

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

M U N I
S C I



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**
 - 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

„Reverzně genetický“ přístup

5'TTATATATATATATATTTAAAAATAAAATAA
AAGAACAAAAAAGAAAAATAAATA...3'

BIOINFORMATIKA

inzerční mutageneze

FUNKČNÍ GENOMIKA



3

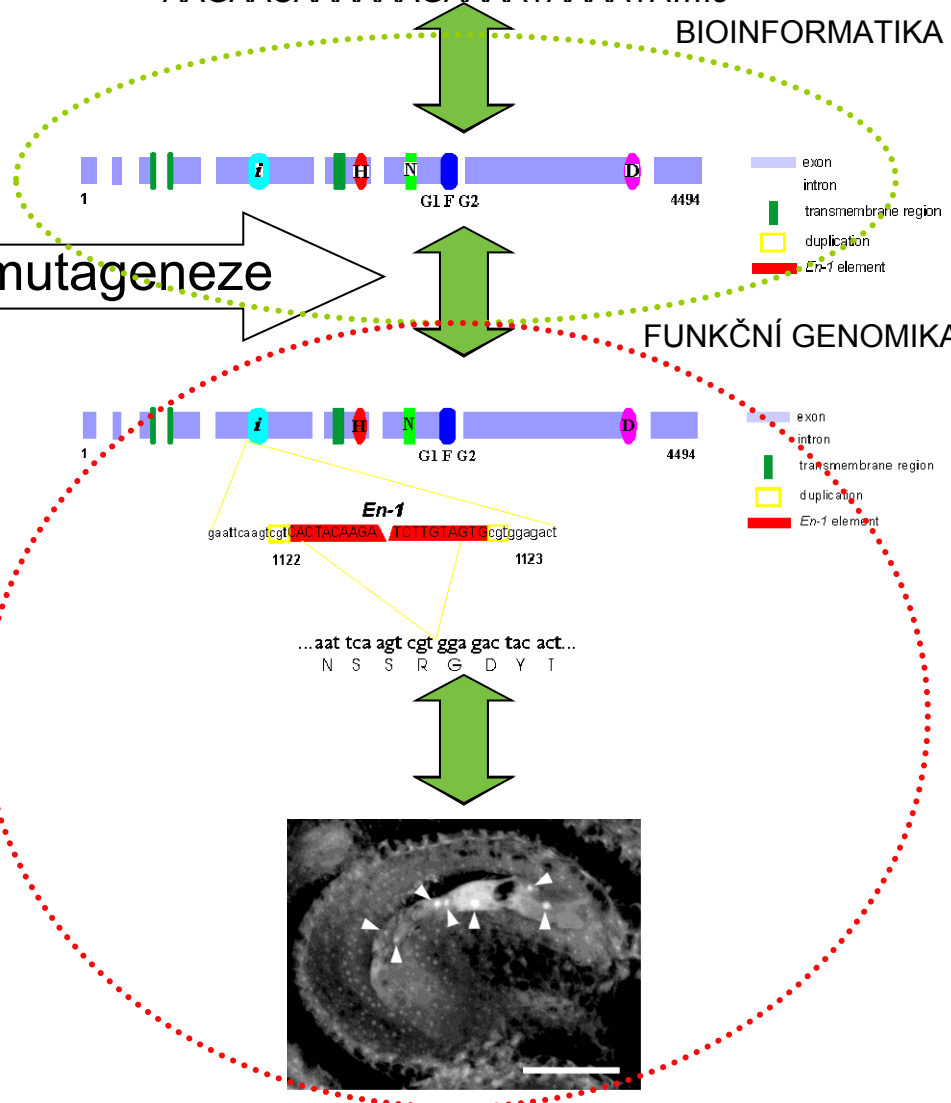
:



1



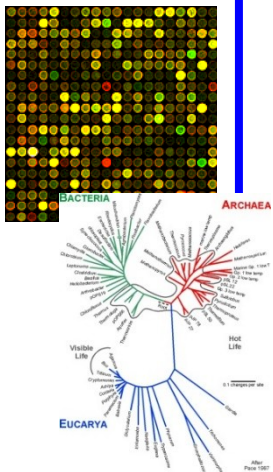
?



Osnova

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

Bioinformatika



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

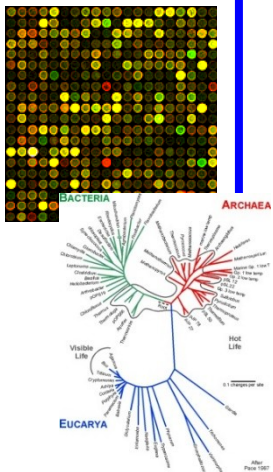
Výzkum, vývoj nebo aplikace výpočetních nástrojů a přístupů za účelem zvyšování rozvoje využití biologických, lékařských, dat o chování nebo zdraví, včetně těch, které umožňují taková data získávat, ukládat, organizovat, archivovat, analyzovat nebo vizualizovat.

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

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.embnnet.org/
BEN	Belgium	http://www.be.embnnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnnet.org/
INFORBIOGEN	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.embnnet.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.embnnet.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.embnnet.org/
BMC	Sweden	http://www.embnnet.se/
SIB	Switzerland	http://www.ch.embnnet.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.unip.edu.ar/embnnet
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.embnnet.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>

The screenshot displays the EBI Services website interface. At the top, there is a navigation bar with links for Services, Research, Training, Industry, and About us. Below this is a large teal banner with the word "Services" and a sub-menu with "Overview", "A to Z", "Service teams", and "Support".

The main content area is titled "Bioinformatics services" and includes a paragraph: "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."

Below this text is a grid of service categories, each with an icon and a brief description:

- 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

At the bottom left of the main content, there is a section for "Programmatic access" with the text: "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".

On the right side of the page, there are several smaller sections:

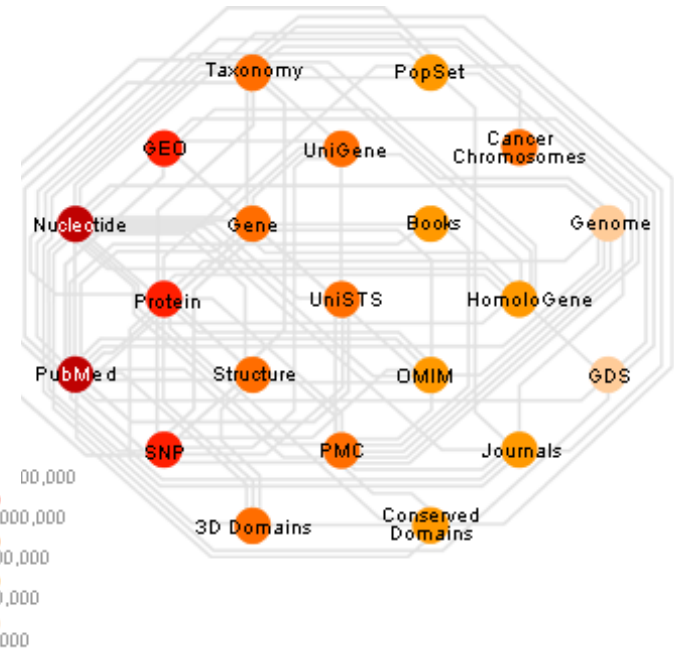
- Popular**: A list of frequently used services including Ensembl, UniProt, PDB, ArrayExpress, ChEMBL, BLAST, Europe PMC, Reactome, Train online, and Support.
- Bioinformatics training**: A section featuring a photograph of people in a meeting.
- Guide to resources**: A section featuring a photograph of a woman smiling.
- Service news**: A section featuring a photograph of a person looking at a screen.

The browser's address bar shows the URL "http://www.ebi.ac.uk/services". The taskbar at the bottom indicates the system is running Windows 7 with various applications open.

Spektrum on-line zdrojů

□ NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' menus, and a 'My NCBI Sign In' link. Below this is a search bar with the text 'All Databases' and a 'Search' button. The main content area is divided into three columns. The left column contains a 'Resource List (A-Z)' with links to various databases and tools. The middle column features a 'Welcome to NCBI' message, a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions', and a 'NCBI YouTube channel' section with a 'GO' button. The right column lists 'Popular Resources' such as PubMed, Bookshelf, and BLAST, and 'NCBI Announcements'.



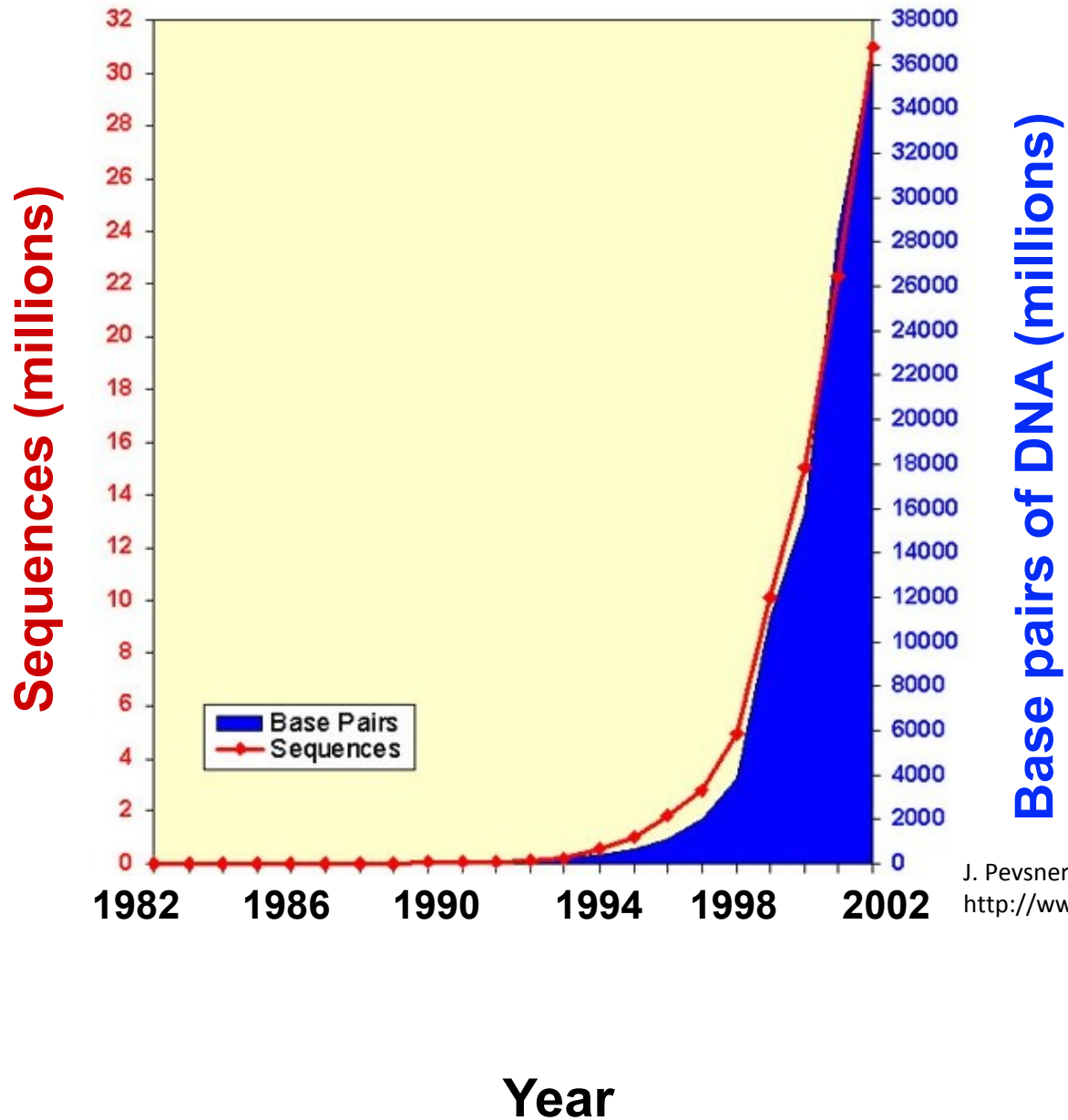
Osnova

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

Primární databáze

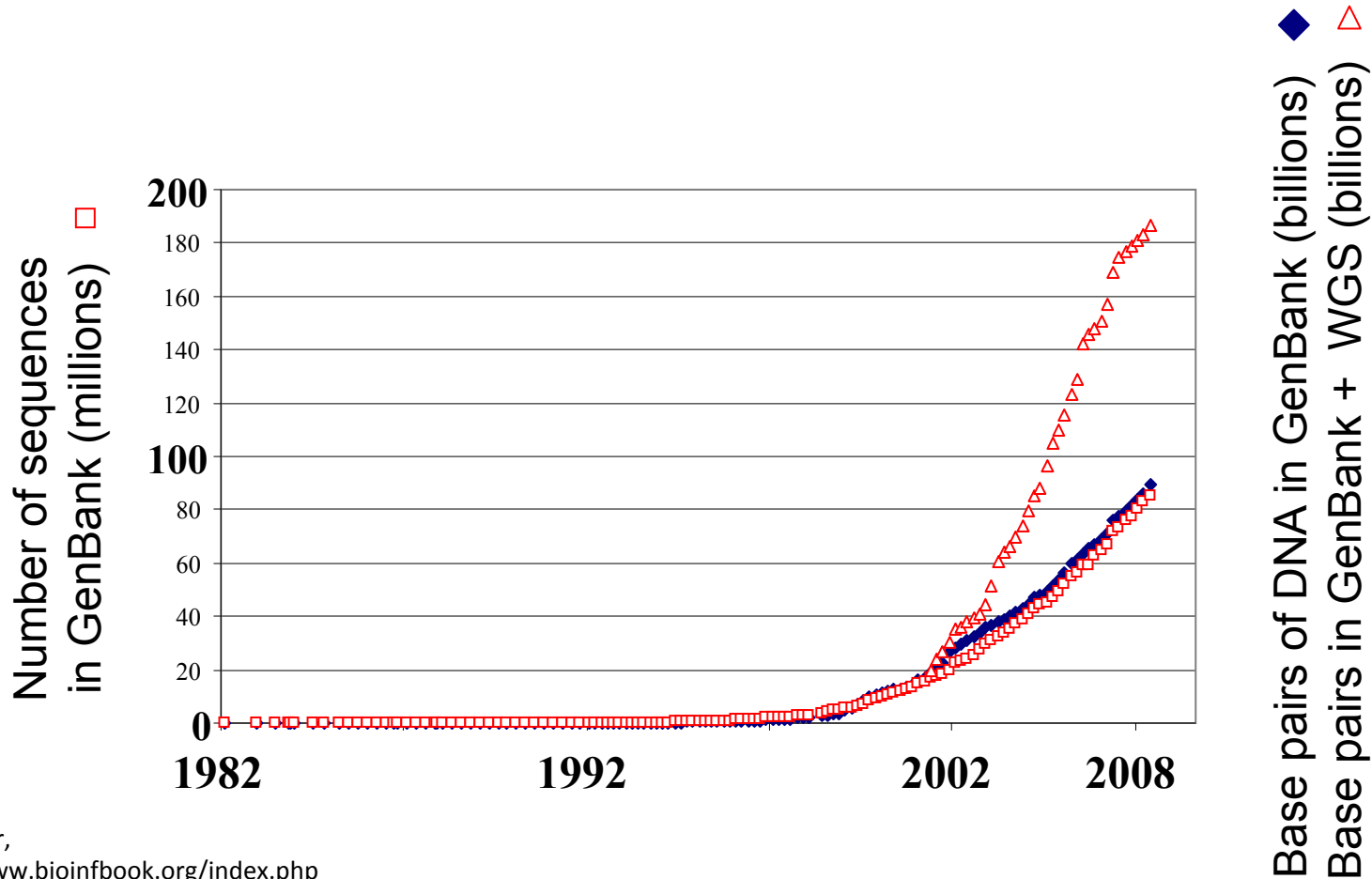
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
 - Sekvence v databázích tzv. „Velké trojky“:
 - **EMBL**
 - <http://www.ebi.ac.uk/embl/>
 - **GenBank,**
 - <https://www.ncbi.nlm.nih.gov/>
 - **DDBJ,**
 - <http://www.ddbj.nig.ac.jp>
 - denně vzájemná výměna a zálohování dat
 - velká datová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

Growth of GenBank



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

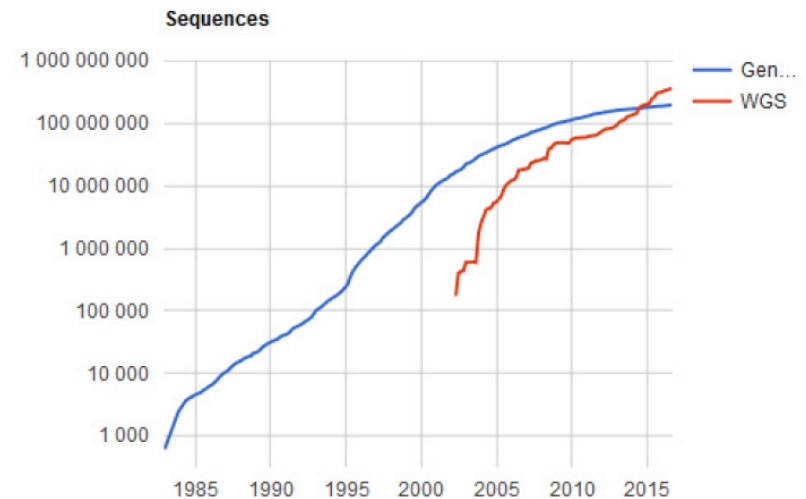
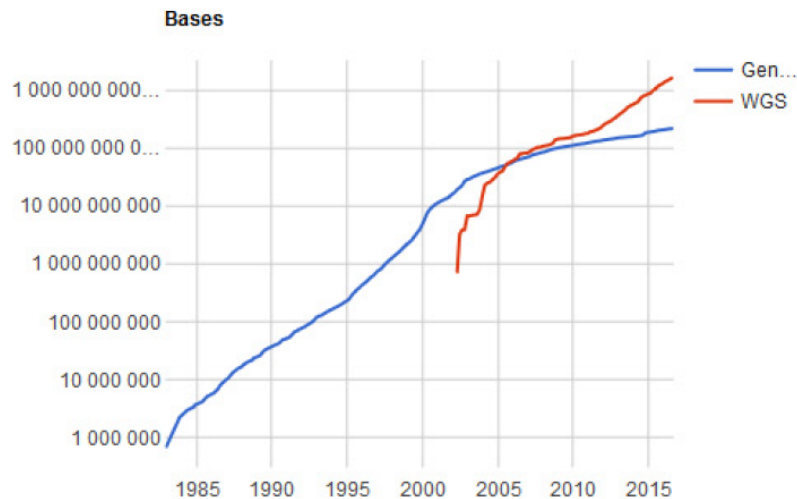
Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases



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

Growth of GenBank

Aug 2016



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

WGS

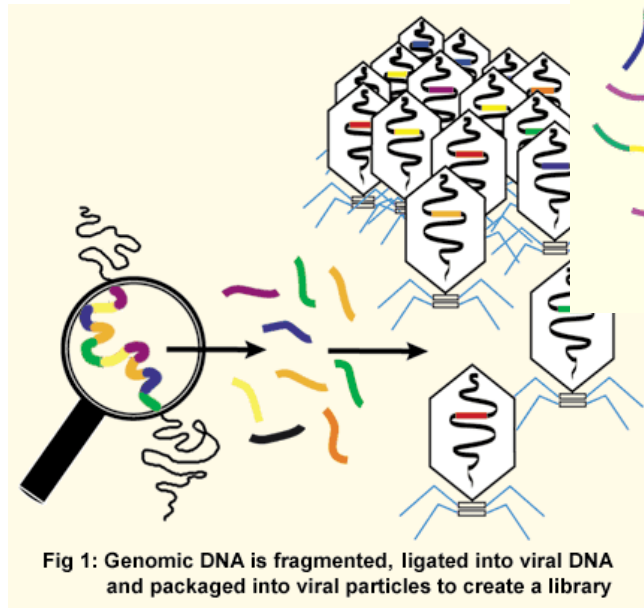
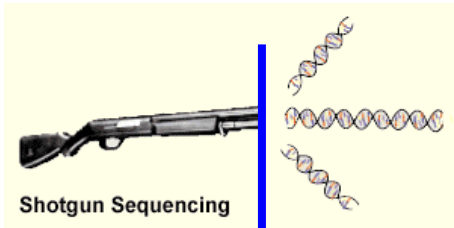


Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

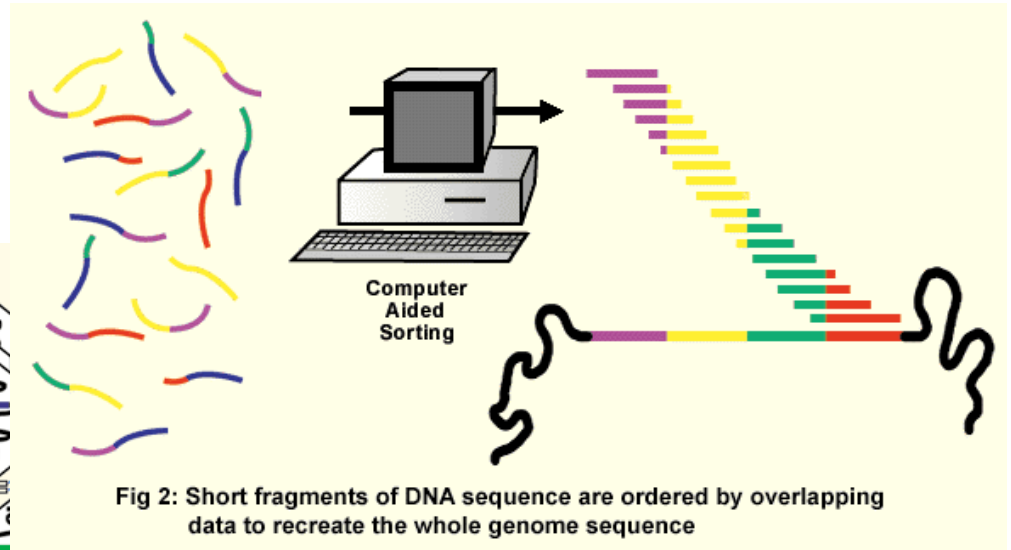
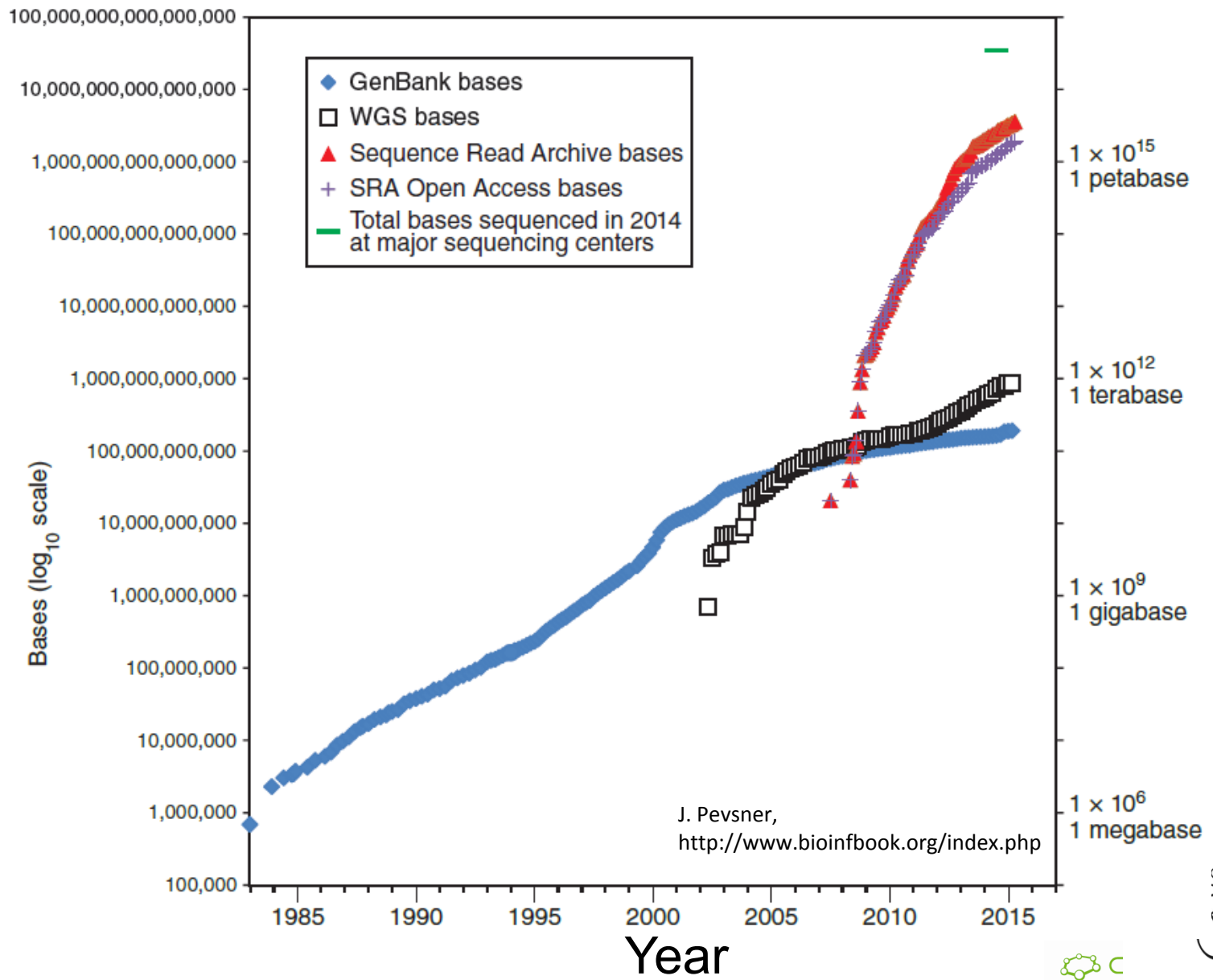


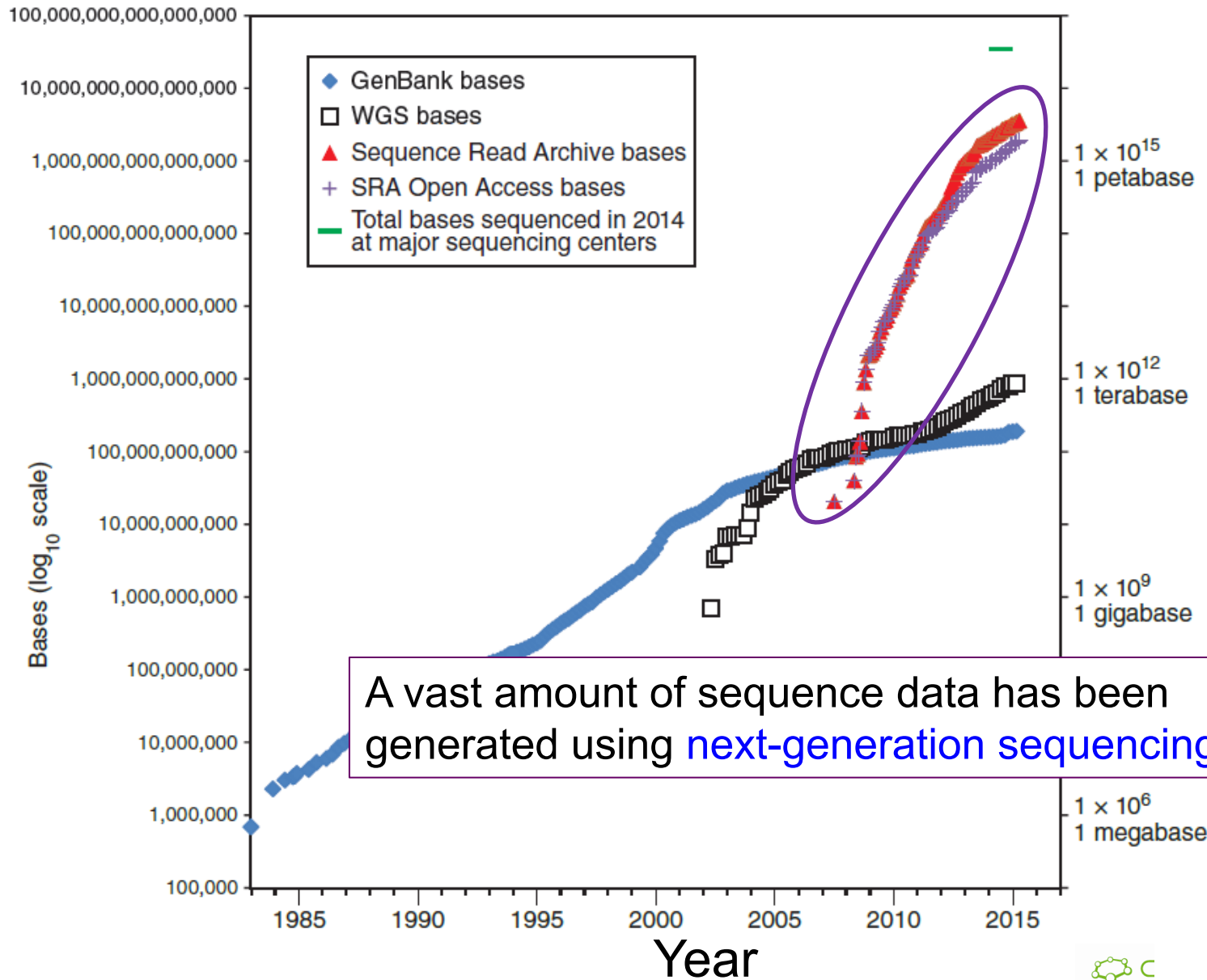
Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

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

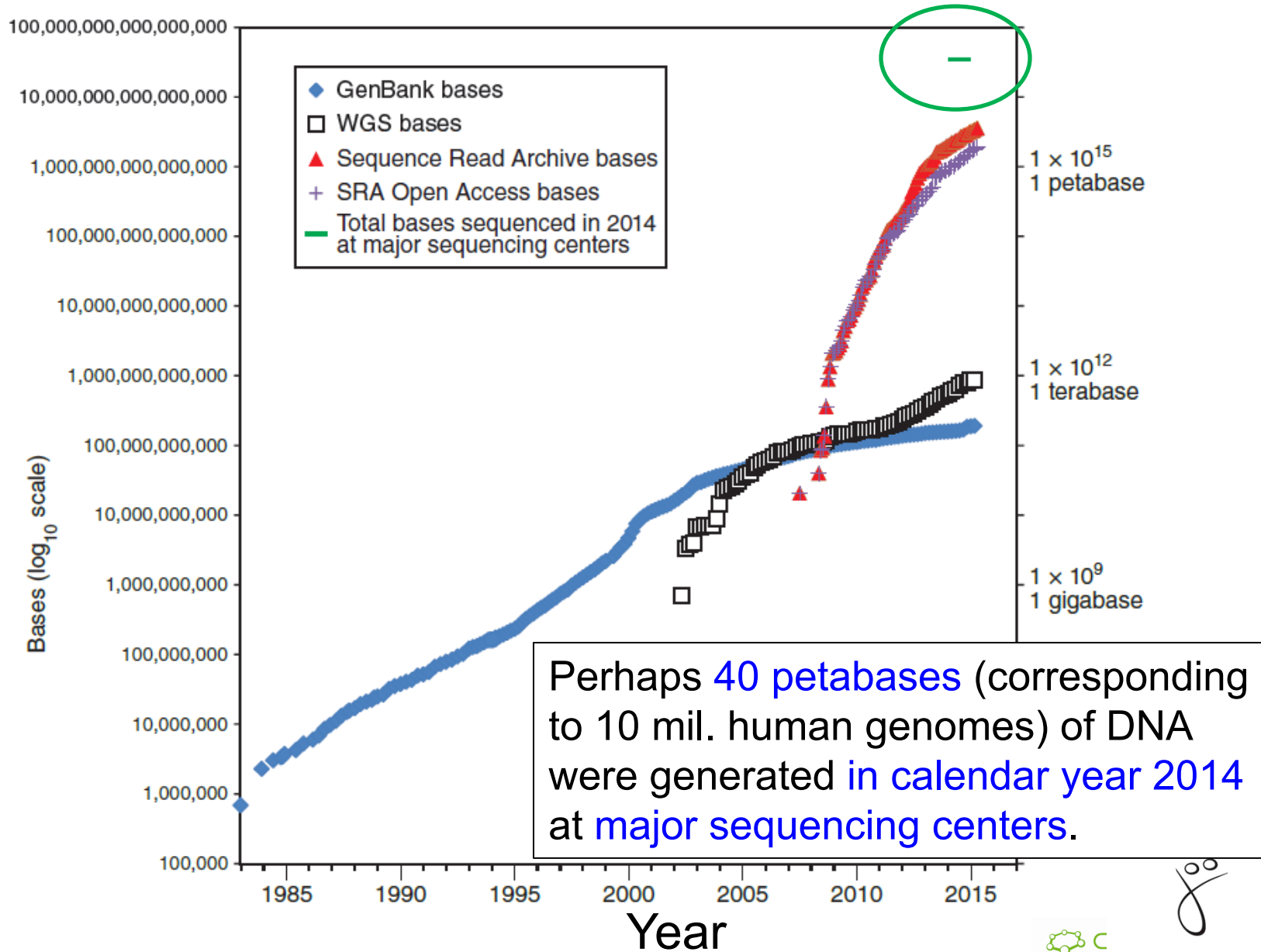
Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



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** (Expressed Sequence Tags)
 - **HGTS** (High Throughput Genome Sequencing)
 - neanotované „surové“ výsledky sekvenačních projektů
 - Referenční sekvence anotovaných genomů
 - **TPAs** (Third Party Annotation)
 - sekvence anotované jinými než původními autory

Primární databáze

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

The screenshot displays the NCBI homepage with the following elements:

- Header:** NCBI logo, "Resources" and "How To" dropdown menus, and "My NCBI Sign In" link.
- Search Bar:** "All Databases" search input and "Search" button.
- Left Navigation Menu:**
 - 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
- Main Content:**
 - Welcome to NCBI:** "The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information." Includes links for "About the NCBI", "Mission", "Organization", "Research", and "RSS Feeds".
 - Get Started:**
 - [Tools](#): Analyze data using NCBI software
 - [Downloads](#): Get NCBI data or software
 - [How-To's](#): Learn how to accomplish specific tasks at NCBI
 - [Submissions](#): Submit data to GenBank or other NCBI databases
- Right Side:**
 - Popular Resources:** PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem.
 - NCBI Announcements:** "New version of GenO available" and "An integrated, downlo for viewing and analy NCBI's July Newslett".
- Footer:** "NCBI YouTube channel" banner with a "GO" button and a video player control bar showing a play button and numbers 1 through 8.

Primární databáze

Firefox - BioProject - NCBI | virA two-component VirA-like sensor... | www.ncbi.nlm.nih.gov/gene/1224316

Summary

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](#)

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]

Genes

virA

Related articles

1. [Sequence analysis of the virA locus of Agrobacterium tumefaciens octopine Ti plasmid pTi15955](#). Schrammeyer 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 2595559.
4. [Analysis of the complete nucleotide sequence of the Agrobacterium tumefaciens virB operon](#). Thompson DJ, et al. Nucleic Acids Res. 1988 May 25. PMID 2837739.

GeneRIF: Gene References Into Functions [What's a GeneRIF?](#)

Submit: [New GeneRIF](#) [Correction](#)

Related sites

- BLAST
- Genome
- BioProject
- Genomic Biology
- GEO
- HomoloGene
- Map Viewer
- OMIM
- Probe
- RefSeq
- UniGene
- UniSTS

Feedback

- Contact Help Desk
- Submit Correction
- Submit GeneRIF

Windows Media | virA two-compo... | Kalendář - Osobn... | Doručená pošta - ... | EndNote X4 - [rea... | C6020_2012_Less... | Adobe Acrobat Pr... | 16:53

Primární databáze

The screenshot displays the NCBI Gene database interface for the gene **NP_059797.1**. The main view shows a genomic map with a scale from 145,400 to 147,600. A red bar represents the gene, with a zoomed-in view showing its structure from 1,400 to 2,000. A tooltip provides the following details:

- 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 tooltip, there are sections for **Bibliography** and **Related articles in PubMed**. The browser's taskbar at the bottom shows several open applications, including Firefox, Windows Media Center, and various utility programs.

Primární databáze

The screenshot displays the NCBI GeneBank interface. At the top, there is a search bar with 'Nucleotide' selected and a search button. Below the search bar, the accession number 'NC_002377.1' is highlighted with a red circle and labeled 'Přístupový kód'. The main content area shows the following details:

LOCUS NC_002377 2490 bp DNA linear BCT 29-DEC-2003
DEFINITION *Agrobacterium tumefaciens* extrachrom plasmid Ti, complete sequence.
ACCESSION [NC_002377](#) REGION: 141684..148183
VERSION NC_002377.1 [GI:10955016](#)
KEYWORDS
SOURCE *Agrobacterium tumefaciens* (Rhizobium radiobacter)

The 'GeneBank Identifier' is highlighted with a red circle and labeled 'GeneBank Identifier'. It is the accession number 'NC_002377.1'.

Below the accession number, there is a list of features including 'source', 'gene', and 'CDS'. The 'source' feature is highlighted with a red circle and labeled 'GeneBank Identifier'. The 'gene' feature is also highlighted with a red circle and labeled 'GeneBank Identifier'. The 'CDS' feature is also highlighted with a red circle and labeled 'GeneBank Identifier'.

Primární databáze

```
/translation="MNGRYSFTRQDFKTGAKPWSILALI VAAIIPAPMAVASWQDNAT
TQAILSQLRSINADSASLQRDVLRAHTGTVANYPPIISRLGALRNLEDLKQLFRQSH
IVREBNAQILLRQLEVLSNSADAAVAAPGQNVRLQDLSLAFTRALSSLPKASTQDT
LKKFTFLASMMQLPLQCPSPALSPFHSLELRLQKQRGLDRAFVRLAREQPTIHSLL
PQVVDLNMVQTSTTARIAEMLQRCLVYLSLKNVRESARIPLSGASVGCCTYIITL
VYLERKKTQMLARLDVYELIKEIGVCFEGEAATSSQAALRIORFPDADTCALAL
VHDERRNAVETPGAKHPKPVWDSVLREIIVESTKADBERATVRIIESKKIVHLPLEIP
GLSILLAHKSTDKLIAVCSLGYQSYEPFCQSGIQLLELATACLCHYIDVRKQTECD
VLARRLRHAQRLEAVGTLAGGIAHEFNILGSELGHARLAQNSVRSVTRRYIDYII
SSGDRAMLIDQILTLRQRKQRMKPFVSESLVTRIAFLRMLPNIPLSFPDQMO
SVIRGSPLELQQVLINICKMASQAMTANQIDIIISQAPLPVKKILAHGVNPPGDYVL
LSISDNHGGIPBAVLPHIPFPFPTTRARNGSTGLGLASVHGHSAPAGYIDVSTVGH
GTRPDYLYLPPSSKPFVNPDSFPGRNKAPRGNGEIVALVEPDDLREA YEDKIAALGYE
FVGFPTFNKIRDWISKGHEADLWVVDQASLPEQSPNSVDLVLKTAISIIGNDLKM
LSRHDVTRDLYLFPKISSR7MAHALTKIKT"
```

ORIGIN

```
1 atgaacggaa gatattcaac gaecggcgac gattttaaga caggcgcgaa gccctgtgtc
61 atattggccc ttaatcgtgc tgaatgatt ttcgcttca tggcggttgc gtcctggcac
121 gacaatcgca ctaccacggc aatcctcagc caactacgat cgtataaagg cgacagcgcc
181 tcaactgacg cggatgtact ccgcgcctac acgggcacgc tggcgaacta ccgcccattc
241 atctccaggc tgggagctct cgggagaagt ctgggaagat tgaagcaatt atttagacaa
301 tctcatattg taagttagag caatgctgct caactgctac gccagctaga agtgctctca
361 aattcggctg acggcggcgt ccgcgccttt ggtgcgcaaa atgtacgctt gcaagattcg
421 ctggccagtt tcaactcgtc tttgagcagt cttccaggaa aagcctcaac cgtacagact
481 ttgaaaaaac caacagaatt ggttagcagt atgctccaat ttcttcggca accaagcccg
541 gctatttcat togagatcag ccttgaacta gagagctcc aaaaaaacg cggcttgtag
601 gaagctcccg tgcgcaact tgcactgtaa ggtccactta tcttatcgct ttgccaacg
661 gtgaajaatc tggtagaact gattcagagc tctgaacccg cagaatbpc gtagatcctg
721 cagcgcagat gtttggaagt ctatagcttg aaaaatgtag aagagcggcg ccgactatc
781 ttctctgggt ccgcttccag ggtctttgct ctctacatca tcaacttagt ctataggeta
841 cgcataaaaa ccgatgggtt agcggcgct ttgattaac aagagotaat caaagacatc
901 gggatgatgt tgaaggtga ggcggccacc acgtcgtccg ccgaagctgc acttcgtatt
961 atccagcgtt tctttgatgc cgtatcgtgc cgtttagctc tagtggacca tgaccgtaga
1021 tggcgtctcg aaacattcgg tgcgaacac ccaaaacctg tgtgggacga cagcgtgcta
1081 cgcgaataag tctctcgtac caaagcggac gaacggcgca cggattcccg catcatatcg
1141 tgcataaaaa tctacattt gctctcggaa atccaggctc tctcgtact actgctcacc
1201 aatccacag ataaactaat tggggttgt tcactgggtt accaaagcta tgcctctoga
1261 ccttgcaag gcgaattca gctcttgaa ctgcgcaacc cctgcctctg tcactatata
1321 gatgtccgc gtaagcagac cgaatcgac gtttggcca gacgatgga gcatgcgcaa
1381 cgccttjagg cagtggtag acttccgcgc ggaatagac atgaattca caactttg
1441 ggtccactcc tccggcaccg agaattaga caaaactcgg tgcctcagac atctgtacc
1501 cgaagatata tgaactatc catttctgca ggcgacagc ccaatgctat tctcgtcag
1561 atcttgaccg tgaecgaaa acaggcgcgc atgatcaagc cacttatggt ctcagcgtt
1621 gtgaccgaaa togetccctt gctacgtatg getctccgcg caaacatoga gcttagttc
1681 agatttgatc aatgcagag cgtgatcgaa ggaagccgc tgaacttea acaggtaacta
1741 ataacatct gcaagaatgc tccccagcc atgactgcaa atggtcaaat cgcactatc
1801 atcagcaaac cttttttacc agttaagaaa atctcggcgc atggtgttat gccacctggc
1861 gactatgttc tctatctat tagcgcaaat ggtggaggca tccccagggc tgtgttacc
1921 caacttttg aacctctct taacacagca gctcgcaacc gtggaacggg tctcggctt
1981 gctctcgtgc atggtcaatc cagcgcgctt cggggttaca togagcttag tcaactgtt
2041 gggcatggga cgcgcttga ctttatctc cctccgttt ctaaggaaac cgtaaactca
2101 gacagtttt tccgcgcaa taagcaccgc cgtggaaacg gggagattg ggcacttct
2161 gacccgatg acctcctcgc gtaggcgtat gaagacaaga tcccgctct aggatctgag
2221 ccgctcgtgt tctacactt taatgaaat ccgcatggga ttccaaaagg caatgaagcc
2281 gatctgctca tggtagacca agcgtctctt cctgaagatc aaagtctcaa tctcgtgat
2341 ttatgctca agaccctc catcatcatt ggcggaatg atctcaaat gaccttca
```

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

The screenshot shows the NCBI RefSeq page for the gene 'two-component VirA-like sensor kinase'. The page is viewed in a Firefox browser window. A yellow oval highlights the search bar area. The main content is organized into sections: 'Genome Annotation', 'Reference assembly', and 'mRNA and Protein(s)'. Under 'Reference assembly', there is a table for genomic data. Under 'mRNA and Protein(s)', there is a table for protein data with descriptions of conserved domains.

two-component VirA-like sensor kinase

[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	HikKA; 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

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

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

Primární databáze

The screenshot displays the NCBI Gene database interface for the gene **NP_059797.1**. The main view shows a genomic map for **NC_002377.1: 145K..148K (2.9Kbp)** with a scale from 145,400 to 147,600. A red bar represents the gene, with a zoomed-in view showing a length of 2,490 bp. A tooltip provides the following details:

- 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 gene entry, there are sections for **Bibliography** and **Related articles in PubMed**. The browser's taskbar at the bottom shows several open applications, including Firefox, Windows Media Center, and various utility programs.

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
ATGAACGGAGATATTCACCGACGGCGCAGGATTTAAGACAGGGCGAAGCCTTGGCTATATTGGCCC
TTATCGTGTGCTGCAATGATTTTCGCGTTTCATGGCGGTTGCGTCTGCGCAGGACAATGGCACTCCCAAGC
AATCTCAGCCACTACGATCGATTAAACCGCCGACAGCGCCTCACTGACGCGGATGACTCCGCGCTCAC
ACGGGCACCGTGGCGAAGTACCGCCCATTTATCCAGGCTGGGAGCTCTGCGGAAGAATCTGGAGATT
TGAAGCAATTTAGACAATCTCATATTTAAGTGAGAGCAATGCTCAACTGCTACGCGAGCTAGA
AGTGTCTTAAATTCGGCTGACGCGCGGTGCGCGCCTTTGGTGGCAAAATGTACGCTCGCAAGATTG
CTGGCCAGTTTCACTCGTCTTTGAGCAGTCTTCCAGGAAAAGCTCAACCGATCAGACTTTAGAAAAAC
CAACAGAAATGGCTAGCATGATGCTCCAATTTCTCGGCAACCAAGCCCGGCTATTTCAATTCGAGATCAG
CCTTGAATAGAGGGCTCAAAAAACAACGGCTCTTGAAGCTCCCGTGGCCTACTTGCAGCTGAA
GGTCCCATTTATCTTTCGCTTTGCCACAGTGAAGATCTGGTGAACATGATTCAGAGCTGACACCG
CAGAAATTCGGGAGATGCTGACGCGGAGTGTGGAGGCTTATAGCTTGAAGATGTAGAGGCGGGAG
CCACGATCTTTTGGGTCGGCTTCAGTGGGTCTTTCCTCTACATCACTCACTTACTATAGGCTA
CGCAAAAAACCGATTGGTTAGCGCGGCTTGAAGTACGAAGAGCTAATCAAGAGATCGGAGTATGTT
TTGAAGGTGAGCGCGCCACCACTCGTCCGCGCAAGCTGCACTTCGATATTATCAGCGCTCTTTTGAATGC
CGATACGTGCGGTTAGCTCTAGTGGACCATGACCGTAGATGGGCTGCGAAACATTCGGTGGCAAAAC
CCAAACCTGTGTGGGACGACAGCGTGTACGCGAAATAGTCTCTCGTACCAAGCGGACGAACGGGCGA
CGGATTCGCGATCATATCGTCAAAAAAATCGTACATTTGCTCTCGAAATTCAGGTCTCTCGATACT
ACTGGCTCAAAATCCACAGATAAATTAATGCGGTTTGTCTCACTGGTTACCAAGCTATCGCCCTCGA
CCTTGCACAGCGGAAATTCAGCTTCTTGAATCGCCACCGCCTGCTGCTACTATATCGATGTTCCGCG
GTAAAGCAGCGAATGCGACGTTTTGGCCAGCAGATTGGAGCTGCGCAACGCTTGGAGCAGTTGGTAC
ACTTCCGCGGGAATAGCACATGAATTAATAACATTTGGGCTCACTCCTCGGCGACGAGAAATAGCA
CAAACTCGGTGTCTCGAACATCTGACCCGAAAGATATATTGACTATATCATTTGCTCAGCGCAGAG
CCATGCTCATTTCGATCAGATCTTGACGCTGAGCGAAACAGGAGCGCATGATCAAGCCATTTAGTGT
CTCAGAGCTGTGACCGAATCGTCCCTTGTACGATGGCTTTCGCGCAAAACGAGCTTAGTTTC
AGATTTGATCAAAATGACAGCGTGTGATCGAAGGAGCCGCTTGAATTCACAGGTAATAAATCAACATCT
CGAGAAATGCTCCCAAGCATGACTGCAAAATGGTCAATCGACATCATCAGCAAGCTTTTTTACC
AGTTAAGAAAATTCGGCGCATGTTGTTATGCCACTGGGACTATGTTCTCTCTATATAGCGACAAT
GGTGGAGCATTCGGAGSCTGTGTACCCCAATTTTGAACCTTCTTACGACAGGACTCGCAACG
GTGAAACGGTCTCGGCTTGTCTGTGATGTCATATCAGCGCGTTTGGGCTTACATCGACTTAG
TTCAACTTGGGATGGGCGGCTTGAATTTATCTCCCTCCGCTCTTCAAGAACCCGTAATCCA
GACATTTTTTTCGCGCAATTAAGCCACCGCTGAAACGGGAGATGTGGCACTTGTGAGCCCGATG
ACCTCTCGGCGAGCGTATGAAGACAAGATCGCCGCTCTAGGATATGAGCCGTTGCTGATCTT
TAATGAAATTCGCGATTGGATTTCAAAGGCAATGAAGCGCATCTGTCATGCTCGCAAGCGCTCTT
CCTGAGATCAAAGTCTTAATTCGCTGGATTTAGTGTCAAGACCGCCTCCATCATCTTTGGGAAAG
ATCTCAAATGACCCCTTCAAGGAGGATGTGACAGGAGACCTTTATCTCCGAGCGGATATCGTCCAG
AATCTGGCGCATCAATCTAACCAAAATCAAGACGTAG
```

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

Sekundární databáze

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

EXPASy Home page	Site Map	Search ExPASy	Contact us	Swiss-Prot	PROSITE	Proteomics tools			
Hosted by SIB Switzerland		Mirror sites:	Australia	Belgium	Canada	China	Korea	Taiwan	USA
Search <input type="text" value="PROSITE"/> for <input type="text"/> <input type="button" value="Go"/> <input type="button" value="Clear"/>									



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	Search Swiss-Prot with a PROSITE entry
<p>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:</p> <pre>MDVKVTKLVASRPVIVFCVLAFLVVFVPECTWISNWRVTR DLVKEVASPTRELDLSTLSVSELENIQKFTYAKTNLSTIGLA RVLDIYITNEDTGFTEIQTLAPLFPVASTLIQVQVSY LISDGLMSFYIASNNTVAVPANSSESSSDYTWYTVV DCLTGRLENGSTESQSLDVTHTDNPQAQSNHYTAPVGT SLOGEDNEMLIQEVVSLYKELVSLGPFVKTLEVLNLSL DLRHEELVWTKEDGTVLVREGSLNDSFPISNGSICPSRHS INLAWQCIPEHCSSGTEVEIKELRYQAFCEVIRVGVFL</pre> <p><input type="button" value="Clear"/></p>	<p>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)</p> <p><input type="text"/></p>
<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 (User Manual) (You may also specify a PROSITE entry in the box to the right)</p> <p><input type="checkbox"/> Exclude patterns with a high probability of occurrence</p> <p>Your e-mail (optional): <input type="text"/> (will send results by e-mail)</p> <p><input type="checkbox"/> plain text output</p> <p><input type="button" value="START THE SCAN"/> <input type="button" value="RESET"/></p>	<p>and specify your search limits:</p> <ul style="list-style-type: none">• The <input checked="" type="checkbox"/> Swiss-Prot <input type="checkbox"/> TrEMBL <input type="checkbox"/> TrEMBLnew <input type="checkbox"/> PDB databases (You may also specify a protein in the box to the left) <input checked="" type="checkbox"/> including splice variants• The following taxa: <input type="text"/> (see NFWT Taxonomy; separate multiple taxa with a semicolon, e.g. <i>Homo sapiens; Drosophila</i>, Not available for PDB.)• Sequences with at least <input type="text"/> hits• At most <input type="text" value="1000"/> matches <p>Advanced options: <input type="checkbox"/> FASTA output <input type="checkbox"/> retrieve complete sequences allow at most <input type="text" value="1"/> X sequence characters to match a conserved position in the pattern match mode <input type="text" value="greedy, overlaps, no includes"/> (for patterns, see help) randomize databases <input type="text" value="no"/> (to test a pattern, see help)</p>

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 nkeesatYeteians

>[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 SsK
219 - 221 SsK
369 - 371 TrR
460 - 462 SgK
513 - 515 SgR
585 - 587 SsR
602 - 604 TgK
652 - 654 TgK
716 - 718 SpR
726 - 728 SpK
747 - 749 TgK
794 - 796 SsR
854 - 856 SsK
864 - 866 SsR
868 - 870 SsR
921 - 923 SpK
957 - 959 SvR
960 - 962 TgR
974 - 976 TgK
997 - 999 SsK
1002 - 1004 TgK
1018 - 1020 SgK
1031 - 1033 TgR
1119 - 1121 SsR

Sekundární databáze

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

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

found: 2 hits in 1 sequence

USERSEQ1 (1122 aa)

```
MMVKVKLVASRPVVFVCLAFVVFECIWNINWRITTENLVKEVASFTEDLRTSLVSEIENIGK
FTYAKTNLSTIGLARVIDSVIINNDTGFTEIQIQAPELLFVAYSTILQVSQVSYISRDLGMLFSYIA
ESNTSVAVFANSSNSRSGDYTWYQTVDQLTGRNGNSTKQSLDVTHTDWFQAAQSNNYTAFV
GTSLGGEDNETLIQSVVSLYSKGLVSLGFPVKLTLEVLNLSLHGEELYMWTKDGTVLVREGSLN
DSFFISNGSICFGRESNLSNSQCIPENCSSSGYEVEIKRLRYQAFCVSEVSGVPLRVYTLMPFNKG
GATRIKHQAEKAKYQLIVVMIFLGFGWVFWVFMQATREMHMRATLINQMEATQQAERKSMNK
SQAFANASHDIRGALAGMKGLIDICRDGVKFGSDVDTLNLQVNVCAKDLVALLNSVLDMSKIESGK
NQIVVEEDFNLSKLLLEDVIDFYHFVAMKRGVDVLDPHDGSVFKFSNVRGDSGRKQIILNLYSNAV
KFTVDGHIAVRANAQRFGSSSIVLASYPKGVSKFVMSMFCNKKEESSTYETEISNSIRNNANTHE
FVFEVDDTGKGIPEMRRKSVFENVVQVRETAQHQGTGLGLGIVQSLVRLMGGSEIRITDKAMGEGK
TCFQFNVLTI+LESFPVSDMKVRQ+EIAGGDV+VSTPNLGLTINTSLGGSMNIRNLSRPFNNCLSSS
PKQEGSRVLLLNKEERRRVTEKYIKNLGIKVTVEKWEHLSYALERLFGFSPQSSMGRAECSLSC
FSSRELFFIGMDGIDRSQLPKRRSISFSAVLLVIDAKTGFPELCLDIVKQFRRLPHGISCKVV
WLNESSTRVSEGDISCSRPLHGSRIMEVLKMLPEFGGTV+LKEPPELQRESLLRHSFVAERSPKH
KVQEEGPPSMFNKLGKRI+MASTDSESETRVKSVRTGRKPIGNPEDEQETS+KPSDDEFLRGRKVLV
VDDNFI+SRKVATGKLRKMGVSEVEQCDSGKEALRLVTEGLTQREEQGSVDKLPFDYIFMDCQMP+EM
DGYEATREIRKVEKSYGVRTPIIAVSGHDPGSEEAR+ETIQAGMDAFLDKSLNQLANVIREIESKRH
```

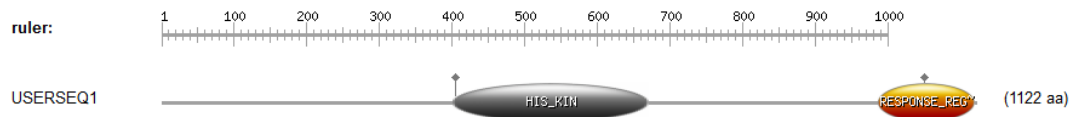
Legend:

disulfide bridge active site other 'ranges' other sites

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

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

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



Sekundární databáze

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

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

found: 2 hits in 1 sequence

USERSEQ1 (1122 aa)

```
MMVKVTKLVASRPVIVVFCVLAFLVVFECIWINSNWRITTEENLVKEVASFTEDLRISLVSEIENIGK
FTYAKTNLSTIGLARVIDSYITNNDIGFTEIQTQIAPLLFVAYSTILQVSVSYISRDGLMFSYIA
ESNTSVAVFANSSNSRSDYTWYQTVVQDLTGRNLGNSTKQSLDVTHTDFWQAAQSNNTYITAFV
GTLGGEDNETLIQSVVSLYSKGLVSLGFFVKTLTEVLNLSNLHGEELMWTKDGTVLVREGSLN
DSFFISNGSICFGRESNLSWSQCIPENCSSSGYEVEIKRLRYQAFCVSEVSGVPLRYTLMFENKG
GATRIKHQAQKAKYQLIVVMI FLGFGWVFWVFMQATRREMHMRATLINQMEATQQAERKSMNK
SQAFANASHDIRGALAGMKGLIDICRDGVVFGSDVDTLQVNVCAKDLVALLNSVLDMNKIESGK
MQLVEEDFNLSKLELDVIDFYHFMKKGVDVLDPHDGSVFKFSNVRGDSGLKQLLNLSVNAV
KFTVDGHIAVRAWAQRPGSNSSVLA SYFKGVSKFVKSMFCNKKEESSTYETEISNSIRNNANTME
FVFEVDDTGKIPMEMRKSVEFENVVQVRETAQGHQGTGLGLGIVQSLVRLMGGEIRITDKAMGEGK
TCFQFNLLTLESPPVSDMKVRQIEAGGDYVSTPNLGLTINTISLGGSMNIRNLSRPNCLSSS
PKQEGSRVLLKNEERRRVTEKYIKNLGIKIVVEKNEHLSVALERLFGFSPQSSMGRACSLSC
PSSRELFFIGMDGIDRSQLPKRRSISFSAVLLVIDAKTGFPELFDIVKQFRGLPHGISCKV
WLNESSTRVSERGDISCSRPLHGSRLMEVLKMLPEFGGTVLKEPTELQRESLLRHSFVAERSPKH
KVQEEGPFSSMFKKLGKRIMASTDSESETRVKSVRTGRKPIGNPEDEQETSCKPSDDEFLLGKRVLV
VDDNFISSRKVATGHLKRMGVSEVEQCDSGKEALRDVTEGLTQREEQSSVDKLPFDHIFMDCQMPEN
DGYEATRIRKVEHSYGVRTPIIAVSGHDPGSEEARETIQAGMDAFLDKSLNQLANVIREIESKRH
```

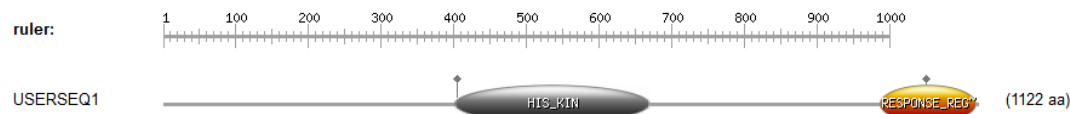
Legend:



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

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

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



Sekundární databáze

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



PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a *SWISS-PROT/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:

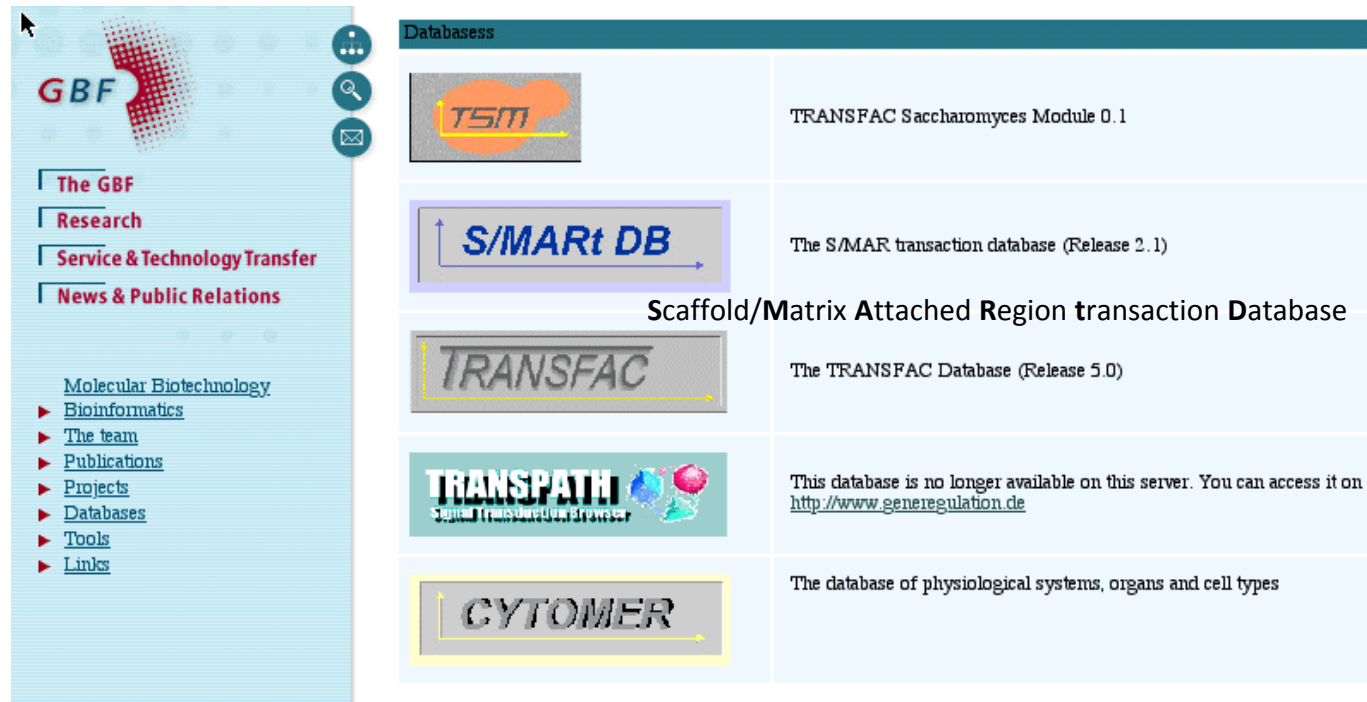
- [By accession number](#)
- [By PRINTS code](#)
- [By database code](#)
- [By text](#)
- [By sequence](#)
- [By title](#)
- [By number of motifs](#)
- [By author](#)
- [By query language](#)

PRINTS search:





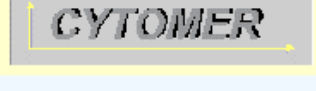
- [Search PRINTS with NEW FingerPRINTScan](#)
- [FPScan](#)
- [GRAPHScan](#)
- [MUI_Scan](#)
- FingerPRINTScan binaries and source are available: contact.scordin@bioinf.man.ac.uk

Sekundární databáze

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



The screenshot shows the GBF website interface. 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 resources.

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

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

PROTEIN DATA BANK

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

[ABOUT PDB](#) | [DATA UNIFORMITY](#) | [RECENT FEATURES](#) | [USER GUIDES](#) |
[FILE FORMATS](#) | [EDUCATION](#) | [STRUCTURAL GENOMICS](#) | [PUBLICATIONS](#) |
[SOFTWARE](#)



[Did you find what you wanted?](#)

Search the Archive

Enter a [PDB ID](#) or keyword

[Query Tutorial](#)

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](#) | [pdb-4 Archive](#)
[Newsletter](#) | [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



Structure Explorer - 1P5Y

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



[View Structure](#)

[Summary Information](#)

[View Structure](#)

[Download/Display File](#)

[Structural Neighbors](#)

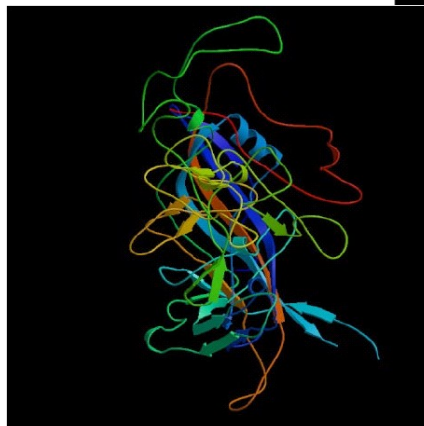
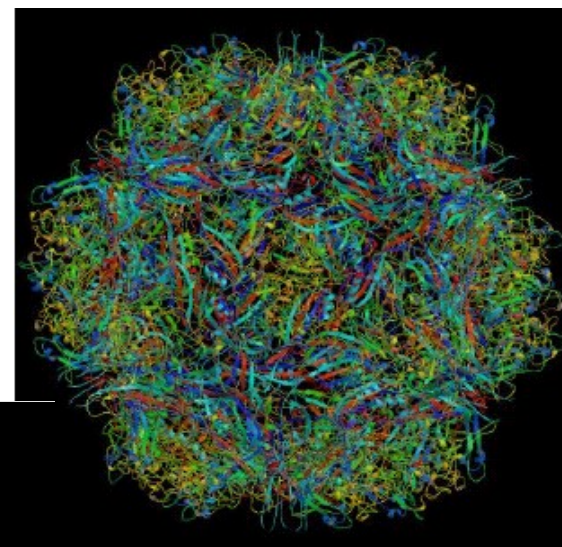
[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

Explore

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

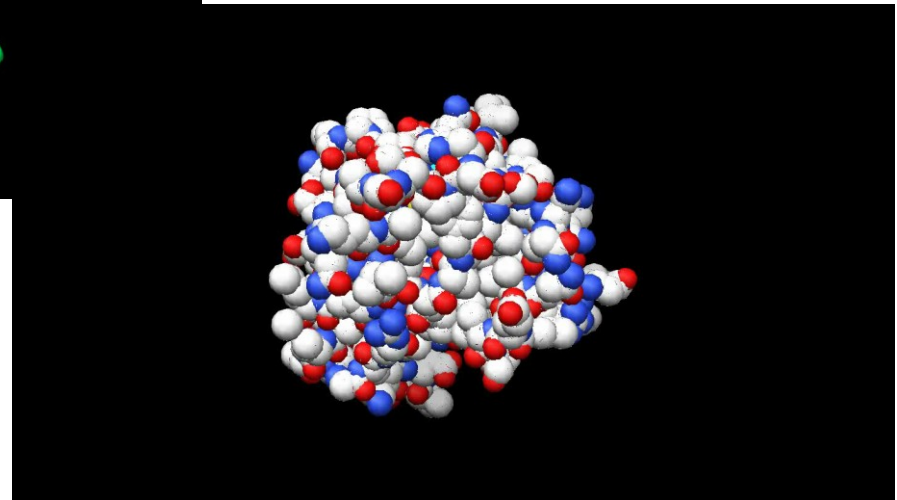


<http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=:pid=173561064349344&bio=1&opt=show&size=500>

12/29/2003

Strukturální databáze

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



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


Osnova

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

Genomové zdroje

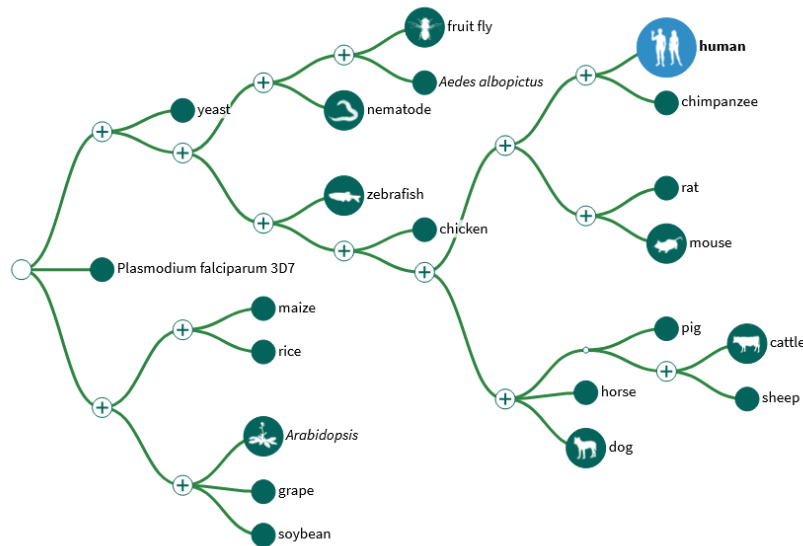
- NCBI Genome Data Viewer <https://www.ncbi.nlm.nih.gov/genome/gdv/>

Genome Data Viewer

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

Select organism

Homo sapiens (human)



Homo sapiens (human) genome



Search in genome

Location, gene or phenotype



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

Assembly

GRCh38.p13

Browse genome

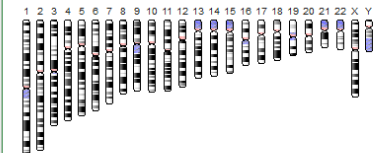
BLAST genome

Assembly details

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

Annotation details

Annotation Release 109
Release date 2020-08-17



Genomové zdroje

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

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

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

[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_g000212	Displays all of the unplaced contig g000212
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 15q11.1 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

U C S C
Homo sapiens
(Graphic courtesy of CGSC)

Genomové zdroje

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

The screenshot displays the UCSC Genome Browser interface for human chromosome 11 (GRCh37/hg19). The main track shows the gene structure for the UBE3A gene, with exons represented by black boxes and introns by lines with arrows. Other tracks include RepeatMasker, RefSeq, and various annotations. A green arrow points to the gene structure track. The interface includes navigation controls, search bars, and a list of available tracks categorized into 'Mapping and Sequencing Tracks' and 'Phenotype and Disease Associations'.

Mapping and Sequencing Tracks

- Base Position
- Chromosome Band
- STS Markers
- FISH Clones
- Recomb Rate
- dsCODE
- Recomb
- ENCODE Pilot
- Map Contigs
- Assembly
- GRC Map Contigs Gap
- Publications
- BAC End Pairs
- Fosmid End Pairs
- GC Percent
- GRC Patch Release
- Hq18 Diff
- GRC Incident
- Hi Seq Depth Wiki Track
- BU ORChID
- Mapability
- Short Match
- Restr Enzymes

Phenotype and Disease Associations

- GAD View DECIPHER
- OMIM AV SNPs
- OMIM Genes
- OMIM Pheno Loc
- COSMIC
- GWAS Catalog
- ISCA
- RGD Human QTL
- RGD Rat QTL
- MGI Mouse QTL
- GeneReviews

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-END##

Transcript exon combination evidence: V00497.1, BU659180.1 [ECO:0000332] ##RefSeq-Attributes-START##

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
	Other Names	GeneReviews	Model Information	Methods	Pathways

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 (SCKA) [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.

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

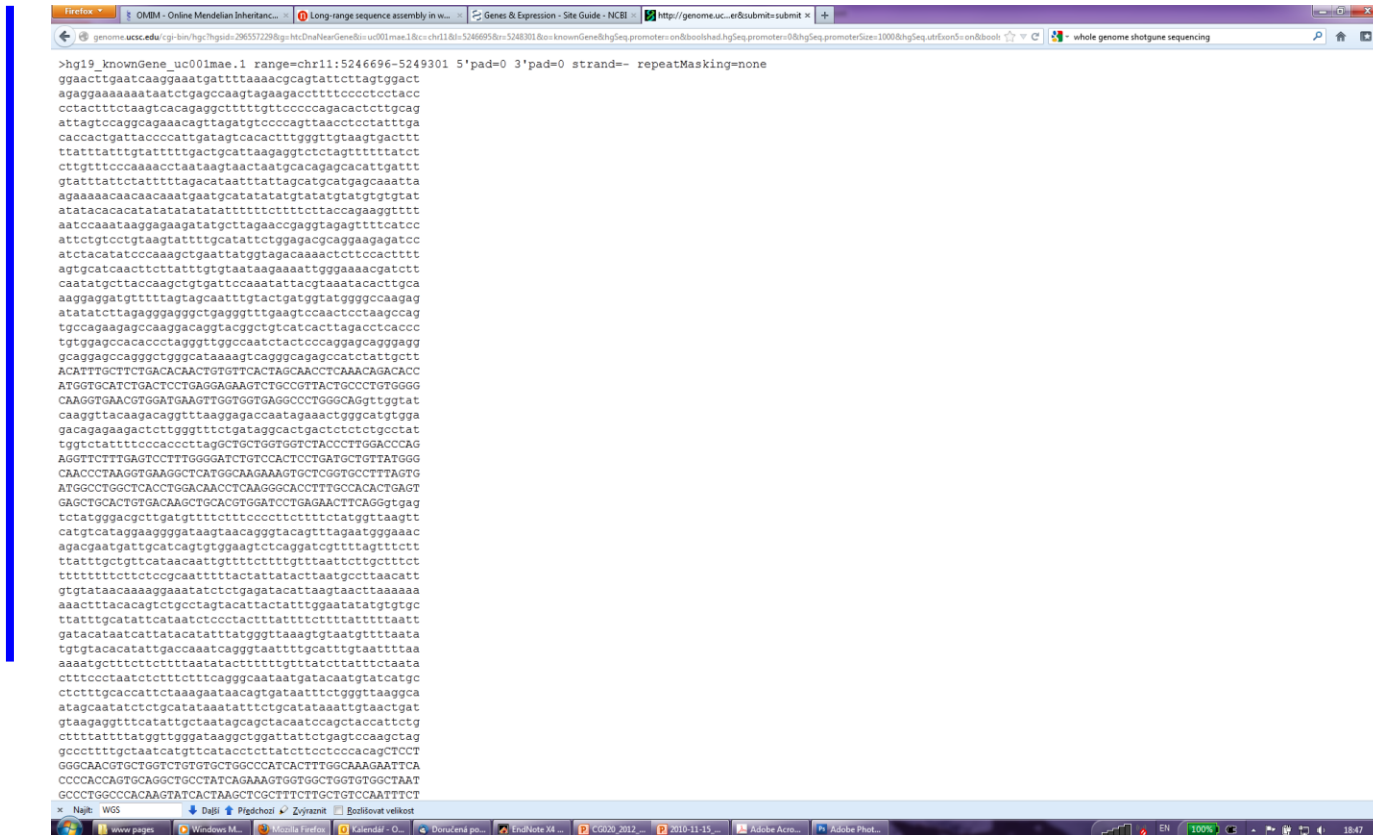
Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

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

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
ggaacttgaatcaaggaatgattttaaaacgcagatctcttagtgact
agaggaaaaaataatctgagccaagtagaagacctttcccctcctacc
cctacttctaagtcacagaggcttttggccccacgacactctgag
atagtcacagcagaacaagtagatgccccagttaacctcctatttga
caccactgattaccaccctgtagagcacaacttgggtgtaagtgactt
ttacttattgtatttttgaactgcataagaggtctcagttttatctc
ctgtttcccaaacctaataagtaactatgacacagacacattgattt
gtattattctatttttagacataatttattagcatgcatgagcaatta
agaaaaaacacaacaatgaatgcatatataatgatatgtatgtgtat
atacacacatatataatatttttttttcttaccagaaggtttt
aatccaaataaggagaagatagtctagaaccgagtagaytttcatcc
attctgctcctgaagtattttgcatattctggagcgcggaagagatcc
atcacatatcccaagctgaattatgtagaacaacctcccactttt
agtgcatacaactcttatttggtaataaagaaatgggaaaaagatctt
caatagcttaccagctgtgattccaaatatacgtaaatacacctgca
aaggagaggtttttagtagcaattgtaactgagtagtggggccaagag
atatacttagaggaggtctgaggttttaggtccaactcttaagcag
tgccagaagaccagaagcaggttagctcacttagacctcacc
tgtggaccacacctagggttggccaactcactccaggagcaggaggg
gcaggagccaaggtctggcataaaatcagggcagagccatctatgctt
ACATTGGCTTGACACACTGTGTTCACTAGCAACCTCACAACAGACACC
ATGGTGATCTGACTCCTGAGGAGAAGCTGCCGTTACTGCCCTGTGGGG
CAAGGTGAACGTGGATGAAGTTGGTGGTAGGCCCTGGGCAggttggtat
caaggttcaagcaggtttaaaggaccaaataagaaactgggcatgtgga
gacagagaagactctgggttctctgtaggcaactgactctctgctat
tggctatttcccacccttaggctgctgggtgctacaccttggaccag
AGGTTCTTTAGTCTTTGGGGATCTGCCACTCTGATGCTGTTATGGG
CAACCTTAAGGTGAAGGCTCATGGCAAGAACTGCTGGGTGCTTTAAGT
ATGCTGCTCACTCAAGGCAACTCAAGGCACTTTGCAACCTGACT
GACTGCACTGTGACAAAGCTGCACCTGGATCTTGAACTTCAGSgtgag
tctatgggagccttgatgttttctcccctcttctctatggttaagt
catgcataggaagggataagtaacaggtacagtttagaagggaac
agacgaatgatgcatcagtggaagctcaggatcgtttagttctctt
ttattggttccataacaattgtttcttggtaactctgcttctct
tttttttctctccgcaattttactatataacttaagcttaacatt
gtgtatacaaaaggaatatctctgagatataagtaacttaaaaaa
aaactttacacagctgcttagtacctactatttggaaatatagtgtg
ttattgcatatcataactcctcacttattttctttatttttaatt
gatcataactatatacatatttgggttaaggttaagtgttctaata
tggtaacatattgccaactcaaggttaagttgcatattgtaatttaaa
baatgcttctcttttaataatacttttggtaactattcttaata
cttcccatactcttcttcagggcaaatgatacaatgtagatgc
ctcttgcacattctaaagaataacagtgataattctgggttaaggca
atagcaatctctgcataataatattctgcatataaattgtaactgat
gtaagaggttctatttggtaatgacagctacaatccagctaccatctg
ctttatttttgggtgggataaggctgagattctgagtccaagctag
gccctttgctaatcatgttatacctctatctctcccacagCTCCT
GGGCAACGTGCTGGTCTGTGCTGCCCACTACTTGGCAAGAAATCA
CCCCACAGTGCAGGCTGCTATCAGAAAGTGGTGGTGGTGGTAAAT
CCTTGGCCCAAGTATCAACTAGCTGCTTCTTCTGCTTCCAATTTCT
```

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" and the address bar shows "www.arabidopsis.org". The website features a navigation menu with links for Home, Help, Contact, About Us, and Login/Register. Below the navigation is a search bar and a menu with options: Search, Browse, Tools, Portals, Download, Submit, News, and ABRC Stocks. The main content area is titled "The Arabidopsis Information Resource" and contains several sections: a detailed description of the resource, information about the Arabidopsis Biological Resource Center, funding details from the Carnegie Institution of Science and the National Science Foundation, and a prominent advertisement for a new online submission form. The submission form advertisement includes a graphic of a laptop with a form titled "TAIR SUBMISSION" and fields for Article #, Gene, Locust, GeneFunction, and Method. It also features a "SUBMIT PAPER" button and a "SUBMIT DATA" button. The text below the graphic reads: "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". To the right of the main content, there are "Breaking News" sections, including "2012 MASC Report Now Available [July 11, 2012]", "New Protein Chip and Cell Cultures at ABRC [May 9, 2012]", and "Share Your Education Resources [February 1, 2012]". The browser's taskbar at the bottom shows several open applications, including Firefox, Windows Explorer, and various office software.

Genomové zdroje

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



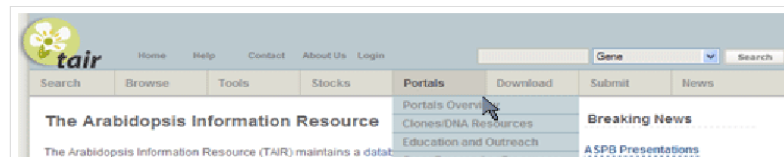
The Arabidopsis Information Resource

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.

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 homologí

Analytické nástroje

□ Globální vs. lokální přiřazení

```

Globální přiřazení
SLAV-----APATNIK-----PIQNYR-I-----AKSETQRYMVIE
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE

Lokální přiřazení
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE
-----NAPATNIKSECVRA-PIQNYRRVEHVRA-----

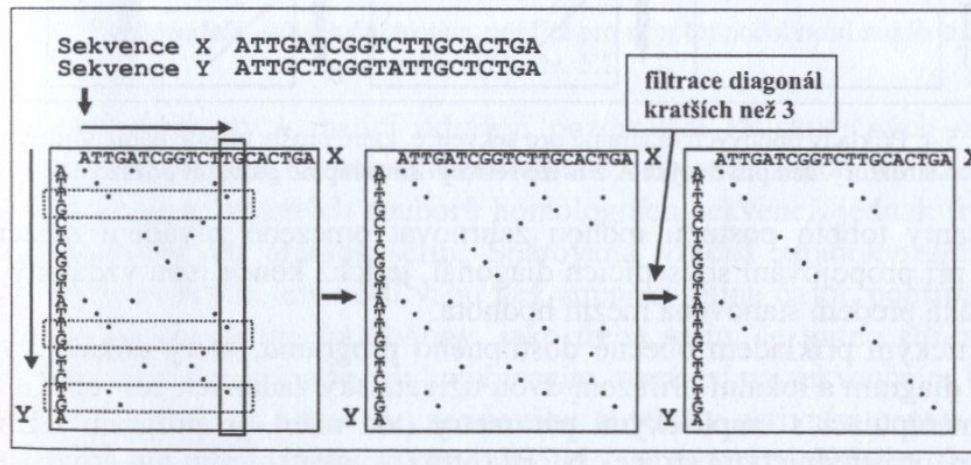
```

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)

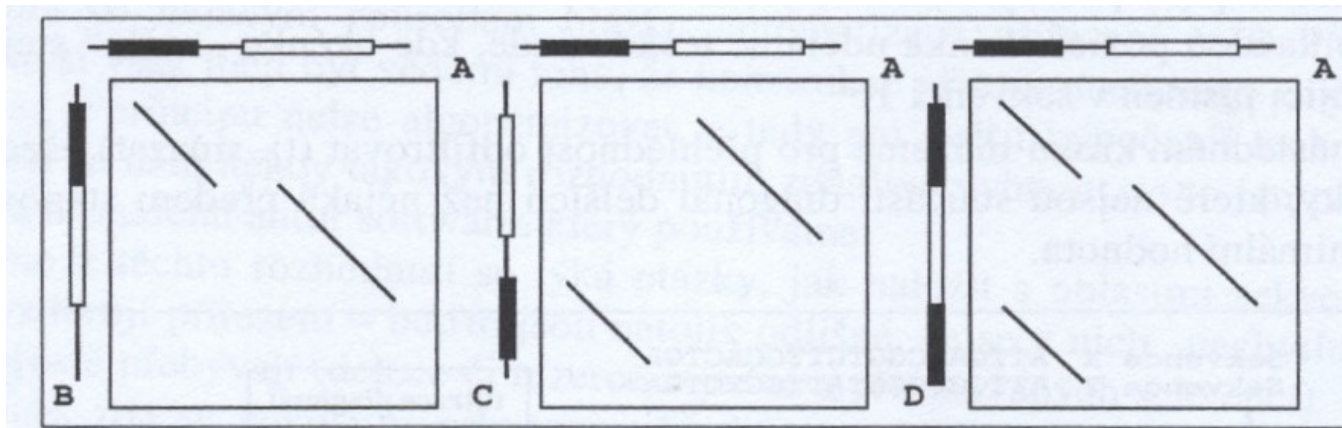


Cvrčková, Úvod do praktické bioinformatiky

- vnesení 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

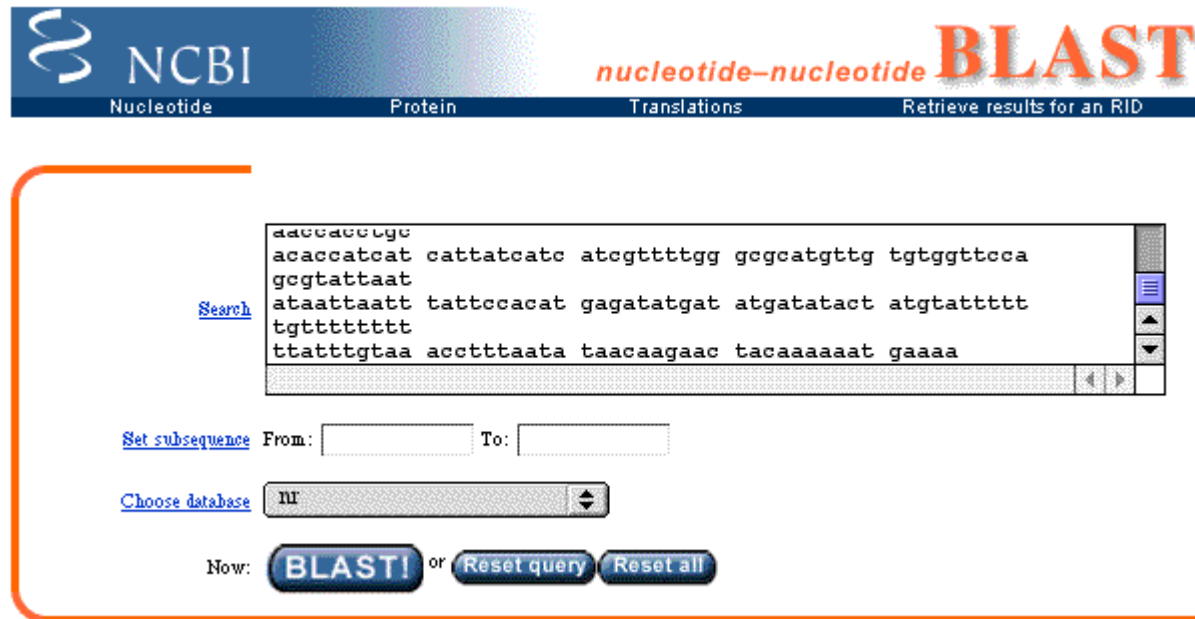


Cvrčková, Úvod do praktické bioinformatiky

- **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/>



The screenshot displays the NCBI BLAST web interface. At the top, the NCBI logo is on the left, and the text "nucleotide-nucleotide BLAST" is on the right. Below the logo, there are four tabs: "Nucleotide", "Protein", "Translations", and "Retrieve results for an RID". The "Nucleotide" tab is selected. The main input area contains a text box with the following sequence: `aaccaccctgc
acaccatcat cattatcacc atcgttttgg ggcgatgttg tgtggtteca
gcgatattaat
ataattaatt tattccacat gagatatgat atgatatact atgtatTTTT
tgTTTTTTTT
ttattttgtaa acctttaata taacaagaac tacaaaaaat gaaaa`. To the left of the text box is a "Search" link. Below the text box are two input fields for "From:" and "To:". Below these is a "Choose database" dropdown menu with "nr" selected. At the bottom, there are three buttons: "BLAST", "Reset query", and "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

Diagram illustrating a 4x4 substitution matrix for nucleotides (A, T, G, C). The diagonal elements (A-A, T-T, G-G, C-C) are 1, representing self-matches. The off-diagonal elements (A-T, A-G, A-C, T-G, T-C, G-C) are 0, representing mismatches. Labels point to the G-A mismatch (0) and the G-G match (1).

Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250

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

```
>gi|5016088|ref|NM_001101.2|
Length = 1793

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

Query: 156 gtcgacaacggctctgcatgtgcaaggccggatttgccggagacgatgctccccgcgcc 215
          |||
Sbjct: 101 gtcgacaacggctccggcatgtgcaaggccggcttcgcgggcgacgatgcccccgggcc 160

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

Query: 276 aaggactcgtacgtgggtgatgagggcagagcaagcgtggtatcctcaccctgaagtac 335
          |||
Sbjct: 221 aaggattcctatgtggggcgacgagggcccagagcaagagaggcatcctcaccctgaagtac 280

Query: 336 cccattgagcaccggtatcgtgaccaactgggacgatatggagaagatctggcaccacacc 395
          |||
Sbjct: 281 cccatcgagcaccggcatcgtcaccaactgggacgacatggagaaaatctggcaccacacc 340
```

- „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

The screenshot displays the NCBI Gene database interface for the gene **NP_059797.1**. The browser address bar shows the URL www.ncbi.nlm.nih.gov/gene/1224316. The main content area shows a genomic track for **NC_002377.1: 145K..148K (2.9Kbp)**. A red bar represents the gene **NP_059797.1**, with a green arrow pointing to it. A tooltip window is open over the gene name, providing 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 gene track, there are sections for **Bibliography** and **Related articles in PubMed**. The Windows taskbar at the bottom shows several open applications, including a web browser, calendar, and Adobe Acrobat.

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
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 1737127	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 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 nukleotidové sekvence](#) přeložené do sekvence [aa](#) v [proteinové databázi](#)

BLAST

Specializované verze

- V současnosti existuje celá řada 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 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** (**P**attern-**H**it **I**nitiated **B**LAST)
 - 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  
MSHIQIPPGTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPEPGPDR  
VADAKGDSESEEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF  
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGNFYVIERGTYDILVTKDNQTRSVGQYDNRGS  
TSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSRMKIVDVIgek  
IYKDGERRITQGEKADSFYIIESGEVSIILRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYEEQLVKMFGSSVDLGNLGQ
```

```
[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

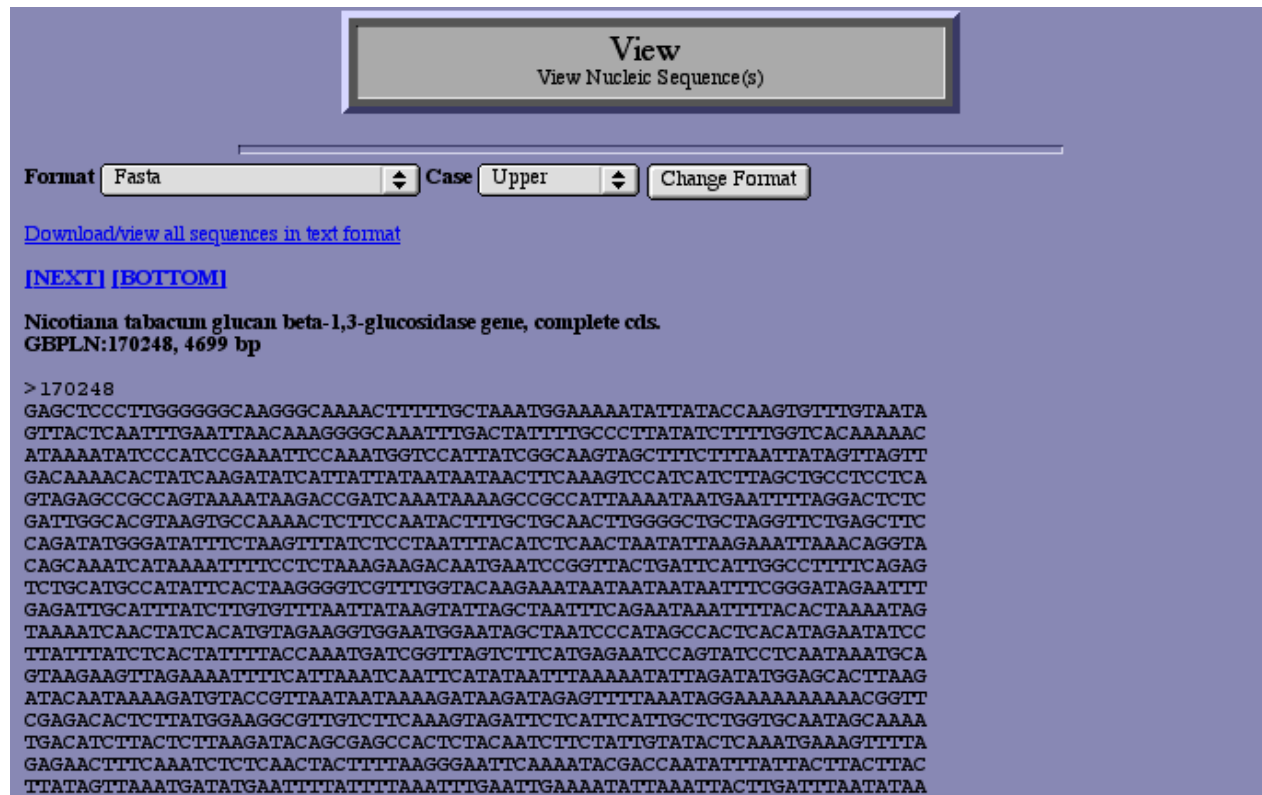
GBPLN:804655 **Hordeum vulgare L. beta-glucosidase (BGQ60) gene, complete cds.**
 GBPLN:170248 **Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds.**

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.
EBI

Analytické nástroje

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



The screenshot shows the 'View' interface of the Biology Workbench. At the top, a box labeled 'View' contains the text 'View Nucleic Sequence(s)'. Below this, there are two dropdown menus: 'Format' set to 'Fasta' and 'Case' set to 'Upper'. A 'Change Format' button is located to the right of the 'Case' dropdown. Below the dropdowns, there is a link: 'Download/view all sequences in text format'. Further down, there are two more links: '[NEXT]' and '[BOTTOM]'. The main content area displays the following text:

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

> 170248
GAGCTCCCTTGGGGGGCAAGGGCAAAACTTTTTGCTAAATGGAAAAATATTATACCAAGTGTITGTAATA
GTTACTCAATTTGAATTAACAAGGGGCAAAATTTGACTATTTTTGCCCTTATATCTTTTTGGTCAAAAAAC
ATAAAATATCCCATCCGAAATTC AAAATGGTCCATTATCGGGCAAGTAGCTTTCTTTAATTTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTTATAATAAATAACTTCAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAAAAGACCGATCAAAATAAAGCCGCCATTAAAAAATGAAITTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACCTTCCAATACTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTAACAGGTA
CAGCAAATCATAAAATTTTCCCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCITTTTCAGAG
TCTGCAATGCCATATTC ACTAAGGGGTCGTTTGGTAC AAGAAATAATAATAAATTTTCGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTAAATPATAAGTATTAGCTAATTTCAAGATAAATTTTACACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACCAAATGATCGGTTAGTCTTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAAATTTTCATTAATCAATTCATATAAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAAGATGTACCGTTAATAATAAAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTTGTCTTCAAGTAGATTCCTATTCTATTGCTCTGGTGCATAGCAAAA
TGACATCTTACTCTTAAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAAATGAAAGTTTTA
GAGAACTTTCAAATCTCTCAACTACTTTTTAAGGGAATTC AAAATACGACCAATTTTTATTACTTACTTAC
TTATAGTTAAAATGATATGAATTTTTATTTTTAAATTTGAATTGAAAATATTAATTTACTTGTATTTAATATAA

Analytické nástroje

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

Regex pattern:

```
ott.{1,32}ott
```

0 sequences were searched

1 match was found

Matches are indicated in blue

```
> 170248
GAGCTCCCTTGGGGGGCCAGGGCAAAACTTTTGGCTAAATGGAAAAATATTATACCAAGTGTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCCACAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTTCCTTTAAATTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAATAATAAACTTCAAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAAATAAGACCGATCAAAATAAAAGCCGCCATTAAAAATAATGAAATTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACCTTCCCAATACTTTGGCTGCAACTTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAAGAAATTAACAGGTA
CAGCAAATCATAAAATTTTCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCATGCCATATTCCTAAGGGGTCGTTTGGTACAGAATAATAATAATAATTTCCGGATAGAAITTT
GAGATTGCATTTATCTTGTGTTTAAATTATAAGTATTAGCTAATTTTCAGAATAAAATTTTACACTAAAAATAG
TAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACC AAATGATCGGTTAGTCTTTCATGAGAATCCAGTATCCCAATAAATGCA
GTAAGAAGTTAGAAAAATTTTCATTTAAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACC GTTAAATAATAAAAGATAAGATAGAGTTTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGGCTTGTCTTCAAAGTAGATTCTCAITTCATTTGCTCTGGTGCAATAGCAAAA
TGACATCTTACTCTTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAAATGAAAGTTTAA
GAGAACTTTCAAATCTCTCAACTACTTTTAAAGGGAATTCAAAATACGACCAATATTTATTACTTTACTTAC
TTATAGTTAAATGATATGAATTTTATTTTAAATTTGAATTTGAAAAATTTAAATTTACTTTGATTTAATATAA
ACAAATAGATATCGCTAAGTATTTACCACAAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT
GATTAAGCAGCTATTCATCTGGTTGTGCAAGGATGAAAGAAAGTAACTAGCTATAATTTCTTTTGTAAAAGT
```

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
ELPFGARAKLFAKWKNIIPSVCSYSI*INKGANLTILPL

E L P W G A R A K L F A K W K N I I P S
1 gagtcoccttggggggcaagggcaaaaactttttgctaaatggaaaaatattataccaagt 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

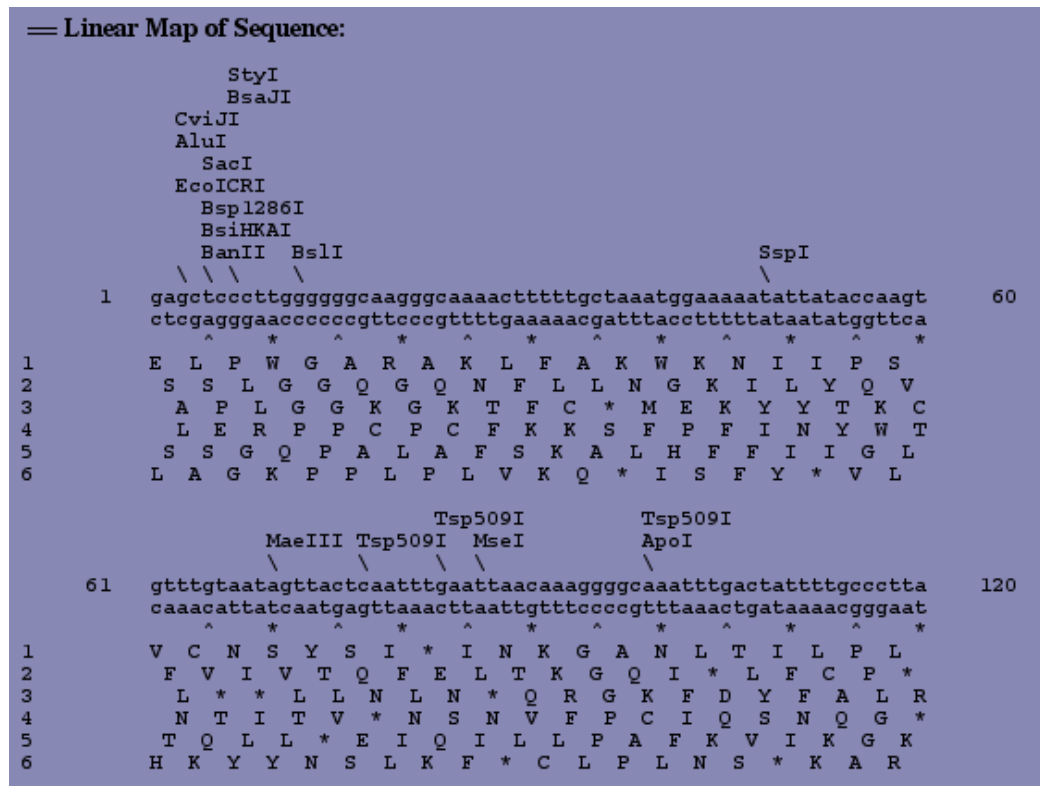
Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. Tran

>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 agtcoccttggggggcaagggcaaaaactttttgctaaatggaaaaatattataccaagt 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/>

Selected Sequence(s)

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

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

```
-----
2580  CTTTGGTTGGTCTGTGCTTGACAACITGGAGTGGAGACTGGGCTAGACTGCCCGTTTGGG  804655

      2650      2660      2670      2680      2690      2700
24  .....                               AATAAGGT  170381
1  .....                               .....  11321163
2430 .....                               CAAAGATT  1702488
1743  CAGTGAATAATCATTGAGAGAACTCGAAAAAGAAAGGAAAAATGGTAAAAAAAATAAATTC  19686
2620  GATGGTCTATGTGGACTTGAATACTGTGAAGAGGTAGGGCAAGGACTGAGGCTTGTGTT  804655

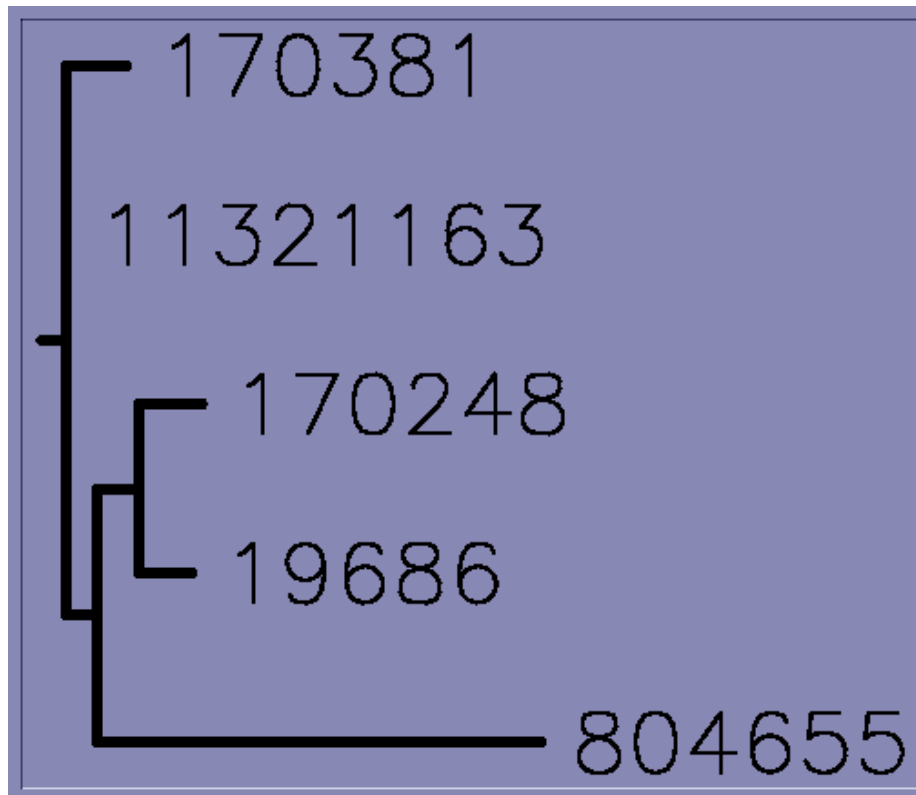
      2710      2720      2730      2740      2750      2760
32  .....  AATAATTCCTCTACGATTACTTGTGGCA.CCAAGATTAGATAC  170381
1  .....  AATAATTCCTCTACGATTACTTGTGGCA.CCAAGATTAGATAC  11321163
2438 .....  AATAATTCCTCTACGATTACTTGTGGCA.CCAAGATTAGATAC  1702488
1803  AGGATGTTTAAATTCCTCTACGATTACTTGTGGCA.CCAAGATTAGATAC  19686
2680  CAAGAAGATGCTCGGAAAGAAGAGCAGGTTGGATGGAAAGAGGATCGGGAGATG  804655

      2770      2780      2790      2800      2810      2820
79  AGCGGTAA...ATAGGTC...TTGTTATGGATATGCGCAAGCACTGGATGAC  170381
1  .....ATAGGTC...TTGTTATGGATATGCGCAAGCACTGGATGAC  11321163
2484  AGCGGTAA...ATAGGTC...TTGTTATGGATATGCGCAAGCACTGGATGAC  1702488
1863  AGCGGTAA...ATAGGTC...TTGTTATGGATATGCGCAAGCACTGGATGAC  19686
2740  AGTGGTCAAGTTGACAAAAAAGATAATGTAATGCTTTTTCATGAAACGGAG  804655

      2830      2840      2850      2860      2870      2880
132  ATTCTAAGTTATAGAGC..TGTACAGTCCAGAAAGATTGAAAGACTGAGGCTTATGA  170381
45  TTGGAAAGTTATAGAGC..TGTACAGTCCAGAAAGATTGAAAGACTGAGGCTTATGA  11321163
2540  TTGGAAAGTTATAGAGC..TGTACAGTCCAGAAAGATTGAAAGACTGAGGCTTATGA  1702488
1919  TTGGAAAGTTATAGAGC..TGTACAGTCCAGAAAGATTGAAAGACTGAGGCTTATGA  19686
2800  ACTTAAAGGCTGTTGTTAAACAGGACCTGGCGAATATGGCAG.CCGCAAAATGGAG  804655
```


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 - NCBI - Mozilla Firefox

http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=65979

Gene

PHACTR4 phosphatase and actin regulator 4 [Homo sapiens]

Gene ID: 65979, updated on 27-Aug-2011

Summary

Official Symbol PHACTR4 provided by HGN

Official Full Name phosphatase and actin regulator 4 provided by HGN

Primary source HGN:25793

Locus tag RP11-442N24_A.1

See related Ensembl:ENS00000204138; 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; Haplorhini; 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

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 a Firefox browser window displaying the OMIM website. The address bar shows 'omim.org/'. The page content includes a navigation menu (Home, About, Statistics, Downloads/API, Help, External Links, Terms of Use, Contact Us), mirror sites (us-east.omim.org, europe.omim.org), and the OMIM logo with the text 'Online Mendelian Inheritance in Man®'. Below the logo, it states 'An Online Catalog of Human Genes and Genetic Disorders' and 'Updated 6 September 2012'. There is a search bar with the text 'Search OMIM' and a 'Search' button. Below the search bar, it says 'Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map'. At the bottom of the page, there are logos for the 'Institute of Genetic Medicine', 'JOHNS HOPKINS MEDICINE', and 'National Human Genome Research Institute'. A footer contains a disclaimer: '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, restričních míst....
 - Další [www genomové nástroje](#)

Diskuse