UNIVERSITY OF
CAMBRIDGE

M. Phil. in COMPUTATIONAL BIOLOGY

Friday, 16 May, 2014   2:00 pm to 4:00 pm

COMPUTATIONAL BIOLOGY

*Attempt **ALL** questions.*

*There are **THREE** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
Calculator - students are permitted to bring an approved calculator.

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Genome Sequence Analysis

Show that in a Markov Chain the distribution of waiting times in any given state is Geometric. What does this imply for inference using a Hidden Markov Model (HMM)? What other limitations are there on using HMMs to infer model parameters and patterns in data?

The Viterbi algorithm applied to a HMM involves storing values of the following variable:

$$\psi_n(i) = \underset{i_{n-1}}{\operatorname{argmax}} \, A_{i_{n-1}i} \, \delta_{n-1}(i_{n-1})$$

What does $A_{i_{n-1}i}$ represent here?

In terms of the number of hidden states and the length of the observed sequence, how many values $\psi_n(i)$ are there in total? If each value can be stored in one byte of memory, and the data contains 200 million observations, what is the maximum number of hidden states we can use if performing this calculation on a machine with 4 GB of memory? (Assume the whole calculation is done in memory, and ignore other storage requirements.)

2 Genome Informatics

1. There are many functional elements encoded in an eukaryotic genome. Complete the following statements:

   (a) There are __a__ and __b__ genes in the genome. An example of the latter are pseudo-genes and genes for certain RNAs, for example __c__ that bind to 3 UTRs in the mRNA of other genes and lead to their degradation. On the other hand, there are structural RNAs like __d__ or __e__, both of which are involved in translation.

   (b) Immediately upstream of a gene is its __a__. This is where the Pol II complex binds, which is a group of proteins that __b__ DNA into RNA. Activation and repression of transcription is mediated by site-specific transcription factors, which can bind in multiple combinations to __c__ in the genome. These sites are not always accessible, as the DNA is wound around __d__. There are chemical marks on those that may determine the level of accessibility, these marks are also called __e__.

2. One of the traditional subjects of genome informatics is gene finding. Especially in the context of the most common gene class (which one?), these methods rely on identifying both indicative signals and content. List five of them and assign them to either signal or content.

3. The interpretation of genome-wide protein binding data is largely dependent on the platforms used in the experiments. Can you explain the reason for the observations described here: “We havent seen any binding of TF X in our yeast promoter array dataset. However, gene expression changes were clearly visible after depletion of X, but for gene Y, only for one particular cDNA on a cDNA array. Later, we thought TF X might bind to genes when we moved to Affymetrix or Nimblegen arrays. However, through studies in higher metazoans we now know it has nothing to do with binding the gene itself, its just the typical first intron binding. We now also know that the transcriptional changes only apply to particular splice forms, as shown by RNA-seq.”

3 Network Biology

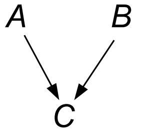
1. For a discrete Bayesian network G a closed form for calculating its likelihood $p(D \mid G)$ for some counting data D exists.

   (a) [10%] The equation for the likelihood as discussed in the lecture and coursework is

$$P(n(D) \mid G, \alpha) = \prod_{i=1}^{N} \prod_{j=1}^{s_i} \left(\frac{\Gamma(\sum_{k=1}^{r_i} \alpha_{ijk})}{\Gamma(\sum_{k=1}^{r_i} (n_{ijk} + \alpha_{ijk}))} \prod_{k=1}^{r_i} \frac{\Gamma(n_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \right)$$

   where α_{ijk} are prior parameters. What do N, n_{ijk}, s_i, r_i stand for? What do the indices i, j, k stand for? What is s_i if i has no parents?

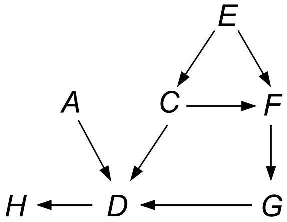
   (b) [20%] Given the following network and table



A	B	C
1	1	0
0	1	1
1	1	1
0	1	1

   what is the likelihood of the network using the above equation and assuming $\alpha_{ijk} = 1$ and that all nodes have levels 0 and 1. Write it down in terms of arguments to the Γ function. You may want to consider the contribution of each node in turn and don't forget that of A and B. Start with putting together a table with i, j, k, and n_{ijk} as columns.

2. Consider the following directed graph



   (a) [5%] List all triplets of nodes forming a v-structure.

   (b) [10%] Confirm or disconfirm the following statements about d-separation in the graph ($\perp$ means d-separated or independent, $\not\perp$ not d-separated or dependent), by listing blocked/unblocked paths:

      i. $A \perp C \mid H$
      ii. $C \not\perp E \mid \{F, H\}$
      iii. $\{D, G\} \perp E \mid \{C, F\}$
      iv. $E \perp A \mid \{C, H\}$

   (c) [15%] Draw either a Bayesian network or an undirected Markov network so that the following independencies coincide exactly with d-separation (Bayes net) or graph separation (Markov net):

i. Four nodes A, B, C, D: $A \perp B \mid \{C, D\}$ and $C \perp D \mid \{A, B\}$, all other possible (conditional) relationships are dependencies.

ii. Three nodes A, B, C: $A \perp B$, all other possible (conditional) relationships are dependencies. If you decide on one type of network (Bayesian or Markov) argue that the other type cannot realise exactly this independency and no more.

3. For the following we assume the likelihood $P(D \mid G)$ of a Bayesian network G for data D is available in closed form.

(a) [10%] List the steps of a Metropolis-Hastings (MH) MCMC algorithm for sampling from $P(G \mid D)$ using the acceptance ratio below.

(b) [10%] Explain the components in the acceptance ratio calculation

$$\alpha = \min(1, \frac{P(D \mid G_{\text{new}})P(G_{\text{new}})\,Q(G_{\text{old}} \mid G_{\text{new}})}{P(D \mid G_{\text{old}})P(G_{\text{old}})\,Q(G_{\text{new}} \mid G_{\text{old}})}$$

(c) [10%] Describe the details of a proposal step adding a directed edge to G_{old}. Explain how to obtain the probability for this proposal.

(d) [10%] Briefly describe a hill-climbing strategy to find a high scoring network and its advantages and disadvantages over an MCMC simulation.

END OF PAPER