Analysis of CpG Islands in Human Sex Chromosomes using Hidden Markov Models

Martí Díez Macià, Ainhoa López Carrasco
and Maria López Moriana

CLUSTERING METHODS AND ALGORITHMS IN GENOMICS AND EVOLUTION /
ALGORITHMS FOR SEQUENCE ANALYSIS IN BIOINFORMATICS

Index

INTRODUCTION

PIPELINE OF THE PROJECT

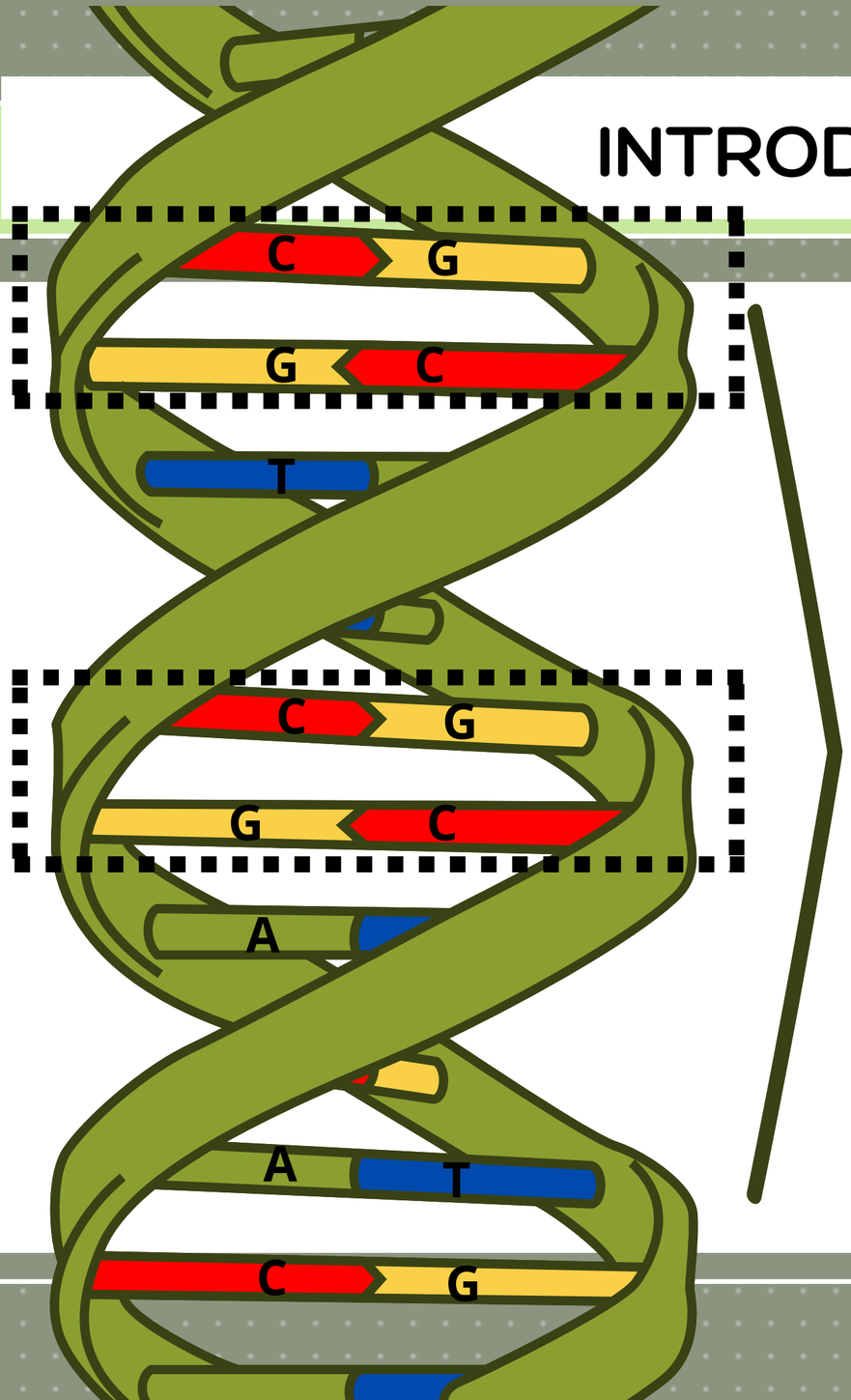
TRANSITION PROBABILITIES AND EVOLUTIONARY MODEL

COMPARATIVE DATA VISUALIZATION

DISCUSSION

CONCLUSION

INTRODUCTION: What are CpG islands?



- DNA regions where there is a high concentration of 'CG' dinucleotides
- Unusual in the genome

Biological importance:

- located near gene promoters,(transcription and expression of that gene)
- Crucial role in gene regulation
- Involved in DNA methylation; Abnormal methylation of CpG islands is associated with various diseases

INTRODUCTION: OUTLINE OF THE PROJECT

1 **Obtain fasta sequences**
(Chr X, Y, 21)

2 **Comparison Criteria in CpG island detection**

- **10% regions quantity CpG island**
- **20% regions size CpG island**

3 **CpG Island counting and mapping**
Coutning CpG regions and storing positions

4 **HMM Application**
Find hidden states

5 **Probability matrix calculation**

6 **Comparative Analysis**

HYPOTHESIS:
The detection and distribution of CpG islands across human chromosomes X, Y are influenced by chromosome size and the defined percentage threshold, affecting their identification and biological implications

METHODOLOGY
TRANSITION PROBABILITIES AND EVOLUTIONARY MODEL

Visual example: CpG **ATCGTCGCGAATCATG**
CCCCCCCCCCCCNNNNN No CpG

Identifying CpG Islands and States

Create function to count CpG islands in sequences in blocks of 500 base pairs (bps)

Functions: "contar_cpg_islands"

10% Threshold

At least 50 'CG' dinucleotides, high concentration of CpG islands.

20% Threshold

At least 100 'CG' dinucleotides, region where larger CpG islands are found

Usage of Hidden Markov Model (HMM)

Find hidden states, and indicate state for each nucleotide on the sequences.

Functions: "obtener_posiciones_nucleotidos_cpg", "obtener_posiciones_nucleotidos_no_cpg" , "etiquetar"

METHODOLOGY
TRANSITION PROBABILITIES AND EVOLUTIONARY MODEL

Probability Matrix

Transition Matrix

```
Normalized transitions:
{'C': {'C': 0.9984823529411765, 'N': 0.0015176470588235294}, 'N': {'C': 1.6507418913926747e-06, 'N': 0.9999983492581086}}
```

Emission Matrix

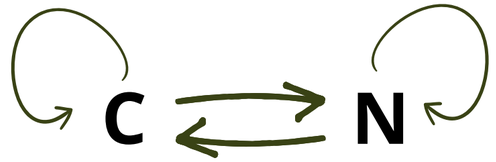
```
Unnormalized emissions:
{'C': {'A': 22607, 'C': 59851, 'G': 62471, 'T': 22678}, 'N': {'A': 46593325, 'C': 30439987, 'G': 30386600,
T': 46672052}}

Normalized emissions:
{'C': {'A': 0.13488100139015674, 'C': 0.3570912909365361, 'G': 0.3727230962907277, 'T': 0.1353046113825795}
 'N': {'A': 0.3023734904177093, 'C': 0.19754428595640458, 'G': 0.19719782402150446, 'T': 0.3028843996043817
```

Using these matrices we were able to map the landscape of CpG islands along chromosomes

Transition Probability in HMM

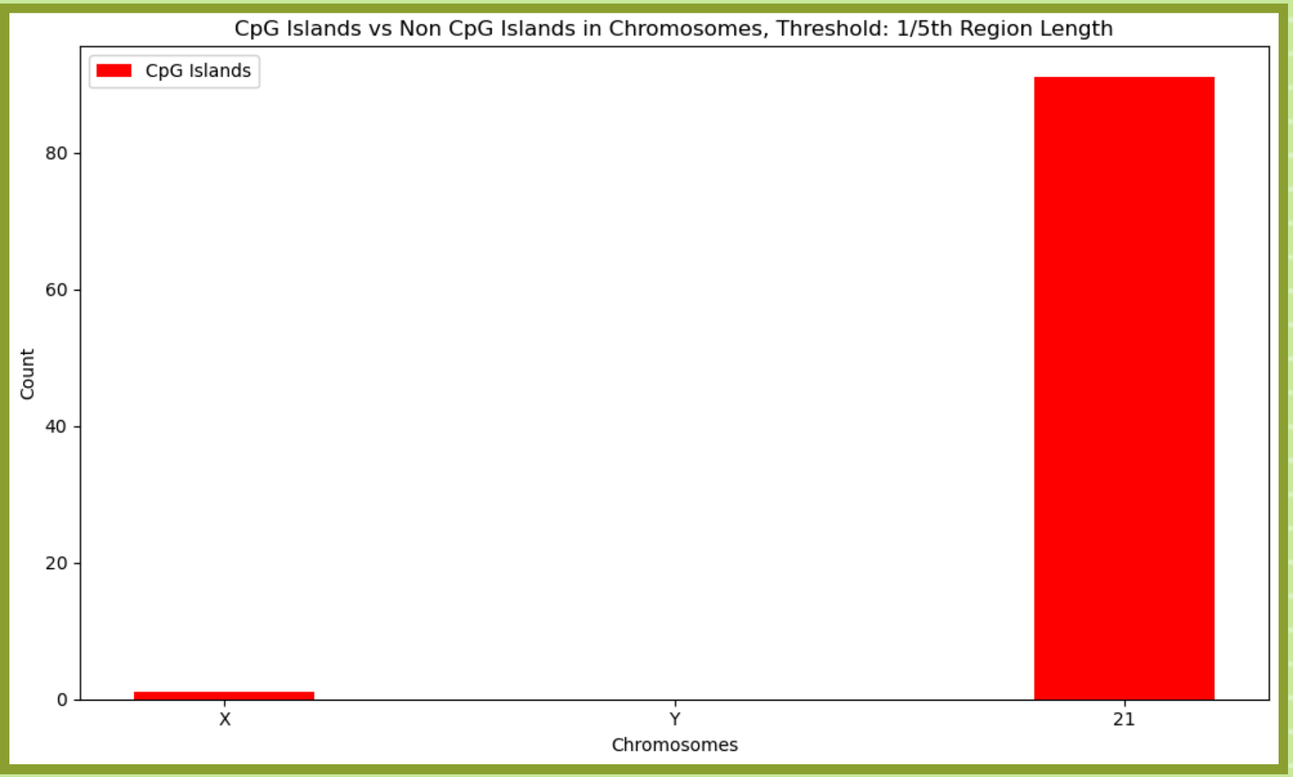
Likelihood of moving from one state to another

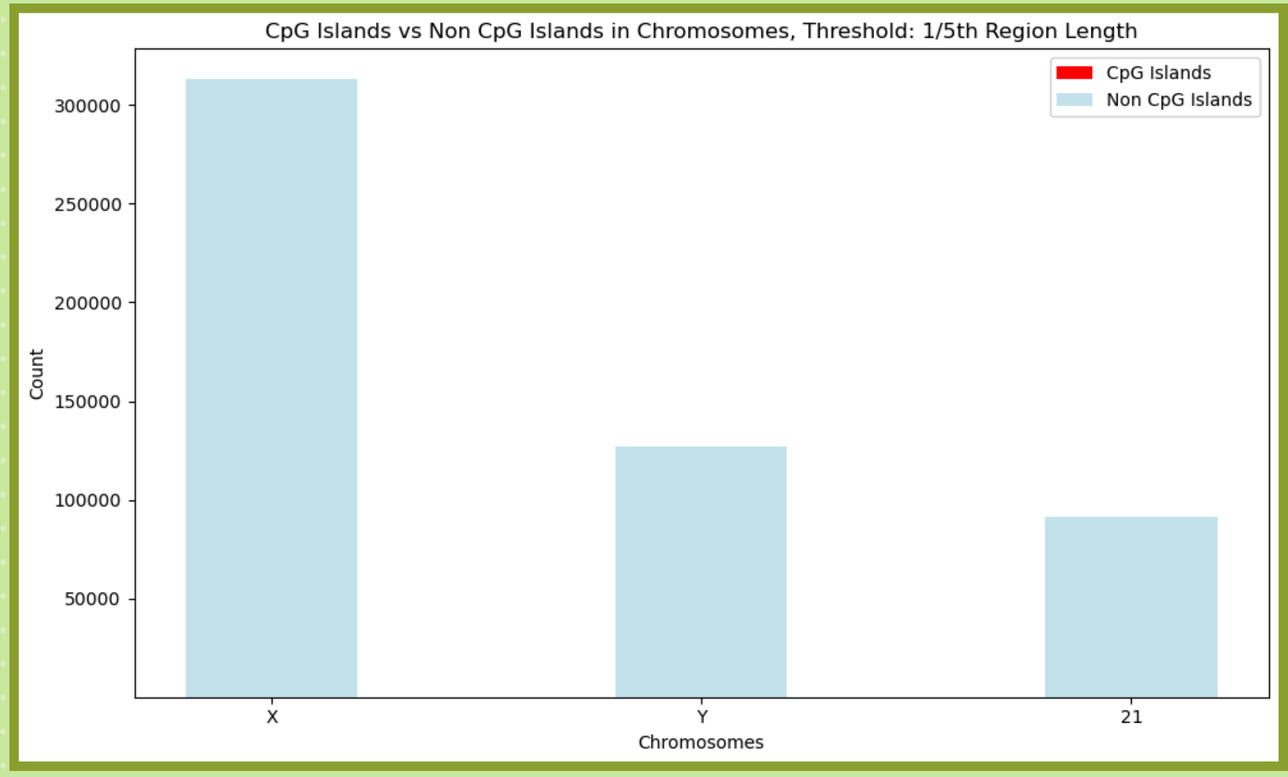


Normalize;
Divide by length to obtain probability

METHODOLOGY
COMPARATIVE DATA VISUALIZATION

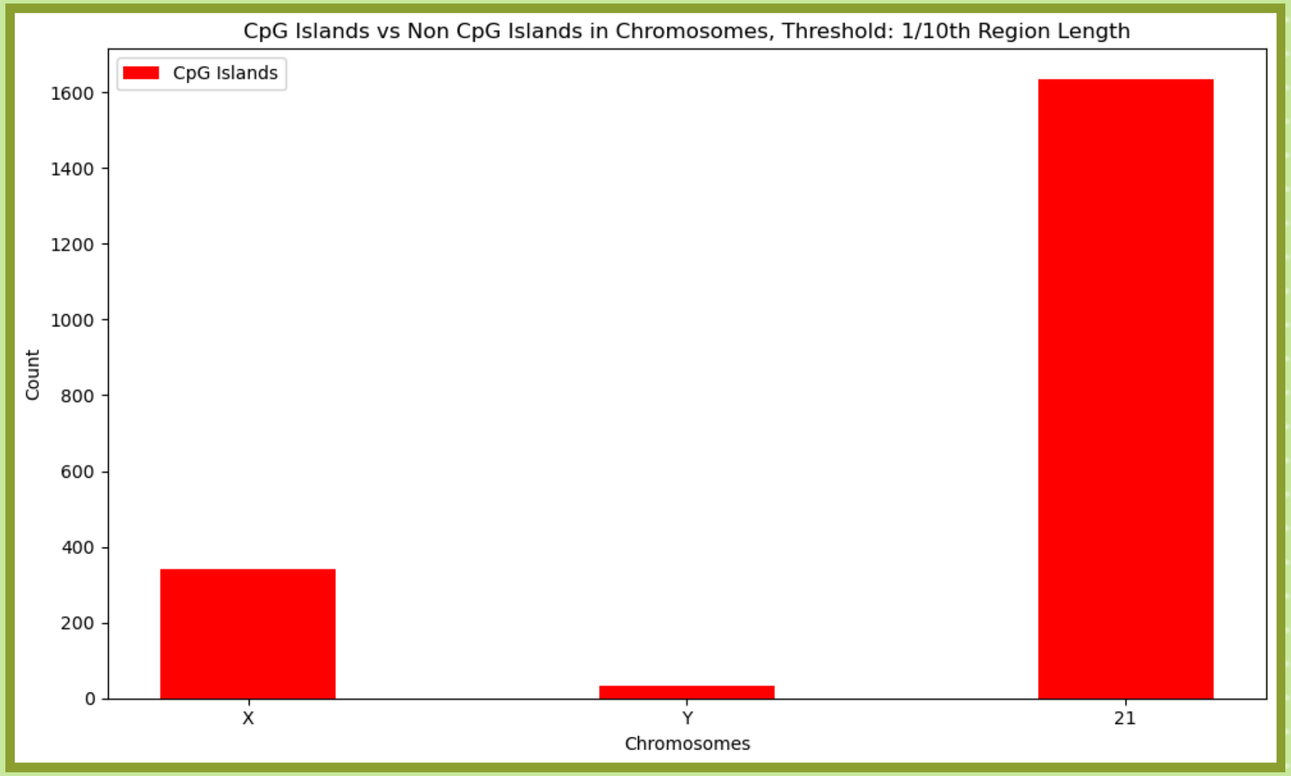
COMPARISON AMONG CHROMOSOMES, 20% THRESHOLD:
CHROMOSOME 21 HIGHER PROPORTION OF CPG ISLANDS THAN CHROMOSOME X
CHROMOSOME Y BARELY HAS CPG REGIONS.

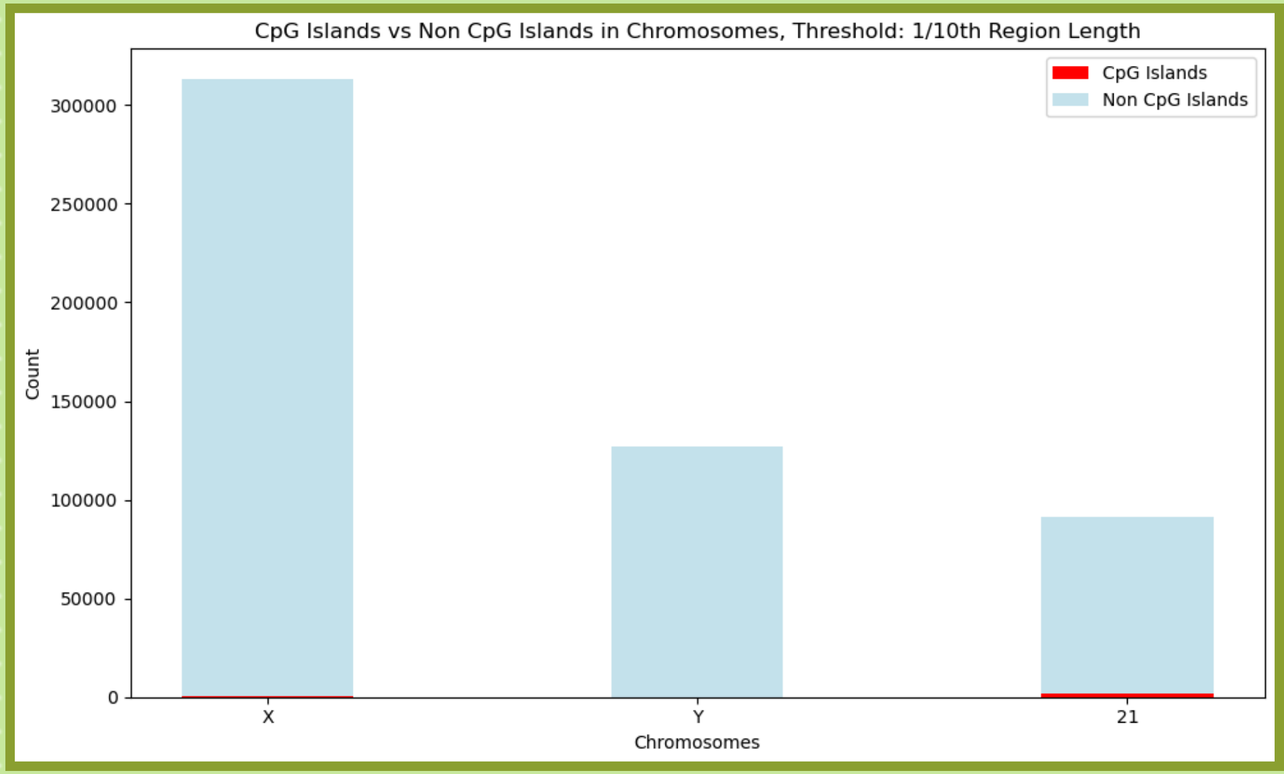




METHODOLOGY
COMPARATIVE DATA VISUALIZATION

COMPARISON AMONG CHROMOSOMES, 10% THRESHOLD:
WE CAN SEE THE SAME PROPORTIONS, BUT MORE EVIDENT IN THE 10% THRESHOLD

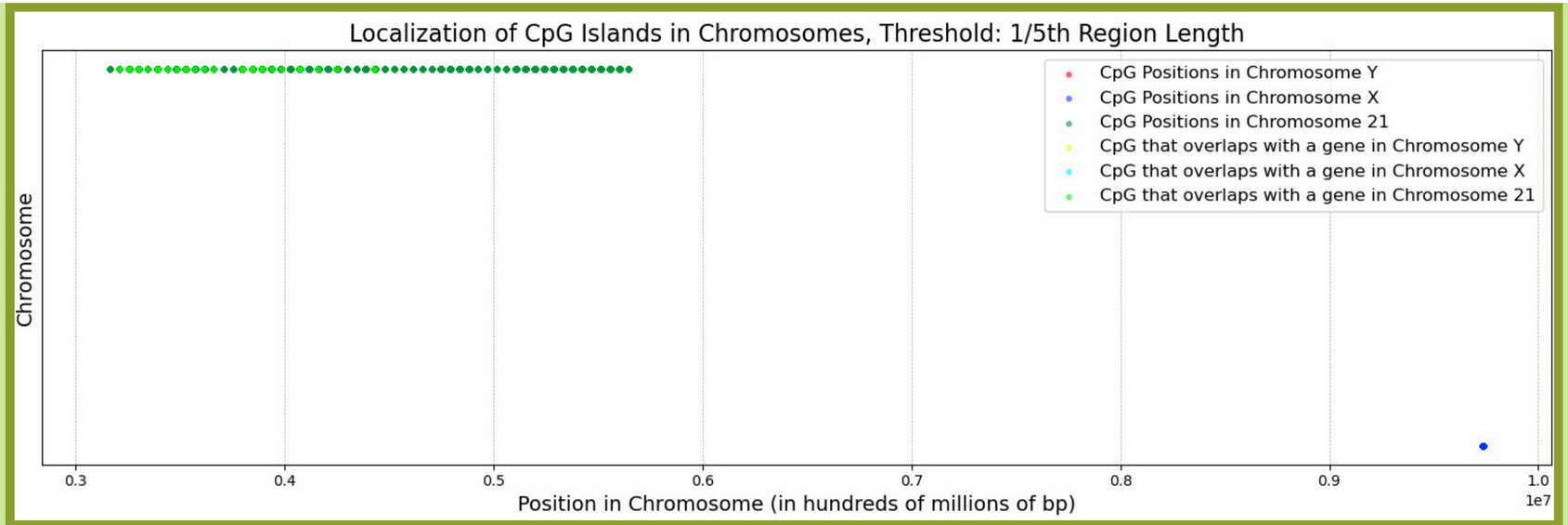


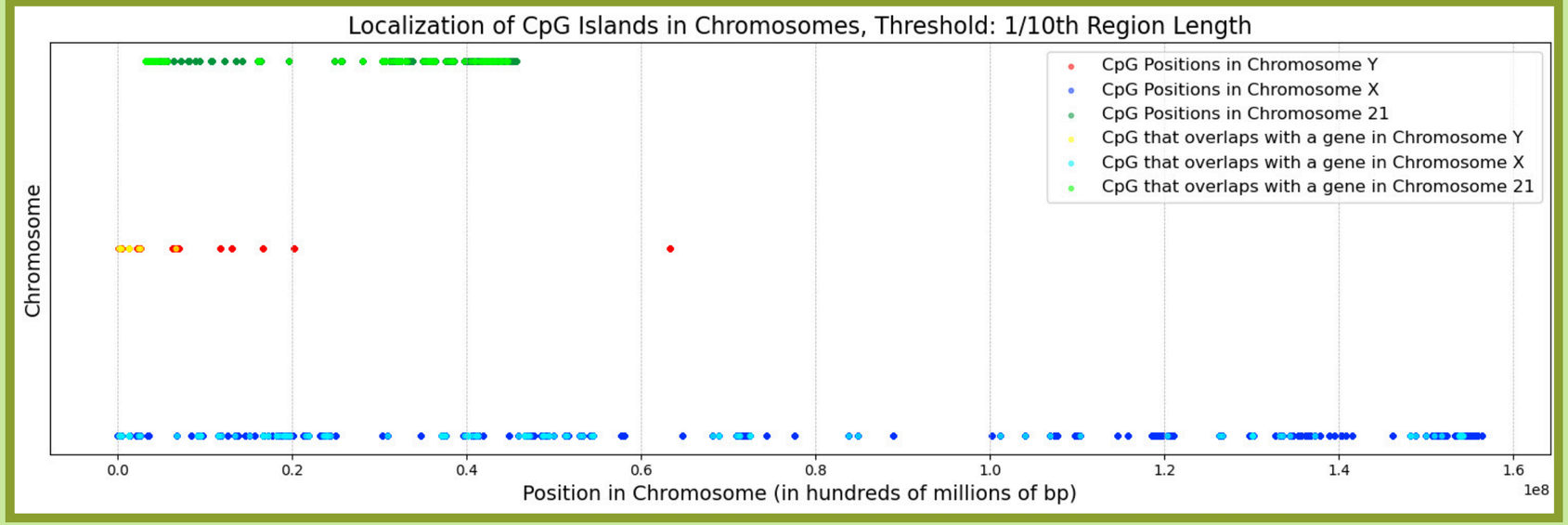


LOCALIZATION CpG AND OVERLAPPED GENES
COMPARATIVE DATA VISUALIZATION

Comparison of the **positions of the CpG** islands found in the Chromosomes (**X**, **Y**, **21**) and the part that would **overlap with a codifying gene.**

Intuition:
CpG islands located near conserved regions, could mean they act as **stable** elements for **consistent gene expression patterns**





DISCUSSION: About the model

Assumptions of the model

- Independence between observations and states.
- Constant transition % between states.

Limitations of the Model

- Computationally intensive.
- Current state depends on the previous one, not what leads to the previous.
- Risk of inaccurate calculations.
- Diversion from actual biological realities.

DISCUSSION: Comparing with articles

Article 1: Prediction of CpG Islands with HMMS - Thierry Grimm

- Algorithmic differences
- Similarity in probabilities

Article 2: Using Hidden Markov Models to Infer Locations of CpG Islands and Promoter Regions - Karthik Mittal

- Different CpG island definition
- Alignment and emission probabilities

Article 3: Correlating CpG islands, motifs, and sequence variants in human chromosome 21 - Nick Cercone, Leah Spontaneo

- Functional focus/promoter detection
- Divergence in CpG island size

CONCLUSION: Key points

CpG prediction
Using HMM we have predicted the location of CpG islands across the selected chromosomes.

Significance in gene regulation
Presence of CpG islands near gene promoters in conserved regions, affecting DNA methylation patterns and gene expression.

Variance of CpG islands between chromosomes
We have found that there are varying amounts of CpG islands.

Analytical approach
We were able to create a model that effectively explains the distribution and appearance of CpG islands using HMM.

Future Research
- Complex interactions between CpG islands and gene regulation.
- Evolutionary significance of CpG islands

Challenges found
Limitations in our model, large amount of data, computational power at hand.

Relevant links for the articles metioned.

Article1: https://github.com/thierrygrimm/cpg-island-hmm/blob/master/Jupyter%20Notebooks/CpG%20islands%20Hidden%20Markov%20Model.ipynb,
Article2: https://medium.com/analytics-vidhya/using-hidden-markov-models-to-infer-locations-of-cpg-islands-and-promoter-regions-480db92b6472,
Article 3: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-S2-S10

Thanks for your attention