Name_____Student ID_____

BioE 131/231 MidTerm Exam

October 20, 2003

1. Monte Carlo (30 points).

(a) Give pseudo-code for a general Simulated Annealing algorithm for a protein chain. The pseudo code should show what variables are initialized, describe a trial move, a cooling schedule, and acceptance criterion.

```
Initialize Rcurr, Ecurr(Rcurr), Tcurr=Thigh
Do I=1,Ntsteps
   Beta=1/(kTcurr)
    Do J=1,NMCsteps
       Make trial move (Rtrial)
       Evaluate Energy(Rtrial)
       Boltz=exp(-Beta*(Energy-Ecurr))
       URN=rand(iseed)
       If (URN .lt. Boltz) then
         Ecurr=Energy
         Rcurr=Rtrial
       Else
          Continue
      End do
    End J
   Tcurr=T-Delta (or some other cooling schedule)
  End I
```

(b) Write down the detailed balance expression and explain why is it important.

$P(x)T(x \rightarrow y)=P(y)T(y \rightarrow x)$

Ensures that moves and their acceptance conform to the correct limiting sampling distribution.

(c) Write down expressions for Monte Carlo and molecular dynamics which show their reliance on a potential energy surface

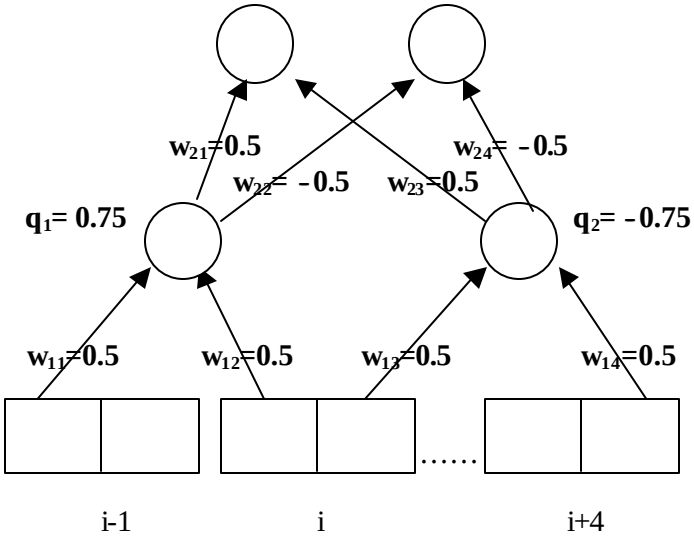
Monte Carlo through $P(x)/P(y)=\exp(-\beta * \Delta E)$

Moelcular dynamics through Hamiltonian in Hamilton's equation of motion or numerically as

$$R(t+\Delta t)=-R(t-\Delta t)+2R(t)+(F(t)/m)(\Delta t)^2$$

where $F(t)=-dE/dR$

2. Supervised Learning (40 points). Neural networks have been used for secondary structure prediction: the prediction of alpha-helix, beta-sheet, or coil given the amino acid sequence as input. Each amino acid is represented by two numbers, the first representing its propensity to be hydrophobic (+1) or hydrophilic (-1), while the second is its propensity to form helix (+1) or not helix (-1). A helix is predicted by the network if the output is (1,-1), β−sheet if output is (-1,1), and coil if (-1,-1). We could design two simple Boolean functions as part of a bigger network with the following connectivity:



(a) It is found that successful helix (and not helix) prediction is enhanced when:

amino acid i OR its i+4 partner has a high helix propensity
AND
amino acid i AND i-1 are similarly hydrophobic.

Label the weights and thresholds on the above network consistent with this observation. Also label the layers on the above network as input, hidden layer, and output. Assume a heaviside response function.

(b) Given the following input values for one pattern:

i-1=(-1,1) i=(-1,-1) i+4=(1,-1)

Feedforward through the above network (with your determined weights and biases from (a)) and give the calculated output for this pattern and corresponding secondary structure definition.

Output: -1, +1: b−sheet

(c) The actual *observed* output for this one pattern is (-1,-1). *Define the error and calculate it.*

$$E=(-1.0-(-1.0))^2+(-1.0-1.0)^2=4.0$$

(d) What could you do to improve the fidelity of the network? Give a formula for the weight update of w21, but do not calculate.

Back-propagation, Hebb's rule.

$$w_{21}(new)=w_{21}(old) -g\,dE/dw_{21}$$

3. Probability and statistics (20 points). Suppose that the population consists of 48% males and 52% females. 1% of females are color blind and 10% of males are color blind.

(a). Construct the conditional probabilities for Boolean variables, M (i.e *Male and not Male)* and CB (i.e. *Colorblind and not Colorblind)*.

P(M)=0.48 **P(not M)=0.52**
P(CB|M)=0.1 **P(CB| notM)=0.01**

(b) A color blind person is chosen at random. What is the probability of this person being male?

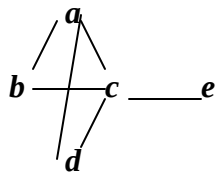
P(M|CB)=P(CB|M)*P(M)/P(CB)
P(CB)=P(CB|M)*P(M)+P(CB|notM)*P(not M)=0.1*0.48+0.01*0.52=0.048+0.0052=0.0532
P[M|CB]=0.048/0.0532=0.90

4. Phylogeny (20 points) Consider a phylogeny construction problem in which there are 5 species (A,B,C,D,E) and 5 binary characters (a,b,c,d,e). The following table gives the character states:

	a	b	c	d	e
A	1	1	1	0	1
B	0	1	0	1	0
C	1	0	1	1	1
D	1	0	1	0	0
E	0	1	1	1	1

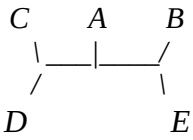
Give a phylogeny consistent with the largest possible number of characters.

(1) *a,b* *b,c* *c,d*
a,c *c,e*
a,d

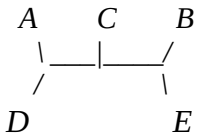


(2) maximum clique of 3

(3) Take a,b,c: *ACD from BE*
ABE from CD
ACDE from B



Take a,c,d: *ACD from BE*
ACDE from B
AD from BCE

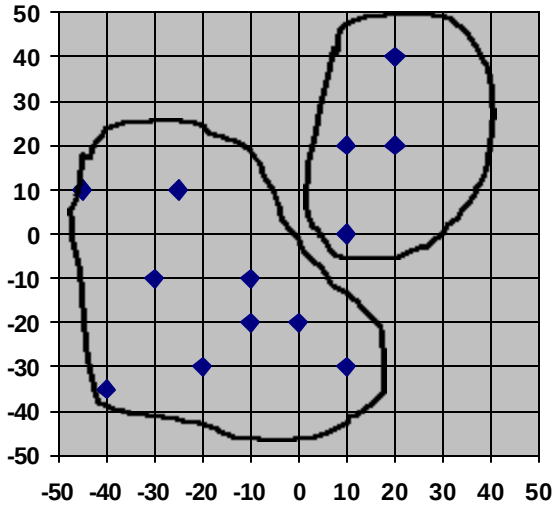


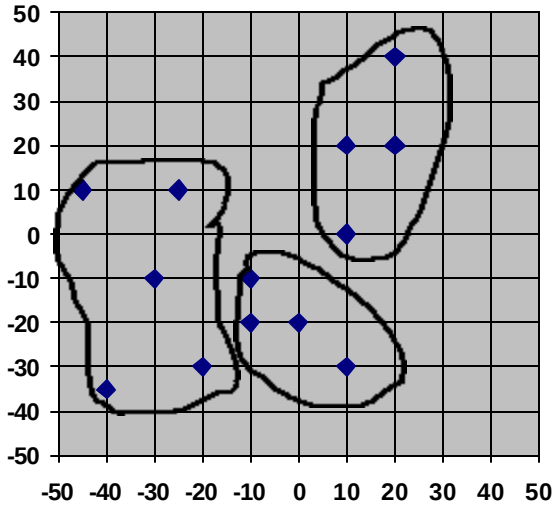
5. Microarrays (30 points). A publicly available microarray data set is given by Golub et al. (1999), who performed a study of gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The training dataset we will work with is a subset of that data, and comprises 9 cases of ALL and 4 cases of AML. The ALL class actually comprises two subclasses: T-cell All (5 cases) and B-cell ALL (4 cases).

(a) What is the purpose of cluster analysis?

The purpose of clustering is that we can sort elements of a set into groups based on their similarity (belonging to the same group) or dissimilarity (belonging to different groups). Clustering seeks a correlative relationship among genes driven by a hypothesis, but sometimes simply looking at different ways of grouping genes to "discover" new hypotheses.

(b) Consider the following hypothetical gene expression patterns exhibited by the 13 leukemia cases. Simply by eye, cluster the genes into two tumor classes (AML and ALL), and show that clustering on the left-hand graph. Simply by eye, cluster genes into three tumor classes (AML, T-ALL, B-ALL) and show that clustering on the right-hand graph.





(c) Describe a partitioning method that allows us to perform the above clustering with a specified number of clusters K=2 or K=3. Give pseudo-code for this type of clustering method.

Partitioning around mediods:

(1) Choose the number of clusters, K
(2) extremize some criteria to find the best division of the data into these K cluster
(3) swap members
(4) select optimal number of clusters by using silhouette plot