Bioinformatics data reduction techniques must be used with caution

July 5 2022, by Mariah Chuprinski

Credit: Pixabay/CC0 Public Domain

In the field of bioinformatics, DNA analysis can be performed with data sketching, a method that systematically reduces the size of a dataset to a smaller sample that allows scientists to analyze and approximate it at greater speeds. While the scalability of this method is appealing, two common tools used for data sketching allow for inaccuracies and inconsistencies in the analysis and results, a team of Penn State researchers found.

The team published their results about their analysis and findings of two different tools—the Jaccard estimator and the MinHash estimator—in *Bioinformatics* and the *Journal of Computational Biology*, respectively.

"The biomedical field has undergone a transformation over the last dozen years, as we now have technology that can sequence DNA from living organisms at an unprecedented scale," said co-principal investigator Paul Medvedev, director of the Center for Computational Biology and Bioinformatics and associate professor of computer science and engineering, biochemistry and molecular biology. "The bottleneck has therefore shifted from collecting data to analyzing it in a statistically rigorous and computationally feasible manner."

In genome sketching, data scientists extract a small but representative set of datapoints, called k-mers, which form a sketch that can be used to estimate the divergence between two genome sequences. The estimated divergence should be nearly identical to the true divergence and within an acceptable confidence interval. The researchers found that, contrary to common assumptions in the field, some sketching strategies used in bioinformatics do not meet these goals.

In *Bioinformatics*, published June 27, researchers discovered that the minimizer Jaccard estimator is biased and inconsistent, meaning that no matter how many data points one puts into the sketch, the estimate of the divergence between two genomes remains inaccurate. The reason for this, according to researchers, is that the sketch is sensitive to the ordering of the data points on the genome in a way that the true divergence is not.

To arrive at these findings, researchers simulated and analyzed E. coli genomes, where they compared the minimizer Jaccard estimate of a substring of E. coli data to the true value that they calculated by hand to find where in the sequence the smaller substring belonged. They did not align, and as a result, the researchers showed that there was a chance the method would not find the right location of the read within the larger genome.

"We came up with an abstract mathematical representation of the problem that is suited for the application of probability tools we wanted to use," said co-principal investigator David Koslicki, associate professor of computer science and engineering and biology. "We worked through that theory and determined if our assumptions of the theoretical structure were accurate. It turned out there was some small bias in the minimizer Jaccard estimator."

"The tools can still be useful for researchers if they do not mind the inconsistencies or small bias that are present, but if they affect what you're trying to measure, there are other sketching techniques we suggest using," Medvedev said.

In the *Journal of Computational Biology* paper, which was published in February, researchers tested the MinHash estimator, another method commonly used for data sketching, for its effectiveness in genomic research. In the study, researchers calculated the statistical properties of sketch datapoints that are affected by evolution.

"We studied how many of these k-mer datapoints get destroyed," Medvedev said. "Once we got those numbers, we were able to develop a confidence interval on the prediction of the estimator."

Confidence intervals determine the probability that a parameter will fall in a certain range of values, or in other words, how accurate a prediction is statistically, according to co-principal investigator Antonio Blanca, assistant professor of computer science and engineering.

"While sketching methods are ubiquitous in bioinformatics, the field has not rigorously studied how sequencing errors and mutations affect many of these methods," Blanca said. "These findings allow researchers to derive statistics and error estimates that they need to use sketching effectively in practice, leading to better measurements of the similarity and differences between organisms and more accurate methods of building DNA sequences."

In August 2021, the researchers presented their JCB paper at the RECOMB Conference, short for Research in Computational Molecular Biology, which took place in Padova, Italy. The *Bioinformatics* paper will be presented at the Intelligent Systems for Molecular Biology conference in July.

More information: Antonio Blanca et al, The Statistics of k-mers from a Sequence Undergoing a Simple Mutation Process Without Spurious Matches, *Journal of Computational Biology* (2022). DOI: 10.1089/cmb.2021.0431

Mahdi Belbasi et al, The minimizer Jaccard estimator is biased and inconsistent, *Bioinformatics* (2022). DOI: 10.1093/bioinformatics/btac244

Provided by Pennsylvania State University

Citation: Bioinformatics data reduction techniques must be used with caution (2022, July 5) retrieved 29 May 2024 from https://phys.org/news/2022-07-bioinformatics-reduction-techniques-caution.html