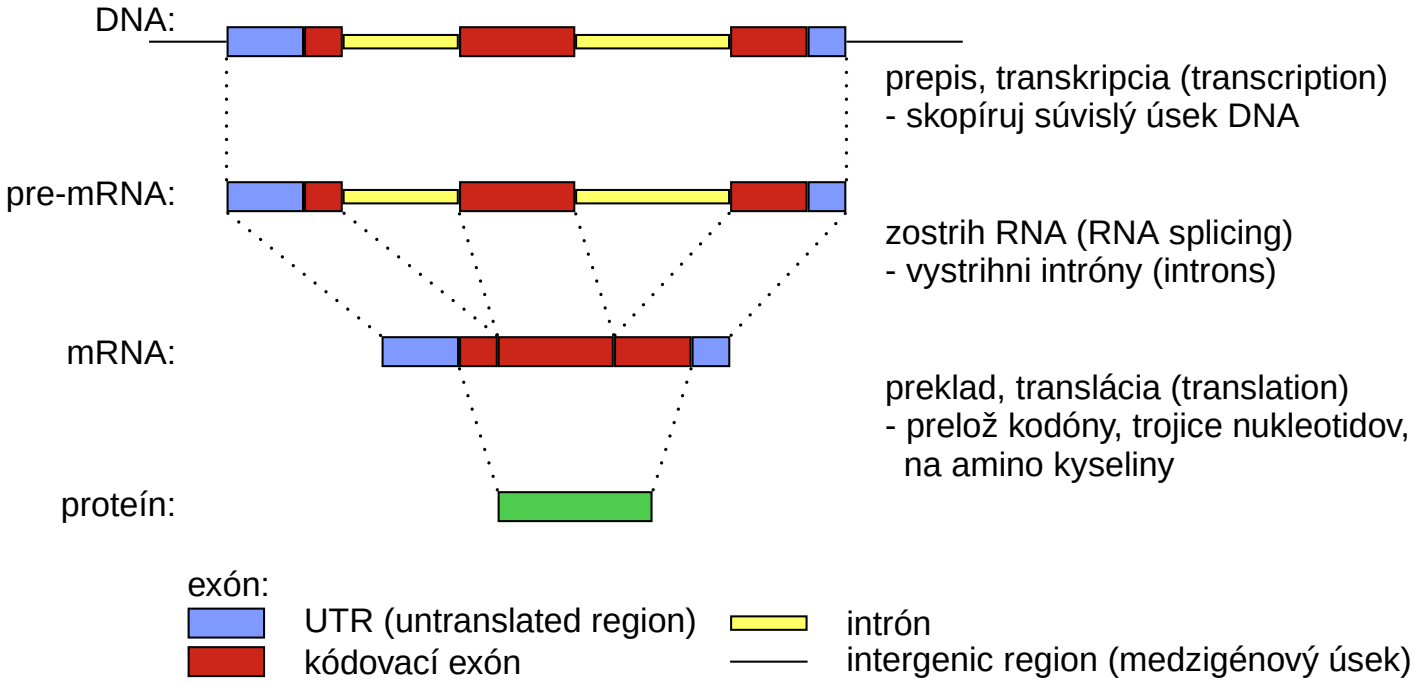
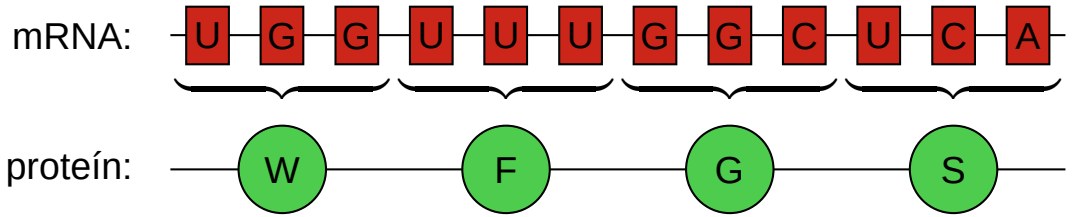
Štruktúra eukaryotických génov

Proces tvorby proteínov:



Translácia: tri bázy mRNA (kodón) → aminokyselina proteínu



Bioinformatický problém: Hľadanie génov

Cieľ: označ každú bázu ako intrón/exón/medzigénový úsek

cggtgaaactgcacgattgttgctggcttaaagatagaccaatcagagtgtgtaacgtca

tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca

tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca

ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta

aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc

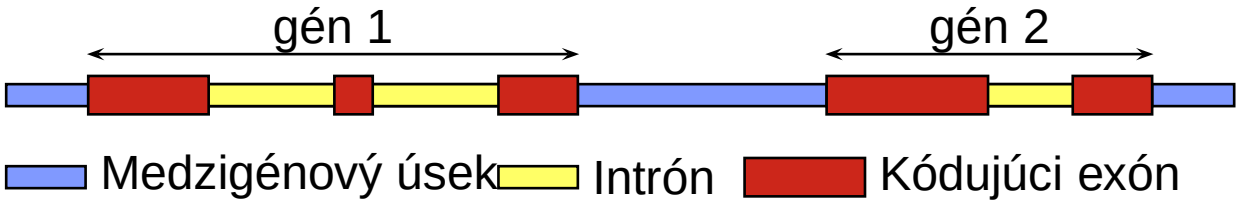
cgggcaccgtggctctgcgcgagatccgccgttatcagaagtccactgaactgcttattc

gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc

gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc

tatttgaggacactaacctgtgcgccatccacgccaagcgcgtcactatcatgcccaagg

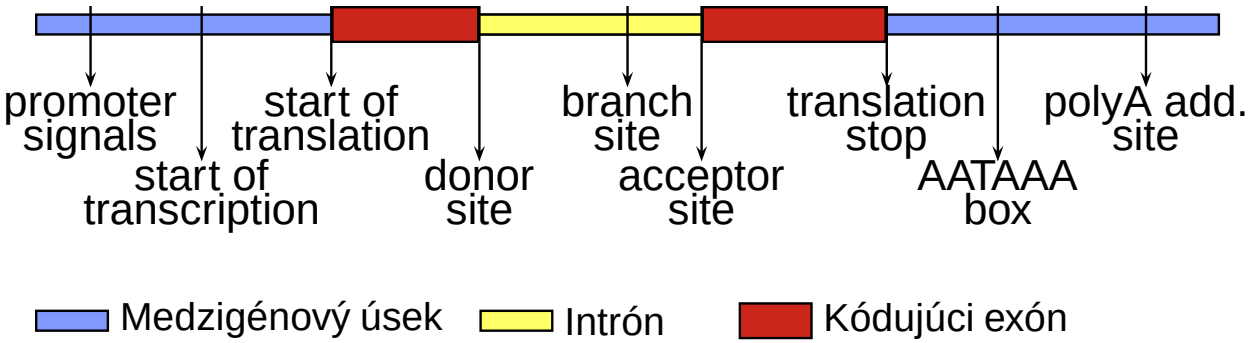
acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac



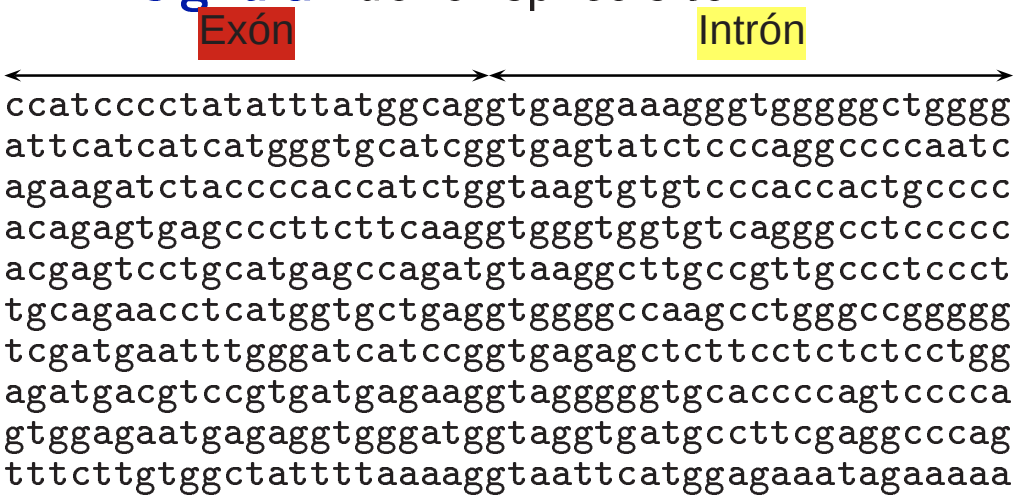
Ako spoznáme gény?

Signály na hraniciach exónov:

krátke reťazce, kde sa viažu komplexy zúčastnujúce sa na expresii génu



Príklad signálu: donor splice site



Ako spoznáme gény?

Zloženie sekvencie:

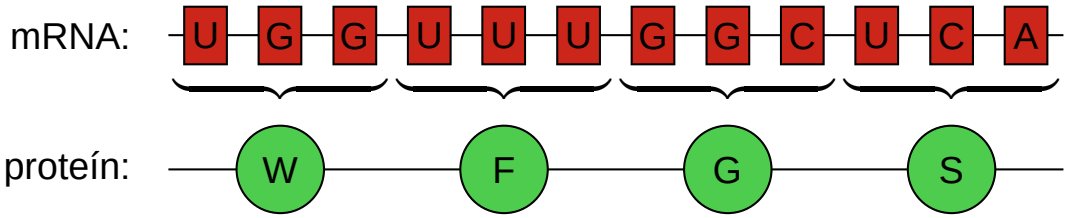
- iná frekvencia k-tic báz v kódujúcich a nekódujúcich úsekoch,
- kódujúce úseky sú 3-periodické,
- stop kodóny (TAA, TGA, TAG) len na konci posledného kódovacieho exónu.

Príklad: ak uvažujeme len jednotlivé bázy, exóny majú viac C a G (ľudský genóm)

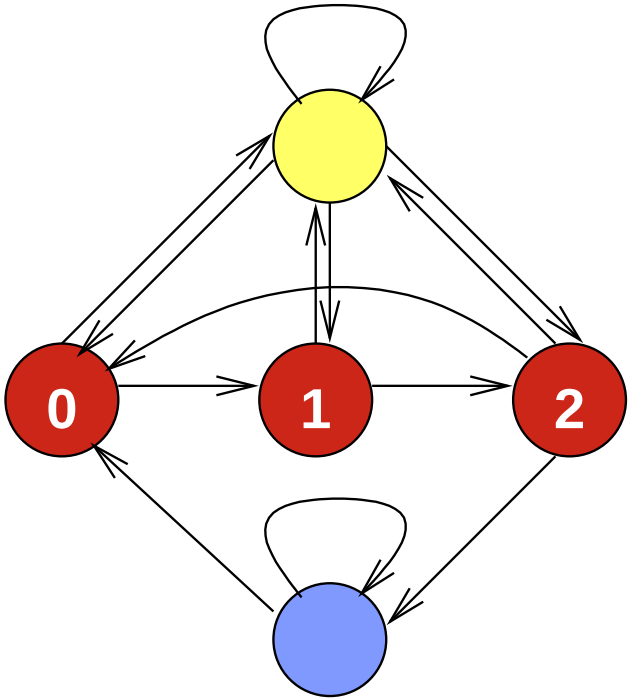
		a	c	g	t
kódujúci exón	0	0.26	0.26	0.32	0.16
	1	0.30	0.24	0.20	0.26
	2	0.17	0.32	0.31	0.20
intrón		0.26	0.22	0.22	0.30
medzig. úsek		0.27	0.23	0.23	0.27

HMM na hľadanie génov: 3-periodické exóny

Kodón (trojica báz) → jedna aminokyselina



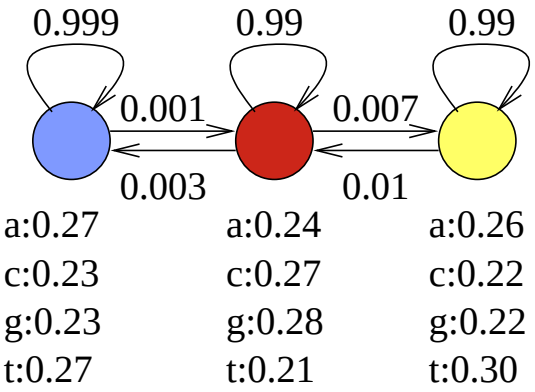
Namiesto jedného stavu pre exón použijeme tri stavy v cykle.

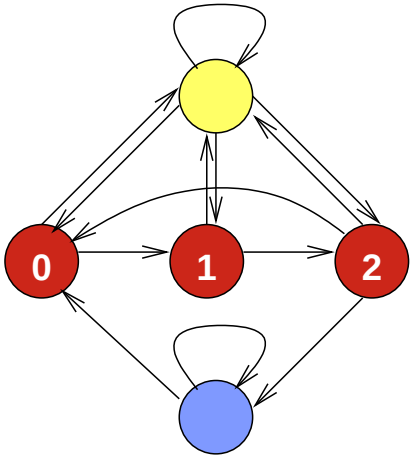


a	0	1	2		
0	0		0		0
1	0	0			0
2		0	0		
					0
		0	0	0	

$\Pr(A_i|A_{i-1})$

Nové stavy majú odlišné emisné pravdepodobnosti



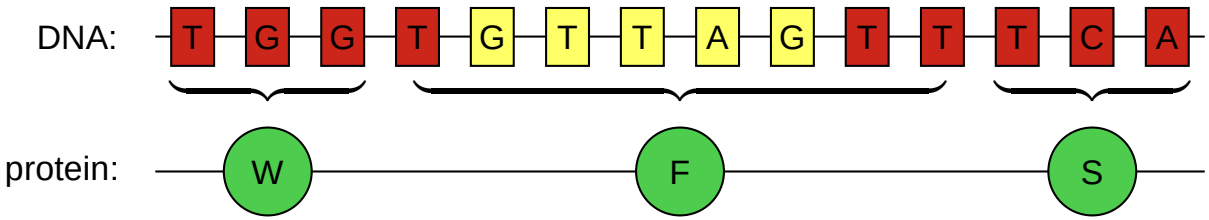


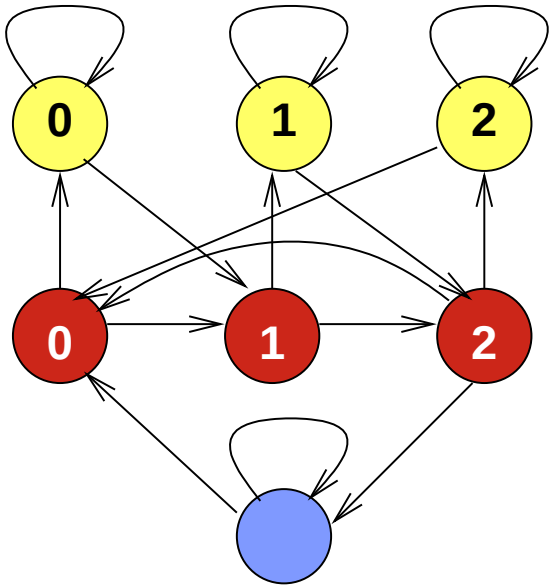
e	a	c	g	t
	0.24	0.27	0.28	0.21
	0.26	0.22	0.22	0.30
	0.27	0.23	0.23	0.27

e	a	c	g	t
0	0.26	0.26	0.32	0.16
1	0.30	0.24	0.20	0.26
2	0.17	0.32	0.31	0.20
	0.26	0.22	0.22	0.30
	0.27	0.23	0.23	0.27

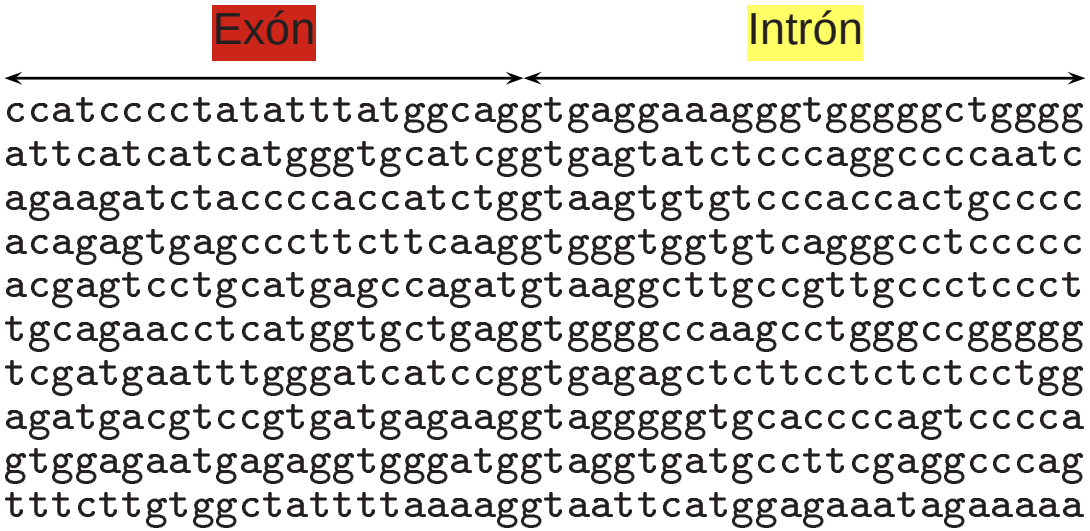
HMM na hľadanie génov: konzistentné kodóny

Intrón môže prerušiť kodón uprostred, chceme pokračovať, kde sme prestali.

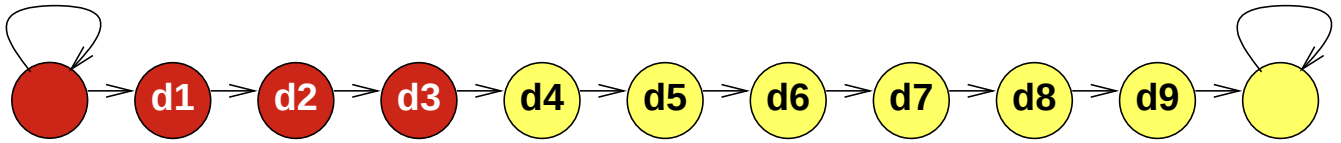




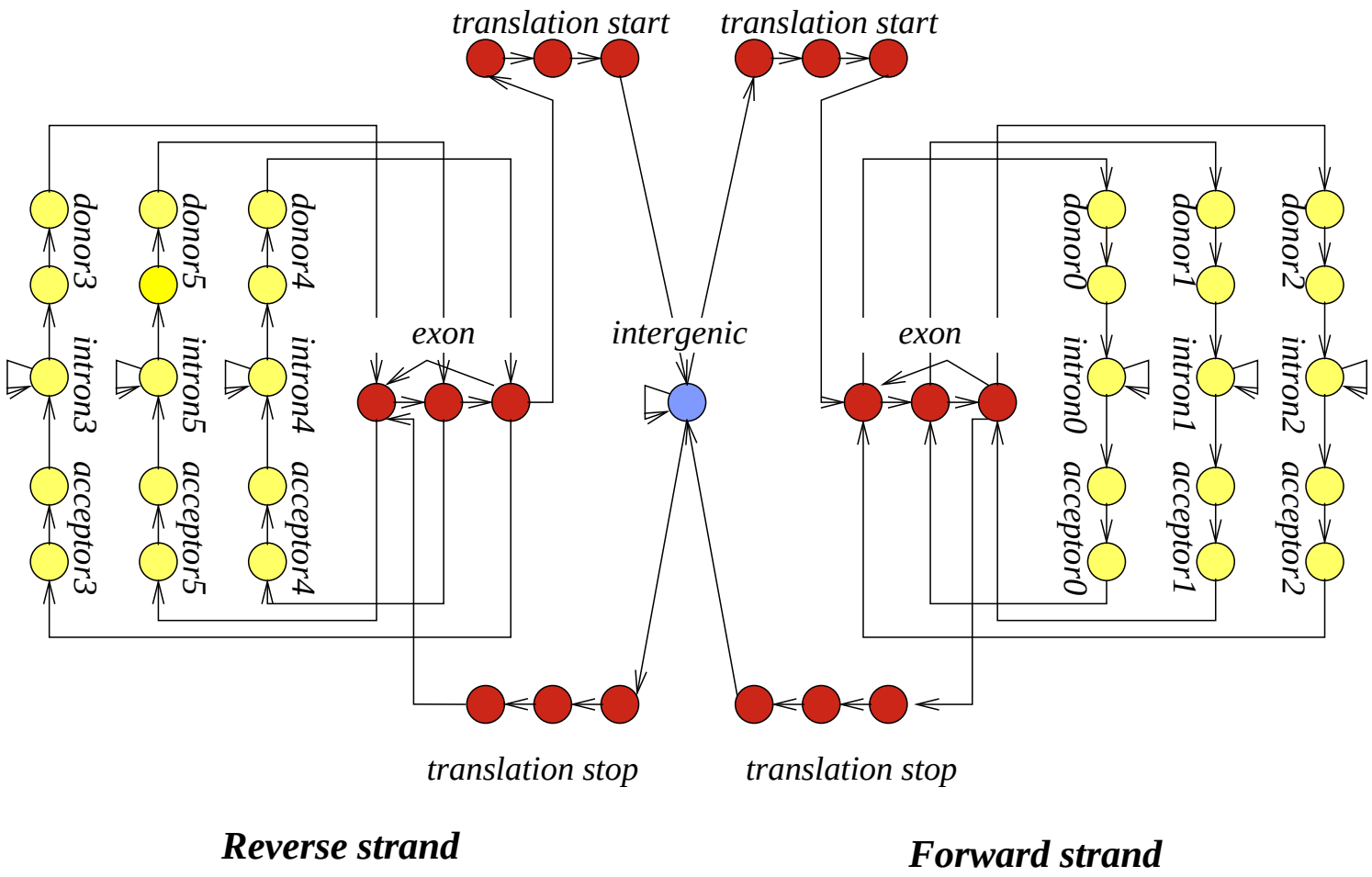
HMM na hľadanie génov: signály



Pridaj sériu stavov medzi exón a intrón:



HMM na hľadanie génov: celkový model



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.