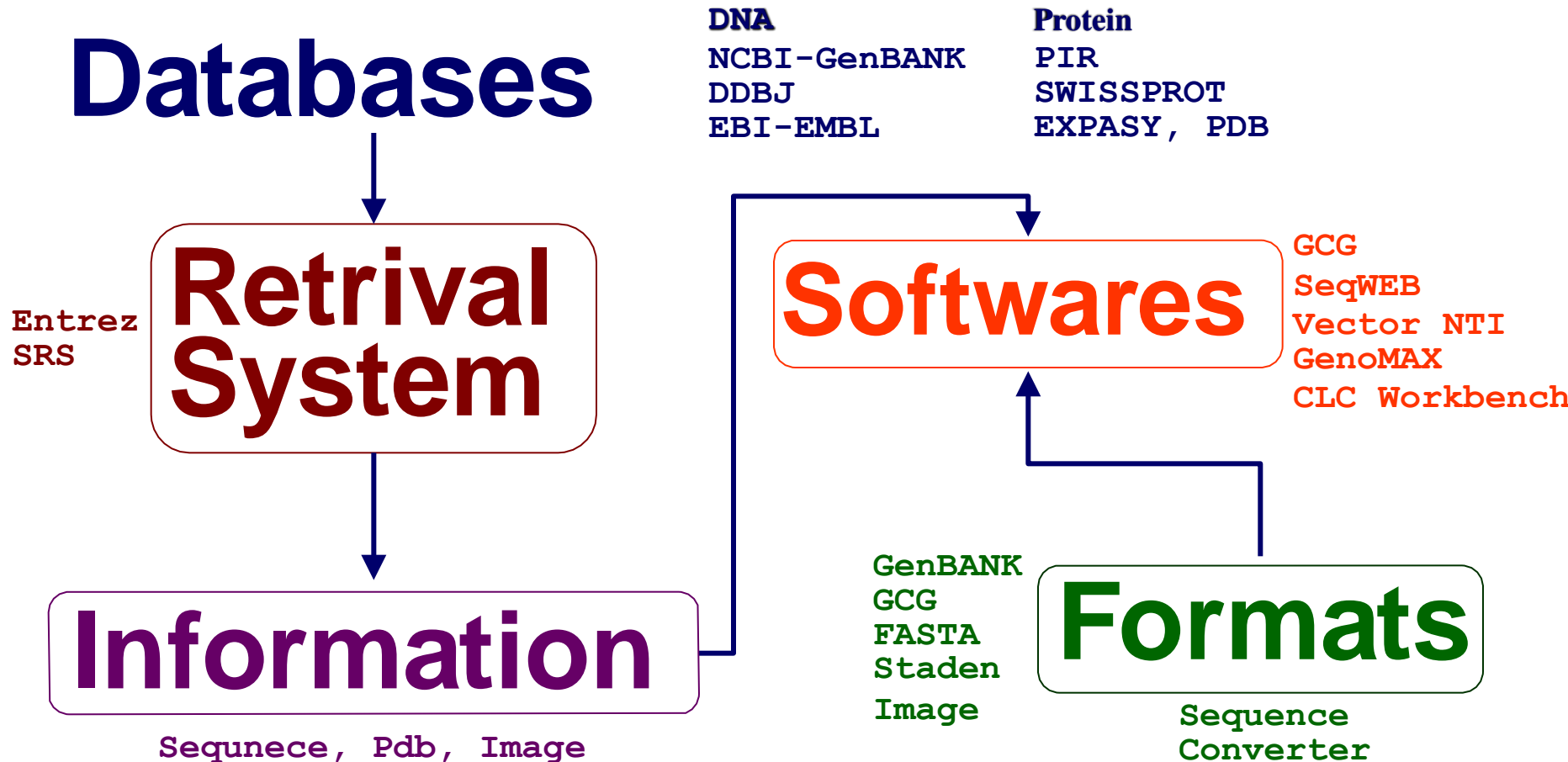


Manipulace se sekvenčními daty

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/>

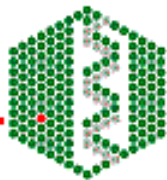
National Institute of Genetics (NIG)

EMBL: <http://www.ebi.ac.uk>

EMBL

European Bioinformatics Institute (EBI)

European Bioinformatics Institute



ExpASY: <http://tw.expasy.org>

Ex p er t P r o t e i n A n a l y s i s S y s t e m

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 (pu <u>R</u> in)
Y	T/C (p <u>Y</u> rimidin)
K	G/T (nukleosid s <u>K</u> eto skupinou)
M	A/C (nukleosid s a <u>M</u> ino skupinou)
S	G/C (silná = <u>S</u> trong vazba)
W	A/T (slabá = <u>W</u> eak 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 příložená

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

PLAIN SEQUENCE FORMAT

Obsahuje pouze IUPAC znaky

Obsahuje jedinou sekvenci

Příklad

```
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAA  
CCTCCCATCCGTGTCTATTGTACCCTGTTGCTTCGGCGGGCCCGC  
CGCTTGTCGGCCGCCGGGGGGGGCGCCTCTGCCCGGGCCCGTG  
CCCGCCGGAGACCCAACACGAACACTGTCTGAAAGCGTGCAGTC  
TGAGTTGATTGAATGCAATCAGTTAAAACTTCAACAATGGATCT
```

FASTA FORMAT

Může obsahovat více sekvencí

Začíná specifickým záhlavím („>“)

Příklad:

```
>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAACT
TTCAACAATGGATCTCTTGGTTCCGGC
```

EMBL FORMAT

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

Sekvence začíná 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 cgcttgctcg ccgccggggg ggcgcctctg      120
      ccccccgggc ccgtgccgcg cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc      180
      tgagttgatt gaatgcaatc agttaaactt 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 tattgtacc  tgttgcttcg gcgggccgcg cgcttgtcgg ccgccggggg ggcgcctctg
            121 cccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc
            181 tgagttgatt gaatgcaatc agttaaact  ttcaacaatg gatctcttgg ttccggc
//
```

Poznámka k používaným fontům

- Proporcionální fonty

- ◆ Arial, Times
- ◆ Každý znak jiná šířka
- ◆ Nevhodné

gaattttttt
cttaaaaaaa

- Neproporcionální fonty

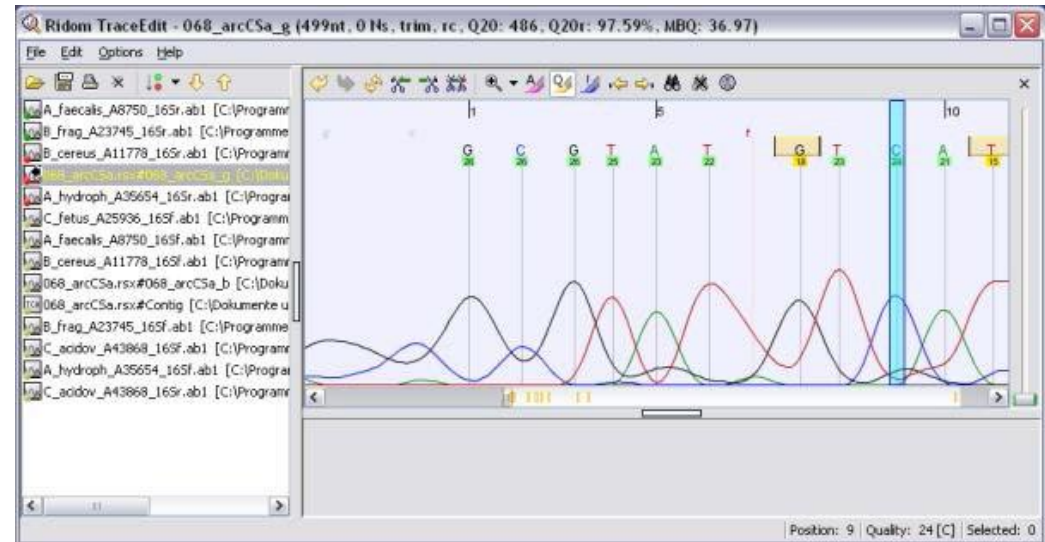
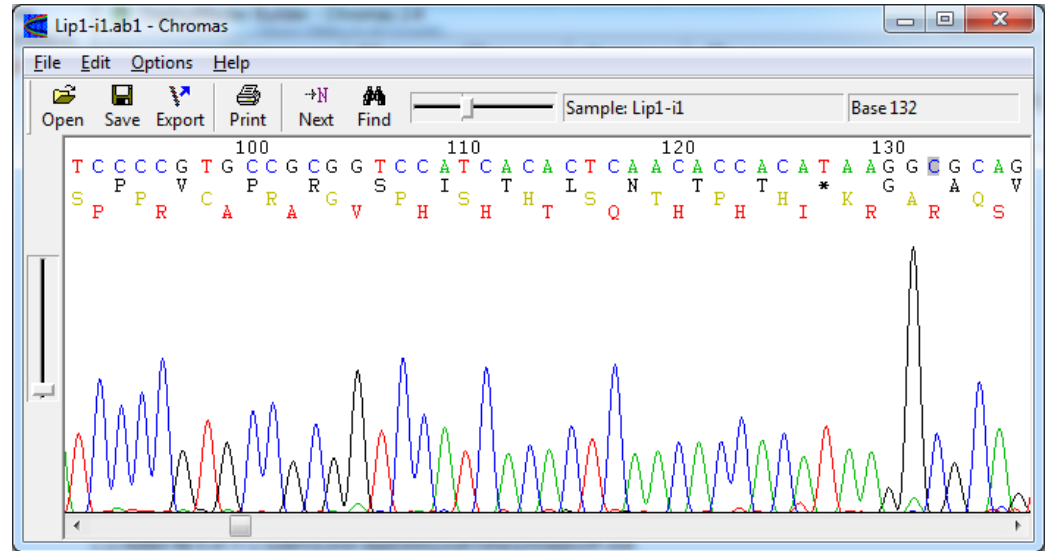
- ◆ Vhodné k použití
- ◆ Všechny znaky stejná šířka
- ◆ Courier, Monospaced

gaatttttttt
cttaaaaaaaa

- 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 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 – výstup tabulka
 - ◆ kompletní sada
 - ◆ komerční sada
 - ◆ které sekvenci neštěpí
 - ◆ které štěpí – počet a pozice rozpoznávacích míst
- Lineární nebo kružnicová mapa sekvence se znázorněním pozice restričních míst
 - ◆ Grafika
 - ◆ Identifikace ORF a translace do proteinu

NEB Cutter

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

NEW ENGLAND BioLabs Inc. NEBcutter

Circular Sequence: L08752 [Help](#) [Comments](#)

Display: - NEB single cutter restriction enzymes
- Main non-overlapping, min. 100 aa ORFs

GC=51%, AT=49%

Cleavage code	Enzyme name code
✂ blunt end cut	Available from NEB
▾ 5' extension	Has other supplier
▾ 3' extension	Not commercially available
▾ cuts 1 strand	*: cleavage affected by CpG meth.
	#: cleavage affected by other meth.
	(enz.name): ambiguous site

ORFs:
a: 286 aa
b: 133 aa
c: 118 aa

2686 bp

**WARNING: Not all enzymes shown
See linear display**

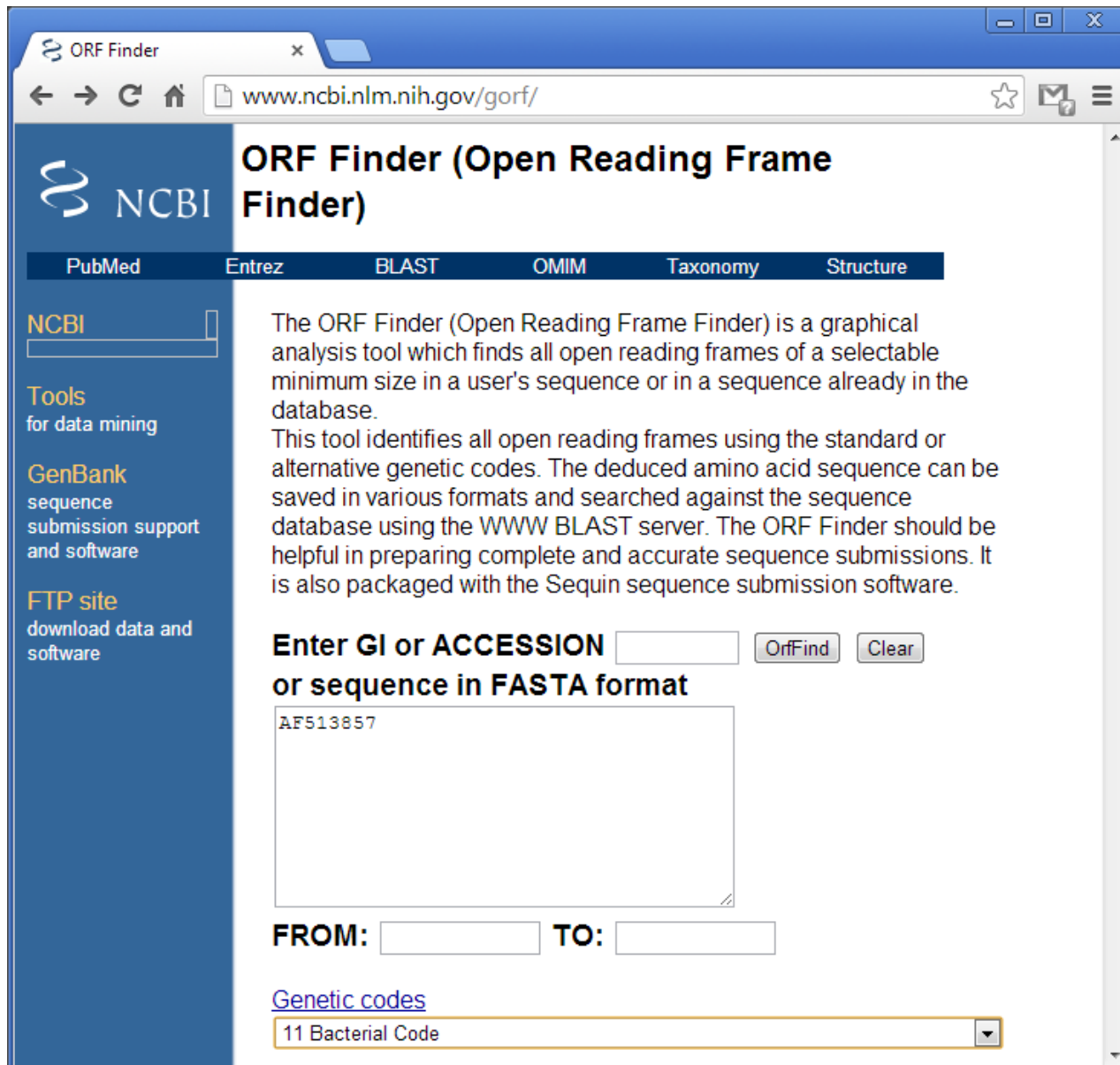
Enzyme sites shown on the circular map include: PciI, AflIII, BspQI, SapI, BsaXI, BseYI, AlwNI, XmnI, *BcgI, ScaI, *BsrFI, BpmI, BsaI, *TspMI, *AvaI, *SmaI, BamHI, XbaI, *SalI, *AccI, *HincII, SbfI, PstI, BfuAI, BspMI, SphI, HindIII, *PluI, *SfoI, *NarI, *KasI, BstAPI, NdeI, Eco0109I, *AatII, *ZraI, SspI.

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>



The screenshot shows the NCBI ORF Finder web interface. The browser address bar displays www.ncbi.nlm.nih.gov/gorf/. The page title is "ORF Finder (Open Reading Frame Finder)". A navigation bar includes links for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area contains a description of the tool and a form for inputting a sequence. The left sidebar provides links to NCBI, Tools for data mining, GenBank, and FTP site.

ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI

Tools for data mining

GenBank sequence submission support and software

FTP site download data and software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.

This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION OrfFind Clear

or sequence in FASTA format

```
AF513857
```

FROM: TO:

[Genetic codes](#)

11 Bacterial Code

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/

The screenshot shows the EMBOSS Transeq web interface in a browser window. The browser's address bar shows the URL http://www.ebi.ac.uk/Tools/st/emboss_transeq/. The page header includes the EMBL-EBI logo and navigation links for Services, Research, Training, Industry, and About us. The main heading is "EMBOSS Transeq". Below the heading are links for "Input form", "Web services", and "Help & Documentation", along with "Share" and "Feedback" options. The breadcrumb trail is "Tools > Sequence Translation > EMBOSS Transeq".

EMBOSS Transeq

EMBOSS Transeq translates nucleic acid sequences to their corresponding peptide sequences. It can translate to the three forward and three reverse frames, and output multiple frame translations at once.

STEP 1 - Enter your input sequence

Enter or paste a set of sequences in any supported format:

Or, upload a file:

- 1
- 2
- 3
- F (Forward three frames)
- 1
- 2
- 3
- R (Reverse three frames)
- 6 (All six frames)

CODON TABLE

ost users and, for that reason, are not visible.

100%

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

```
LLIQQAKSNSDITTPAMPLDTCGAMSQGMIGYWLETEINRILTEMNSDRTVGTIVTRVEVD  
KDDPRFDNPTKPIGPFYTKEEVEELQKEQPDSVFKEDAGRGYRQVVASPLPQSILEHQLI  
QTLADGKNIVIACGGGGIPVIKKENTYEGVEA
```

5'3' Frame 2

```
Y-SNKLNRVTQRRQCHWILVVQCHRV--AIGWKLKSI AF-LK-IVIEL-AQSLHVWK-I  
KMIHDLITQLNQLVLFIRKKKLKKNYKKN SQTQSLKMKQDVVIEK-LRHHYLNLY-NTS-F  
KL-QTVKILSLHAVVAVFQL-KKKIPMKVLK
```

5'3' Frame 3

```
INPTS-IEQ-HNAGNAIGYLWCNVTGYDRLLVGN-NQSHFN-NE---NCRHNR YTCGSR-  
R-STI--PN-TNWSFLYERRS-RITKR TARLSL-RRCRTWL-KSSCVTTTTSIYTRTPVNS  
NFSRR-KYCHC MRWWRYSYKKRKYL-RC-S
```

Příklady translace *in silico*

EMBOSS Sixpack

[Input form](#) | [Web services](#) | [Help & Documentation](#)

[Tools](#) > [Sequence Translation](#) > EMBOSS Sixpack

Results for job emboss_sixpack-l20141006-192122-0940-32029869-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 TtattAaTccaACAaGCTAAaATCGAACaGTGACACAACGCCGCAATGCCATTGGATACT 60
----:----|----:----|----:----|----:----|----:----|----:----|
1 AataaTtAggtTGTtCGATTtaGCTTtGtCACTGTGTTGCGGCCGTTACGGTAACCTATGA 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 TGTGGTGCAATGTCCAGGGTATGATAGGCTATTGGTTGGAACCTGAAATCAATCGCATT 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

The screenshot displays the Vector NTI software interface for a plasmid named pBR322. The main window shows a circular map of the plasmid with a total length of 4361 bp. Key features and restriction sites are labeled:

- SD SEQ
- P3 P
- Cla* I (25)
- Hind* III (30)
- P2 P
- Bam* HI (3)
- TCr
- APr
- pa* LI (4034)
- (3612)

The left sidebar shows the 'General' tab with the following information:

- General
- DNA Plasmid
- ATCC 37017
- length: 4361 bp
- storage type: ...
- form: Circular
- Function:
- CDS [3]
- Misc_features [1]
- Promoters [1]
- RBS [2]

The bottom of the window shows a sequence viewer with the following DNA sequence:

```
151 CTTGGTTATG CCGGTACTGC CGGGCCTCTT GCGGGATATC GTCCATTCCG
    GAACCAATAC GGCCATGACG GCCCGGAGAA CGCCCTATAG CAGGTAAGGC
201 ACAGCATCGC CAGTCACTAT GGC GTGCTGC TAGCGCTATA TGCGTTGATG
    TGTCGTAGCG GTCAGTGATA CCGCACGACG ATCGCGATAT ACGCAACTAC
251 CAATTTCTAT GCGCACCCGT TCTCGGAGCA CTGTCCGACC GCTTTGGCCC
```

The status bar at the bottom indicates 'Ready' and shows a zoom level of '200 bp - 1200 bp' and a current position of '1201 bp'.

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	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

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

Clone Manager (Sci-Ed Software)

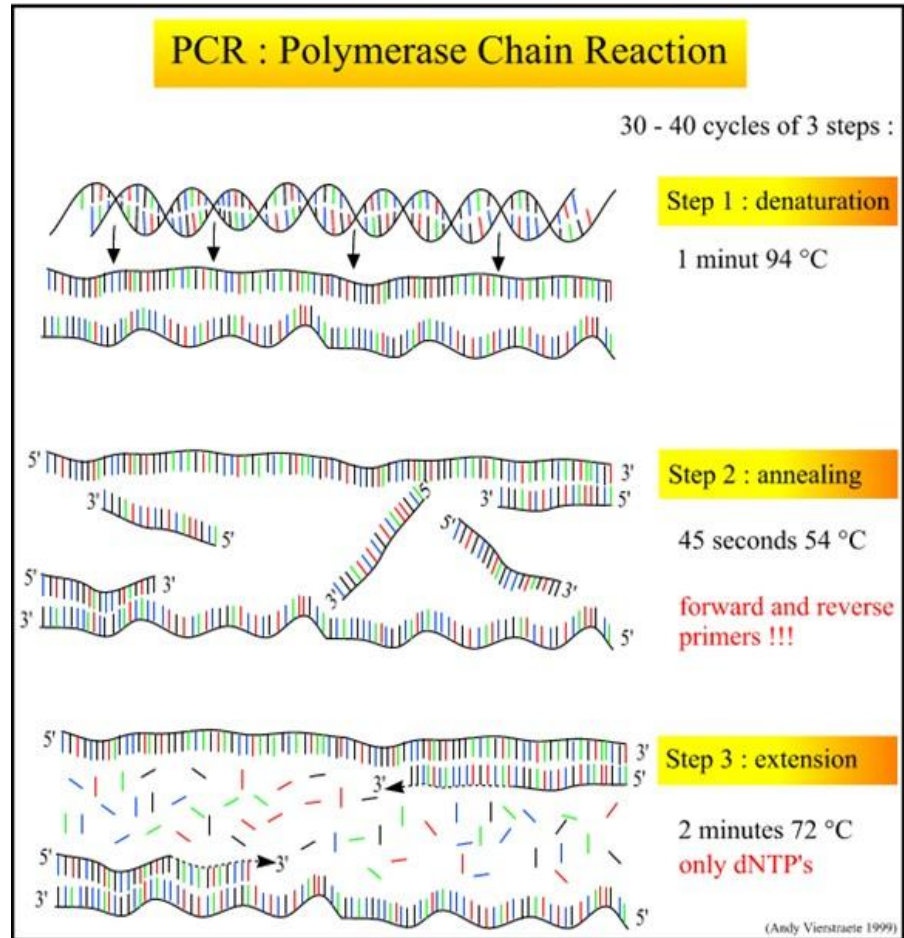
http://www.sci-ed.com/pr_cmbas.htm

The screenshot displays the Clone Manager software interface. The main window shows a circular plasmid map for a 2686bp construct named SYNPU18V. The map is annotated with various restriction enzyme sites. A list of 44 enzyme sites is shown on the right, with the ApoI site highlighted in blue. The ApoI site is located at position 230 and has a 5' overhang.

Name	Pos	Type
ApoI	230	sc 5'
EcoRI	230	sc 5'
BanII	236	sc 3'
Eco53kI	236	sc bl
SacI	236	sc 3'
Acc65I	242	sc 5'
KpnI	242	sc 3'
AvaI	246	sc 5'
SmaI	246	sc bl
XmaI	246	sc 5'
BamHI	251	sc 5'
XbaI	257	sc 5'
AccI	263	sc 5'
HincII	263	sc bl
SalI	263	sc 5'
BspMI	267	sc 5'
SbfI	268	sc 3'
PstI	269	sc 3'
SphI	275	sc 3'

Navrhování sekvencí primerů pro PCR

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



PCR - Syntéza obou řetězců u specifické sekvence



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 příložená 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 T_m
 - ◆ Specificita
 - ◆ Komplementarita primerových sekvencí
 - ◆ Sekvence 3'-konce

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'

TGGCCTAGCATGCT
TGGCCTAGCATGCT

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:

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

$$T_a = T_m - 5 \text{ °C}$$

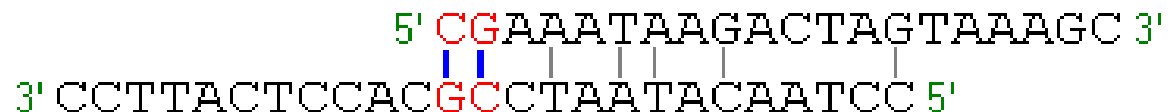
Vnitřní sekvence a struktura primeru

- 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:



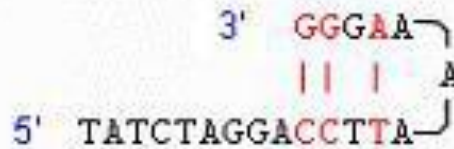
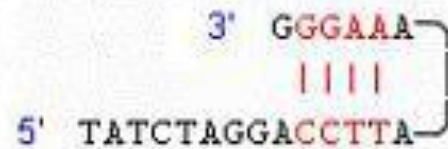
- ◆ Správně navržená dvojice primerů, která vytváří pouze málo stabilní duplex na 5'-konci; na 3'-konci je G nebo C zaručující stabilní párování s templátem:



- ◆ Chybně navržený primer, vytvářející vlásenku:

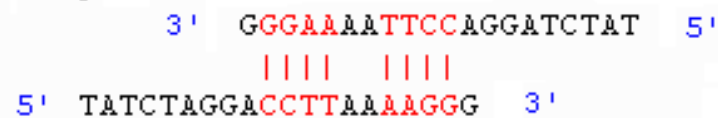


Hairpin

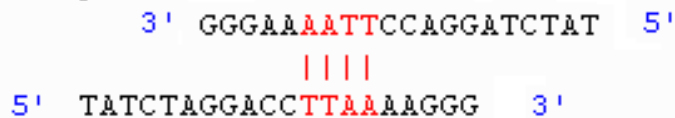


Self-Dimer

8 bp

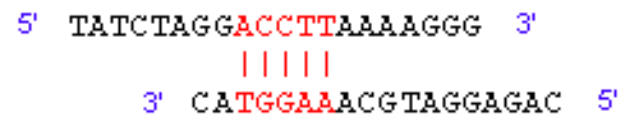


4 bp



Dimer

forward primer



reverse primer

GC svorky a 3'-koncová stabilita

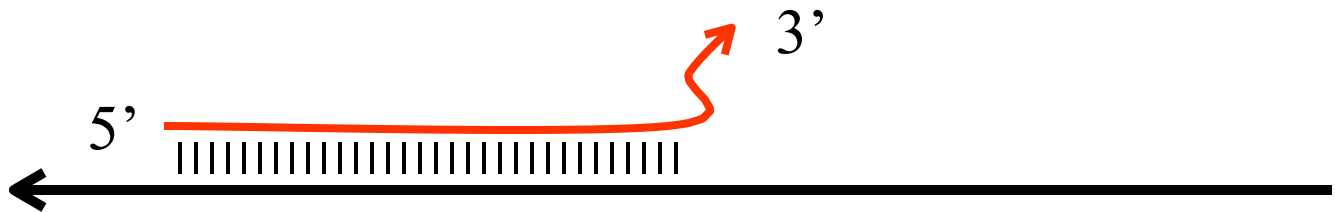
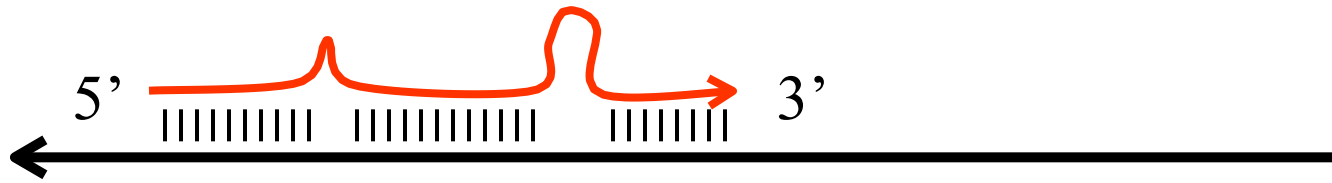
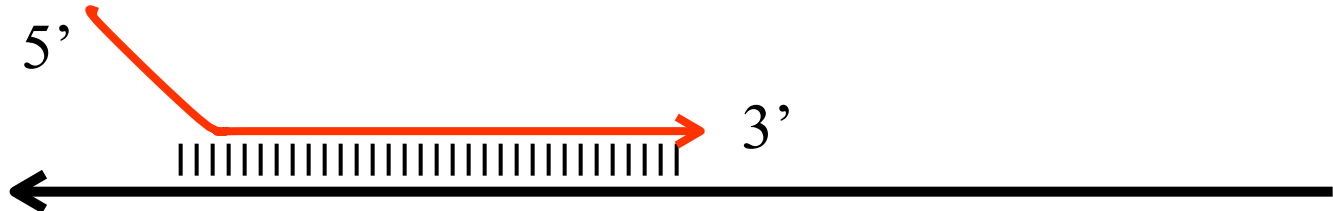
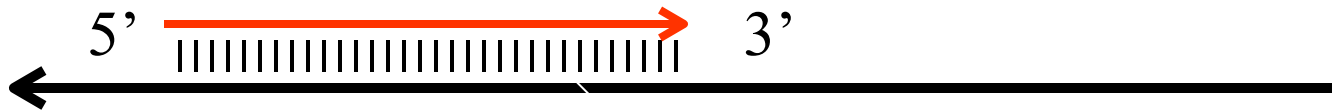
■ GC svorka

- ◆ 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

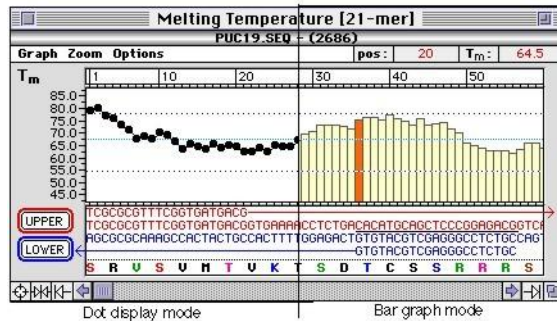
■ Maximální 3'-koncová stabilita

- ◆ Maximalizace ΔG posledních 5 bází na 3'-konci primeru.

Kdy je primer ještě primerem?



Pro návrh primerů se obvykle používá specializovaný software



Current Oligo
pCBlu3_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' GATGAGCATTTCATCAGGCGGG 3'
 P.E.#: 537
 1/E: 4.80 nmol/A₂₆₀
 31.7 μg/A₂₆₀

Selected Primers
pCBlu3_seq

pCBlu3:269U21 Upper Primer

5' CGGGCCAGATCTGGTACCA 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

pCBlu3:817L21 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) GGTTCCTCCAGTCACGACG (6310)3'
 3'(6328) ccaaaaagggtcagtgctgc (6310)5'

Priming efficiency: 205 (above the threshold)

5'(6328) GGTTCCTCCAGTCACGACG (6310)3'
 3'(626) agcaaatggctc--tgc tgc (610)5'

Priming efficiency: 194 (above the threshold)

5'(6328) GGTTCCTCCAGTCACGACG (6310)3'
 3'(808) gtaataggctcagtcctgc (790)5'

Priming efficiency: 185 (above the threshold)

5'(6328) GGTTCCTCCAGTCACGACG (6310)3'
 3'(5125) tc taagtggctcagtg-tgc (5108)5'

Priming efficiency: 121

5'(6328) GGTTCCTCCAGTCACGACG (6310)3'
 3'(5989) agaaaag tgg tc-gc tc tgc (5971)5'

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

Priming efficiency: 76

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

PCR
pCBlu3_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
 - ◆ 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>

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
GAATTCAAAACCCAGCAAAAAGCTGTGAAAAAGCCATTACCAAGTAAAGATAATTTGGCTATATTGTATGGAGAAGGATTCATATTTGTAAGGCG
AATTATTTGGAAAACATCGACATGGTGAAGATTGTCTGTTCTGTTTAAAGTGTGATTAATCAAGCACACTCAAATAGTGTATATAATTAT
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 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

Primer3Plus – rozšířené rozhraní (2007) Primer 3

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

The screenshot shows the Primer3Plus web interface. At the top, there is a navigation bar with links for [Primer3Manager](#), [Help](#), [About](#), and [Source Code](#). The main heading is **Primer3Plus** with the subtitle "pick primers from a DNA sequence".

The **Task:** dropdown menu is set to "Detection". A descriptive text reads: "Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified." There are "Pick Primers" and "Reset Form" buttons.

Navigation tabs include **Main**, **General Settings**, **Advanced Settings**, **Internal Oligo**, **Penalty Weights**, and **Sequence Quality**.

The **Sequence Id:** field is empty. Below it, there is a "Paste source sequence below" link and an "Or upload sequence file:" section with a "Procházet..." button and an "Upload File" button. A large text area for the sequence is currently empty.

Below the text area, there are "Mark selected region:" buttons: "< >", "[]", "{ }", and "Clear". A "Save Sequence" button is also present.

There are three input fields for region specifications:

- Excluded Regions:** < [] >
- Targets:** []
- Included Region:** { }

At the bottom, there are three checkboxes:

- Pick left primer or use left primer below.
- Pick hybridization probe (internal oligo) or use oligo below.
- Pick right primer or use right primer below (5'→3' on opposite strand).

The browser window title is "Primer3Plus" and the address bar shows "http://www.bioinf...". The zoom level is 90%.

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

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

The screenshot shows the GenePipe PrimerZ web application interface. The browser address bar displays the URL `http://genepipe.ngc.sinica.edu.tw/primerz/beginDesign.do`. The page header includes the logo "GenePipe PrimerZ" and the tagline "A high performance bioinformatics pipeline for large-scale human genomic variation studies". Navigation links for "Home", "DB Info", "Blog", "About Us", "Help", "FAQ", and "History" are present. A version selector shows "NCBI 37.3" with a "Switch to NCBI 36.3" option. The main content area is titled "Primer Z: streamlined primer design for promoters, exons and human SNPs" and includes a citation: "Tsai, M.F., Lin, Y.J., Cheng, Y.C., Lee, K.H., Huang, C.C., Chen, Y.T. and Yao, Adam (2007) PrimerZ: streamlined primer design for promoters, exons and human SNPs, Nucleic Acids Research, doi:10.1093/nar/gkm383 Full Text". The interface features several form fields: "Species" (set to "Human"), "Query By" (set to "Promoter regions and exons (NCBI)"), and "Gene Name (NCBI Official Symbol)" (with a "Maximum Genes : 200" limit). There are two radio buttons for "Input Genes" (selected) and "Upload a File", with a "Procházet..." button next to the file upload option. A "Result Preview" section shows a table of results, and an "Upload Parameters" button is located at the bottom right. The status bar at the bottom indicates a zoom level of 90%.

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 ·

ATTACCATTAATTACATACAGATTTTACCTATCCACAATAGTCAGAAAAACAATTGGCATTTCATATACTTTACAGGAAAAAAAAATTCTGTTGTTCCATTTTATGCAGAAGCATATTTTGCTGGTTTGAAAAGATTATGATGCAT
TAATGGTAATTAATGTATGTCTAAAATGGATAGGTGTTATCAGTCTTTTGTGAACCGTAAAGATATGAAATGTCTTTTTTTTAAGACAACAAGGTAAAATACGTCTTCGTATAAAAACGACCAAACCTTTCTAATACTACGTA

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...

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:

Max Product Size: Min Perfect Match: Min Good Match: 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  
gactgcagagaaaggcagggctggttcataacaagctttgtgogtcccaa  
tatgacagctgaagttttccaggggctgatggtgagccagtgagggtaaag  
tacacagaacatcctagagaaacccctcattccttaagattaaaaataaa
```

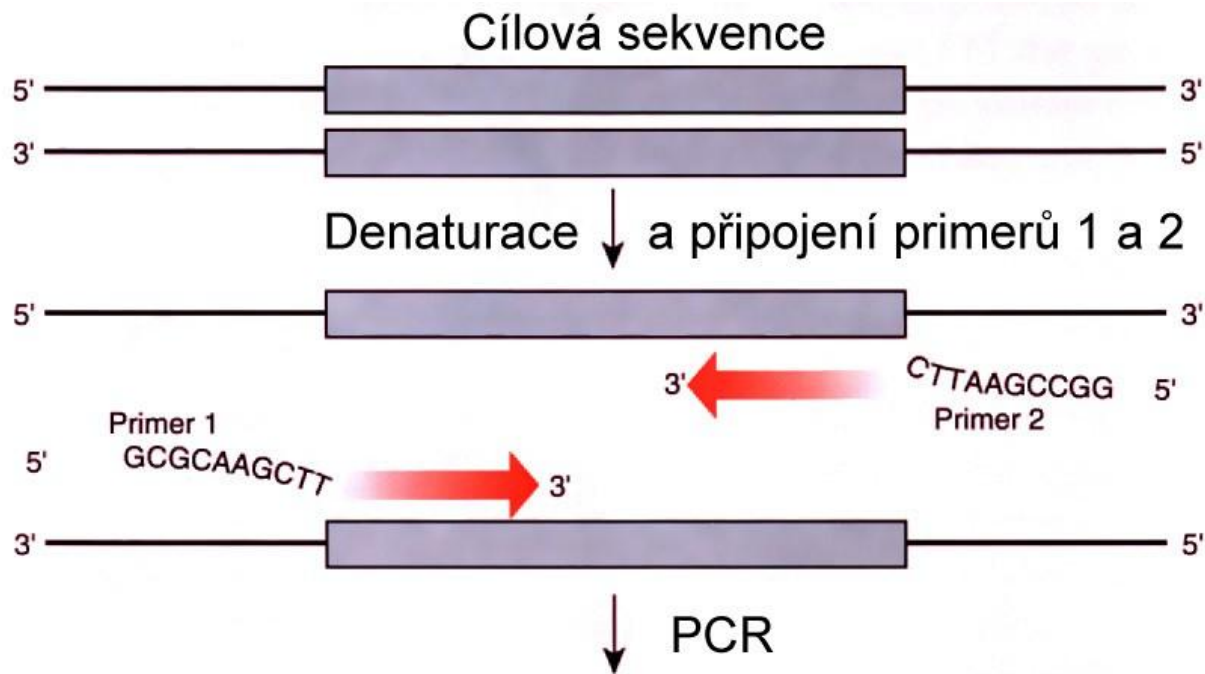
Výsledky

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

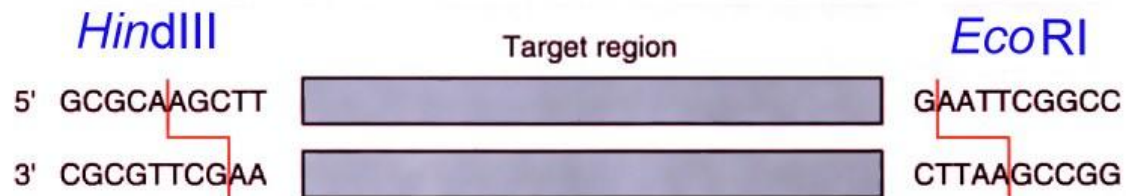
Pokročilý návrh primerů

- 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

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/glo-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

The screenshot shows the NCBI Probe Database website. At the top, there is a search bar with a dropdown menu set to 'Probe' and a 'Search' button. Below the search bar, there are links for 'Limits' and 'Advanced'. A red banner contains a disclaimer: '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](#).' Below the banner, the page is titled 'Welcome to Probe Database' and includes a paragraph describing the database as a public registry of nucleic acid reagents. Two tables follow, providing statistics on probe entries.

Statistics - Probe - NCBI

NCBI Resources How To Sign in to NCBI

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	308,174
Genome Mapping	288,858

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
TagMan Gene Expression (TagMan)	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