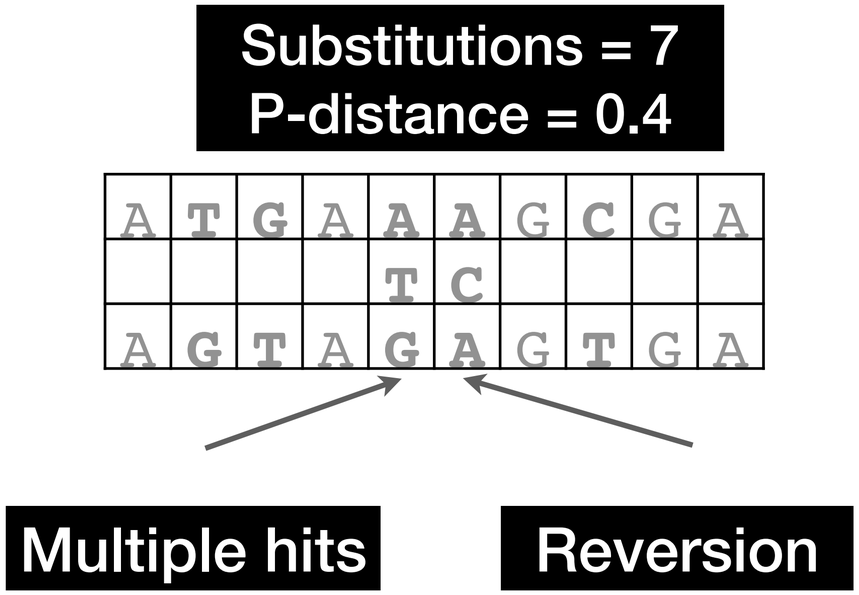
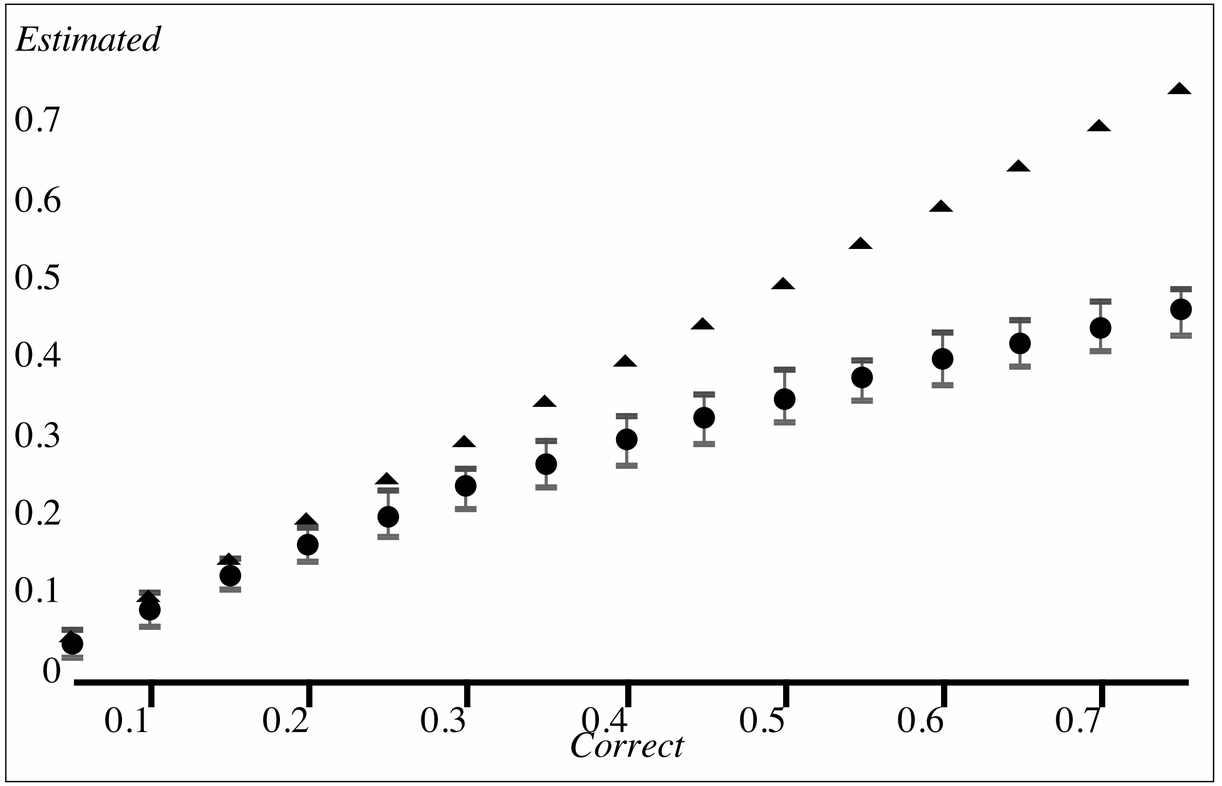
NG86 limitations: multiple substitutions

- How many synonymous and how many non-synonymous substitutions does it take to replace **CCA** with **CAG**?
- **Assume** the shortest path (minimum of 2 substitutions)
  - **CCA (Proline)** $\Longrightarrow$ **CAA (Histidine)** $\Longrightarrow$ **CAG (Glutamine)**
  - **CCA (Proline)** $\Longrightarrow$ **CCG (Proline)** $\Longrightarrow$ **CAG (Glutamine)**
- Average over the two possible paths: `0.5` synonymous and `1.5` non-synonymous substitutions.
- Intuitively, paths should **not** be equiprobable, e.g., because it should be more expensive to route evolution through (presumably) suboptimal intermediate amino-acids.

NG86 limitations: underestimation of substitution counts for higher divergence levels





- Simulated 100 replicates of 1000 nucleotide long sequences for various divergence levels (substitutions/site)
- Plotted simulated divergence vs that estimated by p-distance.

- Even for divergence of `0.25` (1/4 sites have mutation on average), p-distance already significantly underestimates the true level: `0.2125 (0.19-0.241 95% range)`
- Underestimation becomes progressively worse for larger divergence levels