Identification of repeat sequences in large genomes

Michael Ivanitsky, Connor Puritz

The identification of repeat sequences in genomes is crucial to understanding their structures. In particular, self-replicating DNA sequences (known as transposons) make up at least 44% of the human genome. They can often cause harmful mutations when they insert themselves into important regions of genes, but also have potential to be used in targeted gene therapy, making them important topics in genomic research.

Repeated DNA sequences can rarely be directly identified since a given sequence will inevitably vary across separate genomes, and even across repeated sequences in the same genome due to mutations. There are several methods used to account for this. The simplest is to use large, precompiled lists of known repeat sequences to estimate whether a given sequence is just a mutated version of a known one. This works well for humans, as we have had our genome heavily scrutinized. But for mist species, this method is unavailable.

When the first approach is unavailable, we have to analyze genomes *de novo*. To do this, probabilistic models are developed to predict whether a set of sequences are mutations of some canonical repeat sequence. A collection of such sequences is created, giving way to the first approach to be used in the future. For large genomes, though, this method is very slow, and can be unfeasible for many researchers to perform.

Our approach will be a combination of the aforementioned approaches. We hope to construct an artificial neural network trained on known repeat sequences, which will be able to predict whether sequences from the genomes' of related organisms are actually repeat sequences from the original organism. The hope is that by training the neural network to filter out noise caused by mutation, it can identify conserved repeat sequences shared between related organisms.

We believe that a modified form of a long-short term memory neural network would be able to learn to recognize sequences that are likely to be repeat sequences through the use of memory cells that would be able to store larger amounts of data.