Second Midterm Exam

BioE131

November 16, 2011

Email your numbered answers to `ihh@berkeley.edu` as a **single email** with the exact string `MidtermAnswers` in the subject line (no space between the words).

Only your best 5 answers will be counted. The maximum length of your email should be about 1,500 words.

Where appropriate, support your answers with references to the scientific literature.

Answers must be received by 9am the morning after the exam is distributed.

1 RNA folding

1. Write out a form of Nussinov's algorithm [Nussinov et al., 1978] that enforces the constraint that any two (Watson-Crick) paired bases must have at least K intervening bases.
2. What is Kinefold? How does it work?

2 Pairwise alignment

1. In scoring schemes for pairwise sequence alignment, what is a *substitution matrix*?
2. What is a *log odds-ratio substitution matrix*? How does it relate to the probability distribution $p(x, y)$ over columns of the form $\begin{array}{|c|}\hline x \\ \hline y \\ \hline\end{array}$ in a pairwise alignment?
3. What is *Hamming distance*? How does it relate to dynamic programming algorithms for pairwise alignment?
4. How (if at all) does Hamming distance relate to a log odds-ratio substitution matrix?

3 Bayes' Theorem

Let X be a DNA aptamer of uniform composition and length L. Assume that it may be modeled as an I.I.D. sequence.

N mutants of X are generated by process Y, which uses an error-prone polymerase: some number ML of randomly selected sites are mutated (with $0 < M \ll 1$). The polymerase is more likely to induce transitions than transversions (the transition/transversion ratio is R), so $\frac{ML}{1+R}$ of the sites experience transversions. You may assume that the errors induced by the polymerase are independent from one position to the next, so this amounts to a point substitution process with an independent substitution probability at each site.

N mutants of X are generated by a second process, process Z, which uses a computer to mutate a similar number (ML) of randomly selected sites, substituting the existing nucleotide at each site with a *different* nucleotide selected at random from the 3 possibilities. The mutated sequence is then synthesized.

The $2N$ mutant sequences (N from process Y, and N from process Z) are mixed and then tested in a high-throughput assay for binding to a substrate of interest. The best-performing mutant is selected; call this B.

Variant B is sequenced and is found to differ from X at $K < L$ positions, of which $J < K$ are transversions.

Let W denote the process by which B was generated. This could be Y or Z; however, we do not observe this directly.

1. What does "I.I.D." mean?
2. What is the *a priori* probability, $P(W = Y)$, that a mutant in the pool of $2N$ was generated using the polymerase (process Y)?
3. Give an expression for the likelihood $P(B|W = Y)$ in terms of J, K, L, M, N, and/or R.
4. Give an expression for $P(B|W = Z)$. Comment on this expression's usage (or not) of J, K, L, M, N & R.

5. Give an expression for the evidence $P(B)$ in terms of $P(W=Y)$, $P(B|W=Y)$ and $P(B|W=Z)$.

6. Derive the *a posteriori* probability, $P(W=Y|B)$, that mutant B was generated using process Y.

4 Probability and Information

In a given column (C) of a K-row protein multiple alignment, each amino acid x appears $n(x)$ times, so $\sum_x n(x) = K$. Let $p(x)$ be a probability distribution over amino acids. Treat the column as K independent draws from p.

1. What is the log-likelihood $L_C = \log P(C)$ of observing the column?

2. What is the log-likelihood $L_n = \log P(n(1), n(2) \ldots n(K))$ of obtaining the summary counts?

3. Define $L = L_C$ as above. Suppose that the probabilities $p(x)$ are free variables, subject to the constraint that $\sum_x p(x) = 1$. Derive the probabilities $p(x)$ that maximize L. Comment on the interpretation of L in this case.

4. Biologically, what does it mean (a) if L is low for some columns but not others, (b) if L is low for all columns, (c) if L is high for all columns?

5. Describe an information-theoretic statistic that could be used to test for covariation between two columns in the alignment (that is, to reveal whether the amino acid in column C_i is predictive of the amino acid in column C_j).

6. What does this information-theoretic statistic tell you about data compression of this multiple alignment?

5 Multiple alignment

Supporting your answer with reference to at least one paper from the literature (as close as possible to the current state-of-the-art), discuss the time **and** memory complexity of aligning N sequences, each of which has length L.

6 Phylogeny

Consider a stochastic process with parameters $\theta = (\pi_A, \pi_C, \pi_G, \pi_T, \rho)$. The four states of the process are $\{A, C, G, T\}$. At time $t = 0$, the initial state X_0 of the process is sampled from the initial probability distribution $(\pi_A, \pi_C, \pi_G, \pi_T)$. In any subsequent infinitesimal time interval $[t, t+dt)$, the process experiences a "replacement event" with probability $\rho \cdot dt$. (That is, ρ is the expected instantaneous rate of replacement events.) Immediately after a replacement event, i.e. at time $t + dt$, the state of the process X_{t+dt} is sampled from the same initial probability distribution $(\pi_A, \pi_C, \pi_G, \pi_T)$, independently of the state X_t before the replacement event. (There is, consequently, a finite probability that a "replacement event" will be "silent", i.e. $X_{t+dt} = X_t$, yielding no directly observable change.)

1. At any given time t, what is the probability $p_x(t) = P(X_t = x)$ that the state of the system at time t will be x?

2. In a given finite time interval $[0, t)$ what is the expected number of replacement events?

3. If the system is in state x, what is the expected instantaneous rate of silent replacement events?

4. In a given finite time interval $[0, t)$ what is the expected number of silent replacement events?

5. Suppose the units of "time" in this model are to be calibrated that the expected rate of substitution events (i.e. non-silent replacement events) is equal to one. What effective constraint does this place on the parameters θ?

6. This sort of model is known as a discrete-state continuous-time Markov chain. Find the instantaneous rate matrix $\mathbf{R}$ for this model, i.e. the 4×4 matrix such that the matrix differential equation $\frac{d\mathbf{p}}{dt} = \mathbf{pR}$ holds, where

$$\mathbf{p}(t) = \left[\begin{array}{c} P(X_t = A) \\ P(X_t = C) \\ P(X_t = G) \\ P(X_t = T) \end{array} \right]$$

References

[Nussinov et al., 1978] Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J. (1978). Algorithms for loop matchings. *SIAM Journal of Applied Mathematics*, 35:68–82.