PART II B 2010

Module 4G1, May 2010 – Computational and Systems Biology – Answers

Question 1 (Dr Lio)

(a) Gibbs Sampling is an iterative procedure that discards one l-mer after each iteration and replaces it with a new one. Gibbs Sampling proceeds slowly and chooses new l-mers at random increasing the odds that it will converge to the correct solution.

How it works:

1) Randomly choose starting positions $s = (s_1, ..., s_t)$ and form the set of l-mers associated with these starting positions.

2) Randomly choose one of the t sequences.

3) Create a profile p from the other $t - 1$ sequences.

4) For each position in the removed sequence, calculate the probability that the l-mer starting at that position was generated by p.

5) Choose a new starting position for the removed sequence based on the probabilities calculated in step 4.

6) Repeat steps 2-5 until there is no improvement.

[35%]

(b) The sequence logo uses multiple sequence alignment and provides an estimate of sequence conservation. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the information content derived from the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence. [35%]

(c) Association between adjacent bases will lead to association between more distant bases, and an estimate of how far the relations extend may be found from Markov Chain theory. Without invoking any biological mechanism, a Markov chain of order *k* supposes that the base present at a certain position in a sequence depends only on the bases present at the previous *k* positions. In particular, third and four order Markov chain models allow to predict frequencies of sequences which are not under selection. [30%]

Assessor's comment

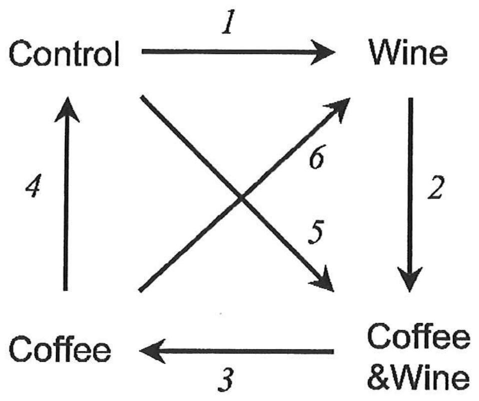
A popular and straightforward question, well-answered by most candidates.

Question 2 (Dr Barbosa-Morais)

A scientist is interested in the changes on gene expression in prostate tumours in old males introduced by the consumption of coffee and red wine. A clinical trial was performed and RNA samples were extracted from biopsies of prostate tumours of: patients who had drunk no coffee nor red wine for a month (Control), patients who had drunk two *espressos* a day but no red wine for a month (Coffee), patients who had drunk two glasses of red wine a day but no coffee for a month (Wine), and patients who had drunk both two *espressos* and two glasses of red wine a day for a month (Coffee&Wine). The samples were hybridised against a total of six two-colour microarrays. The following table summarises the experimental design, specifying the sample types hybridised against each array and the dyes used in labelling each sample:

Array #	Cy3 (green)	Cy5 (red)
1	Control	Wine
2	Wine	Coffee&Wine
3	Coffee&Wine	Coffee
4	Coffee	Control
5	Control	Coffee&Wine
6	Coffee	Wine

(a) Draw a diagram of the layout of this experimental design, using conventional arrows to represent the two-colour arrays. [20%]



By convention, we place the green/Cy3-labelled sample at the tail and the red/Cy5-labelled sample at the head of the arrow.

(b) Determine the design matrix, taking the coffee effect, the wine effect, and the interaction between wine and coffee as your independent parameters. [25%]

A design matrix must be created to reflect the expected log-ratio for each slide. These log-ratios must be a linear combination of independent contrasts of interest (parameters).

We start by representing the effects we expect to measure for each sample type (single-channel representation):

Control = *baseline*
Coffee = *baseline* + *coffee*
Wine = *baseline* + *wine*
Coffee&Wine = *baseline* + *coffee* + *wine* + *interaction*

Here we assume there is a "baseline" gene expression given by the 'Control' samples; for the 'Coffee' samples, gene expression should be given by that *baseline* plus the changes in expression associated with *coffee*; for the 'Wine' samples, it should be given by the *baseline* plus the effect of *wine*; for the 'Coffee&Wine' samples, we expect the expression to be given by the *baseline* plus the effects of both *coffee* and *wine* plus the putative effect of *interaction* between coffee and wine.

We choose *coffee*, *wine*, and *interaction* as our independent parameters and then write what is estimated in each array as combination of the parameters:

Array 1 = Wine – Control = (*baseline* + *wine*) – (*baseline*) = 1 x *wine*
Array 2 = Coffee&Wine – Wine =
= (*baseline* + *coffee* + *wine* + *interaction*) – (*baseline* + *wine*) =
= 1 x *coffee* + 1 x *interaction*
Array 3 = Coffee – Coffee&Wine =
= (*baseline* + *coffee*) – (*baseline* + *coffee* + *wine* + *interaction*) =
= -1 x *wine* + -1 x *interaction*
Array 4 = Control – Coffee = (*baseline*) – (*baseline* + *coffee*) = -1 x *coffee*
Array 5 = Coffee&Wine – Control =
= (*baseline* + *coffee* + *wine* + *interaction*) – (*baseline*) =
= 1 x *coffee* + 1 x *wine* + 1 x *interaction*
Array 6 = Wine – Coffee = (*baseline* + *wine*) – (*baseline* + *coffee*) =
= -1 x *coffee* + 1 x *wine*

We can now use the multipliers to create the design matrix:

$$X = \begin{array}{c} \left[\begin{array}{ccc} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & -1 & -1 \\ -1 & 0 & 0 \\ 1 & 1 & 1 \\ -1 & 1 & 0 \end{array}\right] \end{array} \begin{array}{l} \text{Array 1} \\ \text{Array 2} \\ \text{Array 3} \\ \text{Array 4} \\ \text{Array 5} \\ \text{Array 6} \end{array}$$

(c) Determine which effect(s) this design would allow you to estimate with higher precision. Explain the caveats of these precision estimates by discussing the concept of effective replication. [35%]

Let X be the design matrix and σ the standard deviation between slides for a particular gene. The standard error of the i^{th} parameter estimate is then given by $\sigma.\sqrt{c_i}$, where c_i is the i^{th} diagonal element of the matrix $(X^TX)^{-1}$.

For our example:

$$(X^TX)^{-1} = \begin{bmatrix} 0.5 & 0.25 & -0.5 \\ 0.25 & 0.5 & -0.5 \\ -0.5 & -0.5 & 1 \end{bmatrix}$$

$c_{coffee} = 0.5$

$c_{wine} = 0.5$

$c_{interaction} = 1$

The *coffee* and *wine* effects should therefore be estimated with higher precision than the *interaction* effect (their theoretical standard errors would be smaller by a factor of $\sqrt{2}$).

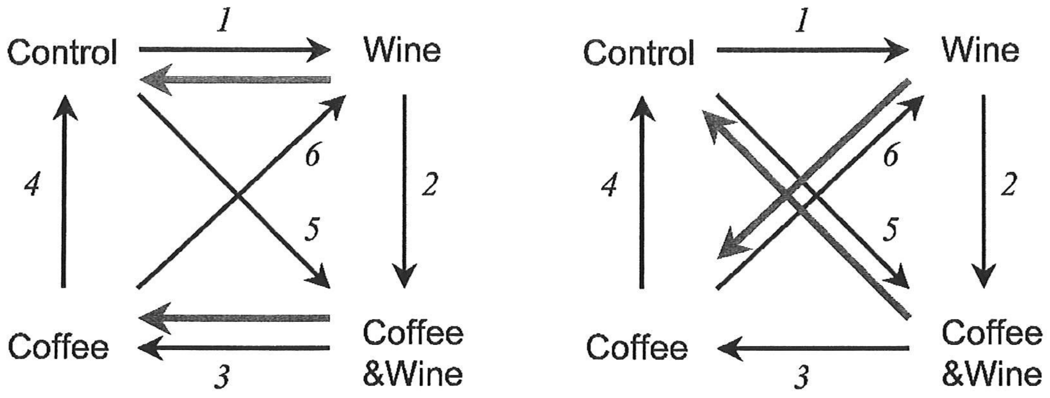
These calculations assume independence of replicates, which does not happen in reality. There is always some correlation between the expression levels of different sample types. Moreover the effective replication for each sample type depends on the correlation between replicates. The more independent the replicates, the higher the effective replication. Having highly correlated replicate arrays is equivalent to having fewer arrays on the estimation of the parameters and their variances.

(d) This design is not completely dye-balanced. Please explain why. How would you balance the design for dye if you could add two arrays and had no restrictions on the amount of any of the samples? [20%]

For each of the sample types, the numbers of replicates labelled with each of two dyes are uneven: for 'Control' and 'Coffee' there are two green-labelled and one red-labelled; for 'Wine' and 'Coffee&Wine' there are two red-labelled and one green-labelled.

If two arrays could be added, the dye-balance would be achieved with any combination of labelling one sample of 'Control' and one sample of 'Coffee' with Cy5 (red) and one sample of 'Wine' and one sample of 'Coffee&Wine' with Cy3 (green).

Below are the diagrams representing the two possible combinations, with the blue arrows depicting the two arrays to be added:



References for all the questions:
. Lecture 3 notes;
. Glonek GF, Solomon PJ. “Factorial and time course designs for cDNA microarray experiments”. *Biostatistics*, 2004 Jan;5(1):89-111. [PMID: 14744830]
. Yang YH, Speed T. “Design issues for cDNA microarray experiments”. *Nature Reviews Genetics*, 2002 Aug;3:579-588. [PMID: 12154381]

References can be found on:
http://www.compbio.group.cam.ac.uk/People/Nuno/SysBio.html

Assessor's comment:

Another popular question without any major problems.

Question 3 (DR. LESTAS)

(a) (i)

$$\frac{d}{dt} P(k,t) = \rho(k-1)\, P(k-1,t) + \lambda(k+1)\, P(k+1,t) - [g(k) + \lambda k] P(k,t)$$

(ii)

$$\frac{d}{dt}\langle x\rangle = \sum_k k \frac{d}{dt} P(k,t)$$

$$= \sum_k (k+1)\rho(k) P(k,t) + \sum_k (k-1)\lambda k P(k,t) \cancel{-[\rho(k)+\lambda k]}$$

$$- \sum_k \left[k \rho(k) + \lambda k^2\right] P(k,t)$$

$$= \sum_k \rho(k)\, P(k,t) - \sum_k \lambda k\, P(k,t)$$

$$= \langle \rho(x)\rangle - \lambda\langle x\rangle$$

At equilibrium $\frac{d}{dt}\langle x\rangle = 0 \Rightarrow \frac{\langle \rho(x)\rangle}{\langle x\rangle} = \lambda$

(iii) $\delta_x^2 = \langle x^2\rangle - \langle x\rangle^2$

$$\frac{d}{dt}\delta_x^2 = \frac{d}{dt}\langle x^2\rangle - 2\langle x\rangle\left(\frac{d}{dt}\langle x\rangle\right)$$

$$= 2\langle x\, g(x)\rangle - 2\lambda\langle x^2\rangle + 2\lambda\langle x\rangle - 2\langle x\rangle\left(\langle g(x)\rangle - \lambda\langle x\rangle\right)$$

$$= 2\langle x\, g(x)\rangle - 2\lambda\left(\langle x^2\rangle - \langle x\rangle^2\right) - 2\langle x\rangle\langle \rho(x)\rangle + 2\lambda\langle x\rangle$$

$$= 2\Big\langle x\left(\rho(\langle x\rangle) + \rho'(\langle x\rangle)(x - \langle x\rangle)\right)\Big\rangle$$

$$- 2\lambda\left(\langle x^2\rangle - \langle x\rangle^2\right) - 2\langle x\rangle\langle \rho(\langle x\rangle) + \rho'(\langle x\rangle)(x - \langle x\rangle)\rangle$$

$$+ 2\lambda\langle x\rangle$$

$$= 2\rho'(\langle x\rangle)\langle (x - \langle x\rangle)^2\rangle - 2\lambda\left(\langle x^2\rangle - \langle x\rangle^2\right) + 2\lambda\langle x\rangle$$

$$= 2\rho'(\langle x\rangle)\,\delta_x^2 - 2\lambda\,\delta_x^2 + 2\lambda\langle x\rangle$$

$$= 2\left(g'(\langle x\rangle) - \lambda\right)\delta_x^2 + 2\lambda\langle x\rangle$$

(b) (i)

$$P(ON)\, b_1 x = P(OFF)\, b_2 = [1 - P(ON)]\, b_2$$

$$\Rightarrow \quad P(ON) = \frac{b_2}{b_1 x + b_2}$$

(ii) From (a) (iii) at equilibrium

$$\sigma_x^2 = \frac{\lambda \langle x \rangle}{\lambda - f'(\langle x \rangle)}$$

$$f(x) = \frac{b_2}{b_1 x + b_2} \quad \Rightarrow \quad f'(x) = - \frac{b_2 b_1}{(b_1 x + b_2)^2}$$

So $$\sigma_x^2 = \frac{\lambda \langle x \rangle}{\lambda + \dfrac{b_1 b_2}{(b_1 \langle x \rangle + b_2)^2}}$$

(iii) linearization : small fluctuations

$$P(ON) = \frac{b_2}{b_2 + b_1 x}$$: ON/OFF process reaches equilibrium much more quickly than transcription

Assessor's comment:

This year, this question showed a dramatic improvement. It was always by far the most unpopular question. Although most students in the course had trouble following some of the mathematical concepts in the course, most questions were fully answered. This question did not involve very complicated mathematics but it is still clear that the students prefer the more descriptive and less mathematical questions.