Preface

Recent years have seen impressive progress in computational biology. To an increasing extent, this is owed to the use of modern machine learning techniques for the analysis of high-dimensional or structured data. In the early days of machine learning in computational biology, a substantial number of relatively straightforward applications of existing techniques to interesting data analysis problems were carried out. However, because the problems that can be dealt with in this way are gradually running out, it has become increasingly important to actively develop learning algorithms that can deal with the difficult aspects of biological data, such as their high dimensionality (e.g., in the case of microarray measurements), their representation as discrete and structured data (e.g., DNA and amino acid sequences), and the need to combine heterogeneous sources of information.

A recent branch of machine learning, called kernel methods, lends itself particularly well to the study of these aspects, making it rather suitable for problems of computational biology. A prominent example of a kernel method is the *support vector machine* (*SVM*). Its basic philosophy, which is shared by other kernel methods, is that with the use of a certain type of similarity measure (called a kernel), the data are implicitly embedded in a high-dimensional feature space, in which linear methods are used for learning and estimation problems. With the construction of various types of kernels, one can take into account particular aspects of problems of computational biology, while the choice of the linear learning method which is carried out in the feature spaces allows one to solve a variety of learning tasks such as pattern recognition, regression estimation, and principal component analysis.

This book provides an in-depth overview of current research in the field of kernel methods and their applications to computational biology. In order to help readers from different backgrounds follow the motivations and technicalities of the research chapters, the first part is made up of two tutorial and one survey chapter. The first chapter, by Alexander Zien, provides a compressed introduction to molecular and computational biology. Mainly directed toward computer scientists and mathematicians willing to get involved in computational biology, it may also provide a useful reference for bioinformaticians. The second chapter, by Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf, is a short introduction to kernel methods. Focusing more on intuitive concepts than on technical details, it is meant to provide a self-contained introduction for the reader new to this field. The third chapter, by William S. Noble, is an in-depth survey of recent applications of

kernel methods in computational biology, apart from the ones covered later in the book.

Following these three introductory chapters, the book is divided into three parts, which roughly correspond to three general trends in current research: kernel design, data integration, and advanced applications of SVMs to computational biology. While the chapters are self-contained and may be read independently of each other, this organization might help the reader compare different approaches focused on related issues and highlight the diversity of applications of kernel methods.

Part II is made up of six contributions that present different ideas or implementations for the design of kernel functions specifically adapted to various biological data. In chapter 4, Christina Leslie, Rui Kuang, and Eleazar Eskin present a family of kernels for strings based on the detection of similar short subsequences that have fast implementations and prove to be useful as kernels for protein sequences for the detection of remote homologies. A fast implementation of some of these kernels is detailed in chapter 5 by S.V.N. Vishwanathan and Alexander J. Smola, using suffix trees and suffix links to speed up the computation. Jean-Philippe Vert, Hiroto Saigo, and Tatsuya Akutsu present in chapter 6 a different kernel for protein sequences derived from measures of sequence similarities based on the detection of local alignments, also tested on a benchmark experiment of remote homology detection. Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi introduce in chapter 7 kernels for graphs, based on the detection of similar paths between graphs, with applications to the classification of chemical compounds. Kernels between nodes of a graph, called diffusion kernels, are then presented in chapter 8 by Risi Kondor and Jean-Philippe Vert, with applications to the comparison of gene expression and metabolic pathways. Finally, a kernel between short amino acid sequences is introduced in chapter 9 by Yann Guermeur, Alain Lifchitz, and Régis Vert, with application to protein secondary structure prediction.

Part III covers different approaches based on kernel methods to learn from heterogeneous information. In chapter 10, Yoshihiro Yamanishi, Jean-Philippe Vert and Minoru Kanehisa propose to detect correlations between heterogeneous data using a generalization of canonical correlation analysis that involves the kernel trick, and illustrate their approaches by the automatic detection of operons in bacterial genomes. A formalism based on semidefinite programming to combine different kernels representing heterogeneous data is presented in chapter 11 by Gert R. G. Lanckriet, Nello Cristianini, Michael I. Jordan, and William S. Noble, with applications in functional genomics. As a third kernel method to learn from heterogeneous data, Taishin Kin, Tsuyoshi Kato, and Koji Tsuda propose in chapter 12 a formalism based on the information geometry of positive semidefinite matrices to integrate several kernels, with applications in structural genomics.

Part IV contains several examples where SVMs are successfully applied to difficult problems in computational biology. Gunnar Rätsch and Sören Sonnenburg focus in chapter 13 on the problem of splice site prediction in genomic sequences, and develop a state-of-the-art algorithm based on SVMs. Chapter 14, by Balaji Krishnapuram, Lawrence Carin, and Alexander Hartemink, and chapter 15, by

Sepp Hochreiter and Klaus Obermayer, both focus on the classification of tissues based on gene profiling experiments and on the problem of gene selection in this context. They come up with two different variants of the SVM algorithm that perform gene selection and tissue classification simultaneously, with very promising experimental results.

The impetus for this book was a workshop entitled "Kernel Methods in Computational Biology" which was held in the Harnack-Haus of the Max Planck Society in Berlin, on April 14, 2003. The one-day workshop brought together the leading proponents of this emerging field, providing a snapshot of the state of the art. Held at the same time as the RECOMB conference, it attracted an audience of 135 registered participants. The program consisted of nine invited talks, three contributed talks, and five posters. The articles in this book are partly based on presentations at the workshop, augmented with several invited papers. All chapters have been carefully peer-reviewed and edited to produce, we hope, a useful vehicle for helping people getting up to speed on an exciting and promising direction in basic research.

We thank everybody who helped make the workshop and this book possible, in particular Sabrina Nielebock for administrative help with the workshop, Karin Bierig for help with the figures, and Arthur Gretton for proofreading.

Bernhard Schölkopf, Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Koji Tsuda, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, and AIST Computational Biology Research Center, Tokyo, Japan

Jean-Philippe Vert, Ecoles des Mines, Paris, France