FOCUS: EDUCATING YOURSELF IN BIOINFORMATICS

Perspectives on an Education in Computational Biology and Medicine

Jill C. Rubinstein

Yale School of Medicine, New Haven, Connecticut

The mainstream application of massively parallel, high-throughput assays in biomedical research has created a demand for scientists educated in Computational Biology and Bioinformatics (CBB†). In response, formalized graduate programs have rapidly evolved over the past decade. Concurrently, there is increasing need for clinicians trained to oversee the responsible translation of CBB research into clinical tools. Physician-scientists with dedicated CBB training can facilitate such translation, positioning themselves at the intersection between computational biomedical research and medicine. This perspective explores key elements of the educational path to such a position, specifically addressing: 1) evolving perceptions of the role of the computational biologist and the impact on training and career opportunities; 2) challenges in and strategies for obtaining the core skill set required of a biomedical researcher in a computational world; and 3) how the combination of CBB with medical training provides a logical foundation for a career in academic medicine and/or biomedical research.

INTRODUCTION

Over the past few decades, high-throughput assays producing large-scale datasets have become mainstream research tools [1-4]. In response to the need for scientists with facility in this multi-dimensional data space, the fields of Computational Biology and Bioinformatics (CBB) have emerged as major players in modern biomedical research [5,6]. Several years have passed since formalized programs in CBB graduated their first PhD students [7-10]. With these scientists finding roles as academic faculty running independent research labs, as principal PIs on major grants, and as

To whom all correspondence should be addressed: Jill Rubinstein, Yale School of Medicine, 367 Cedar St., New Haven, CT 06520; Email: jill.rubinstein@yale.edu.

†Abbreviations: CBB, Computational Biology and Bioinformatics; EMR, electronic medical record.

Keywords: computational biology, bioinformatics, graduate education, MD/PhD
key players within the industrial hierarchy, it is fair to say that the field is firmly established as a legitimate arm of biomedical research.

Despite widespread recognition of the importance of dedicated training in CBB, challenges remain in establishing the specific content of such training in a rapidly evolving field [11,12]. Difficult as this can be for program directors responsible for defining prerequisite and core curricular courses for graduate programs, students also face a degree of uncertainty about how best to ensure their preparedness for the careers that await them. In one sense, this uncertainty reflects the natural and healthy insecurity of the early graduate student who is learning to navigate a new environment, particularly in such a diverse field. Yet, given the youth of the field and the resultant scarcity of program alumni working within any particular CBB specialty, it is understandable that many students would like to identify a core skill set that we feel comprises computational biologists’ most basic common armamentarium. Below, I will describe my version of such a skill set and attempt to pinpoint where in my education I had the greatest opportunity to acquire these skills.

With the decreasing cost and increasing availability of high-dimensional molecular interrogation techniques, the pressure to begin realizing progress not only in the basic sciences but also in the clinical realm is ever growing [13]. The relative paucity of scientists trained to responsibly implement these research techniques and the veritable dearth of active clinicians with experience in this realm presents a very real problem as we increasingly integrate computational biological approaches in basic science, translational research, and clinical applications. Rather than allowing the haphazard implementation of bioinformatics-driven tools in the clinical setting (or, worse yet, allowing for-profit industry to drive the effort), it behooves us to make conscious decisions about the process through which we develop, test, and adopt the use of such tools. In particular, how do we facilitate productive collaboration between CBB researchers and their clinical counterparts, who should be driving these efforts, and what training do such people need?

**A COMPUTATIONAL WHAT NOW? CHANGING PERCEPTIONS AND AWARENESS OF CBB**

In the roughly 10 years since the beginning of my graduate education, there has been a significant shift in the perception of the role of the computational biologist. When I first started, the fields of Computational Biology and Bioinformatics had only recently been recognized as formal disciplines for which organized and directed educational programs should exist. Happily, major research universities across the country had committed to training scientists in CBB, and there were multiple nascent PhD programs from which to choose. Yet, there seemed to remain a degree of skepticism among some traditional basic science researchers regarding the necessity of establishing the field as its own entity. As I toured the major academic centers interviewing for MD/PhD programs, I had the sense that a few of the scientists I encountered regarded the bioinformatician as a member of the support staff whose skill set was needed in the functioning of a modern lab, but who shouldn't aspire to stand alone as an independent researcher.

Although the skepticism I met with on the interview trail did not accurately represent the governing attitude at my chosen institution, the young age of our graduate program could be seen reflected in the confused expression of many who asked about my research interests. CBB as a field of study had not yet been universally recognized. While early bioinformatics research tools such as microarrays were already in widespread use at the time, because they were relatively new and fairly expensive, they remained the domain of a minority group of specialists. Still, while there weren't an endless number, there were ample labs in which to do research rotations and from which to choose a PhD mentor.

My first jolt of recognition that computational biology had hit prime time came a
few years after I started the MD/PhD program. I was sitting in a Department of Surgery Grand Round when a heat map flashed on the screen, and I learned that the presenting surgeon ran a lab that was using microarray data to search for biomarkers for the risk of abdominal aortic aneurysm. Shortly thereafter, while shadowing a general surgeon in the office setting, I learned about Oncotype Dx, a tool for predicting the likelihood of recurrence and the potential benefit of chemotherapy in certain types of breast cancer [14,15]. This was the first time I had seen the clinical implementation of a tool that derived directly from bioinformatics-driven research.

In the past few years, large-scale assays have become such standard tools in biomedical research that I am no longer surprised when presented with high-dimensional data in forums not traditionally associated with CBB. With the increasing availability and decreasing cost of these assays, many scientists are including investigations of this type in their research. As someone invested in CBB as a field, I am nervous that too many scientists are generating mountains of data without first defining clear hypotheses and research questions. At the same time, I am excited by the explosion in opportunities for collaboration. It would seem that those of us with specific training in computational biology are currently in high demand. Having recently completed the residency interview trail, I was struck by the contrasting attitudes with which I was met as a computational biologist now as opposed to during my MD/PhD interviews eight years ago. While some misperceptions still exist about what my research entails, it was clear that most of my interviewers possessed at least a basic understanding of the field and were aware of the value of graduate training in CBB.

WHAT DOES A COMPUTATIONAL BIOLOGIST NEED TO KNOW?

The computational biologist can play a key role in capitalizing on the power of available technologies and in ensuring their responsible application. However, defining the core set of required skills and organizing a curriculum that will prepare scientists for this role is a daunting task. It is imperative that these scientists are properly trained to responsibly design large-scale experiments, analyze immense datasets, and draw biologically and/or clinically meaningful conclusions. Equally important is the ability to ask meaningful questions and communicate effectively with more traditional bench scientists and, perhaps, clinicians.

In my mind, the key elements of my education can be grouped into three broad categories. The most easily identifiable of these is the one that students seem most preoccupied with in the early years. That is, what are the concrete skills I need first to qualify for training in this field and second to function as a computational biologist at the end of my training? The question of prerequisites seems to have become less frequent in recent years with the increasing availability of CBB courses and programs at the undergraduate level, but it is still a difficult one to answer. In the early years of the program, I remember my program director expressing the opinion that it is slightly easier to teach a computer scientist about biology than vice versa, but that a strong background in either was sufficient at least for consideration for our program. These days, students are expected to arrive with a solid foundation in the basics sciences as well as significant programming experience. While these are reasonable requirements for today’s climate, based on my own experience, I would encourage those who aren't sure if they qualify to speak with prospective program directors before they write themselves off.

All too often I hear good students, with genuine interest in the field, dismissing CBB as an option because they lack confidence in their level of computational preparedness. For those who fall into this category, I will share that I was an Architecture major in college. I did complete a master’s degree in Computer Science, but to my chagrin, the few years I spent pursuing that degree had not transformed me into an expert of the type who had been building computers in her parents’ basement since age 9. Still, I
gained the qualifications I needed: I learned to program. Although my relatively limited background did create more work for me at the beginning of training, my skills quickly improved and did so in accord with the specific needs of my research. As for my basic science background, the intensive post-baccalaureate pre-medical program I completed just before beginning graduate school provided a strong foundation for my future studies.

Overall, in thinking about the variety of backgrounds among students in my program, it is clear that there is no one set of prerequisites that constitutes the best preparation. In fact, I consider the varied perspectives arising from our diverse backgrounds to be a great strength of graduate training in CBB. In contrast to this flexibility in specific prerequisites, there are a few concrete computational skills that all students should possess before finishing training. In my mind, the most fundamental toolbox of the computational biologist should include the following minimum competencies:

1. Expertise in a scripting language (ex. Perl, Python)
2. Expertise in a statistical environment (ex. R, Matlab)
3. Facility with database design, management, and use (ex. SQL)
4. Facility with biostatistics
5. Experience in a compiled programming language (ex. C++, Java)

As for how best to attain these skills, it seems dedicated coursework is a necessary first step for most of us. However, in terms of gaining a reasonable comfort level in any of these areas, for me there is no greater motivation than to be confronted with the specific needs that arise in the course of research.

Complementing these hands-on abilities is the second broad category of educational content, perhaps best described as the knowledge base. Examples of this type of core content include sequence alignment approaches, macromolecular simulation, biomedical data modeling, and standardized biomedical ontologies, to name a few. This category gives rise to a language and culture that is specific to CBB. It is the esoteric content, the common ground that allows two computational biologists to speak in shorthand with one another. This is a more difficult concept to define than the specific, nameable skills listed above, but here again the beginnings of this knowledge base can be found in the core CBB coursework.

The coursework should be augmented at first with broad reading of the literature, to be narrowed accordingly as particular research interests arise. Beyond this, I have found it is beneficial to pay deliberate attention to the ways in which researchers communicate with each other, be it in desk-side chats, in lab meetings, or in organizing and delivering talks. During medical school, listening to doctors discuss clinical cases with other doctors was among the most educational ways I could spend my time. It is through such observation, and gradually escalated levels of participation, that I learned to appreciate the clinical nuance and varying priorities that drive patient care for different specialists. I believe this same opportunity exists in graduate education for those who are astute observers.

Finally, there is the educational content I think of as the hidden curriculum. It is both the most abstract and the least specific to any particular field. It includes such abilities as critical thinking, organization, creativity, and scientific discipline. Beyond this list of desirable attributes, here I include self-awareness, communication skills, diplomacy, and academic political savvy. This third category is by far the least tangible, but arguably the most important element of an education. These are exceedingly difficult skills to teach programmatically, but instead seem largely perpetuated by mentorship, role modeling, and institutional culture. In deciding which graduate program to matriculate in, and particularly which lab to join, it is therefore worthwhile to go beyond considerations of specific research strengths and attempt to get a real sense of the governing
culture. In choosing a lab, one might want to look for evidence that students are encouraged to think creatively and rewarded for taking the initiative to pursue their ideas. Part of graduate education is learning to take your place at the table. That is, students should gradually find their voices and understand that they can make valuable contributions to the discussion. The culture of the environment in which we train and the leadership styles to which we are subjected could either promote or delay this process and will inevitably influence the way we carry ourselves when we ascend to the supervisory role. It is therefore critical that we are cognizant of the type of mentorship we are receiving and that we aren’t afraid to seek guidance from those we most respect and wish to emulate.

IS THERE A DOCTOR OF BIOINFORMATICS IN THE HOUSE? INTEGRATING CBB AND MEDICINE

At the beginning of my graduate education, it was explained to me that the nature of MD training is very different from that of PhD training. The medical student is expected to memorize details of pathologies, learn to recognize constellations of signs and symptoms, and apply management paradigms that may or may not be based on a complete understanding of a condition’s pathophysiology. In contrast, the PhD student is encouraged to be more questioning from the outset, read each journal article with a hyper-critical eye, and always attempt to attain an intimate understanding of the underlying mechanisms of the system in study. While there are certainly plenty of times when graduate students accept facts at face value and medical students think critically and creatively about patient care, ultimately the differing priorities of the researcher and the clinician cause each to emphasize the acquisition of different skill sets. It is the goal of MD/PhD training to produce physician-scientists, whose dual educations position them at the boundary between the research and the clinical.

This concept of physician-scientist as the bridge between bedside and bench is far from a new one, but I would argue that there is currently an under-utilized niche for the computational biologist in this dual role. The doctor’s clinical experience provides understanding of the physical manifestation of disease processes and their impact on the patient, as well as insight into the ways our treatments work or don’t work. For those who are paying attention, the clinical arena is a veritable breeding ground for important and interesting research questions, the answers to which can change the way we practice medicine. For the computational biologist, the goal is to devise and implement methods that capitalize on the recent drastically increased availability of massive amounts of data and impressive processing power to address these questions.

One major obstacle computational biologists face in achieving translational progress is the historical lack of standardization in the way data, particularly clinical data, is collected and stored. In the current system, more often than not the clinical data with which we work is entirely divorced from the molecular. While this is partly due to issues of informed consent and patient confidentiality, the missed opportunity we incur by not going through the proper channels to ensure compliance with these important considerations is too great. Beyond issues of confidentiality, perhaps the main barrier here is the relatively slow adoption of the electronic medical record (EMR) into standard practice and the low priority of research functionality in many of the early systems. Maximizing the clinical utility of the EMR while still providing quality research utility is far from a trivial task and is itself an example of an area where computational biologist doctors might make a real impact.

The past decade’s increased adoption of high-throughput technologies has defined an important role for computational biologists in translational research. A few characteristics of CBB make it a very amenable field for dividing time between research and clinical responsibilities. Already accustomed to navigating between the different research cultures of the computationalists, statisti-
cians, and basic scientists, the computational biologist’s experience in bridging communication gaps can be a major asset in facilitating productive collaborations between researchers and clinicians. From a logistical standpoint, the computational biologist has great flexibility in the planning and execution of experiments. The availability of remote access to computational resources means there are fewer constraints in determining exactly when and where the work is carried out. This simple reality greatly eases the burden of balancing time in the lab with the more rigid time constraints imposed by one’s clinical obligations, an important consideration that is easy to overlook while still a student.

CONCLUSIONS AND OUTLOOK

Since its formal beginnings as a field not much more than a decade ago, computational biology has become established as a central discipline in biomedical research [5,6]. While the detailed content of graduate CBB training continues to evolve with the technology, more than a decade of development as an academic field has created an essential common skill set and knowledge base shared by experts in the field. For those wishing to apply their CBB expertise in translational research, it is important to recognize that the climates of MD and PhD training are very different and often emphasize opposing skills. These differing priorities, perceived in the way doctors and researchers think and communicate, create a need for physician-scientists with combined MD/PhD training; while this is true in all areas of biomedical research, recent advances in our technological capability have created a particular need for computational biologists in this role.

There is growing pressure from the health care establishment and also the general public to realize the clinical potential of affordable, large-scale molecular interrogation [13,16,17]. However, the very availability of these high-throughput technologies has perhaps created a false belief that we are further along the path toward individualized medical care than we actually are. I believe the next decade will witness large advances in our molecular-level understanding of complex diseases, which will in turn provide opportunity for significant improvements in our current standards of care. However, there is danger that research efforts driven by those without sufficient understanding of large-scale data and CBB methodology could lead to the development of sub-optimal or even flawed clinical tools. Given the significant pressure from health care institutions, funding agencies, and industry to speed clinical advances with new technologies, there is not only an opportunity but a responsibility for those completing dual training in CBB and medicine to guide the translation of the knowledge gained from high-dimensional assays into clinically informative tools. Existing MD/PhD programs have a strong tradition of training physician-scientists to guide translational research efforts; with the coming of age of graduate education in CBB, I believe we will soon see a steep increase in the number of MD/PhD students choosing this field of research to complement their medical educations and, ultimately, their careers in academic medicine.

REFERENCES

1. Gresham D, Dunham MJ, Botstein D. Comparing whole genomes using DNA microarrays. Nat Rev Genet. 2008;9(4):291-302.
2. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet. 2004;5(5):335-44.
3. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135-45.
4. Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. Genomics. 2009;93(2):105-11.
5. Altman RB. A curriculum for bioinformatics: the time is ripe. Bioinformatics. 1998;14(7):549-50.
6. Azuaje FJ, Heymann M, Ternes AM, Wiebecke-Baldacchino A, Struck D, Moes D, et al. Bioinformatics as a driver, not a passenger, of translational biomedical research: Perspectives from the 6th Benelux bioinformatics conference. J Clin Bioinforma. 2012;2:7.
7. Tan TW, Lim SJ, Khan AM, Ranganathan S. A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the “-omics” era. BMC Genomics. 2009;10(Suppl 3):S36.
8. Altman RB, Klein TE. Biomedical informatics training at Stanford in the 21st century. J Biomed Inform. 2007;40(1):55-8.
9. Gerstein M, Greenbaum D, Cheung K, Miller PL. An interdepartmental Ph.D. program in computational biology and bioinformatics: The Yale perspective. J Biomed Inform. 2007;40(1):73-9.
10. Johnson SB, Friedman RA. Bridging the gap between biological and clinical informatics in a graduate training program. J Biomed Inform. 2007;40(1):59-66.
11. Ranganathan S. Bioinformatics education—perspectives and challenges. PLoS Comput Biol. 2005;1(6):e52.
12. Ranganathan S. Towards a career in bioinformatics. BMC Bioinformatics. 2009;10 (Suppl 15):S1.
13. Butte AJ. Translational bioinformatics: coming of age. J Am Med Inform Assoc. 2008;15(6):709-14.
14. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817-26.
15. Cronin M, Sangli C, Liu ML, Pho M, Dutta D, Nguyen A, et al. Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. Clin Chem. 2007;53(6):1084-91.
16. Zerhouni EA. Translational and clinical science—time for a new vision. N Engl J Med. 2005;353(15):1621-3.
17. Lesko LJ. Personalized medicine: elusive dream or imminent reality? Clin Pharmacol Ther. 2007;81(6):807-16.