IV107 Bioinformatika I

Přednáška 4

Katedra informačních technologií
Masarykova Univerzita Brno

Jaro 2023

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

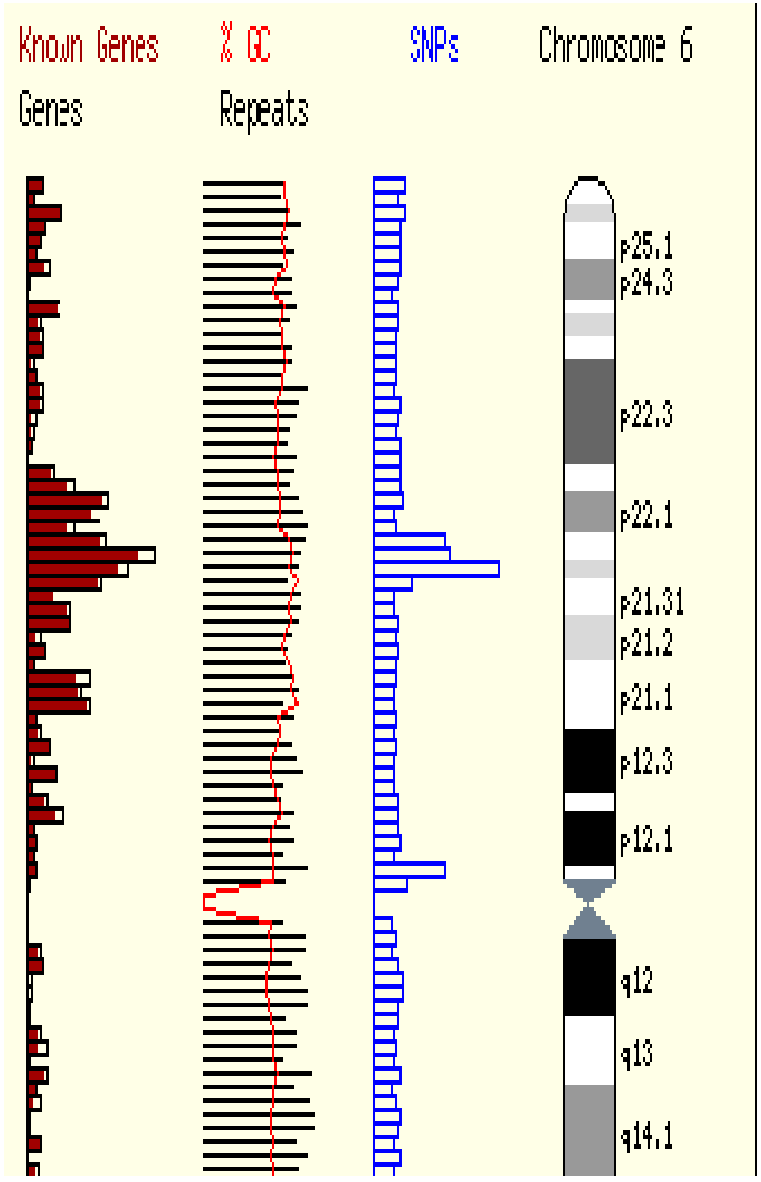
Sekvence DNA

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

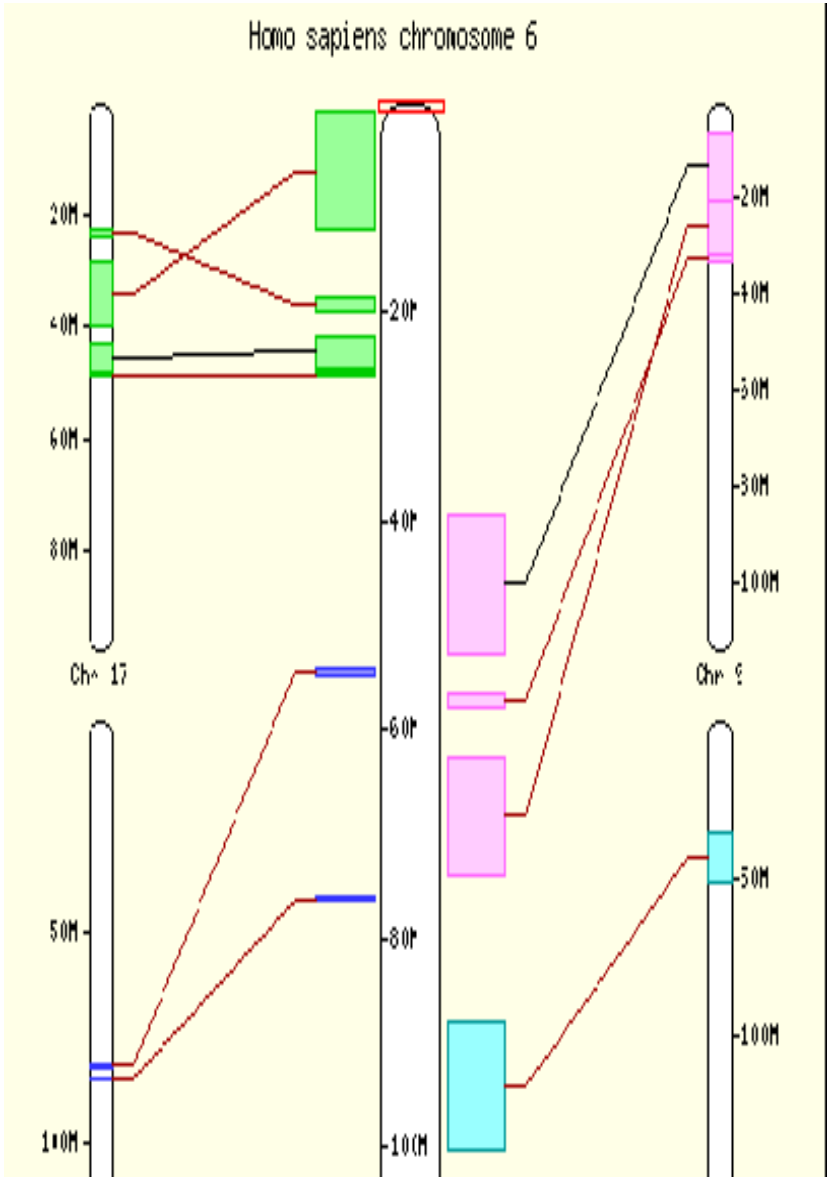
Anotovaná sekvence DNA

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


>P12346 Protein1
MASAQSFYLLDHNQNQNFDDHLAVDIVMILSHERFMN
```

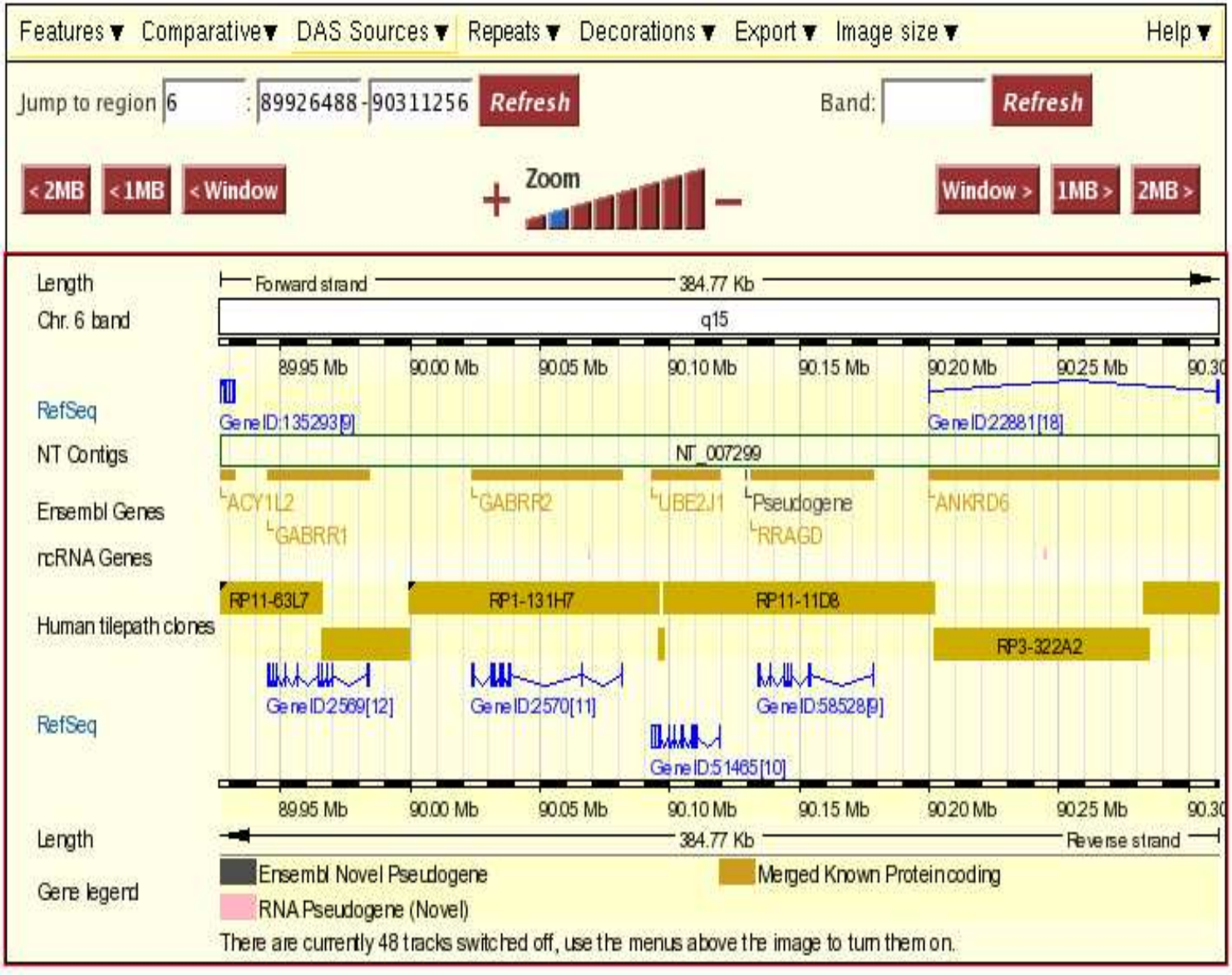
Anotace genomu



Anotace genomu



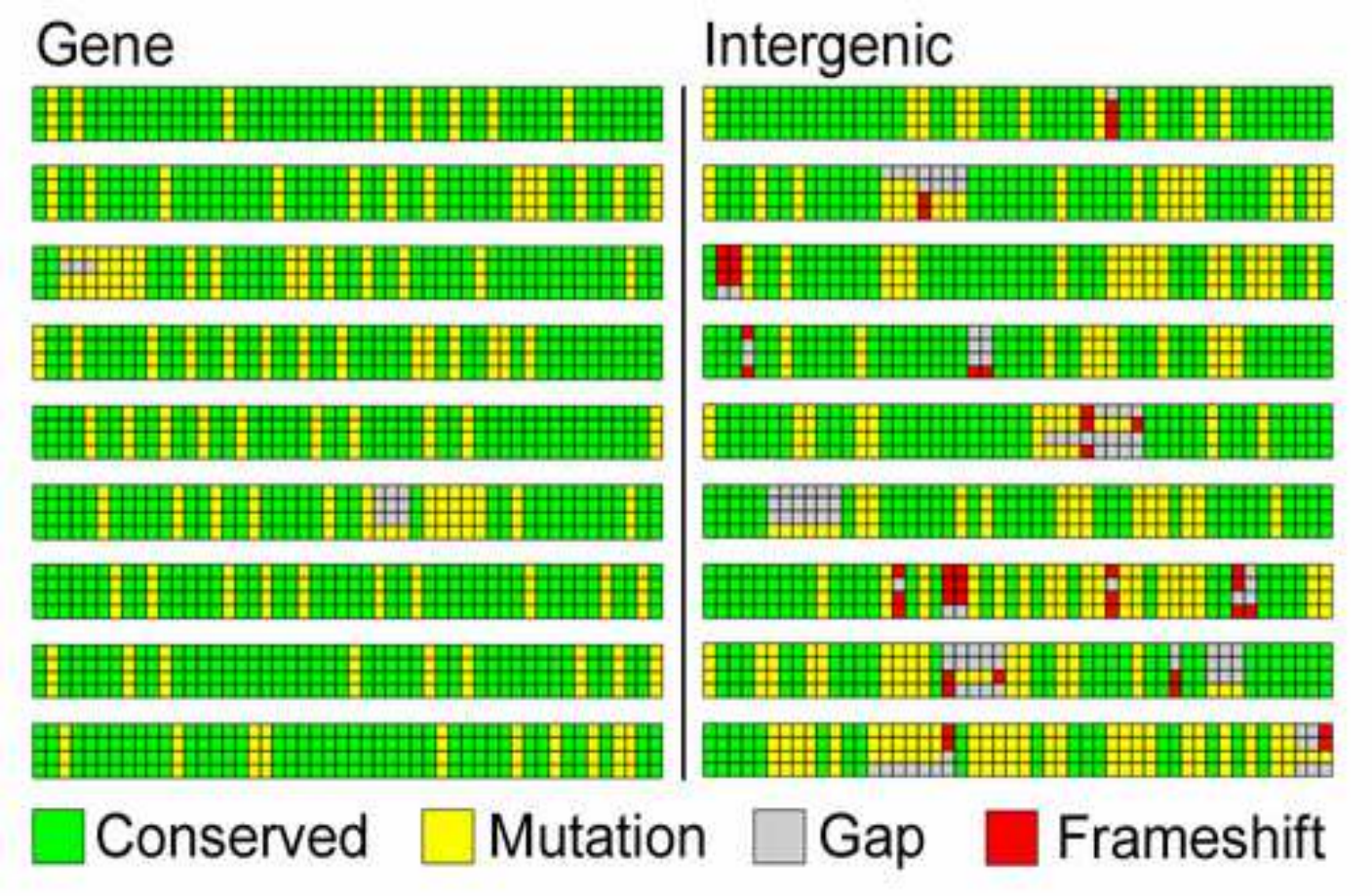
Anotace genomu http://www.ensembl.org/



Způsoby identifikace genů in silico

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

Identifikace genů podle charakteru mutací



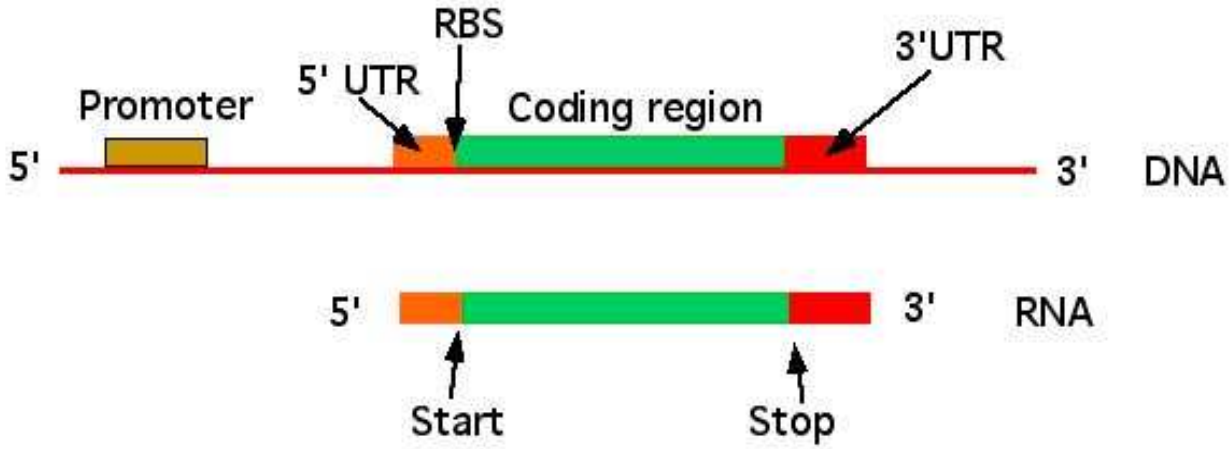
Využití známé struktury genů

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

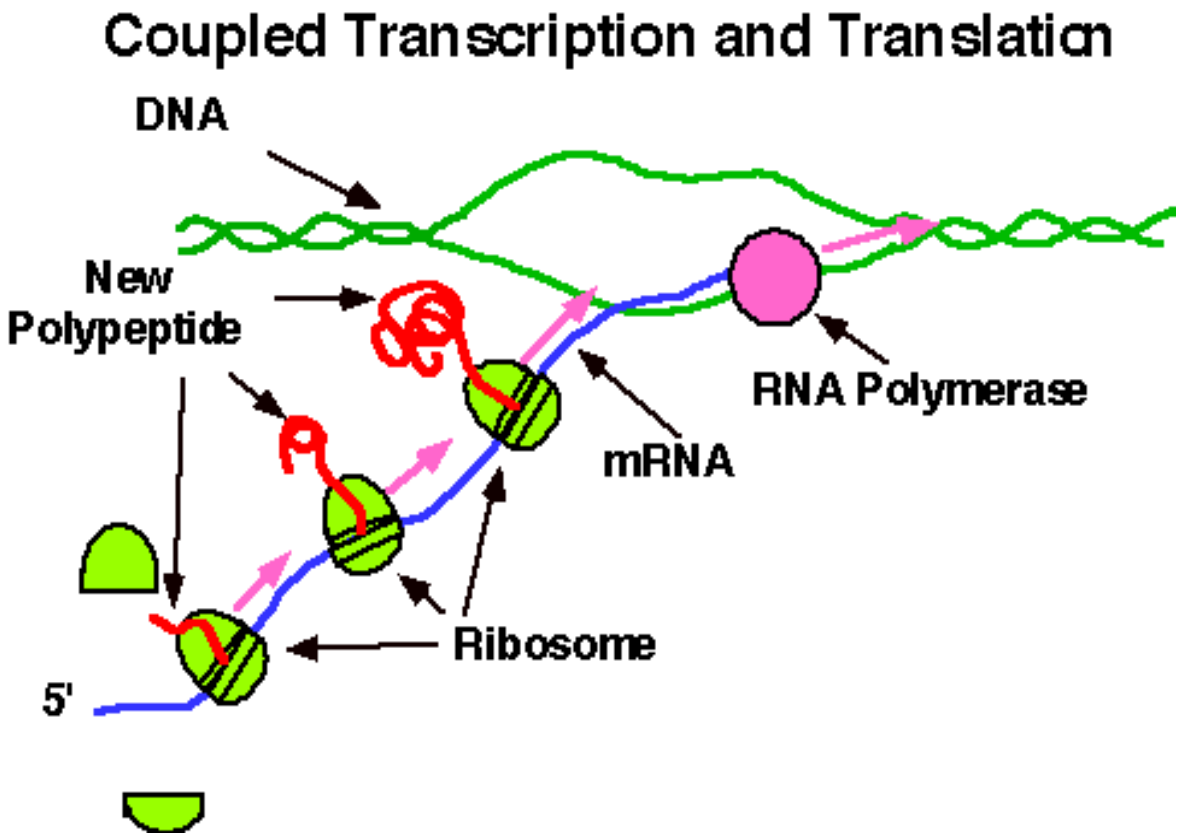
U prokaryotů 95-100% spolehlivost, u složitějších eukaryotů 90% na úrovni bazí, 70% na úrovni exonů/intronů

- 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

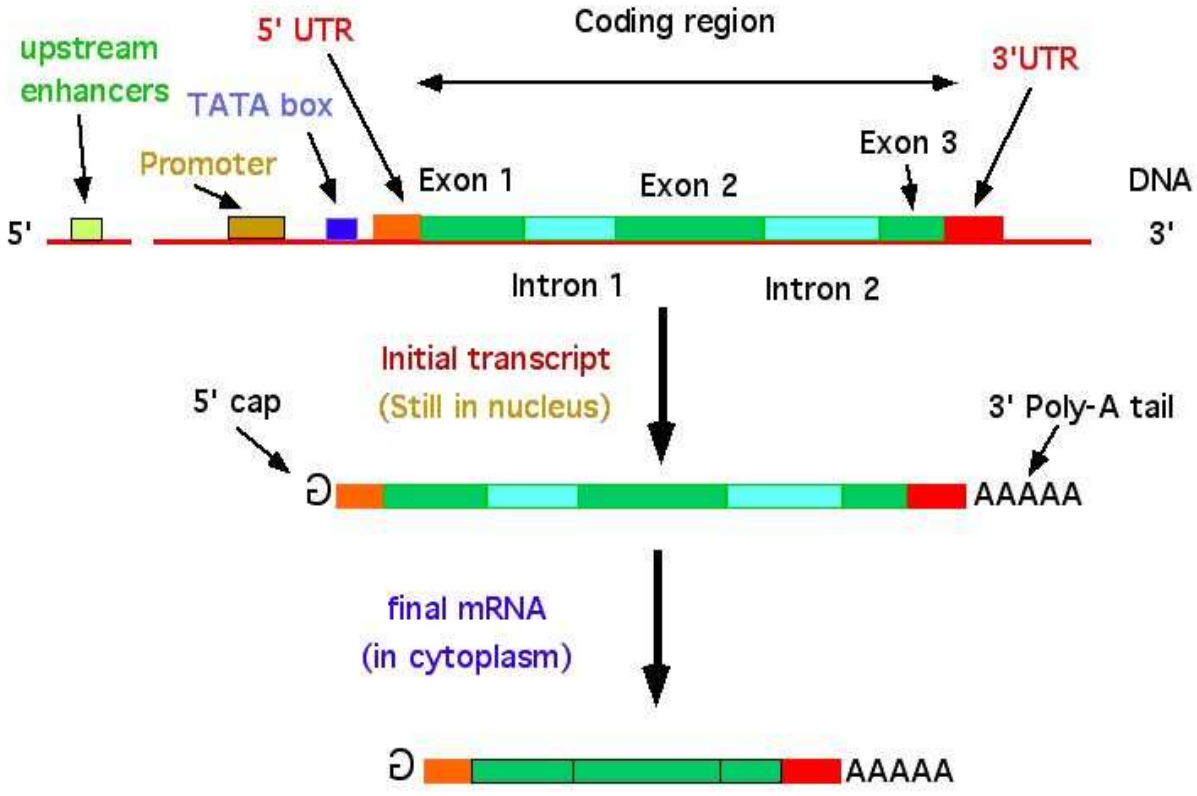
Struktura genu (prokaryotická)



Vztah transkripce a translace u prokaryotů



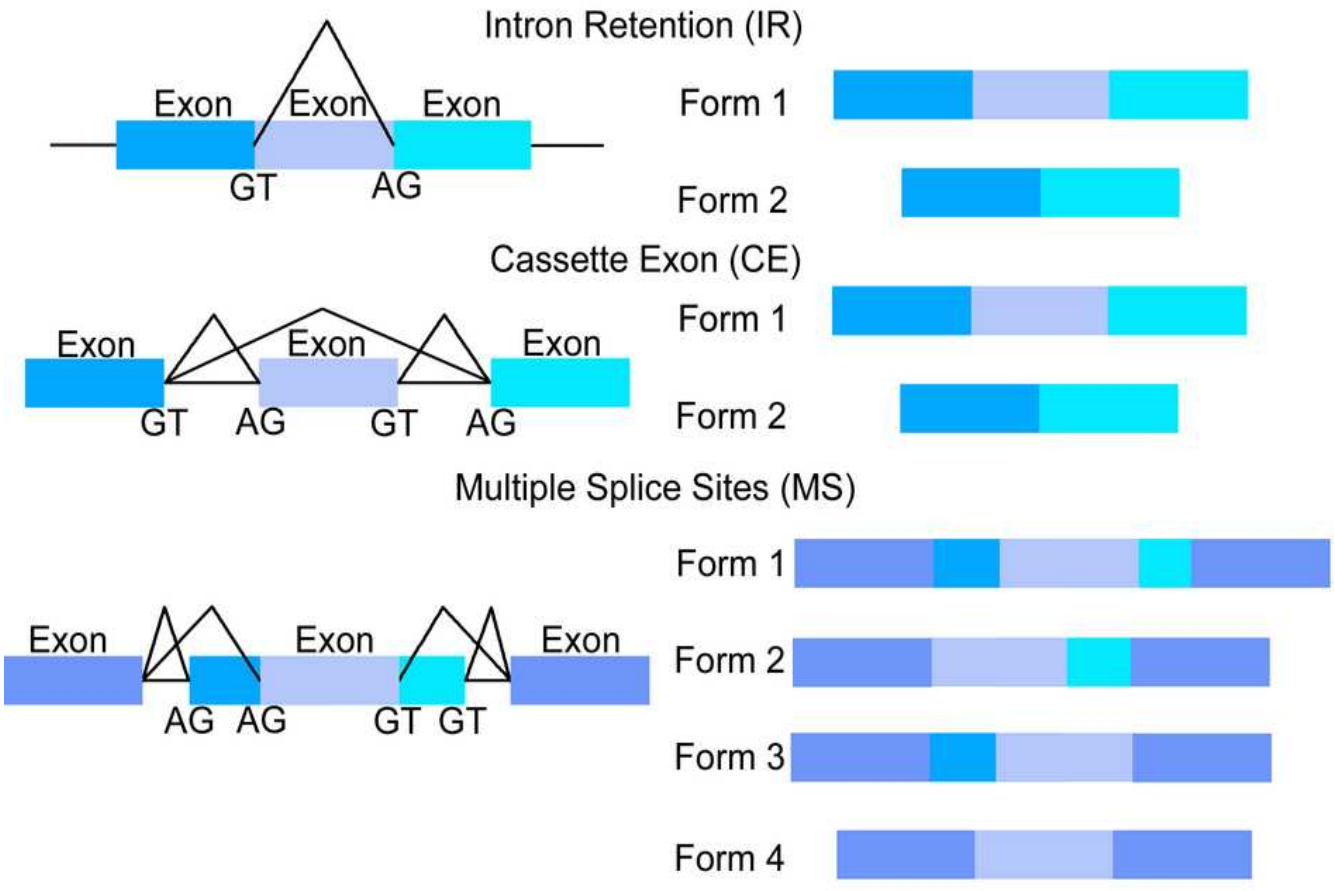
Struktura genu (eukaryotická)



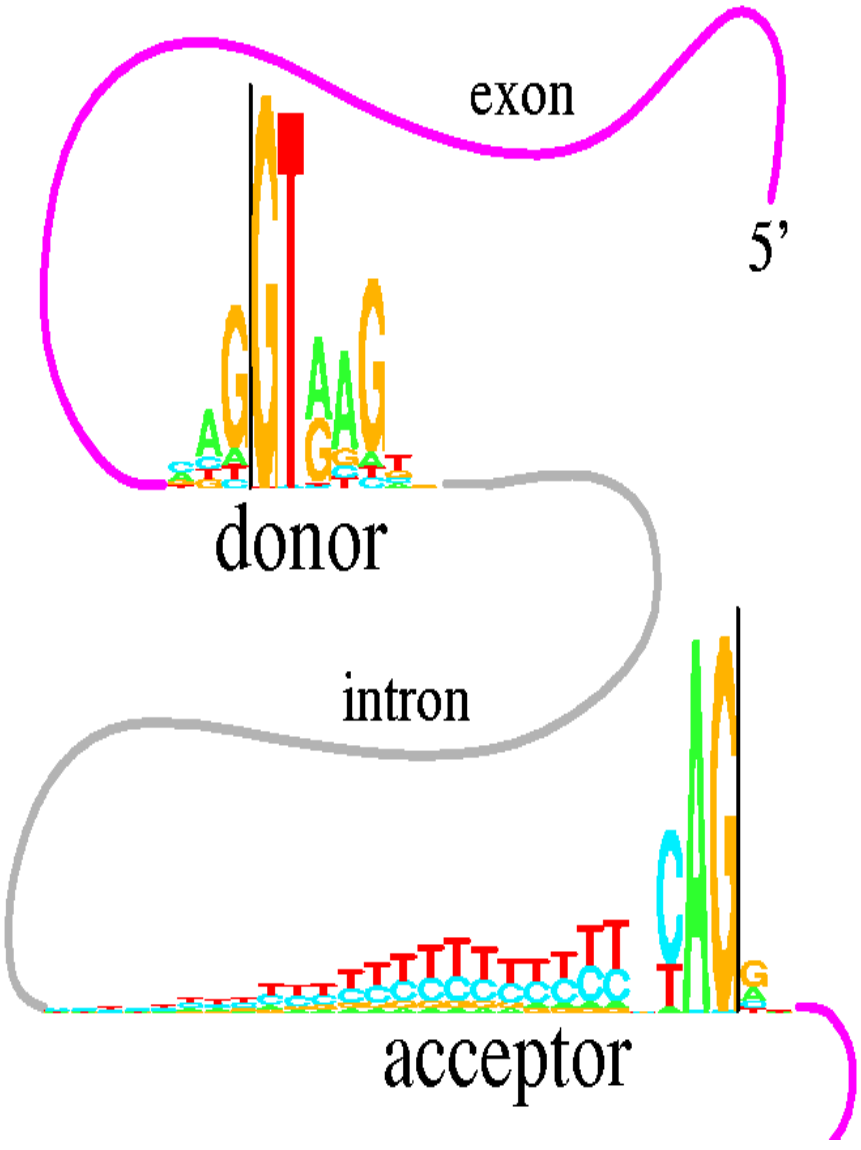
Typické sekvence v eukaryotických genech

- 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)
  - terminátor transkripce

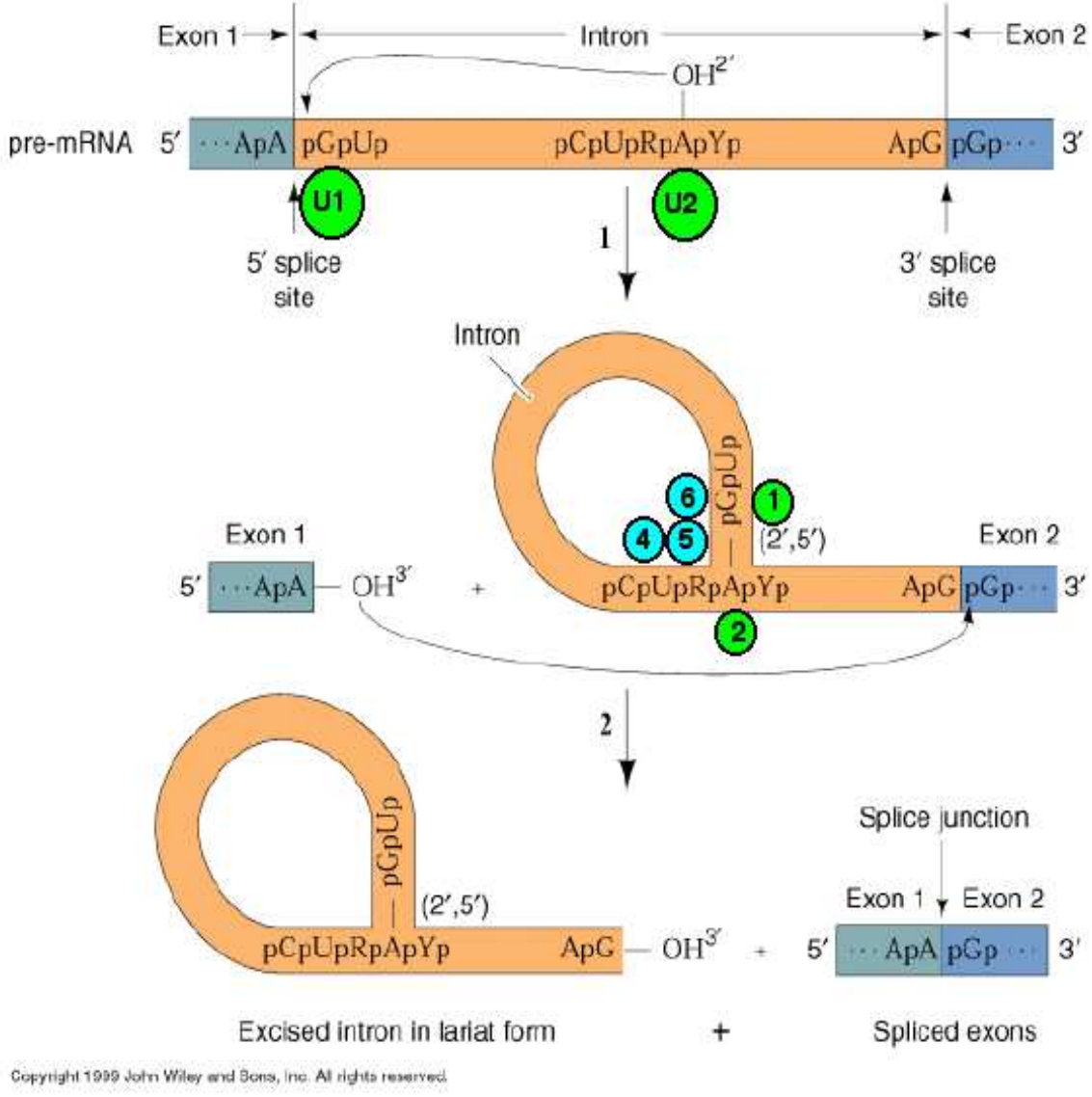
Sestřih mRNA



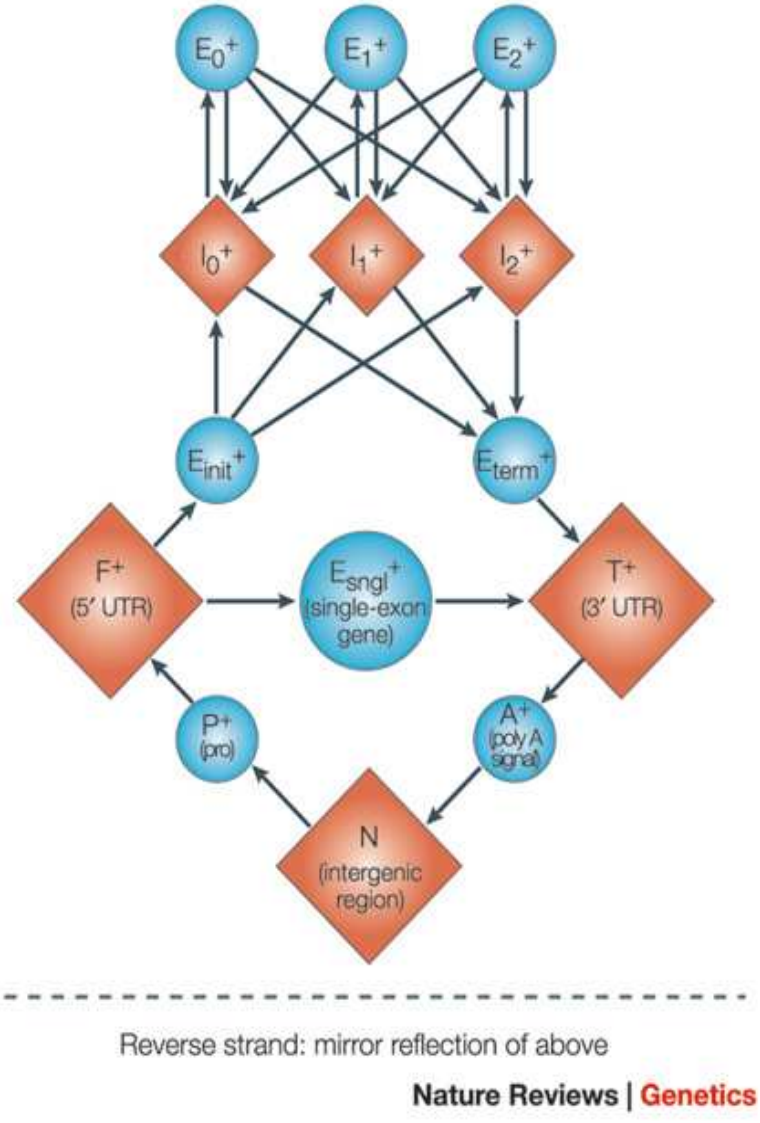
Sekvenční logo intronu



detaily sestřihu



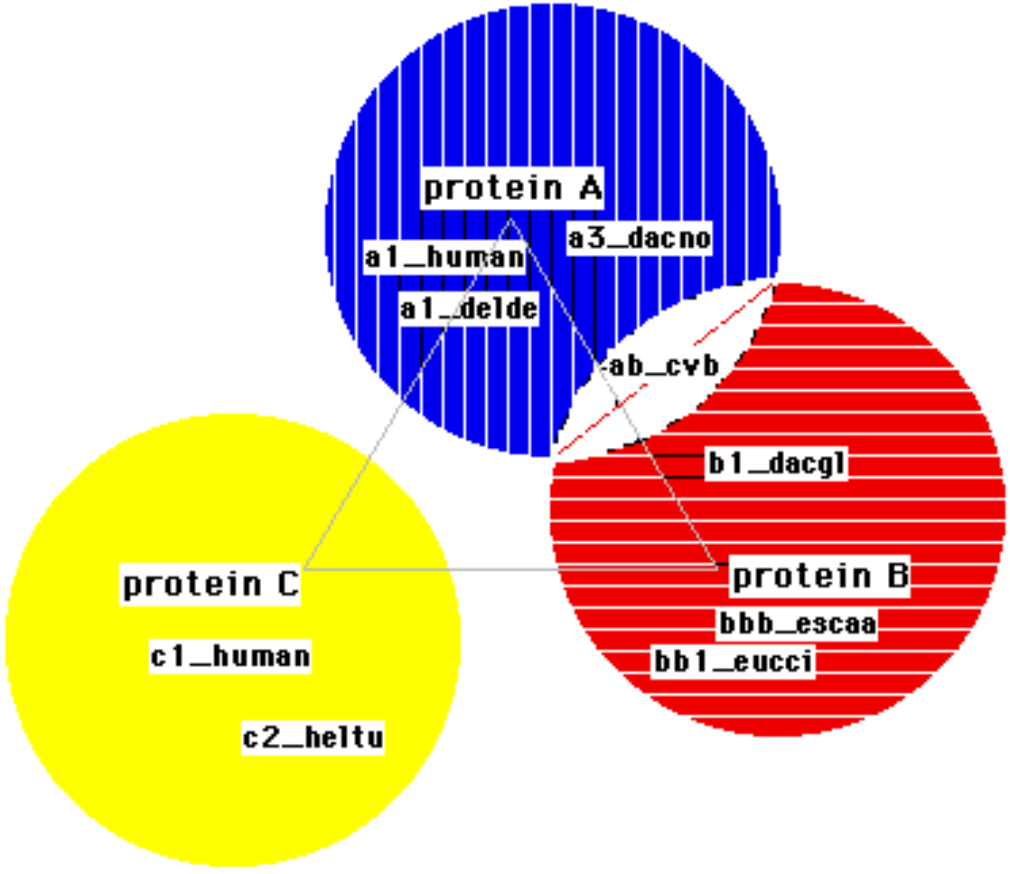
Identifikace genů podle struktury



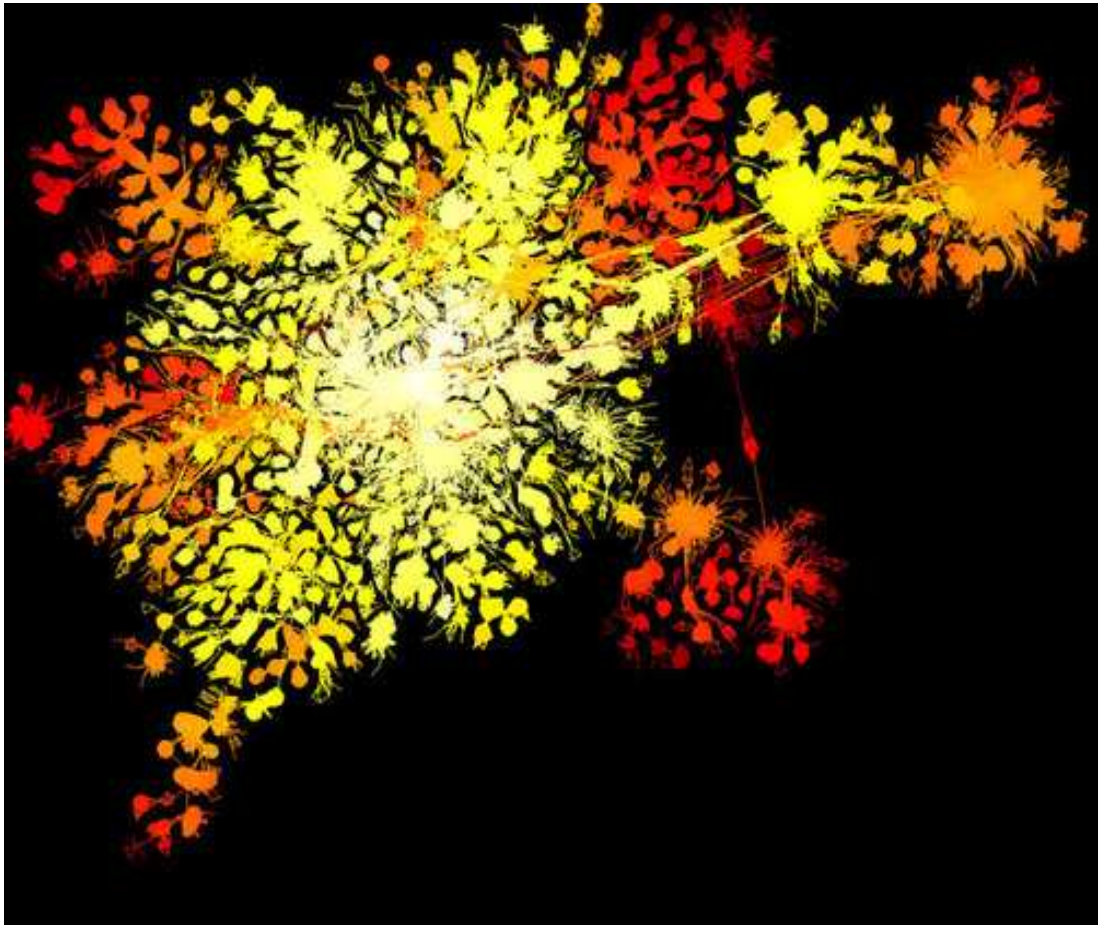
Příbuzné geny mají podobnou funkci i sekvenci

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

Příbuzné geny mají podobnou funkci i sekvenci

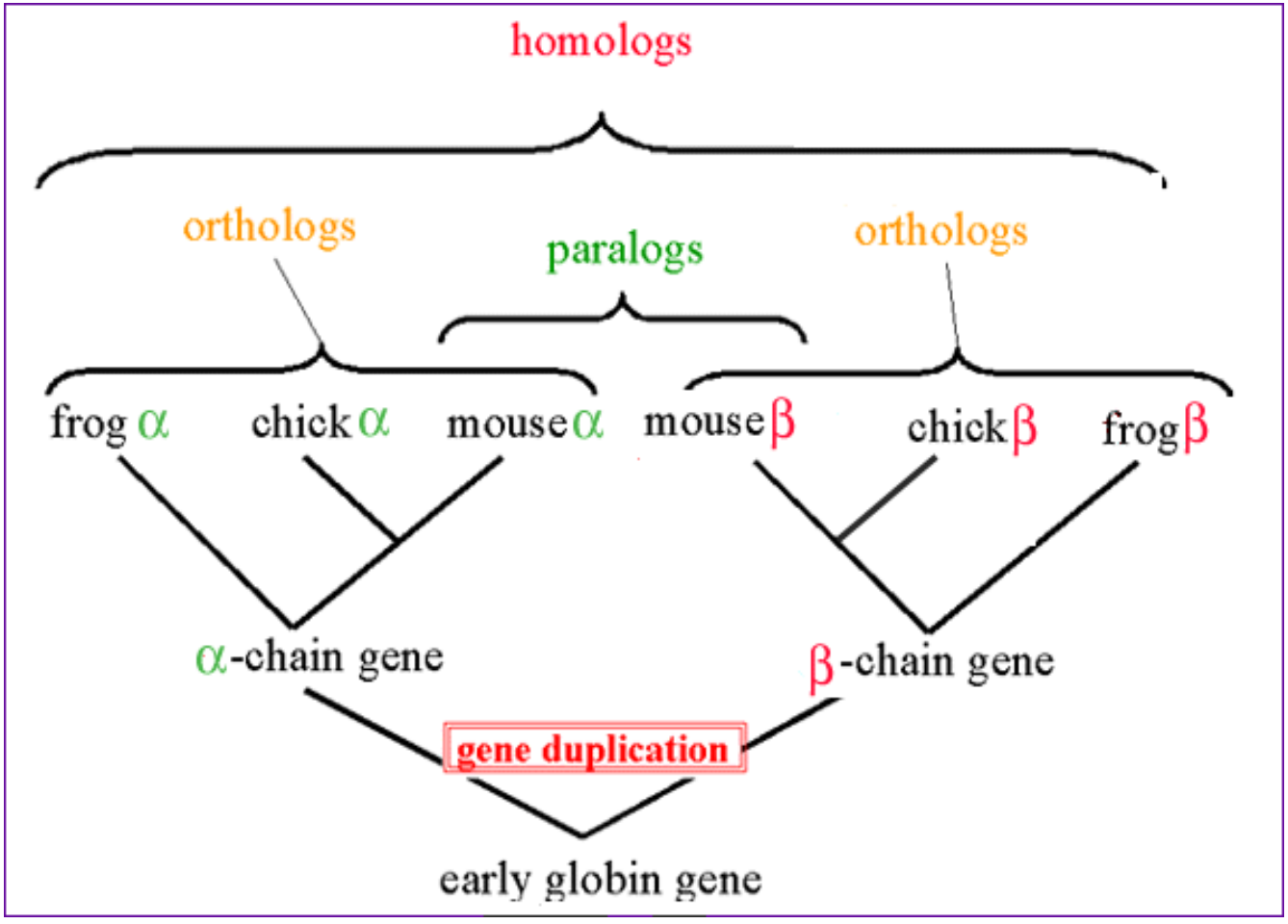


Příbuzné geny mají podobnou funkci i sekvenci



Proteiny přepojené podle sekvnenční podobnosti. Každý z 30727 vrcholů reprezentuje protein, kazdá z 1,206,654 hran podobnost. Seed Magazine, Červenec 2006

homologie


homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene

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.

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

(A) local

```
PI3-kinase  DRHNSNIMVKDDGQLFHIDFG
cAMP PK     DLKPENLLIDQQGYIQVTDFG
```

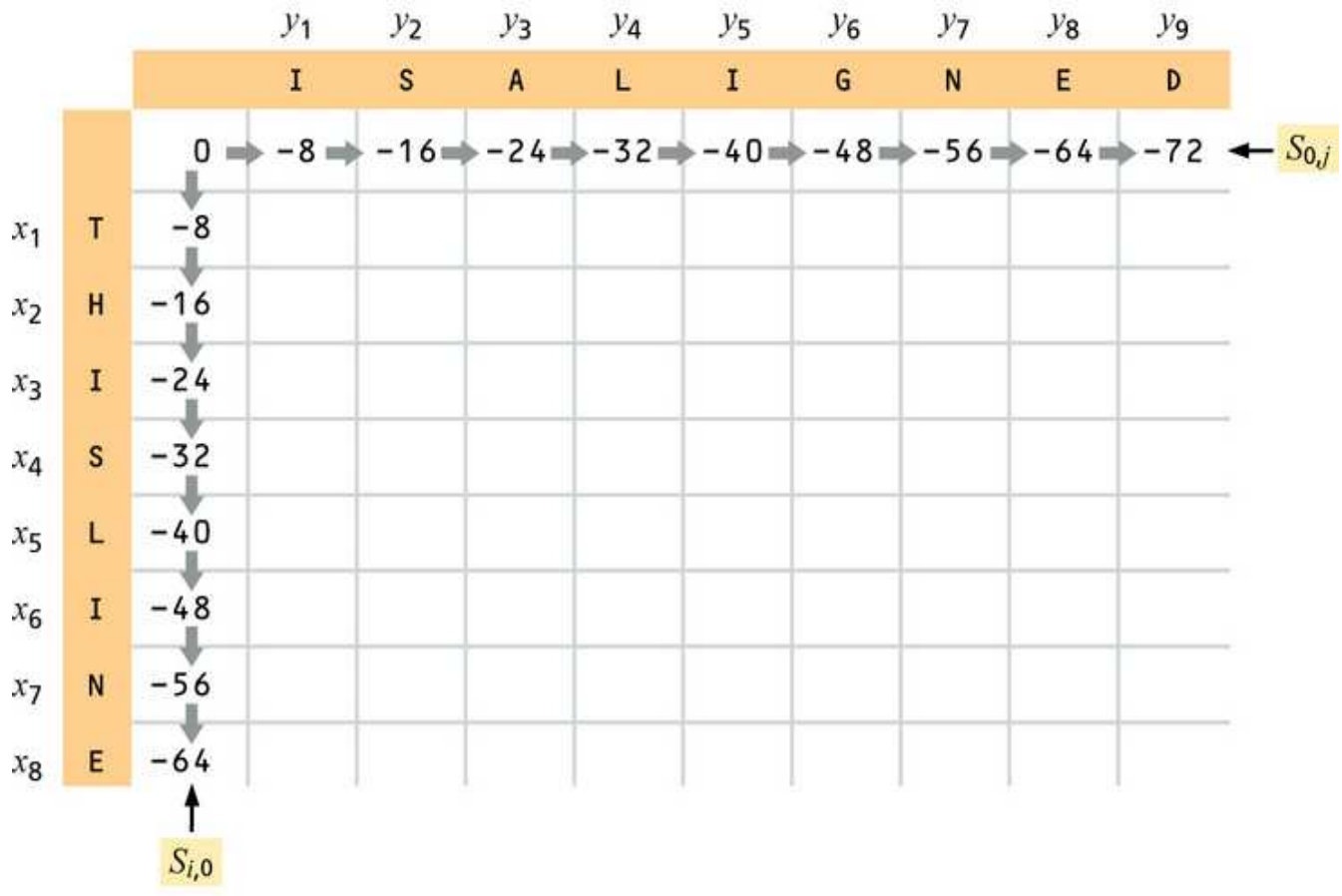
(B) global

```
                    10        20        30        40        50
PI3-kinase  HQLGNLR--LEECRI---MSSAKRPLWLNWENPDIMSELLFQNNEIIFKNGDDLRQDMLT
cAMP PK     GNAAAAKKGXEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVML-
                10        20        30        40        50

                60        70        80        90        100       110
PI3-kinase  LQIIRIME--NIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQ-IQCKGGLKGAL
cAMP PK     ---VKHMETGNHYAMKILDKQKVVK--------LKQIEHTLNEKRILQAVNFPFLVKLEF
               60        70        80                  90        100

                    120       130       140       150       160
PI3-kinase  QFNSHT-LHQWLKDKNKGEIYDAA--IDLFTRSCAGYCVATFILGIGDRHNSNIMVKD-D
cAMP PK     SFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLK
            110       120       130       140       150       160

            170       180       190       200       210       220
PI3-kinase  GQLFHIDFGHFLDHKKKFGYKRERVP-----FVLTQDFL---IVISKGAQECTKTREFE
cAMP PK     PENLLIDQQGYI--QVTDFGFAK-RVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALG
            170       180       190       200       210       220
```

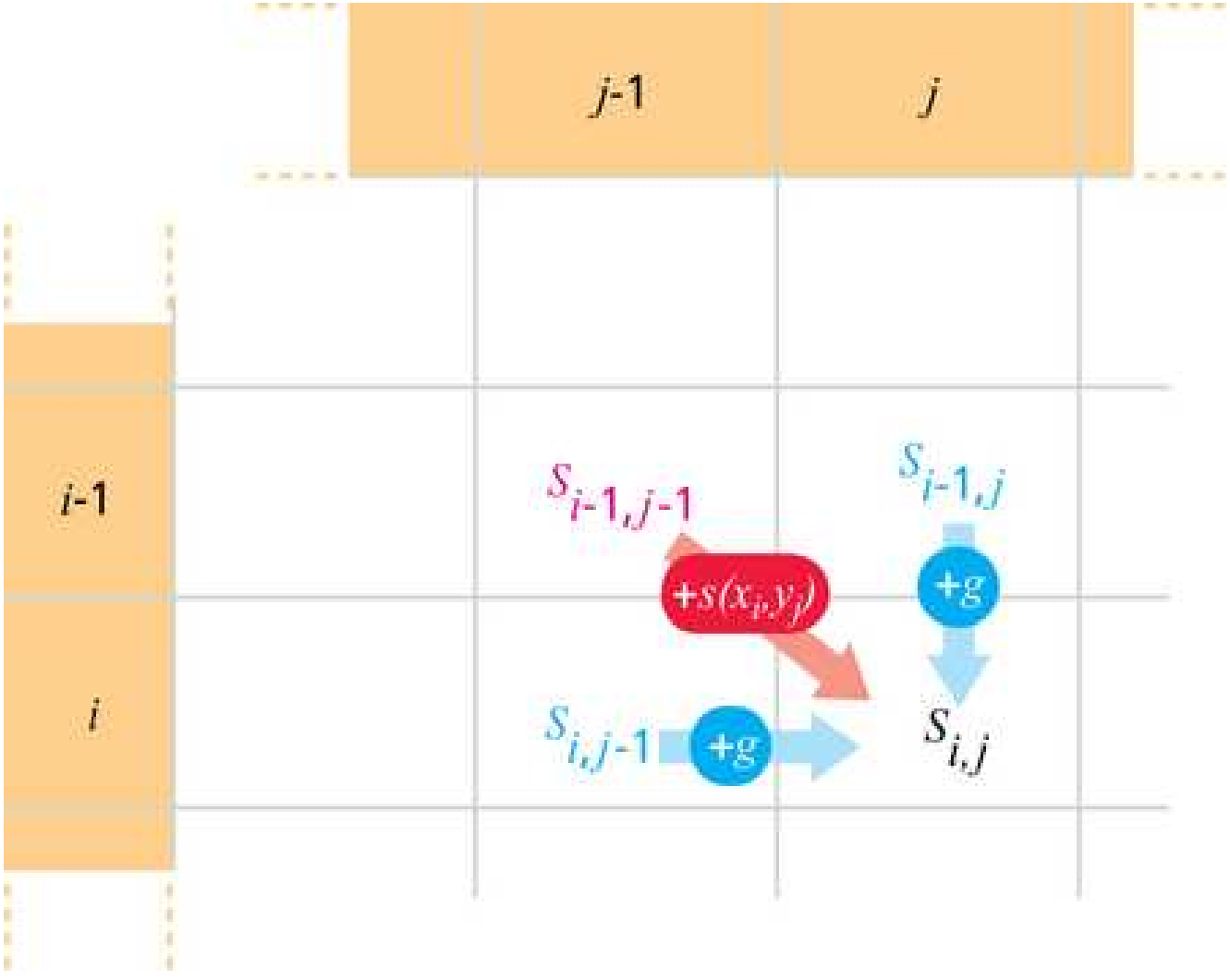
Matice pro hodnocení podobnosti proteinových sekvencí

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
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

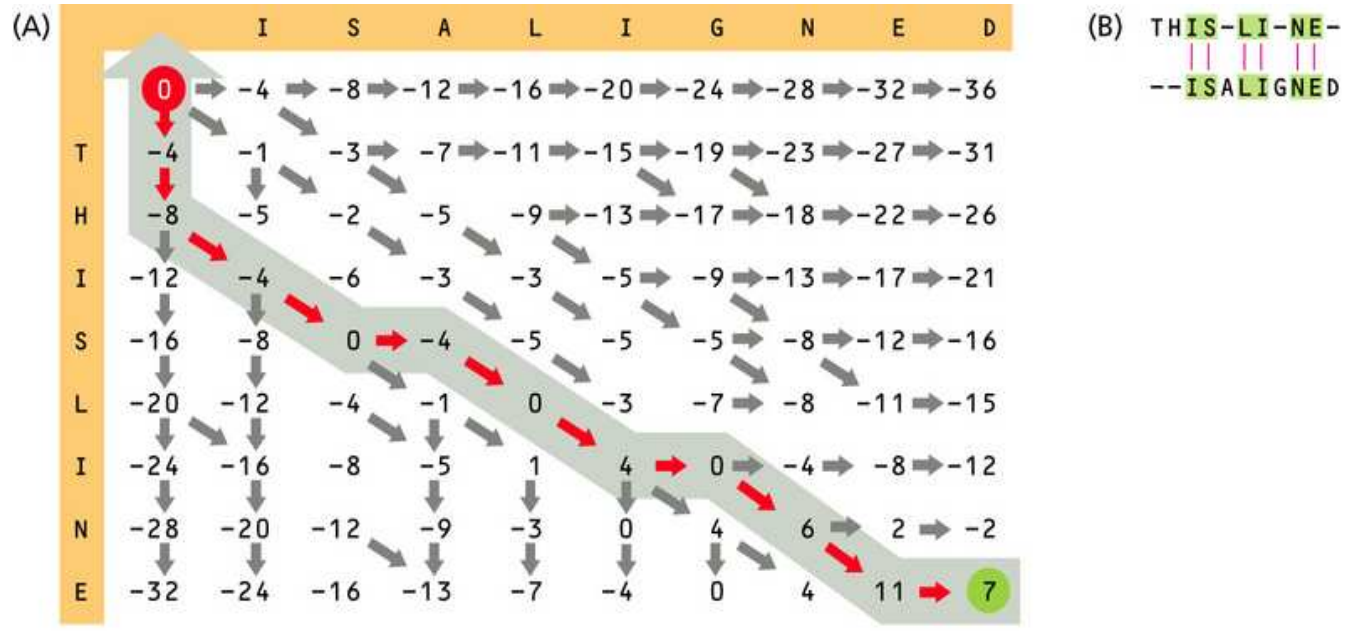
Tabulka pro algoritmus dynamického programování



Tabulka pro algoritmus dynamického programování

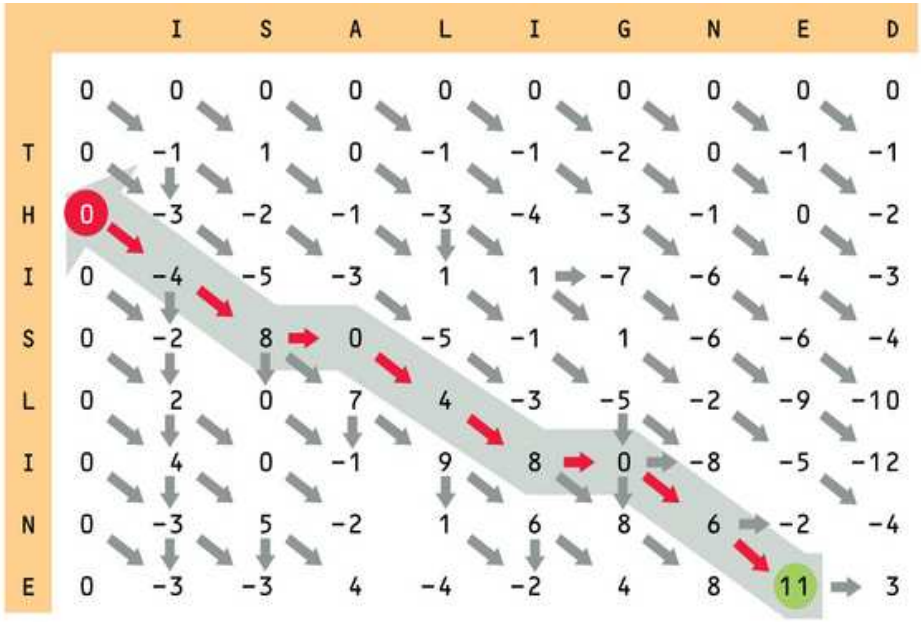


Tabulka pro algoritmus dynamického programování



Tabulka pro algoritmus dynamického programování

(A)

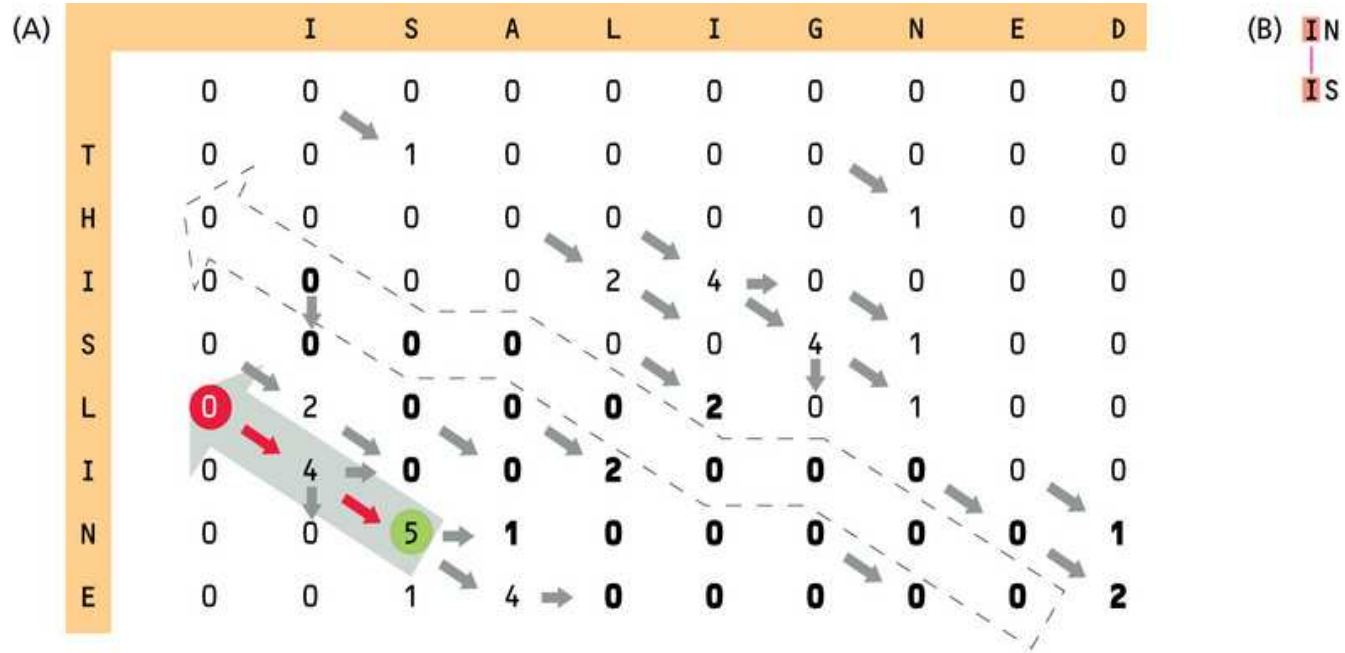


		I	S	A	L	I	G	N	E	D
	0	0	0	0	0	0	0	0	0	0
T	0	-1	1	0	-1	-1	-2	0	-1	-1
H	0	-3	-2	-1	-3	-4	-3	-1	0	-2
I	0	-4	-5	-3	1	1	-7	-6	-4	-3
S	0	-2	8	0	-5	-1	1	-6	-6	-4
L	0	2	0	7	4	-3	-5	-2	-9	-10
I	0	4	0	-1	9	8	0	-8	-5	-12
N	0	-3	5	-2	1	6	8	6	-2	-4
E	0	-3	-3	4	-4	-2	4	8	11	3

(B)

```
THIS-LI-NE-
  ||  ||  ||
--ISALIGNED
```

Tabulka pro algoritmus dynamického programování



Příště

Bioinformatické databázy

Outline

For Further Reading

X