictive patterns.

Spencer V. Muse. The Harvard Business Review recently dubbed data science "the sexiest job of the 21st century." Acknowledging the challenge of defining the field, they suggested that "data scientists' most basic, universal skill is the ability to write code". This claim implies that data

science is grounded in computer science. Reflecting a different perspective, celebrity statistician Nate Silver remarked that data science is a "sexed up term for statistician." The truth likely falling somewhere in the middle, most data scientists would likely agree that they work at the intersection of computer science and statistics, with a heavy dose of discipline-specific knowledge thrown into the mix. The demand for these individuals has presented a workforce and training challenge. The fundamental difficulty is one shared by most emerging interdisciplinary areas: the traditional educational paths in the constituent fields lack the breadth or flexibility to allow students to easily become data scientists. A core set of skills for data science training is beginning to emerge, though. Students must be proficient programmers able to work with large heterogeneous data sets, often distributed across multiple locations. (Note that few traditionally-trained statisticians have those skills.) Students must also be fluent with a wide range of statistical techniques, have a strong knowledge modeling complex data, and be able to combine those skills to build advanced statistical analysis tools. (Abilities rarely found in traditionally-trained computer scientists.)

It is no surprise that the influx of data has created tremendous demand for data scientists. Under the umbrella of "biomedical informatics" we now have specialization in areas including medical informatics, clinical informatics, and bioinformatics. In the same way that one would not expect an endocrinologist to perform well as a cardiologist simply because they are both physicians, one should not expect to place someone trained in, say, bioinformatics into a medical informatics position and get satisfactory performance. While there is certainly a high degree of overlap (e.g. complex, high-order database searches; network construction; data mining and predictive modeling), the details of the needs and tools in each specialty are driven by the fundamentals specific to each, and one can neither be an effective developer or user of the tools without being firmly grounded in the underlying discipline.

Predrag Radivojac. Data science may best be described as a discipline whose intellectual core derives from the interplay between statistics and computer science. Statistics generally provides frameworks for modeling and inference from data. Numerous such approaches have been proposed in both predictive and descriptive scenarios as well as for characterizing inference methods. Computer science, on the other hand, studies computing paradigms for implementing such approaches. It generally provides algorithmic framework for solving statistically formulated problems, given the resources such as a particular computer architecture, clock time and memory. In addition, computer science provides a framework to formally address data management, software engineering and visualization issues. Various concepts from other disciplines also contribute to data science; for example, those from physics, biology, psychology, logic, information theory and others.

A biomedical data scientist must possess core competencies in statistics and computer science, but must also understand the biomedical side of the equation. Biomedical expertise may come from a diverse set of sub-disciplines, including molecular biology, developmental biology, evolutionary biology, biochemistry, analytical chemistry, genetics, pharmacology and neuroscience, but also a combination thereof. Overall, a biomedical data scientist must not only have deep domain expertise and the ability to identify important biomedical problems, but also the ability to formally pose such problems within statistical and computer science frameworks and then properly solve them.

I believe that the core skills of a biomedical data scientist significantly overlap with those of a bioinformatics scientist, but the main difference may come from the emphasis on particular problems rather than the ability of such scientists to be able to tackle them. A data scientist may have a larger and deeper focus on data modeling problems, perhaps of the systems biology or functional genomics flavor, whereas a narrowly defined bioinformatician may be more focused on

algorithmic issues such as sequence analysis. However, a large number of traditional bioinformaticians regularly handle data modeling issues typically by developing and applying machine learning methodologies as well as by creating tools for biologists and medical scientists.

In my view, biomedical data science is a suitable umbrella term for a host of other disciplines that rely on biomedical data to ultimately produce knowledge. At this time, however, it is too early to tell whether it will have transformative impact on biomedical research beyond what other (overlapping) disciplines have already initiated. Consequently, the development of biomedical data science curricula may not require significant restructuring of the more traditional but broadly defined bioinformatics, biomedical informatics, or systems biology curricula on many campuses.

Jason H. Moore. Data science is a rapidly emerging discipline that combines pieces of computer science and statistics to manage and analyze big data across different domains. At face value, this definition is not much different than some definitions of bioinformatics where the big data is coming from biological or biomedical sources. One possible difference between the two disciplines is with regard to the integration of statistics. Bioinformatics has traditionally focused much more on the computational sciences including algorithms, databases, high-performance computing, machine learning and software engineering, for example. This is likely due, in part, to the lack of formal statistics training in computer science curricula. Fully exploiting the potential of big data requires and equal mix of computational and statistical sciences. For example, a working knowledge of statistical inference can significantly complement machine learning approaches to big data where false-positives (type I errors) and false-negatives (type II errors) are common. Similarly, the ability to complement computational methods such as support vector machines with statistical methods such as logistic regression expand the analytical toolbox in useful ways. The demand is there and data scientists are few and far between given the rarity of in depth training in both computational and statistical sciences. Given the need for this unique blend of skills and expertise it might be time for bioinformatics training programs to consider adding additional courses in statistics to the curriculum. These courses would not replace those in algorithms and databases but rather extend the requirements. Additional courses will take additional time to complete. This is likely unappealing to some but may be necessary to fully prepare our graduate students for a world of big data.

References

1. "NIH Names Dr. Philip E. Bourne First Associate Director for Data Science". 2013. http://www.nih.gov/news/health/dec2013/od-09.htm [Feb 13 2014].
2. "Obama Administration Unveils 'Big Data' Initiative: Annnounces $200 Million In New R&D Investments". 2012. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf [Feb 13 2014].
3. "IT Data Scientist - Best Jobs in America 2013". 2013. http://money.cnn.com/pf/best-jobs/2013/snapshots/32.html?iid=BestJobs_fl_list [Feb 13 2014].
4. "Data Scientist: The Sexiest Job of the 21st Century - Harvard Business Review". 2012. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ [Feb 13 2014].
5. R.S. Ledley, L.B. Lusted. "Reasoning Foundations of Medical Diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason". Science, 1959. 130(3366): 9–21. 10.1126/science.130.3366.9.
6. V. Dhar. "Data science and prediction". Commun. ACM. ACM, 2013. 56(12): 64–73. 10.1145/2500499.
7. L. Darden. book: Craver, Carl F. and Lindley Darden (2013), In Search of Biological Mechanisms: Discoveries across the Life Sciences. Chicago, IL: University of Chicago Press. 2013.