CHALMERS UNIVERSITY OF TECHNOLOGY

Examination in Bioinformatics, MVE360

Saturday 3 March 2012, 14:00-18:00

Solutions

Updated 2012-12-07

Question 1. 4 p

a) caaat
b) aa
c) tcca
d) c, a
e) HGTQAAS
f) HXLXTQAAX
g) SLGSGSSTQAAS
h) HTLPTPTTTQAAT

Question 2. 8 p

a)
```
while ( <DATA> ) {
        if ( /TGGAAAG.*TGGAAAG/ ) {
                print;
        }
}
```

b)
```
#!/usr/bin/perl -w

$sequence = "";
$count = 0;

while ( <> ) {
        chomp;
        if ( /^[^>]/ ) {
                $sequence .= $_;
        }
}

while ( $sequence =~ /TGGAAAG(.*)/ ) {
        ++$count;
        $sequence = $1;
}

print "$count\n";
```

Question 3. 7 p

a)
```
GTTAAC
GTTGAC
GTCAAC
GTCGAC
```

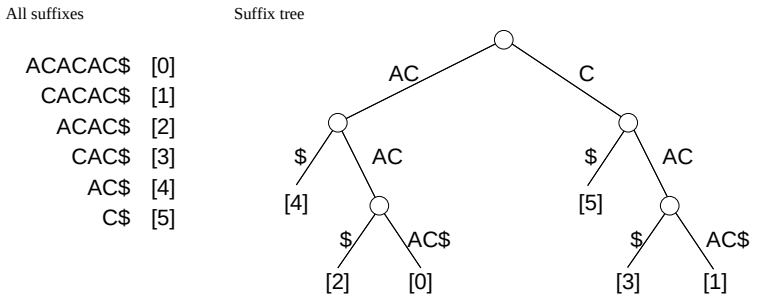
b)
```
#!/usr/bin/perl -w

$s = "AAGTTAACBBBBBGTCGACCCCC";

$s =~ s/(GT[TC])([AG]AC)/$1_$2/g;

@frags = split(/_/, $s);

print"@frags\n";
```

Question 4. 7 p

a) See: http://www.cse.chalmers.se/ kemp/cgi-bin/global_alignment.html
Optimal score is 1, giving 4 global alignments:

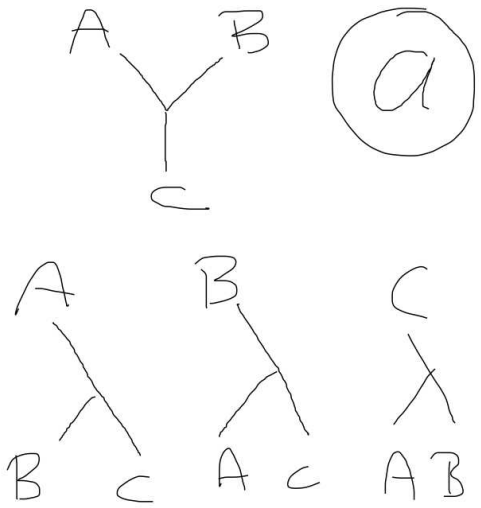
```
CATA-C    CATA-C    CATAC-    CATAC-
  || |      || |      |||       |||
-GTACC    G-TACC    -GTACC    G-TACC
```

b)



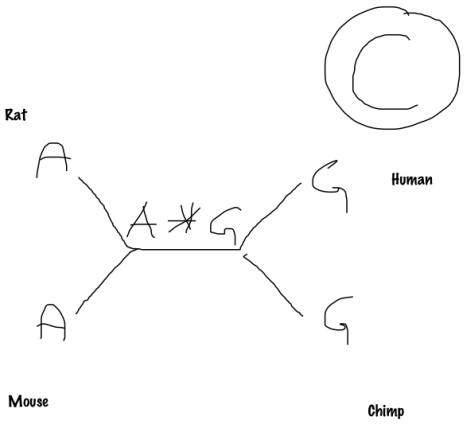
Question 5. 4 p

a) There is only one unrooted tree, and three different rooted trees, see figure



b) "Distance based" refers a method that that relies on the construction of a distance matrix, i.e a matrix showing pairwise distances for all possible pairs of taxonomic units. A typical distance measure is the number of amino acid or nucleotide substitutions between two sequences. The distance matrix is in turn used to construct a tree based on for instance the neighbor-joining method.

c) The tree shown in the figure is the most likely one because the tree may be explained by one mutational change, whereas the other two possible trees requires at least two changes.



Question 6. 14 p

a) Next generation sequencing techniques are massively parallel and can therefore sequence most bases per time unit at a lower cost. In contrast, Sanger sequencing is a serial process where single DNA fragments at one time. The error rate is however higher for next generation DNA sequencing techniques.

b) Massively parallel pyrosequencing, aka 454 sequencing, is a NGS technique with moderate throughput but long reads. The reads have an average sequencing length of up to 400 bases. Common applications are sequencing project where read length is important, for example sequencing of longer amplicons, de novo genome sequencing and de novo RNA sequencing.

Illumina sequencing is based on the sequencing by synthesis-technique and has a very large throughput. One single Illumina runt can result in more than 500 gigabases of data. The read length is however limited, and is usually less than 150 bases. Illumina is therefore applied to sequencing project where massive amount of data is needed and/or read length is not important. This includes metagenomics, genome resequencing and RNA-seq (RNA sequencing).

c) The most common type of sequencing error for massively parallel pyrosequencing is insertions/deletions after homopolymers, i.e. subsequences with repeated single base pairs. The inserted/deleted erroneous base pair is always of the same type as bases in the homopolymer. Longer homopolymers has a higher probability for sequencing errors. The error originates from the quantification of massively parallel pyrosequencing where light is detected from the insertion of complementary bases along single-stranded DNA fragments. For homopolymeric regions, multiple bases are inserted simulatenously and light is emitted with higher intensity. Wrongly quantified intensities lead to insertions/deletions. In theory, substitution errors are very rare in massively parallel pyrosequencing.

Question 7. 6 p

a) Given the current state, the future is independent of the past.

b) A Markov chain is a random process that satisfies the Markov property. A hidden Markov model is a random process comprised of two interrelated processes; a Markov chain that is hidden from the observer, and an observed process that is dependent on the underlying state sequence.

c) In a standard Markov chain, the output in each jump is a single symbol, typically the state. In a Generalized HMM the output is of random length, or duration, where the duration is chosen from some distribution.

d) The Viterbi algorithm is used to predict the most likely state sequence for the given model and the given observed sequence. The forward algorithm is used to compute the probability, or the likelihood, of the observed sequence under the given model. The forward-backward algorithm is used to train the model in an EM-algorithm procedure called the Baum-Welch algorithm.

Question 8. 4 p

a) For the Markov property to hold, we should have that

$$P(X_3|X_2, X_1) = P(X_3|X_2).$$

But, for instance we have $P(X_3 = A|X_2 = A) = 0.5$ while
$P(X_3 = A|X_1 = A, X_2 = A) = 0$, which contradicts the Markov property. Thus, the answer is NO, this is not a Markov chain.

b) Yes, this is a Markov chain. For instance,

$$P(X_4 = A|X_1 = A, X_2 = A, X_3 = C) = P(X_4 = A|X_2 = A, X_3 = C) = 1.$$

Question 9. a) Use the Viterbi algorithm to compute the optimal state sequence.
6 p

$$\delta_i(t) = \max_{1 \le j \le N} \delta_j(t-1)\, a_{ji}\, b_i(Y_t).$$

$$\pi_i = (0.5, 0.5),\ a_{ij} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix},\ b_i(Y_t) = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}$$

$$\begin{aligned}
\delta_A(1) &= 0.5 \cdot 0.6 = 0.3 \\
\delta_B(1) &= 0.5 \cdot 0.3 = 0.15 \\
\psi_A(1) &= A \\
\psi_B(1) &= B \\
\\
\delta_A(2) &= \max\{0.3 \cdot 0.8 \cdot 0.6, 0.15 \cdot 0.2 \cdot 0.6\} = 0.144 \\
\delta_B(2) &= \max\{0.3 \cdot 0.2 \cdot 0.3, 0.15 \cdot 0.8 \cdot 0.3\} = 0.036 \\
\psi_A(2) &= A \\
\psi_B(2) &= B \\
\\
\delta_A(3) &= \max\{0.144 \cdot 0.8 \cdot 0.4, 0.036 \cdot 0.2 \cdot 0.8\} = 0.04608 \\
\delta_B(3) &= \max\{0.144 \cdot 0.2 \cdot 0.7, 0.036 \cdot 0.8 \cdot 0.7\} = 0.02016 \\
\psi_A(3) &= A \\
\psi_B(3) &= A/B \\
\\
\delta_A(4) &= \max\{0.04608 \cdot 0.8 \cdot 0.6, 0.02016 \cdot 0.2 \cdot 0.6\} = 0.02212 \\
\delta_B(4) &= \max\{0.04608 \cdot 0.2 \cdot 0.3, 0.02016 \cdot 0.8 \cdot 0.3\} = 0.00484 \\
\psi_A(4) &= A \\
\psi_B(4) &= B \\
\\
\delta_A(T+1) &= \max\{0.02212 \cdot 0.8, 0.00494 \cdot 0.2\} = 0.0177 \\
\delta_B(T+1) &= \max\{0.02212 \cdot 0.2, 0.00494 \cdot 0.8\} = 0.0044 \\
\psi_A(T+1) &= A \\
\psi_B(T+1) &= A
\end{aligned}$$

Traceback gives the optimal hidden sequence AAAA.

b) The probability of the observed sequence HHTH given the hidden sequence AAAA is

$$P(Y_1^T | X_1^T) = P(H|A)P(H|A)P(T|A)P(H|A) = 0.6 \cdot 0.6 \cdot 0.4 \cdot 0.6 = 0.0867.$$