

CHALMERS UNIVERSITY OF TECHNOLOGY

Examination in Bioinformatics, MVE360

Monday 16 March 2015, 08:30-12:30

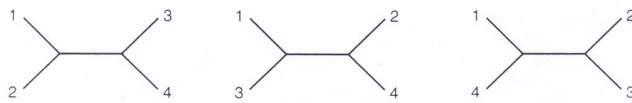
Solutions

Updated 2015-03-24

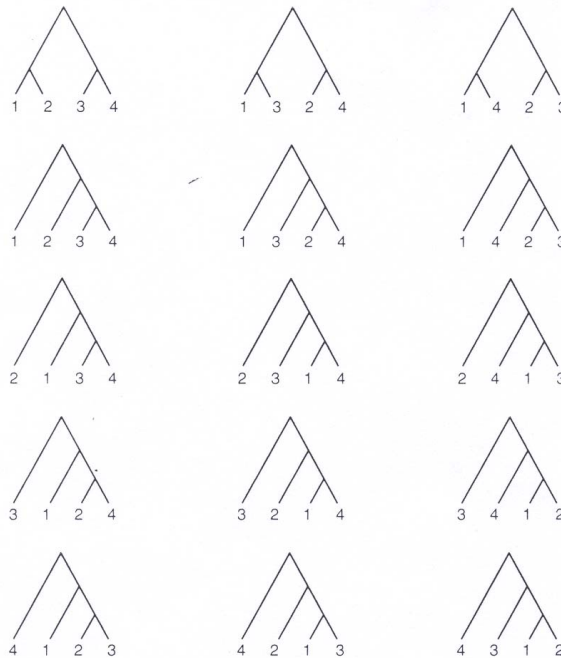
Question 1. a) All possible unrooted and rooted trees:

4 p

(a)



(b)



b) Columns 2 and 5 are informative, as they have

A
A
T
T

and

T
T
C
C,

respectively. For both of these columns the tree (AB)-(CD) will be preferred before the other two possible trees.

Question 2. Optimal score is 5, giving 1 global alignment:

4 p

RFSN-
-YTQC

Score matrix should be given in full.

Question 3. a) aabbacbc

4 p

- b) aabba
- c) acb
- d) ba, cbcc
- e) atgacttctt
- f) atxgtt
- g) aagaaaga
- h) cggccggggg

Question 4. #!/usr/bin/perl

6 p

```
$oc = "";  
  
while ( <> ) {  
    if ( /^OC (.*)/ ) {  
        $s = $1;  
        $s =~ s/[^A-Za-z;]//g;  
        $oc .= $s  
    }  
}  
  
@oc = split( /;/, $oc );  
  
$i = 1;  
foreach $name ( @oc ) {  
    print( "$i $name\n" );  
    ++$i;  
}
```

Question 5. a) *(The solution given here allows motifs involving the same 'aaag' with several different 'cttt's at different separations, and also several 'aaag's with the same 'cttt' at different separations.)*

12 p

```
#!/usr/bin/perl

$sequence = "";

while ( <> ) {
    if ( /^ / ) {
        s/[^a-z]//g;
        $sequence .= $_;
    }
}

for ( $i = 1 ; $i <= 25 ; ++$i ) {
    $count[$i] = 0;
}

while ( $sequence =~ /aaag(.*)/ ) {
    $sequence = $1;
    for ( $i = 1 ; $i <= 25 ; ++$i ) {
        if ( $sequence =~ /^(.{i}cttt)/ ) {
            print( "aaag$i: $i\n" );
            ++$count[$i];
        }
    }
}

for ( $i = 1 ; $i <= 25 ; ++$i ) {
    print "Count $i: " . $count[$i] . "\n";
}

```

```
b) print "Type in a motif (e.g. aaag): ";           # New
$s = <STDIN>;                                       # New
chomp($s);                                         # New
$r = $s;                                           # New
$r =~ tr/acgt/tgca/;                               # New
$r = reverse($r);                                  # New

while ( $sequence =~ /$s(.*)/ ) {                 # Changed
    $sequence = $1;
    for ( $i = 1 ; $i <= 25 ; ++$i ) {
        if ( $sequence =~ /^(.{i}$r)/ ) {         # Changed
            print( "$s$i: $i\n" );                # Changed
            ++$count[$i];
        }
    }
}

```

Question 6. a) The Markov property tell us that $P(X_t|X_1^{t-1}) = P(X_t|X_{t-1})$.
 5 p Remember to explain what it means in words.
 b)

Question 7. Pair HMMs are used in sequence alignment. Describe state space, parameters and draw.
 3 p

Question 8. a)
 8 p b)

Question 9. a) Paired end sequencing refers to the technique where each DNA fragment is sequenced from both ends generating two paired sequence reads with a gap in between. Since the average fragment size normally is known the average distance between the two reads is approximately known and this information can help in alignment or assembly of reads.
 14 p
 b) The DNA from the 11 tumors is sequenced with NGS, either whole genome or only the exons. The raw sequence files are quality controlled and then aligned to a reference genome sequence. Differences between the tumor sequences and the reference sequence are identified. This step is called variant calling and normally involves some statistical method. Known variants are filtered out (known variants can be found in databases such as dbSNP). The remaining variants are candidates for causing tumor development.
 c) To compensate for different amounts of starting material (DNA) loaded to the different samples before sequencing.
 d) Calculate the total amount of reads in each sample:

	Sample1	Sample2	Sample3	Sample4
Gene1	665	292	186	217
Gene2	113	29	11	61
Gene3	450	129	146	104
Gene4	71	48	11	3
Gene5	343	144	137	61
Total	1642	641	492	447

Divide each entry in the matrix with the total amount in that sample.
 Normalized matrix:

0.404998447	0.455072645	0.379069153	0.486135046
0.068757576	0.045192368	0.021593705	0.137194781
0.273976524	0.201190996	0.296943942	0.23226875
0.043448599	0.074390472	0.022854354	0.0071395
0.208818854	0.22415352	0.279538846	0.137261923