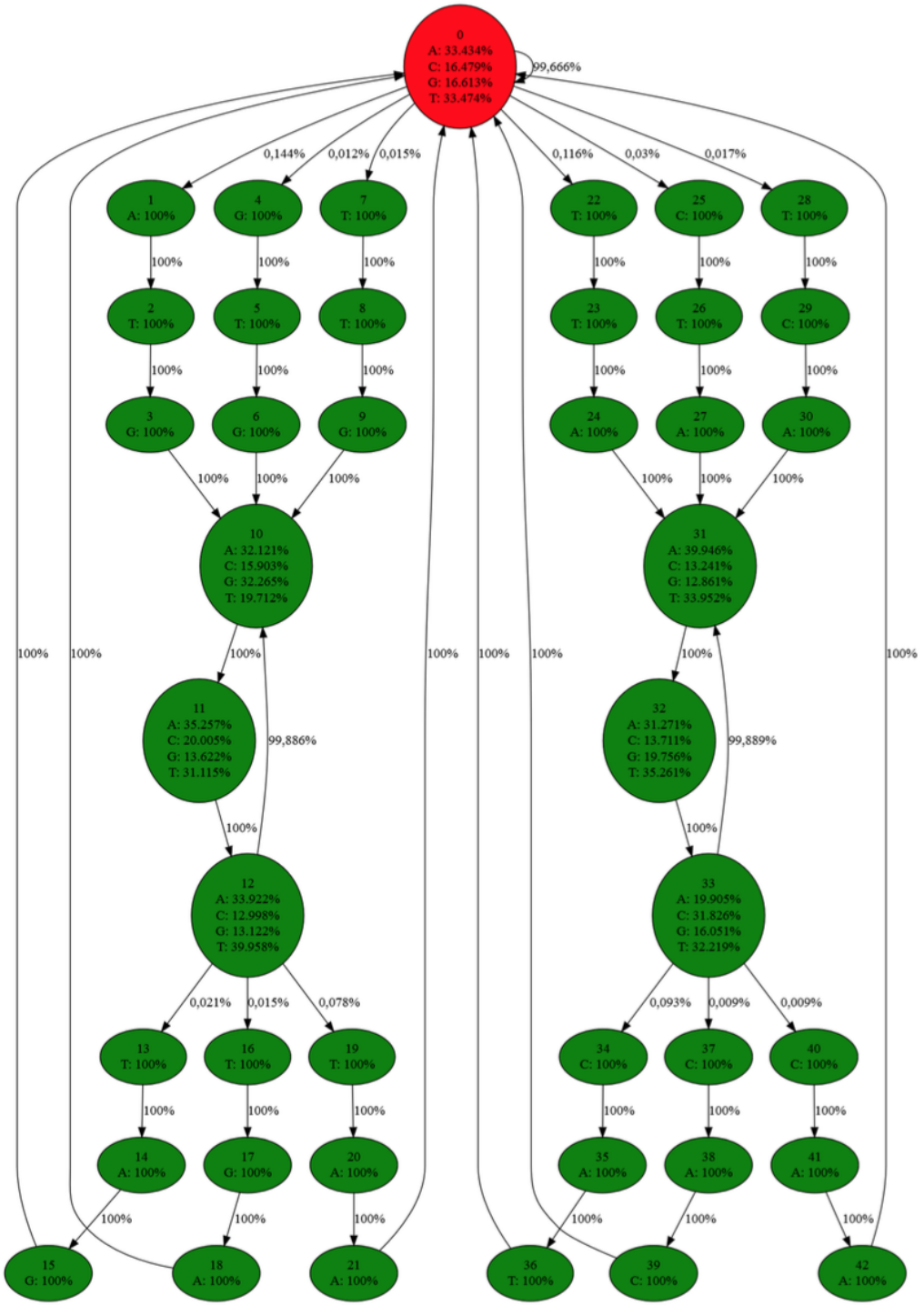
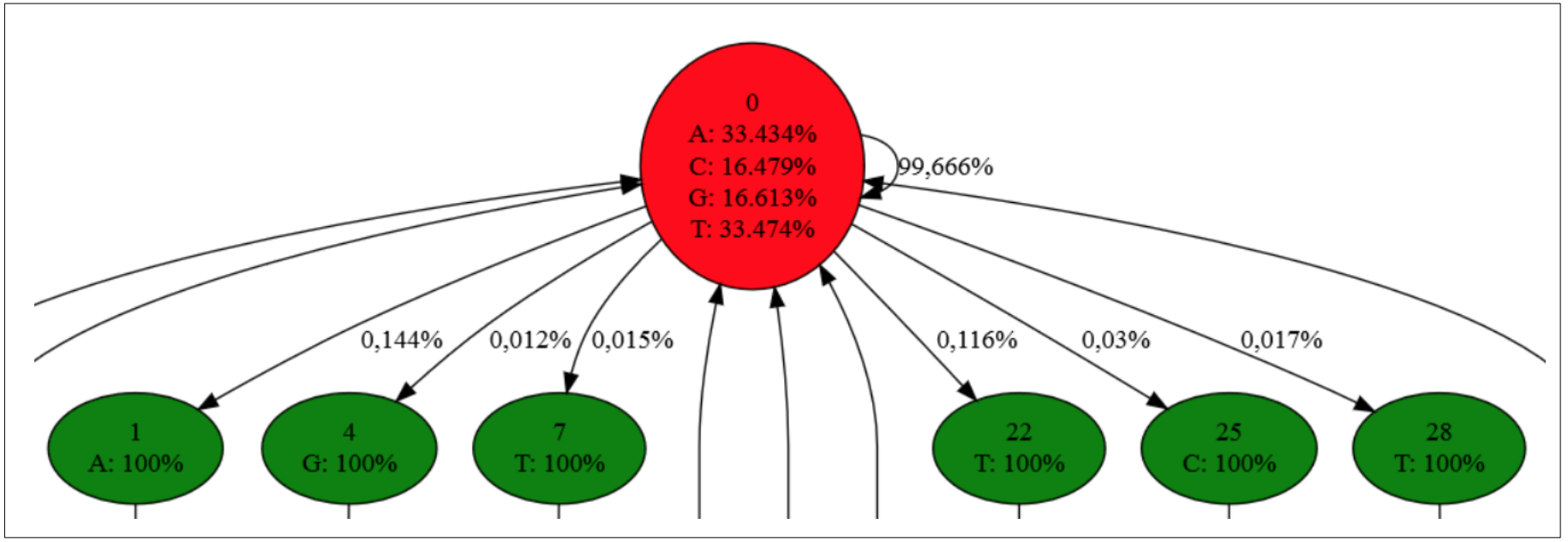
A typical model



Training by counting



In theory, we are given (**X**,**Z**) pairs, but we are given **(X**, "**Z**") pairs, where the NCR-annotations have to be translated to **Z**s.

Also, we may ignore rare start and stop codons, i.e. a gene that starts (or ends) with a ignored start (or stop) codon does not correspond to a path in your model.

Training by counting – Typical solution

To set the transition probabilities:

N → N	N → CCC, where CCC is ATG	N → RRR, where RRR is TTA
	N → CCC, where CCC is GTG	N → RRR, where RRR is CTA
	N → CCC, where CCC is TTG	N → RRR, where RRR is TCA

We count:

#(N->N) = no. of occurrences of “NN” our annotations.

#(N → CCC, were CCC is XYZ) = no. of occurrence “NCCC” in our annotations, where CCC is an annotation of XYZ (in our training data).

#(N → RRR, were RRR is XYZ) = no. of occurrence “NRRR” in our annotations where, RRR is an annotation of XYZ (in our training data).

We compute:

Total = #(N → N) + #(N → CCC where CCC is XYZ) + #(N → RRR, where RRR is XYZ)

We set:

P(N → X) = #(N → X) / Total for each of the 7 transitions