CG020 Genomika
Bi7201 Základy genomiky

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

1

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

2

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

Osnova

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

Přímá vs. reverzní genetika

Revoluce v chápání pojmu genu

Přístupy „klasické" genetiky

„Reverzně genetický" přístup

5'TTATATATATATATTAAAAAATAAAATAAAA
GAACAAAAAAGAAAATAAAATA....3'
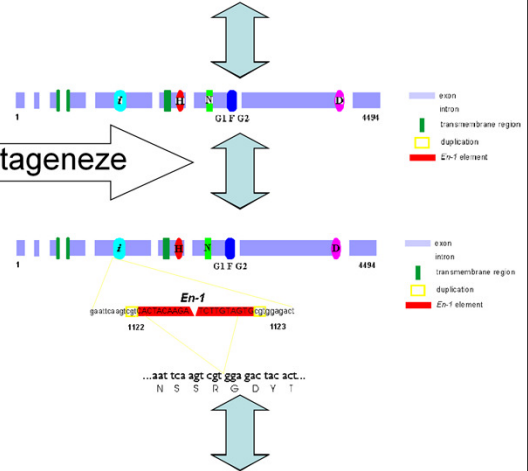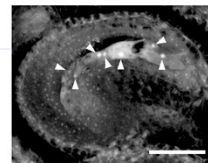
inzerční mutageneze

3 : 1

?

Identifikace role genu
ARR21

• Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*

Identifikace role genu *ARR21*

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

Identifikace role genu *ARR21*

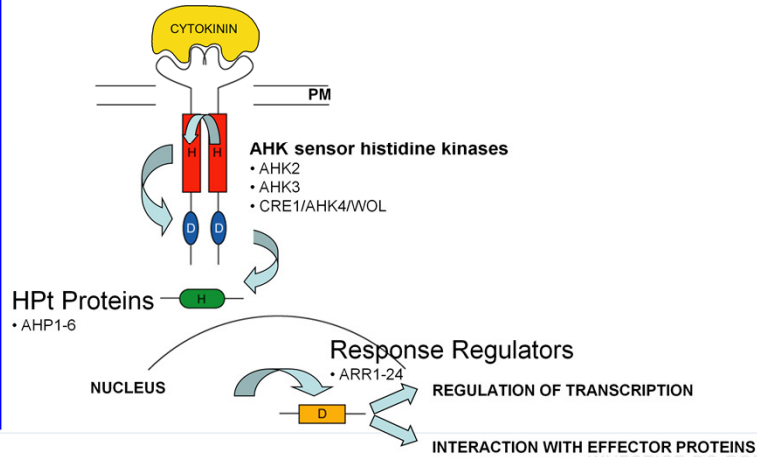• Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*

• Mutant identifikován vyhledáváním v databázi inzerčních mutantů (SINS-sequenced insertion site) pomocí programu BLAST
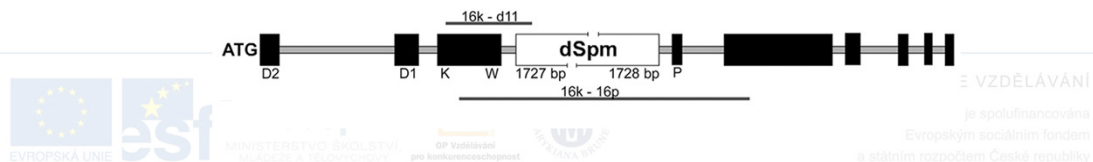
Identifikace role genu
ARR21 – izolace inz. mutanta

- **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**

9

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
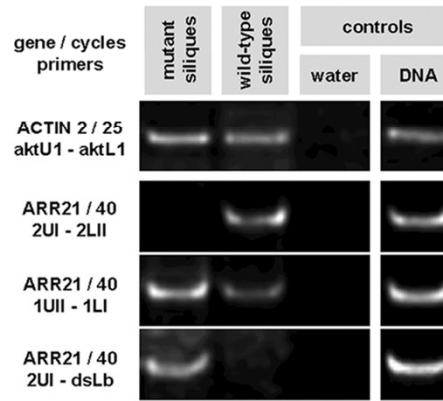
Identifikace role genu
ARR21 – analýza exprese

Standardní typ Inzerční mutant

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

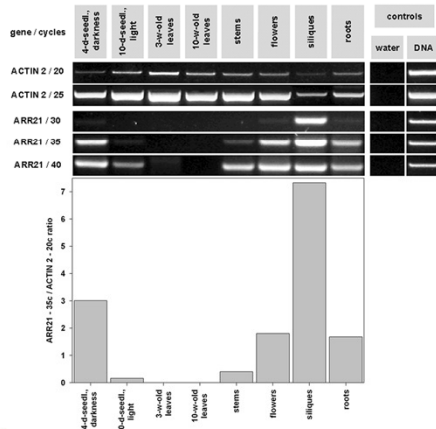• Exprese *ARR21* u standardního typu a Inhibice exprese u inzerčního mutanta potvrzena na úrovni RNA

• Analýza fenotypu inzečního mutanta

Identifikace role genu
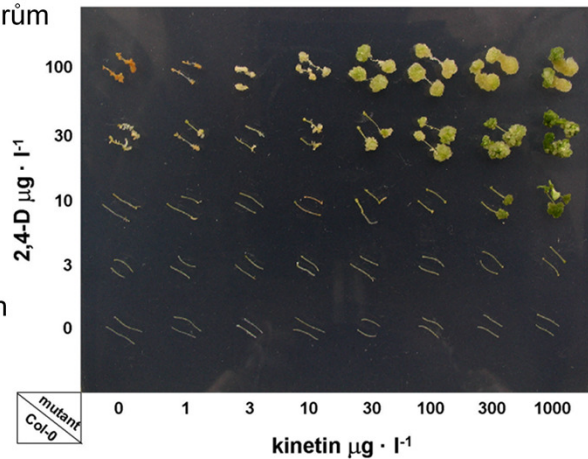## *ARR21* – analýza fenotypu mutanta

- Analýza citlivosti k regulátorům růstu rostlin
  - 2,4-D a kinetin
  - etylén
  - světlo různých vlnových délek

- Doba kvetení i počet semen nezměněn

Identifikace role genu
ARR21 – příčiny absence fenotypu

• Funkční redundance v rámci genové rodiny?

Identifikace role genu
ARR21 – příbuznost ARR genů
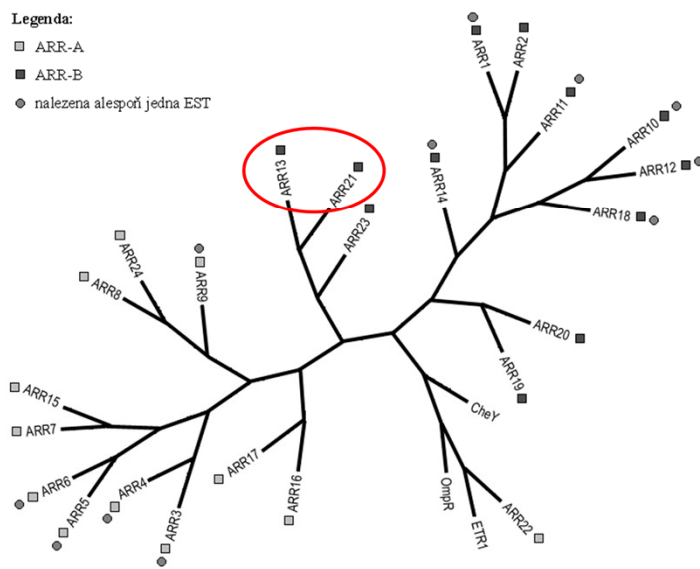
Identifikace role genu
ARR21 – příčiny absence fenotypu

• Funkční redundance v rámci genové rodiny?

• Fenotypový projev pouze za velmi specifických podmínek (?)

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

- Inzerční mutageneze v případě identifikace funkce genu *ARR21* ve vývoji *Arabidopsis* byla neúspěšná, pravděpodobně v důsledku funkční redundance v rámci genové rodiny

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

Sestřih RNA

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

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)

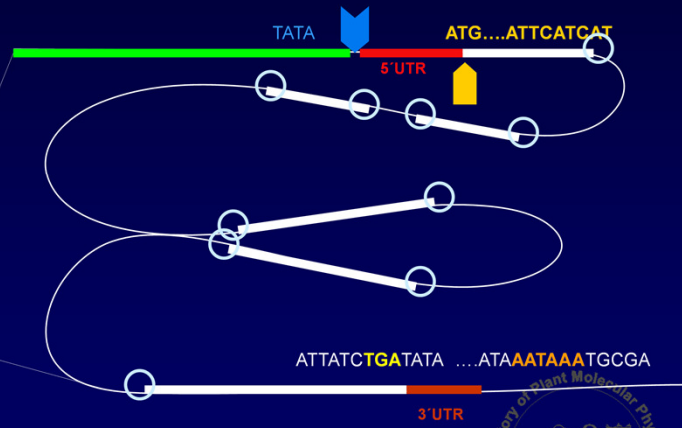INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Predikce míst sestřihu

SplicePredictor

- a method to identify potential splice sites in (plant) pre-mRNA by sequence inspection using Bayesian statistical models
(click here to access the older method using logitlinear models)

Sequences should be in the one-letter-code ({a,b,c,g,h,k,m,n,r,s,t,u,w,y}), upper or lower case; all other characters are ignored during input. Multiple sequence input is accepted in **FASTA** format (sequences separated by identifier lines of the form ">SQ;name_of_sequence comments") or in **GenBank** format.

Paste your genomic DNA sequence here:

```
GAGGAGGCACAAAATGACGAATATACAAAATGATCTTAAACAGCTAAACTATATTGGACATTTTTCGATCTCAGATATA
AAAGATTTCATTCAATATAATACTTGGATAAATACTCTTATTATTTTTCTTTAGTTTATTAAAAAAAACCTCTAATAAAT
ACGAGTTTAAGTCCACAAAATCGCTTAGACTAAAATACACCATATAATTTCAAACGATAAAGTTTACAAAAGTAATATCC
AAGTATCTCATAGTCAACATATATATAGTAATAATTAGTTGACGTATAAGAAAATAAAAATAAATAAATTAGTATCTTAT
TTTGGGTGGTGCTGACTGGTGACTGGTGACTGCAGAATGCTCGGCAAATGGAACCATATCCCAAGACATGGGTTTTAGAT
```

... or upload your sequence file (specify file name):

[Browse...]

... or type in the GenBank accession number of your sequence:

Predikce míst sestřihu

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/)

Predikce míst sestřihu

Predikce míst sestřihu

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       TTCCAAACAC^GTTAATATTT
    1906         0      +       0.99       CGGTGAACGG^GTCAGAACAT
    3582         1      +       1.00       GCCGTTCTAG^GTAATCTTGC  H
    3745         1      +       1.00       TTCCGTCCTG^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       GAAGATTAAG^GTTTTTCTCT

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^GGTTTCGACT
    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
```
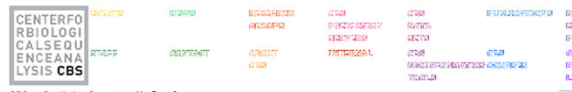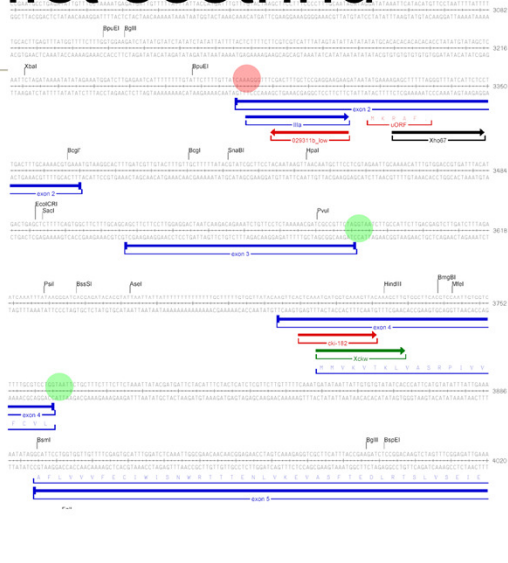
Sestřih RNA a adaptace

- 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

28

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

Identifikace genů *ab initio*

- 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

Identifikace genů *ab initio*

- 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)
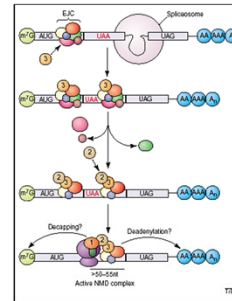
Identifikace genů *ab initio*

- 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/)

Identifikace genů *ab initio*

Identifikace genů *ab initio*

GENSCANW output for sequence CKI1

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

Sequence CKI1 : 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
```
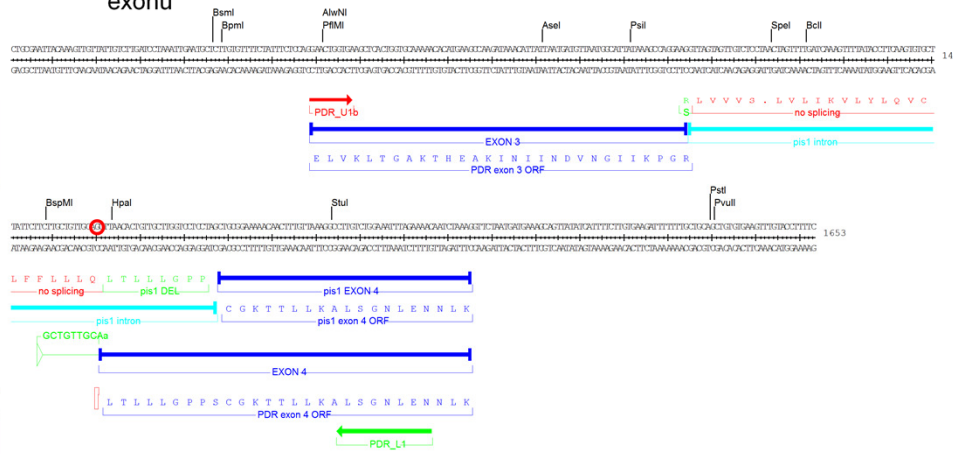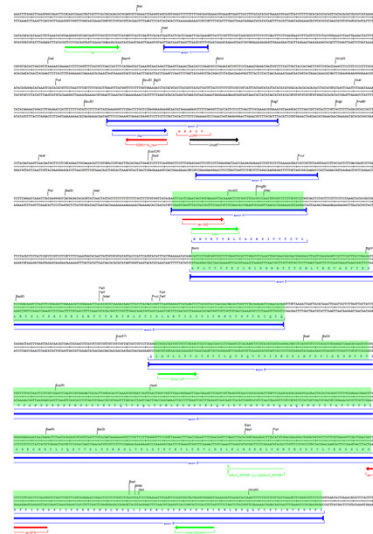
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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) **Ph** : net phase of exon (exon length modulo 3) – the position of the intron towards the ORF of the exon (0, 1 or 2) **I/Ac** : initiation signal or 3' splice site score (tenth bit units) **Do/T** : 5' splice site or termination signal score (tenth bit units) **CodRg** : coding region score (tenth bit units) **P** : probability of exon (sum over all parses containing exon) **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 si te 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 compat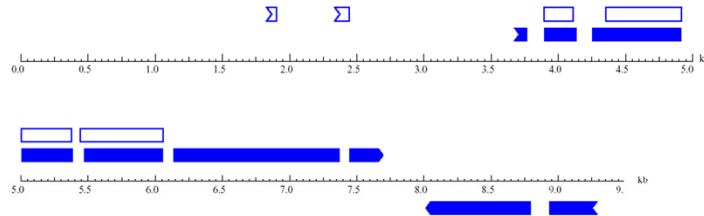ible with the sequence is assigned a probability. The default output of the program is simply the "optimal" (highest probability) parse of the sequence. The exons in this optimal parse are referred to as "optimal exons" and the translation products of the corresponding "optimal genes" are printed as 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.

Identifikace genů *ab initio*

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) **Ph** : net phase of exon (exon length modulo 3) – the position of the intron towards the ORF of the exon (0, 1 or 2) **I/Ac** : initiation signal or 3' splice site score (tenth bit units) **Do/T** : 5' splice site or termination signal score (tenth bit units) **CodRg** : coding region score (tenth bit units) **P** : probability of exon (sum over all parses containing exon) **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 si te 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.
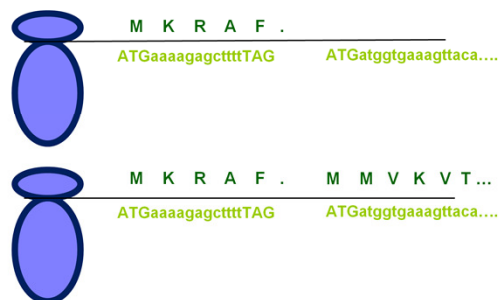
What are the suboptimal exons?

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.

Regulace transkripce

• 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

M K R A F .
ATGaaaagagcttttTAG   ATGatggtgaaagttaca….

M K R A F .   M M V K V T...
ATGaaaagagcttttTAG   ATGatggtgaaagttaca….

Regulace transkripce

• 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
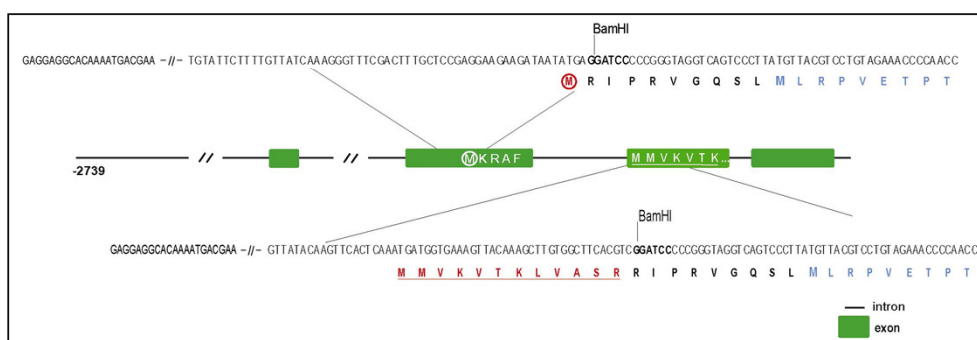
Genové modelování

- 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/)

Identifikace genů *ab initio*

Identifikace genů *ab initio*

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 - -
```

40

Identifikace genů *ab initio*

Result of last submission:
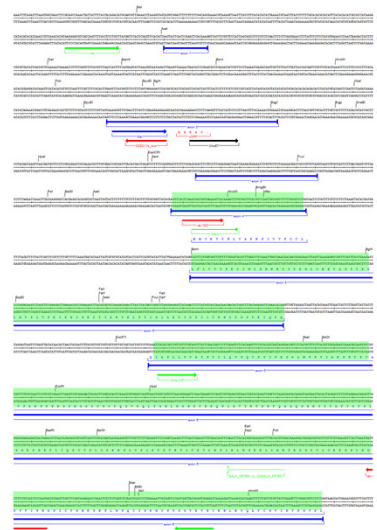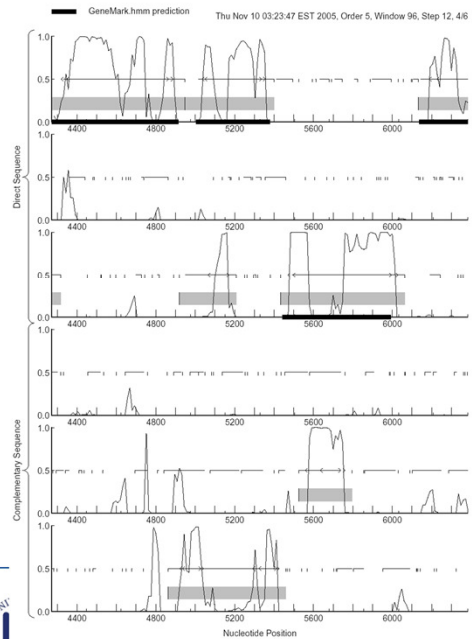
View PDF Graphical Output

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 - -
```


GeneMark.hmm prediction   Thu Nov 10 03:23:47 EST 2005, Order 5, Window 96, Step 12, 4/6

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)

42

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
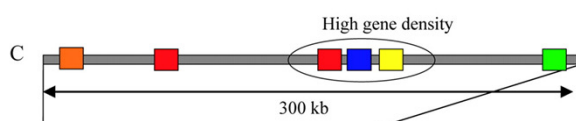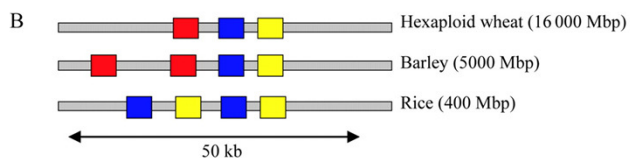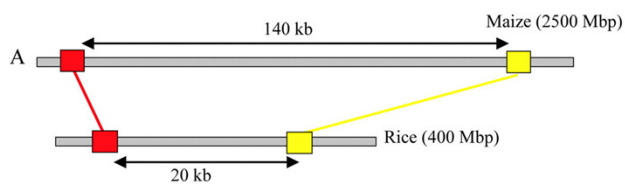
Genomová kolinearita

- 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)
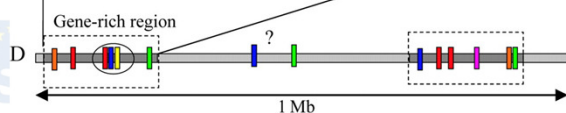
Genomová kolinearita

Feuillet and Keller, 2002

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.)
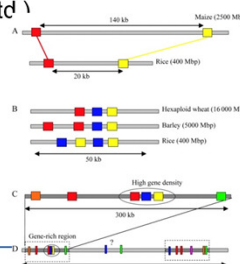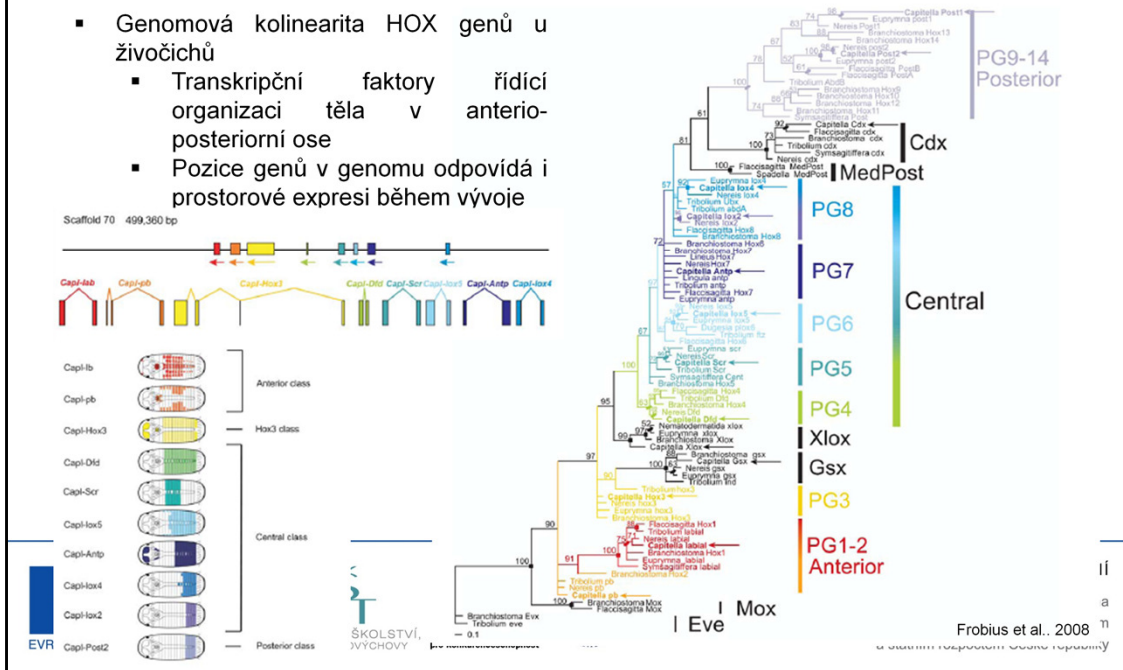
Genomová kolinearita

- Genomová kolinearita HOX genů u živočichů
  - Transkripční faktory řídící organizaci těla v anterio-posteriorní ose
  - Pozice genů v genomu odpovídá i prostorové expresi během vývoje

Frobius et al.. 2008

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

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

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 %

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**
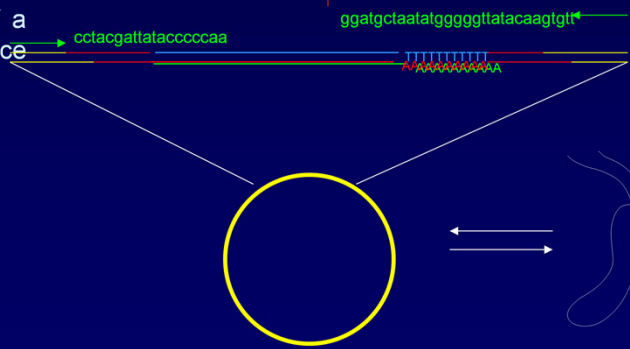 - **EST knihovny**

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

EST knihovny

- příprava EST knihoven
  - izolace mRNA
  - RT PCR
  - 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

Základy genomiky II, Identifikace genů

52

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

53

Přímá a reverzní genetika

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky

    - Změna fenotypu po mutagenezi
        - **Genetika přímá**

    - Identifikace inzerčního mutanta a analýza jeho fenotypu
        - **Genetika reverzní**

    - Analýza exprese daného genu a jeho časoprostorové specifity

Přímá a reverzní genetika - shrnutí

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky

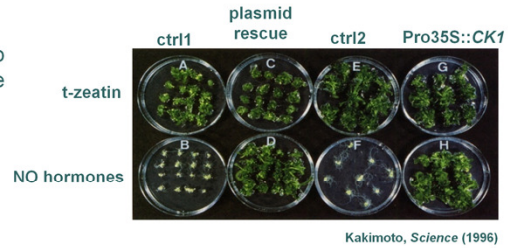 - Změna fenotypu po mutagenezi
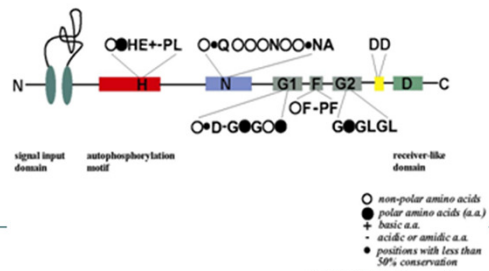   - **Genetika přímá**

Cloning of CKI1

□ CKI1 was identified via activation mutagenesis in *Arabidopsis*

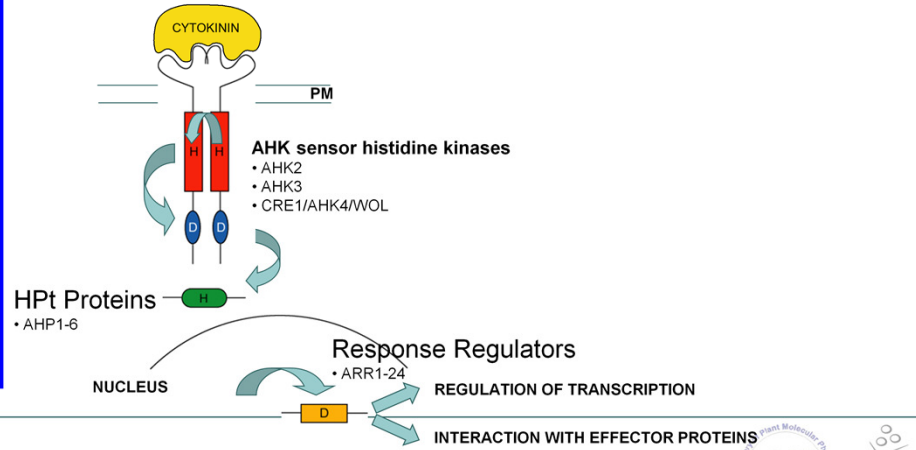- Overexpresion of *CKI1* leads to CK-like response in the hypocotyl explants

plasmid
rescue
ctrl1   ctrl2   Pro35S::*CK1*

t-zeatin

NO hormones

Kakimoto, *Science* (1996)

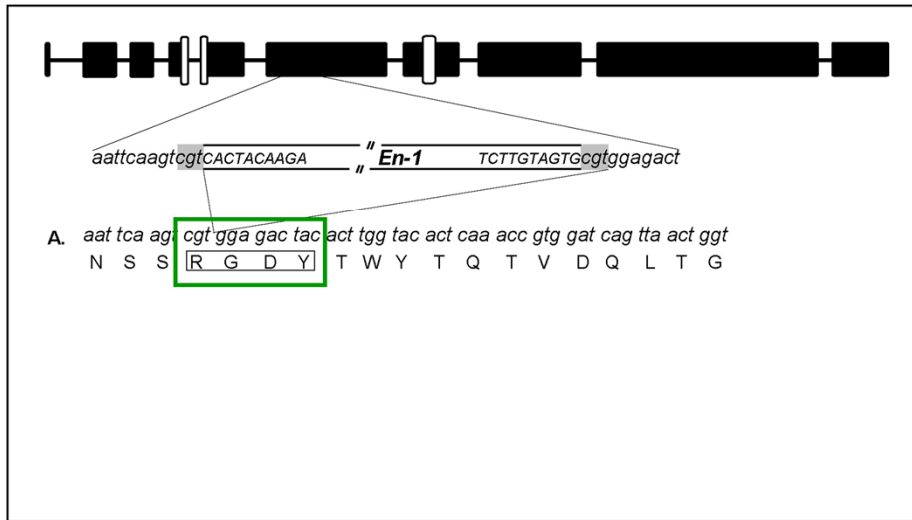- *CKI1* encodes a protein with similarity to bacterial histidine klinases

Přímá a reverzní genetika

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky

  - Změna fenotypu po mutagenezi
    - **Genetika přímá**

  - Identifikace inzerčního mutanta a analýza jeho fenotypu
    - **Genetika reverzní**

Identification of insertional *cki1* mutant allele

A. aat tca ag: cgt gga gac tac act tgg tac act caa acc gtg gat cag tta act ggt
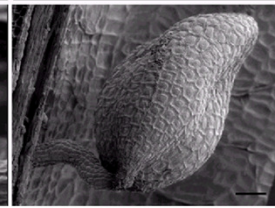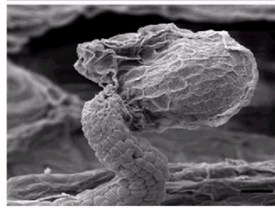   N   S   S   R   G   D   Y   T   W   Y   T   Q   T   V   D   Q   L   T   G

CKI1 Regulates Female Gametophyte Development

☐ CKI1 is necessary for proper megagametogenesis in *Arabidopsis*

CKI1/cki1-i

CKI1/CKI1
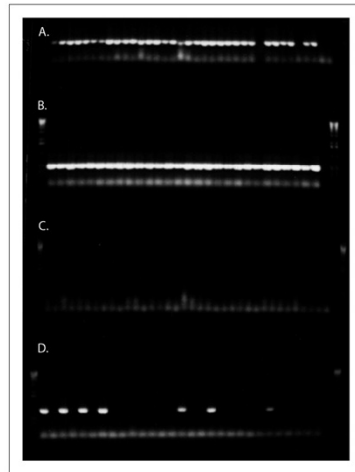

Hejátko et al., *Mol Genet Genomics* (2003)

Hormonal regulations of plant development

60

CKI1 and megagametogenesis

□ *cki1-i* is not transmitted through the female gametophyte

A. ♂ wt x ♀ *CKI1/cki1-i*

⬍ CKI1 specific primers (PCR positive control)
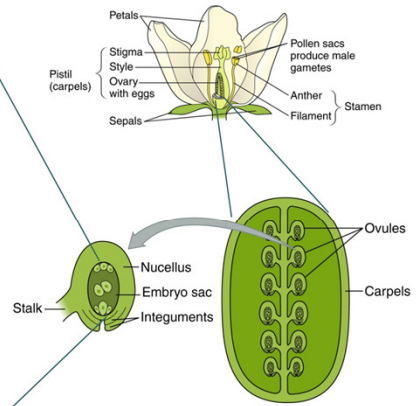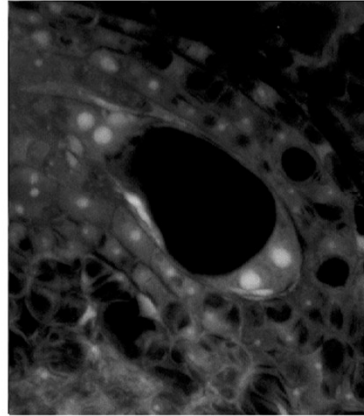
B. ♂ *CKI1/cki1-i* x ♀ wt

C. ♂ wt x ♀ *CKI1/cki1-i*

⬍ *cki1-i* specific primers

D. ♂ *CKI1/cki1-i* x ♀ wt

CKI1 and megagametogenesis

FG
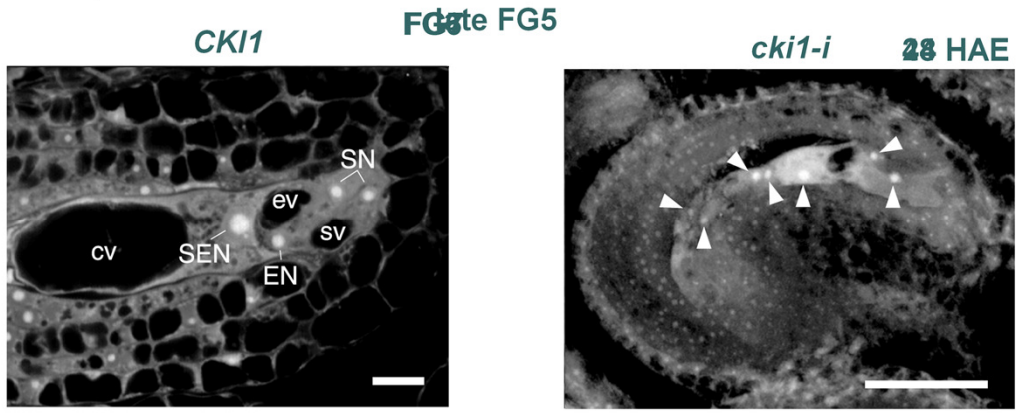
Hormonal regulation of plant development

FG1 FG2 FG3 FG4

Petals
Stigma
Style
Pistil (carpels)
Ovary with eggs
Sepals
Pollen sacs produce male gametes
Anther
Filament
Stamen

Nucellus
Embryo sac
Stalk
Integuments
Ovules
Carpels

CKI1 and megagametogenesis

CKI1   **FG6**   late FG5   *cki1-i*   28 HAE

Hejátko et al., *Mol Genet Genomics* (2003)

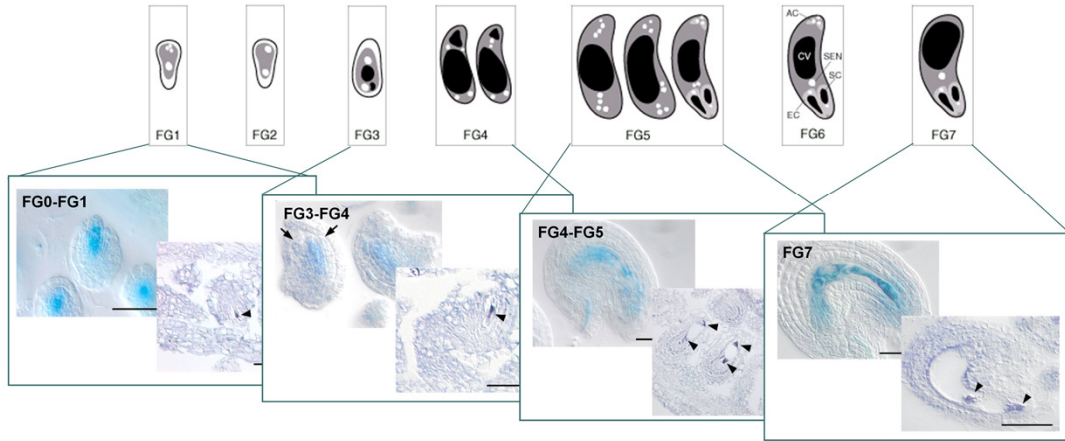Hormonal regulations of plant development

Přímá a reverzní genetika

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky

  - Změna fenotypu po mutagenezi
    - **Genetika přímá**

  - Identifikace inzerčního mutanta a analýza jeho fenotypu
    - **Genetika reverzní**

  - Analýza exprese daného genu a jeho časoprostorové specifity

CKI1 is Expressed Throughout Megagametogenesis

Hejátko et al., *Mol Genet Genomics* (2003)

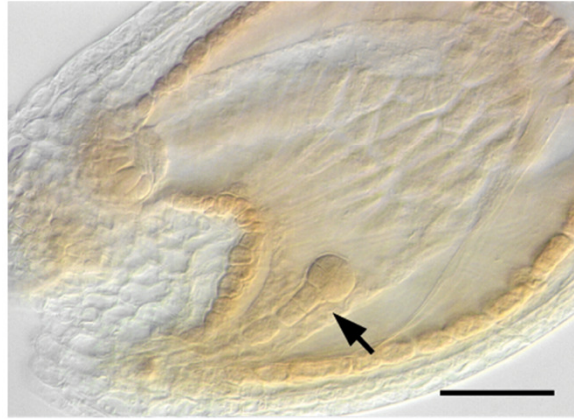Hormonal regulations of plant development

Paternal *CKI1* is Expressed in the *Arabidopsis* Sporophyte Early after Fertilization

♀ wt x ♂ Pro*CKI1:GUS*

72 HAP (hours after pollination)

Hejátko et al., *Mol Genet Genomics* (2003)

Hormonal regulations of plant development

Přímá a reverzní genetika - shrnutí

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky

    - Změna fenotypu po mutagenezi
        - **Genetika přímá**

    - Identifikace inzerčního mutanta a analýza jeho fenotypu
        - **Genetika reverzní**

    - Analýza exprese daného genu a jeho časoprostorové specifity

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

Diskuse