Efficient Minimizer Schemes using Deep Networks

Demetrius Hernandez and Dan DeBlasio
Department of Computer Science, University of Texas at El Paso
dhernandez79@miners.utep.edu, dfdeblasio@utep.edu

Minimizer schemes are sampling methods used in genomic applications to efficiently predict the matching probability of large sequences. Minimizers achieve this sampling by choosing the "minimum" *k*-mer (substring of length exactly *k*) in a window of *w* overlapping *k*-mers, where minimum is defined by a complete ordering over all possible *k*-mers. Despite its widespread application, storing/searching minimizer schemes with complex orderings (those that cannot be encoded into a simple procedure but tend to perform better on standard metrics) pose a challenge because the number of *k*-mers needed to store the ordering grows exponentially. To address this issue, a neural network was implemented to perform the *k*-mer lookup, which once trained can produce a result with little computational cost. The developed network architecture is complex enough to encode existing schemes, but not so complex that training times are extremely high. To determine the ideal architecture a supervised learning approach was used, relying on known optimal minimizer schemes and backpropagation. The produced network will be a drop-in replacement for existing minimizer implementations. By making complex orderings less resource intensive their use will be made more accessible, improving the performance of the current state of the art assemblers.

Keywords: Deep learning, Minimizer schemes, Sequence sketching, Machine Learning