CS176 Fall 2018
Prof. Yosef

Tiffany Chen: cs176-aaw
Nader Namini Asl: cs176-aag

RNA-Sequencing Read-Alignment Project Write-Up

Approach

Our approach to the RNA sequencing project involved first alignment to the annotated transcriptome and then for the un-aligned reads, alignment to the genome. We created all the isoforms of all the genes and tried to align to the reads without any gaps, greedily choosing the mismatches, and bounding them at `MAX_NUM_MISMATCHES`. This meant that we checked the equality of the read sequence and each isoform element-by-element starting at index 0 and counted every inequality as a mismatch and terminated if the number of mismatches exceeded `MAX_NUM_MISMATCHES`. This gave us a fast runtime that was on average less than $0.5s$.
For the reads that did not align to the annotated transcriptome, we tried to align to the genome. After experimentation and evaluating our alignments with the alignment to hidden gene transcriptome, we found it best to limit the number of parts of our alignment to the genome to 1 as that would cover more than half of the hidden alignments. However, to find a *single* ungapped alignment for the reads to genome, we had to consider all the permutations of the read sequence. This meant that we went through the read sequence element by element and considered all the possible permutations of the bases at every index that would result in an exact match in the genome sequence. We achieved this by using a recursive function (`find_max_so_far`) that is documented in `project.py` along with all the helper functions used in the alignment.