IV107 Bioinformatika I

Přednáška 4

Katedra informačních technologií
Masarykova Univerzita Brno

Jaro 2011

Před týdnem

Existují techniky pro manipulaci, modifikaci, kopírování a detekci DNA, RNA a proteinů.

- rekombinace a klonování DNA
- PCR
- hybridizace DNA a RNA
- měření aktivity proteinů
- DNA čipy, microarray, proteinové čipy
- zjišťování sekvence

Outline

Anotace sekvence a genů
Anotace genomu
Identifikace genů
Homologie a podobnost
Příště
Bioinformatické databázy

Sekvence DNA

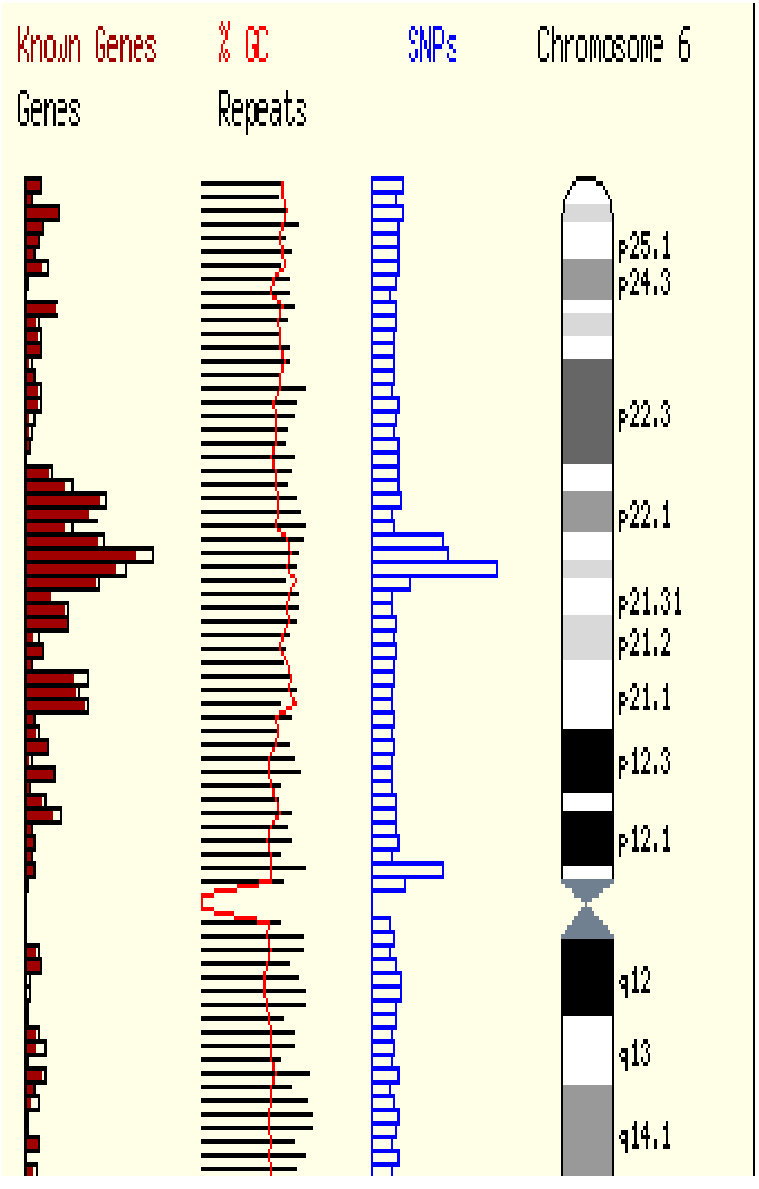

```
>P12345 Yeast chromosome1
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TACAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATTAC
AGATTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
```

Anotovaná sekvence DNA

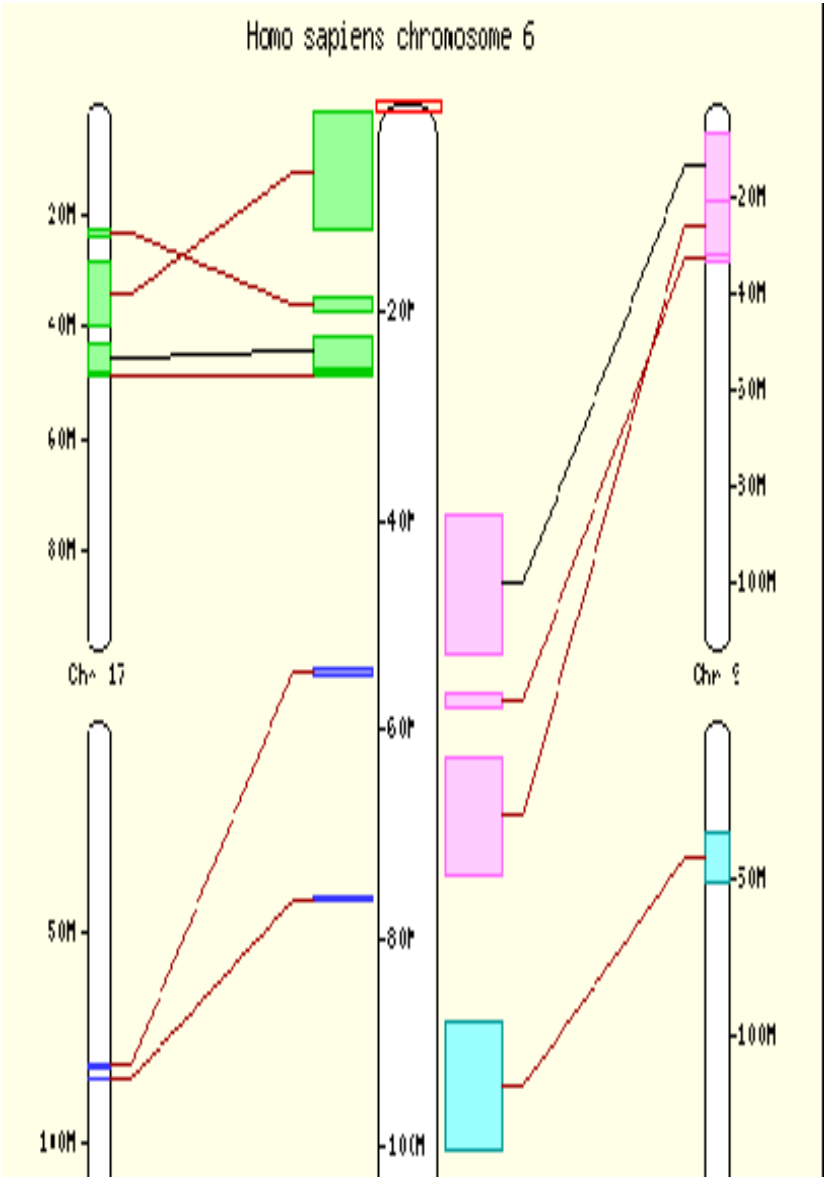

```
>P12345 Gen1 - protein
alkoholdehydrogenáza
TATA      TATAAA
          CGATTGACGATGACGAT
start     ATG
exon1     TACAGATTACAGATTACAGATTAAGATGT
intron1   CAGATTACAGATTACAGATTACACAGATTCA
exon2     AGATTACAGATTACAGATTACAGA
stop      TAA


>P12346 Protein1
MASAQSFYLLDHNQNQNFDDHLAVDIVMILSHERFMN
```

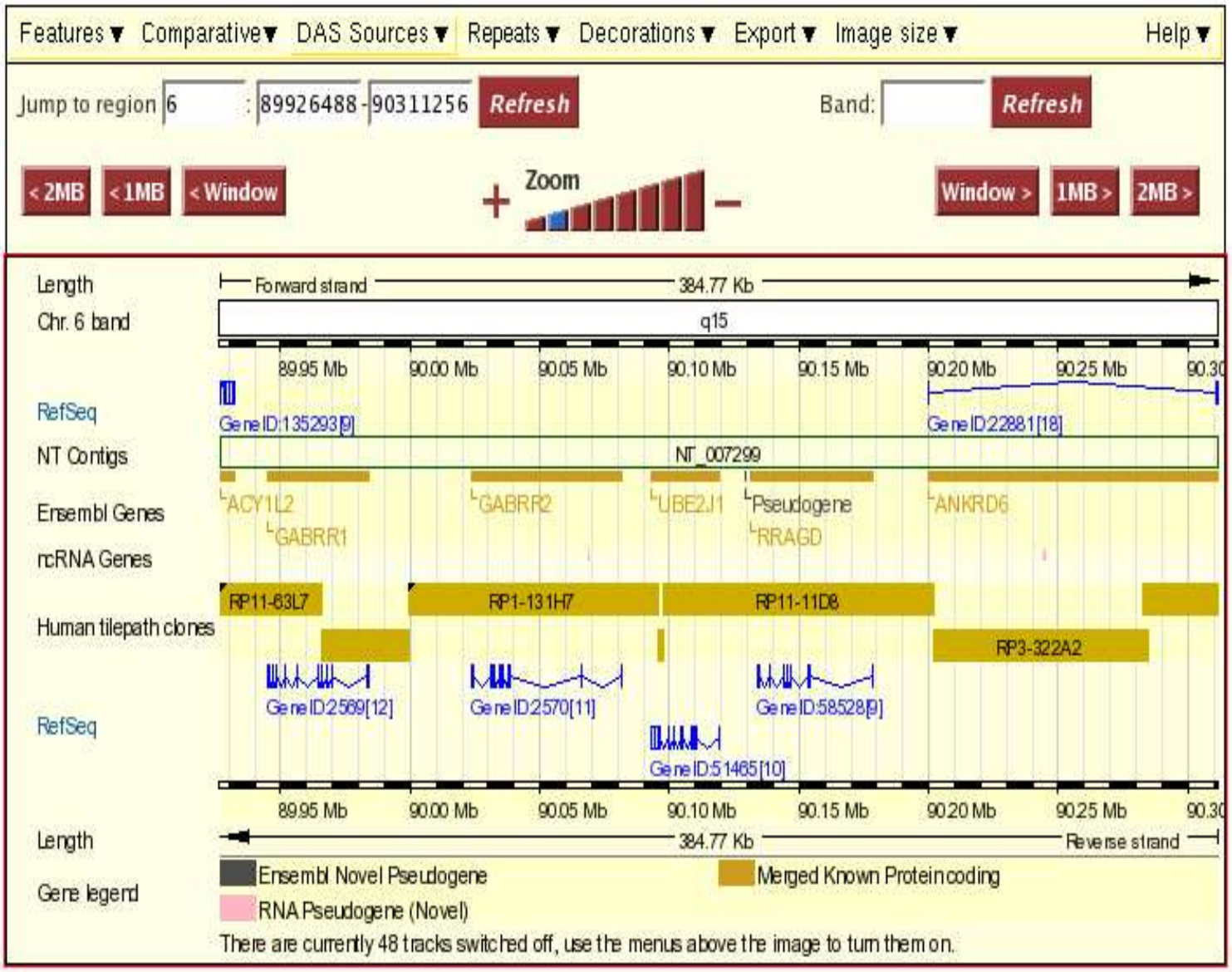
Anotace genomu



Anotace genomu



Anotace genomu http://www.ensembl.org/



Způsoby identifikace genů in silico

- Experimentální metody (cDNA, EST)
- Komparativní metody
  - Selekční tlak
  - Druh zachovaných mutací
- Strukturní metody (GeneMark, GeneScan,GeneID)
- Detekce charakteristických signálů

Identifikace genů podle charakteru mutací

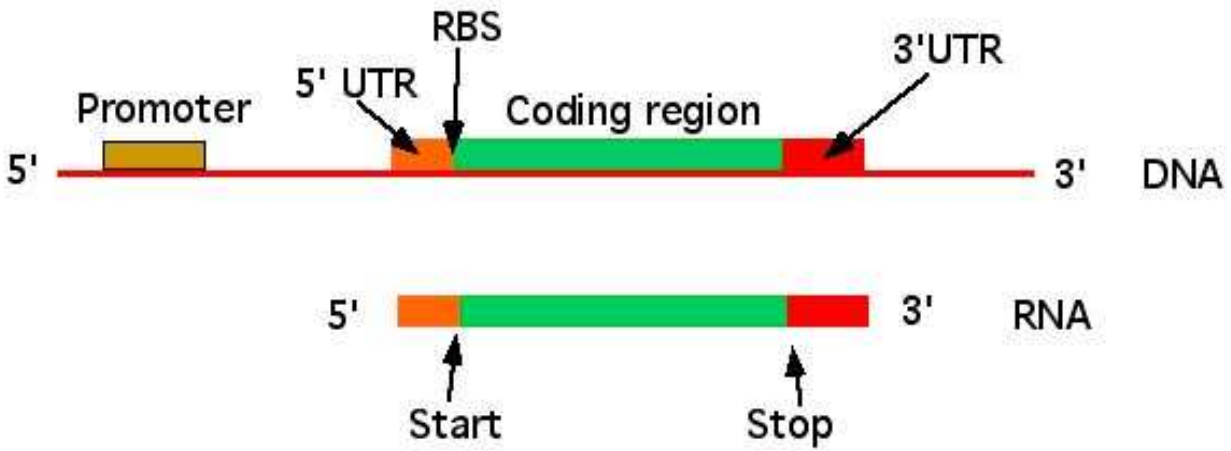
Využití známé struktury genů

- intergenová DNA
- geny
  - kódující protein
    - statistika sekvence
    - ORF
    - exon/intron (u eukaryotů)
    - promotor
  - RNA geny (rRNA, tRNA, jiné)

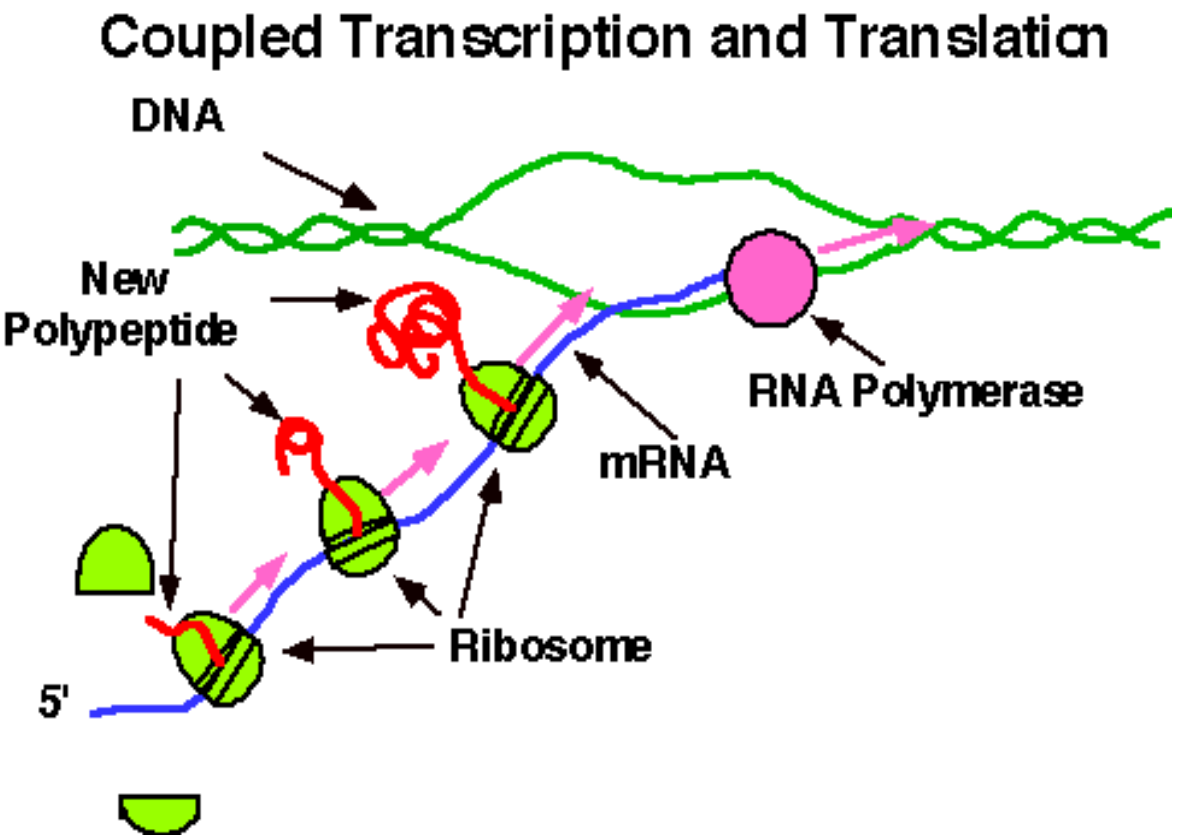
U prokaryotů 95-100% spolehlivost, u složitějších eukaryotů 90% na úrovni bazí, 70% na úrovni exonů/intronů

- existence intronů
- větší genomy
- nízká hustota genů ($<$30%; 3% u Homo sapiens)
- alternativní splicing (zhruba u poloviny genů)
- velké množství repetitivních sekveních
- občasný překryv genů

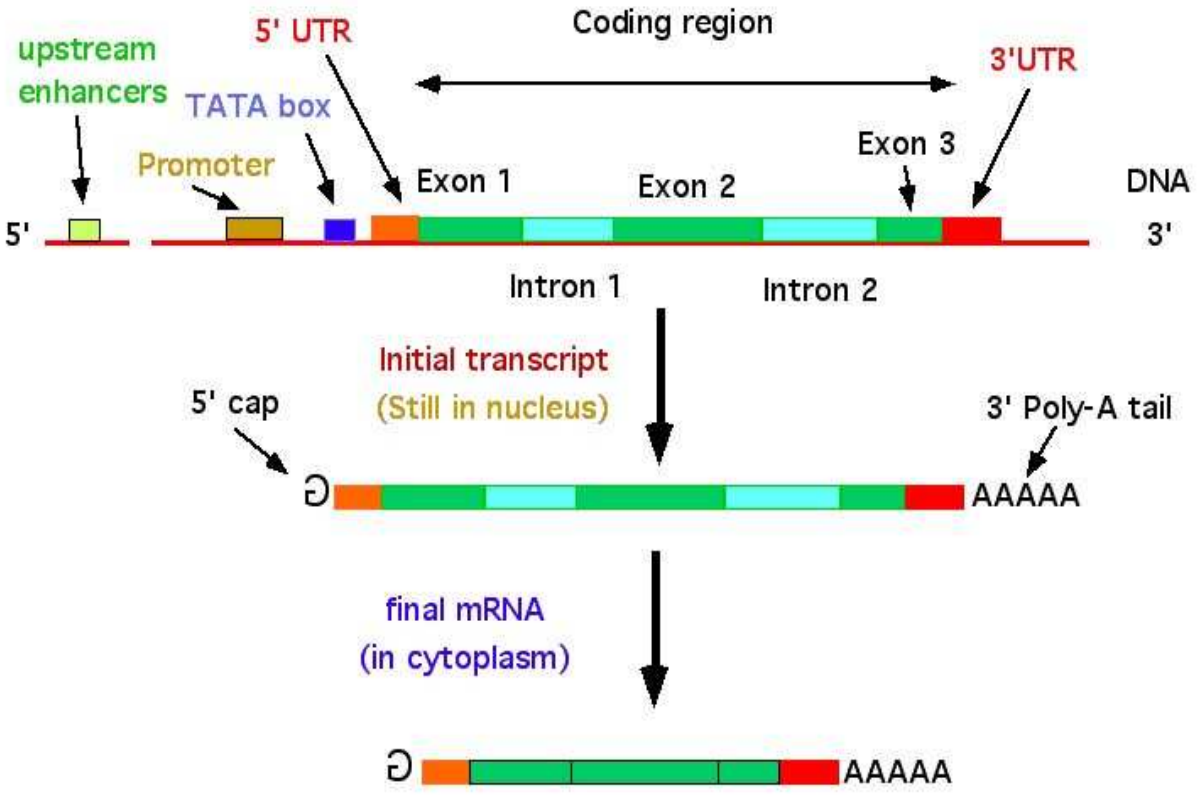
Struktura genu (prokaryotická)



Vztah transkripce a translace u prokaryotů



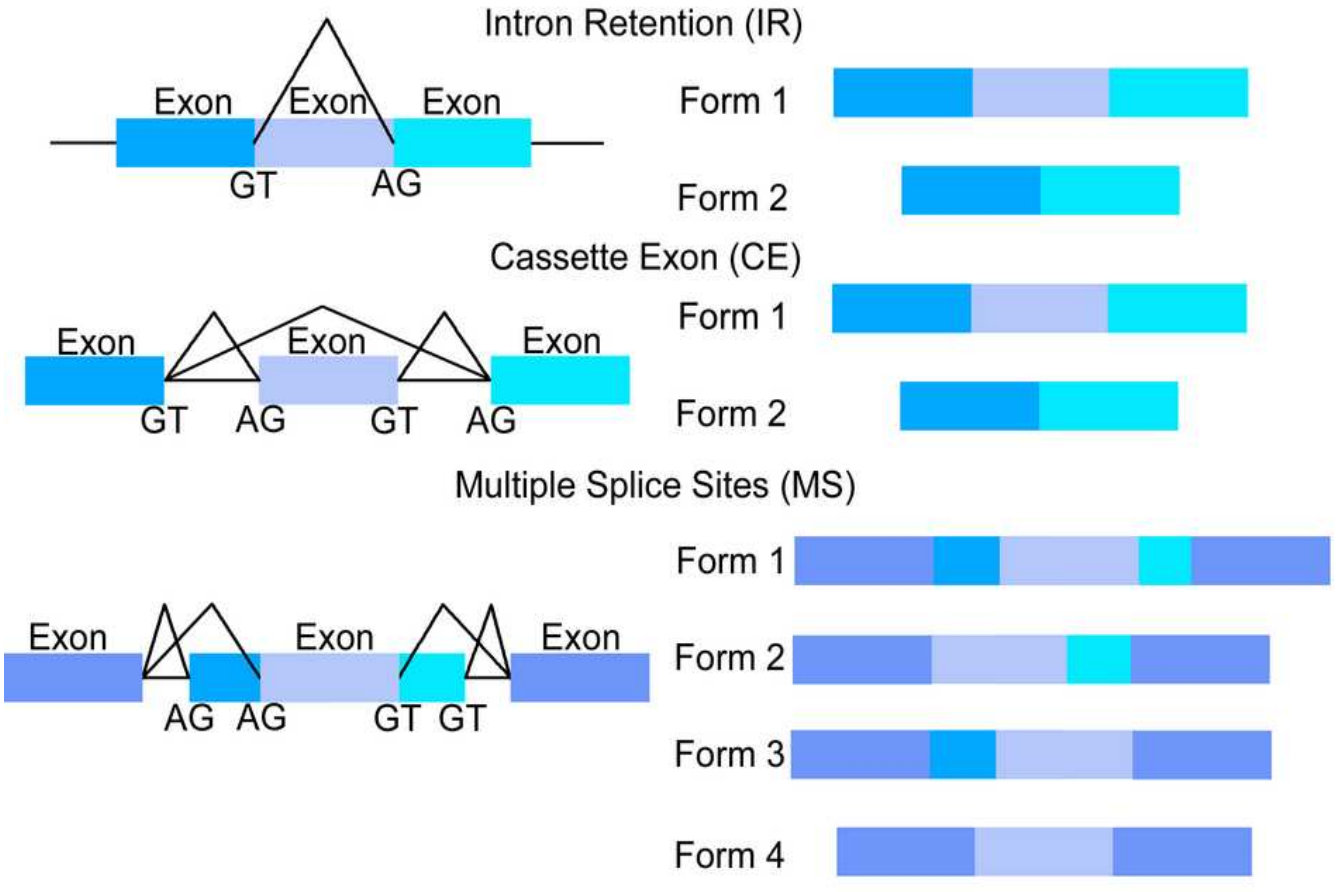
Struktura genu (eukaryotická)



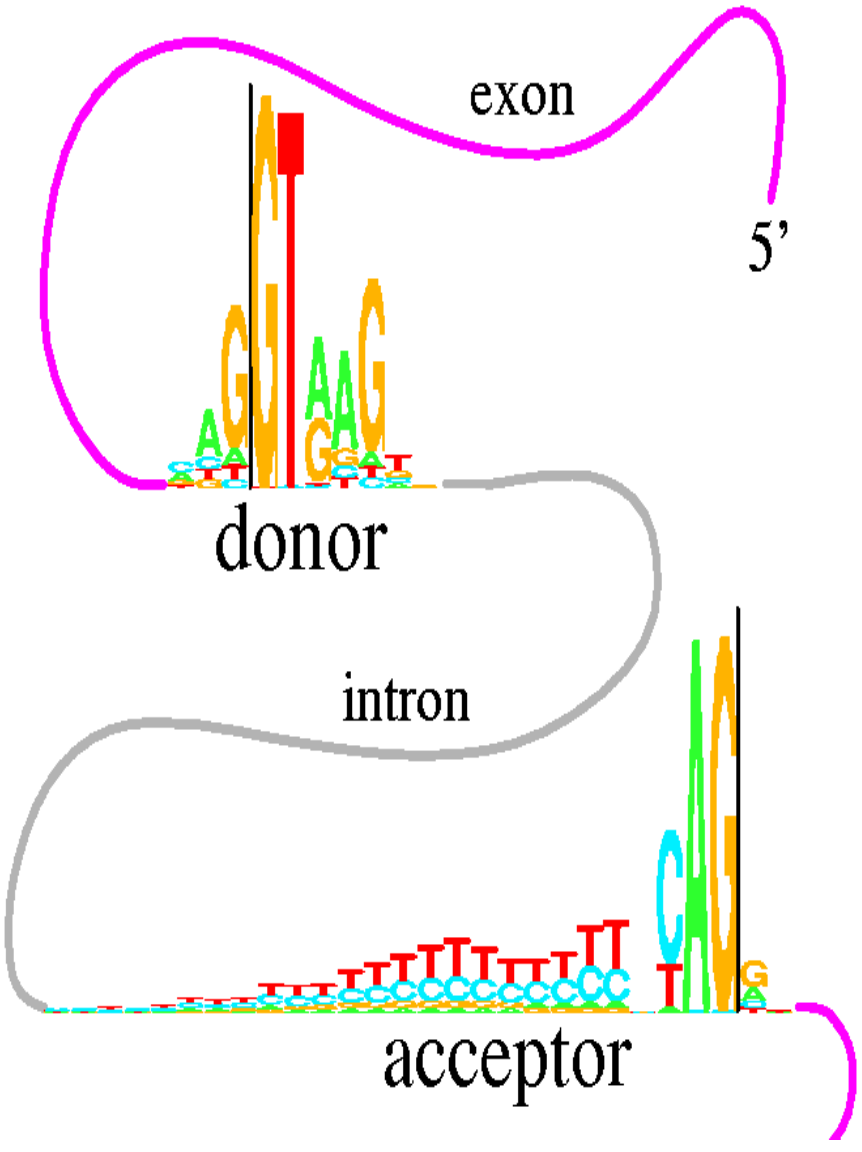
Typické sekvence v eukaryotických genech

- Enhancer
- Promotor
  - vazební místo transkripčního faktoru (aktivátor, represor)
  - TATA-box
- 5′-UTR
  - Začátek transkripce
- Kódující oblast
  - Začátek translace (často ATG)
  - exony
  - introny
    - donor (ag/GTaagt)
    - akceptor (cAG/gt)
    - lariat (CU[AG]A[CU])
  - terminátor translace (stop kodon = UAG—UAA—UGA)
- 3′-UTR
  - polyadenylačnísignál (AATAAA)

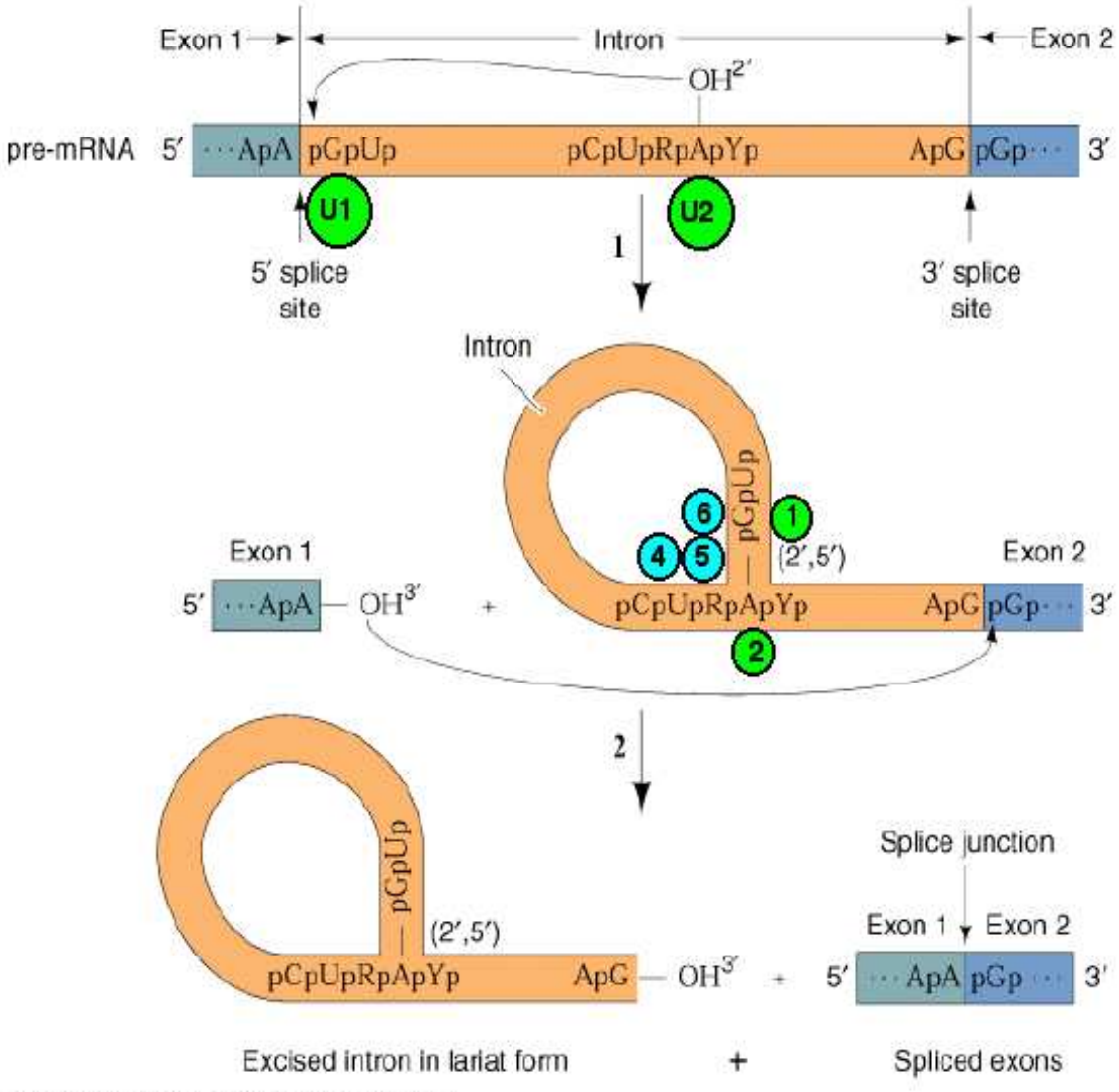
Sestřih mRNA



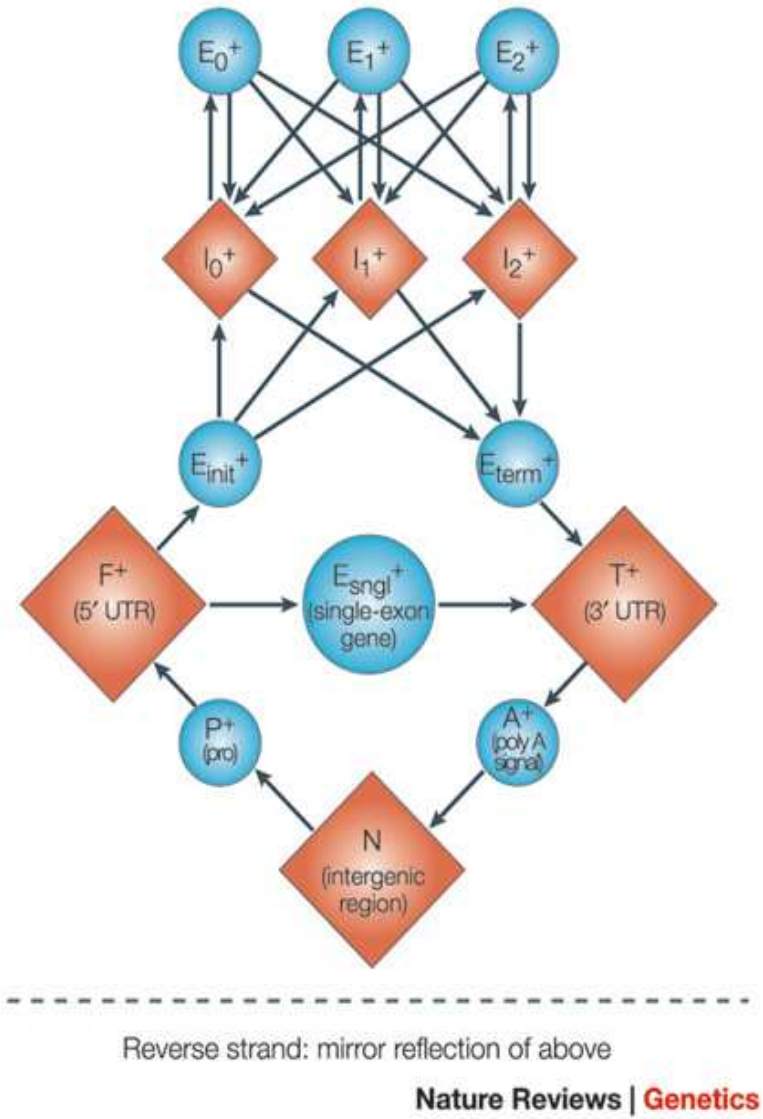
Sekvenční logo intronu



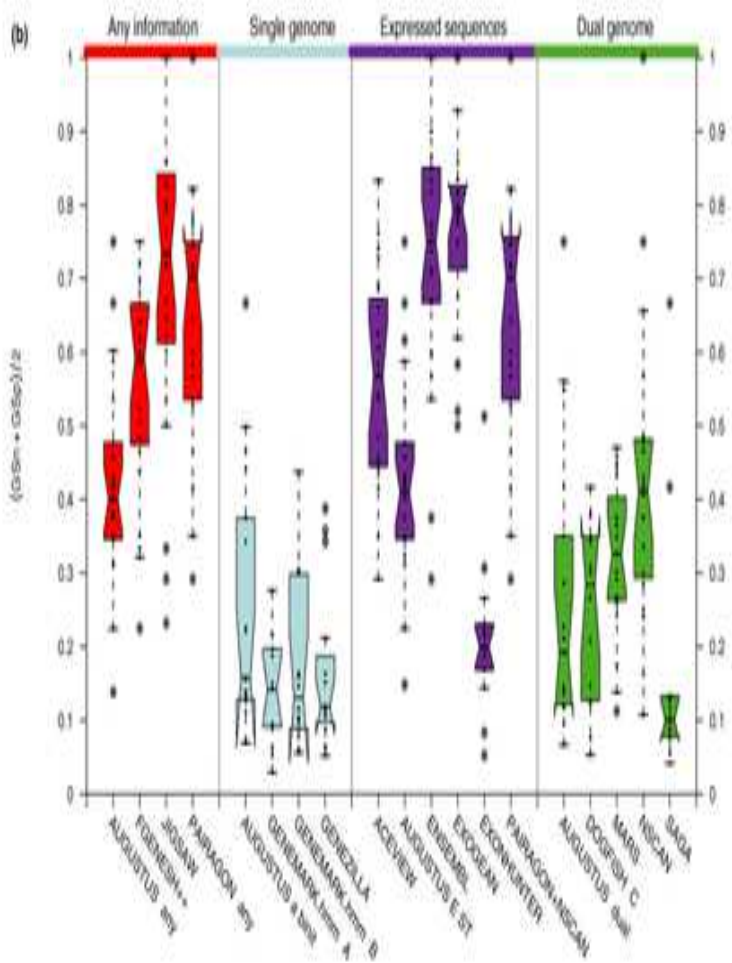
detaily sestřihu



Identifikace genů podle struktury



Úspěšnost identifikace genů

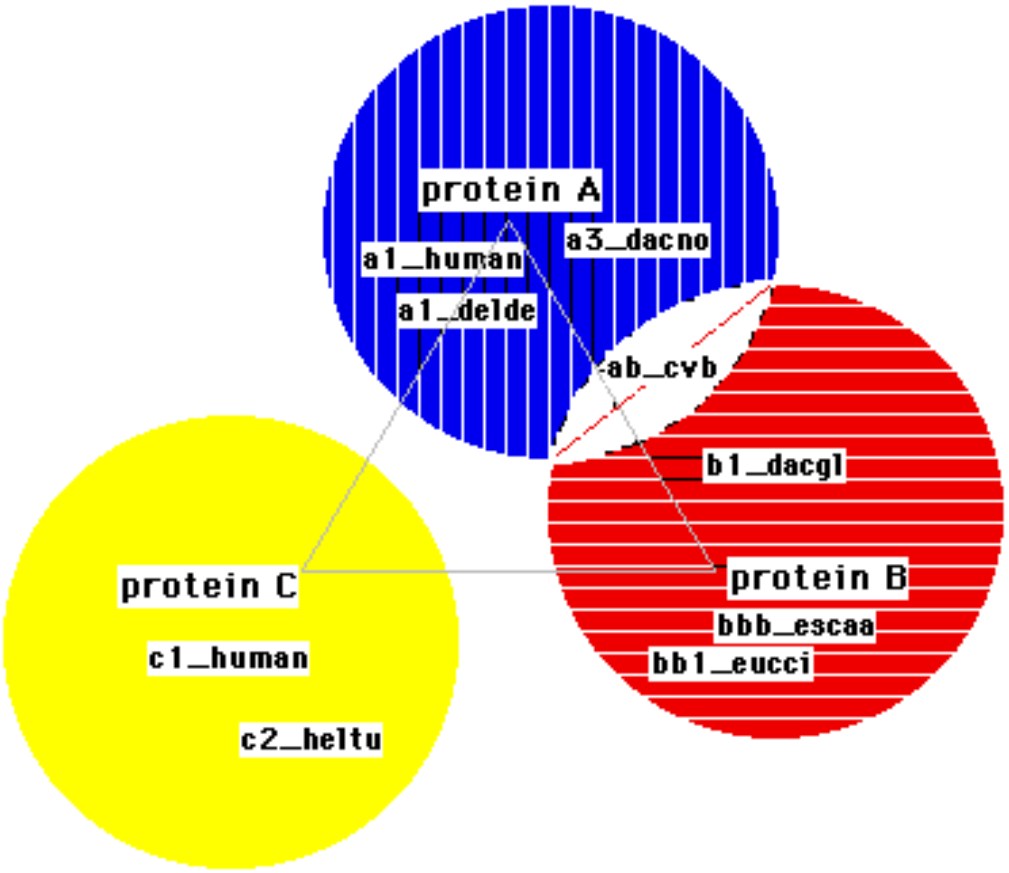


Guigo et al, Genome Biology (2006)

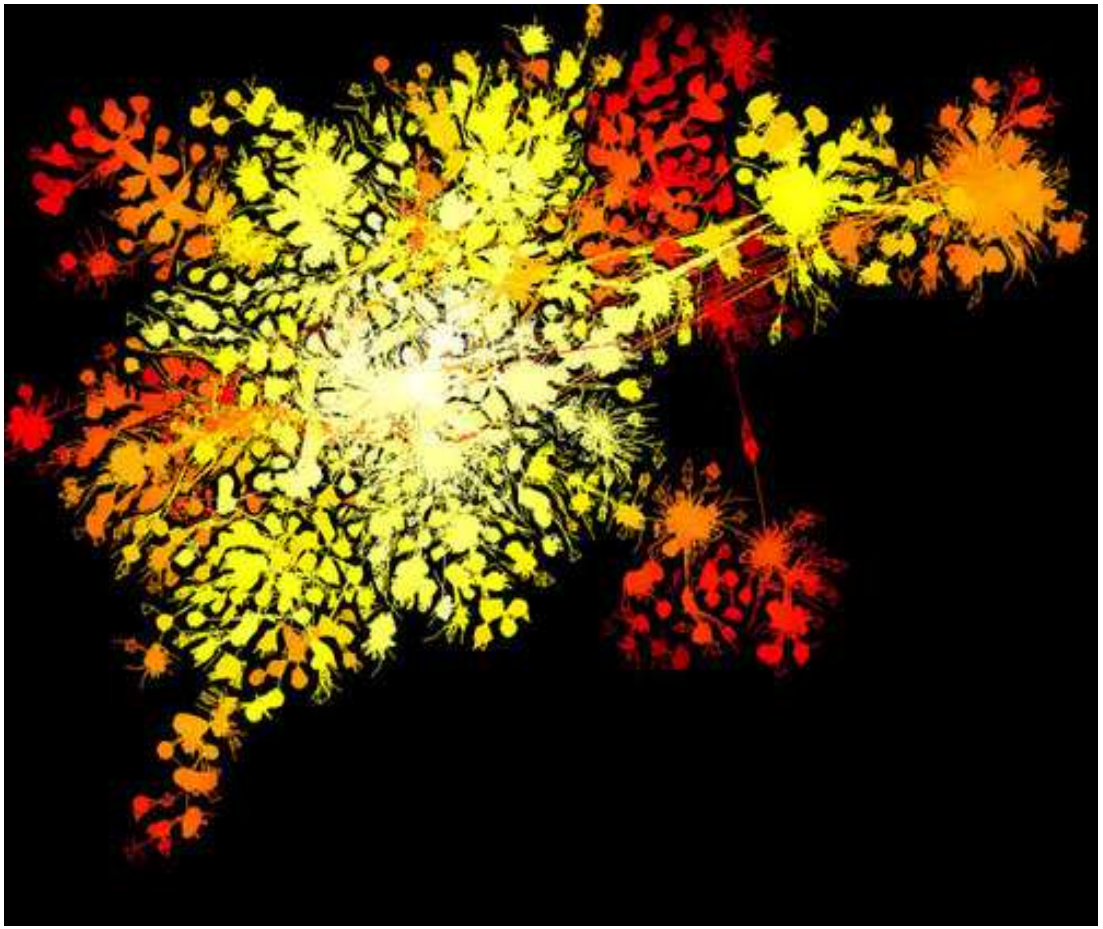
Příbuzné geny mají podobnou funkci i sekvenci

Rost studoval proteiny s různou sekvenční podobností. Zjistil, že když je víc než 30% aminokyselin identických, proteiny mají velmi podobnou strukturu.

Rost - prostor podobnosti

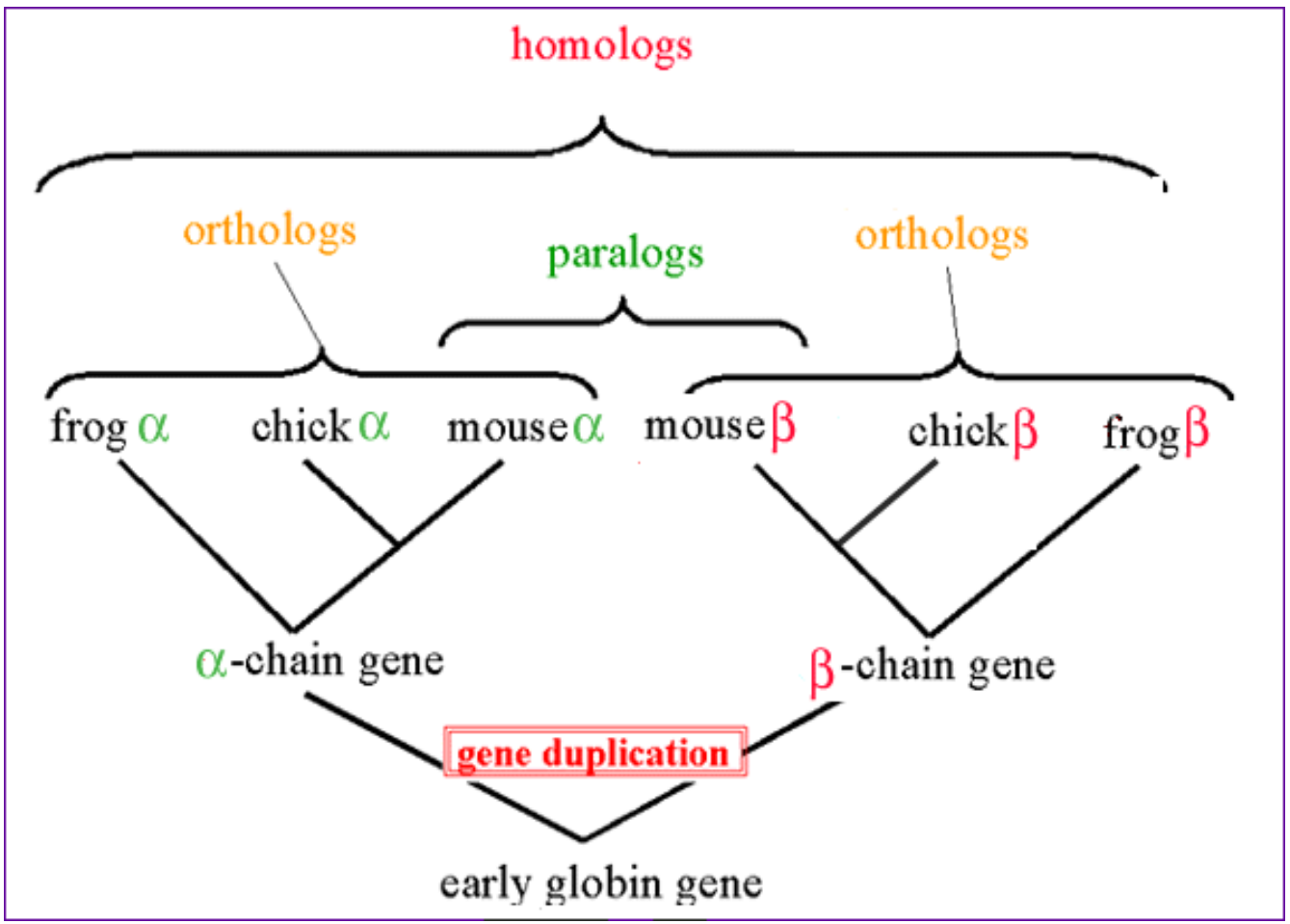


Síť proteinů podle sekvenční podobnosti



Proteiny přepojené podle sekvnenční podobnosti. Každý z 30727 vrcholů reprezentuje protein, kazdá z 1,206,654 hran podobnost. Seed Madazine. Červenec

homologie

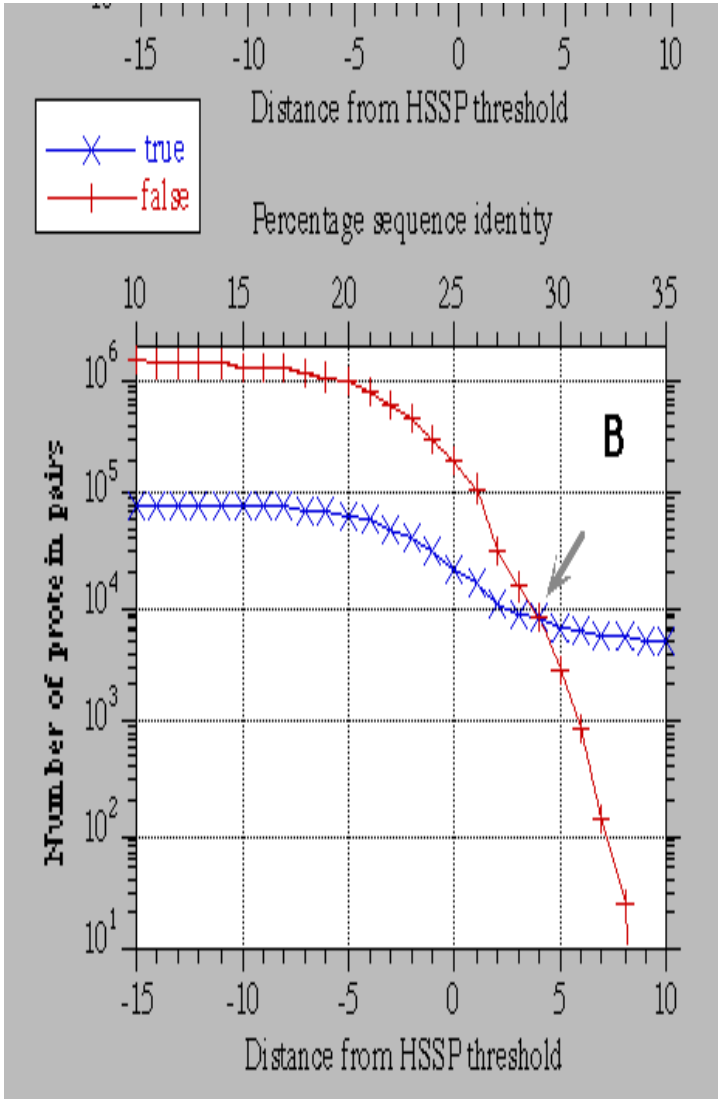


Příbuznost a podobnost sekvencí

- Homologie
  buď je nebo není
- Podobnost
  lze kvantifikovat a stupňovat

Od určitého stupně podobnosti je homologii velmi pravděpodobná. U proteinových sekvencí od cca. 30% identity.

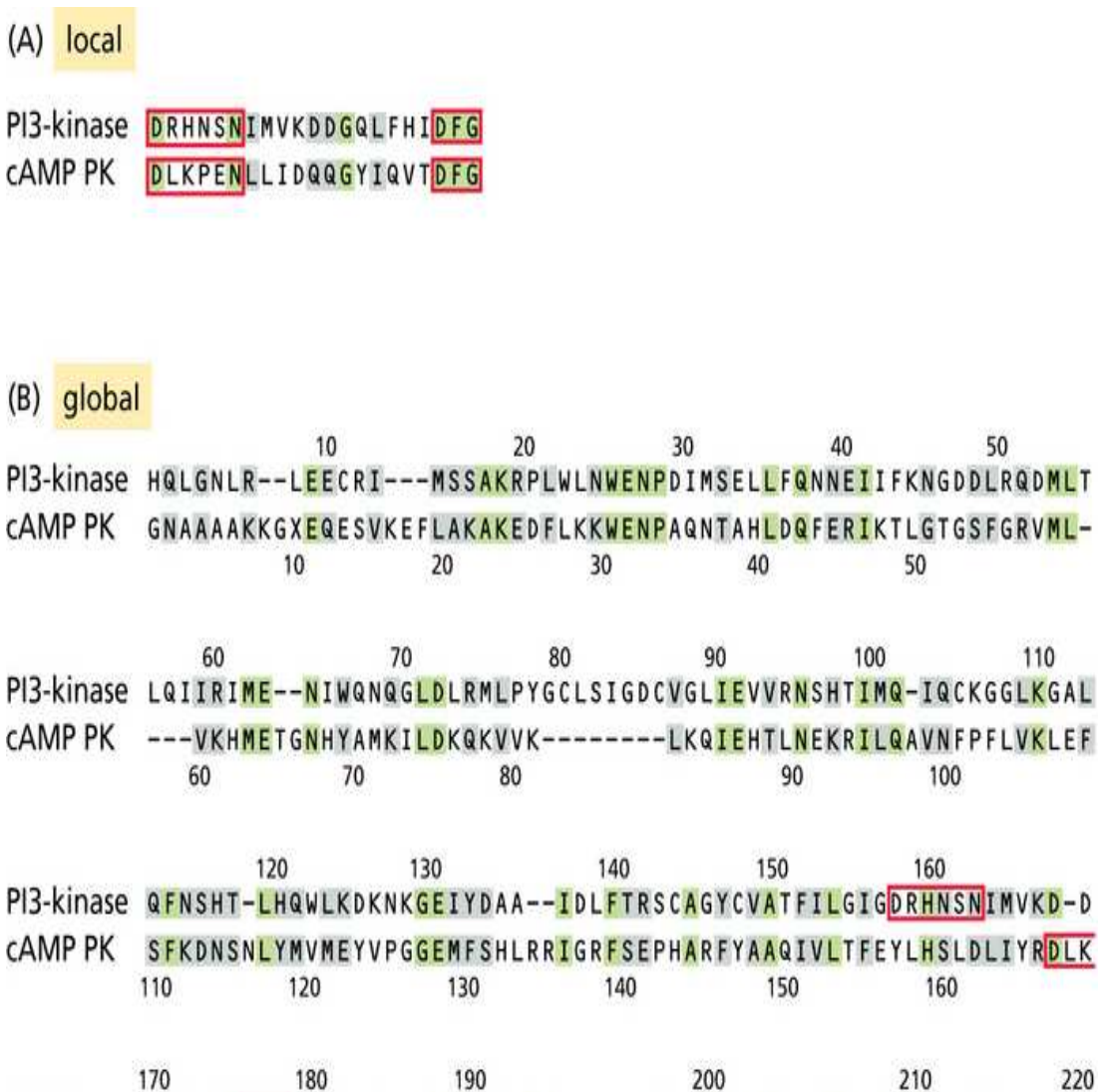
Rost - ”twilight zone”



Podobnost sekvencí

- bez zarovnání (přiložení)
  - např obsah n-gramů
- se zarovnáním (přiložením)
  - stejná délka, pozice si odpovídají
  - libovolná délka, pozice přiřazujeme

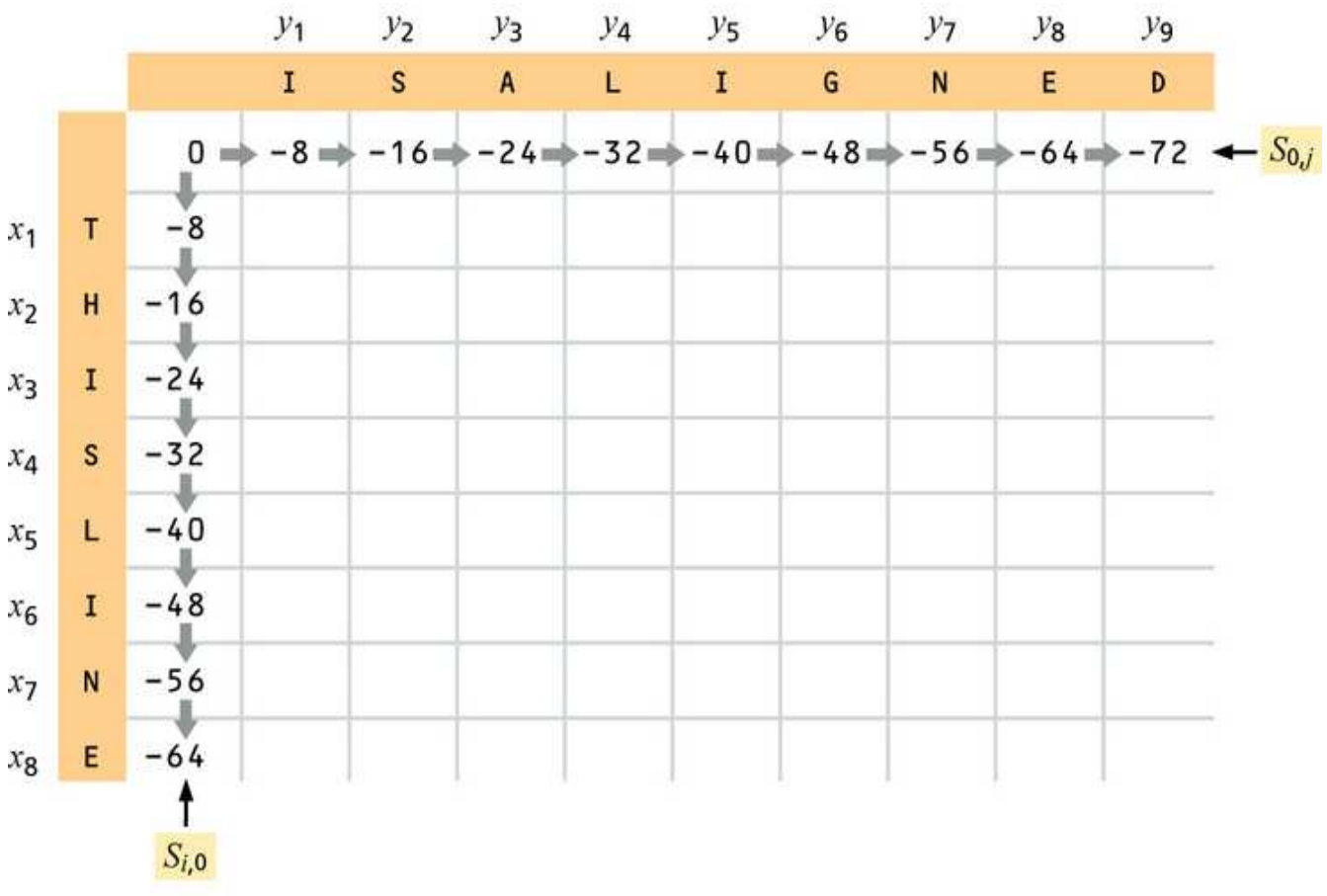
Rozdíl mezi lokálním a globálním porovnáváním



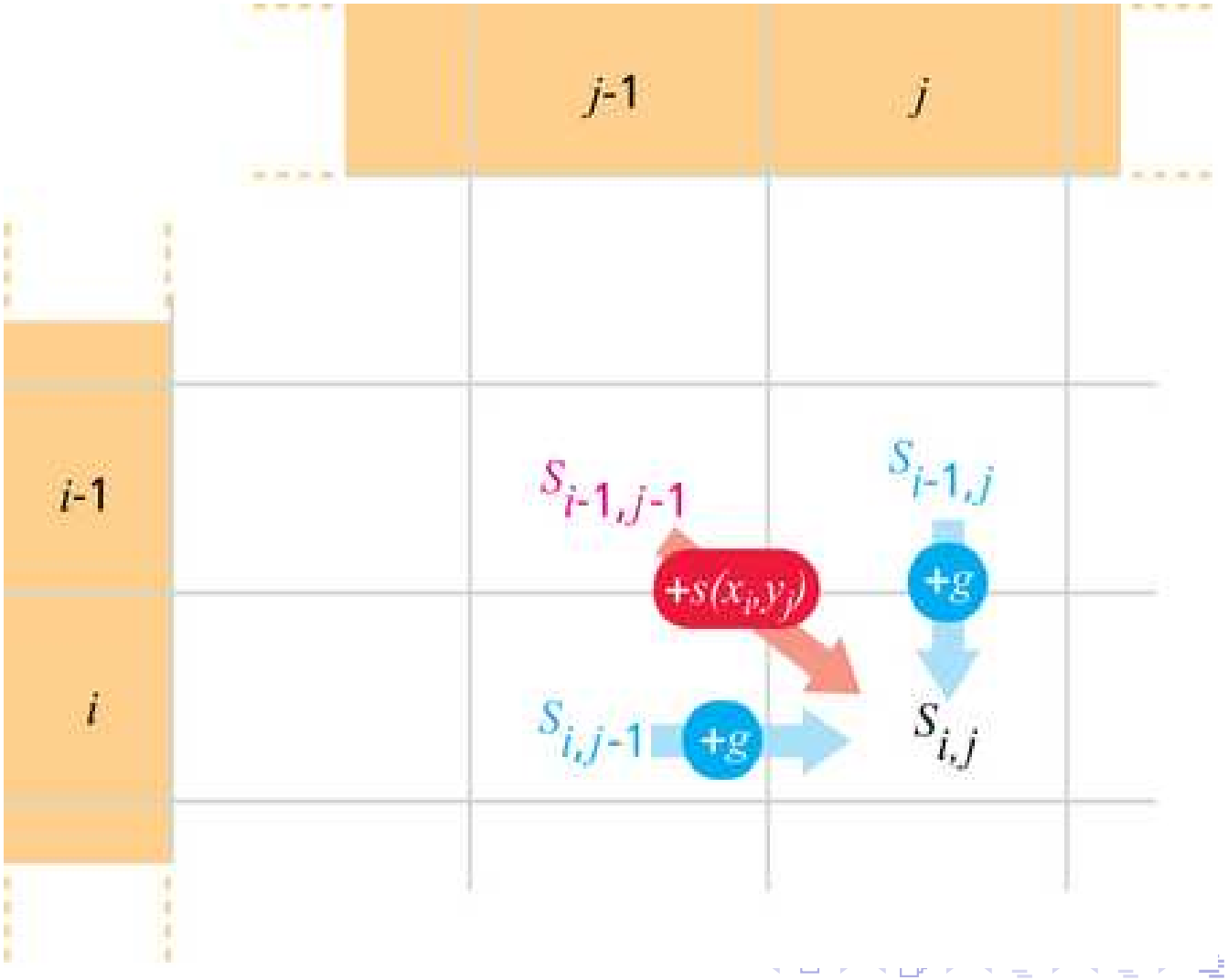
Matice pro hodnocení podobnosti proteinových sekvencí

C	9																		
S	-1	4																	
T	-1	1	5																
P	-3	-1	-1	7															
A	0	1	0	-1	4														
G	-3	0	-2	-2	0	6													
N	-3	1	0	-2	-2	0	6												
D	-3	0	-1	-1	-2	-1	1	6											
E	-4	0	-1	-1	-1	-2	0	2	5										
Q	-3	0	-1	-1	-1	-2	0	0	2	5									
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8								
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5							
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5						
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5					
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4				
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4			
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4		
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

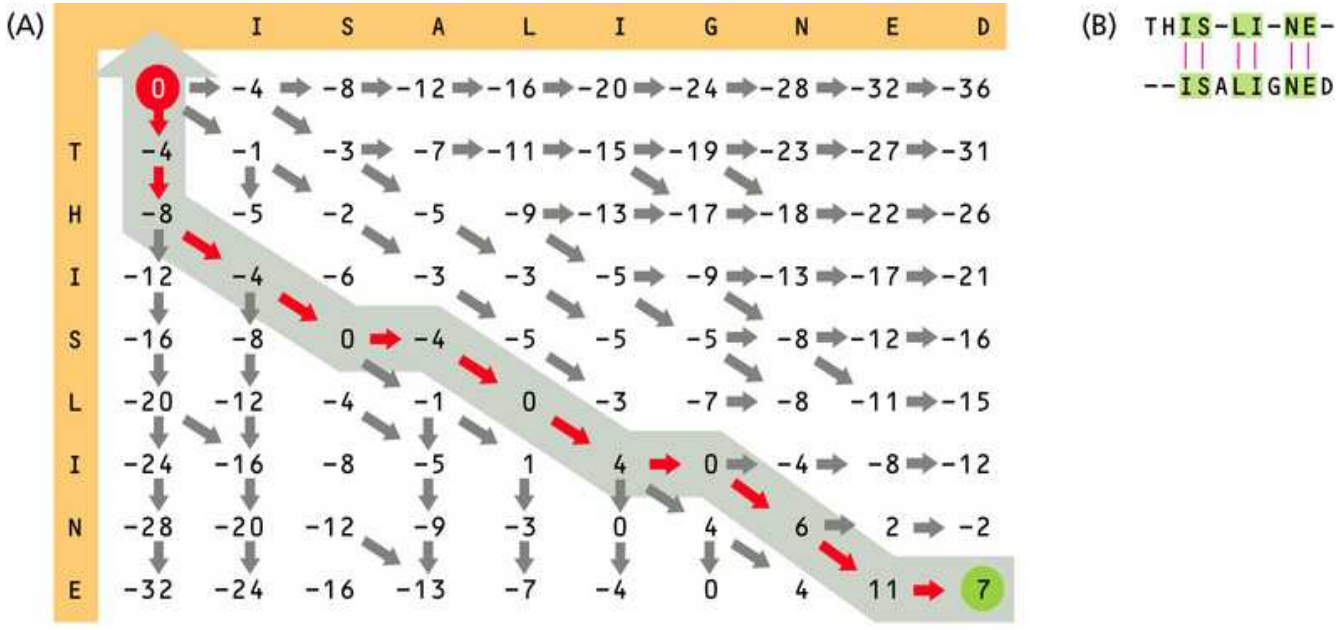
Tabulka pro algoritmus dynamického programování



Tabulka pro algoritmus dynamického programování

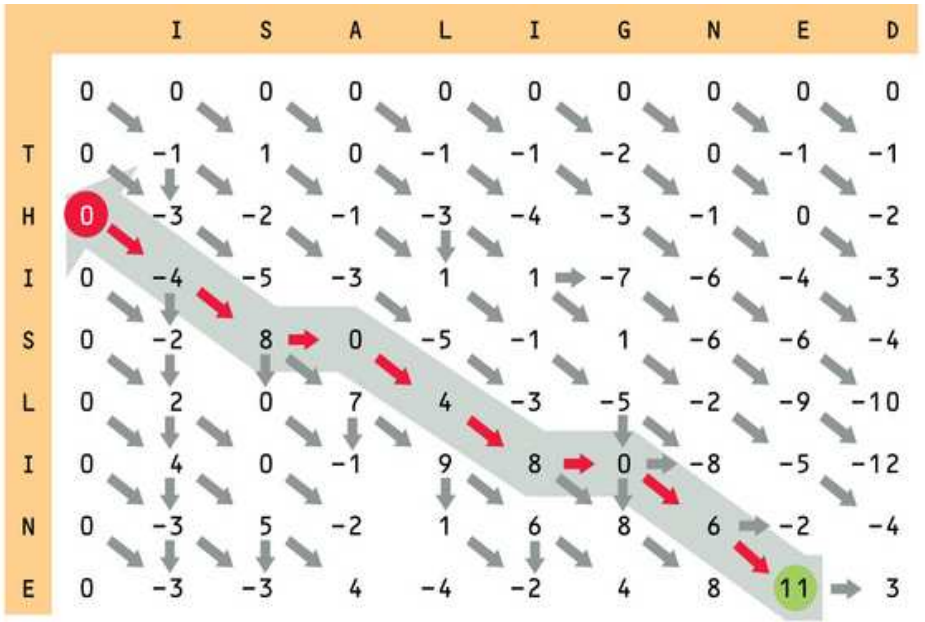


Tabulka pro algoritmus dynamického programování



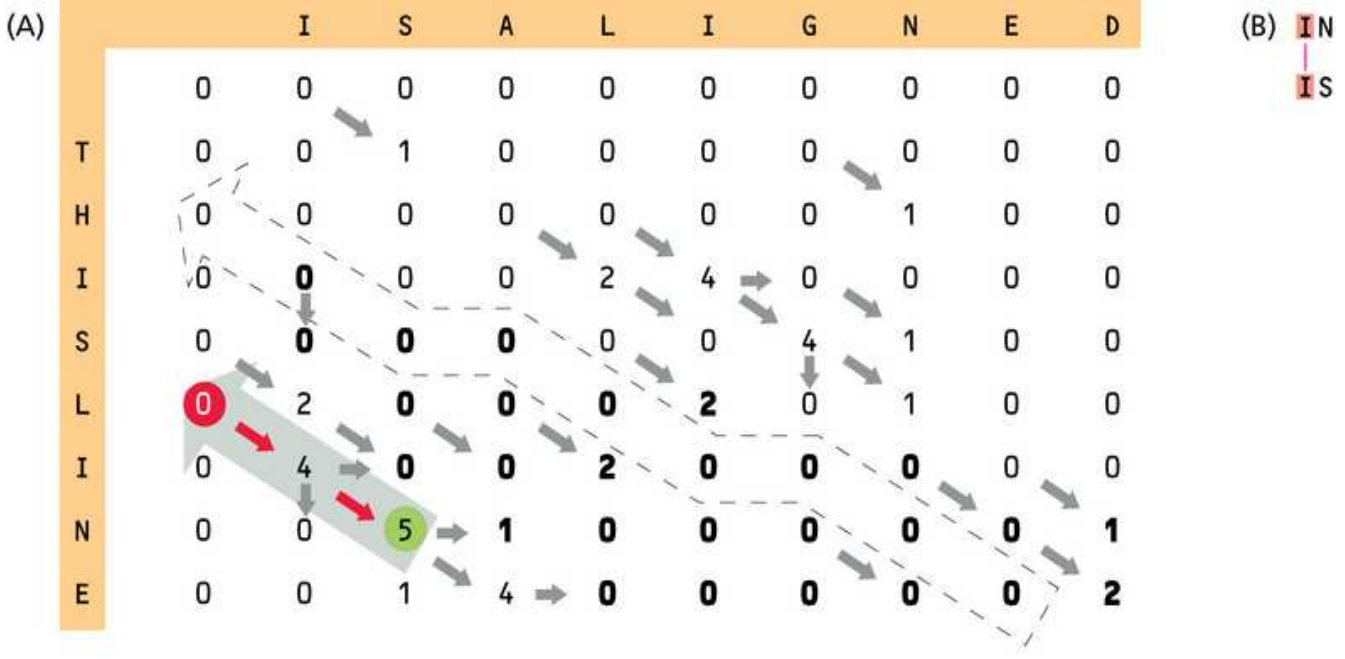
Tabulka pro algoritmus dynamického programování

(A)



(B)

Tabulka pro algoritmus dynamického programování



Bioinformatické databázy

Outline

Dodatek

For Further Reading

X