Stavy vyšších rádov

Rád 0: emisná tabuľka e určuje $\Pr(S_i|A_i)$

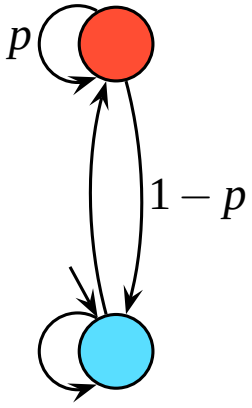
Rád 1: e určuje $\Pr(S_i|A_i, S_{i-1})$

A_i	S_{i-1}	a	c	g	t
	a	0.24	0.23	0.34	0.19
	c	0.30	0.31	0.13	0.26
	g	0.27	0.28	0.28	0.17
	t	0.13	0.28	0.38	0.21
	a	0.30	0.18	0.27	0.25
	c	0.32	0.28	0.06	0.35
	g	0.27	0.22	0.27	0.24
	t	0.20	0.21	0.26	0.33

. . .

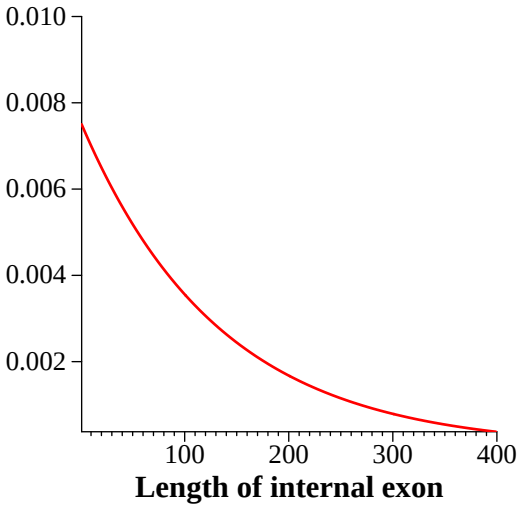
Na charakterizovanie exónov, intrónov atď používame rád 4-5.

Modeling length distributions



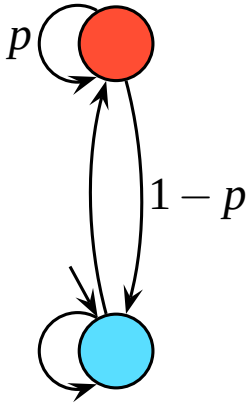
What is the length distribution of red segments generated by the model?

$$\Pr(\text{red segment of length } \ell) = p^{\ell-1}(1-p)$$

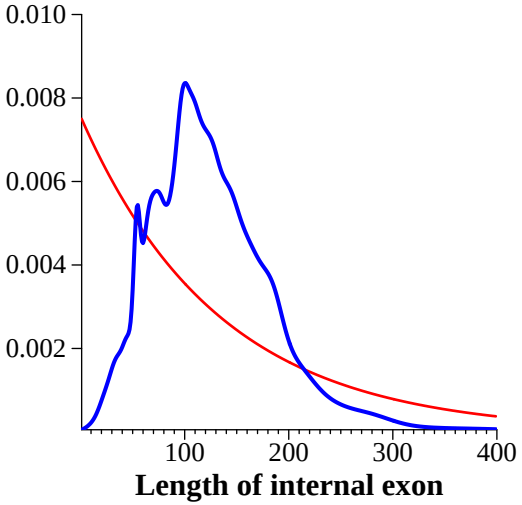


Modeling length distributions

What is the length distribution of red segments?



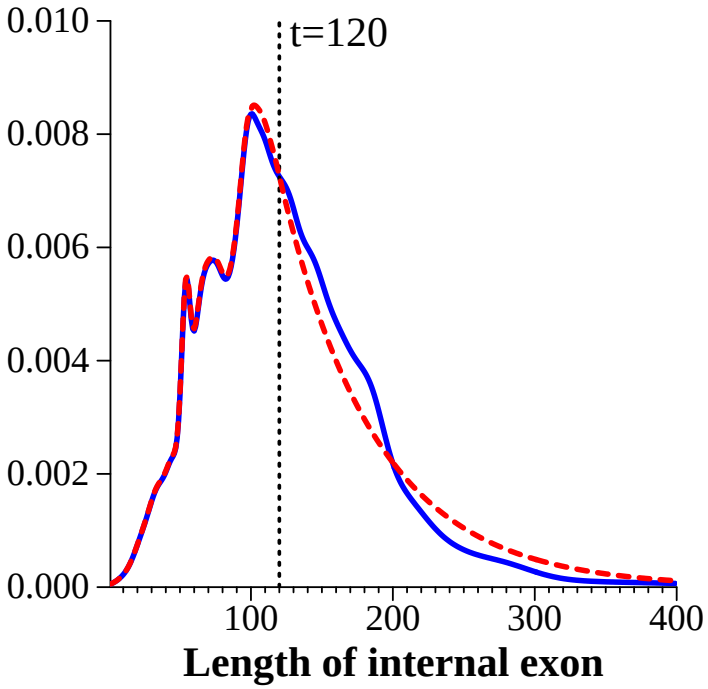
$$\Pr(\text{red segment of length } \ell) = p^{\ell-1}(1-p)$$



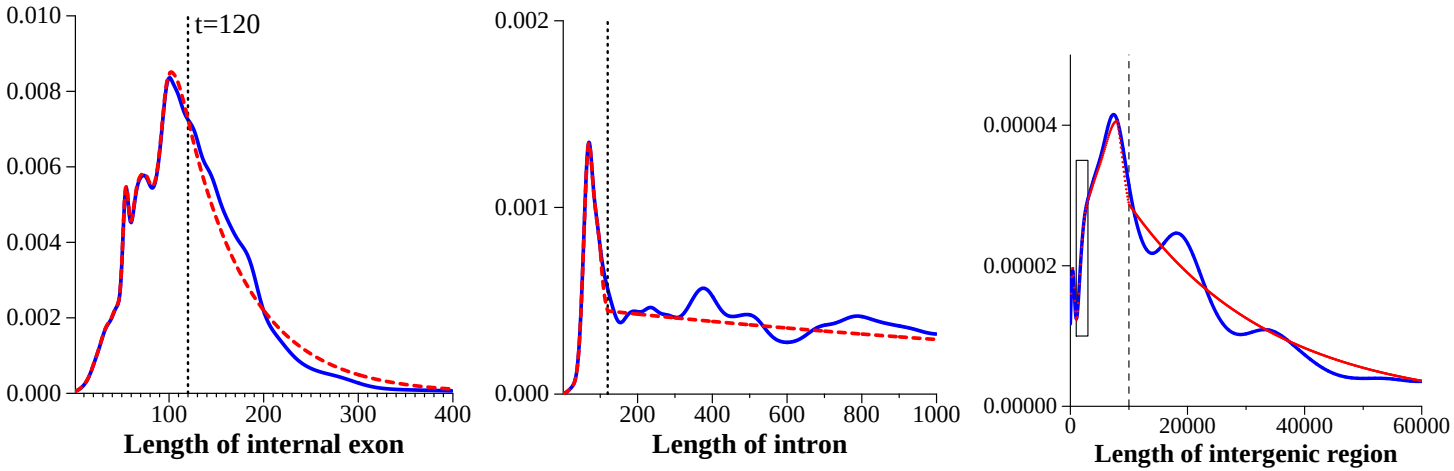
- Geometric distributions: bad model of real world; $O(n)$ time [Viterbi 1967]
- Arbitrary distributions: faithful model; $O(n^2)$ time [Rabiner 1989]
- Will show: geometric tails: better model; $O(nt)$ time.

Geometric tail distributions

- head (lengths $< t$): specify explicitly
- tail (lengths $\geq t$): geometrically decaying



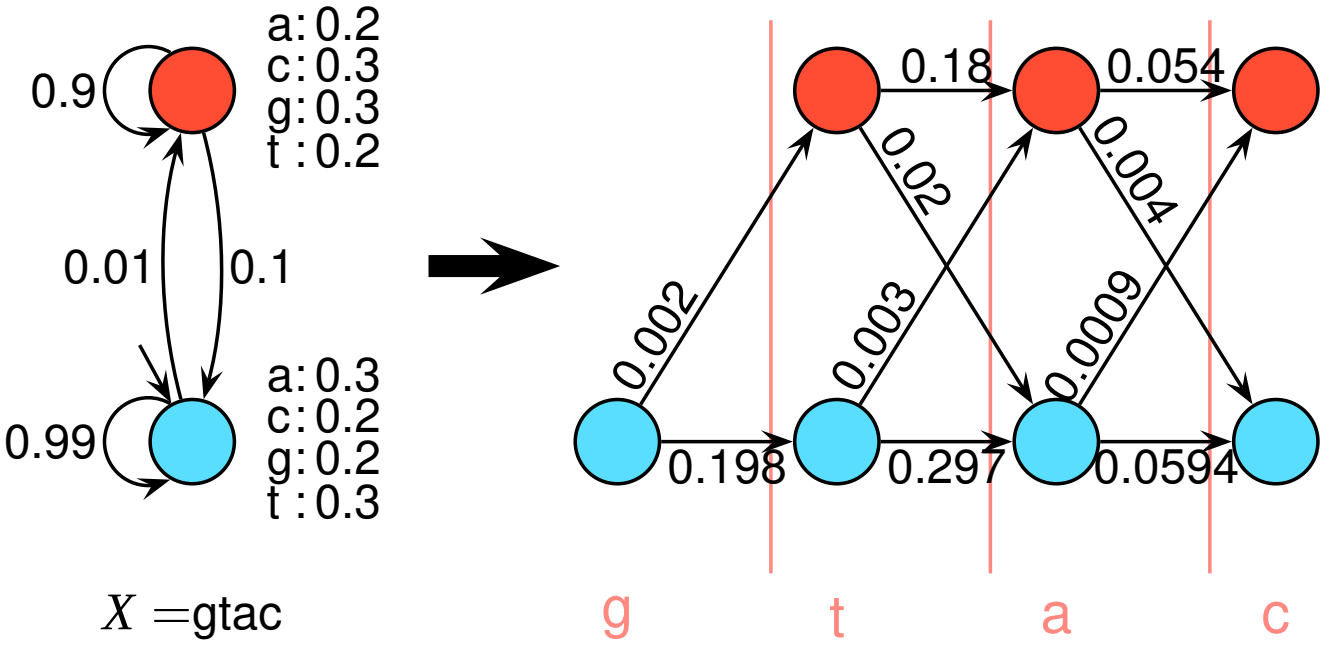
Geometric tail is a good approximation



- $O(nt)$ works for exons and introns
- Intergenic regions ($t \approx 10000$):
  - Use less accurate approximation
  - Better running time: $O(n\sqrt{t})$

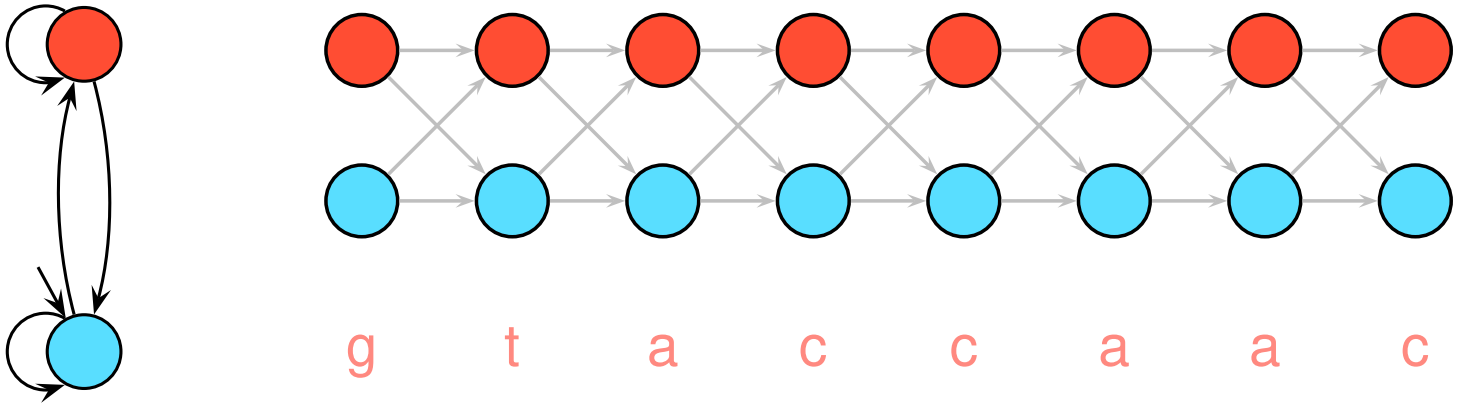
Viterbi algorithm: the most probable state path [Viterbi 1967]

(but geometric length distributions only)

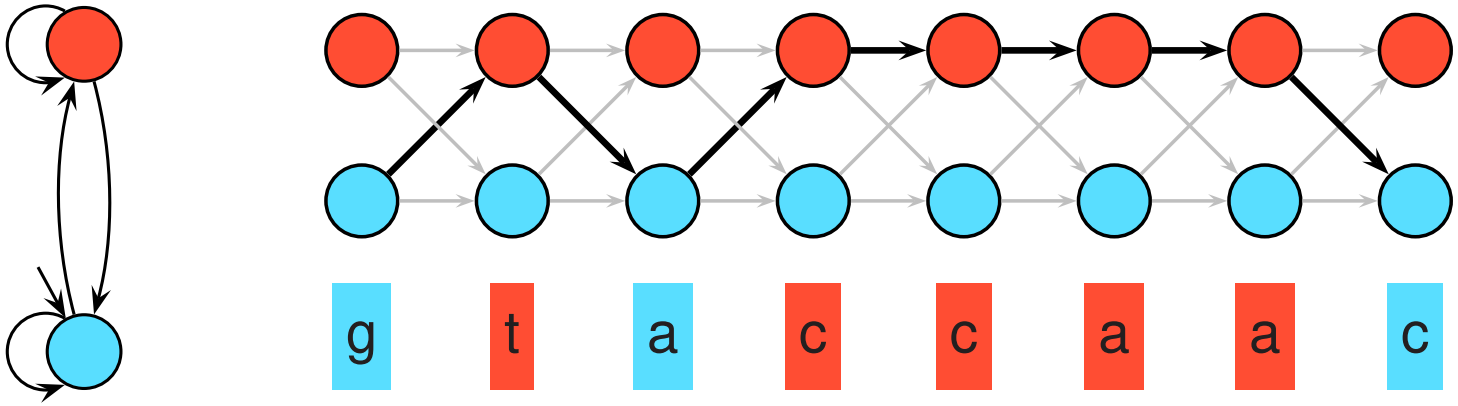


- Take $-\log$ of weights and compute shortest path in DAG
- Running time: $O(n)$

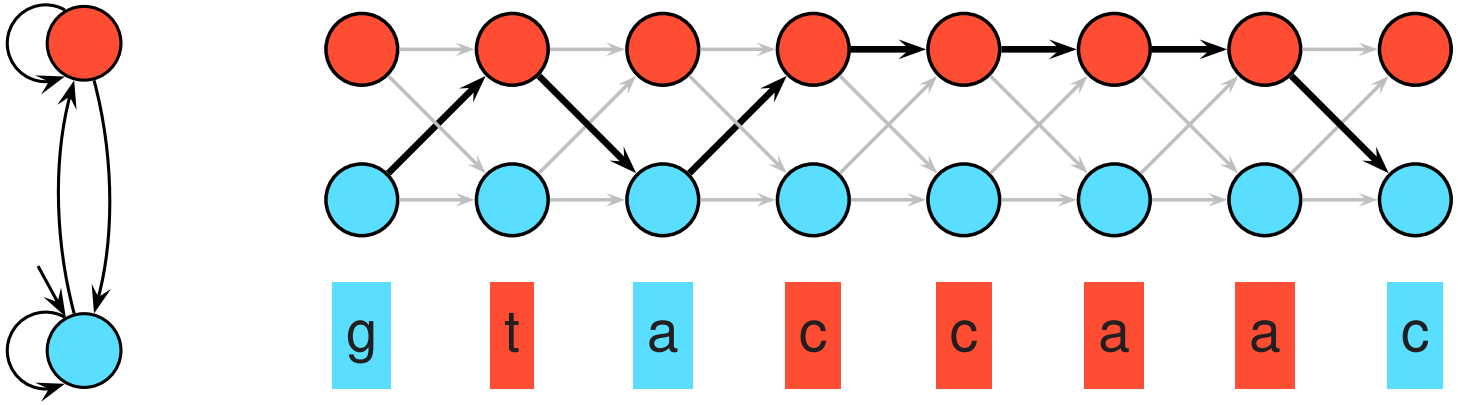
Viterbi algorithm – $O(n)$ time



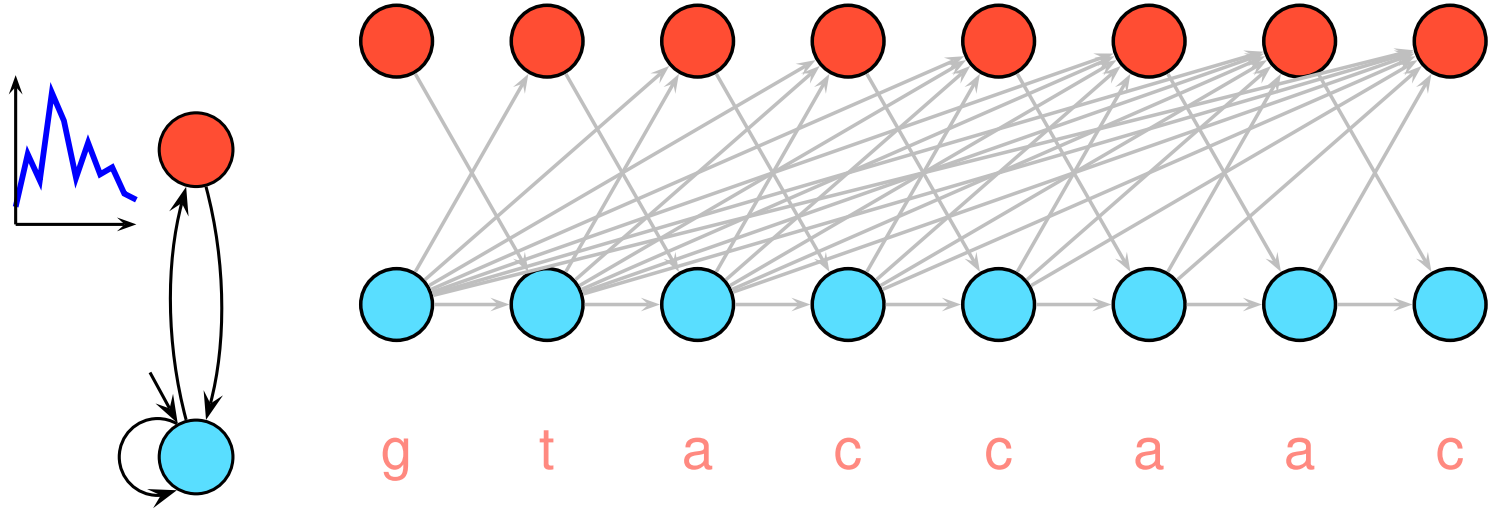
Viterbi algorithm – $O(n)$ time



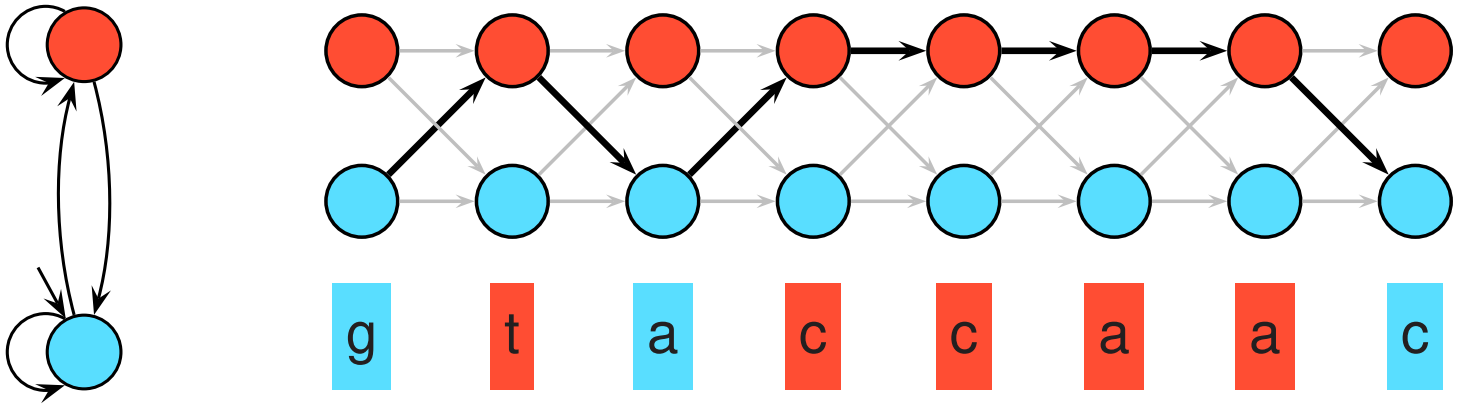
Viterbi algorithm – $O(n)$ time



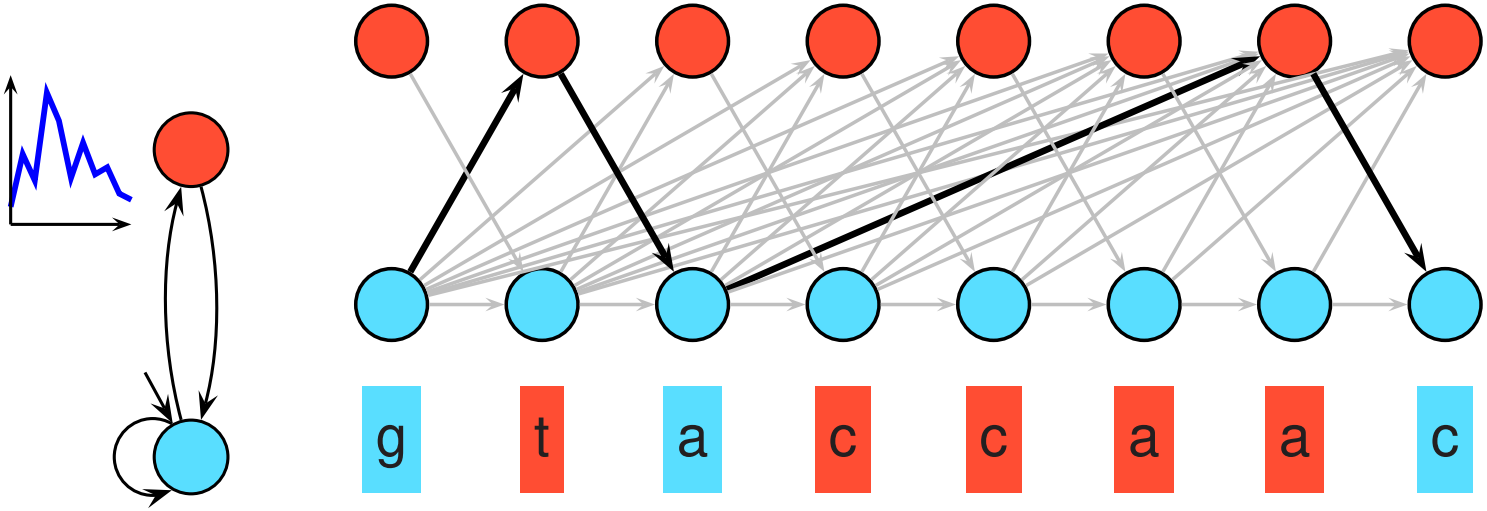
Generalized Viterbi algorithm [Rabiner, 1989]



Viterbi algorithm – $O(n)$ time

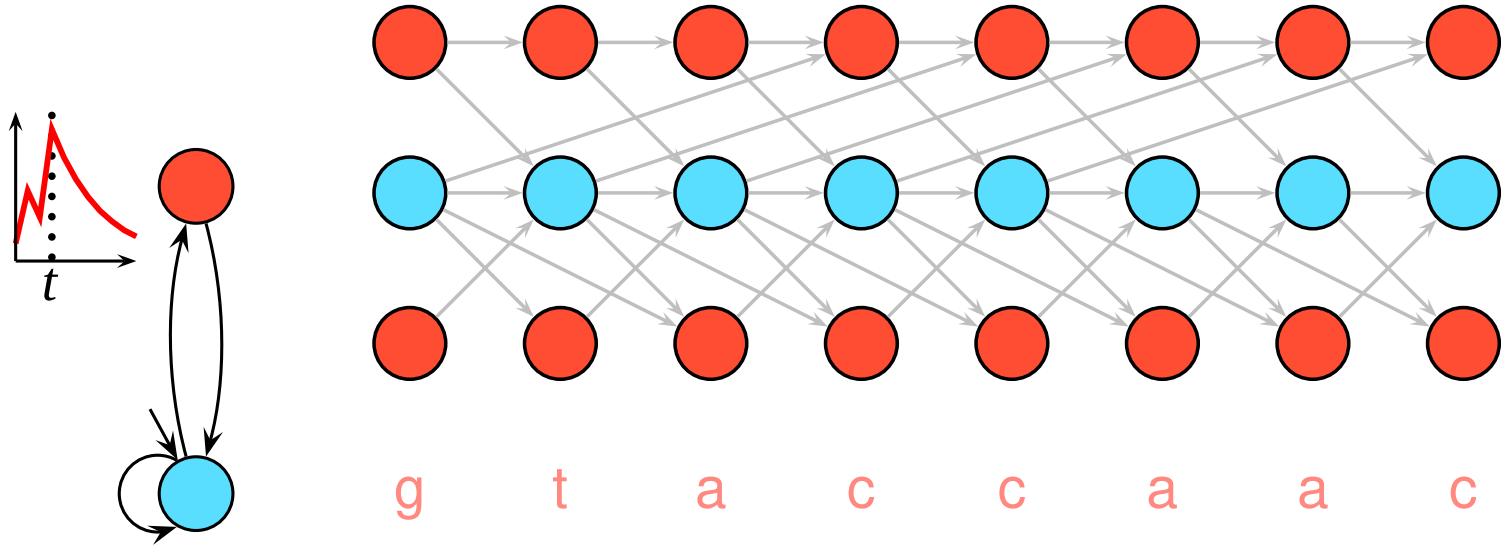


Generalized Viterbi algorithm – $O(n^2)$ time



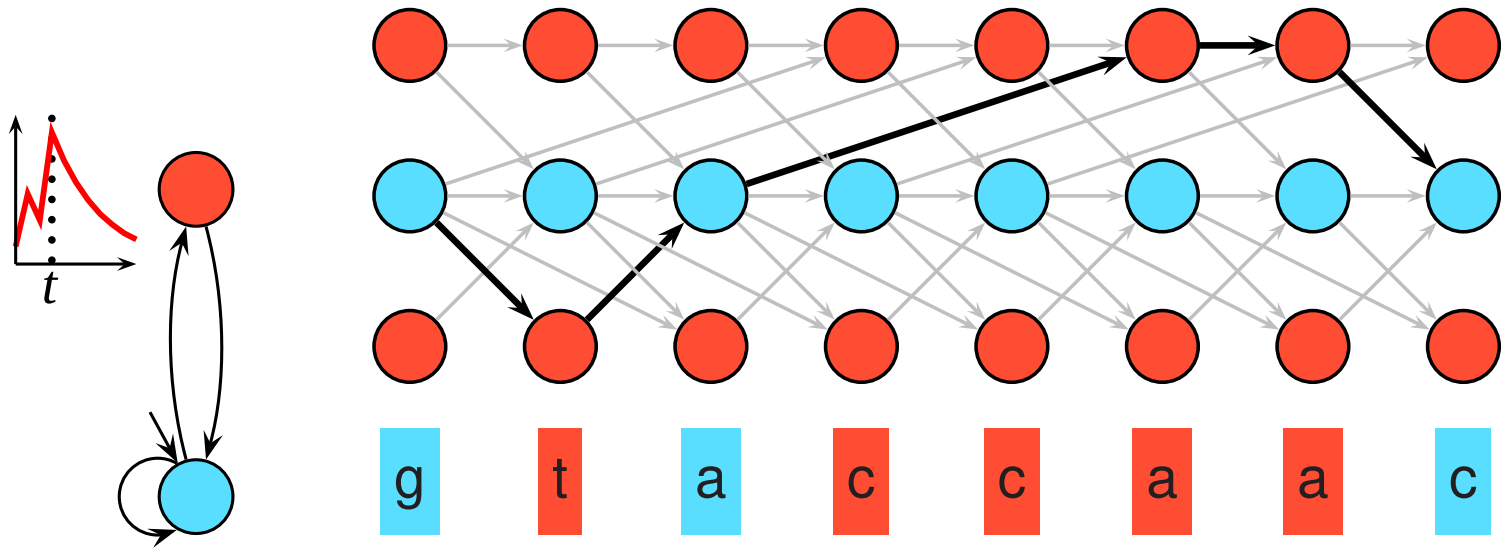
Combining two algorithms

Assumption: Length distribution – geometric tail starting at $t = 3$



Combining two algorithms

Assumption: Length distribution – geometric tail starting at $t = 3$



Running time: $O(nt)$

Modeling length distributions – summary

- Change from $O(n)$ to $O(n^2)$ – any length distribution you want
- Instead: trade-off between model faithfulness and running time
  - Approximate by geometric tail: $O(nt)$ time
  - If t is too large: $O(n\sqrt{t})$ time