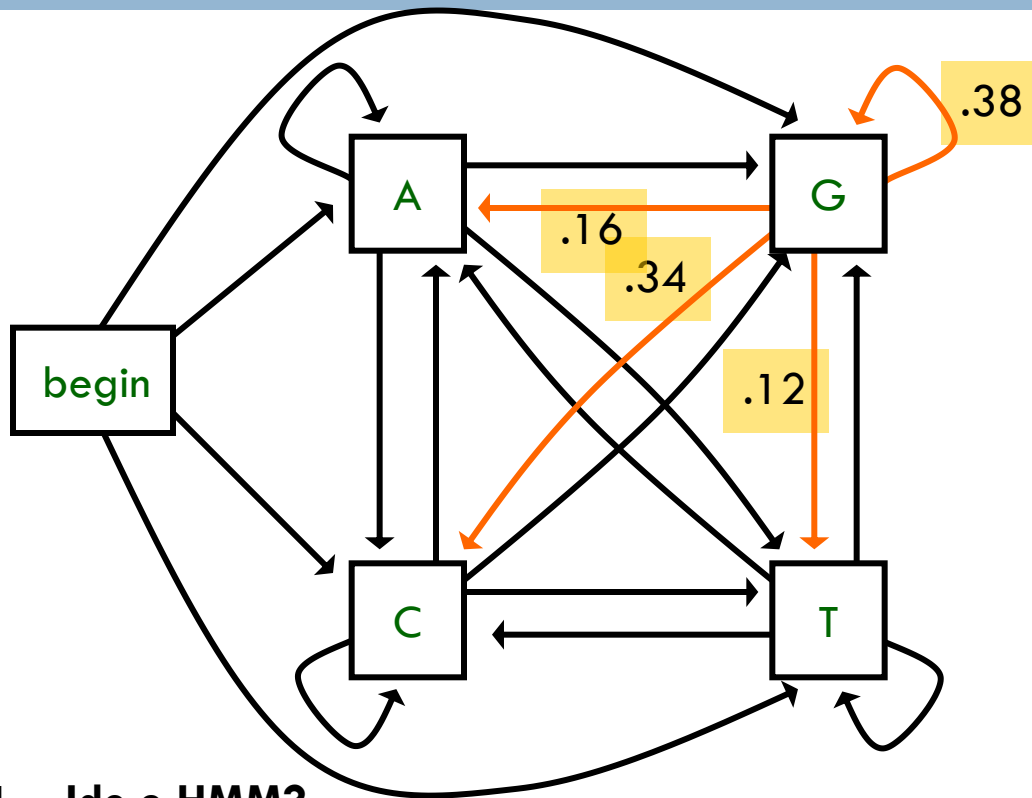


BIOINFORMATIKA – CV. 6

Markovské modely

(Some slides are courtesy of Mark Craven, U. of Wisconsin)

Příklad 1 - jednoduchý MM (1)



transition probabilities

$$P(x_i = a | x_{i-1} = g) = 0.16$$

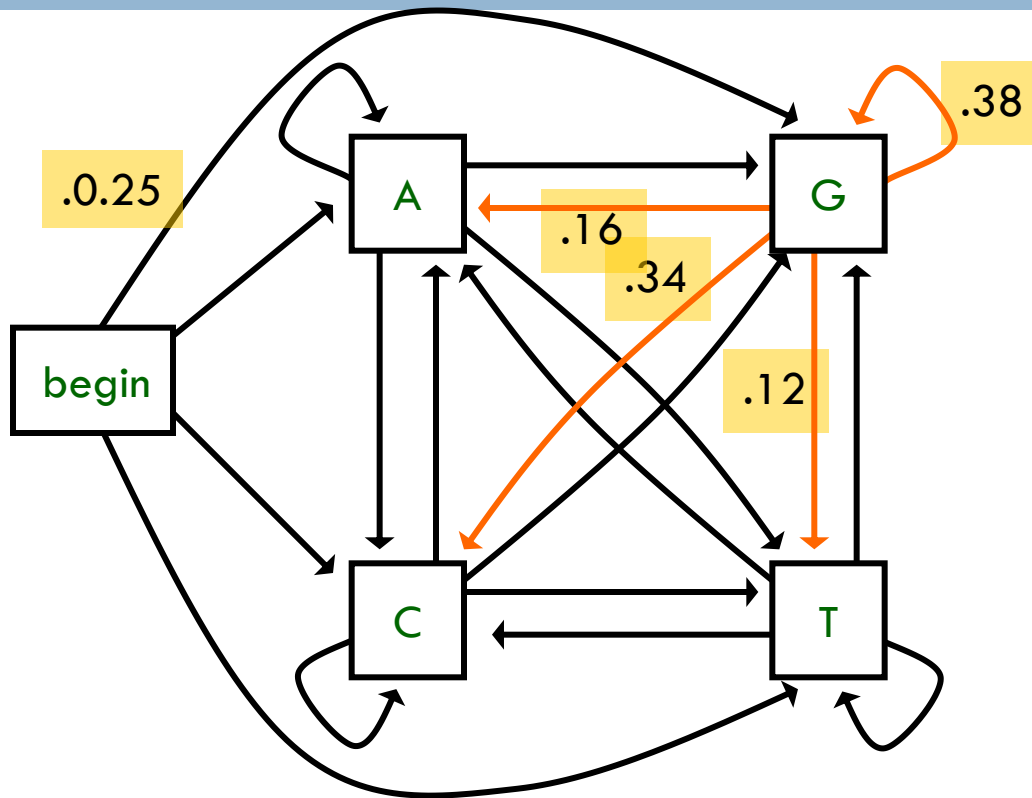
$$P(x_i = c | x_{i-1} = g) = 0.34$$

$$P(x_i = g | x_{i-1} = g) = 0.38$$

$$P(x_i = t | x_{i-1} = g) = 0.12$$

1. Jde o HMM?
2. Vyjádřete pravděpodobnost obecné sekvence symbolů X_1, \dots, X_k
3. Jde o model pro sekvence pevné (avšak volitelné) délky, nebo pro sekvence variabilní délky?
4. Jaké parametry vám chybí pro výpočet pravděpodobnosti sekvence **GGGA**?

Příklad 1 - jednoduchý MM (2)



transition probabilities

$$P(x_i = a \mid x_{i-1} = g) = 0.16$$

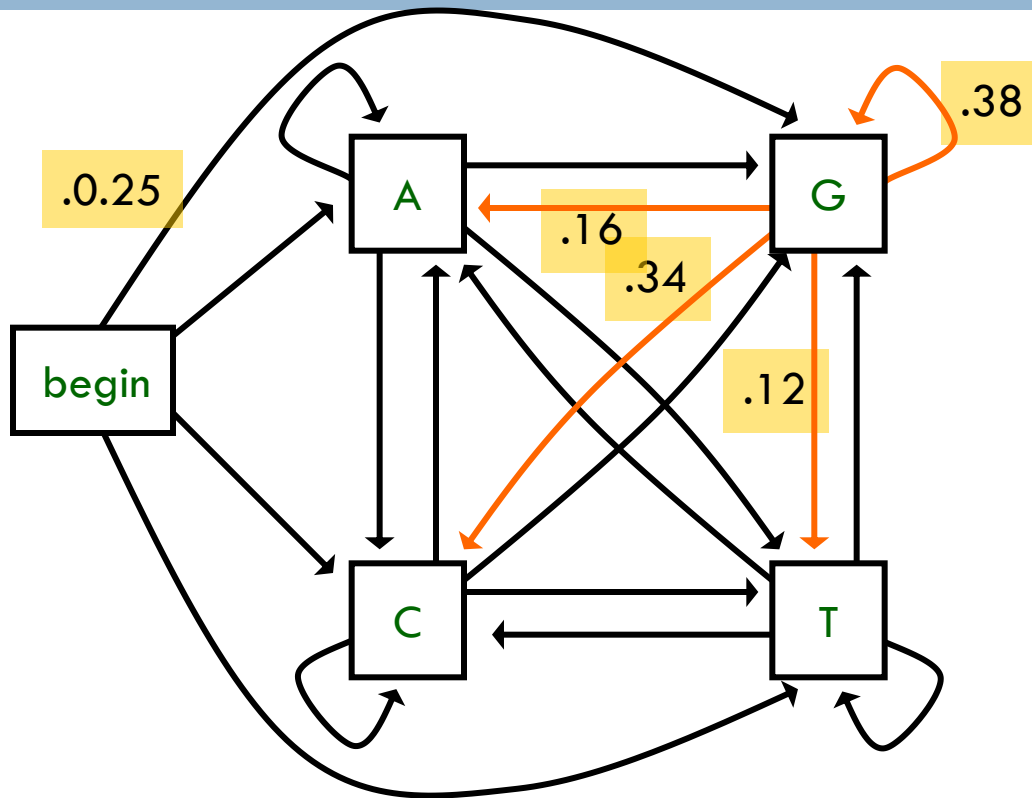
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$

$$P(x_i = g \mid x_{i-1} = g) = 0.38$$

$$P(x_i = t \mid x_{i-1} = g) = 0.12$$

5. Jaká je pravděpodobnost sekvence GGGA?

Příklad 1 - jednoduchý MM (2)



transition probabilities

$$P(x_i = a \mid x_{i-1} = g) = 0.16$$

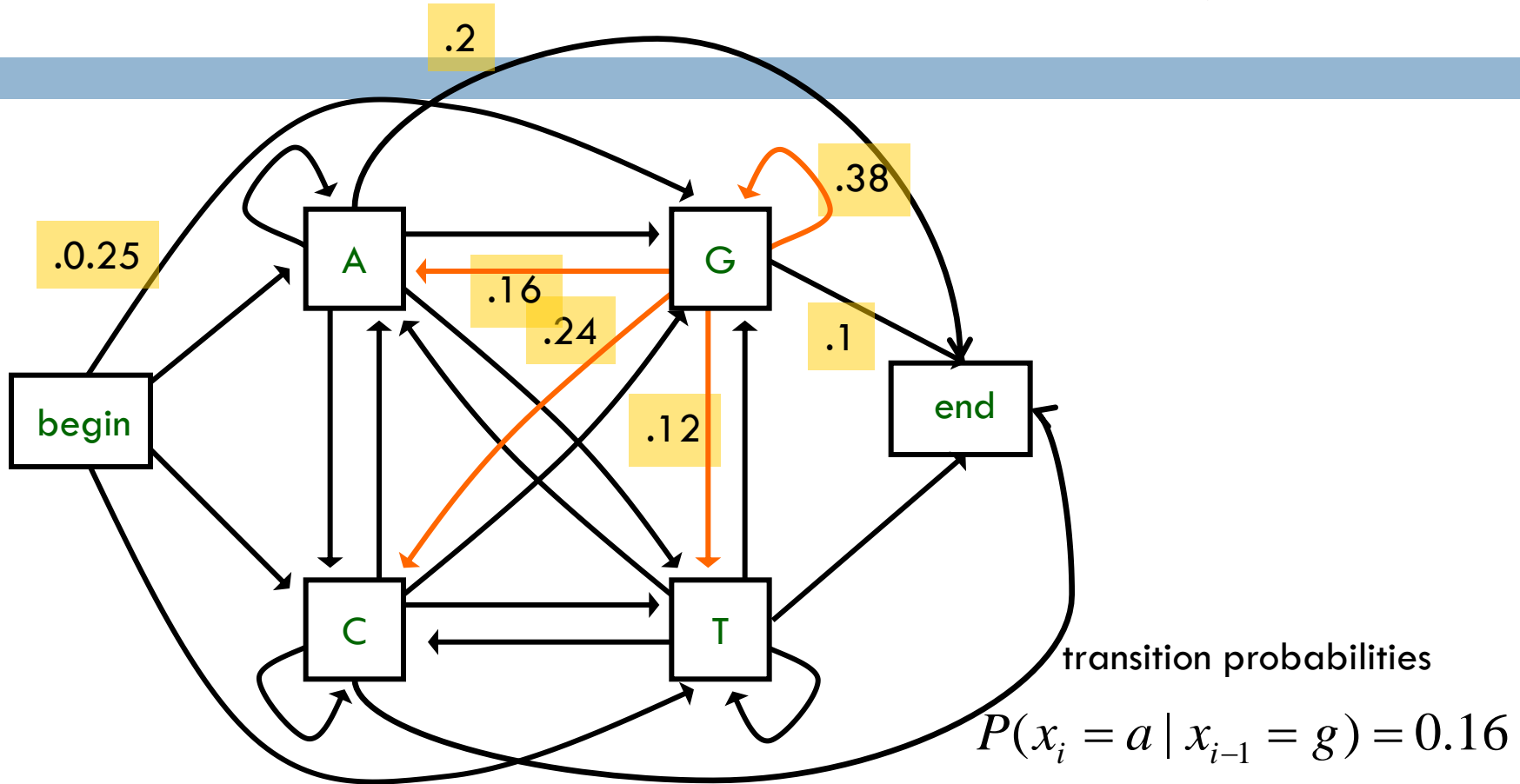
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$

$$P(x_i = g \mid x_{i-1} = g) = 0.38$$

$$P(x_i = t \mid x_{i-1} = g) = 0.12$$

6. Jak byste model upravili, aby modeloval sekvence variabilní délky?

Příklad 1 - jednoduchý MM (3)



$$P(x_i = a | x_{i-1} = g) = 0.16$$

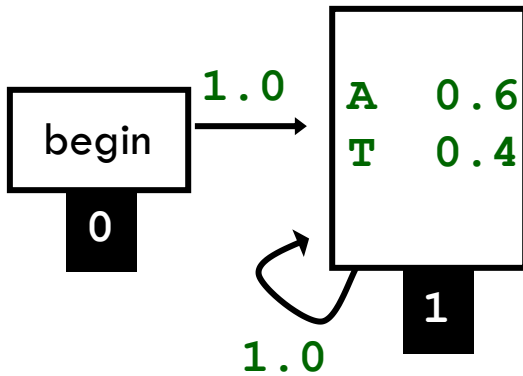
$$P(x_i = c | x_{i-1} = g) = 0.24$$

$$P(x_i = g | x_{i-1} = g) = 0.38$$

$$P(x_i = t | x_{i-1} = g) = 0.12$$

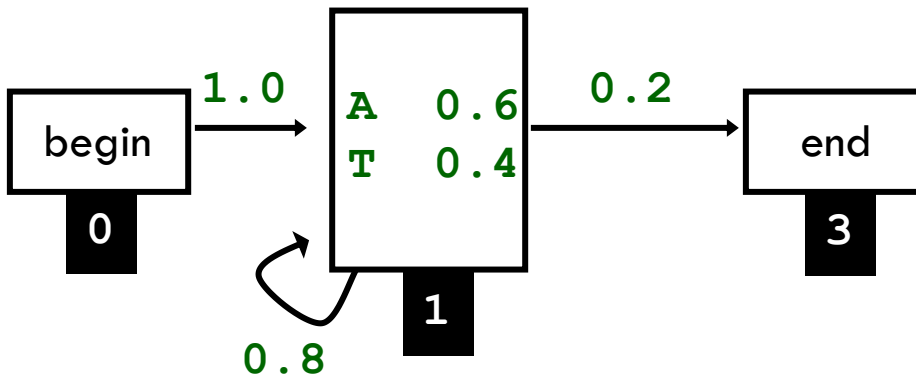
6. Jaká je pravděpodobnost sekvence **GGGA** teď?
(Upravili jsme model pro sekvence variabilní délky.)

Proč potřebujeme stav end?



$$P(A) = 0.6$$
$$P(T) = 0.4$$

$$P(AA) = 0.36$$
$$P(AT) = 0.24$$
$$P(TA) = 0.24$$
$$P(TT) = 0.16$$



$$P(A) = 0.12$$
$$P(T) = 0.08$$

$$P(AA) = 0.0576$$
$$P(AT) = 0.0384$$
$$P(TA) = 0.0384$$
$$P(TT) = 0.0256$$

$$P(L=1) = 0.2$$

$$P(L=2) = 0.16$$

Příklad 2- odhad parametrů MM

- Odhadněte parametry $P(a)$, $P(c)$, $P(g)$, $P(t)$ z následujících dat (spočítejte relativní frekvence):

accgcgctta

gcttagtgac

tagccgttac

$$P(a) = \frac{n_a}{\sum_i n_i}$$

Příklad 2- odhad parametrů MM (1)

- Odhadněte parametry $P(a)$, $P(c)$, $P(g)$, $P(t)$ z následujících dat (spočítejte relativní frekvence):

gccgcgcttg

gcttggtggc

tggccgttgc

$$P(a) = \frac{n_a}{\sum_i n_i}$$

Jednoduché, ale chceme opravdu dostat nulu pro $P(a)$?

$$P(a) = \frac{0}{30} = 0$$

$$P(g) = \frac{13}{30} = 0.433$$

$$P(c) = \frac{9}{30} = 0.3$$

$$P(t) = \frac{8}{30} = 0.267$$

Příklad 2- odhad parametrů MM (2)

- Odhadněte parametry $P(a)$, $P(c)$, $P(g)$, $P(t)$ z následujících dat (spočítejte relativní frekvence):

gccgcgcttg

gcttggtggc

tggccgttgc

$$P(a) = \frac{n_a + 1}{\sum_i (n_i + 1)}$$

Jednoduché, ale chceme opravdu dostat nulu pro $P(a)$?

**Řešení: Použijeme Laplaceův odhad, příp. obecně m-odhad.
Jaké budou odhadnuté parametry teď?**

Příklad 2- odhad parametrů MM (3)

- Odhadněte parametry $P(a)$, $P(c)$, $P(g)$, $P(t)$ z následujících dat (spočítejte relativní frekvence):

gccgcgcttg

gcttggtggc

tggccgttgc

$$P(a) = \frac{n_a + 1}{\sum_i (n_i + 1)}$$

Jednoduché, ale chceme opravdu dostat nulu pro $P(a)$?

**Řešení: Použijeme Laplaceův odhad, příp. obecně m-odhad.
Jaké budou odhadnuté parametry teď'?**

$$P(a) = \frac{0+1}{34}$$

$$P(g) = \frac{13+1}{34}$$

$$P(c) = \frac{9+1}{34}$$

$$P(t) = \frac{8+1}{34}$$

Příklad 2- odhad parametrů MM (4)

- Odhadněte parametry $P(a|g)$, $P(c|g)$, $P(g|g)$, $P(t|g)$ z následujících dat (použijte Laplaceův odhad):

gccgcgcttg

gcttggtggc

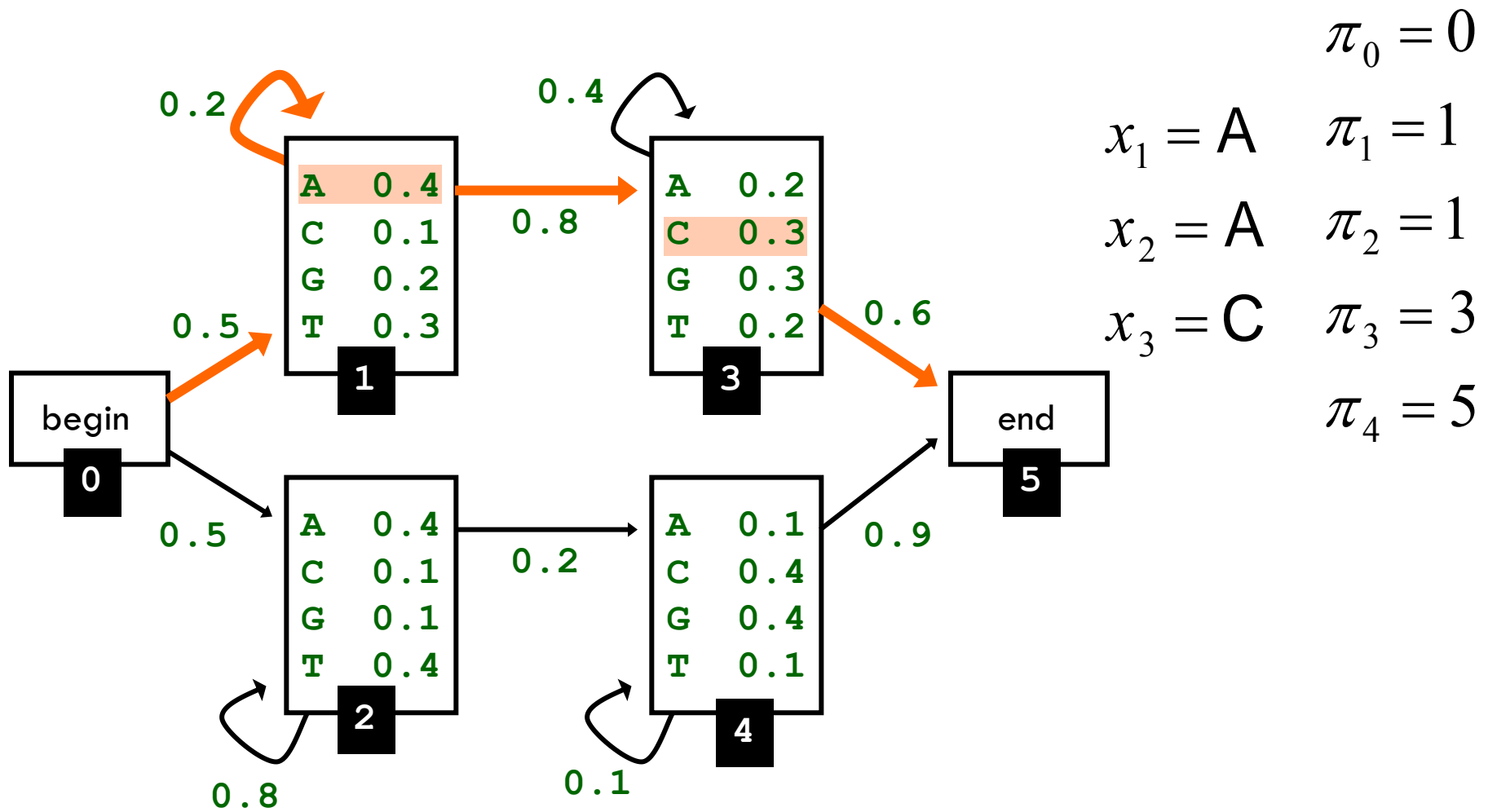
tggccgttgc

Spočítejte odhad parametrů pro případ modelu s fixní délkou sekvencí a s pevnou délkou sekvencí. Jak se odhadování parametrů pro tyto dva příklady liší?

Skryté markovské modely

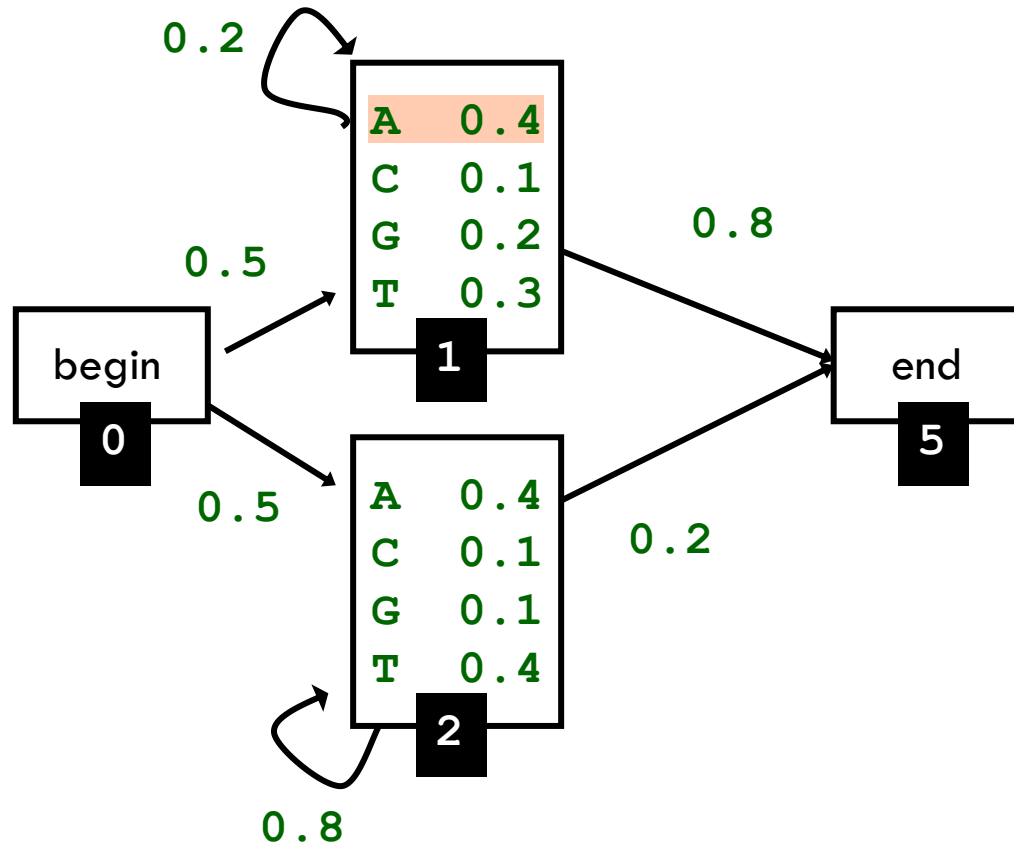
Forward algoritmus, Viterbiho algoritmus

Notace



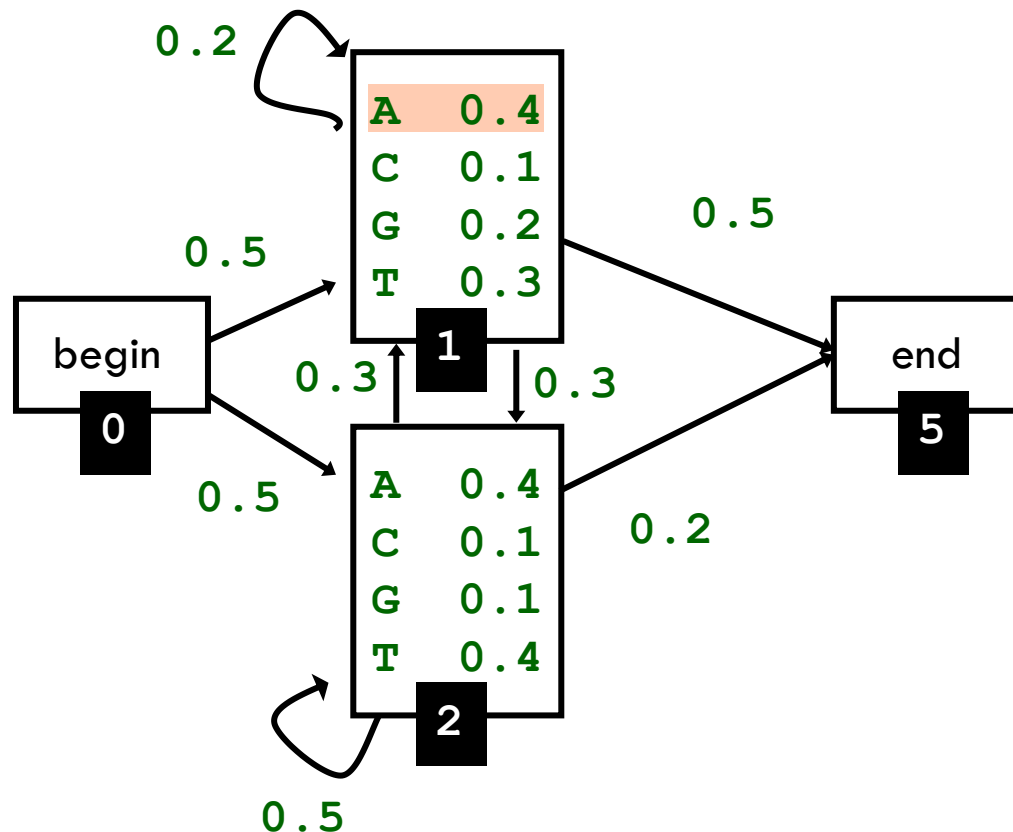
Jednoduchý příklad

- Jaká je pravděpodobnost sekvence AACCC? Vypočítejte ji pomocí výčtu všech relevantních cest (= naivní přístup).

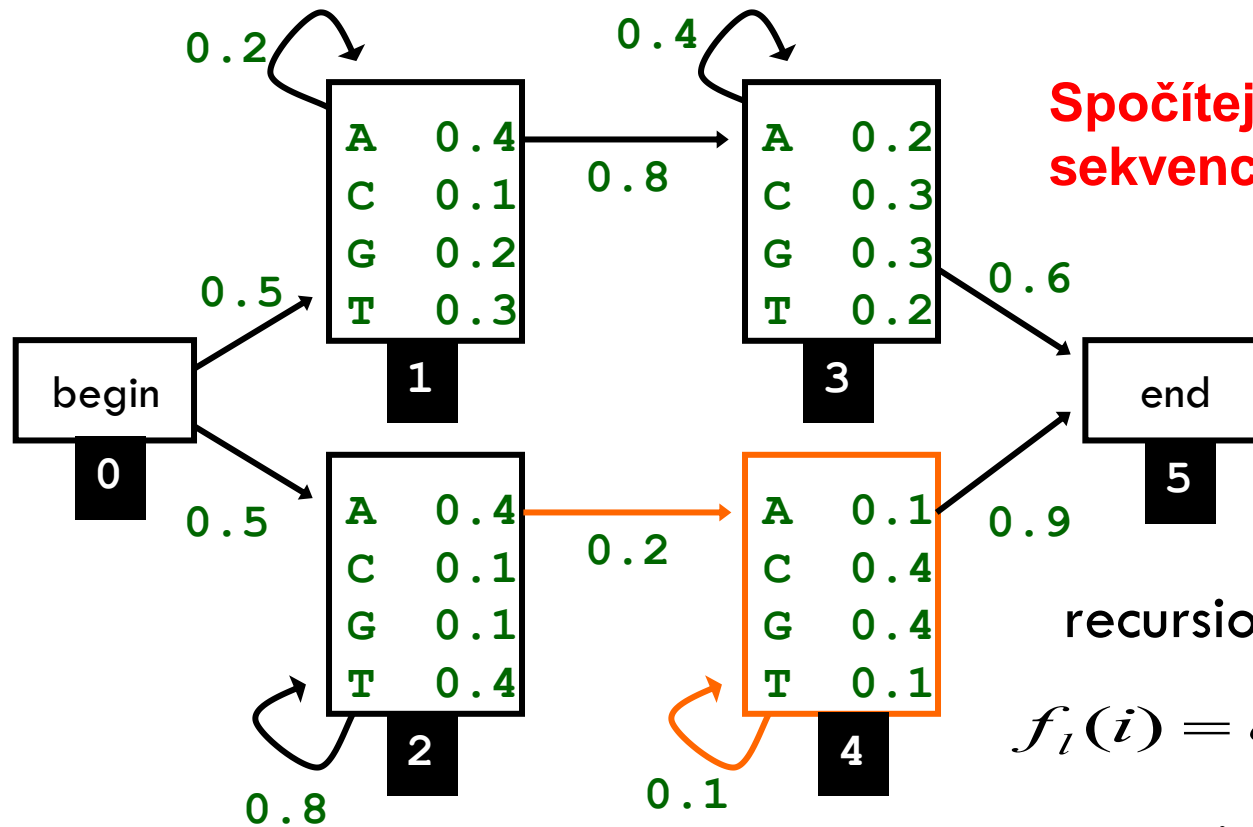


Trošku těžší příklad

- Dokázali byste spočítat pravděpodobnost sekvence AACCC naivním způsobem pro následující model? Kolik cest byste museli vzít v úvahu? Vypište je všechny – pravděpodobnost počítat nemusíte.



Ještě těžší příklad – Forward algoritmus (1)



Spočítejte pravděpodobnost sekvence TAGA

recursion for emitting states:

$$f_l(i) = e_l(i) \sum_k f_k(i-1) a_{kl}$$

recursion for silent states:

$$f_l(i) = \sum_k f_k(i) a_{kl}$$

Ještě těžší příklad – Forward algorithmus (2)

□ Sekvence: **TAGA**

□ Inicializace:

$$f_0(0) = 1 \quad f_1(0) = 0 \quad \dots \quad f_5(0) = 0$$

• Rekurzivní výpočet:

$$\begin{aligned} f_1(1) &= e_1(T) \times (f_0(0)a_{01} + f_1(0)a_{11}) = \\ &= 0.3 \times (1 \times 0.5 + 0 \times 0.2) = 0.15 \end{aligned}$$

$$f_2(1) = 0.4 \times (1 \times 0.5 + 0 \times 0.8)$$

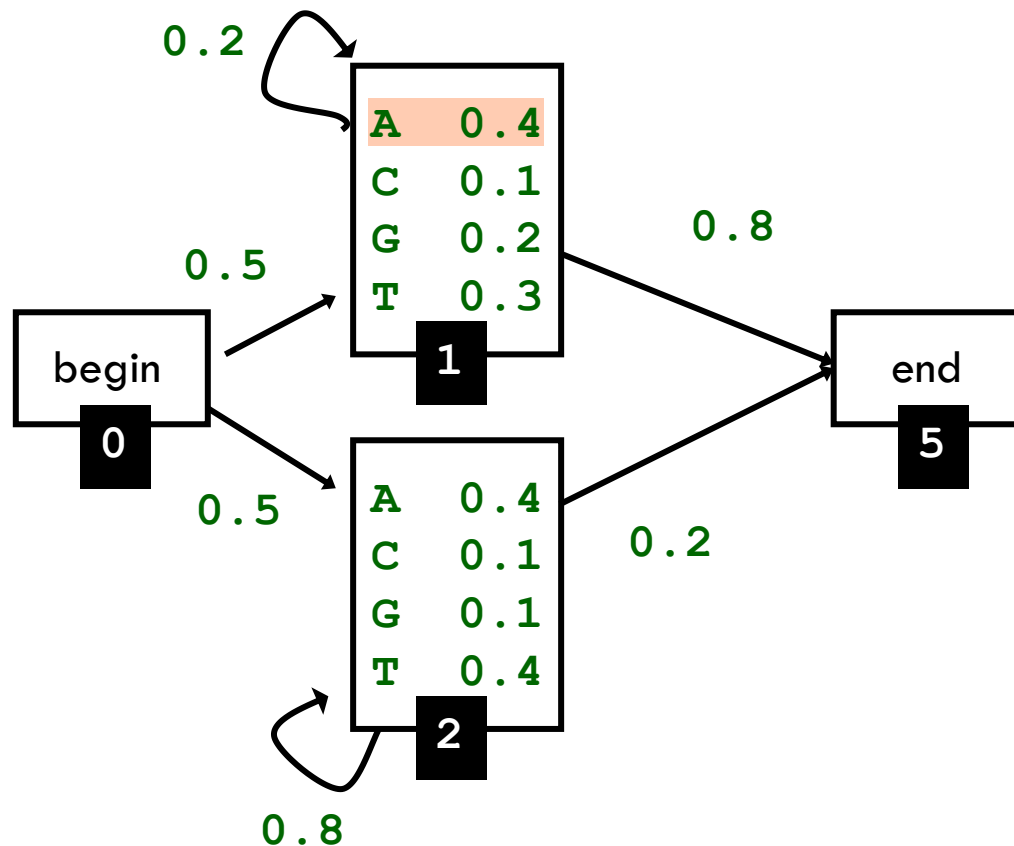
$$\begin{aligned} f_1(2) &= e_1(A) \times (f_0(1)a_{01} + f_1(1)a_{11}) = \\ &= 0.4 \times (0 \times 0.5 + 0.15 \times 0.2) \end{aligned}$$

•••

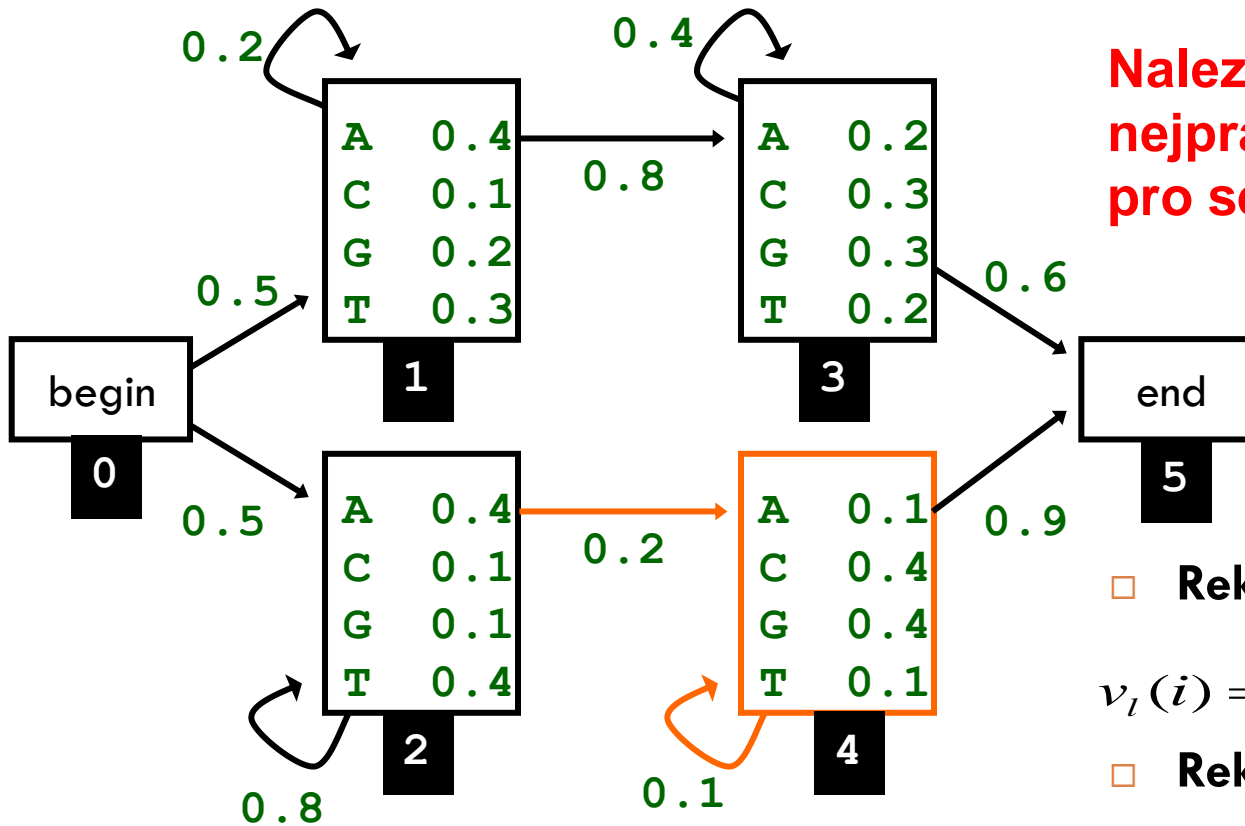
$$P(TAGA) = f_5(4) = (f_3(4)a_{35} + f_4(4)a_{45})$$

Nejpravděpodobnější sled

- Jaký je nejpravděpodobnější sled pro sekvenci AACCC? Určete pomocí výčtu všech relevantních cest (= naivní přístup).



Viterbiho algoritmus



**Nalezněte
nejpravděpodobnější sled
pro sekvenci TAGA**

□ **Rekurze pro emitující stavy:**

$$v_l(i) = e_l(x_i) \max_k [v_k(i-1)a_{kl}]$$

□ **Rekurze pro tiché stavy:**

$$v_l(i) = \max_k [v_k(i)a_{kl}]$$

□ **Nezapomeňte udržovat pointery
během výpočtu.**