Preface

The year 1995 saw the arrival of the first completed microbial genomes, *Haemophilus influenzae* and *Mycoplasma genitalium*. Several years of struggle for a complete genome was ended by the group at The Institute for Genomic Research (TIGR). From today's point of view, 16 years and more than a thousand completed genomes later (including, of course, the human genome), it may be hard to understand how much of an accomplishment this was. At the time, however, most laboratories around the world were still sequencing on slab gels using radioisotopes as the label for the fragments, which represent the DNA sequence.

When we sit down to "browse" genomes today, we do not often remember the days before e-mail and the Internet, or the days before automated DNA sequencing became a commodity. But it is certainly worthwhile to take a look back, as many of the design decisions that were made in the last 16 years influence the way we deal with genomic information today.

When the first genomes were presented at a by-invitation-only meeting in Worcester, England, in 1995, the only tool that was capable of handling such a large file was a word processor. Therefore, the sequence was first presented to the scientists at the meeting as a character file, which was scrolling on the screen behind the speaker. One of the major problems with handling a large DNA sequence file at the time was that most bioinformatics software was only tailored for DNA fragments of a size much less than a complete microbial genome, typically no more than approximately 100 kilobase pairs. The first automated genome analysis and annotation systems were barely emerging in 1995, and thus the handling of a complete genome with a size of more than a million base pairs all at once was impossible.

The Web was a fledgling entity in 1995, with not much power and entirely based on the Hypertext Markup Language (HTML). It became clear very quickly that only large communities of scientists with a diverse

background could really make sense of the genomic information, provided that they were enabled to collaborate, and thus the Web quickly became the vehicle by which genome annotations were created and exchanged among scientists. The first automated genome analysis and annotation systems, which were Web-based, initially produced tabular output that listed the location of potential genes and gene functions, which were predicted mostly by database comparison. It became obvious very early that this was not sufficient for biologists, therefore graphical subsystems were added, which are today part and parcel of all genome analysis and annotation systems and are probably the only part of an automated genome analysis and annotation system that most users ever encounter.

Over time, the Web developed into the massive entity it is today, with many additions to the Web technologies, which were utilized in turn by the developers of today's genome analysis and annotation tools. The three most useful tools in this context were probably (1) the creation of the programming language Java by James Gosling, which allowed the development of truly platform-independent applications; (2) the introduction of Extensible Markup Language (XML), which could adequately be used for the description of biological and medical objects; and (3) the creation of Web services (for example, the BioMOBY system), which made distributed computing simple and easy, and allowed the transparent and seamless integration of new bioinformatics tools into Web-based bioinformatics applications.

DNA sequencing technology has progressed in several iterations to today's level, which is called "next-generation sequencing," but really represents the third or fourth generation of DNA sequencing technologies, with yet another generation just around the corner. The sheer amount of DNA sequence, which can be produced on a single device today, is mind-boggling. It has literally become possible to resequence genomes the size of the human genome within a few hours in a single laboratory and the $1000 human genome is on the horizon. At the time of this writing (2012), very few genome annotation pipelines are capable of dealing with this information volume and new strategies need to be developed to accommodate the needs of today's genome researchers.

In the near future, everyone will be able to carry their genome sequence on some kind of data storage device and diagnostics might become largely based on the results of genomics screens, which will be cheaper than today's advanced imaging technologies (MRI and CT scans, for example). Thoroughly annotated genomic information and the integration

of all information into a single model will be a prerequisite to successful approaches to individualized medicine, the development of advanced crops and the sustainable production of food, medical research and development, and the development of new and sustainable energy sources.

This book attempts to introduce the topic of automated genome analysis and annotation. The initial chapters take the reader through the last 16 years, explaining how the current analysis strategies were developed. This is followed by the introduction of up-to-date tools, which represent today's state of the art. The authors also discuss strategies for the analysis and annotation of next-generation DNA sequencing data. This book is intended to be used by professionals and students interested in entering the field.

We would like to thank Hershel Safer and the editorial team at CRC Press/Taylor & Francis for their patience while creating this book.