

CG020 Genomika

Přednáška 2

Identifikace genů

Jan Hejátko

Funkční genomika a proteomika rostlin,
Středoevropský technologický institut (CEITEC)
a

Národní centrum pro výzkum biomolekul,
Přírodovědecká fakulta,

Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.eu

MUNI
SCI



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 early/late 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)
- Frobisius, 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

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



3



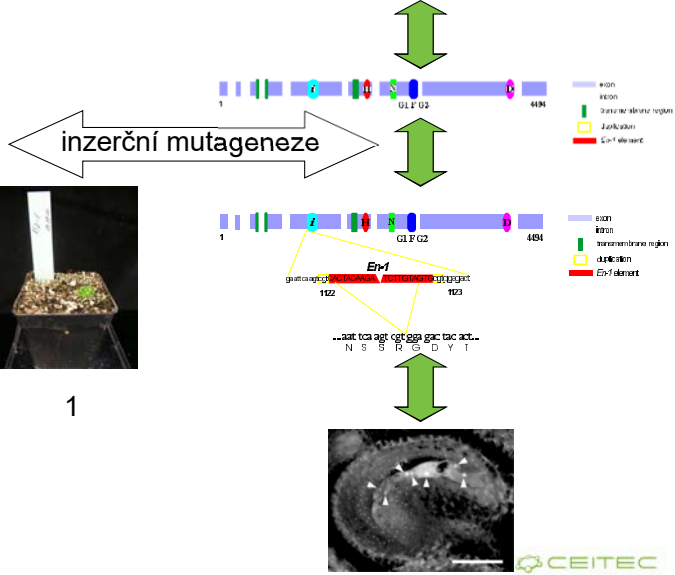
?



1

„Reverzně genetický“ přístup

5'TTATATATATATATATAAAAAATAAAATAA
AAGAACAAAAAGAAAATAAAATA...3'



5

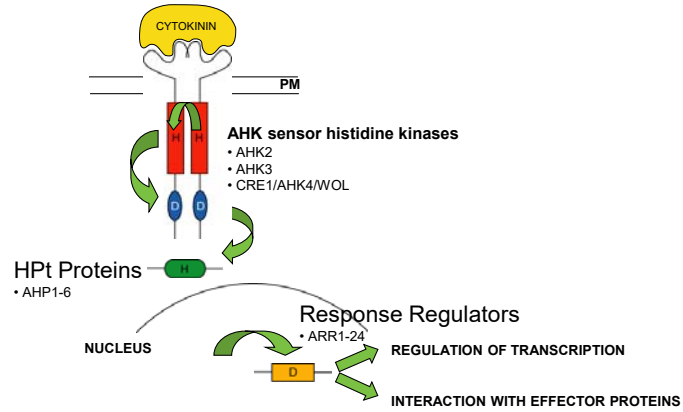
CEITEC

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

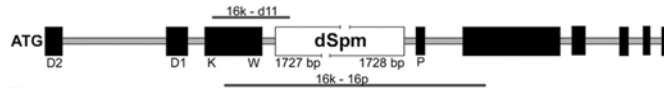
Identifikace role genu *ARR21* – izolace inz. mutanta

- vyhledávání v databázi inzerčních mutantů (SINS)

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

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

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



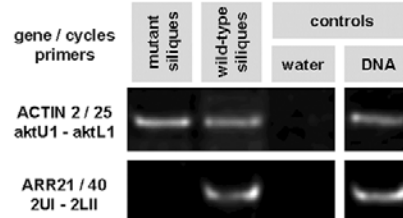
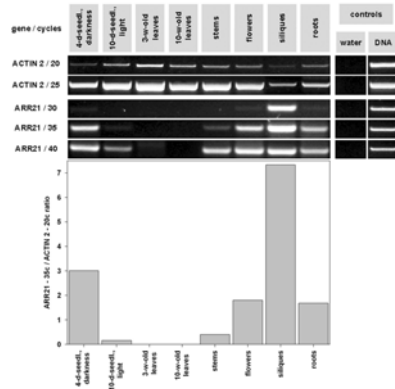
Identifikace role genu *ARR21*

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

Identifikace role genu *ARR21* – analýza exprese

Standardní typ

Inzerční mutant



Identifikace role genu *ARR21*

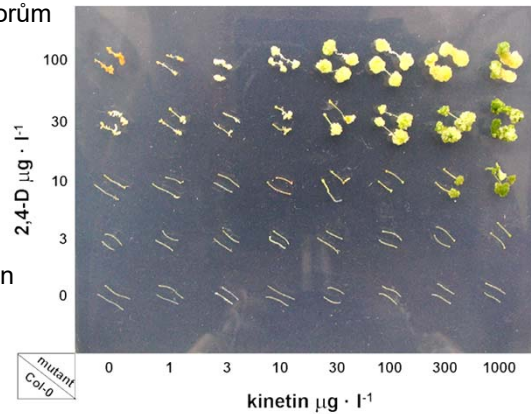
- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*
- Mutant identifikován vyhledáváním v databázi inzerčních mutantů (SINS-sequenced insertion site) pomocí programu BLAST
- Exprese *ARR21* u standardního typu a Inhibice exprese u inzerčního mutantu potvrzena na úrovni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

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



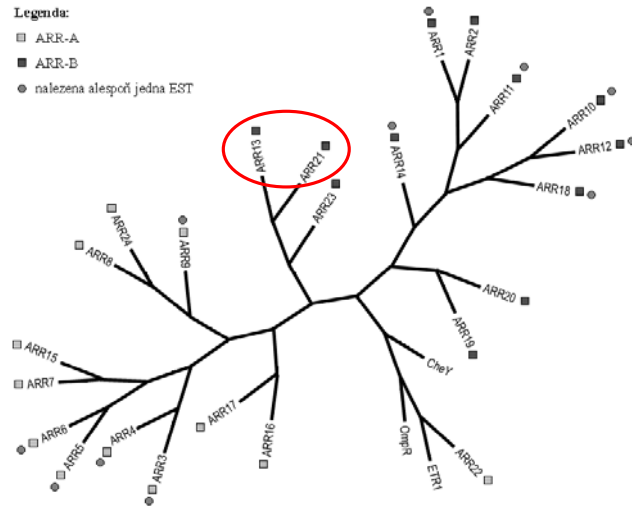
Identifikace role genu *ARR21* – příčiny absence fenotypu

- Funkční redundance v rámci genové rodiny?

Identifikace role genu *ARR21* – příbuznost ARR genů

Legenda:

- ARR-A
- ARR-B
- nalezena alespoň jedna EST



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

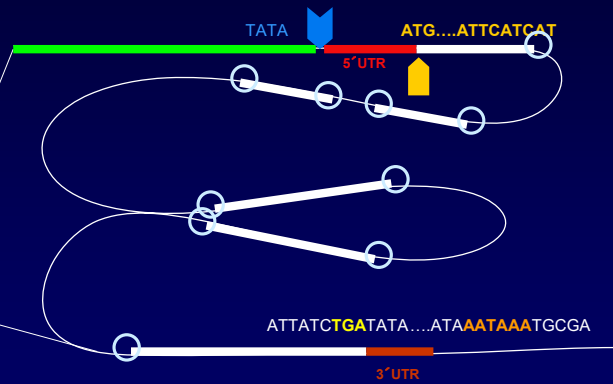
- Gen *ARR21* identifikován pomocí srovnávací analýzy genomu *Arabidopsis*
- Na základě analýzy sekvence byla předpovězena jeho funkce
- Byla prokázána místně specifická exprese genu *ARR21* na úrovni RNA
- Identifikace funkce genu pomocí inzerční mutageneze v případě *ARR21* ve vývoji *Arabidopsis* 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

Struktura genů

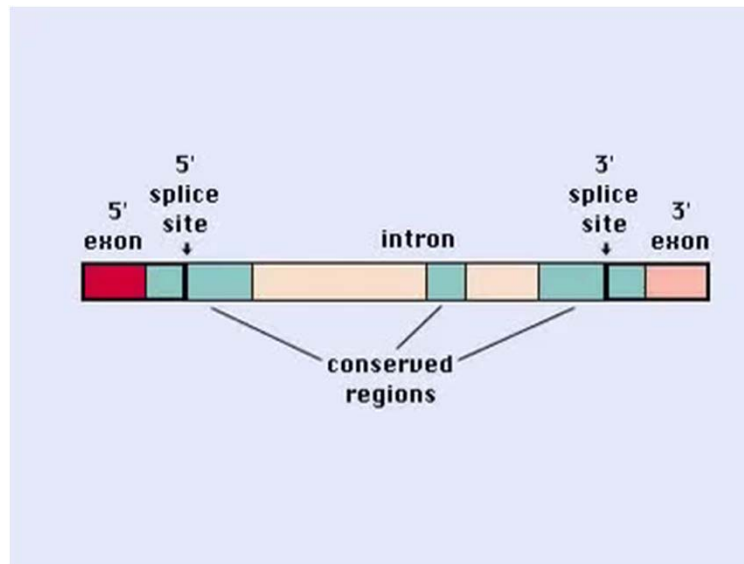
- promotor
- počátek transkripce
- 5' UTR
- počátek translace
- místa sestřihu
- stop kodon
- 3' UTR
- polyadenylační signál



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
- 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

Sestřih RNA



Identifikace Genů *Ab* *Initio*

- programy pro predikci míst sestřihu (specificita přibližně 35%)
 - NetGene2 (<https://services.healthtech.dtu.dk/services/NetGene2-2.42/>)

NetGene 2



CBS >> Prediction Servers >> NetGene2

NetGene2 Server

The NetGene2 server is a service producing neural network predictions of splice sites in human, C. elegans and A. thaliana

[Instructions](#) [Output format](#) [Abstract](#) [Performance](#)

SUBMISSION

Submission of a local file with a single sequence:

File in FASTA format

Human
 C. elegans
 A. thaliana

Submission by pasting a single sequence:

Sequence name

Human
 C. elegans
 A. thaliana

Sequence

```
SAGSAGGCCACAAAATGACGAATATACAAAATGATCTTAAACAGCTAAACTATATGGACATTTTTTCGATC
TCAGATATA
AAAGATTTTCATTCAGATATAACTTGGATAAATCTCTTATATTTTTCTTTAGTTTATTAAMAAAAACCT
CTAATAAAT
ACGAGTTTAAAGTCCACAAAATCGCTTAGACTAAAATACACCATATATTTCAAACGATAAAGTTTACAAAA
```

NOTE: The submitted sequences are kept confidential and will be erased immediately after processing.



NetGene 2

Prediction done

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

The 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'
1764	0	+		0.87	TTCGAACAG	GT	TATATTTT	
1966	0	+		0.99	CGGTGAACGG	GT	AGAACAT	
3562	1	+		1.00	GGCTCTTAG	GT	AACTTGG	
3765	1	+		1.00	TGGGTGGT	GT	AACTTGG	
4134	0	+		0.74	TCAACACAG	GT	TATAAAA	
4619	1	+		0.74	AGCAGAAAG	GT	TGTTC	
4915	0	+		0.94	CTCTCTCTG	GT	AAATAGT	
5356	0	+		0.87	TCTCAACCA	GT	AAATTTT	
5384	1	+		1.00	GATTTGGTT	GT	AAACTCT	
5809	1	+		1.00	TATCTAAG	GT	TGTCCAA	
6057	0	+		1.00	GCAGCTTTG	GT	AAACTCT	
6096	1	+		0.74	CTTCTGCA	GT	AAATCTAG	
7349	0	+		1.00	GGATCTCA	GT	AAATTTAA	
7884	0	+		0.74	GACCAAACT	GT	AAATCAA	
9363	0	+		0.74	GACATTAG	GT	TTTCTT	

Donor splice sites, complement strand

pos	3'>5'	pos	5'>3'	phase	strand	confidence	5'	exon	intron	3'
-----	-------	-----	-------	-------	--------	------------	----	------	--------	----

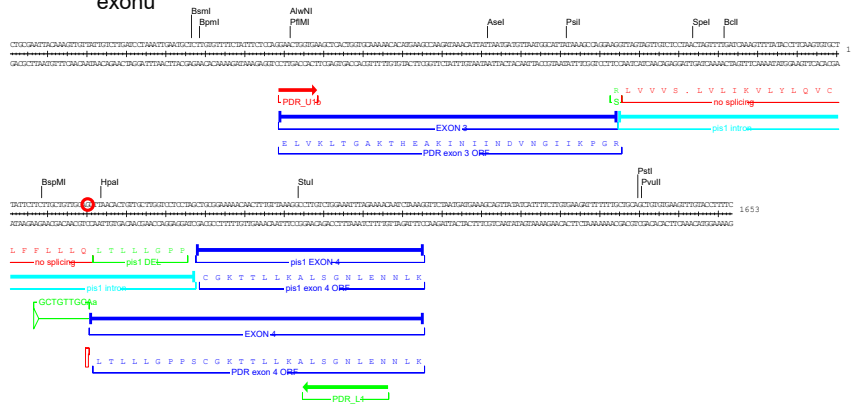
Acceptor splice sites, direct strand

pos	5'>3'	phase	strand	confidence	5'	intron	exon	3'
1213	0	+		0.59	TATTTTTAG	TT	TATGGAC	
1221	2	+		0.87	AGTTATGG	TT	ACAAGATCG	
1373	0	+		0.71	TCTCTCAC	TT	GACACAGAT	
1487	1	+		0.81	ATATTGATG	TT	TGGACATTA	
3264	0	+		0.87	GTATCAAG	TT	GGTTGGCC	
4254	0	+		1.00	TGTCTCAG	TT	ATGACACAT	
4832	2	+		0.54	AAAATGCG	TT	TCCAGTGGC	
5004	0	+		0.94	TTTTTCCG	TT	AGATACAC	
5472	1	+		0.96	AAAATTAC	TT	CTCTGTCAA	
6135	0	+		1.00	ATATTATG	TT	GTAGATTAA	
6490	1	+		0.90	AAAGTTAC	TT	TGTTGGAAA	
6744	0	+		0.59	TGTCAACG	TT	TTCTGTAG	
7447	0	+		0.96	TTCTGCG	TT	ATCCAGAAA	
7780	2	+		0.76	TCCATTTC	TT	ATACAGACA	
7786	2	+		0.92	TCAGATAC	TT	AACACATGCA	



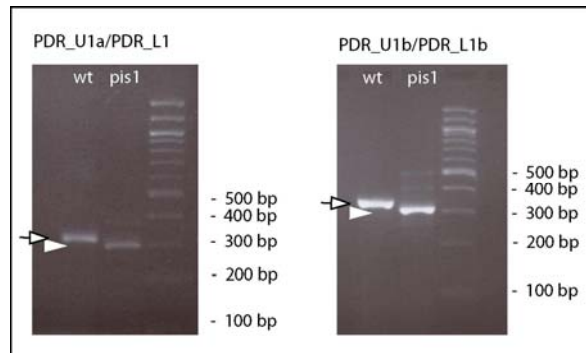
Sestřih RNA a adaptace

- 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



Sestřih RNA a adaptace

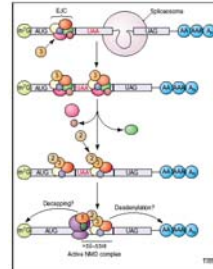
- identifikace mutanta s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5' konci 4. exonu
- analýza pomocí RT PCR prokázala přítomnost fragmentu kratšího než by odpovídalo cDNA po normálním sestřihu



Sestřih RNA a adaptace

- odchylky rozpoznávání míst sestřihu u rostlin v praxi - příklad vývojové plasticity (nejen) rostlin

- identifikace mutantu 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 sestřih s využitím nejbližšího možného místa sestřihu v exonu 4
- existence podobných obranných mechanismů prokázána i u jiných organismů (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://hollywood.mit.edu/GENSCAN.html>)
 - GeneMark.hmm (<http://opal.biology.gatech.edu/GeneMark/>)
- interní:
 - MZEF (<http://rulai.cshl.org/tools/genefinder/>)

GENESCAN

GENSCANW output for sequence CKII

```

GENSCAN 1.0 Date run: 10-Nov-105 Time: 07:26:26
Sequence CKII : 9490 bp : 36,538 C+G : Isochores I f 0 - 43 C+G%
Parameter matrix: Arabidopsis_smar

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
1.01 Init + 3708 3764 57 2 0 63 01 37 0.499 4.03
1.02 Intra + 3894 4133 240 2 0 -3 -7 327 0.713 17.32
2.03 Intra + 4255 4914 660 0 0 86 59 296 0.773 22.97
1.04 Intra + 5005 5383 379 0 1 70 91 343 0.772 31.41
1.05 Intra + 5473 6056 584 2 2 38 99 582 0.722 50.76
1.06 Intra + 6134 7348 1213 0 0 68 108 655 0.977 54.86
1.07 Term + 7448 7660 213 1 0 43 35 212 0.999 12.65
1.08 P1yA + 7910 7915 6
-----
2.03 P1yA - 7976 7971 6
2.02 Term - 8793 8050 257 0 0 107 37 542 0.997 48.46
2.01 Intra - 9253 8936 318 1 0 105 73 386 0.999 41.18

Suboptimal exons with probability > 0.100

Exon Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
-----
S.001 Init + 1847 1905 59 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 Intra + 3894 4110 217 2 1 -3 -34 307 0.177 11.95
S.004 Intra + 4352 4514 163 0 2 75 59 338 0.197 24.20
S.005 Intra + 5005 5379 375 0 0 70 9 335 0.212 22.99
S.006 Intra + 5442 6056 615 2 0 95 99 589 0.208 57.32

```



30

CEITEC

Explanation **Gn.Ex** : gene number, exon number (for reference) **Type** : Init = Initial exon (ATG to 5' splice site) Intra = 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 / initiation site) P1yA = 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.

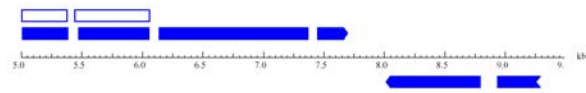
What are the suboptimal exons?

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

GENSCAN predicted genes in sequence 02:56:23

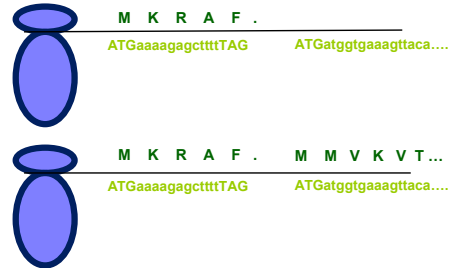


Key: ■ Initial exon ■ Internal exon ■ Terminal exon ■ Single-exon gene ■ Optimal exon □ Suboptimal exon

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 CK11 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://hollywood.mit.edu/GENSCAN.html>)
velice dobrý pro predikci exonů v kódujících oblastech
(testováno na genu *PDR9*, identifikoval všech 23 (!) exonů)
 - **GeneMark.hmm** (<http://opal.biology.gatech.edu/GeneMark/>)
 - **GlimmerHMM** (<https://ccb.jhu.edu/software/glimmerhmm/>)

GENESCAN

GENSCANW output for sequence CKII

```

GENSCAN 1.0 Date: sun 10-Nov-105 Time: 02:26:26
Sequence: CKII : 9490 bp : 36,538 C+G : Isochores 1 f 0 - 43 C+G%
Parameter matrix: Arabidopsis_smar

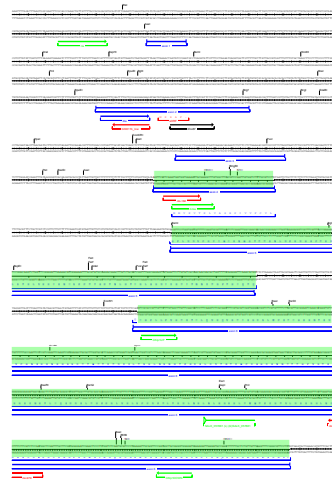
Predicted genes/exons:

Gn.Ex Type S .Begin .End .Len Fr Ph I/Ac Do/T CodRg P... Tscr..
1.00 Prom + 1497 1536 40
1.01 Init + 3708 3764 57 2 0 63 01 37 0.499 4.03
1.02 Intra + 3894 4133 240 2 0 -3 7 327 0.713 17.32
2.03 Intra + 4255 4914 660 0 0 86 59 296 0.773 22.97
1.04 Intra + 5005 5383 379 0 1 70 91 343 0.772 31.41
1.05 Intra + 5473 6056 584 2 2 38 99 582 0.722 50.76
1.06 Intra + 6134 7348 1213 0 0 48 108 655 0.977 84.86
1.07 Term + 7448 7660 213 1 0 43 35 212 0.999 12.65
1.08 P1yA + 7910 7915 6
2.03 P1yA - 7976 7971 6
2.02 Term - 8793 8050 258 0 0 107 37 542 0.997 48.46
2.01 Intra - 9253 8936 318 1 0 105 73 386 0.999 41.18

Suboptimal exons with probability > 0.100

Exon Type S .Begin .End .Len Fr Ph B/Ac Do/T CodRg P... Tscr..
s.001 Init + 1847 1905 59 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 Intra + 3894 4110 217 2 1 -3 -34 307 0.177 11.55
s.004 Intra + 4352 4514 163 0 2 75 59 338 0.197 24.20
s.005 Intra + 5005 5379 375 0 0 70 8 335 0.212 22.99
s.006 Intra + 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) Intra = 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 / initiation site) P1yA = 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 $x \bmod 3$). 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.

GeneMark

Result of last submission:

[View PDF Graphical Output](#)

GeneMarkhm Listing

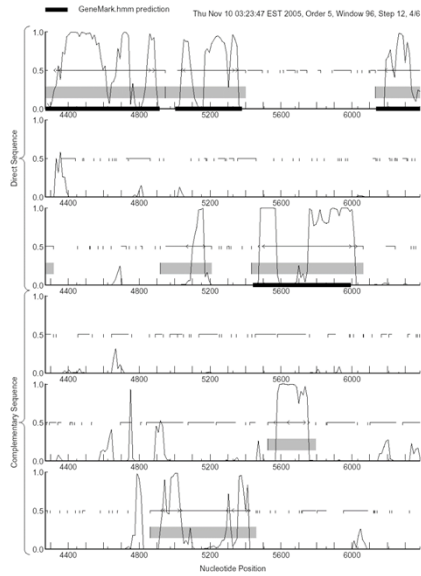
Go to: [GeneMarkhm Protein Translations](#)

Go to: [Job Submission](#)

Eukariotic GeneMarkhm version bp 3.9 April 25, 2008
 Sequence name: CK11
 Sequence length: 5043 bp
 GC content: 38.79%
 Matrices file: /home/genemark/evk_gbm.matrices/mhalima_hmm3.0mod
 Thu Oct 1 11:09:24 2009

Predicted genes/exons

Gene #	Exon #	Strand	Exon Type	Exon Range	Exon Length	Start/End Frame
1	1	+	Initial	969 1025	57	1 0 - -
1	2	+	Internal	1155 1294	140	1 0 - -
1	3	+	Internal	1516 2175	660	1 0 - -
1	4	+	Internal	2256 2544	289	1 1 - -
1	5	+	Internal	2794 3217	424	2 0 - -
1	6	+	Internal	3397 4622	1225	1 0 - -
1	7	+	Terminal	4709 4921	213	1 0 - -



Genové homologie

- vyhledávání genů podle homologí
- porovnávání s EST databázemi
 - BLASTN (<http://www.ncbi.nlm.nih.gov/BLAST/>)
- porovnávání s proteinovými databázemi
 - BLASTX (<http://www.ncbi.nlm.nih.gov/BLAST/>)
 - Genewise (<https://www.ebi.ac.uk/Tools/psa/genewise/>)

porovnávají proteinovou sekvenci s genomovou DNA (po zpětném překladu), je nutná znalost aminokyselinové sekvence
- porovnávání s homologními genomovými sekvencemi z příbuzných druhů
 - VISTA (<http://genome.lbl.gov/vista/index.shtml>)

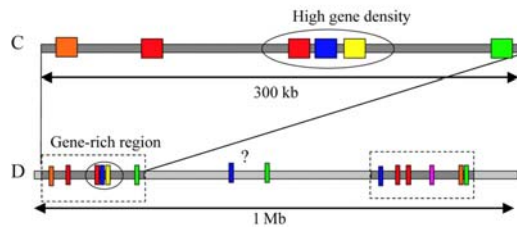
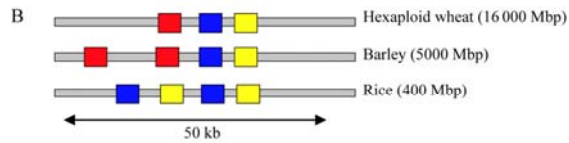
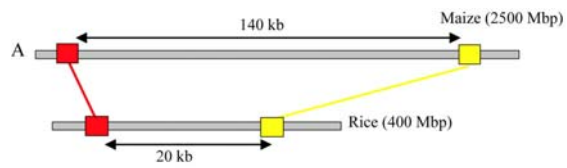
Osnova

- 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 organismů pomocí vyhledávání v databázích
- **Obecné schéma** postupu při využívání **genomové kolinearity** (také „komparativní genomika“) při experimentální identifikaci genů příbuzných organismů:
 - mapování malých genomů s využitím nízkokopiových DNA markerů (např. RFLP)
 - využití těchto markerů k identifikaci orthologních genů (genů se stejnou nebo podobnou funkcí) příbuzného organismu
 - 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)

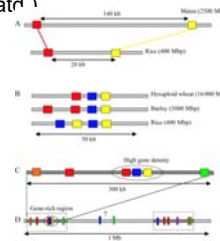
Genomová kolinearita



Feuillet and Keller, 2002

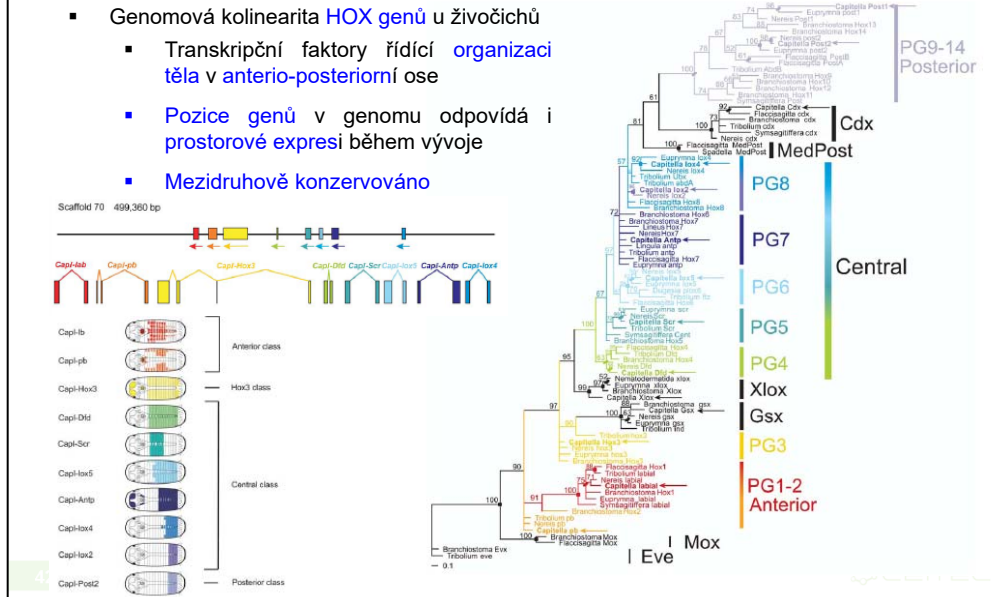
Genomová kolinearita

- zejména využitelné u trav (např. využití příbuznosti u ječmene, pšenice, rýže a kukuřice)
- malé genomové 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
 - Mezidruhově konzervováno**



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, *Cap1-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 phylogenetic tree on the right-hand side shows that the order of the genes on the chromosome 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

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

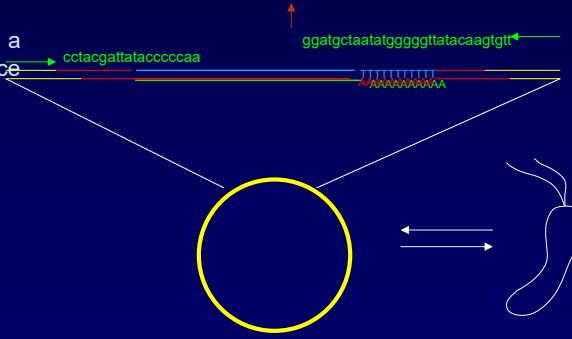
- příprava genově obohacených knihoven pomocí technologie metylačního filtrování
- **Schéma postupu** při přípravě BAC genomových knihoven pomocí metylačního filtrování:
 - příprava genomové DNA bez příměsí organelární DNA (chloroplasty a mitochondrie)
 - fragmentace DNA (1-4 kbp) a ligace adaptorů
 - 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

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



Klíčové koncepty

- **Přímá vs. reverzní genetika**
 - Gen jako faktor určující frekvenci fenotypu vs. fyzická entita, která existuje nezávisle na fenotypu
- **Identifikace genů *ab initio***
 - struktura genů a často i jejich poloha v genomu je konzervovaná
- **Experimentální identifikace genů**
 - lze připravit genově obohacené knihovny
 - EST knihovny umožňují identifikaci transkripčně aktivních genů
 - přímá a reverzní genetika (přednáška 03)

Diskuse