Biological Sequence Analysis (Spring 2015)

Exercise 6 (Thu 19.2, 10-12, B222, Veli Mäkinen)

1. **Detecting large scale variants.**

   Consider different large scale variations in genome, like large deletions, large insertions, gene dublication, copy number variation, inversions, translocations, etc. How they can be identified using read alignment? Is there an advantage of using paired end reads?

2. **Read alignment I.**

   Construct the Burrows-Wheeler transform of `ACATGATCTGCATT` and simulate 1-mismatch backward backtracking search on the corresponding BWT index with the pattern `CAT`.

3. **Read alignment II.**

   Construct the Burrows-Wheeler transform of `ACATGATCTGCATT` and the Burrows-Wheeler transform of the *reverse* of `ACATGATCTGCATT`. Simulate 1-mismatch search on the corresponding BWT indexes using case analysis pruning with the pattern `GTTC`.

4. **Read alignment III.** Give a pseudocode for computing the values $\kappa(i)$ for prefix pruning applied on the prefixes $P_{1..i}$ of the pattern P.

5. **PWM search using indexing.**

   Most sequencing machines give probability for each position inside the read that the measurement was correct. Let $\mathbb{P}(p_i)$ be such a probability for position i containing p_i. Denote $M[c,i] = \mathbb{P}(p_i)$ if $p_i = c$ and $M[c,i] = (1 - \mathbb{P}(p_i))/(\sigma - 1)$ if $c \neq p_i$. Then we have a *positional weight matrix (PWM)* M representing the read (observe that this is a profile HMM without insertion and deletion states). We say that the matrix M *occurs* in position j in a genome sequence $T = t_1 t_2 \cdots t_n$ if $\mathbb{P}(M,T,j) = \prod_{i=1}^{m} M[t_{j+i-1}, i] > t$, where t is a predefined threshold. Explain how backtracking on the BWT index works for finding all occurrences of M in T. Show how to prune branches as soon as they cannot have an occurrence, even if all the remaining positions match $P_{1..j}$ exactly.