Schatz *Genome Biology* 2012, **13**:177
http://genomebiology.com/2012/13/11/177

Genome Biology

COMMENT

Computational thinking in the era of big data biology

Michael C Schatz*

It is fall again, and another class of students has arrived in the Watson School of Biological Sciences at Cold Spring Harbor Laboratory (CSHL). Building on the lab's 100-year history as a leading center for research and education, the Watson School was established in 1998 as a graduate program in biology with a focus on molecular, cellular and structural biology, and neuroscience, cancer, plant biology and genetics. All students in the program complete the same courses, centered around these research topics, with an emphasis on the principles of scientific reasoning and logic, as well as the importance of ethics and effective communication. Three years ago the curriculum was expanded to include a new course on quantitative biology (QB) and I, along with my co-instructor Mickey Atwal and other members of the QB program, have been teaching it ever since.

Quantitative biology is a rather broad and sometimes ill-defined topic of biology - the common theme is not a particular organism or disease or cellular process, but the style of analysis that is applied. It draws on techniques from mathematics, statistics, computer science, physics and other quantitative fields to develop methodologies and answer biological questions. For me, I was trained as a computer scientist and now I apply my skills to answer questions about human genetics and plant biology. Within human genetics I study the origins and dynamics of human diseases such as autism and cancer, and within plant biology I assemble the genomes and transcriptomes of important species and then quantitatively model the molecular basis of their development. These biological questions require sophisticated computational and quantitative systems to answer, and especially to transform or model the raw sequences and phenotypic data into an understanding of how the underlying biology operates.

It should be noted that quantitative techniques have played a critical role in biology and genetics since their earliest days. For example, in the mid-1800s, they formed the basis for Mendel to derive the fundamental laws of heredity by analyzing the transmission of traits among his pea plants, and in the mid-1900s for Luria and Delbruck to solidify the role of selection in evolution by mathematically modeling the emergence of bacteriophage resistance in *Escherichia coli*. Recently, quantitative biology has been escalating in importance because of the explosion of biological data brought on by the dramatic improvements to biotechnology and biological sensors. So much so that these technologies have radically changed the types of questions that we can even ask. Out of shear cost and complexity, just a few years ago it would have been outrageous to propose to sequence the genomes of many hundreds or thousands of people to find out what is unique in the genomes of children with autism compared with their siblings, for example, but today, thanks to million-fold improvements in the cost and throughput of DNA sequencing, we ask and answer these types of questions on a regular basis.

The achievements of 'big data biology' require integration of skills across several fields, many of which have not been a part of a traditional biology education. For example, many disease studies begin with traditional molecular skills to prepare and sequence the samples, but then the analysis becomes computational and quantitative in order to align the reads, detect the variations, and recognize functionally important mutations from the backdrop of normal human variation. Given the vast opportunities in analyzing (and mis-analyzing) high-throughput sequences, networks and other -omics data, it is certain that the role of quantitative analysis in biology will only grow in the future. As such, today's students need to be trained to properly use and understand these instruments of modern biology. I am not alone in having this sentiment, as evident by the growing number of quantitative and computational biology departments around the world. Indeed, the current Howard Hughes Medical Institute report on 'Scientific Foundations for Future Physicians' recommends incoming medical students should be well versed in quantitative techniques, including the core algorithms and statistics for interpreting sequence data [1].

*Correspondence: mschatz@cshl.edu
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

Everyone on campus agrees that the new course is important for enriching the students' education, but the new course also brings a tension that I am sure is not unique to CSHL. The tension centers on how much quantitative education we should expect the students to master, especially since time spent on those topics will probably require a corresponding reduction in other, more traditional, topics. To complicate the situation, the students have a rather mixed background. Some of the students have undergraduate degrees in computer science or applied mathematics, but more commonly they have only had basic training with little programming or advanced quantitative experience. As instructors, we are challenged to structure the course to make it accessible for the quantitative novices while keeping the more experienced students engaged.

The QB course we have developed covers a selection of what we hope are the most important quantitative and computational topics for the students to learn. This includes the basic principles of probability and statistics, information theory, population genetics, chemical kinetics and molecular biophysics, and biological sequence analysis, and an introduction to computer science and programming. The diversity of topics is aligned with the diversity of research across the QB program, and draws on the diverse backgrounds and expertise of the faculty. This is significant because not only do we apply different quantitative methods and focus on different biological systems, but we also have somewhat different approaches to problem solving in general. The computer scientists generally consider different algorithmic and information processing approaches, while physicists might also consider the bio-molecular properties involved or use approaches from statistical mechanics. The lines are often blurry between the approaches, and we all strive to connect the different quantitative techniques to interesting biological questions.

The lecture I am most proud of I call 'Sequence Alignment and Computational Thinking', which attempts to introduce the key ideas of biological sequence alignment along with some of the key ideas from computer science [2]. During the lecture, we discuss the seemingly simple question of how to find 'GATTACA' in the human genome, starting from the very basics up to the state-of-the-art approaches. I find that having this concrete problem to motivate the lecture makes the otherwise abstract ideas much more accessible, and then I can use it as a launch pad into the deeper computational ideas by gradually increasing the level of abstraction and complexity.

The lecture starts with the most basic approach of brute-force searching for GATTACA at every possible position in the genome, then discussing why this approach is too slow to be useful, and then becoming more sophisticated to consider the power of suffix arrays and the Burrows-Wheeler transform (BWT). I strive to keep the lecture as accessible as possible by describing the algorithms and data structures first through metaphors before analyzing their full complexity. For example, I metaphorically describe the suffix array as a 'phonebook of the genome', and how searching for GATTACA can be solved by playing the 'hi-lo game' of binary search. When discussing the BWT, I show how the Last-First property implicitly encodes the suffix array, but I also walk through an example using a simplified representation to highlight the key ideas of how the program Bowtie uses backtracking of the BWT to rapidly find inexact alignments.

My hope is that through these examples the students will learn how to use the program Bowtie in their research, but more importantly I hope to empower the students to understand the key concepts of the Bowtie algorithm. Read mapping programs may come and go, and some of the students may pursue research topics that do not use read mapping at all, but the fundamental concepts behind sorting and searching are universal to all sciences, as are the techniques of abstraction, automation, algorithm, recursion and data structure. These basic problem-solving ideas are often described as the key principles of 'Computational Thinking', championed by Jeannette Wing and others as a universal skill for all scientists [3]. Advocates for computational thinking argue that because these techniques empower people to solve problems in any field of study that would otherwise be too difficult or too complex to answer, they are as fundamental and important to science as basic arithmetic. Also like basic arithmetic, there is no substitute for learning than with practice, and we challenge the students with several homework exercises and exams throughout the course. For these I find it useful to offer optional advanced problems so that more experienced students will be challenged beyond what is reasonable for the novices.

This year we added a new aspect to the QB course and thrust the incoming students into the digital and quantitative realm by pushing them through a 2-day intensive 'bootcamp' during their first week on campus. The bootcamp was designed to be very hands-on, and introduce the students to the key resources and ideas that would be needed for the QB course, and also for solving quantitative and computational problems in genetics, neuroscience, and the rest of the Watson school curriculum. One of the first activities was to strip my desktop computer down to the bare components to introduce the students first-hand to the CPU, RAM and hard drive that they must learn to love and respect during their PhD program. We then went through several exercises together analyzing biological data as a way to introduce

them to working with UNIX, programming in Python and Matlab, and several sequence analysis programs, including in the Galaxy environment [4]. The hands-on aspects of the bootcamp were universally enjoyed, as they showed the novices the steps to solving several interesting problems, and the more experienced students could assist the novices to keep everyone engaged. The Galaxy exercises were so successful I am strongly considering teaching more Galaxy and less Unix next year to ease the introduction to computational sequence analysis for the novices. As an aside, hands-on challenge problems are also a great way to enrich participation at conferences, and I hope to see the tradition of the informatics challenge at 'Beyond the Genome' continue on and grow [5].

The techniques of computational and quantitative biology are as critical or more critical to biology as PCR or molecular cloning. We have a responsibility to teach the students the principles of these techniques and how to properly operate the tools they will use to answer their research questions. Just as it would be naïve to think that after one semester of Italian or Chinese one would be fluent speaking a foreign language, it is naïve to think that after one semester one could become a master programmer or quantitative expert. As an instructor, my goals in the QB course are to draw out the students' intuitions of how to think computationally to solve problems and provide them with some basic tools for tackling those problems. This is reinforced through exercises and exams that connect the methods to biological questions. While teaching, I try to emulate the teachers I have had throughout my education and show off the beautiful elegance hidden in the computational world. I have been very fortunate that many friends, colleagues and teachers have been willing to share their resources with me so that I can often present the core ideas directly from the inventors of those algorithms. I return the favor now and offer my lectures on my website for anyone that may find them useful [6].

Before I sign off, I'd like to thank Justin Kinney and Zach Lippman at CSHL for reviewing the draft of this essay, and James Taylor at Emory University for his helpful discussions. I would also like to thank all of the teachers and mentors that have challenged and inspired me along the way.

Abbreviations

BWT, Burrows-Wheeler transform; CSHL, Cold Spring Harbor Laboratory; QB, quantitative biology.

Published: 29 November 2012

References

1. AAMC-HHMI: **Scientific Foundations for Future Physicians** [http://www.hhmi.org/grants/pdf/08-209_AAMC-HHMI_report.pdf]
2. Schatz M: **Sequence Alignment and Computational Thinking** [http://schatzlab.cshl.edu/teaching/2012/QB%20Bootcamp%202%20-%20Sequence%20Analysis.pdf]
3. Wing JM: **Computational Thinking** [http://www.cs.cmu.edu/~wing/publications/Wing06.pdf]
4. Goecks J, Nekrutenko A, Taylor J, Galaxy Team: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**:R86
5. **Beyond the Genome 2012.** Conference held at Harvard Medical School, Boston, USA, 27-29 September 2012. [http://beyond-the-genome.com]
6. Schatz M: **Schatz Lab** [http://schatzlab.cshl.edu/teaching/]

doi:10.1186/gb-2012-13-11-177
Cite this article as: Schatz MC: **Computational thinking in the era of big data biology.** *Genome Biology* 2012, **13**:177.