Problem Set 4

Problem 1: Clustering (33 points)

Microarray and DNA chip technologies have made it possible to study expression patterns of thousand of genes simultaneously. The amount of data coming out of these efforts is overwhelming. A powerful strategy for analysis of microarray data is the clustering of expression profiles. Expression profiles can be clustered by gene or by condition. Golub *et al.* (*Science*, **286**, 531-7. pdf, supplemental website) clustered different types of leukemia expression data using non-hierarchical Self-organizing Maps (SOMs). Now you will write a Perl program to cluster the same data using an alternative hierarchical clustering algorithm.

I) I) Briefly describe the two major goals of this paper. (2 pts)

a. a. Cancer class discovery (1 pt)
b. b. Cancer class prediction (1 pt)

II) II) Describe the major steps of the SOMs training algorithm without using code. (4 pts)

a. a. Define map: define the topological relations and the number of neurons according to the input data and expected number of clusters (1 pt)
b. b. Initialization: initialize the weight vector with random sample vectors from the training dataset (1 pt)
c. c. Random selection: randomly choose one sample vector from the input dataset, and calculate similarity measure between it and all weight vectors in the map (1 pt)
d. d. Update map: find the weight vector that has the greatest similarity with the input vector, and update the surrounding weight vectors (1 pt)
e. e. repeat step c and d for predefined number of steps

III) III) The authors used Affymetrix GeneChip, which is very different from ratio-based cDNA microarray in the way of measuring expression level of RNA. Data from several different GeneChip microarrays should be normalized before being compared to each other. Describe why normalization is needed, and how the authors normalized their data. (4 pts)

a. a. Affymetrix GeneChip is a ‘one-channel’ platform where only one fluorescent dye is used. Expression levels are determined by the difference of fluorescent intensities between the ‘perfect match’ (PM) probes and ‘mismatch’ (MM) probes for each gene, and absolute values are reported rather than ratios as in the cDNA microarray. The overall brightness (intensity) of a chip may vary from experiment to experiment due to various reasons ranging from sample preparation to chip scanning. Therefore normalization to make all chips into the same brightness is needed to compare these absolute values across different experiments. (2 pts)

b. b. Quoted from http://www-genome.wi.mit.edu/mpr/publications/projects/Leukemia/Files_descriptions.txt: “Intensity values have been re-scaled such that overall intensities for each chip are equivalent. This is done by fitting a linear regression model using the intensities of all genes with "P" (present) calls in both the first sample (baseline) and each of the other samples. The inverse of the "slope" of the linear regression line becomes the (multiplicative) re-scaling factor for the current sample. This is done for every chip (sample) in the dataset except the baseline which gets a re-scaling factor of one.” (2 pts)

IV) IV) A brief summary of the hierarchical clustering algorithm that you are asked to implement can be found here. Your assignment is to cluster the normalized expression data of 50 predictor genes from Golub *et al.* using the single-linkage and complete-linkage Euclidean distance metrics. (11 pts)

a. a. Partial credits are given for the following tasks:

i. i. Reading input data (2 pts)
ii. ii. Constructing distance matrix (2 pts)
iii. iii. Updating distance matrix (3 pts)
iv. iv. Output clustering result (4 pts)

b. b. Use the sample dataset of 5 samples and its clustering result to verify your code. Print the group members and distance matrix at each iteration.

c. c. Please attach your “well-annotated” Perl code to the end of your problem set. You may use this template (ps4-1-template.pl) for your program.

Sample dataset of 5 samples:
http://www.courses.fas.harvard.edu/~bphys101/problemsets/ps4-1-sample.txt

Clustering result of sample dataset using complete-linkage Euclidean distance:
http://www.courses.fas.harvard.edu/~bphys101/problemsets/ps4-1-result.txt

Normalized training dataset:
http://www.courses.fas.harvard.edu/~bphys101/problemsets/ps4-1-train.txt

Clustering result of normalized training dataset using complete/single-linkage Euclidean distance:
http://www.courses.fas.harvard.edu/~bphys101/problemsets/ps4-1-result2.txt

V) V) Decide the number of clusters you want to use in your program, and explain how you decided that number according to the original paper. (1 pt)

2 clusters for AML and ALL (1 pt)

VI) VI) Run your program on the normalized training dataset using the **complete-linkage** (farthest neighbor) Euclidean distance metric. Provide your clustering result (without distance matrix) and compare it with the original paper (table_ALL_AML_predic.txt). If your program does not work, use the provided clustering result above. (4 pt)

```
[Group 1: 10 sample(s)]
28
29
30
31
32
33
34
36
37
38


[Group 2: 28 sample(s)]
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
35
```

AML sample 35 is clustered with ALL samples.

VII) VII) Run your program on the normalized training dataset using the **single-linkage** (nearest neighbor) Euclidean distance metric. Provide your clustering result (without distance matrix) and compare it with the original paper (table_ALL_AML_predic.txt). If your program does not work, use the provided clustering result above. (4 pt)

```
[Group 1: 1 sample(s)]
35

[Group 2: 37 sample(s)]
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
```

34
36
37
38

Cluster sizes: 37, 1
Most of the AML/ALL samples were clustered in one group except for sample 35.

VIII) VIII) Describe and explain any differences or similarities between your results from VI and VII. If your program does not work, use the provided clustering result above. (3 pt)

Single-linkage method was inappropriate for dividing distinct samples into two groups. (1 pt) In single-linkage method the similarity (distance) of the closest pair between two groups (clusters) is used as the similarity measure between these two groups. Since only one small distance can merge very different groups, the resulting clusters tend to be a long chain rather than a several distinct clusters that are merged very late. (2 pts)

Problem 2: Motif searching and functional enrichment (34 pts total)

You will need to read the following paper by Tavazoie et al. to answer the next part: *Nature Genetics **22:**281-5*

If two genes change expression level in the same way in response to a change in conditions, they are often assumed to be related (e.g. co-regulated, or play common roles in cellular processes).

I) I) With reference to table 1 and the "Determination of statistical significance for functional category enrichment" in "Methods" section in Tavazoie *et al.* paper, answer the following questions. (Total 9 points)

a. a. With reference to table 1 and the methods section in Tavazoie *et al.* paper, explain how the statistical significance for functional category enrichment is determined. (3pts)

The basic idea is to make sure that the likelihood of functional category enrichment occurs just by chance is very low. The probability of observing (at least) the number of genes from a particular functional category within each cluster is represented by P value and is calculated based on hypergeometric distribution. If a P value were beyond certain threshold, it would mean that the particular functional category enrichment within the cluster could be by chance, and hence insignificant. If a P value were below the threshold, the probability for the functional enrichment to occur by chance is so low that the enrichment is statistically significant.

b. b. Examine table 1 and figure 1. What are the clusters that show cell cycle periodicity and at which stage of cell cycle genes in each of these clusters may be required? How would you use such information to learn more about a gene with previously unknown function in one of these clusters? (3 pts)

Cluster 2, 7 and 14 show cell cycle periodicity. Cluster 2 expressed higher at G1-S transition; cluster 7 expressed higher in M phase; cluster 14 expressed higher at S-G2 transition. A functionally unknown gene in any of these clusters may be involved in cell division or cell cycle regulation. Depending on where the gene is clustered in, it could be involved in different stage of cell division or cell cycle regulation.

c. c. The total number of genes within a genome is not available for many species, for example, the human genome. Looking at the calculation of the hypergeometric distribution in the methods section, make an argument in favor of using a different number for the variable represented as *g* than the total number of genes within the genome. What number would you use and why? (Hint: think about the clustering step). (3 pts)

Typically only the most variable genes were used in clustering and get counted for variable k. Thus the number of genes used in clustering can be used for the variable represented as g.

You might find Hughes et al., J. Mol. Biol. ***296****: 1205-1214, helpful when answering the following questions.*

II) II) AlignACE uses a Gibbs sampling algorithm to identify over-represented motifs in a set of DNA sequences. The program can be accessed at the following site: http://atlas.med.harvard.edu/cgi-bin/fullanalysis.pl. Here you will use it to analyze the upstream regions of genes present in the Tavazoie *et al.*'s cluster #30 http://arep.med.harvard.edu/network_discovery/clusters_members_distances_annotations.txt (total 12 pts)

Here is the list of gene names in cluster #30 (all you need to do is copy/paste into the "Enter a list of genes below, one gene name per line (Y names only):" filed:

YAL053W
YAL067C
YAR015W
YAR052c
YBL015w
YBR085w
YBR112c
YBR155w
YBR156c
YBR213w
YBR289w
YDL059C
YDL071c
YDR213W
YDR227w
YDR252w
YDR253C
YEL007w
YEL043w
YER042w
YER132c
YFR030W
YGL013C
YGL184C
YGR058W
YGR138C
YGR239C
YHR210C
YIR017C
YJL106W
YJR010W
YJR047C
YJR127C
YJR137C
YKL001C
YKR069W
YLL061w
YLL062c
YLR048w
YLR092w
YLR228C
YLR327C
YLR364W
YMR006C
YMR190C
YMR306C-A
YNL033W
YNL191W
YNL241C
YNL277W
YOL163W
YOR267C
YOR368W
YPL002C
YPL054W
YPL116W

YPL140C
YPL188W
YPR046W
YPR104C

a. a. Report the five best motifs obtained by AlignACE in terms of MAP score (note: AlignACE lists its outcome in the order of decreasing MAP score). (6 pts)

```
Motif 1
AAAAAAAGAAACA 0 606 0
AAAAATAAAAAAA 53 333 1
AAAAAAAGAAAGA 6 610 1
AAAAACATAAATA 43 121 1
AAAAAAAGAAAAA 43 356 1
AAAAAAAAAAAAA 44 194 1
AAAAAGAAAAAAA 3 471 0
AAAAAAAGAAAAA 51 350 1
AAAAAGAAAAAAA 13 190 0
AAAAATAAAAACA 51 530 0
AAAAATACAAATA 15 167 0
AAAAAGACAAACA 15 278 0
AAAAAGAAAAAAA 18 191 0
AAAAAAAAAAAAA 18 375 0
AAAAAAAAAAAAA 23 299 1
AAAAAAAAAAAAA 4 503 0
AAAAAAAAAAATA 26 197 1
AAAAAAAAAAAAA 29 321 1
AAAAAAATAAAAA 29 393 1
AAAAAGACAAAGA 5 244 0
AAAAAAAAAAAAA 29 406 1
AAAAACAGAAATA 5 394 0
GAAAAAAAAAAGA 51 364 1
AAAAAGGGAAATA 35 386 0
AAAAAGGTAAAAA 40 212 0
GAAAAAAAAAATA 42 364 0
AAAAAAGAAAACA 7 389 1
AAAAACGTAAACA 51 652 0
AAAAAAGAAAAGA 12 26 0
GAAAACAAAAATA 47 438 1
AAAAAAGAAAAGA 0 565 0
GAAAAGAAAAATA 5 178 0
AGAAAGATAAAAA 36 292 0
AGAAACAAAAAAA 39 367 1
AGAAAGAAAAAAA 45 472 1
AGAAAGAGAAAAA 33 406 0
GGAAACAAAAATA 4 89 0
AGAAAAGTAAAAA 0 246 0
AGAAACGGAAAGA 24 278 0
AGAAATGTAAACA 51 550 1
AGAAATGTAAATA 32 117 1
GGAAAAAGAAAAA 35 156 0
```

```
AAAAAGAAAAAGG 35 328 0
AAAAAAACAAAAG 0 509 0
AAAAAAAAAAAGG 12 361 1
AAAAAAAAAAAAG 16 168 1
AAAAAAAAAAATG 7 175 0
AAAAAAAAAAAGG 29 369 1
AAAAAAATGAAGA 3 97 1
AAAAAGAAGAACA 29 93 1
AAAAAAATGAAAA 29 442 1
GGAAAGGAAAAAA 29 248 1
GGAAAAGTAAAAA 34 176 0
GGAAAGGAAAAAA 41 163 0
GAAAACAGAAATG 15 203 0
AAAAAAGGAAATG 4 434 1
AAAAAGGTAAAAG 25 341 1
GAAAATAAAAATG 27 190 0
AGAAAGAAAAATG 36 309 0
AGAAAAACAAAAG 31 479 1
GAAAAGAAGAAGA 40 171 1
GAAAAGAAGAAAA 29 845 0
AAAAAAGTGAAAA 47 182 0
AAAAACGGGAAAA 47 206 0
AAAAAAGCGAAGA 12 394 1
AAAAAGGAGAAAA 55 389 0
AAAAAAGAGAAGA 15 348 0
AAAAAGGTGAAAA 26 166 1
AAAAATGTGAAAA 8 284 1
AGAAATAAGAACA 10 532 0
AGAAATATGAATA 28 312 1
AGAAAGAAGAATA 46 243 1
GAAAAGGGAAAGG 27 274 0
GAAAAGGGAAAAG 50 591 0
GGAAACAAAAAAG 39 725 0
GGAAATAAAAAAG 35 183 1
AAAAATACGAAAG 1 281 1
GGAAAAGAAAATG 27 144 0
GGAAACATGAAAG 46 293 0
***** * *** *
MAP Score: 103.131
Specificity score: 2.5e-01
No matching motifs.

Motif 2
GAAATATGCTTAGAATAG 0 105 1
AAAATGAAGGGAGAACTA 0 140 1
AAAAAAAAAAAGAAACAG 0 605 0
AAAAAAAATTACAAAGAG 1 456 1
AAAAAAAATGAAGAACCA 3 96 1
AAAAGAAAAAAAGAAATG 3 465 0
AAAAAAAAAAAAAAAGGA 4 500 0
GAAAAACGCCAAAAAAAG 5 122 0
AAAAGAAAAATAAAATAA 5 172 0
AAAAGACAAAGAAAAAAA 5 238 0
```

```
AAAAGAAAAACAGAAATA 5 394 0
AAAACACGACTGGAAAAA 5 468 1
AAAAAGGTGGAAAAAAAA 6 599 1
GAAAAGTGCCGGAAATTA 6 638 0
AAAACAAACGAAAAAAAA 7 180 0
AAAAAAGAAAACAAACAA 7 389 1
AAAAAAATGTGAAAATCG 8 282 1
AAAATAGCTTGGAAAAAA 11 117 0
AAAAAAGCGAAGAAAACA 12 394 1
AAAAAACAAAACAAAGCA 13 178 0
AAAATAAGCAGGAAATAA 14 237 0
AAAAAAAACACCAAAGAA 15 73 0
GAAACGATCGGGAAACGA 15 180 0
AAAAAGACAAACAAACAA 15 273 0
AAAAAAGAGAAGAAAATA 15 343 0
AAAAAAATAATAAAAAAA 16 157 1
AAAAAAGCAGAAAAAACA 18 167 0
GAAAAAGAAAAAAAAGCA 18 187 0
GAAAAGCAGTTGAAAAAA 19 293 0
AAAAAAGAGTCAGAATGA 21 115 0
GAAATGTGGCCAGAAGCA 21 292 1
AAAATATCAAGCAAACAA 21 461 1
AAAACACGCTTGAAAAAG 22 133 0
AAAAGACTTTTGAAAGTA 22 247 0
AAAATAATACACAAAACA 23 134 0
GAAAAAAAAAAAAAATTA 23 298 1
AAAAAGAAACGGAAAGAA 24 277 0
GAAACAGGTGAAAAATTA 24 461 1
AAAAAGGTGAAAAAAGCG 26 166 1
AAAATAAAAATGAAATGA 27 184 0
AAAACAGGCAACGAACGA 29 289 1
GAAAAACGAAAAAAAAAA 29 313 1
AAAAAAACACAGAAAAGA 29 331 1
AAAAAAAATAAAAAAAAA 29 392 1
AAAAAGTGAAAAAAAATG 29 433 1
GAAAGATCGGTGAAAACA 29 585 0
AAAAAATCTTGGAAAACA 30 117 1
AAAATAACAAGCAAAGGA 31 390 0
AAAAAGGTCGAGAAACCA 33 353 0
GAAAAACGTATAAAATGA 33 394 0
GAAAAAGTAATCAAAAAA 34 3 0
AAAAAGTTAATGAAAAAA 34 210 0
GAAATAAAAAAGAAATTG 35 184 1
GAAAAAGGGTGCAAATAA 35 318 0
AAAAGATCGCAGAAACTA 36 107 1
AAAAAAATTTGAGAAAAG 39 373 1
GAAAGGAGTTGGAAAAAG 39 529 1
AAAAAAGGTAAAAAAAGG 40 208 0
AAAATAATTCAGAAACGA 42 352 0
AAAAAAGAAAAAGAACAA 43 357 1
AAAAAATGACTAAAAAAA 44 201 1
AAAAGAAAGAAAAAAACG 45 469 1
AAAAAAACGGGAAAAGGA 47 203 0
```

```
AAAACAATTCCAGAAAGA 49 213 0
AAAATAAGAGAAAAAAAA 50 90 1
AAAAGATGATAGGAAGAA 50 254 1
GAAAAAAAGAAAAAAGAA 51 349 1
AAAACAAAAATAAAAACA 51 530 0
AAAATATAACGAAAAAAA 51 819 1
GAAAAAAAATCAAAAAGA 52 208 1
GAAAAAAAAATAAAAATA 53 322 1
AAAAAAAGGAGAAAATAA 55 386 0
**** *     ****  *
MAP Score: 70.5503
Specificity score: 1.7e-02
No matching motifs.
```

```
Motif 3
TGCATATATAAATA 1 480 1
TGTATATATATATA 8 360 0
TGGATATATATATA 28 325 1
TGCATACATATATA 14 100 1
TGTATATATATATA 2 359 0
TGAATATATATATA 45 280 0
TGAATAAATAAATA 51 130 1
TGAATATATAAAAA 7 238 1
TGTATATATATATA 21 385 0
TATATATATATATA 28 339 1
TAAATAAATAAAAA 1 270 1
TAAATAGATAAATA 33 233 0
TATATACATACAGA 29 822 0
TACATATATATATA 28 355 0
TACATACATATACA 16 34 0
TAAATAAATAAATA 21 349 1
TAAATAGATAAAAA 51 116 1
TAAATATATATATA 21 365 1
TAGATAGATAAAAA 8 160 0
TGTATATATAAACG 52 25 1
TGTATATATAGACG 3 344 1
GGCATATATATATA 0 725 1
GGTATATATAGATA 14 140 0
GGCATATATAAACA 17 135 1
TAAATAAATACAGG 27 380 0
TATATATATATATG 36 224 0
TGGATAGATATGTA 14 117 1
GATATATATATATA 33 572 1
GATATAGATACACA 37 43 0
GAAATACATAGAAA 49 179 0
TAAATATATACGTA 19 6 1
GGTATACATACATG 22 570 0
GGCATATATATATG 29 772 1
TGAATAAATACGCG 42 123 0
TGTATAGATATGAG 55 472 0
TGAATATATATGTG 52 341 0
GAAATATATATAAG 38 317 1
GGTATATATAAGCA 19 206 1
```

```
            CAAATATATATACA 43 389 1
            ** *** *** * *
            MAP Score: 36.202
            Specificity score: 1.8e-05
            Matching motifs: AT_repeat(0.906376)

            Motif 4
            AAACTGTGGC 10 147 1
            TAAATGTGGC 11 95 1
            AAACTGTGGC 14 317 1
            CAACTGTGGT 17 84 0
            ACAGTGTGGC 21 64 0
            AAAGTGTGGC 21 135 1
            TCACTGTGGC 22 307 0
            CAACTATGGC 26 116 1
            AAACTGTGGT 26 144 1
            CAACTGTGGC 26 275 1
            AAACTGTGGC 28 246 0
            CAACTATGGC 30 212 0
            TAATTGTGGC 31 225 0
            AAAATGTGGC 32 304 1
            AAATTGTGGC 34 145 0
            AAACTGTGGC 34 169 0
            CCACTGTGGC 36 37 1
            CAACTGTGGC 36 77 0
            TAAATGTGGC 38 46 1
            AAACTGTGGC 39 776 0
            AAACTGTGGC 43 326 1
            AAACTGTGGG 44 333 1
            ATACTGTGGC 50 232 0
            **********
            MAP Score: 26.6977
            Specificity score: 2.3e-17
            Matching motifs: MET31_32(0.981925)

Motif 5
ATGTTCACGTG 0 379 0
AATGGCACGTG 0 457 0
AATTTCACGTG 8 249 1
CACGTCACGTG 10 192 0
AATGTCACGTG 14 367 1
CTGGTCACGTG 16 94 1
AAGCTCACGTG 19 89 0
TAAGTCACGTG 26 208 1
AAGGTCACGTG 28 118 1
AAAGTCACGTG 28 135 1
AATATCACGTG 31 167 0
AATTTCACGTG 32 267 1
AATGTCACGTG 33 434 0
AATGTCACGTG 36 26 1
CAGGTCACGTG 36 52 1
ATTTTCACGTG 45 142 1
** ********
```

MAP Score: 20.2411
Specificity score: 7.7e-11
Matching motifs: CBF1(1) PHO4(0.685328)
Due to the randomness of seeding in Gibbs Sampling algorithm, the highest MAP score samples may vary between runs. But the MET31_32 and CBF1(1) PHO4 should be within the first few motifs listed according to MAP scores.

b. b. AlignACE lists its outcome in the order of decreasing MAP score. From the result you obtained by running AlignACE, does the highest MAP score always correlate to most meaningful functional motif? Why or why not? What information do group specificity scores add when trying to infer the “real” cis-regulatory elements? (3pts)

The highest MAP score does not always correlate to most meaningful functional motif. Group specificity should also be significant. The group specificity score gauges how well a given motif targets the upstream regions of the genes used to find it relative to the upstream regions of all genes in the genome; while MAP score only assess how well the motif is over-represented.

c. c. The article by Hughes *et al.* (see above) suggests significance thresholds for the MAP and group specificity scores: >=10 and <=10^{-10}, respectively. Based on these thresholds and the statistics assigned to your motifs (part c), what can you infer about the regulation of genes in the cluster? (3 pts)

Both motif 4 (MAP Score: 26.6977; Specificity score: 2.3e-17) and motif 5 (MAP Score: 20.2411; Specificity score: 7.7e-11) meet the thresholds for MAP score and group specificity scores, although motif 4 scores a little better. Therefore both MET31_32 and CBF1 motifs are candidates that are responsible for co-regulation of genes in this cluster.
Again, due to the randomness of seeding in Gibbs Sampling algorithm, the highest MAP score samples may vary between runs. But the MET31_32 and CBF1(1) PHO4 should be within the first few motifs listed according to MAP scores. CBF1(1) PHO4 may falls after 5th motif, so some students may have only MET31_32 as the meaningful functional motif within the first 5 motifs with the highest MAP score.

III) III) The relevance of motif results (8 pts total)

a. a. Why are motif analyses performed using clusters from gene expression data (such as microarray) instead of whole genome? (2pts)

Because the goal is to find regulatory sequences that control the gene expression, we need to find over-represented motifs in a set of genes that are co-regulated. Co-regulated genes are likely to be co-expressed, and they will be in the same cluster from expression data.

In addition, these regulatory motifs are usually very short, compared to the size of a genome. The probability for such stretches of short sequences occurs simply by chance increases. Motif analyses will have to deal with a much higher background noise if performed on whole genome..

b. b. Do all the genes in same cluster share same motif? Why or why not? (3pts)

Not always. Motifs are supposed to control regulation. Therefore genes co-regulated should share same motif. However not all genes in a cluster are co-regulated. Some may be expressed in similar pattern coincidentally, hence clustered together, but regulated differently and under control of different motifs.

c. c. Do all the genes sharing same motif clustered together? Why or why not? (3pts)

Not always. The expression of a gene could be controlled by multiple motifs. Thus sharing one motif does not guarantee co-regulation. Therefore genes sharing same motif do not all clustered together.

IV) IV) Sequence Logos as visual representations of motifs (5 pts total)
You might find the following URL helpful in answering these questions:
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

a. a. Motifs can be represented visually as sequence logos with nucleotide letters of various sizes at each site. How is each letter's size calculated? What can you conclude if there is only a tiny "A" at a given position in a motif? (3 pts)

(2pts) The height of each letter reflects the sequence conservation at that position. The height is called information content and is measured in bits. The calculation follows:

First uncertainty is computed:

$H(L) = \square_{ACGT} [\, f(b,l)\ log2\ f(b.l) + e(n(l))\,]$
Where
$e(n(l))$ *= a correction for the small sample size n at position l, and*
$f(b,l)$ *= the frequency of base b at position l.*

Next information content (or sequence conservation), which ranges from 0-2, is computed:

$R_{sequence}(L) = 2 - H(L)$

(1pt) If there is only a tiny "A" at a given position in a motif, we know that there is a small probability that "A" is conserved in the given position in the motif. We don't have information about conservation of the other nucleotides.

b. b. Can you reconstruct the actual regulatory sequences that went into building the motif? What information is lost in representing motifs by sequence logos? (1 pts)

(1pt) From a sequence logo alone, you cannot regenerate each binding site that contributed to the logo. Information about each individual sequence is lost.

c. c. How are sequence logos an improvement over consensus sequences? (1 pt)

(1pt) A consensus sequence does not allow variability at each position. Each position is assigned one and only one nucleotide or amino acid, and deviation is not tolerated. A sequence logo, however, reveals strongly conserved positions as well as positions more tolerant of variability. A sequence logo better represents the biological significance of a site.

Problem 3: Markov Chains and Hidden Markov Models (33 pts total)

Mount pages 185-191 and Durbin chapters 3-6 will be helpful.

I) I) Describe an example of a simple Markov chain. It does not have to be a biological example. (2 pts)

Various examples shall be accepted. Give credits as long as the student understands the main point that: it *models a series of observable states, and each state probabilistically depends only on the previous state.*

II) II) What is a Hidden Markov Model? (2 pts)

A hidden Markov model is an extension of simple Markov chain. Here we have a series of observable states that may arise from a set of underlying states. We know the observed states but not the hidden states that produced them. We can, however, determine the probability that an observed state came from a hidden state, based on transition probability from the previous state and the emission probability from the hidden state.

III) III) Suppose you were generating a HMM to predict whether a given sequence most likely came from a CpG island, a non-CpG island, or partially from both.(12 pts total)

What two states will your model consider? (2 pts)What probabilities should your model include? Use variables to represent the probabilities, not actual numbers (i.e. P(A+ ->C-) would represent the probability of generating a C in a non-island following an A in a CpG island). (10 pts)

<u>(2pts) States:</u>
1) in a CpG island
2) not in a CpG island

<u>Probabilities:</u>

(2pt) For first nucleotide:
Pb(A+), Pb (C+), Pb (G+), Pb (T+), Pb (A-), Pb (C-), Pb (G-), Pb (T-)

(2pts) From one nucleotide to another within a CpG island:
P(A+ ->A+), P(A+ ->C+), P(A+ ->G+), P(A+ ->T+),
P(C+ ->A+), P(C+ ->C+), P(C+ ->G+), P(C+ ->T+),
P(G+ ->A+), P(G+ ->C+), P(G+ ->G+), P(G+ ->T+),
P(T+ ->A+), P(T+ ->C+), P(T+ ->G+), P(T+ ->T+)

(2pts) From one nucleotide to another, outside of a CpG island
P(A- ->A-), P(A- ->C-), P(A- ->G-), P(A- ->T-),
P(C- ->A-), P(C- ->C-), P(C- ->G-), P(C- ->T-),
P(G- ->A-), P(G- ->C-), P(G- ->G-), P(G- ->T-),
P(T- ->A-), P(T- ->C-), P(T- ->G-), P(T- ->T-)

(2pts) From a nucleotide with in an island to another in a non-island
P(A+ ->A-), P(A+ ->C-), P(A+ ->G-), P(A+ ->T-),
P(C+ ->A-), P(C+ ->C-), P(C+ ->G-), P(C+ ->T-),
P(G+ ->A-), P(G+ ->C-), P(G+ ->G-), P(G+ ->T-),
P(T+ ->A-), P(T+ ->C-), P(T+ ->G-), P(T+ ->T-)

(2pts) From a non-island nucleotide to one within an island
P(A- ->A+), P(A- ->C+), P(A- ->G-+, P(A- ->T+),
P(C- ->A+), P(C- ->C+), P(C- ->G+), P(C- ->T+),
P(G- ->A+), P(G- ->C+), P(G- ->G+), P(G- ->T+),
P(T- ->A+), P(T- ->C+), P(T- ->G+), P(T- ->T+)

IV) IV) Consider the sequence, "CCGTGC." Based on your HMM described above, how would you compute the probability that the first 3 nucleotides of this sequence came from a CpG island and the remaining 3 nucleotides came from a non-island? Since we have not assigned numbers to the probabilities above, write your answer using variables (i.e. probability = P(A) x P(C) x P(G)). (3 pts)

Probability = P(C+) x P(C+ ->C+) x P(C+ ->G+) x P(G+ ->T-) x P(T- ->C-) x P(C- ->G-)

Besides Durbin chapters 5-6, PFAM related web sites such as http://pfam.wustl.edu/index.html will also help for the next few questions.

One of the most common uses of hidden Markov models for molecular biology is in protein family classification. Suppose we want to find out what the function of a protein might be. Before heading towards the bench, we would like to get as much information as possible from existing information about known proteins. We could do a BLAST search against a protein database, which will give us pairwise alignments of our unknown sequences with every similar protein in the database. However this is not always satisfactory because sometimes our unknown protein is not very similar to any individual protein in the database. An alternative approach would be to gather information from all (or most) sequences in a protein family and compare our unknown protein with such information to examine the likelihood of our unknown sequence being related to that protein family.

V) V) Without getting into details of the probabilistic model, briefly explain how hidden Markov models can be used to classify a new protein into a known family and/or to search a database for new proteins that may belong to a known family. (5pts)

- Build a high quality multiple alignment (may involve expert hand curation) from some relatively well-known sequences for each protein family (such alignment is called a seed alignment).
- Build a profile hiddin Markov model (profile HMM) from each seed alignment.
  (Note: the utility in HMMer package to do this is hmmbuild)
- Use the profile HMM to find all sequences belong to the family in whole protein database and align them (called full alignment). If the

result is unsatisfactory, rebuild seed alignment and/or profile HMM. This process is iterated until satisfactory profile HMM is built.

- The profile HMM can be used to search for the whole protein family to find new or previously unknown members.
  (Note: the utility in HMMer package to do this is hmmsearch)
- One can also search the profile HMM database (aka. pfam) with an unknown sequence to find its matching family (or the lack of).
  (Note: the utility in HMMer package to do this is hmmpfam)

VI) VI) Now let's turn to a practical example. Go to http://www.ncbi.nlm.nih.gov/ (or any of your favorite protein sequence database web sites) and retrieve the protein sequence with accession BAA76778. (2 pts total)

a. a. What functional information (if any) or definition do you get from the annotation in the database entry? (1 pt)

```
KIAA0934 protein [Homo sapiens].
```

b. b. Now do a BLAST search against the non-redundant protein database. What would you conclude from the BLAST search results? You do not need to show the blast output. (1pt)

Not much. Almost all hits with over 25% identities are annotated as "hypothetical" or "unknown" protein.

VII) VII) The profile hidden Markov models (aka Pfam) for most (if not all) protein families are readily available and can be searched with protein sequence queries. You can conveniently perform such search on web sites such as http://pfam.wustl.edu/index.html.

a. a. Do a protein search using the sequence you retrieved in problem VI (BAA76778) as the query. You will need to use the fasta format sequence for the search. Do you believe the Pfam HMM search results? Why? (Hint: Examine the scores and E-values.) (2 pts)

Yes, the result is believable. The scores are all above the Pfam gathering cutoffs (GA). In fact in this case they are all above the Pfam Trusted cutoffs. E-value of e–41 is very significant; E-value of 0.00017 is not that great, but still greater than 0.5. Therefore the pfam search result is significant.

b. b. What would you conclude from the Pfam HMM search results? (Hint: follow the link to the Pfam entry.) (3 pts)

This protein contains one complete and two partial domains of AMP-binding enzyme family. This is a rather diverse family. The query sequence is related to one of the followings: long chain fatty acid Co-A ligase, acetyl-CoA synthetase or various other closely related synthetases.

Note: no point will be deducted if student does not know that luciferase is not a human enzyme.

VIII) VIII) Can you think of an advantage and a disadvantage of protein classification using Pfam compared to BLAST search? (2 pts)

Advantage: The major advantage of Pfam search is better in detecting distant similarities to protein families.

Disadvantage: HMM method is generally much more computationally intensive than BLAST. This is especially a drawback when using profile HMM to search against protein or DNA sequence databases.