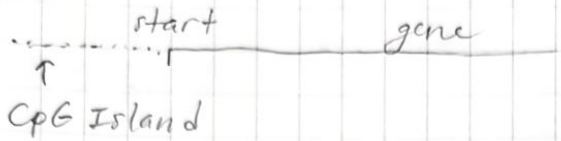
The CpG Island Problem

11/25/2025



CG dinucleotide CpG to distinguish from C-G base pair across the two DNA strands

↖ on the same strand

When CG occurs the C nucleotide (cytosine) is typically modified by methylation. Result methyl-C mutates into T

$C_{\text{H}_3}G \rightarrow TG$ # of CGs decrease

This has consequence that CpG dinucleotides are rarer in the genome than what would have been expected from independent probabilities of C and G

For biological reasons, the methylation process is suppressed in short stretches of the genome, such that around promoters or "start regions" of genes.

In these regions, you see many more CpG dinucleotides than elsewhere. And in fact more C and G nucleotides in general. Such regions are called CpG Islands. They are a few hundred to a few thousand base pairs long.

Problems we want to solve:

1) Is the region from a CpG Island?
2) How would we find the CpG Islands in a sequence?

Markov Chains
HIDDEN