CG020 Genomika

Přednáška 2

Identifikace genů

Jan Hejátko

Funkční genomika a proteomika rostlin,
Mendelovo centrum genomiky a proteomiky rostlin,
Středoevropský technologický institut (CEITEC), Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.muni.cz

EVROPSKÁ UNIE esf MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Literatura

- Zdrojová literatura ke kapitole 2

- Plant Functional Genomics, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey
- Majoros, W.H., Pertea, M., Antonescu, C. and Salzberg, S.L. (2003) GlimmerM, Exonomy, and Unveil: three ab initio eukaryotic genefinders. *Nucleic Acids Research*, **31**(13).
- Singh, G. and Lykke-Andersen, J. (2003) New insights into the formation of active nonsensemediated decay complexes. *TRENDS in Biochemical Sciences*, **28** (464).
- Wang, L. and Wessler, S.R. (1998) Inefficient reinitiation is responsible for upstream open reading frame-mediated translational repression of the maize R gene. *Plant Cell*, **10**, (1733)
- de Souza et al. (1998) Toward a resolution of the introns earlyylate debate: Only phase zero introns are correlated with the structure of ancient proteins *PNAS*, **95**, (5094)
- Feuillet and Keller (2002) Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution *Ann Bot*, 89 (3-10)
- Frobius, A.C., Matus, D.Q., and Seaver, E.C. (2008). Genomic organization and expression demonstrate spatial and temporal Hox gene colinearity in the lophotrochozoan Capitella sp. I. PLoS One 3, e4004

EVROPSKÁ UNIE esf MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Osnova

- Postupy „přímé“ a reverzní genetiky
  - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
  - struktura genů a jejich vyhledávání
  - genomová kolinearita a genová homologie
- Experimentální identifikace genů
  - příprava genově obohacených knihoven pomocí technologie metylačního filtrování
  - EST knihovny
  - přímá a reverzní genetika

EVROPSKÁ UNIE esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Osnova

- Postupy „přímé“ a reverzní genetiky
  - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Přímá vs. reverzní genetika

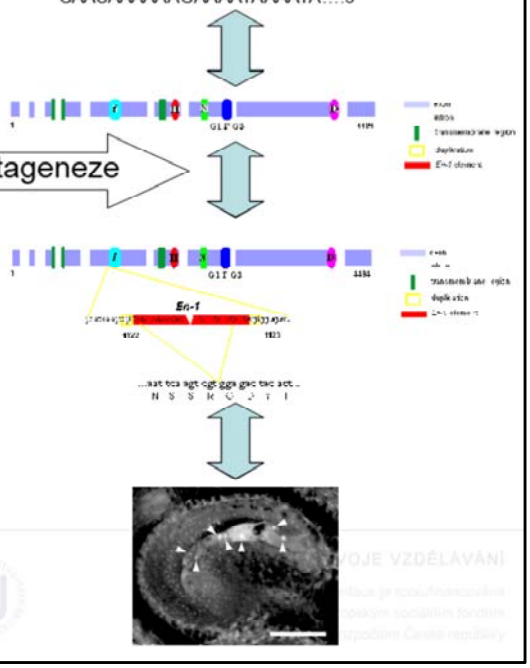
Revoluce v chápání pojmu genu

Přístupy „klasické“ genetiky

„Reverzně genetický“ přístup

5'TTATATATATATATTAAAAAATAAAATAAAA
GAACAAAAAAGAAAATAAAATA....3'

inzerční mutageneze



3 : 1

?

Identifikace role genu *ARR21*

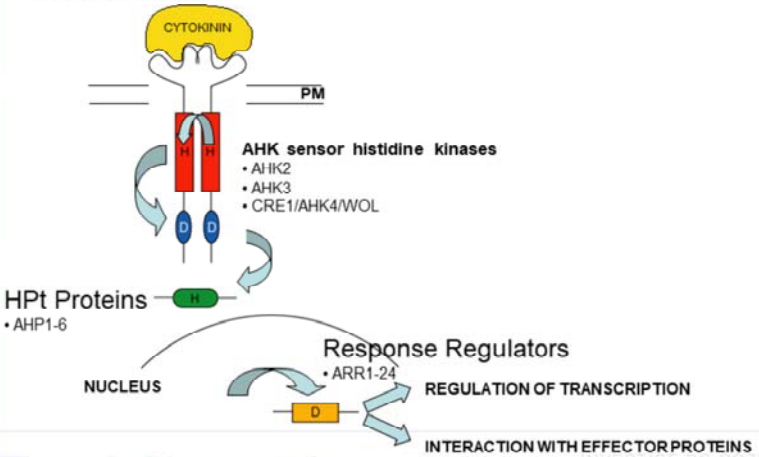
- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21*

Recent Model of the CK Signaling via Multistep Phosphorelay (MSP) Pathway



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21*

- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*
- Mutant identifikován vyhledáváním v databázi inzerčních mutantů (SINS-sequenced insertion site) pomocí programu BLAST

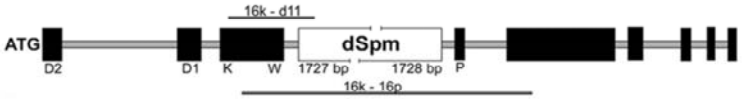
EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21* – izolace inz. mutanta

- **vyhledávání v databázi inzerčních mutantů (SINS)**

```
Insert_SINS: 01_09_64
Query: 80    tcctagcgttcatgagcgtaccatacttgacaanagagaacgtagccagccatttacagg 139
             ||||||||||||||||||||||||||||||||| |||||||||||||||||||||||||
Sbjct: 58319 tcctagcgttcatgagcgtaccatacttgacaagagagaacgtagccagccatttacagg 58378
Arr21:  1830

Insert_SINS: 01_09_64
Query: 140   tttgatatctcttgtcaaaaatgtttttggattttactgt 179
              |||||||||||||||||||||||||||||||||||||||
Sbjct: 58379 tttgatatctcttgtcaaaaatgtttttggattttactgt 58418
Arr21:  1890
```

- **lokalizace inzerce *dSpm* v genomové sekvenci *ARR21* pomocí sekvenace PCR produktů**



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
VZDĚLÁVÁNÍ
Evropským sociálním fondem
a státním rozpočtem České republiky

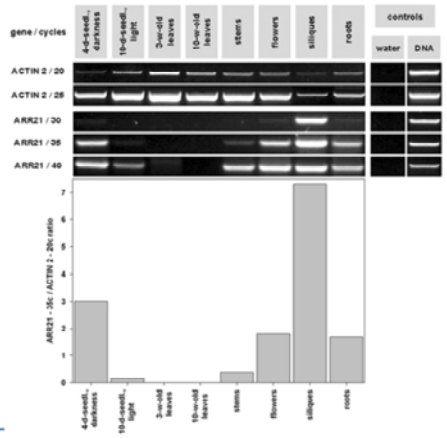
Identifikace role genu *ARR21*

- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*
- Mutant identifikován vyhledáváním v databázi inzerčních mutantů (SINS-sequenced insertion site) pomocí programu BLAST
- Exprese *ARR21* u standardního typu a Inhibice exprese u inzerčního mutanta potvrzena na úrovni RNA

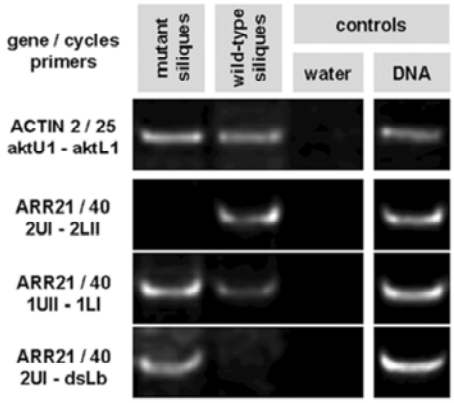
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21* – analýza exprese

Standardní typ



Inzerční mutant



EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21*

- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*
- Mutant identifikován vyhledáváním v databázi inzerčních mutantů (SINS-sequenced insertion site) pomocí programu BLAST
- Exprese *ARR21* u standardního typu a Inhibice exprese u inzerčního mutanta potvrzena na úrovni RNA
- Analýza fenotypu inzečního mutanta

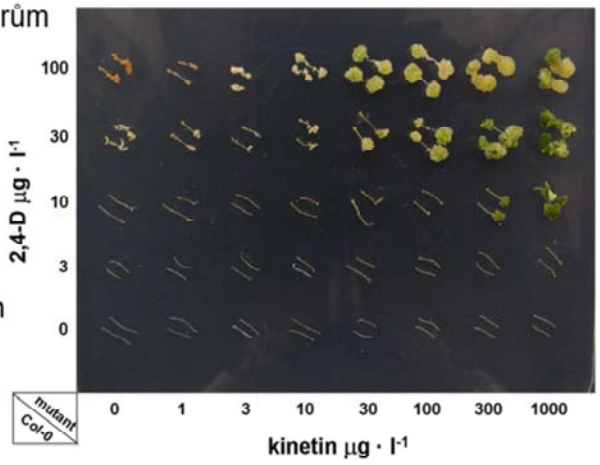
EVROPSKÁ UNIE esf MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21* – analýza fenotypu mutanta

- Analýza citlivosti k regulátorům růstu rostlin
  - 2,4-D a kinetin
  - etylén
  - světlo různých vlnových délek
- Doba kvetení i počet semen nezměněn



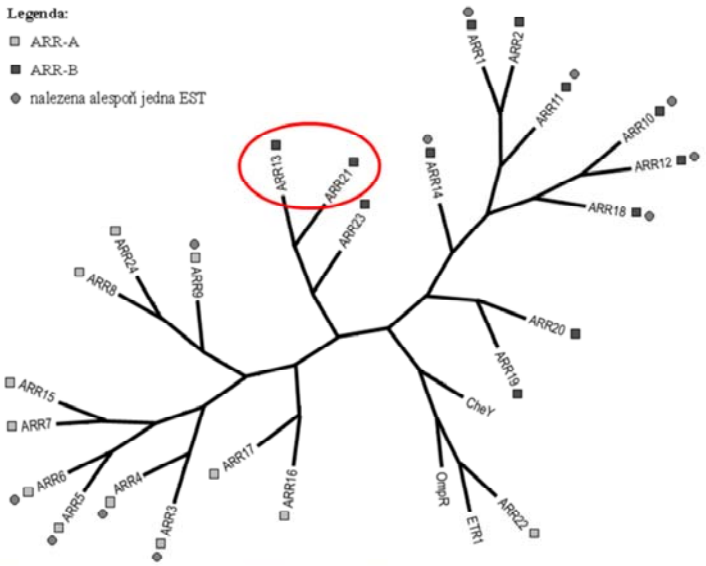
EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21* – příčiny absence fenotypu

- Funkční redundance v rámci genové rodiny?

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21* – příbuznost ARR genů



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
ZVOJE VZDĚLÁVÁNÍ
realizace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21* – příčiny absence fenotypu

- Funkční redundance v rámci genové rodiny?
- Fenotypový projev pouze za velmi specifických podmínek (?)

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace role genu *ARR21* – shrnutí

- Gen *ARR21* identifikován pomocí srovnávací analýzy genomu *Arabidopsis*
- Na základě analýzy sekvence byla předpovězena jeho funkce
- Byla prokázána místně specifická exprese genu *ARR21* na úrovni RNA
- Identifikace funkce genu pomocí inzerční mutageneze v případě *ARR21* ve vývoji *Arabidopsis* byla neúspěšná, pravděpodobně v důsledku funkční redundance v rámci genové rodiny

EVROPSKÁ UNIE esf MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

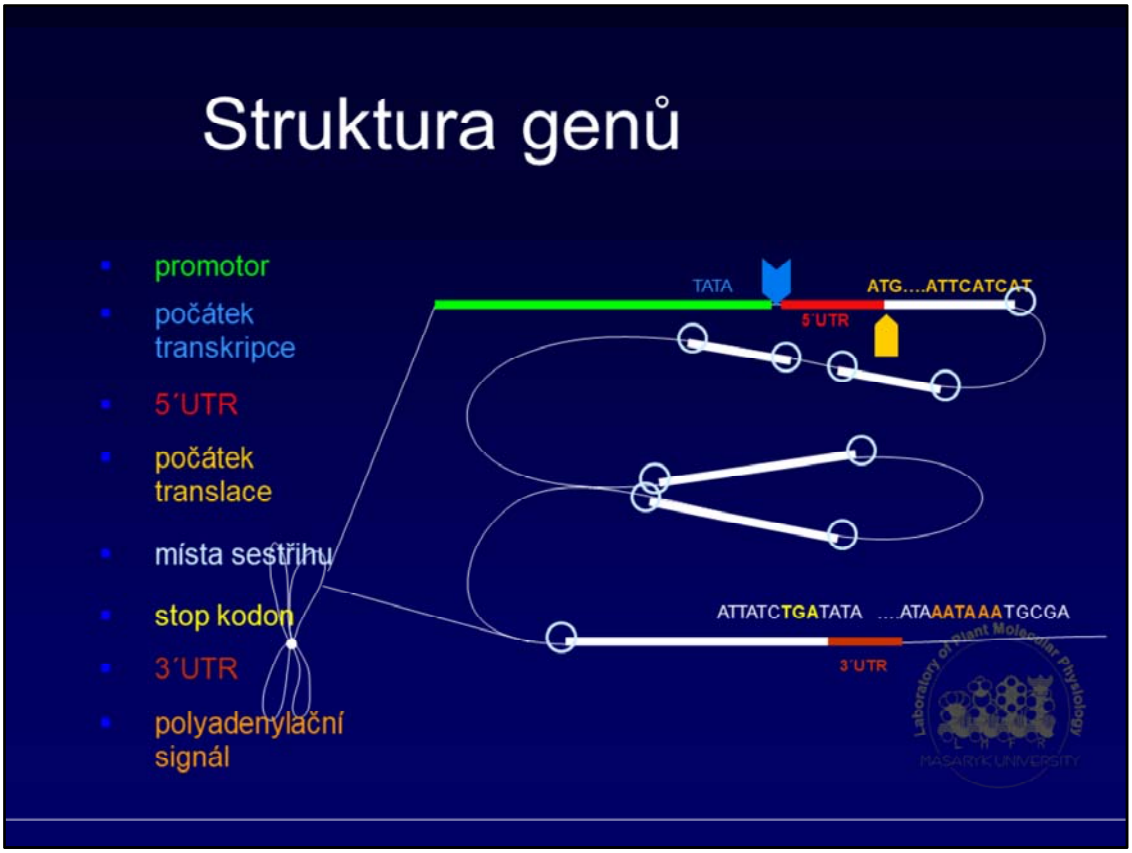
Osnova

- Postupy „přímé“ a reverzní genetiky
  - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
  - struktura genů a jejich vyhledávání

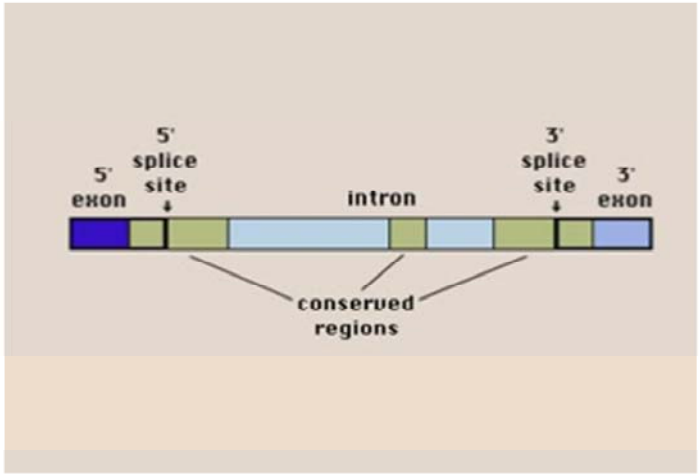
EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Struktura genů

- promotor
- počátek transkripce
- 5´UTR
- počátek translace
- místa sestřihu
- stop kodon
- 3´UTR
- polyadenylační signál



Sestřih RNA



EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Identifikace Genů *Ab Initio*

- zanedbání 5‘ a 3‘ UTR
- identifikace počátku translace (ATG) a stop kodonu (TAG, TAA, TGA)
- nalezení donorových (většinou GT) a akceptorových (AG) míst sestřihu
- většina ORF není skutečně kódujícími sekvencemi – u *Arabidopsis* je asi 350 mil. ORF na každých 900 bp (!)
- využití různých statistických modelů (např. Hidden Markov Model, HMM, viz doporučená studijní literatura, Majoros et al., 2003) k posouzení a ohodnocení váhy identifikovaných donorových a akceptorových míst

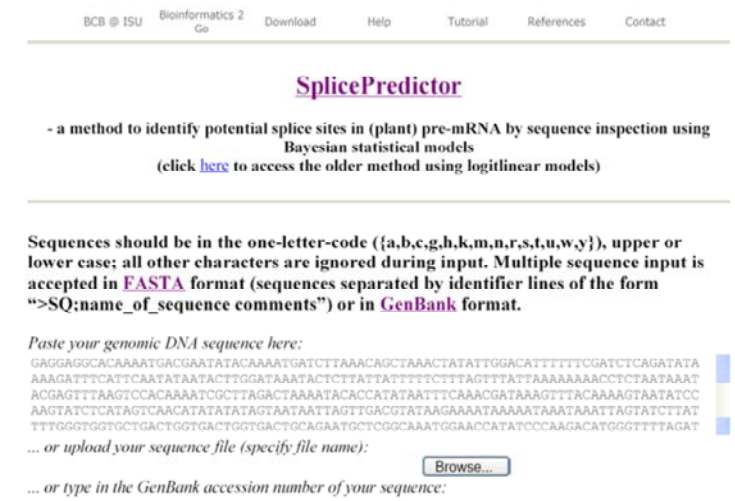
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Predikce míst sestřihu

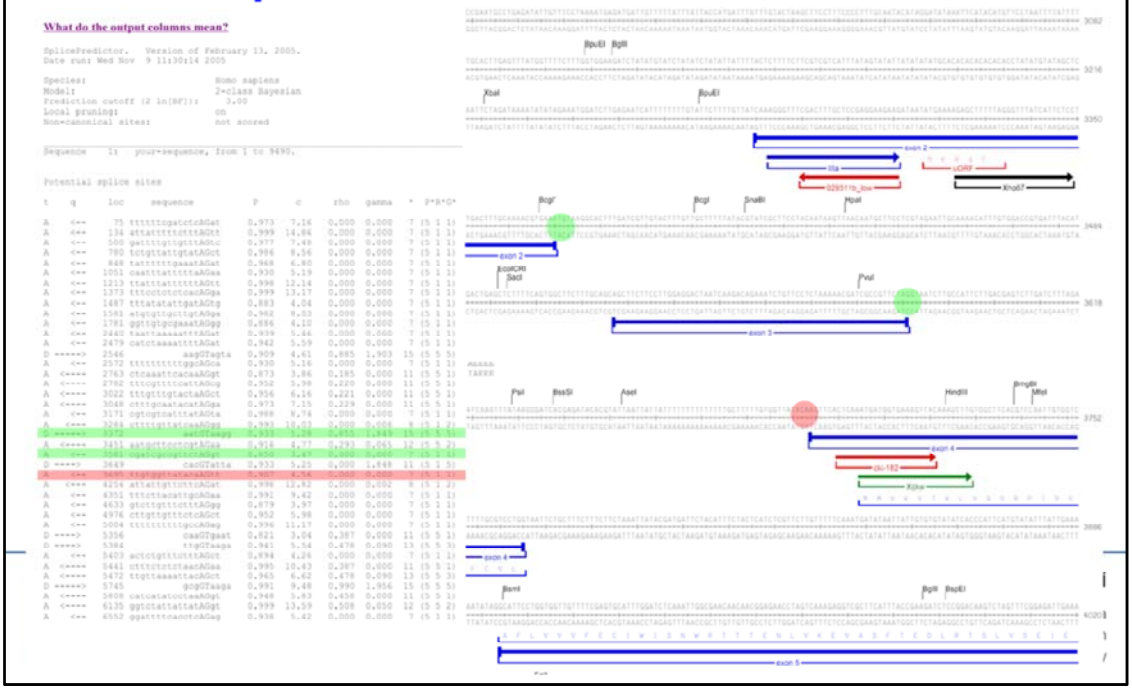
- programy pro predikci míst sestřihu
  (specificita přibližně 35%)
  - GeneSplicer (http://www.tigr.org/tdb/GeneSplicer/gene_spl.html)
  - SplicePredictor (http://deepc2.psi.iastate.edu/cgi-bin/sp.cgi)

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

SplicePredictor



SplicePredictor

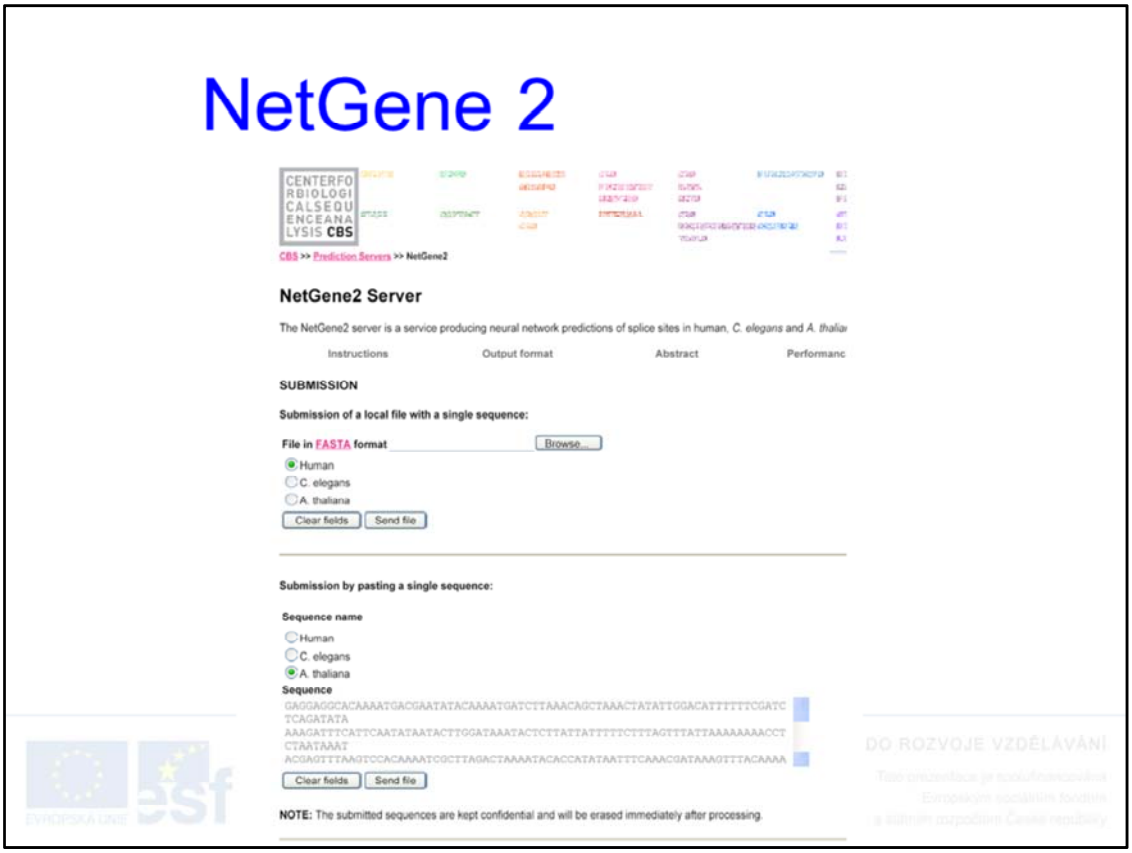


Identifikace Genů *Ab Initio*

- programy pro predikci míst sestřihu (specificita přibližně 35%)
  - GeneSplicer (http://www.tigr.org/tdb/GeneSplicer/gene_spl.html)
  - SplicePredictor (http://deepc2.psi.iastate.edu/cgi-bin/sp.cgi)
  - NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

NetGene 2



EVROPSKÁ UNIE
esf
DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

NetGene 2

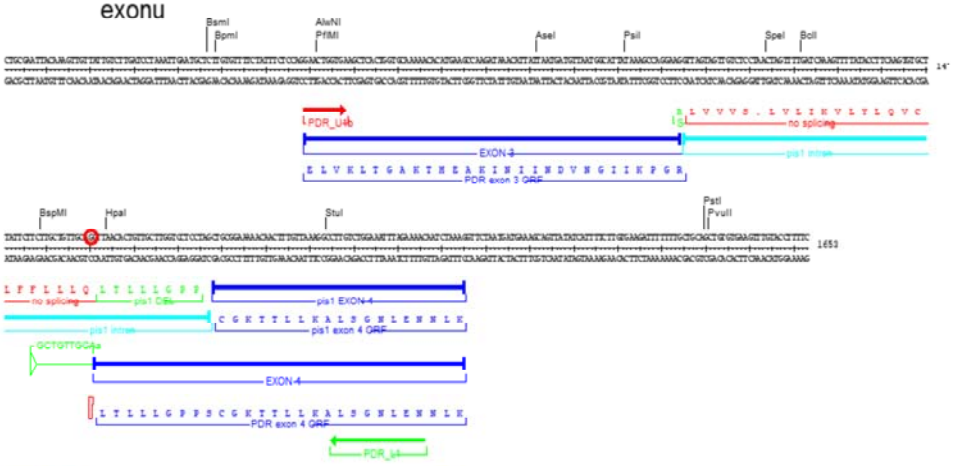
Prediction done

```
********** NetGene2 v. 2.4 **********

The sequence: Sequence has the following composition:

Length: 9490 nucleotides.
31.8% A, 17.0% C, 19.6% G, 31.7% T, 0.0% X, 36.5% G+C

Donor splice sites, direct strand
---------------------------------
                pos 5'->3'  phase strand  confidence  5'     exon intron    3'
                      1704      0      +        0.87       TTCCAAACAG^GTTAATATTT
                      1906      0      +        0.99       CGGTGAACGG^GTCAGAACAT
                      3582      1      +        1.00       GCCGTTCCAG^GTAATCCTGC H
                      3710      1      +        1.00       TTGCTTCCTG^GTAATTCTGC H
                      4134      0      +        0.74       TCAAACACAG^GTTGTTAAAA
                      4619      1      +        0.74       AGCAAGAAAG^GTCTTGTTTC
                      4915      0      +        0.94       CGTTCCTCTG^GTAAATACTG
                      5356      0      +        0.87       TCTCAACCAA^GTGAATGTTT
                      5384      1      +        1.00       GATTTGGTTG^GTAAGACTCT H
                      5809      1      +        1.00       TATCCTAAAG^GTGTGTCCAA
                      6057      0      +        1.00       GCAGTCTTTG^GTAAGCTACT H
                      6096      1      +        0.74       CTCTTCACAA^GTAAATCTAG
                      7369      0      +        1.00       GGACTGCCAA^GTAAGTTTAA H
                      7886      0      +        0.74       GAACAAAATG^GTTAGATGAA
                      9323      0      +        0.74       GAAGATTAGG^GTTTTTCTGT

Donor splice sites, complement strand
-------------------------------------
   pos 3'->5'  pos 5'->3'  phase strand  confidence  5'     exon intron    3'

Acceptor splice sites, direct strand
------------------------------------
                pos 5'->3'  phase strand  confidence  5'     intron exon    3'
                      1213      0      +        0.59       TATTTTTTAG^TTATGGAGAC
                      1221      2      +        0.87       AGTTATGGAG^ACAAGAATCG
                      1373      0      +        0.71       TCTCTCACAG^GACACAGAAT
                      1487      1      +        0.81       ATATTGATAG^TGGGACATTA
                      3284      0      +        0.87       GTTATCAAAG^GGTTCTCACC
                      4254      0      +        1.00       TGTTCTTCAG^ATCGCACCAT H
                      4832      2      +        0.54       AAAATTGCAG^TTCCAGTGGC
                      5004      0      +        0.94       TTTTTGCCAG^AGATACACAC
                      5472      1      +        0.96       AAAATTACAG^CTCTGCTCAA
                      6135      0      +        1.00       ATTATTATAG^GTAAGATTAA H
                      6490      1      +        0.90       AAAGTTACAG^TGGTGGAGAA
                      6744      0      +        0.59       TGTCAAACAG^TTTCGTAGAG
                      7447      0      +        0.96       TTCTGCACAG^ATGCCAGAAA
                      7780      2      +        0.76       TCCATTTCAG^ATACAGAACA
                      7786      2      +        0.92       TCAGATACAG^AACACATGCA
```

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Sestřih RNA a adaptace

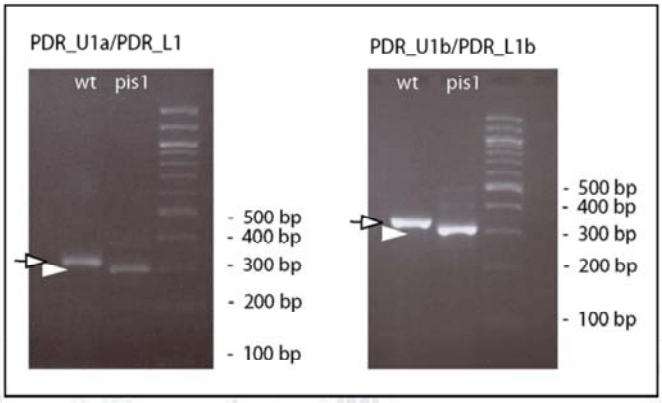
- odchylky rozpoznávání míst sestřihu u rostlin v praxi - příklad vývojové plasticity (nejen) rostlin
  - identifikace mutanta s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5' konci 4. exonu



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
a státním rozpočtem České republiky

Sestřih RNA a adaptace

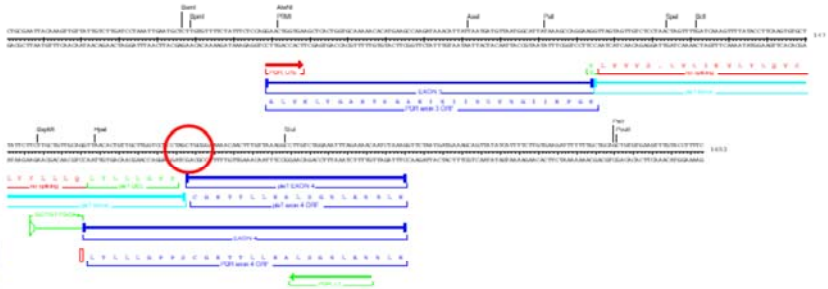
- identifikace mutanta s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5' konci 4. exonu
- analýza pomocí RT PCR prokázala přítomnost fragmentu kratšího než by odpovídalo cDNA po normálním sestřihu



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
ROZVOJE VZDĚLÁVÁNÍ
Evropským sociálním fondu
a státním rozpočtem České republiky

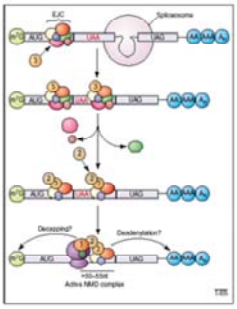
Sestřih RNA a adaptace

- odchylky rozpoznávání míst sestřihu u rostlin v praxi - příklad vývojové plasticity (nejen) rostlin
  - identifikace mutanta s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5‘ konci 4. exonu
  - analýza pomocí RT PCR prokázala přítomnost fragmentu kratšího než by odpovídalo cDNA po normálním sestřihu
  - sekvenace tohoto fragmentu pak ukázala na alternativní sesřih s využitím nejbližšího možného místa sestřihu v exonu 4



Sestřih RNA a adaptace

- odchylky rozpoznávání míst sestřihu u rostlin v praxi - příklad vývojové plasticity (nejen) rostlin
  - identifikace mutanta s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5' konci 4. exonu
  - analýza pomocí RT PCR prokázala přítomnost fragmentu kratšího než by odpovídalo cDNA po normálním sestřihu
  - sekvenace tohoto fragmentu pak ukázala na alternativní sesřih s využitím nejbližšího možného místa sestřihu v exonu 4



  - existence podobných obranných mechanizmů prokázána i u jiných organizmů (např. nestabilita mutantní mRNA se vznikem předčasného stopkodonu (> 50-55 bp před normálním stop kodonem) u eukaryot, viz doporučená studijní literatura, Singh and Lykke-Andersen, 2003)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

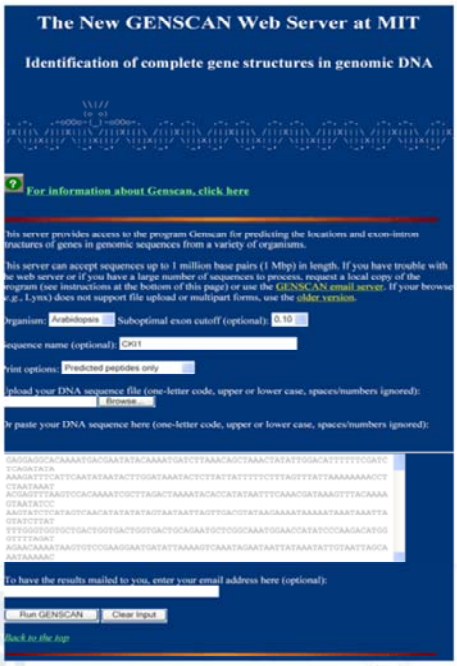
Identifikace genů *ab initio*

- programy pro predikci exonů
  - 4 typy exonů (podle polohy):
    - iniciační
    - vnitřní
    - terminální
    - jednoduché
  - programy kromě rozpoznávání míst sestřihu zohledňují i strukturu jednotlivých typů exonů

- iniciační:
  - Genescan (http://genes.mit.edu/GENSCAN.html)
  - GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/)
- interní:
  - MZEF (http://rulai.cshl.org/tools/genefinder/)

EVROPSKÁ UNIE esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

GENESCAN



The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

For information about Genscan, click here

his server provides access to the program Genscan for predicting the locations and exon-intron tructures of genes in genomic sequences from a variety of organisms.

his server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with he web server or if you have a large number of sequences to process, request a local copy of the rogram (see instructions at the bottom of this page) or use the GENSCAN email server. If your browse e.g., Lynx) does not support file upload or multipart forms, use the older version.

rganism: Arabidopsis Suboptimal exon cutoff (optional): 0.10

equence name (optional): CK01

rint options: Predicted peptides only

pload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

Browse...

r paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

To have the results mailed to you, enter your email address here (optional):

Run GENSCAN Clear Input

Back to the top

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

GENESCAN

GENSCANW output for sequence CKII

GENSCAN 1.0 Date run: 10-Nov-105 Time: 02:24:26

Sequence CKII : 9490 bp : 36.53% C+G : Isochore 1 (0 - 43 C+G%)

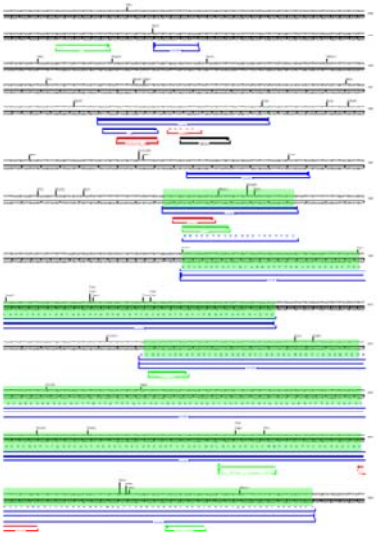
Parameter matrix: Arabidopsis.smat

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.00	Prom	+	1497	1536	40							-3.85
1.01	Init	+	3708	3764	57	2	0	63	51	37	0.499	4.03
1.02	Intr	+	3894	4133	240	2	0	-3	7	327	0.713	17.32
1.03	Intr	+	4255	4914	660	0	0	86	59	296	0.771	22.57
1.04	Intr	+	5005	5383	379	0	1	70	91	343	0.772	31.41
1.05	Intr	+	5473	6056	584	2	2	38	99	582	0.722	50.76
1.06	Intr	+	6136	7368	1233	0	0	68	108	655	0.977	56.86
1.07	Term	+	7448	7660	213	1	0	43	35	212	0.999	12.65
1.08	PlyA	+	7910	7915	6							-0.45
2.03	PlyA	-	7976	7971	6							-4.83
2.02	Term	-	8793	8050	744	0	0	107	37	542	0.997	48.46
2.01	Init	-	9253	8936	318	1	0	105	73	386	0.999	41.18

Suboptimal exons with probability > 0.100

Exnum	Type	S	.Begin	...End	.Len	Fr	Ph	B/Ac	Do/T	CodRg	P....	Tscr..
S.001	Init	+	1867	1905	39	0	0	64	40	57	0.298	3.74
S.002	Init	+	2374	2442	69	0	0	55	95	-11	0.132	2.40
S.003	Intr	+	3894	4110	217	2	1	-3	-34	307	0.177	11.55
S.004	Intr	+	4352	4914	563	0	2	75	59	338	0.187	26.20
S.005	Intr	+	5005	5379	375	0	0	70	8	335	0.212	22.99
S.006	Intr	+	5442	6056	615	2	0	95	99	589	0.208	57.32



Explanation Gn.Ex : gene number, exon number (for reference) **Type** : Init = Initial exon (ATG to 5' splice site) Intr = Internal exon (3' splice site to 5' splice site) Term = Terminal exon (3' splice site to stop codon) Sngl = Single-exon gene (ATG to stop) Prom = Promoter (TATA box / initation site) PlyA = poly-A signal (consensus: AATAAA) **S** : DNA strand (+ = input strand; - = opposite strand) **Begin** : beginning of exon or signal (numbered on input strand) **End** : end point of exon or signal (numbered on input strand) **Len** : length of exon or signal (bp) **Fr** : reading frame (a forward strand codon ending at x has frame x mod 3). For example, if nucleotides 1,2,3 of the sequence are read as a codon, that's called reading frame 0. If 2,3,4 are read as a codon, that's reading frame 1. If 3,4,5 are read as a codon, that's reading frame 2, and so on. This information, together with the starting and ending positions of the exon, is sufficient to give the amino acid sequence encoded by the exon. Another use of the reading frame is that if you see two adjacent predicted exons separated by a relatively short intron which share the same reading frame, it may be worth looking at the possibility that the intervening intron is not correct, i.e. that the two exons plus the intervening intron might form one long exon (assuming there are no inframe stops in the intron, of course). **Ph** : net phase of exon (exon length modulo 3). For example, an exon of length 15 bp has net phase 0 since 15 is divisible by 3, an exon of length 16 bp has net phase 1 because 16 divided by 3 leaves a remainder of 1, an exon of length 17 bp has net phase 2, and an exon of length 18 bp has net phase 0 again. The point of this is that exons whose net phase is 0 can be omitted from the gene without disrupting the reading frame: such exons are candidates for being either 1) incorrect, or 2) alternatively spliced. **I/Ac** : initiation signal or 3' splice site score (tenth bit units; x 10). If below zero, probably not a real acceptor site. **Do/T** : 5' splice site or termination signal score (tenth bit units; x 10) If below zero, probably not a real donor site. **CodRg** : coding region score (tenth bit units) **P** : probability of exon (sum over all parses containing exon). This quantity is close to the actual probability that the predicted exon is correct. **Tscr** : exon score (depends on length, I/Ac, Do/T and CodRg scores).

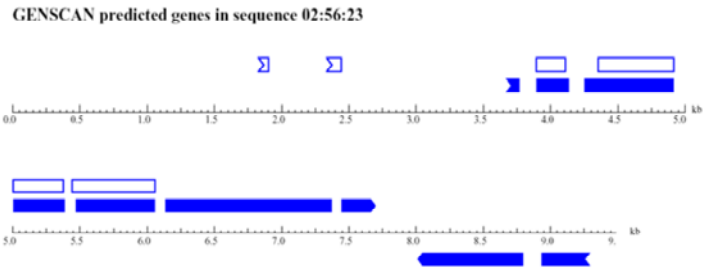
Comments The SCORE of a predicted feature (e.g., exon or splice site) is a log-odds measure of the quality of the feature based on local sequence properties. For example, a predicted 5' splice site with score > 100 is strong; 50-100 is moderate; 0-50 is weak; and below 0 is poor (more than likely not a real donor site). The PROBABILITY of a predicted exon is the estimated probability under GENSCAN's model of genomic sequence structure that the exon is correct. This probability depends in general on global as well as local sequence properties, e.g., it depends on how well the exon fits with neighboring exons. It has been shown that predicted exons with higher probabilities are more likely to be correct than those with lower probabilities.

What are the suboptimal exons?

Under the probabilistic model of gene structural and compositional properties used by GENSCAN, each possible "parse" (gene structure description) which is compatible with the sequence is assigned a probability. The default output of the program is simply the "optimal" (highest probability) parse of the sequence. The exons in this optimal parse are referred to as "optimal exons" and the translation products of the corresponding "optimal genes" are printed as GENSCAN predicted peptides. (All the data in our J Mol Biol paper and on the other GENSCAN web pages refer exclusively to the optimal parse/optimal exons.) Of course, the optimal parse does not always correspond to the actual (biological) parse of the sequence, that is, the actual set of exons/genes present. In addition, there may be more than one parse which can be considered "correct", for example, in the case of a gene which is alternatively transcribed, translated or spliced. For both of these reasons, it may be of interest to consider "suboptimal" ("near-optimal") exons as well, i.e. exons which have reasonably high probability but are not present in the optimal parse. Specifically, for every potential exon E in the sequence, the probability P(E) is defined as the sum of the probabilities under the model of all possible "parses" (gene structures) which contain the exact exon E in the correct reading frame. (This quantity is calculated as described on the GENSCAN exon probability page.) Given a probability cutoff C, suboptimal exons are those potential exons with P(E) > C which are not present in the optimal parse.

Suboptimal exons have a variety of potential uses. First, suboptimal exons sometimes correspond to real exons which were missed for whatever reason by the optimal parse of the sequence. Second, regions of a prediction which contain multiple overlapping and/or incompatible optimal and suboptimal exons may in some cases indicate alternatively spliced regions of a gene (Burge & Karlin, in preparation). The probability cutoff C used to determine which potential exons qualify as suboptimal exons can be set to any of a range of values between 0.01 and 1.00. The default value on the web page is 1.00, meaning that no suboptimal exons are printed. For most applications, a cutoff value of about 0.10 is recommended. Setting the value much lower than 0.10 will often lead to an explosion in the number of suboptimal exons, most of which will probably not be useful. On the other hand, if the value is set much higher than 0.10, then potentially interesting suboptimal exons may be missed.

GENESCAN

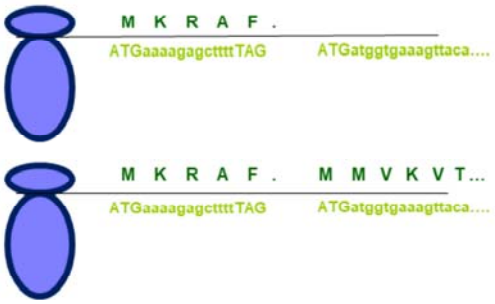


EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
OZVOJE VZDĚLÁVÁNÍ
Evropským sociálním fondu
a státním rozpočtem České republiky

Regulace translace

• Funkční význam sestřihu v nepřekládaných oblastech - důležitá regulační součást genů

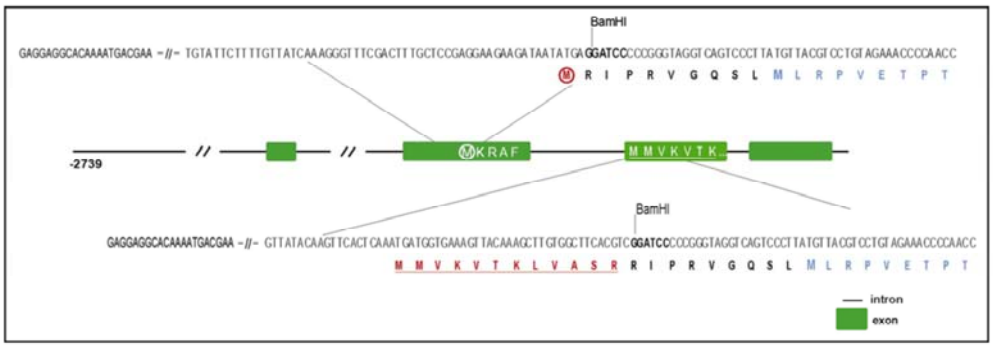
- Translační represe prostřednictvím krátkých ORF v 5'UTR
- Identifikováno např. u kukuřice (Wang and Wessler, 1998, viz doporučená lit.)
- V případě CKI1 pokus prokázat tento způsob regulace genové exprese pomocí transgenních linií nesoucích *uidA* pod kontrolou dvou verzí promotoru, zatím nepotvrzeno



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Regulace translace

- Funkční význam sestřihu v nepřekládaných oblastech - důležitá regulační součást genů
  - V případě CKI1 pokus prokázat tento způsob regulace genové exprese pomocí transgenních linií nesoucích *uidA* pod kontrolou dvou verzí promotoru, zatím nepotvrzeno



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Genové modelování

- programy pro genové modelování
  - zohledňují také další parametry, např. návaznost ORF
    - Genescan (http://genes.mit.edu/GENSCAN.html)
      velice dobrý pro predikci exonů v kódujích oblastech
      (testováno na genu *PDR9*, identifikoval všech 23 (!) exonů
    - GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/)
    - GlimmerHMM (http://http://ccb.jhu.edu/software/glimmerhmm/

EVROPSKÁ UNIE esf MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY OP Vzdělávání pro konkurenceschopnost UNIVERSITAS MASARYKIANA BRUNENSIS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

GeneMark

GeneMark™
A family of gene prediction programs provided by Mark Borodovsky's Bioinformatics Group at the Georgia Institute of Technology, Atlanta, Georgia.
What's New: - November, 2005
Prokaryotes: predicted gene database.
Prokaryotes: models for GeneMark and GeneMark.hmm.
Supported by NIH
Gene Prediction in Bacteria and Archaea
For bacterial and archaeal gene prediction, you can use the parallel combination of the GeneMark and GeneMark.hmm programs here.
If the DNA sequence of interest belongs to a species whose name is not in the list of available models, you should use either the Heuristic models option or, if the sequence is longer than 1 Mb, generate models with the self-training program GeneMarkS. Both options will allow you to generate models and then to use GeneMark.hmm and GeneMark in parallel.
Gene Prediction in Eukaryotes
For eukaryotic gene prediction, you can use the parallel combination of the GeneMark and GeneMark.hmm programs here.
Gene Prediction in EST and cDNA
To analyze ESTs and cDNAs, please follow this link.
Gene Prediction in Viruses
For viral gene prediction, or to access our virus database VIOLIN, please follow this link.
What the programs do:
Borodovsky Group
Gene Prediction Programs
GeneMark
GeneMark.hmm
Frame-by-Frame
GeneMarkS
Heuristic models
Statistics
Documented GeneMark.* usage
Help
References
Papers
FAQ
Contact
Databases of predicted genes
Prokaryotes New!
Viruses/Phages (VIOLIN)
Bioinformatics Resources
Links
Bioinformatics Studies at Georgia Tech
MS Degree Program
PhD Program
Lectures
Seminars
Center for Bioinformatics and
Eukaryotic GeneMark.hmm(1,2) (Reload this page)
References:
[1]Borodovsky M. and Lukashin A. (unpublished)
[2]Lomsadze A., Ter-Hovhannisyan V., Chernoff Y. and Borodovsky M., "Gene identification in novel eukaryotic genomes by self-training algorithm", Nucleic Acids Research, 2005, Vol. 33, No. 20, 6494-6506
Accuracy comparison
UPDATE October 2005: Added pre-built models of eukaryotic GeneMark.hmm ES-3.0 (E - eukaryotic; S - self-training; 3.0 - the version)
Listing of previous updates
Input Sequence
Title (optional):
Exit
Sequence:
Sequence file upload:
Prohlížet...
Species: Athaliana ES-30
Model description
Output Options
Email Address: (required for graphical output or sequences longer than 400000 bp)
Generate PDF graphics (screen)
Generate PostScript graphics (email)
Print GeneMark 2.4 predictions in addition to GeneMark.hmm predictions
Translate predicted genes into proteins
Run
Default
Start GeneMark.hmm
EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERZITA KARLOVA V PRAZE
LÁVÁNÍ
Evropským sociálním fondem
a státním rozpočtem České republiky

GeneMark

Result of last submission:

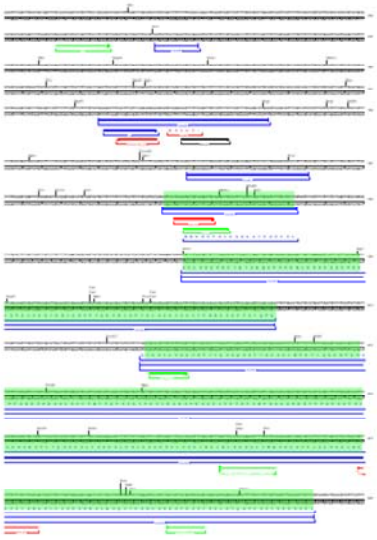
View PDF Graphical Output

GeneMark.hmm Listing

Go to: GeneMark.hmm Protein Translations

Go to: Job Submission

```
Eukariotyc GeneMark.hmm version bp 3.9 April 25, 2008
Sequence name: CKI1
Sequence length: 5043 bp
G+C content: 38.79%
Matrices file: /home/genmark/euk_ghm.matrices/athaliana_hmm3.0mod
Thu Oct  1 11:09:24 2009

Predicted genes/exons

Gene Exon Strand Exon          Exon Range          Exon      Start/End
  #    #         Type                              Length      Frame

  1    1   +   Initial          969 1025 57 1 3 - -
  1    2   +   Internal        1155       1394        240        1 3 - -
  1    3   +   Internal        1516       2175        660        1 3 - -
  1    4   +   Internal        2266       2644        379        1 1 - -
  1    5   +   Internal        2734       3317        584        2 3 - -
  1    6   +   Internal        3397       4629       1233        1 3 - -
  1    7   +   Terminal        4709       4921        213        1 3 - -
```



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
/ZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

GeneMark

Result of last submission:

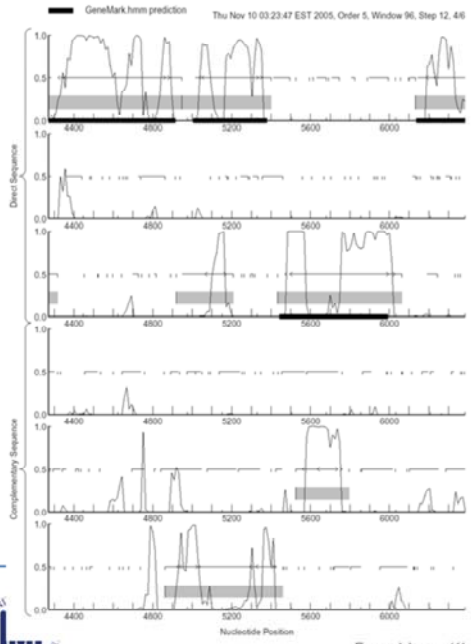
View PDF Graphical Output

GeneMark.hmm Listing

Go to: GeneMark.hmm Protein Translations

Go to: Job Submission

```
Dukariotyc GeneMark.hmm version bp 3.9 April 25, 2008
Sequence name: CKI1
Sequence length: 5043 bp
G+C content: 38.79%
Matrices file: /home/genmark/euk_ghm.matrices/athaliana_hmm3.0mod
Thu Oct  1 11:09:24 2009

Predicted genes/exons

Gene Exon Strand Exon          Exon Range          Exon      Start/End
  #    #         Type                              Length      Frame

  1    1   +   Initial          969 1025 57 1 3 - -
  1    2   +   Internal        1155       1394       240        1 3 - -
  1    3   +   Internal        1516       2175       660        1 3 - -
  1    4   +   Internal        2266       2644       379        1 1 - -
  1    5   +   Internal        2734       3317       584        2 3 - -
  1    6   +   Internal        3397       4629      1233        1 3 - -
  1    7   +   Terminal        4709       4921       213        1 3 - -
```



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
...ÁVÁNÍ
...ancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Genové homologie

- vyhledávání genů podle homologií
  - porovnávání s EST databázemi
    - BLASTN (http://www.ncbi.nlm.nih.gov/BLAST/, http://workbench.sdsc.edu/)
  - porovnávání s proteinovými databázemi
    - BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/, http://workbench.sdsc.edu/)
    - Genewise (http://www.ebi.ac.uk/Wise2/)

      porovnávají proteinovou sekvenci s genomovou DNA (po zpětném překladu), je nutná znalost aminokyselinové sekvence
  - porovnávání s homologními genomovými sekvencemi z příbuzných druhů
    - VISTA/AVID (http://www.lbl.gov/Tech-Transfer/techs/lbnl1690.html)

EVROPSKÁ UNIE esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Osnova

- Postupy „přímé“ a reverzní genetiky
  - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
  - struktura genů a jejich vyhledávání
  - genomová kolinearita a genová homologie

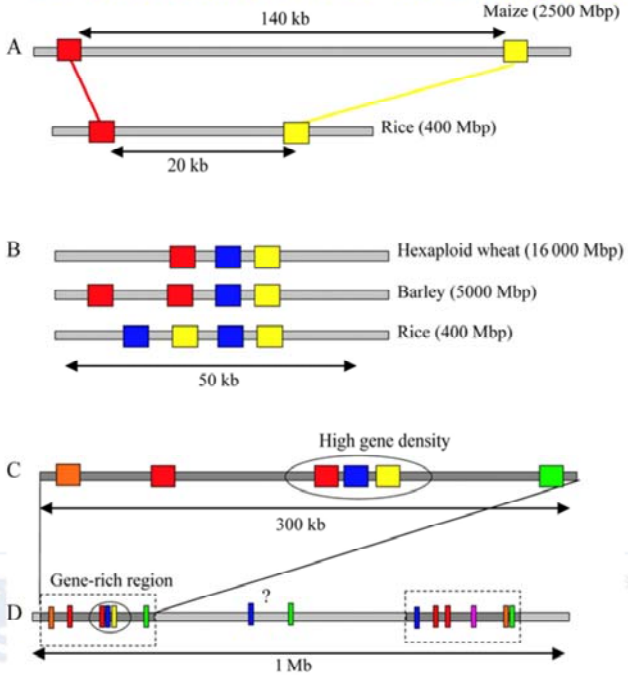
EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Genomová kolinearita

- genomy příbuzných druhů se přes značné odlišnosti vyznačují podobnostmi v uspořádání i sekvencích, možnost využití při identifikaci genů u příbuzných organizmů pomocí vyhledávání v databázích
- obecné schéma postupu při využívání geonomové kolinearity (také „komparativní genomika“) při experimentální identifikaci genů příbuzných organizmů:
  - mapování malých genomů s využitím nízkokopiových DNA markerů (např. RFLP)
  - využití těchto markerů k identifikaci orthologních genů (genů se stejnou nebo podobnou funkcí) příbuzného organizmu
  - malý genom (např. rýže, 466 Mbp) může sloužit jako vodítko, kdy jsou identifikovány molekulární nízkokopiové markery (např. RFLP) ve vazbě s genem zájmu a tyto oblasti jsou pak použity jako sonda při vyhledávání v BAC knihovnách při identifikaci orthologních oblastí velkých genomů (např. ječmene nebo pšenice, 5000, resp. 16000 Mbp)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Genomová kolinearita

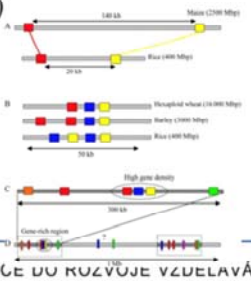


Feuillet and Keller, 2002

ŠTICE DO ROZVOJE VZDĚLÁVÁNÍ

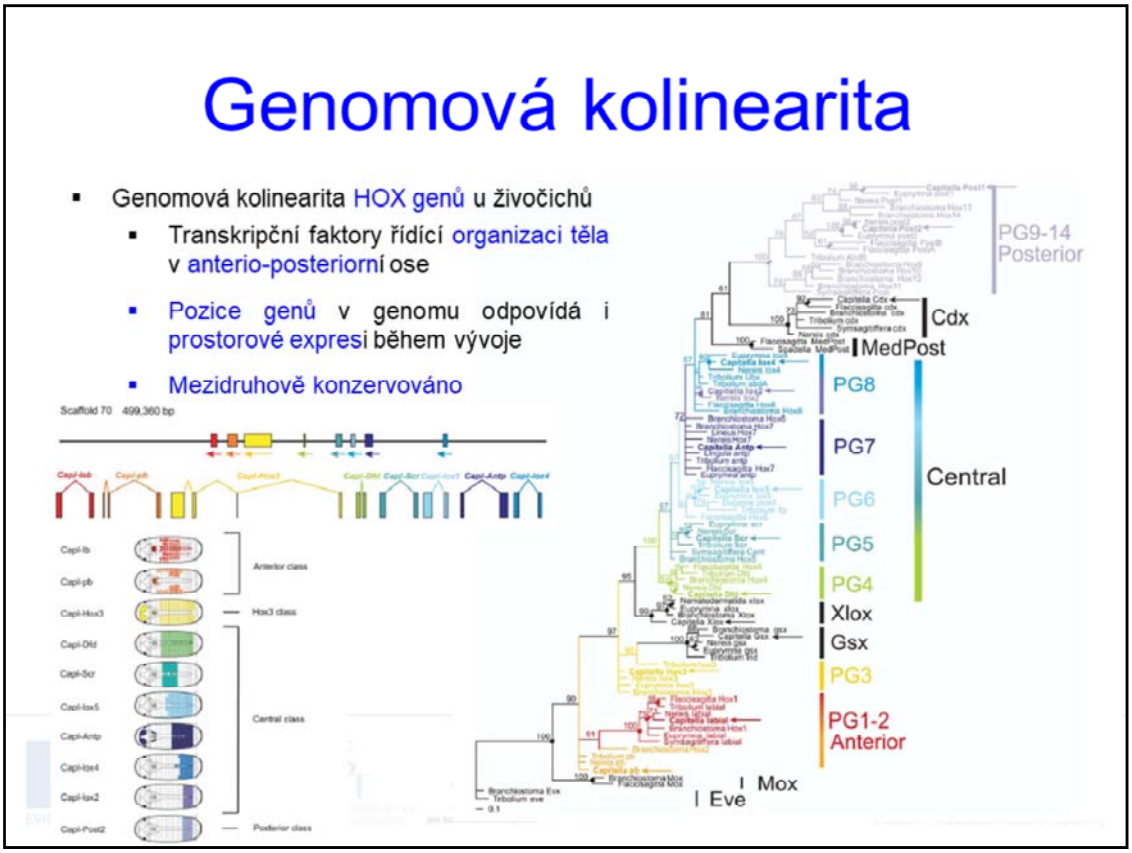
Genomová kolinearita

- zejména využitelné u trav (např. využití příbuznosti u ječmene, pšenice, rýže a kukuřice)
- malé geonomové přestavby (dalece, duplikace, inverze a translokace menší než několik cM) jsou pak detekovány podrobnou sekvenční komparativní analýzou
- během evoluce dochází u příbuzných druhů k odchylkám především v nekódujících oblastech (invaze retrotranspozonů atd.)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky



Genomic organization of the Capitella sp. I Hox cluster. A total of 11 Capitella sp. I Hox genes are distributed among three scaffolds. Black lines depict two scaffolds, which contain 10 of the Capitella sp. I Hox genes. The eleventh gene, CapI-Post1, is located on a separate scaffold surrounded by ORFs of non-Hox genes (unpublished data). No predicted ORFs were identified between adjacent linked Hox genes. Transcription units are shown as boxes denoting exons, connected by lines that denote introns. Transcription orientation is denoted by arrows beneath each box. Color coding is the same as that used in on the right-hand side for each ortholog.

The phylogenic tree on the right-hand side shows that the order of the genes on the chromozome is retained in several species (genome colinearity).

Osnova

- Postupy „přímé“ a reverzní genetiky
  - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
  - struktura genů a jejich vyhledávání
  - genomová kolinearita a genová homologie
- Experimentální identifikace genů
  - příprava genově obohacených knihoven pomocí technologie metylačního filtrování

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Metylační filtrování

- příprava genově obohacených knihoven pomocí technologie metylačního filtrování
  - geny jsou (většinou!) hypometylované, kdežto nekódující oblasti jsou metylované
  - využití bakteriálního RM systému, který rozpoznává metylovanou DNA pomocí rest. enzymů McrA a McrBC
    - McrBC rozpoznává v DNA metylovaný cytozin, který předchází purin (G nebo A)
    - pro štěpení je nutná vzdálenost těchto míst z 40-2000 bp

EVROPSKÁ UNIE esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Metylační filtrování

- příprava genově obohacených knihoven pomocí technologie metylačního filtrování
  - schéma postupu při přípravě BAC genomových knihoven pomocí metylačního filtrování:
    - příprava genomové DNA bez příměsí organelární DNA (chloroplasty a mitochondrie)
    - fragmentace DNA (1-4 kbp) a ligace adaptorů
    - příprava BAC knihovny v *mcrBC*+ kmeni *E. coli*
    - selekce pozitivních klonů
  - omezené využití: obohacení o kódující DNA o pouze cca 5-10 %

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

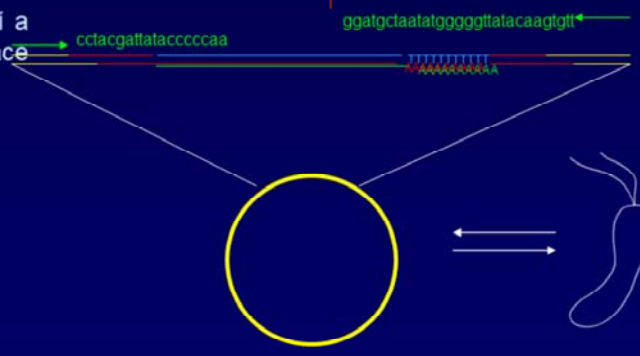
Osnova

- Postupy „přímé“ a reverzní genetiky
  - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
  - struktura genů a jejich vyhledávání
  - genomová kolinearita a genová homologie
- Experimentální identifikace genů
  - příprava genově obohacených knihoven pomocí technologie metylačního filtrování
  - EST knihovny

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

EST knihovny

- příprava EST knihoven
  - izolace mRNA
  - RT
  - ligace linkerů a syntéza druhého řetězce cDNA
  - klonování do vhodného bakteriálního vektoru
  - transformace do bakterií a izolace DNA (amplifikace DNA)
  - sekvenace s použitím primerů specifických pro použitý plasmid
  - uložení výsledků sekvenace do veřejné databáze



Shrnutí

- Postupy „přímé“ a reverzní genetiky
  - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
  - struktura genů a jejich vyhledávání
  - genomová kolinearita a genová homologie
- Experimentální identifikace genů
  - příprava genově obohacených knihoven pomocí technologie metylačního filtrování
  - EST knihovny
  - přímá a reverzní genetika (přednáška 03)

EVROPSKÁ UNIE esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Diskuse

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky