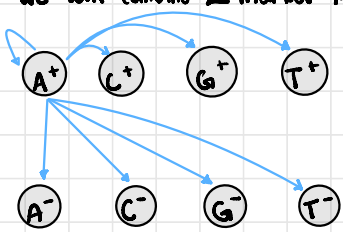
CpG ISLANDS

- in the human genome, when C and G occur consecutively (denoted CpG), the C nucleotide is typically chemically modified by methylation, resulting in methyl-C
- Such methyl-C is mutated with high probability into nucleotide T

→ CpG's are rarer in the genome than would be expected by chance

- HOWEVER, near the beginning of a gene, methylation is SUPPRESSED, so CpGs are enriched compared to the rest of the genome. These regions are 100s–1000s of basepairs long.
- Question: Given a DNA region, how can we decide whether it is a CpG island, or whether it contains such islands?
- solution: we will combine 2 markov models: the "+" model [CpG island] and the "−" model [not CpG island]



need to define transition probabilities between all states in both "+" and "−" models

Hidden states: $A^+, C^+, G^+, T^+, A^-, C^-, G^-, T^-$

emissions: $A^+ \to A,\ C^+ \to C,\ G^+ \to G,\ T^+ \to T,\ A^- \to A,\ C^- \to C,\ G^- \to G,\ T^- \to T$

For the "+" model: look in the CpG islands database to compute frequencies of one nucleotide being followed by another

	A^+	C^+	G^+	T^+
A^+	.18	.274	.426	.12
C^+	.17	.368	.274	.188
G^+	.16	.34	.375	.125
T^+	.079	.355	.384	.182

For the "−" model:

	A^-	C^-	G^-	T^-
A^-	.3	.21	.29	.21
C^-	.32	.298	.078	.30
G^-	.248	.246	.298	.208
T^-	.177	.239	.292	.292

notice how CG is way more probable in "+" model than "−" model

▷ defining transitions between "+" and "−" states is a little trickier to compute from data, so let's just assume that the probability of transitioning between models is quite low.

ex suppose we have a sequence: C G C G

possible state sequences:

- $C^+ \ G^+ \ C^+ \ G^+$ ← X
- $C^- \ G^- \ C^- \ G^-$ ← Y
- $C^+ \ G^- \ C^+ \ G^-$ ← Z

$P(X) > P(Y) > P(Z)$

→ so of the 3 state sequences (X, Y, and Z), X is the most probable.

Think about how this relates to the Viterbi Algorithm