Leveraging Sequence Classification by Taxonomy-Based Multitask Learning

Christian Widmer[1], Jose Leiva[1,2], Yasemin Altun[2], and Gunnar Rätsch[1]

[1] Friedrich Miescher Laboratory, Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany
[2] Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany

Abstract. In this work we consider an inference task that biologists are very good at: deciphering biological processes by bringing together knowledge that has been obtained by experiments using various organisms, while respecting the differences and commonalities of these organisms. We look at this problem from an sequence analysis point of view, where we aim at solving the same classification task in different organisms. We investigate the challenge of combining information from several organisms, whereas we consider the relation between the organisms to be defined by a tree structure derived from their phylogeny. Multitask learning, a machine learning technique that recently received considerable attention, considers the problem of learning across tasks that are related to each other. We treat each organism as one task and present three novel multitask learning methods to handle situations in which the relationships among tasks can be described by a hierarchy. These algorithms are designed for large-scale applications and are therefore applicable to problems with a large number of training examples, which are frequently encountered in sequence analysis. We perform experimental analyses on synthetic data sets in order to illustrate the properties of our algorithms. Moreover, we consider a problem from genomic sequence analysis, namely splice site recognition, to illustrate the usefulness of our approach. We show that intelligently combining data from 15 eukaryotic organisms can indeed significantly improve the prediction performance compared to traditional learning approaches. On a broader perspective, we expect that algorithms like the ones presented in this work have the potential to complement and enrich the strategy of homology-based sequence analysis that are currently the quasi-standard in biological sequence analysis.

1 Introduction

Over a decade ago, an eight-year lasting collaborative effort resulted in the first completely sequenced genome of a multi-cellular organism, the free-living nematode *Caenorhabditis elegans*. Today, more than 50 eukaryotic genomes have been sequenced and several hundred more are underway. The genome sequences are the basis for much of the research on the molecular processes in these organisms.

B. Berger (Ed.): RECOMB 2010, LNBI 6044, pp. 522–534, 2010.

Typically, the more closely related the organisms are, the more similar are these processes. For some organisms, certain biochemical experiments for the analysis of particular processes can be performed more readily than for others (i.e. a large part of biological understanding was obtained from experiments based on a few model organisms such as yeast). This understanding can then be transferred to other organisms, for instance by verifying or refining models of the processes—often at a fraction of the original cost. This is but one example of a situation, where transfer of knowledge across organisms is very fruitful.

In computational biology we often study the problem of building statistical models from data in order to predict, analyze, and ultimately understand biological systems. Regardless of the problem at hand, be it the recognition of sequence signals such as splice sites, the prediction of protein-protein interactions, or the modeling of metabolic networks, we frequently have access to data sets for multiple organisms. Thus, our goal is to develop methods that aim at taking advantage of the data from different organisms in order to improve the performance of the statistical models built for all organisms. We argue that, when building a predictor for a given organism, data from other organisms should be incorporated to the extent of the relation between the organisms.

Since it is assumed that all life can be traced back to an ancient common ancestor, all organisms can ultimately be related by phylogeny. Furthermore, if two organisms share a sufficiently long evolutionary history before divergence, it can be expected that certain biological mechanisms (e.g., splicing) are conserved to some degree. Thus, it is reasonable to assume that we can leverage data from other organisms to enhance model quality for the organism of interest. In bioinformatics, this is traditionally done by considering sequence homology. This approach, however, is limited to almost exact correspondences of sequences between one or several biological sequences, while it fails to capture other features such as sequence composition that can be used to build an accurate model.

A family of machine learning methods, commonly referred to as *domain adaptation* or *transfer learning*, investigates the application of a predictor trained with data from a given domain to data from a different one (see e.g., [3,5,11]). Furthermore, the so-called *multitask learning* techniques consider the problem of simultaneously obtaining predictors from different domains by exploiting the fact that the domains are related (see e.g., [1,6]). Most of these methods assume uniform relations across domains/tasks.[1] However, it is conceivable that sharing information between closely related domains is more beneficial than sharing between domains that are only distantly related (according to a given criterion). Hence, it is important to take into account the degree of relatedness among the domains when obtaining the set of models. Here, we investigate multitask learning scenarios where we are given *a priori* information about a hierarchy that relates the domains at hand, which is often the case for biological problems. In particular, we treat each organism as a domain and employ the hierarchy given by the phylogeny. The fact that the availability of data describing the same biological mechanism in several organisms is a reoccurring theme makes

[1] We use the terms *task* and *domain* interchangeably.