Introduction

The Breakage-Fusion-Bridge cycle was proposed by Barbara McClintock as a mechanism for genomic variation. Here, we explore the implications of an abstract model of the BFB cycle.

1. Start with a string $x = 1, 2, \ldots, n$, representing chromosomal ‘segments’ from the centromere to just before the telomere.

2. In the following, strings $x^0 = x, x^1, \ldots$ will represent abstractions of the genome after BFB cycle. Let $\text{suf}(y)$ describe a suffix of string $y$, and $-y$ describes a reversal of string $y$ with the sign inverted on each symbol. Then, for all $t$,

   $$x^t = x^{t-1} \cdot -\text{suf}(x^{t-1})$$

   where the suffix is chosen arbitrarily (excluding the entire string), and ‘$\cdot$’ implies concatenation. For example, the following is possible:

   $x^0 = 1, 2, 3, 4, 5$
   $x^1 = 1, 2, 3, 4, 5, -5, -4$
   $x^2 = 1, 2, 3, 4, 5, -5, -4, 4, 5, -5, -4, -3, -2$

Questions

1. Let $C^t[i]$ denote the copy number of the $i$-th segment after the $t$-th cycle. Thus, $C^0[i] = 1$ for all $i$ in the example above, but $C^1[4] = C^1[5] = 2$. Design and implement a fast algorithm that simulates BFB cycles, and returns the expected count of each segment at the end of the $t$-th iteration. Run your algorithm for $n = 10$, and describe (empirically) the growth of expected $C^t[i]$ for each $i$. You should submit a pseudo-code description of the algorithm, and plots of expected counts after 20-30 cycles.

2. Given a string $x \in [1, \ldots, n]^l$ of length $l$, design and implement an algorithm to test if it is a BFB string. Your code will be tested on examples.

3. Generate 3 random binary strings of length $n$ corresponding to an individual haplotype strains with $n$ SNVs. We could see this data, for example with viral or other microbial strains in a sample. However, unlike Haplotype assembly, we can no longer assume that one string is the bitwise complement of the other. Partial reads are sampled from these strings. Assume that each (gapped) read sample samples two bits with a gap of 1-3 bits, and a bit is read incorrectly with probability $\varepsilon$. Design and implement an MCMC algorithm to identify the 3 strains. Use $\varepsilon = 0.02$. Your submission must have the following:

   (a) Describe a formula for computing the likelihood of the data, given haplotype strains.
   (b) Describe the MCMC algorithm along with an argument that upon convergence, you will sample from the posterior likelihood distribution.
   (c) Provide your best answers for the data sets provided along with a count of number of iterations.
   (d) **Extra credit 10 pts.** Design and implement a dynamic programming algorithm, and compare against MCMC. What are the pros and cons of using D.P.? If you are not submitting A5, you can still implement the dynamic programming code to earn extra credit. Note that we have not discussed dynamic programming in class, so you will need to use techniques learned in a regular Algorithms class.