

# Computing the transition probabilities

- In order to recover transition probabilities  $\mathbf{T}(t)$  from the rate matrix  $\mathbf{Q}$ , one computes the matrix exponential  $\mathbf{T}(t) = \exp(\mathbf{Q}t)$ , same as with standard nucleotide models, e.g. HKY85 or GTR
- Because the computational complexity of matrix exponentiation scales as the cube of the matrix dimension, codon based models require roughly  $(61/4)^3 \approx 3500$  more operations than nucleotide models
- This explains why codon probabilistic models were not introduced until the 1990s, even though they are relatively straightforward extensions of 4x4 nucleotide models

# Multiple substitutions

- The model assumes that point mutations alter one nucleotide at a time, hence most of the instantaneous rates (3134/3761 or 84.2% in the case of the universal genetic code) are 0.
- This restriction, however, does not mean that the model disallows any substitutions that involve multiple nucleotides (e.g., **ACT**⇒**AGG**).
  - This can be further relaxed with models supporting multiple nucleotide changes.
- Such substitutions must simply be realized via several single nucleotide steps, e.g., **ACT**⇒**AGT**⇒**AGG**
- In fact the  $(i, j)$  element of  $\mathbf{T}(t) = \exp(\mathbf{Q}t)$  sums the probabilities of all such possible pathways of duration  $t$ , including reversions
- Compare this to the naive NG86 parsimony approach.