

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

Bioinformatické databázy

Sequence 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

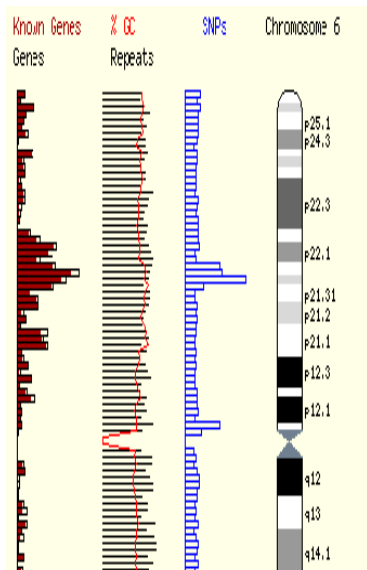
stop TAA

>P12346 Protein1

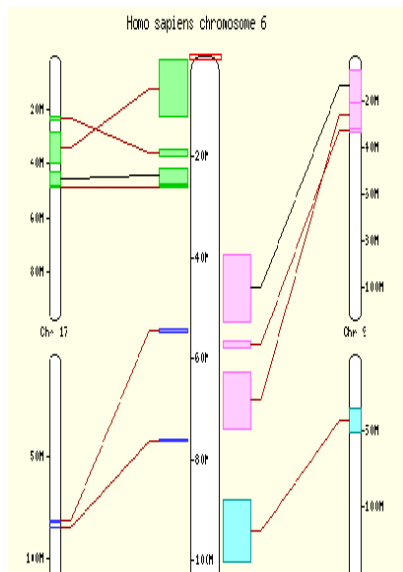
MASAQSFYLLDHNQNQNFDDHLAVDIVMILSHERFMN



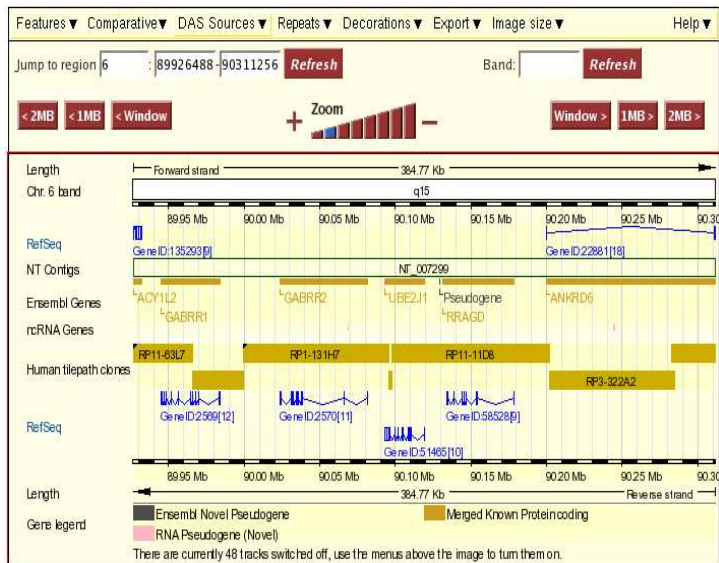
Anotace genomu



Anotace genomu



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

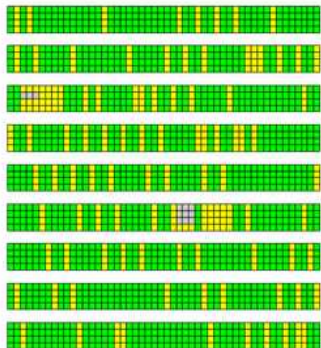


Způsoby identifikace genů in silico

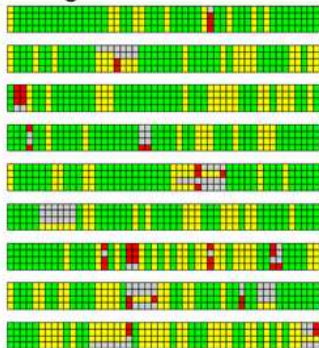
- ▶ Experimentální metody (cDNA, EST)
- ▶ Komparativní metody
 - ▶ Selekční tlak
 - ▶ Druh zachovaných mutací
- ▶ Strukturní metody (GeneMark, GeneScan, GeneID)
- ▶ Detekce charakteristických signálů

Identifikace genů podle charakteru mutací

Gene



Intergenic



Conserved Mutation Gap Frameshift

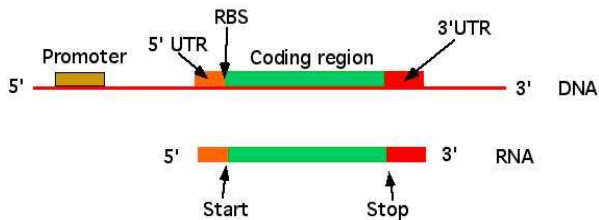
Využití známé struktury genů

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

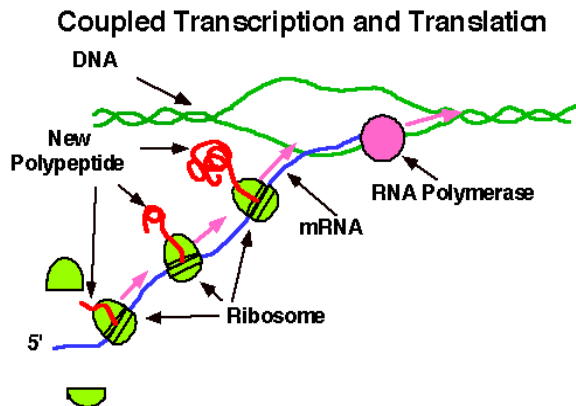
U prokaryotů 95-100% spolehlivost, u složitějších eukaryotů 90% na úrovni bazí, 70% na úrovni exonů/intronů

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

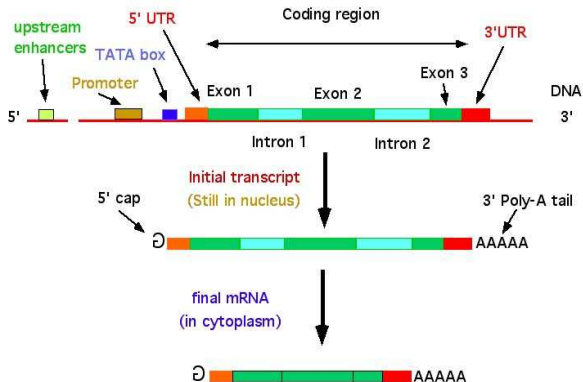
Struktura genu (prokaryotická)



Vztah transkripce a translace u prokaryotů



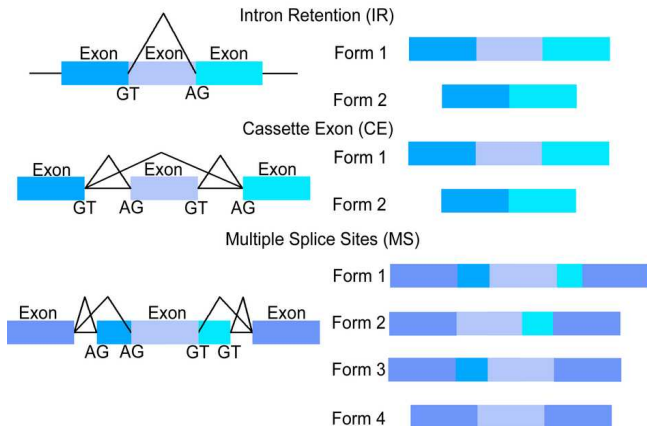
Struktura genu (eukaryotická)



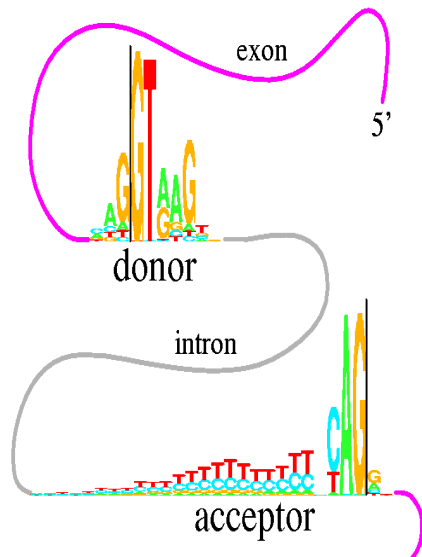
Typické sekvence v eukaryotických genech

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

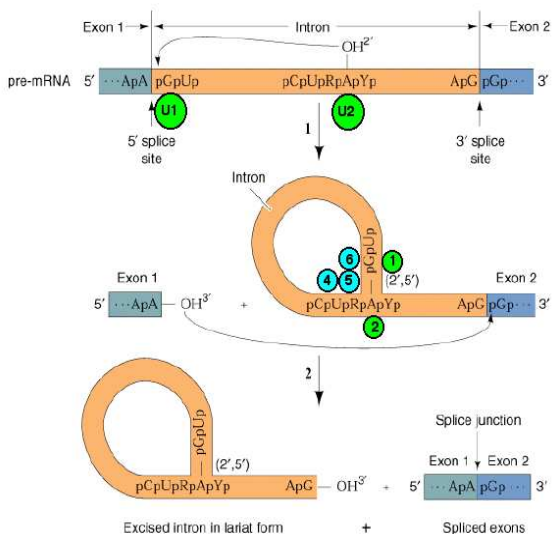
Sestřih mRNA



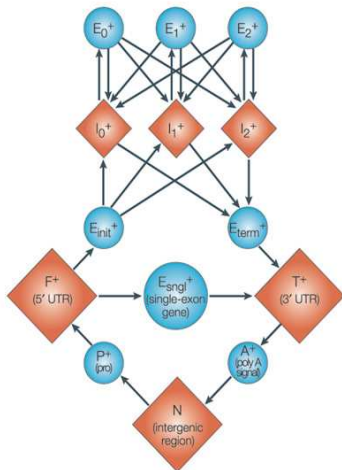
Sekvenční logo intronu



detaily sestřihu



Identifikace genů podle struktury

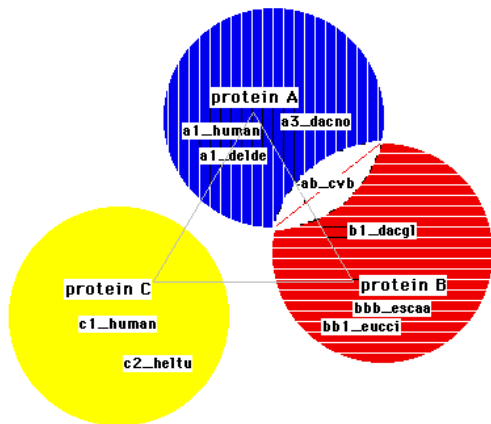


Reverse strand: mirror reflection of above

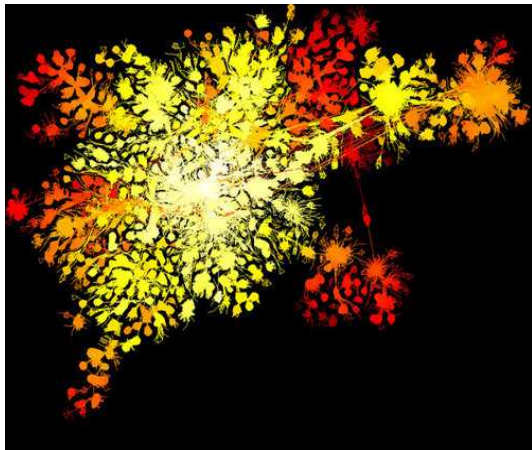
Příbuzné geny mají podobnou funkci i sekvenci

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

Rost - prostor podobnosti

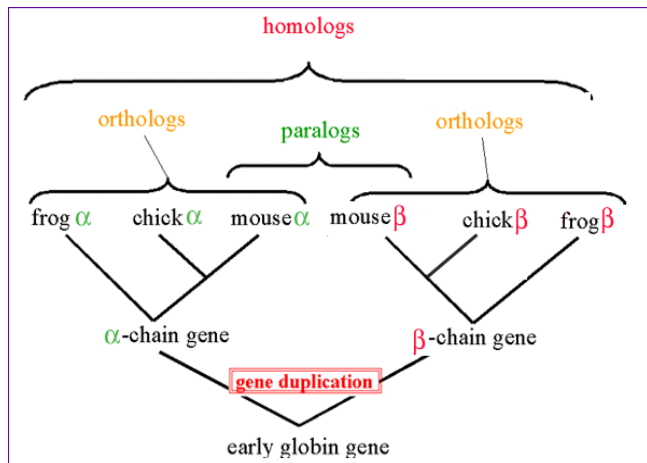


Síť proteinů podle sekvenční podobnosti



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

homologie

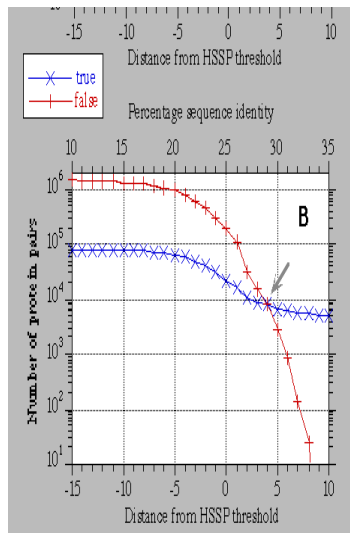


Příbuznost a podobnost sekvencí

- ▶ 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 porovnáním

(A) local

PI3-kinase DRHNSNIMVKDDGQLFHI DFG

cAMP PK DLKPENLLIDQQGYIQVT DFG

(B) global

```
PI3-kinase  HQLGNLR--LEECRI---MSSAKRPLWLNWENPDIMSELFQNNEIFKNGDDLRRQDMLT
cAMP PK      GNAAAAKKGXEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVML-
```

```
PI3-kinase  LQIIRIME--NIWQNGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQ-IQCKGGLK GAL
cAMP PK      ---VKHMETGNHYAMKILDKQKVVK-----LKQIEHTLNEKRILQAVNFPFLVKLEF
```

```
PI3-kinase  QFNSHT-LHQWLKDKNGEIIDAA--IDLFRSCAGYCVATFILGIGDRHNSNIMVKD-D
cAMP PK      SFKDNSNLYMVMVEYVPGGEMFSLRRITGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLK
```

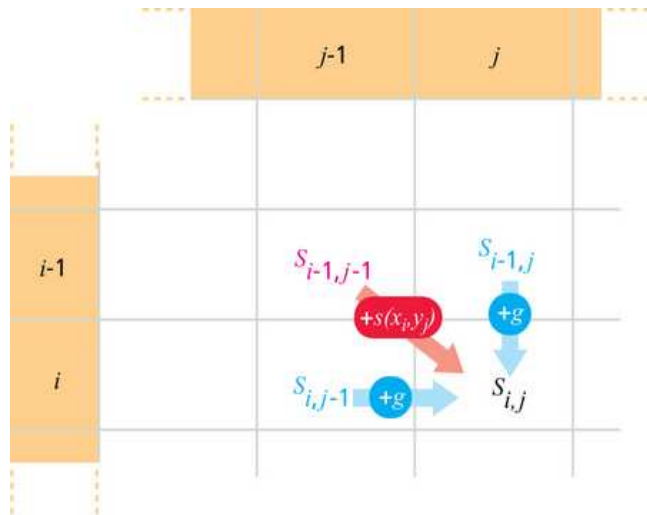
```
PI3-kinase  GQLFHI DFGHFLDHKKKFGYKRERVP-----FVLTQDFL---IVISKAQECTKTREFE
cAMP PK      PENLLIDQQGYI--QVTD FGF AKA-RVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALG
```


Tabulka pro algoritmus dynamického programování

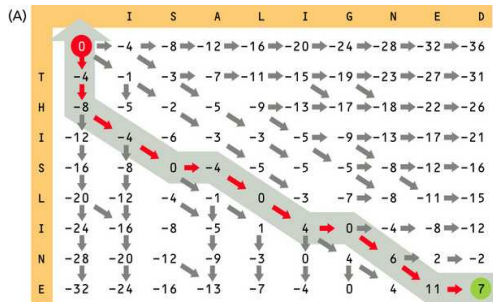
	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	
	I	S	A	L	I	G	N	E	D	
x_0	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
x_1	T	-8								
x_2	H	-16								
x_3	I	-24								
x_4	S	-32								
x_5	L	-40								
x_6	I	-48								
x_7	N	-56								
x_8	E	-64								

Annotations: $S_{0,j}$ points to the top row of values. $S_{i,0}$ points to the first column of values.

Tabulka pro algoritmus dynamického programování

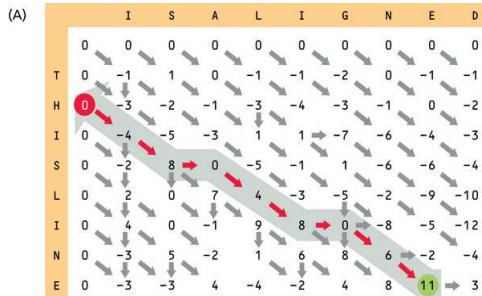


Tabulka pro algoritmus dynamického programování



(B) THIS-LI-NE-
--ISALIGNED

Tabulka pro algoritmus dynamického programování



(B) T H I S - L I - N E -
 | | | | | | | |
 - - I S A L I G N E D

Tabulka pro algoritmus dynamického programování

(A)

	I	S	A	L	I	G	N	E	D
T	0	0	0	0	0	0	0	0	0
H	0	0	1	0	0	0	0	0	0
I	0	0	0	0	2	4	0	0	0
S	0	0	0	0	0	0	4	1	0
L	0	2	0	0	0	2	0	1	0
I	0	4	0	0	2	0	0	0	0
N	0	0	5	1	0	0	0	0	1
E	0	0	1	4	0	0	0	0	2

(B) I N
I S

Příště

Bioinformatické databáze



Outline

Dodatek

For Further Reading

X

