

Preface

In putting together this textbook, we hope that students from a range of fields – including biology, computer science, engineering, physics, mathematics, and statistics – benefit by having a convenient starting point for learning most of the core concepts and many useful practical skills in the field of bioinformatics, also known as computational biology.

Students interested in bioinformatics often ask about how should they acquire training in such an interdisciplinary field as this one. In an ideal world, students would become experts in all the fields mentioned above, but this is actually not necessary and realistically too much to ask. All that is required is to combine their scientific interests with a foundation in biology and any single quantitative field of their choosing. While the most common combination is to mix biology with computer science, incredible discoveries have been made through finding creative intersections with any number of quantitative fields. Indeed, many of these quantitative fields typically overlap a great deal, especially given their foundational use of mathematics and computer programming. These natural relationships between fields provide the foundation for integrating diverse expertise and insights, especially when in the context of performing bioinformatic analyses.

While bioinformatics is often considered an independent subfield of biology, it is likely that the next generation of biologists will not consider bioinformatics as being separate and will instead consider gaining bioinformatics and data science skills as naturally as they learn how to use a pipette. They will learn how to program a computer, likely starting in elementary school. Other data science knowledge areas, such as math, statistics, machine learning, data processing, and data visualization will also be part of any core curriculum. Indeed, the children of one of the editors recently learned how to construct bar plots and other data charts in kindergarten! The same editor is teaching programming in R (an important data science programming language) to all incoming biology graduate students at his university starting this year.

As bioinformatics and data science become more naturally integrated in biology, it is worth noting that these fields actively espouse a culture of open science. This culture is motivated by thinking about why we do science in the first place. We may be curious or like problem solving. We could also be motivated by the benefits to humanity that scientific advances bring, such as tangible health and economic benefits. Whatever the motivating factor, it is clear that the most efficient way to solve hard problems is to work together as a team, in a complementary fashion and without duplication of effort. The only way to make sure this works effectively is to efficiently share knowledge and coordinate work across disciplines and research groups. Presenting scientific results in a reproducible way, such as freely sharing the code and data underlying the results, is also critical. Fortunately, there are an increasing number of resources that can help facilitate these goals, including the bioRxiv preprint server, where papers can be shared before the very long process of peer review is completed; GitHub, for sharing computer code; and data science notebook technology that helps combine code, figures, and text in a way that makes it easier to share reproducible and reusable results.

We hope this textbook helps catalyze this transition of biology to a quantitative, data science-intensive field. As biological research advances become ever more built on interdisciplinary, open, and team science, progress will dramatically speed up, laying the groundwork for fantastic new discoveries in the future.

We also deeply thank all of the chapter authors for contributing their knowledge and time to help the many future readers of this book learn how to apply the myriad bioinformatic techniques covered within these pages to their own research questions.

Andreas D. Baxevas

Gary D. Bader

David S. Wishart