Real-time pathogenicity prediction during genome sequencing of novel viruses and bacteria

Jakub M. Bartoszewicz[1-3], Ulrich Genske[1-4], Bernhard Y. Renard[1-2] jakub.bartoszewicz@hpi.de

[1]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, [2]Robert Koch Institute, [3]Free University of Berlin, [4]Charité – Universitätsmedizin Berlin

HPI Hasso Plattner Institut

Background

DNA **sequencing** is the state-of-the-art for open-view pathogen detection, generating millions of short DNA sequences per sample.

Targeted diagnostic assays are unavailable for **novel** pathogens at first.

The standard analysis is mapping: matching DNA reads against a database of **known** pathogen genomes.

Problem 1: novel, divergent threats may be **undetectable**

➡ **ResNets** predict if reads originate from novel pathogens

Problem 2: relatively **long** turnaround times

➡ **Real-time, selective** analysis of partial results

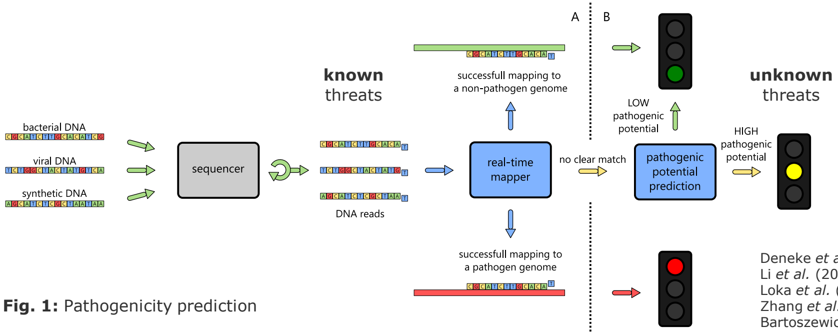

Fig. 1: Pathogenicity prediction

Results

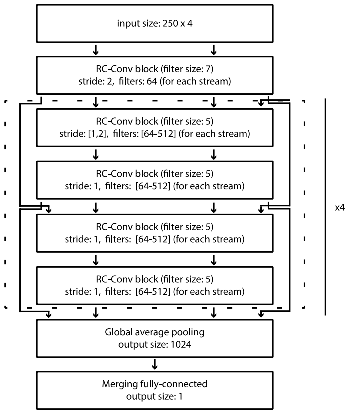
input size: 250 x 4
RC-Conv block (filter size: 7)
stride: 2, filters: 64 (for each stream)
RC-Conv block (filter size: 5)
stride: [1,2], filters: [64-512] (for each stream)
RC-Conv block (filter size: 5)
stride: 1, filters: [64-512] (for each stream)
RC-Conv block (filter size: 5)
stride: 1, filters: [64-512] (for each stream)
RC-Conv block (filter size: 5)
stride: 1, filters: [64-512] (for each stream)
x4
Global average pooling
output size: 1024
Merging fully-connected
output size: 1

Fig. 2: ResNet architecture

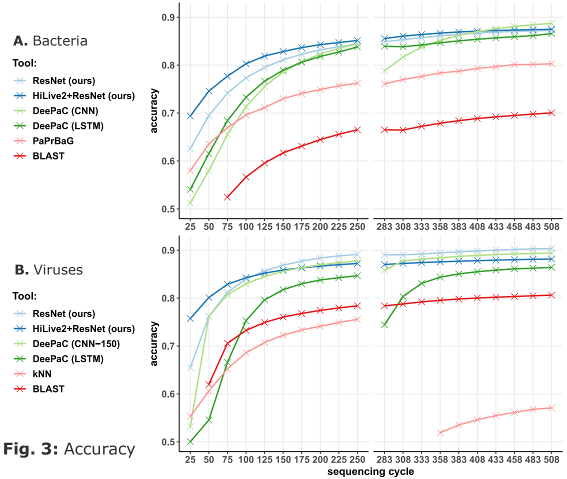

Fig. 3: Accuracy

Tab. 1: Recall	***Staphylococcus aureus*** *(not in training DB)*		**SARS-CoV-2** *(not in training DB)*	
	Nanopore, 250bp	Illumina, 250bp	Nanopore, 250bp	Illumina, 50bp
ResNet (ours)	**94.7**	**97.2**	**52.7**	**51.3**
mapping	3.3	1.6	4.6	0.6

Deneke *et al.* (2017), PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data, *Scientific Reports,* 7:39194.
Li *et al.* (2018), Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics,* 34(18):3094-3100.
Loka *et al.* (2019), Reliable variant calling during runtime of Illumina sequencing, *Scientific Reports,* 9(1):1-8.
Zhang *et al.* (2019), Rapid identification of human-infecting viruses, *Transboundary and Emerging Diseases*, 66(6):2517-2522.
Bartoszewicz *et al.* (2020), DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks, *Bioinformatics*, 36(1):81-89.
Bartoszewicz *et al.* (2021), Interpretable detection of novel human viruses from genome sequencing data, *NAR Genomics and Bioinformatics*, 3(1):lqab004.
Bartoszewicz *et al.* (2021), A deep learning framework for real-time detection of novel pathogens during sequencing, *bioRxiv*, 2021.01.26.428301.