Advanced Algorithms in Bioinformatics (P4)

Sequence and Structure Analysis

Freie Universität Berlin, Institut für Informatik
Prof. Dr. Knut Reinert, Sandro Andreotti
Sommersemester 2009

8. Exercise sheet, 13. June 2009
Discussion: 19. June 2009

Exercise 1.

Prove the lemmata used for the Manhattan distance and the sum-of-pairs distance in the chaining problem, as discussed in the lecture.

Exercise 2.

(Random background in motif search)
We consider a random sequence $T = T[1 \,..\, n]$ where each $T[i]$ is chosen uniformly at random from the nucleotide alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, and the choice of $T[i]$ is independent from all other $T[j]$, where $j \neq i$.

1. Compute the probability that $T[1 \,..\, 3]$ contains $P = \mathtt{AT}$ without substitutions.
2. Compute the probability that $T[1 \,..\, 3]$ contains $P = \mathtt{AA}$ without substitutions.
3. Compute the expected number of occurences without substitutions of $P = \mathtt{AA}$ in $T[1 \,..\, n]$. Would it be different for $P = \mathtt{AT}$?

Exercise 3.

Mutual information content
Consider the following alignment:

1	2	3	4	5	6	7	8	9	10	11	12
C	A	G	C	A	A	C	A	G	A	A	U
G	A	C	C	A	C	G	A	C	C	A	G
C	A	C	C	A	G	C	A	G	G	A	C
G	A	G	C	A	U	G	A	C	U	A	A
(	.	)	(	.	)	(	.	)	(	.	)

Compute the RNA sequence logo. Compare with the results you obtain at *http://www.cbs.dtu.dk/gorodkin/appl/slogo.html*.
Compute also the mutual information content for all column pairs. Discuss the differences.

Exercise 4.

Context free RNA grammars

Consider the hairpin loop CFG from the lecture:

a) Write derivations for $s_1 =$ `CAGGAAACUG` and $s_2 =$ `GCUGCAAAGC`.

b) Write a *regular* grammar that generates s_1 and s_2 but not `GCUGCAACUG`.

c) Consider the complete language generated by the CFG from the lecture. Write a *regular* grammar that generates exactly the same language. Does this seem like a good idea?

Exercise 5.

CFGs, SCFGs and stochastic regular grammars

a) G-U pairs are accepted in base paired RNA stems but occur with lower frequency than G-C and A-U Watson-Crick pairs. Transform the hairpin loop context-free grammar from the lecture into a SCFG, allowing G-U pairs in the stem with half the probability of a Watson-Crick pair.

b) Consider a simple HMM that models two kinds of base composition in DNA. The model has two states fully interconnected by four state transitions. State 1 emits GC-rich sequence with probabilities $(p_a, p_c, p_g, p_t) = (0.1, 0.4, 0.4, 0.1)$ and state 2 emits AT-rich sequence with probabilities $(p_a, p_c, p_g, p_t) = (0.3, 0.2, 0.2, 0.3)$. (1) Draw this HMM. (2) Set the transition probabilities so that the expected length of a run of state 1 is 1000 bases, and the expected length of a run of state 2 is 100 bases. (3) Give the same model in stochastic regular grammar form with terminals, nonterminals and production rules with their associated probabilities.

c) Convert the production rule $W \longrightarrow aWbW$ to Chomsky normal form. If the probability of of the original production is p, show the probabilities for the productions in normal form version.