

**UIC BIOINFORMATICS  
PH.D. QUALIFYING EXAM**

**JUNE 2, 2003, 8:00 AM – 12:00 PM**

**Name: \_\_\_\_\_**

There are 9 problems to this exam. Work 6 problems, but do not choose both Problems 3 and Problem 8. Grade will be given based on the 5 problems with best scores. Do all written work on the exam. Circle the number of the 6 problems that you wish to be graded.

You may use a non-programmable calculator. You are not allowed to use programmable calculators.

1. (a) (5 points.) For the multiple alignment:

ACA---ATG  
TCAACTATC  
ACAC--AGC  
AGA---ATC  
ACCG--ATC

- i. Use match state(s) and insertion state(s) to design an architecture of a hidden Markov model that characterizes this sequence family.
- ii. What are the parameters associated with your hidden Markov model? What is your estimated value for each of these parameters?
- iii. What are the probabilities of sequences TGCT--AGG and ACAC--ATC belonging to this sequence family? What conclusion can you draw from these probabilities?

Please show your work.

- (b) (5 points.) Align the sequences  $A = \mathbf{ggaatgg}$  and  $B = \mathbf{atg}$ , using dynamic programming for global alignment and the following simple scoring scheme: match = 0, mismatch = 20, insertion or deletion = 25. Show:
- i. The matrix of dynamic programming after initialization of the first row and first column.
  - ii. The complete filled matrix.
  - iii. The trace-back path on the matrix.
  - iv. List all alignments with optimal score.



2. (a) (3 points.) Assuming nucleotide substitution can be modeled as a reversible continuous time Markov process, the instantaneous rates are known, and substitutions are site independent. Given the topology of a phylogenetic tree  $T$  of 10 taxa shown in Fig 3, where each taxon (sequence) has  $m$  nucleotides, and the branch lengths are known (*eg.*, denote the edge length between node 1 and 10 as  $d_{1,10}$ ), write down explicitly the full likelihood function of the phylogenetic tree. Do not write in recursive form.

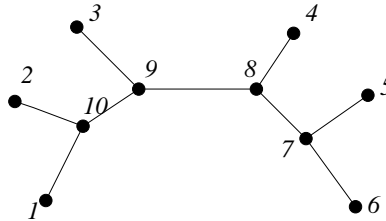


Figure 1:

- (b) (4 points.) How many unrooted binary phylogenetic trees a taxa of  $n$  sequences can have? Show your proof.
- (c) (3 points.) In Metropolis Monte Carlo method, the final equilibrium distribution reached after the chain convergency is the desired target contribution:

$$\int \pi(\mathbf{x})A(\mathbf{x}, \mathbf{y})d\mathbf{x} = \pi(\mathbf{y}),$$

where  $\mathbf{x}$  is the state variable,  $A(\mathbf{x}, \mathbf{y}) = T(\mathbf{x}, \mathbf{y}) \cdot r(\mathbf{x}, \mathbf{y})$  is the actual transition function, the product of the proposal function  $T(\mathbf{x}, \mathbf{y})$ , and an acceptance-rejection rule  $r(\mathbf{x}, \mathbf{y})$ . The proposal function  $T(\mathbf{x}, \mathbf{y})$  suggests a possible move from  $\mathbf{x}$  to  $\mathbf{y}$ . The acceptance-rejection rule decides whether the proposed move to  $\mathbf{y}$  will be accepted: Draw a random number  $u$  from the uniform distribution  $U[0, 1]$ . If  $u \leq r(\mathbf{x}, \mathbf{y})$ , the move is accepted and  $\mathbf{y}$  is taken as the new position. Otherwise stay with  $\mathbf{x}$ .

In the original Metropolis Monte Carlo method, the proposal function is symmetric:  $T(\mathbf{x}, \mathbf{y}) = T(\mathbf{y}, \mathbf{x})$ , and the acceptance-rejection rule is:  $r(\mathbf{x}, \mathbf{y}) = \min\{1, \pi(\mathbf{y})/\pi(\mathbf{x})\}$ . Since the target distribution is the Boltzmann distribution  $\pi(\mathbf{x}) \sim \exp(-h(\mathbf{x}))$ , where  $h(\mathbf{x})$  is an energy function, the acceptance rule is often written as:  $u \leq r(\mathbf{x}, \mathbf{y}) = \exp(-[h(\mathbf{y}) - h(\mathbf{x})])$ .

Hastings first realized that the proposal distribution does not need to be symmetric, but can be arbitrarily chosen so long as the condition of detailed balance is satisfied. His generalization leads to the modified acceptance rule:

$$u \leq r(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})T(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})T(\mathbf{x}, \mathbf{y})}\right\},$$

so more flexible and efficient sampling strategy can be developed that still generates samples following the desired target distribution. Answer the following questions and show your proofs:

- i. Show that Hasting's rule satisfies the detailed balance condition  $\pi(\mathbf{x})A(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})A(\mathbf{y}, \mathbf{x})$ .
- ii. Why does Hasting's rule works? That is, why is the equilibrium distribution the same as the desired target distribution?
- iii. According to the same principle, will the following trial acceptance rule work?

$$u \leq r(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{x})T(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{y})T(\mathbf{y}, \mathbf{x}) + \pi(\mathbf{y})T(\mathbf{x}, \mathbf{y})}\right\}$$

- iv. How about the next rule below?

$$u \leq r(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})T(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{y})T(\mathbf{y}, \mathbf{x}) + \pi(\mathbf{x})T(\mathbf{x}, \mathbf{y})}\right\}.$$



3. (a) (5 points. You cannot choose both this problem and Problem 8). Let  $X$ ,  $Y$  and  $Z$  be independent and uniformly distributed on  $[0, 1]$ .
- i. Find the joint density function of  $XY$  and  $Z^2$ .
  - ii. Show the simplest functional form you can find for  $\mathbb{P}(XY < Z^2)$ .
- (b) (5 points). Let  $X$  and  $Y$  have the bivariate normal distribution with zero means, unit variances, and correlation  $\rho$ . Find the joint density function of  $X + Y$  and  $X - Y$ , and their marginal density functions.



4. (10 points). Consider the maximization of the function  $p(\theta_1, \theta_2)$ :

$$p(\theta_1, \theta_2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{(x_i - \theta_1)^2}{2\theta_2}\right),$$

where  $x_i$  ( $i = 1, \dots, n$ ) are given.

(a) (5 points.) Derive the first order optimization condition, and represent the local optimal solution  $(\theta_1^*, \theta_2^*)$  with  $x_i$  ( $i = 1, \dots, n$ ).

(b) (5 points.) Let

$$(x_1, \dots, x_{15}) = (-0.29, 1.57, -0.55, 0.31, 1.46, -3.03, 1.09, 0.24, -1.19, 0.16, -1.43, \\ -1.08, -0.46, 0.62, 0.53).$$

Calculate the optimal solution of the function  $p(\theta_1, \theta_2)$ .



5. (10 points). Given a set of points  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  in some  $d$ -dimensional space, consider a cost-based clustering method. The objective is to minimize the sum-of-squared error criterion  $J_e$ , defined as

$$J_e = \sum_{i=1}^c \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2,$$

where  $c$  is the number of clusters,  $D_i$  ( $i = 1, \dots, c$ ) are the corresponding clusters and  $\mathbf{m}_i$  is the mean of each cluster, i.e.,

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}_i, \quad n_i : \text{number of points in } D_i.$$

The following is an iterative optimization algorithm to minimize  $J_e$  starting from randomly selected cluster means.

**Algorithm : Basic Iterative Minimum-Squared-Error Clustering**

**Step 0.** Initialize  $c, \mathbf{m}_1, \dots, \mathbf{m}_c$ ;

**Step 1** Repeat

randomly select a sample  $\hat{\mathbf{x}}$ ;

$i \leftarrow \operatorname{argmin}_{i'} \|\mathbf{m}_{i'} - \hat{\mathbf{x}}\|$

if  $n_i \neq 1$  then compute

$$\rho_j = \begin{cases} \frac{n_j}{n_j+1} \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2, & j \neq i \\ \frac{n_j}{n_j-1} \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2, & j = i \end{cases}$$

if there is an  $k \neq i$  such that  $\rho_k \leq \rho_j$  for all  $j$ , then transfer  $\hat{\mathbf{x}}$  to  $D_k$ ;

Recompute  $J_e, \mathbf{m}_i, \mathbf{m}_k$ ;

Until no change in  $J_e$  in  $n$  attempts;

**Step 2** Return  $\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_c$ .

Demonstrate that  $\rho_k \leq \rho_j$  for all  $j$  in **Step 1** above is the condition of improving the criterion  $J_e$  when transferring  $\hat{\mathbf{x}}$  from  $D_i$  to  $D_k$ .



6. (10 points). The Fibonacci numbers were originally defined by the Italian mathematician Fibonacci as the series given by the following recurrence relation :

$$F_n = F_{n-1} + F_{n-2}$$

with basis cases  $F_0 = 0$  and  $F_1 = 1$ . Thus  $F_2 = 1, F_3 = 2$ , and the series continues  $\{3, 5, 8, 13, 21, 34, 55, 89, 144, \dots\}$  and  $F_n/F_{n-1} \cong (1 + \sqrt{5})/2 \cong 1.61803$ .

- (a) (2 points.) Write a recursive program (in Pseudocode) to compute the  $n$ th Fibonacci number.
- (b) (4 points.) Evaluate the number of recursive calls needed for the computation of  $F_n$  in terms of  $n$ .
- (c) (4 points.) Write a linear time algorithm for the same computation.



7. (10 points). Comparing Molecular dynamics (MD) and Monte Carlo (MC) simulation:
- (a) (2.5 points). With continuous potentials, molecular dynamics trajectories are generated using finite difference techniques. Derive the verlet algorithm, and explain how to calculate velocity and acceleration.
  - (b) (1.5 points). How to choose the time step in molecular dynamics simulations.
  - (c) (3.0 points). Explain the Metropolis method in Monte Carlo simulation. Using lattice model of the polymers as example, describe the procedure of Metropolis method including the choice of move set and temperature selection.
  - (d) (3.0 points). How to decide which method we should use? Give two cases MD is better than MC, and two cases MC is better than MD.



8. (10 points. You cannot choose both this problem and Problem 3).

- (a) (2.0 points). Explain the genetic code; translation; transcription.
- (b) (1.0 points). Formation of peptide bond
- (c) (3.0 points). List and describe the forces that stabilize proteins native structures.
- (d) (2.0 points). Describe the secondary structures of protein.
- (e) (2.0 points). Draw the  $\phi$  and  $\psi$  angles.



9. (10 points). The random variable  $x$  has pdf as following:

$$p(x; \theta) = \theta^2 \cdot x \cdot \exp(-\theta x) \cdot u(x)$$

where  $u(x)$  is the unit step function:

$$u(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x < 0 \end{cases}$$

Given  $N$  measurements,  $x_1, x_2, \dots, x_N$  of  $x$ , calculate the maximum likelihood estimate of  $\theta$ .

