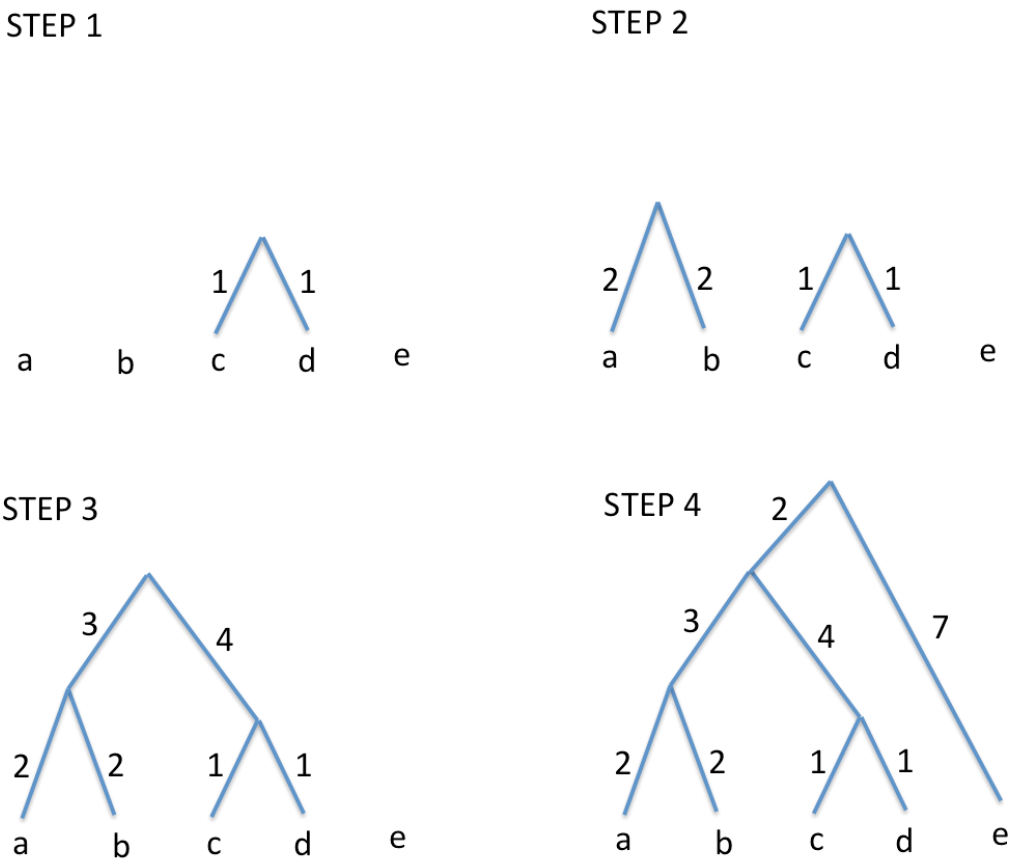
Genomics and Bioinformatics

Exam correction

December 17, 2013

Question 1 - Phylogenetic trees



Question 2 - Linear models

1. Assign value 0 to "untreated" and 1 to "treated". The model can be written $Y = \mathrm{a} + bT + \varepsilon$, where $Y \in \mathbb{R}^{12}$ are the probe intensities, $\mathrm{a} = (a, ..., a)' \in \mathbb{R}^{12}$ is the intercept, $b \in \mathbb{R}$ is the coefficient measuring the effect on Y of varying T, $T \in \mathbb{R}^{12}$ is the binary vector "treated/untreated", $\varepsilon \in \mathbb{R}^{12}$ is the measurement error.

In matrix form:

$$\begin{pmatrix} 166 \\ 121 \\ 166 \\ 270 \\ 39 \\ 121 \\ 10 \\ 14 \\ 24 \\ 14 \\ 10 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{pmatrix},$$

or $Y = X\beta + \varepsilon$.

2.
   - The p-value 0.00488 from the table is by definition the probability to observe an even bigger effect in the future, given our data, under the null hypothesis. That is, it will happen randomly 5 times out of 1000 under the null, which can -arguably- be considered as a rare event. So we can consider significant the effect of B.
   - The expected increase in intensity is the estimate of b, 57.83.

3. The normalized table:

	P1	P2	P3	P4	P5	P6
A	90	60	90	140	20	60
B	60	60	90	140	20	90
C	60	90	140	90	60	20

Question 3 - Transcription

1. RNA-seq detects fragments of mature RNA, which do not contain introns due to splicing. On the other hand, NET-seq is able to detect nascent trancripts before being spliced and modified. This is why we can see peaks in introns in the NET-seq data, but not the RNA-seq data.

2. We can deduce from figure 2d that there is a positive correlation between histone H4 hyperacetylation and antisense transcription.

3.
   - Loss of RCO1 should lead to more acetylation, and therefore more antisense transcription.
   - Figure 3b shows this effect on a genome-wide level. We can see that loss of RCO1 leads to a generally higher antisense/sense ratio.

Question 4 - DNA Binding

1. Consensus: `CATG` (score=4)

2. The next best sequences are `CCTG`, `CAGG` (score=3.5)

3. The scores are, with the best score and motif in red:

A	**T**	**A**	**G**	**C**	**C**	**T**	**A**	**G**
−6.28	0.5	−7.28	−3.64	1.5	−4.28			

4. The scores of the motifs are $W(\texttt{CATG}) = 4$ and $W(\texttt{CTGG}) = -0.64$ therefore the ratio of the number of proteins binding to each motif is:

$$\frac{n(\texttt{CTGG})}{n(\texttt{CATG})} = \frac{1000 \cdot e^{-0.64}}{e^4} = 1000 \cdot e^{-4.64} .$$

This leads to the following proportion

$$\frac{n(\texttt{CATG})}{n(\texttt{CATG}) + n(\texttt{CTGG})} = \frac{1}{1 + 1000 \cdot e^{-4.64}} = \frac{1}{11}.$$