Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Dipl.-Inform. Klaus-Bernd Schürmann

Foundations of Sequence Analysis

Winter semester 2003/2004

Exercises

Exercise 5, Discussion: 12/15/2003.

1. (Fasta Similarity Model)

   Calculate the Fasta score for the sequences

   (a) $u = AGTGCACACATC$ and $t = ATCACACTTAGC$ for $q = 1,\ 2,\ 3$,

   (b) and $u = AGCGATAG$ and $t = AGTGACAG$ for $q = 2,\ 3$.

2. (BlastP Similarity Model)

   Let a $\mathcal{A} =$\{C, G\} be a subset of the DNA alphabet. Given a unit score function

   $$\sigma(\alpha \to \beta) = \begin{cases} 1 & : \quad \alpha, \beta \in \mathcal{A} \wedge \alpha = \beta \\ 0 & : \quad otherwise \end{cases}$$

   (a) Given a query string $w = GGCCGC$, construct the DFA concerning the BlastP Similarity Model for $q = 4$ and $k = 3$.

   (b) For arbitrary alphabet, describe some BlastP automata in your own words:

   - What is the minimal automaton for $k = 0$?
   - What is the minimal automaton for $k > q$?
   - What is the minimal automaton for $k = 1$?
   - What is the minimal automaton for $k = q$?

3. (Suffix Trees)

   (a) Draw the suffix tree for the sequence $u = ACGCGACG$.

   (b) Count the number of different substrings of u. How can the suffix tree be used for this task?