Mystery Organism Project

Hunter Casillas

Abstract

The world of computing has huge benefits on the biological world. We can use the power of computers to compute the origin of replication for an organism. I calculated possible dnaA boxes and accounted for a hamming distance of 1, GC disparity, and reverse compliments. Through this implementation, I was able to find the most likely dnaA box in the mystery organism and the actual origin of replication in E. coli.

Methods

In order to calculate the minimum skew for E. coli from the supplied fasta file, I first had to convert my existing functions/scripts to read fasta files. I did this by importing SeqIO from Bio which gave me the sequence quickly and easily. After refactoring some previous functions and implementing my frequent words function to work with reverse compliments, I was able to find the minimum skew and most frequent 9mers with a window length of 500 and 1 mismatch. The minimum skew position and dnaA box/origin of replication matched the ones in the book so I knew I was on the right path.

After following the tutorial in the book for finding dnaA boxes in E. coli, I was able to apply my functions to the mystery organism. The first thing I did was find the minimum skew of said organism. My function returned 2 possible positions that were only 2 indexes apart. These positions were so close to each other that it was safe to assume I could use either of the skews to attempt to find the dnaA boxes for the mystery organism.

Next, I used the minimum skew and ran my frequent words function on it in a few different varieties. Using varying window lengths and having my minimum skew position be the start, middle, and end of my window I was able to calculate a number of likely dnaA boxes for the mystery organism.

Results – E. coli

The minimum skew for the E. coli sequence I calculated was 3923620. The dnaA boxes I found were located near the minimum skew, with a window length of 500 starting at the minimum skew. My function returned a list of possible dnaA boxes shown below. Out of that list, the correct origin of TTATCCACA was found just like in the book! For E. coli, I found that the 9-mer 'TTATTCACA', its reverse complement, or its hamming distance 1 neighbors appeared frequently within a 500 nucleotide window.

Possible dnaA boxes returned for E. coli:

['AAGAGATCT', 'AAGGATCCT', 'AATGATCCG', 'AGAACAACA', 'AGATCTCTT', 'AGCTGGGAT', 'AGGATCAAC', 'AGGATCCTT', 'ATCCCAGCT', 'CAGAAGATC', 'CCAGGATCC', 'CGGATCATT', 'CTGGGATCA', 'CTGTTGATC', 'GATCAACAG', 'GATCCCAGC', 'GATCTTCTG', 'GCTGGGATC', 'GGATCCTGG', 'GGTTATCCA', 'GTGGATAAC', 'GTTATCCAC', 'GTTGATCCT', 'TCTGGATAA', 'TGATCAACA', 'TGATCCCAG', 'TGGATAACC', 'TGTGAATAA', 'TGTGGATAA', 'TGTTGATCA', 'TGTTGTTCT', 'TTATCCACA', 'TTATCCAGA', 'TTATTCACA']

Results – Mystery organism

The minimum skew for the mystery organism sequence I calculated was 3744269. The dnaA boxes I found were located near the minimum skew, with a window length of 500 with varying positions around the minimum skew. For example I used the minimum skew as my starting, ending and midpoint indexes. My function returned a few lists of possible dnaA boxes shown below. Out of that list, it appears that the 9-mer 'CGGCGCCGG' its reverse complement, or its hamming distance 1 neighbors appeared the most often within a 500 nucleotide window.

Possible dnaA boxes returned for the mystery organism:

['CCAGGATCC', 'CCCGGATCC', 'CCGGATCCG', 'CGGATCCGG', 'GGATCCGGG', 'GGATCCTGG']

['AGCTTCCGG', 'CCGGAAGCT']

['AACACGATC', 'AACCAGATC', 'AAGCCGATC', 'ACGGCGCCG', 'CCGGCGCCG', 'CGGCGCCGA', 'CGGCGCCGC', 'CGGCGCCGG', 'CGGCGCCGT', 'GATCGGCTT', 'GATCGTGTT', 'GATCTGGTT', 'GCGGCGCCG', 'TCGGCGCCG']

['CCAGGATCC', 'GGATCCTGG']

Discussion

In E. Coli I found the best candidate for the dnaA box was 'TTATTCACA', this also happens to be the origin of replication according to the book so it’s safe to assume that my program finds accurate possible dnaA boxes for sequences.

Using the minimum skew, reverse complement, and d-neighborhood techniques, I found a list of accurate possible dnaA boxes for the mystery organism. Although there are multiple possible candidates for a DnaA box, I determined that 'CGGCGCCGG' is the most likely candidate. While we cannot know for certain without experimentally confirming the result, this one seemed the most likely due to its location in relation to the minimum skew as well as having the highest frequency.