Chapter 1.24
Kernel Methods in Genomics and Computational Biology

Jean-Philippe Vert
Ecole des Mines de Paris, France

ABSTRACT

Support vector machines and kernel methods are increasingly popular in genomics and computational biology due to their good performance in real-world applications and strong modularity that makes them suitable to a wide range of problems, from the classification of tumors to the automatic annotation of proteins. Their ability to work in a high dimension and process nonvectorial data, and the natural framework they provide to integrate heterogeneous data are particularly relevant to various problems arising in computational biology. In this chapter, we survey some of the most prominent applications published so far, highlighting the particular developments in kernel methods triggered by problems in biology, and mention a few promising research directions likely to expand in the future.

INTRODUCTION

Recent years have witnessed a dramatic evolution in many fields of life science with the apparition and rapid spread of so-called high-throughput technologies, which generate huge amounts of data to characterize various aspects of biological samples or phenomena. To name just a few, DNA sequencing technologies have already provided the whole genome of several hundreds of species, including the human genome (International Human Genome Sequencing Consortium, 2001; Venter, 2001). DNA microarrays (Schena, Shalon, Davis, & Brown, 1995), that allow the monitoring of the expression level of tens of thousands of transcripts simultaneously, opened the door to functional genomics, the elucidation of the functions of the genes found in the genomes (DeRisi, Iyer, & Brown, 1997). Recent advances in ionization technology have boosted large-scale capabilities in mass spectrometry and the rapidly growing field

of proteomics, focusing on the systematic, large-scale analysis of proteins (Aebersold & Mann, 2003). As biology suddenly entered this new era characterized by the relatively cheap and easy generation of huge amounts of data, the urgent need for efficient methods to represent, store, process, analyze, and finally make sense out of these data triggered the parallel development of numerous data-analysis algorithms in computational biology. Among them, kernel methods in general and support vector machines (SVMs) in particular have quickly gained popularity for problems involving the classification and analysis of high-dimensional or complex data. Half a decade after the first pioneering papers (Haussler, 1999; T. S. Jaakkola, Diekhans, & Haussler, 1999; Mukherjee, Tamayo, Mesirov, Slonim, Verri, & Poggio, 1998), these methods have been applied to a variety of problems in computational biology, with more than 100 research papers published in 2004 alone. The main reasons behind this fast development involve, beyond the generally good performances of SVM on real-world problems and the ease of use provided by current implementations, (a) the particular capability of SVM to resist high-dimensional and noisy data, typically produced by various high-throughput technologies, (b) the possibility to model linear as well as nonlinear relationships between variables of interest, and (c) the possibility to process nonvectorial data, such as biological sequences, protein structures, or gene networks, and to easily fuse heterogeneous data thanks to the use of kernels. More than a mere application of well-established methods to new data sets, the use of kernel methods in computational biology has been accompanied by new developments to match the specificities and the needs of the field, such as methods for feature selection in combination with the classification of high-dimensional data, the invention of string kernels to process biological sequences, or the development of methods to learn from several kernels simultaneously. In order to illustrate some of the most prominent applications of kernel methods in computational biology and the specific developments they triggered, this chapter focuses on selected applications related to the manipulation of high-dimensional data, the classification of biological sequences, and a few less developed but promising applications. This chapter is therefore not intended to be an exhaustive survey, but rather to illustrate with some examples why and how kernel methods have invaded the field of computational biology so rapidly. The interested reader will find more references in the book by Schölkopf, Tsuda, and Vert (2004) dedicated to the topic. Several kernels for structured data, such as sequences or trees, widely developed and used in computational biology, are also presented in detail in the book by Shawe-Taylor and Cristianini (2004).

CLASSIFICATION OF HIGH-DIMENSIONAL DATA

Several recent technologies, such as DNA microarrays, mass spectrometry, or various miniaturized assays, provide thousands of quantitative parameters to characterize biological samples or phenomena. Mathematically speaking, the results of such experiments can be represented by high-dimensional vectors, and many applications involve the supervised classification of such data. Classifying data in high dimension with a limited number of training examples is a challenging task that most statistical procedures have difficulties in dealing with, due in particular to the risk of overfitting the training data. The theoretical foundations of SVM and related methods, however, suggest that their use of regularization allows them to better resist to the curse of dimension than other methods. SVMs were therefore naturally tested on a variety of data sets involving the classification of high-dimensional data, in particular, for the analysis of tumor samples from gene expression data, and novel algorithms were developed in the framework of kernel methods to select a few