Towards Large-scale Machine Learning for Metagenomics Sequence Classification

Kévin Vervier [1,2,3], Jean-Philippe Vert [1,2], Maud Tournoud [3],
Jean-Baptiste Veyrieras [3] and Pierre Mahé [3*]

Introduction

Assigning taxonomic labels to DNA sequences and estimating a microbial abundance profile from high throughput sequencing data is one of the main challenges in metagenomics, often referred to as taxonomic binning (Kunin et al., 2008). Two main computational strategies have been proposed to perform this task: (i) similarity-based approaches, where the DNA sequence is searched against a reference sequence database with sequence alignment tools like BLAST (Huson et al., 2007) or TMAP (Homer et al., 2010), and (ii) compositional approaches, where a machine learning model such as a naive Bayes (NB) classifier (Wang et al., 2007) or a support vector machine (SVM, McHardy et al., 2006; Patil et al., 2012) is trained to label the sequence based on the set of k-mers it contains. Since the taxonomic classification of a sequence by compositional approaches is only based on the set of k-mers it contains, they offer significant gain in terms of classification time over similarity-based approaches. Nevetheless, it seems that similarity-based and compositional approaches achieve comparable performances in terms of classification accuracy (Parks et al., 2011; Patil et al., 2012).

Compositional approaches must be trained on a set of sequences with known taxonomic labels, typically obtained by sampling fragments from reference genomes. In the case of NB classifiers, explicit sampling of fragments from reference genomes is not needed to train the model: instead, a global profile of k-mer abundance from each reference genome is sufficient to estimate the parameters of the NB model, leading to simple and fast implementations (Wang et al., 2007). On the other hand, in the case of SVM, an explicit sampling of fragments from reference genomes to train the model based on the k-mer content of each fragment is needed. For example, Patil et al. (2012) sampled approximately 10,000 fragments from 1768 genomes to train a structured SVM (based on a k-mer representation with $k = 4, 5, 6$), and reported an accuracy competitive with similarity-based approaches.

Increasing the number of fragments sampled to train a SVM may improve its accuracy, and allow to investigate larger values of k. However it also raises computational challenges, as it involves machine learning problems where a model must be trained from potentially millions or billions of training examples, each represented by a vector in 10^9 dimensions for, e.g., $k = 15$. This is out of reach of most standard implementations of SVM. In this work, we investigate the potential of modern, large-scale SVM implementations for taxonomic label assignment. We demonstrate in particular that increasing the number of fragments used to train the SVM has a significant impact on the accuracy of the model, and allows to estimate models based on longer k-mers. In this study, we evaluated tools for compositional read classification based on SVM Liblinear (Fan et al., 2008) and Vowpal Wabbit (Langford et al., 2007). The corresponding machine learning problem is a multiclass classification learned on millions of examples represented in the feature space $\mathbb{R}^{4^k}$, whose dimensionality exponentially grows with k.

Results

We implemented a multiclass SVM model for taxonomic label assigment based on two state-of-the-art large-scale SVM implementations, SVM Liblinear (Fan et al., 2008) and Vowpal Wabbit (Langford et al.,

[*1] Data and Knowledge Lab., bioMérieux, 69280 Marcy-l'Étoile, France , [2] Center for Computational Biology, Mines ParisTech, 77300 Fontainebleau, France , [3] Institut Curie, U900 INSERM, 75005 Paris, France

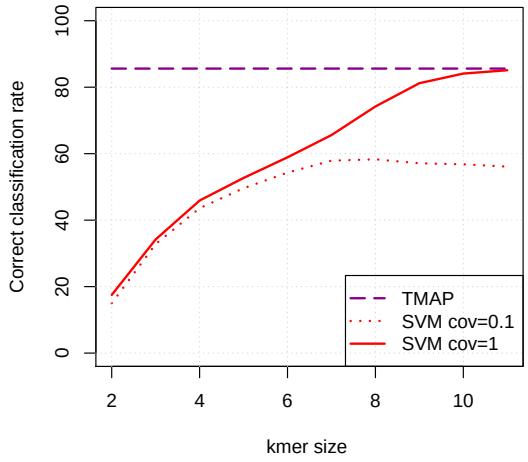

Figure 1: **Comparison between SVM and TMAP on simulated shotgun data.** This figure gives correct classification rate for two SVMs (red) trained from fragments covering each reference genome with a mean coverage of 0.1 (dotted line) and 1 (solid line). Performances are reported as a function of k-mer sizes and compared to the baseline value given by alignment-based method TMAP (purple dashed line).

2007). For a given length of k-mers, each sequence x is represented by the vector $\phi(x) \in \mathbb{R}^{4^k}$ of counts of k-mers it contains, and we estimate a linear SVM in this space. We considered a reference database with 356 complete genome sequences covering 52 bacterial species, and simulated a test set of 134,539 Roche 454 reads from a bronchoalveolar lavage model (Erb-Downward et al., 2011) with a mean length equal to 450 bp[1]. To train the SVM, we randomly sampled 450 bp fragments from each genome in order to cover each genome with a mean coverage of 0.1 or 1. This led to a training set of about 200k fragments at coverage 0.1, and of about 2 millions fragments at coverage 1. We compared the performance of a multiclass SVM trained on these two datasets, for different values of k, with the state-of-the-art TMAP alignment-based method.

Figure 1 shows the performance of the different methods in terms of correct species-level classification rate. We see that when the number of training fragments is large enough, SVM benefit from longer k-mer size up to $k = 11$, the largest value we tested in this preliminary study. We see also that at this level (corresponding to $4^{11} \simeq 4.10^6$ dimensions), increasing the training set size from 200k to 2 millions examples increases the performance by about 30%. For $k = 11$ and 2 millions training examples, the performance of the multiclass SVM is similar to TMAP, the similarity-based method. In terms of speed, the best SVM was 22 times faster than TMAP and took 3 minutes to classify the 134k test sequences on a single core. These first results demonstrate the potential of SVM-based methods with massive training set for sequence classification in metagenomics.

References

J. R Erb-Downward et al. *PloS one*, 6(2):e16384, 2011.

R.-E. Fan et al. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

N. Homer et al. https://github.com/iontorrent/TMAP.

D. H Huson et al. *Genome research*, 17(3):377–386, 2007.

V. Kunin et al. *Microbiology and Molecular Biology Reviews*, 72(4):557–578, 2008.

J. Langford et al. Technical Report, Yahoo, 2007.

A. C. McHardy et al . *Nature methods*, 4(1):63–72, 2006.

K. R. Patil et al. *PloS one*, 7(6):e38581, 2012.

Q. Wang et al. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.

D. H. Parks et al. *BMC Bioinformatics*, 12:328, 2011.

[1]A similar experiment involving reads of length 200 bp simulated with an Ion Torrent sequencing error model led to similar results.