Preface

The recent accumulation of information from genomes, including their sequences, has resulted not only in new attempts to answer old questions and solve longstanding issues in biology, but also in the formulation of novel hypotheses that arise precisely from this wealth of data. The storage, processing, description, transmission, connection, and analysis of these data has prompted bioinformatics to become one the most relevant applied sciences for this new century, walking hand-in-hand with modern molecular biology and clearly impacting areas like biotechnology and biomedicine.

Bioinformatics skills have now become essential for many scientists working with DNA sequences. With this idea in mind, this book aims to provide practical guidance and troubleshooting advice for the computational analysis of DNA sequences, covering a range of issues and methods that unveil the multitude of applications and relevance that Bioinformatics has today. The analysis of protein sequences has been purposely excluded to gain focus. Individual book chapters are oriented toward the description of the use of specific bioinformatics tools, accompanied by practical examples, a discussion on the interpretation of results, and specific comments on strengths and limitations of the methods and tools. In a sense, chapters could be seen as enriched task-oriented manuals that will direct the reader in completing specific bioinformatics analyses.

The target audience for this book is biochemists, and molecular and evolutionary biologists that want to learn how to analyze DNA sequences in a simple but meaningful fashion. Readers do not need a special background in statistics, mathematics, or computer science, just a basic knowledge of molecular biology and genetics. All the tools described in the book are free and all of them can be downloaded or accessed through the web. Most chapters could be used for practical advanced undergraduate or graduate-level courses in bioinformatics and molecular evolution.

The book could not start in another place than describing one of the most widespread bioinformatics tool: BLAST (Basic Local Alignment Search Tool). Indeed, one of the first steps in the analysis of DNA sequences is their collection. Therefore, **Chapter 1** guides the reader through the recognition of similar sequences using BLAST. Next, the use of OrthologID for understanding the nature of this similarity is described in **Chapter 2**, followed by a **Chapter 3** about one of the most important stages in most bioinformatics pipelines, the alignment, which shows the basis and the application of the program MAFFT. The next set of chapters is intimately related to the study of molecular evolution. Indeed, the DNA sequences that we see today are the result of this process. In **Chapter 4**, SeqVis is used to detect compositional changes in DNA sequences through time, while **Chapter 5** is focused on the selection of models of nucleotide substitution using jModelTest. Precisely the use of these models for phylogenetic reconstruction is described in **Chapter 6**, which capitalizes upon the estimation of maximum likelihood phylogenetic trees with Phyml. Indeed, the estimation of phylogenies is often the first step in many evolutionary analyses. How to combine multiple trees in a single supertree is the basis of **Chapter 7**, which explains

the use of the program Clann. Next, **Chapters 8 and 9** are centered on the characterization of two key evolutionary processes acting on DNA sequences. The use of the server Datamonkey for the detection of selection is described in **Chapter 8**, while **Chapter 9** shows the nuts and bolts of the detection of recombination using RDP3.

The study of codon usage, which has provided many important insights at the genomic scale, is deciphered in **Chapter 10** using CodonExplorer, an interactive data base, while **Chapter 11** explains how differences in the genetic code can be detected using GenDecoder. The next chapters are related to the annotation of genomes, an essential requisite for many other analyses. In **Chapter 12**, we learn how to predict genes using GeneID, while in **Chapter 13** the identification of regulatory motifs with A-Glam is described. **Chapter 14** then explains the use of the UCSC genome browser and its applications, for example, to characterize a gene or to explore conserved elements. The discovery of single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) with bioinformatics tools SNPServer, dbSNP, and SSR Taxonomy Tree is the subject of **Chapter 15**, and **Chapter 16** highlights the use of Censor and RepeatMasker for the detection and characterization of transposable of sequences in eukaryotic genomes. To end the book, **Chapter 17** explains how to make the most of DnaSP for the analysis of DNA sequences in populations.

I am very grateful to all the authors, the fundamental piece, who have put a lot of effort replying patiently to all my queries. Hopefully, the result has been a set of clear and useful chapters that will be of help to other scientists. I want to thank all of them for sharing their time, wisdom and expertise. Finally, I want to thank John Walker, the editor of the series, for his continuous advice.

Vigo, July 2008   *David Posada*