PhD Proposal :
String embeddings for large-scale machine learning in genomics

Description

The cost of DNA sequencing has been divided by 100,000 in the last 10 years[1]. It is now so cheap that it has quickly become a routine technique to characterize the genomic content of biological samples with numerous applications in health[2], food or energy[3]. The output of a typical DNA sequencing experiment is a set of billions of short sequences, called *reads*, of lengths 100∼300 in the {A,C,G,T} alphabet ; these billions of reads are then automatically processed and analyzed by computers to get some biological information such as the presence of particular bacterial species in a sample, or of a specific mutation in a cancer.

As the throughput of DNA sequencing continues to increase at a fast rate, the major bottleneck in many applications involving DNA sequencing is quickly becoming computational. **The goal of this PhD project is to advance the state-of-the-art and propose new solutions for *storing* and *analyzing* efficiently the billions of reads produced by each experiment.**

More precisely, we will focus on two important applications of DNA sequencing :

- *metagenomics*, where the goal is to assign each read to a bacterial species in order to quantify the species that may be present in the sample analyzed ;
- *RNA-seq*, where the goal is to assign each read to a gene, in order to quantify the level of expression of all genes in the sample analyzed.

The basic problem to be solved in both applications is to assign each read to one among a set of known, longer target sequences (bacterial genomes or gene sequences). Standard techniques to solve that problem try to *align* each read to each target, using tools such as BLAST[4], BWA[5] or BOWTIE[6]. However, the computational cost of these techniques becomes prohibitive with current large sequence datasets, and faster alternative have been proposed recently. In particular, the problem can be reformulated as a supervised multiclass classification problem and solved by machine learning techniques such as naive Bayes[7] or support vector machines (SVM)[8]. We recently showed that large-scale machine learning techniques are competitive in

---

[1] https://www.genome.gov/sequencingcostsdata/
[2] https://cancergenome.nih.gov
[3] http://www.hydrocarbonmetagenomics.com
[4] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.
[5] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
[6] Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359.
[7] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Applied and environmental microbiology, 73(16):5261–5267, 2007.
[8] Kaustubh R Patil, Peter Haider, Phillip B Pope, Peter J Turnbaugh, Mark Morrison, Tobias Scheffer, and Alice C McHardy. Taxonomic metagenome sequence assignment with structured output models. Nature methods, 8(3):191–192, 2011.

accuracy and much better in computational cost that alignment-based methods for metagenomics applications[9].

The standard approach to solve the machine learning formulation is to represent each read as a fixed-length vector and then to train a linear classifier. A typical representation is to count the number of occurrences of each *k*-mer in a read, and to store these counts in a 4^*k* – dimensional vector, where *k* is an integer between 8 and 15. Recently, different representations using gapped *k*-mers and locality-sensitivity hashing (LSH) have been proposed and led to promising results[10], suggesting that there exists room for improvement in the way we represent reads as vectors for large-scale machine learning.

In this context the PhD candidate will investigate and propose new ways to represent DNA sequencing reads, that would lead to both (i) a compact representation for efficient storage and fast processing, and (ii) good performance in read classification for metagenomics and RNA-seq applications. Techniques to be investigated will include, in particular :

- Random features to approximate string kernels[11]
- LSH-based representations, including minHash[12]
- Deep learning-based representations, including convolutional[13] and recurrent neural networks

Application

PhD supervised by Jean-Philippe Vert (MINES ParisTech / Institut Curie / ENS Paris)

PhD fellowship of MINES ParisTech (about 1,690 €/month, net)

The ideal candidate should have a background in statistical machine learning, and a keen interest in biological applications (but no prior background in biology is needed).

To apply : send CV, master grades and contact informations of two persons I could reach for recommendation to jean-philippe.vert@mines-paristech.fr **before July 6, 2017.**

Do not hesitate to contact me for further information.

[9] Vervier K., Mahé, P.,Tournoud, M., Veyrieras, J.-B., and Vert, J.-P. (2016). Large-scale machine learning for metagenomics sequence classification. Bioinformatics, 32(7) :1023-1032.
[10] Luo, Y., Yu, Y, Zeng, J., Berger, B. and Peng, J. (2017) Metagenomic binning through low density hashing. biRxiv 133116.
[11] Rahimi, and Recht, B. (2007) Random features for large-scale machine learning. In NIPS 2007.
[12] Indyk, P. and Motwani, R. 1998. Approximate nearest neighbor: Towards removing the curse of dimensionality. In *Proceedings of the Symposium on Theory of Computing*.
[13] Zhang, X., Zhao, J., and LeCun, Y. (2016). Character-level Convolutional Networks for Text Classification. arXiv 1509:01626.