

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

IV107 Bioinformatika I

Přednáška 4

Katedra informačních technologií
Masarykova Univerzita Brno

Jaro 2009

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 databáze

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 databáze

Anotovaná sekvence DNA

>P12345 Gen1 - protein
alkoholdehydrogenáza

```
TATA      TATAAA
           CGATTGACGATGACGAT
start     ATG
exon1     TACAGATTACAGATTACAGATTAAGATGT
intron1   CAGATTACAGATTACAGATTACACAGATTCA
exon2     AGATTACAGATTACAGATTACAGA
stop      TAA
```

>P12346 Protein1
MASAQSFYLLDHNQNQNFDDHLAVDIVMILSHERFMN

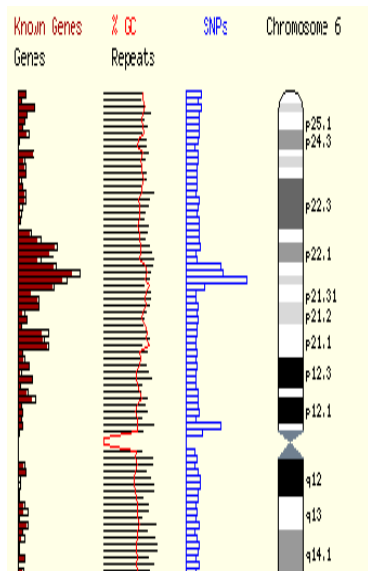
Anotace sekvence a genů

Anotace genomu
Identifikace genů
Homologie a podobnost

Příště

Bioinformatické databázy

Anotace genomu



Anotace sekvence a genů

Anotace genomu

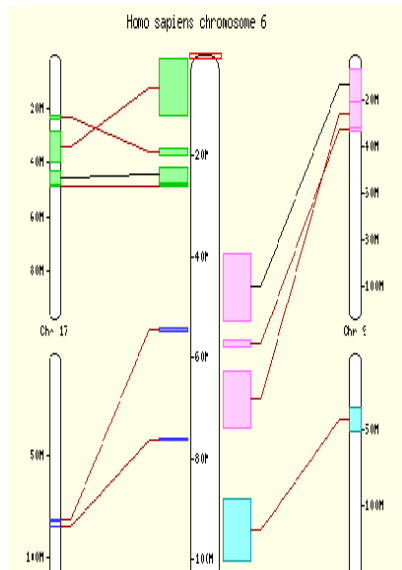
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

Anotace genomu



Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

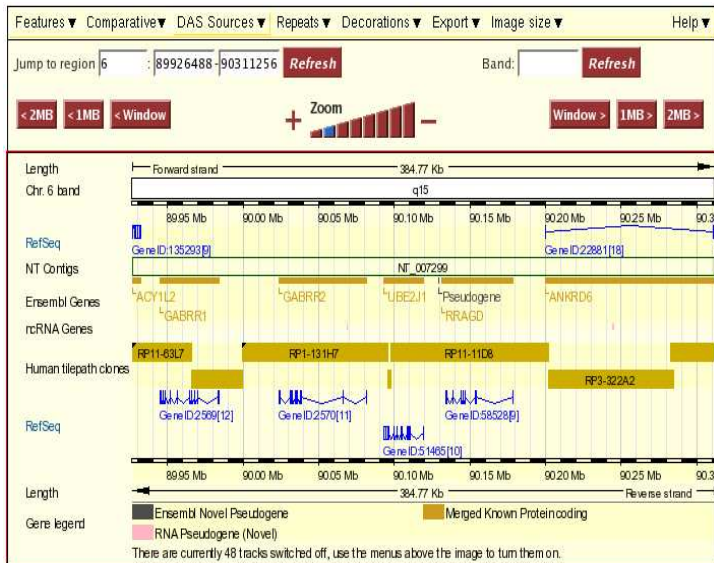
Anotace genomu <http://www.ensembl.org/>

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 databáze

Způsoby identifikace genů in silico

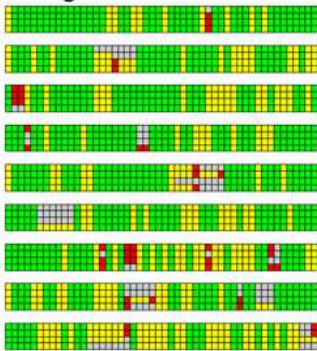
- ▶ Experimentální metody (cDNA, EST)
- ▶ Komparativní metody
 - ▶ Selekční tlak
 - ▶ Druh zachovaných mutací
- ▶ Strukturní metody (GeneMark, GeneScan, GeneID)
- ▶ Detekce charakteristických signálů





Identifikace genů podle charakteru mutací

Gene



Intergenic



 Conserved  Mutation  Gap  Frameshift

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

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

Anotace sekvence a genů

Anotace genomu

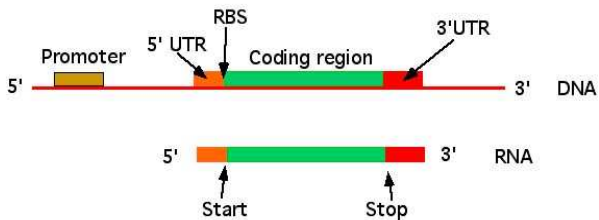
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

Struktura genu (prokaryotická)



Anotace sekvence a
genů

Anotace genomu

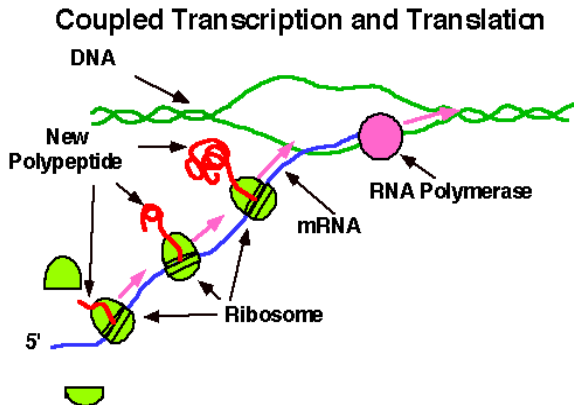
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Vztah transkripce a translace u prokaryotů



Anotace sekvence a
genů

Anotace genomu

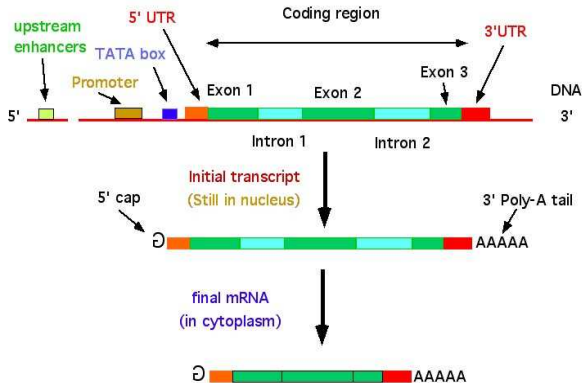
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

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

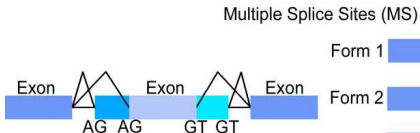
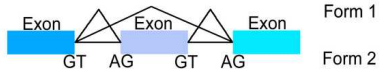
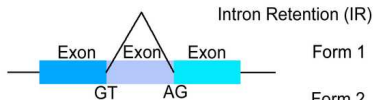
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

Sestřih mRNA



Anotace sekvence a
genů

Anotace genomu

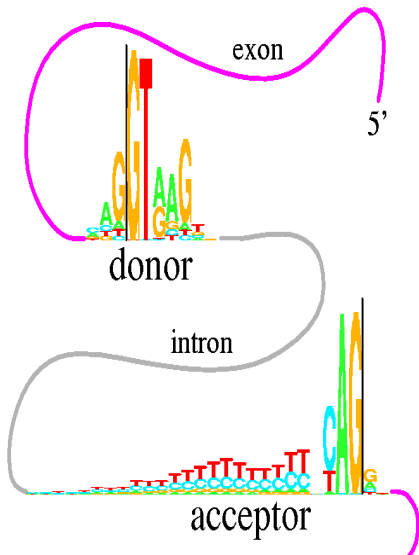
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Sekvenční logo intronu



Anotace sekvence a genů

Anotace genomu

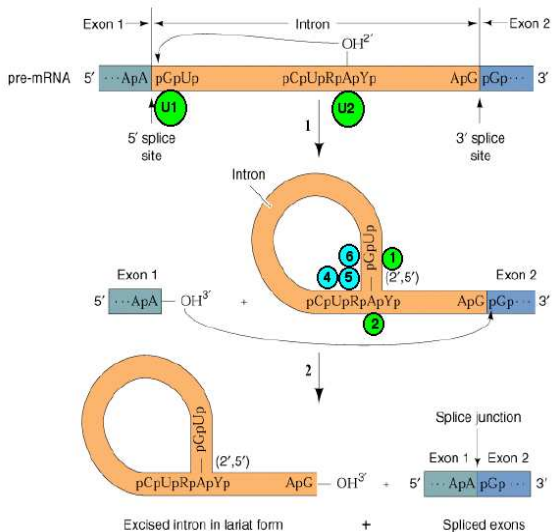
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

detaily sestřihu



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

Anotace sekvence a genů

Anotace genomu

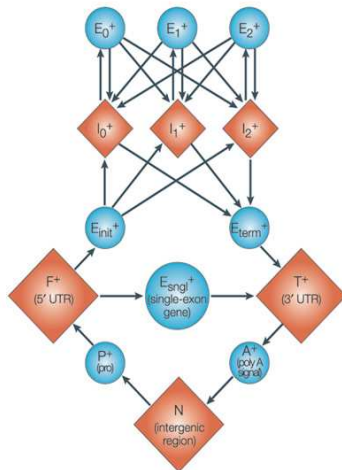
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Identifikace genů podle struktury



Reverse strand: mirror reflection of above

Nature Reviews | Genetics

Anotace sekvence a genů

Anotace genomu

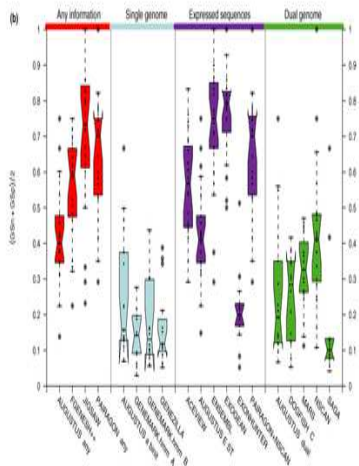
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Úspěšnost identifikace genů



Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

Guigo et al, Genome Biology (2006)

Anotace sekvence a
genů

Anotace genomu

Identifikace genů

Homologie a podobnost

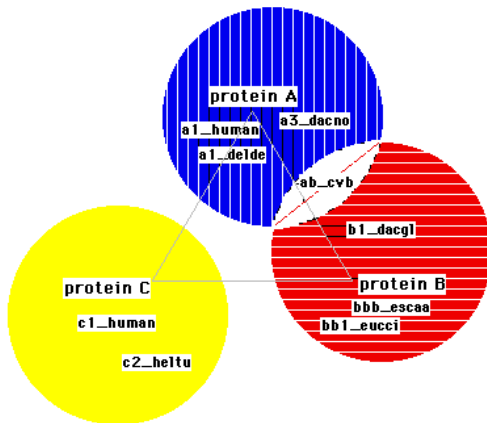
Příště

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 - prostor podobnosti



Anotace sekvence a genů

Anotace genomu

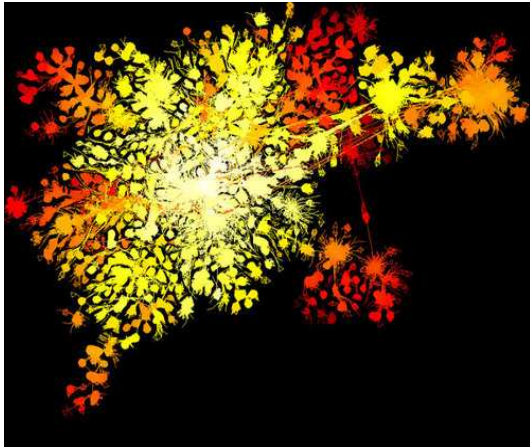
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Sít 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

Anotace sekvence a genů

Anotace genomu

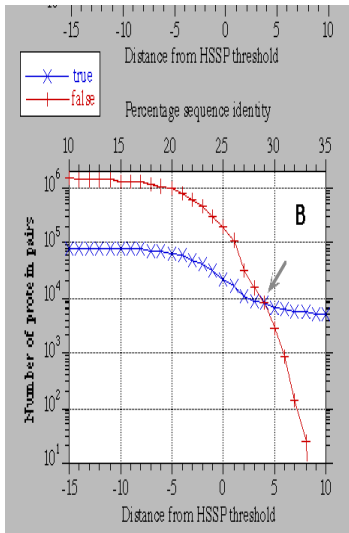
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Rost - "twilight zone"



Anotace sekvence a
genů

Anotace genomu

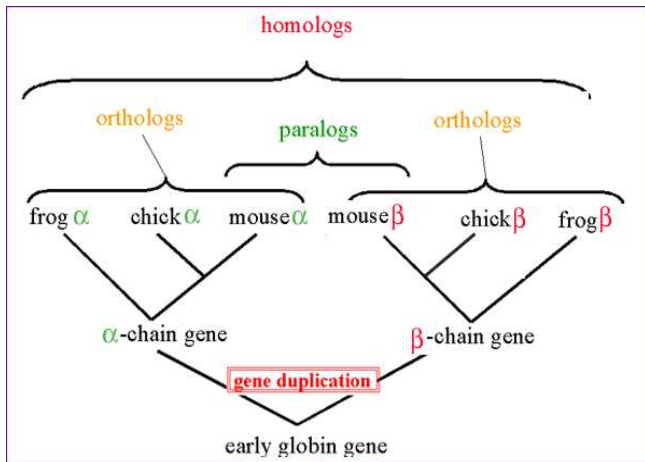
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

homologie



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

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

Rozdíl mezi lokálním a globálním porovnáním

(A) local

PI3-kinase DRHNSNIMVKDDGQLFHI DFG

cAMP PK DLKPENLLIDQQGYIQVT DFG

(B) global

PI3-kinase HQLGNLR--L EECRI---MSSAKRPLWLNWENP DIMSEL FQNNEIIFKNGDDL RQDMLT
cAMP PK GNAAA KKGXEQESVKEFLAKAKEDFLK KWENPAQNTAHL DQFERIKTLGTGSFGRVML-

PI3-kinase LQIIRIME--NIWQNGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQ-IQCKGGLK GAL
cAMP PK ---VKHMETGNHYAMKILDKQKVVK-----LQIEHTLNEKRILQAVNFPFLVKLEF

PI3-kinase QFN SHT-LHQWLKDKNKGEIYDAA--IDL FTRSCAGYCVATFILGIGDRHNSNIMVKD-D
cAMP PK SFDNSNLYMVMEYVPGGEMFSLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLK

PI3-kinase GQLFHI DFGHFLDHKKKFGYKRERVP-----FVLTQDFL---IVISKGAQECTKTREFE
cAMP PK PENLLIDQQGYI--QVTD FGF AK-RVKGRTWXLGCTPEYLAPEIILSKGYNKAVDWWALG

Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

Matice pro hodnocení podobnosti proteinových sekvencí

Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

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

		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										
		0	-8	-16	-24	-32	-40	-48	-56	-64	-72	$S_{0,j}$
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										
		$S_{i,0}$										

Anotace sekvence a genů

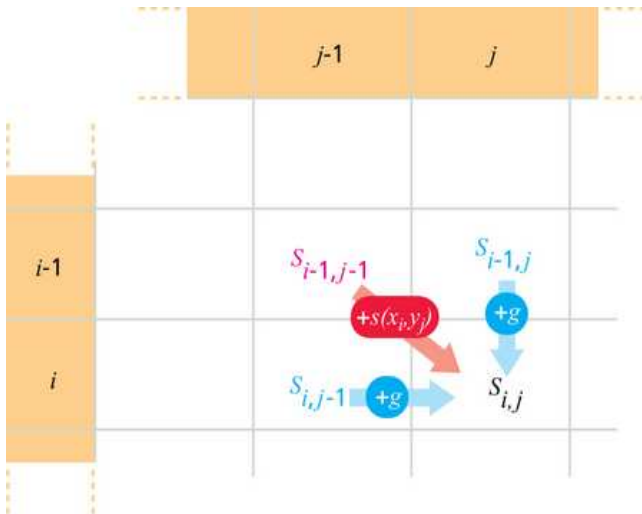
Anotace genomu
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

Tabulka pro algoritmus dynamického programování



Anotace sekvence a
genů

Anotace genomu

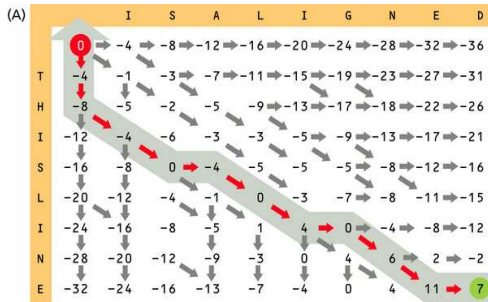
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Tabulka pro algoritmus dynamického programování



Anotace sekvence a
genů

Anotace genomu

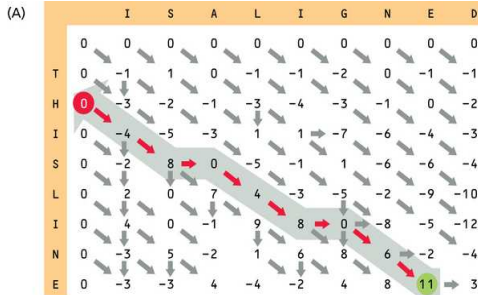
Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databázy

Tabulka pro algoritmus dynamického programování



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

Anotace sekvence a genů

Anotace genomu

Identifikace genů

Homologie a podobnost

Příště

Bioinformatické databáze

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	0	0	0	0	0	1	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



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

Příště Bioinformatické databázy

Outline

Dodatek

Dodatek

For Further Reading

Dodatek

For Further Reading

For Further Reading

X