

CpG ISLANDS

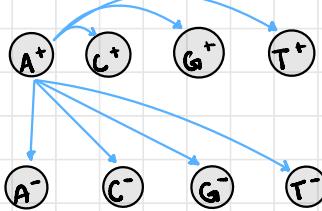


- in the human genome, when C and G occur consecutively (denoted CpG), the C nucleotide is typically chemically modified by methylation, resulting in methyl-C
- Such methyl-C is mutated with high probability into nucleotide T
 - \rightarrow CpGs are rarer in the genome than would be expected by chance
- 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?

[CpG island]

[not CpG island]

Solution: we will combine 2 markov Models: the "+" model and the "-" model



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

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

Emissions: A , C , G , T

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

For the "-" model:

	A^-	C^-	G^-	T^-
A^-	.3	.21	.29	.21
C^-	.32	.29	.28	.30
G^-	.284	.246	.298	.208
T^-	.177	.231	.292	.292

	A^-	C^-	G^-	T^-
A^+	.16	.124	.116	.12
C^+	.17	.368	.274	.198
G^+	.16	.34	.375	.125
T^+	.079	.355	.385	.182

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

and CT is more probable in "-" model

\triangleright 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

$C^+ \quad G^+ \quad C^- \quad G^- \quad \leftarrow X$

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

possible state sequences:

$C^- \quad G^- \quad C^- \quad G^- \quad \leftarrow Y$

$C^+ \quad G^- \quad C^- \quad G^- \quad \leftarrow Z$

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

Think about how this relates to the Viterbi Algorithm