

# 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'TTATATATATATATATTAATAAAATAAAATAAAA  
GAACAAAAAGAAAATAAAATA....3'



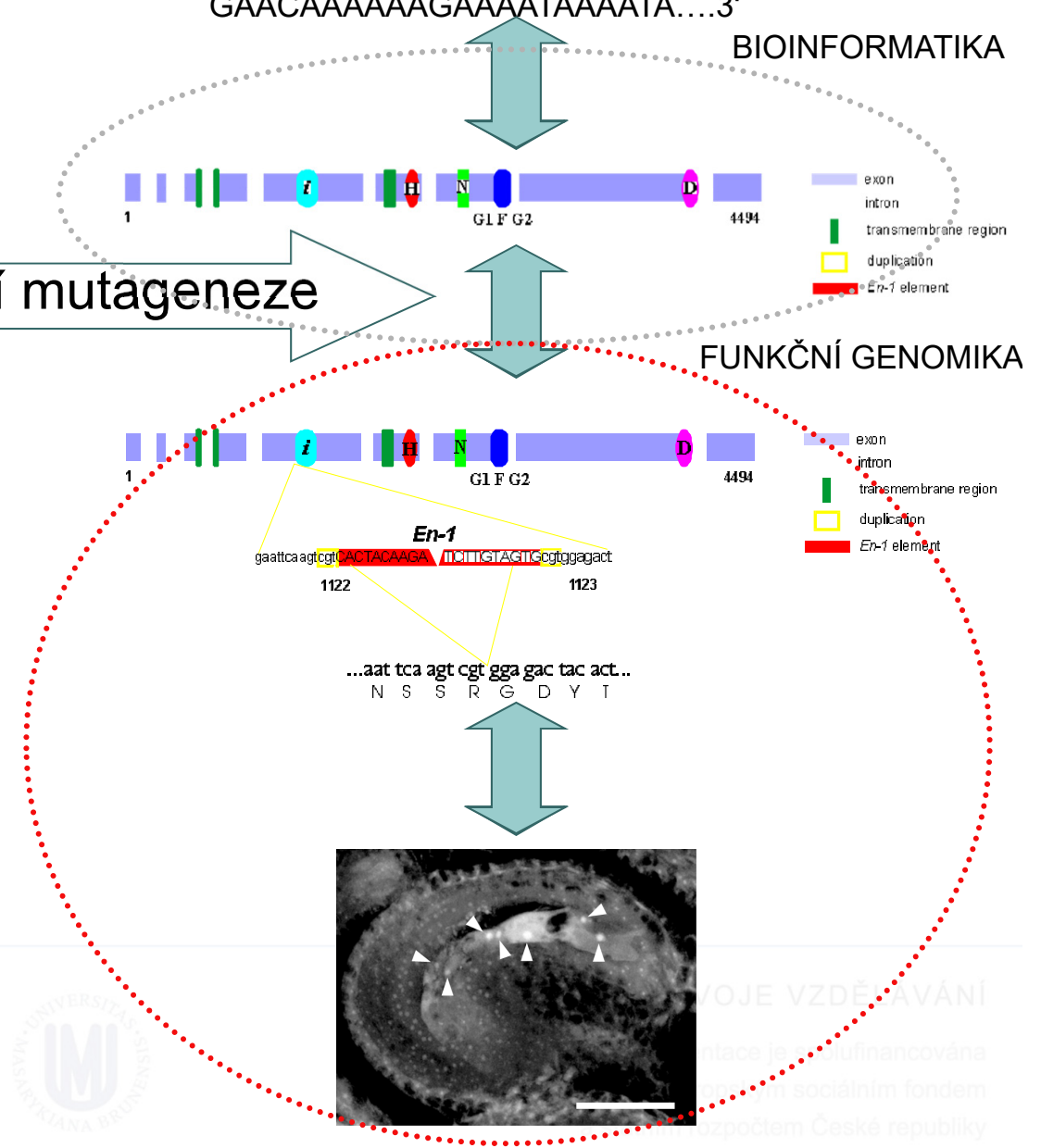
3

:

1



?



EVROPSKÁ UNIE



MLÁDEŽE A TĚLOVÝCHOVY

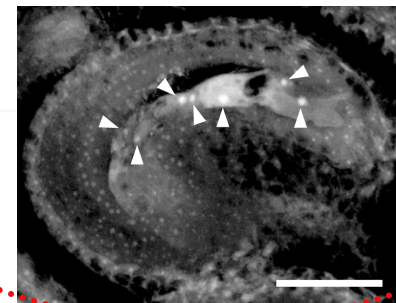
pro konkurenceschopnost



vání



UNIVERSITY MASARYKŮVA BRNO

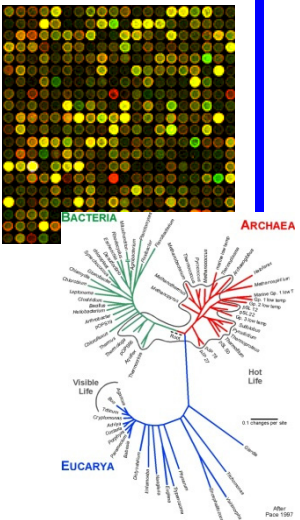


VOJE VZDĚÁVÁNÍ  
ntace je financována  
sociálním fondem  
zpočetm České republiky

# Osnova

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

# 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.

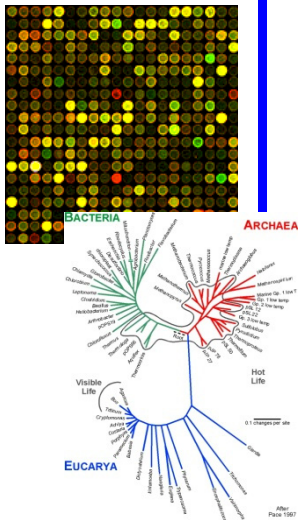
J. Pevsner,  
<http://www.bioinfbook.org/index.php>



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

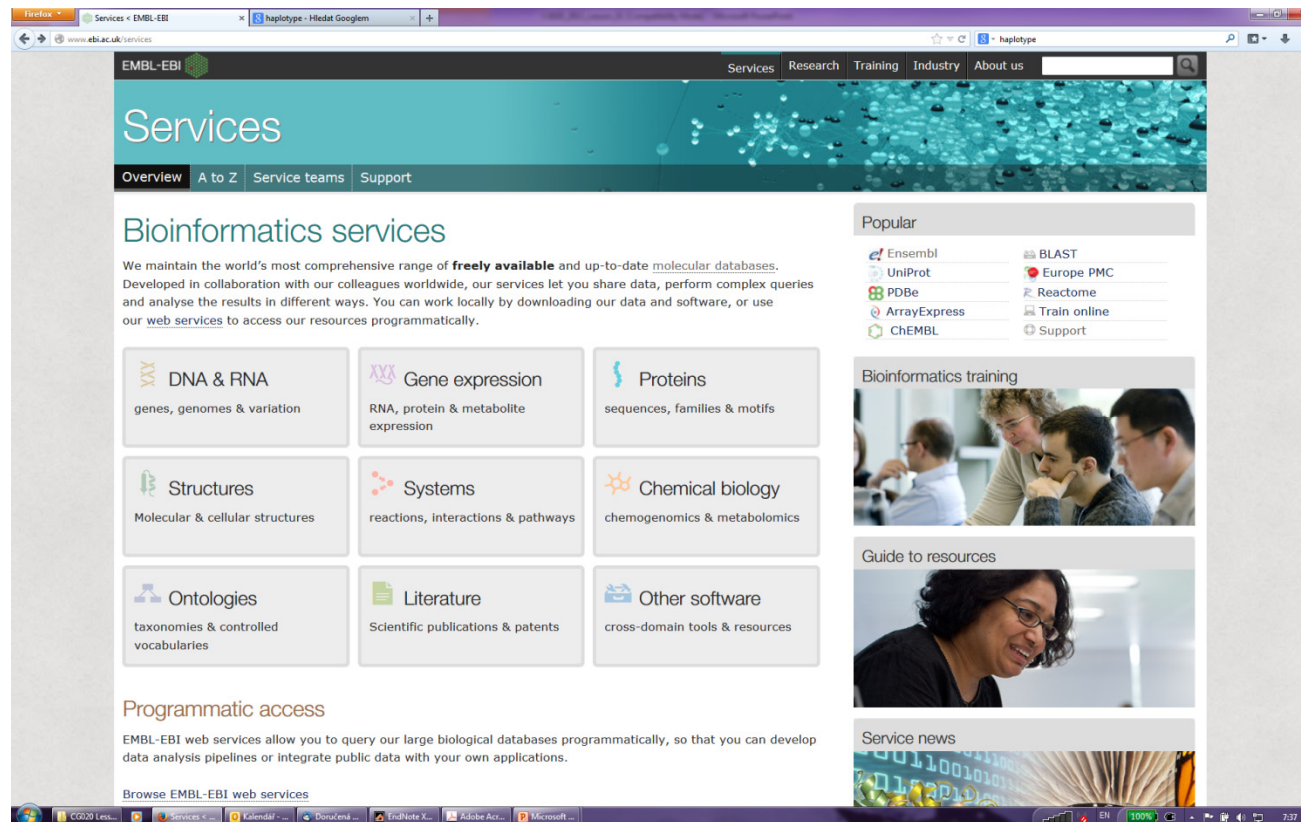
Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

# 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>

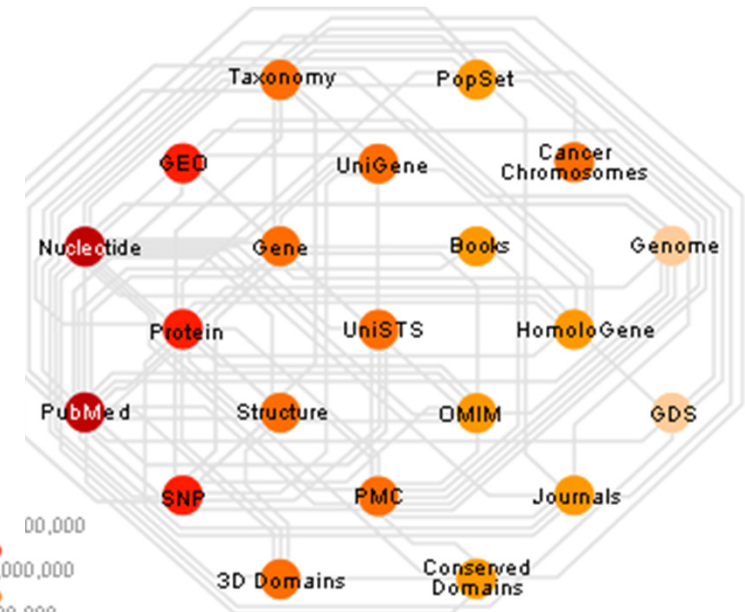




# Spektrum on-line zdrojů

□ NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI homepage with a search bar at the top. The main content area is divided into three columns. The left column contains a 'Resource List (A-Z)' with categories like 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The middle column features a 'Welcome to NCBI' message, a 'Get Started' section with links for Tools, Downloads, How-To's, and Submissions, and a 'NCBI YouTube channel' advertisement. The right column lists 'Popular Resources' such as PubMed, Booksshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem, along with 'NCBI Announcements'.



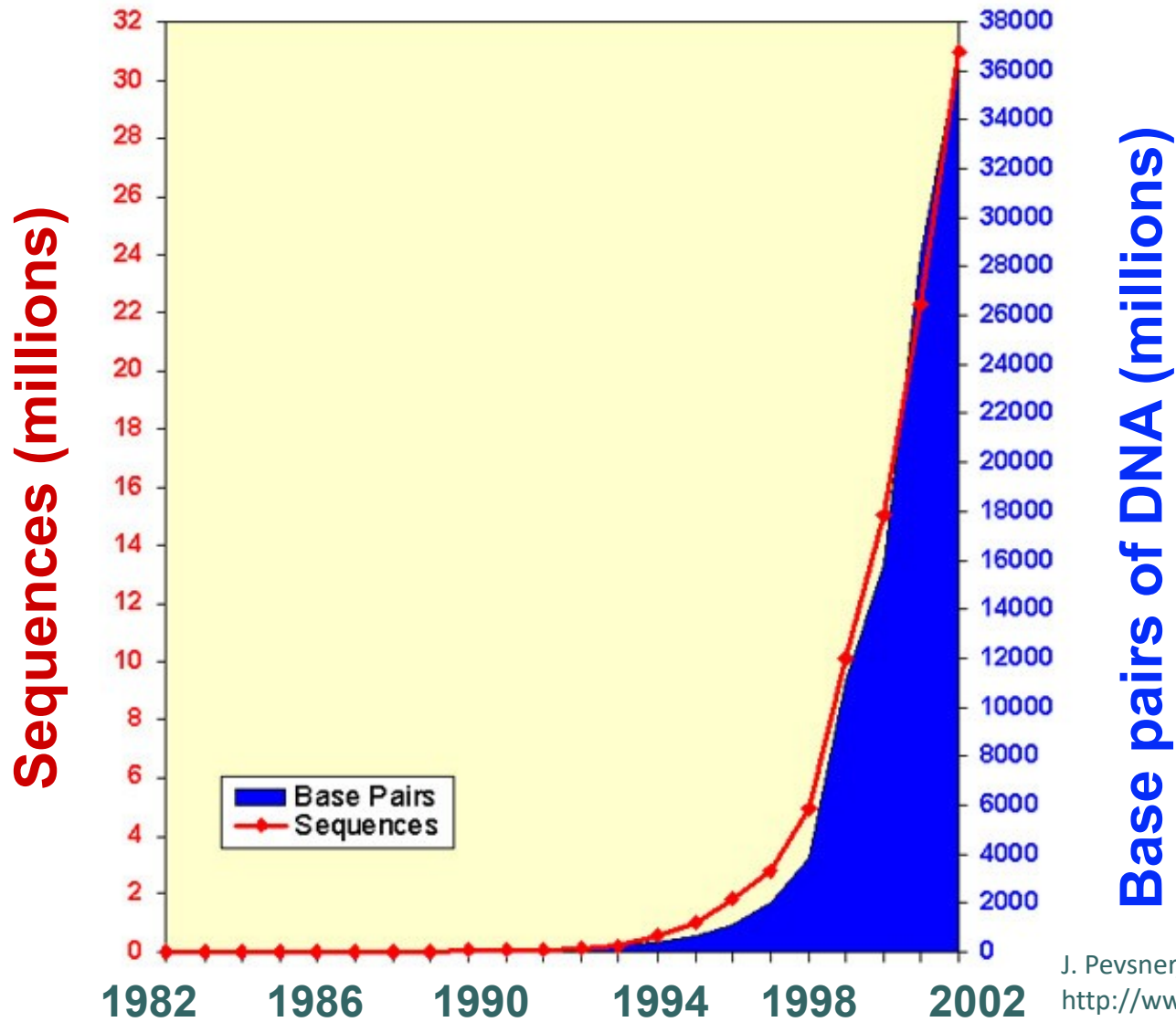
# Osnova

- 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)

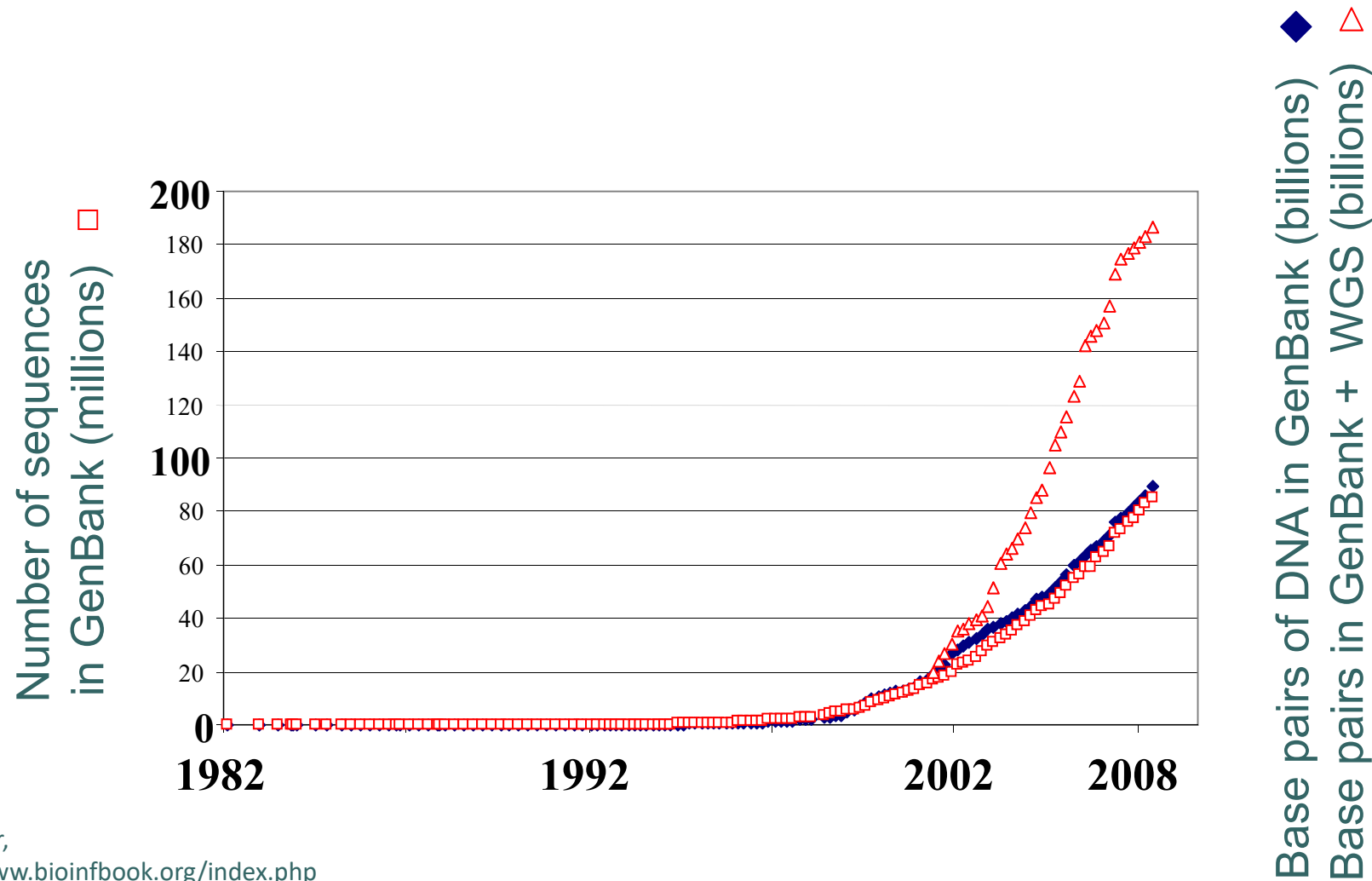
# Growth of GenBank



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>

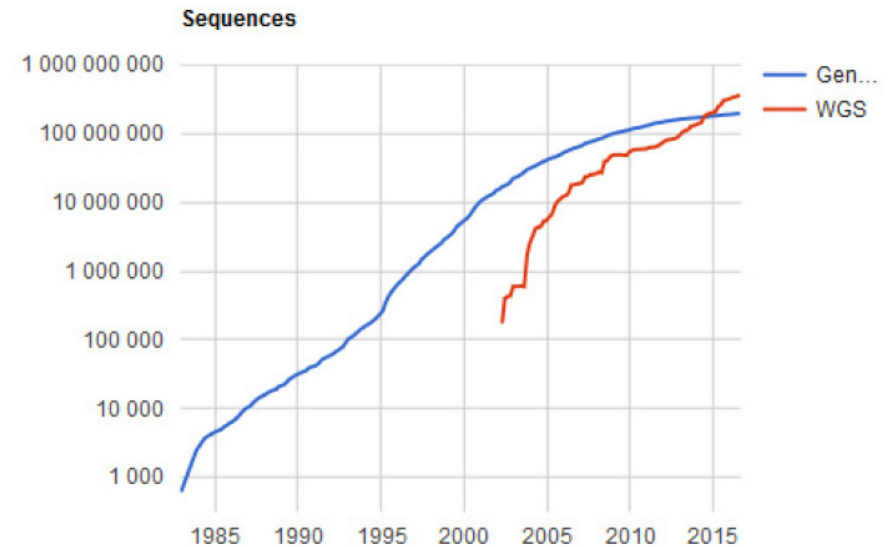
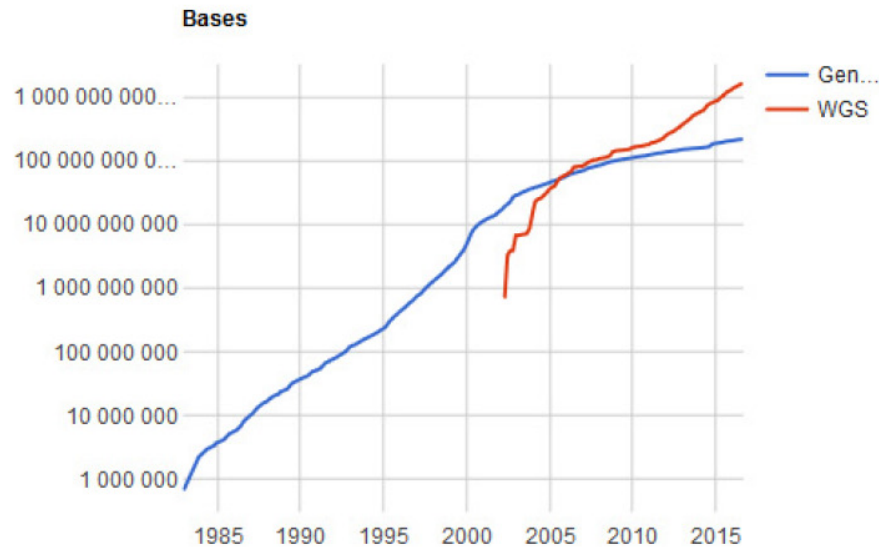


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

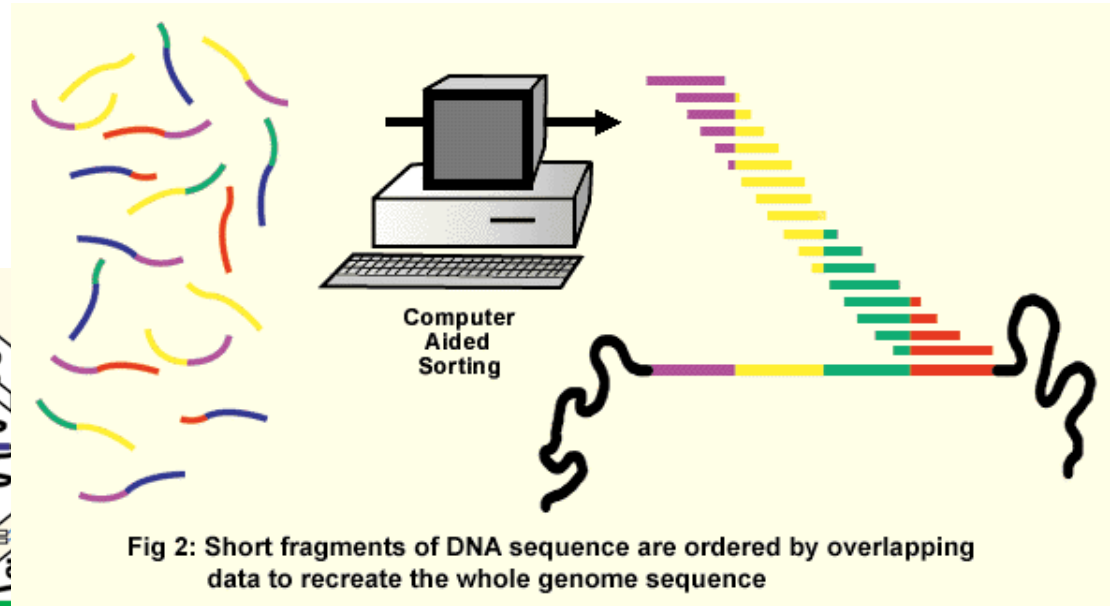
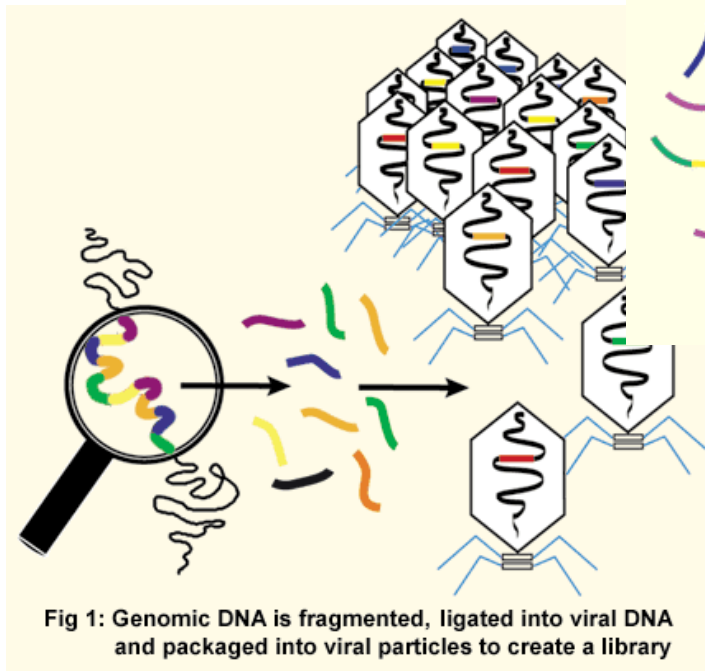
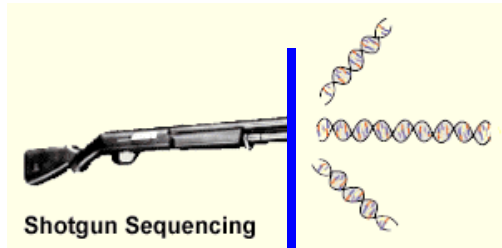
# Growth of GenBank

## Aug 2016



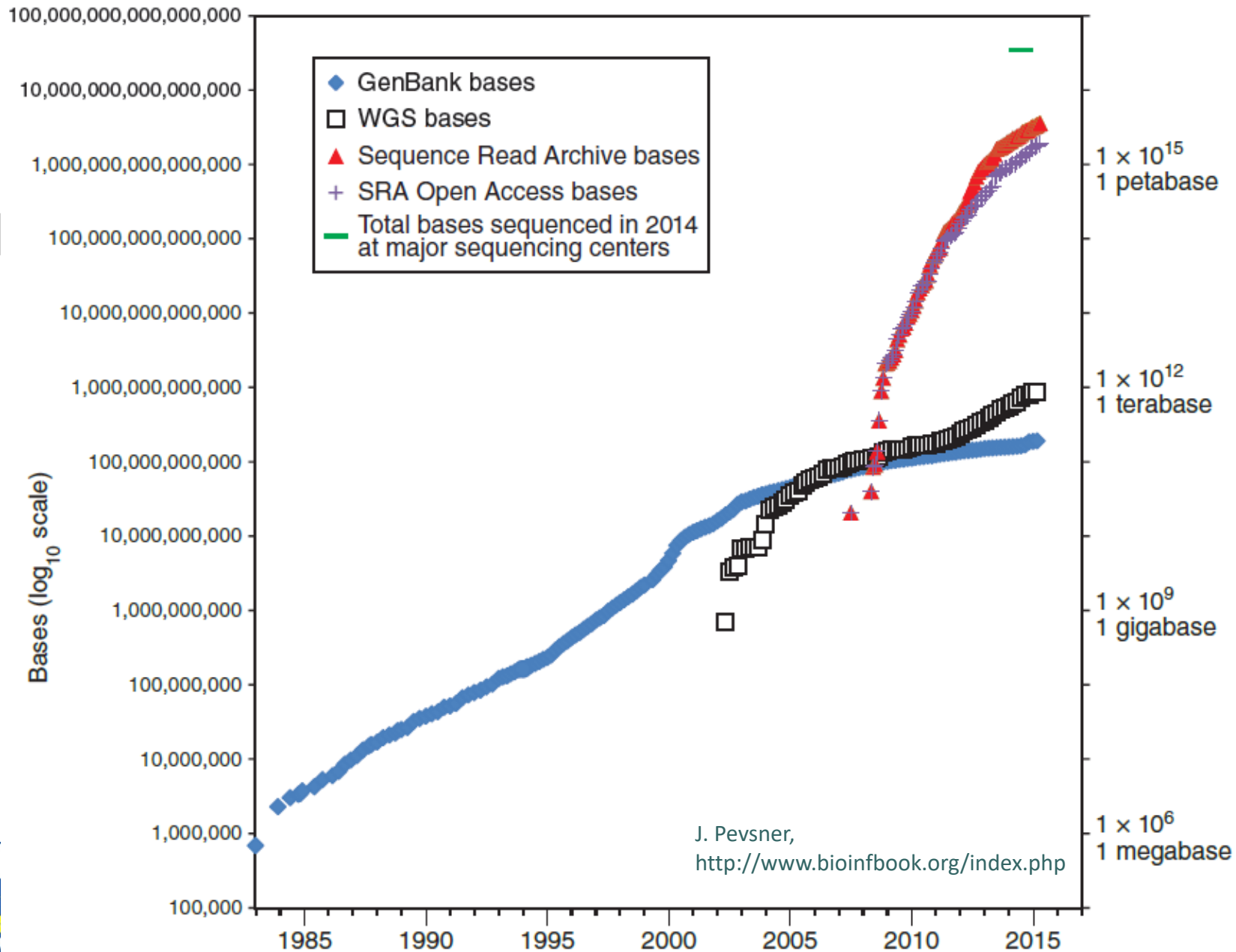
- Prosinec **1982** 680 338 bp, 606 sekvencí
- Duben **2002**  $19 \times 10^9$  bp,  $17 \times 10^6$  sekvencí + WGS  $692 \times 10^6$  bp, 172 768 sekvencí
- Srpen **2016**  $218 \times 10^9$  bp,  $196 \times 10^6$  sekvencí + WGS  $1,6 \times 10^{12}$  bp,  $360 \times 10^6$  sekvencí

# WGS



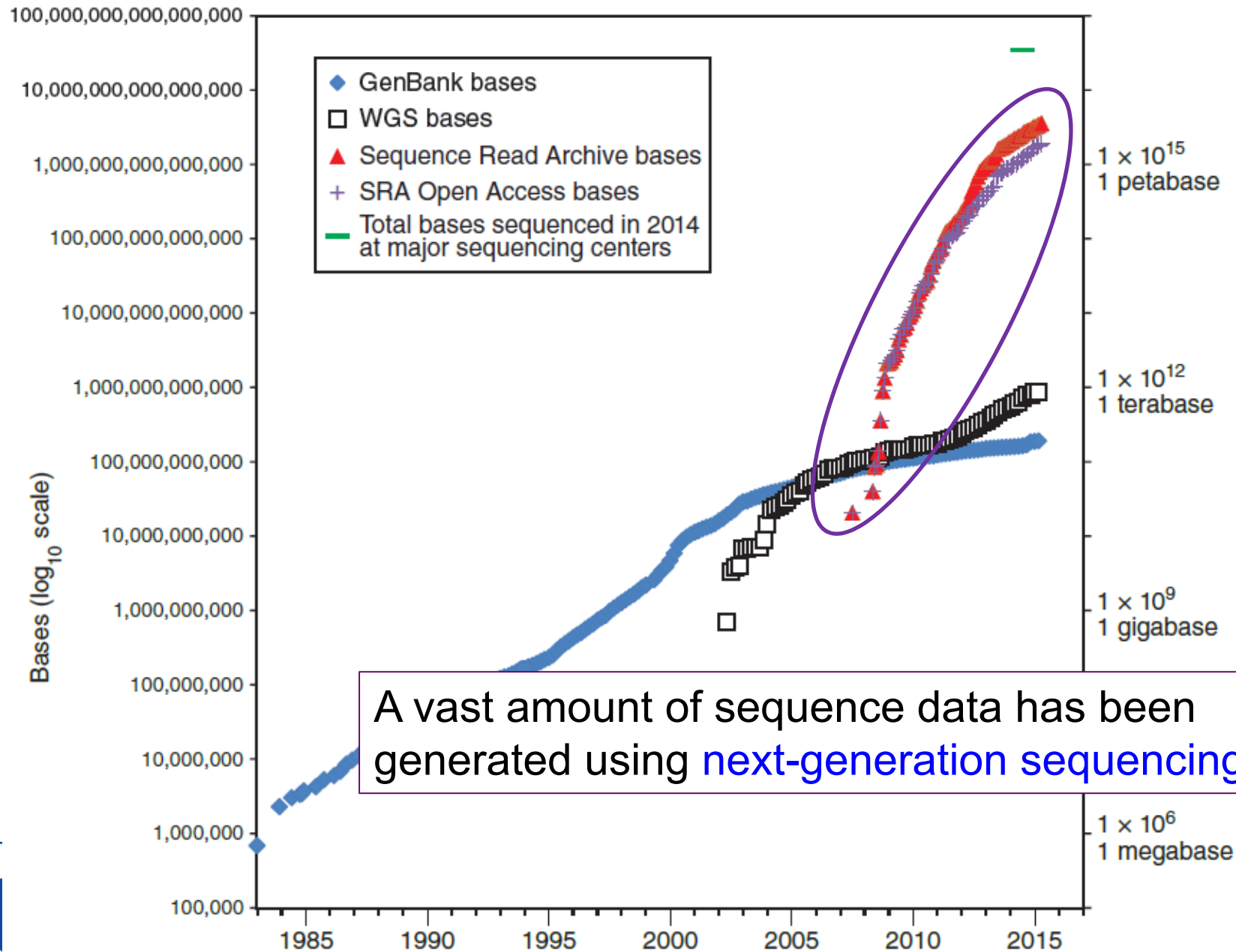
Interactive concepts in biochemistry, Rodney Boyer, Wiley, 2002, <http://www.wiley.com//college/boyer/0470003790/>

# Growth of DNA Sequence in Repositories

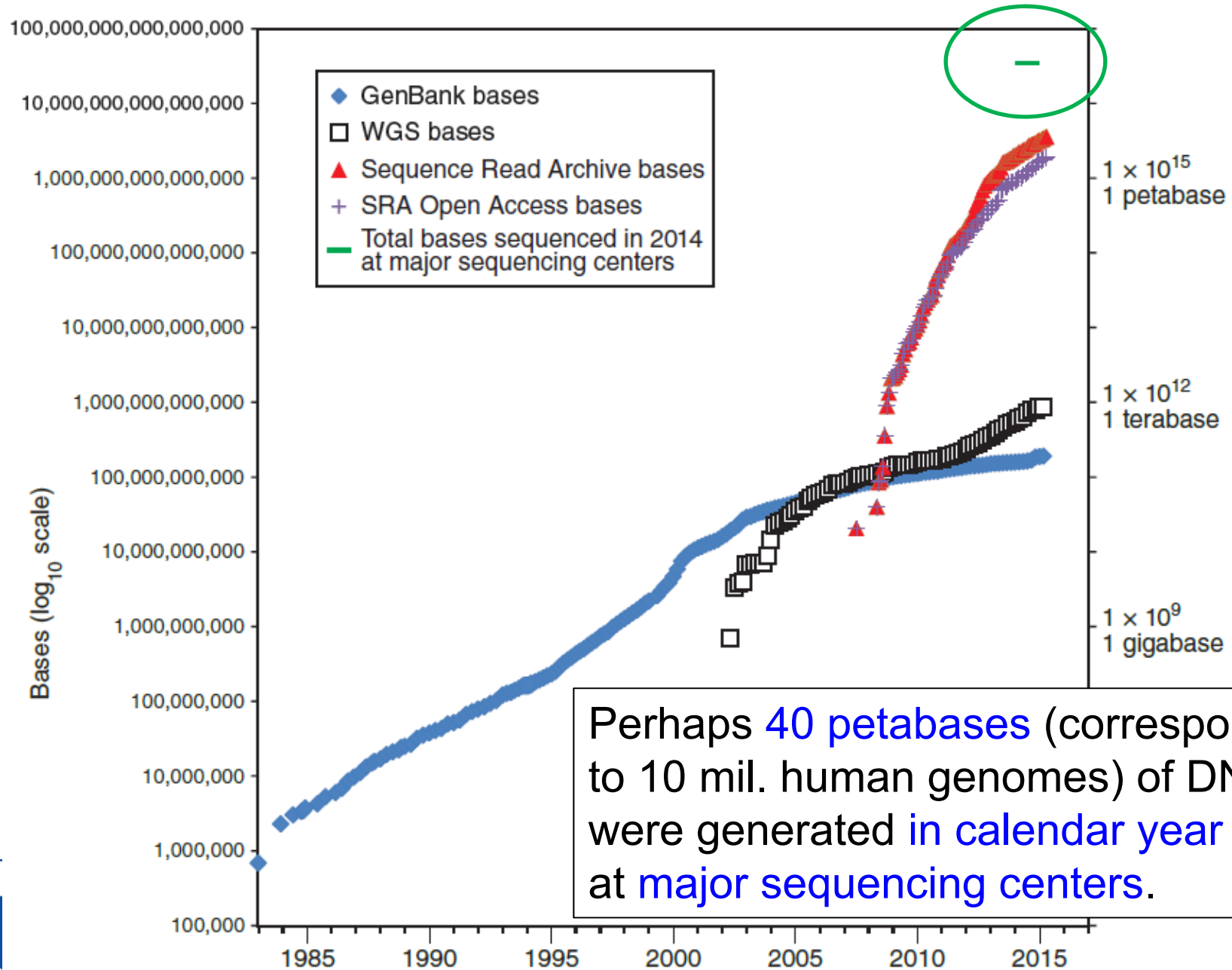




# Growth of DNA Sequence in Repositories



# 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**notation)
    - 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 central 'Welcome to NCBI' section. The search bar contains 'All Databases' and a 'Search' button. The navigation menu includes 'NCBI Home', 'Resource List (A-Z)', and various biological categories. The central section features a 'Welcome to NCBI' message, a 'Get Started' section with links to Tools, Downloads, How-To's, and Submissions, and a 'Popular Resources' list including PubMed, Bookshelf, and BLAST. A YouTube channel banner is also visible at the bottom of the main content area.

# Primární databáze

**Gene symbol** virA  
**Gene description** two-component VirA-like sensor kinase  
**Locus tag** pTl\_125  
**Gene type** protein coding  
**RefSeq status** PROVISIONAL  
**Organism** *Agrobacterium tumefaciens* (old-name: *Agrobacterium tumefaciens*, qb-synonym: *Rhizobium radiobacter*)  
**Lineage** Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiaceae; Rhizobium/Agrobacterium group; Agrobacterium; Agrobacterium tumefaciens complex

**Genomic context**  
Location: plasmid: Ti  
Sequence: NC\_002377.1 (145694..148183)

**Genomic regions, transcripts, and products**  
Genomic Sequence: NC\_002377

**Sequence Viewer**  
NC\_002377.1: 145K-148K (3.2Kbp) Find on Sequence: [145,400 | 145,600 | 145,800 | 146 K | 146,200 | 146,400 | 146,600 | 146,800 | 147 K | 147,200 | 147,400 | 147,600 | 147,800 | 148 K | 148,200 | 148,400]

**Gene Details (circled in yellow):**  
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**  
1. Sequence analysis of the virA gene from *Agrobacterium tumefaciens* octopine  $\square$  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.

**GeneRIFs: Gene References Into Functions** What's a GeneRIF?  
Submit: [New GeneRIF](#) [Correction](#)



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](#)  
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**

# Primární databáze

**Přístupový kód**

**GeneBank Identifier**

NCBI  
Nucleotide  
Search for [ ] for [ ] Go Clear  
Preview/Index  
History  
Clipboard  
Details  
Links

LOCUS NC\_002377.1 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)  
FARRAND, S.K., ZHU, J., 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 [AF242981](#).  
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="MP\_059797.1"  
/db\_xref="GI:10955141"



# Primární databáze

```
/translation="MNGRYSPTQDFKTKAKPWSILALIYAAMI FAFMAVASWQDNMT  
TQAILSQLRINADSASLQRDVLAHTCTVANYRPI I SRLGALRKNLEDLKQLFRQSH  
IVSEENRQQLRQLEVSLSMSADAAVAAPGQNVRLQDSIASPTRALSSLPKASTDQT  
LEKPTLEASMMQLRQSPASISFPI SLELEELQKQRLDEAPVVLAREGPI ILSLL  
PQVKDLVNMQISTDIAEMLQRCLVYSLKNVEERSARIPLSASVGLCLYIITL  
VYLRKKTDWLARRLDYELIKEI GVCFFGSRATSSQAALRI IQRPFDADTCALAL  
VDHDERWAVETFGAKHFKPVWDSVLRRI VSRTKADEBRATVFR IISKKIVHLFLHIP  
GLSILLAHKSTDKLIAVCSLGYQSYRFPFCQGEIQLLELATACLCHYIDVRRKQTECD  
VLARELEHAQRLAVGTLAGGIAHFNNILGSELGHAELAQNSVSRTEVTRRYIDYII  
SSGDRAMLIIDQILTLRKRQEMIKPPSVSELVTEIAPLRLMALPPNIELSFRPDQMC  
SVI EGSPLRLQQLINICKNASQAMTANQIDII IIGQAPLPVKKILAHGVMPFGDYVL  
LSISDNQGGIPRAVLPHI FEPFPTTRARNGGTGLGLASVHGHISAPAGYIDVSTVGH  
GTRFDIYLPSPSKKFPVNPDSFFGRNKA PRGNGEIVALVFPDDLREAYRDKI AALGYE  
PVGFRTPNKRDIWISKGNEDLVMDQASLPEDQSPNSVDLVKTA SIIIGGNDLKM  
LSREDVT RDLYLFPKPISSRTMAHALTKIKT"
```

```
ORIGIN  
1 atgaacggaa gatattcacc gaecggcgag gattttaaga caggcgcgaa gcccttggct  
61 atattggccc ttatcgttgc tgaatgatt ttocggttca tggcggttgc gtccttggcag  
121 gacaatgcca ctaccacgga aatcctcage caactacgat cgat taacgc cgacagcgcc  
181 tcaactgacg gogatgact cecgectcac acgggcaacg tggcgaacta ccgccccatt  
241 atctccaggg tgggagctct gcggaagaat ctggaagatt tgaagcaatt atttagacaa  
301 tctcatattg taagtgagag caatgctget caactgctac gccagctaga agtgtctcta  
361 aatctggctg acgcgcggtg cgcgcctttt ggtgocgaaa atgtacgctt gcaagattcg  
421 ctggcgcagt tcaactcgtg tttgagcagt ctccacgaaa aagcctcaac cgatcagact  
481 ttgaaaaaac caacagaatt ggotagcagt atgctccaat ttcttggca accaaagccg  
541 gctatttcat togagatcag ccttgaacta gagaggctcc aaaaacaacg cggcttggat  
601 gaagctcccg tgcgcatact tgcacgtgaa ggtccattta tcttctcgtt tttgccacag  
661 gtgaaagatc tgggtaacat gatcagacgc tctgacacgc cagaatctgc gtagatgctg  
721 cagcgcgagt gtttgagggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc  
781 ttctttgggt cgccttcagt gggctcttgc ctctacatca tcaccttagt ctataggcta  
841 cgcacacaaa cogatgggtt agcgcggcgt ttgatatacg aagagctaat caaagagatc  
901 ggagtagtgt ttgaaggtga ggcggccacc acgtcgtccg cgcacagctc actctgtatt  
961 atcagcgcct tcttgatgc cgtacgttgc gcttagctc tagtggacca tgacgttaga  
1021 tgggctgtcg aaacattcgg tgcgaaacac caaaacctgt tctgggacga cagcgtgcta  
1081 cgcgaaatag tctctcgtac caaagcggac gaacgggcca cggatctcgc catcatatcg  
1141 tgcacacaaa tctacatttt gctctcgtac atctcagctc tctcgtactc actggtctcc  
1201 aaatccacag ataaactaat tgcggtttgt tcaactgggtt accaaagcta tgcgcctcga  
1261 ccttgcacag gogaacttca gctcttggaa ctgcacacgc cctgctctgt tcatatatac  
1321 gatgttcggc gtaagcagac cgaatgcgac gttttggcca gacgatgga gcatgcgcaa  
1381 cgccttgagg cagttggtac acttgcggcg ggaatgacac atgaatttaa taacattttg  
1441 ggctcaatcc tgggcaacgc agaattagca caaaactcgg tctctcgaac atctgtcacc  
1501 cgaagatata ttgactatat catttctgca ggccacagag ccatgctcat tctcgtcag  
1561 atcttgacgc tgagccgaaa acaggagcgc atgatcaagc catttagtgt ctcagagctt  
1621 gtgaccgaaa tgcctccttt gctacgtatg gctcttcgcg caaacatoga gcttagtttc  
1681 agatttgatc aaatgcagag cgtgatcgaa ggaagccgcg ttgaacttca acaggtacta  
1741 ataacatct gcaagaatgc tcccaagcc atgacgtcaa atggtcaaat cgcacatcct  
1801 atcagccaag cttttttacc agttaagaaa atcttggcgc atggtgttat gccocctggc  
1861 gactatgttc tctatctatc tagcgaacat ggtggaggca tcccgaggcc tgtgttacc  
1921 cacatttttg aacctctctt tacgacacga gctgcacacg gtggaacggg tctcggcctt  
1981 gcttctgtgc atggtcatal cagcgcgctt gcgggttaca tgcagcttag ttoactgttt  
2041 gggcatggga cgcgcttga ctttatctc cctcctgtct ctaaagaaec cgtaaactca  
2101 gacagttttt bccggccgaa taaggccacc cgtggaacgc gggagattgt ggcactgttt  
2161 gacccgatg acctcctgag gtaggcctat gaagacaaga tgcgcgctct aggatagag  
2221 ccggtcgggt ttctgactct taatgaactt cgcgatggga tttcaaaagg caatgaagcc  
2281 gatctggtca tggctcagca agcgtctctt cctgaagatc aaagtcttaa tctcgtggat  
2341 ttagtgtca agacgcctc catcatcatt ggcggaatg atctcaaat gacccttca
```



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           |
| NT_030059 | Genomic contig                         |
| Rs7079946 | dbSNP (single nucleotide polymorphism) |

**DNA**

|           |   |
|-----------|---|
| N91759.1  | An expressed sequence tag (1 of 170)    |
| NM_006744 | RefSeq DNA sequence (from a transcript) |

**RNA**

|           |                                    |
|-----------|------------------------------------|
| NP_007635 | RefSeq protein                     |
| AAC02945  | GenBank protein                    |
| Q28369    | SwissProt protein                  |
| 1KT7      | Protein Data Bank structure record |

**Protein**

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>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# 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
ATGAACGGAAGATATTCACCGACGCGGCAGGATTTAAGACAGCGCGGAAGCCTTGGTCTATATGGGCC
TTATCGTTGCTGCAATGATTTTCGCGTTTCATGGCGGTTGCGTCTGGCAGGACAAATCGACTACCCAGGC
AATCCTCAGCCAACACGATCGATTAACCGCGACAGCCCTCACTGCAGCGCGATGACTCCGCGCTCAC
ACGGCACCGTGGCGAATACCGCCCATTTATCTCCAGGCTGGAGCTCTGGGAAGAAATCGAAAGATT
TGAAGCAATTTAGCAATCTCATATTTGAAGTGAAGCAATGCTGCTCAACTGCTACGCGACGTAGA
AGTGTCTAAATTCGGCTGACGCGCGGCTCGCCGCTTTGGTGGCAAAATGTACGCTGAAAGATTG
CTGGCCAGTTTCACTCGTGGCTTGGAGCTTCCAGGAAAAGCCTCAACCGATCAGACTTTAGAAAAAC
CAACAGAATTGGTAGCATGATGCTCCAATTTCTCGGCAACCAAGCCGGCTATTTCAATCGAGATCAG
CCTTGAAGTGAAGAGGCTCCAAAAACAACCGCGCTTGTATGAAGTCCCGTGGCATACTTGCACGTGAA
GGTCCCAATTTATCGCTTTGGCCAGAGTGAAGATCTGGTGAACATGATTCAGACGCTCTGACACCG
CAGAAATTCGGAGATGCTGCAGCGAGTGTGGAGGTCTATAGCTTGAATAATGTAGAGGAGCGGAG
CGCAGTATCTTTCTGGTCCGCTTCAGTGGGTCTTGGCTCTACATCATCACTTGTCTATAGGCTA
CGCAAAAAACCGATTGGTTAGCGCGGCTTAGATTACGAAGAGCTAATCAAGAGATCGGAGTAGTGT
TTGAAGTGAAGCGGCCACCACTGCTCGCGCAAGCTGCATTCGTATTATTCAGCGCTTTTGGATGC
CGATACGTCGCGCTTAGCTTAGTGGACCATGACCGTAGTGGGCTGTGCAAAACATTCGTTGCAAAAC
CCAAAACTGTGGGACGACAGCGTGTACGGCAATAGTCTCTGTACCAAGCGGACGACGCGGCGA
CGGTATTCGCATCATCTGCGAAAAAATCGTACATTTGCCTCTCGAAATCCAGGTCTCTCGATACT
ACTGGCTCAAAATCCACAGATAAATAATTTGGCTTTGTTCACTGGTATCCAAAGCTATCGCCCTCGA
CCTTGCCAAAGCGAAATTCAGCTTCTTGAAGTCCGACCGCTGCTCTGACTATATCGATGTTGGCG
GTAAGCAGACCGAATGCGACGTTTGGCCAGACGATTGGAGCATGCGCAACGCTTGGAGCAGTTGGTAC
ACTTCCGCGGGAATAGCACATGAATTAATAACATTTTGGCTCAATCTCGGGCAGCAGAAATAGCA
CAAACTCGGTCTCGAACATCTGTACCCGAAGATATATGACTATATCATTTCTGTCAGGCGACAGAG
CCATGCTCATATCGATCAGATCTTGAAGCTGAGCGGAAACAGGAGCGCATGATCAAGCCATTTAGTGT
CTCAGAGTTGTGACCGAAATCGTCCCTTGTACTGATGGCTTCCGCAAAACATCGAGCTTAGTTTC
AGATTTGATCAAAATGACAGCGGTGATCGAAGGAAGCCCGCTTGAAGTCAACAGGTACTAATTAACATCT
GCAAGATGCTTCCAAAGCCATGACTGCAATGGTCAATCGACATCATCATAGCCAAAGCTTTTTTACC
AGTTAAGAAAATTCGGCGCATGGTGTATGCCACCTGGCGACTATGTTCTCCTATCTATAGCGACAAAT
GGTGGAGGCAATCCCGAGGCTGTGTACCCACATTTTGAACCTTCTTACGACACGAGCTCGCAACG
GTGGAACGGGCTCTGGCCCTGCTCTGTGTCATGTTGATATCAGCGGCTTGGCGGTTACATCGAGCTTAG
TTCAACTTGGGATGGAGCGGCTTGCATTTATCTCCCTCGCTTCTAAGGAACCCGTAATTTCCA
GACAGTTTTTTCGGCCGCAATAAGGCACCGCTGGAAAACGGGAGATTGTGGCACTTTTGGAGCCGATG
ACCTCCTGGGGAGGCGTATGAAACAAGATCGCCCTCTAGGATATGAGCGGTCGGTTTTTCTGATCCTT
TAATGAAATTCGGATTTGGATTTCAAAAGCAATGAAGCCGATCTGGTCAAGTGTGCAACAAAGCTCTCT
CCTGAAGATCAAGTCTTAATCCGTTGATTTAGTGTCAAGACCGGCTCCATCATCATTTGGCGAAATG
ATCTCAAAATGACCCCTTCAAGGGAGGATGTGACCGGAGCTTTATCTCCGAAGCCGATATCGTCCAG
AACTATGGCGCATGCAATCTCAACAAATCAAGACGTAG
```

Change region shown

Whole sequence  
Selected region  
from: 145694 to: 148183  
Update View

Customize view

Analyze this sequence

Run BLAST  
Pick Primers  
Highlight Sequence Features  
Find in this Sequence

Related information

BioProject  
Full text in PMC  
Gene  
Genome  
Identical GenBank Sequence  
Protein  
Protein Clusters  
PubMed  
PubMed (Weighted)  
Taxonomy

Recent activity

Turn Off Clear

- Agrobacterium tumefaciens plasmid Ti, complete sequence Nucleotide
- virA [Agrobacterium tumefaciens] Gene
- virA [Agrobacterium tumefaciens str. C58] Gene




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>MMVKVTKLYASPTVFPVLAFLVVPFCTWISNMTTTE<br/>DLVKEVASFTEDLRLSLVSEIENIGKPTVAKTHLSTGLA<br/>RVIDEYITNNDGPTFLIQQLAFLPLFVAVSTILQVQGVY<br/>ISRDGIMPSYIARNTSVAIVFANSSNSRGGDTWYQTV<br/>DQLTGRLRNGNSTRSQSLDVTHTWQQAQSHNYTTPVGT<br/>ELGGEDMETLIQSVVSLYSKGLVSLGFPFRTITVNLGL<br/>NLHRELIMWTEDVLYVSEISLNDSPFISGSIQFGRRE<br/>NSLWQICIPENCSSGVEVLRRLRYQAQPCSVIRVSGVPL</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): _____ (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<br/>(You may also specify a protein in the box to the left)<br/><input checked="" type="checkbox"/> including splice variants</li><li>• The following taxa: _____<br/>(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 _____ hits</li><li>• At most 1000 matches</li></ul> <p>Advanced options: <input type="checkbox"/> FASTA output <input type="checkbox"/> retrieve complete sequences</p> <p>allow at most _____ X sequence characters to match a conserved position in the pattern</p> <p><a href="#">match mode</a>: greedy, overlaps, no includes (for patterns, see <a href="#">help</a>)</p> <p><a href="#">randomize databases</a>: 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 strukturálních *motivů* získaných srovnáním primárních dat (sekvencí)
- **PROSITE**, <http://www.expasy.org/prosite/>

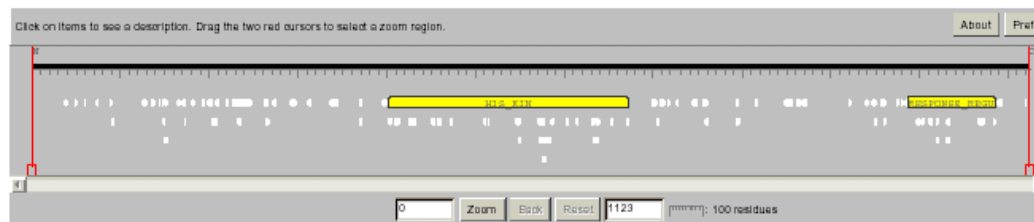
>[PDOC50109 PS50109 HIS\\_KIN](#) Histidine kinase domain [profile].

```
402 - 671 NASHDIRGALAGMEGLIDICRDGVKPGSDVDTTINQVMVCAKDLVALLNSVLEMSKIESG
      KMQLVRHDFNLSKLLLEDVIDFHPVAMKKGVVLDPHDgavEKPSNVRGDSGRKQILN
      NLVSNARVFTVD--GHIAVRWAQrpgensaavlasypkgvskfvkkmfckkkesaatye
      teianairnnaTMEFVFEVDDTGKGIHMEHRKSVPRNYVQVREtAQGHQGTGLGLGI VQ
      SLVRLMG3EIRITDKAMGeKGTCPQPNVLLTT
```

>[PDOC50110 PS50110 RESPONSE\\_REGULATORY](#) Response regulatory domain [profile].

```
987 - 1085 RVLVVDNPFISRRKVTGKLLKMGVSeVEQCDSGKEALRLVTEGLtqreeggsvdklpFDY
      IFMDQMPEMDGYRATREIRkvekSYGVRTPITAVSGHD-----
```

Graphical summary of hits (*java applet*)



98 hits with 12 PROSITE entries

|                                  |                          |                               |                            |                            |                         |                                  |
|----------------------------------|--------------------------|-------------------------------|----------------------------|----------------------------|-------------------------|----------------------------------|
| <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> |
|----------------------------------|--------------------------|-------------------------------|----------------------------|----------------------------|-------------------------|----------------------------------|

# 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:

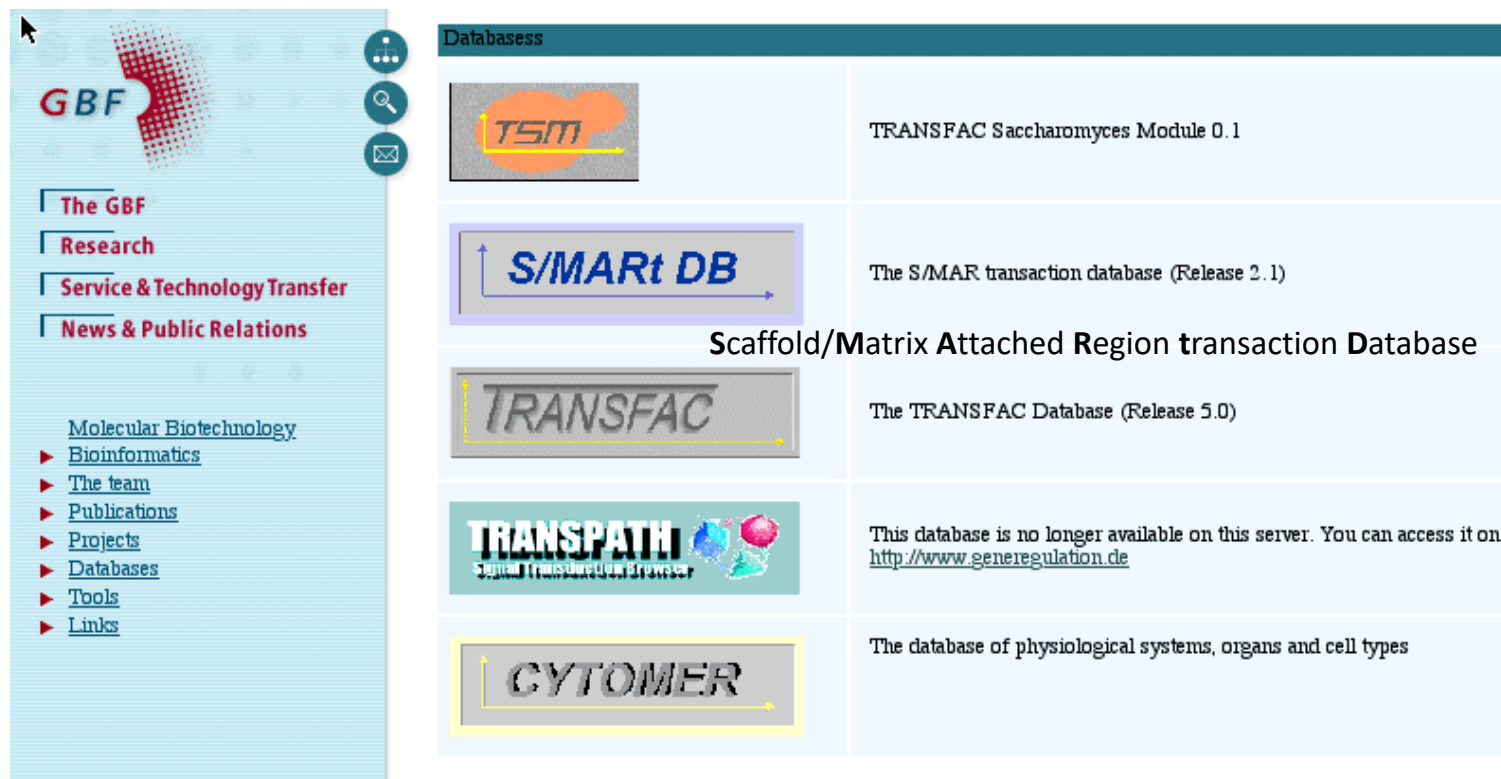
- [By accession number](#)
- [By PRINTS code](#)
- [By database code](#)
- [By text](#)
- [By sequence](#)
- [By title](#)
- [By number of motifs](#)
- [By author](#)
- [By query language](#)

#### PRINTS search:





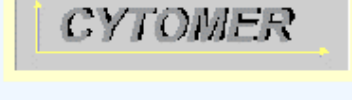
- [Search PRINTS with NEW FingerPRINTScan](#)
- [FPScan](#)
- [GRAPHScan](#)
- [MULScan](#)
- FingerPRINTScan binaries and source are available: [contact.scordis@bioinf.man.ac.uk](mailto:contact.scordis@bioinf.man.ac.uk)

# Sekundární databáze

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



The screenshot shows the website interface for TRANSFAC. On the left is a navigation menu with the GBF logo and links for '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.

| Databases  |   |
|--|---|
|    | TRANSFAC Saccharomyces Module 0.1   |
|    | The S/MAR transaction database (Release 2.1)<br><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:** Enter a PDB ID or keyword, Query Tutorial, Find a structure button, checkboxes for query by PDB id only, match exact word, and remove sequence homologues.
- PDB Mirrors:** List of mirror sites including San Diego Supercomputer Center, Rutgers University, National Institute of Standards and Technology, Cambridge Crystallographic Data Centre, UK, National University of Singapore, Osaka University, Japan, Universidade Federal de Minas Gerais, Brazil, and Max Delbrück Center for Molecular Medicine, Germany.
- News:** 23-Dec-2002 Happy Holidays from the PDB! The PDB staff wish to extend our best wishes to the community for a happy holiday season and a wonderful new year!

# Strukturální databáze

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

Structure Explorer - 1P5Y

**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](#)

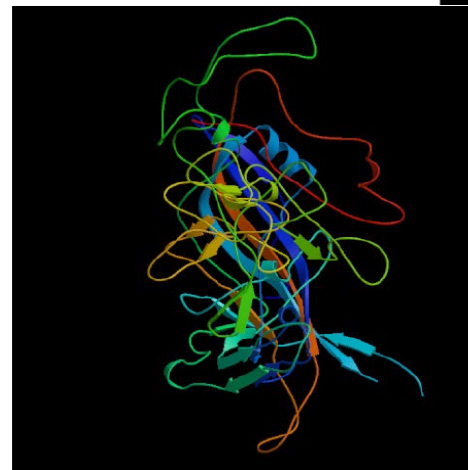
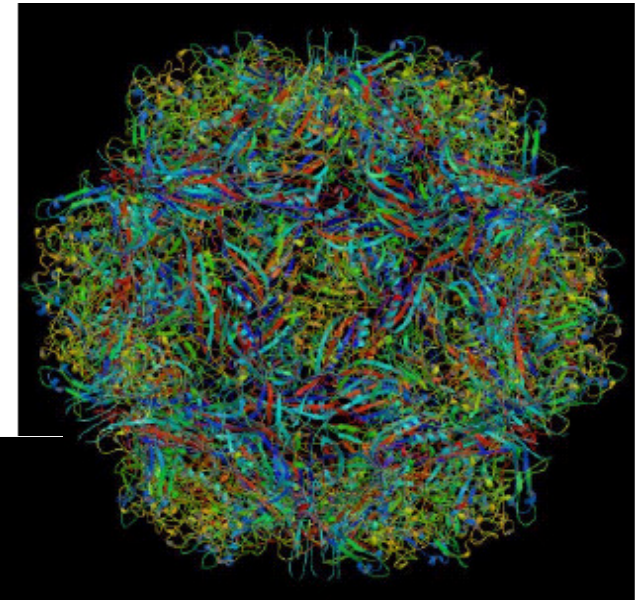
[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

Explore

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

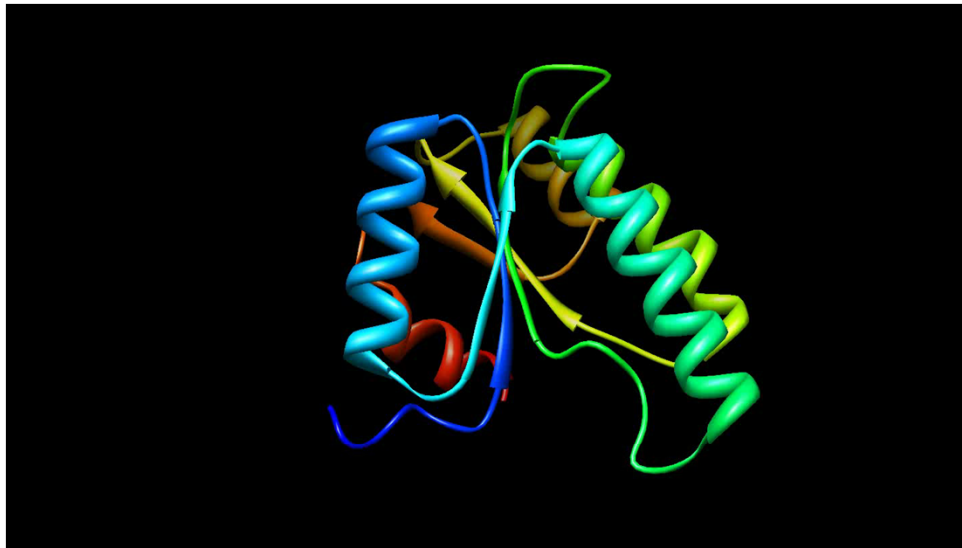


<http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=;pid=173561064349344&bio=1&opt=show&size=500>

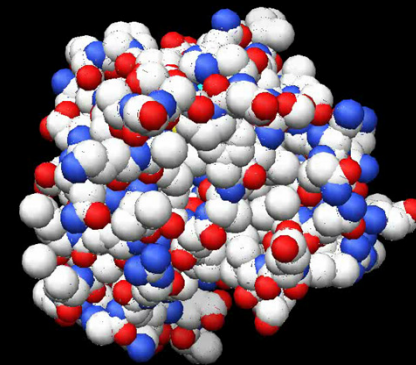
12/29/2003

# 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

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

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

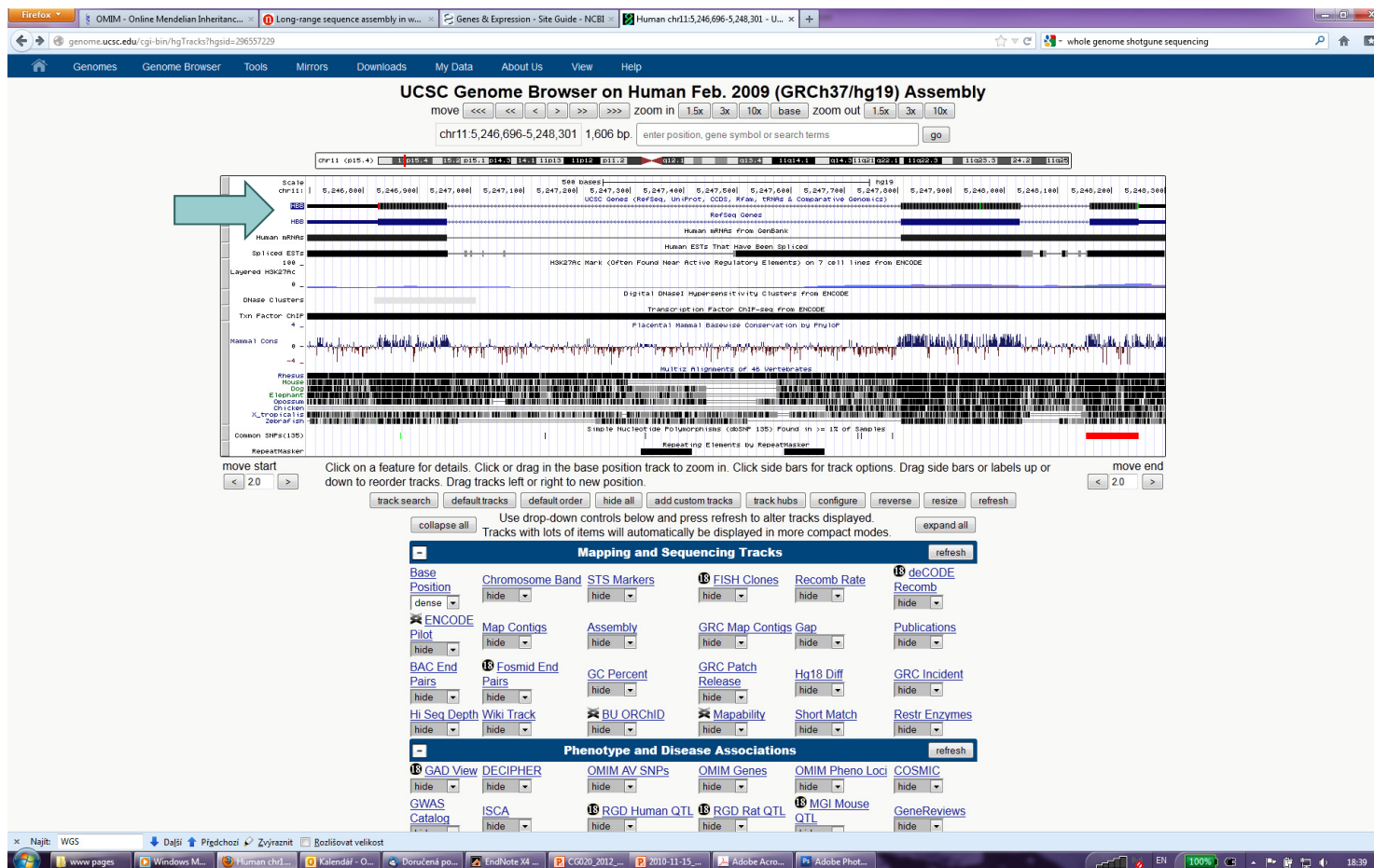
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the [User's 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<br>15q11;15q13<br>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  |

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

# 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>

Genomic Sequence Near Gene

### Get Genomic Sequence Near Gene

Note: if you would prefer to get DNA for more than one feature of this track at a time, try the [Table Browser](#) using the output format sequence.

**Sequence Retrieval Region Options:**

- Promoter/Upstream by 1000 bases
- 5' UTR Exons
- CDS Exons
- 3' UTR Exons
- Introns
- Downstream by 1000 bases
- One FASTA record per gene.
- One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra downstream (3')
- Split UTR and CDS parts of an exon into separate FASTA records

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

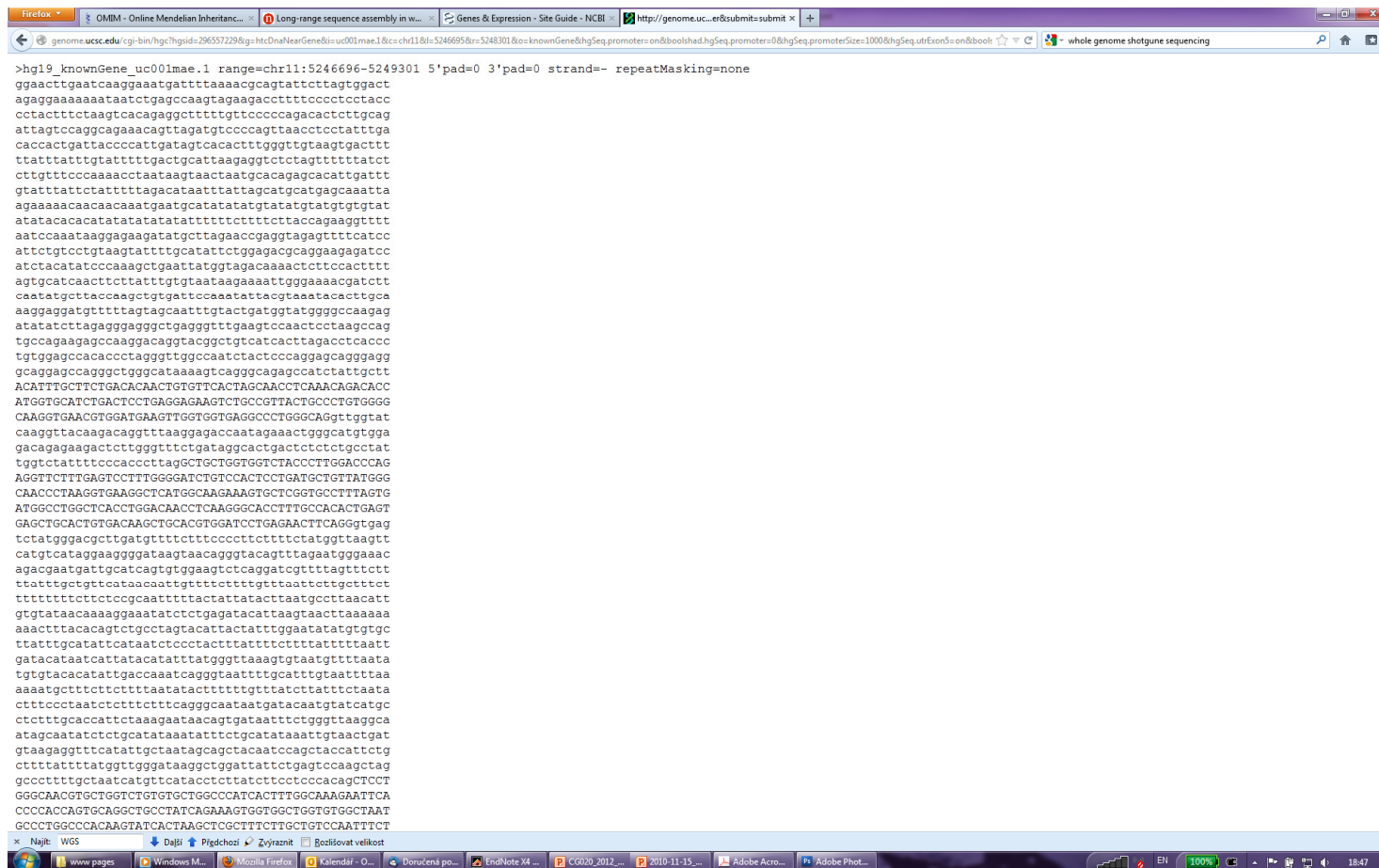
**Sequence Formatting Options:**

- Exons in upper case, everything else in lower case.
- CDS in upper case, UTR in lower case.
- All upper case.
- All lower case.
- Mask repeats:  to lower case  to N

submit

# 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 "TAIR - Home Page". 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 address bar shows "www.arabidopsis.org" and the status bar at the bottom indicates the system time as 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

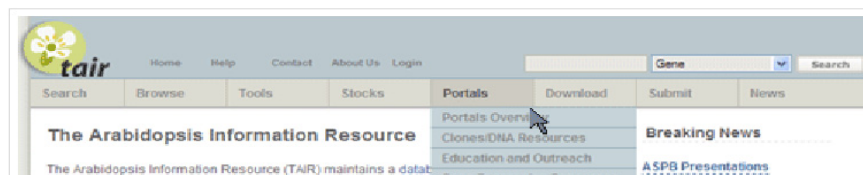
[October 15, 2006]  
Search for **genes**, **germplasms**, and **polymorphisms** using associated phenotype, and see improved phenotype data display in results and detail pages.

### 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.



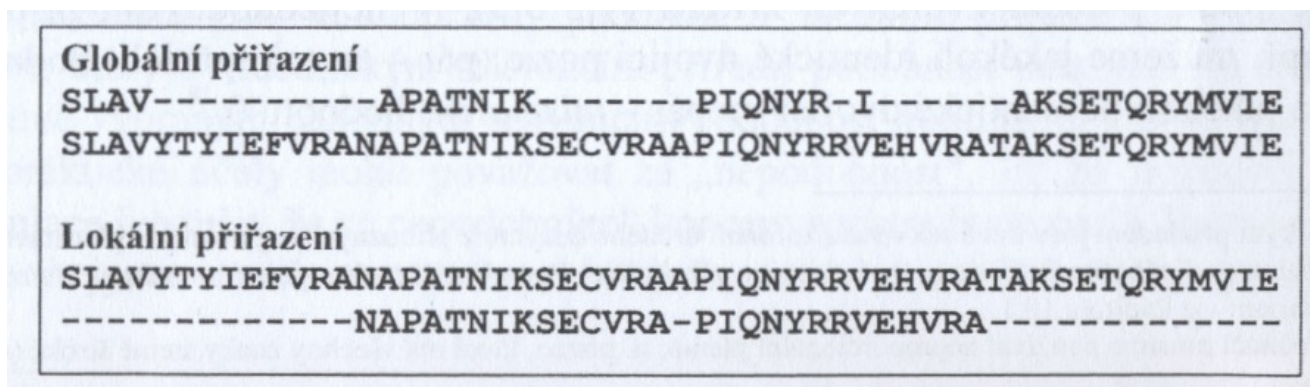
# Osnova

- 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

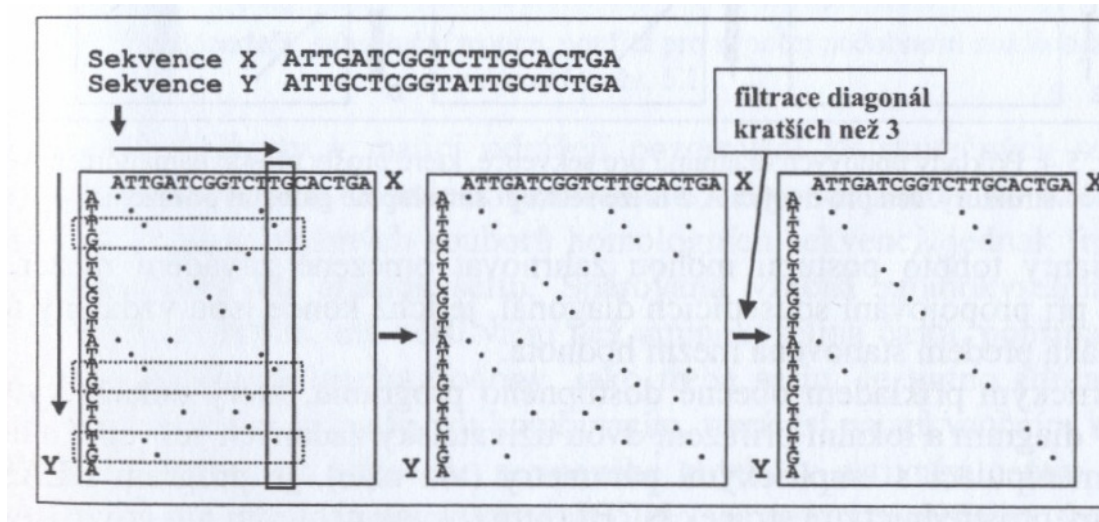


Cvrčková, Úvod do praktické bioinformatiky

- **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)

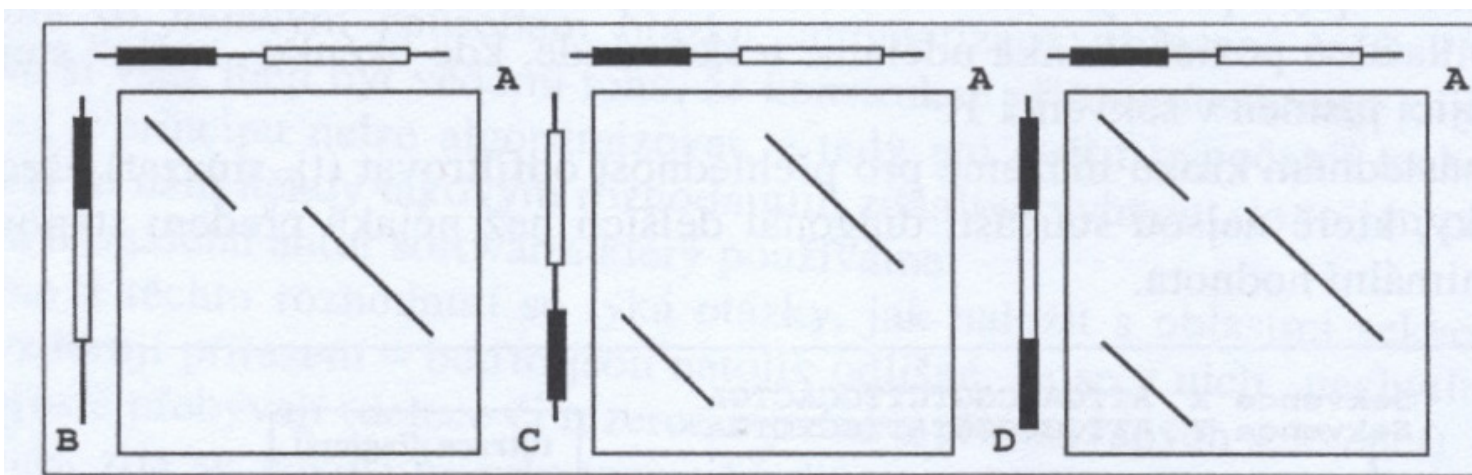


Cvrčková, Úvod do praktické bioinformatiky

- 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](#)

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

[Set subsequence](#) From:  To:

[Choose database](#) nr

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 \(145,694..148,183\)](#)  
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**

# 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  
VADAKGDSESEEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQBACKDILLF  
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.

# 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 AAATGGTCCATTATCGGCCAAGTAGCTTTC TTTAATTATAGTTAGTT  
GACAAAACACTATCAAGATATCATTATATAATAATAACATCAAGTCCATCATCTTAGCTGCCTCCTCA  
GTAGAGCCGCCAGTAAAAAAGACCGGATCAAAATAAAGCCGCCATTAAAAAATGAATTTTAGGACTCTC  
GATTGGCACGTAAGTGCCAAAACCTTCCAAATCTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC  
CAGATATGGGATATTTCTAAGTTTTATCTCTAAATTTACATCTCAACTAATATTAAGAAATTAACAGGTA  
CAGCAAATCATAAAATTTCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG  
TCTGCATGCCATATTC ACTAAGGGGTCGTTTGGTAC AAGAAATAATAATAAATTTTCGGGATAGAATTT  
GAGATTGCATTTATCTTGTGTTTTAATTATAAGTATTAGCTAATTT CAGAATAAAATTTTACTAAAATAG  
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCATAGCCACTCACATAGAATATCC  
TTATTTATCTCACTATTTTACC AAATGATCGGTTAGTCTTTCATGAGAATCCAGTATCCTCAATAAATGCA  
GTAAGAAGTTAGAAAATTTTCATTAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG  
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAACGGTT  
CGAGACACTCTTATGGAAGGCGTTTGTCTTCAAAGTAGATTCTCATTCAATTGCTCTGGTGC AATAGCAAAA  
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA  
GAGAATTTCAAATCTCTCAACTACTTTTAAGGGAATTC AAAATACGACC AATATTTATTACTTACTTAC  
TTATAGTTAAATGATATGAATTTTTATTTAAATTTGAATTGAAAAATTAATAATTACTTGTATTAATATAA



# 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
CAGCAATCATAAAATTTTCTCTAAAGAAAGACAATGAATCCGGTACTGATTCATTGGCTTTTTCAGAG
TCTGCATGCCATATTTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAATAATTTTCGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTAAATATAAGTATTAGCTAATTTTCAAGAATAAATTTTACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACC AAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAAATGCA
GTAAGAAGTTAGAAAAATTTTCATTAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTAAATAGGAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTGTCTTCAAAGTAGATTTCTCATTCATTGCTCTGGTGC AATAGC AAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTTACAACTTCTATTGTATACTCAAAATGAAAGTTTAA
GAGAACTTTTCAAATCTCTCAACTACTTTTAAAGGGAATTC AAAATACGACCAATATTTATTACTTACT
TTATAGTTAAATGATATGAATTTTAAATTTGAAATTTGAAAATATTAATTTACTTGTATTAATATAA
ACAATAGATATCGCTAAGTATTTACCACAAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT
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   gtttgtaatagttactcaatttgaattaacaaaggggcaaatttgactattttgcctta 120
```

## Frame 2, 1 stop codon

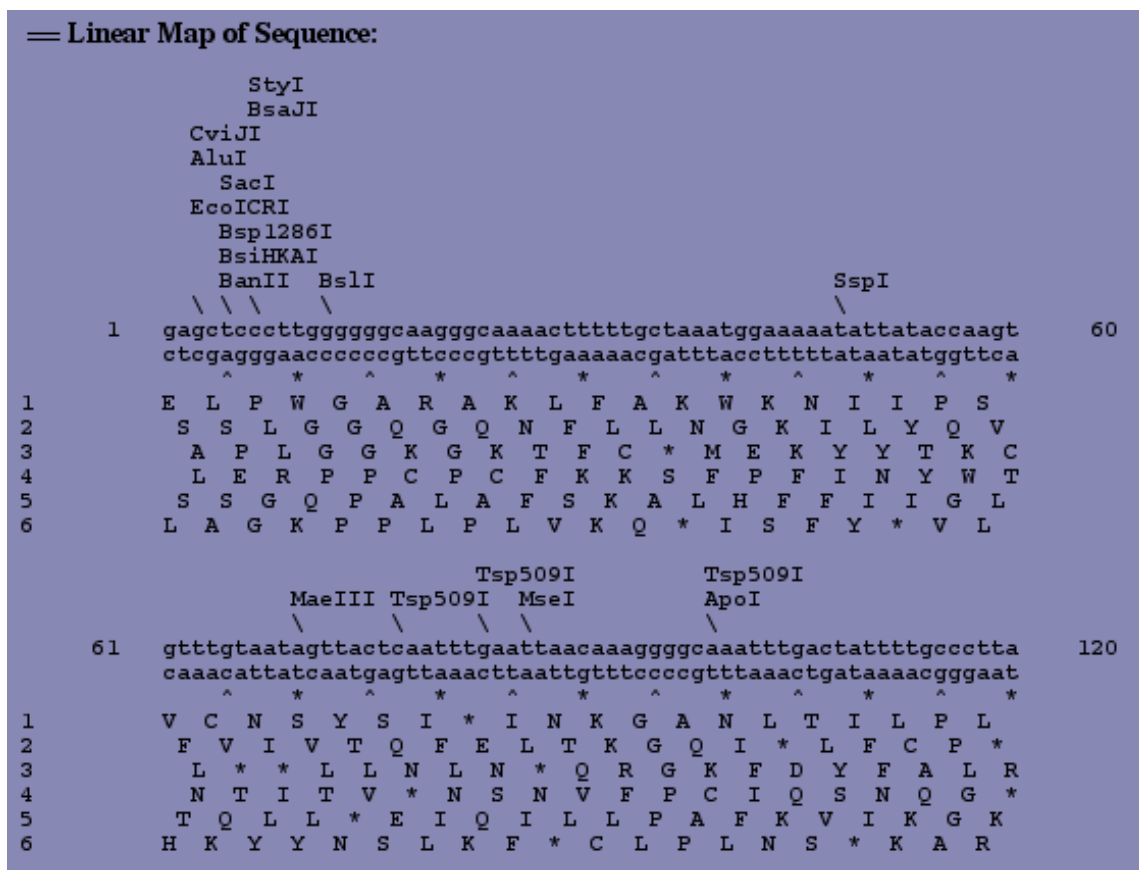
Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. Tran

```
>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   tttgtaatagttactcaatttgaattaacaaaggggcaaatttgactattttgcctta 120
```

# Analytické nástroje

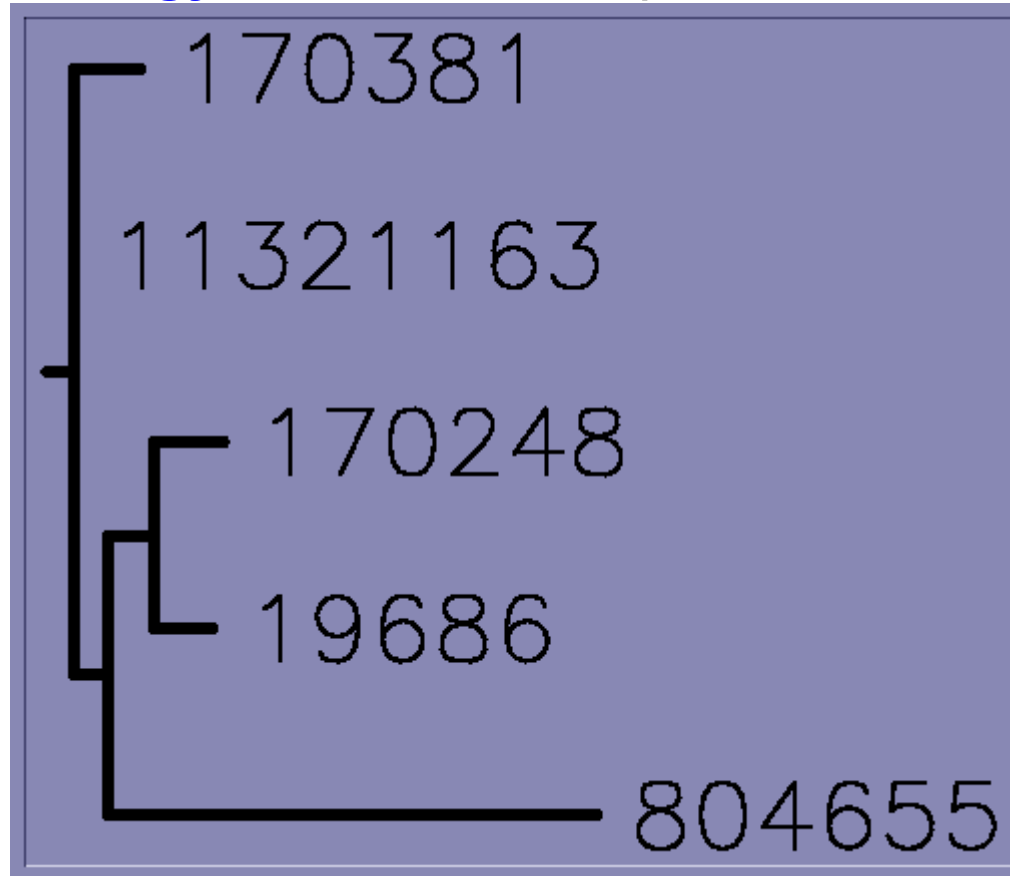
- **Biology Workbench** <http://workbench.sdsc.edu/>





# 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

PHACTR4 phosphatase and actin regulator 4 [Homo sapiens]  
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)

[See PHACTR4 in MapViewer](#)

**Genomic regions, transcripts, and products**

**Genomic Sequence** NC\_000001 chromosome 1 reference GRCh37.p5 Primary Assembly

[Go to reference sequence details](#)

[Go to nucleotide](#) [Graphics](#) [FASTA](#) [GenBank](#)

**Table of contents**

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Interactions
- General gene info
- General protein info
- Reference sequences
- Related sequences
- Additional links

**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