Your Name:

Notes

1. Be sure to write your name.
2. If you write nothing, you get no credit, so do write your thoughts. Do not write gibberish.
3. Show your computation. A correct idea is worth more than a correct answer.

Grade:

<table>
<thead>
<tr>
<th>Problem</th>
<th>Total</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Please do not fill in
Questions

1. You have just taken a job as a microarray designer for Affymetrix, the world’s foremost manufacturer of oligonucleotide arrays. As discussed in class, Affymetrix uses photolithography to print oligos. Your boss Steven Fodor has just asked you to design the next generation oligonucleotide array for analysis of the expression level of every human gene.

   (a) Stipulate the number of spots on your array and the length of the oligos. Given the known size of the human genome ($3 \cdot 10^9$bp), justify your choice for oligo length based on theoretical calculations of the shortest oligo that would be likely to hybridize to a single RNA species (show your work).

   (b) Limited only by the wavelength of light (recall what type of light is used), what is the theoretical maximum number of oligonucleotide features that can be printed on an area of 1 square cm?
(c) How low would the error rate of oligo synthesis have to be (expected number of mistakes per base pair per site) in order to synthesize a 20mer so that 99% of the oligos are correct? What about for a 40mer? 100mer? An entire synthetic human genome?

A "correct" oligo means that all bps of the oligo are exactly as designed, with no mistakes during synthesis. A mistake is defined as either a substitution or an indel.
2. Define what a type II* restriction enzyme is, and draw a diagram of why it is essential to the SAGE technique.
In this problem you will develop and implement a bi-clustering algorithm. Define a bi-cluster \((I, J)\) as a subset \(J\) of experiments and a subset \(I\) of genes. Define bi-clustering as a partition of the data set into bi-clusters. In other words, each gene or experiment is assigned to exactly one bi-cluster. Let \(H(I, J)\) denote the cost of the b-cluster \((I, J)\). We seek to find large bi-clusters, while minimizing cost, as follows:

\[
H(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{IJ} + a_{Ij})^2
\]

where:

- \(a_{ij}\) is the log expression of gene \(i\) at experiment \(j\).
- \(a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}\) is the average log-expression value of gene \(i\) over all experiments in \(J\).
- \(a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij}\) is the average log-expression of all genes in \(I\), for the \(j\)-th experiment.
- \(a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij}\) is the average log-expression value of a cell, computed over all cells in the bi-cluster \((I, J)\).

Our goal will be to find Bi-clusters with the maximal number of genes \(|I|\) and experiments \(|J|\) that achieve \(H(I, J) < \delta\) where \(\delta\) is a user defined threshold.

(a) Motivate the selection of \(H(I, J)\) as the value to minimize. In particular, what does a value of 0 mean? Are there always Bi-clusters with a value of 0?

(b) Propose a heuristic method to find the largest bi-cluster with \(H(I, J) < \delta\), and use it to identify the best \(K\) clusters. Recall that genes and experiments are not shared by bi-clusters.

(c) Describe the complexity of your algorithm in terms of the number of genes \(n\), and the number of experiments, \(m\).
4. Consider a set of points from two classes, (+) and (−). The classes are considered to be ‘well-separated’ by the hyperplane $\beta^T x = \beta_0$ if the following holds.

(a) For all + points $x$, $\beta^T x \geq \beta_0 + 1$.
(b) For all − points $x$, $\beta^T x \leq \beta_0 - 1$.

Assume the points in the two classes are on a 4-dim. space, and are well-separated by two planes.

**Hyperplane 1:** $3x_1 + 2x_2 + x_3 - 2x_4 = 1$

**Hyperplane 2:** $5x_1 + x_2 + 3x_3 - x_4 = 2$

Which hyper-plane will you choose, and why?
5. Consider a peptide $T$ with $n$ residues. $T$ generates a tandem mass spectrum $S$ with $m$ peaks. Each peak is either a $b$-ion, or a noise peak. If $T$ can have at most 2 modifications, describe an algorithm to identify the sites on $T$ that are modified. If you like, you can use the following notation:

(a) $m(A)$ is the residue mass of amino-acid $A$.

(b) The mass of a $b$-ion peak $p$ corresponding to the first $k$ residues is given by $m(p) = 1 + \sum_{i=1}^{k} m(T_i)$.

(c) The parent mass of the spectrum $S$ is given by $M = (\sum_{i=1}^{n} T(i)) + 1 + \Delta$, where $\Delta$ is the sum of all modification mass values on $T$.

You are also allowed to use your own notation, but please define it clearly.
6. Consider a BW transform with the following data structures

- array $\text{Pos}\left[\sigma\right]$: The index at which $\sigma$ appears as the first symbol in the lexicographically sorted suffixes.
- $B[i]$: the BW transform, which is the ‘previous’ symbol of the i-th lexicographically sorted suffix.
- $\text{Occ}\left[\sigma, i\right]$: number of occurrences of $\sigma$ in $B[0 \ldots i]$. 

The function $\text{FindPrev}(B,i)$ returns the symbol that occurs before $B[i]$ in the original string. Write an expression for $\text{FindPrev}(B,i)$ using only the provided data structures.