

Species Classification through Deep Learning

Gabrielle Strandquist, Yahya Bokhari, and Tomasz Arodz

Department of Computer Science, College of Engineering, Virginia Commonwealth University, USA

Introduction

Data Avalanche: Advances in sequencing technology lead to exponentially increasing data volume **GenBank**: 722 million sequenced sequences (Oct 2018) Data growing faster than computer speeds: Data volume doubles every 12 months Moore's Law: 18-month doubling time



DNA reads: Sequencers produce short DNA fragments called *reads* instead of a single long nucleotide sequence

- Metagenomic analyses depend on taxonomic labeling of reads.
- 360 times more bacterial

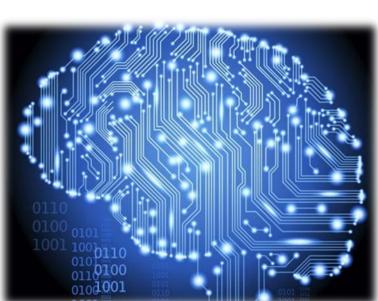
genes than human genes

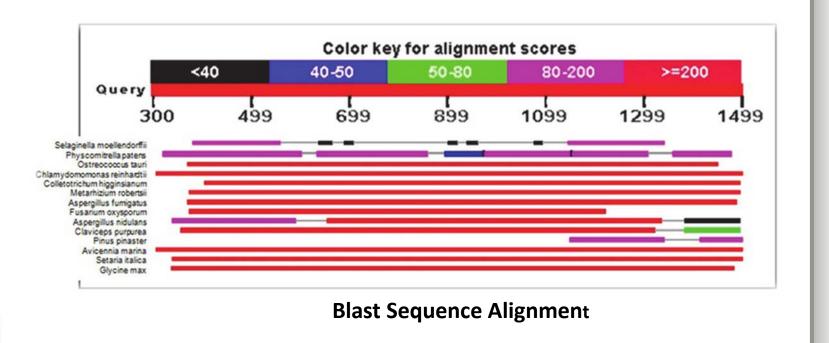
Research Problem:

How to quickly assign millions of DNA reads to their correct taxa?

Existing methods:

- Sequence Alignment is accurate but extremely
- Statistical Models based on k-mer profiles are faster but less accurate





New Approach:

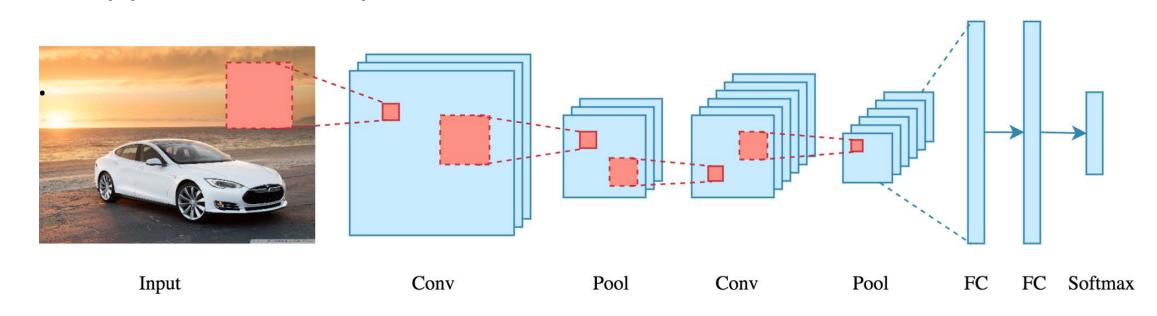
Deep Neural Networks that can "learn" correct taxonomy from DNA reads

Methods

DeepMetaReader, method 1: a taxonomic classifier that uses a deep neural network to learn descriptors from microbial sequences. DeepMetaReader uses convolutional, pooling and dense layers to find generalized patterns in DNA.

Convolution:

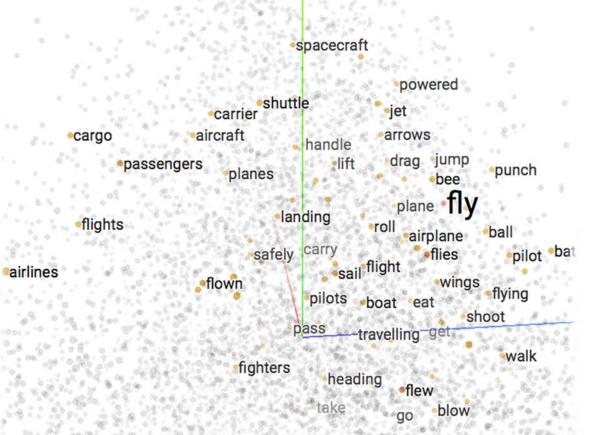
Patterns learned irrespective of where they appear in the input



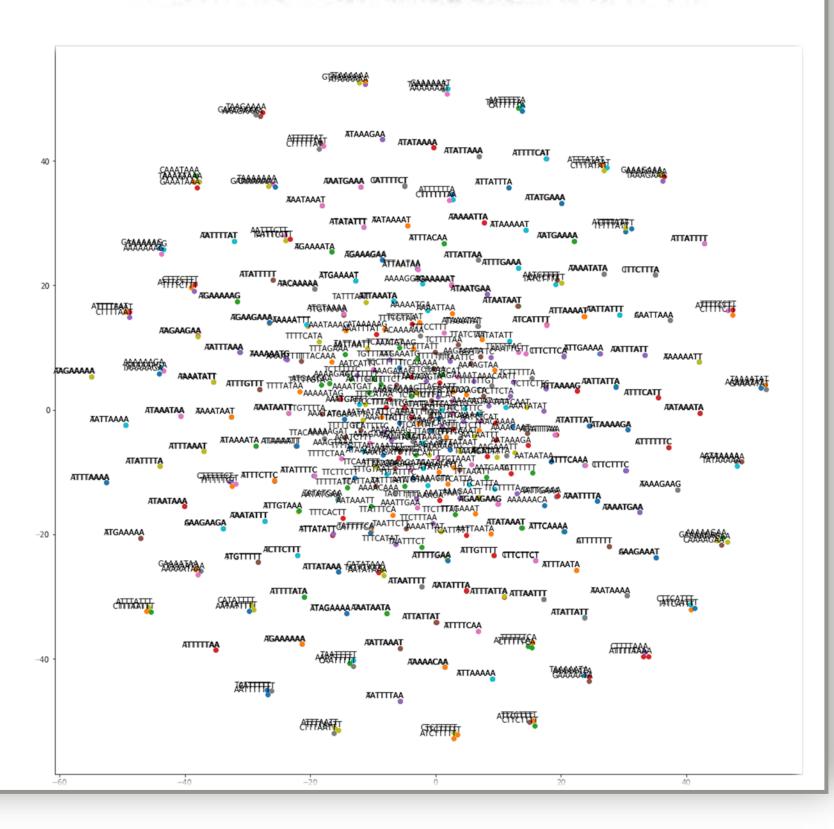
kmer2Vec, method 2:

- Inspired by Natural Language Processing techniques
- Relates words to context, word and context are nucleotide sequences

Word2vec in 3 dimensions:



kmer2vec in 16 dimensions:

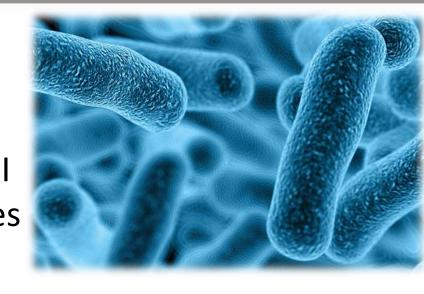


Results

DeepMetaReader

Data used:

- Over 400 whole bacterial genomes from NCBI
- Over 1 million reads generated from genomes
- Classified up to 68 different species



Success criteria:

- Precision: fraction of correct read-to-species assignments over total number of read-to-species assignments
- Sensitivity: ratio between number of correct read-to-species assignments and number of reads to be classified

Dataset	# species	Accuracy[%]	Precision	Sensitivity
EnSVMB	68	90.20	89.05	91.86
SimHC.20.50	19	93.06	92.14	94.04
HiSeq	10	97.34	97.12	97.58

kmer2vec

Data used:

- Whole bacterial genomes from NCBI
- Initial results for *B. anthracis* (47 genomes) vs. *E. coli* (474 genomes). One embedding model for *B.a.*, one for *E.c.*

Success criteria:

Contingency table: fraction of reads that are preferred by the correct model vs. preferred by the other model

	B. anthracis reads	E. coli reads
B. anthracis model preferred	100%	1.5%
E. coli model preferred	0%	98.5%

Conclusions

In the datasets we tested, DeepMetaReader surpasses existing methods while maintaining memory footprint low enough to fit into consumer-level GPU cards. We plan further studies on the kmer2vec embeddings, as a method for obtaining very compact representations that will allow for expanding the number of distinct taxa the tool can be trained to recognize.