CG020 Genomika

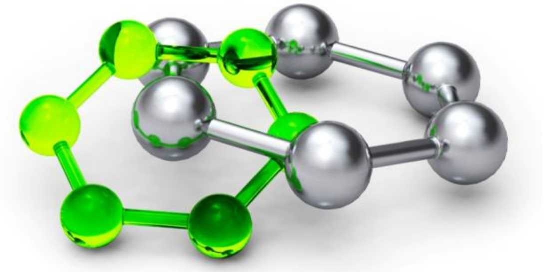
Přednáška 1

Úvod do bioinformatiky

Jan Hejátko

Funkční genomika a proteomika rostlin,
Středoevropský technologický institut (CEITEC)
a
Národní centrum pro výzkum biomolekul,
Přírodovědecká fakulta,

Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.eu



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, restrikčních 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**
  - **RNA interference a genomové editování**
- ☐ **Kapitola 06**
  - **Genová exprese a chemická genetika**
- ☐ **Kapitola 07**
  - **Protein-proteinové interakce a jejich analýza**
- ☐ **Kapitola 08**
  - **Současné metody sekvenování DNA**

Schéma předmětu

- **Kapitola 09**
  - **Struktura genomů**
- **Kapitola 10**
  - **Evoluce genomů**
- **Kapitola 11**
  - **Genomika a systémová biologie**
- **Kapitola 12**
  - **Praktické aspekty funkční genomiky**
  - **Modelové organismy**
  - **PCR**

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

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

3 : 1

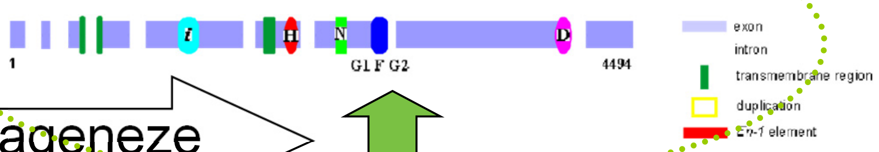
?

inzerční mutageneze

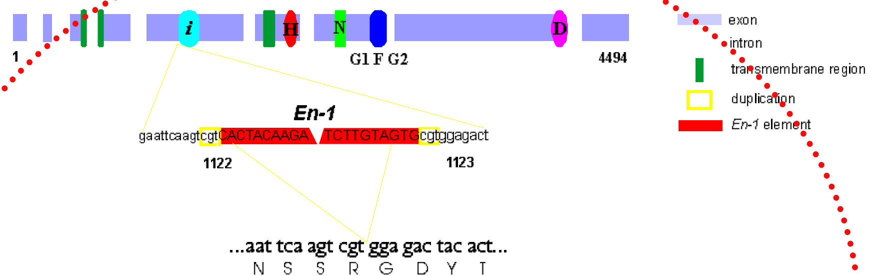
„Reverzně genetický“ přístup

5‘TTATATATATATATTAAAAAATAAAATAA AAGAACAAAAAAGAAAATAAAATA….3‘

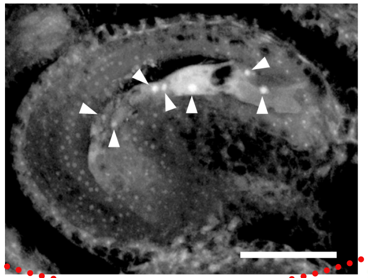
BIOINFORMATIKA



FUNKČNÍ GENOMIKA



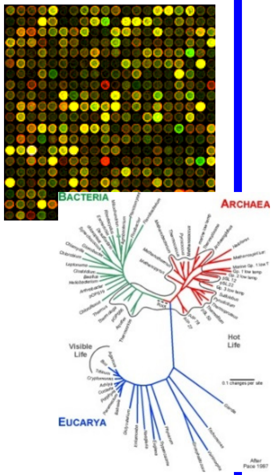
…aat tca agt cgt gga gac tac act…
N S S R G D Y T



Osnova

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

Bioinformatika



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

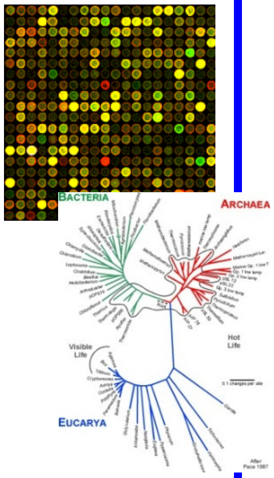
Výzkum, vývoj nebo aplikace výpočetních nástrojů a přístupů za účelem zvyšování rozvoje využití biologických, lékařských, dat o chování nebo zdraví, včetně těch, které umožňují taková data získávat, ukládat, organizovat, archivovat, analyzovat nebo vizualizovat.

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

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

Množství informace v DNA

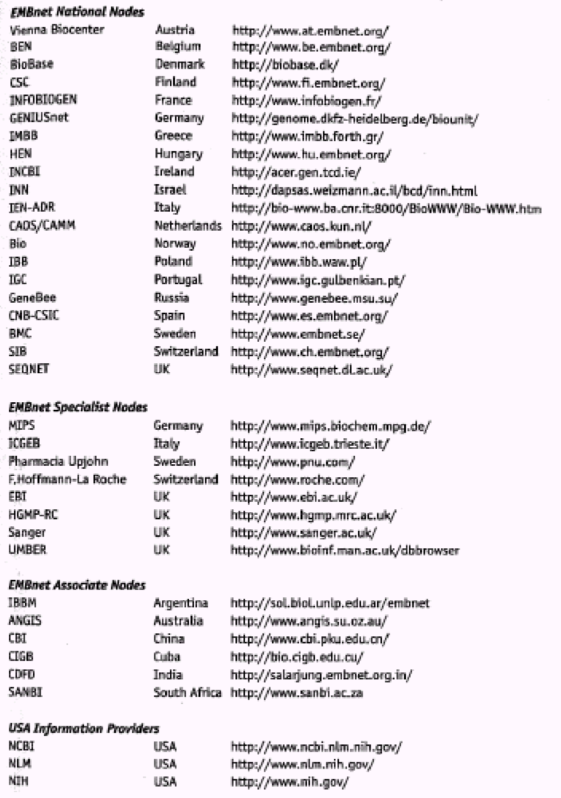
- Every **bp**= **4 bits**
- Human genome = ~**3 billion bp**
    - $= 4 \times 3 \times 10^9$
    - $=1.2 \times 10^{10}$ bits
    - $=1.5 \times 10^9$ bytes (**1.5 GB**)
- This amount of information is contained in a cell nucleus with 10µm diameter

- There is ~**2m** of DNA in **every somatic human cell**
    - **Each human** in composed of about **10^{12} cells**
    - Thus every human contains 2 × 1 **10^{12}** of DNA
        - **$=2 \times 10^9$km** of DNA
    - Distance from the sun to Uranus = 2.8 × 109km
    - Each single human contains enough DNA to stretch from the sun to Uranus

Osnova

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

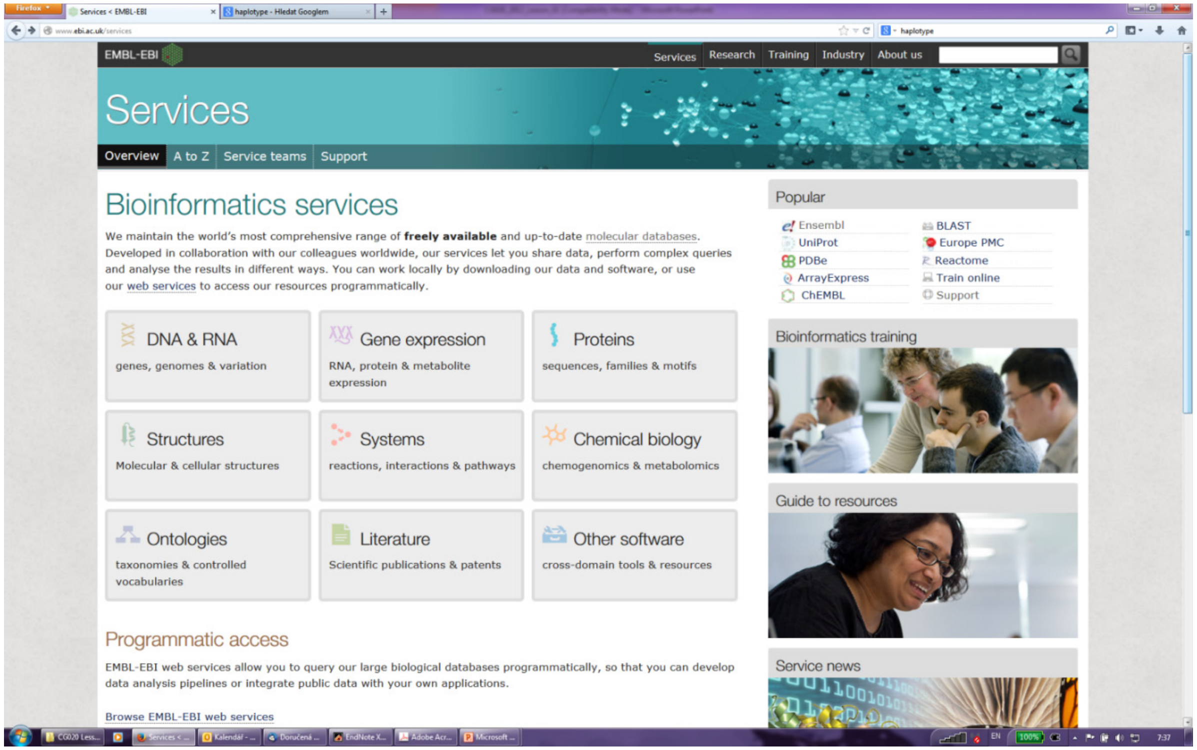
Spektrum on-line zdrojů



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

Spektrum on-line zdrojů

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



Bioinformatics services

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.

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

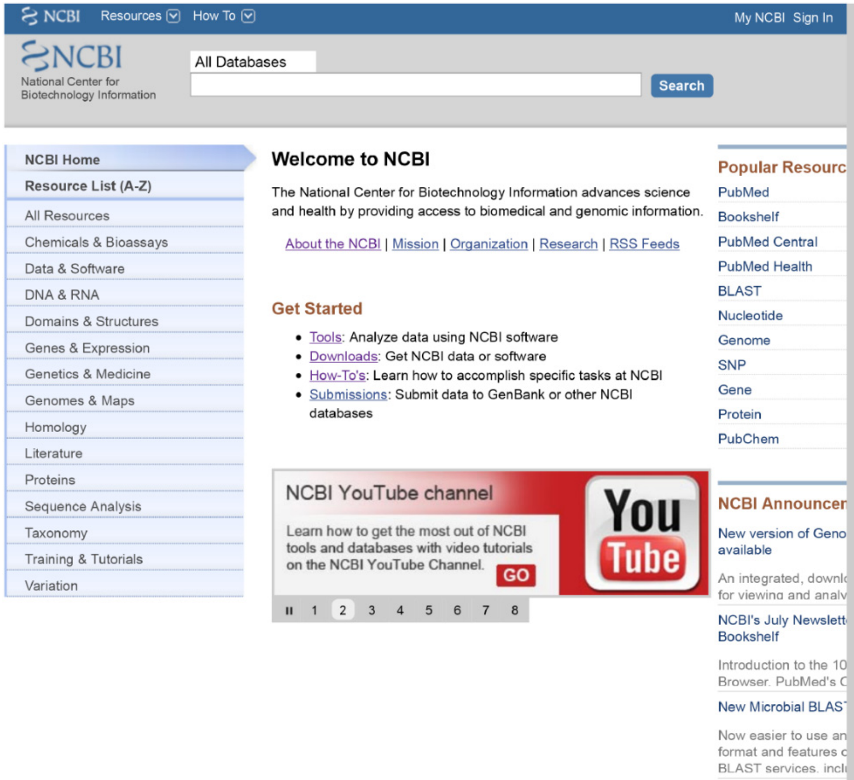
Programmatic access

EMBL-EBI web services allow you to query our large biological databases programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications.

Browse EMBL-EBI web services

Spektrum on-line zdrojů

- NCBI http://www.ncbi.nlm.nih.gov/



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

NCBI YouTube channel

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel.

Popular Resourc

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

NCBI Announcem

New version of Geno available

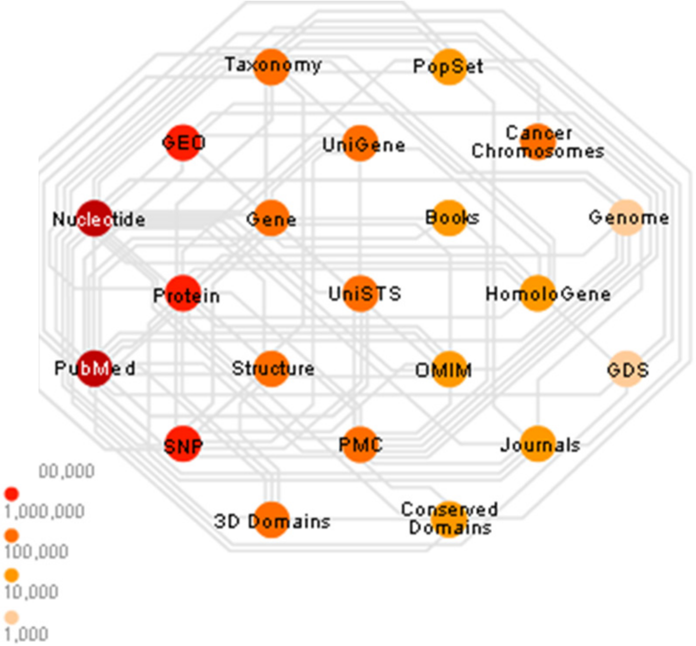
An integrated, downl for viewing and analy

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, incl



Osnova

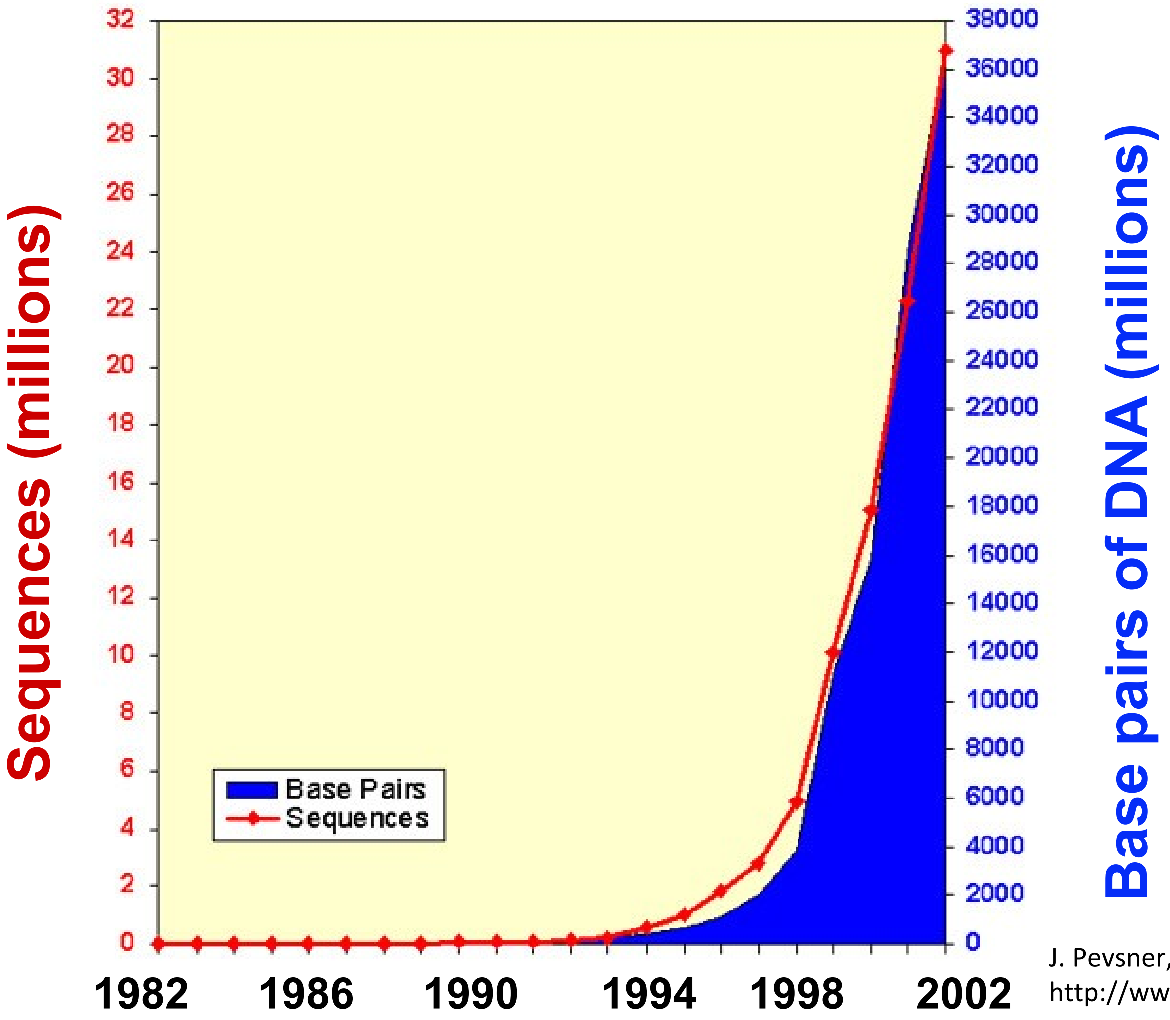
- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
    - Spektrum „on-line“ zdrojů
    - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze

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

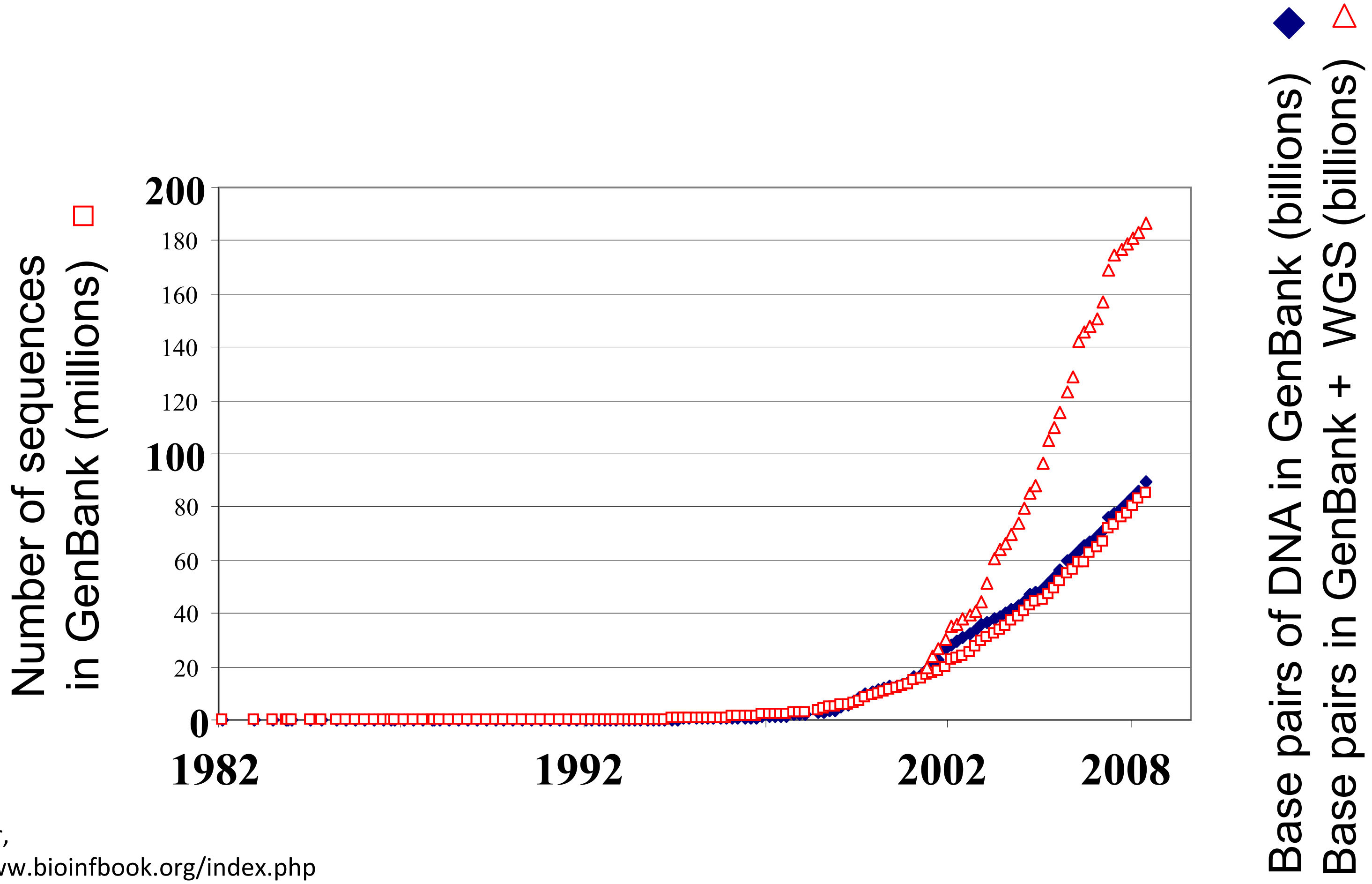
- denně vzájemná výměna a zálohování dat
- velká datová náročnost (kapacita i software)

Growth of GenBank



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

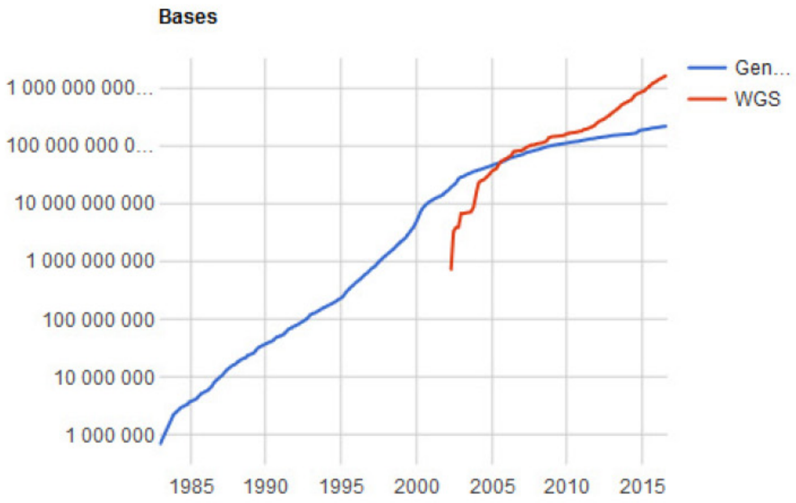
Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases

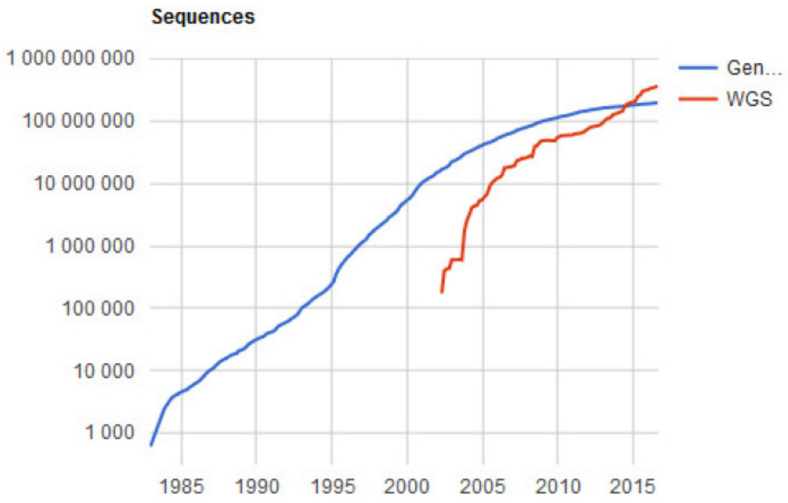


J. Pevsner,
http://www.bioinfbook.org/index.php

Growth of GenBank

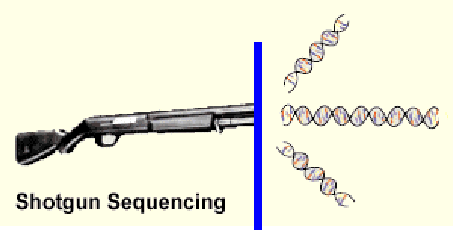
Aug 2016


Bases


Sequences

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

WGS



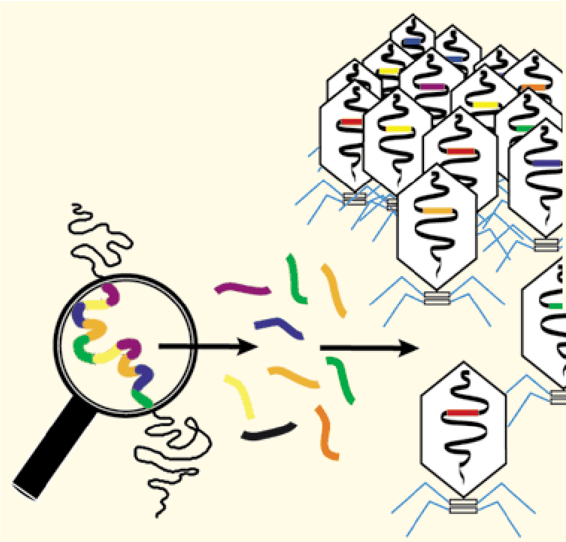


Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

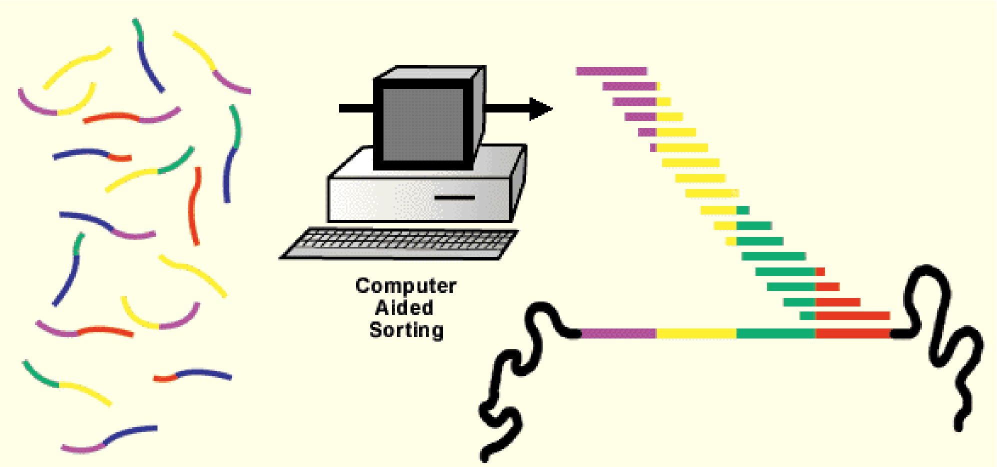
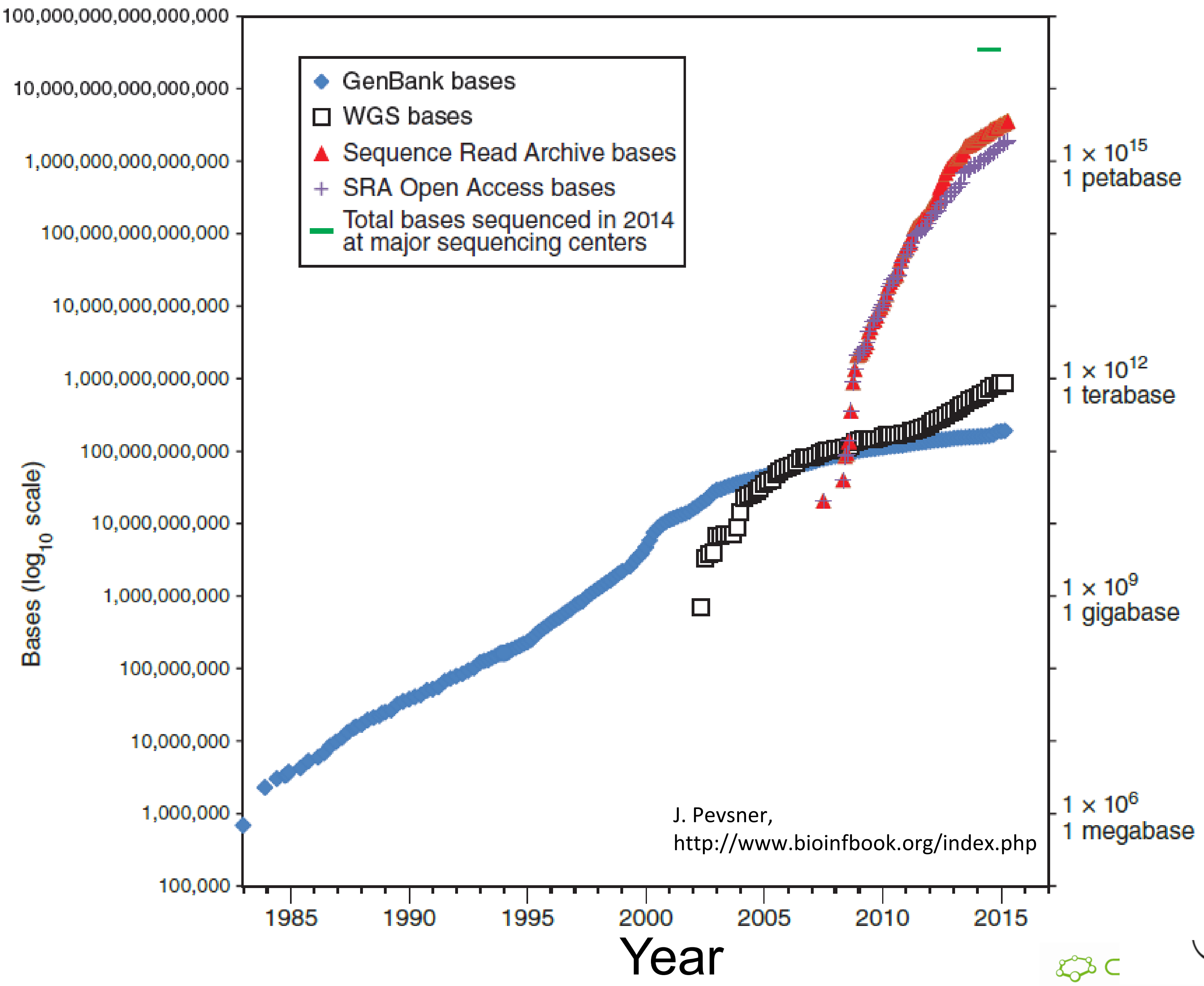


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

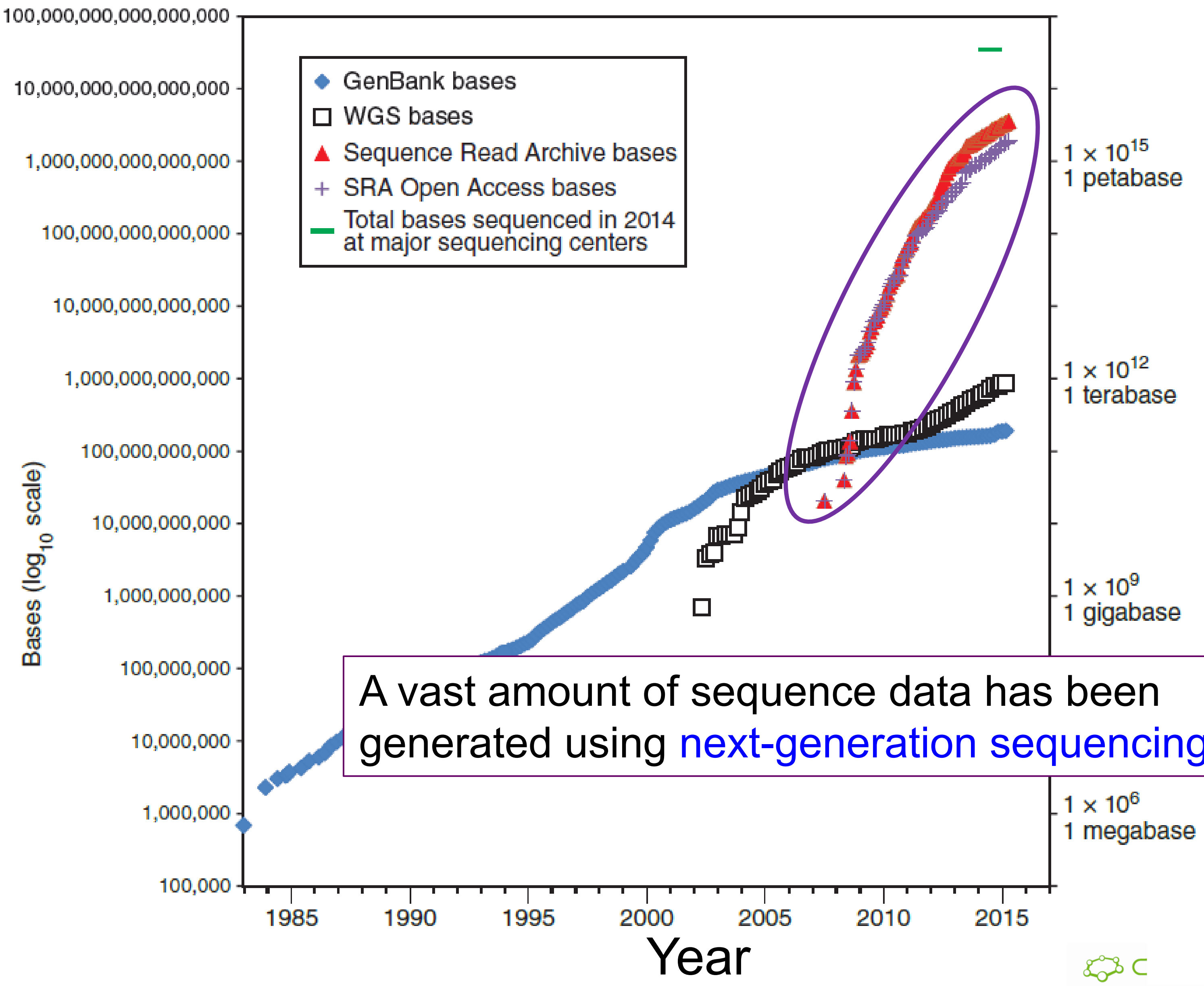
Interactive concepts in biochemistry, Rodney Boyer, Wiley, 2002, http://www.wiley.com//college/boyer/0470003790/

Growth of DNA Sequence in Repositories



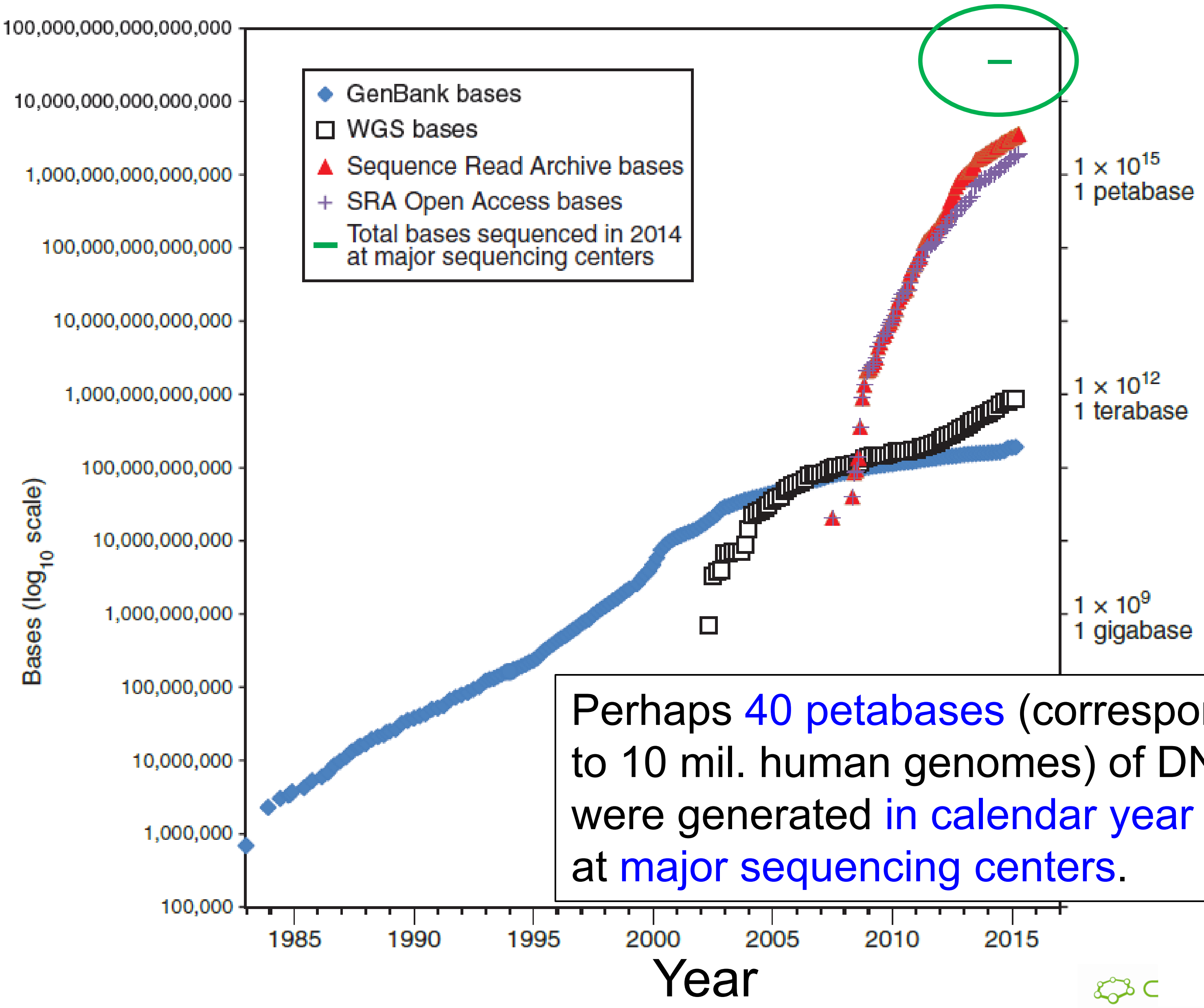
J. Pevsner,
http://www.bioinfbook.org/index.php

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.

B&FG 3e
Fig. 2-3
Page 22

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

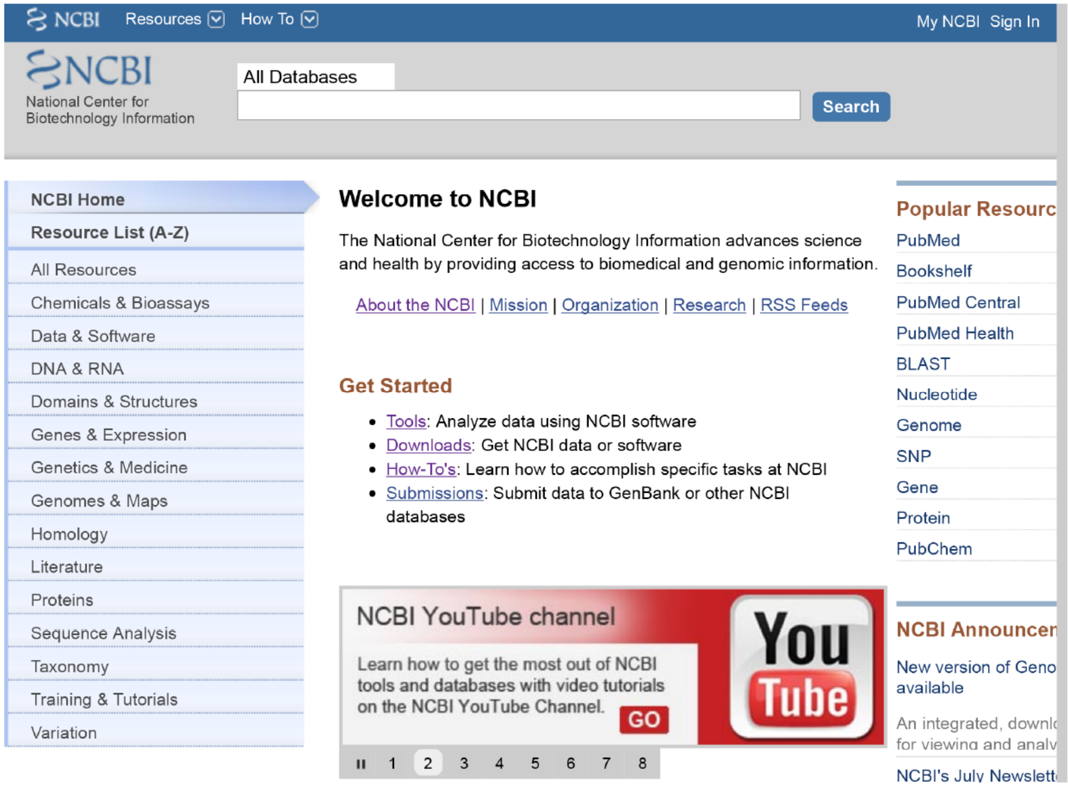
- Typy sekvencí v primárních databázích
  - Standardní nukleotidové sekvence získané kvalitním sekvencováním
  - **EST**s (**E**xpressed **S**equence **T**ags)
  - **HGTS** (**H**igh **T**hroughput **G**enome **S**equencing)

    - neanotované „surové“ výsledky sekvenačních projektů
  - Referenční sekvence anotovaných genomů
  - **TPA**s (**T**hird **P**arty **A**nnotation)

    - sekvence anotované jinými než původními autory

Primární databáze

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



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

NCBI YouTube channel

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel.

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

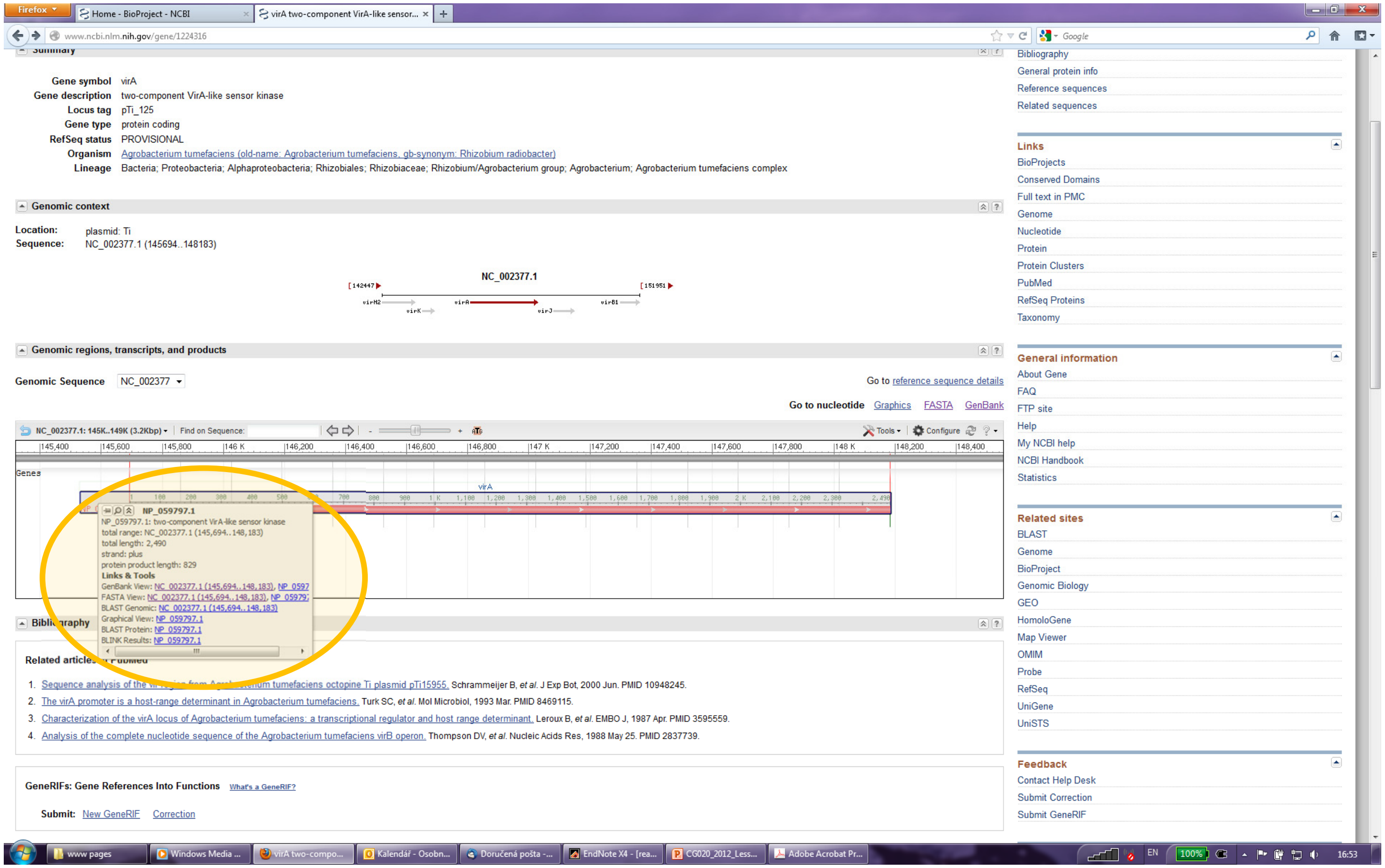
NCBI Announcen

New version of Geno available

An integrated, downl for viewing and analy

NCBI's July Newslett

Primární databáze



Gene symbol	virA
Gene description	two-component VirA-like sensor kinase
Locus tag	pTi_125
Gene type	protein coding
RefSeq status	PROVISIONAL
Organism	Agrobacterium tumefaciens (old-name: Agrobacterium tumefaciens, gb-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

Go to reference sequence details

Go to nucleotide Graphics FASTA GenBank

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_0597
FASTA View: NC_002377.1 (145,694..148,183), NP_05979
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

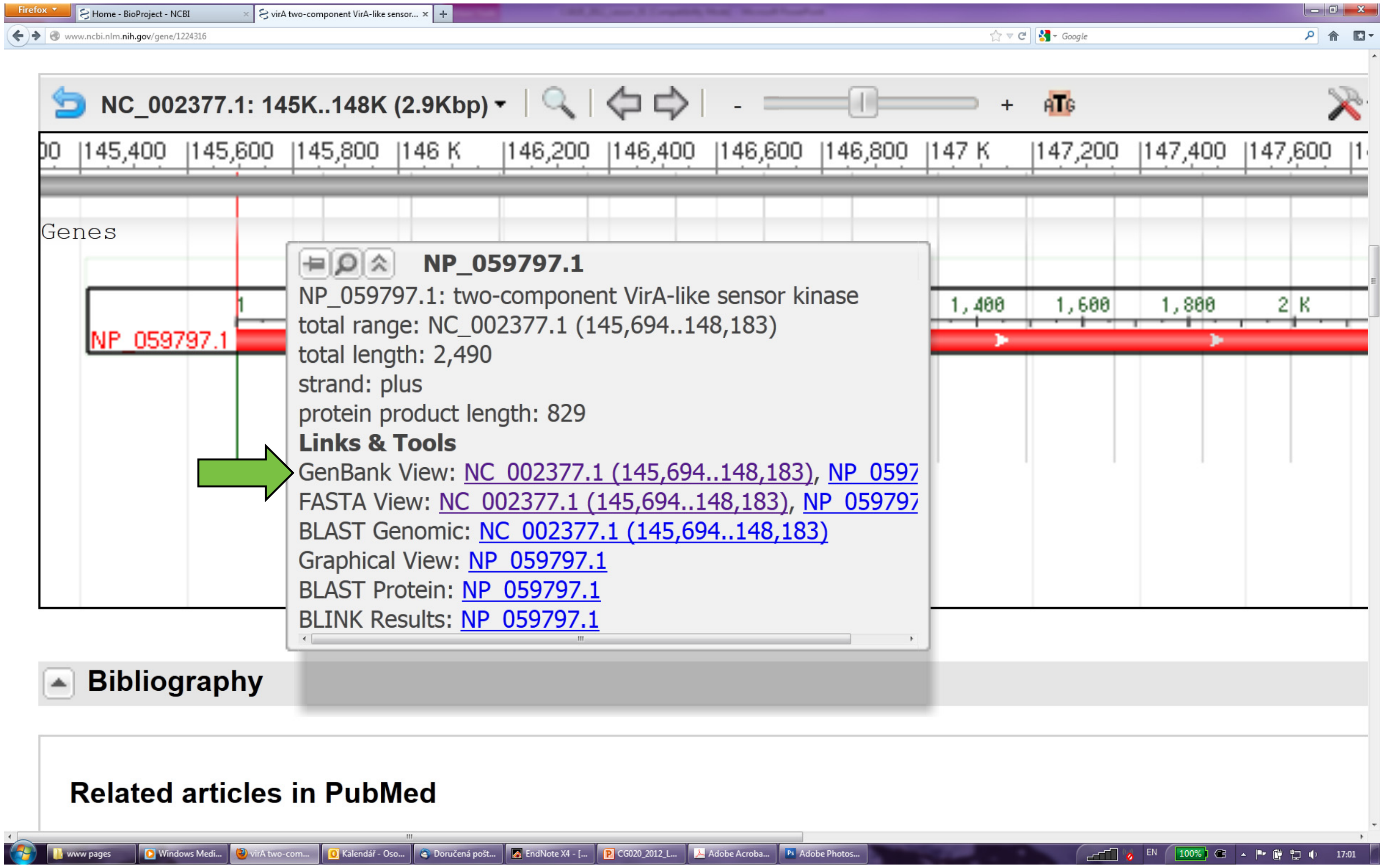
Related articles in PubMed

1. Sequence analysis of the vir-region from Agrobacterium tumefaciens octopine Ti plasmid pTi15955. 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 virB 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

Primární databáze



NP_059797.1

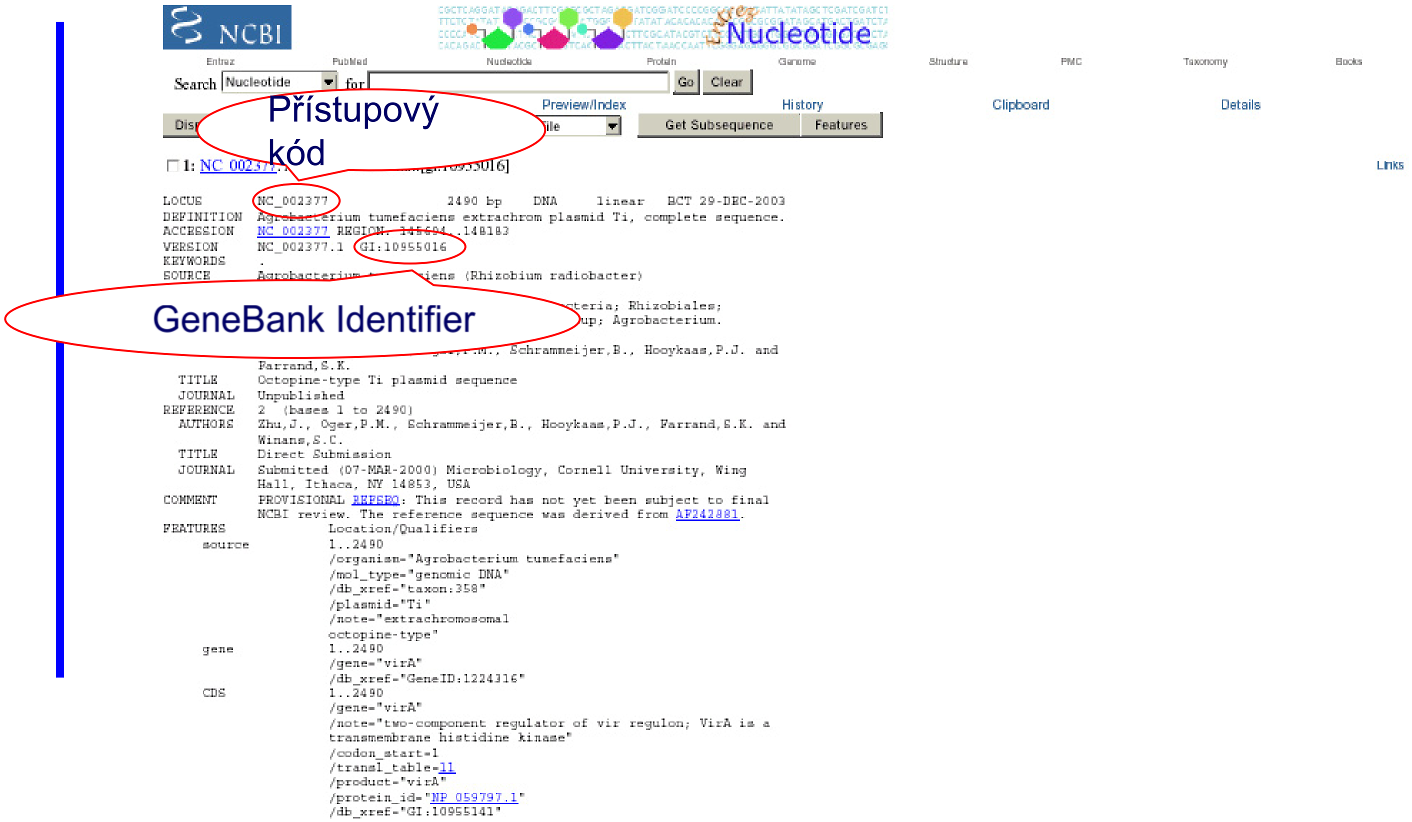
NP_059797.1: two-component VirA-like sensor kinase
total range: NC_002377.1 (145,694..148,183)
total length: 2,490
strand: plus
protein product length: 829

Links & Tools

GenBank View: NC_002377.1 (145,694..148,183), NP_0597
FASTA View: NC_002377.1 (145,694..148,183), NP_059797
BLAST Genomic: NC_002377.1 (145,694..148,183)
Graphical View: NP_059797.1
BLAST Protein: NP_059797.1
BLINK Results: NP_059797.1

Related articles in PubMed

Primární databáze



Přístupový kód

GeneBank Identifier

```
☐ 1: NC_00237                  [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         iens (Rhizobium radiobacter)
                                         teria; Rhizobiales;
                                         up; Agrobacterium.

                       ,P.M., Schrammeijer,B., Hooykaas,P.J. and
            Farrand,S.K.
  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="MNGRYSPTRQDFKTGAKPWSILALIVAAMIFAFMAVASWQDNAT
                     TQAILSQLRSINADSASLQRDVLRAHTGTVANYRPIISRLGALRKNLEDLKQLFRQSH
                     IVSESNAAQLLRQLEVSLNSADAAVAAFGAQNVRLQDSLASFTRALSSLPGKASTDQT
                     LEKPTELASMMLQFLRQPSPAISFEISLELERLQKQRGLDEAPVRILAREGPIILSLL
                     PQVKDLVNMIQTSDTAEIAEMLQRECLEVYSLKNVEERSARIFLGSASVGLCLYIITL
                     VYRLRKKTDWLARRLDYEELIKEIGVCFEGEAATTSSAQAALRIIQRFFDADTCALAL
                     VDHDRRWAVETFGAKHPKPVWDDSVLREIVSRTKADERATVFRIISSKKIVHLPLEIP
                     GLSILLAHKSTDKLIAVCSLGYQSYRPRPCQGEIQLLELATACLCHYIDVRRKQTECD
                     VLARRLEHAQRLEAVGTLAGGIAHEFNNILGSILGHAELAQNSVSRTSVTRRYIDYII
                     SSGDRAMLIIDQILTLSRKQERMIKPFSVSELVTEIAPLLRMALPPNIELSFRFDQMQ
                     SVIEGSPLELQQVLINICKNASQAMTANGQIDIIISQAFLPVKKILAHGVMPPGDYVL
                     LSISDNGGGIPEAVLPHIFEPFFTTRARNGGTGLGLASVHGHISAFAGYIDVESTVGH
                     GTRFDIYLPPSSKEPVNPDSFFGRNKAPRGNGEIVALVEPDDLLREAYEDKIAALGYE
                     PVGFRTFNEIRDWISKGNEADLVMVDQASLPEDQSPNSVDLVLKTASIIIGGNDLKMT
                     LSREDVTRDLYLPKPISSRTMAHAILTKIKT"
ORIGIN
        1 atgaacggaa gatattcacc gacgcggcag gattttaaga caggcgcgaa gccttggtct
       61 atattggccc ttatcgttgc tgcaatgatt ttcgcgttca tggcggttgc gtcctggcag
      121 gacaatgcga ctacccaggc aatcctcagc caactacgat cgattaacgc cgacagcgcc
      181 tcactgcagc gcgatgtact ccgcgctcac acgggcaccg tggcgaacta ccgccccatt
      241 atctccaggc tgggagctct gcggaagaat ctggaagatt tgaagcaatt atttagacaa
      301 tctcatattg taagtgagag caatgctgct caactgctac gccagctaga agtgtctcta
      361 aattcggctg acgcggcggt cgccgccttt ggtgcgcaaa atgtacgcct gcaagattcg
      421 ctggccagtt tcactcgtgc tttgagcagt cttccaggaa aagcctcaac cgatcagact
      481 ttagaaaaac caacagaatt ggctagcatg atgctccaat ttcttcggca accaagcccg
      541 gctatttcat tcgagatcag ccttgaacta gagaggctcc aaaaacaacg cggtcttgat
      601 gaagctcccg tgcgcatact tgcacgtgaa ggtcccatta tcttatcgct tttgccacag
      661 gtgaaagatc tggtgaacat gattcagacg tctgacaccg cagaaattgc ggagatgctg
      721 cagcgcgagt gtttggaggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc
      781 tttcttgggt ccgcttcagt gggtctttgc ctctacatca tcaccttagt ctataggcta
      841 cgcaaaaaaa ccgattggtt agcgcggcgt ttagattacg aagagctaat caaagagatc
      901 ggagtatgtt ttgaaggtga ggcggccacc acgtcgtccg cgcaagctgc acttcgtatt
      961 attcagcgct tctttgatgc cgatacgtgc gcgttagctc tagtggacca tgaccgtaga
     1021 tgggctgtcg aaacattcgg tgcgaaacac ccaaaacctg tgtgggacga cagcgtgcta
     1081 cgcgaaatag tctctcgtac caaagcggac gaacgggcga cggtattccg catcatatcg
     1141 tcgaaaaaaa tcgtacattt gcctctcgaa attccaggtc tctcgatact actggctcac
     1201 aaatccacag ataaactaat tgcggtttgt tcactgggtt accaaagcta tcgccctcga
     1261 ccttgccaag gcgaaattca gcttcttgaa ctcgccaccg cctgcctctg tcactatatc
     1321 gatgttcggc gtaagcagac cgaatgcgac gttttggcca gacgattgga gcatgcgcaa
     1381 cgccttgagg cagttggtac acttgccggc ggaatagcac atgaatttaa taacattttg
     1441 ggctcaatcc tcgggcacgc agaattagca caaaactcgg tgtctcgaac atctgtcacc
     1501 cgaagatata ttgactatat catttcgtca ggcgacagag ccatgctcat tatcgatcag
     1561 atcttgacgc tgagccgaaa acaggagcgc atgatcaagc catttagtgt ctcagagctt
     1621 gtgaccgaaa tcgctccctt gctacgtatg gctcttccgc caaacatcga gcttagtttc
     1681 agatttgatc aaatgcagag cgtgatcgaa ggaagcccgc ttgaacttca acaggtacta
     1741 attaacatct gcaagaatgc ttcccaagcc atgactgcaa atggtcaaat cgacatcatc
     1801 atcagccaag ctttttttacc agttaagaaa attctggcgc atggtgttat gccacctggc
     1861 gactatgttc tcctatctat tagcgacaat ggtggaggca ttcccgaggc tgtgttaccc
     1921 cacattttttg aacccttctt tacgacacga gctcgcaacg gtggaacggg tctcggcctt
     1981 gcttctgtgc atggtcatat cagcgcgttt gcgggttaca tcgacgttag ttcaactgtt
     2041 gggcatggga cgcgctttga catttatctc cctccgtctt ctaaggaacc cgtaaatcca
     2101 gacagttttt tcggccgcaa taaggcaccg cgtggaaacg gggagattgt ggcacttgtt
     2161 gagcccgatg acctcctgcg ggaggcgtat gaagacaaga tcgccgctct aggatatgag
     2221 ccggtcggtt ttcgtacctt taatgaaatt cgcgattgga tttcaaaagg caatgaagcc
     2281 gatctggtca tggtcgacca agcgtctctt cctgaagatc aaagtcctaa ttccgtggat
     2341 ttagtgctca agaccgcctc catcatcatt ggcggaaatg atctcaaaat gaccctttca
```

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	**DNA**
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	**RNA**
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	**Protein**
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

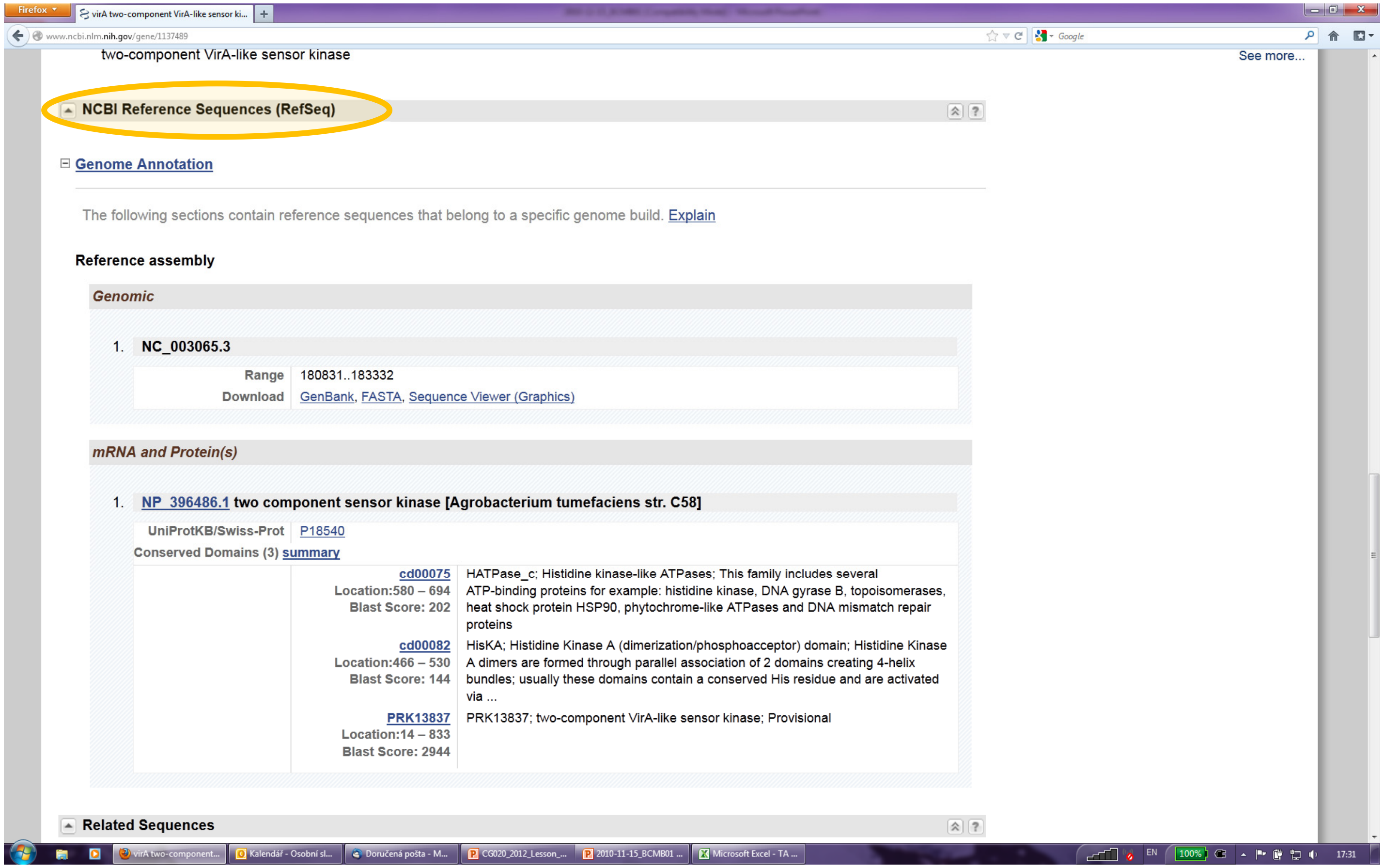
RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon "reference" version of a sequence.

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

RefSeq



two-component VirA-like sensor kinase

See more...

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

cd00075 **Location:580 – 694** **Blast Score: 202**	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
cd00082 **Location:466 – 530** **Blast Score: 144**	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 ...
PRK13837 **Location:14 – 833** **Blast Score: 2944**	PRK13837; two-component VirA-like sensor kinase; Provisional

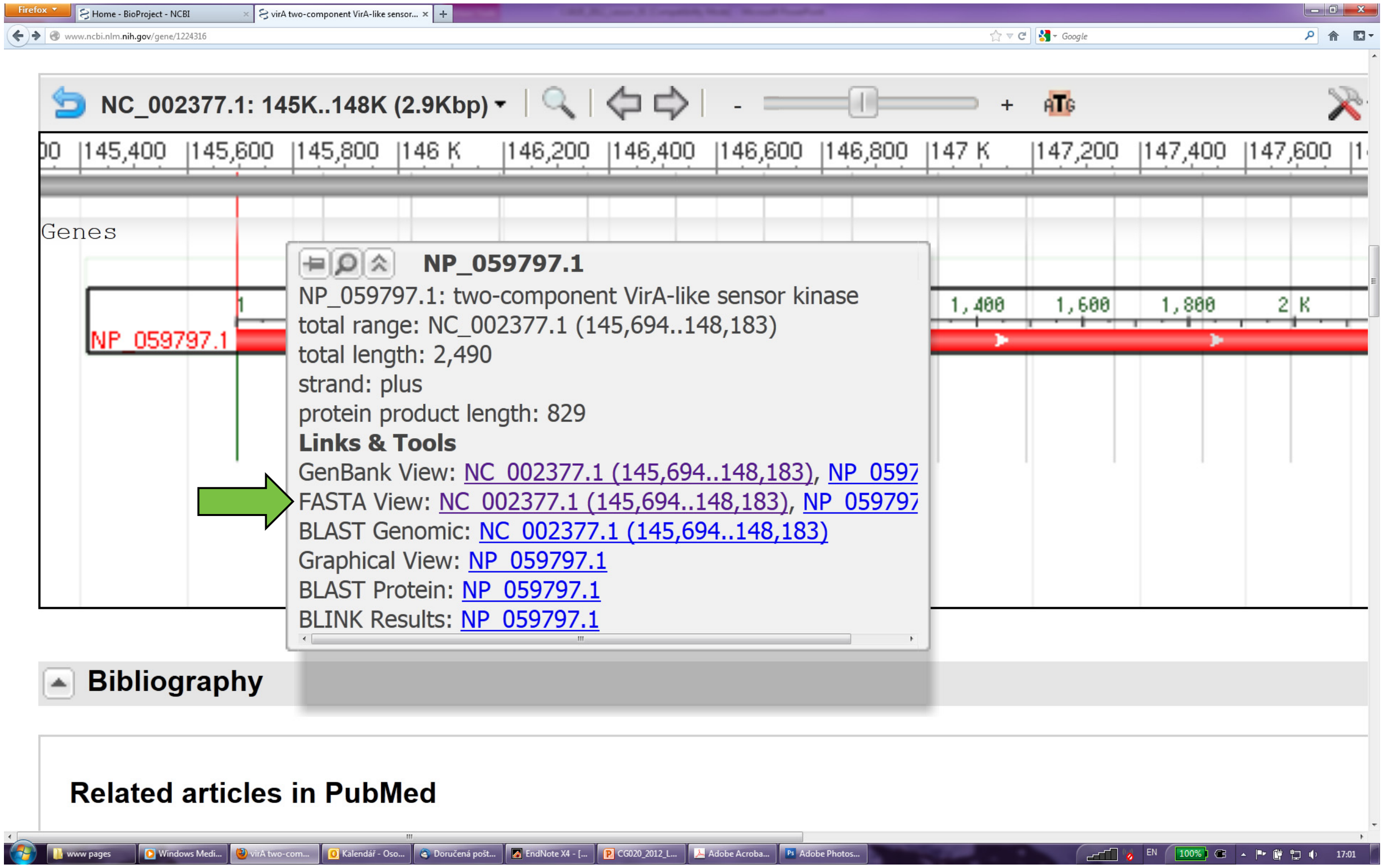
Related Sequences

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

Accession	Molecule	Method	Note
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



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_0597

FASTA View: NC_002377.1 (145,694..148,183), NP_059797

BLAST Genomic: NC_002377.1 (145,694..148,183)

Graphical View: NP_059797.1

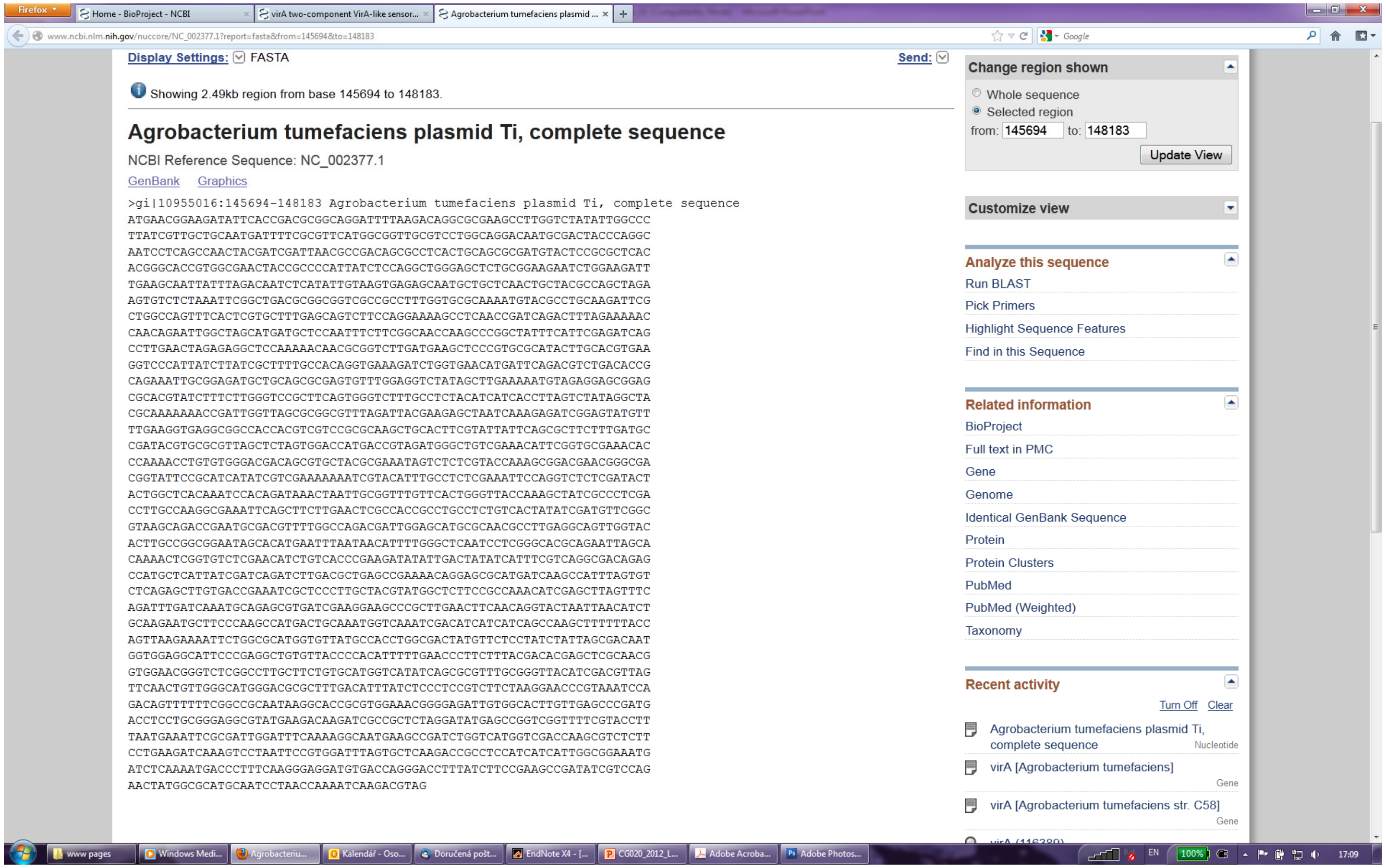
BLAST Protein: NP_059797.1

BLINK Results: NP_059797.1

Bibliography

Related articles in PubMed

Primární databáze



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

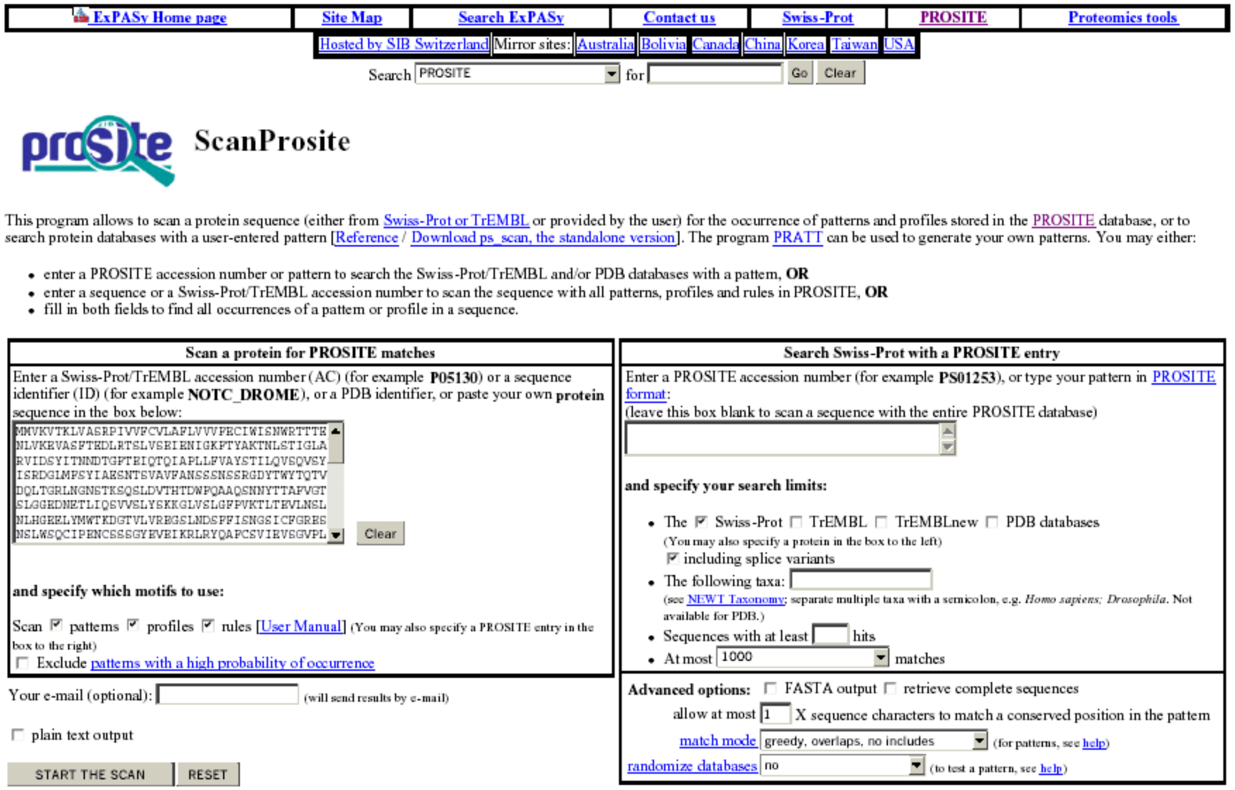
```
>gi|10955016:145694-148183 Agrobacterium tumefaciens plasmid Ti, complete sequence
ATGAACGGAAGATATTCACCGACGCGGCAGGATTTTAAGACAGGCGCGAAGCCTTGGTCTATATTGGCCC
TTATCGTTGCTGCAATGATTTTCGCGTTCATGGCGGTTGCGTCCTGGCAGGACAATGCGACTACCCAGGC
AATCCTCAGCCAACTACGATCGATTAACGCCGACAGCGCCTCACTGCAGCGCGATGTACTCCGCGCTCAC
ACGGGCACCGTGGCGAACTACCGCCCCATTATCTCCAGGCTGGGAGCTCTGCGGAAGAATCTGGAAGATT
TGAAGCAATTATTTAGACAATCTCATATTGTAAGTGAGAGCAATGCTGCTCAACTGCTACGCCAGCTAGA
AGTGTCTCTAAATTCGGCTGACGCGGCGGTCGCCGCCTTTGGTGCGCAAAATGTACGCCTGCAAGATTCG
CTGGCCAGTTTCACTCGTGCTTTGAGCAGTCTTCCAGGAAAAGCCTCAACCGATCAGACTTTAGAAAAAC
CAACAGAATTGGCTAGCATGATGCTCCAATTTCTTCGGCAACCAAGCCCGGCTATTTCATTCGAGATCAG
CCTTGAACTAGAGAGGCTCCAAAAACAACGCGGTCTTGATGAAGCTCCCGTGCGCATACTTGCACGTGAA
GGTCCCATTATCTTATCGCTTTTGCCACAGGTGAAAGATCTGGTGAACATGATTCAGACGTCTGACACCG
CAGAAATTGCGGAGATGCTGCAGCGCGAGTGTTTGGAGGTCTATAGCTTGAAAAATGTAGAGGAGCGGAG
CGCACGTATCTTTCTTGGGTCCGCTTCAGTGGGTCTTTGCCTCTACATCATCACCTTAGTCTATAGGCTA
CGCAAAAAAACCGATTGGTTAGCGCGGCGTTTAGATTACGAAGAGCTAATCAAAGAGATCGGAGTATGTT
TTGAAGGTGAGGCGGCCACCACGTCGTCCGCGCAAGCTGCACTTCGTATTATTCAGCGCTTCTTTGATGC
CGATACGTGCGCGTTAGCTCTAGTGGACCATGACCGTAGATGGGCTGTCGAAACATTCGGTGCGAAACAC
CCAAAACCTGTGTGGGACGACAGCGTGCTACGCGAAATAGTCTCTCGTACCAAAGCGGACGAACGGGCGA
CGGTATTCCGCATCATATCGTCGAAAAAAATCGTACATTTGCCTCTCGAAATTCCAGGTCTCTCGATACT
ACTGGCTCACAAATCCACAGATAAACTAATTGCGGTTTGTTCACTGGGTTACCAAAGCTATCGCCCTCGA
CCTTGCCAAGGCGAAATTCAGCTTCTTGAACTCGCCACCGCCTGCCTCTGTCACTATATCGATGTTCGGC
GTAAGCAGACCGAATGCGACGTTTTGGCCAGACGATTGGAGCATGCGCAACGCCTTGAGGCAGTTGGTAC
ACTTGCCGGCGGAATAGCACATGAATTTAATAACATTTTGGGCTCAATCCTCGGGCACGCAGAATTAGCA
CAAAACTCGGTGTCTCGAACATCTGTCACCCGAAGATATATTGACTATATCATTTCGTCAGGCGACAGAG
CCATGCTCATTATCGATCAGATCTTGACGCTGAGCCGAAAACAGGAGCGCATGATCAAGCCATTTAGTGT
CTCAGAGCTTGTGACCGAAATCGCTCCCTTGCTACGTATGGCTCTTCCGCCAAACATCGAGCTTAGTTTC
AGATTTGATCAAATGCAGAGCGTGATCGAAGGAAGCCCGCTTGAACTTCAACAGGTACTAATTAACATCT
GCAAGAATGCTTCCCAAGCCATGACTGCAAATGGTCAAATCGACATCATCATCAGCCAAGCTTTTTTACC
AGTTAAGAAAATTCTGGCGCATGGTGTTATGCCACCTGGCGACTATGTTCTCCTATCTATTAGCGACAAT
GGTGGAGGCATTCCCGAGGCTGTGTTACCCCACATTTTTGAACCCTTCTTTACGACACGAGCTCGCAACG
GTGGAACGGGTCTCGGCCTTGCTTCTGTGCATGGTCATATCAGCGCGTTTGCGGGTTACATCGACGTTAG
TTCAACTGTTGGGCATGGGACGCGCTTTGACATTTATCTCCCTCCGTCTTCTAAGGAACCCGTAAATCCA
GACAGTTTTTTCGGCCGCAATAAGGCACCGCGTGGAAACGGGGAGATTGTGGCACTTGTTGAGCCCGATG
ACCTCCTGCGGGAGGCGTATGAAGACAAGATCGCCGCTCTAGGATATGAGCCGGTCGGTTTTCGTACCTT
TAATGAAATTCGCGATTGGATTTCAAAAGGCAATGAAGCCGATCTGGTCATGGTCGACCAAGCGTCTCTT
CCTGAAGATCAAAGTCCTAATTCCGTGGATTTAGTGCTCAAGACCGCCTCCATCATCATTGGCGGAAATG
ATCTCAAAATGACCCTTTCAAGGGAGGATGTGACCAGGGACCTTTATCTTCCGAAGCCGATATCGTCCAG
AACTATGGCGCATGCAATCCTAACCAAAATCAAGACGTAG
```

Change region shown

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

Sekundární databáze

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



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 PRATT 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

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **NOTC_DROME**), or a PDB identifier, or paste your own **protein** sequence in the box below:

```
MMVKVTKLVASRPIVVFCVLAFLVVVFECIWISNWRTTTE
NLVKEVASFTEDLRTSLVSEIENIGKFTYAKTNLSTIGLA
RVIDSYITNNDTGFTEIQTQIAPLLFVAYSTILQVSQVSY
ISRDGLMFSYIAESNTSVAVFANSSSNSSRGDYTWYTQTV
DQLTGRLNGNSTKSQSLDVTHTDWFQAAQSNNYTTAFVGT
SLGGEDNETLIQSVVSLYSKKGLVSLGFPVKTLTEVLNSL
NLHGEELYMWTKDGTVLVREGSLNDSFFISNGSICFGRES
NSLWSQCIPENCSSSGYEVEIKRLRYQAFCSVIEVSGVPL
```

and specify which motifs to use:

Scan ☑ patterns ☑ profiles ☑ rules [User Manual] (You may also specify a PROSITE entry in the box to the right)

☐ Exclude patterns with a high probability of occurrence

Your e-mail (optional): (will send results by e-mail)

☐ plain text output

Search Swiss-Prot with a PROSITE entry

Enter a PROSITE accession number (for example **PS01253**), or type your pattern in PROSITE format:
(leave this box blank to scan a sequence with the entire PROSITE database)

and specify your search limits:

- The ☑ Swiss-Prot ☐ TrEMBL ☐ TrEMBLnew ☐ PDB databases
  (You may also specify a protein in the box to the left)
  ☑ including splice variants
- The following taxa:
  (see NEWT Taxonomy; separate multiple taxa with a semicolon, e.g. *Homo sapiens; Drosophila*. Not available for PDB.)
- Sequences with at least hits
- At most 1000 matches

Advanced options: ☐ FASTA output ☐ retrieve complete sequences

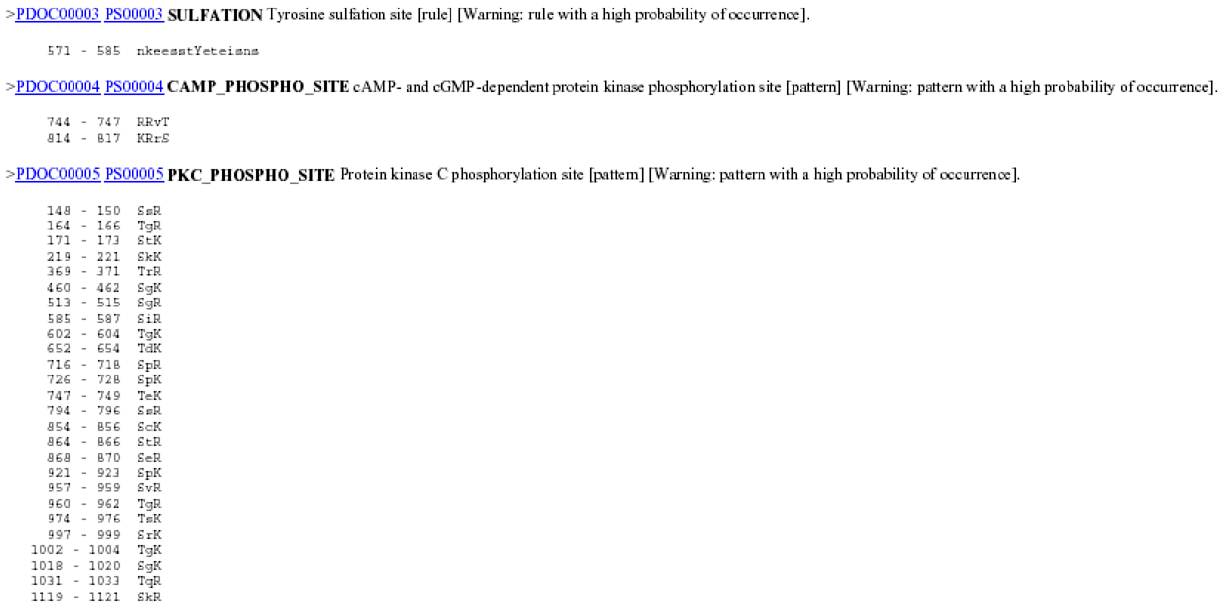
allow at most 1 X sequence characters to match a conserved position in the pattern

match mode: greedy, overlaps, no includes (for patterns, see help)

randomize databases: no (to test a pattern, see help)

Sekundární databáze

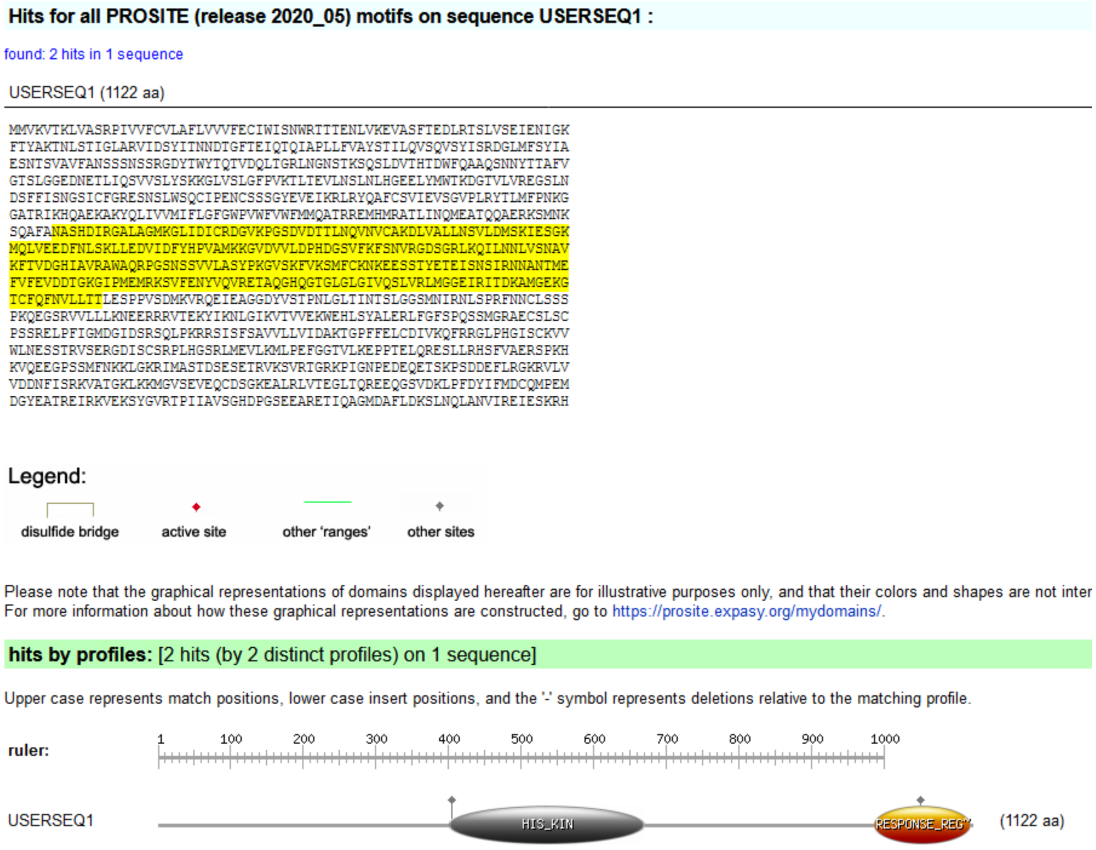
- **Databáze funkčních** nebo **strukturních motivů** získaných srovnáváním primárních dat (sekvencí)
- **PROSITE**, http://www.expasy.org/prosite/



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

     571 - 585   nkeesstYeteisns

>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   SiR
     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   SeR
     921 - 923   SpK
     957 - 959   SvR
     960 - 962   TgR
     974 - 976   TsK
     997 - 999   SrK
    1002 - 1004  TgK
    1018 - 1020  SgK
    1031 - 1033  TqR
    1119 - 1121  SkR
```

Sekundární databáze

- **Databáze funkčních** nebo **strukturních motivů** získaných srovnáváním 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)

```
MMVKVTKLVASRPIVVFCVLAFLVVVFECIWISNWRTTTENLVKEVASFTEDLRTSLVSEIENIGK
FTYAKTNLSTIGLARVIDSYITNNDTGFTEIQTQIAPLLFVAYSTILQVSQVSYISRDGLMFSYIA
ESNTSVAVFANSSSNSSRGDYTWYTQTVDQLTGRLNGNSTKSQSLDVTHTDWFQAAQSNNYTTAFV
GTSLGGEDNETLIQSVVSLYSKKGLVSLGFPVKTLTEVLNSLNLHGEELYMWTKDGTVLVREGSLN
DSFFISNGSICFGRESNSLWSQCIPENCSSSGYEVEIKRLRYQAFCSVIEVSGVPLRYTLMFPNKG
GATRIKHQAEKAKYQLIVVMIFLGFGWPVWFVWFMMQATRREMHMRATLINQMEATQQAERKSMNK
SQAFANASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVNVCAKDLVALLNSVLDMSKIESGK
MQLVEEDFNLSKLLEDVIDFYHPVAMKKGVDVVLDPHDGSVFKFSNVRGDSGRLKQILNNLVSNAV
KFTVDGHIAVRAWAQRPGSNSSVVLASYPKGVSKFVKSMFCKNKEESSTYETEISNSIRNNANTME
FVFEVDDTGKGIPMEMRKSVFENYVQVRETAQGHQGTGLGLGIVQSLVRLMGGEIRITDKAMGEKG
TCFQFNVLLTTLESPPVSDMKVRQEIEAGGDYVSTPNLGLTINTSLGGSMNIRNLSPRFNNCLSSS
PKQEGSRVVLLLKNEERRRVTEKYIKNLGIKVTVVEKWEHLSYALERLFGFSPQSSMGRAECSLSC
PSSRELPFIGMDGIDSRSQLPKRRSISFSAVVLLVIDAKTGPFFELCDIVKQFRRGLPHGISCKVV
WLNESSTRVSERGDISCSRPLHGSRLMEVLKMLPEFGGTVLKEPPTELQRESLLRHSFVAERSPKH
KVQEEGPSSMFNKKLGKRIMASTDSESETRVKSVRTGRKPIGNPEDEQETSKPSDDEFLRGKRVLV
VDDNFISRKVATGKLKKMGVSEVEQCDSGKEALRLVTEGLTQREEQGSVDKLPFDYIFMDCQMPEM
DGYEATREIRKVEKSYGVRTPIIAVSGHDPGSEEARETIQAGMDAFLDKSLNQLANVIREIESKRH
```

Legend:

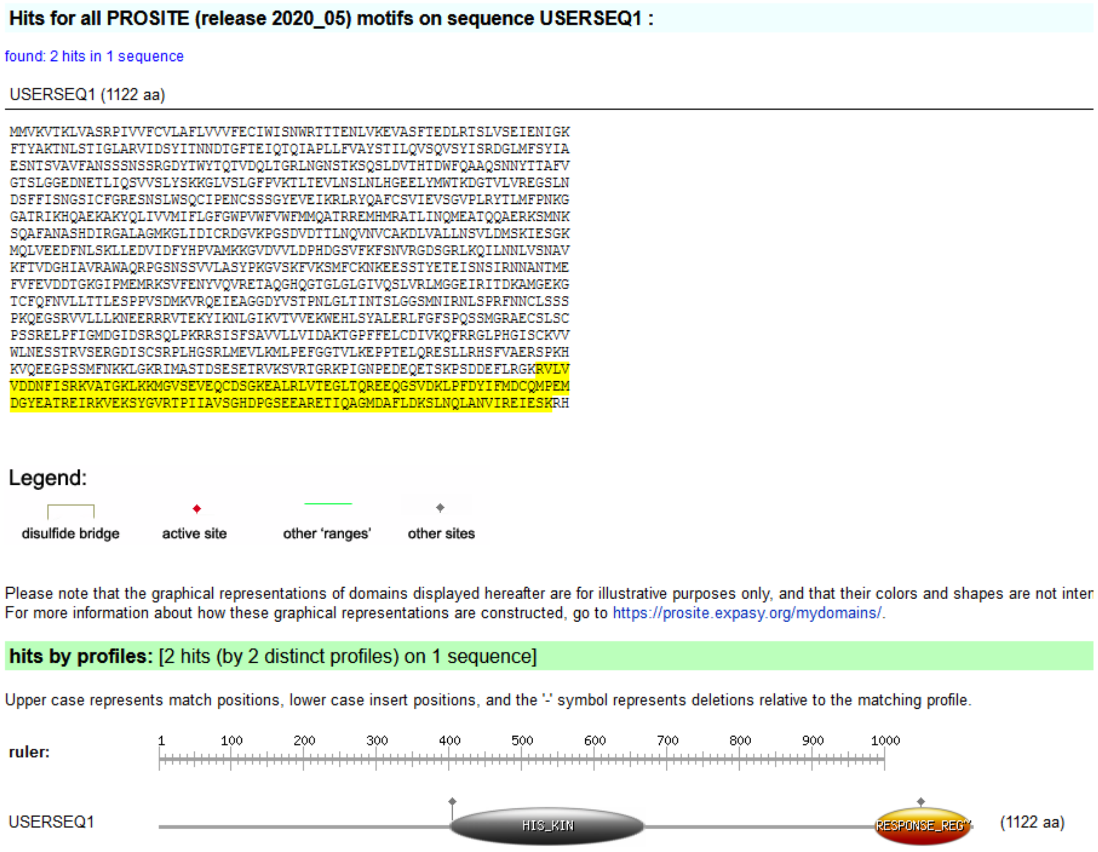
Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter
For more information about how these graphical representations are constructed, go to https://prosite.expasy.org/mydomains/.

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 srovnáváním 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)

```
MMVKVTKLVASRPIVVFCVLAFLVVVFECIWISNWRTTTENLVKEVASFTEDLRTSLVSEIENIGK
FTYAKTNLSTIGLARVIDSYITNNDTGFTEIQTQIAPLLFVAYSTILQVSQVSYISRDGLMFSYIA
ESNTSVAVFANSSSNSSRGDYTWYTQTVDQLTGRLNGNSTKSQSLDVTHTDWFQAAQSNNYTTAFV
GTSLGGEDNETLIQSVVSLYSKKGLVSLGFPVKTLTEVLNSLNLHGEELYMWTKDGTVLVREGSLN
DSFFISNGSICFGRESNSLWSQCIPENCSSSGYEVEIKRLRYQAFCSVIEVSGVPLRYTLMFPNKG
GATRIKHQAEKAKYQLIVVMIFLGFGWPVWFVWFMMQATRREMHMRATLINQMEATQQAERKSMNK
SQAFANASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVNVCAKDLVALLNSVLDMSKIESGK
MQLVEEDFNLSKLLEDVIDFYHPVAMKKGVDVVLDPHDGSVFKFSNVRGDSGRLKQILNNLVSNAV
KFTVDGHIAVRAWAQRPGSNSSVVLASYPKGVSKFVKSMFCKNKEESSTYETEISNSIRNNANTME
FVFEVDDTGKGIPMEMRKSVFENYVQVRETAQGHQGTGLGLGIVQSLVRLMGGEIRITDKAMGEKG
TCFQFNVLLTTLESPPVSDMKVRQEIEAGGDYVSTPNLGLTINTSLGGSMNIRNLSPRFNNCLSSS
PKQEGSRVVLLLKNEERRRVTEKYIKNLGIKVTVVEKWEHLSYALERLFGFSPQSSMGRAECSLSC
PSSRELPFIGMDGIDSRSQLPKRRSISFSAVVLLVIDAKTGPFFELCDIVKQFRRGLPHGISCKVV
WLNESSTRVSERGDISCSRPLHGSRLMEVLKMLPEFGGTVLKEPPTELQRESLLRHSFVAERSPKH
KVQEEGPSSMFNKKLGKRIMASTDSESETRVKSVRTGRKPIGNPEDEQETSKPSDDEFLRGKRVLV
VDDNFISRKVATGKLKKMGVSEVEQCDSGKEALRLVTEGLTQREEQGSVDKLPFDYIFMDCQMPEM
DGYEATREIRKVEKSYGVRTPIIAVSGHDPGSEEARETIQAGMDAFLDKSLNQLANVIREIESKRH
```

Legend:

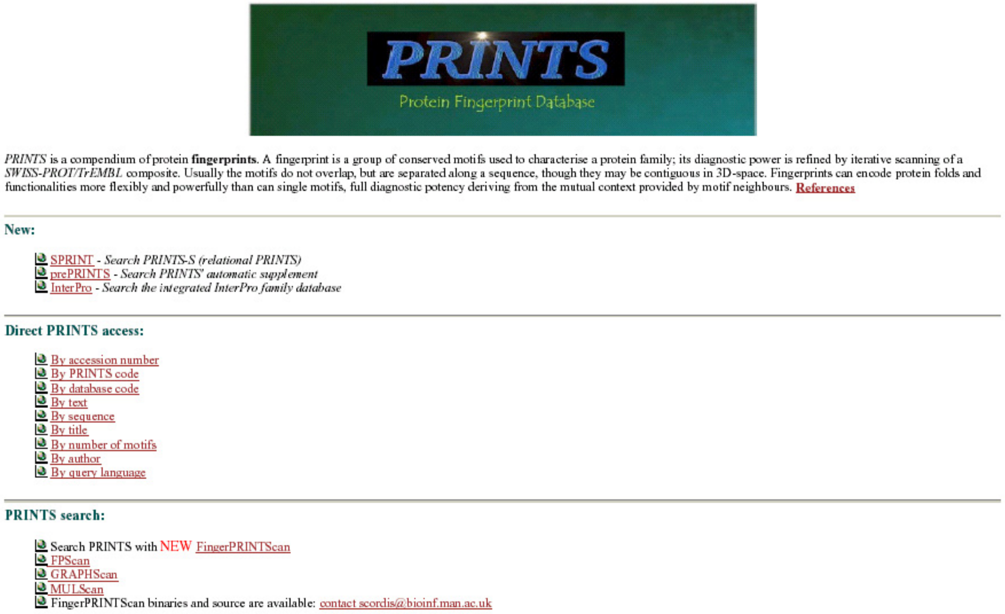
Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter
For more information about how these graphical representations are constructed, go to https://prosite.expasy.org/mydomains/.

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 srovnáváním 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/TrEMBL* 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:

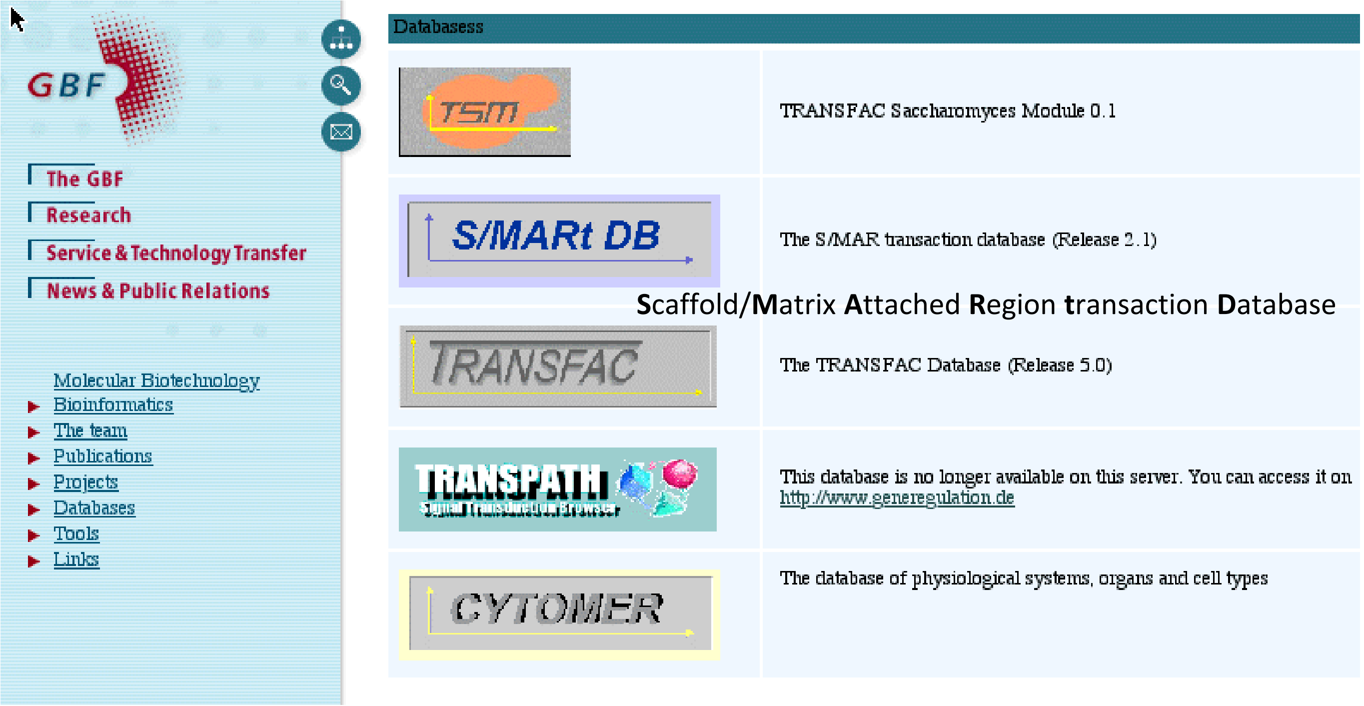
- By accession number
- By PRINTS code
- By database code
- By text
- By sequence
- By title
- By number of motifs
- By author
- By query language

PRINTS search:

- Search PRINTS with NEW FingerPRINTScan
- FPScan
- GRAPHScan
- MULScan
- FingerPRINTScan binaries and source are available: contact scordis@bioinf.man.ac.uk

Sekundární databáze

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



- The GBF
- Research
- Service & Technology Transfer
- News & Public Relations

Molecular Biotechnology
- Bioinformatics
- The team
- Publications
- Projects
- Databases
- Tools
- Links

Databasess	
TSM	TRANSFAC Saccharomyces Module 0.1
S/MARt DB	The S/MAR transaction database (Release 2.1)
TRANSFAC	The TRANSFAC Database (Release 5.0)
TRANSPATH Signal Transduction Browser	This database is no longer available on this server. You can access it on http://www.generegulation.de
CYTOMER	The database of physiological systems, organs and cell types

Scaffold/**M**atrix **A**ttached **R**egion **t**ransaction **D**atabase

Strukturální databáze

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

DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features
BETA mmCIF files

Current Holdings

19623 Structures
Last Update: 30-Dec-2002
PDB Statistics

Molecule of the Month: Cytochrome c

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Did you find what you wanted?

ABOUT PDB | DATA UNIFORMITY | RECENT FEATURES | USER GUIDES | FILE FORMATS | EDUCATION | STRUCTURAL GENOMICS | PUBLICATIONS | SOFTWARE

Search the Archive

Enter a PDB ID or keyword Query Tutorial

Find a structure

☐ query by PDB id only ☐ match exact word
☐ remove sequence homologues

SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

News Complete News Newsletter pdb-l Archive Subscribe

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!

PDB Mirrors

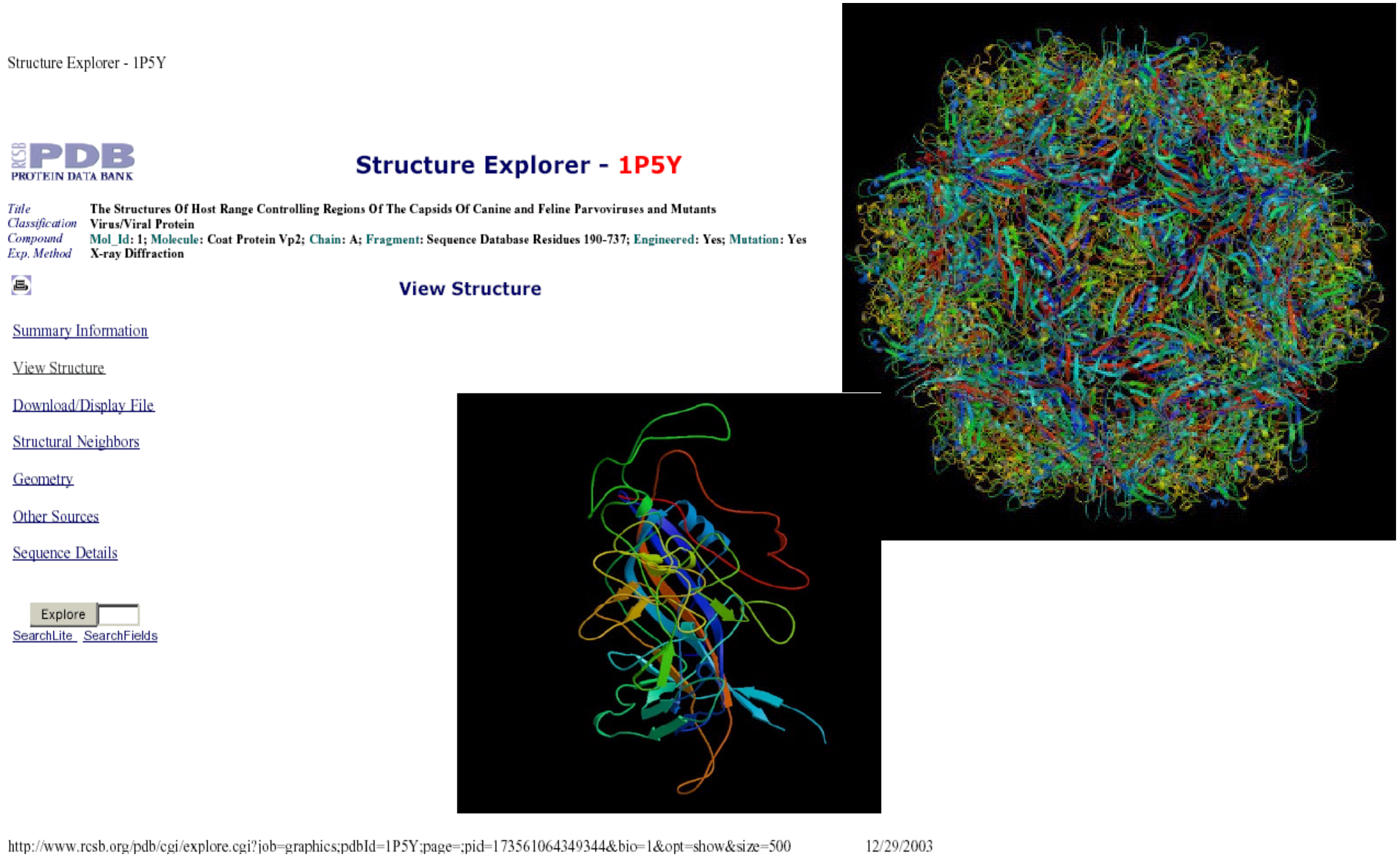
Please bookmark a mirror site

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
Max Delbrück Center for Molecular Medicine, Germany

OTHER SITES

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

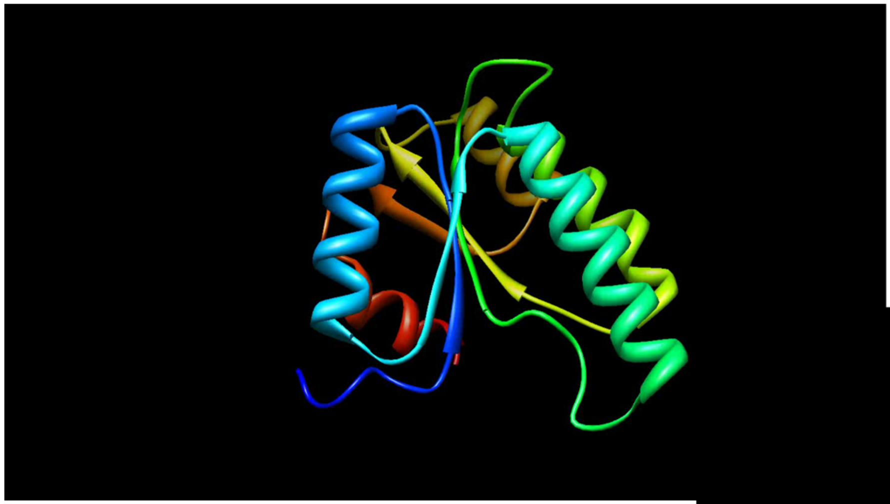
Other Sources

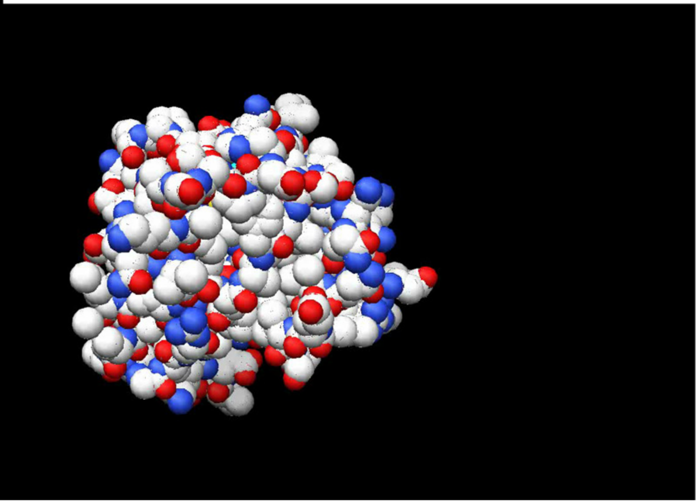
Sequence Details

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/





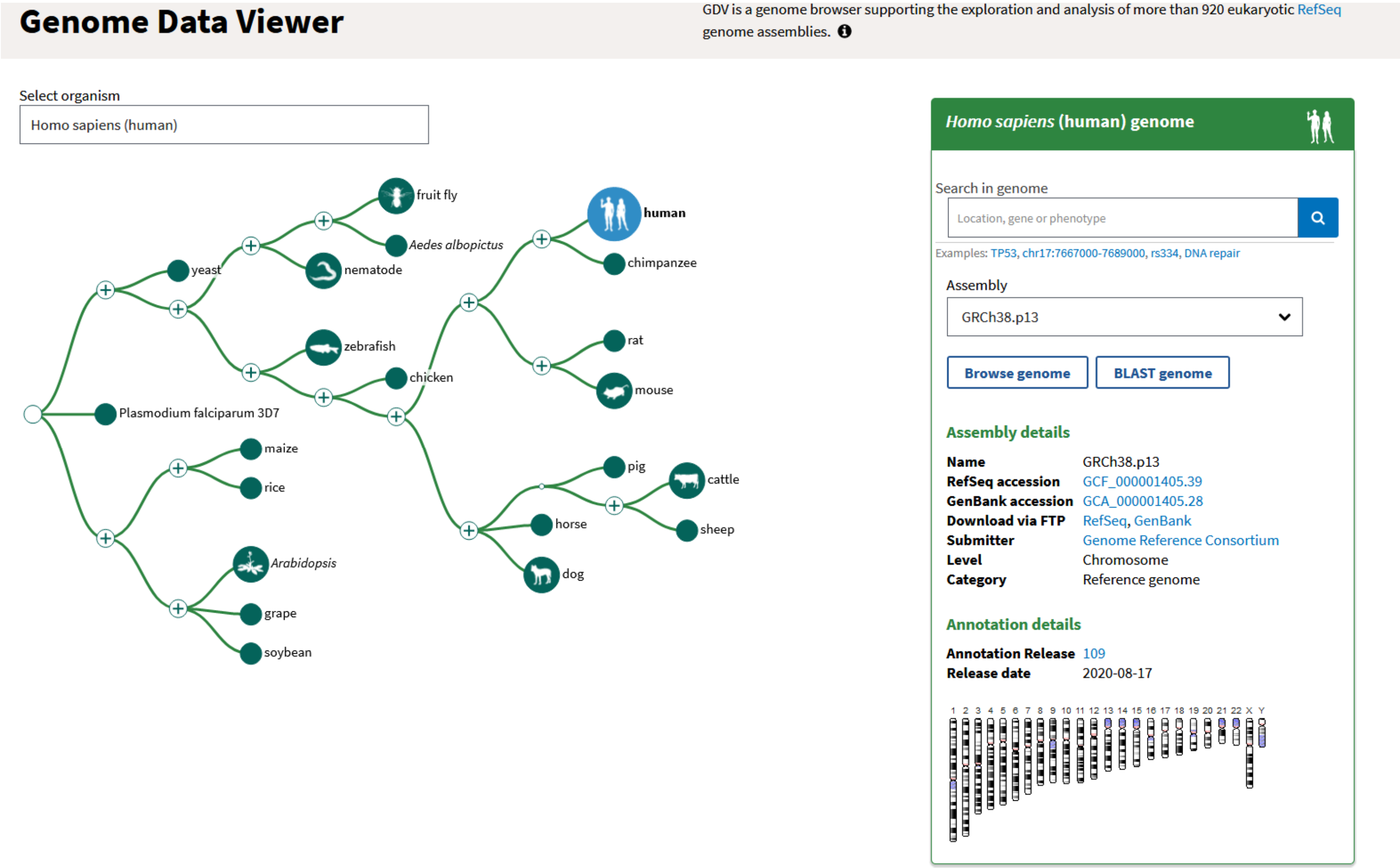
Pekárová et al., *Plant Journal* (2011)

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

Assembly details

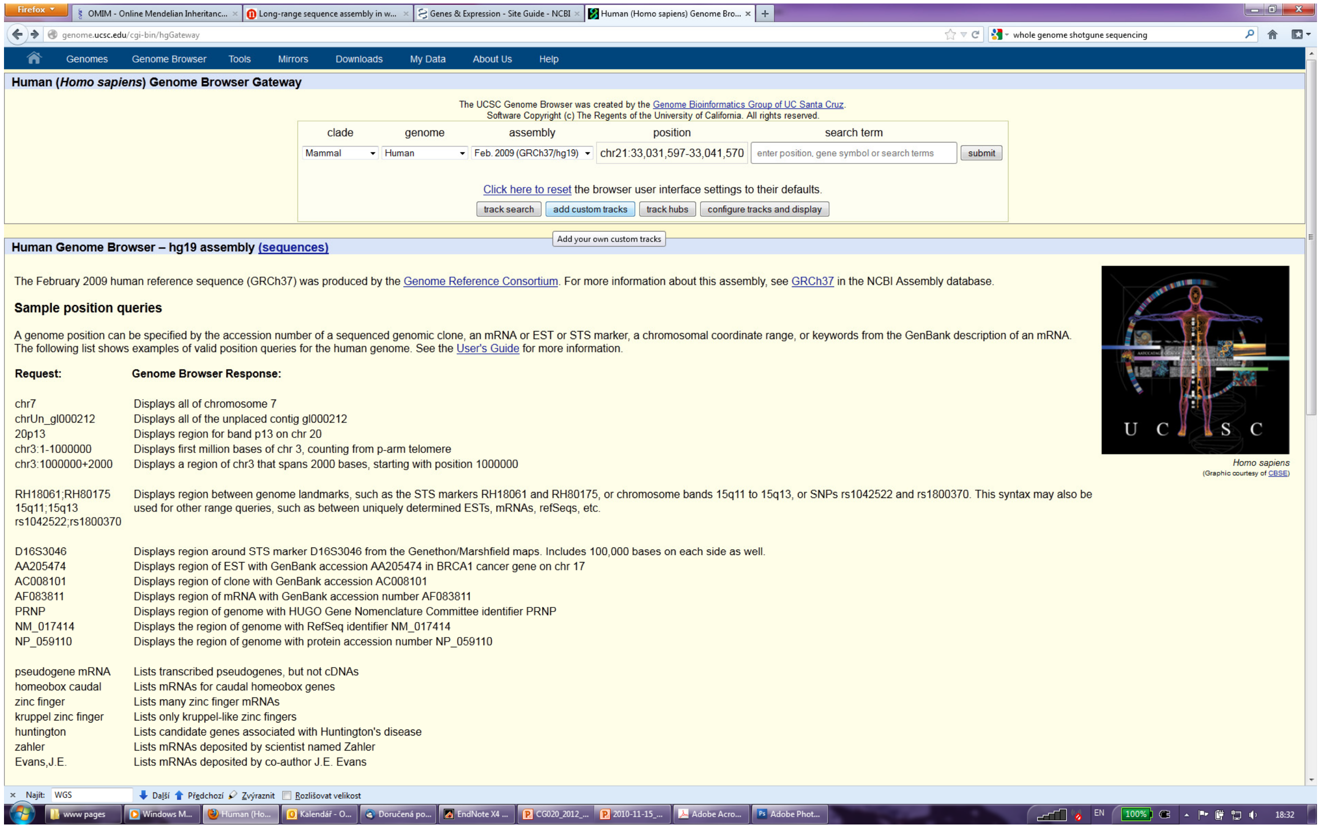
Name	GRCh38.p13
RefSeq accession	GCF_000001405.39
GenBank accession	GCA_000001405.28
Download via FTP	RefSeq, GenBank
Submitter	Genome Reference Consortium
Level	Chromosome
Category	Reference genome

Annotation details

Annotation Release	109
Release date	2020-08-17

Genomové zdroje

- **Genome Browser Gateway** https://genome.ucsc.edu/



Human (*Homo sapiens*) Genome Browser Gateway

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.

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

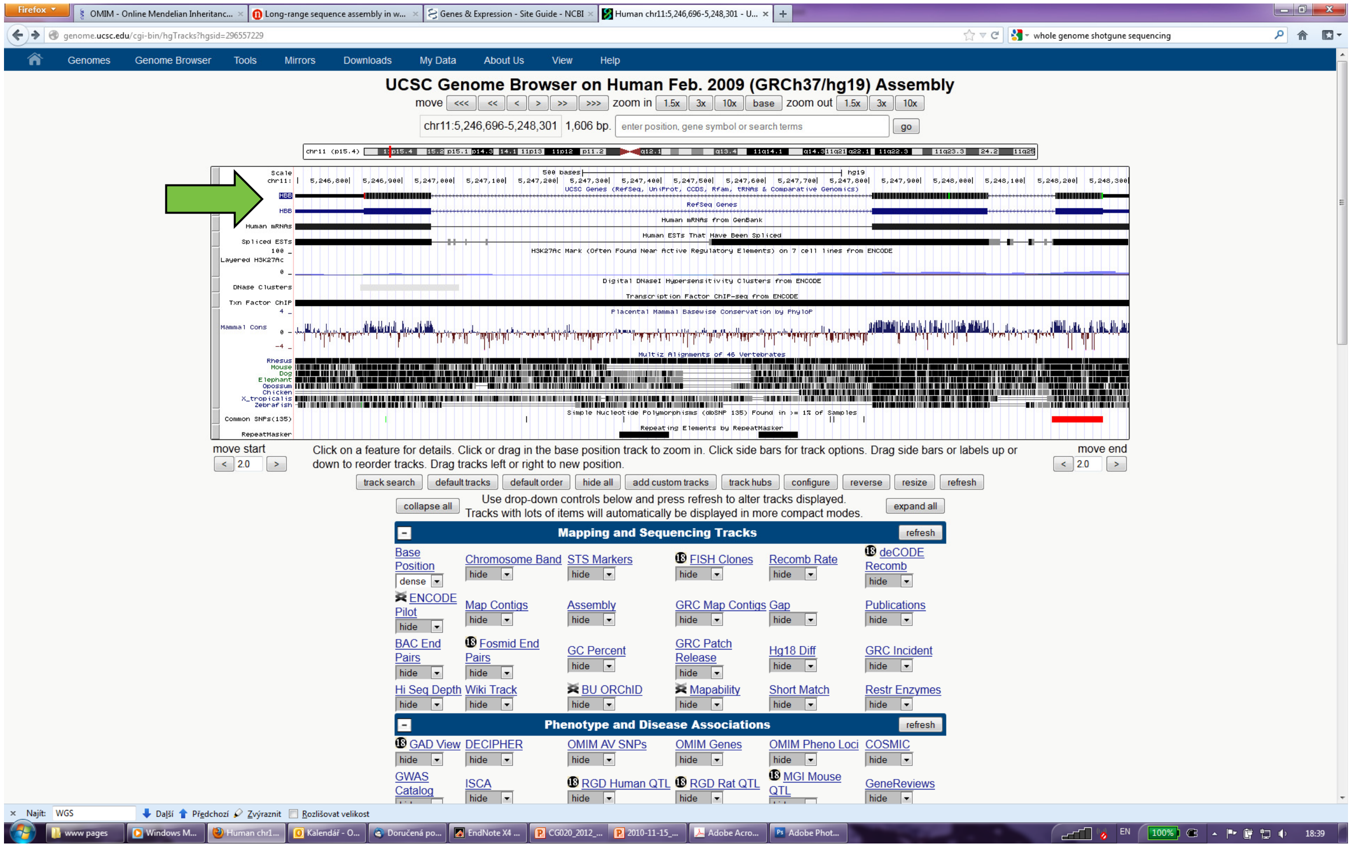
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's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

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
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

Homo sapiens
(Graphic courtesy of CBSE)

Genomové zdroje

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



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

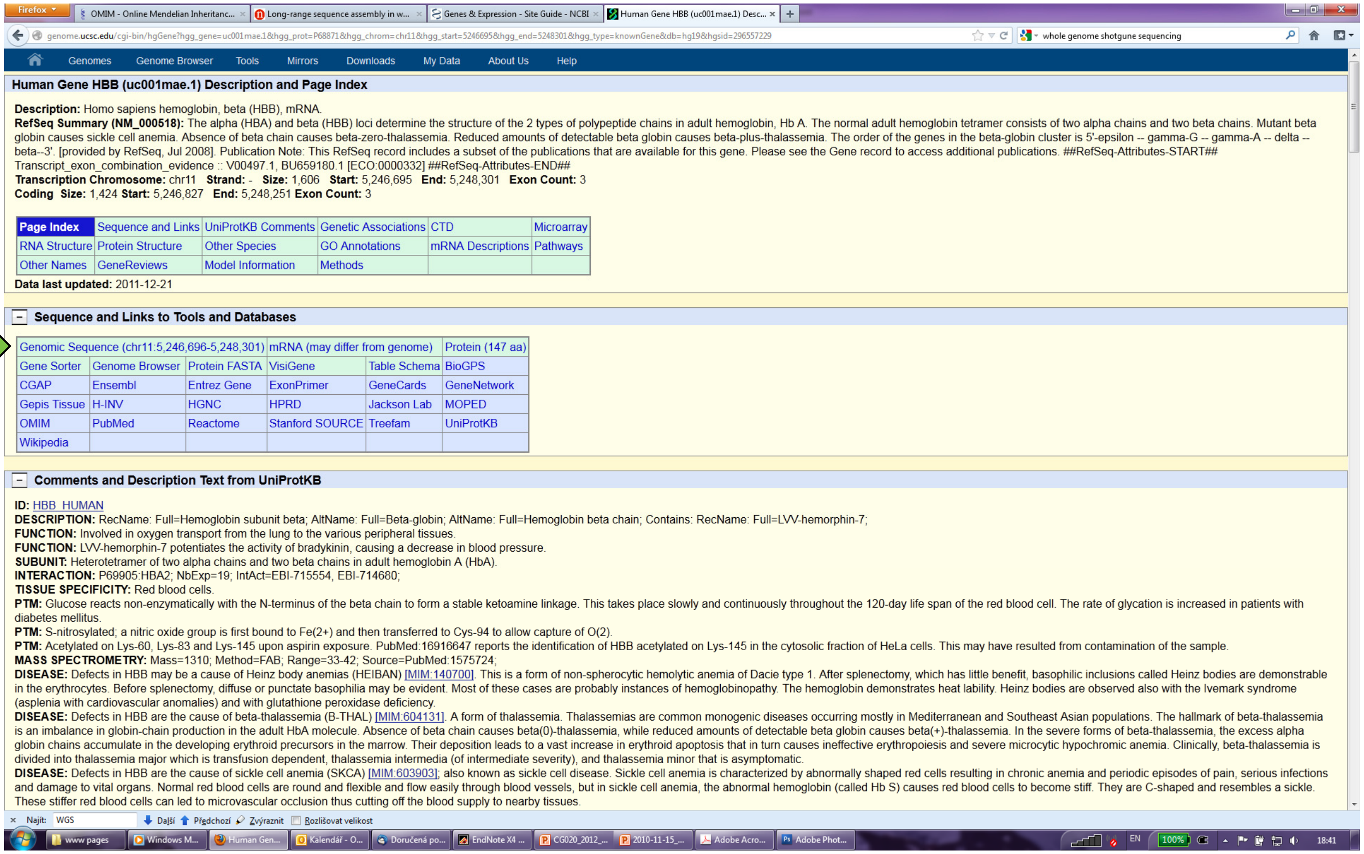
chr11:5,246,696-5,248,301 1,606 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

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

Page Index	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

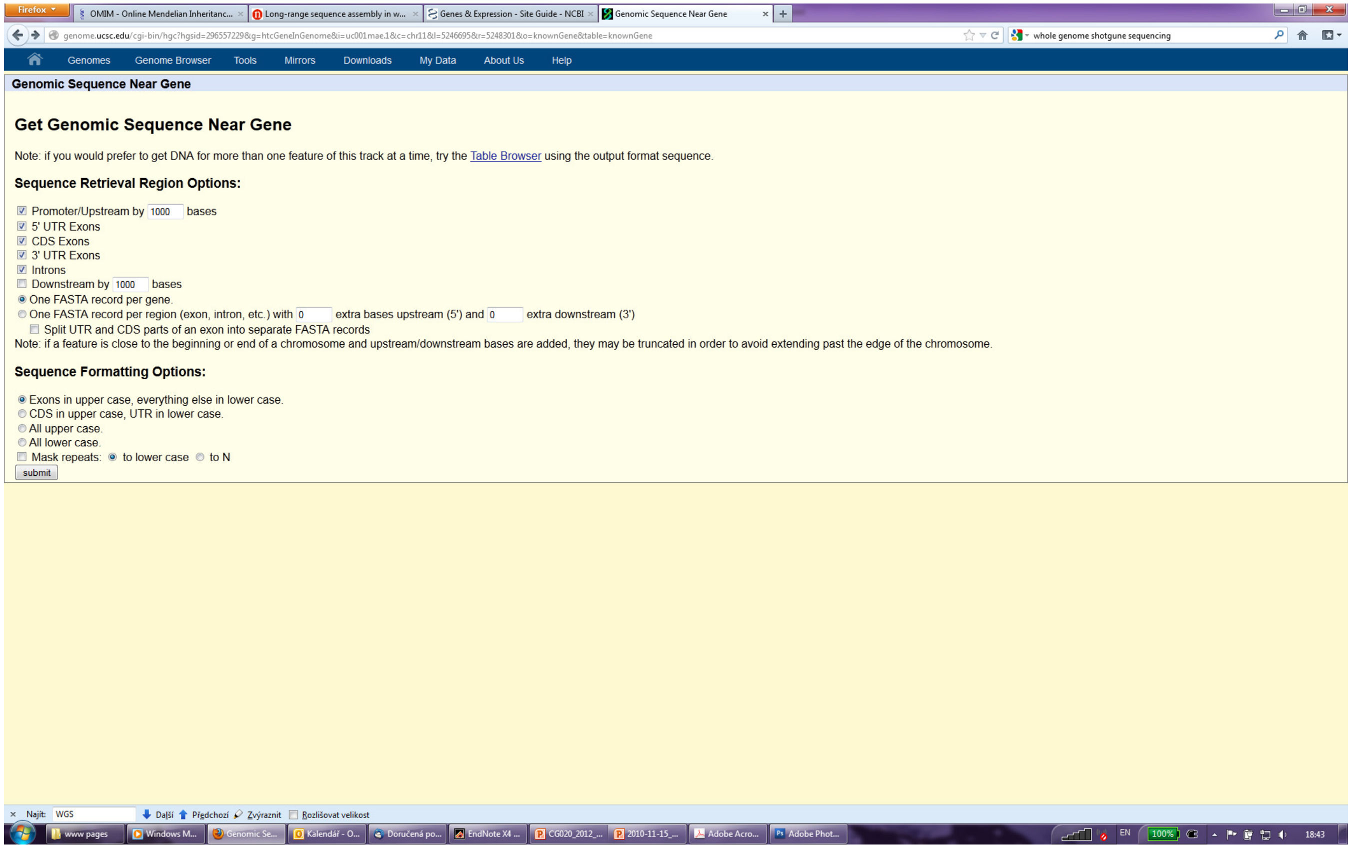
Genomic Sequence (chr11:5,246,696-5,248,301)			mRNA (may differ from genome)		Protein (147 aa)
Gene Sorter	Genome Browser	Protein FASTA	VisiGene	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	H-INV	HGNC	HPRD	Jackson Lab	MOPED
OMIM	PubMed	Reactome	Stanford SOURCE	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: Heterotetramer 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 led to microvascular occlusion thus cutting off the blood supply to nearby tissues.

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.

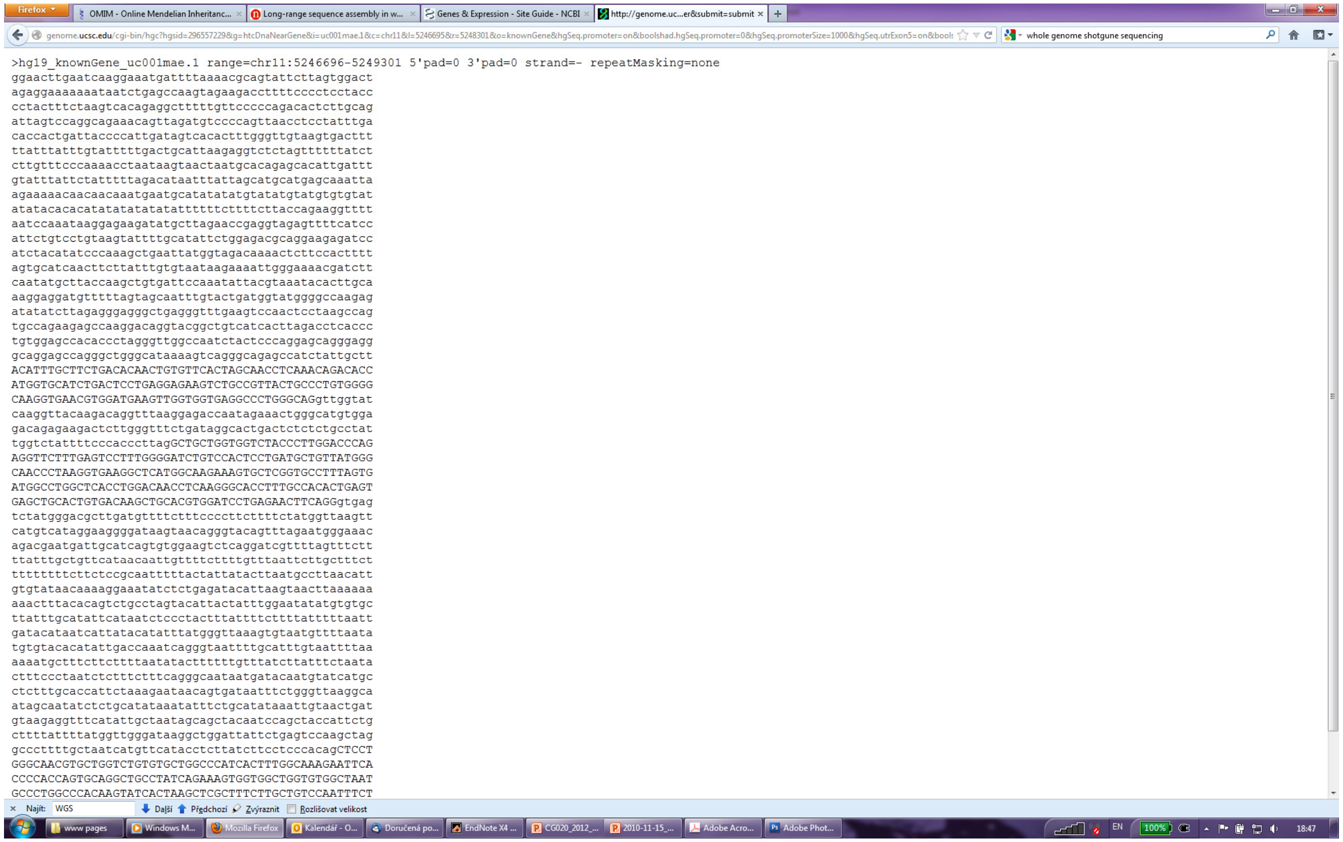
Sequence Formatting Options:

● 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

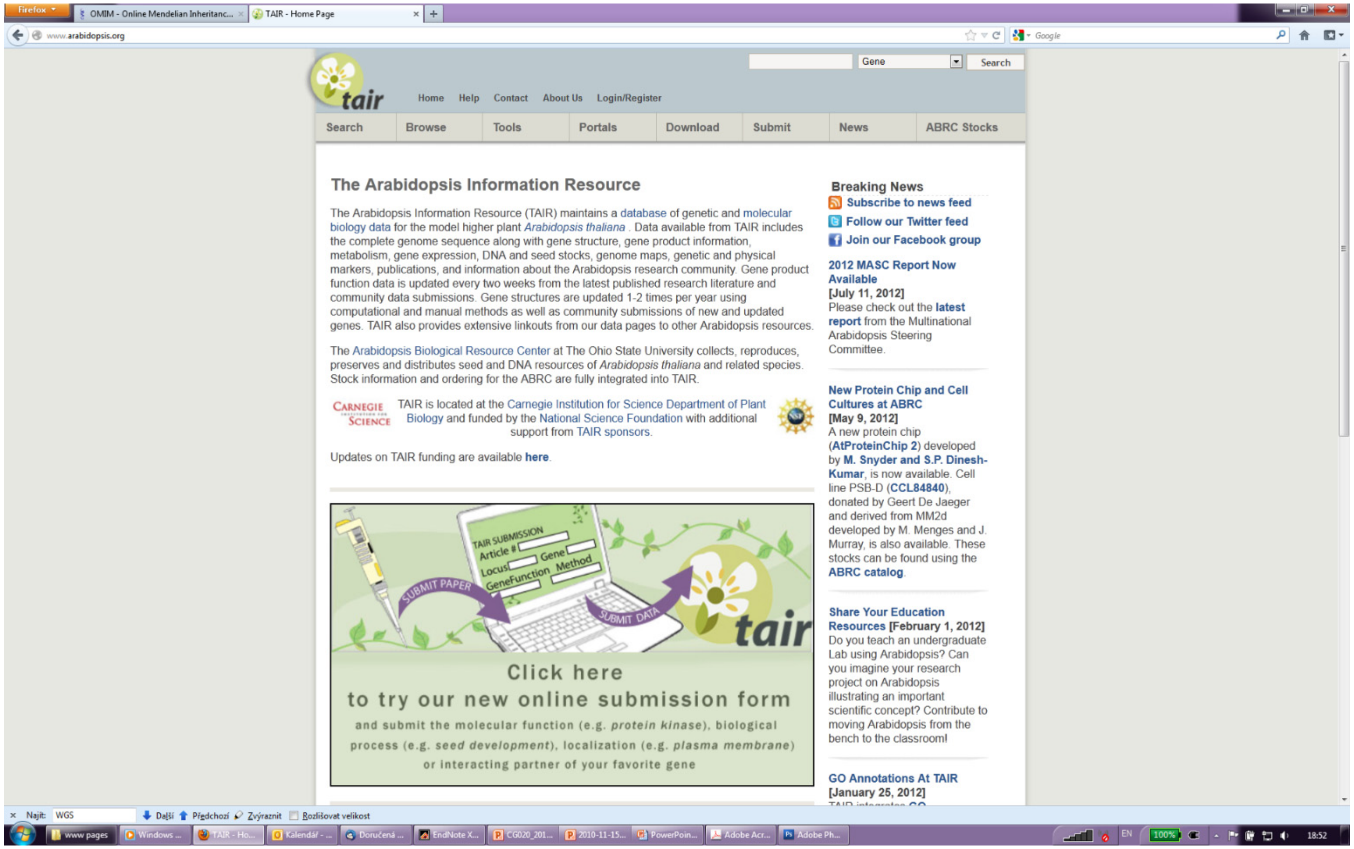
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



```
>hg19_knownGene_uc001mae.1 range=chr11:5246696-5249301 5'pad=0 3'pad=0 strand=- repeatMasking=none
ggaacttgaatcaaggaaatgattttaaaacgcagtattcttagtggact
agaggaaaaaaataatctgagccaagtagaagaccttttcccctcctacc
cctactttctaagtcacagaggctttttgttcccccagacactcttgcag
attagtccaggcagaaacagttagatgtccccagttaacctcctatttga
caccactgattaccccattgatagtcacactttgggttgtaagtgacttt
ttatttatttgtatttttgactgcattaagaggtctctagttttttatct
cttgtttcccaaaacctaataagtaactaatgcacagagcacattgattt
gtatttattctattttttagacataatttattagcatgcatgagcaaatta
agaaaaacaacaacaaatgaatgcatatatatgtatatgtatgtgtgtat
atatacacacatatatatatatatttttttcttttcttaccagaaggtttt
aatccaaataaggagaagatatgcttagaaccgaggtagagttttcatcc
attctgtcctgtaagtattttgcatattctggagacgcaggaagagatcc
atctacatatcccaaagctgaattatggtagacaaaactcttccactttt
agtgcatcaacttcttatttgtgtaataagaaaattgggaaaacgatctt
caatatgcttaccaagctgtgattccaaatattacgtaaatacacttgca
aaggaggatgtttttagtagcaatttgtactgatggtatggggccaagag
atatatcttagagggagggctgagggtttgaagtccaactcctaagccag
tgccagaagagccaaggacaggtacggctgtcatcacttagacctcaccc
tgtggagccacaccctagggttggccaatctactcccaggagcagggagg
gcaggagccagggctgggcataaaagtcagggcagagccatctattgctt
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACC
ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGG
CAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGgttggtat
caaggttacaagacaggtttaaggagaccaatagaaactgggcatgtgga
gacagagaagactcttgggtttctgataggcactgactctctctgcctat
tggtctattttcccacccttagGCTGCTGGTGGTCTACCCTTGGACCCAG
AGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTG
ATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGT
GAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGgtgag
tctatgggacgcttgatgttttctttccccttcttttctatggttaagtt
catgtcataggaaggggataagtaacagggtacagtttagaatgggaaac
agacgaatgattgcatcagtgtggaagtctcaggatcgttttagtttctt
ttatttgctgttcataacaattgttttcttttgtttaattcttgctttct
ttttttttcttctccgcaatttttactattatacttaatgccttaacatt
gtgtataacaaaaggaaatatctctgagatacattaagtaacttaaaaaa
aaactttacacagtctgcctagtacattactatttggaatatatgtgtgc
ttatttgcatattcataatctccctactttattttcttttatttttaatt
gatacataatcattatacatatttatgggttaaagtgtaatgttttaata
tgtgtacacatattgaccaaatcagggtaattttgcatttgtaattttaa
aaaatgctttcttcttttaatatactttttttgtttatcttatttctaata
ctttccctaatctctttctttcagggcaataatgatacaatgtatcatgc
ctctttgcaccattctaaagaataacagtgataatttctgggttaaggca
atagcaatatctctgcatataaatatttctgcatataaattgtaactgat
gtaagaggtttcatattgctaatagcagctacaatccagctaccattctg
cttttattttatggttgggataaggctggattattctgagtccaagctag
gccctttttgctaatcatgttcatacctcttatcttcctcccacagCTCCT
GGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCT
```

Genomové zdroje

- **T**he **A**rabidopsis **I**nformation **R**esource (**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.

TAIR is located at the Carnegie Institution for Science Department of Plant Biology and funded by the National Science Foundation with additional support from TAIR sponsors.

Updates on TAIR funding are available **here**.

Click here

to try our new online submission form

and submit the molecular function (e.g. *protein kinase*), biological process (e.g. *seed development*), localization (e.g. *plasma membrane*) or interacting partner of your favorite gene

Breaking News

Subscribe to news feed

Follow our Twitter feed

Join our Facebook group

2012 MASC Report Now Available
[July 11, 2012]
Please check out the **latest report** from the Multinational Arabidopsis Steering Committee.

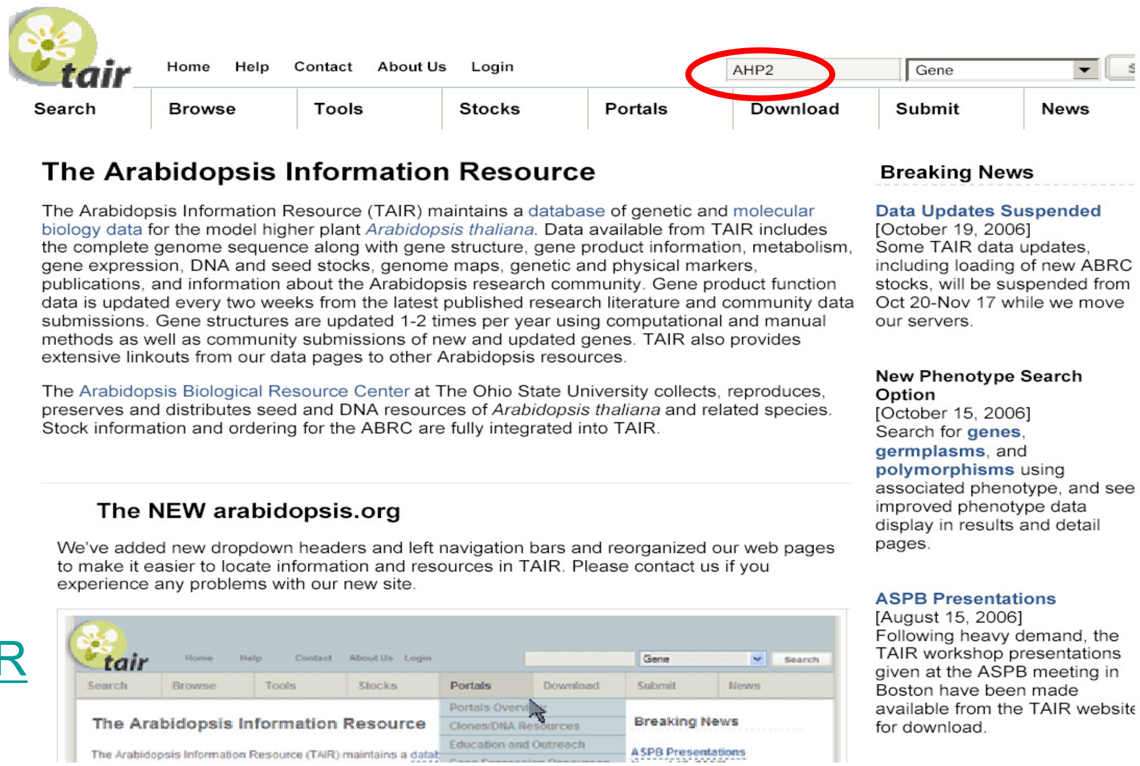
New Protein Chip and Cell Cultures at ABRC
[May 9, 2012]
A new protein chip (**AtProteinChip 2**) developed by **M. Snyder and S.P. Dinesh-Kumar**, is now available. Cell line PSB-D (**CCL84840**), donated by Geert De Jaeger and derived from MM2d developed by M. Menges and J. Murray, is also available. These stocks can be found using the **ABRC catalog**.

Share Your Education Resources [February 1, 2012]
Do you teach an undergraduate Lab using Arabidopsis? Can you imagine your research project on Arabidopsis illustrating an important scientific concept? Contribute to moving Arabidopsis from the bench to the classroom!

GO Annotations At TAIR
[January 25, 2012]

Genomové zdroje

- **T**he **A**rabidopsis **I**nformation **R**esource (**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.

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.

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.

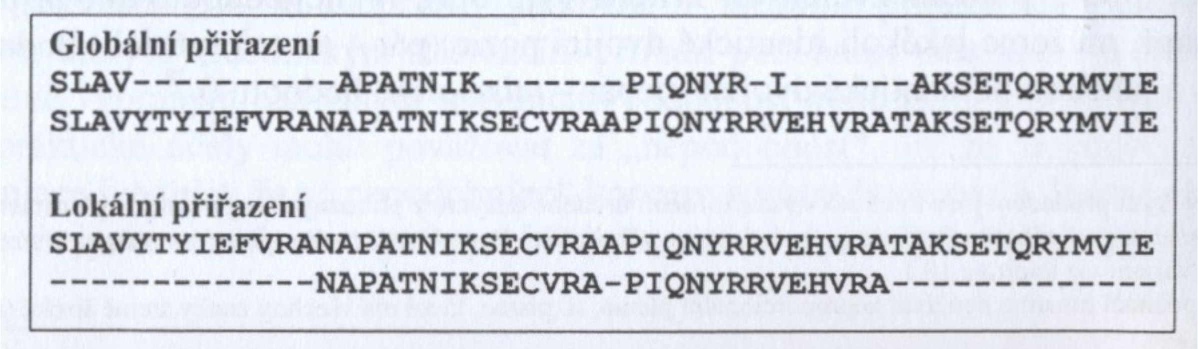
[AHP2 @ TAIR]()

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



Globální přiřazení

```
SLAV----------APATNIK-------PIQNYR-I------AKSETQRYMVIE
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE
```

Lokální přiřazení

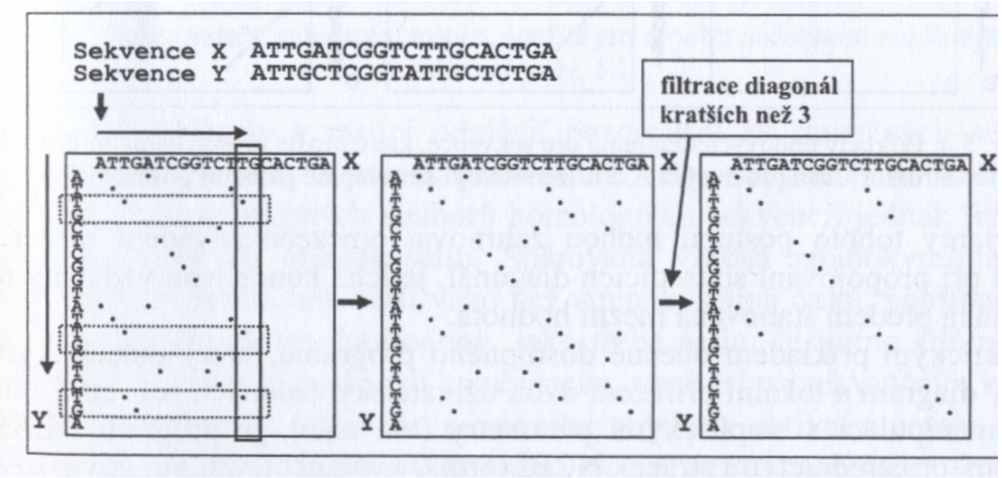
```
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE
-------------NAPATNIKSECVRA-PIQNYRRVEHVRA-------------
```

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)

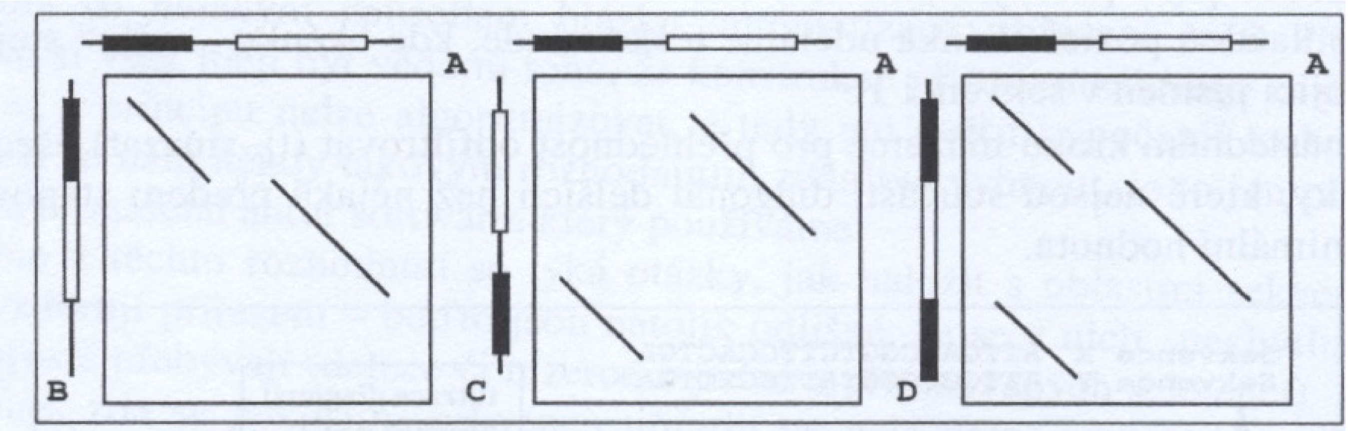


Cvrčková, Úvod do praktické bioinformatiky

  - 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

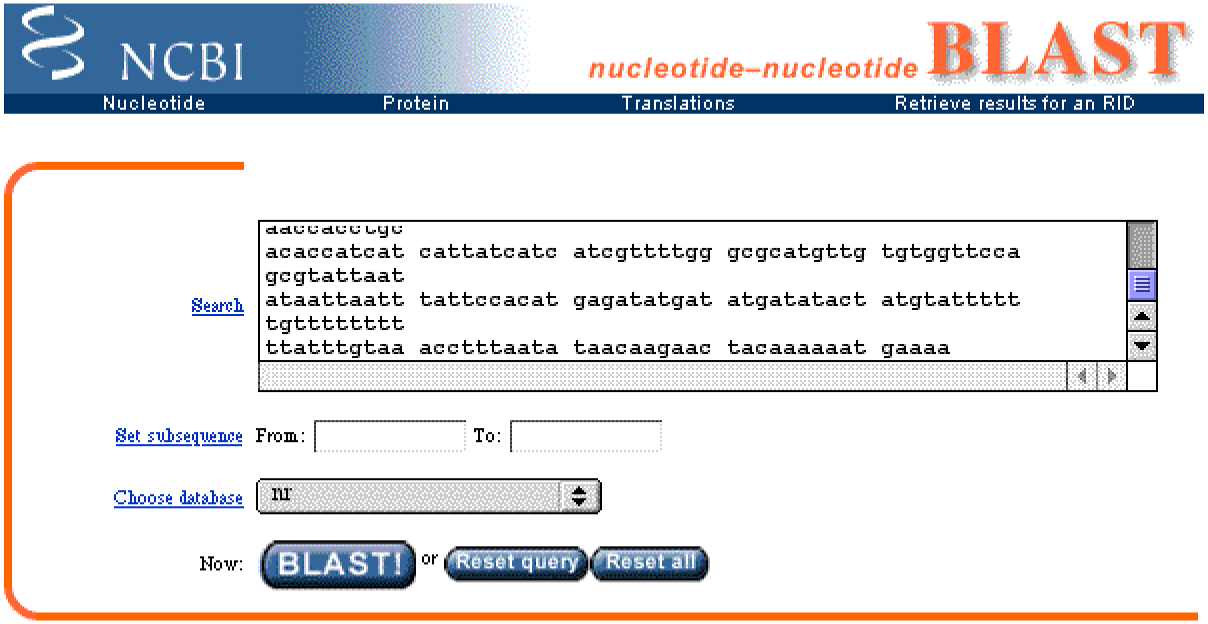


Cvrčková, Úvod do praktické bioinformatiky

- 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 srovnávání programem BLAST2 (viz dále)

Analytické nástroje

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



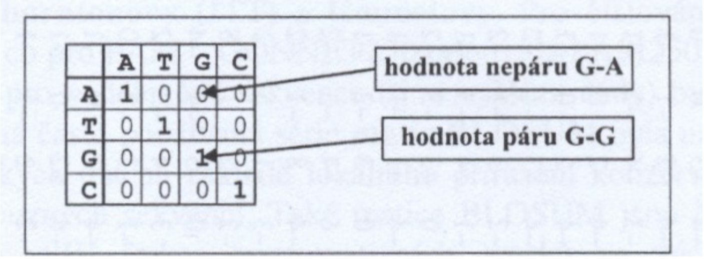
```
aaccacctgc
acaccatcat cattatcatc atcgttttgg gcgcatgttg tgtggttcca
gcgtattaat
ataattaatt tattccacat gagatatgat atgatatact atgtattttt
tgttttttt
ttatttgtaa acctttaata taacaagaac tacaaaaaat gaaaa
```

Choose database: nr

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool

- 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

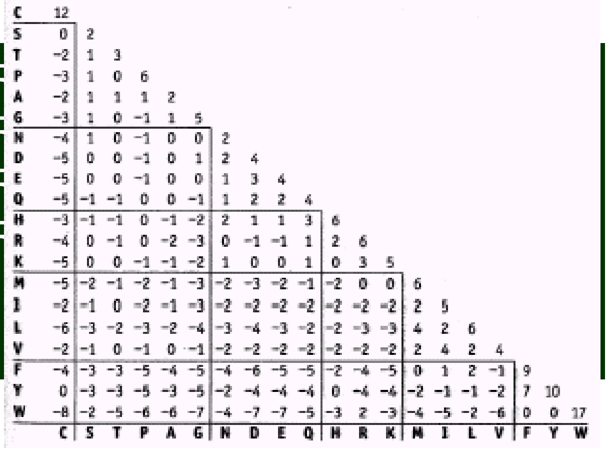


	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250

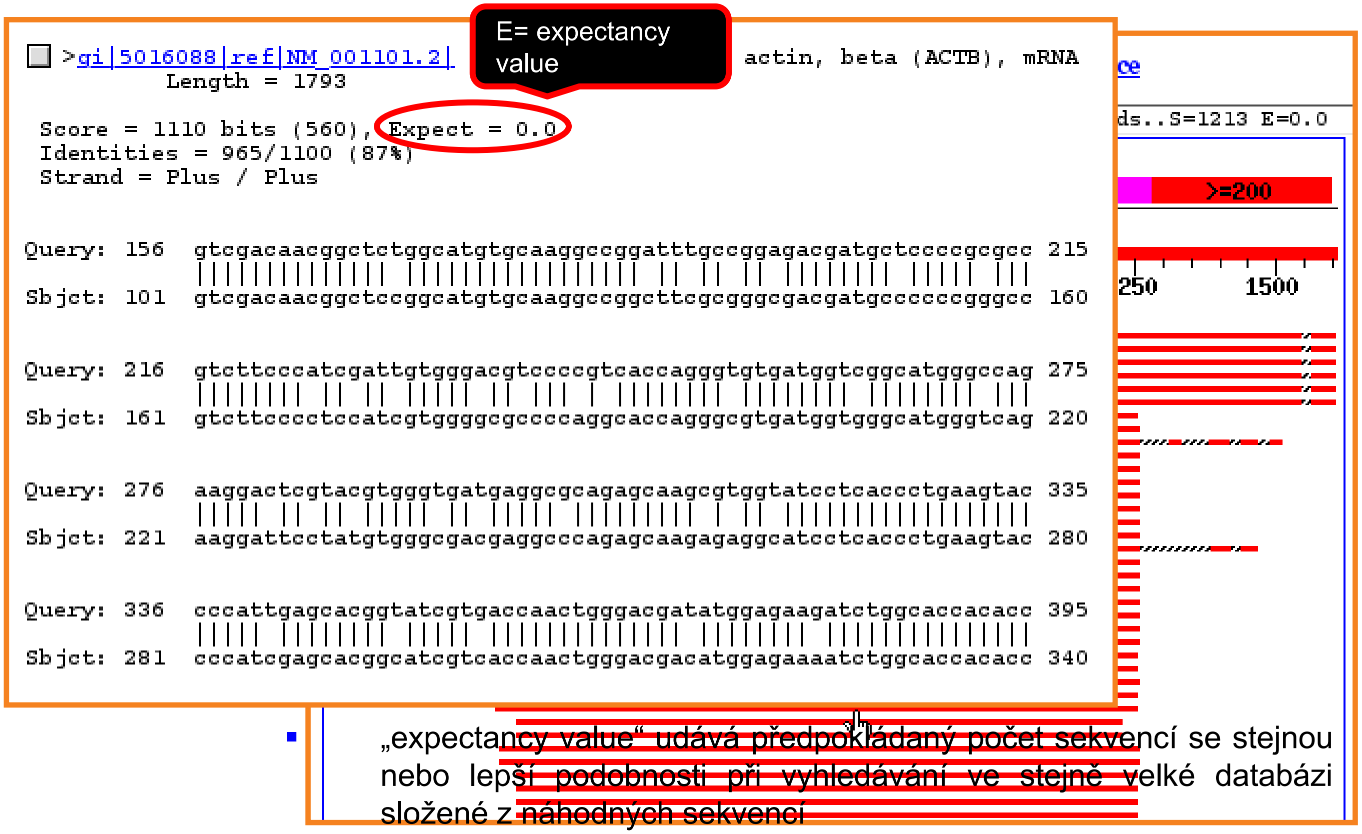
MF
MF
MF
MF



C	12																			
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	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool

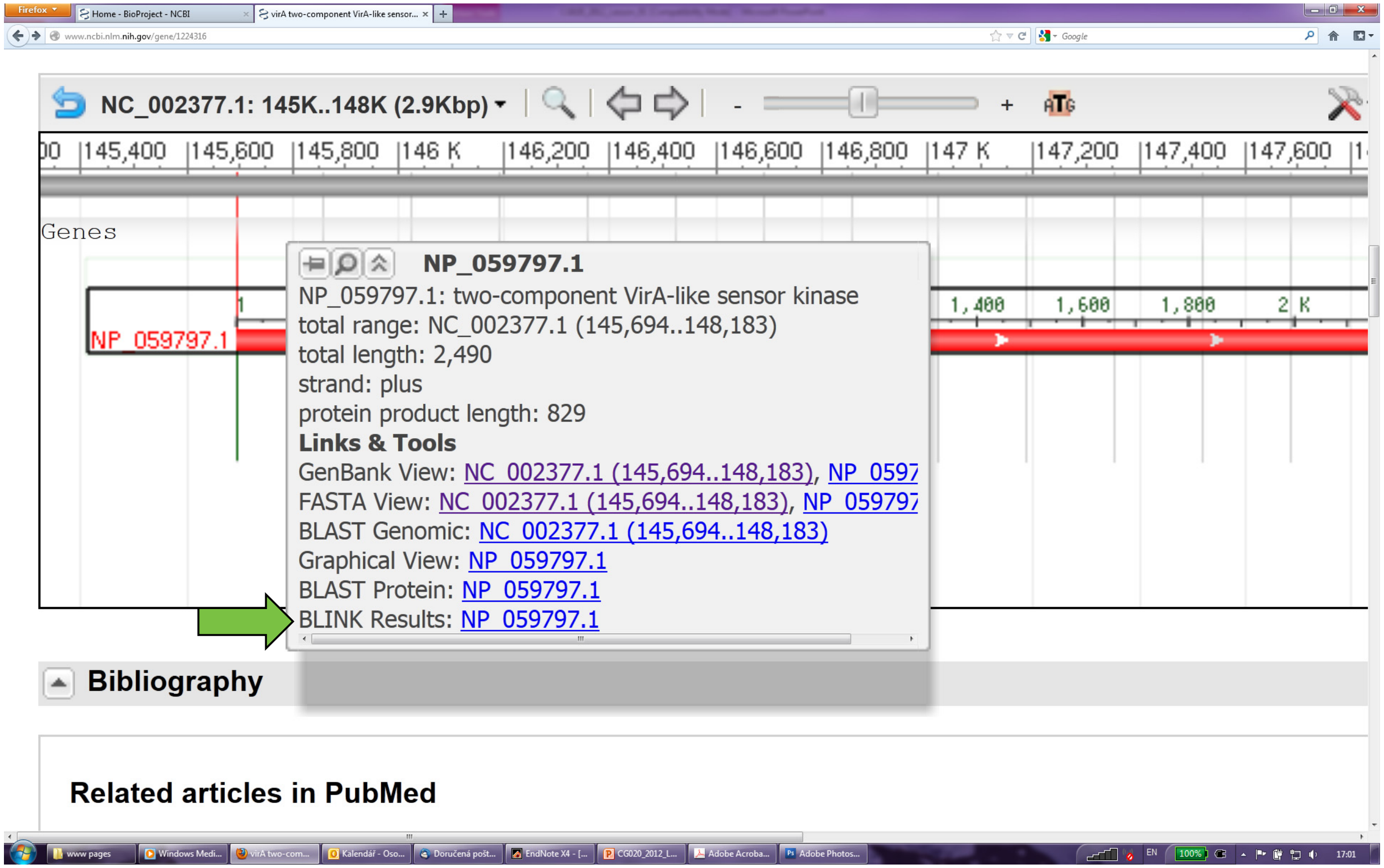


E= expectancy value

```
☐ >gi|5016088|ref|NM_001101.2|                         actin, beta (ACTB), mRNA
          Length = 1793

  Score = 1110 bits (560), Expect = 0.0
  Identities = 965/1100 (87%)
  Strand = Plus / Plus


Query: 156   gtcgacaacggctctggcatgtgcaaggccggatttgccggagacgatgctccccgcgcc 215
              |||||||||||||| ||||||||||||||||| || || || |||||||| ||||| |||
Sbjct: 101   gtcgacaacggctccggcatgtgcaaggccggcttcgcgggcgacgatgccccccgggcc 160


Query: 216   gtcttcccatcgattgtgggacgtccccgtcaccagggtgtgatggtcggcatgggccag 275
              |||||||| || || ||||| || ||| | |||||||| |||||||| |||||||| |||
Sbjct: 161   gtcttcccctccatcgtggggcgccccaggcaccagggcgtgatggtgggcatgggtcag 220


Query: 276   aaggactcgtacgtgggtgatgaggcgcagagcaagcgtggtatcctcaccctgaagtac 335
              ||||| || || ||||| || ||||| |||||||||| | || ||||||||||||||||||
Sbjct: 221   aaggattcctatgtgggcgacgaggcccagagcaagagaggcatcctcaccctgaagtac 280


Query: 336   cccattgagcacggtatcgtgaccaactgggacgatatggagaagatctggcaccacacc 395
              ||||| |||||||| ||||| |||||||||||||| |||||||| ||||||||||||||||
Sbjct: 281   cccatcgagcacggcatcgtcaccaactgggacgacatggagaaaatctggcaccacacc 340
```

- „expectancy value“ udává předpokládaný počet sekvencí se stejnou nebo lepší podobnosti 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



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

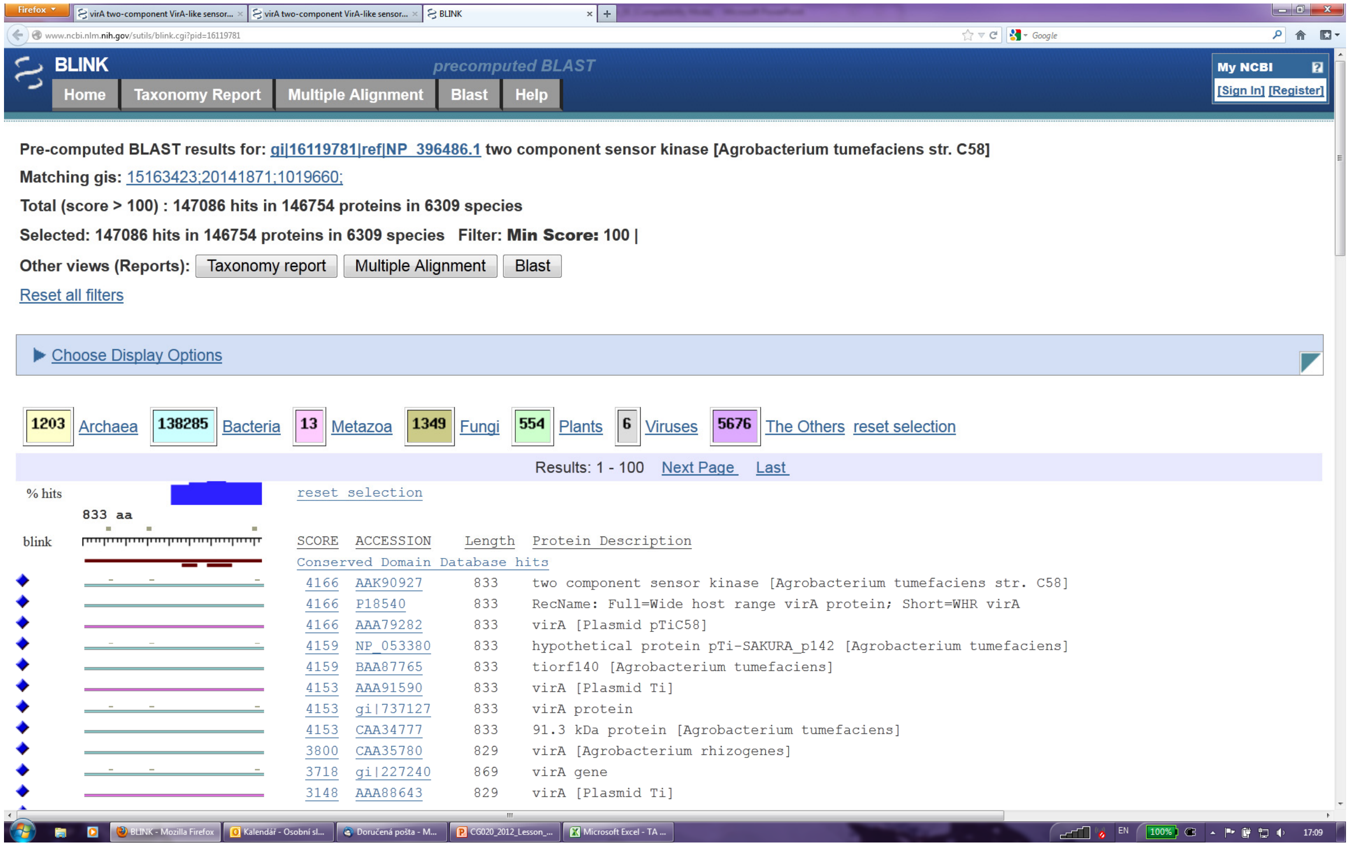
GenBank View: NC_002377.1 (145,694..148,183), NP_0597
FASTA View: NC_002377.1 (145,694..148,183), NP_059797
BLAST Genomic: NC_002377.1 (145,694..148,183)
Graphical View: NP_059797.1
BLAST Protein: NP_059797.1
BLINK Results: NP_059797.1

Bibliography

Related articles in PubMed

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool



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

reset selection

Conserved Domain Database hits

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

BLAST

Specializované verze

- ☐ V současnosti existuje celá řad 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 nukleotidové sekvence přeložené do sekvence aa v proteinové databázi

BLAST

Specializované verze

- ☐ V současnosti existuje celá řad specializovaných verzí programu BLAST

- **TBLASTN**
  - vyhledávání podobnosti proteinové sekvence v nukleotidové databázi přeložené do sekvence aa

- **TBLASTX**
  - vyhledávání k sekvenci nukleotidů přeložené do sekvence aa v databázi nukleotidových sekvencí přeložených do sekvence aa

BLAST

Specializované verze

☐ V současnosti existuje celá řad specializovaných verzí programu BLAST

- PSI-BLAST (**P**osition-**S**pecific **It**erated **BLAST**)
  - 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 (Position Specific Substitution Matrix)
  - 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

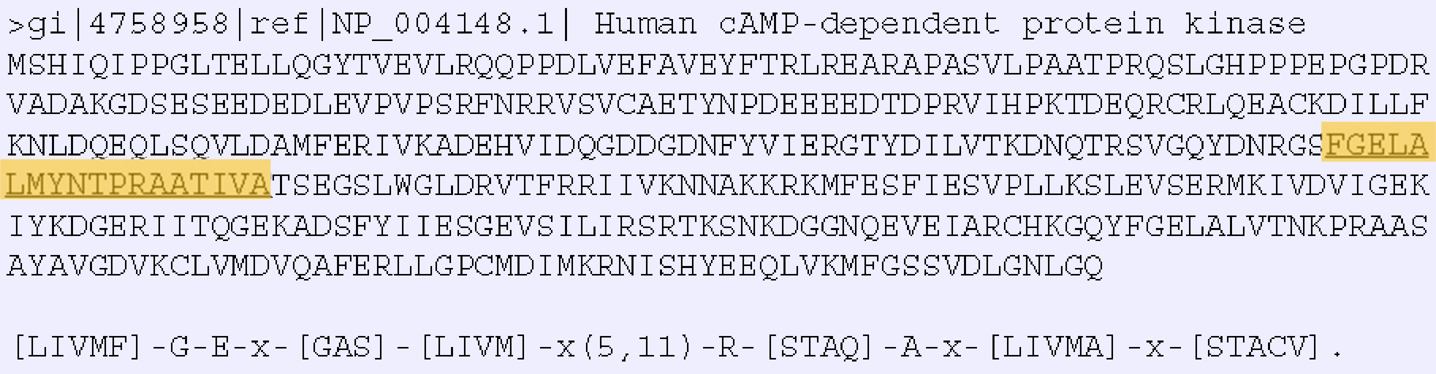
- V současnosti existuje celá řada specializovaných verzí programu BLAST

  - **PHI-BLAST** (**P**attern-**H**it **I**nitiated **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



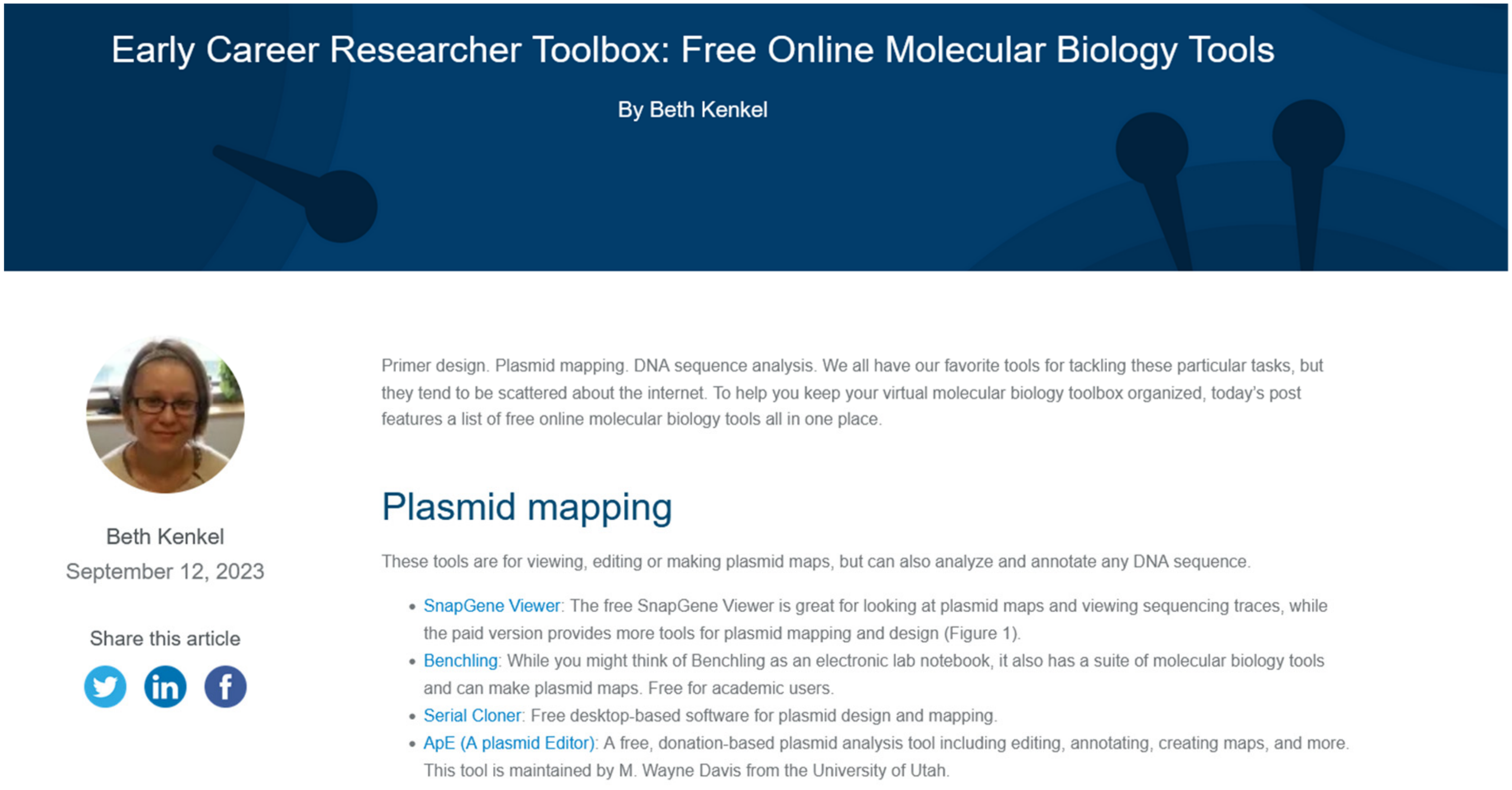
```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAATPRQSLGHPPPEPGPDR
VADAKGDSESEEDEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA
LMYNTPRAATIVATSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIGEK
IYKDGERIITQGEKADSFYIIESGEVSILIRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNISHYEEQLVKMFGSSVDLGNLGQ

[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, restrikčních míst….

Analytické nástroje

https://blog.addgene.org/free-online-molecular-biology-tools



Early Career Researcher Toolbox: Free Online Molecular Biology Tools

By Beth Kenkel

Beth Kenkel
September 12, 2023

Share this article

Primer design. Plasmid mapping. DNA sequence analysis. We all have our favorite tools for tackling these particular tasks, but they tend to be scattered about the internet. To help you keep your virtual molecular biology toolbox organized, today's post features a list of free online molecular biology tools all in one place.

Plasmid mapping

These tools are for viewing, editing or making plasmid maps, but can also analyze and annotate any DNA sequence.

- SnapGene Viewer: The free SnapGene Viewer is great for looking at plasmid maps and viewing sequencing traces, while the paid version provides more tools for plasmid mapping and design (Figure 1).
- Benchling: While you might think of Benchling as an electronic lab notebook, it also has a suite of molecular biology tools and can make plasmid maps. Free for academic users.
- Serial Cloner: Free desktop-based software for plasmid design and mapping.
- ApE (A plasmid Editor): A free, donation-based plasmid analysis tool including editing, annotating, creating maps, and more. This tool is maintained by M. Wayne Davis from the University of Utah.

SnapGene

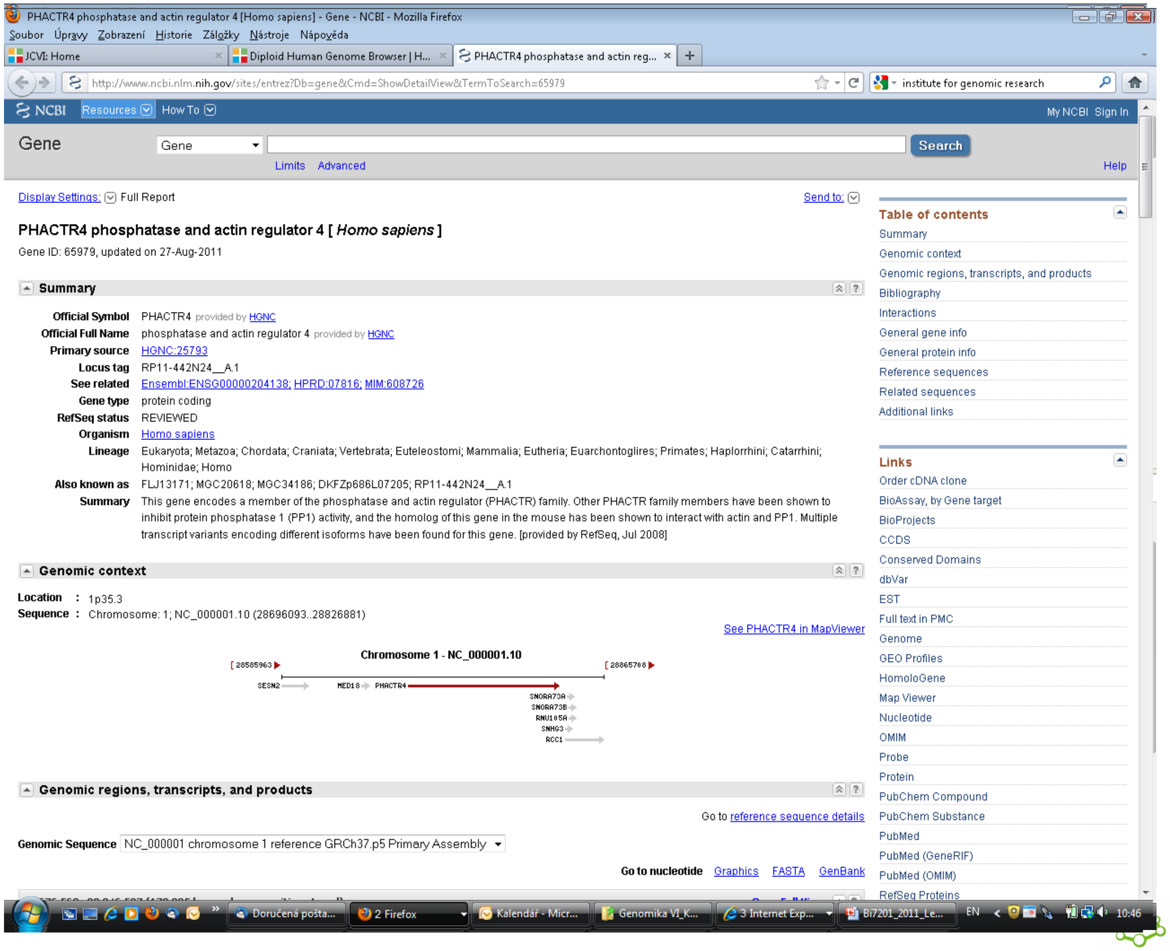
https://www.snapgene.com/snapgene-viewer/download

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, restrikčních míst….
  - Další www genomové nástroje

Další WWW zdroje

- TIGR (**T**he **I**nstitute for **G**enomic **R**esearch), 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:07816; MIM:608726
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; 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 (28696093..28826881)

See PHACTR4 in MapViewer

Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence NC_000001 chromosome 1 reference GRCh37.p5 Primary Assembly

Go to nucleotide Graphics FASTA GenBank

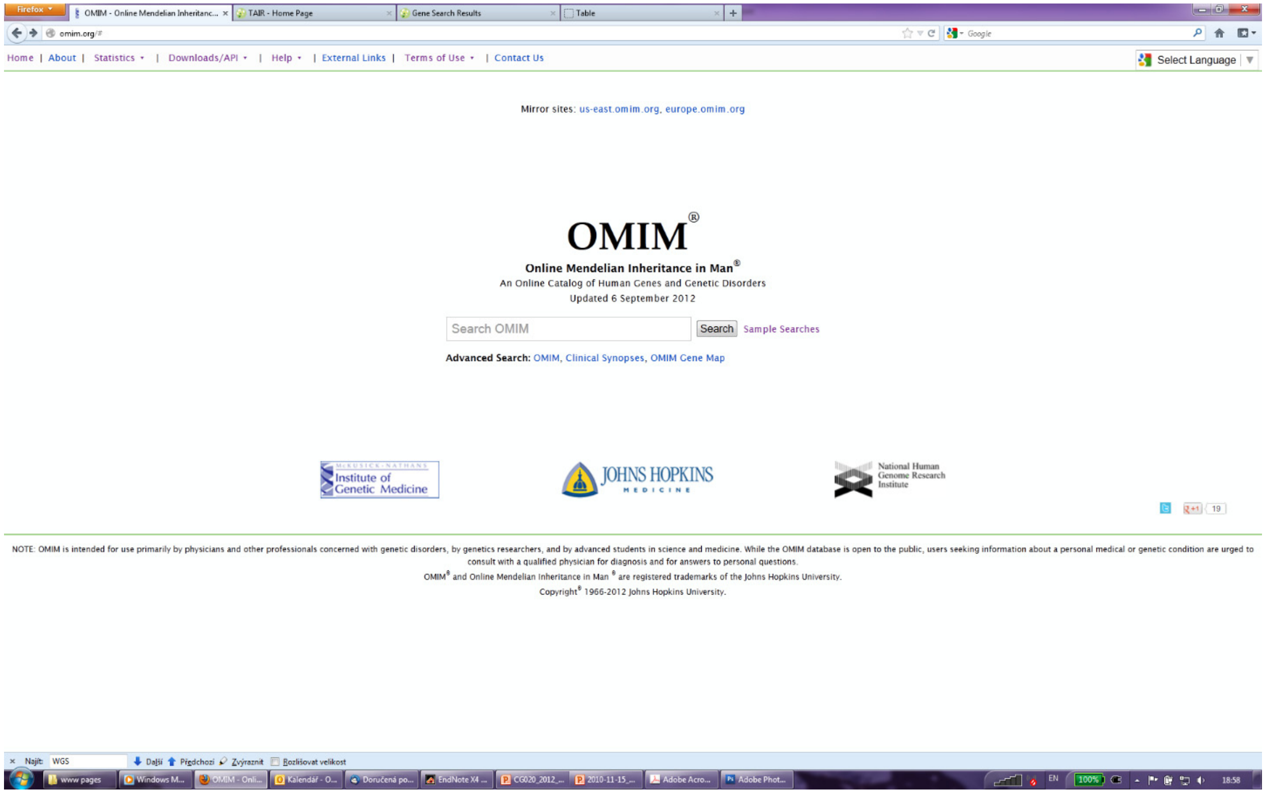
Table of contents

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

- **O**nline **M**endelian **I**nheritance in **M**an (OMIM) http://www.omim.org/



Mirror sites: us-east.omim.org, europe.omim.org

OMIM®

Online Mendelian Inheritance in Man®

An Online Catalog of Human Genes and Genetic Disorders

Updated 6 September 2012

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® and Online Mendelian Inheritance in Man® are registered trademarks of the Johns Hopkins University.

Copyright® 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, restrikčních míst….
  - Další www genomové nástroje

Diskuse