Advanced Algorithms in Bioinformatics (P4)

Sequence and Structure Analysis

Freie Universität Berlin, Institut für Informatik
Prof. Dr. Knut Reinert, Sandro Andreotti, Markus Bauer, Andreas Döring

Sommersemester 2008

2. Review, 2008-06-25

Name, Vorname	Matrikelnummer

Zur Bearbeitung des Reviews stehen Ihnen 50 Minuten zur Verfügung. Jeder Punkt entspricht in etwa einer Minute.

Geben Sie auf diesem Titelblatt und auf allen eventuell zusätzlich abgegebenen Blättern ihren Namen und ihre Immatrikulationsnummer an.

Schreiben Sie ihre Lösungen direkt auf die entsprechenden Aufgabenbögen. Sollte dort der Platz nicht ausreichen und Sie weitere Blätter benötigen, vermerken Sie dies bitte, damit wir auch den Rest ihrer Antwort finden und bei der Bewertung berücksichtigen können. Am Ende des Reviews sind sämtliche Aufgabenblätter wieder abzugeben.

Ergebnis:

Aufgabe	maximal	erreicht
1	14	
2	12	
3	11	
4	13	
$\sum$	50	

Exercise 1. 4+4+2+4=14 Punkte

1	2	3	4
C	A	G	A
G	A	C	C
C	A	G	G
G	A	C	U

1. Compute the sequence logo of the above alignment (in bits). Assume the a priori distribution as $p_A = p_G = p_C = p_U = 0.25$.
2. Compute the mutual information content $H(1,3)$ and $H(1,2)$ (in bits).
3. Which production rules are allowed for regular grammars and which for context free grammars?
4. Which production rules are allowed for a CFG to be in Chomsky normal form (CNF)? Convert the production rule $W \rightarrow aWbWc$ into CNF.

Exercise 2. 4 + 2 + 2 + 2 + 2 Punkte

1. What information do the inside, outside and CYK algorithm compute? What are their counterparts for HMMs?

2. Given the following SCFG:
   $S \to SV,\ S \to SVN,\ V \to VN,\ N \to NT,\ T \to PN$
   $S \to$ *wir*, $V \to$ *werden*, $N \to$ *europameister*, $P \to$ *in*, $N \to$ *wien*
   $t_S(S, V) = 0.25,\ e_S(wir) = 0.5,\ t_V(V, N) = 0.5,\ t_N(N, T) = 0.5,\ e_N(wien) = 0.25$

   - Complete the set of production probabilities.
   - Transform this SCFG into Chomsky normal form (preserving probabilities).
   - Draw all possible parse trees generating the sequence *"wir werden europameister in wien"*.
   - Compute the values $\alpha(1, 5, S)$ (result of inside algorithm) and $\gamma(1, 5, S)$ (result of CYK algorithm).

Exercise 3. 4+4+3=11 Punkte

1. For symmetric set and Hausdorff distance give example RNA structures where the distance measure performs badly (argue why).
2. Given two sequences A and B with three segment matches $S_1(A_{ij}B_{kl}), S_2(A_{i'j'}B_{k'l'})$, and $S_3(A_{i''j''}B_{k''l''})$ with $i < i' < j < j' < i'' < j''$ and $k < l < k' < l' = k'' < l''$. Visualize the situation and draw the minimal refinement. Aassume all segment matches correspond to direct matches (no reversed, no gaps in alignment).
3. When is a set of matches S called resolved (precise definition as presented in the script/lecture)?

Exercise 4. 6+4+3=13 Punkte

1. Consider a random iid distributed sequence $T = T[1 .. n]$ from alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$
   - Compute the probability that $T[1 .. 4]$ contains $P = \mathtt{ATG}$ without substitutions.
   - Compute the probability that $T[1 .. 4]$ contains $P = \mathtt{AAA}$ without substitutions.
   - Compute the expected number of occurences without substitutions for each pattern in $T[1 .. 4]$.
   - How does the variance differ between the two patterns (do not compute, just qualitatively)?
2. EM algorithm:
   - What happens in the two steps (E-step and M-step) of the EM algorithm in the general case?
   - Summarize the steps in the MEME motif finding algortihm. What is the observed data and the missing information?
3. Give a formula for the probability that a given l-mer occurs with **up to** d substitutions at a given position of a random DNA sequence.