

NEWS RELEASE 17-APR-2019

New software tool could provide answers to some of life's most intriguing questions

UNIVERSITY OF WATERLOO

A University of Waterloo researcher has spearheaded the development of a software tool that can provide conclusive answers to some of the world's most fascinating questions.

The tool, which combines supervised machine learning with digital signal processing (ML-DSP), could for the first time make it possible to definitively answer questions such as how many different species exist on Earth and in the oceans. How are existing, newly-discovered, and extinct species related to each other? What are the bacterial origins of human mitochondrial DNA? Do the DNA of a parasite and its host have a similar genomic signature?

The tool also has the potential to positively impact the personalized medicine industry by identifying the specific strain of a virus and thus allowing for precise drugs to be developed and prescribed to treat it.

ML-DSP is an alignment-free software tool which works by transforming a DNA sequence into a digital (numerical) signal, and uses digital signal processing methods to process and distinguish these signals from each other.

"With this method even if we only have small fragments of DNA we can still classify DNA sequences, regardless of their origin, or whether they are natural, synthetic, or computer-generated," said Lila Kari, a professor in Waterloo's Faculty of Mathematics. "Another important potential application of this tool is in the healthcare sector, as in this era of personalized medicine we can classify viruses and customize the treatment of a particular patient depending on the specific strain of the virus that affects them."

In the study, researchers performed a quantitative comparison with other state-of-the-art classification software tools on two small benchmark datasets and one large 4,322 vertebrate mitochondrial genome dataset. "Our results show that ML-DSP overwhelmingly outperforms alignment-based software in terms of processing time, while having classification accuracies that are comparable in the case of small datasets and superior in the case of large datasets," Kari said. "Compared with other alignment-free software, ML-DSP has significantly better classification accuracy and is overall faster."

The authors also conducted preliminary experiments indicating the potential of ML-DSP to be used for other datasets, by classifying 4,271 complete dengue virus genomes into subtypes with 100 per cent accuracy, and 4,710 bacterial genomes into divisions with 95.5 per cent accuracy.

###

A paper detailing the new software tool, titled ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels, which was authored by Kari together with Western University PhD candidate Gurjit Randhawa and Dr Kathleen Hill, an Associate Professor in the Department of Biology at We

Disclaimer: AAAS and EurekAlert! are not responsible for the accuracy of news releases posted to EurekAlert! by contributing institutions or for the use of any information through the EurekAlert system.

Media Contact

Matthew Grant
matthew.grant@uwaterloo.ca
226-929-7627

 @uWaterlooNews

<http://www.uwaterloo.ca/> 
