Preface

Bioinformatics is the management and analysis of data for the life sciences. As such, it is inherently interdisciplinary, drawing on techniques from Computer Science, Statistics, and Mathematics and bringing them to bear on problems in Biology. Moreover, its subject matter is as broad as Biology itself. Users and developers of Bioinformatics methods come from all of these fields. Molecular biologists are some of the major users of Bioinformatics, but its techniques are applicable across a range of life sciences. Other users include geneticists, microbiologists, biochemists, plant and agricultural scientists, medical researchers, and evolution researchers.

The ongoing exponential expansion of data for the life sciences is both the major challenge and the *raison d'être* for twenty-first century Bioinformatics. To give one example among many, the completion and success of the human genome sequencing project, far from being the end of the sequencing era, motivated a proliferation of new sequencing projects. And it is not only the quantity of data that is expanding; new types of biological data continue to be introduced as a result of technological development and a growing understanding of biological systems.

Bioinformatics describes a selection of methods from across this vast and expanding discipline. The methods are some of the most useful and widely applicable in the field. Most users and developers of Bioinformatics methods will find something of value to their own specialties here, and will benefit from the knowledge and experience of its 86 contributing authors. Developers will find them useful as components of larger methods, and as sources of inspiration for new methods. Volume I, Section IV in particular is aimed at developers; it describes some of the "meta-methods"—widely applicable mathematical and computational methods that inform and lie behind other more specialized methods—that have been successfully used by bioinformaticians. For users of Bioinformatics, this book provides methods that can be applied as is, or with minor variations to many specific problems. The Notes section in each chapter provides valuable insights into important variations and when to use them. It also discusses problems that can arise and how to fix them. This work is also intended to serve as an entry point for those who are just beginning to discover and use methods in Bioinformatics. As such, this book is also intended for students and early career researchers.

As with other volumes in the Methods in Molecular Biology™ series, the intention of this book is to provide the kind of detailed description and implementation advice that is crucial for getting optimal results out of any given method, yet which often is not incorporated into journal publications. Thus, this series provides a forum for the communication of accumulated practical experience.

The work is divided into two volumes, with data, sequence analysis, and evolution the subjects of the first volume, and structure, function, and application the subjects of the second. The second volume also presents a number of "meta-methods": techniques that will be of particular interest to developers of bioinformatic methods and tools.

Within Volume I, Section I deals with data and databases. It contains chapters on a selection of methods involving the generation and organization of data, including

sequence data, RNA and protein structures, microarray expression data, and functional annotations.

Section II presents a selection of methods in sequence analysis, beginning with multiple sequence alignment. Most of the chapters in this section deal with methods for discovering the functional components of genomes, whether genes, alternative splice sites, non-coding RNAs, or regulatory motifs.

Section III presents several of the most useful and interesting methods in phylogenetics and evolution. The wide variety of topics treated in this section is indicative of the breadth of evolution research. It includes chapters on some of the most basic issues in phylogenetics: modelling of evolution and inferring trees. It also includes chapters on drawing inferences about various kinds of ancestral states, systems, and events, including gene order, recombination events and genome rearrangements, ancestral interaction networks, lateral gene transfers, and patterns of migration. It concludes with a chapter discussing some of the achievements and challenges of algorithm development in phylogenetics.

In Volume II, Section I, some methods pertinent to the prediction of protein and RNA structures are presented. Methods for the analysis and classification of structures are also discussed.

Methods for inferring the function of previously identified genomic elements (chiefly protein-coding genes) are presented in Volume II, Section II. This is another very diverse subject area, and the variety of methods presented reflects this. Some well-known techniques for identifying function, based on homology, "Rosetta stone" genes, gene neighbors, phylogenetic profiling, and phylogenetic shadowing are discussed, alongside methods for identifying regulatory sequences, patterns of expression, and participation in complexes. The section concludes with a discussion of a technique for integrating multiple data types to increase the confidence with which functional predictions can be made. This section, taken as a whole, highlights the opportunities for development in the area of functional inference.

Some medical applications, chiefly diagnostics and drug discovery, are described in Volume II, Section III. The importance of microarray expression data as a diagnostic tool is a theme of this section, as is the danger of over-interpreting such data. The case study presented in the final chapter highlights the need for computational diagnostics to be biologically informed.

The final section presents just a few of the "meta-methods" that developers of Bioinformatics methods have found useful. For the purpose of designing algorithms, it is as important for bioinformaticians to be aware of the concept of *fixed parameter tractability* as it is for them to understand NP-completeness, since these concepts often determine the types of algorithms appropriate to a particular problem. *Clustering* is a ubiquitous problem in Bioinformatics, as is the need to *visualize* data. The need to interact with massive data bases and multiple software entities makes the development of *computational pipelines* an important issue for many bioinformaticians. Finally, the chapter on *text mining* discusses techniques for addressing the special problems of interacting with and extracting information from the vast biological literature.

Jonathan M. Keith

http://www.springer.com/978-1-58829-707-5

Bioinformatics
Volume I: Data, Sequence Analysis and Evolution
Keith, J.M. (Ed.)
2008, XII, 562 p. 136 illus., 3 illus. in color., Hardcover
ISBN: 978-1-58829-707-5
A product of Humana Press