Compression of Gene Coding Sequences

MohammadReza Ghodsi

April 22, 2008

The gene coding sequences are believed to be the most informative part of the genome. These sequences are often stored as a sequence of letters, each representing a nucleotide and each three of which correspond to an amino acid.

The genetic code has some redundancy. There are 4^3 possible codons but there are only 20 amino acids. Furthermore some codons might appear more frequently than others and also there might be repetitive pattern in the amino acid sequence of a protein. All of these suggest that we might be able to use lossless compression methods to considerably reduce storage requirements of coding sequences.

I am going to implement a compression algorithm designed specifically with gene coding sequences in mind. In particular I want to use a combination of the following two algorithms.

Ziv-Lempel This is one of the most popular compression algorithms. I am planning to use suffix trees to implement this algorithm.

Huffman coding This algorithm will produce a prefix-free code for codons. More frequent codons will be expressed using shorter strings of bits.

I hope that my implementation would have a better compression ratio that general purpose compression tools that use the same techniques (e.g. gzip).

One application of this tool might be in distributed computing efforts which process large amount of genomic data such as Folding@home. In such applications compressing the sequences will reduce the amount of total bandwidth used to transmit them over the network.