

# CG020 Genomika

## Přednáška 1

### Úvod do bioinformatiky

Jan Hejátko

**Funkční genomika a proteomika rostlin,**  
Mendelovo centrum genomiky a proteomiky rostlin,  
Středoevropský technologický institut (CEITEC), Masarykova univerzita, Brno  
[hejatko@sci.muni.cz](mailto:hejatko@sci.muni.cz), [www.ceitec.muni.cz](http://www.ceitec.muni.cz)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst....
  - Další [www genomové nástroje](#)

# Schéma předmětu

- **Kapitola 01**
  - Úvod do bioinformatiky
  
- **Kapitola 02**
  - Identifikace genů
  
- **Kapitola 03**
  - Přístupy reverzní genetiky
  
- **Kapitola 04**
  - Přístupy genetiky přímé

# Schéma předmětu

- **Kapitola 05**
  - Přístupy funkční genomiky
  
- **Kapitola 06**
  - Protein-protein interakce a jejich analýza
  
- **Kapitola 07**
  - Současné metody sekvenování DNA
  
- **Kapitola 08**
  - Struktura genomů

# Schéma předmětu

- **Kapitola 09**
  - Evoluce genomů
  
- **Kapitola 10**
  - Genomika a systémová biologie
  
- **Kapitola 11**
  - Praktické aspekty funkční genomiky
  - Modelové organismy
  - PCR
  - Zásady navrhování primerů

# Literatura

- Literární zdroje pro kapitolu 01:
  - **Bioinformatics and Functional Genomics**, 3rd Edition, Jonathan Pevsner, Wiley-Blackwell, 2015  
<http://www.bioinfbook.org/php/?q=book3>
  - **Úvod do praktické bioinformatiky**, Fatima Cvrčková, 2006, Academia, Praha
  - **Plant Functional Genomics**, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey

# Osnova

- Schéma předmětu
- Definice



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# GENOMIKA-co to je?

- *Sensu lato* (v širším pojetí) zkoumá **STRUKTURU** a **FUNKCI** genomů
  - Předpokladem je znalost genomu (sekvencí)-práce s databázemi
- *Sensu stricto* (v užším pojetí) zkoumá **FUNKCI** jednotlivých genů - **FUNKČNÍ GENOMIKA**
  - používá zejména přístupy **REVERZNÍ GENETIKY**



# GENOMIKA-co to je?

## role BIOINFORMATIKY ve FUNKČNÍ GENOMICE

Přístupy „klasické“ genetiky

„Reverzně genetický“ přístup

5'TTATATATATATATATTAAAAAATAAAATAAAA  
GAACAAAAAGAAAATAAAATA....3'



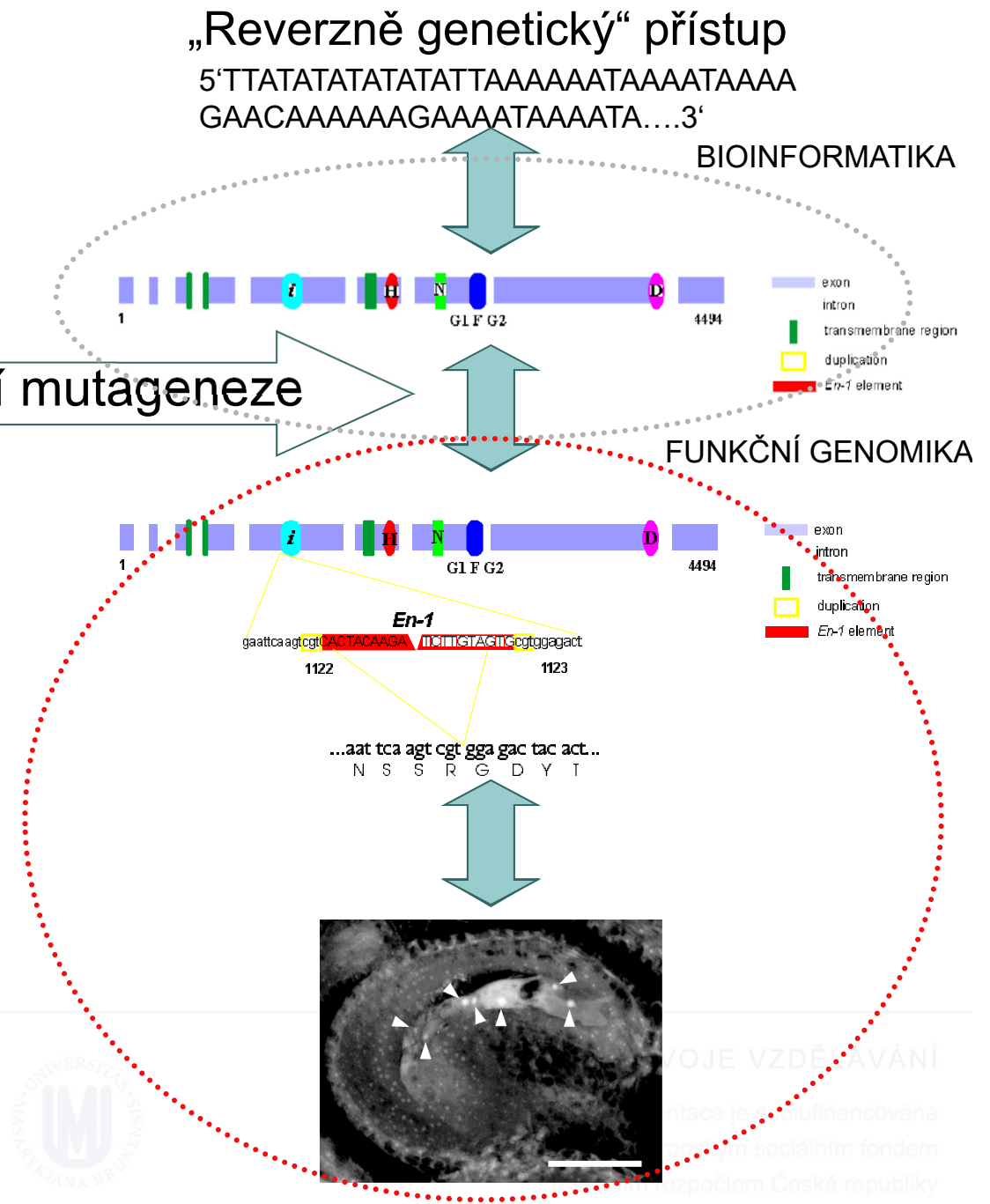
3

:

1



?



EVROPSKÁ UNIE



MLÁDEŽE A TĚLOVÝCHOVY

pro konkurenceschopnost



vání



UNIVERSITY MASARYKŮV  
BRNO

VOJE VZDĚÁVÁNÍ

ntace je financována

sociálním fondem

zpočetm České republiky

# Osnova

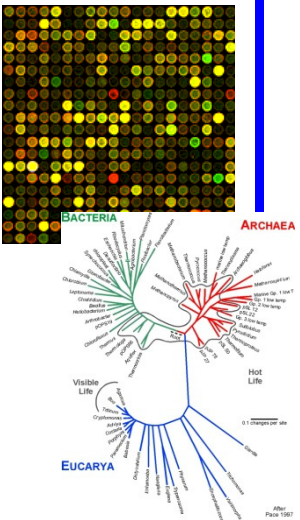
- Schéma předmětu
- Definice
- **Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY**



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# Bioinformatika



- **Definice bioinformatiky** (podle NIH vědeckého a technologického konsorcia pro biomedicínské informace)

**Výzkum, vývoj nebo aplikace výpočetních nástrojů a přístupů za účelem zvyšování rozvoje využití biologických, lékařských, dat o chování nebo zdraví, včetně těch, které umožňují taková data získávat, ukládat, organizovat, archivovat, analyzovat nebo vizualizovat.**

# What is Bioinformatics?

- Interface of **biology** and **computers**
- Analysis of **proteins, genes** and **genomes** using **computer algorithms** and **computer databases**
- **Genomics** is the **analysis of genomes**. The **tools of bioinformatics** are used **to make sense** of the **billions of base pairs of DNA** that are sequenced by genomics projects.

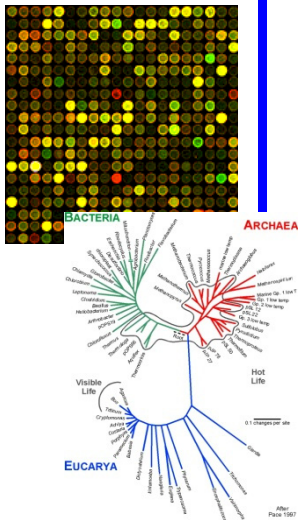
J. Pevsner,  
<http://www.bioinfbook.org/index.php>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# Bioinformatika



- **Bioinformatika ve funkční genomice**
  - **Zpracování a analýza sekvenačních dat**
    - Identifikace referenčních sekvencí
    - Identifikace genů
    - Identifikace homologů, ortologů a paralogů
    - Korelační analýzy mezi genomy a fenotypy (včetně člověka)
  - **Zpracování a analýza transkripčních dat**
    - Transkripční profilování pomocí DNA čipů nebo next-gen sekvenování
  - **Vyhodnocování experimentálních dat a predikce nových regulací v přístupech systémové biologie**
    - Matematické modelování genových regulačních sítí

# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# Spektrum on-line zdrojů

<b>EMBNet National Nodes</b>		
Vienna Biocenter	Austria	<a href="http://www.at.embnet.org/">http://www.at.embnet.org/</a>
BEN	Belgium	<a href="http://www.be.embnet.org/">http://www.be.embnet.org/</a>
BioBase	Denmark	<a href="http://biobase.dk/">http://biobase.dk/</a>
CSC	Finland	<a href="http://www.fi.embnet.org/">http://www.fi.embnet.org/</a>
INFOTIAGEN	France	<a href="http://www.infobiogen.fr/">http://www.infobiogen.fr/</a>
GENIUSnet	Germany	<a href="http://genome.dkfz-heidelberg.de/biounit/">http://genome.dkfz-heidelberg.de/biounit/</a>
IMBB	Greece	<a href="http://www.imbb.forth.gr/">http://www.imbb.forth.gr/</a>
HEN	Hungary	<a href="http://www.hu.embnet.org/">http://www.hu.embnet.org/</a>
INCEBI	Ireland	<a href="http://acer.gen.tcd.ie/">http://acer.gen.tcd.ie/</a>
INN	Israel	<a href="http://dapsas.weizmann.ac.il/bcd/inn.html">http://dapsas.weizmann.ac.il/bcd/inn.html</a>
IEN-ADR	Italy	<a href="http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm">http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm</a>
CAOS/CAMM	Netherlands	<a href="http://www.caos.kun.nl/">http://www.caos.kun.nl/</a>
Bio	Norway	<a href="http://www.no.embnet.org/">http://www.no.embnet.org/</a>
IBB	Poland	<a href="http://www.ibb.waw.pl/">http://www.ibb.waw.pl/</a>
IGC	Portugal	<a href="http://www.igc.gulbenkian.pt/">http://www.igc.gulbenkian.pt/</a>
GeneBee	Russia	<a href="http://www.genebee.msu.su/">http://www.genebee.msu.su/</a>
CNB-CSIC	Spain	<a href="http://www.es.embnet.org/">http://www.es.embnet.org/</a>
BMC	Sweden	<a href="http://www.embnet.se/">http://www.embnet.se/</a>
SIB	Switzerland	<a href="http://www.ch.embnet.org/">http://www.ch.embnet.org/</a>
SEQNET	UK	<a href="http://www.seqnet.dl.ac.uk/">http://www.seqnet.dl.ac.uk/</a>
<b>EMBNet Specialist Nodes</b>		
MIPS	Germany	<a href="http://www.mips.biochem.mpg.de/">http://www.mips.biochem.mpg.de/</a>
ICGEB	Italy	<a href="http://www.icgeb.trieste.it/">http://www.icgeb.trieste.it/</a>
Pharmacia Upjohn	Sweden	<a href="http://www.pnu.com/">http://www.pnu.com/</a>
F.Hoffmann-La Roche	Switzerland	<a href="http://www.roche.com/">http://www.roche.com/</a>
EBI	UK	<a href="http://www.ebi.ac.uk/">http://www.ebi.ac.uk/</a>
HGMP-RC	UK	<a href="http://www.hgmp.mrc.ac.uk/">http://www.hgmp.mrc.ac.uk/</a>
Sanger	UK	<a href="http://www.sanger.ac.uk/">http://www.sanger.ac.uk/</a>
UMBER	UK	<a href="http://www.bioinf.man.ac.uk/dbbrowser">http://www.bioinf.man.ac.uk/dbbrowser</a>
<b>EMBNet Associate Nodes</b>		
IBBM	Argentina	<a href="http://sol.biol.unlp.edu.ar/embnet">http://sol.biol.unlp.edu.ar/embnet</a>
ANGES	Australia	<a href="http://www.angis.su.oz.au/">http://www.angis.su.oz.au/</a>
CBI	China	<a href="http://www.cbi.pku.edu.cn/">http://www.cbi.pku.edu.cn/</a>
CIGB	Cuba	<a href="http://bio.cigb.edu.cu/">http://bio.cigb.edu.cu/</a>
CDFD	India	<a href="http://salarjung.embnet.org.in/">http://salarjung.embnet.org.in/</a>
SANBI	South Africa	<a href="http://www.sanbi.ac.za">http://www.sanbi.ac.za</a>
<b>USA Information Providers</b>		
NCBI	USA	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>
NLM	USA	<a href="http://www.nlm.nih.gov/">http://www.nlm.nih.gov/</a>
NIH	USA	<a href="http://www.nih.gov/">http://www.nih.gov/</a>

# Spektrum on-line zdrojů

- EBI <http://www.ebi.ac.uk/services>

The screenshot displays the EBI Services website interface. The main heading is "Services" with a sub-menu including "Overview", "A to Z", "Service teams", and "Support". The primary section is "Bioinformatics services", which states: "We maintain the world's most comprehensive range of **freely available** and up-to-date **molecular databases**. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our **web services** to access our resources programmatically."

Services are categorized into several boxes:

- DNA & RNA**: genes, genomes & variation
- Gene expression**: RNA, protein & metabolite expression
- Proteins**: sequences, families & motifs
- Structures**: Molecular & cellular structures
- Systems**: reactions, interactions & pathways
- Chemical biology**: chemogenomics & metabolomics
- Ontologies**: taxonomies & controlled vocabularies
- Literature**: Scientific publications & patents
- Other software**: cross-domain tools & resources

Additional sections include:

- Popular**: Ensembl, UniProt, PDBe, ArrayExpress, ChEMBL, BLAST, Europe PMC, Reactome, Train online, Support.
- Bioinformatics training**: Image of people in a meeting.
- Guide to resources**: Image of a woman looking at a screen.
- Service news**: Image of a book and a laptop.

At the bottom, there is a section for "Programmatic access" and "Browse EMBL-EBI web services". The browser's taskbar at the bottom shows various open applications like "EndNote X...", "Adobe Act...", and "Microsoft...".

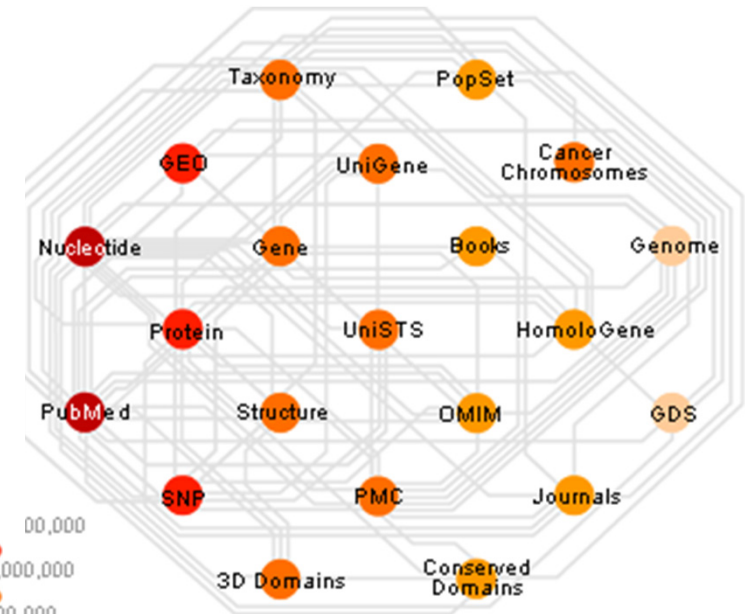


# Spektrum on-line zdrojů

□ NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI homepage with the following elements:

- Search Bar:** "All Databases" search field with a "Search" button.
- Navigation Menu (Left):** NCBI Home, Resource List (A-Z), All Resources, Chemicals & Bioassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, Variation.
- Welcome to NCBI:** "The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information." Includes links for "About the NCBI", "Mission", "Organization", "Research", and "RSS Feeds".
- Get Started:**
  - Tools:** Analyze data using NCBI software
  - Downloads:** Get NCBI data or software
  - How-To's:** Learn how to accomplish specific tasks at NCBI
  - Submissions:** Submit data to GenBank or other NCBI databases
- NCBI YouTube channel:** "Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel." Includes a "GO" button and a YouTube logo.
- Popular Resources:** PubMed, Books, UniGene, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem.
- NCBI Announcements:** "New version of Gen... available", "An integrated, downlo... for viewing and analy...", "NCBI's July Newslett... Bookshelf", "Introduction to the 10... Browser. PubMed's C...", "New Microbial BLAS...", "Now easier to use an... format and features c... BLAST services. inclu..."



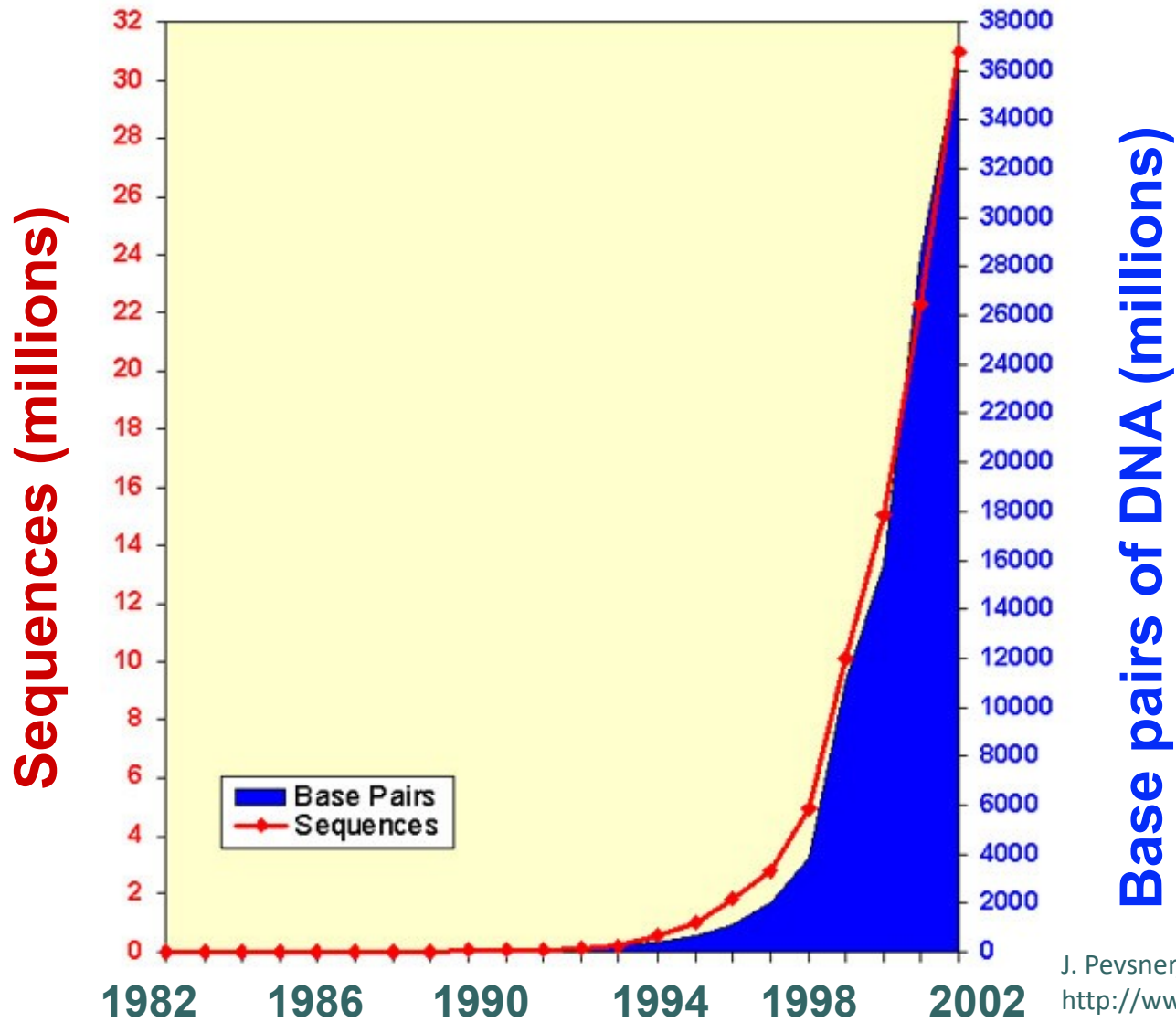
# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze

# Primární databáze

- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - Sekvence v databázích tzv. „Velké trojky“:
    - EMBL
      - <http://www.ebi.ac.uk/embl/>
    - GenBank,
      - <https://www.ncbi.nlm.nih.gov/>
    - DDBJ,
      - <http://www.ddbj.nig.ac.jp>
  - denně vzájemná výměna a zálohování dat
  - velká datová náročnost (kapacita i software)

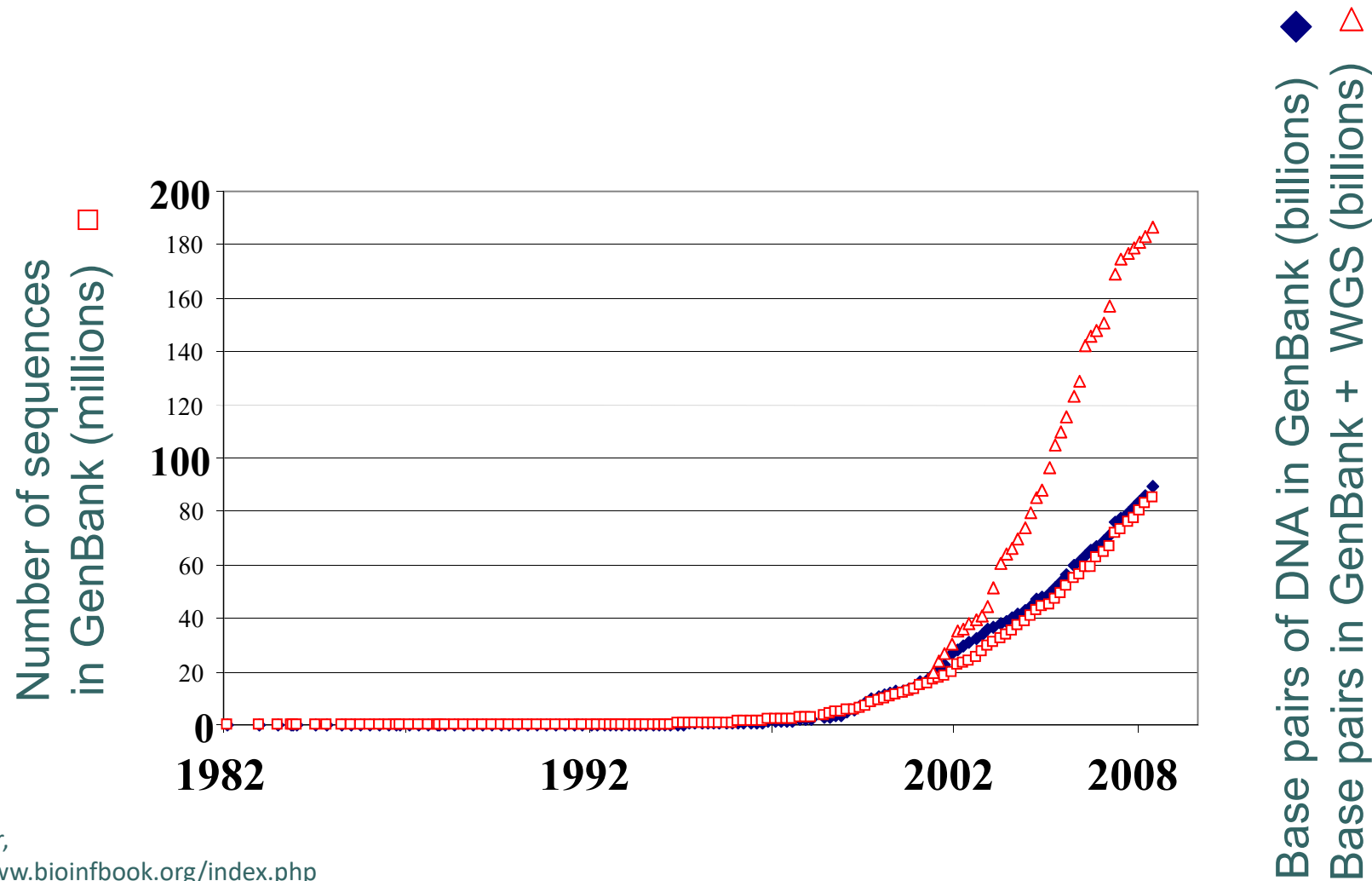
# Growth of GenBank



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
 Evropským sociálním fondem  
 a státním rozpočtem České republiky

# Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases



J. Pevsner,  
<http://www.bioinfbook.org/index.php>

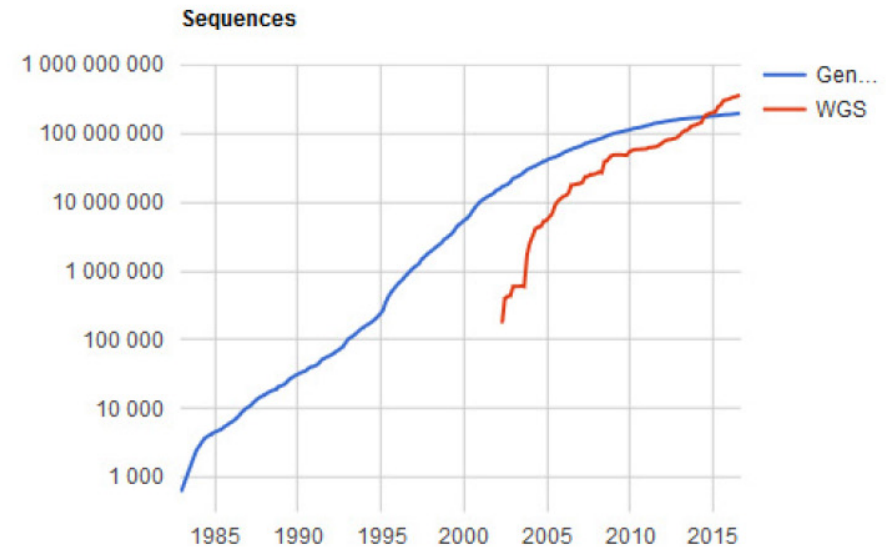
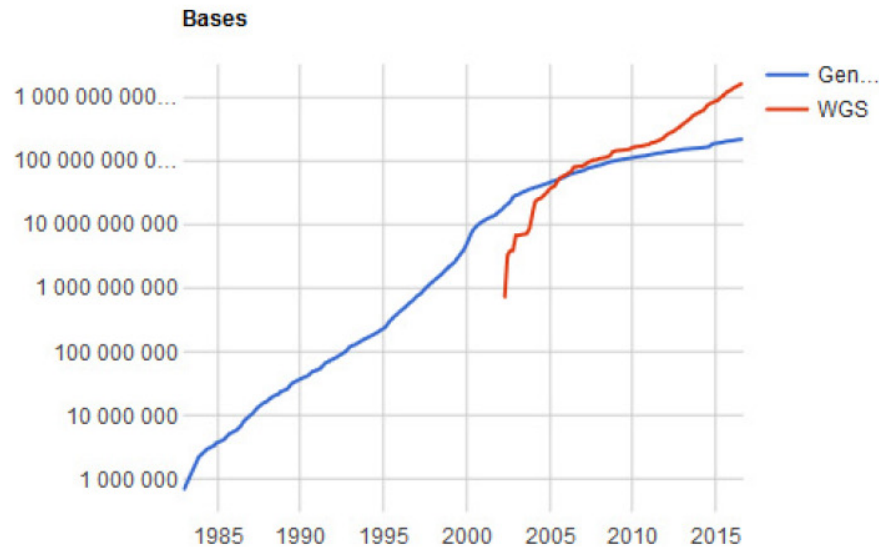


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
 Evropským sociálním fondem  
 a státním rozpočtem České republiky

# Growth of GenBank

## Aug 2016



- Prosinec **1982** 680 338 bp, 606 sekvencí
- Duben **2002**  $19 \times 10^9$  bp,  $17 \times 10^6$  sekvencí + WGS  $692 \times 10^6$  bp, 172 768 sekvencí
- Srpen **2016**  $218 \times 10^9$  bp,  $196 \times 10^6$  sekvencí + WGS  $1,6 \times 10^{12}$  bp,  $360 \times 10^6$  sekvencí

# WGS

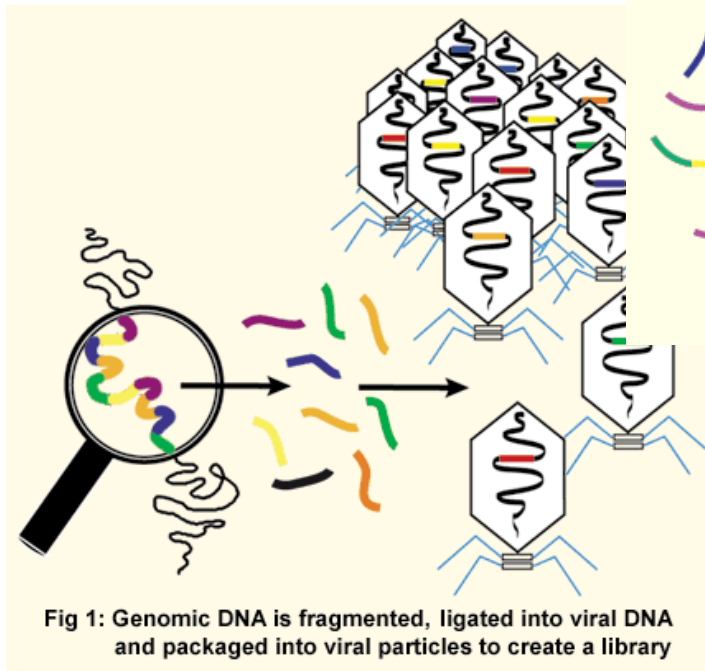
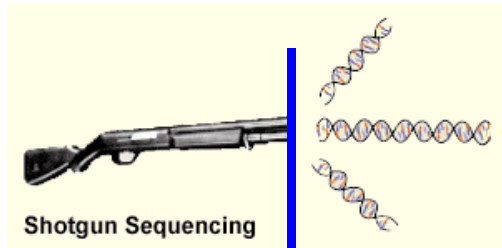


Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

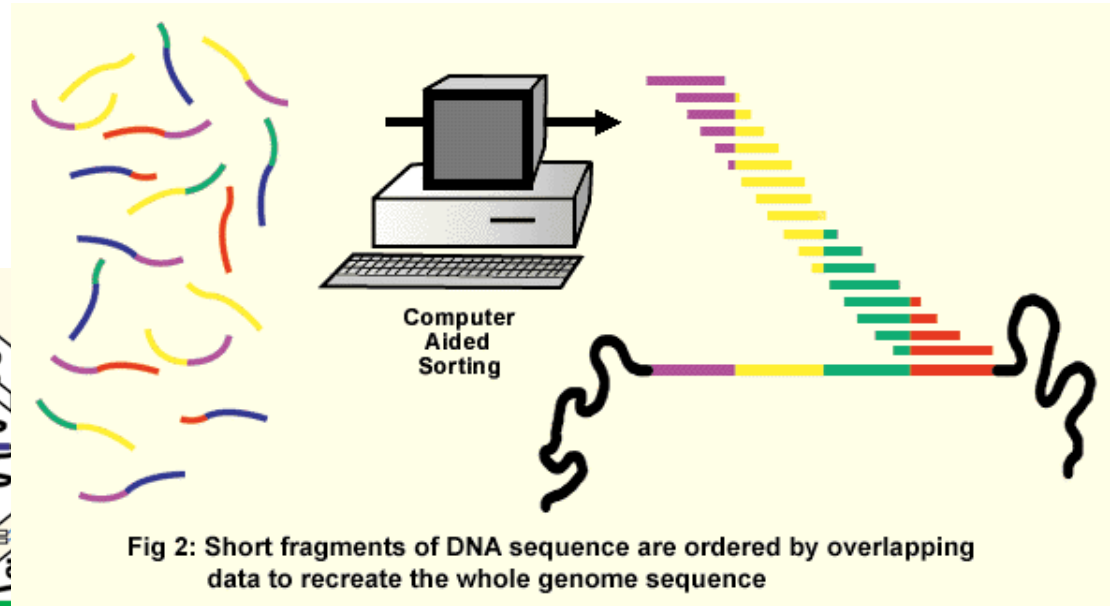
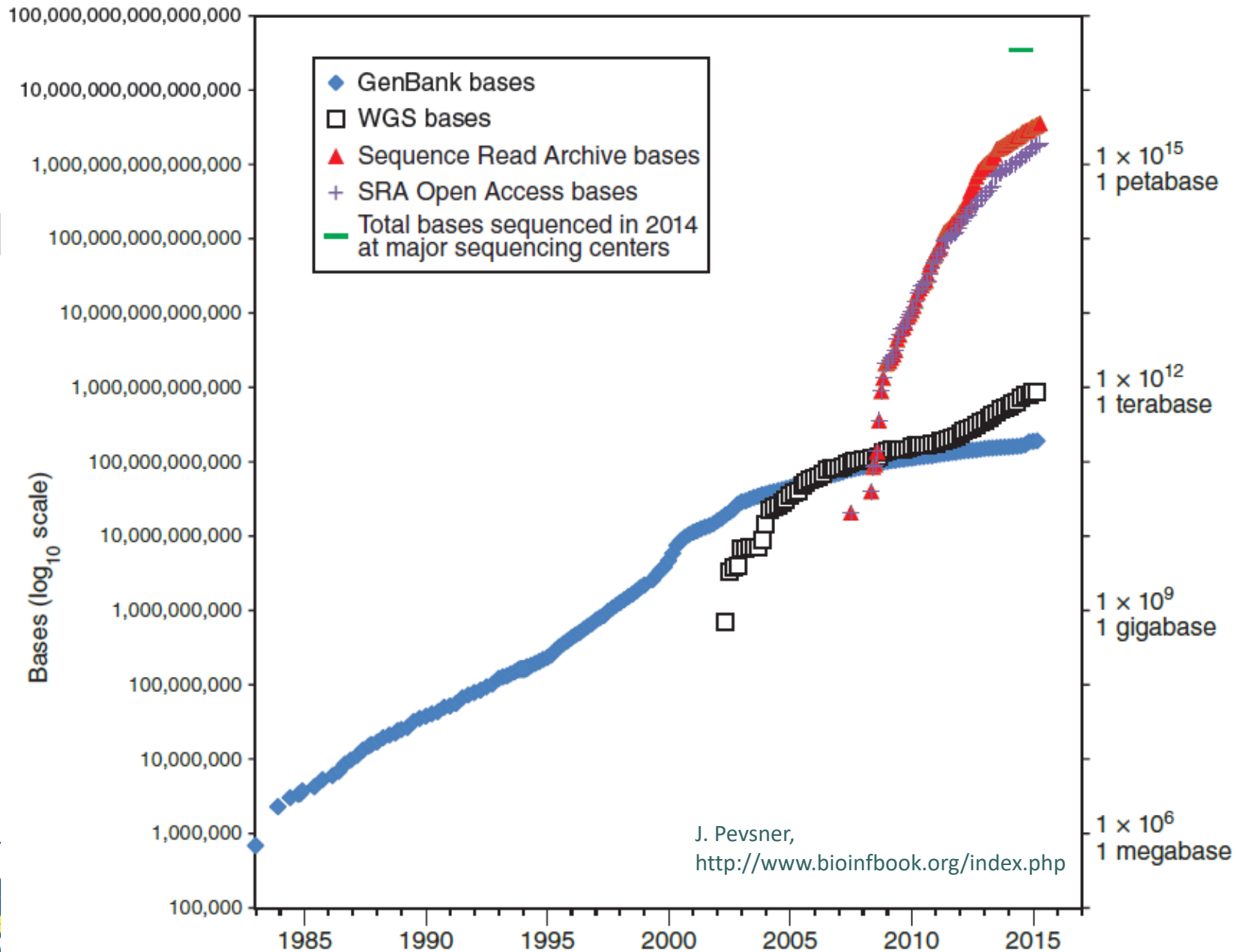


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

Interactive concepts in biochemistry, Rodney Boyer, Wiley, 2002, <http://www.wiley.com//college/boyer/0470003790/>

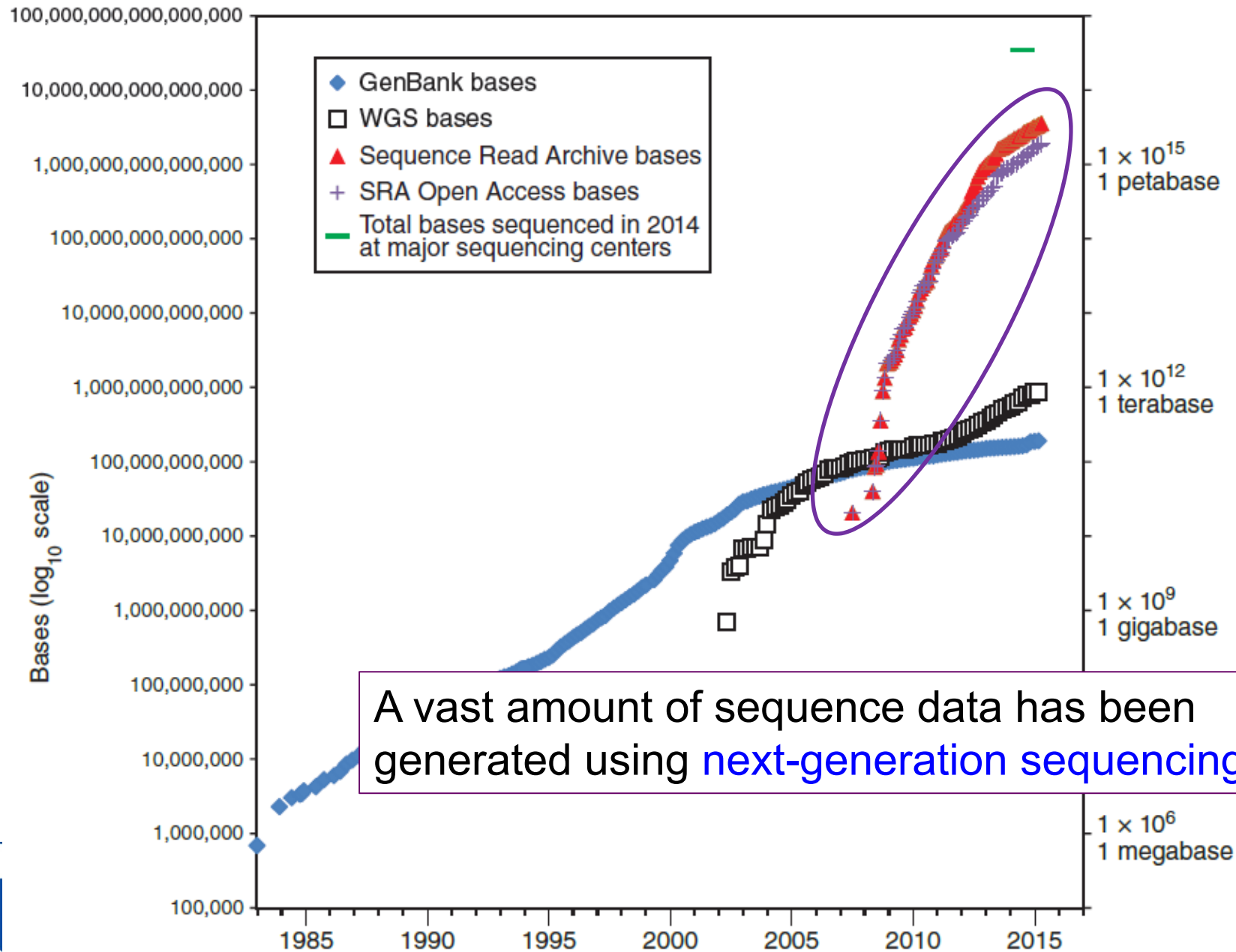
# Growth of DNA Sequence in Repositories



J. Pevsner,  
<http://www.bioinfbook.org/index.php>

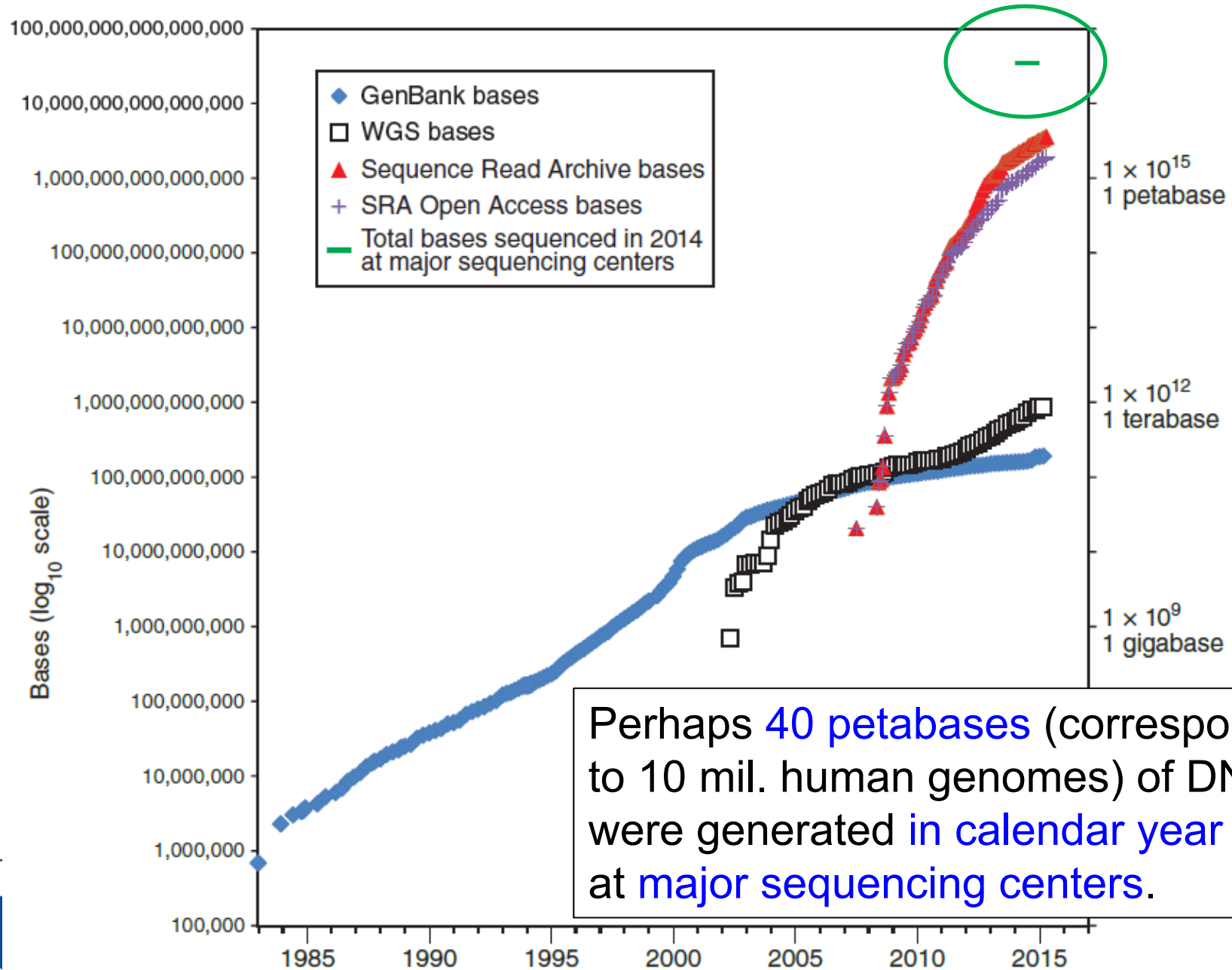


# Growth of DNA Sequence in Repositories



A vast amount of sequence data has been generated using **next-generation sequencing**.

# Growth of DNA Sequence in Repositories



Perhaps 40 petabases (corresponding to 10 mil. human genomes) of DNA were generated in calendar year 2014 at major sequencing centers.

# Primární databáze

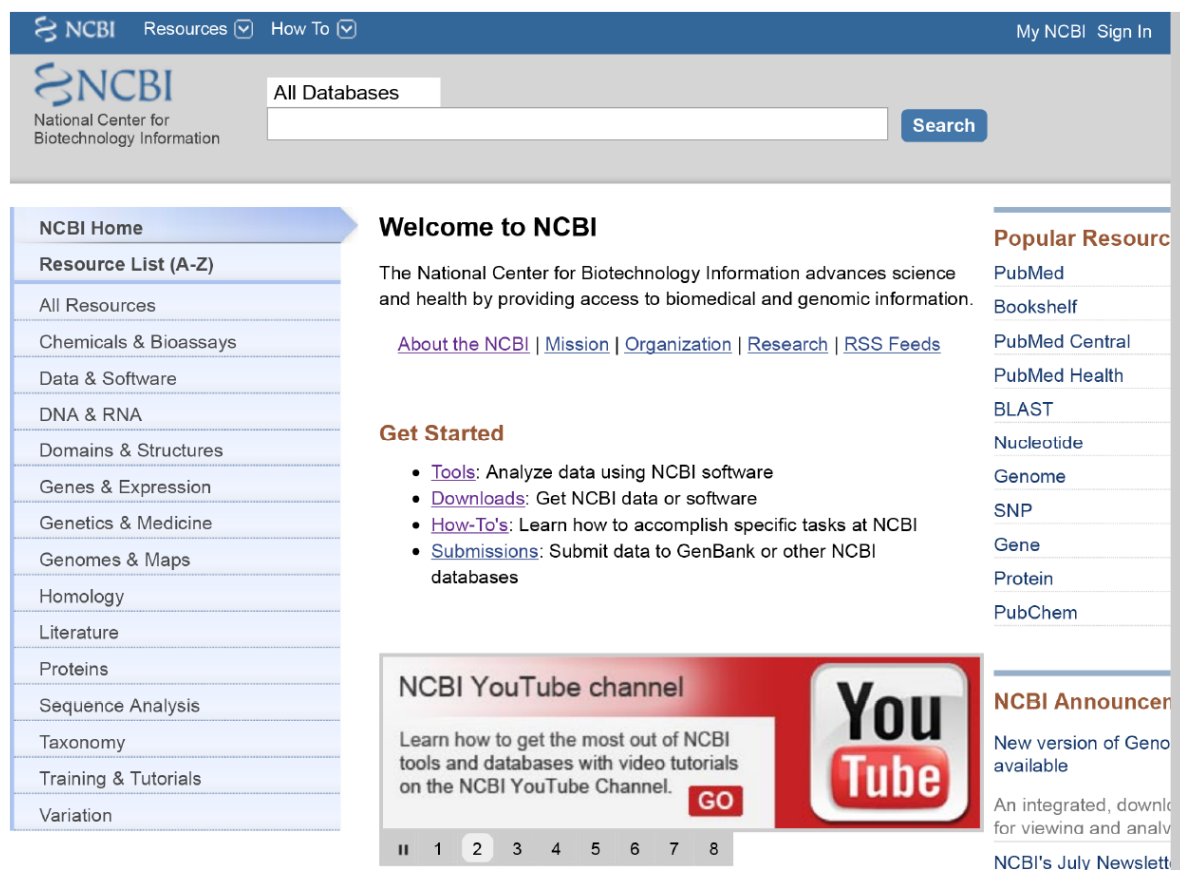
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - **Proteinové sekvence:**
    - PIR, <http://pir.georgetown.edu/>
    - MIPS, <http://www.mips.biochem.mpg.de>
    - SWISS-PROT, <http://www.expasy.org/sprot/>

# Primární databáze

- Typy sekvencí v primárních databázích
  - Standardní nukleotidové sekvence získané kvalitním sekvencováním
  - **ESTs** (**E**xpressed **S**equences **T**ags)
  - **HGTS** (**H**igh **T**hroughput **G**enome **S**equencing)
    - neanotované „surové“ výsledky sekvenačních projektů
  - Referenční sekvence anotovaných genomů
  - **TPAs** (**T**hird **P**arty **A**nnotation)
    - sekvence anotované jinými než původními autory

# Primární databáze

GenBank (NCBI) <https://www.ncbi.nlm.nih.gov/>



The screenshot shows the NCBI homepage with a navigation menu on the left, a search bar at the top, and a main content area with a 'Welcome to NCBI' message and a 'Get Started' section. A 'Popular Resources' list is on the right, and a 'NCBI Announcer' section is at the bottom right. A YouTube channel banner is also visible.

**NCBI Home**

- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

**Welcome to NCBI**

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

**Get Started**

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

**Popular Resources**

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

**NCBI Announcer**

New version of GenBank available

An integrated, downloadable tool for viewing and analyzing sequence data

NCBI's July Newsletter

**NCBI YouTube channel**

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel. [GO](#)

# Primární databáze

The screenshot displays the NCBI Gene database entry for the *virA* gene. The page is organized into several sections:

- Gene summary:** Includes the gene symbol (*virA*), description (two-component VirA-like sensor kinase), locus tag (pTl\_125), gene type (protein coding), RefSeq status (PROVISIONAL), organism (*Agrobacterium tumefaciens*), and lineage.
- Genomic context:** Shows the location of the gene on plasmid Ti (NC\_002377.1) with coordinates 145694-148183. A diagram illustrates the gene's position relative to other features like *virK*, *virB*, and *virC*.
- Genomic regions, transcripts, and products:** Provides a genomic sequence viewer for NC\_002377.1. A yellow circle highlights the 'Bibliography' section, which lists several scientific papers related to the gene.
- General information:** Includes links to various resources like BioProjects, Conserved Domains, and RefSeq Proteins.
- Related sites:** Lists external databases such as BLAST, GenBank, and UniGene.

The 'Bibliography' section, highlighted by a yellow circle, contains the following entries:

1. [Sequence analysis of the \*virA\* locus from \*Agrobacterium tumefaciens\* octopine  \$\Pi\$  plasmid pTl15955](#), Schrammeijer B, et al. J Exp Bot. 2000 Jun. PMID 10948245.
2. [The \*virA\* promoter is a host-range determinant in \*Agrobacterium tumefaciens\*](#), Turk SC, et al. Mol Microbiol. 1993 Mar. PMID 8469115.
3. [Characterization of the \*virA\* locus of \*Agrobacterium tumefaciens\*: a transcriptional regulator and host range determinant](#), Leroux B, et al. EMBO J. 1987 Apr. PMID 3595559.
4. [Analysis of the complete nucleotide sequence of the \*Agrobacterium tumefaciens\* \*virD\* operon](#), Thompson DV, et al. Nucleic Acids Res. 1988 May 25. PMID 2837739.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# Primární databáze

NC\_002377.1: 145K..148K (2.9Kbp)

Genes

**NP\_059797.1**

NP\_059797.1: two-component VirA-like sensor kinase  
total range: NC\_002377.1 (145,694..148,183)  
total length: 2,490  
strand: plus  
protein product length: 829

**Links & Tools**

GenBank View: [NC\\_002377.1 \(145,694..148,183\)](#), [NP\\_059797.1 \(145,694..148,183\)](#)  
FASTA View: [NC\\_002377.1 \(145,694..148,183\)](#), [NP\\_059797.1 \(145,694..148,183\)](#)  
BLAST Genomic: [NC\\_002377.1 \(145,694..148,183\)](#)  
Graphical View: [NP\\_059797.1](#)  
BLAST Protein: [NP\\_059797.1](#)  
BLINK Results: [NP\\_059797.1](#)

**Bibliography**

**Related articles in PubMed**

# Primární databáze

NCBI Nucleotide

Search Nucleotide for [ ] Go Clear

Dist. Preview/Index History

Get Subsequence Features

Clipboard Details Links

NC\_002377.1 [GI:10955016]

LOCUS NC\_002377 2490 bp DNA linear BCT 29-DEC-2003

DEFINITION Agrobacterium tumefaciens extrachrom plasmid Ti, complete sequence.

ACCESSION [NC\\_002377](#) REGION: 145694..148183

VERSION NC\_002377.1 GI:10955016

KEYWORDS

SOURCE Agrobacterium tumefaciens (Rhizobium radiobacter)

ORGANISM Agrobacterium tumefaciens; Rhizobiales; Rhizobium; Agrobacterium.

REVIEWED

Farrand, S.K., Oger, P.M., Schrammeijer, B., Hooykaas, P.J. and Winans, S.C.

TITLE Octopine-type Ti plasmid sequence

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 2490)

AUTHORS Zhu, J., Oger, P.M., Schrammeijer, B., Hooykaas, P.J., Farrand, S.K. and Winans, S.C.

TITLE Direct Submission

JOURNAL Submitted (07-MAR-2000) Microbiology, Cornell University, Wing Hall, Ithaca, NY 14853, USA

COMMENT PROVISIONAL REFSEQ: This record has not yet been subject to final NCBI review. The reference sequence was derived from [AF242881](#).

FEATURES

Location/Qualifiers

source 1..2490

/organism="Agrobacterium tumefaciens"

/mol\_type="genomic DNA"

/db\_xref="taxon:358"

/plasmid="Ti"

/note="extrachromosomal octopine-type"

gene 1..2490

/gene="virA"

/db\_xref="GeneID:1224316"

CDS 1..2490

/gene="virA"

/note="two-component regulator of vir regulon; VirA is a transmembrane histidine kinase"

/codon\_start=1

/transl\_table=11

/product="virA"

/protein\_id="NP\_059797.1"

/db\_xref="GI:10955141"



# Primární databáze

```
/translation="MNGRYSPTQDFKTKAKPWSILALIYAAMI FAFMAVASWQDNMT  
TQAILSQLRINADSASLQRDVLRHTCTVANYRPI I SRLGALRKNLEDLKQLFRQSH  
IVSEENRQQLRQLEVSLMSADAAVAAPGQNVRLQDSIASPTRALSSLPKASTDQT  
LEKPTLEASMMQLRQSPASISPI SLELEELQKQRLDEAFVILAREGPI ILSLL  
PQVKDLVNMQISTDAIEMLQRCLVYSLKNVEERSARIPLSSASVGLCLYIITL  
VYLRKKTDWLARRLDYELIKEI GVCFFGSRATSSQAALRI IQRPFDADTCALAL  
VDHDERWAVETFGAKHFKPVWDSVLRRIVSRTKADEBRATVFR IISKKIVHLFLHIP  
GLSILLAHKSTDKLIAVCSLGYQSYRFPFCQGETQLLELATACLCHYIDVRRKQTECD  
VLARLEHAQRLAVGTLAGGIAHFNNILGSLGHAEALQNSVSRTEVTRRYIDYII  
SSGDRAMLIIDQILTLRKRQEMIKPPSVSELVTEIAPLRLMALPPNIELSFRPDQMC  
SVI EGSPLRLQQLINICKNASQAMTANQIDII IIGQAPLPVKKILAHGVMPFGDYVL  
LSISDNQGGIPRAVLPHI FEPFPTTRARNGGTGLGLASVHGHISAPAGYIDVSTVGH  
GTRFDIYLPSPSKKFPVNPDSFFGRNKA PRGNGHI VALVFPDDLREAYRDKI AALGYE  
PVGFRTPNKRDIWISKGNEADLVMDQASLPEDQSPNSVDLVLKTA SIIIGGNDLKM  
LSREDVT RDLYLFPKPISSRTMAHALTKIKT"
```

```
ORIGIN  
1 atgaacggaa gatattcacc gaecggcgag gattttaaga caggcgcgaa gccctggctt  
61 atattggccc ttatcgttgc tgaatgatt ttocggttca tggcggttgc gtcctggcag  
121 gacaatgcca ctaccacgga aatcctcage caactacgat cgat taacgc cgacagcgcc  
181 tcactgcagc gogatgact cecgectcac acgggcaacg tggcgaacta ccgccccatt  
241 atctccaggg tgggagctct gcggaagaat ctggaagatt tgaagcaatt atttagacaa  
301 tctcatattg taagtgcagc caatgctgct caactgctac gccagctaga agtgtctcta  
361 aatctggctg acgcgcgctg cgcgccttt ggtgcgcaaa atgtacgct gcaagattcg  
421 ctggcagctt tcactcgtgc tttgagcagt ctccagaa aagcctcaac cgatcagact  
481 ttgaaaaaac caacagaatt ggttagcagt atgctccaat ttcttggca accaaagccg  
541 gctatttcat toagatcag ccttgaacta gagaggctcc aaaaacaacg cggctttgat  
601 gaagctcccg tgcgcaact tgcacgtgaa ggtcccatla tcttateget tttgccacag  
661 gtgaaagatc tgggtaacat gatcagacg tctgacacgc cagaatgac gtagatgctg  
721 cagcgcgagt gtttgagggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc  
781 ttctttgggt cegcttcagt gggctcttgc ctctacatca tcacctagt ctataggcta  
841 cgcacacaaa cagattggtt agcgcggcgt ttgatatac aagagctaat caaagagatc  
901 ggagtagttt ttgaagtgta ggcggccacc acgtcgtccg cgcacagctc actctgtatt  
961 atcagcgcct tcttgatgc cgtacgtgc cgttagctc tagtggacca tgacgtaga  
1021 tgggctgtcg aaacattcgg tgcgaaacac caaaacctg tctgggacga cagcgtgcta  
1081 cgcgaaatag tctctcgtac caaagcggac gaacgggcca cggtatccg catcatatcg  
1141 tgcacacaaa tctacattt gccctctgaa atccagctc tctcgatact actggtctac  
1201 aaatccacag ataaactaat tgcggtttgt tcaactgggtt accaaagcta tgcctctoga  
1261 ccttgccaag gcaaaatca gctcttggaa ctgcacacgc cctgctctg tcatatatac  
1321 gatgttcggc gtaagcagac cgaatgcgac gttttggcca cagcattgga gcatgcgcaa  
1381 cgccttgagg cagttggtac acttgcggcg ggaatgacac atgaatttaa taacattttg  
1441 ggctcaatcc tgggcaacgc agaattagca caaaactcgg tctctgaaac atctgtcacc  
1501 cgaagatata ttgactatat cattctgcca ggcgacagag ccatgctcat tctgatcag  
1561 atcttgacgc tgagccgaaa acaggagcgc atgatcaagc catttagtgt ctcagagctt  
1621 gtgaccgaaa tgcctcctt gctacgtatg gctctccgc caaacatoga gcttagtttc  
1681 agatttgatc aaatgcagag cgtgatcgaa ggaagccgcg ttgaacttca acaggtacta  
1741 ataacatct gcaagaatgc tcccaagcc atgacgcaa atggtcaaat cgcacatcct  
1801 atcagccaag cttttttacc agttaagaaa atctggcgc atggtttat gccocctggc  
1861 gactatgttc tctatctat tagcgaacat ggtggaggca tcccgagcgc tgtttacc  
1921 cacatttttg aacctctct taagacagca gctgcacgc gttgaaacgg tctcggcctt  
1981 gctctgtgct atggtcatat cagcgcgctt gcgggttaca toagcttag ttoactgtt  
2041 gggcatggga cgcgcttga catttatctc cctcgtctt ctaaagaaec cgtaaatcna  
2101 gacagttttt bccggccgaa taaggccacc cgtgaaacgc gggagattgt ggcactgtt  
2161 gacccgatg acctcctgag gtagcgtat gaagacaaga tgcgcgctc aggatagag  
2221 ccggtcggtt tctgacctt taatgaatt cgcgatggga tttcaaaagg caatgaagcc  
2281 gatctggtca tggctgacca agcgtctctt cctgaagatc aaagtcttaa tctcgtggat  
2341 ttagtgtca agacgcctc catcatcatt ggcggaatg atctcaaat gccctctta
```



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# What is an **Accession Number**?

An accession number is label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	<b>DNA</b>
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	<b>RNA</b>
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	<b>Protein</b>
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

J. Pevsner,  
<http://www.bioinfbook.org/index.php>

# NCBI's important **RefSeq** project: best **representative sequences**

**RefSeq** (accessible via the main page of NCBI) provides an **expertly curated accession number** that corresponds to the most stable, agreed-upon “reference” version of a sequence.

RefSeq identifiers include the following formats:

Complete genome	NC_#####
Complete chromosome	NC_#####
Genomic contig	NT_#####
mRNA (DNA format)	NM_##### e.g. NM_006744
Protein	NP_##### e.g. NP_006735

J. Pevsner,  
<http://www.bioinfbook.org/index.php>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# RefSeq

two-component VirA-like sensor kinase

**NCBI Reference Sequences (RefSeq)**

**Genome Annotation**

The following sections contain reference sequences that belong to a specific genome build. [Explain](#)

**Reference assembly**

**Genomic**

1. **NC\_003065.3**

Range: 180831..183332  
Download: [GenBank](#), [FASTA](#), [Sequence Viewer \(Graphics\)](#)

**mRNA and Protein(s)**

1. **NP\_396486.1** two component sensor kinase [Agrobacterium tumefaciens str. C58]

UniProtKB/Swiss-Prot: [P18540](#)

Conserved Domains (3) [summary](#)

<a href="#">cd00075</a>	HATPase_c: Histidine kinase-like ATPases; This family includes several ATP-binding proteins for example: histidine kinase, DNA gyrase B, topoisomerases, heat shock protein HSP90, phytochrome-like ATPases and DNA mismatch repair proteins
<a href="#">cd00082</a>	HisKA: Histidine Kinase A (dimerization/phosphoacceptor) domain; Histidine Kinase A dimers are formed through parallel association of 2 domains creating 4-helix bundles; usually these domains contain a conserved His residue and are activated via ...
<a href="#">PRK13837</a>	PRK13837; two-component VirA-like sensor kinase; Provisional

**Related Sequences**

# NCBI's RefSeq project: many accession number formats for genomic, mRNA, protein sequences

<u>Accession</u>	<u>Molecule</u>	<u>Method</u>	<u>Note</u>
AC_123456	Genomic	Mixed	Alternate complete genomic
AP_123456	Protein	Mixed	Protein products; alternate
NC_123456	Genomic	Mixed	Complete genomic molecules
NG_123456	Genomic	Mixed	Incomplete genomic regions
NM_123456	mRNA	Mixed	Transcript products; mRNA
NM_123456789	mRNA	Mixed	Transcript products; 9-digit
NP_123456	Protein	Mixed	Protein products;
NP_123456789	Protein	Curation	Protein products; 9-digit
NR_123456	RNA	Mixed	Non-coding transcripts
NT_123456	Genomic	Automated	Genomic assemblies
NW_123456	Genomic	Automated	Genomic assemblies
NZ_ABCD12345678	Genomic	Automated	Whole genome shotgun data
XM_123456	mRNA	Automated	Transcript products
XP_123456	Protein	Automated	Protein products
XR_123456	RNA	Automated	Transcript products
YP_123456	Protein	Auto. & Curated	Protein products
ZP_12345678	Protein	Automated	Protein products

J. Pevsner,  
<http://www.bioinfbook.org/index.php>

# Primární databáze

The screenshot displays the NCBI GenBank database interface for the gene **NP\_059797.1**. The main view shows a genomic map with a red bar representing the gene's location on the chromosome. A detailed popup window provides the following information:

- NP\_059797.1**
- NP\_059797.1: two-component VirA-like sensor kinase
- total range: NC\_002377.1 (145,694..148,183)
- total length: 2,490
- strand: plus
- protein product length: 829
- Links & Tools**
- GenBank View: [NC\\_002377.1 \(145,694..148,183\)](#), [NP\\_059797.1](#)
- FASTA View: [NC\\_002377.1 \(145,694..148,183\)](#), [NP\\_059797.1](#)
- BLAST Genomic: [NC\\_002377.1 \(145,694..148,183\)](#)
- Graphical View: [NP\\_059797.1](#)
- BLAST Protein: [NP\\_059797.1](#)
- BLINK Results: [NP\\_059797.1](#)

Below the popup, there are sections for **Bibliography** and **Related articles in PubMed**. The browser window shows the URL [www.ncbi.nlm.nih.gov/geo/1224316](http://www.ncbi.nlm.nih.gov/geo/1224316).

# Primární databáze

Display Settings: FASTA

Showing 2.49kb region from base 145694 to 148183.

### Agrobacterium tumefaciens plasmid Ti, complete sequence

NCBI Reference Sequence: NC\_002377.1

[GenBank](#) [Graphics](#)

```
>gi|10955016:145694-148183 Agrobacterium tumefaciens plasmid Ti, complete sequence
ATGAACGGAAGATATTCACCGACGCGGCAGGATTTAAGACAGCGCGAAGCCTTGGTCTATATGGGCC
TTATCGTTGCTGCAATGATTTTCGCGTTTCATGGCGGTTGCGTCTGGCAGGACAAATCGACTACCCAGGC
AATCTCAGCCAACACGATCGATTAACCGCGACAGCCCTCACTGACAGCGGATGACTCCGCGCTCAC
ACGGCACCGTGGCGAATACCGCCCATTTATCTCCAGGCTGGAGCTCTGGGAAGAAATCGAAAGATT
TGAAGCAATTTAGCAATCTCATATTTGAAGTGAAGCAATGCTGCTCACTGCTACGCGAGCTAGA
AGTGTCTAAATTCGGCTGACGCGCGGCTCGCCGCTTTGGTGGCAAAATGTACGCTGAAAGATTG
CTGGCCAGTTTCACTCGTGTGTTGAGCAGTCTCCGGAAGAGCCTCAACCGATCAGACTTTAGAAAAAC
CAACAGAATTGGTAGCATGATGCTCCAATTTCTTCGGCAACCAAGCCGGCTATTTCAATCGAGATCAG
CCTTGAAGTGAAGAGGCTCCAAAAACAACCGCGCTTGTATGAAGTCCCGTGGCATACTTGACCTGAA
GGTCCCAATTTATCGCTTTTGGCCAGAGTGAAGATCTGGTGAACATGATTCAGACGCTCTGACACCG
CAGAAATTCGGGATGCTGACGCGGAGTGTGGAGGTCTATAGCTTGAATAATGTAGAGGAGCGGAG
CGCAGTATCTTTCTGGTCCGCTTCAGTGGGTCTTGGCTCTACATCATCACTTGTCTATAGGCTA
CGCAAAAAACCGATTGGTTAGCGCGGCTTTAGATTACGAAGAGCTAATCAAGAGATCGGAGTATGTT
TTGAAGTGAAGCGGCCACCACTGCTCGCGCAAGCTGCATTCGTATTATTCAGCGCTTTTGGATGC
CGATACGTCGCGCTTAGCTTAGTGGACCATGACCGTAGAGGGCTGTGCAAAACATTCGTTGCAAAAC
CCAAAACTGTGGGACGACAGCGTGTACGGCAATAGTCTCTGTACCAAGCGGACGACGCGGCGA
CGGTATTCGCATCATCTGCGAAAAAATCGTACATTTGCCTCTCGAAATCCAGGTCTCTCGATACT
ACTGGCTCAAAATCCACAGATAAATAATTTGGCTTTGTTCACTGGTATCCAAAGCTATCGCCCTCGA
CCTTGCCAGGCGAAATTCAGCTTCTTGAAGTCCGACCGCTGCTCTGACTATATCGATGTTGCGG
GTAAGCAGACCGAATGCGACGTTTGGCCAGACGATTGGAGCATGCGCAACGCTTGAGGCACTTGGTAC
ACTTCCGCGCGGAATAGCACATGAATTAATAACATTTTGGCTCAATCTCGGGCAGCAGAAATAGCA
CAAACTCGGTCTCGAACATCTGTACCCGAAGATATATGACTATATCATTTCTGTCAGGCGACAGAG
CCATGCTCATATCGATCAGATCTTGAAGCTGAGCGGAAACAGGAGCGCATGATCAAGCCATTTAGTGT
CTCAGAGTGTGACCGAAATCGTCCCTTGTCTAGCTATGGCTTCCGCAAAACATCGAGCTTAGTTTC
AGATTTGATCAAAATGACAGCGGTGATCGAAGGAAGCCCGCTTGAAGTCAACAGGTACTAATTAACATCT
GCAAGATGCTTCCAGCCATGACTGCAATGGTCAATCGACATCATCATAGCCAAAGCTTTTTTACC
AGTTAAGAAAATTCGGCGCATGGTGTATGCCACCTGGCGACTATGTTCTCCTATCTATAGCGCAAT
GGTGGAGGCAATCCCGAGGCTGTGTACCCACATTTTGAACCTTCTTACGACAGAGCTCGCAACG
GTGGAACGGGTCTCGCCCTGCTCTGTGTCATGTTGATATCAGCGGCTTTGGCGGTTACATCGAGCTTAG
TTCAACTGTGGCATGGACGCGCTTGCATTTATCTCCCTCGCTTCTAAGGAACCCGTAATCCCA
GACAGTTTTTTCGGCCGCAATAAGGCACCGCTGGAAACGGGAGATTGTGGCACTTTGTGAGCCGATG
ACCTCCTGGGGAGGCGTATGAAACAAGATCGCCGCTTAGGATATGAGCGGTCGGTTTTCTGTAACCTT
TAATGAAATTCGGATTTGGATTTCAAAAGCAATGAAGCCGATCTGGTCAAGTGTGCAAGAGCTCTCTT
CCTGAAGATCAAGTCTTAATCCGTTGATTTAGTGTCAAGACCGCTCCATCATCATTTGGCGAAATG
ATCTCAAAATGACCCCTTCAAGGGAGGATGTGACCGGAGCTTTATCTCCGAAAGCGGATATCGTCCAG
AATATGGCGCATGCAATCTAACAAAATCAAGACGTAG
```




INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

<a href="#">EXPASY Home page</a>	<a href="#">Site Map</a>	<a href="#">Search ExPASy</a>	<a href="#">Contact us</a>	<a href="#">Swiss-Prot</a>	<a href="#">PROSITE</a>	<a href="#">Proteomics tools</a>			
Hosted by SIB Switzerland		Mirror sites:	<a href="#">Australia</a>	<a href="#">Bolivia</a>	<a href="#">Canada</a>	<a href="#">China</a>	<a href="#">Korea</a>	<a href="#">Taiwan</a>	<a href="#">USA</a>
Search		PROSITE	for		Go	Clear			

 ScanProsite

This program allows to scan a protein sequence (either from [Swiss-Prot](#) or [TrEMBL](#) or provided by the user) for the occurrence of patterns and profiles stored in the [PROSITE](#) database, or to search protein databases with a user-entered pattern ([Reference](#) / [Download ps\\_scan, the standalone version](#)). The program [PRATI](#) can be used to generate your own patterns. You may either:

- enter a PROSITE accession number or pattern to search the Swiss-Prot/TrEMBL and/or PDB databases with a pattern, **OR**
- enter a sequence or a Swiss-Prot/TrEMBL accession number to scan the sequence with all patterns, profiles and rules in PROSITE, **OR**
- fill in both fields to find all occurrences of a pattern or profile in a sequence.

Scan a protein for PROSITE matches	Search Swiss-Prot with a PROSITE entry
<p>Enter a Swiss-Prot/TrEMBL accession number (AC) (for example <b>P0130</b>) or a sequence identifier (ID) (for example <b>NOTC_DROME</b>), or a PDB identifier, or paste your own protein sequence in the box below:</p> <pre>MMVKVTKLYASPTVTPCVLAPLVVPECTWISNMTTTE BLVKEVASFTEDLRLSLVSEIENIGKPTVAKTHLSTIGLA RVIDEYITNNDTQPTFIQTQIALPLFVAVSTILQVQVSY ISRDGIMPSYIARNTSVAVFANSSNSRGGDTWYQTV DQLTGRLRNGNSTRSQSLDVTHTWQQAQSHNYTTPVGT ELGGEDMETLIQSVVLSYKGLVSLGFPFRTITVNLGL BLRHELIYMTKDTLVYSEISLNSFPFISGSIQFGRSE NSLWQICIPENCSSGVEVEIKRLRYQAPCSYIRVSGVPL</pre> <p>and specify which motifs to use:</p> <p>Scan <input checked="" type="checkbox"/> patterns <input checked="" type="checkbox"/> profiles <input checked="" type="checkbox"/> rules <a href="#">[User Manual]</a> (You may also specify a PROSITE entry in the box to the right)</p> <p><input type="checkbox"/> Exclude patterns with a high probability of occurrence</p> <p>Your e-mail (optional): <input type="text"/> (will send results by e-mail)</p> <p><input type="checkbox"/> plain text output</p> <p><input type="button" value="START THE SCAN"/> <input type="button" value="RESET"/></p>	<p>Enter a PROSITE accession number (for example <b>PS01253</b>), or type your pattern in <a href="#">PROSITE format</a>:</p> <p>(leave this box blank to scan a sequence with the entire PROSITE database)</p> <p>and specify your search limits:</p> <ul style="list-style-type: none"><li>• The <input checked="" type="checkbox"/> Swiss-Prot <input type="checkbox"/> TrEMBL <input type="checkbox"/> TrEMBLnew <input type="checkbox"/> PDB databases (You may also specify a protein in the box to the left) <input checked="" type="checkbox"/> including splice variants</li><li>• The following taxa: <input type="text"/> (see <a href="#">NEWT Taxonomy</a>; separate multiple taxa with a semicolon, e.g. <i>Homo sapiens; Drosophila</i>. Not available for PDB.)</li><li>• Sequences with at least <input type="text"/> hits</li><li>• At most <input type="text"/> matches</li></ul> <p>Advanced options: <input type="checkbox"/> FASTA output <input type="checkbox"/> retrieve complete sequences</p> <p>allow at most <input type="text"/> X sequence characters to match a conserved position in the pattern</p> <p><a href="#">match mode</a>: <input type="text"/> greedy, overlaps, no includes (for patterns, see <a href="#">help</a>)</p> <p><a href="#">randomize databases</a>: <input type="text"/> no (to test a pattern, see <a href="#">help</a>)</p>



# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

>[PDOC00003 PS00003](#) SULFATION Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].

```
571 - 585 nkeesstYeteians
```

>[PDOC00004 PS00004](#) CAMP\_PHOSPHO\_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

```
744 - 747 RRvT  
814 - 817 KRrS
```

>[PDOC00005 PS00005](#) PKC\_PHOSPHO\_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

```
148 - 150 SsR  
164 - 166 TgR  
171 - 173 StK  
219 - 221 SkK  
369 - 371 TrR  
460 - 462 SgK  
513 - 515 SgR  
585 - 587 S1R  
602 - 604 TgK  
652 - 654 TdK  
716 - 718 SpR  
726 - 728 SpK  
747 - 749 TeK  
794 - 796 SsR  
854 - 856 ScK  
864 - 866 StR  
868 - 870 SsR  
921 - 923 SpK  
957 - 959 SvR  
960 - 962 TgR  
974 - 976 TeK  
997 - 999 SrK  
1002 - 1004 TgK  
1018 - 1020 SgK  
1031 - 1033 TgR  
1119 - 1121 SkR
```

# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

Hits for all PROSITE (release 2020\_05) motifs on sequence USERSEQ1 :

found: 2 hits in 1 sequence

USERSEQ1 (1122 aa)

```
MMVKVTKLVASRPVIVFCVLAFLVWVFECIWISNWRTTENLVKEVASFTEDLRTSLVSEIENIGK
FTYAKTNLSTIGLARVIDSYITNNDTGFTEIQTQIAPLLFVAYSTILQVSQVSYISRDLMFYSIA
ESNTSVAVFANSSSSSRGDYTWYQTVDQLTGRNLNGNSTKQSLDVTHTIDWFQAQSNNTTAFV
GTSLGGEDNETLIQSVVSLYSKKGVLVSLGFFVKTLTEVLNLSNLHGEELVMWTKDGTVLVREGSLN
DSFFISNGSICFGRESNLSWSQCIPENCSSSGYEVEIKRLRYQAFCSVIEVSGVPLRYTLMFNKG
GATRIKHQAEKAKYQLIVVMIFLGFGWVFWFVFMQATRREMHRATLINQMEATQQAERKSMNK
SQAFANASHDIRGALAGMKGLIDICRDGVKPGSDVDTILNQVNVCCARDLVALLNSVLDMSKIESGK
MQLVEEDFNLSKLLDVIDFYHPVAMKKGVDVLDPHDGSVFKFSNVRGDSGRKQLLNLSNAV
KFTVDGHIARAWAQRPGSNSSVVLASYPKGVSKFVKSMFCNKKEESSSTYETEISNSIRNNANIME
FVFEVDDTGKGIPEMRKSVFENYVQVRETAQGHQGTGLGLGIVQSLVRLMGGELRITDKAMGEKG
TCFQENVLLTILESPVSDMKVRQEIEAGGDYVSTPNLGLTINTSLGSMNIRNLSPRFNCLSSS
PKQEGSRVLLKNEERRRVTEKYIKNLGKIVTVVEKWEHLSYALERLFGFSPPQSSMGRACSLSC
PSSRELFFIGMDGIDSRSQLPKRRSISFSAVLLVIDAKTGPFPELCDIVKQFRRLPHGISCKVV
WLNESSTRVSEKGDISCSRPLHGSRLMEVLKMLPEFGGTVLKEPPELQRESLLRHSFVAERSPKH
KVQEEGFPSSMFNKLKGRIMASTDSESETRVKSVRTGRKPIGNPEDEQETSXPSDDEFLRGKRVLV
VDDNFI SRKVATGKLGKMGVSEVEQCDSGKEALRLVTEGLTQREEQGSVDKLPFDYIFMDCQMPEM
DGYEATREIRKVEKSYGVRTPIIAVSGHDPGSEEAETIQAGMDAFLDKSLNQLANVIREIESKRH
```

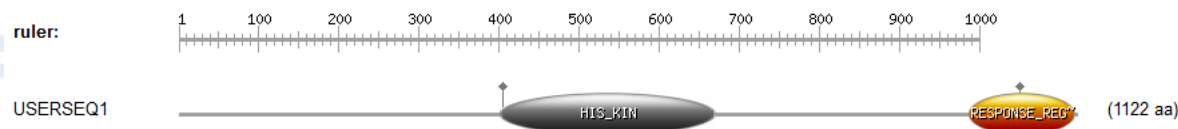
Legend:



Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter. For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/mydomains/>.

hits by profiles: [2 hits (by 2 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

Hits for all PROSITE (release 2020\_05) motifs on sequence USERSEQ1 :

found: 2 hits in 1 sequence

USERSEQ1 (1122 aa)

```
MMVKVTKLIVASRPVIVFCVLAFLVVFECIWNWRTITTEINLVKEVASFTEDLRTSLVSEIENIGK
FTYAKTNLSIIGLARVIDSYITNNDTGFTETIQIAPLLFVAYSTILQVSVSYISRDLMFYSYA
ESNTISVAVFANSSNSRSGDYTWYQTVDQLTGRNLGNSTKQSLDVIHTDWFQAAQSNNTAFV
GTSLGGEDNETLIQSVVSLYSKKGVLVSLGFPVKLTLEVLNLSLHGEELYMTKDGTVLVREGSLN
DSFFISNGSICFGRESNLSWSQCIPENCSSSGYEVEIKRLRYQAFCSVIEVSGVPLRYTLMFENK
GATRIKHQAEKAKYQLIVVMI FLGFGNPFVWFVFMQATREMHMRATLINQMEATQQAERKSMNK
SQAFANASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVNVCAKDLVALNLSVLDMSKIESGK
MQLVEEDFNLSKLLLEDVIDFYHPVAMKKGVDVLDPHDGSVFKFSNVRGDSGRKQLLNLSNAV
KFTVDGHIAVRAWAQRPGSNSSVVLASYPKGVSKFVKMFCKNKEESSTYETEISNIRNNANTME
FVFEVDITGKGI PMEMRKS VFENYVQVRETAQGHQGTGLGLGIVQSLVRLMGGEIRITDKAMGK
TCFQFNVLTLTLESPFVSDMKVRQIEAGGDYVSTPNLGLTINTSLGGSMNIRNLSPRFNNCLSSS
PKQGSRVLLLNKNEERRRVTKEYIKNLGIKVTVVEKWEHLSVALERLFGFSPQSSMGRACSLSC
PSSREL PFIGMDGIDSRSQLPKRRSISFSAVVLLVIDAKTGPFFELCDIVKQFRRGLPHGISCKV
WLNESSTRVSERGDISCSRPLHGSRLMEVLRMLPEFGGTVLKEPPTLQRESLLRHSFVAERSPKH
KVQEEGPPSMFNKLGKRMASSTDESETRVKSVRTGRKPIGNPEDEQETS KPSDDEFLRGRVLLV
VDDNFISRKVAIGKLLKMGVSEVEQCDSGKEALRLVTEGLTQREEQGSDKLPFDYIFMDCQMPM
DGYEATREIRKVEKSYGVRTPIIAVSGHDPGSEEARETIQAGMDAFLDKSLNQLANVIREIESKRH
```

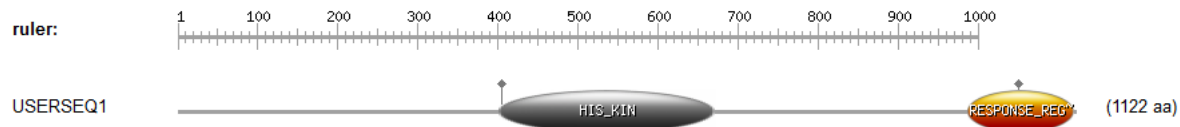
Legend:



Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter. For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/mydomains/>.

hits by profiles: [2 hits (by 2 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '\*' symbol represents deletions relative to the matching profile.



# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí)
- **PRINTS**, <http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>



PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a SWISS-PROT/EMBL composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. [References](#)

#### New:

- [SPRINT](#) - Search PRINTS-S (relational PRINTS)
- [prePRINTS](#) - Search PRINTS' automatic supplement
- [InterPro](#) - Search the integrated InterPro family database

#### Direct PRINTS access:

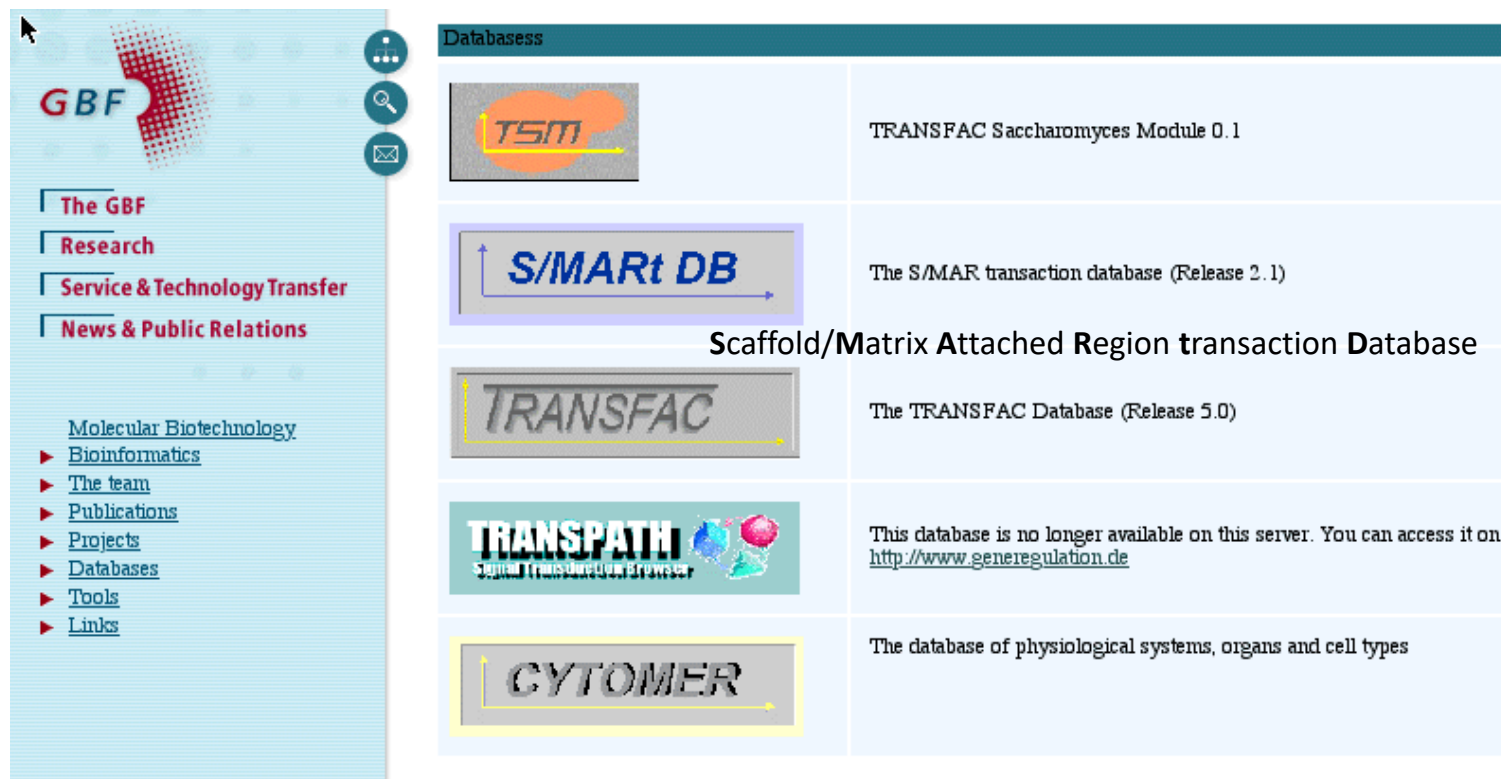
- [By accession number](#)
- [By PRINTS code](#)
- [By database code](#)
- [By text](#)
- [By sequence](#)
- [By title](#)
- [By number of motifs](#)
- [By author](#)
- [By query language](#)

#### PRINTS search:





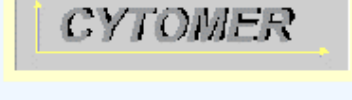
- [Search PRINTS with NEW FingerPRINTScan](#)
- [FPScan](#)
- [GRAPHScan](#)
- [MULScan](#)
- FingerPRINTScan binaries and source are available: [contact.scordis@bioinf.man.ac.uk](mailto:contact.scordis@bioinf.man.ac.uk)

# Sekundární databáze

- **TRANSFAC** <http://www.gene-regulation.com/>



The screenshot shows the website interface for TRANSFAC. On the left is a navigation menu for GBF (Gene Bioinformatics Foundation) with categories like 'The GBF', 'Research', 'Service & Technology Transfer', and 'News & Public Relations'. Below these are links for 'Molecular Biotechnology', 'Bioinformatics', 'The team', 'Publications', 'Projects', 'Databases', 'Tools', and 'Links'. The main content area is titled 'Databases' and contains a table of database entries:

Database Logo	Description
	TRANSFAC Saccharomyces Module 0.1
	The S/MAR transaction database (Release 2.1) <b>Scaffold/Matrix Attached Region transaction Database</b>
	The TRANSFAC Database (Release 5.0)
	This database is no longer available on this server. You can access it on <a href="http://www.generegulation.de">http://www.generegulation.de</a>
	The database of physiological systems, organs and cell types

# Strukturální databáze

- PDB <http://www.rcsb.org/pdb/>

The screenshot shows the PDB website with the following elements:

- Navigation Links:** [DEPOSIT data](#), [DOWNLOAD files](#), [browse LINKS](#), [BETA TEST new features](#), [BETA mmCIF files](#)
- Current Holdings:** 19623 Structures, Last Update: 30-Dec-2002, PDB Statistics
- Search the Archive:** Includes a search box, "Find a structure" button, and options for "query by PDB id only", "match exact word", and "remove sequence homologues".
- PDB Mirrors:** Lists various international mirrors such as San Diego Supercomputer Center, Rutgers University, etc.
- News:** A section for "23-Dec-2002 Happy Holidays from the PDB!" with a small image of lit candles.

# Strukturální databáze

- **PDB** <http://www.rcsb.org/pdb/>

Structure Explorer - 1P5Y

**RCSB**  
**PDB**  
PROTEIN DATA BANK

## Structure Explorer - 1P5Y

*Title* The Structures Of Host Range Controlling Regions Of The Capsids Of Canine and Feline Parvoviruses and Mutants  
*Classification* Virus/Viral Protein  
*Compound* Mol. Id: 1; Molecule: Coat Protein Vp2; Chain: A; Fragment: Sequence Database Residues 190-737; Engineered: Yes; Mutation: Yes  
*Exp. Method* X-ray Diffraction



[View Structure](#)

[Summary Information](#)

[View Structure](#)

[Download/Display File](#)

[Structural Neighbors](#)

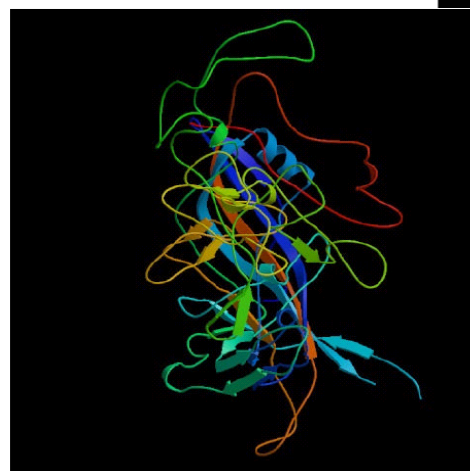
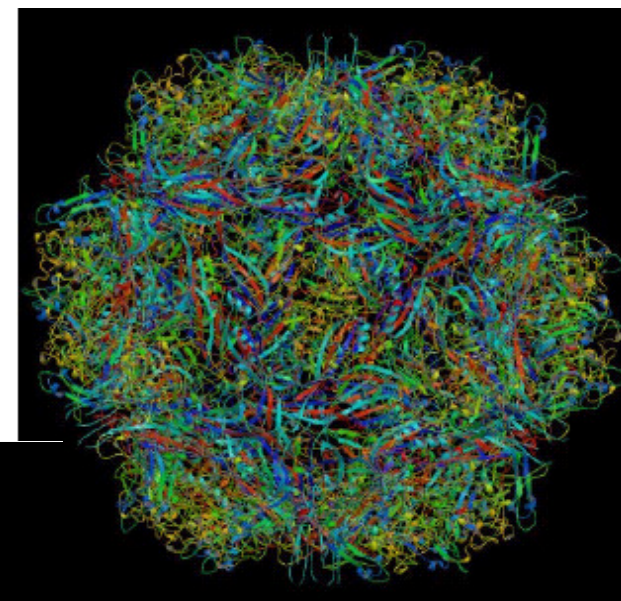
[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

Explore

[SearchLite](#) [SearchFields](#)

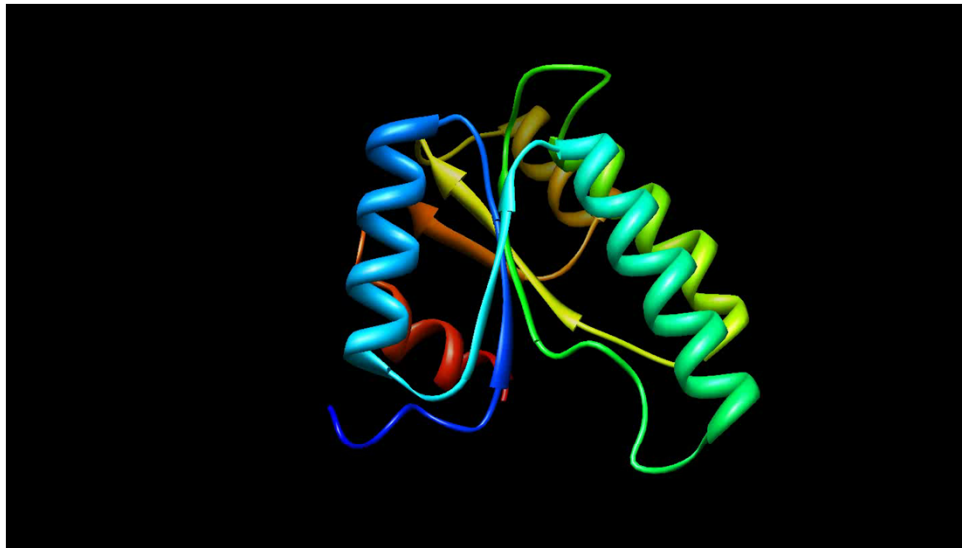


<http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=;pid=173561064349344&bio=1&opt=show&size=500>

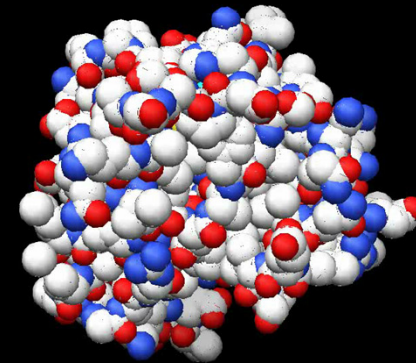
12/29/2003

# Strukturální databáze

- **PDB** <http://www.rcsb.org/pdb/>



Pekárová et al., *Plant Journal* (2011)





# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje

# Genomové zdroje

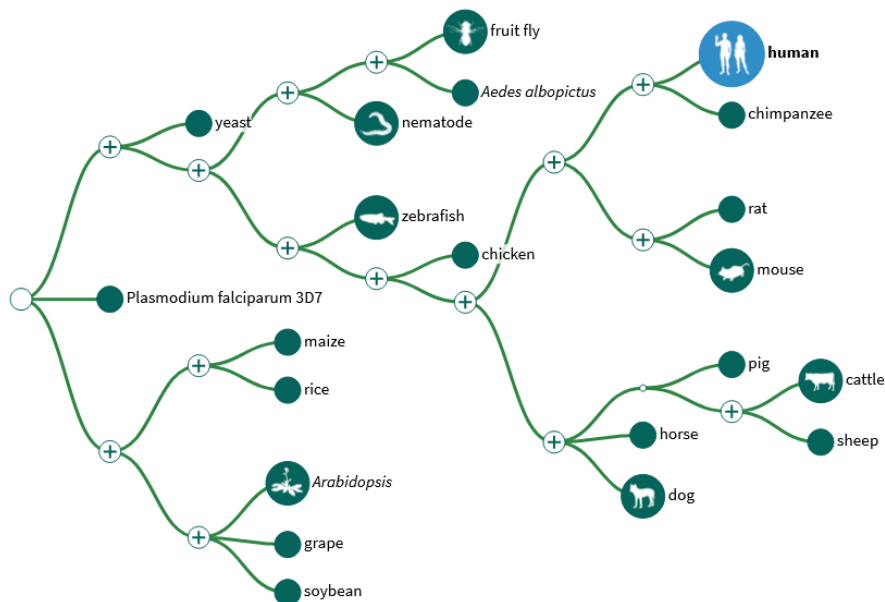
- **NCBI Genome Data Viewer** <https://www.ncbi.nlm.nih.gov/genome/gdv/>

## Genome Data Viewer

GDV is a genome browser supporting the exploration and analysis of more than 920 eukaryotic RefSeq genome assemblies. ⓘ

Select organism

Homo sapiens (human)



### Homo sapiens (human) genome

Search in genome

Location, gene or phenotype

Examples: TP53, chr17:7667000-7689000, rs334, DNA repair

Assembly

GRCh38.p13

[Browse genome](#)

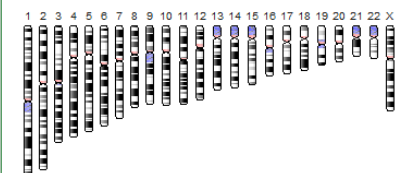
[BLAST genome](#)

#### Assembly details

**Name** GRCh38.p13  
**RefSeq accession** [GCF\\_000001405.39](#)  
**GenBank accession** [GCA\\_000001405.28](#)  
**Download via FTP** [RefSeq](#), [GenBank](#)  
**Submitter** [Genome Reference Consortium](#)  
**Level** Chromosome  
**Category** Reference genome

#### Annotation details

**Annotation Release** [109](#)  
**Release date** 2020-08-17



# Genomové zdroje

## □ Genome Browser Gateway <https://genome.ucsc.edu/>

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).  
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position	search term
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570	enter position, gene symbol or search terms

[Click here to reset](#) the browser user interface settings to their defaults.

[track search](#) [add custom tracks](#) [track hubs](#) [configure tracks and display](#)

[Add your own custom tracks](#)

### Human Genome Browser - hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the [Genome Reference Consortium](#). For more information about this assembly, see [GRCh37](#) in the NCBI Assembly database.

#### Sample position queries

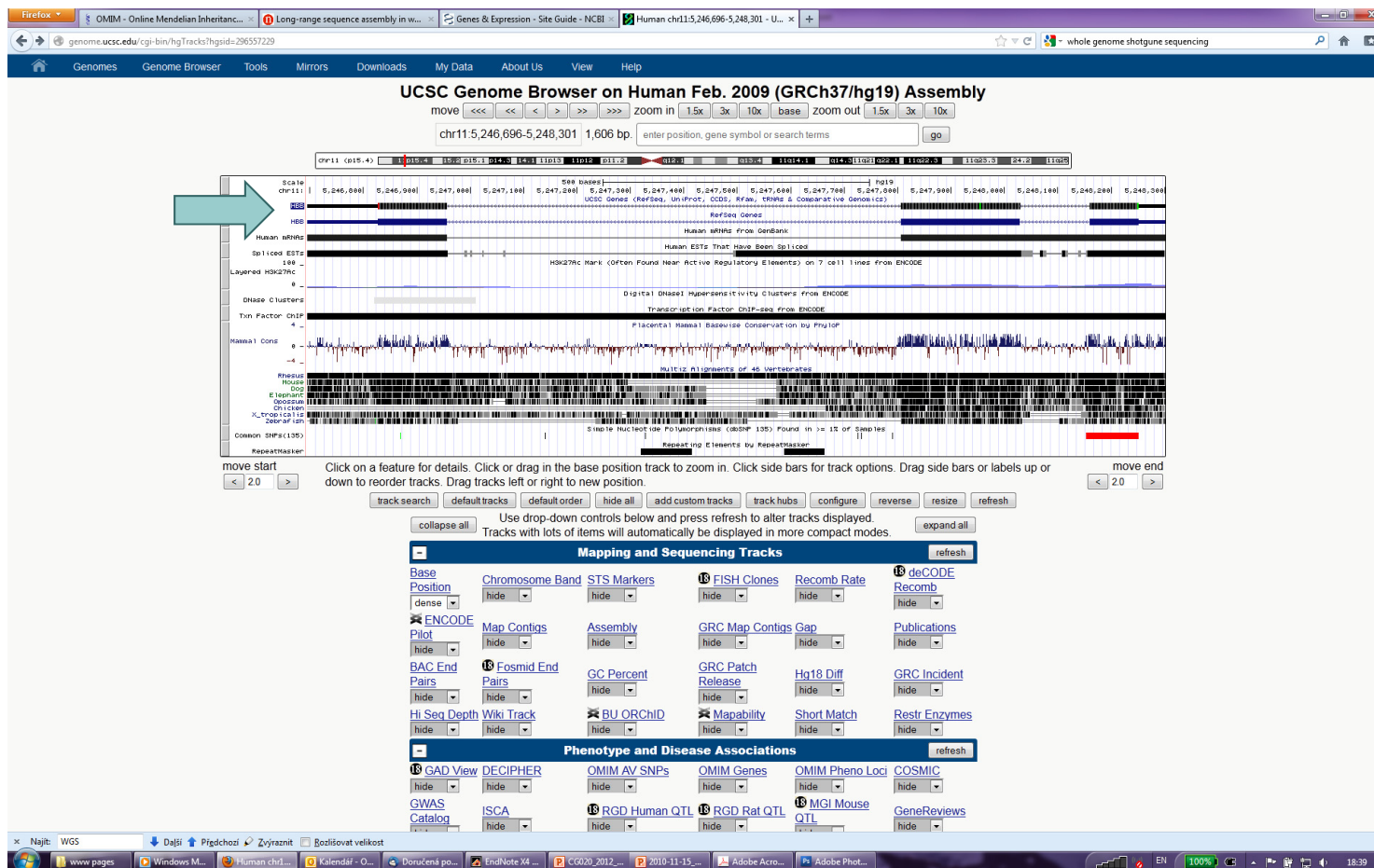
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the [User's Guide](#) for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gI000212	Displays all of the unplaced contig gi000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175 15q11;15q13 rs1042522;rs1800370	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17
AC008101	Displays region of clone with GenBank accession AC008101
AF083811	Displays region of mRNA with GenBank accession number AF083811
PRNP	Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs
homeobox caudal	Lists mRNAs for caudal homeobox genes
zinc finger	Lists many zinc finger mRNAs
kruppel zinc finger	Lists only kruppel-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease
zahler	Lists mRNAs deposited by scientist named Zahler
Evans, J.E.	Lists mRNAs deposited by co-author J.E. Evans

**UCSC**  
Homo sapiens  
(Graphic courtesy of [NCBI](#))

# Genomové zdroje

- Human Genome Browser <http://genome.ucsc.edu/cgi-bin/hgGateway>



# Genomové zdroje

## □ Human Genome Browser <http://genome.ucsc.edu/cgi-bin/hgGateway>

**Human Gene HBB (uc001mae.1) Description and Page Index**

**Description:** Homo sapiens hemoglobin, beta (HBB), mRNA.

**RefSeq Summary (NM\_000518):** The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3' [provided by RefSeq, Jul 2008]. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. ##RefSeq-Attributes-START##

Transcript\_exon\_combination\_evidence :: V00497.1, BU659180.1 [ECO:0000332] ##RefSeq-Attributes-END##

**Transcription Chromosome:** chr11 **Strand:** - **Size:** 1,606 **Start:** 5,246,695 **End:** 5,248,301 **Exon Count:** 3

**Coding Size:** 1,424 **Start:** 5,246,827 **End:** 5,248,251 **Exon Count:** 3

<b>Page Index</b>	Sequence and Links	UniProtKB Comments	Genetic Associations	CTD	Microarray
RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Pathways
Other Names	GeneReviews	Model Information	Methods		

Data last updated: 2011-12-21

**Sequence and Links to Tools and Databases**

Genomic Sequence (chr11:5,246,696-5,248,301)	mRNA (may differ from genome)	Protein (147 aa)			
Gene Sorter	Genome Browser	Protein FASTA	VisiGene	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	H-INV	HGNC	HPRD	Jackson Lab	MOPED
OMIM	PubMed	Reactome	Stanford SOURCE	Treefam	UniProtKB
Wikipedia					

**Comments and Description Text from UniProtKB**

**ID:** HBB\_HUMAN

**DESCRIPTION:** RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain; Contains: RecName: Full=LVV-hemorphin-7;

**FUNCTION:** Involved in oxygen transport from the lung to the various peripheral tissues.

**FUNCTION:** LVV-hemorphin-7 potentiates the activity of bradykinin, causing a decrease in blood pressure.

**SUBUNIT:** Helotetramer of two alpha chains and two beta chains in adult hemoglobin A (HbA).

**INTERACTION:** P69905:HBA2; NbExp=19; IntAct=EBI-715554; EBI-714680;

**TISSUE SPECIFICITY:** Red blood cells.

**PTM:** Glucose reacts non-enzymatically with the N-terminus of the beta chain to form a stable ketoamine linkage. This takes place slowly and continuously throughout the 120-day life span of the red blood cell. The rate of glycation is increased in patients with diabetes mellitus.

**PTM:** S-nitrosylated; a nitric oxide group is first bound to Fe(2+) and then transferred to Cys-94 to allow capture of O(2).

**PTM:** Acetylated on Lys-60, Lys-83 and Lys-145 upon aspirin exposure. PubMed:16916647 reports the identification of HBB acetylated on Lys-145 in the cytosolic fraction of HeLa cells. This may have resulted from contamination of the sample.

**MASS SPECTROMETRY:** Mass=1310; Method=FAB; Range=33-42; Source=PubMed:1575724;

**DISEASE:** Defects in HBB may be a cause of Heinz body anemias (HEIBAN) [MIM:140700]. This is a form of non-spherocytic hemolytic anemia of Dacie type 1. After splenectomy, which has little benefit, basophilic inclusions called Heinz bodies are demonstrable in the erythrocytes. Before splenectomy, diffuse or punctate basophilia may be evident. Most of these cases are probably instances of hemoglobinopathy. The hemoglobin demonstrates heat lability. Heinz bodies are observed also with the Ivemark syndrome (asplenia with cardiovascular anomalies) and with glutathione peroxidase deficiency.

**DISEASE:** Defects in HBB are the cause of beta-thalassemia (B-THAL) [MIM:604131]. A form of thalassemia. Thalassemias are common monogenic diseases occurring mostly in Mediterranean and Southeast Asian populations. The hallmark of beta-thalassemia is an imbalance in globin-chain production in the adult HbA molecule. Absence of beta chain causes beta(0)-thalassemia, while reduced amounts of detectable beta globin causes beta(+)-thalassemia. In the severe forms of beta-thalassemia, the excess alpha globin chains accumulate in the developing erythroid precursors in the marrow. Their deposition leads to a vast increase in erythroid apoptosis that in turn causes ineffective erythropoiesis and severe microcytic hypochromic anemia. Clinically, beta-thalassemia is divided into thalassemia major which is transfusion dependent, thalassemia intermedia (of intermediate severity), and thalassemia minor that is asymptomatic.

**DISEASE:** Defects in HBB are the cause of sickle cell anemia (SKCA) [MIM:603903]; also known as sickle cell disease. Sickle cell anemia is characterized by abnormally shaped red cells resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resembles a sickle. These stiffer red blood cells can lead to microvascular occlusion thus cutting off the blood supply to nearby tissues.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

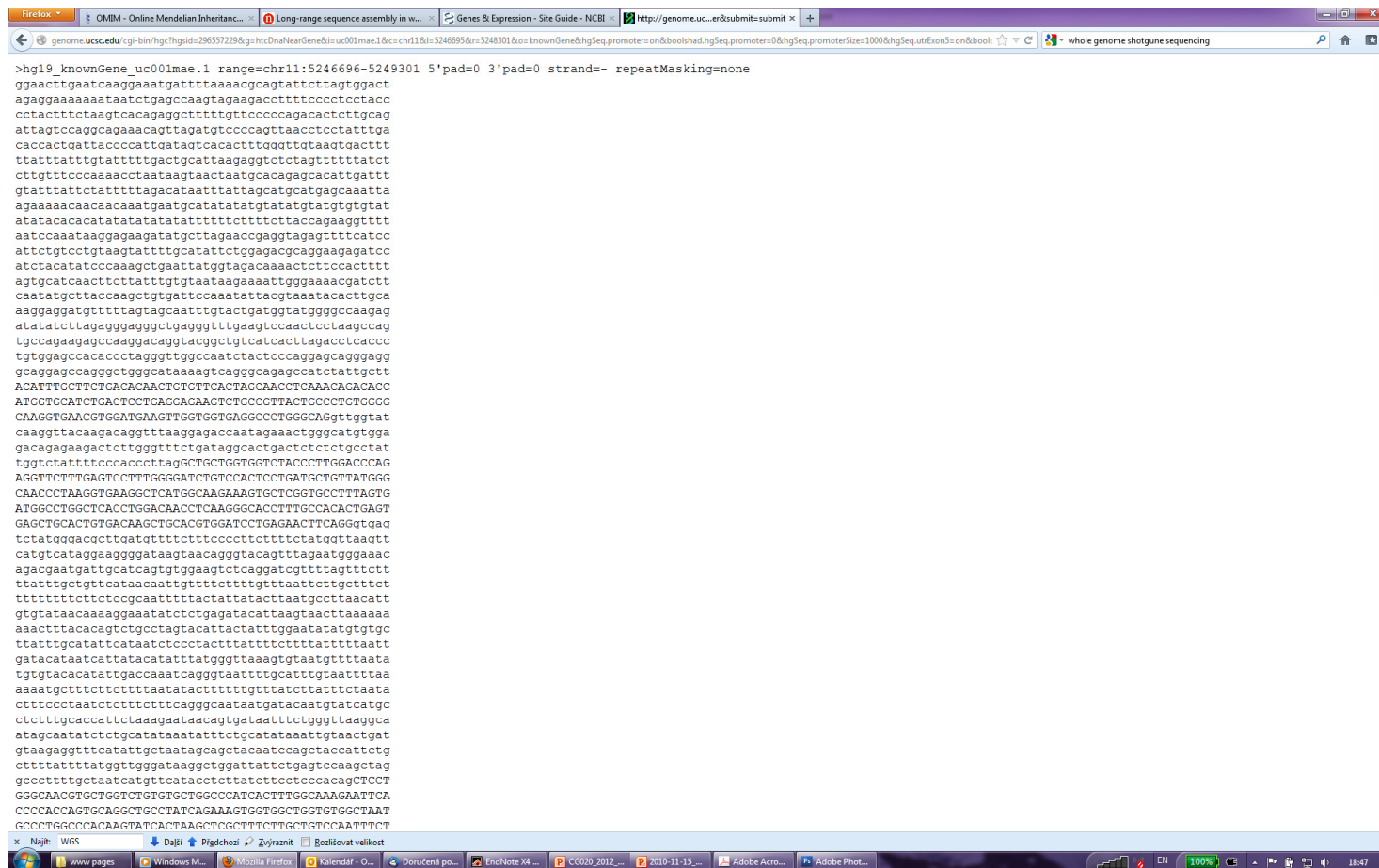
# Genomové zdroje

- **Human Genome Browser** <http://genome.ucsc.edu/cgi-bin/hgGateway>

The screenshot displays the 'Genomic Sequence Near Gene' interface. It includes a navigation menu at the top with links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, About Us, and Help. The main content area is titled 'Get Genomic Sequence Near Gene' and contains a note about using the 'Table Browser' for multiple features. Below this, there are two sections of options: 'Sequence Retrieval Region Options' and 'Sequence Formatting Options'. The 'Sequence Retrieval Region Options' section has several checked checkboxes for 'Promoter/Upstream by 1000 bases', '5' UTR Exons', 'CDS Exons', '3' UTR Exons', and 'Introns'. There are also input fields for 'Downstream by 1000 bases', 'One FASTA record per gene', and 'Split UTR and CDS parts of an exon into separate FASTA records'. The 'Sequence Formatting Options' section has radio buttons for 'Exons in upper case, everything else in lower case', 'CDS in upper case, UTR in lower case', 'All upper case', and 'All lower case'. There are also radio buttons for 'Mask repeats' with options 'to lower case' and 'to N'. A 'submit' button is located at the bottom of the options section.

# Genomové zdroje

- **Human Genome Browser** <http://genome.ucsc.edu/cgi-bin/hgGateway>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# Genomové zdroje

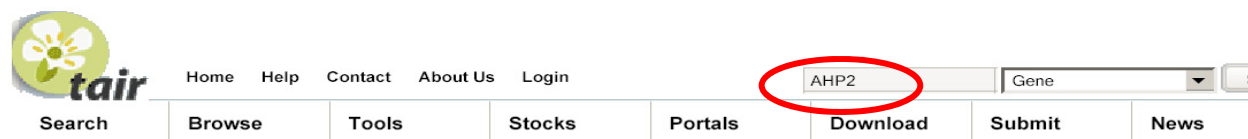
- The Arabidopsis Information Resource (TAIR) <http://www.arabidopsis.org>

The screenshot shows the TAIR website homepage. The browser window title is "OMIM - Online Mendelian Inheritance... | TAIR - Home Page". The address bar shows "www.arabidopsis.org". The website features a search bar at the top right and a navigation menu with options like Home, Help, Contact, About Us, and Login/Register. Below the navigation, there are tabs for Search, Browse, Tools, Portals, Download, Submit, News, and ABRC Stocks. The main content area is titled "The Arabidopsis Information Resource" and includes a detailed description of the resource, a "Breaking News" section with links to subscribe to a news feed, follow on Twitter, and join a Facebook group, and a "2012 MASC Report Now Available" section. There is also a "New Protein Chip and Cell Cultures at ABRC" section and a "Share Your Education Resources" section. A large banner at the bottom of the main content area promotes a new online submission form, with a "Click here" button and text describing the form's capabilities. The browser's status bar at the bottom shows the system tray with various icons and the time 18:52.



# Genomové zdroje

- The Arabidopsis Information Resource (TAIR) <http://www.arabidopsis.org>



## The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a [database](#) of genetic and [molecular biology data](#) for the model higher plant *Arabidopsis thaliana*. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of *Arabidopsis thaliana* and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

## Breaking News

### Data Updates Suspended

[October 19, 2006]  
Some TAIR data updates, including loading of new ABRC stocks, will be suspended from Oct 20-Nov 17 while we move our servers.

### New Phenotype Search Option

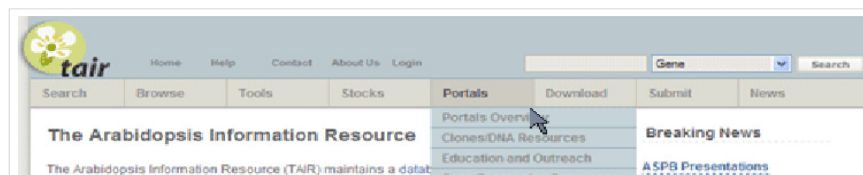
[October 15, 2006]  
Search for **genes**, **germplasms**, and **polymorphisms** using associated phenotype, and see improved phenotype data display in results and detail pages.

### ASPB Presentations

[August 15, 2006]  
Following heavy demand, the TAIR workshop presentations given at the ASPB meeting in Boston have been made available from the TAIR website for download.

## The NEW arabidopsis.org

We've added new dropdown headers and left navigation bars and reorganized our web pages to make it easier to locate information and resources in TAIR. Please contact us if you experience any problems with our new site.

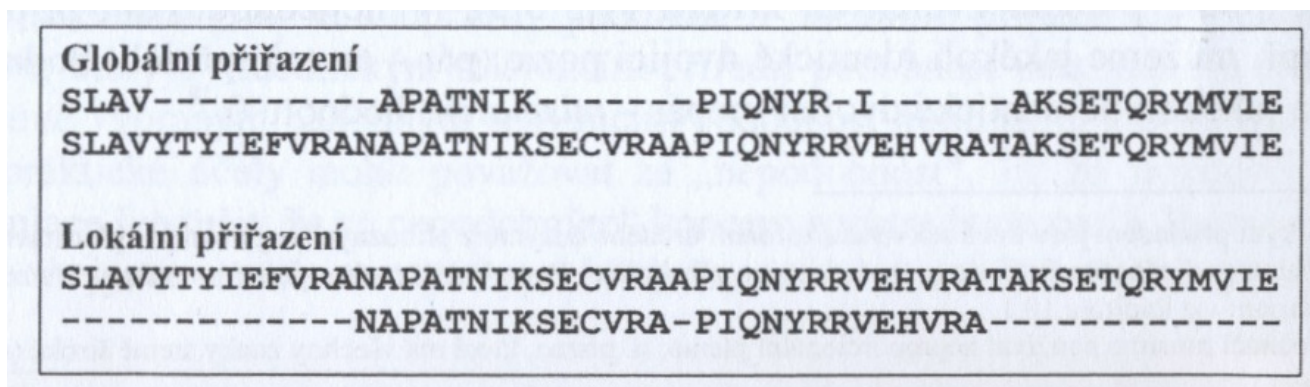


# Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií

# Analytické nástroje

## □ Globální vs. lokální přiřazení

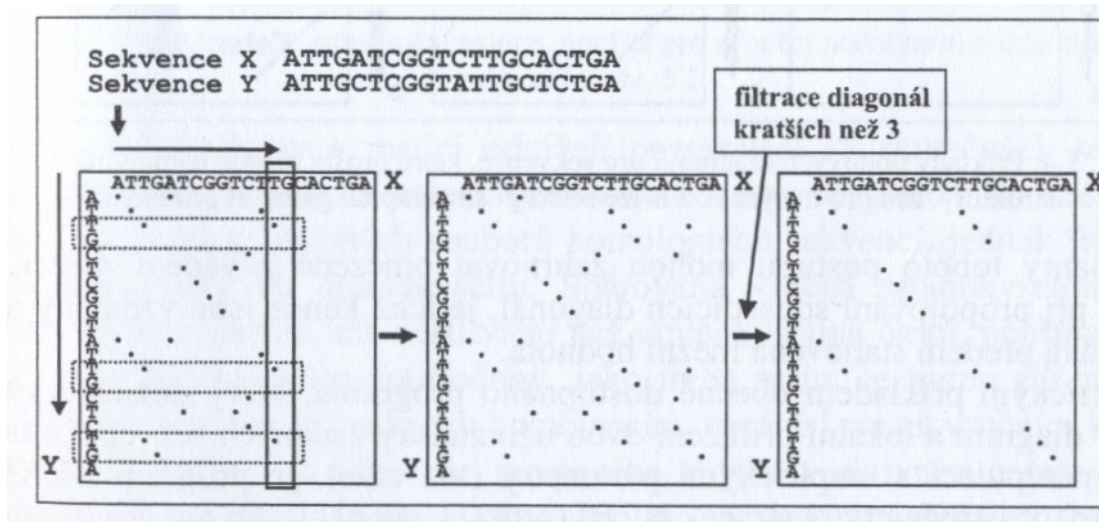


Cvrčková, Úvod do praktické bioinformatiky

- **Globální přiřazení** pouze u sekvencí, které jsou si **podobné a podobné délky** (za cenu vnášení mezer do jedné nebo obou sekvencí)
- Globální přiřazení se používá především v případě **mnohačetného přiřazování** (CLUSTALW, viz dále)
- **Lokální přiřazení** umožní identifikaci a srovnání i v případě porovnávání pouze **úseků sekvencí** s významnou mírou podobnosti, např. i při záměně pořadí proteinových domén během evoluce

# Analytické nástroje

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)

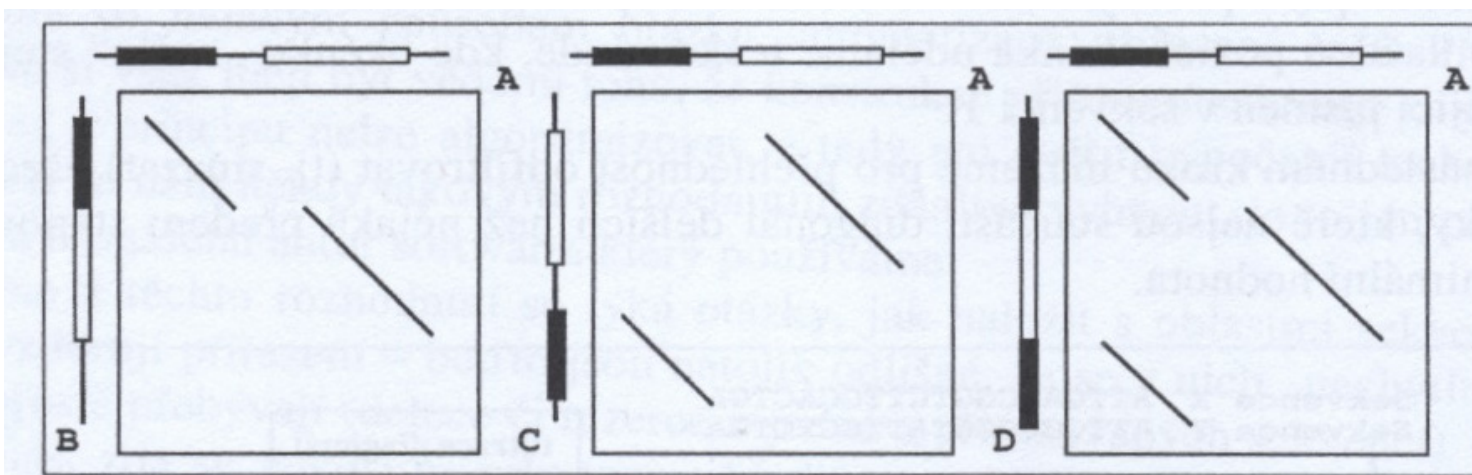


Cvrčková, Úvod do praktické bioinformatiky

- vynesení sekvencí proti sobě
- identifikace shody v okně o dané velikosti (např. 2 bp)
- „odfiltrování“ diagonál o délce menší než je mezní hodnota (threshold)

# Analytické nástroje

- příklady srovnání sekvencí pomocí bodového diagramu



Cvrčková, Úvod do praktické bioinformatiky

- globálně lze srovnávat pouze sekvence A, B
- ostatní sekvence prošly během evoluce **záměnou domén** a je nutné je porovnávat **lokálně**
- **bodový diagram** lze získat pomocí srovnávacího programu **BLAST2** (viz dále)

# Analytické nástroje

- **BLAST** <http://ncbi.nlm.nih.gov/BLAST/>

NCBI *nucleotide-nucleotide* **BLAST**  
Nucleotide Protein Translations Retrieve results for an RID

[Search](#)

```
aacacccg  
acaccatcat cattatcacc atcgttttgg ggcgatggtg tgtgggtcca  
gogtattaat  
ataattaatt tattccacat gagatatgat atgatatact atgtattttt  
tgtttttttt  
ttatttgtaa acctttaata taacaagaac tacaaaaaat gaaaa
```

[Set subsequence](#) From:  To:

[Choose database](#)

Now: **BLAST!** or **Reset query** **Reset all**

# BLAST

Basic Local Alignment Search Tool

- Velikost vyhledávacího slova (word size): 10-11 bp, resp. 2-3 aa
  - Primární podobnosti (seed matches)
  - Rozšiřování oblasti homologie doprava i doleva
- Hodnocení homologie pomocí matice PAM (Point Accepted Mutation) nebo BLOSUM (BLOCKS Substitution Matrix)
- Zobrazení výsledků

	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

hodnota nepáru G-A

hodnota páru G-G

Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250

C	12	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
C	12	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

# BLAST

## Basic Local Alignment Search Tool



- „expectancy value“ udává předpokládaný počet sekvencí se stejnou nebo lepší podobností při vyhledávání ve stejně velké databázi složené z náhodných sekvencí
- výsledek udává frakci totožných a u proteinů i podobných pozic, příp. počet vložených mezer



# Primární databáze

NC\_002377.1: 145K..148K (2.9Kbp)

Genes

**NP\_059797.1**

NP\_059797.1: two-component VirA-like sensor kinase  
total range: NC\_002377.1 (145,694..148,183)  
total length: 2,490  
strand: plus  
protein product length: 829

**Links & Tools**

GenBank View: [NC\\_002377.1 \(145,694..148,183\)](#), [NP\\_059797.1](#)  
FASTA View: [NC\\_002377.1 \(145,694..148,183\)](#), [NP\\_059797.1](#)  
BLAST Genomic: [NC\\_002377.1 \(145,694..148,183\)](#)  
Graphical View: [NP\\_059797.1](#)  
BLAST Protein: [NP\\_059797.1](#)  
BLINK Results: [NP\\_059797.1](#)

**Bibliography**

**Related articles in PubMed**

# BLAST

## Basic Local Alignment Search Tool

BLINK *precomputed BLAST*

Home Taxonomy Report Multiple Alignment Blast Help

My NCBI [Sign In] [Register]

Pre-computed BLAST results for: [gi|16119781|ref|NP\\_396486.1](#) two component sensor kinase [Agrobacterium tumefaciens str. C58]

Matching gis: [15163423:20141871:1019660](#)

Total (score > 100) : 147086 hits in 146754 proteins in 6309 species

Selected: 147086 hits in 146754 proteins in 6309 species Filter: **Min Score: 100** |

Other views (Reports): [Taxonomy report](#) [Multiple Alignment](#) [Blast](#)

[Reset all filters](#)

Choose Display Options

1203 Archaea 138285 Bacteria 13 Metazoa 1349 Fungi 554 Plants 6 Viruses 5676 The Others [reset selection](#)

Results: 1 - 100 [Next Page](#) [Last](#)

% hits [reset selection](#)

833 aa

blink

SCORE	ACCESSION	Length	Protein Description
<b>Conserved Domain Database hits</b>			
4166	<a href="#">AAK90927</a>	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]
4166	<a href="#">P18540</a>	833	RecName: Full=Wide host range virA protein; Short=WHR virA
4166	<a href="#">AAA79282</a>	833	virA [Plasmid pTiC58]
4159	<a href="#">NP_053380</a>	833	hypothetical protein pTi-SAKURA_p142 [Agrobacterium tumefaciens]
4159	<a href="#">BAA87765</a>	833	tiorf140 [Agrobacterium tumefaciens]
4153	<a href="#">AAA91590</a>	833	virA [Plasmid Ti]
4153	<a href="#">gi 737127</a>	833	virA protein
4153	<a href="#">CAA34777</a>	833	91.3 kDa protein [Agrobacterium tumefaciens]
3800	<a href="#">CAA35780</a>	829	virA [Agrobacterium rhizogenes]
3718	<a href="#">gi 227240</a>	869	virA gene
3148	<a href="#">AAA88643</a>	829	virA [Plasmid Ti]

# BLAST

## Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
  - vyhledávání podle zdroje (organismu) sekvencí, např. známých genomů mikroorganismů
  - **BLASTP**
    - vyhledávání podobnosti k **proteinu** v **databázi proteinových sekvencí**
  - **BLASTN**
    - vyhledávání podobnosti k **nukleotidové sekvenci** v **databázi nukleotidových sekvencí**
    - další varianty jako např. **MEGABLAST** pro identifikaci totožných nebo velice podobných sekvencí (vyhledává dlouhé podobné úseky nukl. sekvencí)
  - **BLASTX**
    - vyhledávání **podobnosti k proteinu** v **databázi nukleotidových sekvencí přeložených do sekvence aa**



# BLAST

## Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
  - **TBLASTN**
    - vyhledávání k sekvenci nukleotidů přeložené do sekvence aa v databázi proteinů
  - **TBLASTX**
    - vyhledávání k sekvenci nukleotidů přeložené do sekvence aa v databázi nukleotidových sekvencí přeložených do sekvence aa

# BLAST

## Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
  - **PSI-BLAST** (**P**osition-**S**pecific **I**terated **B**LAST)
    - Prvním krokem je standardní BLAST, při kterém PSI-BLAST identifikuje skupinu podobných sekvencí s E hodnotou lepší než minimální hodnota (standardně 0,005)
    - PSI-BLAST vytváří pro každé přiřazení tzv. **PSSM** (**P**osition **S**pecific **S**ubstitution **M**atrix)
    - PSSM matice zohledňuje výskyt jedné aminokyseliny ve stejné pozici se zvýšenou frekvencí u sekvencí identifikovaných jako podobné v prvním kole pomocí BLAST, což může znamenat funkční konzervovanost



# BLAST

## Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
  - **PHI-BLAST** (Pattern-Hit Initiated **BLAST**)
    - Určen k identifikaci specifické sekvence, např. motivu (pattern) v sekvenci podobných proteinových sekvencí
    - Sekvenci motivu je třeba vložit pomocí **speciálního syntaxu**
      - [LVIMF] znamená buď Leu, Val, Ile, Met nebo Phe
      - - je oddělovník (neznamená nic)
      - x(5) znamená 5 jakýchkoliv aminokyselin
      - x(3, 5) znamená 3 až 5 jakýchkoliv aminokyselin

# BLAST

## Specializované verze

### □ Příklad vyhledávání pomocí PHI-BLAST

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase  
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPEPGPDR  
VADAKGDESEEEDEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQBACKDILLF  
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA  
LMYNTPRAATIVA TSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIgek  
IYKDGERIITQGEKADSFYIESGEVSI LIRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYEEQLVKMFGSSVDLGNLQ
```

```
[LIVMF] -G-E-x- [GAS] - [LIVM] -x(5,11) -R- [STAQ] -A-x- [LIVMA] -x- [STACV] .
```

# Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst....



# Analytické nástroje

- **Biology Workbench** <http://workbench.sdsc.edu/>

**Biology WorkBench**  
click here to toggle between menus and buttons  
**WE Moved!** <http://workbench.sdsc.edu/>  
Version 3.2

Session Tools Protein Tools **Nucleic Tools** Alignment Tools Structure Tools (Alpha)

beta-glucosidase

GBPLN:804655 **Hordeum vulgare L. beta-glucosidase (BGQ60) gene, complete cds.**  
 GBPLN:170248 **Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds.**

Select All Deselect All Ndjinn BATCH Add Edit Delete Copy View Download ViewRecords  
BL2SEQ BL2SEQX BLASTN BLASTX TBLASTX FASTA FASTX FASTY SSEARCH CLUSTALW  
CLUSTALWPROF ALIGN LALIGN LFASTA PATTERNMATCHDB PATTERNMATCH TACG PRIMER3  
NASTATS BESTSCOR PFSCAN PRIMERCHECK PRIMERTM SIXFRAME REVCOMP RANDSEQ

Copyright (C) 1999, Board of Trustees of the University of Illinois.  
SDSC

# Analytické nástroje

- **Biology Workbench** <http://workbench.sdsc.edu/>

View  
View Nucleic Sequence(s)

Format  Case

[Download/view all sequences in text format](#)

[\[NEXT\]](#) [\[BOTTOM\]](#)

Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds.  
GBPLN:170248, 4699 bp

>170248  
GAGCTCCCTTGGGGGGCAAGGGCAAAAACTTTTGCTAAATGGAAAAATATTATACCAAGTGTGTAATA  
GTTACTCAATTTGAATTAACAAAGGGGCAAAATTTGACTATTTTGCCCTTATATCTTTTGGTCACAAAAAC  
ATAAAATATCCCATCCGAAATTC AAATGGTCCATTATCGGCCAAGTAGCTTTCTTTAATTATAGTTAGTT  
GACAAAACACTATCAAGATATCATTATATAATAATAACATCAAGTCCATCATCTTAGCTGCCTCCTCA  
GTAGAGCCGCCAGTAAAAAAGACCGATCAAAATAAAGCCGCCATTAATAAATGAATTTTAGGACTCTC  
GATTGGCACGTAAGTGCCAAAACCTTTCCAATACCTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC  
CAGATATGGGATATTTCTAAGTTTTATCTCTAAATTTACATCTCAACTAATATTAAGAAATTAACAGGTA  
CAGCAAAATCATAAAATTTTCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG  
TCTGCATGCCATATTCCTAAGGGGTCGTTTGGTACAAAGAAATAATAATAAATTTTCGGGATAGAATTT  
GAGATTGCATTTATCTTGTGTTTTAATTATAAGTATTAGCTAATTTTCAGAAATAAATTTTACTAAAATAG  
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCATAGCCACTCACATAGAATATCC  
TTATTTATCTCACTATTTTACC AAATGATCGGTTAGTCTTTCATGAGAATCCAGTATCCTCAATAAATGCA  
GTAAGAAGTTAGAAAATTTTCAATTAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG  
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAACGGTT  
CGAGACACTCTTATGGAAGGCGTTTGTCTTCAAAGTAGATTCTCATTCAATTGCTCTGGTGCATAGCAAAA  
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA  
GAGAATTTCAAATCTCTCAACTACTTTTAAGGGAATTCAAAATACGACC AATATTTATTACTTACTTAC  
TTATAGTTAAATGATATGAATTTTTATTTAAATTTGAATTGAAAAATTAATAATTACTTGAATTAATATAA

# Analytické nástroje

- **Biology Workbench** <http://workbench.sdsc.edu/>

## Regex pattern:

```
ctt. {1,32}ctt
```

0 sequences were searched

1 match was found

Matches are indicated in blue

```
>170248
```

```
GAGCTCCCTTGGGGGGCAGGGGCAAACTTTTGGCTAAATGGAAAAATATTATACCAAGTGTGTTGTAATA  
GTTACTCAATTTGAATTAACAAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCACAAAAAC  
ATAAAATATCCCATCCGAAATTC AAAATGGTCCATTATCGGCAAGTAGCTTTCTTTAAATATAGTTAGTT  
GACAAAACACTATCAAGATATCATTATTATAATAATAAACTTCAAGTCCATCATCTTAGCTGCCTCCTCA  
GTAGAGCCGCCAGTAAAAAAGACCAGATCAAAATAAAGCCGCCATTAAAAATAATGAATTTTAGGACTCTC  
GATTGGCACGTAAGTGCCAAAACCTTCCAAACTTTTGGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC  
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTAACAGGTA  
CAGCAAATCATAAAATTTTCTCTAAAGAAAGACAATGAATCCGGTACTGATTCATTGGCCTTTTAGAG  
TCTGCATGCCATATTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAAATTTTCGGGATAGAATTT  
GAGATTGCATTTATCTTGTGTTTAAATATAAGTATTAGCTAATTTTACAATAAAATTTTACTAAAATAG  
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC  
TTATTTATCTCACTATTTTACC AAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA  
GTAAGAAGTTAGAAAAATTTTCATTAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG  
ATACAATAAAGATGTACCGTTAATAATAAAGATAAGATAGAGTTTAAATAGGAAAAAAAACGGTT  
CGAGACACTCTTATGGAAGGCGTTGCTTCAAAGTAGATTTCTCATTCATTGCTCTGGTGC AATAGC AAAA  
TGACATCTTACTCTTAAGATACAGCGAGCCACTTACAACTTCTATTGTATACTCAAAATGAAAGTTTAA  
GAGAACTTTTCAAATCTCAACTACTTTAAGGGAATTC AAAATACGACCAATATTTATTACTTACT  
TTATAGTTAAATGATATGAATTTTAAATTTGAAATGAAAATATTAATTTACTTGTATTAATATAA  
ACAATAGATATCGCTAAGTATTTACCACAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT  
GATTAAGCAGCTATTCATCTGGTTTGTGCAGGATGAAAGAAAGTAACTAGCTATAATTTCTMMTGTAAGT
```

# Analytické nástroje

- **Biology Workbench** <http://workbench.sdsc.edu/>

## Frame 1, 1 stop codon

Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. Tran

```
>170248 Translated - Frame 1  
ELPWGARAKLFAKWKNIIIPSVCSYSI*INKGANLTILPL
```

```
      E L P W G A R A K L F A K W K N I I P S  
1    gagctcccttggggggcaagggcaaaactttttgctaaatggaaaaatattataccaagt 60  
      V C N S Y S I * I N K G A N L T I L P L  
61   gtttgtaatagttactcaatttgaattaacaaaggggcaatttgactattttgcctta 120
```

## Frame 2, 1 stop codon

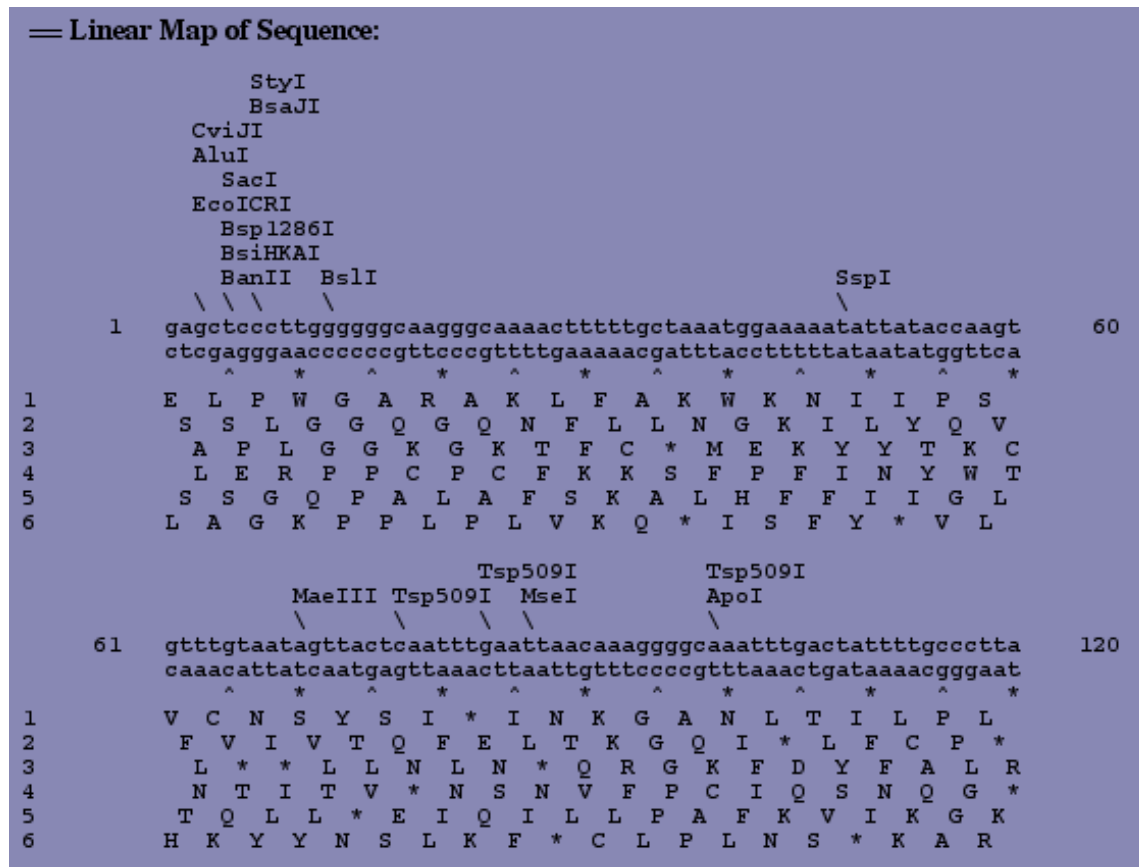
Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. Tran

```
>170248 Translated - Frame 2  
SSLGGQGQNFLLNGKILYQVFVIVTQFELTKGQI*LFCP
```

```
      S S L G G Q G Q N F L L N G K I L Y Q V  
2    agctcccttggggggcaagggcaaaactttttgctaaatggaaaaatattataccaagtg 61  
      F V I V T Q F E L T K G Q I * L F C P  
62   tttgtaatagttactcaatttgaattaacaaaggggcaatttgactattttgcctta 120
```

# Analytické nástroje

- **Biology Workbench** <http://workbench.sdsc.edu/>



# Analytické nástroje

- o **Biology Workbench** <http://workbench.sdsc.edu/>

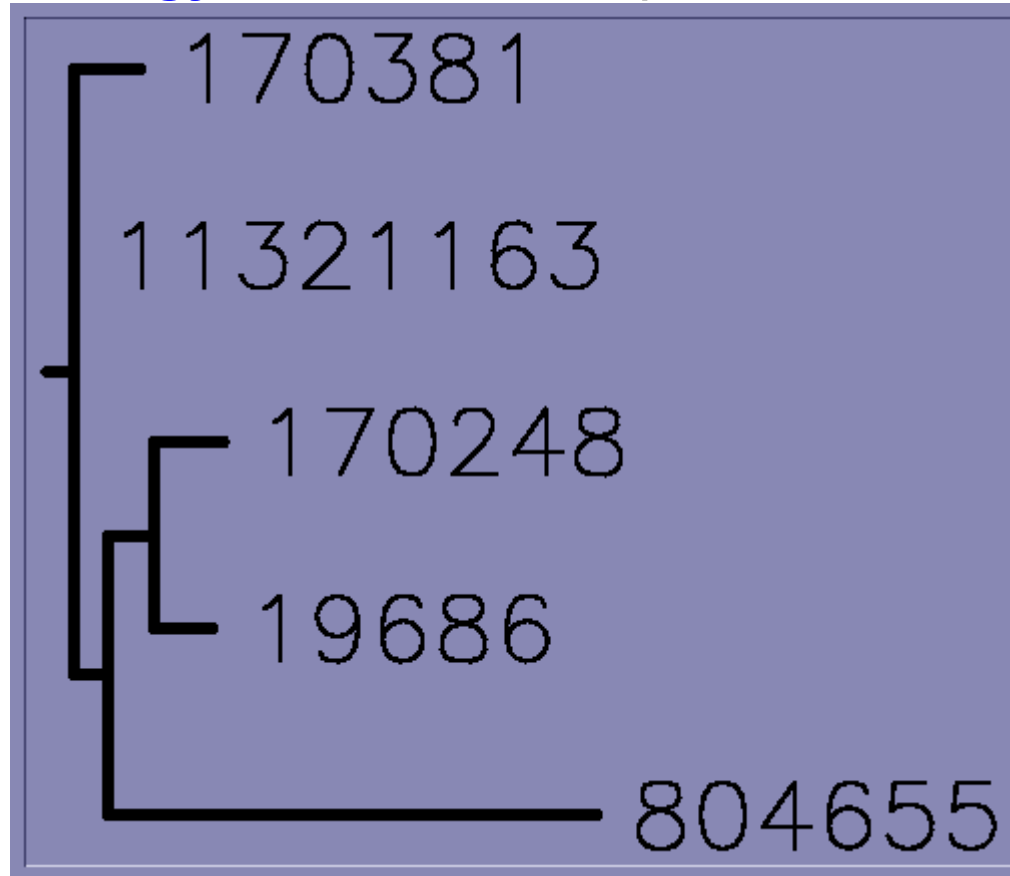
**Selected Sequence(s)**  
• Lycopersicon esculentum beta-1,3-glucanase mRNA, complete cds.,  
Capsicum annuum clone GC170 beta-1,3-glucanase-like protein gene,  
Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds.,  
Nicotiana plumbaginifolia beta-(1,3)-glucanase gene for a vacuolar,  
Hordeum vulgare L. beta-glucosidase (BGQ60) gene, complete cds.

[Download a PostScript version of the output](#)

The screenshot displays a sequence alignment interface. At the top, there is a list of 'Selected Sequence(s)' including various beta-1,3-glucanase genes from different species. Below this, a link to 'Download a PostScript version of the output' is visible. The main area shows a multiple sequence alignment of DNA sequences. The sequences are arranged in columns, with positions 2560, 2700, 2710, 2720, 2730, 2740, 2750, 2760, 2770, 2780, 2790, 2800, 2810, and 2820 marked. Several regions are highlighted in red and yellow, indicating areas of interest or conservation. The sequences are numbered on the left and right sides, ranging from 2560 to 2800 on the left and 804855 to 170381 on the right. The highlighted regions include motifs such as 'AAATGGCT', 'GAAAGATT', 'ATTAATGCTTCTACGATTCTTGTGCGGA', 'GCAACATTEAGATAG', 'AGGGTFAA', 'ATAGGTG', 'TTGTTATGCAATGAGC', 'AACAAGTTCGATGAG', 'T...ATGGGTG', 'TTGGTATGCAATGAGC', 'AACAAGTTCGATGAG', 'AGGGTFAATCAATAGGTG', 'TTGGTATGCAATGAGC', 'AACAAGTTCGATGAG', 'AGGGTCAATGATAGGTG', 'TTGGTATGCAATGAGC', 'AACAAGTTCGATGAG', 'AGGTGCTGAGCTTCAATAAAAAAGATAATGTAATGCTTTATAGAAAGTGAAG', and 'ATGTTAAGTATAAAG', 'TTTACAAGTGGAGAACATTAACAAGTTCAGGCTTTATGA', 'TTTCCAAGTATAAAG', 'TTTACAAGTCAAGAACATTAACAAGTTCAGGCTTTATGA', 'TTTCCAAGTATAAAG', 'TTTACAAGTCAAGAACATTAACAAGTTCAGGCTTTATGA', and 'TTTCCAAGTATAAAG'.


# Analytické nástroje

- **Biology Workbench** <http://workbench.sdsc.edu/>



# Analytické nástroje

- Virtual PCR (VPCR) <http://grup.cribi.unipd.it/cgi-bin/mateo/vpcr2.cgi>

**SEARCH**  [ABOUT](#) [DOWNLOAD](#) [LINKS](#)

VPCR 2.0 (WWW interface) - Please, enter nucleotide primer sequences ([IUB codes](#) allowed for degenerate primers). VPCR 2.0 searches the specified database for matches to the primers. If matches are found within 10000 bases, a PCR simulation model predicts amplification. Calculated PCR products are displayed within a minute.

NOTE: Abilities of VPCR 2.0 are still limited by BLAST capabilities and settings, as well as inability of our current software to deal with more than a couple thousand matches per primer. For example, using primers shorter or roughly equal to our 11-base word size misses most matches. Primers with overrepresented sequences cause problems as well. We are now busy solving most of these problems, please, be patient. If you have a minute, please, let us know what kind of expectations you have for VPCR 2.0 etc. Currently, this address is for testing VPCR 2.0, stable features will be installed on [VPCR 2.0 Homepage](#).

Search using  in the database for

Primer 1

Primer 2

Primer 3

Primer 4


Primer 5

Primer 6

Primer 7

Primer 8

Annealing temperature





# Analytické nástroje

- **Virtual PCR (VPCR)** <http://grup.cribi.unipd.it/cgi-bin/mateo/vpcr2.cgi>



# Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst....
  - Další www genomové nástroje

# Další WWW zdroje

- TIGR (The Institute for Genomic Research), <http://www.tigr.org/software/>
  - Recently part of the J. Craig Venter Institute

The screenshot displays the NCBI Gene database entry for PHACTR4 phosphatase and actin regulator 4 [Homo sapiens]. The page is viewed in a Mozilla Firefox browser. The main content includes:

- Gene ID:** 65979, updated on 27-Aug-2011
- Summary:**
  - Official Symbol:** PHACTR4 provided by HGNC
  - Official Full Name:** phosphatase and actin regulator 4 provided by HGNC
  - Primary source:** HGNC:25793
  - Locus tag:** RP11-442N24\_A.1
  - See related:** Ensembl:ENSG00000204138; HPRD:07818; MIM:608726
  - Gene type:** protein coding
  - RefSeq status:** REVIEWED
  - Organism:** Homo sapiens
  - Lineage:** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Iliomniidae; Homo
  - Also known as:** FLJ13171; MGC20618; MGC34186; DKFZp686L07205; RP11-442N24\_\_A.1
  - Summary:** This gene encodes a member of the phosphatase and actin regulator (PHACTR) family. Other PHACTR family members have been shown to inhibit protein phosphatase 1 (PP1) activity, and the homolog of this gene in the mouse has been shown to interact with actin and PP1. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]
- Genomic context:**
  - Location:** 1p35.3
  - Sequence:** Chromosome: 1; NC\_000001.10 (28866093..28866881)
  - Diagram:** A genomic map of Chromosome 1 - NC\_000001.10 showing the location of PHACTR4 and other genes (SESN2, MED19, SNORA73A, SNORA73B, RNUL15A, ZWUC3, RRC1) in the region from 28555963 to 28865716.
- Genomic regions, transcripts, and products:**
  - Genomic Sequence:** NC\_000001 chromosome 1 reference GRCh37.p5 Primary Assembly
  - Links:** Order cDNA clone, BioAssay, by Gene target, BioProjects, CCDS, Conserved Domains, dbVar, EST, Full text in PMC, Genome, GEO Profiles, HomoloGene, Map Viewer, Nucleotide, OMIM, Probe, Protein, PubChem Compound, PubChem Substance, PubMed, PubMed (GeneRIF), PubMed (OMIM), RefSeq Proteins.

# Další WWW zdroje

- Online Mendelian Inheritance in Man (OMIM) <http://www.omim.org/>




The screenshot shows the OMIM website in a Firefox browser window. The address bar displays "omim.org/#". The page content includes the OMIM logo, the text "Online Mendelian Inheritance in Man", and a search bar. The footer contains a disclaimer and copyright information.

Mirror sites: [us-east.omim.org](http://us-east.omim.org), [europe.omim.org](http://europe.omim.org)

**OMIM<sup>®</sup>**  
Online Mendelian Inheritance in Man<sup>®</sup>  
An Online Catalog of Human Genes and Genetic Disorders  
Updated 6 September 2012

Search OMIM   [Sample Searches](#)

Advanced Search: [OMIM](#), [Clinical Synopses](#), [OMIM Gene Map](#)

NOTE: OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions.

OMIM<sup>®</sup> and Online Mendelian Inheritance in Man<sup>®</sup> are registered trademarks of the Johns Hopkins University.  
Copyright<sup>®</sup> 1966-2012 Johns Hopkins University.

# Shrnutí

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst....
  - Další [www genomové nástroje](#)

# Diskuse



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky