

Teaching Bioinformatics: Needs and Challenges

Introduction

Bioinformatics is a new discipline that addresses the need to manage and interpret the data that in the past decade was massively generated by genomic research. This discipline represents the convergence of genomics, biotechnology and information technology, and encompasses analysis and interpretation of data, modeling of biological phenomena, and development of algorithms and statistics. I would like to discuss here the needs and challenges of teaching bioinformatics, with an ultimate objective of laying the grounds for the development of a *white paper* representing a consensus of thought in the subject. There are many aspects that make this a particularly difficult task. Bioinformatics is by nature a cross-disciplinary field that began in the 1960s with the efforts of Margaret O. Dayhoff, Walter M. Fitch, Russell F. Doolittle and others (e.g., [1]) and has matured into a fully developed discipline [2]. However, bioinformatics is wide-encompassing and is therefore difficult to define. For many, including myself, it is still a nebulous term that encompasses molecular evolution, biological modeling, biophysics, and systems biology. For others, it is plainly computational science applied to a biological system. I will discuss these different views in detail. Bioinformatics is also a thriving field that is currently in the forefront of science and technology. Our society is investing heavily in the acquisition, transfer and exploitation of data, and bioinformatics is at the center stage of activities that focus on the living world. It is currently a hot commodity, and students in bioinformatics will benefit from employment demand in government, the private sector, and academia.

Defining the discipline

With the advent of computers, humans have become 'data gatherers', measuring every aspect of our life with inferences derived from these activities. In this new culture, everything can and will become data (from internet traffic and consumer taste to the mapping of galaxies or human behavior). Everything can be measured (in pixels, Hertz, nucleotide bases, etc), turned into collections of numbers that can be stored (generally in bytes of information), archived in databases, disseminated (through cable or wireless conduits), and analyzed. We are expecting giant pay-offs from our data: proactive control of our world (from earthquakes and disease to finance and social stability), and clear understanding of chemical, biological and cosmological processes. Ultimately, we expect a better life. Unfortunately, data brings clutter and noise and its interpretation cannot keep pace with its accumulation. One problem with data is its multi-dimensionality and how to uncover underlying signal (patterns) in the most parsimonious way (generally using nonlinear approaches [3-5]). Another problem relates to what we do with the data. Scientific discovery is driven by falsifiability and imagination [6] and not by purely logical processes that turn observations into understanding. Data will not generate knowledge if we use inductive principles.

The gathering, archival, dissemination, modeling, and analysis of biological data falls within a relatively young field of scientific inquiry, currently known as 'bioinformatics', 'computational biology', 'biomolecular informatics', or 'computational molecular biology'. Some terms are more restrictive than others and some also refer to the use of biological macromolecules as computing devices (e.g., computational molecular biology). I have

chosen to refer to this data-driven field as bioinformatics. Bioinformatics was spurred by wide accessibility of computers with increased compute power and by the advent of *genomics*. Genomics made it possible to acquire nucleic acid sequence and structural information from a wide range of genomes at an unprecedented pace and made this information accessible to further analysis and experimentation. For example, sequences were matched to those coding for globular proteins of known structure (defined by crystallography) and were used in high-throughput combinatorial approaches (such as DNA microarrays) to study patterns of gene expression. Inferences from sequences and biochemical data were used to construct metabolic networks. These activities have generated terabytes of data that are now being analyzed with computer, statistical, and machine learning techniques. The sheer number of sequences and information derived from these endeavors has given the false impression that imagination and hypothesis do not play a role in acquisition of biological knowledge. However, bioinformatics becomes only a science when fueled by hypothesis-driven research and within the context of the complex and ever-changing living world [7]. Devoid of inductive attributes, bioinformatics can offer powerful ways to understand life's complex behavior and this message needs to be conveyed clearly in teaching.

The science that relates to bioinformatics has many components. It usually relates to biological molecules and therefore requires knowledge in the fields of biochemistry, molecular biology, molecular evolution, thermodynamics, biophysics, molecular engineering, and statistical mechanics, to name a few. It requires the use of computer science, mathematical, and statistical principles. Experimentation is usually *in silico* but *in vitro* experiments are often part of bioinformatics research (e.g. the study of *in vitro* RNA evolution). One illustrative example is the folding of RNA into a spatial structure. RNA folding is believed to follow a relatively simple biophysically grounded model in which a genotype (heritable repository of information) is mapped into a phenotype (the physical, organizational and behavioral manifestation of life) and expresses a rich statistical structure that includes neutral networks and congruent changes in genetic variability and molecular plasticity [8]. This biophysical model connects with the real world and acts as a geometric, kinetic and thermodynamic scaffold for the actual three-dimensional (3D) structure of RNA. Schultes and Bartel [9] recently demonstrated in an elegant *in vitro* experiment the existence of neutral paths in RNA sequence space. They constructed a ribozyme sequence that spanned two evolutionarily unrelated RNA molecules, one with self-cleaving and the other with self-ligating activities, and were able to trace back the mutations in paths through neutral space. This impressive result shows that there can be convergence between theory and experimental science as it relates to the characterization of a macromolecular universe. Bioinformatics is therefore in the cross roads of experimental and theoretical science. Bioinformatics is not only about modeling or data 'mining', it is about understanding the molecular world that fuels life from evolutionary and mechanistic perspectives. It is truly inter-disciplinary and is changing. Much like biotechnology and genomics, bioinformatics is moving from applied to basic science, from developing tools to developing hypotheses.

There are clearly two unifying themes in bioinformatics: *evolution* and *complexity*. One adds the dynamic-temporal dimension to biology. The other adds holistic views on how biological systems interact and operate. These two themes are not mutually exclusive; in fact, they complement each other under the wide umbrella of *genomics*. I will return to these two unifying themes later in my discussion.

Defining teaching needs and a clientele

Bioinformatics represents the cross roads of many areas of expertise. Consequently, teaching bioinformatics appropriately constitutes a considerable challenge. It is particularly difficult because audiences, teaching needs, and teaching objectives can vary considerably. To meet the needs of computer scientists, basic aspects of the chemistry of life need to be offered, including an introduction to procedures, challenges and scientific thought behind molecular biology, genomics, and biotechnology. To meet the needs of biologists, basic skills and principles in computer science and information technology need to be offered, including the development and application of programming skills. While many aspects have to be adequately addressed (e.g. biochemical or statistical principles), the background of the audience will be broad and will demand coverage of certain subject matter necessarily at an elementary level. Students in other areas of inquiry, such as biophysics, mathematics, and statistics, may also be drawn to the new field. This adds to the complexity of how to deliver courses that address a highly diverse audience (in interests and knowledge) and at the same time are effective and challenging. Moreover, different academic settings impose varying requirements. For example, students in a Crop Sciences Department will benefit immensely from courses in bioinformatics, especially if their careers target sectors that are actively pursuing bioinformatic R&D endeavors (e.g. ag-biotech, agrochemical, biomaterials, and nutraceutical sectors). These courses need to deliver concepts and hands-on experience at a fast pace and in the course of one or few semesters. The coverage of subject matter needs to be comprehensive but elemental. In contrast, students in a Computer Sciences Department may need courses that are fine-tuned and tailored to accommodate for example statistical and machine-learning aspects of bioinformatics. There is no need for a fast pace and hands-on approach right away, but the course will have to de-emphasize the biological component. How do we accomplish this in an effective way? How can individual courses interface with more elaborate teaching programs in bioinformatics?

In academia, a subject needs a clientele. Luckily, bioinformatics is currently a trendy subject and many in different fields are drawn to this new area of science. Unfortunately, the audience will have widely different backgrounds and widely different expectations. As mentioned earlier, there will be students from diverse backgrounds looking for different challenges. For many, bioinformatics is a venue to develop challenging computer applications, to develop comprehensive and complex databases, to develop the most sophisticated algorithm, or to generate a model that covers the most variables and approaches reality. For others, bioinformatics is a set of tools that can help them progress in their individual areas of interest in biology, complementing perhaps the study of their favorite gene or biological process or helping them achieve a broader view. Yet for others, it represents a competitive edge that will further their careers. Different interests will result in different levels of motivation and different expectations. There may be many that will want to focus their studies exclusively in bioinformatics, seeking for teaching programs that provide deep coverage of subject matter. In contrast, those that see bioinformatics as a complementary activity will only expect broad coverage. These students would want to have rudiments in all aspects of bioinformatics, but without details. When needed and as they progress in their professional careers, they will explore those aspects in bioinformatics that will be useful for them.

Defining a teaching strategy

Many institutions in the academic transect are focusing on research and teaching programs in bioinformatics. Usually this comes as part of 'post-genomic' initiatives that seek to develop new trends in future science. With teaching, the approach can be at different levels. In some cases a major in bioinformatics is justified when a critical number of faculty in the subject can be reached on campus. This is generally difficult because the expertise in the field is currently being developed, and most self-made and first-generation bioinformaticians are generally captured by government and industry. In other cases, programs in biology that 'emphasize' computational biology and use the expertise of resident faculty can be very effective, but programs need to be carefully tuned. Similarly, individual courses can be added to well establish curricula and this can be done effectively in a variety of programs across campus. These last two strategies accomplish different goals, but both have the common theme of strengthening individual teaching programs.

A completely different strategy is to establish a *skeletal set of courses* capable of providing a general background in all aspects of bioinformatics. This set of courses can then link to other that are more specialized and are housed in different departments across campus. This approach has its advantages. A broad background in bioinformatics from a start can provide students with a better understanding of what are the options to pursue in the field. Those with inclinations towards the *evolution*-unifying theme can then opt to strengthen coursework that relate to genomics (e.g., in functional, structural, comparative genomics) and evolutionary biology (e.g., phylogenomics, population biology). Those with inclinations towards the *complexity*-unifying theme can then opt for alternatives that stress for example biochemistry and biophysics (e.g., biomolecular engineering, nanobiotechnology, metabolomics, integrative biology or systems biology). Those that are pursuing aspects of bioinformatics from a computer science or statistical perspective will receive basic biology instruction that is tailored to bioinformatics science, while those in the biological sciences will expand their knowledge base and will be primed to pursue more focused and specialized courses in statistics, biophysics, computer science, or biomolecular engineering. It is my opinion this is a viable strategy that can save teaching resources and can be quite effective.

The objectives of the skeletal coursework should address aspects as varied as comparative and functional genomics, molecular evolution and phylogenetic reconstruction, cladistic, phenetic and phylogenomic analysis, alignment theory, biomolecular structure and its prediction, molecular modeling and thermodynamics, structural genomics and definition of molecular universes, information theory and machine learning, networks and complex behavior. Basic concepts in computer science, statistics, and probability theory need to be adequately introduced. Fitting so many aspects in two or three basic courses that would define the *evolution*- and *complexity*-unifying themes may be quite a challenge, perhaps illustrated in a hypothetical introductory course that could be designed to fit perhaps the 200 and 300 level boundary. I will label this course *Genomes and Bioinformatics*. For lack of other models, I will tailor this course to match the level of students majoring in Crop or Animal Sciences. The backbone of this course is composed of 10 sections, each section requiring on average 3-4 contact hours. Sections include quite catchy titles: (1) *Biology in the computer age*, (2) *The molecules of life*, (3) *Molecular and genome biology*, (4) *Introduction to the bioinformatics workstation*, (5) *Databases*, (6) *Bioinformatic tools*, (7) *Molecular evolution, phylogenetic analysis, and phylogenomics*, (8) *The Bayesian*

probabilistic framework, (9) *Bioinformatics and systems biology*, and (10) *From concepts to working code*. A single book cannot be used as reference, so the course in its preliminary offerings will have to use a combination of books and hand out materials. The course is advertised as an introductory course intended for students interested in biology, genomics and informatics. It includes some hands on exercises (DNA sequence alignment, RNA folding, phylogenetic analysis, and PERL scripting), a project on networks (in which each student studies a system and tests if it exhibits network behavior), and review sessions. While the title of the course *Genomes and Bioinformatics* appears to provide foundations for genomics from an informational perspective, the course is set up broadly and attempts to cover all aspects of informatics, but under the wide umbrella of *genomics*. It is therefore valuable and complements courses that stress molecular genetics, and functional and comparative genomics. It will be however difficult to introduce so many basic concepts spanning so many fields in such a short time frame.

The introductory and tentative *Genomes and Bioinformatics* could be followed by two or more specialized courses of a series (perhaps with tentative titles *Bioinformatics: molecular evolution*, *Bioinformatics: statistics*, and *Bioinformatics: molecular structure*), and/or immediately by courses provided by resident faculty in widely different departments. These courses can also strengthen bioinformatic 'emphasis' programs or other initiatives that focus for example on biomolecular engineering. It could also represent a "priming" scheme for the design of a more wide-encompassing bioinformatics initiative on campus.

Conclusion

I have here presented my personal views on how to rationalize teaching needs and challenges in the emerging area of bioinformatics. I hope this exercise will promote constructive discussion on how we can be effective and competitive in the teaching of this new discipline. Bioinformatics is evolving and its current standing resembles *genomics* a decade ago. So we are to expect radical developments and changes in this discipline in the years to come.

References

1. Dayhoff MO (1969) Computer analysis of protein evolution. *Sci Am* 221:86-95.
2. Boguski MS (1998) *Trends Guide to Bioinformatics*, Elsevier Science, pp. 1-3.
3. Strogatz SH (2001) Exploring complex networks. *Nature* 410:268-276.
4. JB Tenenbaum, V de Silva & JC Langford (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319-2323.
5. Donoho DL, Elad M (2003) Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proc Natl Acad Sci USA* 100:2197-2202.
6. Popper KR (1969) *The logic of scientific discovery*. Hutchison, London.
7. Allen J (2001) Bioinformatics and discovery: induction beckons again. *BioEssays* 23:104-107.
8. Fontana W (2002) Modelling 'evo-devo' with RNA. *BioEssays* 24:1164-1177
9. Schultes EA, Bartel DP (2000) One sequence, two ribozymes: implication for the emergence of new ribozyme folds. *Science* 289:448-452

Gustavo Caetano-Anollés
Associate Professor of Bioinformatics
Department of Crop Sciences
E-mail: gca@uiuc.edu