Journal of Cellular Biochemistry 80:212–215 (2000)

The Making of a Bioinformatician

Eric E. Schadt

Department of Biomathematics, University of California, Los Angeles, California 90095; Departments of Bioinformatics and Biomathematics, Roche Bioscience, Palo Alto, California 94304

INTRODUCTION

As university research centers, biotech companies, and pharmaceutical companies, among many others, struggle to cope with the sea of biological data being generated by high-throughput genomics technologies, universities are rushing to establish programs to train the next generation of scientists capable of handling and analyzing (mining) these large stores of biological data. These new research strategies and methodologies are fundamentally transforming biology into a more quantitative science that will come to demand the same mathematical, statistical, and data modeling sophistication physicists have known for years [Reichhardt 1999]. As biologists begin to appreciate that it is simply not enough to store terabytes of data (and beyond) without paying particular attention to serious data modeling issues, the demand for scientists capable of modeling these large stores of data and of developing statistical tools to analyze these data will ever increase. In many molecular biology labs around the country, the computer science, mathematical, and statistical skills do not exist to even begin thinking about how data should be stored, let alone how the data should be analyzed. What has become apparent from the early efforts applied to storing biological data is that storing gigabytes of data in flat files or heavily normalized relational databases serves only to archive the data and, perhaps, to allow a small subset of relatively simple questions to be asked of the data. However, what awaits to be discovered in all of these sequence, gene expression, and proteomics data is information that, when associated with other orthogonal biological factors, will elucidate the complexity of living systems. Therefore, the opportunities to devise intelligent data models and to develop methods to mine the data and to then formulate mathematical models that allow proper interpretation of these data, abound as the biological sciences become poised to revolutionize everything from identifying genes associated with complex diseases to rapidly elucidating protein structures, functions, and interactions between other proteins and genes. When I, for the first time, began to appreciate this vision, espoused by leaders in the world of research biology, on what the future of biology held, it became apparent to me that future successes in biology would come to depend on sophisticated mathematical and statistical methodologies, and that my computer science and math training would find utility among the most pressing problems facing biologists; a view that is now widely echoed in many official reports discussing how to meet the challenges facing biological research in the Golden Age of biology [see, e.g., the Biotechnology National Science and Technology Council 1998 report]. To become an active participant in the biological revolution, I had to make the decision to extend my cross-disciplinary training to include biology, so that I could begin to learn a common language that would enable me to work in the sort of collaborative environment I believe will come to define success in future biological research endeavors.

A Bioinformatician Emerges

My academic pursuit of mathematical biology and the computational components I feel comprise an integral part of bioinformatics, was set into motion one afternoon as I decided to take a break from my algebraic topology studies, where I was investigating the use of the five-lemma in establishing general properties of relative homology groups, to attend a seminar on nonlinear dynamics in biology at the Institute of Theoretical Dynamics at UC Davis. I had entered a Ph.D. pure mathematics program with an admittedly naïve plan to develop myself deeply as a pure mathematician, and then to transition into an applied area of science upon completion of my Ph.D. pure mathematics training. I had not been in the pure mathematics program long, however, before realizing I was falling victim to its addictive nature. Pure mathematics is a discipline focused on the pursuit of deep and often very subtle connections that serve to link important mathematical ideas together, and one quickly begins to appreciate the profoundness of the more serious connections (in the context of mathematics) and the beauty of these relationships, despite the fact that they are not usually very practically useful. The sacrifice one has to make for this addiction proved too great for me as I struggled to maintain my motivation to actually use mathematics to model interesting physical phenomenon, and after attending several "biomath" seminars, I began my search for mathematical programs more supportive of the biological interests I had developed through these seminars. While my understanding of many of the current biological problems in genetics, genomics, and proteomics was, at the time, a bit primitive, my strong computer science and applied math training as an undergraduate presented an arsenal of tools I thought could be potentially very useful to the coming biological data revolution. In researching various programs, what became important to me was to be viewed, in the end, as more than a servant to the ideas of researchers in the biological sciences. I wanted to attain a comparable level of training received by biologists, so that I could formulate biological problems of interest on my own, and not only be fluent in the language of biology, but develop the insights and intuitions into many of the pressing biological problems that would enable me to form strong cross-disciplinary collaborations with researchers in biology, mathematics/statistics, and computer science. I sought to be an integral part of the loftiest of pursuits in the life sciences, which included seeking to understand the complexities of living systems.

Choosing the Right Academic Program and Tailoring a Curriculum to Suit My Interests

I was immediately attracted to the biomathematics program at UCLA for its mathematical rigor (alas, I was unwilling to completely give up my mathematical pursuits) and for its requirement that I achieve Ph.D. candidacy in a field of biology, which was a more demanding step other programs I had investigated were not willing to take, but which I thought necessary to ensure I would be able to approach biological problems as a biologist, not only as a mathematician and/or computer scientist. I was immediately drawn to a notable mathematician in the biomathematics department at UCLA, Ken Lange, who had also departed the pure math world for a very successful academic career in mathematical biology, a career that has produced fundamental contributions to many different areas of statistical genetics, including, but not limited to, construction of radiation hybrid maps, analysis of human pedigrees, haplotype reconstruction and phylogeny reconstruction. I saw Professor Lange's deep mathematical training, his solid grasp of biology (genetics), and his push to not only design mathematical models to interpret biological data, but to implement these models in software, making them generally available to researchers all over the world, as a model for me to follow in my pursuit to become a topnotch mathematical biologist. My fascination with the fundamental nature of life drew me almost immediately to genetics and molecular evolution, and I was afforded the opportunity by the biomathematics department to tailor my curriculum to these interests (a flexibility that should not go unnoticed by other academic programs seeking to train students in a highly cross-disciplinary field like bioinformatics).

My decision to pursue a Ph.D. in biomathematics did not come without some huge penalties. My computer science and applied math undergraduate degree provided for no biology or chemistry training, and I had only my dissertation proposal to submit before achieving candidacy in pure mathematics, which did little to ease the rigorous requirements imposed by the biomathematics program at UCLA. Upon entering the biomathematics program, I was inundated with graduate level courses in biomathematics and with a slew of undergraduate biology and chemistry courses. The decision to forgo general chemistry and introductory genetics and to jump head first into organic chemistry/biochemistry and human genetics courses, made for an interesting first year as I struggled to read general chemistry, introductory genetics, and introductory molecular biology texts while pushing to keep up in my human genetics, organic chemistry, biochemistry, advanced graduate probability and biomedical data analysis courses. Only by belonging to a department fully supportive of such a rigorous and diverse curriculum was I able to make it out the other end intact. Truly, one of my fortunes in starting down what I hope to be a successful career in bioinformatics was training in a department that fully appreciated the mathematical, statistical, computational, and biological skills needed to successfully compete in the new age of biology.

Nearing the End, Was the Transition into Computational Biology Worth It?

The end of my academic training currently involves writing a dissertation on advances in maximum likelihood methods for reconstructing evolutionary trees [based on work by Schadt et al. 1998]. Because of the economic hardships that can befall graduate students studying and raising a small family on fellowships providing stipends well below the poverty line, and because many of the high-throughput genomics technologies and very expensive commercial sequence databases were available to me only through private industry, I made the jump early to Roche Bioscience as a research scientist to develop methods to analyze gene expression array data and to provide statistical genetics support in human and mouse genetics studies. My training in computer science, mathematics, and biology were an instant hit at Roche as I quickly adapted to the demands of the pharmaceutical drug discovery process. My progress at Roche has been rapid, and I am now in charge of numerous projects, including gene expression experimental design and analysis methods development, developing approaches to identify disease susceptibility genes using mouse models for common human diseases, developing computational and information management components needed to intelligently store and mine large data stores, and developing methods to reconstruct evolutionary trees. My projects have included developing several academic collaborations with outstanding researchers at UCLA, UCSF, Johns Hopkins, and Harvard (for many of these collaborations, competitive NIH, Biostar, and LSI proposals have been submitted and are awaiting review), and these collaborations, if successful, will result in publications for much of the collaborative work, which will hopefully allow me to remain a competitive candidate for faculty positions, should I choose to pursue that career path. Every experience I had as a graduate student at UCLA has proven valuable to me at Roche, as I strive to meet the ever-growing computational biology demands upon which the pharmaceutical companies are coming to depend.

The path I have followed has not been problem-free, however, nor will I ever be allowed to become content in what I learned through years and years of education, if I have any hope of maintaining a competitive edge in the midst of the biological revolution. Current academic programs, including the biomathematics program I am coming out of, do not do enough to familiarize students with the vast array of electronic biological data that are publicly available to academic researchers, nor do they make an adequate attempt at training scientists to use publicly available genome analysis software, let alone teaching these scientists the theoretical underpinnings upon which the software are based. The current perception among many academic units and industry leaders focusing on bioinformatics seems to be that of placing computer-savvy biologists capable of navigating the various databases of biological information and of using/developing software tools to manipulate and analyze these data, in a position where they are often regarded as second-class citizens with respect to the biological mission of the laboratory to which they belong. Furthermore, those with strong computer science skills are often harnessed for IT/IM support with respect to a lab's bioinformatics efforts. While this trend is natural given the explosion of data in the field and the struggle many labs face in bringing these data under control, it would be shortsighted for any institution to ignore the strong computational and statistical/mathematical components upon which I believe modern biology will come to depend. Current perceptions of the computer-savvy and statistically sophisticated biological researcher must advance beyond those that view these researchers as only providing a service, and instead, come to view them as an invaluable resource upon which modern biological research depends. In the end, the ability to develop algorithms to extract biological information from sequence data, expression data, and other large-scale genomics data will be as important as any of the other more classic components necessary for successful biological research.

SUMMARY

My research plans for the future are ambitious and will demand developing collaborations with experts across a number of departments, including, but not limited to, genetics, molecular biology, molecular evolution, mathematics, statistics, and computer science/informatics. Developing sophisticated algorithms to mine the sea of data coming out of the large-scale genomics efforts, developing ways to visualize these data, and mathematically modeling the underlying biological processes giving rise to these data, will help elucidate our understanding of the complex processes of living systems. I believe the strong cross-disciplinary collaborations necessary in carrying these projects forward will come to define what separates top-notch life sciences research institutions from all of the rest. Furthermore, the future successes of these types of projects will require training the next generation of scientists from a variety of academic departments, to speak a common language that will enable them to develop successful collaborations. Given the successes I have already been able to realize in my young career as a computational biologist, I consider myself extremely fortunate to have had the opportunity to develop the necessary cross-disciplinary skills through the biomathematics department at UCLA, and am confident I will continue to develop my bioinformatics skills and make significant contributions to life sciences research.

REFERENCES

Biotechnology National Science and Technology Council. 1998. Bioinformatics in the twenty-first century. Bioinformatics Workshop, March 3–4.

Reichhardt T. 1999. It's sink or swim as a tidal wave of data approaches. Nature 399:517–520.

Schadt E, Sinsheimer J, Lange K. 1998. Computational advances in maximum likelihood methods for molecular phylogeny. Genome Research 8:222–233.