Identifying Change-Points in the Bacteriophage Lambda

Luis Garreta

April 1, 2017

Doctorado en Ingeniería
Pontificia Universidad Javeriana – Cali

Description

This homework looks at some statistics about the DNA content of the Lambda Phage and shows an example of segmentation of a sequence.

Background

Change-Points

Most genomes, including human genome, have a characteristic structure which takes the form of segmental patterns of variation in various properties. This structure may be caused by diverse factors as the division of genomes into regions of distinct function, the contingent evolutionary processes that gave rise to genomes, or by a combination of both. These segmental patterns are located into regions of approximately uniform statistical behavior and are called ***change-points,*** such as the GpC islands of DNA sequences with high G+C content. Determining the change-points between segments can help to identify important biological signals that can be used for discovering functional components of a genome, understanding the evolutionary processes involved, and fully describing genomic architecture.

One of the most frequently invoked approaches to study sequence characteristics to interpret a yet uncharacterized genome is to study a region by moving a window along the sequence and computing its local properties, as the G+C contents above. This kind of approaches are simply and ease to implement, but they are sensitive to window size with increasing stochastic variations for smaller windows and with diminishing resolution for larger ones. All of this results in imprecise detection of locations of transitions from one property to another. Moreover, probabilistic approaches as the hidden Markov models (HMMs), with a strong theoretical underpinning and a clear way for modeling problems, are often used to search for optimal partitioning of a sequence into classes with distinctive properties.

Bacteriophage lambda

Bacteriophage lambda infects the bacterium Escherichia coli as other phages do with other bacteria. It contains about 48502 bases and the DNA sequence can be obtained from the GenBank database with the accession number NC_001416. It has long stretches of either very GC-rich (mostly in the first half of the genome) or very AT-rich sequence (mostly in the second half of the genome).

Problems

Problem 1

Infer which state of the HMM is most likely to have generated each nucleotide position in the Bacteriophage lambda genome sequence. Use a HMM with two different states, “AT-rich” and “GC-rich”, based on the following values :

- For the AT-rich state, set pA= 0.27, pC= 0.2084, pG= 0.198, and pT= 0.3236.
- For the GC-rich state, set pA= 0.2462, pC= 0.2476, pG= 0.2985, and pT= 0.2077.

Set the probability of switching from the AT-rich state to the GC-rich state to be 0.0002, and the probability of switching from the GC-rich state to the AT-rich state to be 0.0002. What is the most probable state path?

Problem 2

Infer which state of the HMM is most likely to have generated each nucleotide position in the Bacteriophage lambda genome sequence. Use a HMM with four different states, "A-rich", "C-rich", "G-rich" and "T-rich", based on the following values:

- For the A-rich state, set pA= 0.3236, pC= 0.2084, pG= 0.198, and pT= 0.27.
- For the C-rich state, set pA= 0.2462, pC= 0.2985, pG= 0.2476, and pT= 0.2077.
- For the G-rich state, set pA= 0.2462, pC= 0.2476, pG= 0.2985, and pT= 0.2077.
- For the T-rich state, set pA= 0.27, pC= 0.2084, pG= 0.198, and pT= 0.3236.

Set the probability of switching from the A-rich state to any of the three other states to be 6.666667e-05. Likewise, set the probability of switching from the C-rich/G-rich/T-rich state to any of the three other states to be 6.666667e-05. What is the most probable state path? Do you find differences between these results and the results from simply using a two-state HMM (as in Problem 1)?

Plots

Create two plots showing the local fluctuations in the frequencies of nucleotides by using a sliding window of 2000bp. The first plot will show the frequencies for the A, C, G, and T. And the second will show the frequencies for the GC-rich and AT-rich contents. As an example, the following plots in R show the frequencies for the A, C, G, and T content, and the GC-rich-AT-rich contect of a large sequence (~50000 bp) with a window of 3000bp.