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, but are also considered to be important for the regulation of DNA transcription, making them important targets of genomic research once identified.

Repeated DNA sequences cannot normally be directly identified since a given sequence will inevitably vary across genomes from separate individuals, and even in repeated sequences across the same genome, due to mutations. There are several methods used to account for this. The simplest is to simply rely on previously compiled collections of known repeat sequences for a species, or to estimate them from a large collection of sequenced genomes. This method is rarely applicable for species other than humans, making it a poor choice for a generalized procedure.

A second approach is to analyze genomes *de novo*, and use probabilistic models to build a collection of likely repeated sequences. This method tends to be unreasonably slow for large genomes, though, and are often not practical for researchers investigating larger organisms.

We hope to use an artificial neural network, first to quickly analyze genomes for occurrences of known repeated sequences, and eventually to be able to quickly analyze genomes *de novo* through the use of memory neurons. We also hope to be able to lean something about the propagation of transposable elements through the structure of a trained neural network.

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 transposable elements through the use of memory cells that would be able to store larger amounts of data.