

Práce se sekvenčními daty (DNA)

II. část

Nejčastější typy analýz sekvencí DNA

1. Vyhledání otevřených čtecích rámců (ORF)
2. Analýza využití kodonů
3. Párové přiložení sekvencí, stanovení identity a podobnosti
4. Vyhledání motivů v sekvencích
5. Restrikční analýza *in silico*
6. Návrh sekvencí oligonukleotidů
 - primery pro PCR
 - primery pro sekvenování
 - hybridizační sondy

Vyhledání otevřených čtecích rámců (ORF)

- **ORF (Open Reading Frame)**
Sada překládaných kodonů mezi iniciačním a terminačním kodonem
- **ORF Finder**
<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>
 - ◆ Výsledek je závislý na použitém genetickém kódu
 - ◆ U prokaryot, které nemají introny je základem hledání genů
 - ◆ U eukaryot zpravidla využíváme analýzu sekvencí komplementární DNA (cDNA)

Analýza využití kodonů (codon usage)

- Využití synonymních kodonů
 - ◆ není náhodné
 - ◆ je rozdílné u různých genomů, které mají určité preferované kodony pro určité aminokyseliny

- Databáze využití kodonů

<http://www.kazusa.or.jp/codon/>

Gly	GCG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	COC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Nejčastější typy vyhledávání v sekvencích

- restriční místa
- repetice DNA
 - ◆ přímé
 - ◆ obrácené (vlásenky, vlásenky se smyčkou)
- konsenzní vzory
- uživatelsky definované vzory

Restrikční analýza *in silico*

- Restrikční endonukleázy třídy II
 - ◆ Sekvenčně specifické endonukleázy, které štěpí DNA v rozpoznávaných sekvencích
 - ◆ Přehled dostupný v databázi REBASE- Restriction Enzyme Database
<http://rebase.neb.com/rebase/rebase.html>
 - ◆ Sekvence rozpoznávacích míst
 - ◆ Producent enzymu
 - ◆ Reference
 - ◆ Komerční dostupnost
 - ◆ Sekvence genů
 - ◆ Krystalografická data
 - ◆ Citlivost k metylaci
 - ◆ REBpredictor – predikce rozpoznávací sekvence u nových enzymů
 - ◆ Rebase genomes – identifikace genů pro RE v genomech

Software pro restriční mapování

- Konstrukce restričních map na základě analýzy sekvence DNA – vyhledání restričních míst
 - ◆ Nezbytný předpoklad pro klonování
 - ◆ Interpretace RFLP polymorfizmů
 - ◆ Simulace výsledků gelové elektroforézy restričních fragmentů
- Virtuální klonování
- Vytvoření kvalitní grafiky ilustrující restriční mapy
 - ◆ RestrictionMapper (<http://www.restrictionmapper.org/>)
 - ◆ WebCutter (<http://www.firstmarket.com/cutter/cut2.html>)
 - ◆ NEB Cutter v2.0 (<http://tools.neb.com/NEBcutter2/>)
 - ◆ EMBOSS Restrict (<http://bioweb.pasteur.fr/seqanal/interfaces/restrict.html>)
 - ◆ Restriction Maps (<http://arbl.cvmbs.colostate.edu/molkit/mapper/index.html>)
 - ◆ pDRAW32 (<http://www.acaclone.com/>)

Výsledky restriční analýzy *in silico*

- Enzymy, které sekvenci neštěpí
- Enzymy, které štěpí – počet a pozice rozpoznávacích míst
- Lineární nebo kružnicová mapa sekvence se znázorněním pozice restričních míst
 - ◆ Grafika
 - ◆ Identifikace ORF a translace do proteinu

Klonování *in silico*, konstrukce vektorů

- Kombinace segmentů sekvencí
 - ◆ známé/neznámé funkce
- Plazmidy
 - ◆ přebírané z databáze
 - ◆ zpravidla známé funkce
- Inzerty – obvykle nové sekvence
 - ◆ charakterizované restriční mapou
 - ◆ charakterizované sekvencí DNA
 - ◆ charakterizované funkcí
- Nomenklatura pro konstrukty není stanovena

Vector NTI - [pBR322]

File Edit View Molecule Analyze Gel DB List Tools Window Help

General

- DNA Plasmid
- ATCC 37017
- length: 4361 bp
- storage type: Circular
- form: Circular

Function:

- CDS (3)
- Misc_features (1)
- Promoters (1)
- RBS (2)

SD SEQ
P3 P
P2 P
APr
TCr
Cla I (25)
Hind III (30)
Bam HI (3)
Pst I (4034)

pBR322
4361 bp

```

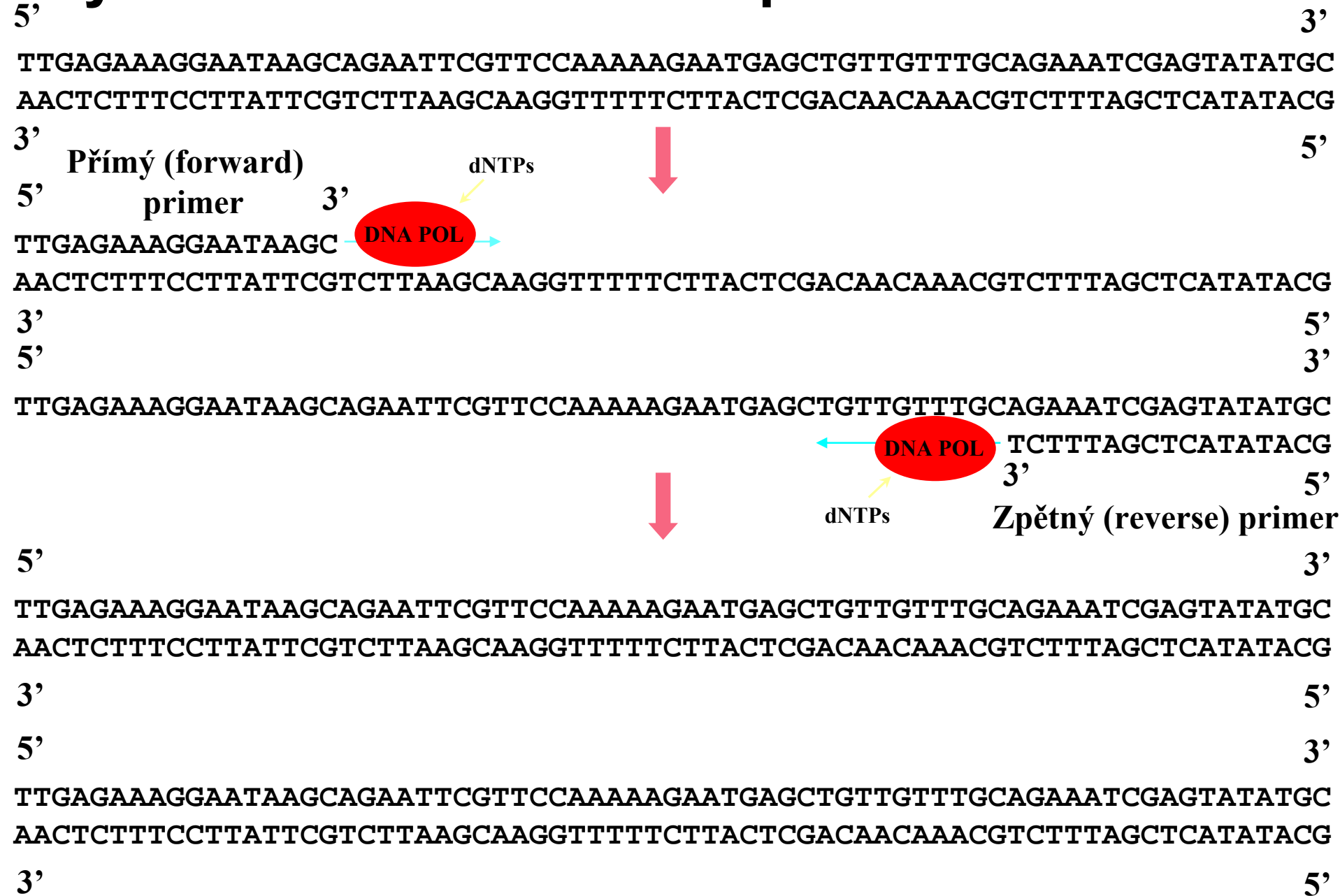
151 CTTGGTTATG CCGGTACTGC CGGGCCTCTT GCGGGATATC GTCCATTCCG
    GAACCAATAC GGCCATGACG GCCCGGAGAA CGCCCTATAG CAGGTAAGCC
201 ACAGCATCGC CAGTCACTAT GGCGTGCTGC TAGCGCTATA TGC GTTGATG
    TGTCGTAGCG GTCAGTGATA CCGCACGACG ATCGCGATAT ACGCAACTAC
251 CAATTTCTAT GCGCACCCGT TCTCGGAGCA CTGTCCGACC GCTTTGGCCG
  
```

Ready 200 bp - 1200 bp 1201 bp (E)

Navrhování sekvencí primerů

- Polymerázová řetězová reakce
- Modifikované oligonukleotidy na 5'-konci pro klonování
- Oligonukleotidy jako hybridizační sondy pro real-time PCR

Syntéza obou řetězců u specifické sekvence



Design primeru pro PCR

- Relativně snadná výpočetní záležitost –
prohledávání sekvence a identifikace krátkých
sekvencí splňujících určitá kritéria
 - ◆ Délka primeru
 - ◆ Obsah G+C
 - ◆ Teplota T_m
 - ◆ Specificita
 - ◆ Komplementarita primerových sekvencí
 - ◆ Sekvence 3'-konce

Jedinečnost primeru

- Na jedinečnost primeru a jeho hybridizační vlastnosti (annealing) má vliv délka primeru a velikost templátové DNA
 - ◆ Délka (17 – 28 bází dlouhé)
- Možná hybridizační místa primeru by se také neměla nacházet na DNA tvořících případné kontaminace vzorků

Templátová DNA

5' . . . TCAACTTAGCATGATCGGGTA . . . GTAGCAGTTGACTGTACA ACTCAGCAA . . . 3'

ϠLITVLCGL
CAGTCAACTGCTAC
ϠLITV LCGL

AGTTG
A

Primer 1 5' -TGCTAAGTTG-3'

Není jedinečný!

Primer 2 5' -CAGTCAACTGCTAC-3'

Jedinečný!

Zastoupení bází

- Zastoupení bází ovlivňuje vlastnosti hybridizace a reasociace primeru
- Žádoucí je náhodná distribuce bází bez oblastí bohatých na AT nebo GC
- Obvyklý obsah G+C, který poskytuje stabilní hybridy je 40-60 %, ale závisí také na obsahu G+C templátu

Templátová DNA

5' . . . TCAACTTAGCATGATCGGGCA . . . AAGATGCACGGGCCTGTACACAA . . . 3'
TGCCTGATCATGCT TGCCTGATCATGCT

Teplota T_m (Melting temperature)

- ◆ mají T_m teplotu 50 – 65 °C

$$T_a = 0,3 \times T_m^{\text{Primer}} + 0,7 \times T_m^{\text{Produkt}} - 25$$

kde T_m^{Primer} je hodnota T_m nejméně stabilního páru primer-matrice a T_m^{Produkt} je hodnota T_m amplifikačního produktu.

- Orientačně lze vypočítat T_a podle vztahu:

$$T_m = 2(A+T) + 4(G+C)$$

$$T_a = T_m - 5 \text{ °C}$$

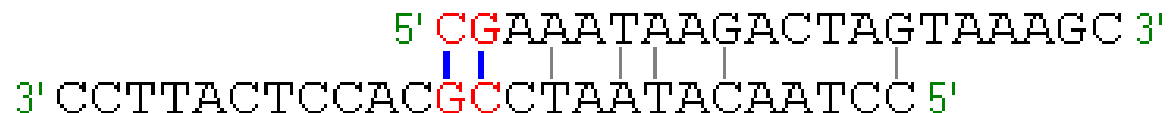
Vnitřní sekvence a struktura primeru

- nejsou komplementární navzájem na 3'-koncích, takže nevytvářejí navzájem nebo samy se sebou duplexy
- neobsahují vnitřní sekundární struktury

- ◆ Chybně navržená dvojice primerů, která vytváří stabilní duplex na 3'-konci:



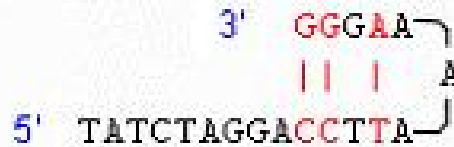
- ◆ Správně navržená dvojice primerů, která vytváří pouze málo stabilní duplex na 5'-konci; na 3'-konci je G nebo C zaručující stabilní párování s templátem:



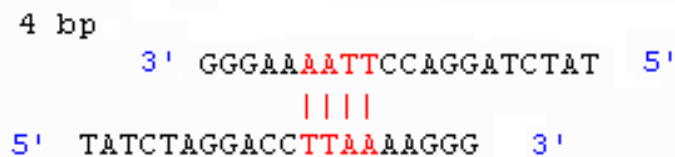
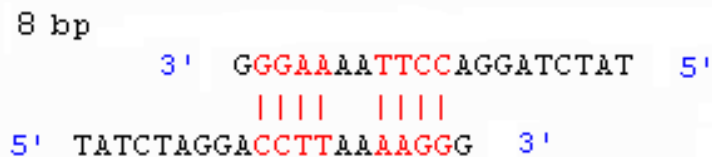
- ◆ Chybně navržený primer, vytvářející vlásenku:



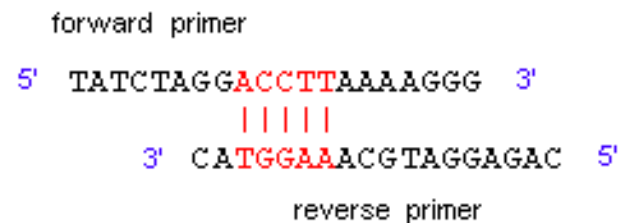
Hairpin



Self-Dimer



Dimer



Jedinečnost primerů

- ◆ na matricové DNA nemají falešná vazebná místa

- Nesprávně navržený primer s falešnými vazebnými místy na templátové DNA:

```
5'(1029) AAGGCTAGAGAAAAATATGG (1048)3'
          | | | | | | | | | | | | | | | | | |
3'(948) tttcctacccttttt-tacc (966)5'
```

```
5'(1029) AAGGCTAGAGAAAAATATGG (1048)3'
          | | | | | | | | | | | | | | | | | |
3'(1191) tttgtattgcattatataacc (1210)5'
```

```
5'(1029) AAGGCTAGAGAAAAATATGG (1048)3'
          | | | | | | | | | | | | | | | | | |
3'(395) tccatttttctttttatcct (414)5'
```

- Správně navržený primer, který nemá falešná vazebná místa na templátu:

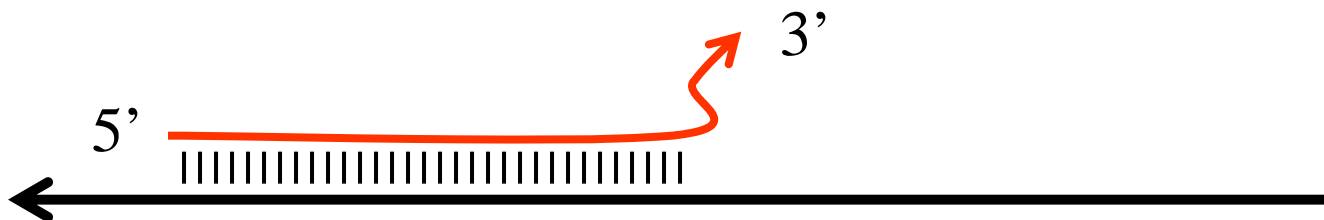
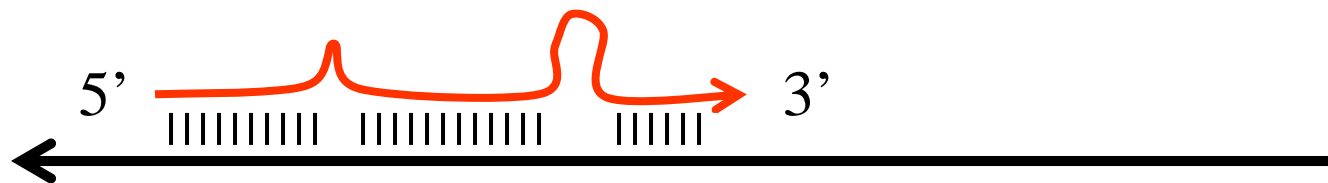
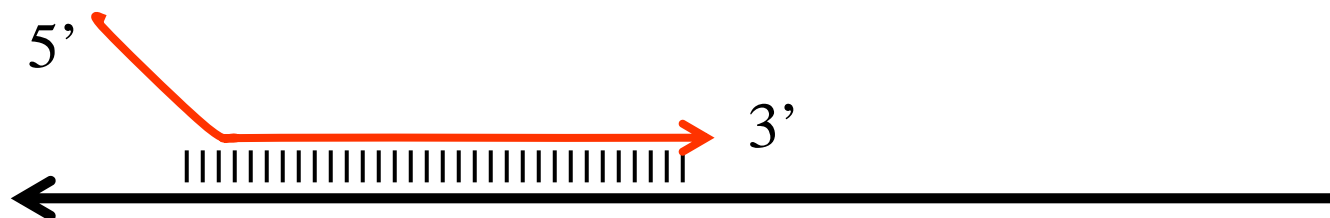
```
5'(2476) CCTAACATAATCCGCACCTCATTC (2452)3'
          | | | | | | | | | | | | | | | | | |
3'(787) taaatctattagtttacacataacc (811)5'
```

```
5'(2476) CCTAACATAATCCGCACCTCATTC (2452)3'
          | | | | | | | | | | | | | | | | | |
3'(3211) caattgtaactataactgcggttatc (3235)5'
```

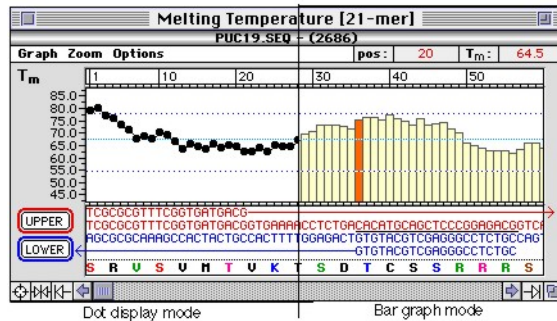
```
5'(2476) CCTAACATAATCCGCACCTCATTC (2452)3'
          | | | | | | | | | | | | | | | | | |
3'(1194) gtattgcattatataacctctgtag (1218)5'
```

```
5'(2476) CCTAACATAATCCGCACCTCATTC (2452)3'
          | | | | | | | | | | | | | | | | | |
3'(1469) atattgta-tatacgaactaaatct (1492)5'
```

Kdy je primer ještě primerem?



Pro návrh primerů se obvykle používá specializovaný software



Current Oligo
pCBlu5_seq

Sequence Length: 1842

Current Oligo (+ strand)

5' CCCGCCTGATGAATGCTCATC 3'
Length: 21-mer
5' Position: 1373
T_m: 72.1 °C
ΔG (25 °C): -42.7 kcal/mol
Degeneracy: 1
P.E.#: 492
1/E: 5.30 nmol/A₂₆₀
34.0 μg/A₂₆₀

Current Oligo (- strand)

5' GATGAGCATTTCATCAGGCG66 3'
P.E.#: 537
1/E: 4.80 nmol/A₂₆₀
31.7 μg/A₂₆₀

Selected Primers
pCBlu5_seq

pCBlu5:269U21 Upper Primer

5' CGGGCCAGATCTGGTACCCA 3'
Length: 21-mer
5' Position: 269
T_m: 76.9 °C
ΔG (25 °C): -46.1 kcal/mol
Degeneracy: 1
P.E.#: 542/542
1/E: 5.12 nmol/A₂₆₀
33.1 μg/A₂₆₀

pCBlu5:817L21 Lower Primer

5' TACCGGGTTGGACTCAAGACG 3'
Length: 21-mer
3' Position: 817
T_m: 69.5 °C
ΔG (25 °C): -41.4 kcal/mol
Degeneracy: 1
P.E.#: 502/502
1/E: 4.89 nmol/A₂₆₀
32.0 μg/A₂₆₀

Lower Primer False Priming Sites
M13MP18

Lower Primer - M13MP18:6310L19 (positive strand)
Priming efficiency of the perfect match is 428 (above the threshold)

Priming efficiency: 428 (above the threshold)

5'(6328) GGTTTTCCAGTCACGAGC (6310)3'
3'(6328) ccaaaaagggtcagtgctgc (6310)5'

Priming efficiency: 205 (above the threshold)

5'(6328) GGTTTTCCAGTCACGAGC (6310)3'
3'(626) agcaaatggtc--tgctgc (610)5'

Priming efficiency: 194 (above the threshold)

5'(6328) GGTTTTCCAGTCACGAGC (6310)3'
3'(808) gtaataatgggtcagtgctgc (790)5'

Priming efficiency: 185 (above the threshold)

5'(6328) GGTTTTCCAGTCACGAGC (6310)3'
3'(5125) tc taagtgggtcagtg--tgctgc (5108)5'

Priming efficiency: 121

5'(6328) GGTTTTCCAGTCACGAGC (6310)3'
3'(5989) agaaaatgggtc--gctcgc (5971)5'

Lower Primer - M13MP18:6310L19 (negative strand)
Priming efficiency of the perfect match is 428 (above the threshold)

Priming efficiency: 76

5'(6328) GGTTTTCCAGTCACGAGC (6310)3'
3'(5744) ccaaaaagcgggaac tgc (5762)5'

PCR
pCBlu5_seq

Optimal Annealing Temperature: 58.3° (Max: 72.0°)

	Position and Length	T _m [°C]	GC [%]	P.E.#
Product	1352	88.0	51.3	-----
Upper Primer	37 21	72.2	47.6	452
Lower Primer	1368 21	79.9	57.1	506

Product T_m - Upper Primer T_m: 15.8
Primers T_m difference: 7.6

	Concentration	
Upper Primer	200.0	nM
Lower Primer	200.0	nM
Monovalent Cation	50.0	mM
Free Mg[2+]	0.7	mM

Total Na[+] Equivalent: 155.8

Terminal stability of the Lower Primer is too high.

Počítačový návrh primerů

- Umožňuje řada molekulárně biologických programů
- Některé jsou volně dostupné na internetu
 - ◆ GCG
 - ◆ Primer3
 - ◆ OligoExplorer
 - ◆ BioTools
 - ◆ WebPrimer
 - ◆ PrimerBlast
- Kalkulátory vlastností primerů
 - ◆ Oligo Analyzer
(<http://eu.idtdna.com/SciTools/SciTools.aspx?cat=DesignAnalyze>)
 - ◆ BioMath (<http://www.promeqa.com/biomath/calc11.htm>)

Primer 3 (<http://frodo.wi.mit.edu/primer3/input.htm>)

Primer3 Input (version 0.4.0) - Mozilla Firefox

Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda

http://frodo.wi.mit.edu/primer3/input.htm

Primer3 Input (version 0.4.0)

Primer3 (v. 0.4.0) Pick primers from a DNA sequence.

[Checks for mispriming in template.](#) [disclaimer](#) [Primer3 Home](#)
[Primer3plus interface](#) [cautions](#) [FAQ/WIKI](#)

Paste source sequence below (5'→3', string of ACGTNacgtn -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a [Mispriming Library \(repeat library\)](#):

```
>SA44kb001 [org=Staphylococcus aureus] [strain=CCM 885] [clone=7/IV] Staphylococcus aureuss
EcoRI-clone from common 44 kb SmaI fragment
GAATTCAAAACCCAGCAAAAAGCTGTGAAAAAGCCATTACCAAGTAAAGATAATTTGGCTATATTGTATGGAGAAGGATTTTCATATTTGTAAGGCG
AATTATTTGGAAAACATCGACATGGTGAAGATTGTCTGTTCTGTTTAAAGGTTTTAAGTGATTAATCAAGCACACTCAAATAGTGTTATAATTAT
AAATGAATATGGTTTGGATAAGTCTGAGACAATGCATGTTTCAGGCTTTAATTGTGTATAAAAGTTTTGGTGATTGCATAAGAGATGGCGGTAATA
AATGTTATTATTAAGTGTGCACGCAGTATCATTAGTTATAAAAATGTAGCTGTTAAAAAGTCAAAAATACATCGAATGTAGTTAGGCATATAATATA
```

<input checked="" type="checkbox"/> Pick left primer, or use left primer below:	<input type="checkbox"/> Pick hybridization probe (internal oligo), or use oligo below:	<input checked="" type="checkbox"/> Pick right primer, or use right primer below (5' to 3' on opposite strand):
<input type="text"/>	<input type="text"/>	<input type="text"/>

Sequence Id: A string to identify your output.

Targets: E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the [source sequence](#) with [and]: e.g. ...ATCT[CCCC]TCAT.. means that primers must flank the central CCCC.

Hotovo

Pick Primers Reset Form

Sequence Id: A string to identify your output.

Targets: E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the [source sequence](#) with [and]: e.g. ...ATCT[CCCC]TCAT.. means that primers must flank the central CCCC.

Excluded Regions: E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. Or mark the [source sequence](#) with < and >: e.g. ...ATCT<CCCC>TCAT.. forbids primers in the central CCCC.

Product Size Ranges 150-250 100-300 301-400 401-500 501-600 601-700 701-850 851-1000

Number To Return Max 3' Stability

Max Repeat Mispriming Pair Max Repeat Mispriming

Max Template Mispriming Pair Max Template Mispriming

Pick Primers Reset Form

General Primer Picking Conditions

Primer Size Min: Opt: Max:

Primer Tm Min: Opt: Max: Max Tm Difference: Table of thermodynamic parameters: ▼

Product Tm Min: Opt: Max:

Primer GC% Min: Opt: Max:

Oligo

Oligo 7 Demo - Human eIF-4E.seq

File Edit Analyze Search Select Change View Window Help

Sequence

File: Human eIF-4E.seq

DNA Sequence		Selected Oligo	Position	Length
Sequence Length:	1868 nt	<input checked="" type="checkbox"/> Forward Primer	997	22
Reading Frame:	+1	<input checked="" type="checkbox"/> Reverse Primer	1061	21
Current Oligo Length:	21 nt	<input checked="" type="checkbox"/> Upper Oligo	956	21
Position:	956	<input type="checkbox"/> Lower Oligo	---	---
t_m :	49.1°C	<input checked="" type="checkbox"/> PCR Product	[85,---] nt	

#	Feature	Location
1	source	-18..1850

pos: tm:

950 960 970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080

ACATACAGATTTTACCTATCC · TGGCATTTCATATACTTTACAGG ·

ATTACCATTAATTACATACAGATTTTACCTATCCACAATAGTCAGAAAAACAATTGGCATTTCATATACTTTACAGGAAAAAAAAATTCTGTTGTTCCATTTTATGCAGAAGCATATTTTCTGCTGTTTGAAAAGATTATGATGCAT
TAATGGTAATTAATGTATGTCTAAAAATGGATAGGTGTTATCAGTCTTTTGTGAACCGTAAAGATATGAAATGTCCTTTTTTTAAGACAACAAGGTAATAACGTCTTCGTATAAAAACGACCAAACCTTTCTAATACTACGTA

CGACCAAACCTTTCTAATACTA

I T I N Y I Q I L P I H N S Q K T T W H F Y T L Q E K K F C C S I L C R S I F C W F E R L - C I

Ready...

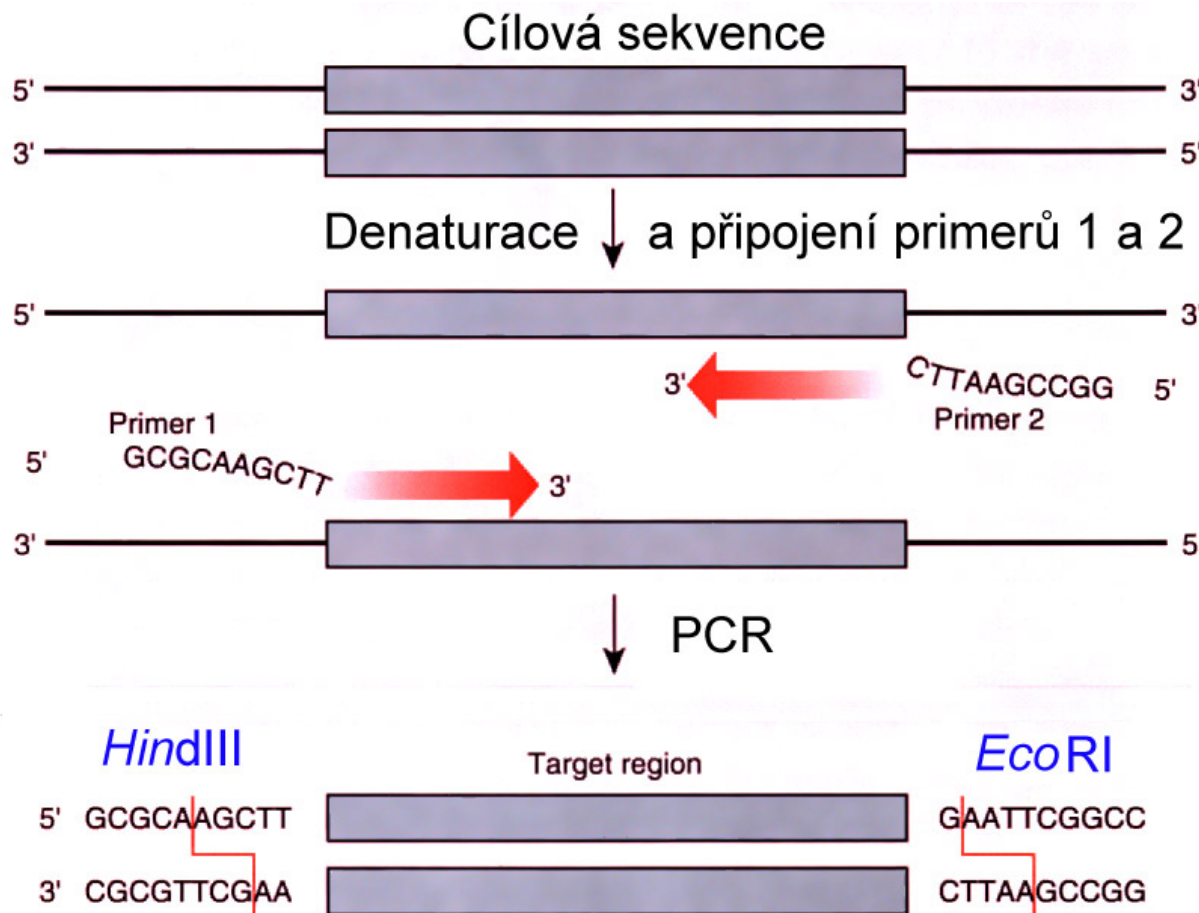
Výsledky

- Výběr optimálního páru primerů
- Sekvence primerů
- Délka primerů a hodnota T_m
- Velikost produktu
- Posouzení sekundárních struktur
- Podmínky reakce
- Alternativní primery

Pokročilý návrh primerů

- Alelově specifické primery
- Molekulární diagnostika
- Vícenásobné detekce - primery pro multiplex PCR
 - ◆ Zajištění kompatibility primerů v reakci
- Konsenzní primery
 - ◆ Pro klonování
 - ◆ Pro PCR-RFLP (např. 16S rRNA)
 - ◆ Vyžaduje identifikaci konzervativních oblastí na základě mnohonásobných přiložení sekvencí (multiple alignment)
- Primery pro modifikaci konců produktů PCR

Modifikace konců DNA, Připojení sekvencí prostřednictvím 5'-konců primerů



„sticky foot“

- Přidávané sekvence
 - ◆ RE místa
 - ◆ Promotory
 - ◆ Terminátory
 - ◆ Translační signály

Další technologie vyžadující návrh oligonukleotidů

- Real-time PCR
 - ◆ TaqMan
 - ◆ Molecular Beacons
- Primer extension
- Sekvenování
 - ◆ Sangerovo sekvenování
 - ◆ Pyrosekvenování
- Ligázová řetězová reakce
- Oligonukleotidy pro microarrays