Bioinformatics

B

Mark A. Ragan
The University of Queensland, Australia

INTRODUCTION

Over the past 25 years, bioinformatics has emerged as new discipline at the interface of molecular bioscience with mathematics, computer science and information technology. Bioinformatics is driven by data arising from new high-throughput technologies in molecular bioscience including DNA and genome sequencing, gene expression analysis, protein and RNA structure characterisation, and bio-imaging. To enable biological discovery, bioinformatics draws on and extends technologies for data capture, management, integration and mining, computing, and communication technology including the Internet. The rise of genomics, from the initial bacterial and model-organism projects to the Human Genome Projects and the thousands of genome projects that have followed, has been a key driver for bioinformatics, which in turn enabled these projects to be completed and their results applied. Genomics, however, was never an end unto itself, but rather was intended to enable the understanding of complex biological systems. Bioinformatics continues to evolve in support of its constituent domains and, increasingly, their integration into genome-scale molecular systems biology.

This article presents bioinformatics first from the perspective of computer science and IT, then from the perspective of bioscience. In practice, these perspectives often merge, making bioinformatics a rich, vibrant area of multidisciplinary research and application.

BACKGROUND

The term *bioinformatics* was introduced in 1970 in reference to the study of informatic processes in biological systems (Hogeweg, 2011). In this original usage, *bioinformatics* encompassed "how living systems gather, process, store and use information" (Nurse, 2008). Never widely adopted, this usage was superseded in the late 1980s when bioinformatics, as presently understood, emerged as a new field at the interface of molecular bioscience with computer science and information technology (Dickson, 1987). Today bioinformatics builds on mathematics, statistics and algorithmics, and finds applications across the biosciences particularly in genomics, proteomics, structural biology and molecular systems biology. Biology is increasingly an information science, with bioinformatics a key enabling technology.

Other disciplines have developed at the bioscience —computer science— IT interface, and there is little consensus on where boundaries should be drawn among them. Bioinformatics is sometimes said to focus on the development and application of methods and software tools to acquire, manage, analyse and/or visualise biological data, whereas *computational biology* is more the application of these methods and tools to theoretical or applied biological questions (Huerta *et al.*, 2000). *Biomathematics* or *mathematical biology* involves the development or use of mathematical modeling or simulation, while *biostatistics* emphasises experimental design and statistical analysis. *Molecular systems biology* focuses on the inference or analysis of networks of genes, proteins and/or other cellular molecules, while *synthetic biology* applies these technologies to design and engineer new biological functions or organisms.

BIOINFORMATICS FROM THE PERSPECTIVE OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

One way of exploring the interface between molecular bioscience and IT is to track experimental data from its generation, capture and retrieval, to its aggregation and dissemination *via* international data services, to its subsequent analysis. Here I deconstruct data analysis into data models, algorithms, analytical methods and

DOI: 10.4018/978-1-4666-5888-2.ch038

software, workflows and visualisation. This trajectory is common to most or all experimental data in the sciences, although bioinformatics is notable for its culture of open data, well-established data formats and standards, and large international data repositories and data services.

Data generation, storage and retrieval: Instruments and experiments generate diverse data types in molecular bioscience. Capturing these primary data and the associated metadata, and managing their storage and retrieval, are primary activities in bioinformatics. The quantities of data generated by DNA-sequencing platforms, in particular, are such that raw data are no longer archived; rather, bioinformatic methods are used to assess quality and extract summaries. Data formats are specific to experimental technologies and, to some extent, instrument manufacturers. In some areas of molecular bioscience, standards have been developed to ensure that data can be interpreted unambiguously and, in principle, the experiment can be reproduced. For example, the MIAME (Minimum Information About a Microarray Experiment) standard (Brazma *et al.*, 2001) specifies how experimental design, laboratory protocols, biological samples, microarray platforms, and raw and processed data must be described, and recommends the use of certain data formats and ontologies. A corresponding standard, MIABi, has been proposed for the description of bioinformatics investigations (Tan *et al.*, 2010).

Public data resources: Newly generated biomolecular data (*e.g.* DNA and protein sequences, protein structures, gene-expression data) are submitted to public data repositories, where they are assigned unique persistent identifiers; all major bioscience journals require new data to be so identified. The main international data collection centres data are the US National Centre for Biotechnology Information (NCBI), the EMBL European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ). Individually and in collaboration with each other, these centres carry out further quality control on incoming data, conduct research in bioinformatics, and provide comprehensive online data services (*e.g.* search, retrieval, integrative analyses over multiple data sources, and links to journal articles and patents) which are cost-free at the "point of use" for the international research community. Other public data resources serve specific areas of molecular bioscience, *e.g.* protein structure. Increasingly, the largest projects in molecular bioscience maintain their own public data resources: examples include The Cancer Genome Atlas and the International Cancer Genome Consortium. Both large and small data sources are reviewed in the annual Database issue of the journal *Nucleic Acids Research*.

Data formats and models: The need for data integration and re-use has driven the development of standard data presentation formats, notably the FASTA format (Lipman & Pearson, 1985). However, no single data model has been universally adopted across all bioinformatics applications. Many public data services provide flat (ascii) files, but relational (MySQL), semantic (RDF), Web Services and hybrid approaches are also in use. Third-party tools such as SRS or BioMart are often used to integrate and index multiple related collections for combined used. With public data now in the tens of petabytes and growing rapidly, bioinformatics has entered the era of Big Data.

Algorithms and computation: The molecular biosciences offer diverse and difficult challenges, against which a wide range of algorithmic approaches have been deployed. Gene and protein sequences map naturally to strings, regulatory signals to motifs, phylogenies to trees, lateral genetic transfer to edits on trees, genetic regulatory networks and protein interactions to networks, and so on. Thus operations in bioinformatics can be recast as known or new problems in *e.g.* string matching, motif discovery, classification or graph theory. Many problems mapped in this way are NP-hard, *e.g.* maximum clique, vertex cover, and Steiner tree (Karp, 1972), or are suspected of being so. Increasingly, however, many are found to be fixed-parameter tractable, allowing high-quality solutions even on very large bioscience data. Heuristics are important in bioinformatics as well: the sequence alignment problem drove the early development of dynamic programming, phylogenetics is an important application domain for Markov chain Monte Carlo in conjunction with Bayesian approaches, and problems in gene regulation, protein localisation and biomolecular networks continue to provide challenges for machine learning. Even so, key problems in bioinformatics continue to pose algorithmic and computational challenges for large data, requiring large shared memory and/or high-capacity input/output.