Computational Molecular Biology and Bioinformatics

Gene Finding

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

August, 2024

Basics

Genes are contiguous segment of DNA sequences that encodes for an RNA product. It is known that genes have some features that controls this process. Therefore, a gene should be separable from a non-gene segment.

Gene finding or gene prediction is a learning process through which we can recognize whether a given DNA sequence is a gene or not. This helps to identify new genes from unknown regions of sequences.

We use classification methods that, unlike clustering methods, uses the class labels of the samples to learn the patterns from their features.

What is a hidden Markov model (HMM)?

An HMM is defined by a quintuplet $< Q, V, p, A, E >$, where each entity is defined as follows.

- Q denotes a finite set of states.
- V denotes a finite set of observation symbols per state
- p denotes the initial state probabilities.
- TM denotes the transition probability matrix. TM_{ij} represents the probability of transitioning from i to j $(i, j \in Q)$, i.e. $TM_{ij} = P(X_n = j | X_{n-1} = i)$.
- EM denotes the emission probability matrix. EM_{ij} represents the observability of a state i $(i \in Q)$ as j $(j \in V)$ at time t, i.e. $EM_{ij} = P(V_j \text{ at time t} | Q_t = i)$.

Note: Only emitted symbols are observable by the system but not the underlying random walk between states (they remain hidden).

Finding CpG islands using hidden Markov models

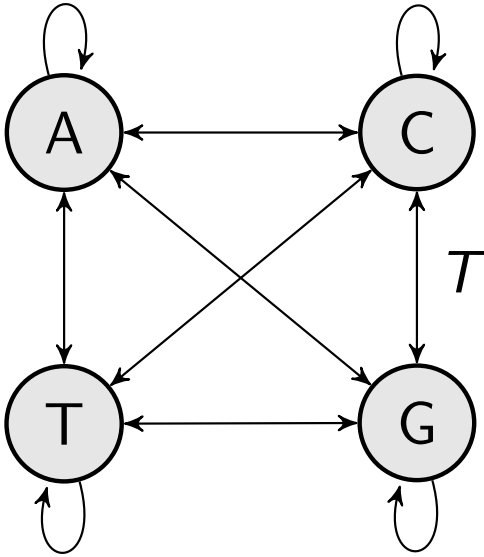
A CpG island is defined as a DNA region having GC content more than than 50% and an observed CpG versus expected CpG ratio ≥ 0.6. CpG islands are present in the promoters of many genes.

Methylation causes the methyl-C to get mutated to T in CpG, thereby making CpG dinucleotides much rarer across the genome. But interestingly, the methylation process gets suppressed around the promoters of many genes. So, CpG dinucleotides are much more frequent in promoters than elsewhere.

Problems:

1. Given a short sequence, does it come from a CpG island or not?
2. How to find the CpG islands in a long sequence?

Finding CpG islands using hidden Markov models



TM_{CG}/TM_{GC} (transition probabilities between C-G)

$TM =$

	A	**T**	**C**	**G**
A	0.11	0.23	0.34	0.26
T	0.14	0.31	0.21	0.24
C	0.07	0.51	0.16	0.23
G	0.19	0.21	0.28	0.29

Finding CpG islands using hidden Markov models

Training Set: Set of DNA sequences with known CpG islands.

Derive two Markov chain models as follows.

- '+' model – from the CpG islands
- '-' model – from the remainder of sequence

Compute the transition probabilities for each model as follows.

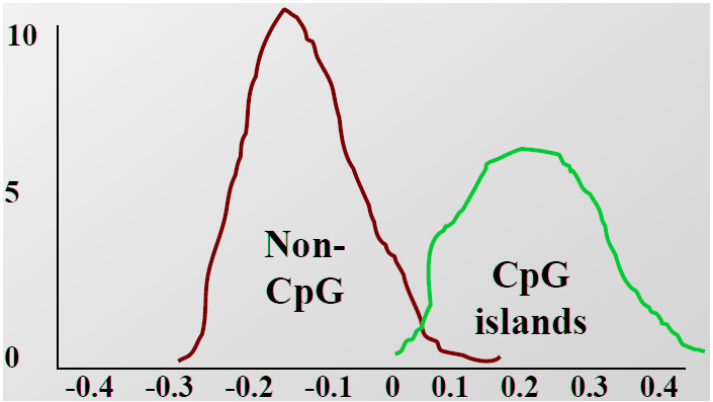
- $TM_{ij}^{+} = \frac{C_{ij}^{+}}{\sum_{j*} C_{ij*}^{+}}$, where C_{ij}^{+} denotes the count of the state i following state j *inside* the CpG islands
- $TM_{ij}^{-} = \frac{C_{ij}^{-}}{\sum_{j*} C_{ij*}}$, where C_{ij}^{-} denotes the count of the state i following state j *outside* the CpG islands

Finding CpG islands using hidden Markov models

Addressing Problem 1: Given a short sequence, does it come from a CpG island or not?

To use these models ('+' and '-') for discrimination, calculate the log-odds ratio of the sequence S (of length L) as follows.

$$LOR(S) = \log \frac{P(S|\text{'+' model})}{P(S|\text{'-' model})} = \sum_{n=1}^{L} \frac{TM^{+}_{S_{n-1}S_n}}{TM^{-}_{S_{n-1}S_n}}$$



Histogram of log-odds scores obtained for CpG islands and the rest

Finding CpG islands using hidden Markov models

Addressing Problem 2: How to find the CpG islands in a long sequence?

Calculate the log-odds score (LOR) for a window of, say, 100 nucleotides around every nucleotide, plot it, and predict CpG islands as ones with positive values. However, defining an appropriate window size is a challenge.

Finding CpG islands using hidden Markov models

Addressing Problem 2: How to find the CpG islands in a long sequence?

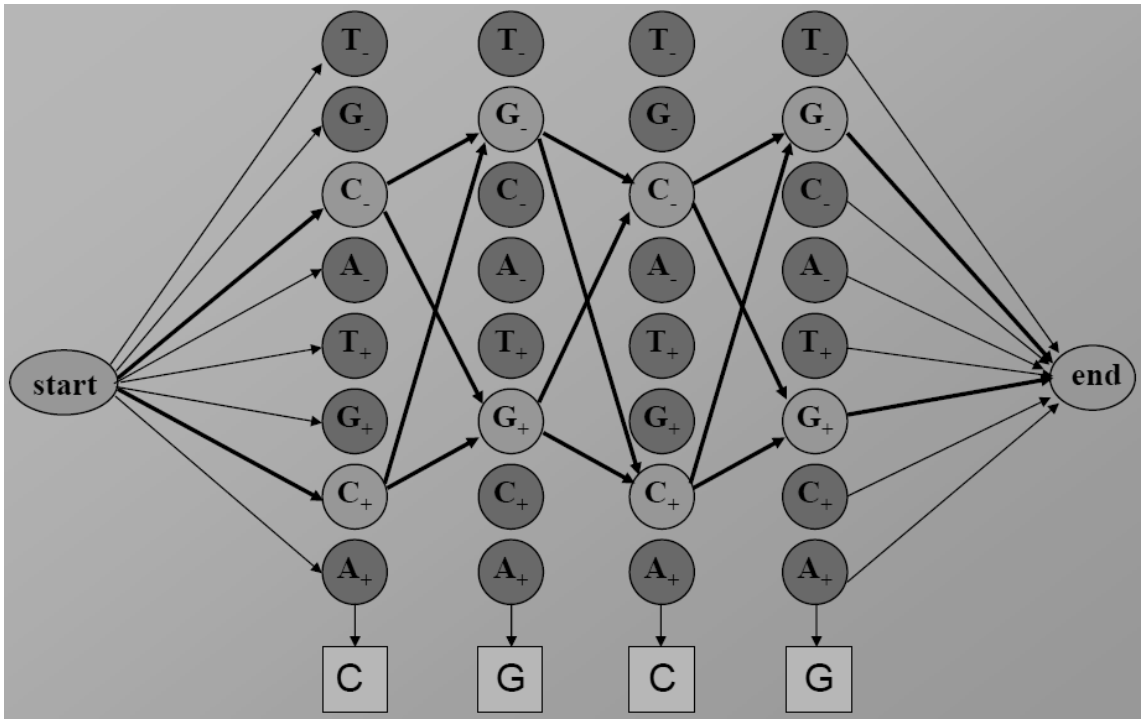
We can build a single HMM, combining both the Markov chains shown earlier, as follows.

- States: A_+, C_+, G_+, T_+, A_-, C_-, G_-, T_-
- Emission probabilities: Distinct for the '+' and the '-' states

Finding CpG islands using hidden Markov models

Addressing Problem 2: How to find the CpG islands in a long sequence?

We have to find out the most likely state path given the observed emissions.



Schematic view of labeled data

Consider the sequence segments as samples and the inherent or dynamic characteristics (e.g., expression profiles, n-mer frequencies, etc.) of the samples as features. Then, the labeled data for classification (training data) appears as follows.

Sample	**Feature 1**	**Feature 2**	**...**	**Feature m**	**Label**
S1	3.2	1.9	...	3.2	Gene
S2	1.2	3.7	...	5.3	Non-gene
S3	1	3.1	...	6.2	Non-gene
...	...	...	...	...	...
Sn	2.2	3.4	...	1.8	Gene

Features for gene finding

The following features are often considered for finding genes:

- The frequencies of n-mers.
- Splice junctions.
- Start and stop codons.
- Promoters such as TATA boxes, TF binding sites, and CpG islands.

Some popular gene prediction softwares are AUGUSTUS, FRAMED, GeneMark, etc.

k-nearest neighbor classification

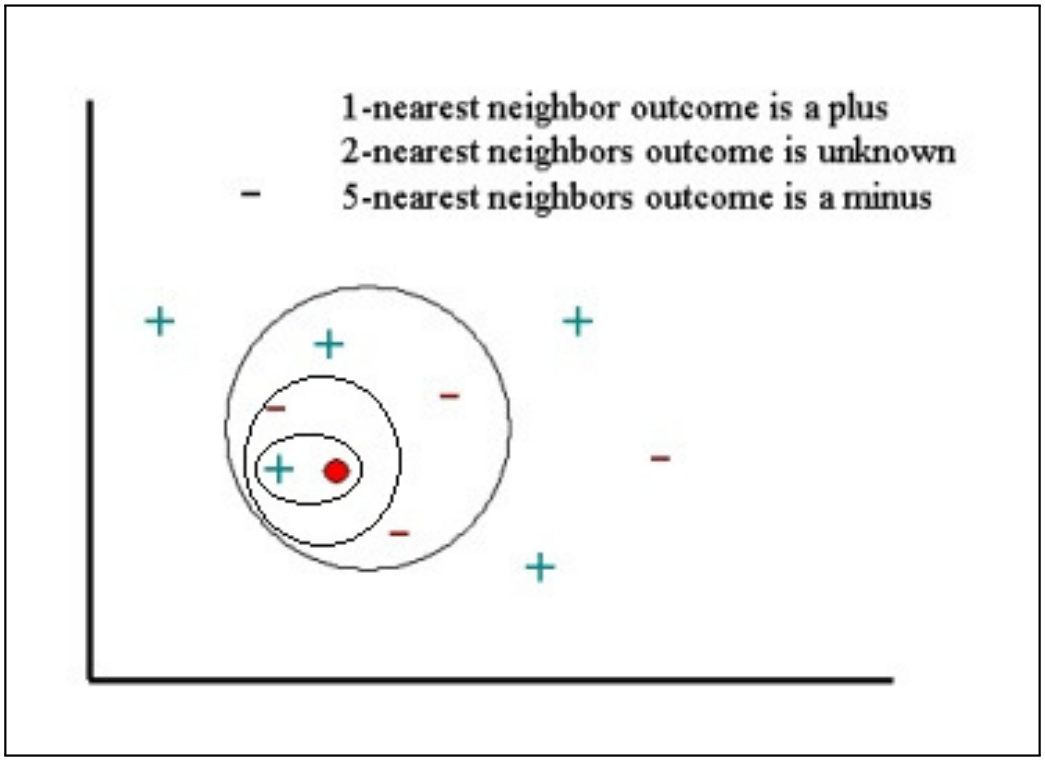
Step 1: Given a query sample S* and its feature values (test data) to be classified, we identify the k samples from the training data that are nearest to S*.
Step 2: The distance is calculated using one of the following measures:

- Euclidean Distance
- Minkowski Distance
- Mahalanobis Distance

Step 3: Return the class that represents the maximum of the k instances.

k-nearest neighbor classification

An example of k-nearest neighbor classification with 10 training samples is shown below.



k-nearest neighbor classification

The choice of k is crucial for this algorithm. A smaller k induces higher variance (less stable) and a larger k induces higher bias (less precise).

The value of k is generally considered as a function of the input size (i.e., the number of samples). The best alternative is $k = \sqrt{n}$, where n denotes the number of samples.

Hands-on

1. Download the following paper and do the following:
   Bruna, T., Lomsadze, A. and Borodovsky, M., GeneMark-ETP: automatic gene finding in eukaryotic genomes in consistency with extrinsic data. BioRxiv, pp. 1-41, 2024.
   i) Get the implementation from:
   `https://exon.gatech.edu/GeneMark/license_download.cgi`
   ii) Get the benchmark datasets from:
   `https://github.com/gatech-genemark/GeneMark-ETP-exp`
   iii) Could you identify some limitations of this implementation? How can you overcome that? Any suggestions?
2. The following paper gives new insights about gene regulation. Based on this, revisit the problem of gene finding:
   Duttke, S.H., et al., Position-dependent function of human sequence-specific transcription factors. Nature, 631:891-898, 2024.