MUNI
SCI

MASARYK UNIVERSITY
FACULTY OF SCIENCE
DEPARTMENT OF EXPERIMENTAL BIOLOGY
LABORATORY OF MICROBIAL MOLECULAR DIAGNOSTICS

Bi5000 – Bioinformatika
Lekce 02

Manipulace se sekvenčními daty

Sylabus výuky předmětů Bi5000 Bioinformatika a Bi5000c Bioinformatika - cvičení v semestru podzim 2024

Přednášky:		**úterý 11:00-12:50** prezenčně v učebně B11-306
	Kontakt:	prof. Mgr. Jiří Damborský, Dr. (1441@mail.muni.cz)
		prof. RNDr. Roman Pantůček, Ph.D. (pantucek@sci.muni.cz)
Cvičení:		**středa 10:00-17:00** prezenčně v učebně B09-316, dle seminárních skupin
	Kontakt:	Ing. Miloš Musil (imusilm@fit.vutbr.cz)
		prof. RNDr. Roman Pantůček, Ph.D.

Datum	č.	Předmět	Lekce
17.09.2024			Výuka se nekoná z důvodu povodní
18.09.2024			
24.09.2024	2	Bi5000	Manipulace se sekvenčními daty
25.09.2024		Bi5000c	Cvičení - textové vyhledávání v databázích
01.10.2024	1	Bi5000	Bioinformatika – základní definice, molekulárně biologické databáze
02.10.2024		Bi5000c	Cvičení – formáty sekvencí, manipulace se sekvenčními daty
08.10.2024	3	Bi5000	Posuzování podobnosti sekvencí nukleových kyselin a proteinů
09.10.2024		Bi5000c	Cvičení – párové přiložení sekvencí, BLAST
15.10.2024	4	Bi5000	Mnohonásobné přiložení sekvencí a fylogeneze
16.10.2024		Bi5000c	Cvičení - mnohonásobné přiložení a fylogeneze
22.10.2024	5	Bi5000	Genomové projekty, sekvenování nové generace
23.10.2024		Bi5000c	Cvičení – návrh oligonukleotidů pro PCR, sekvenování, klonování a mutagenezi
29.10.2024	6	Bi5000	Počítačové vyhledávání genů a srovnávací genomika
30.10.2024		Bi5000c	Cvičení – hledání prokaryotických genů, práce s programy pro srovnávací genomiku
05.11.2024	7	Bi5000	Analýza sekvencí proteinů
06.11.2024		Bi5000c	Cvičení – analýza sekvencí proteinů
12.11.2024	8	Bi5000	Strukturní databáze
13.11.2024		Bi5000c	Cvičení – základní strukturní analýzy, vizualizace proteinových struktur v PyMOL
19.11.2024	9	Bi5000	Predikce struktury proteinů
20.11.2024		Bi5000c	Cvičení – predikce struktury proteinů
26.11.2024	10	Bi5000	Příprava anotované sekvence DNA pro zaslání do databáze
27.11.2024		Bi5000c	Cvičení – NGS data a lokální anotace dat
03.12.2024		Bi5000	1. Předtermín
04.12.2024		Bi5000c	Zápočtový test na PC (odpovědník v ISu)
10.12.2024		Bi5000	2. Předtermín, další termíny v lednu
11.12.2024		Bi5000c	Opravný zápočtový test na PC (odpovědník v ISu)

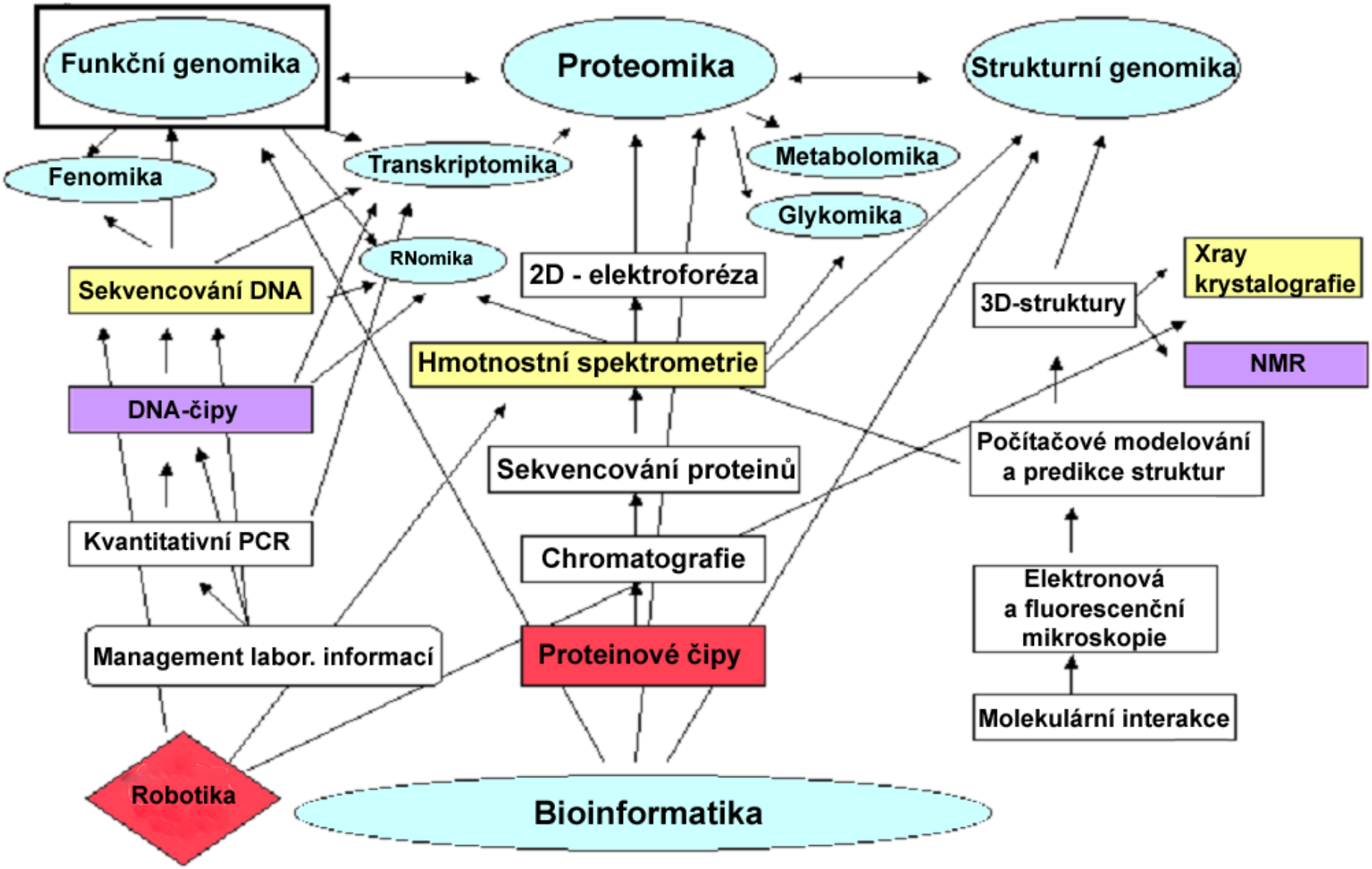
Bioinformatika je disciplína na rozhraní počítačových věd, informačních technologií, matematiky a biologie

- Termín bioinformatika se objevil poprvé v roce 1991
- Představuje spojení technologií z oblastí
  molekulární biologie
  informačních technologií
- Bioinformatika zahrnuje
  studium a analýzu
  praktické uchovávání
  vyhledávání a zobrazování
  modelování biologických dat
- Výpočetní nástroje umožňujících analýzu dat a stanovení jejich vzájemných vztahů
- Dramatický nárůst množství dat a tím současně zvyšující se obtížnost jejich zkoumání a hodnocení ve vztahu k biologickým otázkám

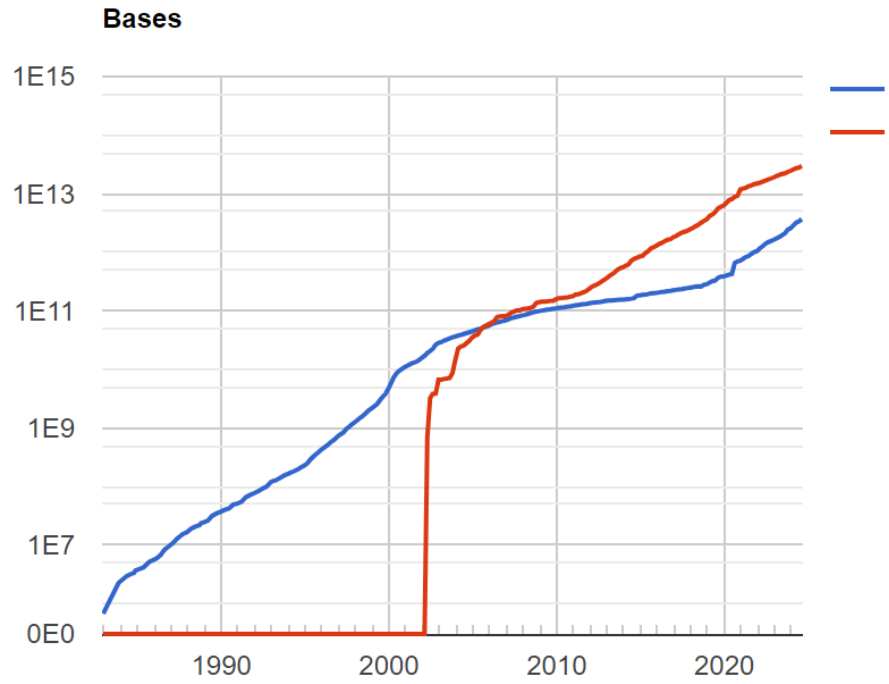
Základní zdroje a aplikace bioinformatiky

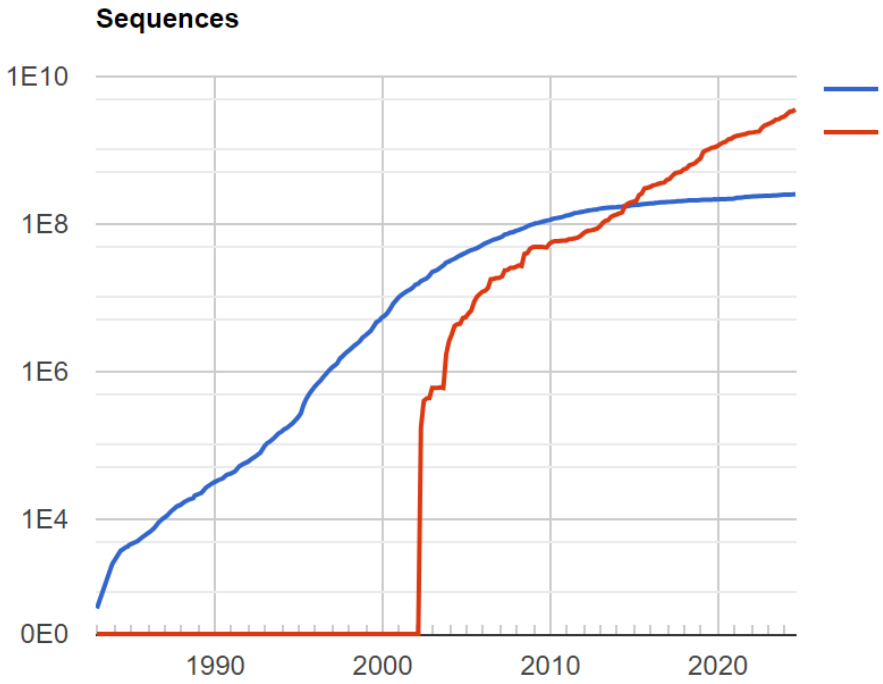
Výpočetní základy	Zdroje dat	Aplikace bioinformatiky
Algoritmy	Obecně dostupné databáze	Získávání dat
Grafika, vizualizace		Nástroje pro přístup k databázím
Zpracování signálu		Mapování a srovnávání genomů
Architektura hardwaru		Sekvenční přiložení, assembly
Informační teorie		Identifikace genů
Správa databází		Funkční identifikace proteinů
Statistika		Molekulární evoluce
Simulace		Molekulární modelování
Umělá inteligence		Predikce struktur
Zpracování obrazu	Zpracování laboratorních dat	Srovnávání struktur
Robotika		Stanovení makromolekulárních struktur
Softwarové inženýrství		Vývoj léčiv na základě struktur

„...omiky“ v molekulární biologii



Trend nárůstu množství dat v bioinformatických databázích





Zdroj: https://www.ncbi.nlm.nih.gov/genbank/statistics/

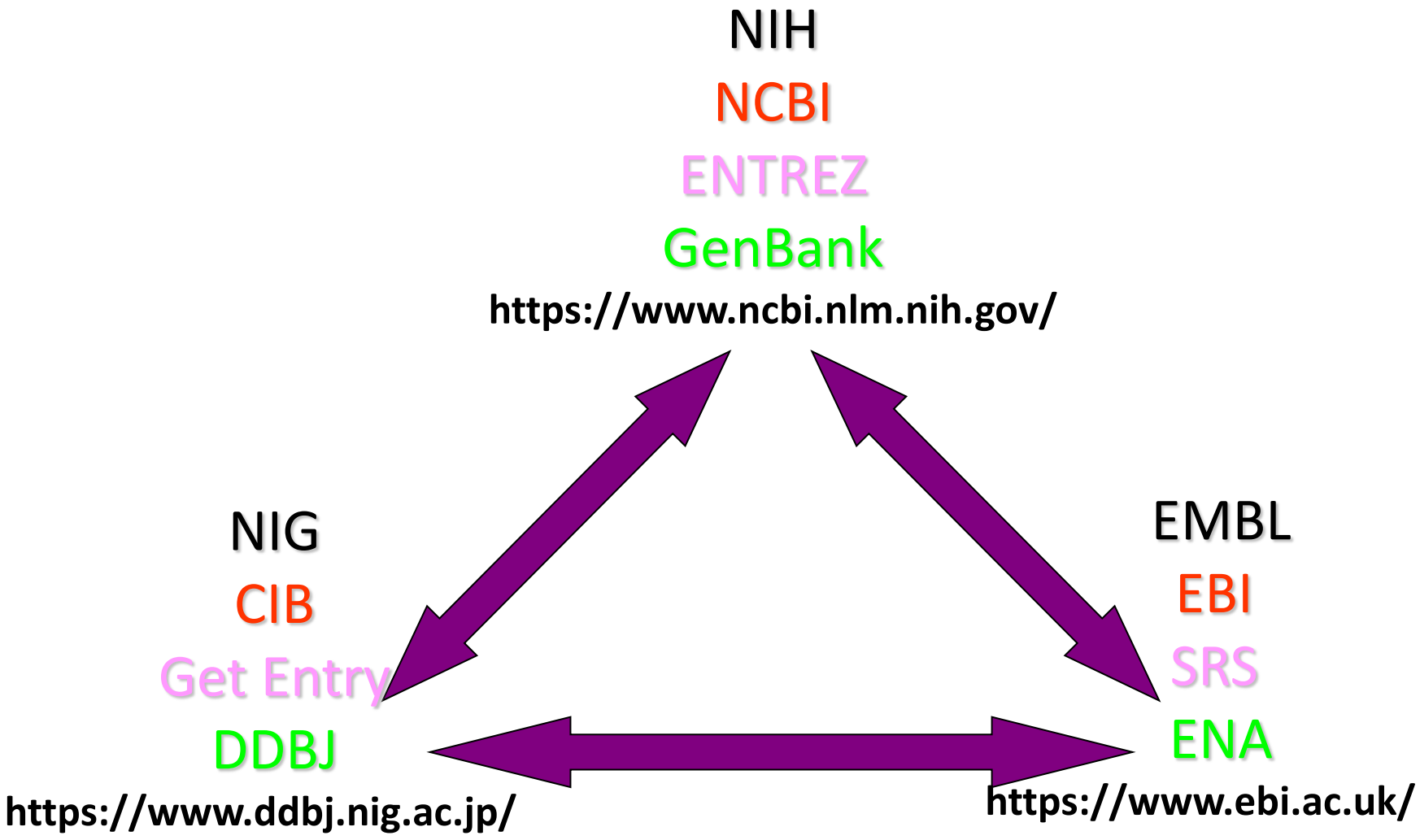
Typy jednoduchých bioinformatických manipulací

1. Přístup k datům a databáze (lekce - úvod)
2. Zápis sekvencí, hledání podobnosti
3. Konverze dat a formátů
4. Sestavení kompletních sekvencí ze sekvenačních dat (lekce – anotace)
5. Výpočetní analýza sekvencí
6. Návrh oligonukleotidů

Nejdůležitější databáze sekvencí nukleových kyselin a proteinů

- V každém ze tří hlavních bioinformatických center je spravována **genomová databáze** sekvencí nukleových kyselin a odpovídajících, z nich přeložených proteinů.
  - **EMBL Nucleotide Sequence Database / European Nucleotide Archive** (v rámci institutu EBI) – 1980
  - **GenBank** (v rámci institutu NCBI) – 1982
  - **DDBJ** (The DNA Data Bank of Japan) - 1984
- Tři samostatné báze vznikly v důsledku potřeby rychlé dostupnosti databáze sekvencí na jednotlivých kontinentech v době, kdy ještě nebyly rozvinuté vysokorychlostní komunikační sítě.

Mezinárodní spolupráce sekvenčních databází (velká trojka)

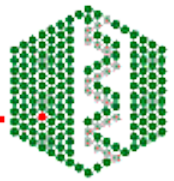


Sdílení dat v základních databázích

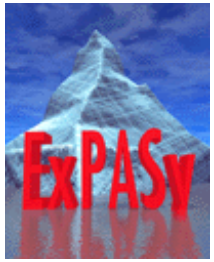
V každém z bioinformatických center jsou dostupné jednoduché nástroje pro manipulaci s daty

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

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



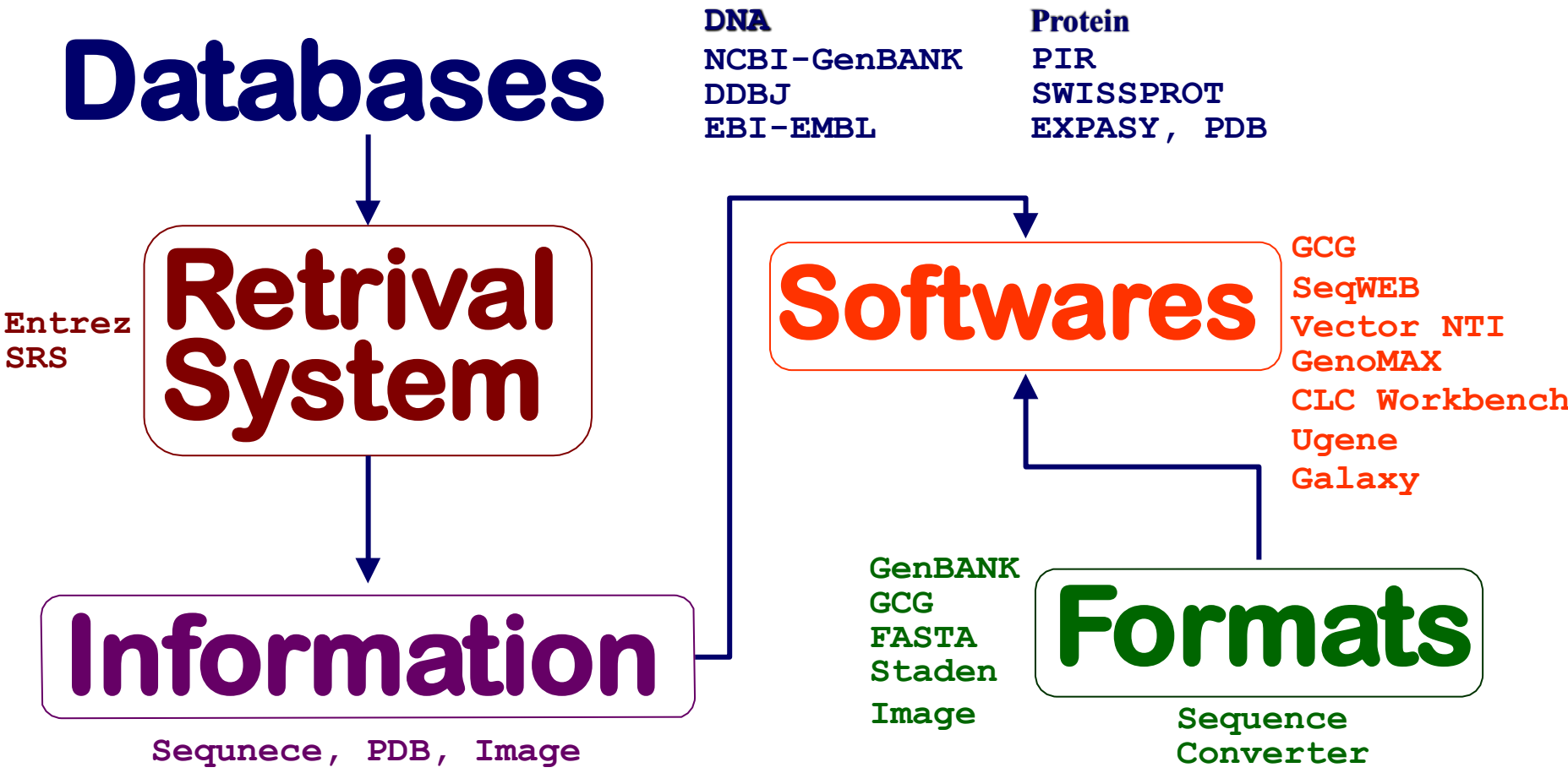
DDBJ: *http://www.ddbj.nig.ac.jp/*
National Institute of Genetics (NIG)



ExPASy:
Expert Protein Analysis System

PDB: *https://www.rcsb.org/*
RCSB Protein Data Bank (RCSB PDB)

Získání dat a manipulace se sekvencemi



2. 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**
    - 5′CAAACGTCGTCTA 3′
  - **protein od N-konce k C-konci**
    - (NH_2-)MKRLSALGPGGLTRR(–COOH)
- **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

Využití zápisu s degenerovanými nukleotidy

TACGGT
TATAAT
TATAAT
GATACT
TATGAT
TATATT

Konsenzní sekvence: **TATAAT**

Degenerovaná sekv.: **KAYRNT**

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

Běžné formáty sekvencí

- Prostý text
- FASTA
- FASTQ
- Genbank
- EMBL
- GCG
- PIR
- ASN1
- Výstupní data sekvenování: ABI, AB1, SCF, SFF, BAM, SAM, FASTF aj.

PLAIN SEQUENCE FORMAT

Obsahuje pouze IUPAC znaky

Obsahuje jedinou sekvenci

Příklad

```
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAA
CCTCCCATCCGTGTCTATTGTAC
```

Použití: pro zápis krátké sekvence např. v textu nebo obrázku

FASTA FORMAT

Může obsahovat více sekvencí

Začíná specifickým záhlavím „>“, za kterým následuje definice

Příklad:

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

>LinB_protein
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGR
LIACDLIGMGDSDKLDPSGPERYAYAEHRDYLDALWEALDLGDRVVLVVHDWGSALG
FDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQAFRSQAGEELVLQDNVFVE
QVLPGLILRPLSEAEMAAYREPFLAAGEARRPTLSWPRQIPIAGTPADVVAIARDYA
GWLSESPIPKLFINAEPGALTTGRMRDFCRTWPNQTEITVAGAHFIQEDSPDEIGAA
IAAFVRRLRPA
```

Použití: univerzální formát pro zápis sekvence vhodný jako vstupní data pro většinu software.

FastQ FORMAT

Záhlaví obsahuje automaticky generovaný identifikátor klastru ze sekvenování.
Následuje primární sekvence
a informaci o kvalitě stanovení sekvence

Příklad:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Použití: výstupní data při sekvenování nové generace

KLÍČ K IDENTIFIKÁTORŮM KVALITY:

Nejnižší kvalita nejvyšší kvalita

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Viz též Phred Quality Score

GENBANK FORMAT

Začíná řádkem s názvem pole LOCUS
Obsahuje mnoho různých deskriptorů
Začátek primární sekvence je vyznačen ORIGIN a sekvence je ukončena „//“

Zpravidla obsahuje i anotované sekvence proteinů a odkazy do dalších databází

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
//
```

EMBL FORMAT

Začíná řádkem s jedinečným identifikátorem (ID), následuje anotace¨.
Obsahuje mnoho různých deskriptorů
Sekvence začíná symboly SQ a sekvence je ukončena „//“

Zpravidla obsahuje i anotované sekvence proteinů a odkazy do dalších databází

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
//
```

Formáty sekvencí obsahující mnohonásobná přiložení

- Sekvenční přiložení umožňuje srovnat podobné sekvence
- K dispozici celá řada formátů, obvykle přizpůsobených používaným programům
  - Multi FASTA
  - Phylip
  - PAUP / NEXUS
  - Clustal
  - MSF

CLUSTAL/MUSCLE MULTIPLE FORMAT

Začíná řádkem s definicí
Vkládá mezery do sekvence tak, aby při mnohonásobném přiložení byly identické zbytky nad sebou

Konzervované pozice se stejným typem zbytku označuje na posledním řádku hvězdičkou
Podobné typy zbytků mohou některé formáty znázorňovat dvoutečkou nebo tečkou

Příklad:

```
Moorella_thermoacetica_ATCC_39073_-_rna.40                ----------GTTTGATCCTGGCTCAGGACAAACGCTGGCGGCGTGCCTAACACATGCAA 50
Ammonifex_degensii_KC4_-_rna.5                            -------AGGGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAACACATGCAA 53
Ammonifex_degensii_KC4_-_rna.31                           -------AGGGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAACACATGCAA 53
Candidatus_Desulforudis_audaxviator_MP104C_-_DAUD_RS00700 TTTATGGAGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAACACATGCAA 60
Candidatus_Desulforudis_audaxviator_MP104C_-_DAUD_RS06920 TTTATGGAGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAACACATGCAA 60
                                                          **********:**********:**************************

Moorella_thermoacetica_ATCC_39073_-_rna.40                GTCGAGCGGTCTTTAATTGGGGAAATCTTCGGATGGAACCGATTAAAGATAGCGGCGGAC 110
Ammonifex_degensii_KC4_-_rna.5                            GTCGAGCGGGCTT------GTCAGGGCCTTGTGT----CCTGGCAAGTTGAGCGGCGGAC 103
Ammonifex_degensii_KC4_-_rna.31                           GTCGAGCGGGCTT------GTCAGGGCCTTGTGT----CCTGGCAAGTTGAGCGGCGGAC 103
Candidatus_Desulforudis_audaxviator_MP104C_-_DAUD_RS00700 GTCGTGCGA---TTGAGAGGTGAGCATCTCACTT----CTCAA------GAGCGGCGGAC 107
Candidatus_Desulforudis_audaxviator_MP104C_-_DAUD_RS06920 GTCGTGCGA---TTGAGAGGTGAGCATCTCACTT----CTCAA------GAGCGGCGGAC 107
                                                          **** ***: ****:*  **  *:  ::*::  *    *: : :**:   **********
```

Poznámka k používaným fontům

- Proporcionální fonty
  - Arial, Times
  - Každý znak - jiná šířka
  - Nevhodné pro zápis sekvence

gaatttttt

cttaaaaaaa

- Neproporcionální fonty
  - Vhodné pro zápis sekvence
  - Všechny znaky stejná šíka
  - Courier, Monospaced

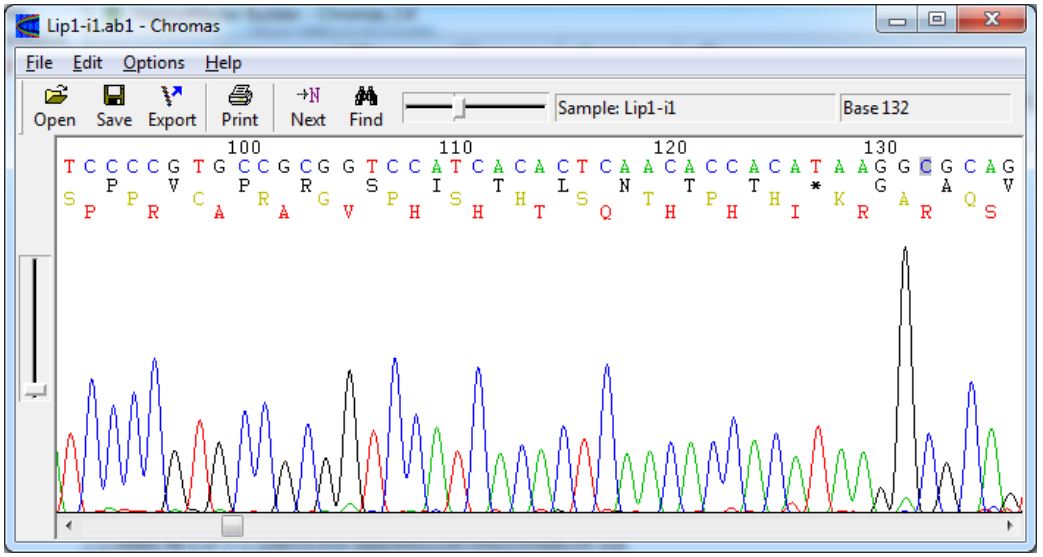
gaatttttttt

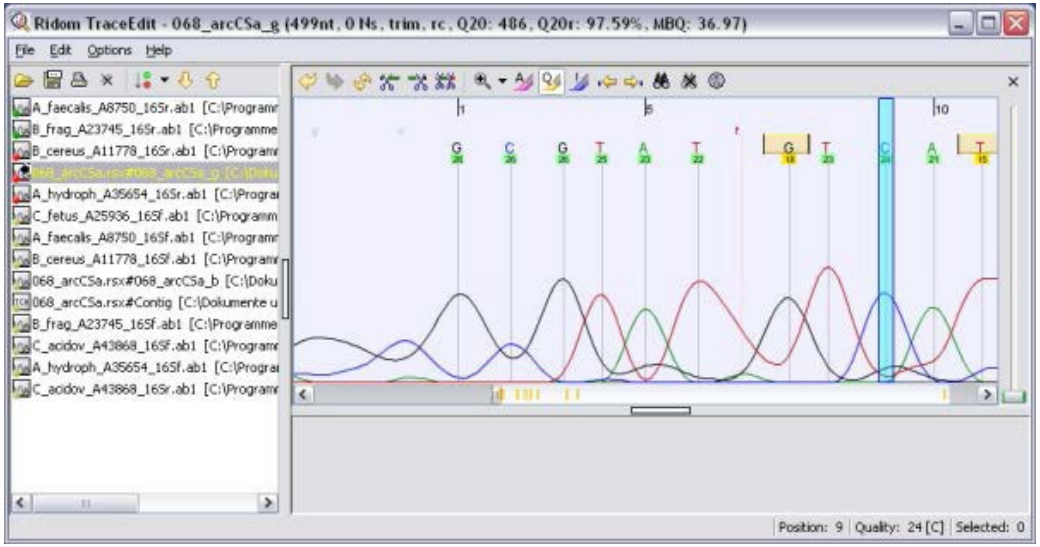
cttaaaaaaaa

- 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, např.
  - Chromas Lite
  - ABIView
  - Ridom Trace Edit
- Export
  - FASTA
  - Prostý text
- Formáty z NGS vyžadují složitější zpracování





Jednoduché formáty sekvencí mají omezení a neobsahují

- Data o expresi genů
- Variace a polymorfismy
- Specifické informace o zdroji sekvence (organismu, klonech, …)
- Odkazy na další informace
- Informace o kvalitě

3. Konverze dat a hledání motivů

- **Převod informace mezi řetězci**
  - Reverse-complement
- **Hledání motivů**
  - Přesné
  - Podobné
- **Přepis a překlad podle ústředního dogmatu**
  - Transkripce
  - Translace – genetický kód
- **Sekvenční přiložení**
  - Párové, stanovení identity a podobnosti
  - Mnohonásobné, identifikace konzervativních motivů
- **Spojování, rozdělování**
  - Restrikční štěpění
  - Klonování *in silico,* konstrukce vektorů a rekombinantní DNA pro přípravu proteinů
- **Assembly – kompletace a sestavení genomů**

Příklady nástrojů pro konverzi formátů

- UNIX-GCG
  - To Genbank, To Fasta….
  - From Genbank, From Fasta…
- SEQRET
  - https://www.ebi.ac.uk/Tools/sfc/
- SMS – The Sequence Manipulation Suite v2
  - http://www.bioinformatics.org/sms2/
    - EMBL to FASTA
    - GenBank to FASTA
    - Reverse Complement – převod mezi řetězci
    - Filtrování znaků
- Vzájemná konverze - Sanger, IonTorrent, Illumina, (ONT)
  - Biopython
  - EMBOSS
  - BioPerl
  - Samtools

Převod informace mezi řetězci

Nástroj **Reverse Complement**

http://www.bioinformatics.org/sms2/rev_comp.html

Převod mezi dvěma řetězci (pozitivní-kódující-horní/negativní-antikódující-spodní)

5′ CCCCATGTTT 3′

3′ GGGGTACAAA 5′

>Sample sequence

5′ **CCCCATGTTT** 3′

>Reverse complement dle pravidel o párování bazí

5′ **AAACATGGGG** 3′

>Reverse – nemá biologický význam

5′ **TTTGTACCCC** 3′

>Complement – nemá biologický význam

5′ **GGGGTACAAA** 3′

Hledání motivů v sekvencích

Hledání slov = uspořádaná množina znaků

GAATTC

GARYTC

GAAN$_{(1-50)}$TTC

- **Přesně definovaný motiv**
- **Degenerované symboly**
- **Povolený počet neshod**
- **Motivy od sebe vzdálené x zbytků**

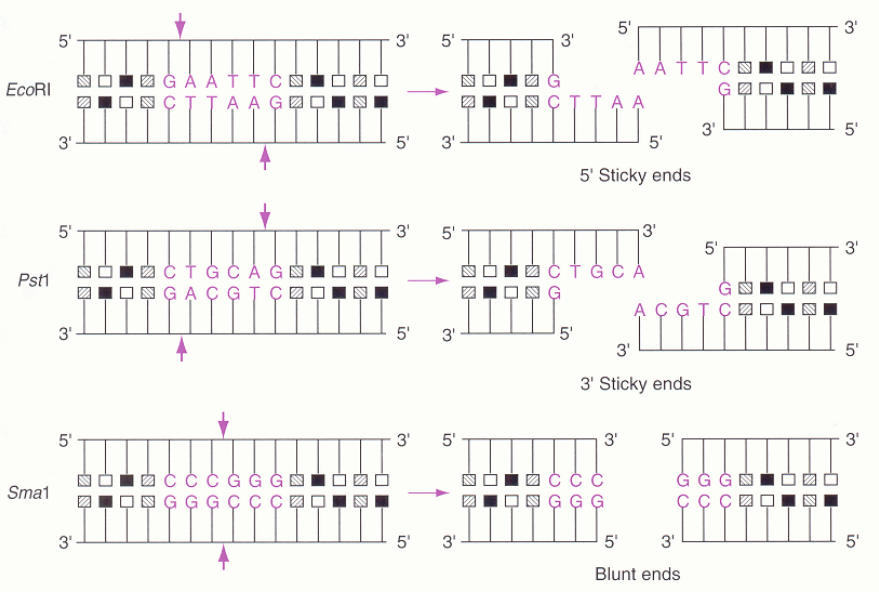
Standardní příklady hledání

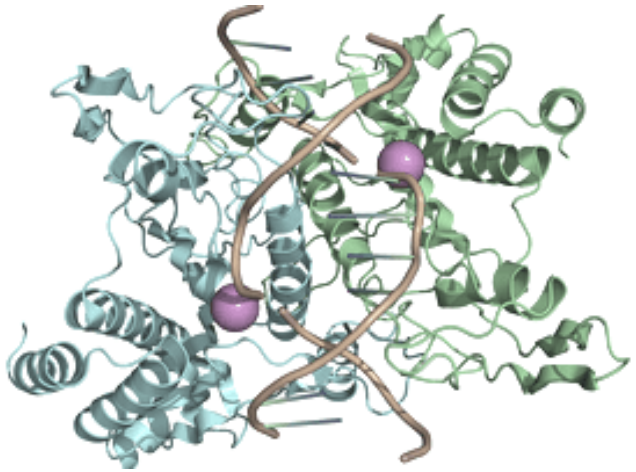
- Restrikční místa
- Repetice
  - Přímé
  - Obrácené (vlásenky se smyčkou)
- Konsenzní vzory
- Uživatelem definované vzory
- Otevřené čtecí rámce (START …. STOP)

- Základ pro hledání genů a funkčních oblastí

Restrikční endonukleázy třídy II mají praktické využití

- Vážou se na specifické (4-6 pb) sekvence nukleotidů
- Katalyzují štěpení dvou řetězců molekuly DNA uvnitř vazebného místa nebo v jeho bezprostředním sousedství
- Produkty štěpení RE
  - tupé konce (po štěpení obou řetězců ve stejném místě)
  - přečnívající konce (po štěpení řetězců v různých místech, která jsou obvykle vzdálena 1-4 nukleotidy)
    - - 5´přečnívající
    - - 3´přečnívající





Restrikční analýza *in silico*

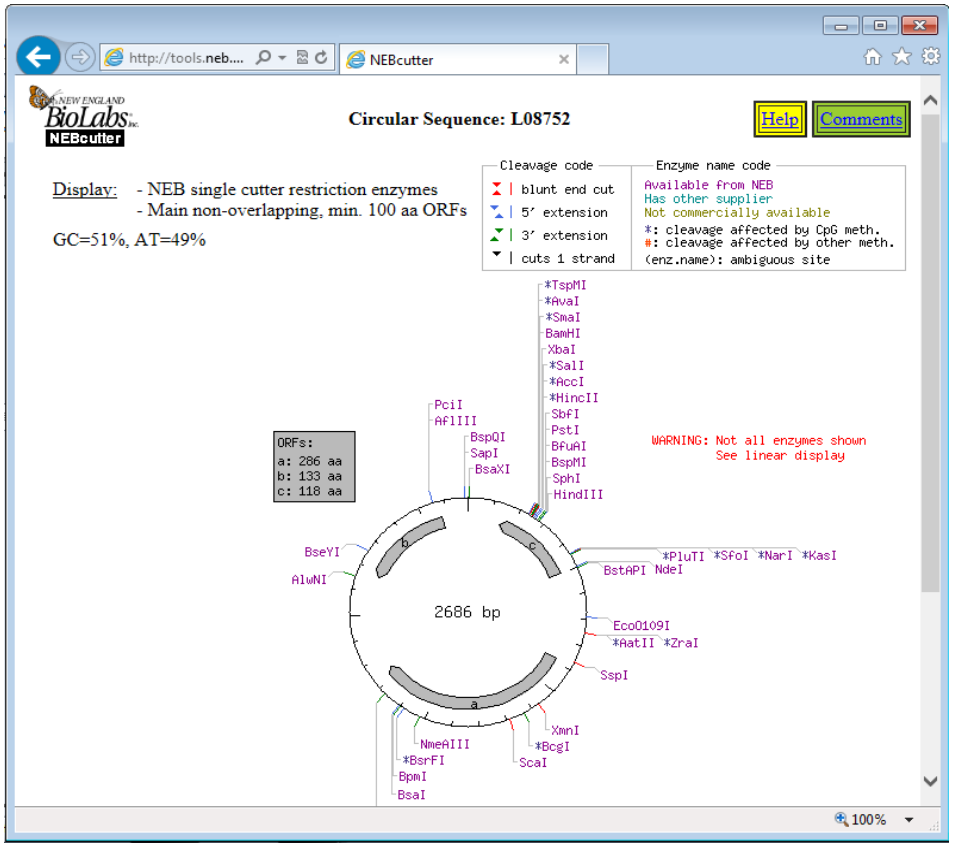
- Restrikční endonukleázy třídy II
  - Sekvenčně specifické endonukleázy, které štěpí DNA v rozpoznávaných sekvencích
  - Přehled dostupný v databázi REBASE- Restriction Enzyme Database http://rebase.neb.com/rebase/rebase.html
    - Sekvence rozpoznávacích míst
    - Producent enzymu
    - Reference
    - 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

- Provádí hledání restrikřních míst na základě analýzy sekvence DNA
- Konstrukce restrikčních map
  - 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
  - NEB Cutter v3.0 (https://nc3.neb.com/NEBcutter/)
  - EMBOSS Restrict (https://www.bioinformatics.nl/cgi-bin/emboss/restrict)
  - pDRAW32 freeware (http://www.acaclone.com/)

NEB Cutter

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



- 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

Vyhledání otevřených čtecích rámců

- **ORF (Open Reading Frame)**
  Sada překládaných kodonů mezi iniciačním a terminačním kodonem
- Výsledek je závislý na použitém genetickém kódu
  - Databáze genetických kódů v NCBI
  - https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi
- U prokaryot, které nemají introny je základem hledání genů
- U eukaryot zpravidla využíváme analýzu ORF u komplementární DNA (cDNA) vzniklé reverzní transkripcí z mRNA

ORF Finder (Open Reading Frame Finder)

https://www.ncbi.nlm.nih.gov/orffinder/

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

```
>bacteriophage 3A
TCGCTTTAAACCCTTCATGCCTTCTTGGACACCTAAATGGTCTAATTTCAGCTCCAA@GGTCATGCCTTCTACTT
TTCATATTAACCTCCTTTCTAGCTTCCAAAAAGTT@TTCTTAAATCCGTACCTGTAATGACTTTTTGTTCACTTT
TCTTCAGTCTCTT@CTTTATTCTCTTCATTAAGTATTTCTAAAAGTTTTACATACGGCTGTTTTCTGACTT@CAG
GTCCACCCATACTGCTCCATACAGAAACGTTGTATTTTCTTAATGTTCG@ATAAAATATCTTTTATTGAGATTGT
TGTCTTTCCCATCTCTTCTGGTTCAGT@TTCTGAATCTTCTTCATCTTCACCATTGATTTCTCGAAATATATCTT
GCTTT@TGTATAAGTTTTAGTGCTCATCTTGTTCAAAACATCTTCTTCAGTCAATCCTTCATC@TTTAAATAAAT
AATAACTGTCGCTCTTTTTGTCTCATTTTTGTTGCGTTAGG@TACTTCTTTTTTATTCTCTTGATTTACTAATTC
```

From: To:

Choose Search Parameters

Minimal ORF length (nt): 75

Genetic code: 11. Bacterial, Archaeal and Plant Plastid

ORF start codon to use:

- (•) "ATG" only
- ( ) "ATG" and alternative initiation codons
- ( ) Any sense codon

Ignore nested ORFs: ☐

Start Search / Clear

Submit Clear

ORF Finder (Open Reading Frame Finder)

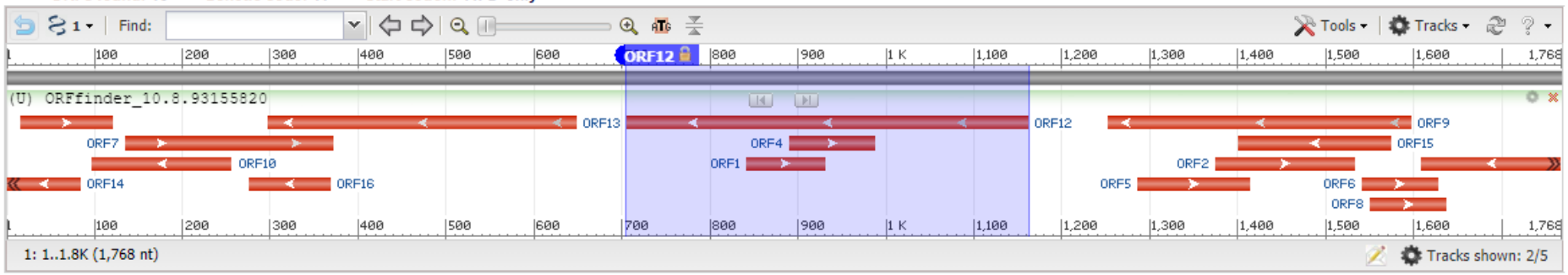
https://www.ncbi.nlm.nih.gov/orffinder/

Open Reading Frame Viewer

Help

Sequence

ORFs found: 16 Genetic code: 11 Start codon: 'ATG' only



Six-frame translation...

ORF12 (151 aa) Display ORF as... Mark

```
>lcl|ORF12
MTKTLKVYKGDDVVASEQGEGKVSVTLSNLEADTTYPKGTYQVAWEENGK
ESSKVDVPQFKTNPILVSGVSFTPETKSITVNADDNVEPNIAPSTATNKT
LKYTSEHPEFVTVDERTGAIHGVAEGTSVITATSTDGSDKSGQITVTVTN
G
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF12	-	**2**	**1161**	**706**	**456 | 151**
ORF13	-	2	648	298	351 | 116
ORF9	-	1	1597	1253	345 | 114
ORF7	+	3	135	371	237 | 78
ORF15	-	3	1574	1401	174 | 57
ORF2	+	1	1375	1533	159 | 52
ORF10	-	1	256	98	159 | 52
ORF11	-	2	<1767	1609	159 | 52
ORF5	+	2	1286	1414	129 | 42
ORF3	+	2	17	121	105 | 34

ORF12

SmartBLAST

BLAST

Marked set (0)

SmartBLAST best hit titles...

BLAST

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

Translace *in silico*

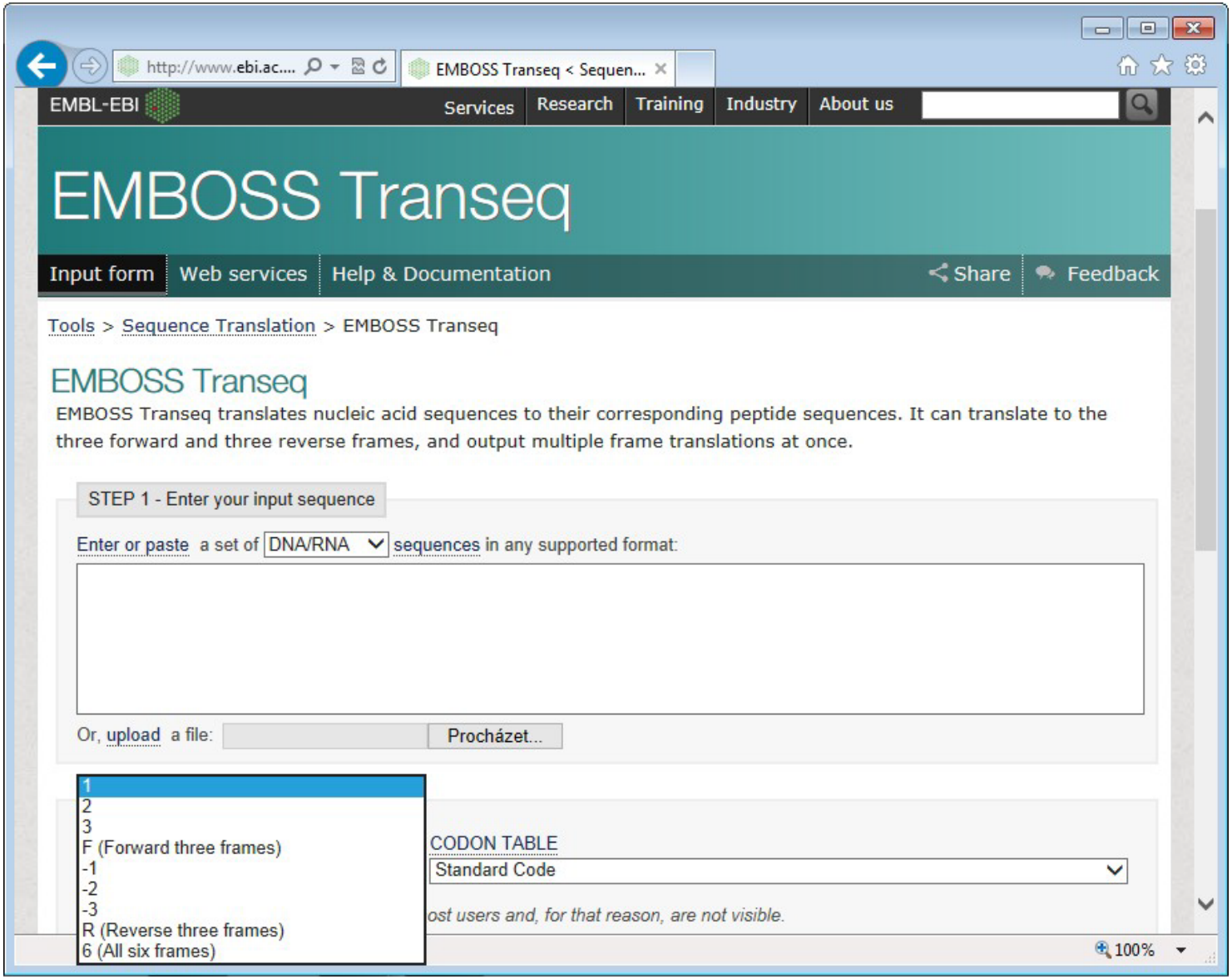
Překlad genetické informace z DNA do proteinu

- 6 možných čtecích rámců ve dvouřetězcové DNA
- Vymezené oblasti - exony
- Výběr genetického kódu
- Nástroje: EMBOSS Transeq, EMBOSS Sixpack, aj.

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

EMBOSS Transeq

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



Příklady translace *in silico*

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
```

Princip translace 6 čtecích rámců

```
STANDARDNÍ GENETICKÝ KÓD
Aminokyseliny  = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
         Start = ----------------------------------M---------------------------
         Báze1 = UUUUUUUUUUUUUUUUCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
         Báze2 = UUUUCCCCAAAAGGGGUUUUCCCCAAAAGGGGUUUUCCCCAAAAGGGGUUUUCCCCAAAAGGGG
         Báze3 = UCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAG
```

Rámec +1

DNA:	ATG	GAT	GTA	AAA	GAA	ATT	AAT	TAA
mRNA:	AUG	GAU	GUA	AAA	GAA	AUU	AAU	UAA
Protein:	M	D	V	K	E	I	N	*

Rámec +2

DNA:	A	TGG	ATG	TAA	AAG	AAA	TTA	ATT	AA
mRNA:	A	UGG	AUG	UAA	AAG	AAA	UUA	AUU	AA
Protein:		W	M	*	K	K	L	I	

Rámec +3

DNA:	AT	GGA	TGT	AAA	AGA	AAT	TAA	TTA	A
mRNA:	AU	GGA	UGU	AAA	AGA	AAU	UAA	UUA	A
Protein:		G	C	K	R	N	*	L	

Princip translace 6 čtecích rámců

```
STANDARDNÍ GENETICKÝ KÓD
Aminokyseliny  = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
         Start = ----------------------------------M---------------------------
         Báze1 = UUUUUUUUUUUUUUUUCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
         Báze2 = UUUUCCCCAAAAGGGGUUUUCCCCAAAAGGGGUUUUCCCCAAAAGGGGUUUUCCCCAAAAGGGG
         Báze3 = UCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAGUCAG
```

Rámec -1 (4)

DNA:	ATG	GAT	GTA	AAA	GAA	ATT	AAT	TAA
mRNA:	UUA	AUU	AAU	UUC	UUU	UAC	AUC	CAU
Protein:	L	I	N	F	F	Y	I	H

Rámec -2 (5)

DNA:	A	TGG	ATG	TAA	AAG	AAA	TTA	ATT	AA
mRNA:	UU	AAU	UAA	UUU	CUU	UUA	CAU	CCA	U
Protein:		N	*	F	L	L	H	P	

Rámec -3 (6)

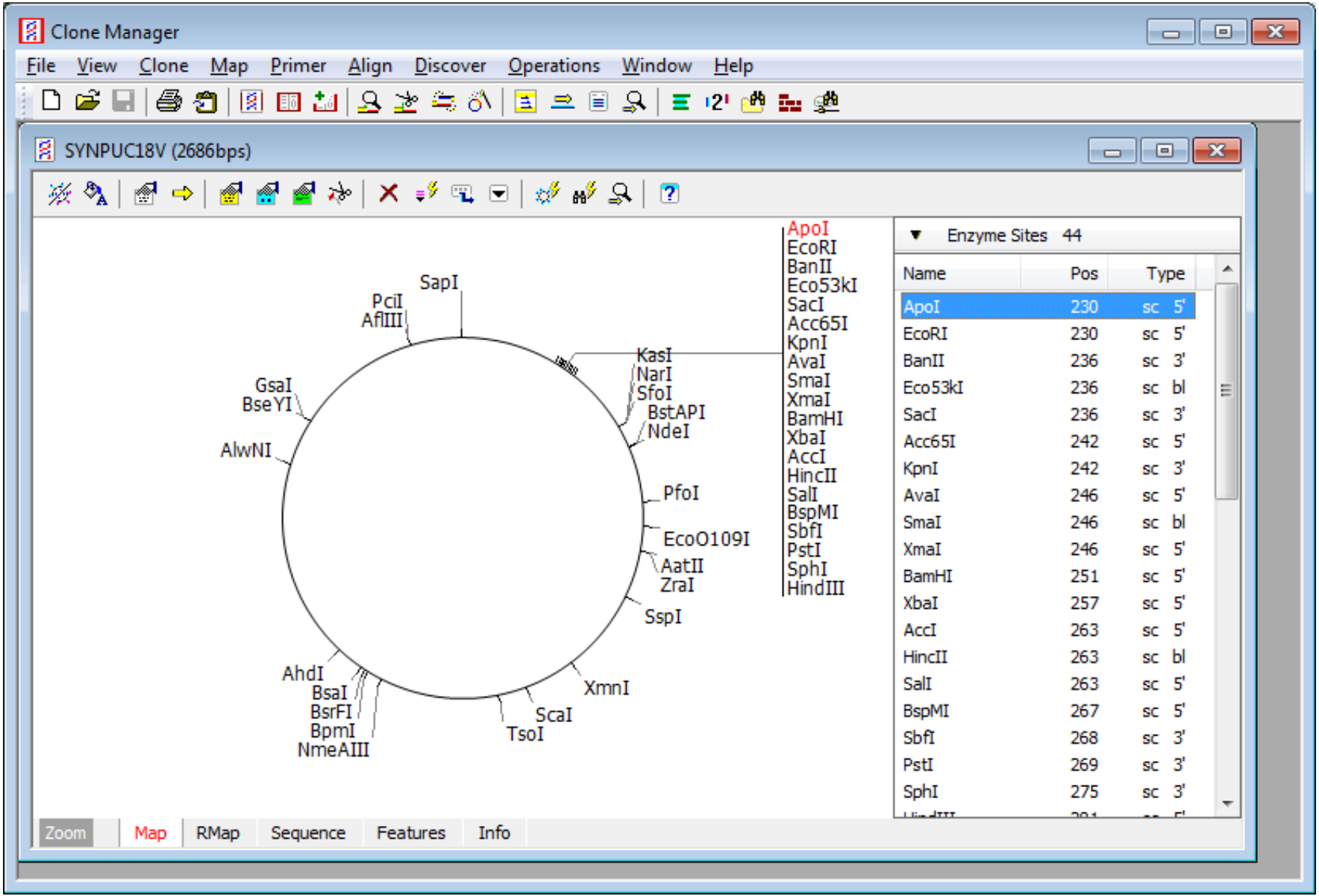
DNA:	AT	GGA	TGT	AAA	AGA	AAT	TAA	TTA	A
mRNA:	U	UAA	UUA	AUU	UCU	UUU	ACA	UCC	AU
Protein:		*	L	I	S	F	T	S	

Klonování *in silico*, konstrukce vektorů

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

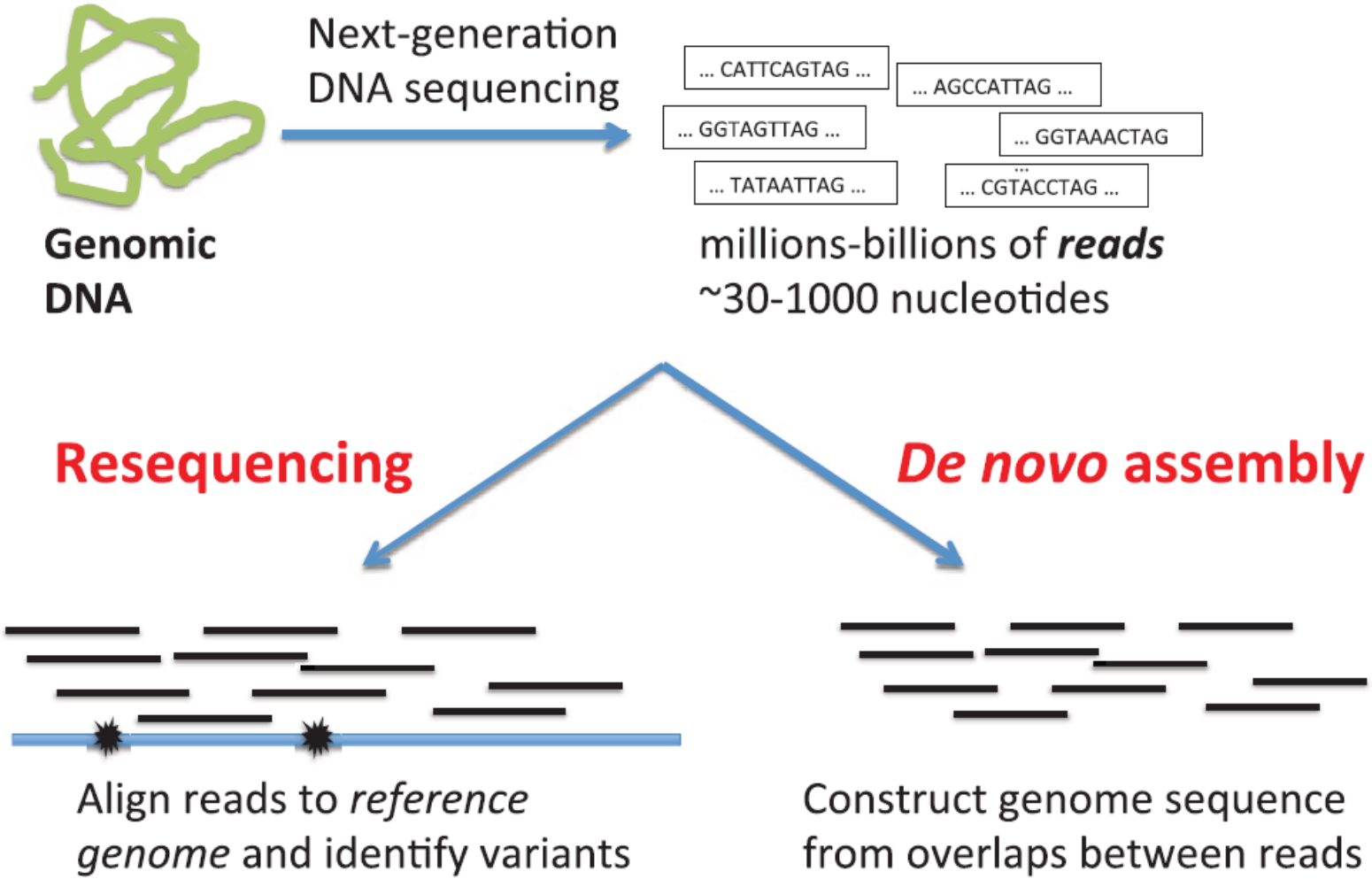
Clone Manager (Sci-Ed Software)

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

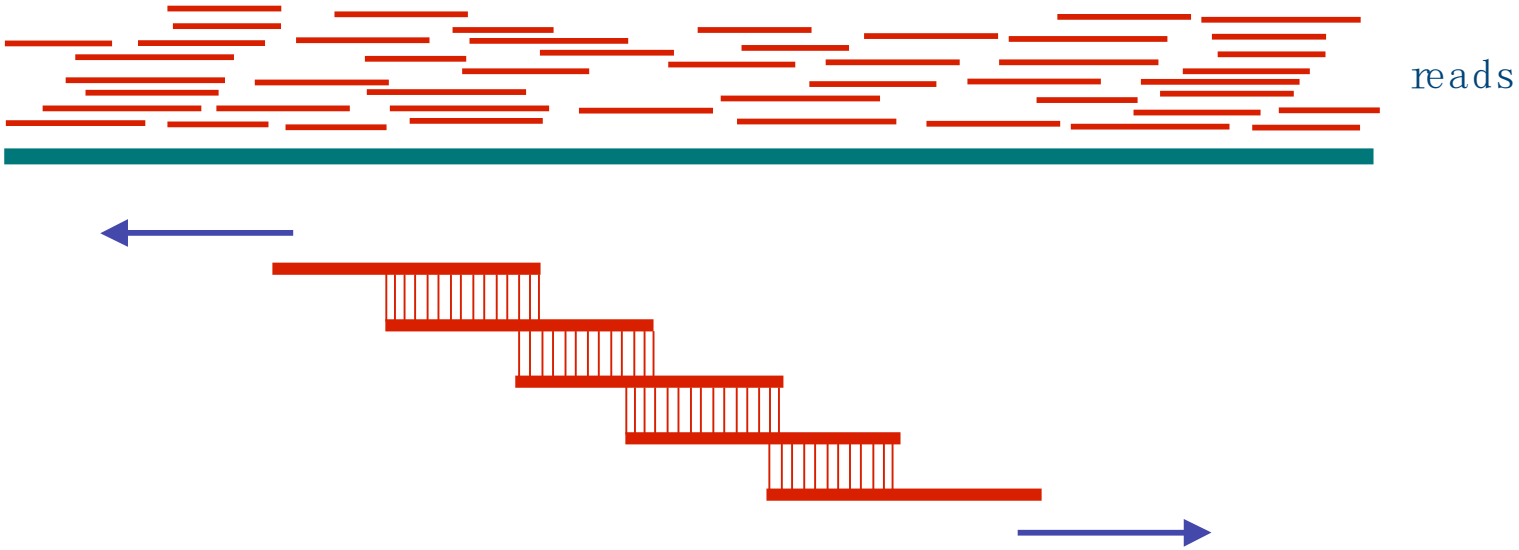


Assembly/ kompletace a sestavení

Resekvenování vs. *de novo* sekvenování



Princip assembly



Pokrytí oblastí >x-násobnou redundancí

Identifikace překryvů, sekvenční přiložení

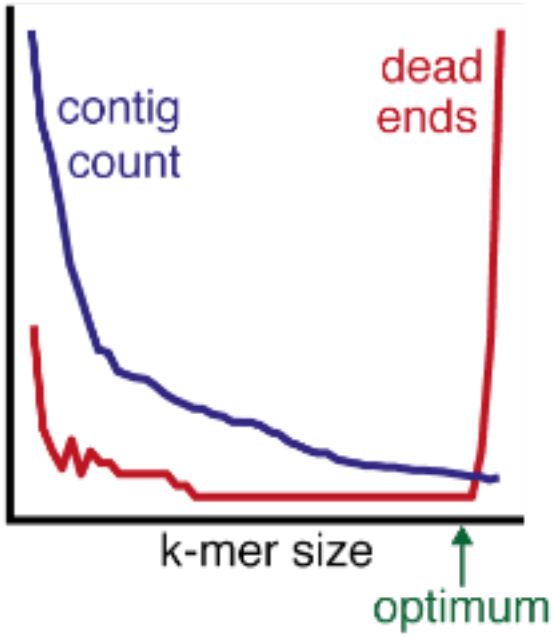
a rekonstrukce sekvence

De novo assembly

- Nezávislé na referenčním genomu
- Parametry
  - Délka čtení
  - Pokrytí genomu (coverage)
- Velké množství dostupných algoritmů
  - Znakové metody
  - Grafové metody
- Výpočetně náročné
- Zpravidla vyžaduje optimalizaci pro každou platformu sekvenování

Princip hledání překryvů

- Vytvoření všech *k*-merů ve čteních, (např. *k* ~ 24)
- Roztřídění čtení do skupin, které sdílejí *k*-mer
- Přiložení párů, které sdílejí *k*-mer
- Mapování a rozšíření sekvenčních přiložení



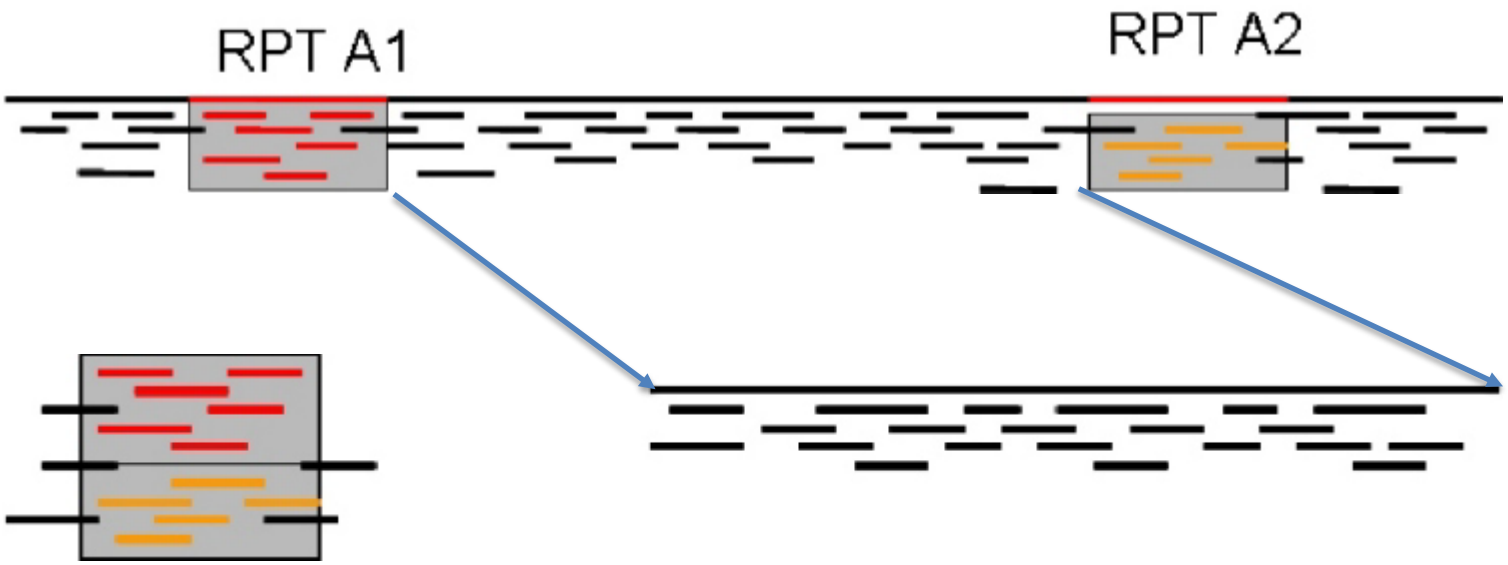
Mapování jednotlivých čtení k referenci

Vytvoření sekvenčního přiložení z jednotlivých čtení

- Hloubka pročtení sekvence odráží kvalitu
- Umožňuje vyřešit neshody
  - Chybně stanovené báze
  - Homopolymerní oblasti
- Umožňuje kvantifikovat polymorfizmy

Repetice jsou příčinou rozdělení genomů do kontigů

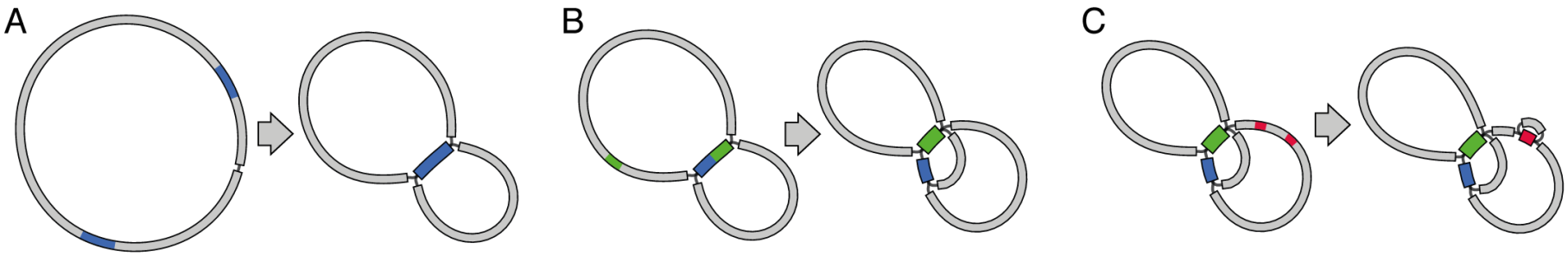
Jestliže čtení je kratší než repetice→ nemožnost sestavení sekvence



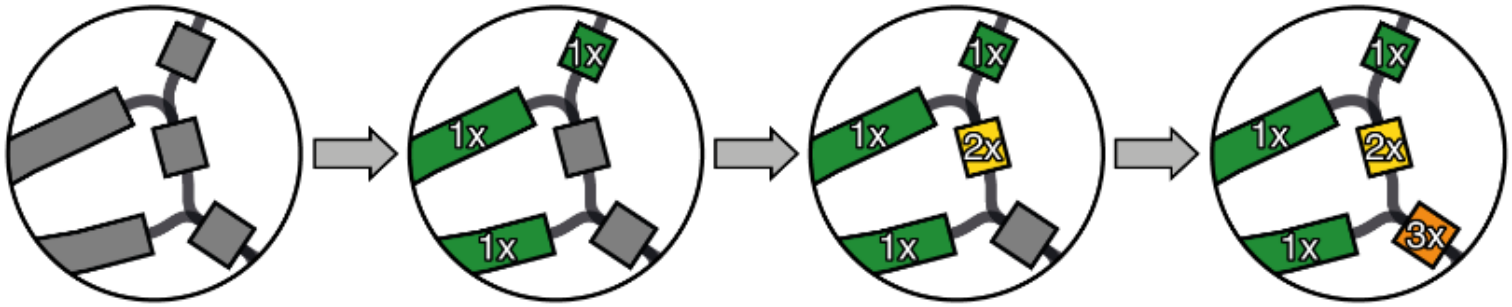
Čtení z mnoha podobných repetic vedou k vytvoření kontigů s pozměněnou strukturou

Kontig tvořený jedinečnou sekvencí, ohraničený repetitivními sekvencemi

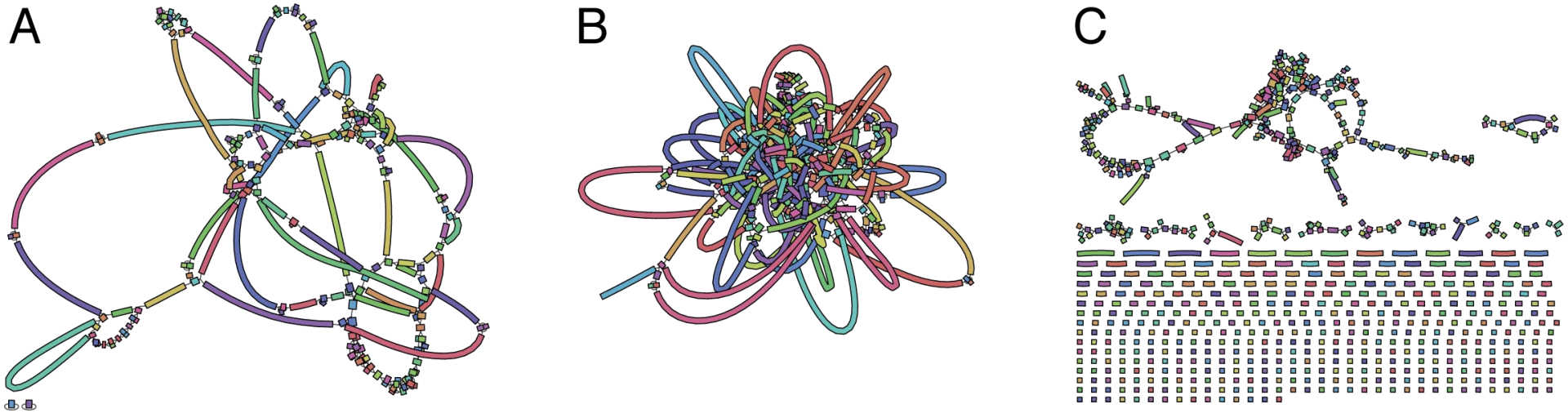
Znázornění repetic v de Bruijnově grafu



- Krátká čtení, hlavní příčina omezení kompletního sestavení
- Stejná sekvence se vyskytuje v genomu vícekrát
- Délka čtení není schopna překlenout tuto repetici
- Pokrytí může indikovat multiplicitu



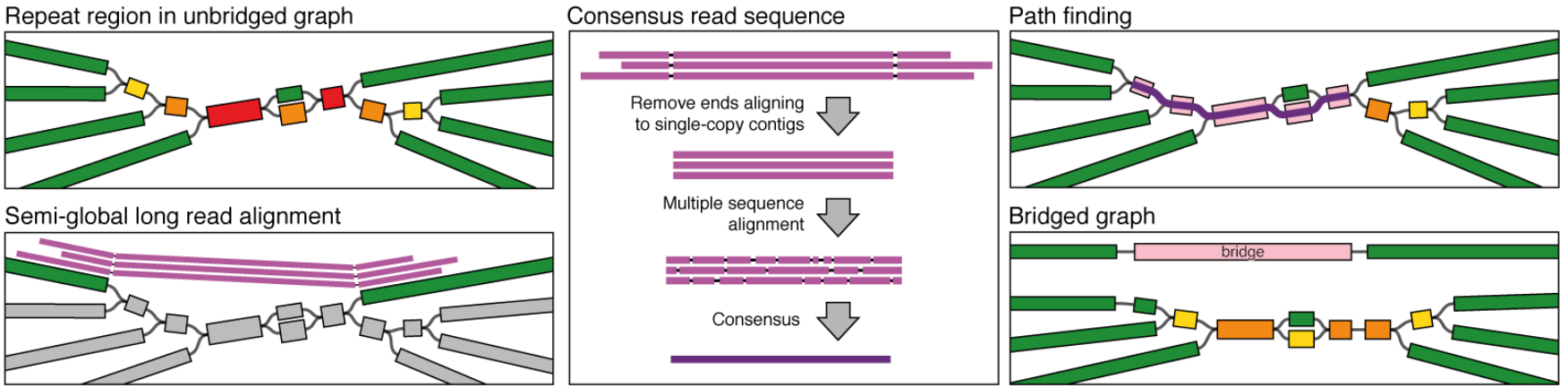
Příklad de Bruijnova grafu u mikrobiálního genomu (Illumina)

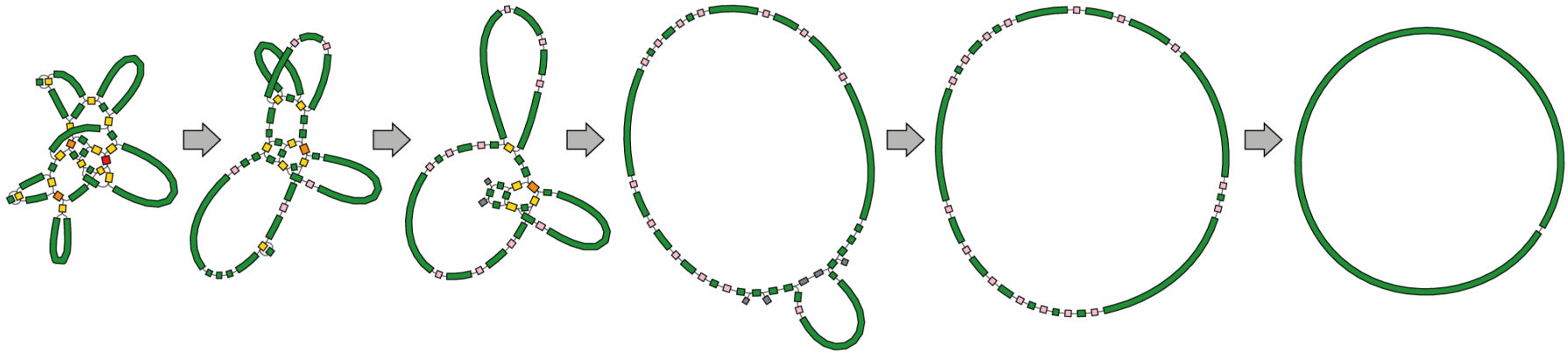


A – kvalitní sestavení
B – sestavení vyžadující optimalizaci, kombinace dlouhých a krátkých kontigů
C – nekvalitní sestavení vycházející z nekvalitních dat, velké množství nezařazených krátkých kontigů

Hybridní assembly a bridging

- Kombinace krátkých čtení (Illumina, IonTorrent) a dlouhých čtení (PacBio, Nanopore) umožňuje hybridní assembly
- Dlouhá čtení: hledání cesty mezi repeticemi





4. Výpočetní analýza sekvencí

- Počet residuí
- Frekvence residuí
- Frekvence oligonukleotidů
- Analýza využití kodonů
- Design oligonukleotidů a primerů

Analýza využití kodonů (codon usage)

- Využití synonymních kodonů
  - není náhodné
  - je rozdílné u různých genomů, které mají určité preferované kodony pro určité aminokyseliny
  - může být problémem při expresi rekombinantních proteinů
- 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

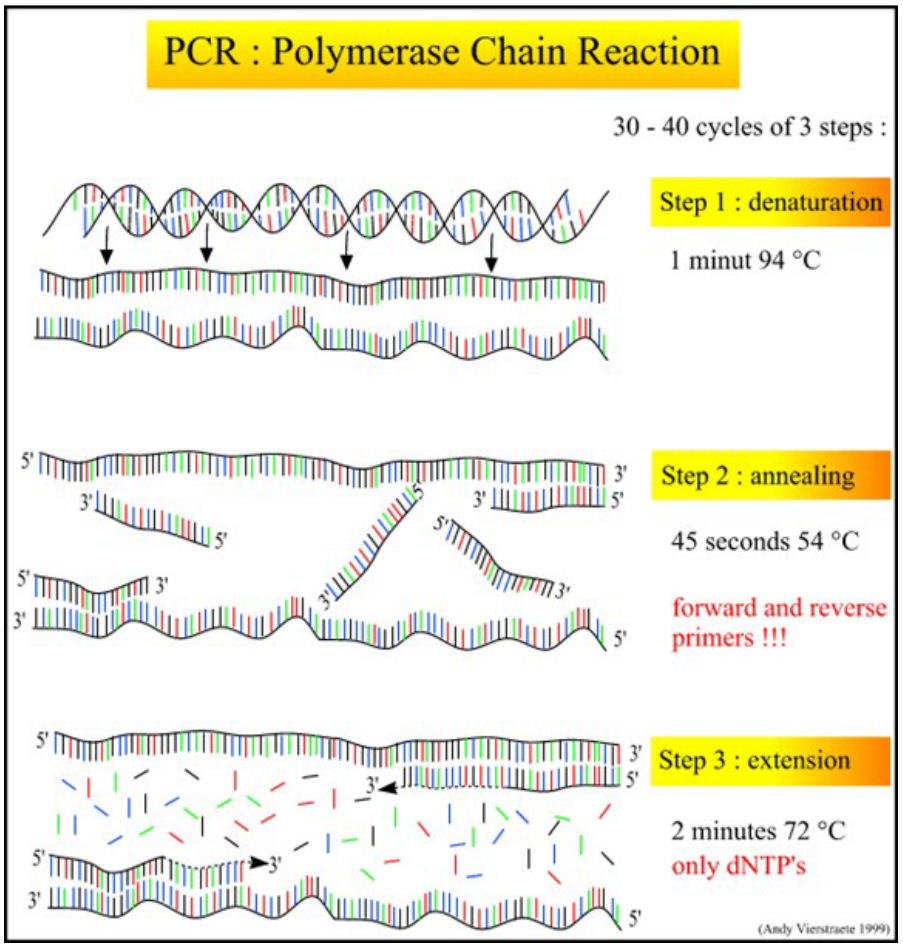
Analýza využití kodonů (codon usage)

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

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

Výběr vhodné strategie před návrhem primerů

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

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 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′

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

- Stabilita duplexů oligonukleotidů a vlásenek se uvádí v ΔG (kcal/mol)
- Oligonukleotidy nejsou komplementární navzájem na 3'-koncích, takže nevytvářejí navzájem nebo samy se sebou duplexy

- Netvoří vnitřní sekundární struktury
  - Chybně navržená dvojice primerů, která vytváří stabilní duplex na 3'-konci:

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

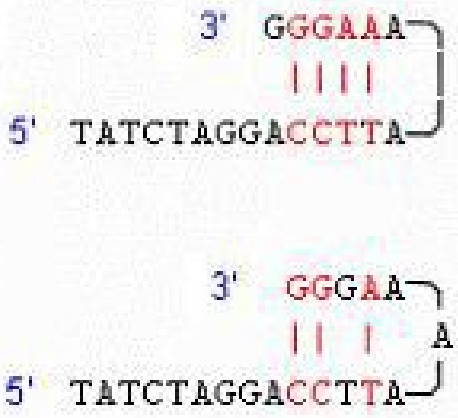
  - Správně navržená dvojice primerů, která vytváří pouze málo stabilní duplex na 5'-konci; na 3'-konci je G nebo C zaručující stabilní párování s templátem:

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

  - Chybně navržený primer, vytvářející vlásenku:

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

Hairpin



Self-Dimer

8 bp

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

4 bp

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

Dimer

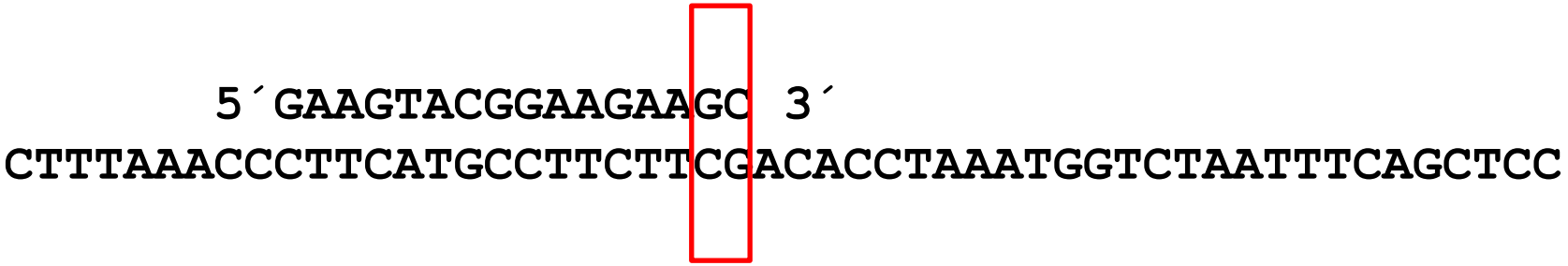
forward primer

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

reverse primer

GC svorky a 3´- koncová stabilita

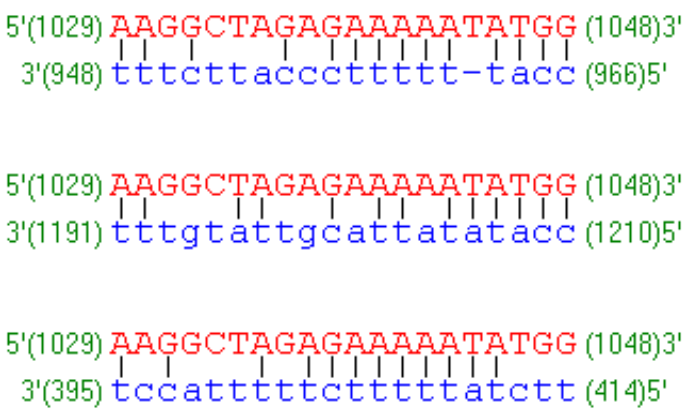
- GC svorka
  - Přítomnost G nebo C mezi posledním 4 bázemi na 3’-konci primeru
  - Zásadní pro zvýšení prevence falešného prodlužování a zvýšení specifičnosti primeru
  - >3 G nebo C v blízkosti 3’-konce jsou však nežádoucí



Jedinečnost primerů

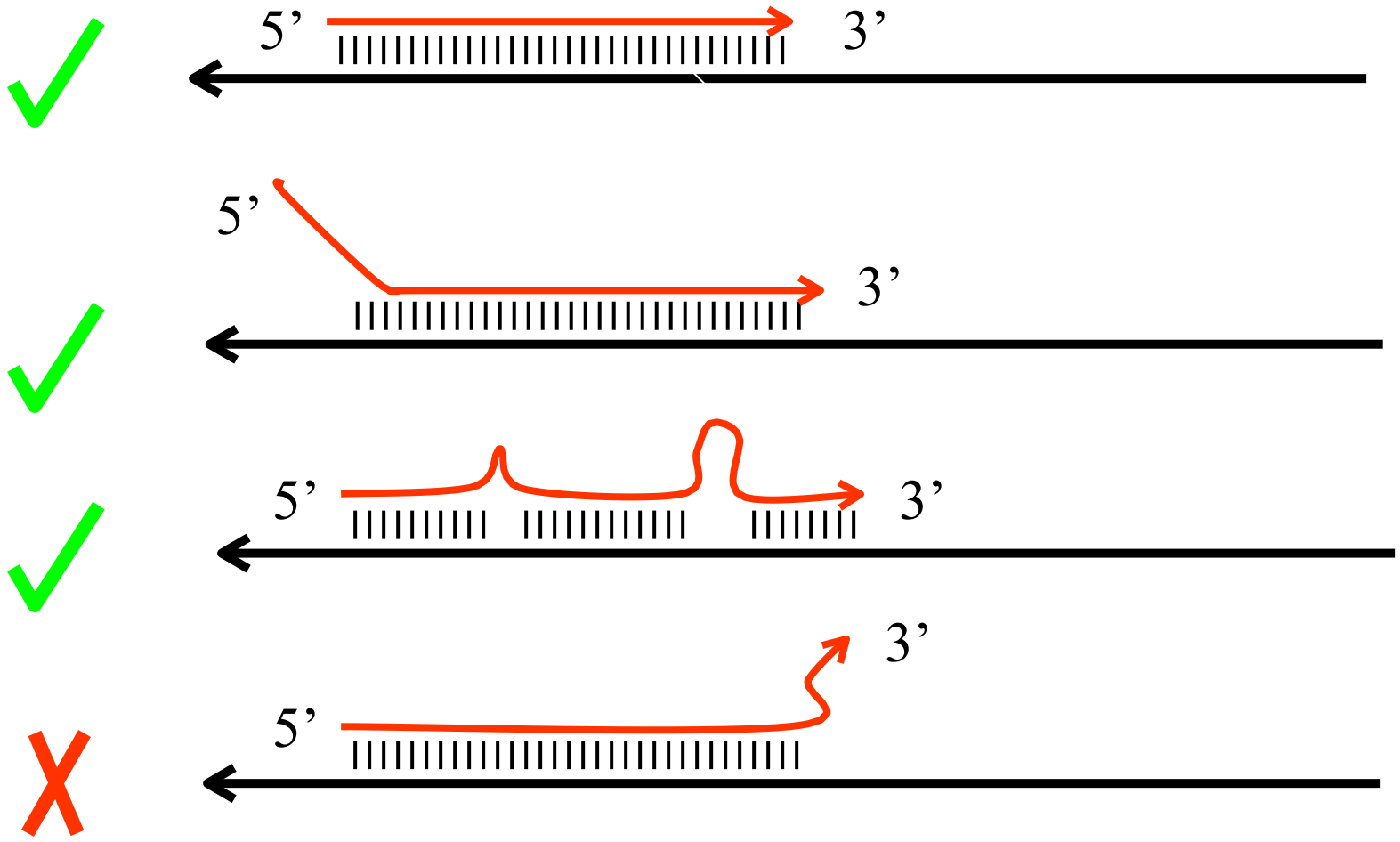
na matricové DNA nemají falešná vazebná místa

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

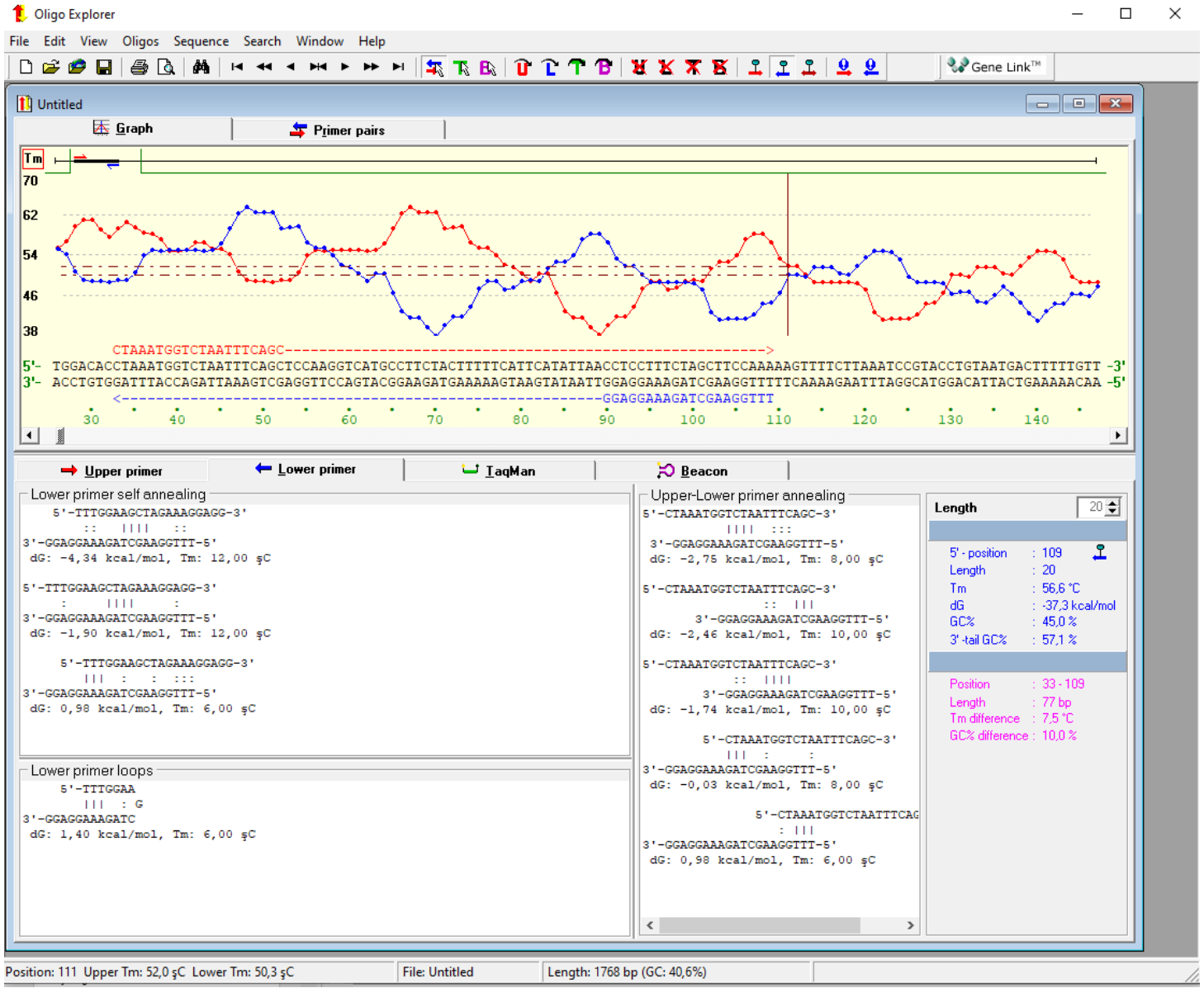


- Správně navržený primer, který nemá falešná vazebná místa na templátu:

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
  - Primer3 (http://primer3.sourceforge.net/webif.php)
  - Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi)
  - PrimerZ – design oligo pro lidsképromotory, exony a SNP (http://genepipe.ngc.sinica.edu.tw/primerz/beginDesign.do)
  - 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

Oligo Calculator

Oligo Calc: Oligonucleotide Properties Calculator

Enter Oligonucleotide Sequence Below
OD calculations are for single-stranded DNA or RNA

Nucleotide base codes

Reverse Complement Strand(5' to 3') is:

5' modification (if any) | 3' modification (if any) | Select molecule: ssDNA

50 nM Primer
50 mM Salt (Na^+)
1 Measured Absorbance at 260 nanometers

Calculate | Swap Strands | BLAST | mfold

Physical Constants

Length: 0 Molecular Weight: [4] GC content: %

1 ml of a sol'n with an Absorbance of 1 at 260 nm
is microMolar [5] and contains micrograms.

Melting Temperature (T_M) Calculations

[1] °C (Basic)
[2] °C (Salt Adjusted)
[3] °C (Nearest Neighbor)

Thermodynamic Constants Conditions: 1 M NaCl at 25°C at pH 7.

RlnK cal/(°K*mol) deltaH Kcal/mol
deltaG Kcal/mol deltaS cal/(°K*mol)

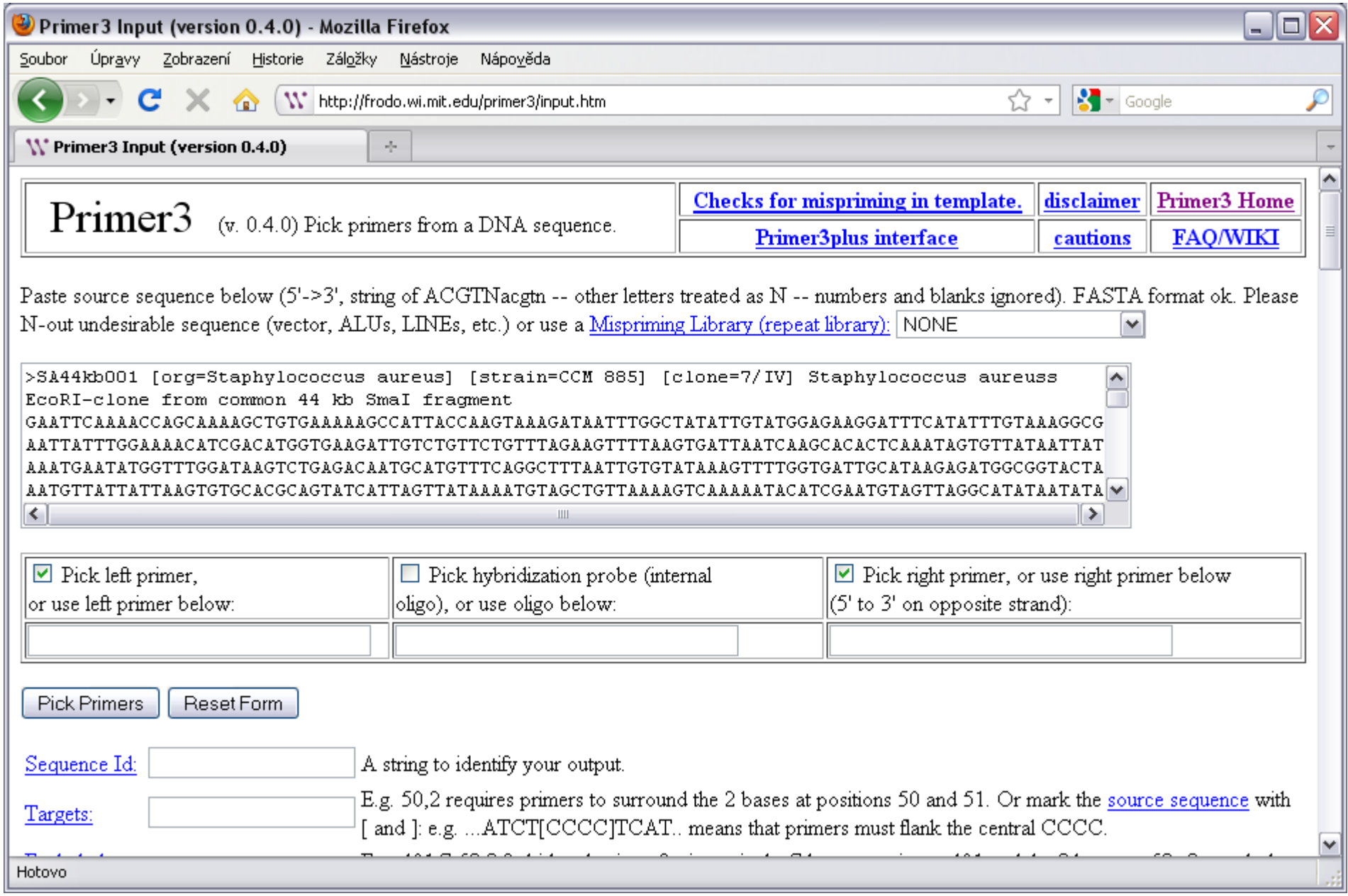
Deprecated Hairpin/self dimerization calculations

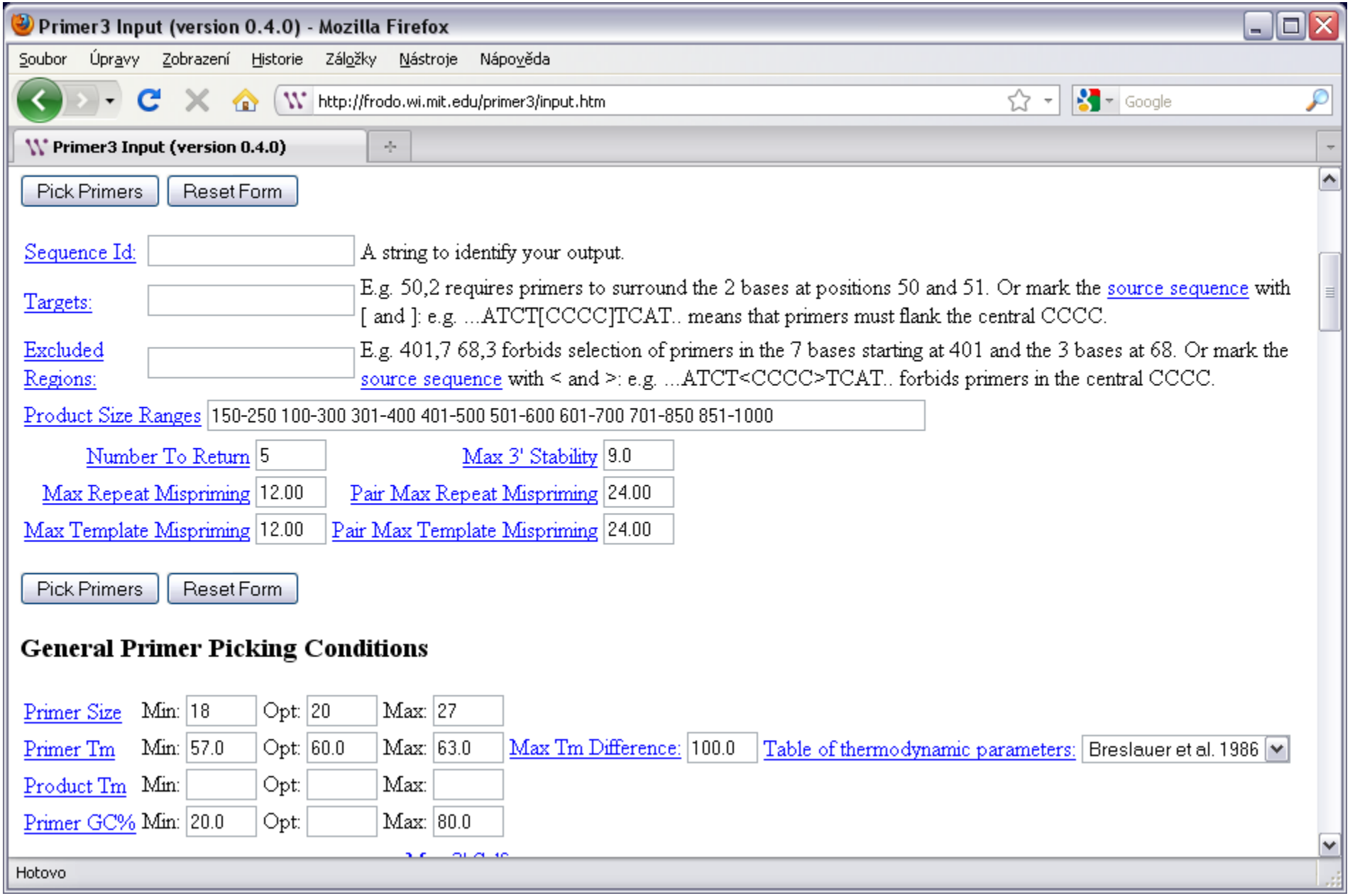
5 (Minimum base pairs required for single primer self-dimerization)
4 (Minimum base pairs required for a hairpin)

Check Self-Complementarity

Citation: Kibbe WA. 'OligoCalc: an online oligonucleotide properties calculator'. (2007)

Primer 3 http://primer3.sourceforge.net/webif.php




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

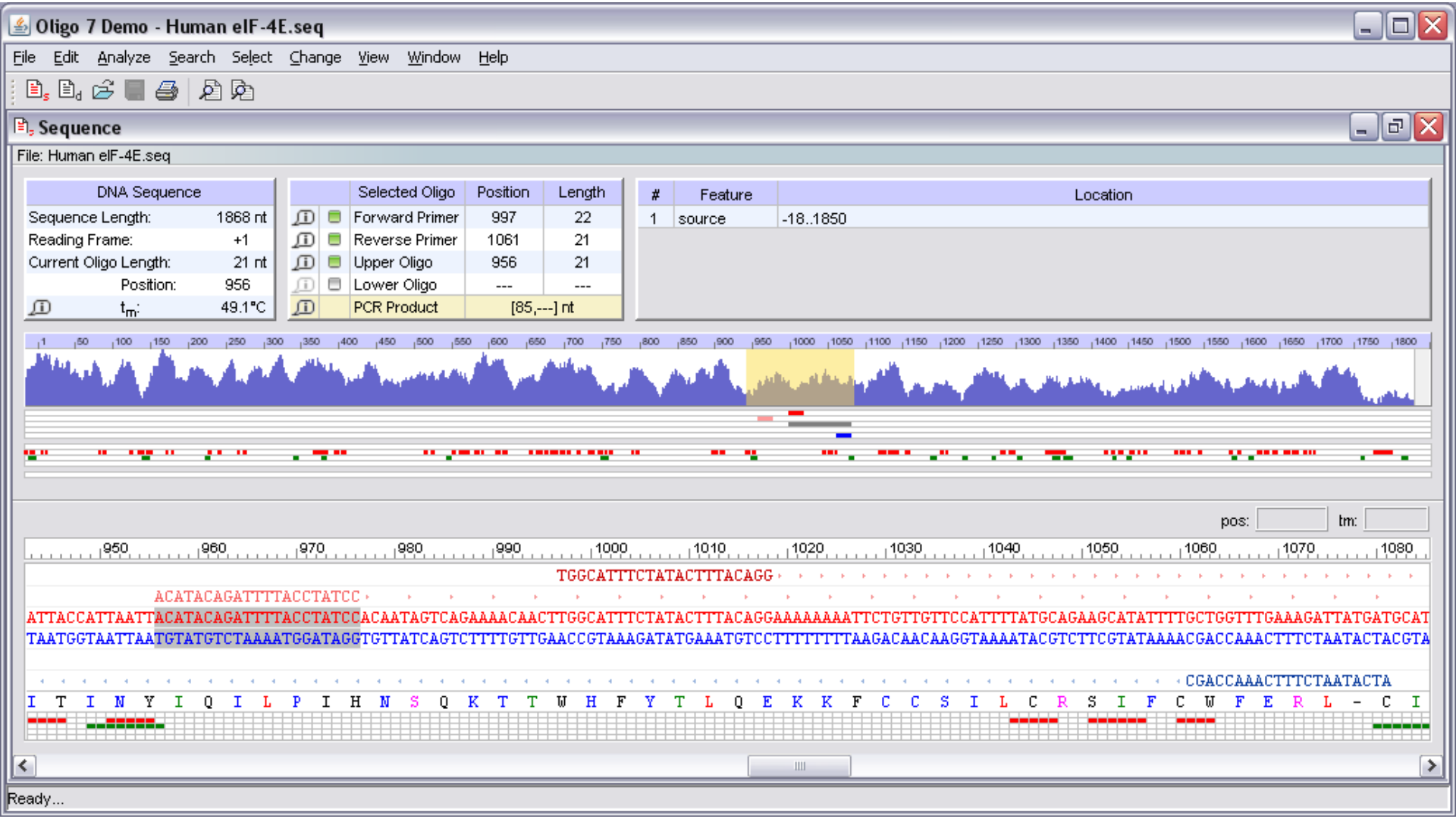
OLIGO	start	len	tm	gc%	any	3'	seq
LEFT PRIMER	159	25	57.21	32.00	6.00	2.00	AATCAAGCACACTCAAATAGTGTTA
RIGHT PRIMER	429	25	58.40	36.00	4.00	3.00	AACTCCTATGAAGACAACCTTTTTC

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 – příklad komerčního software

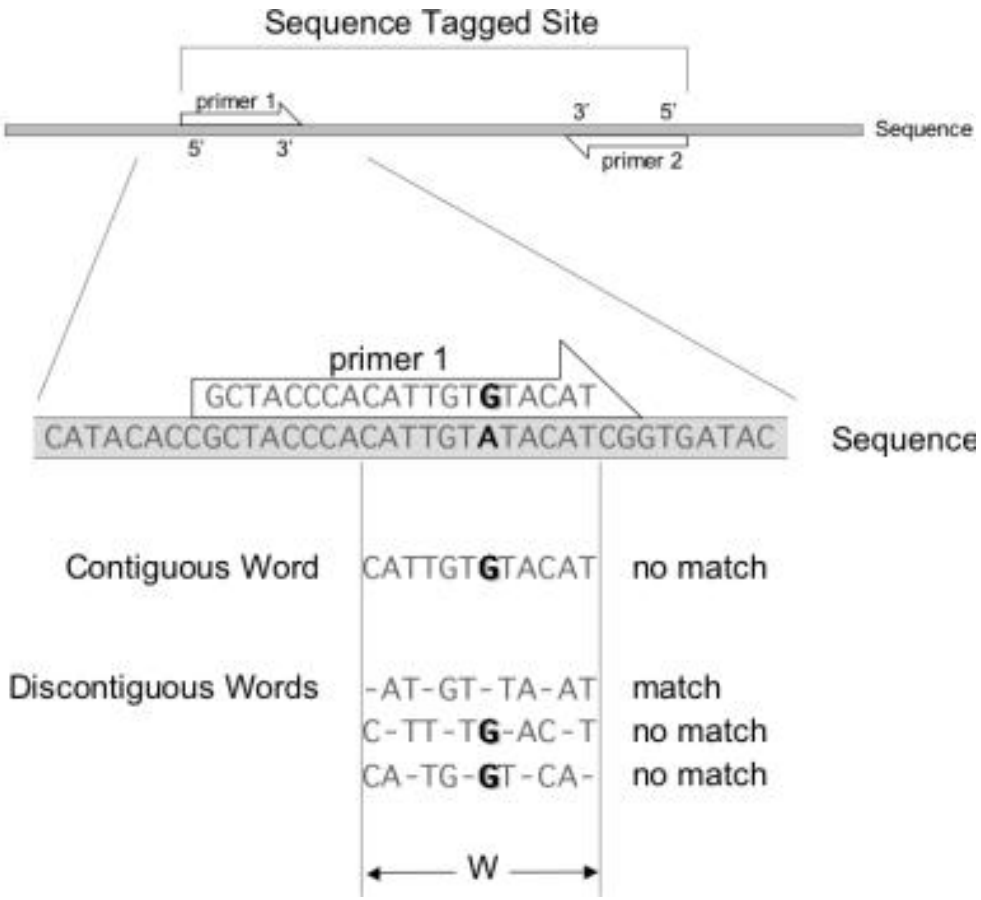


Výsledky poskytované programy pro návrh oligonukleotidů

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

Electronic PCR (e-PCR)

- Většina programů pro design PCR-primerů je omezená délkou templátu
- Posouzení falešných vazebných míst ideálně vyžaduje práci s celogenomovou sekvencí nebo částí databáze
- e-PCR označuje výpočetní postup, který se používá k prohledávání sekvencí DNA s cílem nalezení jedinečných míst, které odpovídají výsledkům PCR
- Nástroje lze využít pro eliminaci falešných pozitivních výsledků nebo design degenerovaných univerzálních oligonukleotidů



NCBI Primer-Blast

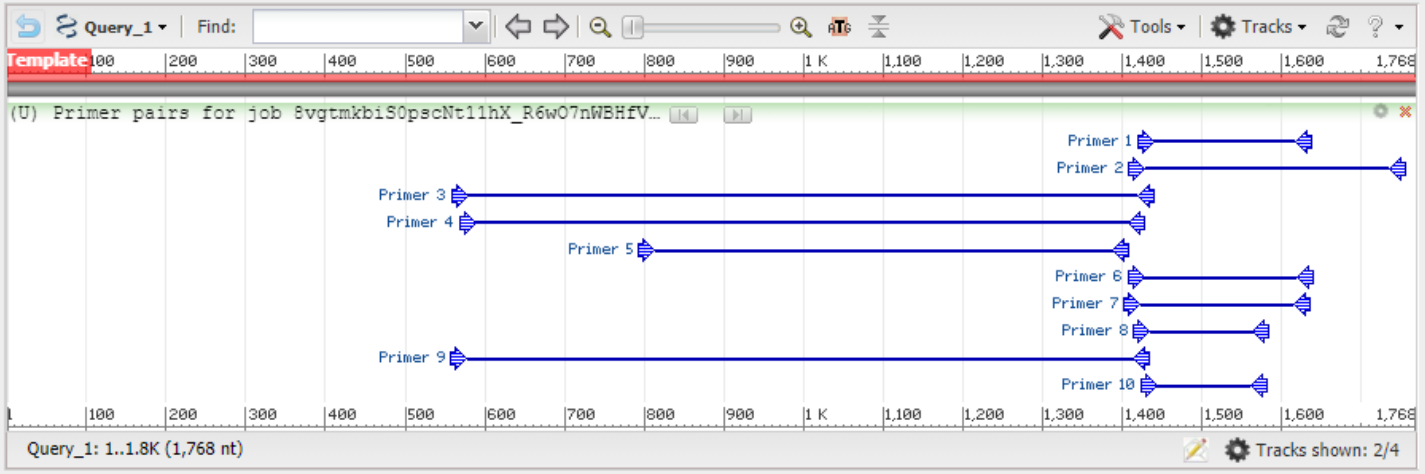
U.S. National Library of Medicine | NCBI | Sign in to NCBI

Primer-BLAST » JOB ID:8vgtmkbiS0pscNt11hX_R6wO7nWBHfVogA

Primer-BLAST Results

Input PCR template	lcl|Query_1
Range	1 - 1768
Specificity of primers	Primer pairs are specific to input template as no other targets were found in selected database: Nucleotide collection (nt) (Organism limited to Staphylococcus aureus subsp. aureus)
Other reports	▷Search Summary

Graphical view of primer pairs



Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AAAACCCTGGCGTTACCCAA	Plus	20	1420	1439	60.11	50.00	5.00	3.00
Reverse primer	CTATGCGGCATCAGAGCAGA	Minus	20	1637	1618	59.97	55.00	7.00	2.00
Product length	218								

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: Mouse | Assembly: Jul. 2007 | Forward Primer: TGCACCACCAaCTGCTT | Reverse Primer: 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

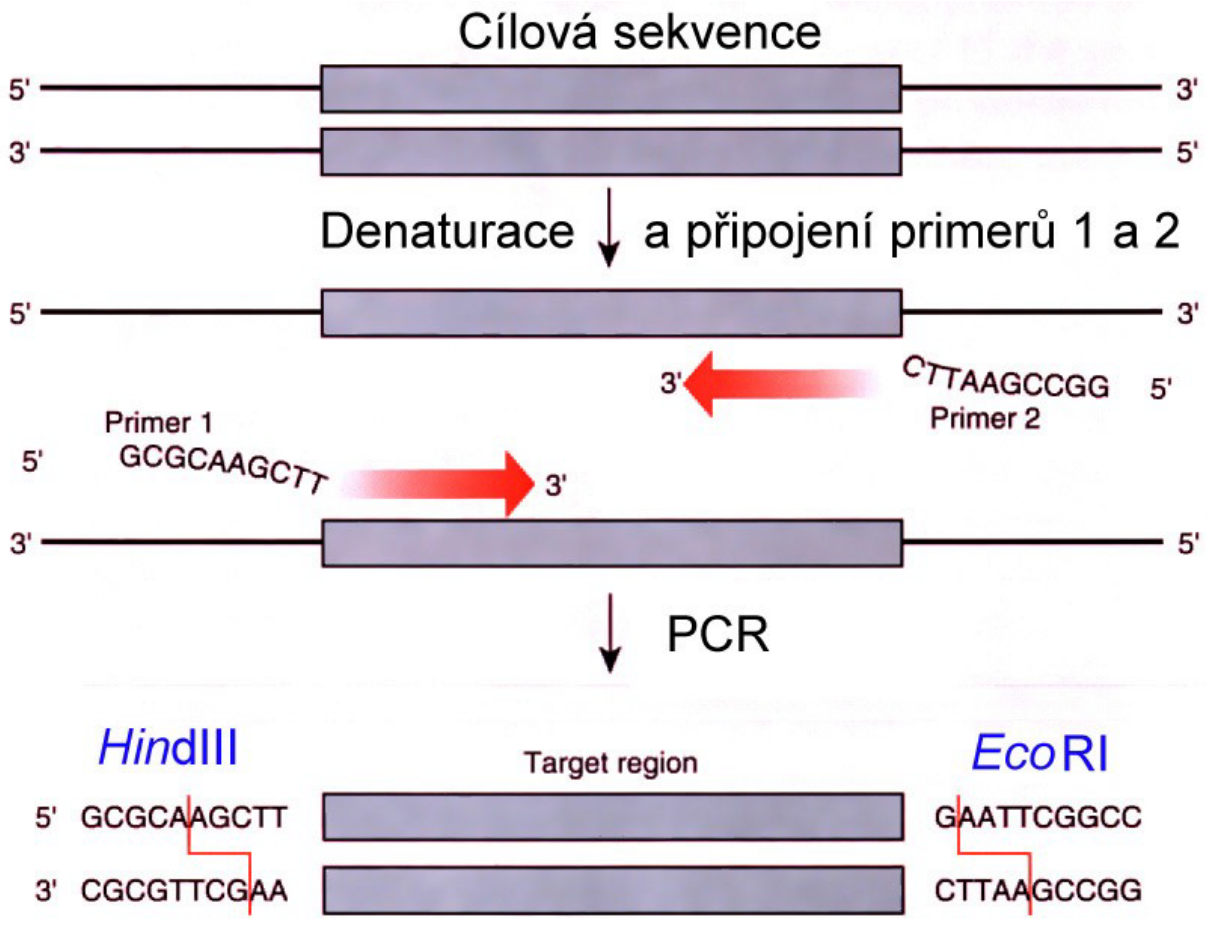
When successful, the search returns a sequence output file in fasta format containing all sequence in the database that lie between and include the primer pair. The fasta header describes the region in the database and the primers. The fasta body is capitalized in areas where the primer sequence matches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

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

Pokročilý návrh primerů

- Alelově specifické primery
- Molekulární diagnostika
  - Vícenásobné detekce - primery pro multiplex PCR
  - Zajištění kompatibility primerů v reakci
- Konsenzní primery
  - Vyžaduje identifikaci konzervativních oblastí na základě mnohonásobných přiložení sekvencí (multiple alignment)
  - Pro klonování
  - Pro mutagenezi
- 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

- NCBI/ Primer-BLAST
- 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,…)

Nejčastěji používané softwarové balíky pro manipulaci se sekvencemi

- Geneious (Biomatters, Inc., New Zealand)
- Ugene (http://ugene.net/) freeware
- CLC Genomics Workbench (CLC bio, Cambridge)
- Vector NTI® (Life Technologies, Carlsbad, CA)
- Bioinformatics Toolbox rozšíření pro MATLAB®
- Hitachi DNASIS® MAX Sequence Analysis Software (Helixx Technologies, Inc., Canada)
- DNASTAR Lasergene (DNASTAR, Inc., Madison, WI)
- Accelrys GCG Package (Accelrys Inc., San Diego)