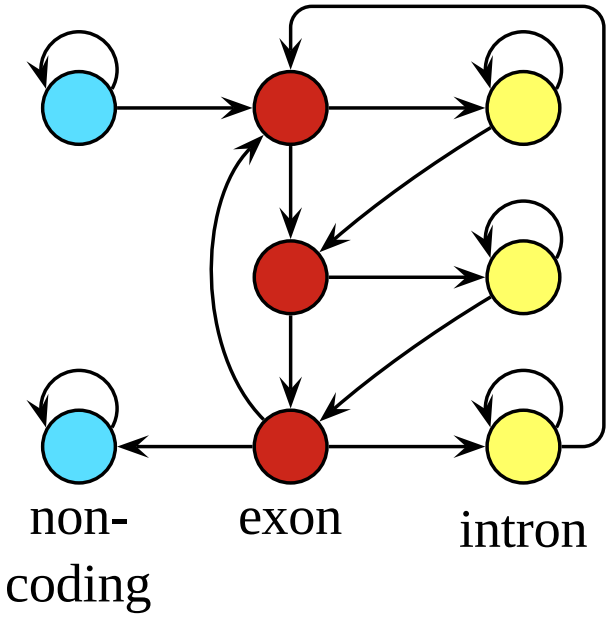
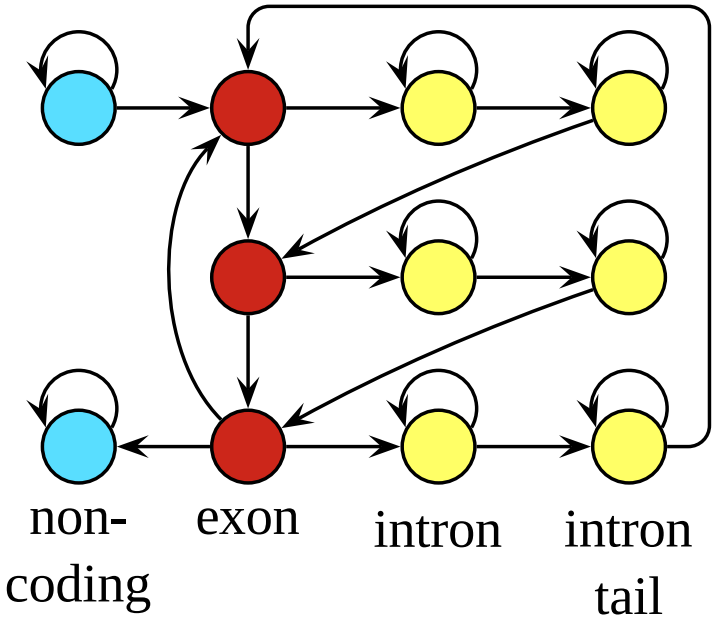
Najpravdepodobnejšia anotácia na zložitejších HMM

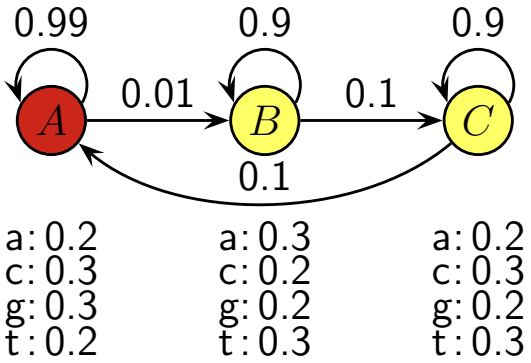
Môžeme mať viacero stavov s tým istým významom.

Príklad: koniec intrónu je obohatený o C a T





Cesta vs. anotácia



Sekvencia $X = x_1 \dots x_n \in \{a, c, g, t\}^*$

Cesta $\pi = \pi_1 \dots \pi_n \in \{A, B, C\}^*$

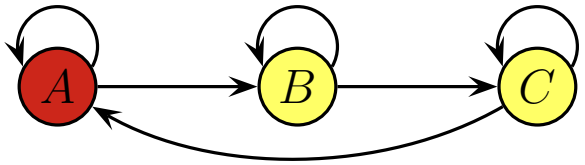
Annotácia $A = a_1 \dots a_n \in \{\blacksquare, \square\}^*$

AAAAAAAAAAAAAAAAAABBBBBBBBBCCCCCCCAAAAA
gttccgctgtgctgttttgtaggctcgcactaaggtcta

$$P(X, \pi) = s(\pi_1) e(\pi_1, x_1) \prod_{i=2}^{n} t(\pi_{i-1}, \pi_i) e(\pi_i, x_i)$$

$$P(X, A) = \sum_{\pi : \mathcal{A}(\pi) = A} P(X, \pi)$$

Cesta vs. anotácia



Sekvencia $X = x_1 \dots x_n \in \{a, c, g, t\}^*$

Cesta $\pi = \pi_1 \dots \pi_n \in \{A, B, C\}^*$

Annotácia $A = a_1 \dots a_n \in \{\blacksquare, \square\}^*$

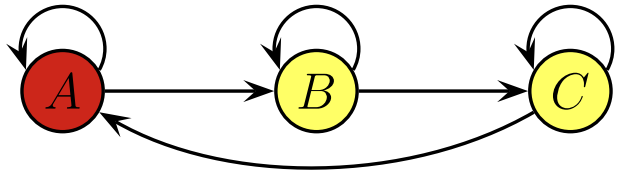
$$P(X, A) = \sum_{\pi : \mathcal{A}(\pi) = A} P(X, \pi)$$

Viterbiho algoritmus hľadá najpravdepodobnejšiu cestu

$\pi^* = \arg\max_\pi P(X, \pi)$

My chceme najpravdepodobnejšiu anotáciu $A^* = \arg\max_A P(X, A)$

Cesta vs. anotácia



p_1 : tcggaagtcta ctggtggcaaggcgc cacgc aaacagttggccacta aggcagcccgcat

p_2 : tcggaagtctact||ggtggcaaggcgccacgcaaacagttggccactaaggcagcccgcat

p_3 : tcggaagtctactggtggcaaggc||gccacgcaaacagttggccactaaggcagcccgcat

p_4 : tcggaagtctactggtggcaaggcgccacg||caaacagttggccactaaggcagcccgcat

...

p_k : tcggaagtctactggtggcaaggcgccacgcaaacagttggcca||ctaaggcagcccgcat

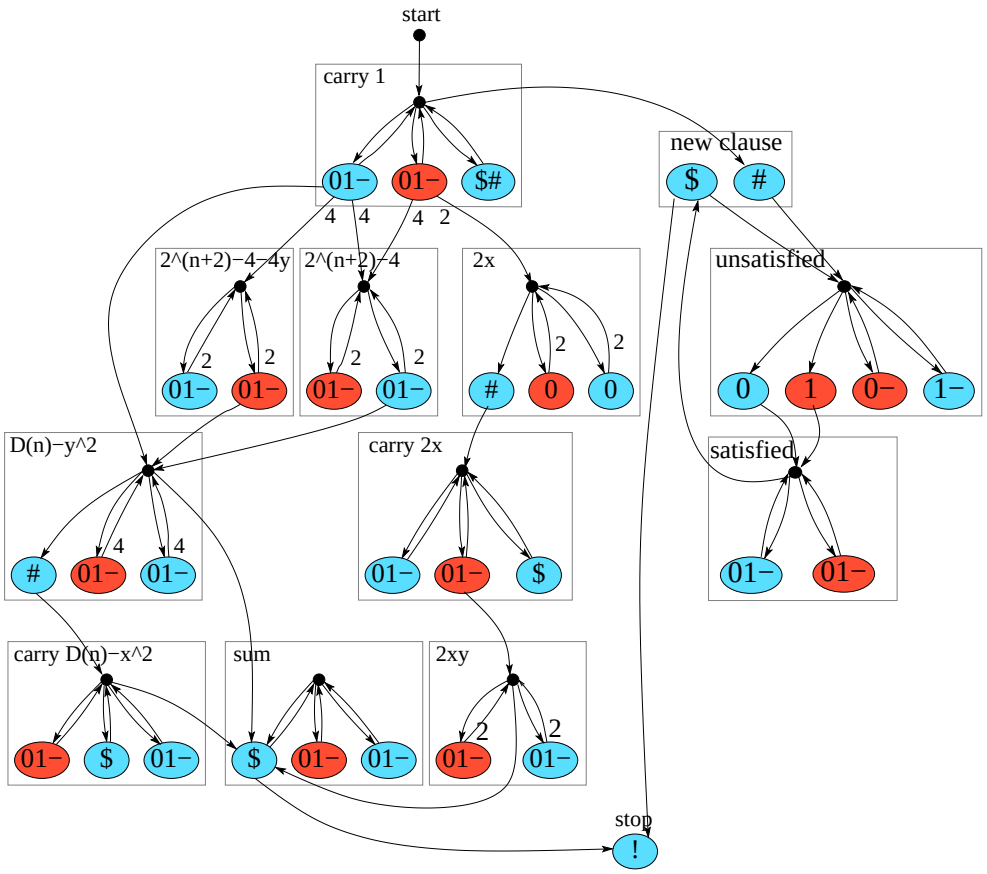
Viterbiho algoritmus: $\arg\max\{p_1, p_2, \ldots, p_k\}$

Najpravdepodobnejšia anotácia: $\arg\max\{p_1, p_2 + p_3 + \cdots + p_k\}$

Problém nájdenia najpravdepodobnejšej anotácie je NP-ťažký

[Lyngso and Pedersen 2002; B., Brown, Vinař 2007]

. . . nepodarí sa nám teda asi nájsť efektívny algoritmus



Logická formula

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

zakódovaná ako sekvencia

$X = \texttt{000\#1--\$-10\$!}$

najpravdepodobnejšia anotácia

$\Rightarrow$

ohodnotenie premenných

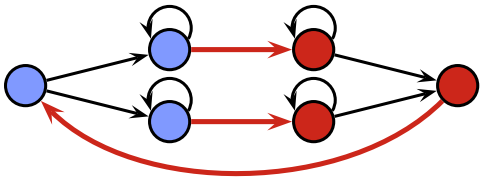
Pre niektoré HMM sa najpravdepodobnejšia anotácia dá spočítať v polynomiálnom čase

Všetky HMMs

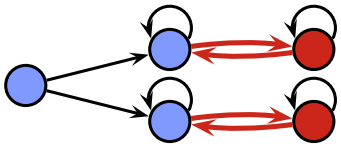
Viterbiho alg. $O(nm^2)$

$O(n^2m^3)$

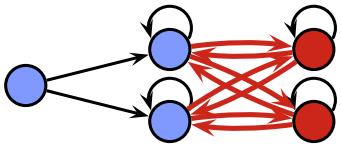
$O(n^3m^5)$



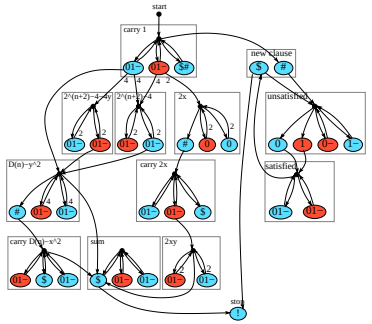
tiež $\in P$



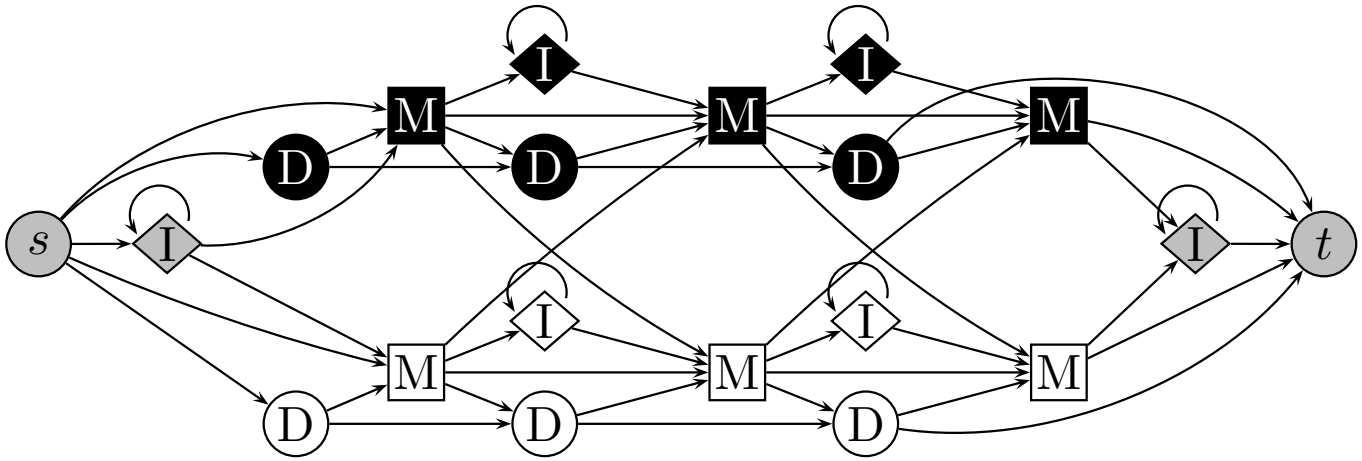
?



NP-ťažké



Annotation issues in jumping HMMs



State path: alignment of sequence to subtype profiles

Annotation: segments of inputs emitted by subtype profiles

Problems with most probable annotation:

— probably hard to decode

— many annotations with slightly shifted boundaries

Change the objective function for decoding

Gain function [Hamada et al. 2009]

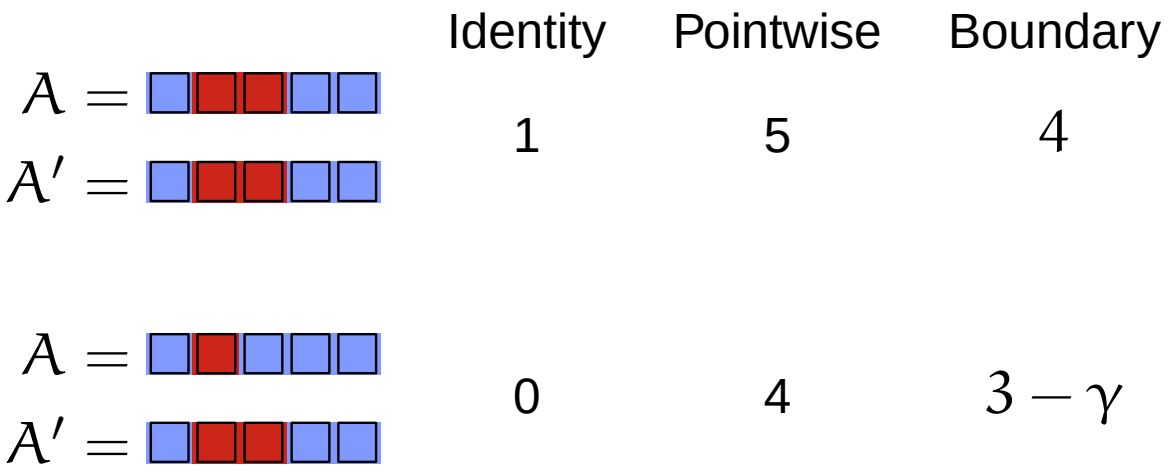
$G(A, A')$ measures accuracy of A wrt. correct annotation A'

Examples:

Identity: score 1 iff A completely correct, 0 otherwise

Pointwise: score +1 for every correct label in A

Boundary: score +1 for every correct boundary, $-\gamma$ for incorrect boundary



	Identity	Pointwise	Boundary
A = ■■■■■ A' = ■■■■■	1	5	4
A = ■■■■■ A' = ■■■■■	0	4	$3-\gamma$

Optimizing expected gain

Goal: find annotation A that maximizes

$$E_{A'|X}[G(A, A')] = \sum_{A'} G(A, A')P(A'|X)$$

Identity gain function: Viterbi algorithm

Pointwise gain function: Posterior decoding (forward-backward)

Boundary gain function: [Gross et al. 2007]

The choice of gain function is application-dependent