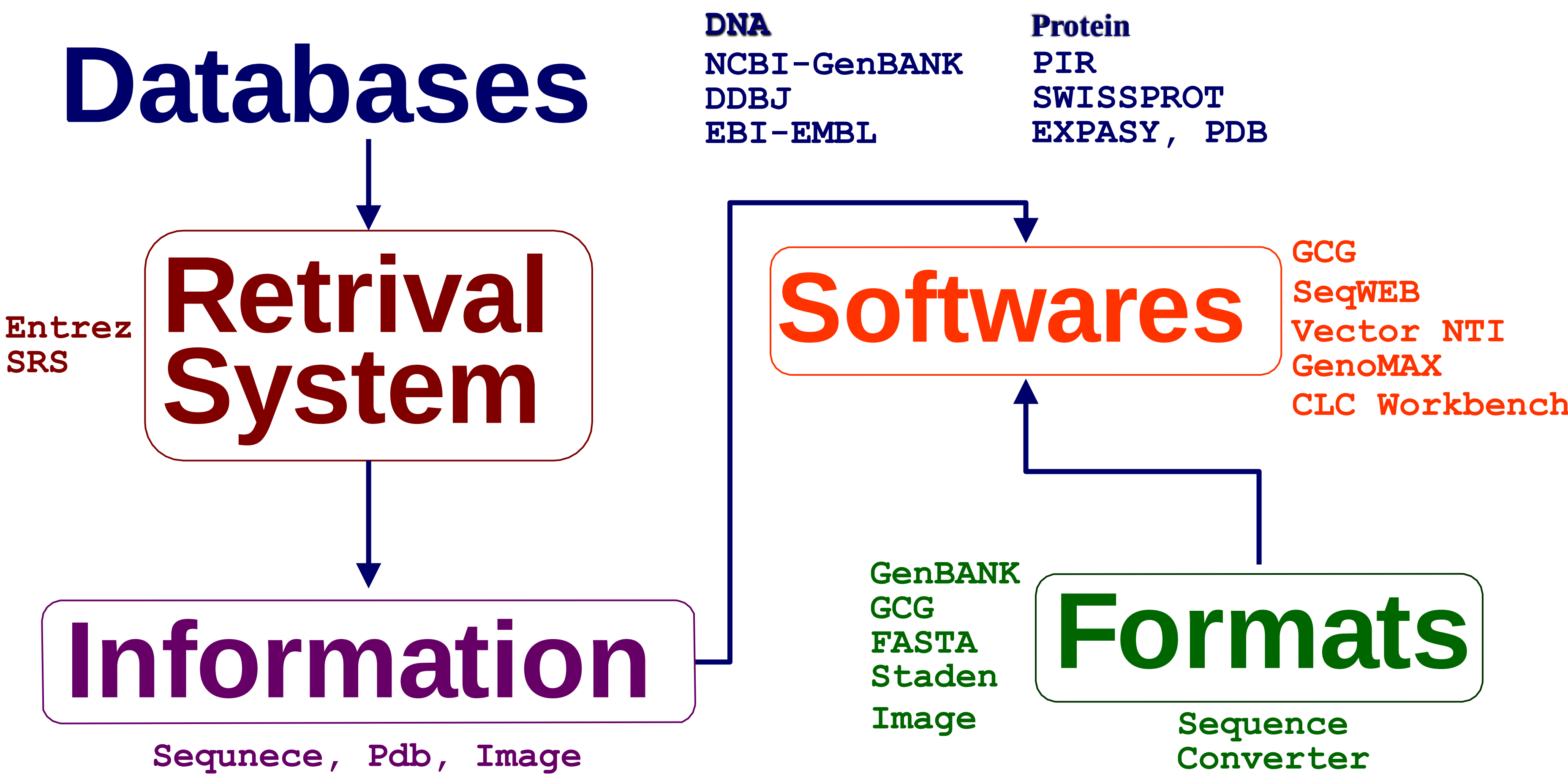
Manipulace se sekvenčními daty

Získání a manipulace se sekvencemi



Sdílení dat v základních databázích

GenBank: *http://www.ncbi.nlm.nih.gov/*
National Center for Biotechnology Information (NCBI)

DDBJ: *http://www.ddbj.nig.ac.jp/*
National Institute of Genetics (NIG)

EMBL: *http://www.ebi.ac.uk*
European Bioinformatics Institute (EBI)

EMBL
European Bioinformatics Institute

ExPASy: *http://tw.expasy.org*
Expert Protein Analysis System

Zápis sekvence

- **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**

Standardní kódy pro sekvence nukleových kyselin podle IUB/IUPAC

A	adenosin
C	cytidin
G	guanidin
T	thymidin
U	uridin
R	G/A (puRin)
Y	T/C (pYrimidin)
K	G/T (nukleosid s Keto skupinou)
M	A/C (nukleosid s aMino skupinou)
S	G/C (silná = Strong vazba)
W	A/T (slabá = Weak vazba)
B	G/T/C (not A)
D	G/A/T (not C)
H	A/C/T (not G)
V	G/C/A (not T)
N	A/G/C/T (jakýkoli)
-	mezera (gap) neurčené délky

Standardní kódy pro sekvence aminokyselin podle IUB/IUPAC

A	alanin
B	kys. asparagová nebo asparagin
C	cystein
D	kys. asparagová
E	kys. glutamová
F	fenylalanin
G	glycin
H	histidin
I	isoleucin
K	lysin
L	leucin
M	metionin
N	asparagin
P	prolin
Q	glutamin
R	arginin
S	serin
T	treonin
U	selenocystein
V	valin
W	tryptofan
Y	tyrosin
Z	kys. glutamová nebo glutamin
X	jakákoli aminokyselina
*****	translační stop (terminační kodon)
-	mezera (gap) neurčené délky

Soubory se sekvencemi DNA/proteinů

- Rozdílné formáty
  - Informační obsah
- Konverze
- Použití
- Editace

Běžné formáty sekvencí

http://orion.sci.muni.cz/kgmb/bioinformat/seq_samples.htm

- FASTA
- Genbank
- EMBL
- GCG
- PIR
- ASN1
- IG(Intelligenetics)
- Text

Formáty sekvencí obsahující mnohonásobná přiložení

- Multi FASTA
- Phylip
- PAUP / NEXUS
- Clustal
- MSF

PLAIN SEQUENCE FORMAT

Obsahuje pouze IUPAC znaky

Obsahuje jedinou sekvenci

Příklad

```
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAA
CCTCCCATCCGTGTCTATTGTACCCTGTTGCTTCGGCGGGCCCGC
CGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTG
CCCGCCGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTC
TGAGTTGATTGAATGCAATCAGTTAAAACTTTCAACAATGGATCT
```

FASTA FORMAT

Může obsahovat více sekvencí

Začíná specifickým záhlavím („>“)

Příklad:

```
>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC
```

EMBL FORMAT

Začíná řádkem s jedinečným identifikátorem (ID), následuje anotace¨.

Sekvence zašíná symboly SQ a sekvence je ukončena „//“

Může obsahovat více sekvencí

Příklad:

```
ID   AA03518    standard; DNA; FUN; 237 BP.
XX
AC   U03518;
XX
DE   Aspergillus awamori internal transcribed spacer 1 (ITS1) and 18S
DE   rRNA and 5.8S rRNA genes, partial sequence.
XX
SQ   Sequence 237 BP; 41 A; 77 C; 67 G; 52 T; 0 other;
     aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc        60
     tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg       120
     ccccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc       180
     tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc          237
//
```

GENBANK FORMAT

Začíná řádkem LOCUS

Začátek sekvence je vyznačen ORIGIN a sekvence je ukončena „//“

Příklad:

```
LOCUS       AAU03518      237 bp    DNA             PLN       04-FEB-1995
DEFINITION  Aspergillus awamori internal transcribed spacer 1 (ITS1) and 18S
            rRNA and 5.8S rRNA genes, partial sequence.
ACCESSION   U03518
VERSION     U03518.1  GI 1235658
BASE COUNT       41 a     77 c     67 g     52 t
ORIGIN
        1 aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc
       61 tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg
      121 ccccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc
      181 tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc
//
```

Poznámka k používaným fontům

- Proporcionální fonty
  - Arial, Times
  - Každý znak jiná šířka
  - Nevhodné
- Neproporcionální fonty
  - Vhodné k použití
  - Všechny znaky stejná šíka
  - Courier, Monospaced

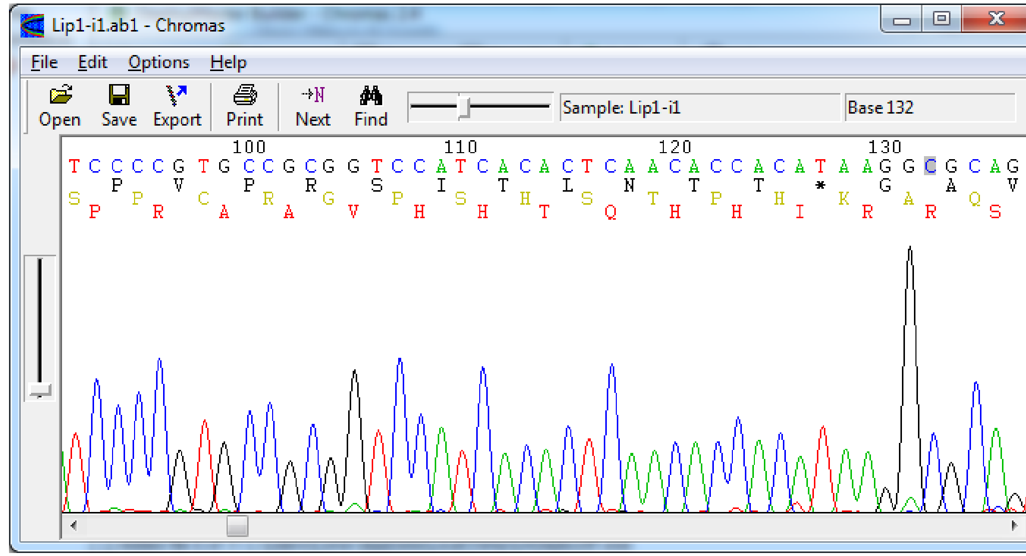
gaattttttt
cttaaaaaaa

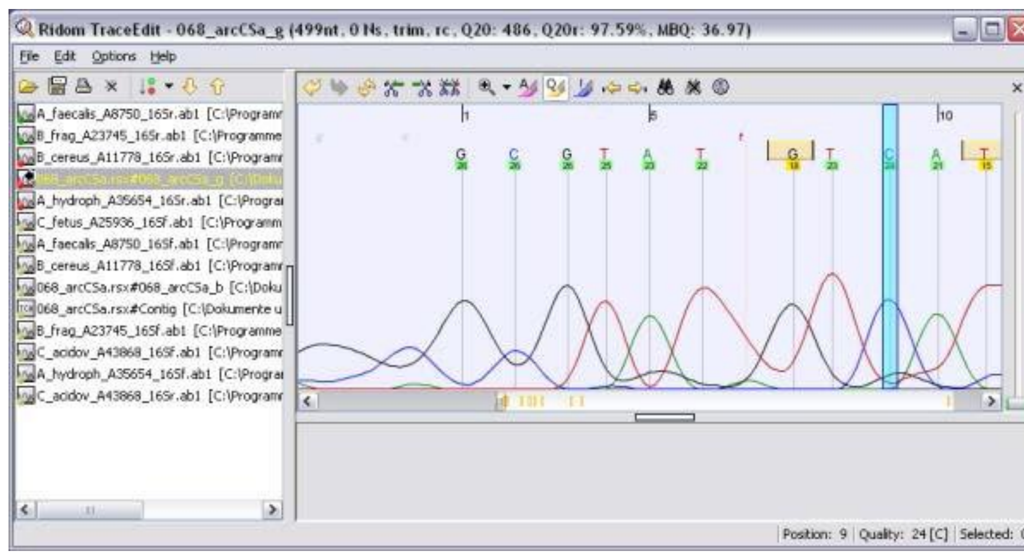
```
gaattttttt
cttaaaaaaa
```

- K editaci jsou vhodné editory, které neukládají informace o formátu textu (Notepad, vývojářské editory – PSPad, aj.)
- Některé formáty jako např. GCG obsahují vnitřní kontrolní součty

Surová data – elektroforetogramy ze sekvenování v kapiláře

- Různé formáty
  - *.abi
  - *.ab1
  - *.scf
- Prohlížeče
  - Chromas
  - ABIView
  - Ridom Trace Edit
- Export
  - FASTA
  - Prostý text





Jednoduché formáty sekvencí mají omezení a neobsahují

- Data o expresi genů
- Variace a polymorfismy
- WWW odkazy na další informace
- Specifické informace o klonech

Konverze formátů sekvencí

- UNIX-GCG
  - To Genbank, To Fasta….
  - From Genbank, From Fasta…
- READSEQ, SEQRET
  - http://www.ebi.ac.uk/Tools/sfc/readseq/
- SMS – The Sequence Manipulation Suite v2
  - http://www.bioinformatics.org/sms2/
    - EMBL to FASTA
    - GenBank to FASTA
    - Reverse Complement
    - Filter DNA / Protein

Příklady jednoduchých manipulací

- Spojování / rozdělování
  - subsekvence
- Převod na reverzní komplementární
  - Ne reverzní
  - Ne komplementární
- Translace
  - Volba vhodného genetického kódu
- Hledání
  - Přesné versus podobné

Hledání motivů

Hledání slov = uspořádaná množina znaků

GAATTC

GARYTC

GAAN(1-50)TTC

Standardní příklady hledání

- Restrikční místa
- Repetice
  - přímé
  - Obrácené (vlásenky se smyčkou)
- Konsenzní vzory
- Uživatelem definované vzory
- Otevřené čtecí rámce

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

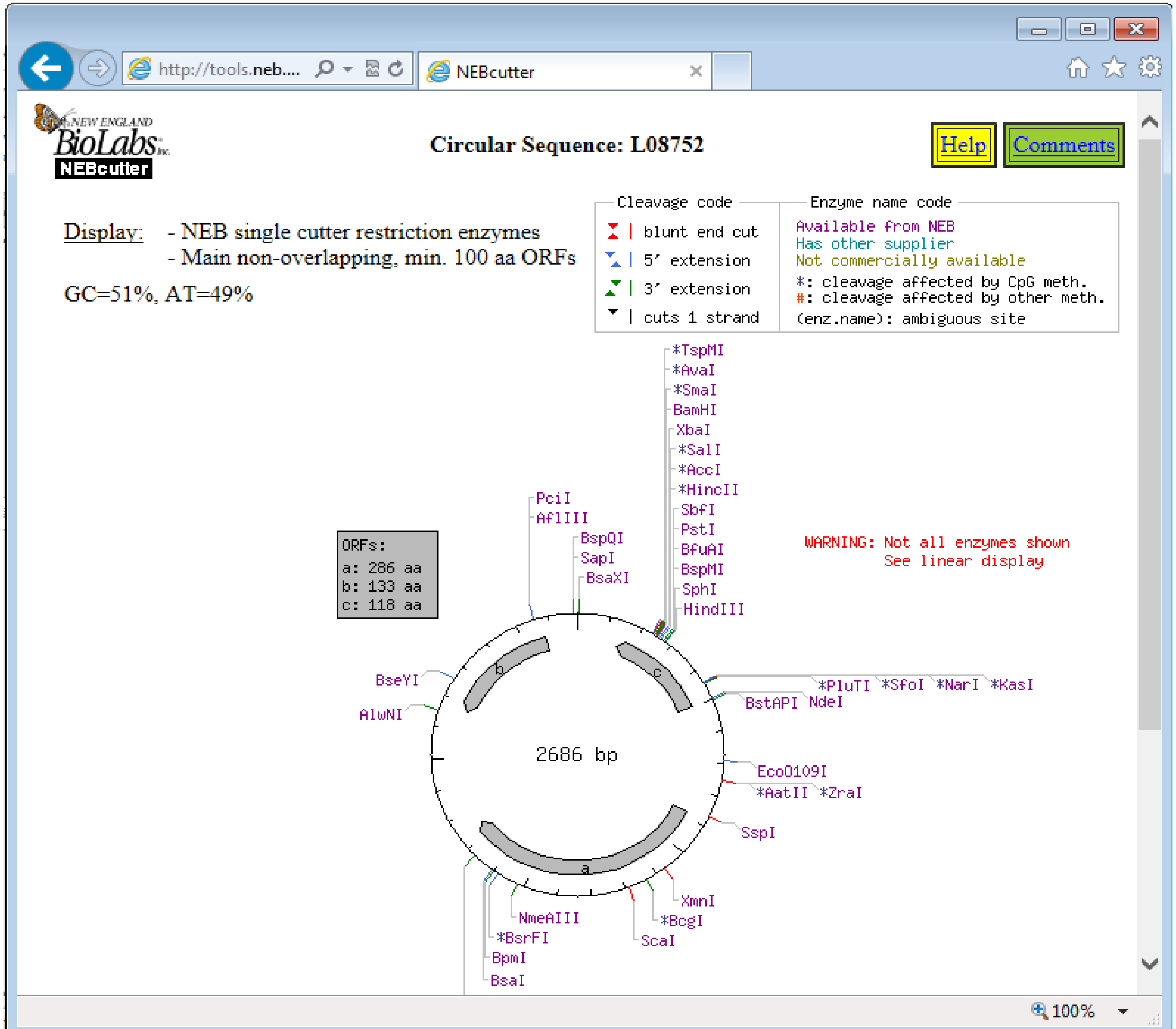
- Konstrukce restrikčních map na základě analýzy sekvence DNA – vyhledání restrikčních míst
  - Nezbytný předpoklad pro klonování
  - Interpretace RFLP polymorfizmů
  - Simulace výsledků gelové elektroforézy restrikčních fragmentů
- Virtuální klonování
- Vytvoření kvalitní grafiky ilustrující restrikční 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 – výstup tabulka
  - kompletní sada
  - komerční sada
  - které sekvenci neštěpí
  - které štěpí – počet a pozice rozpoznávacích míst
- Lineární nebo kružnicová mapa sekvence se znázorněním pozice restrikčních míst
  - Grafika
  - Identifikace ORF a translace do proteinu

NEB Cutter

http://tools.neb.com/NEBcutter2/

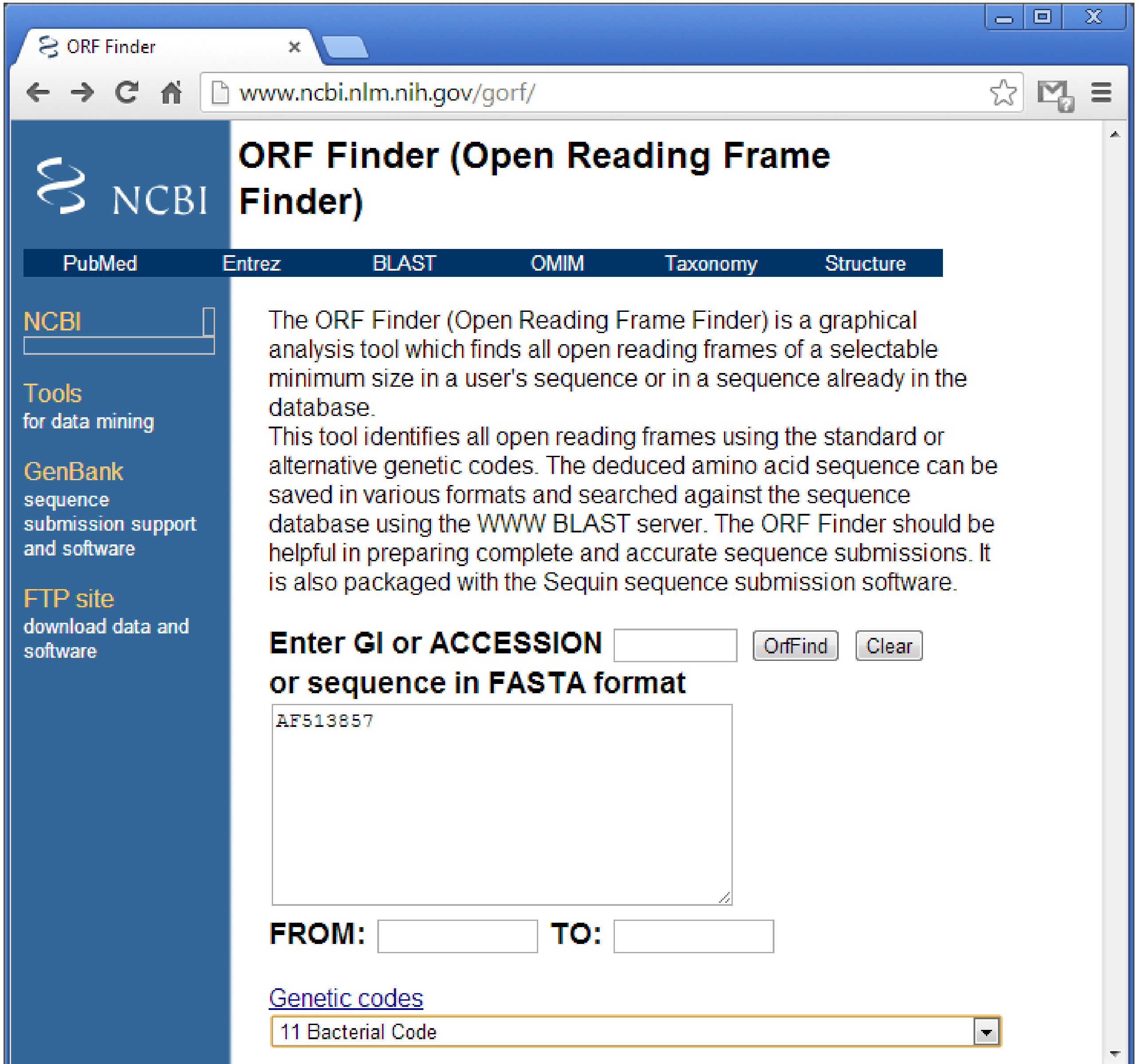


Vyhledání otevřených čtecích rámců

- **ORF (Open Reading Frame)**
  Sada překládaných kodonů mezi iniciačním a terminačním kodonem
- 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)

ORF Finder (Open Reading Frame Finder)

http://www.ncbi.nlm.nih.gov/gorf/gorf.html

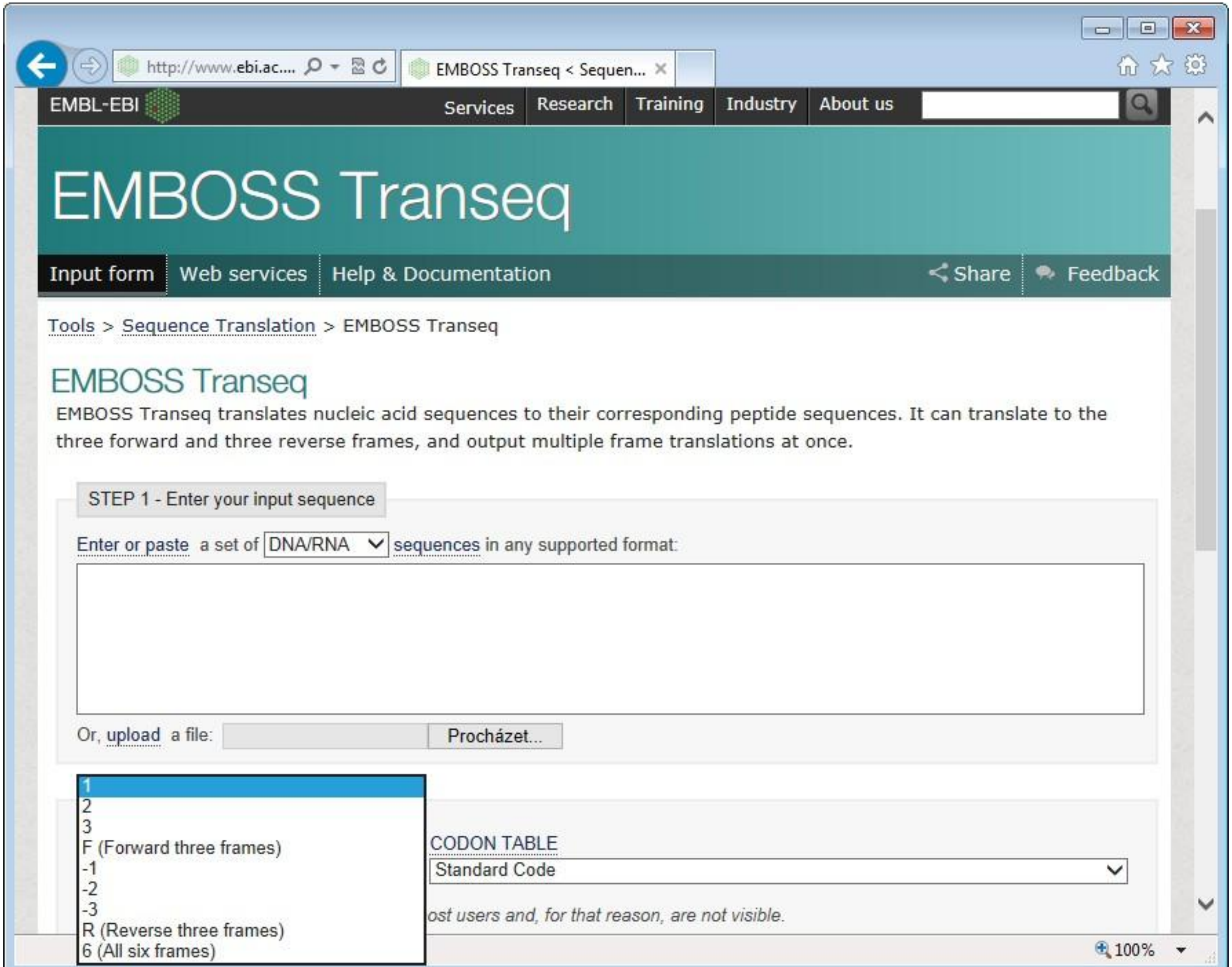


Translace *in silico*

- 6 možných čtecích rámců
- Vymezené oblasti - exony
- Jaký genetický kód?
  - Databáze genetických kódů v NCBI
  - http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

EMBOSS Transeq

http://www.ebi.ac.uk/Tools/st/emboss_transeq/



Příklady translace *in silico*

ExPASy
Bioinformatics Resource Portal

Translate Tool - Results of translation

Open reading frames are highlighted in red. Please select one of the following frames

5'3' Frame 1

```
LLIQQAKSNSDTTPAMPLDTCGAMSQGMIGYWLETEINRILTEMNSDRTVGTIVTRVEVD
KDDPRFDNPTKPIGPFYTKEEVEELQKEQPDSVFKEDAGRGYRKVVASPLPQSILEHQLI
QTLADGKNIVIACGGGGIPVIKKENTYEGVEA
```

5'3' Frame 2

```
Y-SNKLNRTVTQRRQCHWILVVQCHRV--AIGWKLKSIAF-LK-IVIEL-AQSLHVWK-I
KMIHDLITQLNQLVLFIRKKKLKNYKKNSQTQSLKKMQDVVIEK-LRHHYLNLY-NTS-F
KL-QTVKILSLHAVVAVFQL-KKKIPMKVLK
```

5'3' Frame 3

```
INPTS-IEQ-HNAGNAIGYLWCNVTGYDRLLVGN-NQSHFN-NE---NCRHNRYTCGSR-
R-STI--PN-TNWSFLYERRS-RITKRTARLSL-RRCRTWL-KSSCVTTTSIYTRTPVNS
NFSRR-KYCHCMRWWRYSSYKKRKYL-RC-S
```

Příklady translace *in silico*

EMBOSS Sixpack

Input form | Web services | Help & Documentation

Tools > Sequence Translation > EMBOSS Sixpack

Results for job emboss_sixpack-I20141006-192122-0940-32029869-oy

Result Summary | Tool Output | Submission Details

Download Sixpack File

```
EMBOSS_001


      L  L  I  Q  Q  A  K  S  N  S  D  T  T  P  A  M  P  L  D  T     F1
       Y  *  S  N  K  L  N  R  T  V  T  Q  R  R  Q  C  H  W  I  L    F2
        I  N  P  T  S  *  I  E  Q  *  H  N  A  G  N  A  I  G  Y  L   F3
   1 TtattAaTccaACAaGCTAAatCGAACaGTGACACAACGCCGGCAATGCCATTGGATACT 60
     ----:----|----:----|----:----|----:----|----:----|----:----|
   1 AataaTtAggtTGTtCGATTtaGCTTGtCACTGTGTTGCGGCCGTTACGGTAACCTATGA 60
       X  N  I  W  C  A  L  D  F  L  S  V  V  G  A  I  G  N  S  V    F6
      X  I  L  G  V  L  *  I  S  C  H  C  L  A  P  L  A  M  P  Y     F5
        *  *  D  L  L  S  F  R  V  T  V  C  R  R  C  H  W  Q  I  S   F4


      C  G  A  M  S  Q  G  M  I  G  Y  W  L  E  T  E  I  N  R  I     F1
       V  V  Q  C  H  R  V  *  *  A  I  G  W  K  L  K  S  I  A  F    F2
        W  C  N  V  T  G  Y  D  R  L  L  V  G  N  *  N  Q  S  H  F   F3
  61 TGTGGTGCAATGTCACAGGGTATGATAGGCTATTGGTTGGAAACTGAAATCAATCGCATT 120
     ----:----|----:----|----:----|----:----|----:----|----:----|
  61 ACACCACGTTACAGTGTCCCATACTATCCGATAACCAACCTTTGACTTTAGTTAGCGTAA 120
       Q  P  A  I  D  C  P  I  I  P  *  Q  N  S  V  S  I  L  R  M    F6
      K  H  H  L  T  V  P  Y  S  L  S  N  T  P  F  Q  F  *  D  C     F5
        T  T  C  H  *  L  T  H  Y  A  I  P  Q  F  S  F  D  I  A  N   F4
```

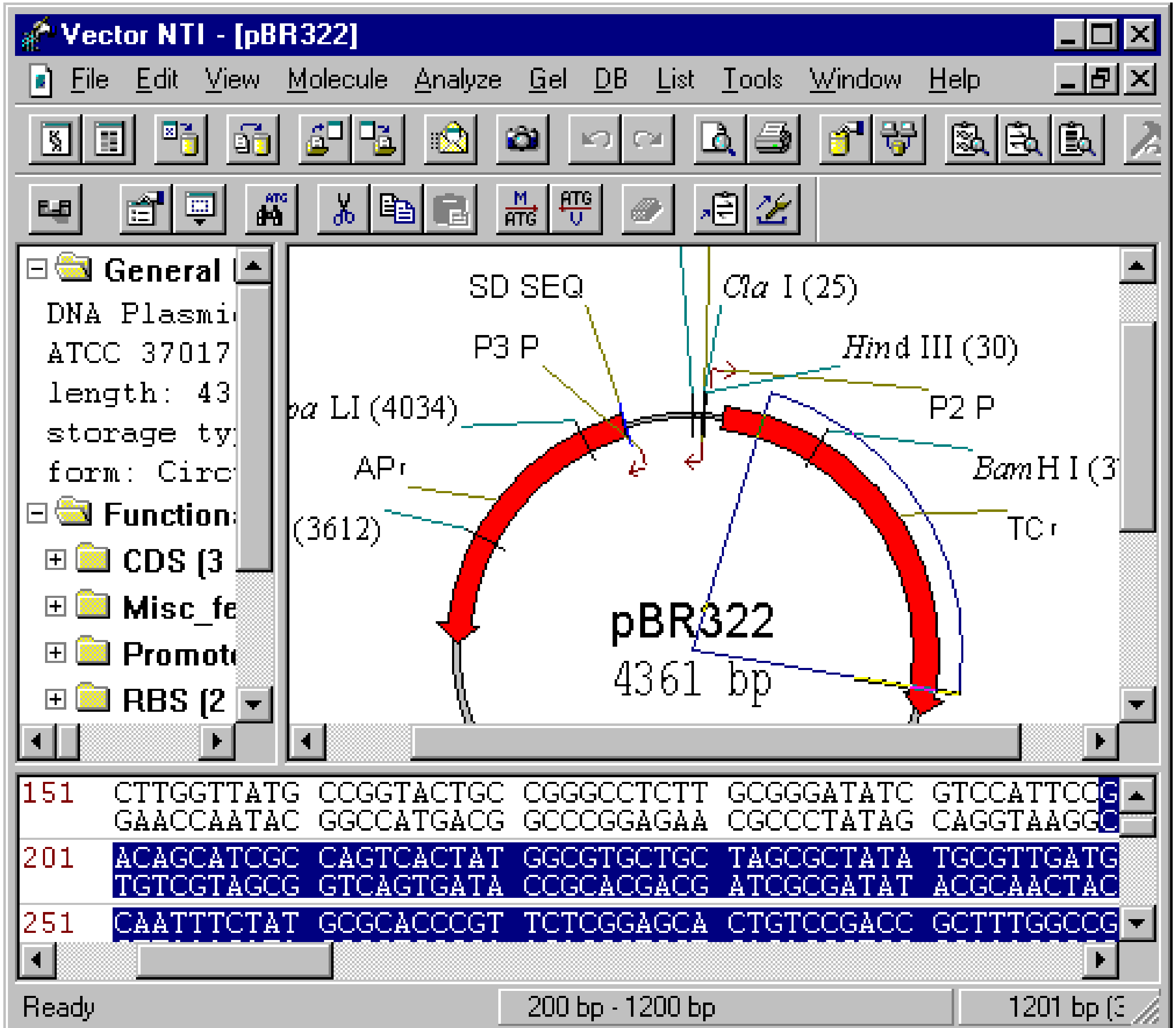
Další typy analýz sekvencí DNA

1. Analýza využití kodonů
2. Klonování *in silico*, konstrukce vektorů
3. Návrh sekvencí oligonukleotidů
   - primery pro PCR
   - primery pro sekvenování
   - hybridizační sondy
4. Párové přiložení sekvencí, stanovení identity a podobnosti
5. Mnohonásobné přiložení sekvencí

Nejčastěji používané softwarové balíky pro manipulaci se sekvencemi a jejich analýzu

- Accelrys GCG Package (Accelrys Inc., San Diego, CA)
- Vector NTI® (Life Technologies, Carlsbad, CA)
- CLC Genomics Workbench (CLC bio, Cambridge, MA)
- The Bioinformatics Toolbox rozšíření pro MATLAB®
- Hitachi DNASIS® MAX Sequence Analysis Software (Helixx Technologies, Inc., Canada)
- DNASTAR Lasergene (DNASTAR, Inc., Madison, WI)

Příklad software Vector NTI



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

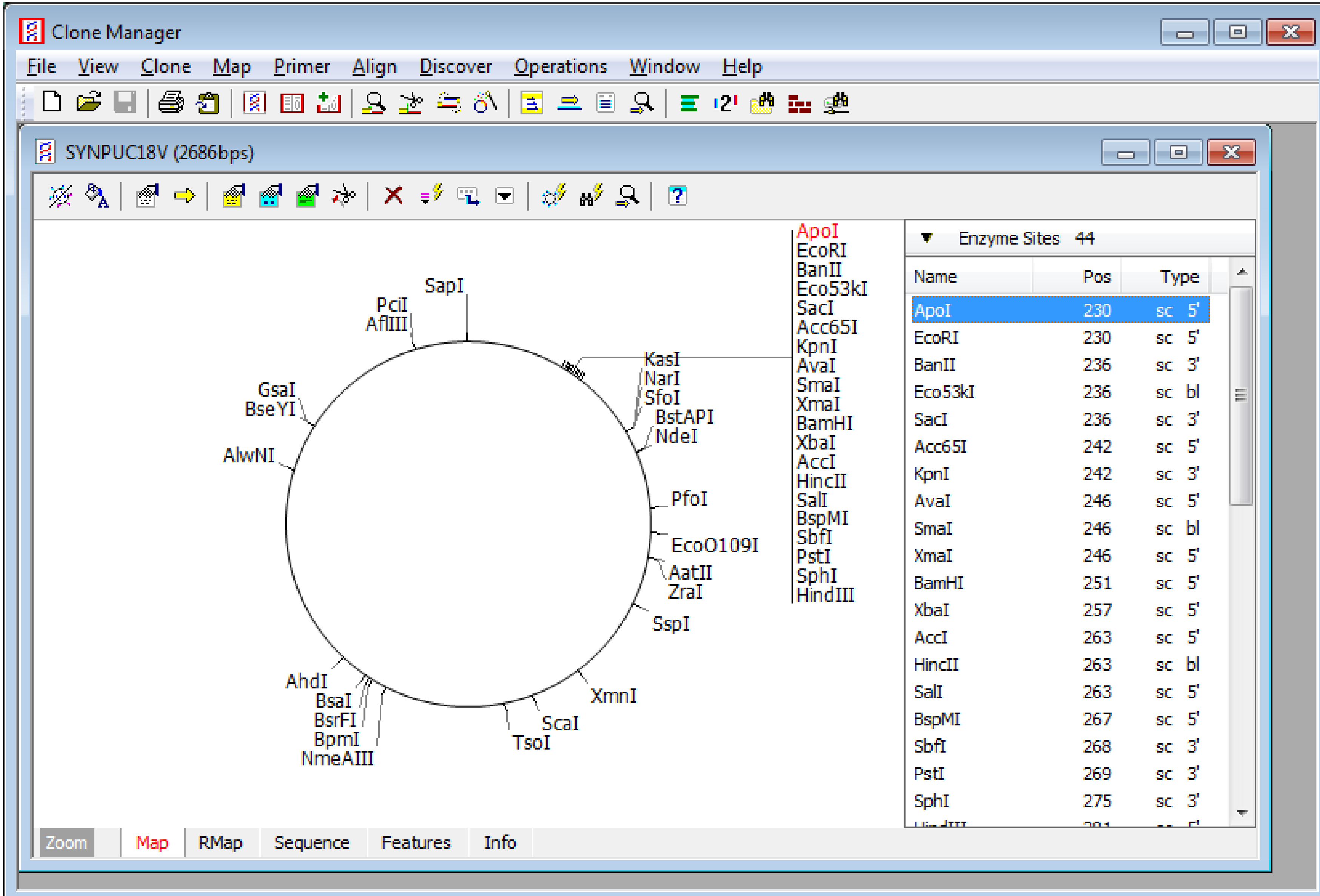
Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	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

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

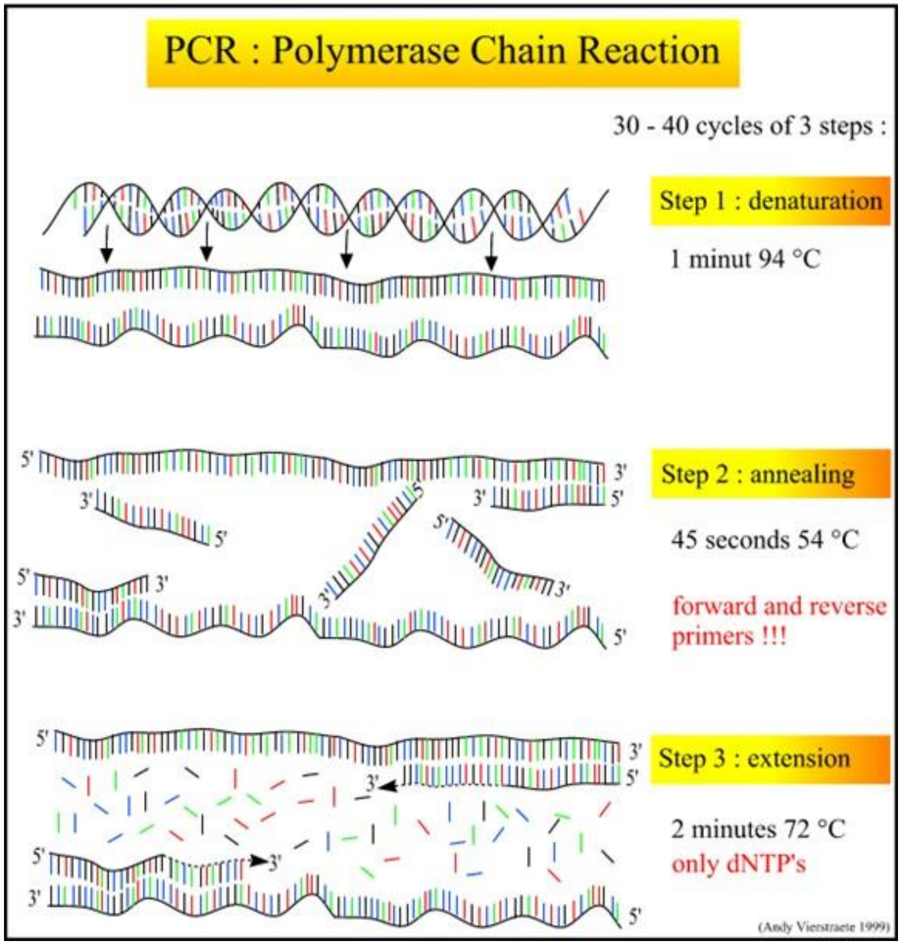
Clone Manager (Sci-Ed Software)

http://www.scied.com/pr_cmbas.htm



Navrhování sekvencí primerů pro PCR

- Standardní primery
- Modifikované oligonukleotidy na 5'-konci pro klonování
- Oligonukleotidy jako hybridizační sondy pro real-time PCR
  - specifičnost
  - jedinečnost



PCR - Syntéza obou řetězců u specifické sekvence

5’ 3’

TTGAGAAAGGAATAAGCAGAATTCGTTCCAAAAAGAATGAGCTGTTGTTTGCAGAAATCGAGTATATGC

AACTCTTTCCTTATTCGTCTTAAGCAAGGTTTTTCTTACTCGACAACAAACGTCTTTAGCTCATATACG

3’ 5’

Přímý (forward) primer

dNTPs

5’ 3’

TTGAGAAAGGAATAAGC — DNA POL →

AACTCTTTCCTTATTCGTCTTAAGCAAGGTTTTTCTTACTCGACAACAAACGTCTTTAGCTCATATACG

3’ 5’

5’ 3’

TTGAGAAAGGAATAAGCAGAATTCGTTCCAAAAAGAATGAGCTGTTGTTTGCAGAAATCGAGTATATGC

← DNA POL - TCTTTAGCTCATATACG

3’ 5’

dNTPs

Zpětný (reverse) primer

5’ 3’

TTGAGAAAGGAATAAGCAGAATTCGTTCCAAAAAGAATGAGCTGTTGTTTGCAGAAATCGAGTATATGC

AACTCTTTCCTTATTCGTCTTAAGCAAGGTTTTTCTTACTCGACAACAAACGTCTTTAGCTCATATACG

3’ 5’

5’ 3’

TTGAGAAAGGAATAAGCAGAATTCGTTCCAAAAAGAATGAGCTGTTGTTTGCAGAAATCGAGTATATGC

AACTCTTTCCTTATTCGTCTTAAGCAAGGTTTTTCTTACTCGACAACAAACGTCTTTAGCTCATATACG

3’ 5’

Výběr vhodné strategie před návrhem primerů

- K čemu jsou primery určeny
  - Standardní end-point PCR
  - Sekvenování
  - Detekce jednonukleotidových polymorfizmů (SNP) nebo variací
  - Studium metylace
  - Real-time PCR
  - Sondy pro microarray
  - Degenerovaná PCR
  - Multiplex PCR
- Z jakých dat vycházíme
  - Jednoduchá sekvence DNA / proteinu
  - Sekvenční přiložení DNA / proteinu
  - GenBank ID/Gene ID/rsSNP ID

Pravidla pro 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

Templátová DNA

5′...TCAACTTAGCATGATCGGGTA...GTAGCAGTTGACTGTACAACTCAGCAA...3′

TGCTAAGTTG   CAGTCAACTGCTAC   TGCT A AGTTG

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′

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:

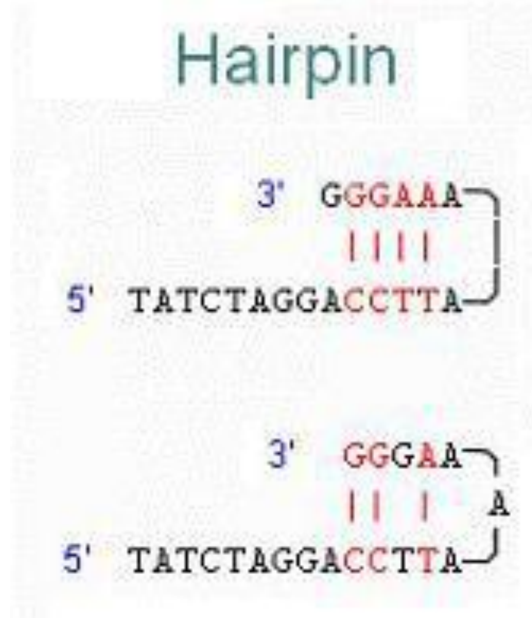
```
5' ATTCAACCGTTCAAACAAGCCC 3'
                  ||||||
                3' GTTCGGCCTACCTTTATTTCTC 5'
```

  - 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:

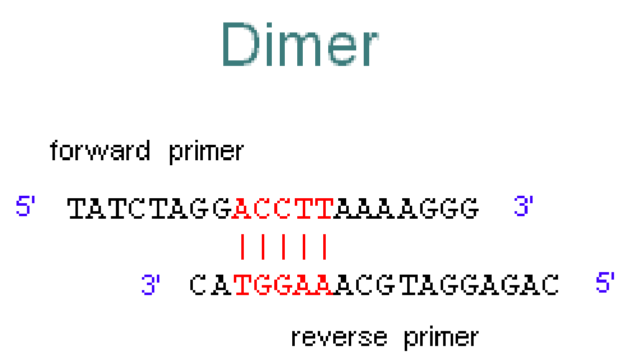
```
           5' CGAAATAAGACTAGTAAAGC 3'
              ||  |  || |    |
3' CCTTACTCCACGCCTAATACAATCC 5'
```

  - Chybně navržený primer, vytvářející vlásenku:

```
5' TTTTTCAAGG
    |||      C
 3' AAAAGAGAT
```


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

GC svorky a 3´- koncová stabilita

- GC svorka
  - Přítomnost G nebo C mezi posledním 4 bázemi na 3’-konci primeru
  - Zásadní pro zvýšení prevence falešného prodlužování a zvýšení specifičnosti primeru
  - >3 G nebo C v blízkosti 3’-konce jsou však nežádoucí
- Maximální 3’-koncová stabilita
  - Maximalizace ΔG posledních 5 bází na 3’-konci primeru.

Jedinečnost primerů

- na matricové DNA nemají falešná vazebná místa

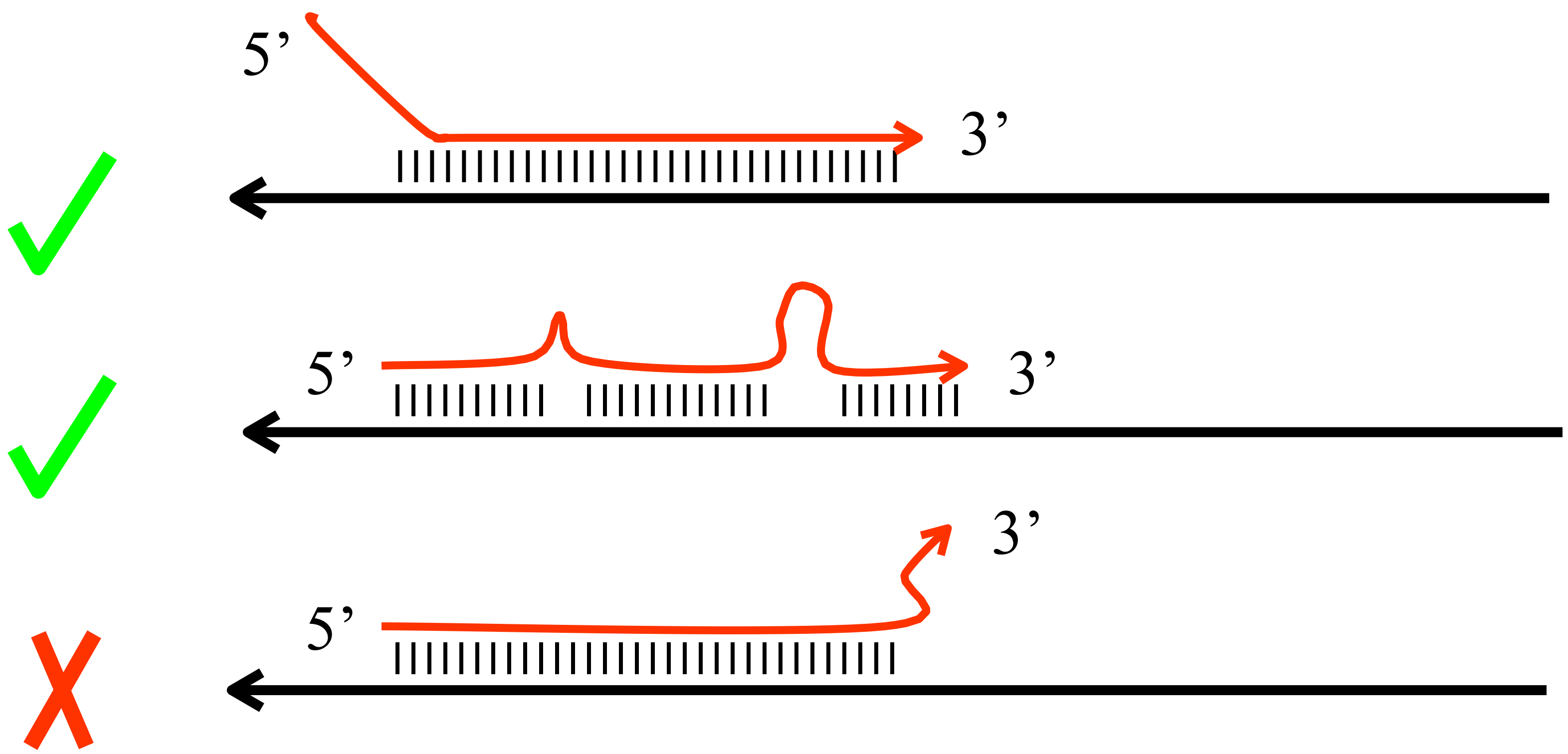
- Nesprávně navržený primer s falešnými vazebnými místy na templátové DNA:

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

5'(1029) AAGGCTAGAGAAAAATATGG (1048)3'
         ||   ||  | || ||||||
3'(1191) tttgtattgcattatatacc (1210)5'

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

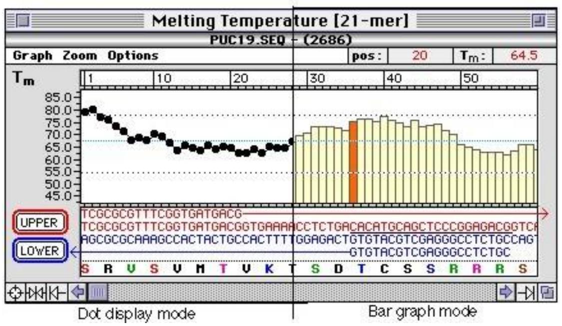
- Správně navržený primer, který nemá falešná vazebná místa na templátu:

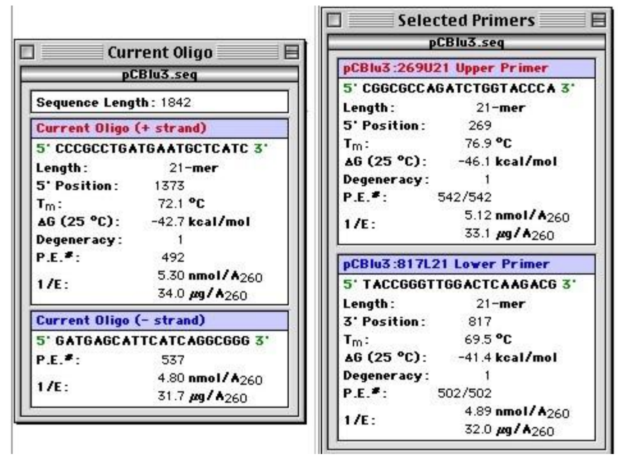
```
5'(2476) CCTAACATAATCCGCACCTCATTCC (2452)3'
           | | ||||||        |||
 3'(787) taaatctattagtttacacataacc (811)5'

5'(2476) CCTAACATAATCCGCACCTCATTCC (2452)3'
           ||||||         | || |
3'(3211) caattgtaactataactgcgttatc (3235)5'

5'(2476) CCTAACATAATCCGCACCTCATTCC (2452)3'
         | |||| ||||             |
3'(1194) gtattgcattatatacctctgttag (1218)5'

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

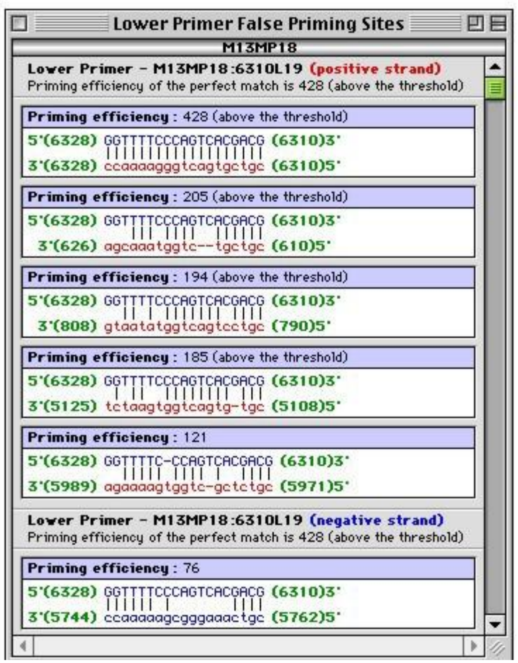
Kdy je primer ještě primerem?

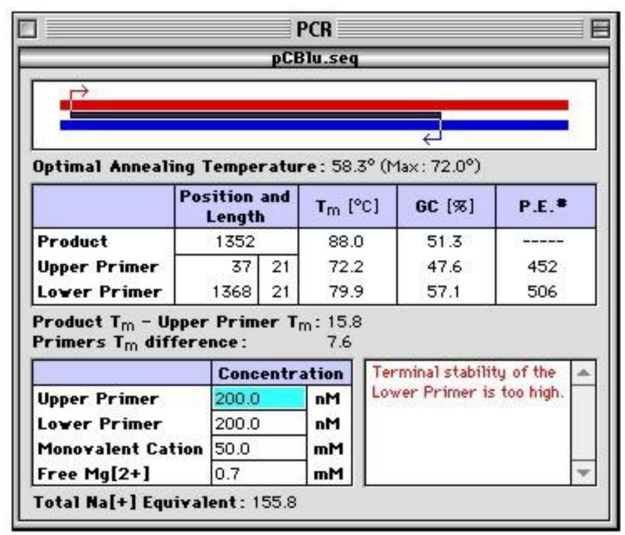


Pro návrh primerů se obvykle používá specializovaný software





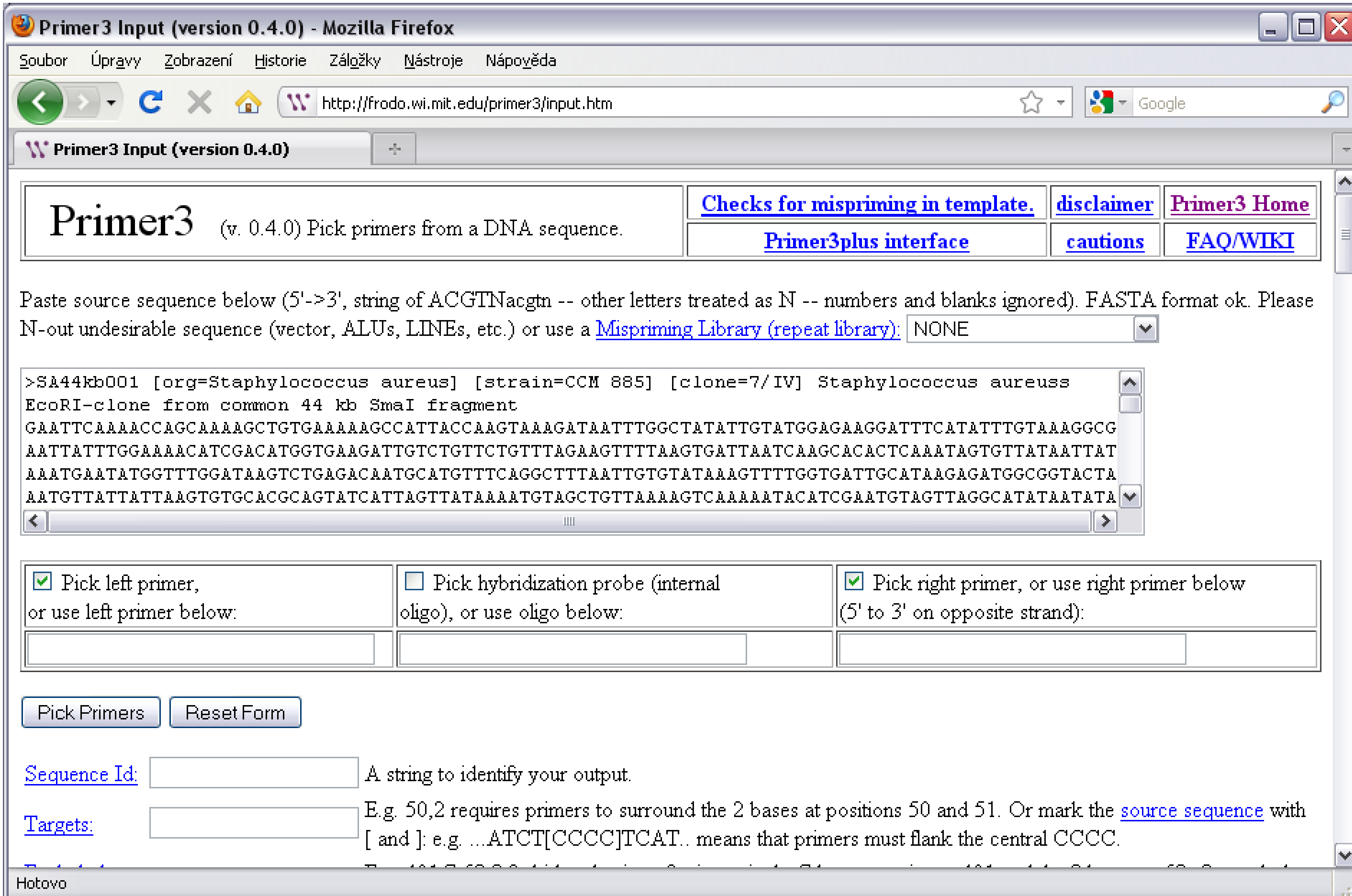


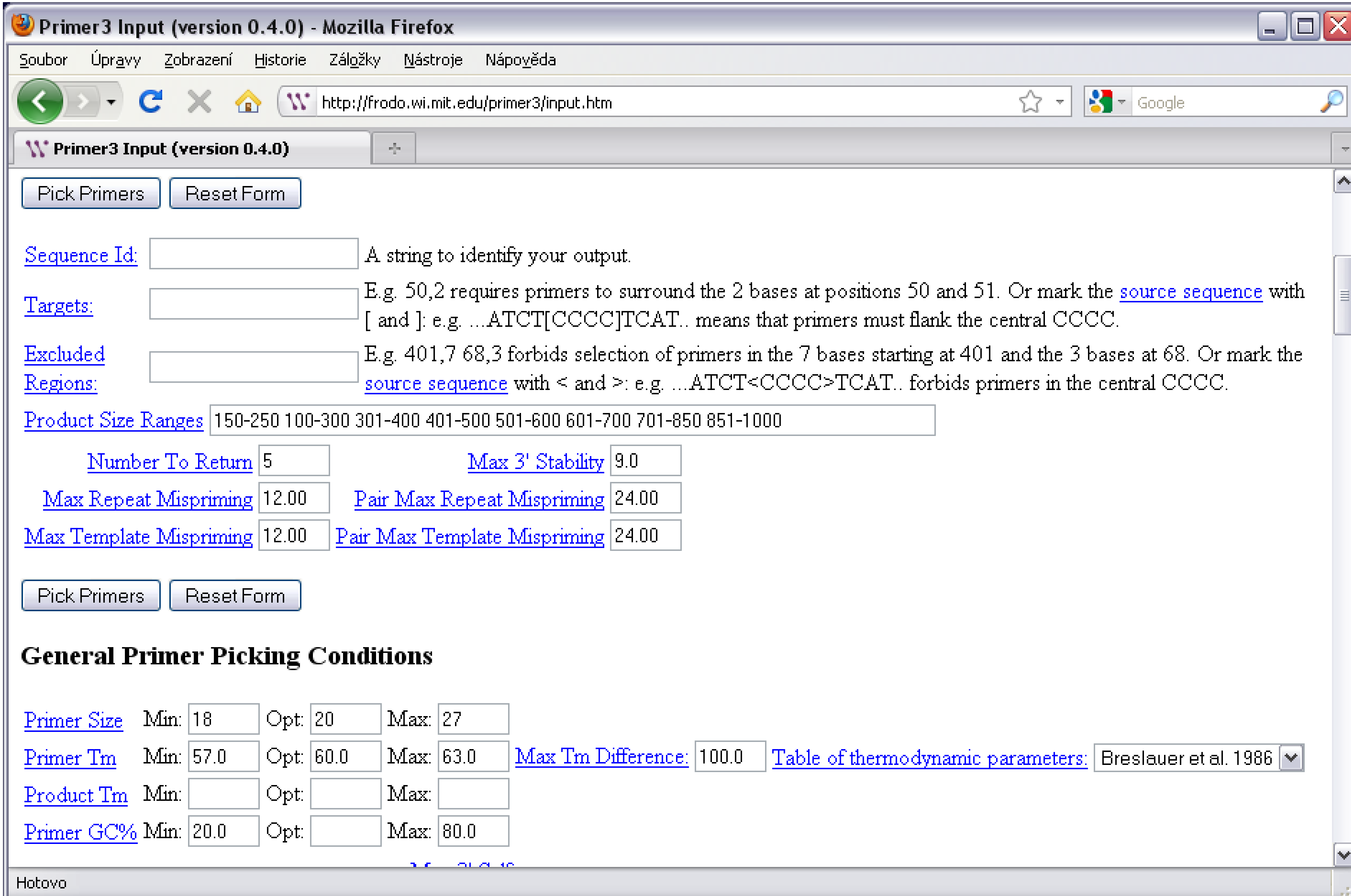


Počítačový návrh primerů

- Umoňuje řada molekulárně biologických programů
- Některé jsou volně dostupné na internetu
  - Primer3
  - Primer3Plus
  - PrimerZ
  - PerlPrimer
  - BioTools
  - WebPrimer
- Kalkulátory vlastností primerů
  - IDT Oligo Analyzer (http://eu.idtdna.com/SciTools/SciTools.aspx?cat=DesignAnalyze)
  - BioMath (http://www.promega.com/biomath/calc11.htm)
  - PrimerBlast
  - UCSC In-Silico PCR
  - AutoDimer

Primer 3 [(http://frodo.wi.mit.edu/primer3/input.htm)](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

Primer3 Output (primer3_results.cgi release 0.4.0) - Mozilla Firefox

Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda

http://frodo.wi.mit.edu/cgi-bin/primer3-web-cgi-bin-0.4.0/primer3_results.cgi

Google

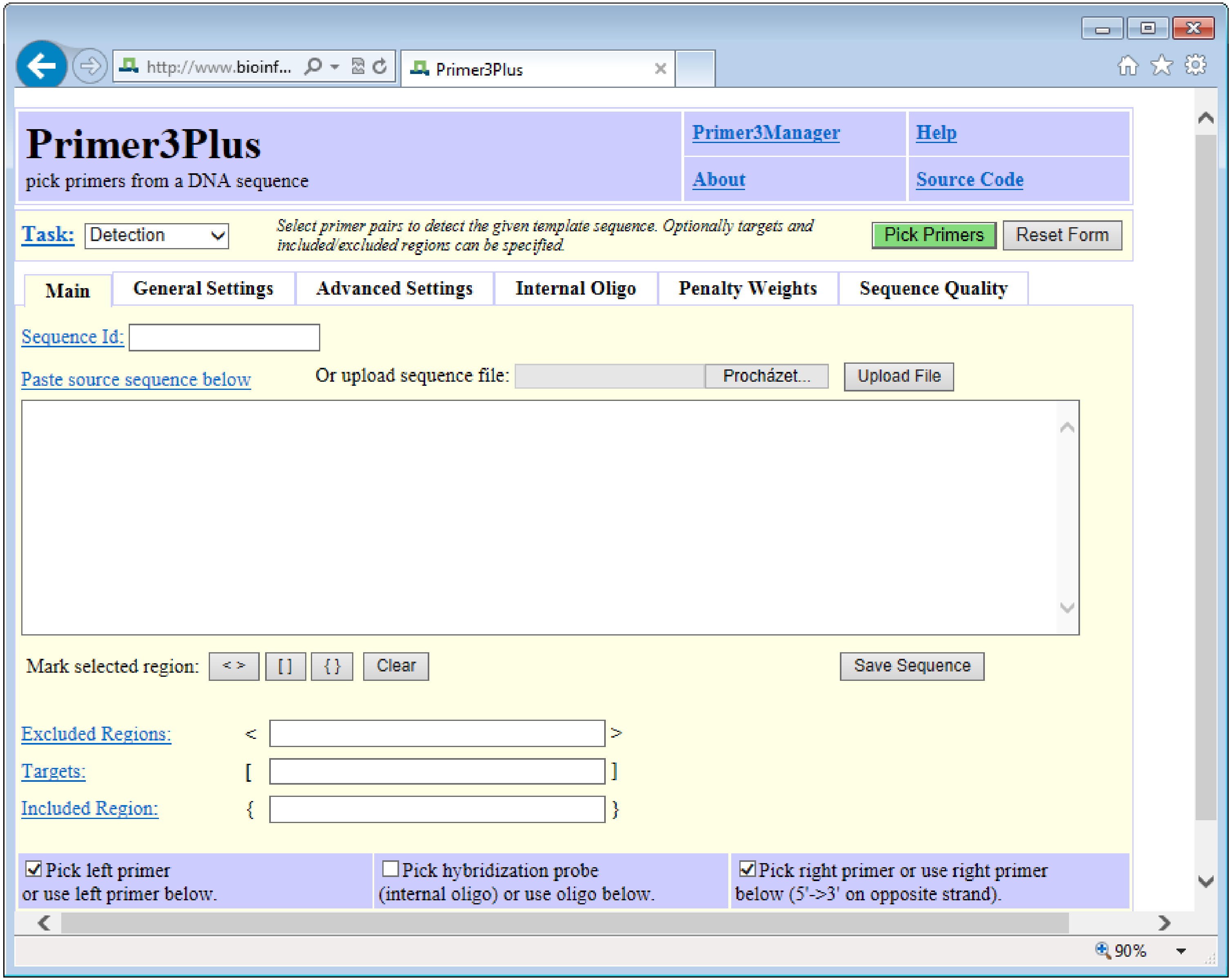
Primer3 Output (primer3_results.cgi...

```
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
      ***************************************     <<<<<<<<<<<<<<<<
```

Hotovo

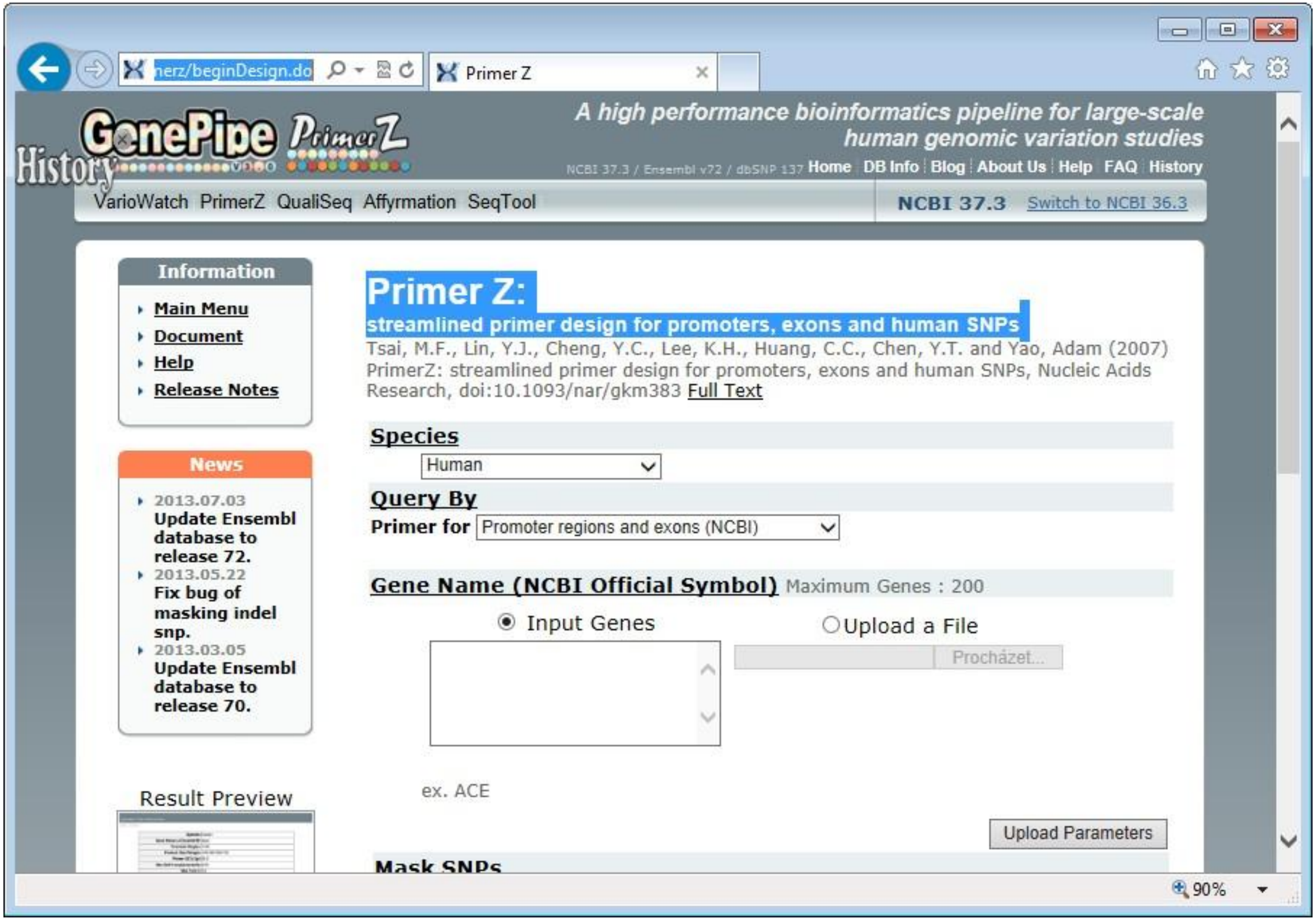
Primer3Plus – rozšířené rozhraní (2007) Primer 3

http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi

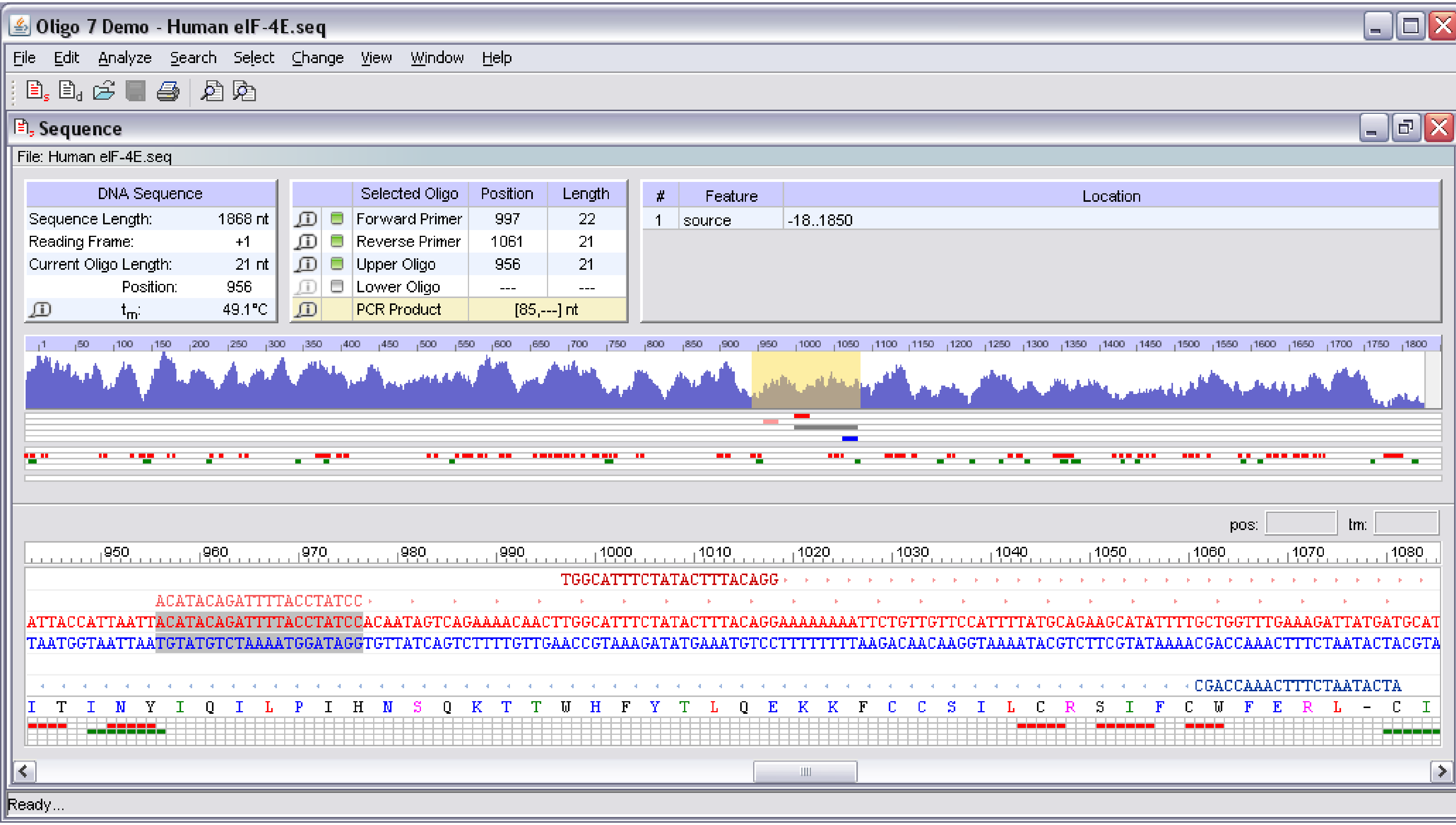


Primer Z:
streamlined primer design for promoters, exons and human SNPs

http://genepipe.ngc.sinica.edu.tw/primerz/beginDesign.do



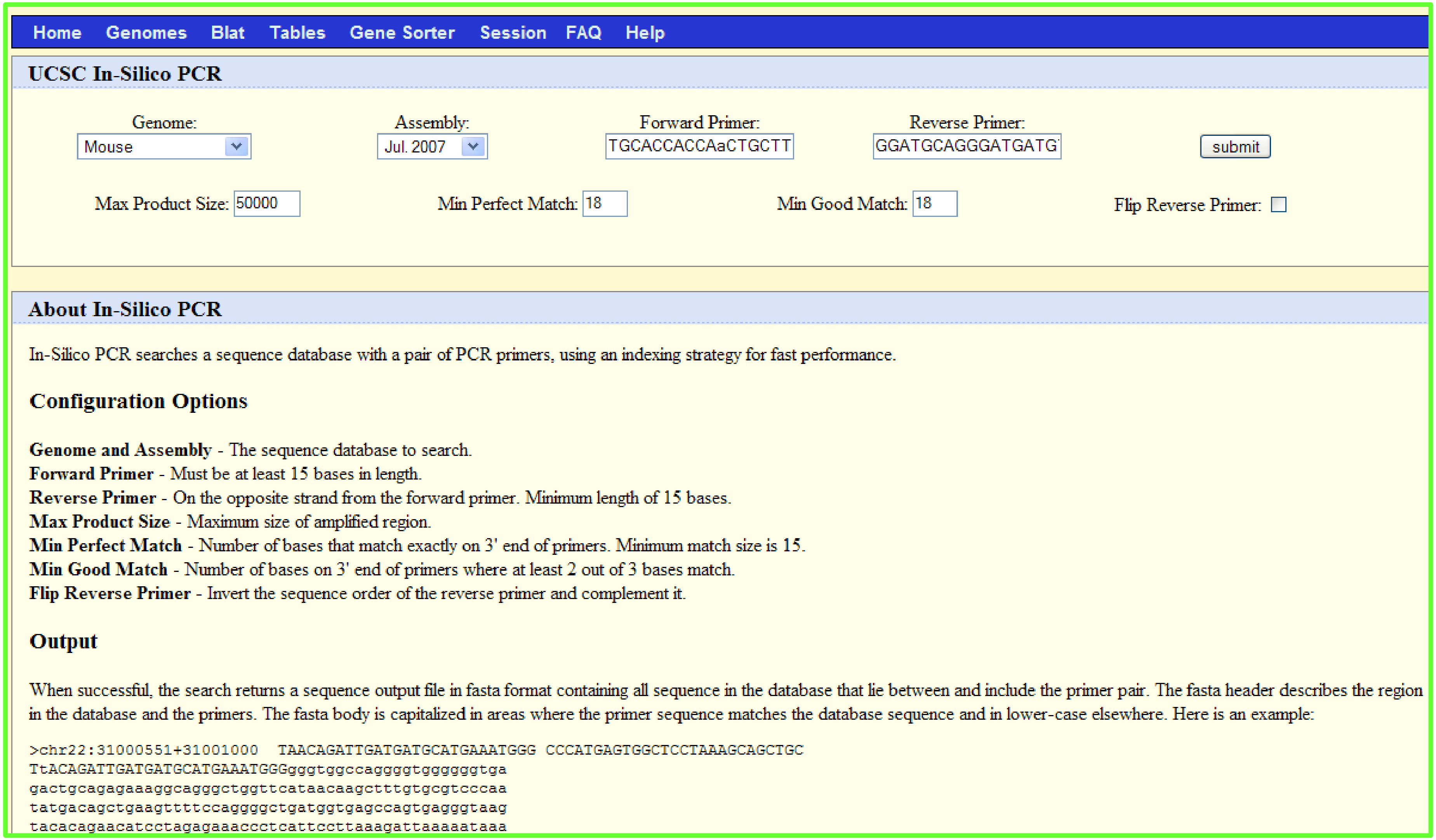
Oligo



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
CGACCAAACTTTCTAATACTA
Ready...

PCR Primer Mapping – UCSC In-Silico PCR

http://genome.ucsc.edu/cgi-bin/hgPcr?db=mm9


Home Genomes Blat Tables Gene Sorter Session FAQ Help

UCSC In-Silico PCR

Genome:	Assembly:	Forward Primer:	Reverse Primer:	
Mouse	Jul. 2007	TGCACCACCAaCTGCTT	GGATGCAGGGATGATG	submit

Max Product Size: 50000 Min Perfect Match: 18 Min Good Match: 18 Flip Reverse Primer:

About In-Silico PCR

In-Silico PCR searches a sequence database with a pair of PCR primers, using an indexing strategy for fast performance.

Configuration Options

Genome and Assembly - The sequence database to search.
Forward Primer - Must be at least 15 bases in length.
Reverse Primer - On the opposite strand from the forward primer. Minimum length of 15 bases.
Max Product Size - Maximum size of amplified region.
Min Perfect Match - Number of bases that match exactly on 3' end of primers. Minimum match size is 15.
Min Good Match - Number of bases on 3' end of primers where at least 2 out of 3 bases match.
Flip Reverse Primer - Invert the sequence order of the reverse primer and complement it.

Output

When successful, the search returns a sequence output file in fasta format containing all sequence in the database that lie between and include the primer pair. The fasta header describes the region in the database and the primers. The fasta body is capitalized in areas where the primer sequence matches the database sequence and in lower-case elsewhere. Here is an example:

```
>chr22:31000551+31001000  TAACAGATTGATGATGCATGAAATGGG CCCATGAGTGGCTCCTAAAGCAGCTGC
TtACAGATTGATGATGCATGAAATGGGgggtggccaggggtggggggtga
gactgcagagaaaggcagggctggttcataacaagctttgtgcgtcccaa
tatgacagctgaagttttccaggggctgatggtgagccagtgagggtaag
tacacagaacatcctagagaaaccctcattccttaaagattaaaaataaa
```

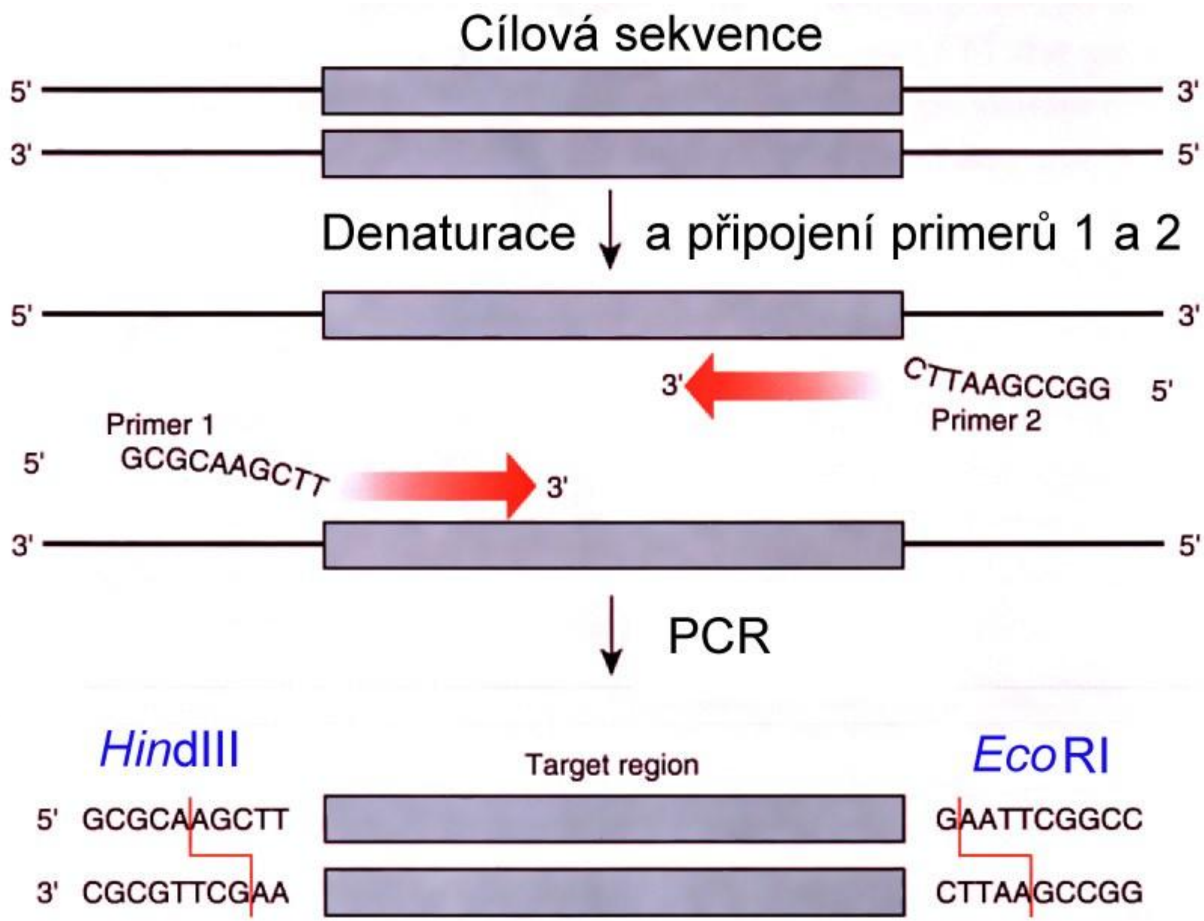
Výsledky

- 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

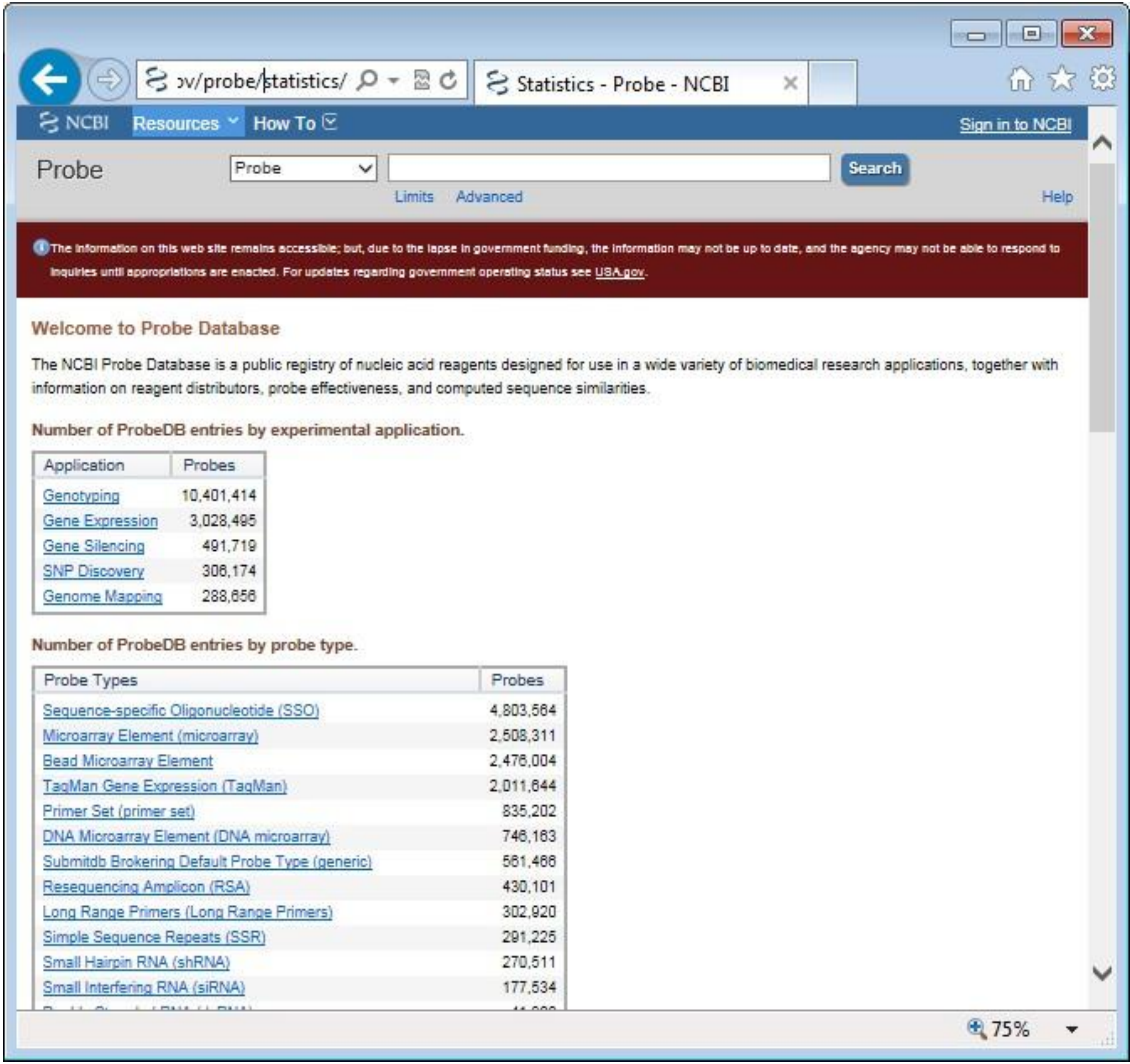
Zdroje pro návrh multiplex PCR

- MultiPLX (http://bioinfo.ebc.ee/multiplx/)
- PrimerStation (http://ps.cb.k.u-tokyo.ac.jp/index.html)
  - Lidský genom
  - Specifikace exonů
  - Vyloučení variabilních oblastí se SNP
- Oligo Explorer (http://www.genelink.com/tools/gl-oe.asp)
  - Posouzení dimerů primerů v multiplexovém uspořádání

Webové zdroje pro design primerů pro real-time PCR

- NCBI Probe Database
- RTPrimerDB
- Primer Bank
- qPrimerDepot
- PCR-QPPD
- PerlPrimer
- Komerční databáze (např. ROCHE,…)

NCBI Probe



ov/probe/statistics/

Statistics - Probe - NCBI

NCBI Resources How To Sign in to NCBI

Probe | Probe | Search

Limits Advanced Help

The information on this web site remains accessible; but, due to the lapse in government funding, the information may not be up to date, and the agency may not be able to respond to inquiries until appropriations are enacted. For updates regarding government operating status see USA.gov.

Welcome to Probe Database

The NCBI Probe Database is a public registry of nucleic acid reagents designed for use in a wide variety of biomedical research applications, together with information on reagent distributors, probe effectiveness, and computed sequence similarities.

Number of ProbeDB entries by experimental application.

Application	Probes
Genotyping	10,401,414
Gene Expression	3,028,495
Gene Silencing	491,719
SNP Discovery	306,174
Genome Mapping	288,656

Number of ProbeDB entries by probe type.

Probe Types	Probes
Sequence-specific Oligonucleotide (SSO)	4,803,564
Microarray Element (microarray)	2,508,311
Bead Microarray Element	2,476,004
TaqMan Gene Expression (TaqMan)	2,011,644
Primer Set (primer set)	835,202
DNA Microarray Element (DNA microarray)	746,163
Submitdb Brokering Default Probe Type (generic)	561,466
Resequencing Amplicon (RSA)	430,101
Long Range Primers (Long Range Primers)	302,920
Simple Sequence Repeats (SSR)	291,225
Small Hairpin RNA (shRNA)	270,511
Small Interfering RNA (siRNA)	177,534

75%

Další technologie vyžadující návrh oligonukleotidů

- Real-time PCR
  - TaqMan
  - Molecular Beacons
- Primer extension
- Sekvenování
  - Sangerovo sekvenování
  - Pyrosekvenování
- Ligázová řetězová reakce
- Microarrays