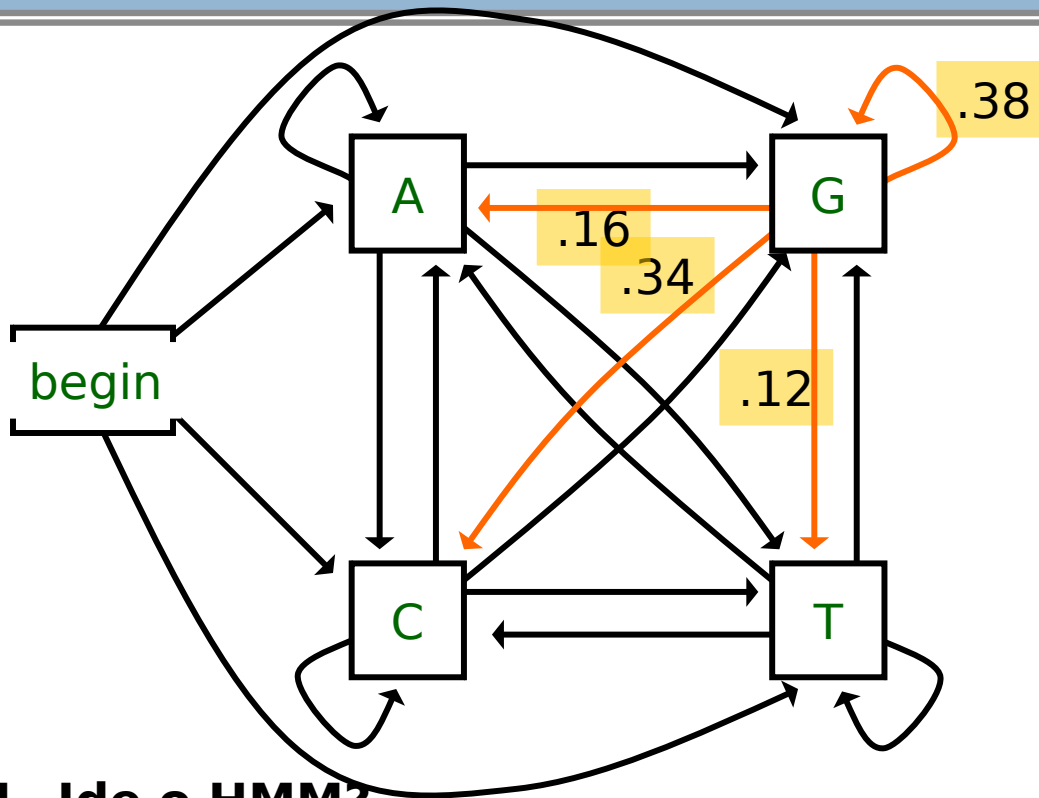


BIOINFORMATIKA – CV. 1

**Krátký úvod do markovských modelů
(Some slides are courtesy of Mark Craven, U. of Wisconsin)**

Příklad 1- jednoduchý MM (1)



transition probabilities

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

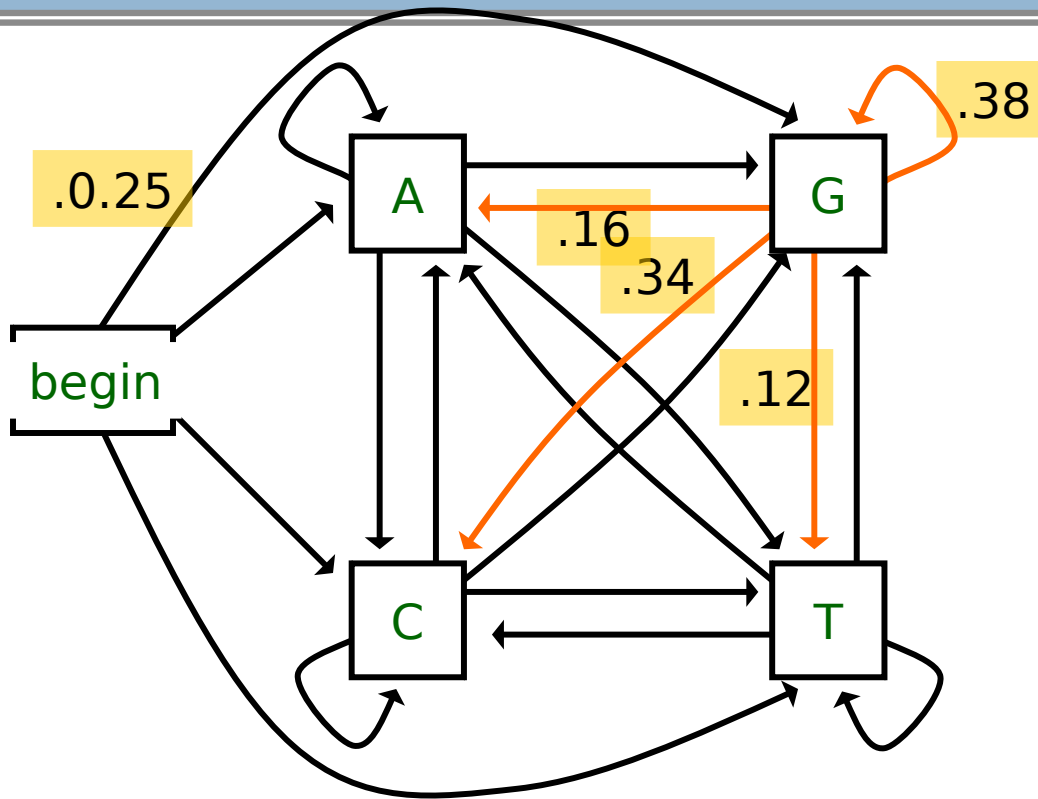
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$

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

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

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

Příklad 1- jednoduchý MM (2)



transition probabilities

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

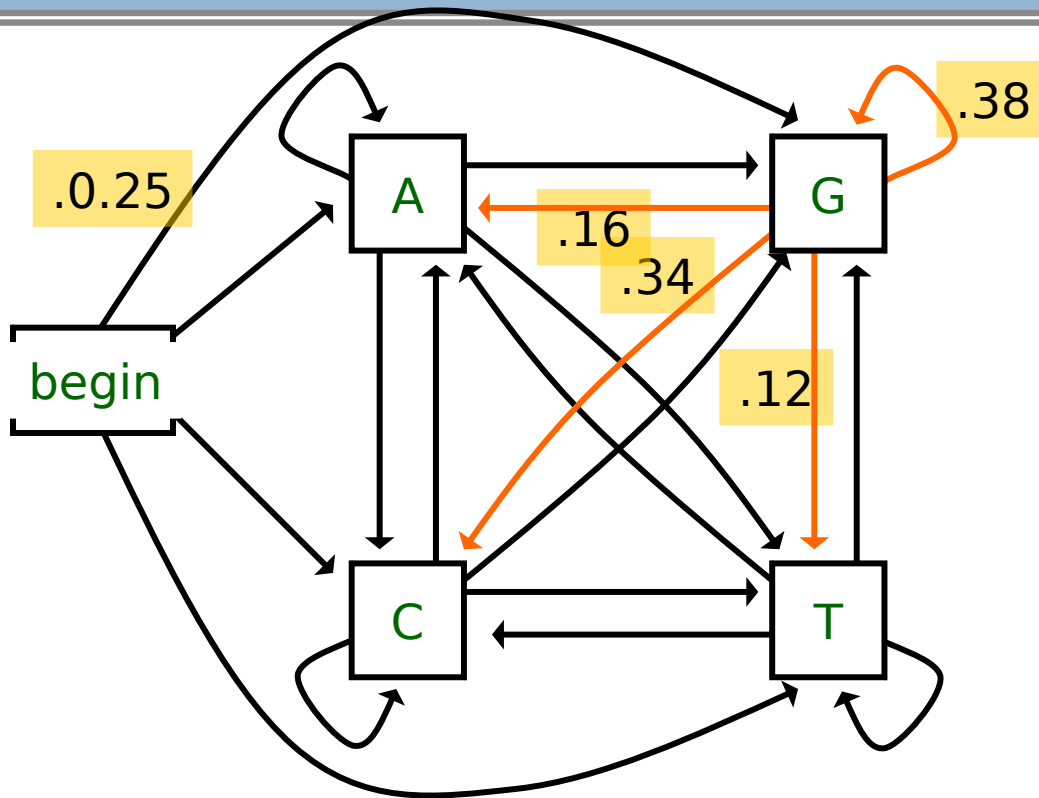
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$

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

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

5. Jaká je pravděpodobnost sekvence **GGGA**?

Příklad 1- jednoduchý MM (2)



transition probabilities

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

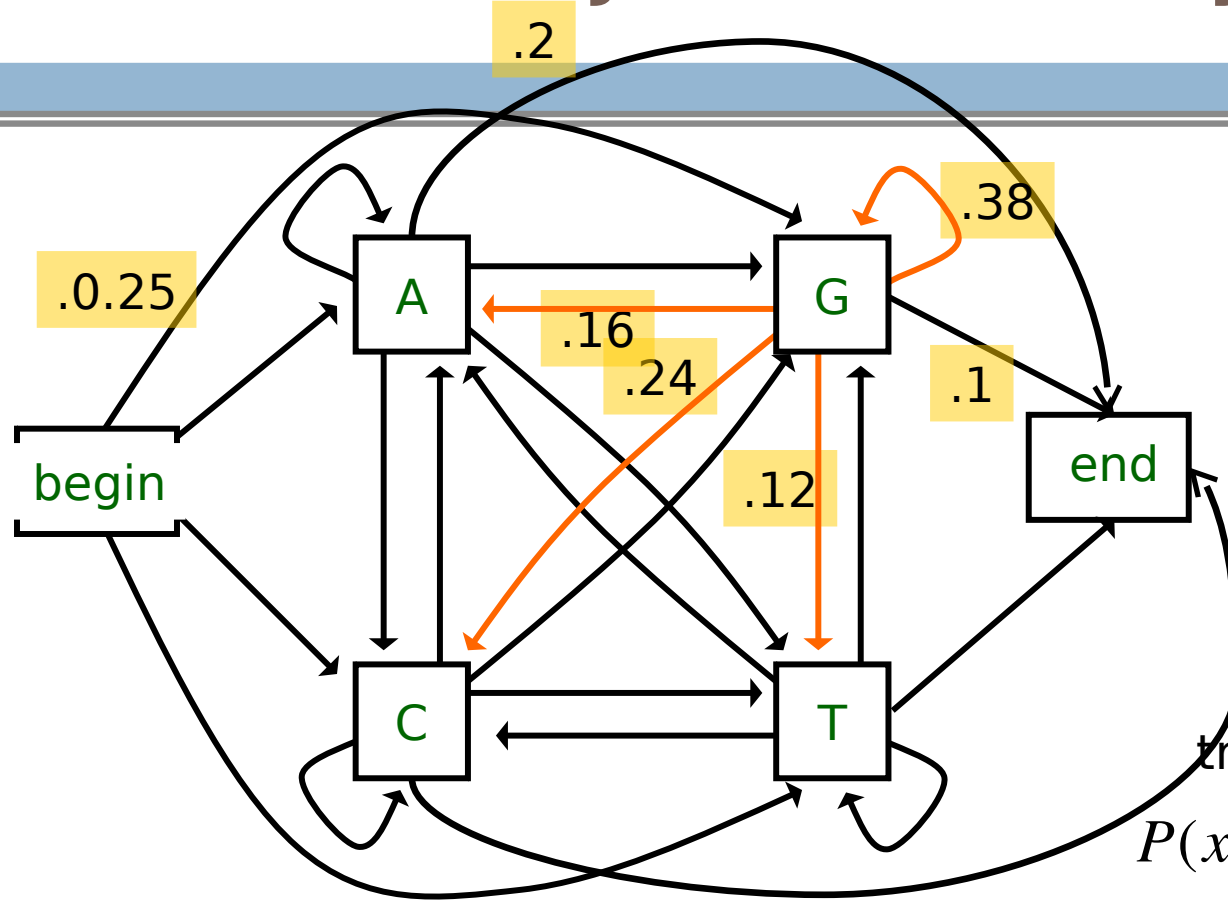
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$

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

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

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

Příklad 1- jednoduchý MM (3)



transition probabilities

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

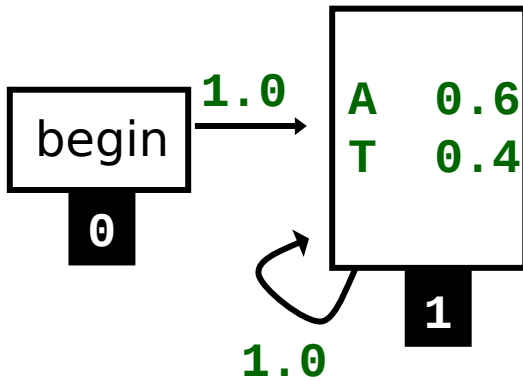
$$P(x_i = c | x_{i-1} = g) = 0.34$$

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

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

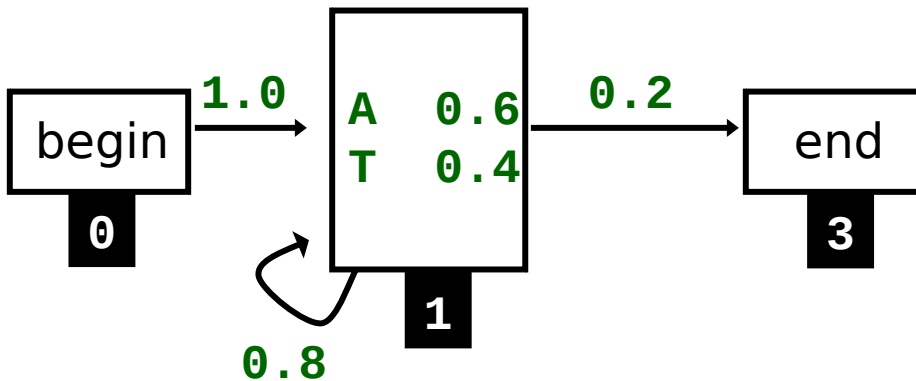
6. Jaká je pravděpodobnost sekvence GGGA teď? (Upravili jsme model pro sekvence variabilní délky.)

Proč potřebujeme stav end?



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

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



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

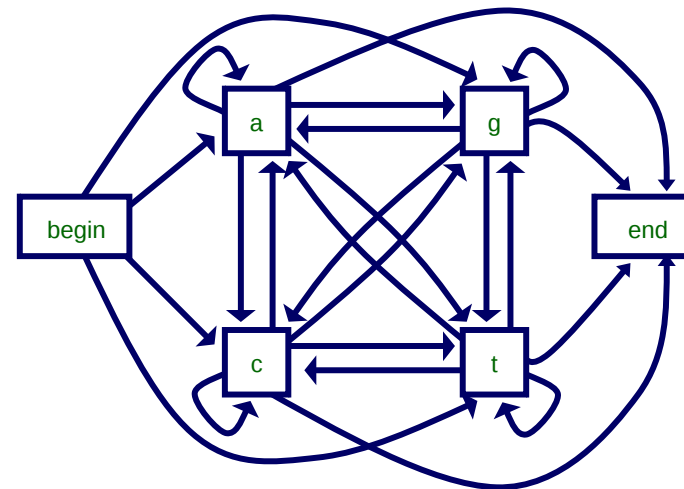
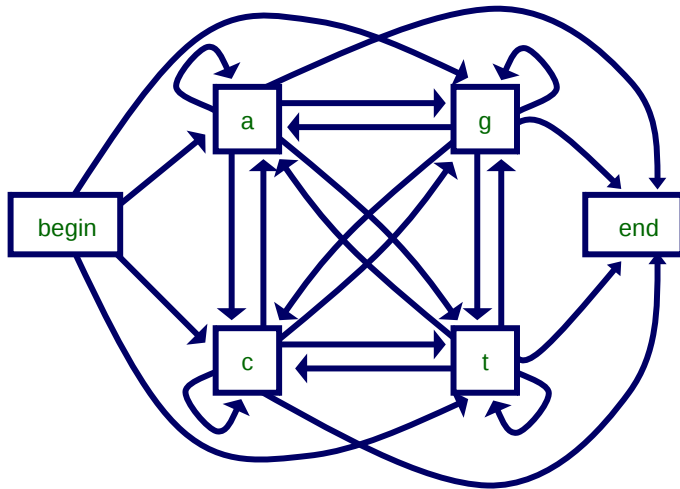
$$P(AA) = 0.0576$$
$$P(AT) = 0.0384$$
$$P(TA) = 0.0384$$
$$P(TT) = 0.0256$$

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

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

CpG islands as a classification task

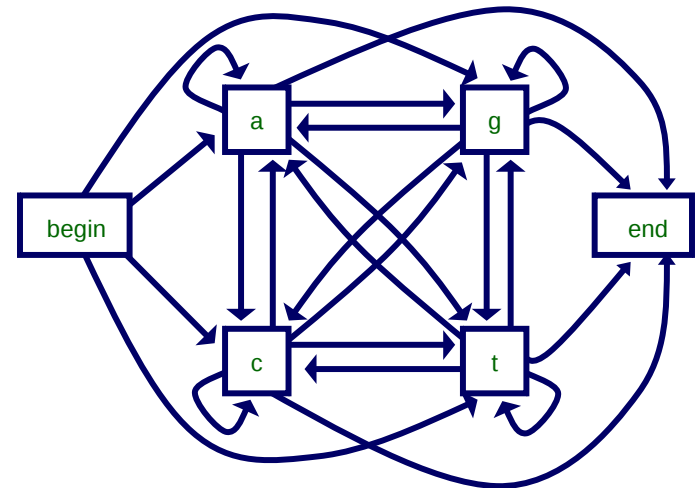
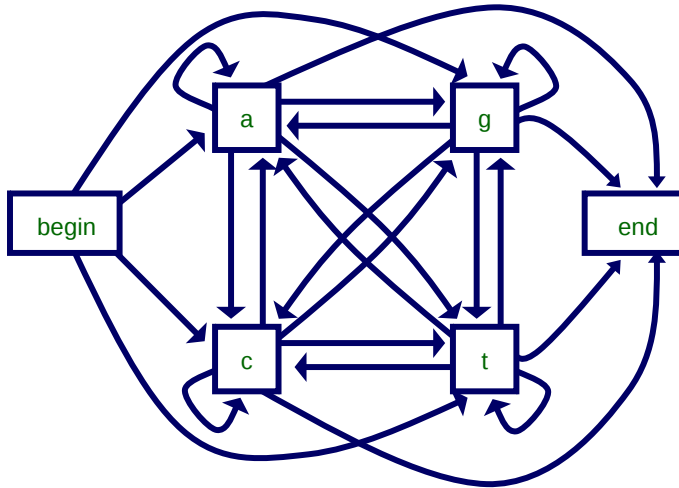
1. train two Markov models: one to represent CpG island sequence regions, another to represent other sequence regions (*null*)



2. given a test sequence, use two models to
 - determine probability that sequence is a CpG island
 - classify the sequence (*CpG* or *null*)

CpG islands as a classification task

1. train two Markov models: one to represent CpG island sequence regions, another to represent other sequence regions (*null*)



2. given a test sequence, use two models to
 - determine probability that sequence is a CpG island
 - classify the sequence (*CpG* or *null*)

Příklad 2- odhad parametrů MM

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

gccgcgcttg

gcttggtggc

tggccgttgc

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

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 -oc
které 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?

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

Příště...