# Práce se sekvenčními daty (DNA)

# Práce se sekvenčními daty

- **Sekvence –** zápis posloupnosti jednoznačných znaků odpovídajících jednotlivým zbytkům (monomerům), které se nacházejí v odpovídající posloupnosti v dané makromolekule
  - **DNA nebo RNA od 5‘-konce k 3‘-konci**
  - **protein od N-konce k C-konci**
- **používají se jednopísmenové kódy dle pravidel IUPAC (tabulka)**
  http://orion.sci.muni.cz/kgmb/bioinformat/seq_samples.htm

# Příklady formátů sekvencí

- surová data – elektroforetogramy ve formátu *.abi, *.ab1 nebo *.scf
  - prohlížeče Chromas, ABIView
- formát FASTA *.nt, *.aa nebo *.fa
- formát GenBank
- software pro převod formátů sekvencí, formátování a manipulaci s daty, např. SMS – The Sequence Manipulation Suite v2 http://www.bioinformatics.org/sms2/

# 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/

The Human Codon Usage Table

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gly | GGG | 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 | CGC | 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

- restrikční 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 restrikční mapování

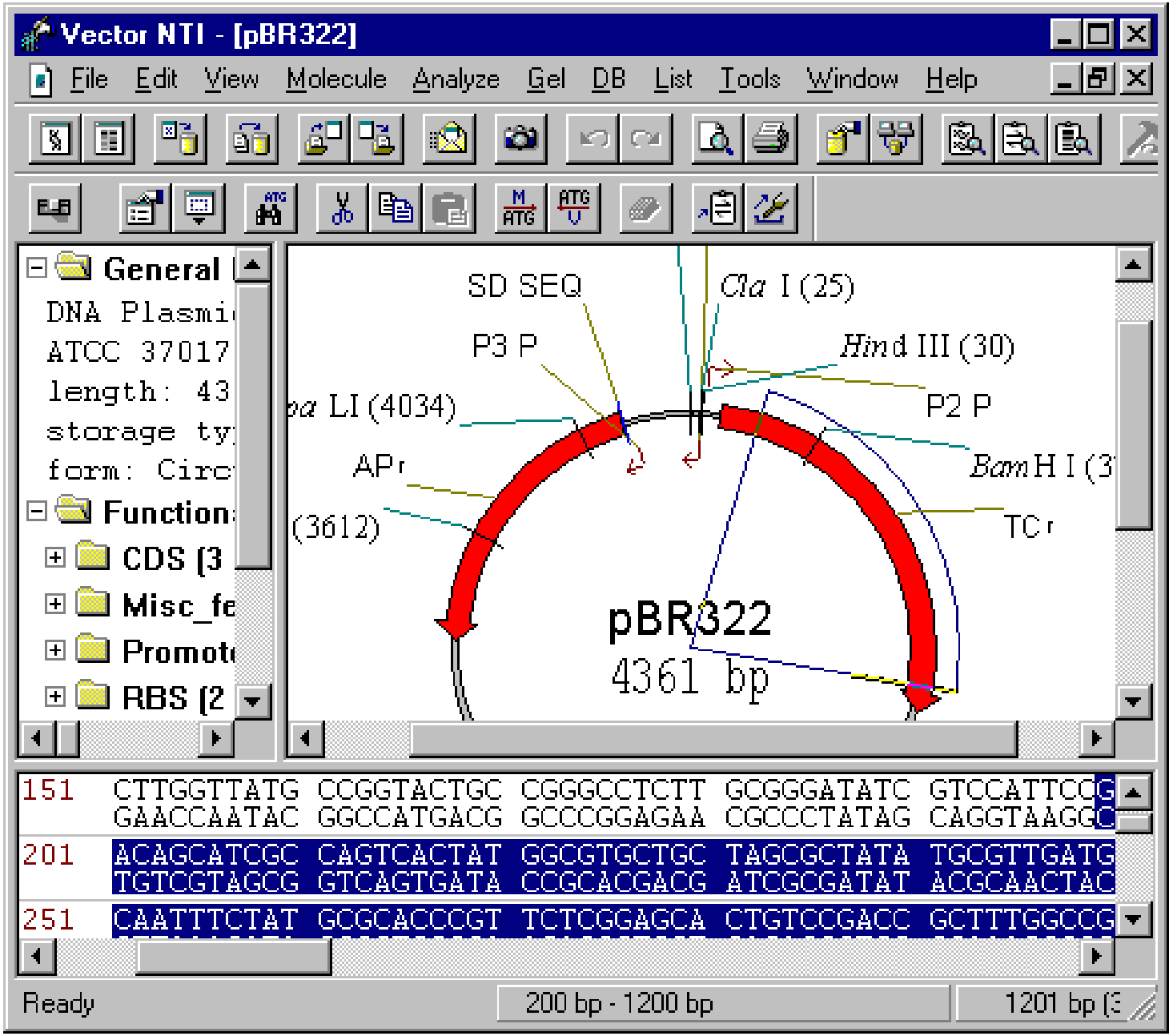
- Konstrukce restrikčních map na základě analýzy sekvence DNA – vyhledání restrikčních míst
  - Nezbytný předpoklad pro klonování
  - Interpretace RFLP polymorfizmů
  - Simulace výsledků gelové elektroforézy restrikčních fragmentů
- Virtuální klonování
- Vytvoření kvalitní grafiky ilustrující restrikční 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 restrikční analýzy *in silico*

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

# Klonování *in silico*, konstrukce vektorů

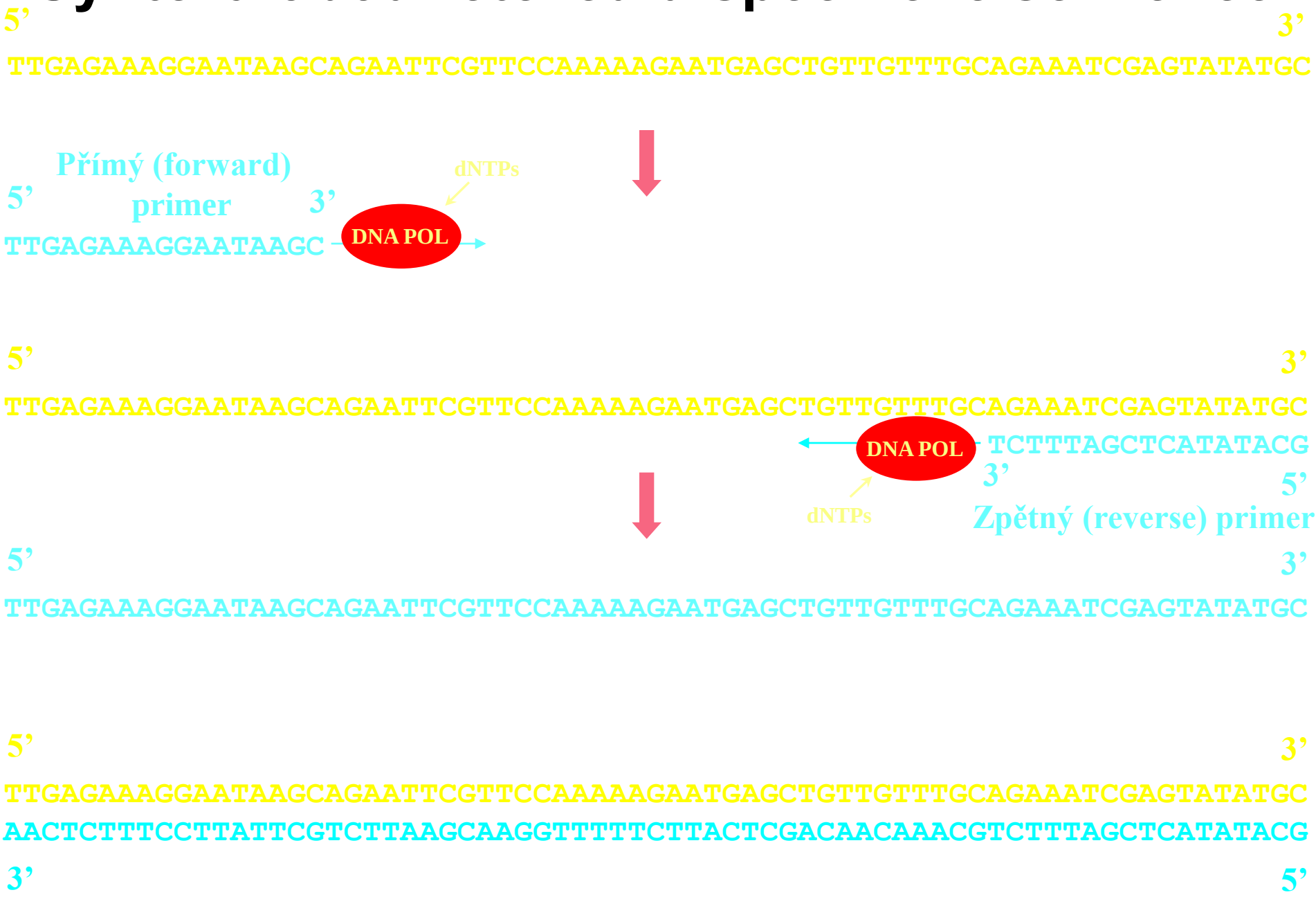
- Kombinace segmentů sekvencí
  - známé/neznámé funkce
- Plazmidy
  - přebírané z databáze
  - zpravidla známé funkce
- Inzerty – obvykle nové sekvence
  - charakterizované restrikční mapou
  - charakterizované sekvencí DNA
  - charakterizované funkcí
- Nomenklatura pro konstrukty není stanovena


Vector NTI - [pBR322]
File Edit View Molecule Analyze Gel DB List Tools Window Help
General
DNA Plasmi
ATCC 37017
length: 43
storage ty
form: Circ
Function
CDS [3
Misc_fe
Promot
RBS [2
SD SEQ
Cla I (25)
P3 P
Hind III (30)
P2 P
BamH I (3
APr
TCr
(3612)
pBR322
4361 bp
151 CTTGGTTATG CCGGTACTGC CGGGCCTCTT GCGGGATATC GTCCATTCCG
GAACCAATAC GGCCATGACG GCCCGGAGAA CGCCCTATAG CAGGTAAGGC
201 ACAGCATCGC CAGTCACTAT GGCGTGCTGC TAGCGCTATA TGCGTTGATG
TGTCGTAGCG GTCAGTGATA CCGCACGACG ATCGCGATAT ACGCAACTAC
251 CAATTTCTAT GCGCACCCGT TCTCGGAGCA CTGTCCGACC GCTTTGGCCG
Ready
200 bp - 1200 bp
1201 bp (3

# 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 Tm
  - 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

**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

# Teplota Tm (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:

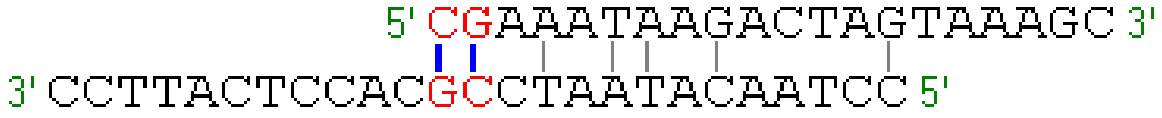
$$T_m = 2(A+T) + 4(G+C)$$

$$T_a = T_m - 5\ °C$$

# 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

```
      3'  GGGAAA
          ||||
5'  TATCTAGGACCTTA

      3'  GGGAA
          || |  A
5'  TATCTAGGACCTTA
```

## Self-Dimer

8 bp

```
      3'  GGGAAAATTCCAGGATCTAT  5'
          ||||  ||||
5'  TATCTAGGACCTTAAAAGGG  3'
```

4 bp

```
      3'  GGGAAAATTCCAGGATCTAT  5'
          ||||
5'  TATCTAGGACCTTAAAAGGG  3'
```

## Dimer

forward primer

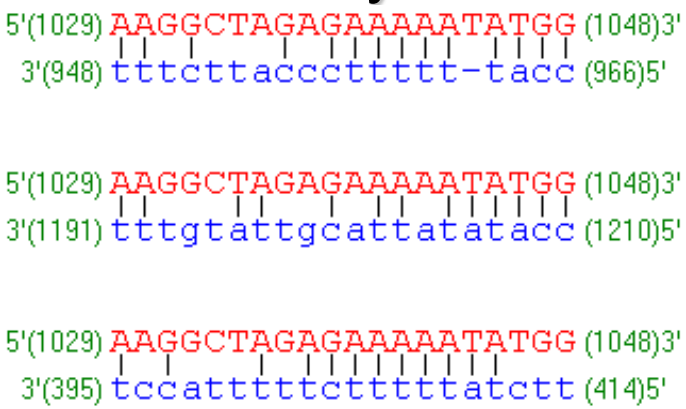
```
5'  TATCTAGGACCTTAAAAGGG  3'
          |||||
      3'  CATGGAAACGTAGGAGAC  5'
```

reverse primer

# Jedinečnost primerů

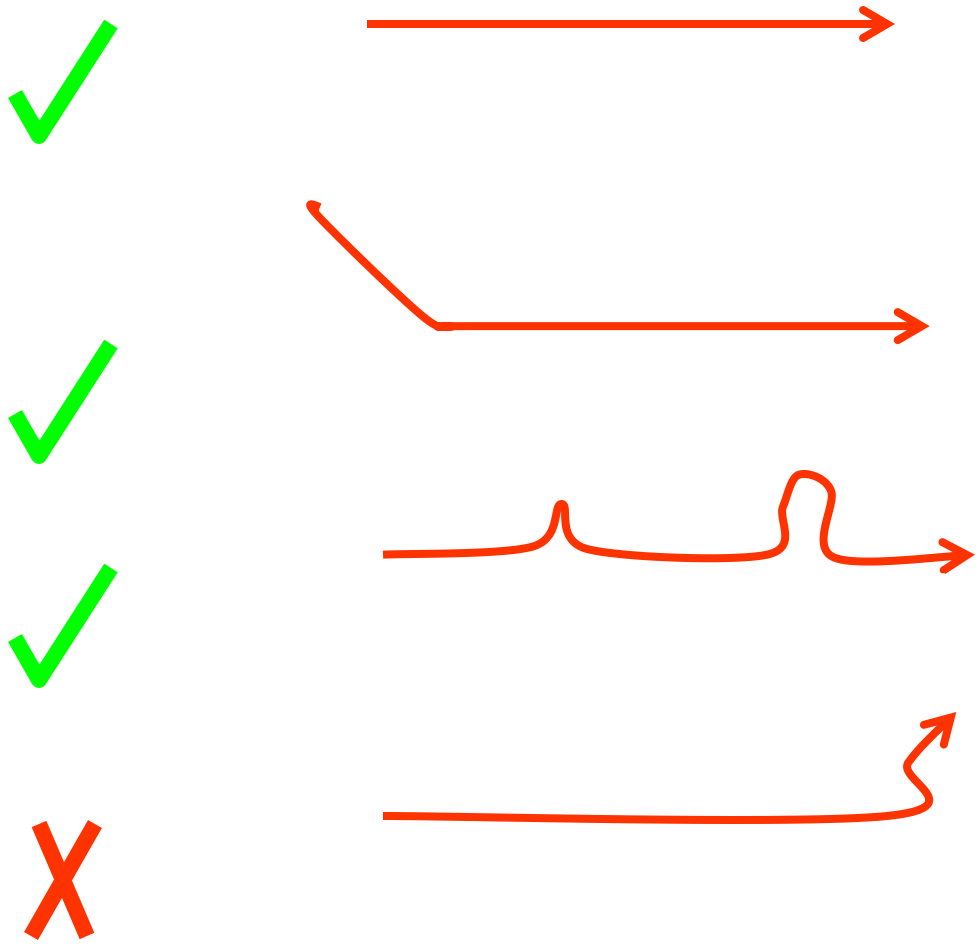
- na matricové DNA nemají falešná vazebná místa

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

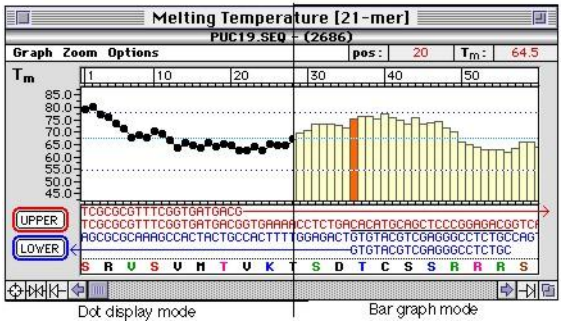


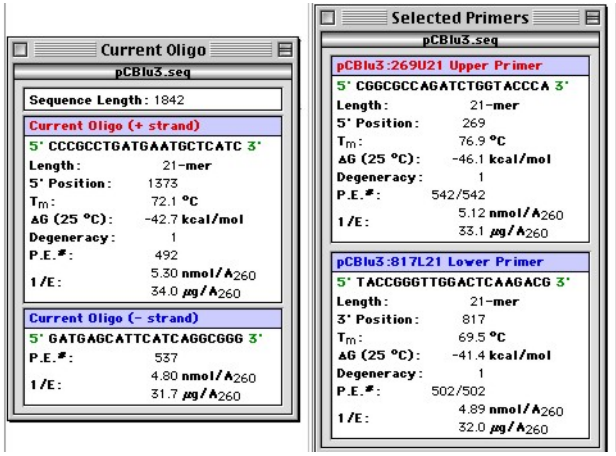

- Správně navržený primer, který nemá falešná vazebná místa na templátu:

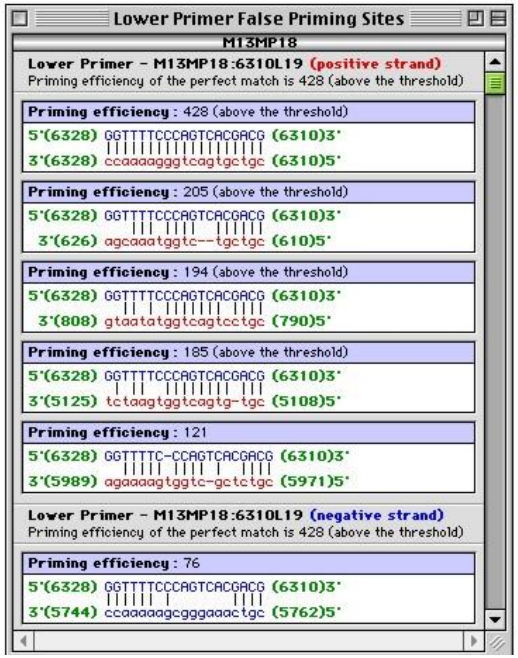
# Kdy je primer ještě primerem?

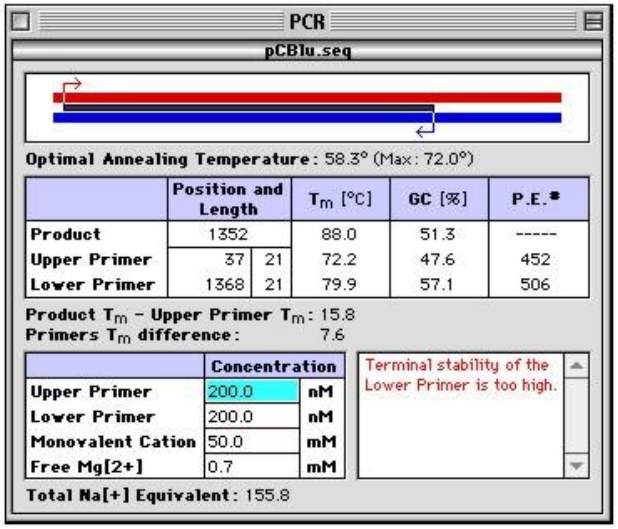


# Pro návrh primerů se obvykle používá specializovaný software



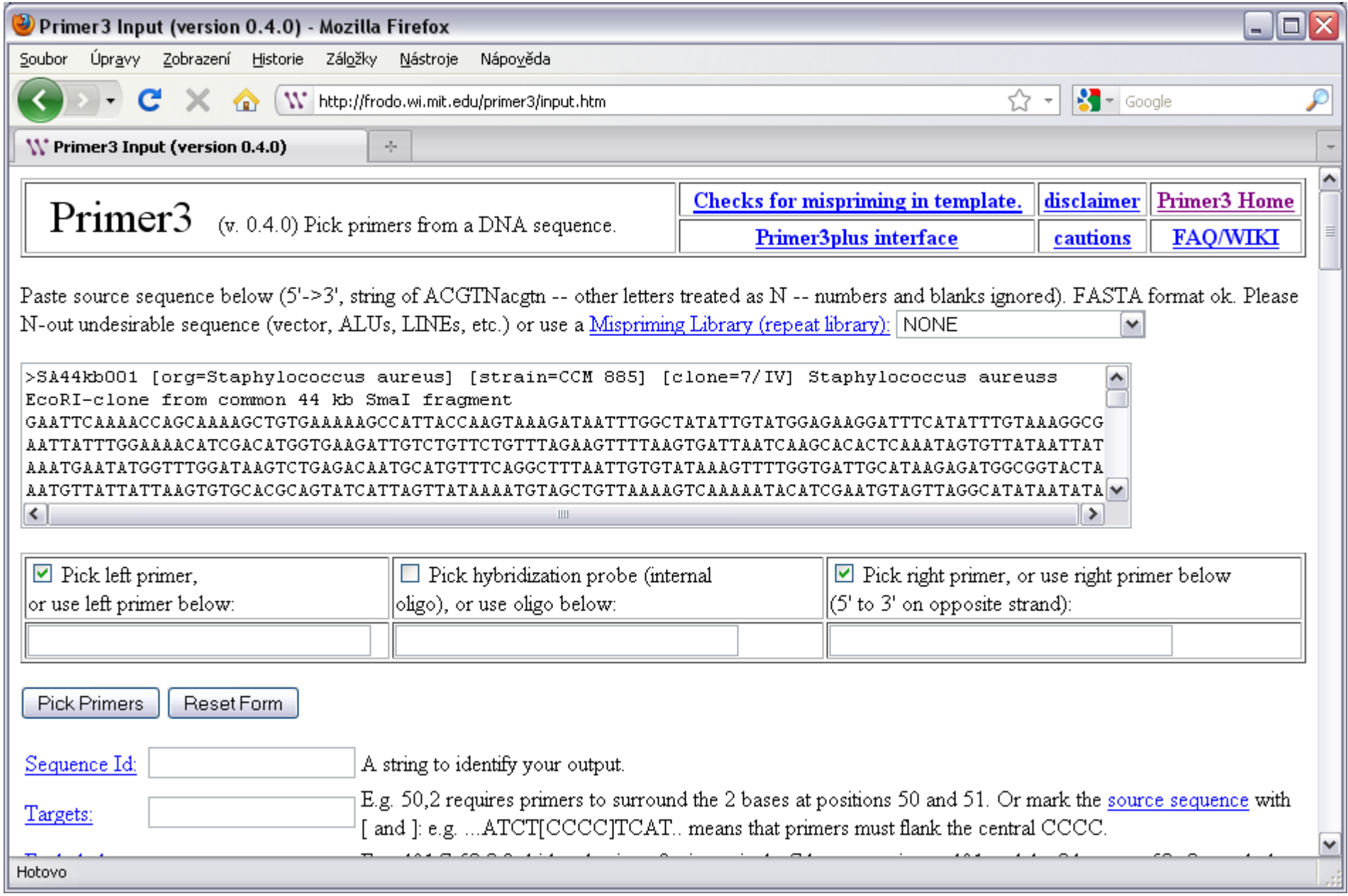



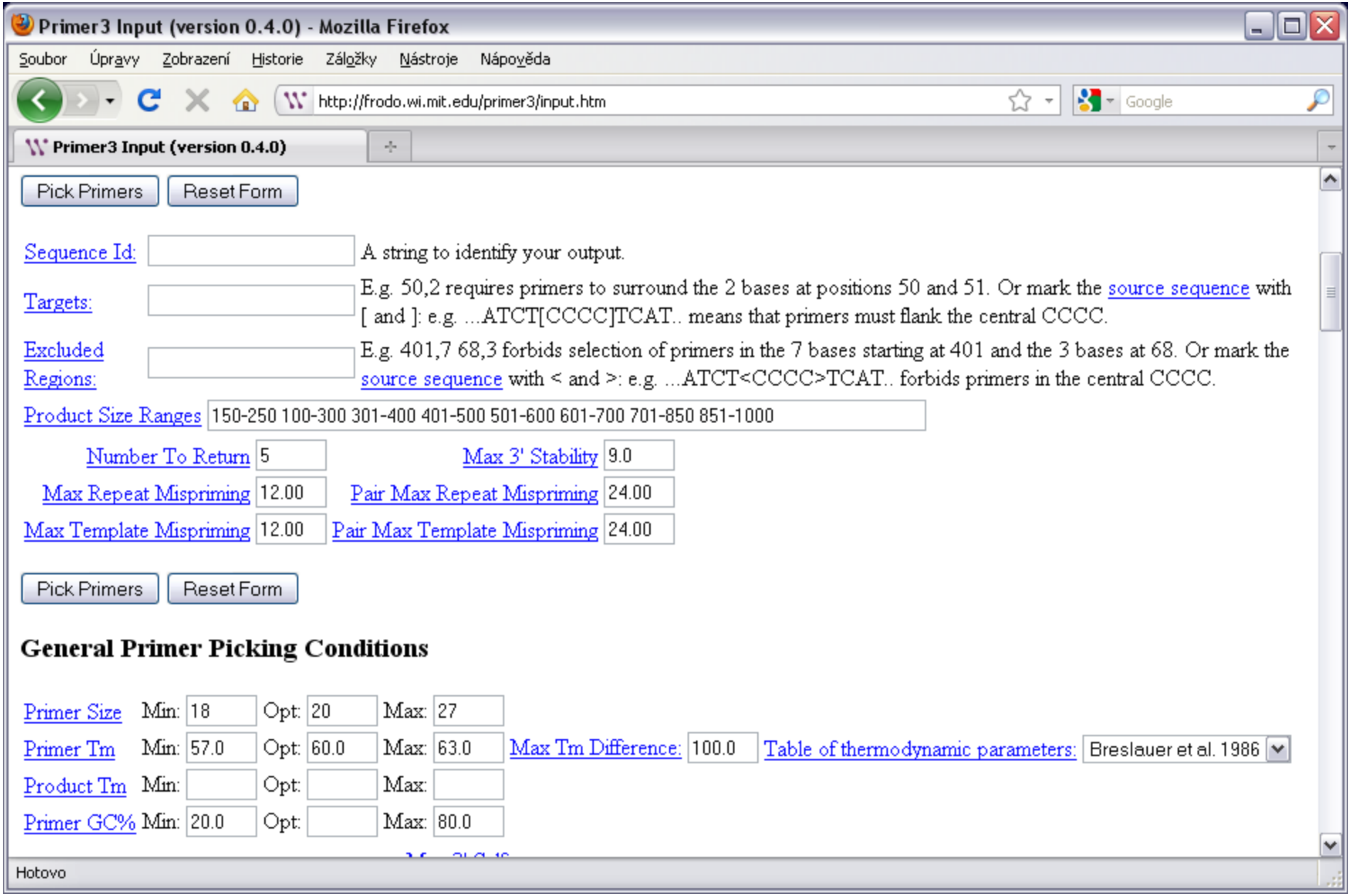



# Počítačový návrh primerů

- Umoňuje 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.promega.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
Google
Primer3 Input (version 0.4.0)
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
5
Max 3' Stability
9.0
Max Repeat Mispriming
12.00
Pair Max Repeat Mispriming
24.00
Max Template Mispriming
12.00
Pair Max Template Mispriming
24.00
Pick Primers
Reset Form
General Primer Picking Conditions
Primer Size Min: 18 Opt: 20 Max: 27
Primer Tm Min: 57.0 Opt: 60.0 Max: 63.0
Max Tm Difference: 100.0
Table of thermodynamic parameters: Breslauer et al. 1986
Product Tm Min: Opt: Max:
Primer GC% Min: 20.0 Opt: Max: 80.0
Hotovo

PRIMER PICKING RESULTS FOR SA44kb001 [org=Staphylococcus aureus] [strain=CCM 885] [clone=7/IV] Staphylococcus aure

No mispriming library specified
Using 1-based sequence positions

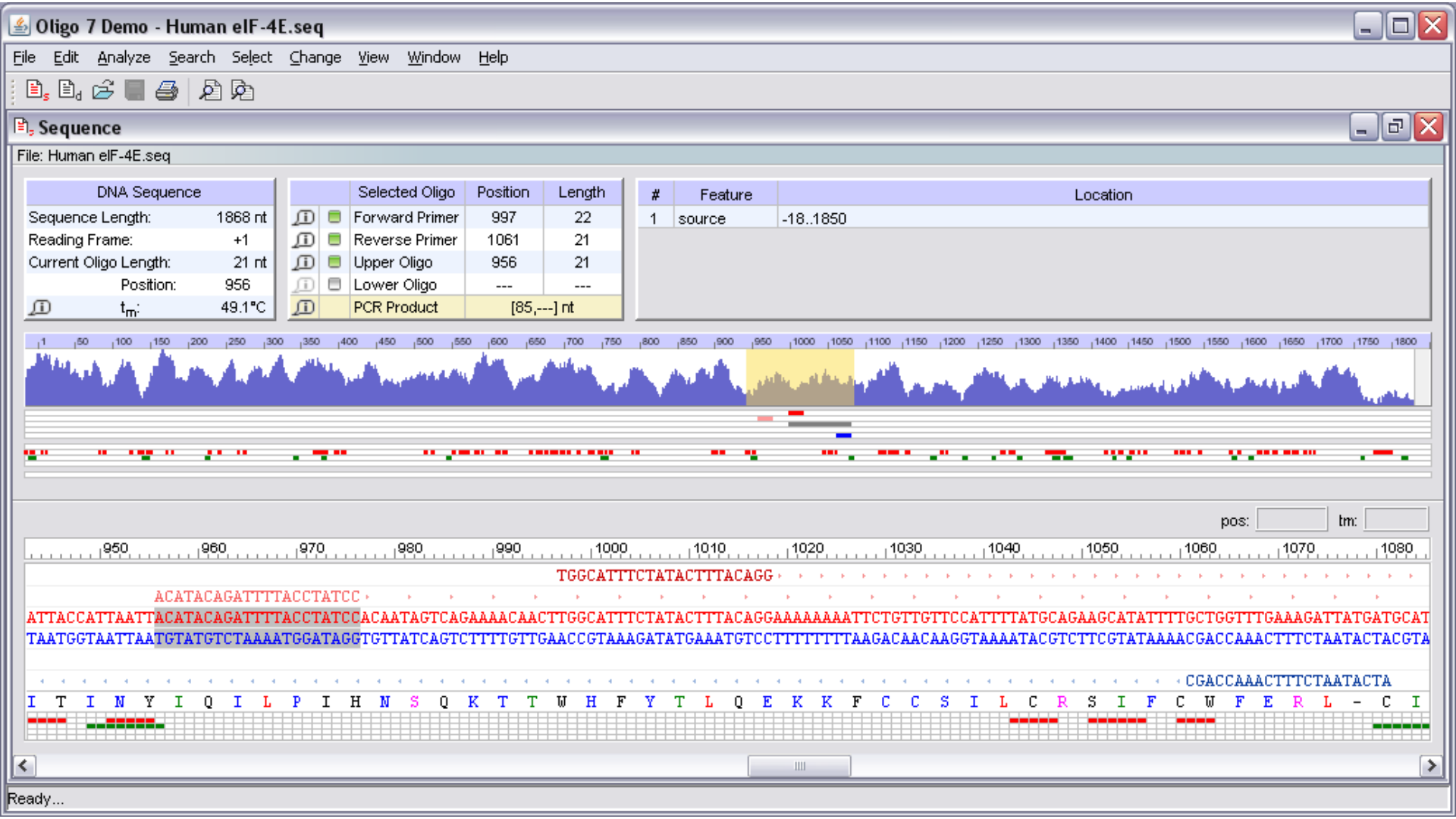
| OLIGO | start | len | tm | gc% | any | 3' | seq |
|---|---|---|---|---|---|---|---|
| LEFT PRIMER | 159 | 25 | 57.21 | 32.00 | 6.00 | 2.00 | AATCAAGCACACTCAAATAGTGTTA |
| RIGHT PRIMER | 429 | 25 | 58.40 | 36.00 | 4.00 | 3.00 | AACTCCTATGAAGACAACCTTTTTC |

SEQUENCE SIZE: 2052
INCLUDED REGION SIZE: 2052

PRODUCT SIZE: 271, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00
TARGETS (start, len)*: 200,200

```
    1 GAATTCAAAACCAGCAAAAGCTGTGAAAAAGCCATTACCAAGTAAAGATAATTTGGCTAT


   61 ATTGTATGGAGAAGGATTTCATATTTGTAAAGGCGAATTATTTGGAAAACATCGACATGG


  121 TGAAGATTGTCTGTTCTGTTTAGAAGTTTTAAGTGATTAATCAAGCACACTCAAATAGTG
                                        >>>>>>>>>>>>>>>>>>>>>>


  181 TTATAATTATAAATGAATATGGTTTGGATAAGTCTGAGACAATGCATGTTTCAGGCTTTA
      >>>               ****************************************


  241 ATTGTGTATAAAGTTTTGGTGATTGCATAAGAGATGGCGGTACTAAATGTTATTATTAAG
      ************************************************************


  301 TGTGCACGCAGTATCATTAGTTATAAAATGTAGCTGTTAAAAGTCAAAAATACATCGAAT
      ************************************************************


  361 GTAGTTAGGCATATAATATAAAAAGAGTTTTCAATTACTCAATAGAAAAAGGTTGTCTTC
      ****************************************     <<<<<<<<<<<<<<<
```

# 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
Sequence Length: 1868 nt
Reading Frame: +1
Current Oligo Length: 21 nt
Position: 956
tm: 49.1°C
Selected Oligo Position Length
Forward Primer 997 22
Reverse Primer 1061 21
Upper Oligo 956 21
Lower Oligo --- ---
PCR Product [85,---] nt
# Feature Location
1 source -18..1850
pos: tm:
TGGCATTTCTATACTTTACAGG
ACATACAGATTTTACCTATCC
ATTACCATTAATTACATACAGATTTTACCTATCCACAATAGTCAGAAAACAACTTGGCATTTCTATACTTTACAGGAAAAAAAATTCTGTTGTTCCATTTTATGCAGAAGCATATTTTGCTGGTTTGAAAGATTATGATGCAT
TAATGGTAATTAATGTATGTCTAAAATGGATAGGTGTTATCAGTCTTTTGTTGAACCGTAAAGATATGAAATGTCCTTTTTTTTAAGACAACAAGGTAAAATACGTCTTCGTATAAAACGACCAAACTTTCTAATACTACGTA
CGACCAAACTTTCTAATACTA
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...

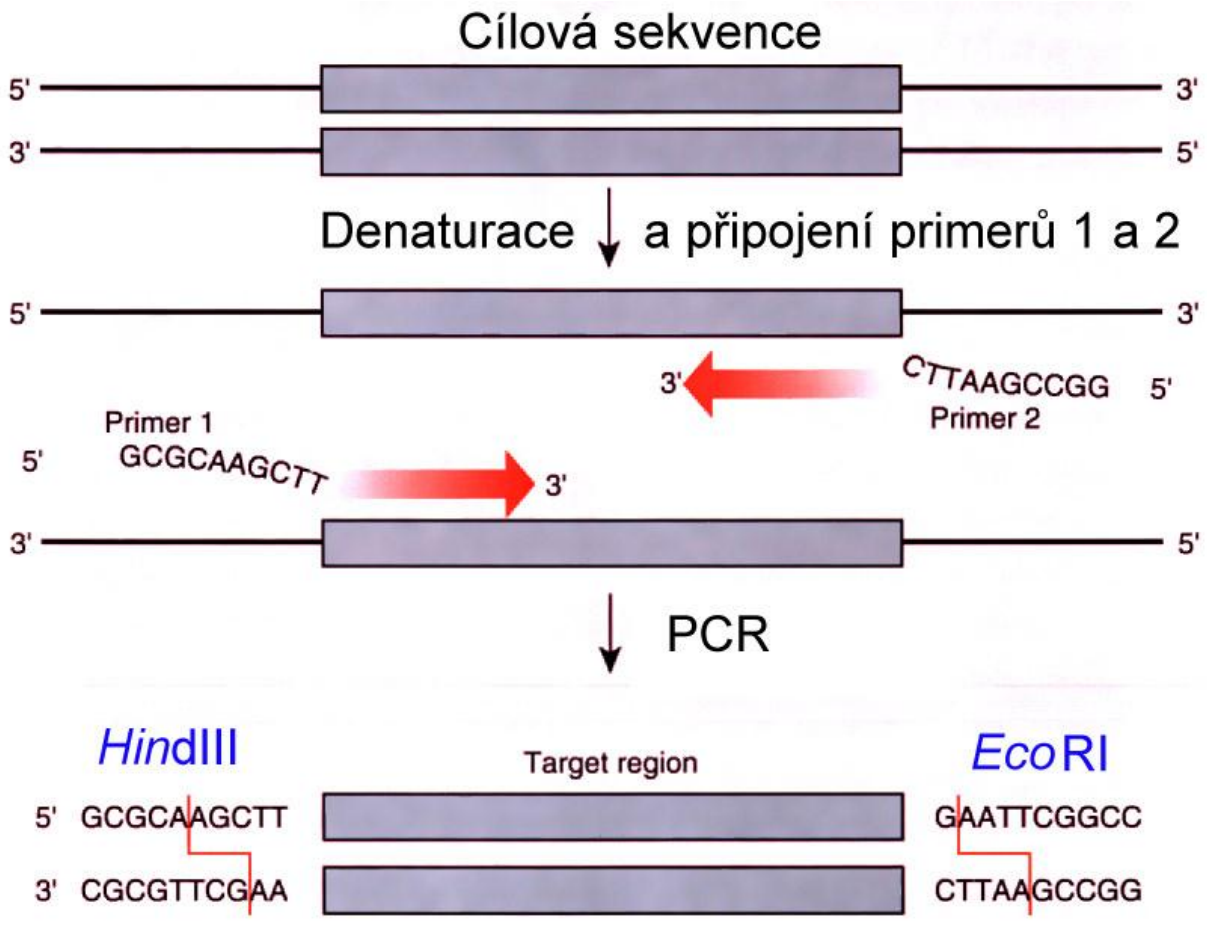
# Výsledky

- Výběr optimálního páru primerů
- Sekvence primerů
- Délka primerů a hodnota Tm
- 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





- 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