

02-510/710: Computational Genomics, Spring 2025

HW4: Dimension reduction and motif-finding

Version: 1

Due: 23:59 EST, Apr 7, 2025 on Gradescope

Topics in this assignment:

1. Dimension reduction
2. Motif-Finding

What to hand in.

- One write-up (in pdf format) addressing each of following questions.
- All source code. If the skeleton is provided, you just need to complete the script and send it back. Your code is tested by autograder, please be careful with your main script name and output format.

Submit the following file which contain the completed code and the pdf file to gradescope separately.

./S2025HW4.pdf

Please note that all the solutions must be your own. We will check for plagiarism after the final submission.

1. [40 pts] dimensionality reduction

In this problem, we will explore linear and nonlinear dimension reduction methods. Assume you have a high-dimensional dataset from gene expression, with each sample containing the expression levels of 1000 genes. The dataset is known to be divided into two main classes (e.g., tumor and normal cells), but their specific characteristics are not yet clear. Data can be found from `provided_data/gene_expression.csv`. Please answer the following questions.

- (a) Data Preparation and PCA. Standardize the gene expression data for each sample. After applying PCA, calculate the explained variance ratio of the first two principal components.

Solution

PC1: 10.98% of variance PC2: 1.52% of variance

- (b) There are 2 key hyperparameters for t-SNE: perplexity and learning rate. Please briefly describe their potential impact on the dimensionality reduction results.

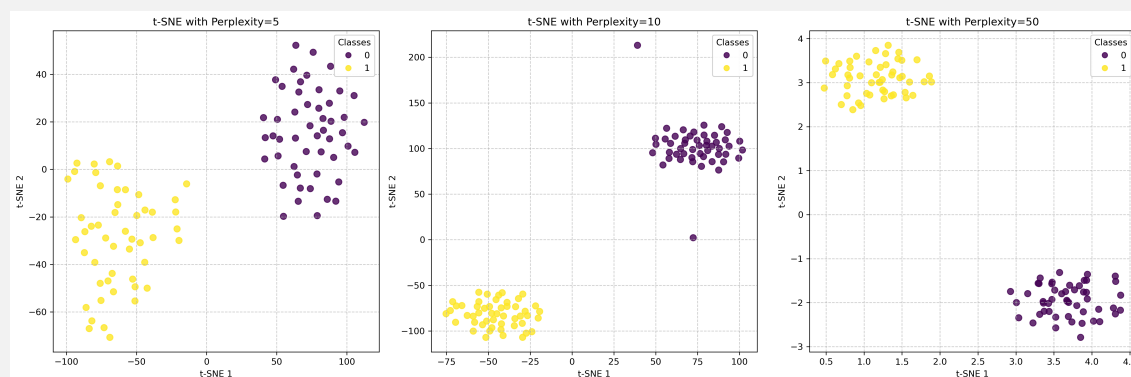
Solution

Perplexity: affects how well you can distinguish between tumor and normal cell clusters. If it is too low, there could be fake sub-clusters within each cell type; if it is too high, the boundary between tumor and normal might become blurred.

Learning rate: sets the size of the steps the algorithm takes when arranging the points. If the learning rate is too low, the process can be very slow or get stuck. If it's too high, the algorithm might jump around too much and miss a good layout.

- (c) t-SNE Implementation. Implement t-SNE on the standardized gene expression data. Experiment with 3 different perplexity values (e.g., 5, 10, 50) and a fixed learning rate (200). Please visualize the results using scatter plots. Additionally, use a quantitative metric such as the silhouette score to evaluate which perplexity setting provides the best separation between “normal” and “tumor” samples.

Solution



The included figure shows the tSNE results with perplexity values 5 (left), 10 (middle), and 50 (right). The highest silhouette score was 0.8835, for perplexity=50, thus providing the best score and best separation.

2. [60 points] Motif Finding

In lecture we were introduced to an algorithm used for finding motifs in DNA sequences based on Expectation Maximization (EM). This algorithm forms the basis of the MEME Suite (**M**ultiple **E**M for **M**otif **E**lucidation), one of the most widely used softwares in genomics. Several good papers are available for understanding the algorithm, including the ones [here](#) and [here](#).

Consider a biological motif of length W , $M = (M_1, \dots, M_W)$, where $M_i \in \{A, C, G, T\}$. Our model for biological motifs is that each M_i is a Multinoulli-distributed random with its own probability distribution over the nucleotides A, C, G , and T . We can equivalently represent this motif as a position weight matrix $(M)_{ij}$, for which

$$M_{ij} = \mathbb{P}(M_j = i),$$

where $j = 1, \dots, W$ and $i \in \{A, C, G, T\}$. Consider the PWM below for a motif of length $W = 6$:

$$M = \begin{matrix} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{bmatrix} 0.8 & 0.1 & 0 & 0.9 & 0 & 0.3 \\ 0 & 0.4 & 0.05 & 0.03 & 0.1 & 0.2 \\ 0.2 & 0 & 0.95 & 0.02 & 0.1 & 0.1 \\ 0 & 0.5 & 0 & 0.05 & 0.8 & 0.4 \end{bmatrix} \end{matrix} \quad (1)$$

One of the key assumptions we make when modeling motifs is that $M_i \perp M_j$ for $i \neq j$; that is, the distributions of the nucleotides in each position of the motif are independent of one another.

- (a) Find the probability $\mathbb{P}(M = ACCTTA)$.

Solution

0.000192

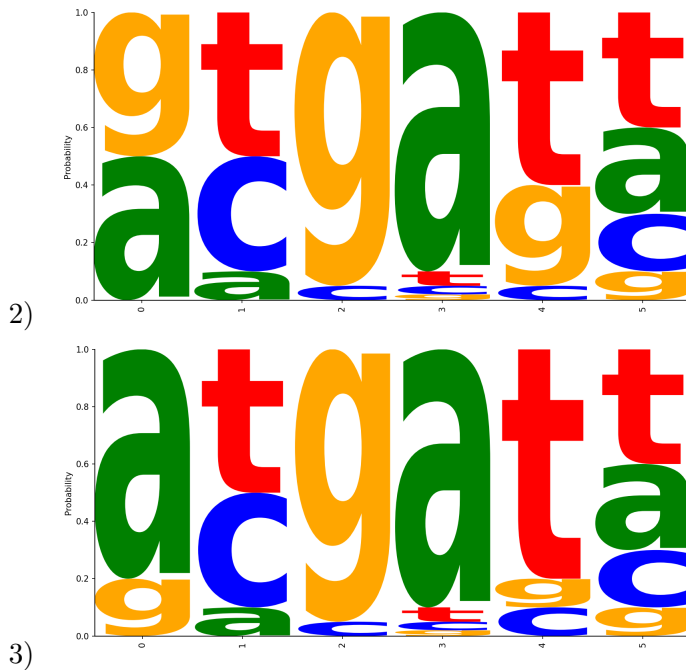
- (b) Provide the Consensus Sequence corresponding to the PWM above; i.e., the sequence $m = (m_1, \dots, m_5)$ such that $\mathbb{P}(M = m)$ is maximized.

Solution

Consensus sequence: ATGATT, $P = 0.1094$

- (c) Which of the following sequence logos accurately represents the PWM? Briefly describe why the other logos are incorrect.





Solution

3 is correct. 1 is incorrect because it shows G as the highest probability nucleotide in M_6 (should be T). 2 is incorrect because it shows G as the highest probability nucleotide in M_1 (should be A).

- (d) The height of the sequence logo is often scaled by the information content at a given position i . The information content is given by $R_i = \log_2(4) - (H_i + e_n)$, where H_i is the entropy of the position, and $H_i = -\sum_j [M_{i,j} * \log_2 M_{i,j}]$. Which position in the PWM has the greatest entropy? How is entropy related to conservation?

Solution

M_6 has the greatest entropy (1.8464 bits) as it has a more uniform distribution of nucleotides. Entropy is inversely related to conservation. Low entropy means that one nucleotide is strongly preferred, and thus will be more conserved. High entropy means that the distribution of nucleotides is uniform and there is low conservation since the position is flexible about which nucleotide appears.

- (e) One of the advantages of the MEME suite is that, for a given DNA sequence, it can detect the most likely positions to find the motif learned in the PWM. To do so, it converts the PWM into a matrix of log likelihood ratios using the formula $LLR(s) = \log_2[\frac{Pr(s|M)}{Pr(s|M_{bg})}]$. The log likelihood score that a given position i is a start position of the motif is equivalent to $S(X) = \sum_{i=x_i}^{i=x_i+w} \log_2[\frac{M_{ij}}{M_{ij}^{bg}}]$.

One method of determining if a position is included in the motif is to use a decision rule. For example, we could use : 'X is a true instance of M if $LLR(X) > 0$, or equivalently, if $S(X) > 0$ '. Is this a statistically sound approach? Why or why not?

Solution

No, $LLR > 0$ doesn't tell you if the result is statistically significant. Many sequences might have slightly positive LLR by random chance.

(f) An alternative approach to determine where or not a position i contains the motif is to use hypothesis testing to calculate the p-value for $S(p)$. We can define the hypotheses as:

- H_0 : X is drawn from the background distribution M^{bg}
- H_1 : X is drawn from the motif distribution M

Briefly describe a method for calculating the test statistic. Hint: one definition of p-value is the probability of receiving a value as or even more extreme than $S(X)$ under the null hypothesis.

Solution

We could generate random sequences from the background model, calculate their LLR scores, and determine what percentage of these random scores are as high or higher than our observed score. This percentage would be the p-value, the probability of seeing such a strong motif match by random chance. Unlike a fixed threshold approach, this method would provide a statistical measure of significance that accounts for the actual distribution of scores.