

IV107 Bioinformatika I

Přednáška 4

Katedra informačních technologií
Masarykova Univerzita Brno

Jaro 2008

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

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

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Outline

Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Bioinformatické databázy

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

Anotace sekvence a genů

[Anotace genomu](#)
[Identifikace genů](#)
[Homologie a podobnost](#)

Příště

[Bioinformatické databázy](#)

Anotovaná sekvence DNA

>P12345 Gen1 - protein

alkoholdehydrogenáza

TATA TATAAA

CGATTGACGATGACGAT

start ATG

exon1 TACAGATTACAGATTACAGATTAAGATGT

intron1 CAGATTACAGATTACAGATTACACAGATTCA

exon2 AGATTACAGATTACAGATTACAGA

stop TAA

>P12346 Protein1

MASAQSFYLLDHNQNQNFDDHLAVDIVMILSHERFMN

Anotace sekvence a
genů

Anotace genomu

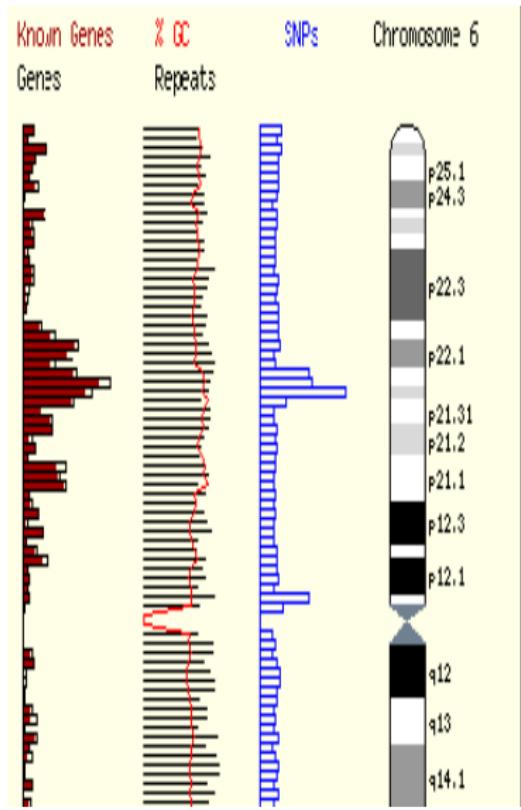
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Anotace genomu



Anotace sekvence a
genů

Anotace genomu

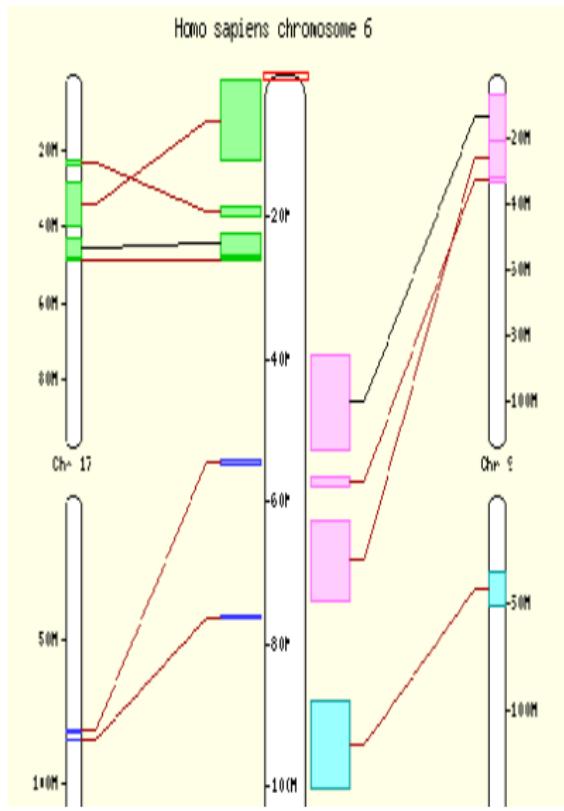
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Anotace genomu



Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

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

Features ▾ Comparative ▾ DAS Sources ▾ Repeats ▾ Decorations ▾ Export ▾ Image size ▾ Help ▾

Jump to region 6 : 89926488 - 90311256 Refresh Band: Refresh

< 2MB < 1MB < Window + Zoom - Window > 1MB > 2MB >

Length Chr. 6 band Forward strand 384.77 Kb q15

Chr. 6 band RefSeq NT Contigs Ensembl Genes rRNA Genes Human tilepath clones RefSeq Length Gene legend

89.95 Mb 90.00 Mb 90.05 Mb 90.10 Mb 90.15 Mb 90.20 Mb 90.25 Mb 90.30

Gene ID:135293[9] Gene ID:22881[18]

NT_007299

ACY1L2 GABRR1 GABRR2 UBE2J1 Pseudogene RRGD ANKRD6

RP11-63L7 RP1-131H7 RP11-11D8 RP3-322A2

Gene ID:2589[12] Gene ID:2570[11] Gene ID:58528[9] Gene ID:51465[10]

89.95 Mb 90.00 Mb 90.05 Mb 90.10 Mb 90.15 Mb 90.20 Mb 90.25 Mb 90.30

384.77 Kb Reverse strand

[Ensembl Novel Pseudogene] [Merged Known Protein coding]
 [RNA Pseudogene (Novel)]

There are currently 48 tracks switched off, use the menus above the image to turn them on.

Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Způsoby identifikace genů in silico

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

Anotace sekvence a
genů

Anotace genomu

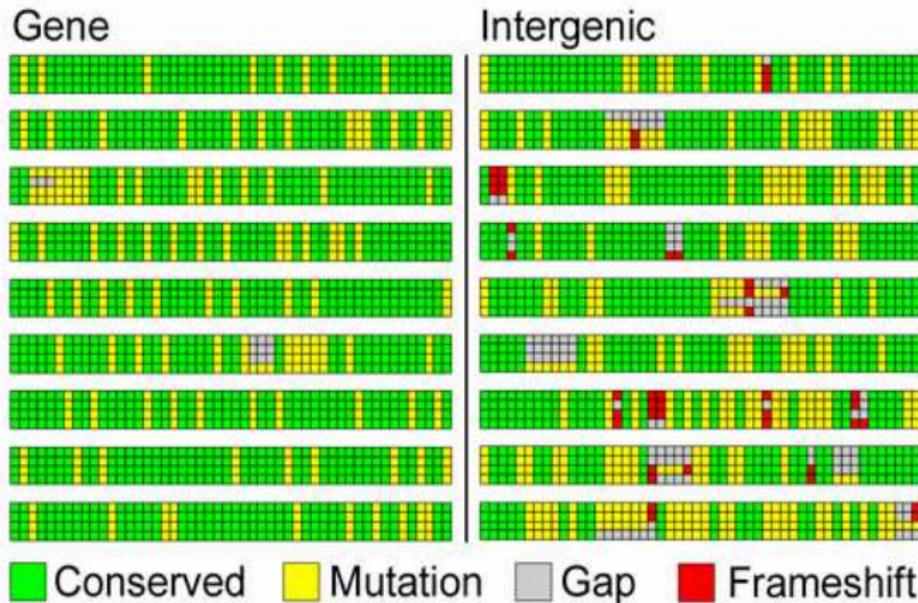
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Identifikace genů podle charakteru mutací



Využití známé struktury genů

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

Anotace sekvence a
genů

Anotace genomu

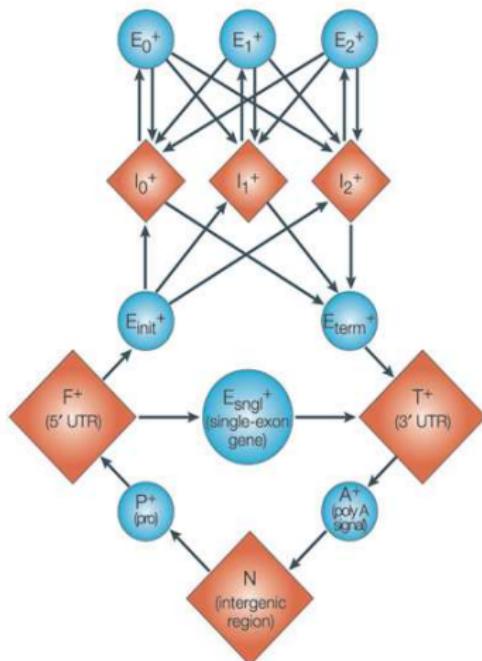
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Identifikace genů podle struktury



Reverse strand: mirror reflection of above

Nature Reviews | Genetics

Anotace sekvence a
genů

Anotace genomu

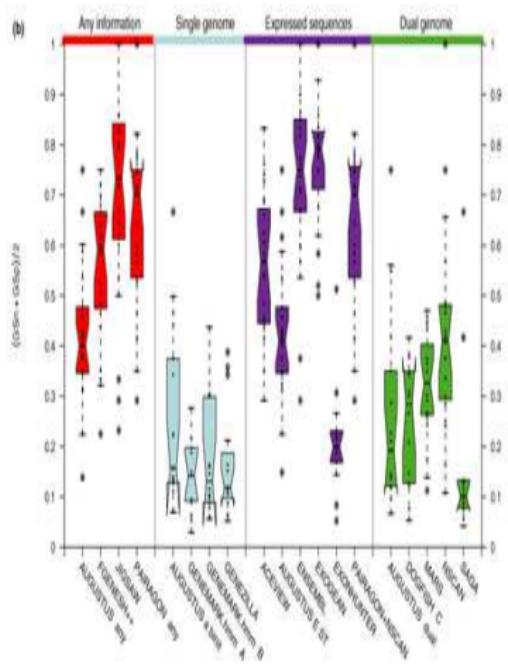
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Úspěšnost identifikace genů



Guigo et al, Genome Biology (2006)

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

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

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Struktura genu (prokaryotická)

Anotace sekvence a genů

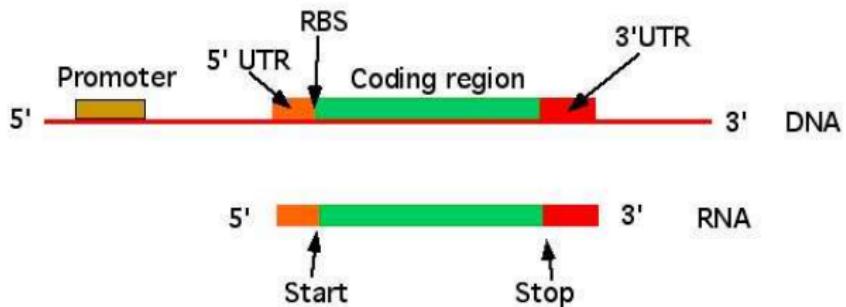
Anotace genomu

Identifikace genů

Homologie a podobnost

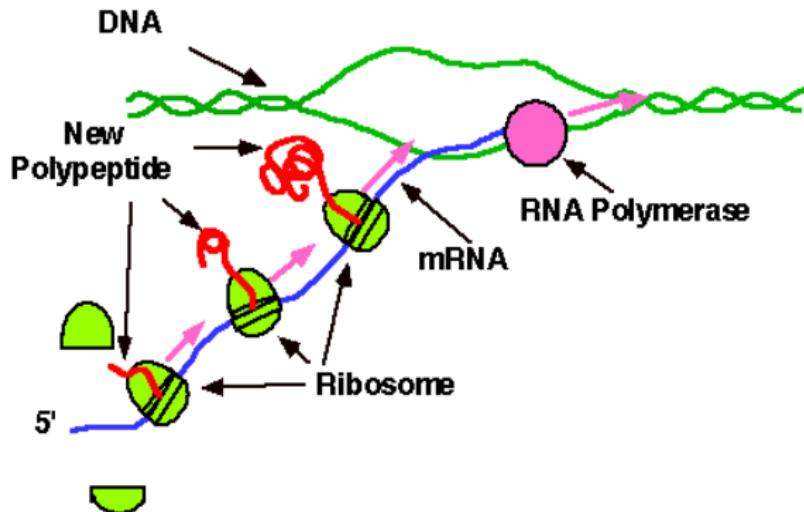
Příště

Bioinformatické databázy

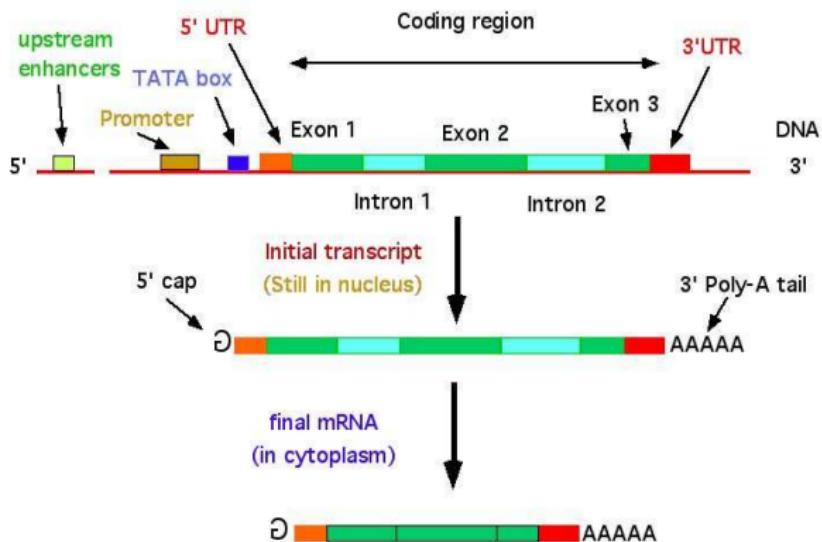


Vztah transkripce a translace u prokaryotů

Coupled Transcription and Translation



Struktura genu (eukaryotická)



Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

- ▶ 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
 - ▶ polyadenylační signál (AATAAA)
 - ▶ terminátor transkripce

Anotace sekvence a genů

Anotace genomu

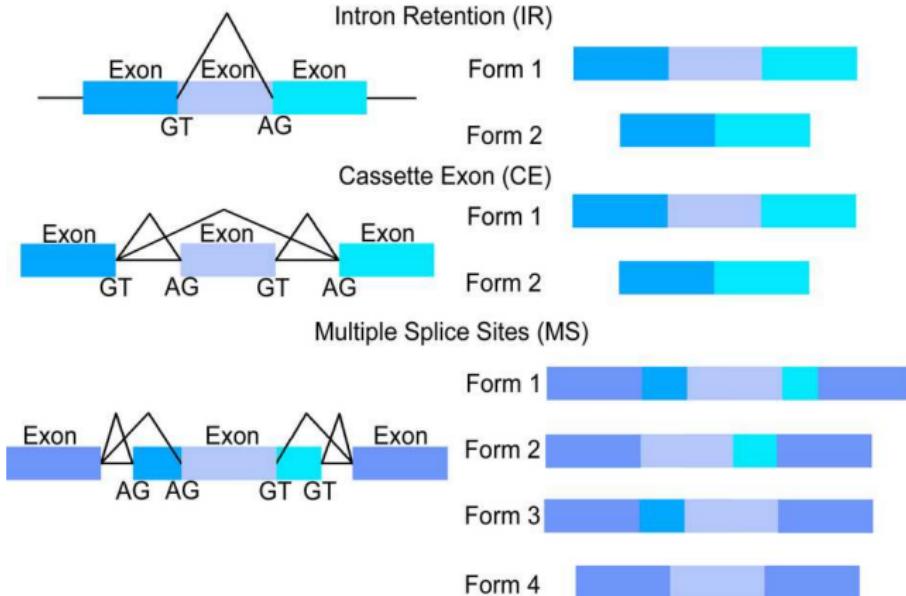
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Sestřih mRNA



Anotace sekvence a
genů

Anotace genomu

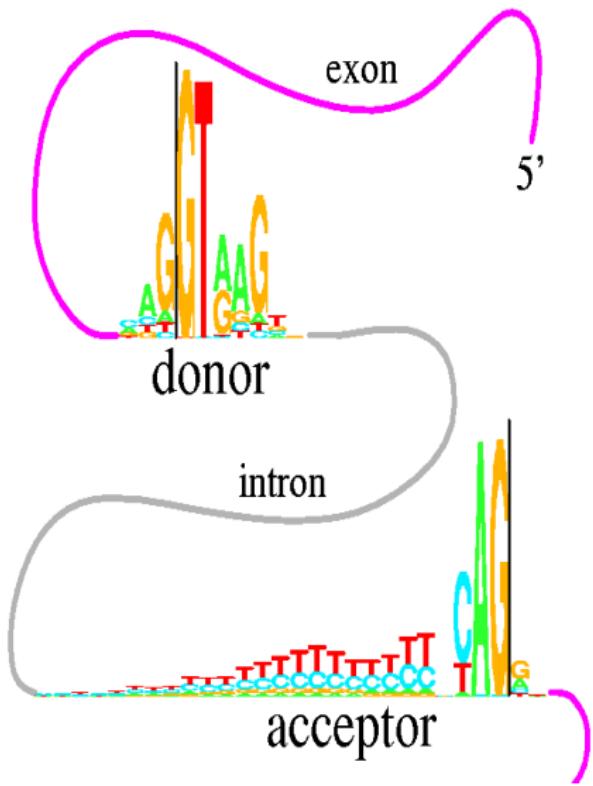
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Sekvenční logo intronu



Anotace sekvence a
genů

Anotace genomu

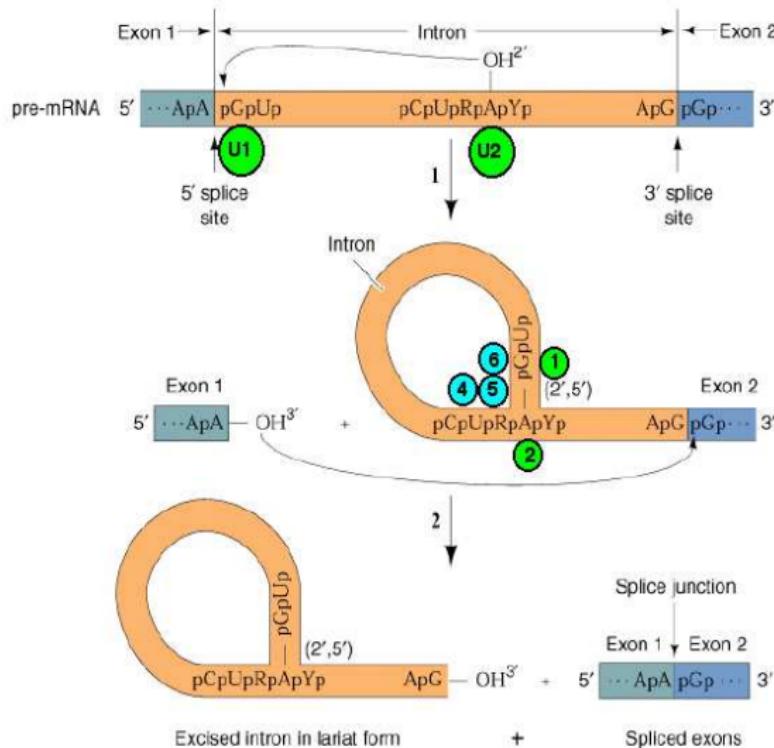
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

detailey sestřihu



Copyright 1999 John Wiley and Sons, Inc. All rights reserved.

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

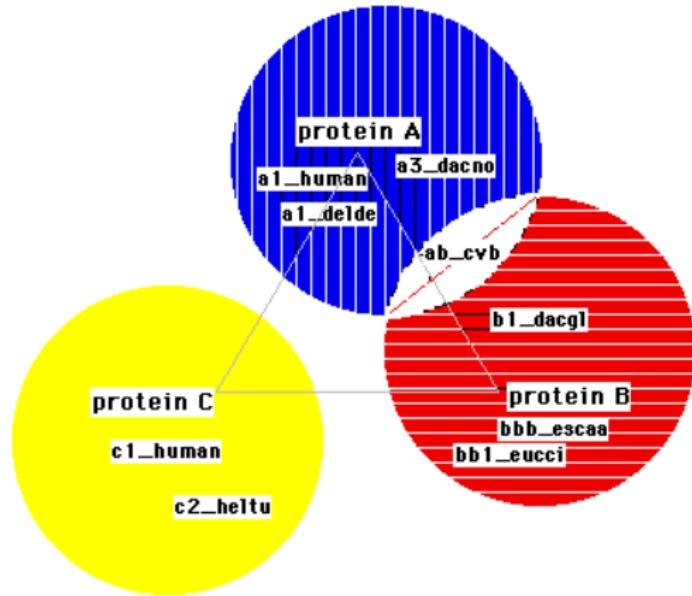
Homologie a podobnost

Příště

Bioinformatické databázy

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



Anotace sekvence a
genů

Anotace genomu

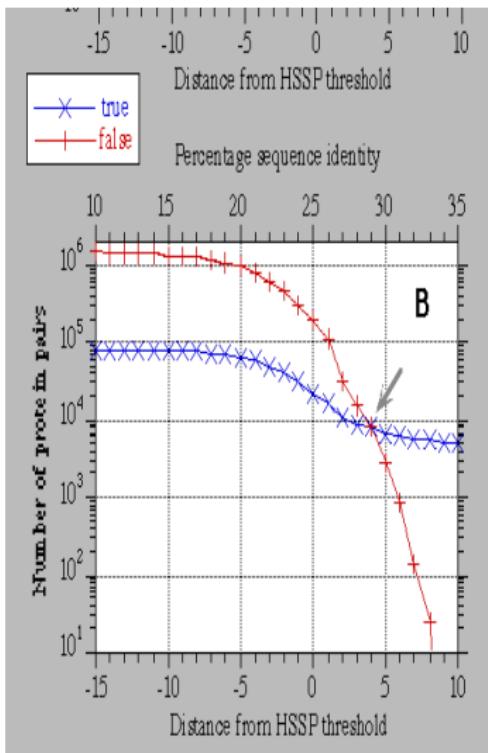
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Rost



Anotace sekvence a
genů

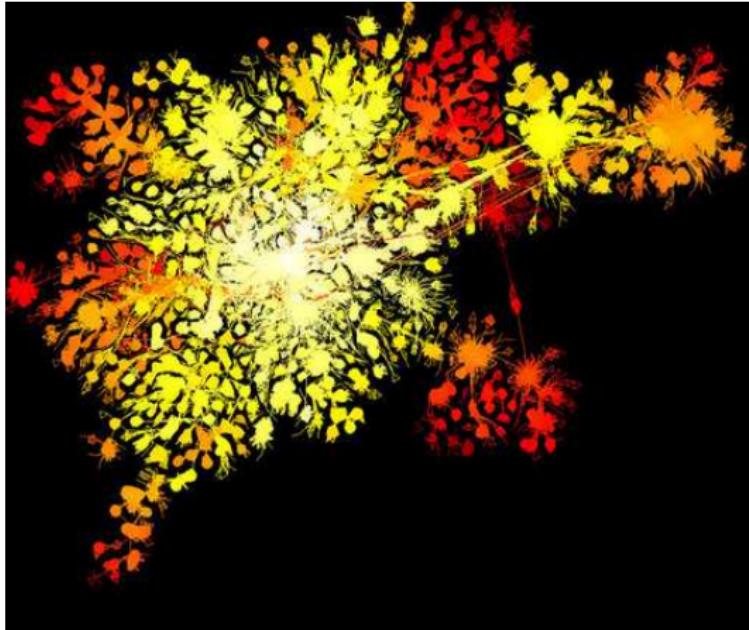
Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy



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

Anotace sekvence a
genů

Anotace genomu

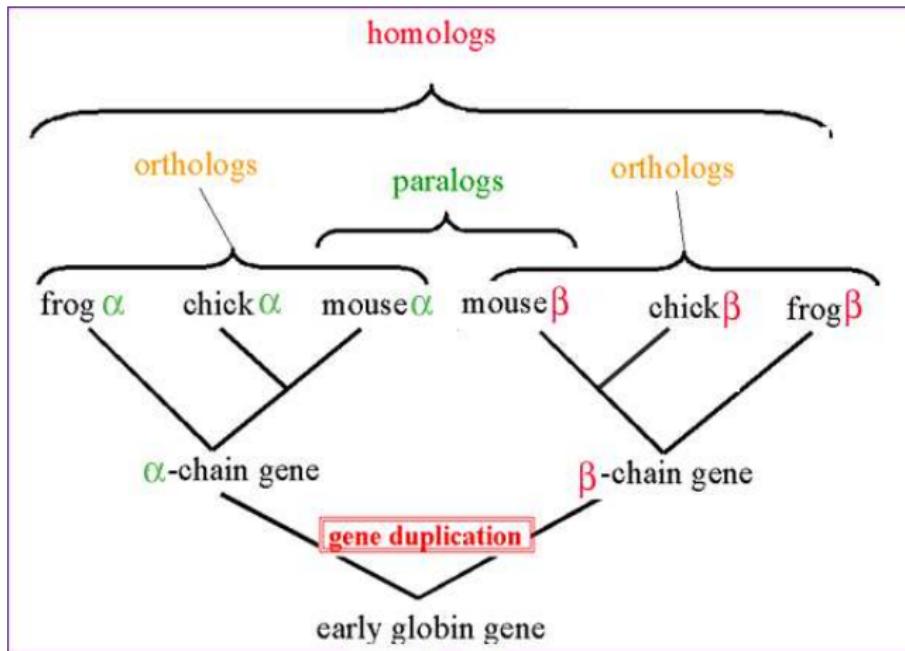
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

detailejší sestřihu



Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

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.

Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Podobnost sekvencí

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

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Rozdíl mezi lokálním a globálním porovnáváním

(A) local

PI3-kinase DRHNSNIMVKDDGQLFHI DFG
cAMP PK DLKPENLIDQQGYIQVT DFG

(B) global

PI3-kinase HQLGNLRL--LEECRI---MSSAKRPLWLWNWENPDIMSELL FQNNEIIFKNGDDL RQDMLT
cAMP PK GNAAAAKGXEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMIL
10 20 30 40 50

PI3-kinase LQIIRIME--NIWQNQGLDLRMLPYGCLSIGDCVGILIEVVRSHTIMQ-IQCKGGLK GAL
cAMP PK ---VKHMETGNHYAMKILDKQKVVK-----LKQIEHTLNEKRILQAVNFPFLVKLEF
60 70 80 90 100 110
60 70 80 90 100

PI3-kinase QFNSHT-LHQWLKDKNKG EYDAA--IDL FTRSCAGYCVATFILGIG DRHNSNIMVKD-D
cAMP PK SFKD NSNLYMVMEYVPG GEMFSHLRRIGRFSEPHARFYAAQIVLT FEYLHSLDLIYR D LK
110 120 130 140 150 160

PI3-kinase GQLFHIDFGHFLDHKKKKFGYKRERVP----FVL TDFL---IVISKGAQECTKTREFE
cAMP PK PENLLIDQQGYI--QVT DFGFAK-RVKGRTWXLCGTPEYLAPEIILSKGYNKA VDWALG
170 180 190 200 210 220

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

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

Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Tabulka pro algoritmus dynamického programování

$y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad y_6 \quad y_7 \quad y_8 \quad y_9$

	I	S	A	L	I	G	N	E	D
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							

$\leftarrow S_{0,j}$

 $\uparrow S_{i,0}$

Anotace sekvence a genů

Anotace genomu

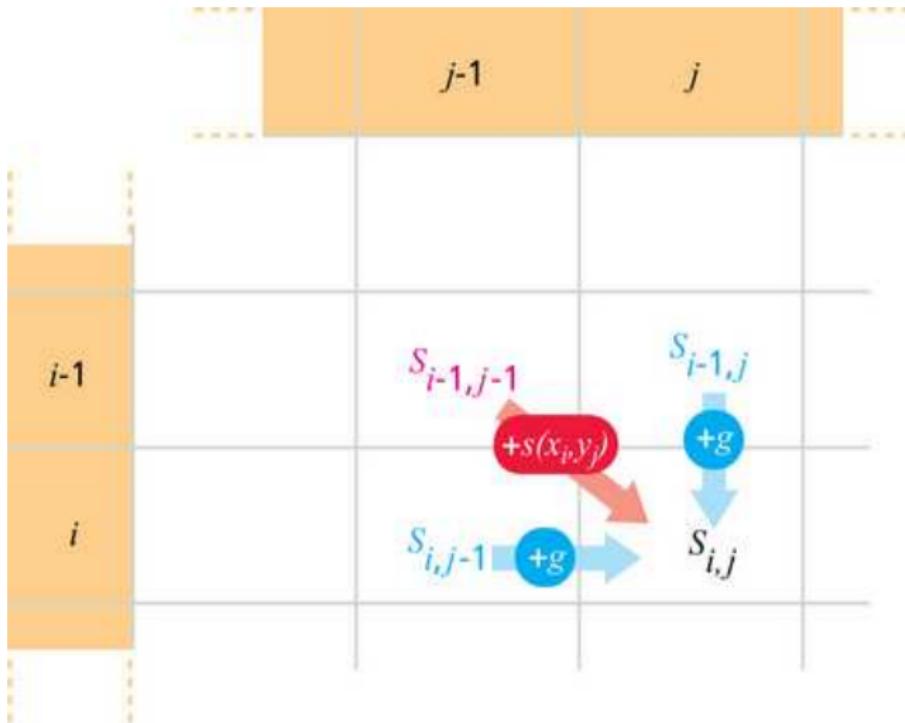
Identifikace a

Homologie a podobnost

Příště

Bioinformatické databázy

Tabulka pro algoritmus dynamického programování

Anotace sekvence a
genů

Anotace genomu

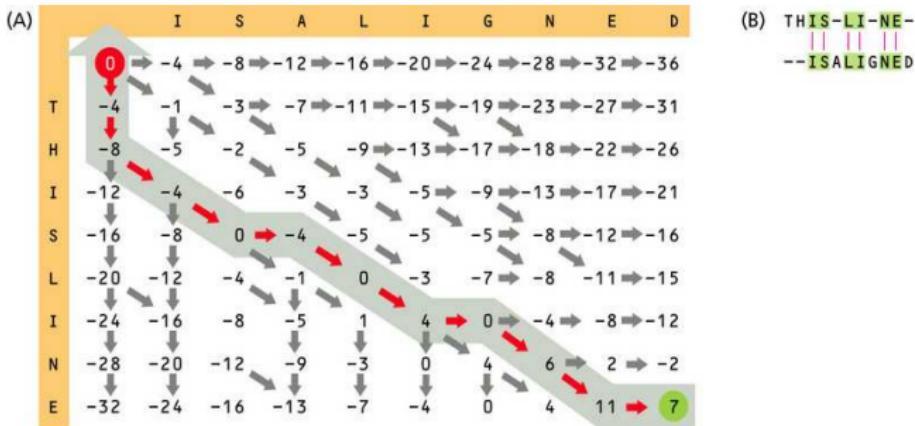
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Tabulka pro algoritmus dynamického programování



Anotace sekvence a genů

Anotace genomu

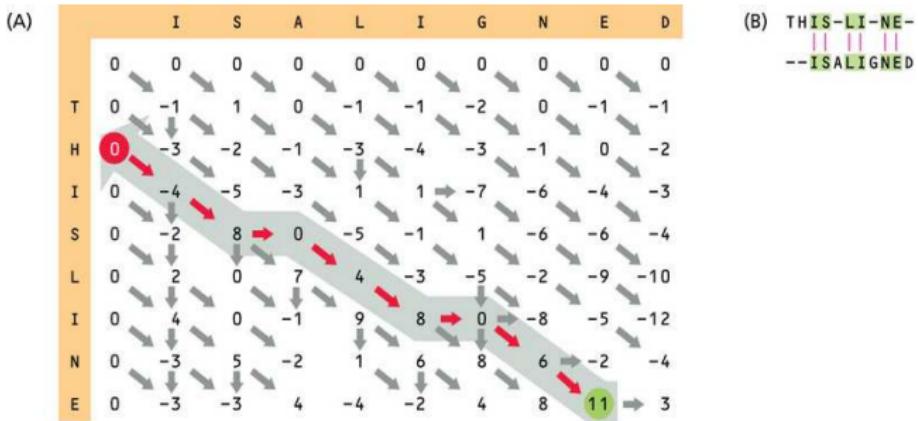
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Tabulka pro algoritmus dynamického programování



Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

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	0	0	0	0	2	0	1	0	
I	0	0	0	0	0	0	0	0	0	
N	0	0	0	0	0	0	0	0	0	
E	0	0	0	0	0	0	0	0	2	

(B) IN
IS

Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Příště Bioinformatické databázy

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Outline

Dodatek

Dodatek

For Further Reading

Dodatek

For Further Reading

For Further Reading
X