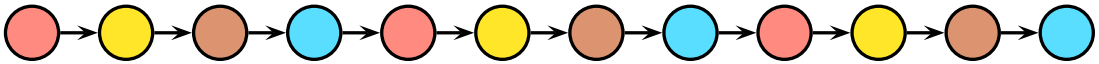
Problém najpravdepodobnejšej anotácie pomocou HMM (skrytého Markovovského modelu)

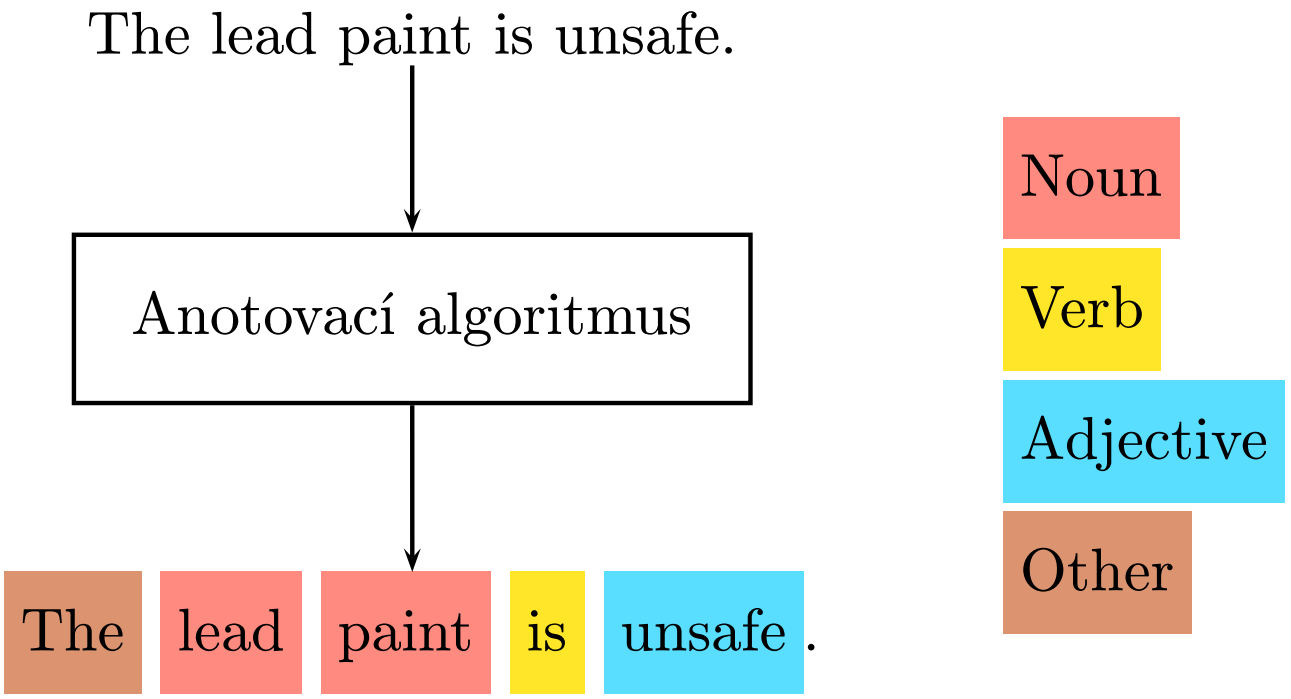


Broňa Brejová
Spoluatori Dan Brown a Tomáš Vinař

Anotácia reťazca

Ofarbi jednotlivé znaky podľa významu (závisí od kontextu)

Príklad 1: určovanie slovných druhov (tagging)



(Zdroj: Andrew McCallum)

Hľadanie génov

cggtgaaactgcacgattgttgctggcttaaagatagaccaatcagagtgtgtaacgtca

tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca

tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca

ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta

aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc

cgggcaccgtggctctgcgcgagatccgccgttatcagaagtccactgaactgcttattc

gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc

gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc

tatttgaggacactaacctgtgcgccatccacgccaagcgcgtcactatcatgcccaagg

acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac

ggtccaagcaaaggctcttttcagagccaccacctttcaagtaaagtagctgtaagaaa

ccaatttaagacaaaagggaatgcattgggagcacttttcgttttaatgctactgaaggc

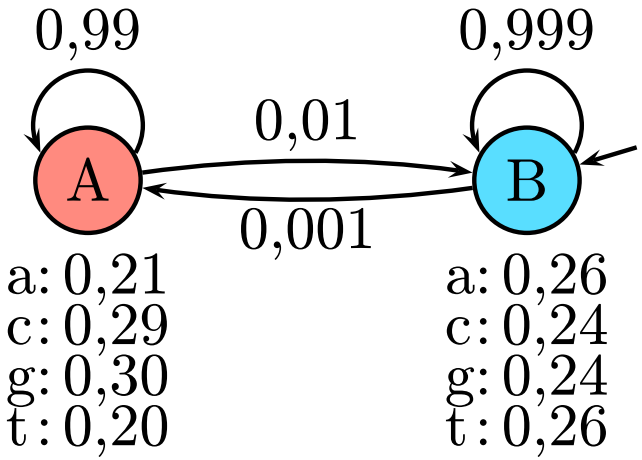
Gén | Medzigénový úsek

Skrytý Markovovský model (Hidden Markov model, HMM)

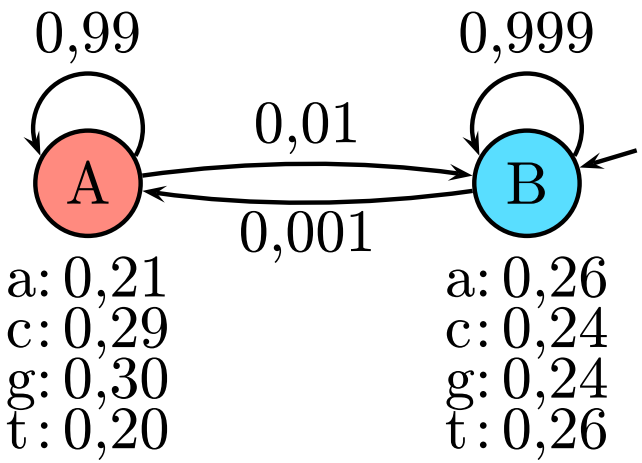
Definuje pravdepodobnosť $\Pr(X, A)$, že vygenerujeme reťazec X s anotáciou A.

"Najlepšia" anotácia reťazca X: $\arg\max_A \Pr(X, A)$

Štruktúra HMM



Štruktúra HMM



Reťazec: $\{a, c, g, t\}^*$

$$X = x_1, x_2, \ldots, x_n$$

Postupnosť stavov: $\{A, B\}^*$

$$S = s_1, s_2, \ldots, s_n$$

Anotácia: $\{$■, ■$\}^*$

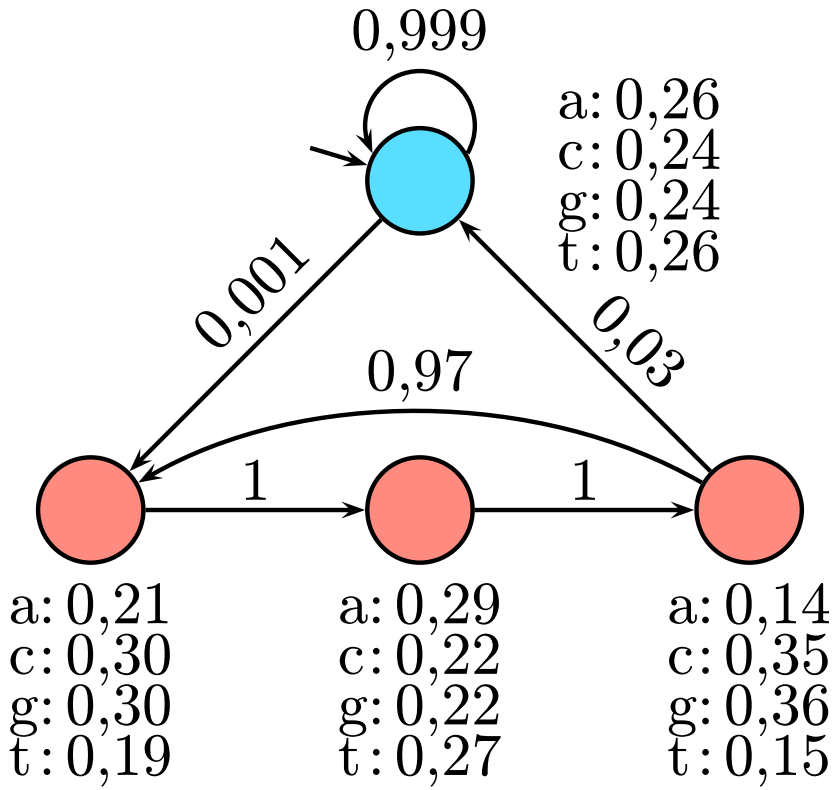
$$A = a_1, a_2, \ldots, a_n$$

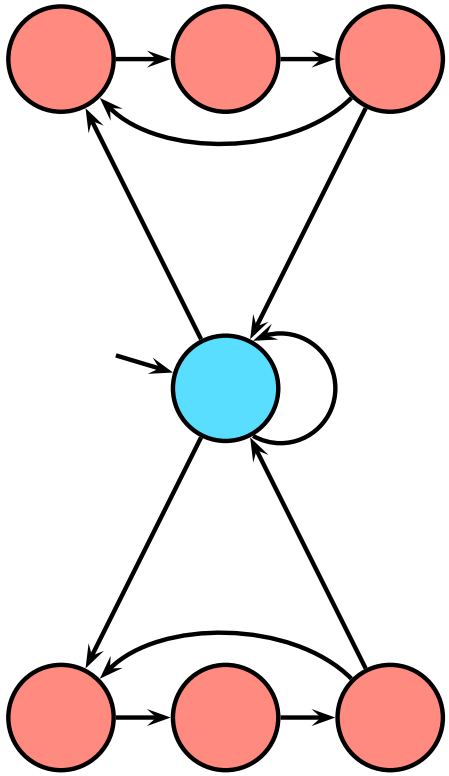
$$a_i = \text{farba } s_i$$

$$\begin{aligned} \Pr(X, S) &= \Pr(x_1|s_1)\Pr(s_1 \to s_2)\Pr(x_2|s_2)\cdots \\ &\quad \Pr(s_{n-1} \to s_n)\Pr(x_n|s_n) \end{aligned}$$

$$\Pr(X, A) = \sum_{S:S \Rightarrow A} \Pr(X, S)$$

Zložitejšie HMM





Problém najpravdepodobnejšej anotácie

Dané: HMM (graf, pravdepodobnosti, farby), reťazec X

Nájdi: $A = \arg\max_A \Pr(X, A) = \arg\max_A \sum_{S \Rightarrow A} \Pr(X, S)$

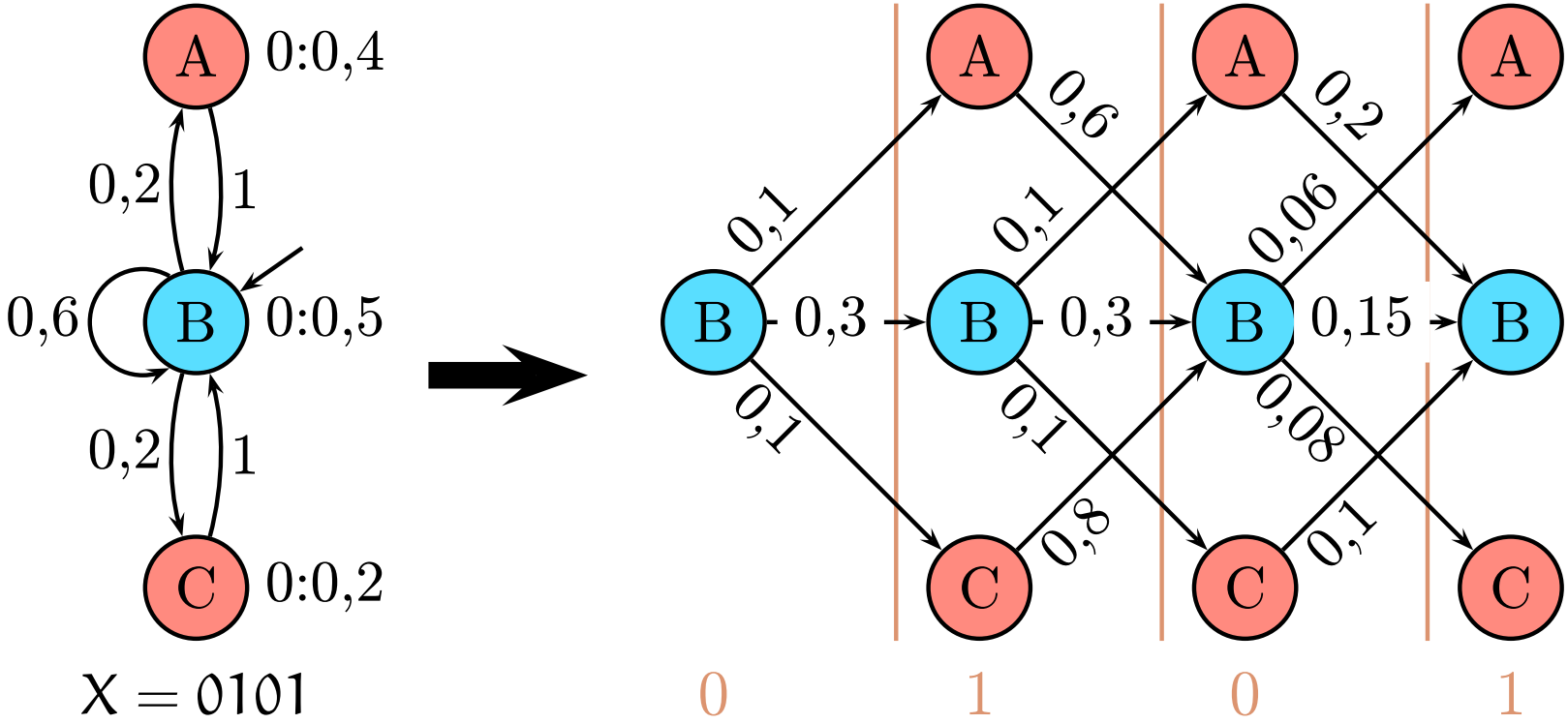
Ťažký problém:

- NP-ťažké pre vstup HMM a X [Lyngsø, Pedersen 2002]
- NP-ťažké pre jeden pevný HMM a vstup X [BBV 2004]

Pre niektoré HMM polynomiálne:

- Ak pre každé A jediné S [Viterbi 1967]
- Všeobecnejšia trieda HMM [BBV 2004]

Zjednodušme problém



S ..... cesta

Pr(X, S) ..... súčin cien hrán

A ..... ofarbenie stĺpcov

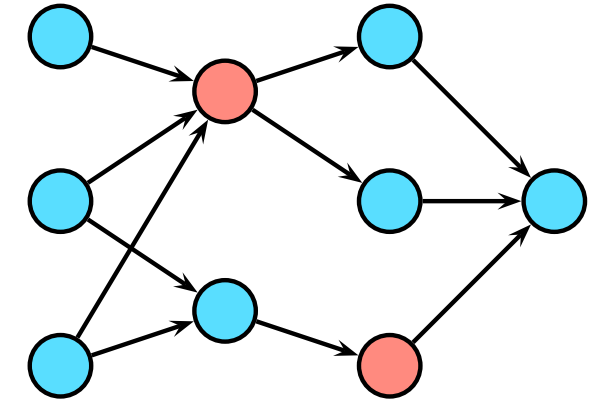
Pr(X, A) ..... súčet zodpovedajúcich ciest

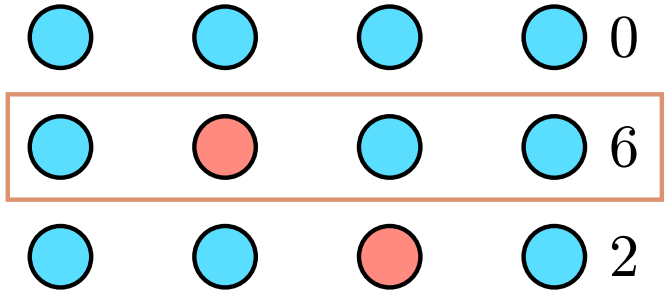
Zjednodušme problém ešte viac

. . . zrušme ceny hrán

Problém ofarbenia vrstiev

- Orientovaný graf
- Rozdelený do vrstiev s konšt. počtom vrcholov
- Vrcholy ofarbené 2 farbami
- Nájdi ofarbenie vrstiev, ktorému zodpovedá najviac ciest (resp. $\geq \mathsf{T}$ ciest)





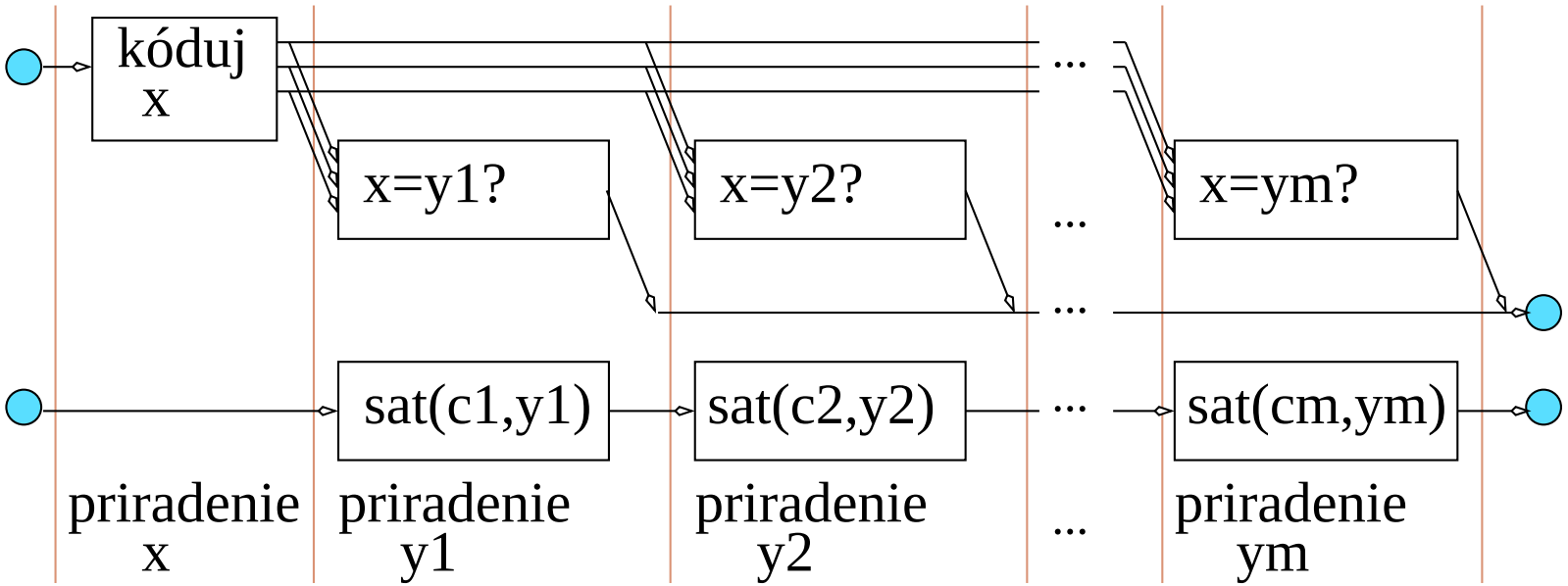
Redukcia zo SAT

Formula v CNF s m klauzulami a n premennými

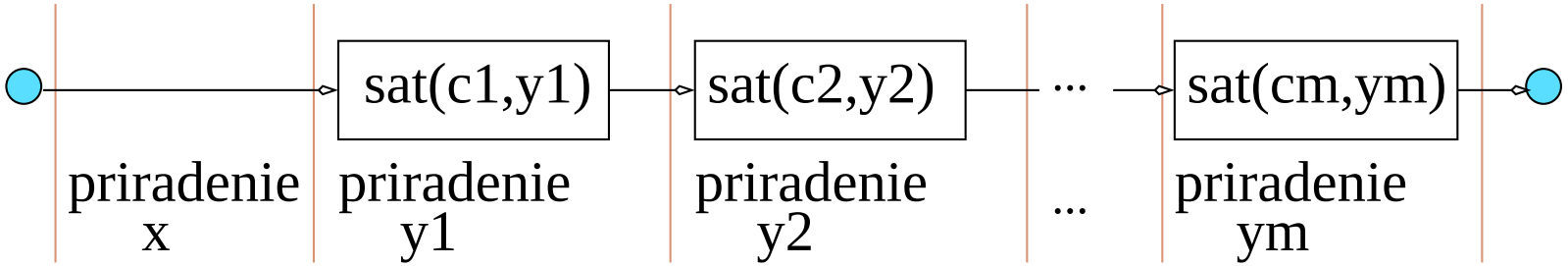
$c_1 \wedge c_2 \wedge \ldots c_m$.

Pravdivostné priradenie zodpovedá ofarbeniu vrstiev

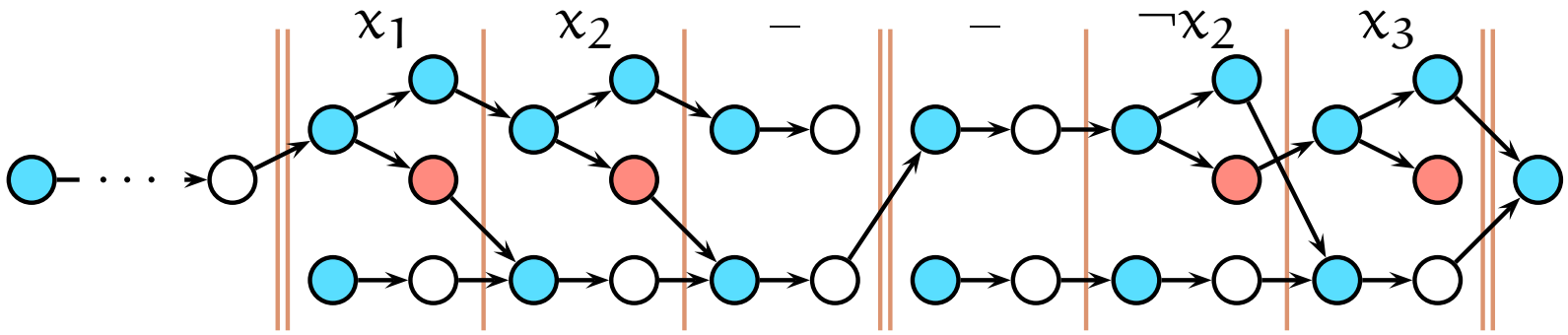
0 1



Je klauzula c_i splnená pre priradenie y_i?



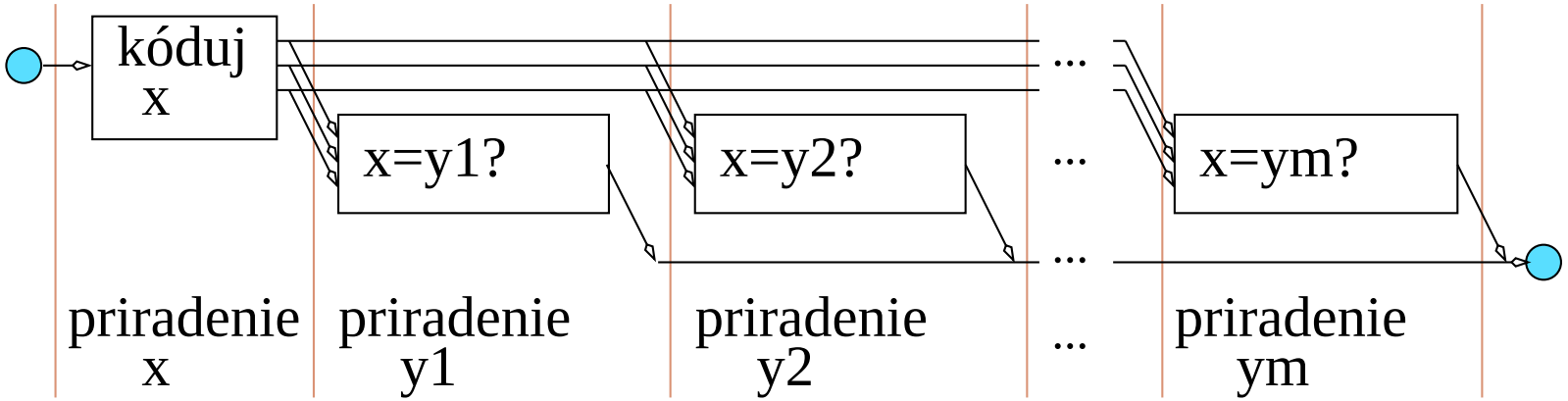
$$(x_1 \vee x_2) \wedge (\neg x_2 \vee x_3)$$

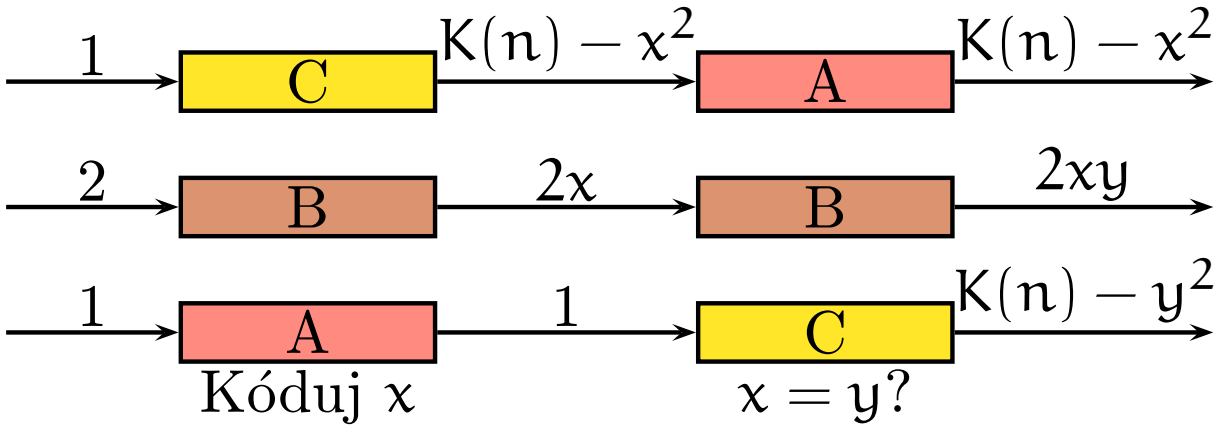


Testovanie rovnosti

Ofarbenie vrstiev = pravdivostné priradenie = binárne číslo

0 1



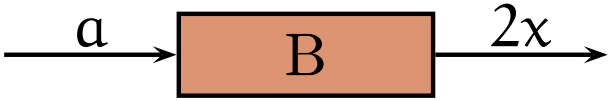


Spolu:

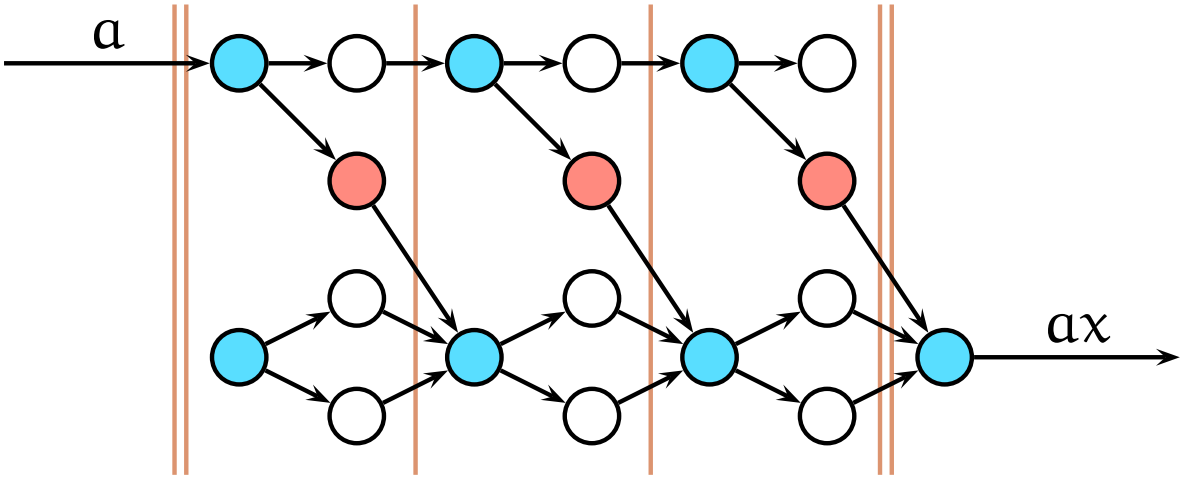
$2K(n) - (x-y)^2$

Vytvorenie čísla v binárnej sústave

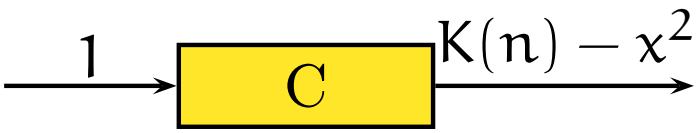
$2n$ vrstiev – n bitov: 0 1



$n = 3$:



Výpočet $K(n) - x^2$

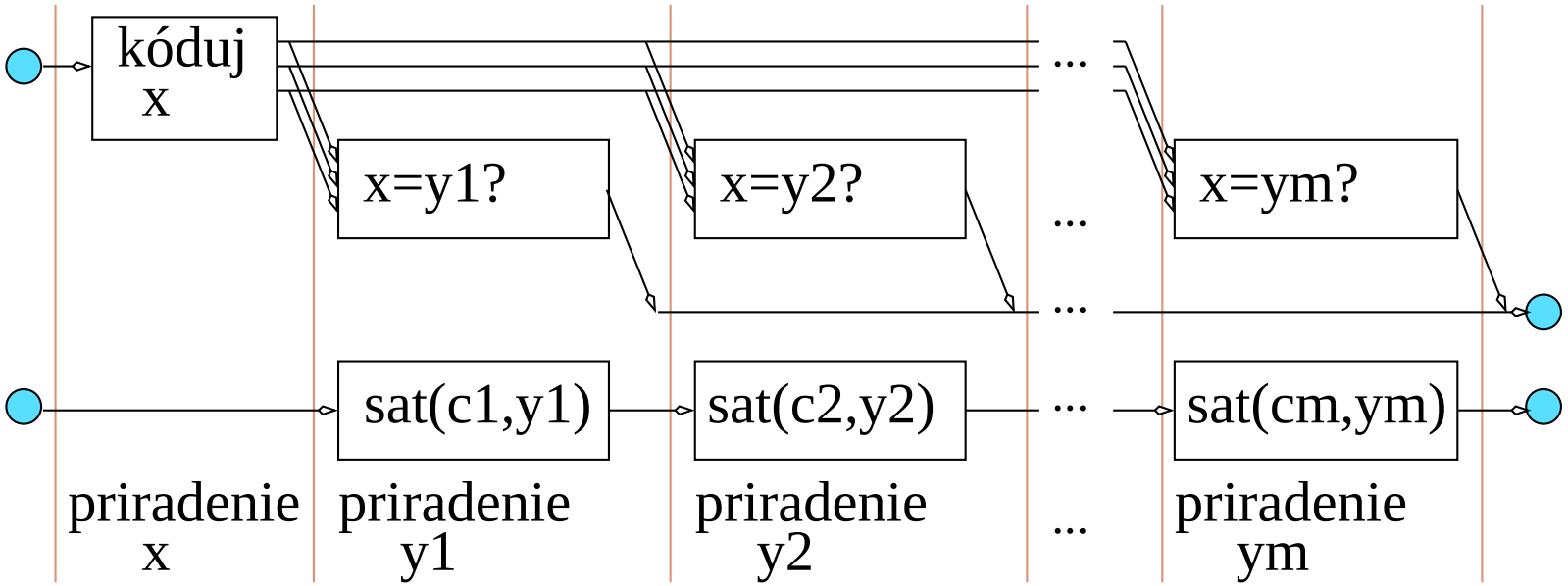


Rekurencie pre pridanie k-teho bitu

Prvých k bitov tvorí číslo $z = 2y + t$ (t je k-ty bit)

$$
\begin{aligned}
U_k = \quad & 2^{k+2} - 4 && = 2U_{k-1} + 4 \\
V_{z,k} = \quad & U_k - 4z && = \begin{cases} 2V_{y,k-1} + 4, & \text{if } t = 0 \\ 2V_{y,k-1}, & \text{if } t = 1 \end{cases} \\
W_{z,k} = & \underbrace{(2^k - 1)^2}_{K(k)} - z^2 && = \begin{cases} 4W_{y,k-1} + U_{k-1} + 1, & \text{if } t = 0 \\ 4W_{y,k-1} + V_{y,k-1}, & \text{if } t = 1 \end{cases}
\end{aligned}
$$

Problém ofarbenia vrstiev je NP-ťažký



- Ak $x = y_1 = \cdots = y_m$:
  - — ak x spĺňa formulu: $2mK(n) + 1$ ciest
  - — ak x nespĺňa formulu: $2mK(n)$ ciest
- Inak: $\leq 2mK(n)$ciest

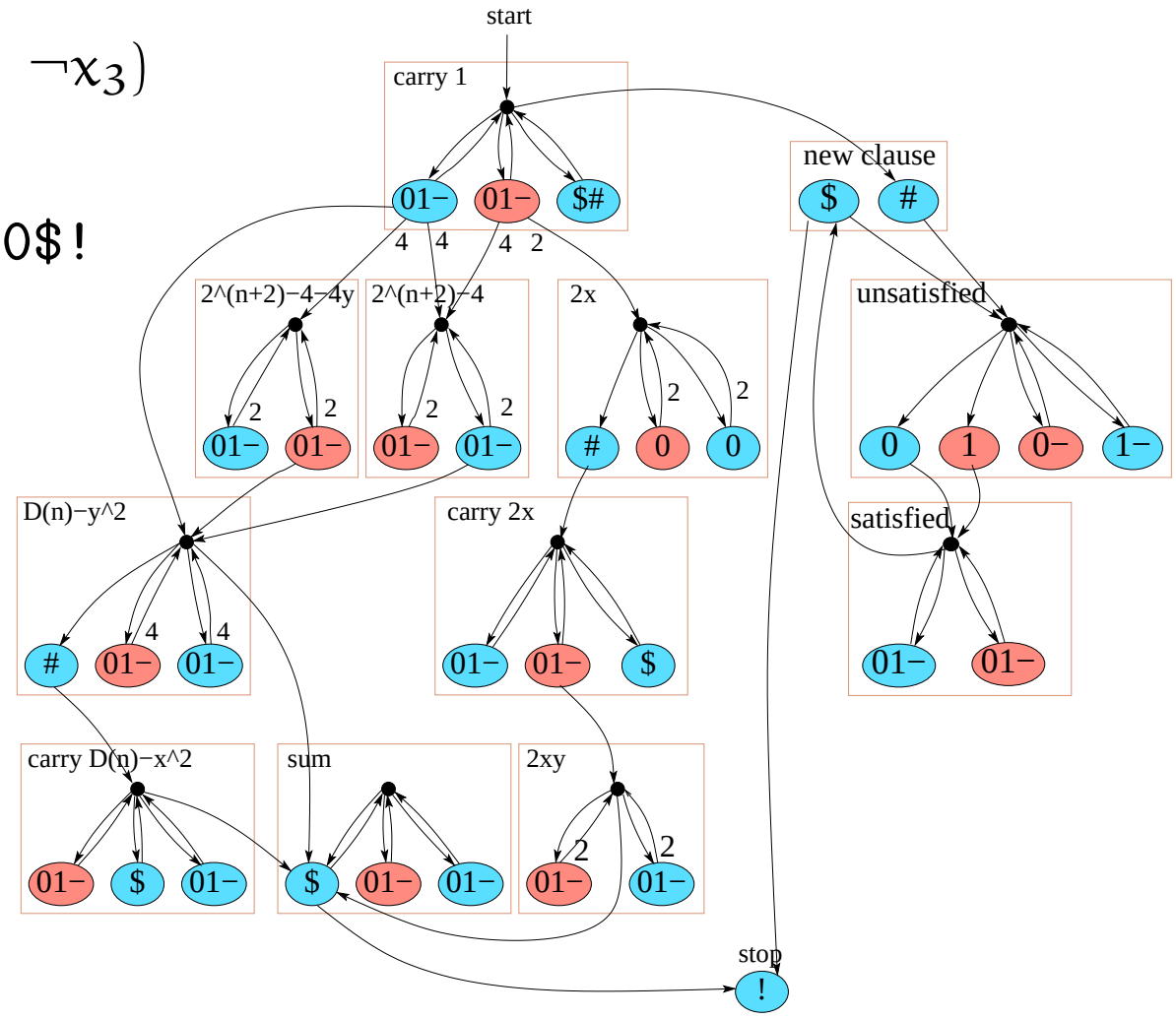
Každá vrstva má najviac 29 vrcholov

Problém najpravdepodobnejšej anotácie je NP-ťažký pre tento HMM

$x_1 \wedge (x_2 \vee \neg x_3)$

zodpovedá

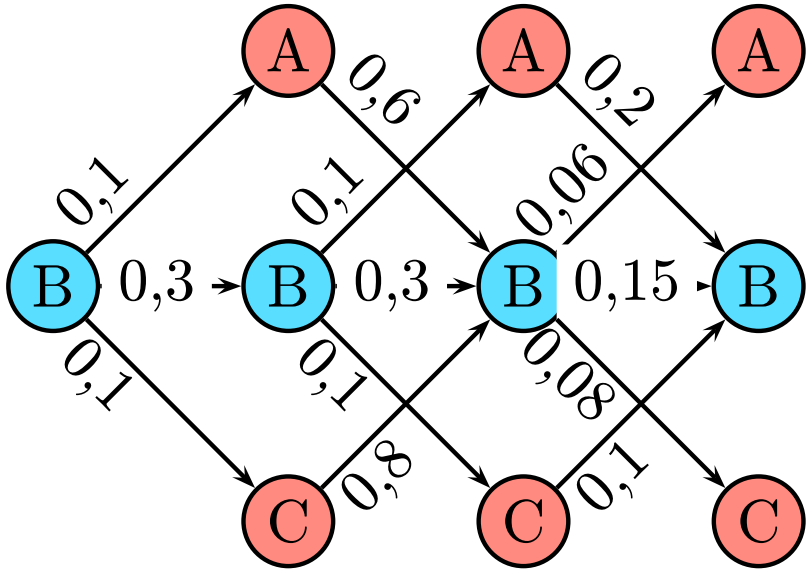
X =000#1---10!



Algoritmická časť

Efektívne algoritmy pre najpravdepodobnejšiu anotáciu v špeciálnych typoch HMM

Viterbiho algoritmus: nájde $\arg\max_S \Pr(X, S)$



Dyn. programovanie:

$D[v]$ = najlepšia cesta do v

$D[v] = \max_{u \to v} D[u] \cdot c(u, v)$

Čas $O(nm^2)$

$n = |X|$

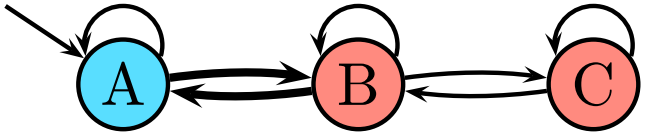
m = počet stavov

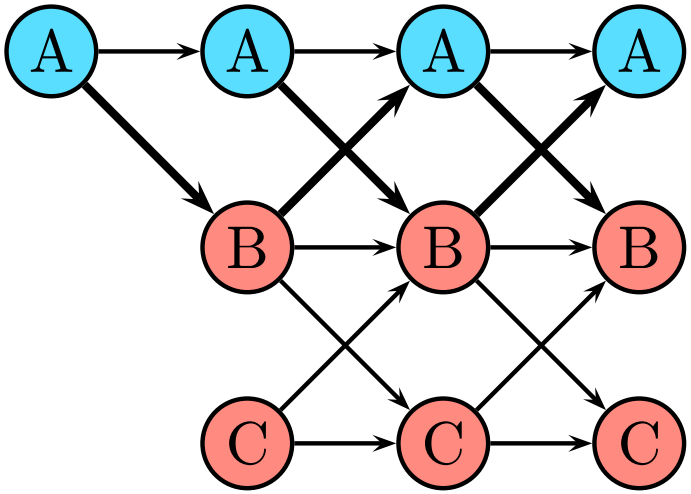
Ak každá anotácia A má iné S, nájde aj $\arg\max_A \Pr(X, A)$

Rozšírený Viterbiho algoritmus

Kritická hrana: spája dve rôzne farby

Rozšírená anotácia R: postupnosť farieb stavov a kritických hrán





Postupnosť stavov:

$S = A, B, C, B$

Rozšírená anotácia:

R = ■, ■, ■, ■, (A → B)

Anotácia:

A = ■, ■, ■, ■

Cieľ: najpravdepodobnejšia rozšírená anotácia R

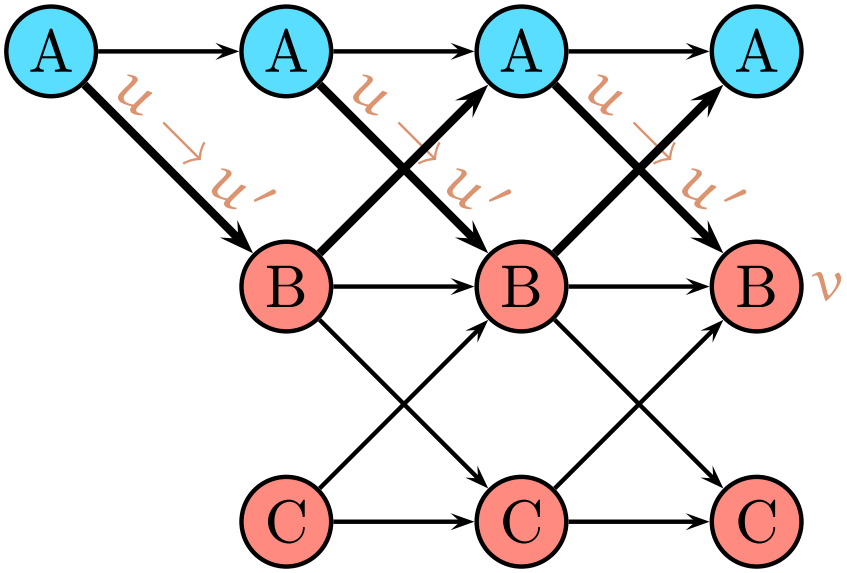
Rozšírený Viterbiho algoritmus

$D[v]$ = najlepšia rozšírená anotácia končiaca vo vrchole v

$D[v] = \max_{u \to u', f(u')=f(v), f(u) \neq f(v)} D[u] \cdot c(u, u') \cdot D'[u', v]$

$D'[s, t]$ = súčet cien jednofarebných ciest z s do t

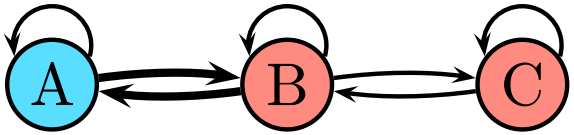
$D'[s, t] = \sum_{s \to s', f(s)=f(s')} c(s, s') \cdot D'[s', t]$



Čas $O(n^2 m^3)$

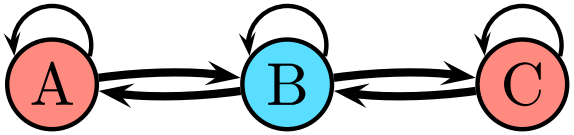
Rozšírený Viterbiho algoritmus (RVA)

Kedy nám RVA nájde najlepšiu anotáciu A?



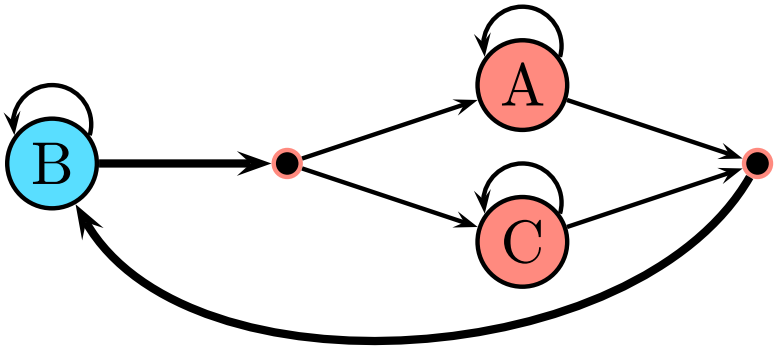
Pre každé A jediné R

$\Rightarrow$ môžeme použiť RVA.



$A =$ ■, ■, ■ má dve možné R

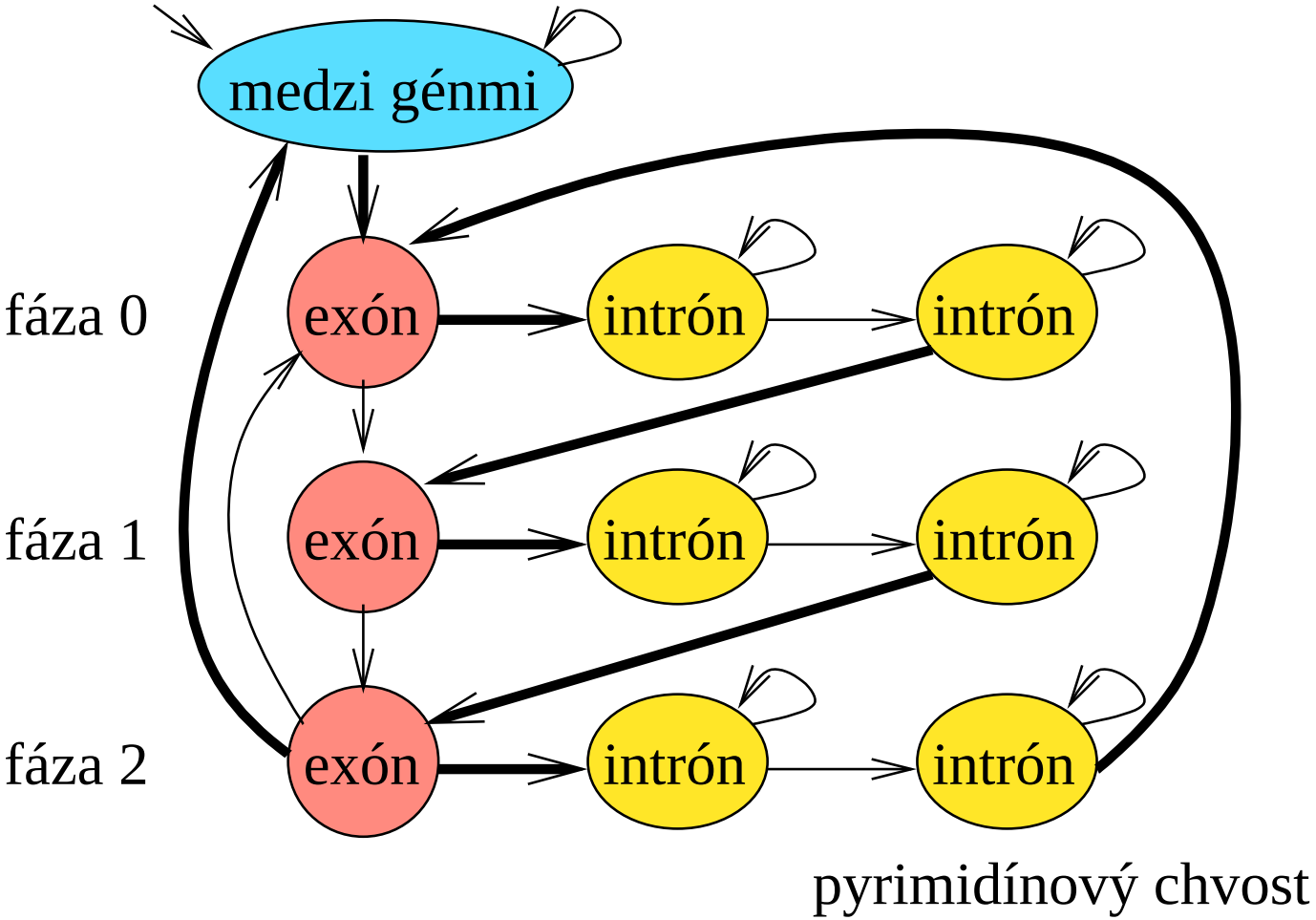
$\Rightarrow$ nemôžeme použiť RVA



Pridáme stavy, ktoré nič negenerujú

Môžeme použiť RVA

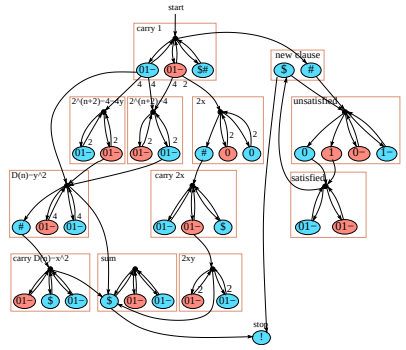
Naša pôvodná motivácia



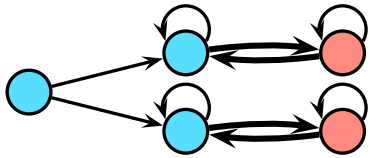
Zhrnutie

Všetky HMM

NP-ťažké

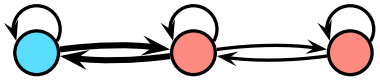


?



Viterbi alg. $O(nm^2)$

RVA $O(n^2m^3)$



ešte rozšír $O(n^3m^5)$