BBS741 Project #4

Deadline: January 4, 2019 @ 11:59 pm
Email to BBS741.2018@gmail.com
with LastName_FirstName_Project1 as subject line

Part 0 (0 points)

We will be using the HMMLearn Python package to train HMMs as a part of this project (https://hmmlearn.readthedocs.io/en/latest/).

Overview of data:

As mentioned in lecture, MMs and HMMs may be used to predict the locations of CpG islands, or regions where CpG dinucleotides are more prevalent than background, in the genome. We have a list of known CpG islands for the hg38 human genome assembly downloaded from UCSC, with associated reference sequence. These come in two formats: one with surrounding negative regions (coordinates in cpg.withbg.bed, sequences in cpg.withbg.fa) and one with just the known positive regions (cpg.fa). We also have randomly shuffled negative regions (cpg.shuffled.fa) and a few test sequences which have several CpG islands within them at mystery locations (cpg.test.fa).
https://www.dropbox.com/sh/79xsktgfestkcxs/AADKldOsThOCJaFxjVvGDlwra?dl=0

Part I - Using a Markov Model to annotate known regions (20 points)

Using the HMMLearn package, initialize two Markov Models, one for positive CpG islands and one for negative non-CpG islands, using the transition probability matrices given on slide 11 of the HMM lecture. Use the MM to score 500-1,000 regions from cpg.fa and cpg.shuffled.fa with a log-likelihood score (see slide 11 from the lecture). You will need to convert the sequences from the FASTA file to a list of integer symbols before annotating regions (see the tutorial/documentation here https://hmmlearn.readthedocs.io/en/latest/tutorial.html#training-hmm-parameters-and-inferring-the-hidden-states). Make a histogram of the scores generated for each of the two sets.

- Do the two sets segrate into distinct classes?
- Is there a score that seems reasonable to use as a threshold for CpG island regions based on this histogram?

Part II - Using a Markov Model to annotate unknown regions (20 points)

Using a sliding window approach and the MM models and threshold you chose in **Part I**, annotate CpG island locations within the test regions. You may need to play with different sliding window sizes and plot the log likelihood scores against coordinates to see what sort of window size is appropriate for scoring. Count the number of CpG islands within each region and obtain their coordinates. You can use the known CpG island locations from cpg.withbg.bed and cpg.withbg.fa to validate your results if you want.

- How many CpG islands do each of the test regions contain?
- What are the coordinates of the CpG islands within these test regions?

Part III - Training a Hidden Markov Model (30 points)

Use HMMLearn to train a Hidden Markov Model to distinguish CpG island regions from non-CpG island regions. You will need to concatenate the training sequences together as described in the tutorial (link provided in **Part I**). You may also need to repeat the training step several times to obtain the most predictive model, as the training algorithm can get stuck in local minima (see the tutorial). You may design this model with any number of states you like, and may constrain any parameters you like during training (https://hmmlearn.readthedocs.io/en/latest/api.html explains how to fix parameters such as emission probabilities in the documentation for the HMM constructor). You may train with data from any of the four FASTA files provided. You do not need to use entire FASTA files train your model; you may subsample to speed up the training process.

- Describe the model design you used (how many states did you choose, did you fix any parameters such as emission probabilities during training, what biological interpretation did you expect your states to have, etc.)
- What transition probabilities and emission probabilities does your final model have? Do the values make sense?

Part IV - Using the HMM to annotate known and unknown regions (30 points)

Using the HMM trained in **Part III**, obtain a predicted sequence of states (CpG and not CpG) for 5-10 of the known regions with surrounding background. Given the known coordinates for the CpG islands within these regions, compute how well the predicted states match the true CpG locations. Next, obtain a predicted sequence of states for the test sequences, and compute how well the state sequences agree with the coordinates the MM predicted.

- What percentage of the states in the sequences predicted by the HMM were correct for the subset of known regions?
- What are the coordinates of the HMM-predicted CpG islands within the test regions?
- What percentage basepairs had the same state predicted by the HMM and the MM?

Grading Rubric

Submissions received after **January 4, 2019 @ 11:59 pm** will earn ½ credit

Part I (20 pts)	Script correctly implements MM	10 pts
	Script plots accurate histogram of log-likelihoods	5 pts
	Student identifies reasonable threshold for distinguishing CpG from non-CpG	5 pts
Part II (20 pts)	Script annotates sliding windows within test	13 pts

	regions	
	Script outputs reasonably accurate counts and coordinates of CpG islands within test regions	7 pts
Part III (30 pts)	Clear description of model parameter selection	10 pts
	Script correctly trains HMM on CpG and negative sequences	10 pts
	Script outputs reasonable transition probabilities	2.5 pts
	Script outputs reasonable emission probabilities for each state	2.5 pts
Part IV (30 pts)	Script annotates known and unknown sequences with CpG island locations	20 pts
	Script accurately compares HMM-predicted states with known CpG island locations	5 pts
	Script accurately compares HMM-predicted and MM-predicted CpG island locations	5 pts
Total		**100 pts**