Biological Sequence Analysis (Spring 2015)

Exercise 7 (Thu 26.2, 10-12, B222, Veli Mäkinen)

1. **Suffix tree applications I.**

   Visualize suffix tree on a concatenation of two strings of your choice and mark the nodes corresponding to maximal unique matches. Choose an example with multiple nodes with two leaf children, of which some correspond to maximal unique matches, and some not.

2. **Suffix tree applications II.**

   Order k *de Bruijn graph* on text $T = t_1 t_2 \cdots t_n$ is a graph whose vertices are $(k-1)$-mers occurring in T and arcs are k-mers connecting $(k-1)$-mers consecutive in T. More precisely, let $\texttt{label}(v) = \alpha_1 \alpha_2 \cdots \alpha_{k-1}$ denote the $(k-1)$-mer of vertex v and $\texttt{label}(w) = \beta_1 \beta_2 \cdots \beta_{k-1}$ the $(k-1)$-mer of vertex w. There is an arc e from vertex v to w with $\texttt{label}(e) = \alpha_1 \alpha_2 \cdots \alpha_{k-1} \beta_{k-1}$ iff $\alpha_2 \alpha_3 \cdots \alpha_{k-1} = \beta_1 \beta_2 \cdots \beta_{k-2}$ and $\texttt{label}(e)$ is a substring of T.

   a) Show that the de Bruijn graph can be built in linear time, given the suffix tree of T.

   b) *(Voluntary extra assignment)* Show that the *k-mer index* can be built in linear time, given the suffix tree of T. Recall that a k-mer index associates to each k-mer of T the sorted list of occurrence positions.

3. **Bidirectional BWT index applications.**

   Simulate Algorithm 16 at page 230 (maximal repeat computation using the bidirectional BWT index) on text `TACAGACAC`.

4. **Positional BWT.**

   Read the paper

   - Richard Durbin. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics*, 30(9): 1266-1272 (2014). `http://dx.doi.org/10.1093/bioinformatics/btu014`.

   What is the difference between PBWT and normal BWT? How is the input to the algorithm produced?

5. **Feedback.**

   a) Fill the course feedback form: `https://ilmo.cs.helsinki.fi/kurssit/servlet/Valinta`.

   b) Choose one assignment from the course from which you felt you learnt the most.

   c) What was the most difficult topic to follow?

   d) What topics you would like to learn more about?

   e) What topics you would not recommend to keep?