

CG020 Genomika

Bi7201 Základy genomiky

Přednáška 1

Úvod do bioinformatiky

Jan Hejátko

Funkční genomika a proteomika rostlin,
Mendelovo centrum genomiky a proteomiky rostlin,
Středoevropský technologický institut (CEITEC), Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.muni.cz



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Osnova

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

Schéma předmětu

- **Kapitola 01** (CG020 , Bi7201)
 - Úvod do bioinformatiky

- **Kapitola 02** (CG020 , Bi7201)
 - Identifikace genů

- **Kapitola 03** (CG020 , Bi7201)
 - Přístupy reverzní genetiky

- **Kapitola 04** (CG020 , Bi7201)
 - Přístupy genetiky přímé

Schéma předmětu

- **Kapitola 05** (CG020 , Bi7201)
 - Přístupy funkční genomiky

- **Kapitola 06** (CG020 , Bi7201)
 - Protein-protein interakce a jejich analýza

- **Kapitola 07** (CG020)
 - Moderní postupy funkční genomiky

- **Kapitola 08** (CG020)
 - Strukturní genomika

Schéma předmětu

- **Kapitola 09** (CG020)
 - Lokalizace genů a genových produktů v buňce

- **Kapitola 10** (CG020)
 - Genomika a systémová biologie

- **Kapitola 11** (CG020)
 - Praktické aspekty funkční genomiky

- **Kapitola 12** (CG020)
 - Lokalizace genů a genových produktů v buňce

Literatura

- Zdrojová literatura ke kapitole I:
 - **Bioinformatics and Functional Genomics**, 2009, Jonathan Pevsner, Willey-Blackwell, Hoboken, New Jersey
<http://www.bioinfbook.org/index.php>
 - **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?

- V širším pojetí-zkoumá **STRUKTURU** a **FUNKCI** genomů
 - Předpokladem je znalost genomu (sekvencí)-práce s databázemi
- 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

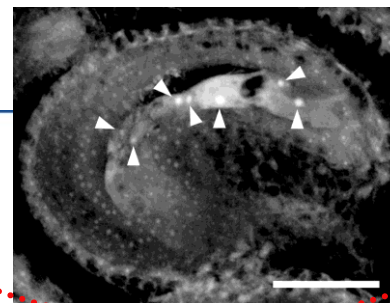
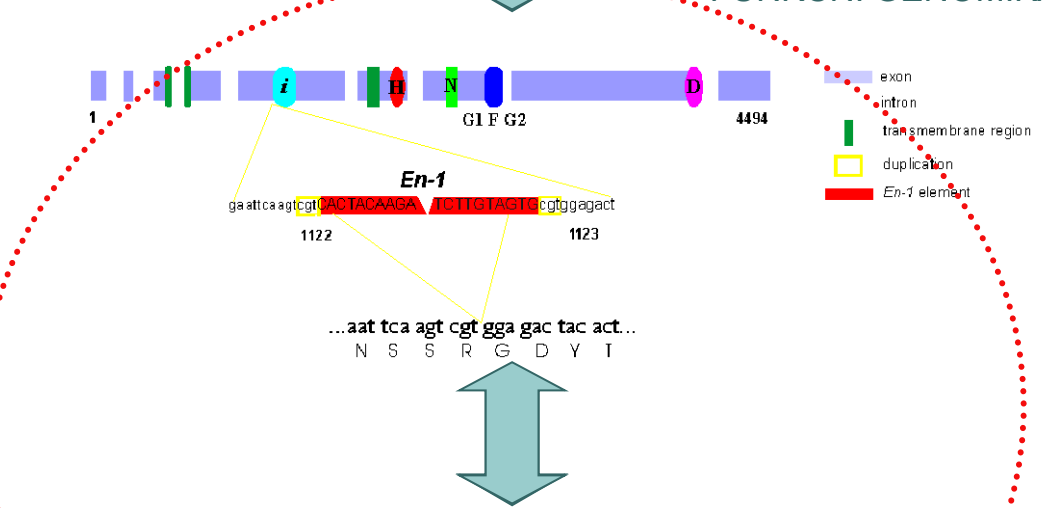
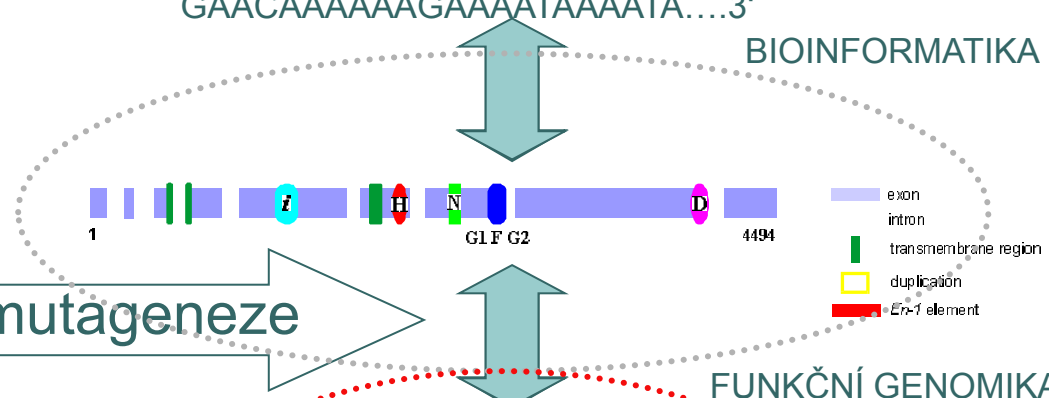
5'TTATATATATATATTAATAAAATAAAATAAAA
GAACAAAAAGAAAATAAAATA...3'



3

:

1



VOJE VZDĚLÁVÁNÍ

entace je spolufinancována
ropským sociálním fondem
a státním rozpočtem České republiky



EVROPSKÁ UNIE



MLÁDEŽE A TĚLOVÝCHOVY

pro konkurenceschopnost

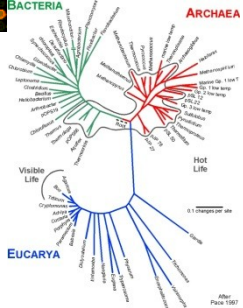
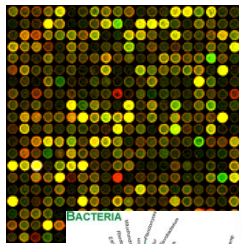


Osnova

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

Bioinformatika

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



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

What is bioinformatics?

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

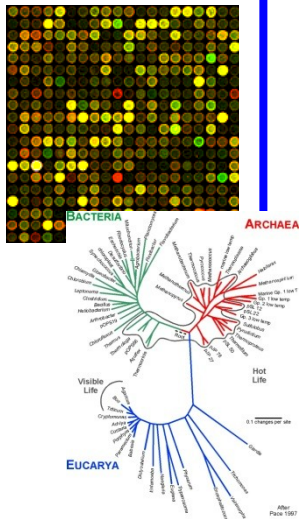
J. Pevsner,
<http://www.bioinfbook.org/index.php>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Bioinformatika



- **Bioinformatika ve funkční genomice**
 - **Zpracování a analýza sekvenačních dat**
 - Identifikace referenčních sekvencí
 - Identifikace genů
 - Identifikace homologů, ortologů a paralogů
 - Korelační analýzy mezi sekvencemi 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.embnet.org/
BEN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFODIOGEN	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/
CDFO	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>

The image shows a screenshot of the EMBL-EBI website interface. At the top, there is a search bar with the text "Nucleotide sequences" and a "Go" button. Below the search bar, the EMBL-EBI logo is displayed, along with the text "European Bioinformatics Institute". A navigation menu includes links for "EBI Home", "About EBI", "Research", "Services", "Toolbox", "Databases", "Downloads", and "Submissions".

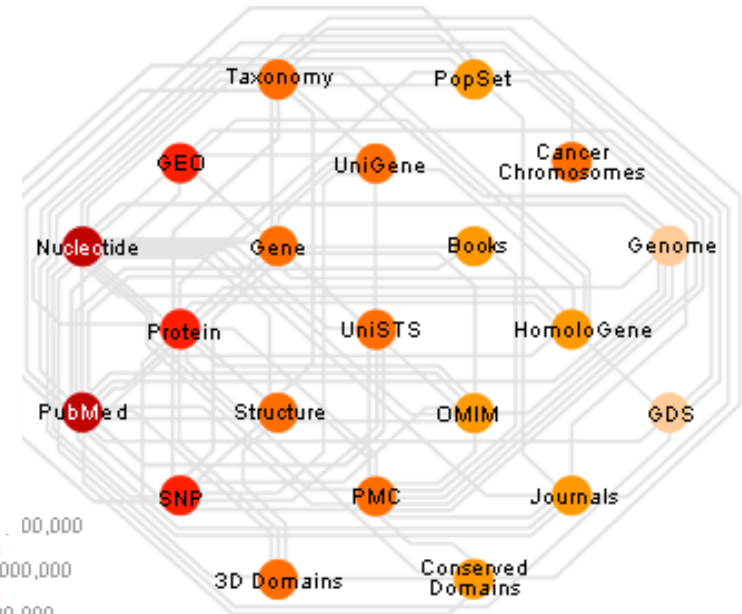
A tree diagram is overlaid on the screenshot, illustrating the structure of the services. The tree is rooted at "Services Overview" and branches into several categories:

- Databases**
 - Database Browsing
 - SRS
 - Nucleotide Databases
 - EMBL Nucleo. Sequence
 - Ensembl
 - Genomes Server
 - Genome MOT
 - EMBL-Align
 - Simple Queries
 - dbSTS Queries
 - Parasites
 - Mutations
 - IMG
 - Protein Databases
 - SWISS-PROT
 - TrEMBL
 - InterPro
 - CluSTR
 - IPI
 - GQA
 - Proteome Analysis
 - HPI
 - IntEnz
- Toolbox**
 - Homology & Similarity
 - Fasta
 - WU-Blast2
 - NCBI-Blast2
 - Blast2_EVEC
 - Genome/Proteome Fasta
 - MPsrch
 - Scanps2.3
 - Parasite-Blast
 - EGI-Blast
 - SNP-Fasta3 Server
 - Prot. Function. Analysis
 - CluSTR Search
 - InterPro Scan
 - FingerPRINTScan
 - ppsearch
 - Gene Quiz
 - Pratt
 - Radar
- Submissions**
 - EMBL via WEBIN
 - EMBL-Info. Submitters
 - SWISS-PROT
 - Webin-Align
 - PDB-AutoDep.
 - MIAMEExpress
 - IMG/HLA
 - Sequin Software
- Downloads**
 - FTP Server
 - Database Repository
 - Software Repository
 - Downloads Help Files
- Services Help**

Spektrum on-line zdrojů

□ NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI website interface. At the top, there's a navigation bar with 'NCBI Resources' and 'How To' menus. Below it is a search bar labeled 'All Databases'. The main content area is divided into several sections: a left-hand navigation menu, a central 'Welcome to NCBI' section with introductory text and links, a 'Get Started' section with bullet points for Tools, Downloads, How-To's, and Submissions, a 'Popular Resources' list on the right, and an 'NCBI YouTube channel' section at the bottom with a video player and a 'GO' button.



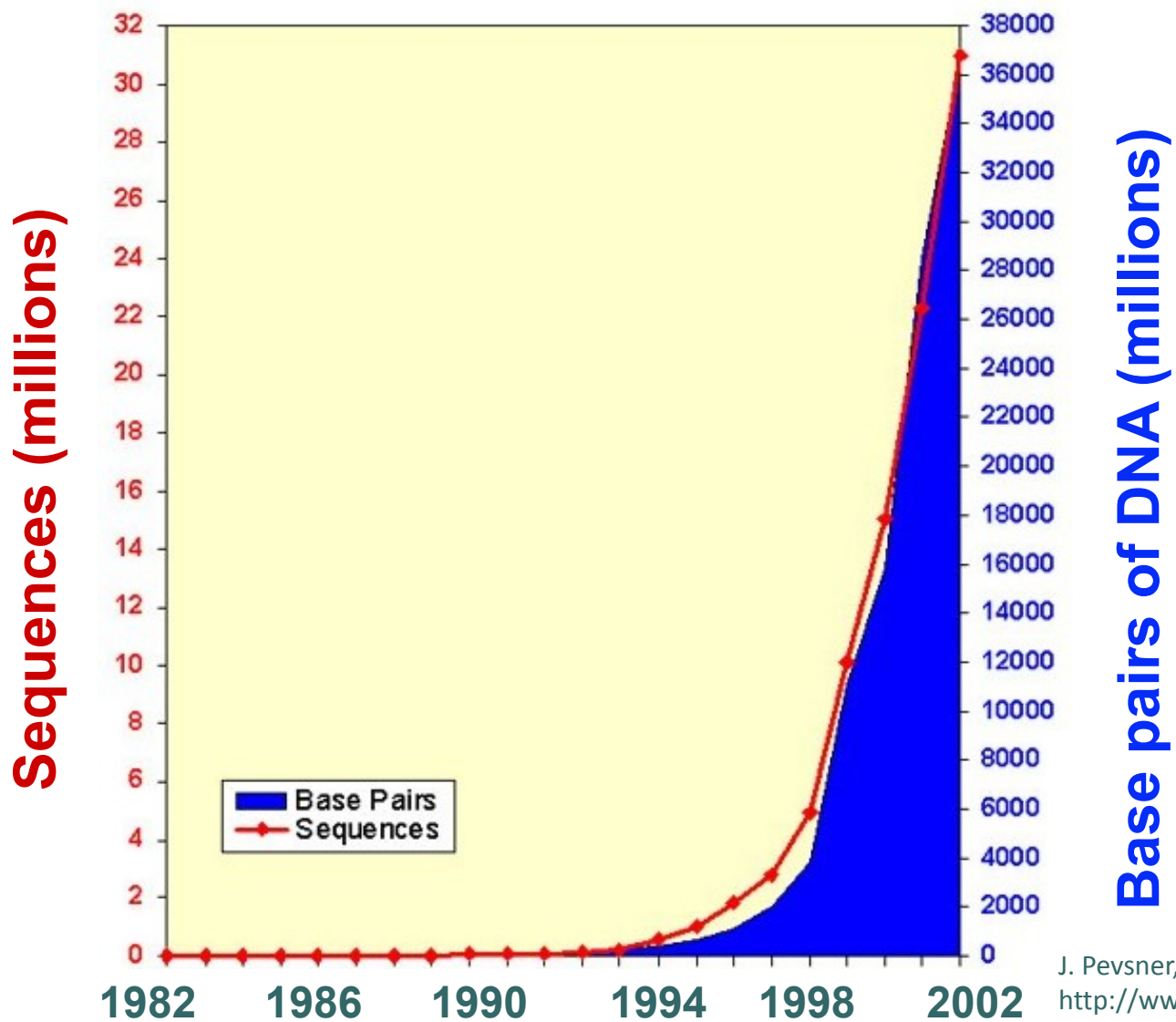
Osnova

- 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, <http://www.ncbi.nih.gov/Genbank/GenbankSearch.html>
 - DDBJ, <http://www.ddbj.nig.ac.jp>
 - denně vzájemná výměna a zálohování dat
 - velká datová náročnost (kapacita i software)
 - září 2003 27,2 x 10⁶ záznamů o zhruba 33 x 10⁹ bp
 - srpen 2005 100 x 10⁹ bp ze 165.000 organizmů

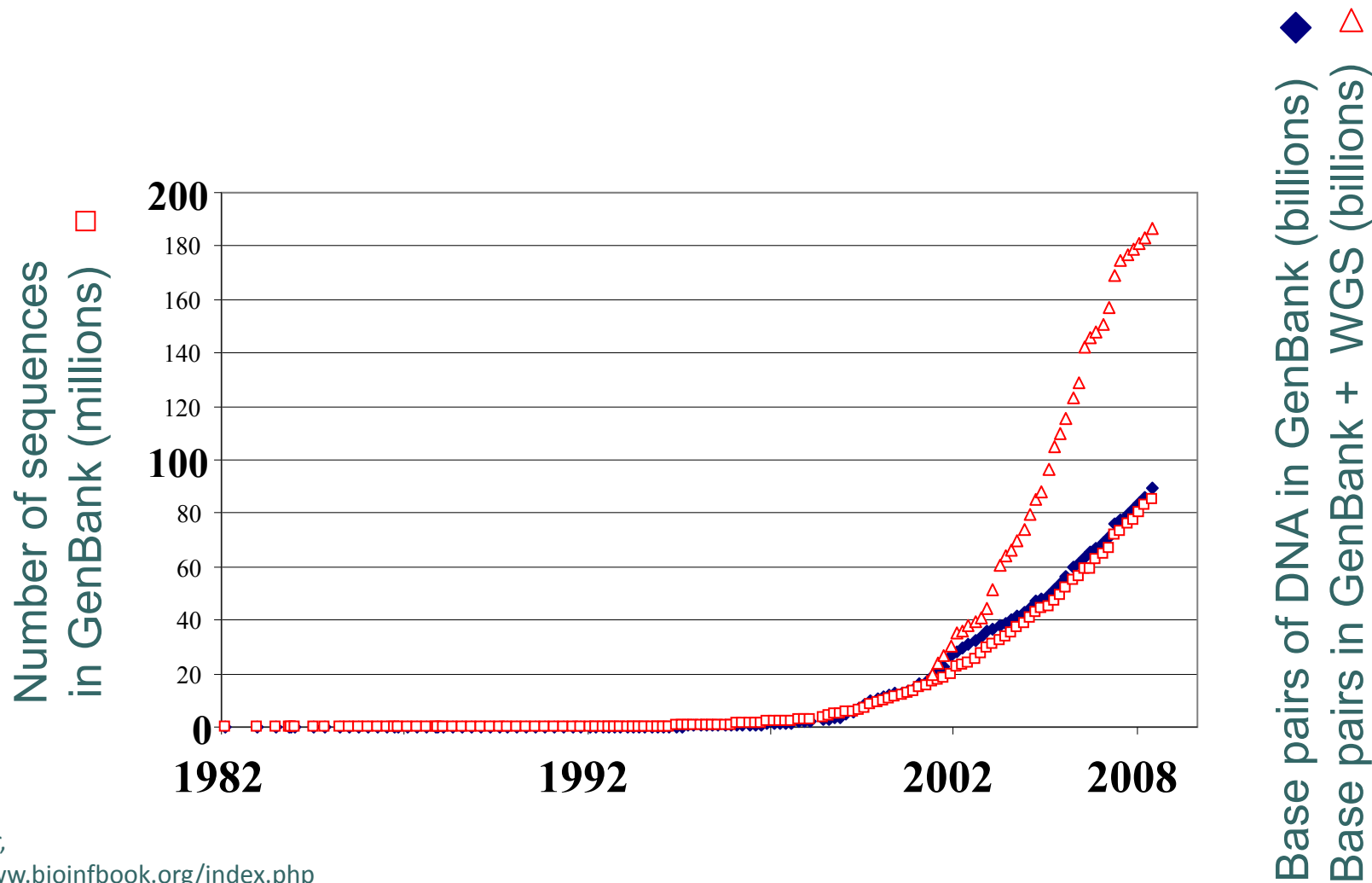
Growth of GenBank



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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



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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

WGS

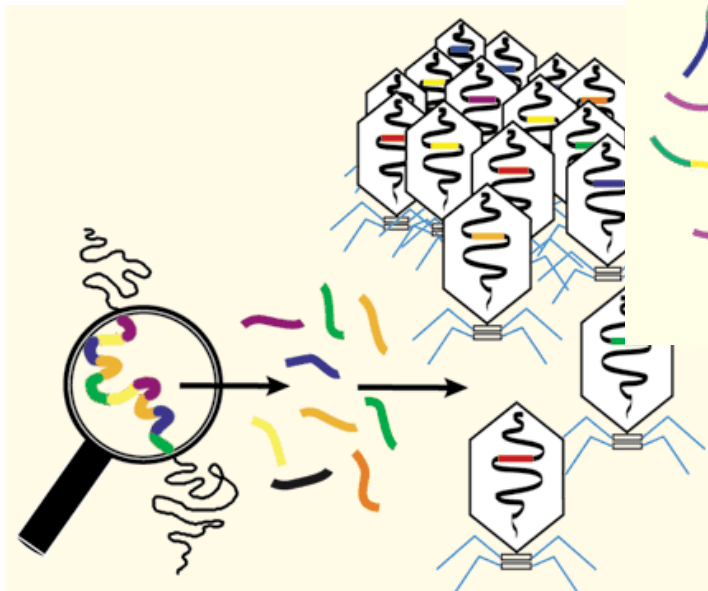
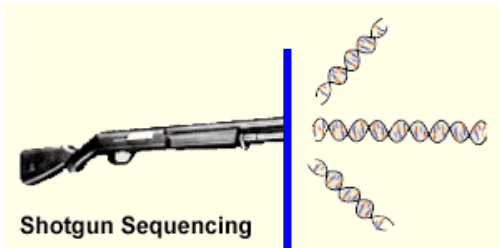


Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

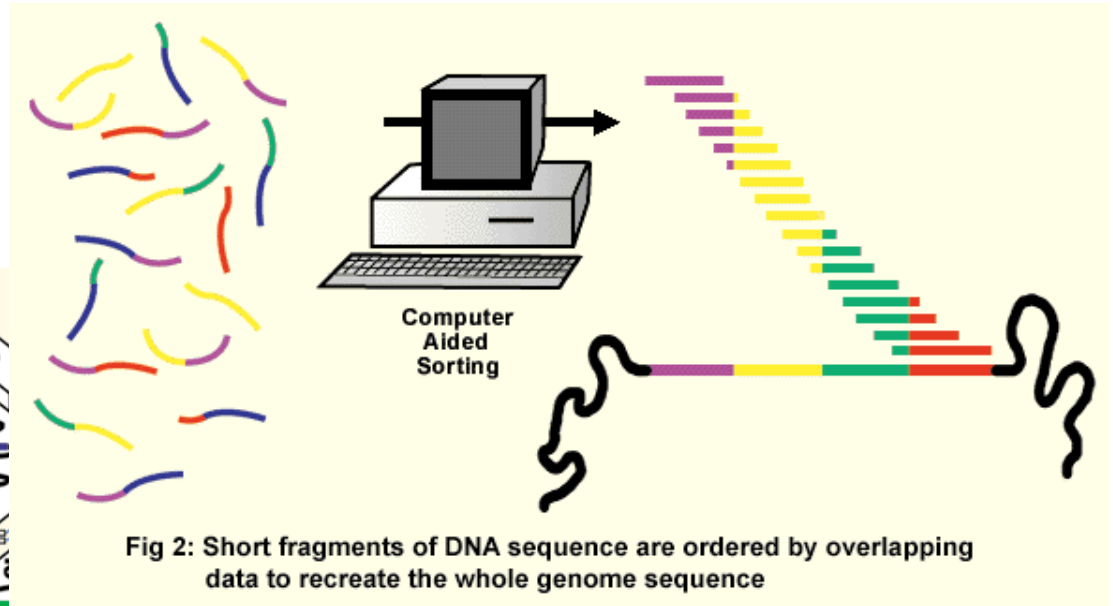
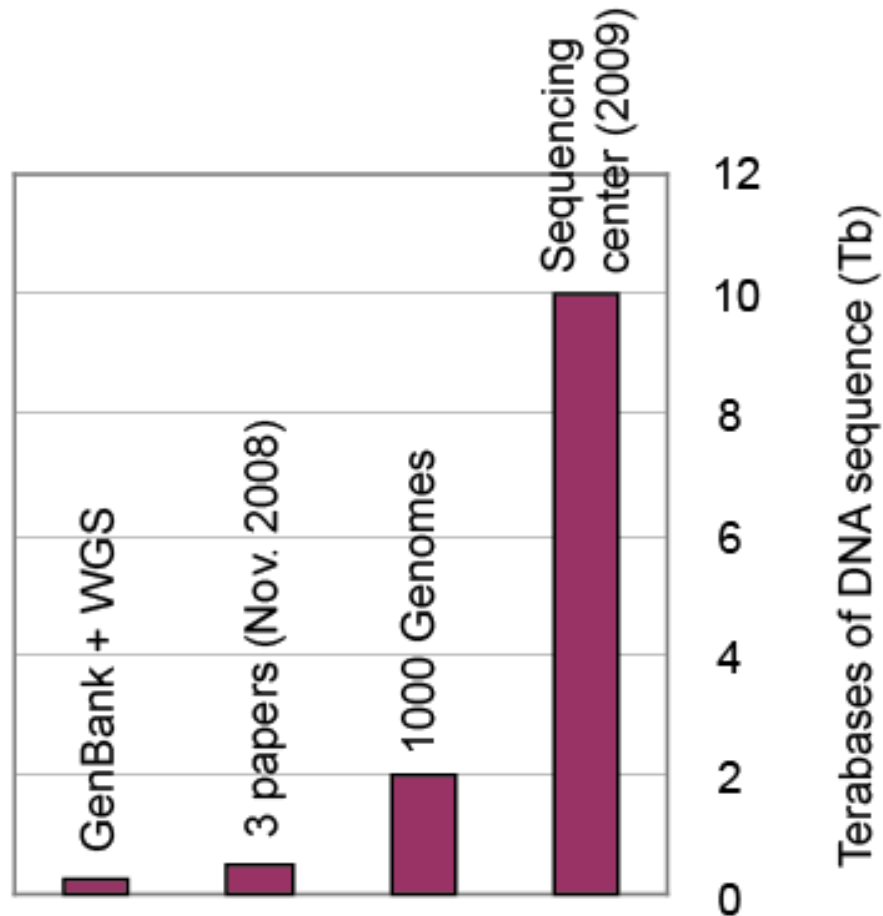


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

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

Arrival of next-generation sequencing: In two years we have gone from 0.2 terabases to 71 terabases (71,000 gigabases) (November 2010)



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

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) <http://www.ncbi.nlm.nih.gov/>



The screenshot shows the NCBI homepage with a navigation menu on the left, a search bar at the top, and a main content area with sections for 'Welcome to NCBI', 'Get Started', and 'Popular Resources'. The search bar contains 'All Databases' and a 'Search' button. The 'Welcome to NCBI' section includes a description and links for 'About the NCBI', 'Mission', 'Organization', 'Research', and 'RSS Feeds'. The 'Get Started' section lists 'Tools', 'Downloads', 'How-To's', and 'Submissions'. The 'Popular Resources' section lists 'PubMed', 'Bookshelf', 'PubMed Central', 'PubMed Health', 'BLAST', 'Nucleotide', 'Genome', 'SNP', 'Gene', 'Protein', and 'PubChem'. The 'NCBI Announcements' section mentions a new version of GenBank and an integrated download for viewing and analysis. The 'NCBI YouTube channel' section features a video player with a 'GO' button and a 'YouTube' logo.

Primární databáze

Gene symbol virA
Gene description two-component VirA-like sensor kinase
Locus tag pTl_125
Gene type protein coding
RefSeq status PROVISIONAL
Organism *Agrobacterium tumefaciens* (old-name: *Agrobacterium tumefaciens*, 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_002377.1

Related articles

1. [Sequence analysis of the virA locus of Agrobacterium tumefaciens octopine Ti plasmid pTl15955](#), 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 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

The screenshot displays a web browser window with a genomic map. The main view shows a genomic region from 145,400 to 147,600 bp. A red bar represents the gene NP_059797.1. A tooltip provides detailed information about this gene:

- 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 view, there are sections for "Bibliography" and "Related articles in PubMed".

Primární databáze

NCBI

Search Nucleotide for [] Go Clear

Preview/Index History

Dist [] 1: **NC_002377.1** [GI:10955016]

LOCUS **NC_002377** 2490 bp DNA linear BCT 29-DEC-2003

DEFINITION *Agrobacterium tumefaciens* extrachrom plasmid Ti, complete sequence.

ACCESSION **NC_002377** REGION: 145694..148183

VERSION NC_002377.1 GI:10955016

KEYWORDS

SOURCE *Agrobacterium tumefaciens* (Rhizobium radiobacter)

ORIGIN

GeneBank Identifier

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 **REPERQ**: This record has not yet been subject to final NCBI review. The reference sequence was derived from [AF242981](#).

FEATURES

Location/Qualifiers

source 1..2490

/organism="Agrobacterium tumefaciens"

/mol_type="genomic DNA"

/db_xref="taxon:359"

/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="MNGRYSPTRODFKTKGAKPWSILALIYAAMIAPFMAVASWQDNMT  
TQAILSQLRSINADSASLQRDVLRAHTCTVANYRPIISRLGALRKNLEDLKLQFRQSH  
IVSEBNAQLLRQLEVSLSADAAVAAPGAQNVRLQDSLASPTRALSSLPKASTDQT  
LEKPTELASMMQLQFLRQSPAISFPISELELELQKQRLDEAFVILAREGPIILSL  
PQVKDLVNMQISDTAEIAEMLQRECLVYSLKNVEERSARIPLSSASVGLCLYIITL  
VYLRKKTDWLARRLDYELIKEIIVCFBGEAATSSAQALRIIQRPFDADTCALAL  
VDHRRWAVETFOAKHFKPVWDSVLRIIVSRTEKADERATVFRILSSKIVHLFLHIP  
GLSILLAHKSTDKLIAVCSLGYQSYRFPFCQGETQLLELATACLCHYIDVRRKQTECD  
VLARRLEHAQRLAVGTLAGGIAHFNNILGSLGHAELAQNSVSRTEVTRRYIDYII  
SSGDRAMLIIDQILTLRKRQEMIKPPSVSELVTEIAFLRMLPFPNIELEFRPDQMC  
SVI EGSPLRLQQLINICKNASQAMTANQIDIIIGQAPLPVKKILAHGVMPFGDYVL  
LSISDNQGGIPRAVLPHIFEPFPTTRARNGTGLGLASVHGHSAPAGYIDVSTVGH  
GTRFDIYLPFSSKFPVNPDSFFGRNKAPEGRNGEIVALVEPDDLREAYRDKIAALGYE  
PVGFRTFNKIRDWISKGNADLVMVDQASLPEDQSPNSVDLVLKTAIIIGNDLKM  
LSREDVTADLYLFPKPISSRTMAHAHLTKIKT"
```

```
ORIGIN  
1 atgaacggaa gatattcacc gaecggcgag gattttaaga caggcgcgaa gcccttggct  
61 atattggccc ttatcgttgc tgcaatgatt ttcoggttca tggcgggttc gtccttggcag  
121 gacaatgcca ctaccacgag aatcctcage caactacgat cgat taacgc cgcacagcgc  
181 tcaactgcaag gogatgtact cecgectcac acgggcaacgc tggcgaacta ccgccccatt  
241 atctccaggg tgggagctct gcggaagaat ctggaagatt tgaagcaatt atttagacaa  
301 tctcatattg taagtgagag caatgtgctc caactgtcac gccagctaga agtgtctcta  
361 aatcgggttg acgcgcggtg cgcgcctttt ggtgcgcaaa atgtacgctc gcaagattcg  
421 ctggcagctt tcaactcgtg tttgagcagt ctccagggaa aagcctcaac cgatcagact  
481 ttgaaaaaac caacagaatt ggttagcagt atgctccaat tctctggca accaaagccg  
541 gctatttcat togagatcag ccttgaacta gagaggctcc aaaaacaagc cggcttggat  
601 gaagctcccg tgcgcaact tgcacgtgaa ggtcccatta tcttatcgtc tttgccacag  
661 gtgaaagatc tgggtaacat gattcagcgc tctgacacgc cagaatctgc gggagatgctg  
721 cagcgcgagt gtttggaggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc  
781 ttcttgggtt cccctcagat gggctcttgc ctctacatca tcaccttagt ctataggcta  
841 cgcacacaaa cagatgggtt agcgcggcgt ttgatatacg aagagctaat caaagagatc  
901 ggagtagttt ttgaaagtga ggcggccacc acgtcgtccg cgcacagctc actctgtatt  
961 atcagcgcct tcttggatgc cgtatcgtgc cgtttagctc tagtggacca tgacgttaga  
1021 tgggctgtcg aaacattcgg tgcgaaaac ccaaacctgt tctgggacga cagcgtgcta  
1081 cgcgaaatag tctctcgtac caaagcggac gaacggggca cggatctcgc catcatatcg  
1141 tgcacacaaa togtacattt gccctcogaa atccagctc tctcgatact actggctcac  
1201 aaatccacag ataaactaat tgcggtttgt tcaactgggtt accaaagcta tgcctctoga  
1261 ccttgcacag gcaaaatca gctcttggaa ctgcacacgc cctgcctctg tcatatatac  
1321 gatgttcggc gtaagcagac cgaatgagac gttttggcca gacgatgga gcatgcccga  
1381 cgccttgagg cagttggtac acttgcgggc ggaatgacac atgaatttaa taacattttg  
1441 ggctcaatcc tgggcaacgc agaattagca caaacctcgg tctctogaac atctgtcac  
1501 cgaagatata ttgactatat cattctgtca ggcgacagag ccatgctcat tctgatcag  
1561 atcttgacgc tgcgcgaaa acagggagcg atgatcaagc catttagtgt ctcagagctt  
1621 gtgaccgaaa toctcctctt gctacgtatg gctctccgc caaacatoga gcttagtttc  
1681 agatttgac aaatcgagag cgtgatcgaa ggaagccgcg ttgaacttca acaggtacta  
1741 attaacatct gcaagaatgc tcccaagcc atgactgcaa atggtcaaat cgcacatcct  
1801 atcagccaag cttttttacc agttaagaaa attctggcgc atggtgttat gccocctggc  
1861 gactatgttc tctatctat tagcgaacat ggtggaggca tcccgagcgc tgtgttacc  
1921 cacatttttg aacctctctt tacgacacga gctgcacacg gtggaaacgg tctcggcctt  
1981 gcttctgtgc atggtcatat cagcgcgctt gcgggttaca togacgttag ttoactgtt  
2041 gggcatggga cgcgcttga catttatctc cctccgtctt ctaaagaaec cgtaaatcna  
2101 gacagttttt bccggccgaa taaggccacc cgtggaacgc gggagattgt ggcactgtt  
2161 gagccgatg acctcctgag gtaggcgtat gaagacaaga tcccgctctc aggatagag  
2221 cggctcgggt ttctgacctt taatgaattt cgcgatggga tttcaaaagg caatgaagc  
2281 gatctggtca tggctgacaa agcgtctctt cctgaagatc aaagtctaa tccctggat  
2341 ttagtgtca agacgcctc catcatcatt ggcggaatg atctcaaat gacccttca
```


What is an accession number?

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

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

X02775	GenBank genomic DNA sequence
NT_030059	Genomic contig
Rs7079946	dbSNP (single nucleotide polymorphism)

DNA

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

RNA

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

protein

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

NCBI's important RefSeq project: best representative sequences

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

RefSeq identifiers include the following formats:

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

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

RefSeq

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. The main content area is titled 'two-component VirA-like sensor kinase' and includes a 'Genome Annotation' section. Below this, there is a 'Reference assembly' section with a 'Genomic' subsection. The first entry is 'NC_003065.3' with a range of 180831..183332 and download links for GenBank, FASTA, and Sequence Viewer. The 'mRNA and Protein(s)' section lists 'NP_396486.1 two component sensor kinase [Agrobacterium tumefaciens str. C58]'. This entry includes UniProtKB/Swiss-Prot ID P18540 and three conserved domains: cd00075 (HATPase_c), cd00082 (HisKA), and PRK13837 (two-component VirA-like sensor kinase). The PRK13837 domain has a location of 14-833 and a Blast Score of 2944. The 'Related Sequences' section is partially visible at the bottom.

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

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

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze

NC_002377.1: 145K..148K (2.9Kbp)

Genes

NP_059797.1

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

Links & Tools

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

Bibliography

Related articles in PubMed

Primární databáze

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
ATGAACGGGAAGATATTCACCGACGCGGCAGGATTTAAGACAGCGCGGAAGCCTTGGCTATATTTGGCCC
TTATCGTTGCTGCAATGATTTTCGCGTTTCATGGCGTTTGGCTCCTGGCAGGACAAATGCGACTACCCAGGC
AATCCTCAGCCAACATCAGATCGATTAACGCCGACAGCGCCTCACTGCAGCGCGATGACTCCGCGCTCAC
ACGGGACCGTGGCGAATACCGCCCATTTATCTCCAGGCTGGGAGCTCTGCGGAAGAAATCTGAAAGATT
TGAAGCAATTTAGCAATCTCATATTTGAAGTGAAGCAATGCTGCTCACTGCTACGCGCAGCTAGA
AGTGTCTTAAATTCGGCTGACGCGCGCTCGCCGCTTTGGTGGCAAAATGTACGCTCAAAATTCG
CTGGCCAGTTTCACTCGTGTCTTGGAGCTTCCAGGAAAAGCCTCAACCGATCAGACTTTAGAAAAAC
CAACAGAATTTGGTAGCATGATGCTCCAATTTCTTCGGCAACCAAGCCCGCTATTTCACTCGAGATCAG
CCTTGAATAGAGAGGCTCCAAAAACAACGCGCTTGTATGAAGCTCCCGTGGCAGACTTGCACGTGAA
GGTCCCATTTATCGCTTTTGGCCAGGTAAGGATCTGGTGAACATGATTCAGAGCTCTGACACCG
CAGAAATTCGGAGATGCTGCAGCGGAGTGTGGAGGCTATAGCTTGAATAATGTAGAGGACGGAG
CGCAGCTATCTTTCTGGTCCGCTTCAGTGGTCTTTGCCTCTACATCATCACTTAGTCTATAGGCTA
CGCAAAAAACCGATTTGGTGTAGCGCGGCTTTAGATTAGAAAGAGCTAATCAAGAGATCGGAGTATGTT
TTGAAGTGAAGCGGCCACCACTGCTCGCGCAAGCTGCATTTGATTTATTCAGCGCTTTTGGATGC
CGATACGTCGCGCTTAGCTTAGTGGACCATGACCGTAGTGGGCTGTGAAAATTTGCTGCGAAACAC
CCAAAACCTGTGTGGGACGACAGCGTGTACGCGAAATAGTCTCTGTAACAAAGCGGACGAAACGGCGCA
CGGATTCGCGATCATGCTCGAAAAAATCGTACATTTGCTCTCGAAATTCAGGCTCTCTGATACT
ACTGGCTCAAAATCCACAGATAAATTAATGCGCTTTGCTCACTGGGTACCAAGCTTACGCGCTCGA
CCTTGCCAGGGGAAATTCAGCTTCTTGAATCGCCACCGCTGCTCTGACTATATCGATGTTGGCG
GTAAGCAGACCGAATGCGACGTTTGGCCAGACGATTGGAGCATGCGCAACGCTTGAGGCGATTTGGTAC
ACTTCCGCGCGAATAGCACATGAATTAATAACATTTTGGGCTCAATCTCGGGCAGCAGAAATAGCA
CAAACTCGGTCTCGAACATCTGTACCCGAAAGATATATGACTATATCATTTCTGTCAGGCGACAGAG
CCATGCTCATATCGATCAGATCTTACGCTGAGCGGAAACAGGAGCGCATGATCAAGCCATTTAGTGT
CTCAGACTGTGACCGAAATCGCTCCCTTGTACGATGGCTTCCGCCAAACATCGAGCTTAGTTC
AGATTTGATCAAAATGCGAGCGTGTGCAAGGAAAGCCCGCTTGAACCTCAACAGGTACTAATTAACAT
GCAAGATGCTTCCAAAGCCATGACTGCAAAATGTTCAAAATCGACATCATCAGCCAAAGCTTTTTTACC
AGTTAAGAAAATTCGGCGCATGGTGTATGCCACCTGGCGACTATGTTCTCCTATCTATAGCGACAAT
GGTGGAGGATTCGGAGGCTGTGTACCCACATTTTGAACCTTCTTACGACAGAGCTCGCAACG
GTGAAACGGGCTCGGCTGCTGTGTGATGTTGATCATGACGCGCTTTCGGGTTTACATCGAGCTTAG
TTCAACTGTGGGATGGGACCGCTTTGACATTTATCTCCCTCCGCTTCTAAGGAACCCGTAATCCA
GACAGTTTTTTCGGCCGCAATAAGGCAACCGCGTGGAAACGGGAGATTGTGGCACTTTGTAGCCCGATG
ACCTCCTCGGGGAGGCGTATGAAACAAGATCGCCCTCTAGGATATGAGCGGTCGGTTTTTCTGATCCCT
TAATGAAATTCGCGATTGGATTTCAAAAGGCAATGAAGCCGATCTGGTCATGTTGACCAAGCGCTCTCT
CCTGAAGATCAAACTCCTAATTCCTGATTTAGTGTCAAGACCGCTCCATCATATTTGGCGAAATG
ATCTCAAAATGACCCCTTCAAGGGAGGATGTGACCGGACCTTTATCTCCGAGCGGATATCGTCCAG
AACTATGGCGCATGCAATCTAACAAAATCAAGACGATG
```



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Sekundární databáze

- databáze funkčních nebo strukturálních *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 [PRATI](#) can be used to generate your own patterns. You may either:

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

Scan a protein for PROSITE matches	Search Swiss-Prot with a PROSITE entry
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: <input type="text" value="MNVKVTLYASEPTVFPVCLAPLVVPECTWISNMTTTE"/> <input type="text" value="NLVKEVASFTDRLTSLVGERIENIKGPTVAKTHLSTGLA"/> <input type="text" value="RVIDSYTNNDTGPTFRIQTQIAFLPLFVAVSTILQVQSVY"/> <input type="text" value="ISRDGIMPFSYIARSNTSVAVAFANSSNSRQDWTYTQTV"/> <input type="text" value="DQLTGRLRHNGHSTRSQSLDVHTDWFQAQASHNYTTAPVGT"/> <input type="text" value="SLGSEMETLIQSVLSYERGLVSLGFPFKTLTEVNL"/> <input type="text" value="NLHRELIMWTEDVLYRELSLNDFFISGSIQFEE"/> <input type="text" value="NSLMSQCIPENCSSGVEVEIKRLRYQAQPCSVIRVSGVPL"/> <input type="button" value="Clear"/>	Enter a PROSITE accession number (for example PS01253), or type your pattern in PROSITE format : <input type="text"/> (leave this box blank to scan a sequence with the entire PROSITE database)
and specify which motifs to use: 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) <input type="checkbox"/> Exclude patterns with a high probability of occurrence	and specify your search limits: <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 NCBI 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"/> matches
Your e-mail (optional): <input type="text"/> (will send results by e-mail) <input type="checkbox"/> plain text output <input type="button" value="START THE SCAN"/> <input type="button" value="RESET"/>	Advanced options: <input type="checkbox"/> FASTA output <input type="checkbox"/> retrieve complete sequences allow at most <input type="text"/> 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)

Sekundární databáze

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

>[PDOC0003 PS00003](#) SULFATION Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].

```
571 - 585 nkeesstYeteisns
```

>[PDOC0004 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
```

>[PDOC0005 PS00005](#) PKC_PHOSPHO_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

```
148 - 150 S#R  
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 S#R  
854 - 856 ScK  
864 - 866 StR  
868 - 870 SeR  
921 - 923 SpK  
957 - 959 SvR  
960 - 962 TgR  
974 - 976 TeK  
997 - 999 SrK  
1002 - 1004 TgK  
1018 - 1020 SgK  
1031 - 1033 TgR  
1119 - 1121 SkR
```

Sekundární databáze

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

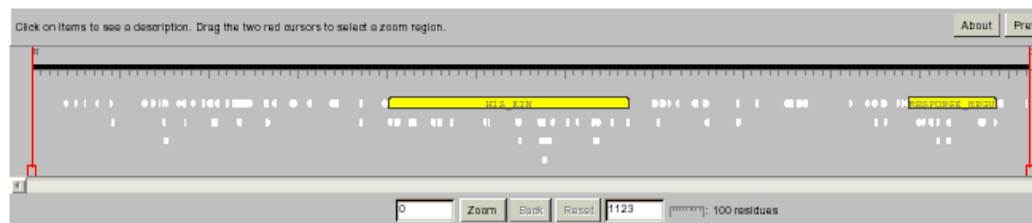
>[PDOC50109 PS50109 HIS_KIN](#) Histidine kinase domain [profile].

```
402 - 671 NASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVMVCAKDLVALLNSVLEMSKIESG
KMQLVREDFNLSKLLLEDVIDFYHFMKKGVDVLDPHDgveEKPSNVRGDSGRLLKQILN
NLVSNVAVRFTVD--GHIAVRAWAQrpgensavvlasyppgvkfvkkmfckkkesatye
teienairnnaTMEFVFEVDDTGKGIHMEMRKSVPBNYVQVREtAQGHQGTGLGLGIVQ
SLVRLMGGEIRITDKAMGeKGTCPQPNVLLTT
```

>[PDOC50110 PS50110 RESPONSE_REGULATORY](#) Response regulatory domain [profile].

```
987 - 1085 RVLVDDNPISRKVTGKLNKMGVSeVEQCDSGKEALRLVTEGLtqreeggsvdklpPDY
IFMDQMPEMDGYRATREIRkvekSYGVRTPIITAVSGHD-----
```

Graphical summary of hits (*java applet*)



98 hits with 12 PROSITE entries

Expasy Home page	Site Map	Search Expasy	Contact us	Swiss-Prot	PROSITE	Proteomics tools
----------------------------------	--------------------------	-------------------------------	----------------------------	----------------------------	-------------------------	----------------------------------

Sekundární databáze

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



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

New:

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

Direct PRINTS access:

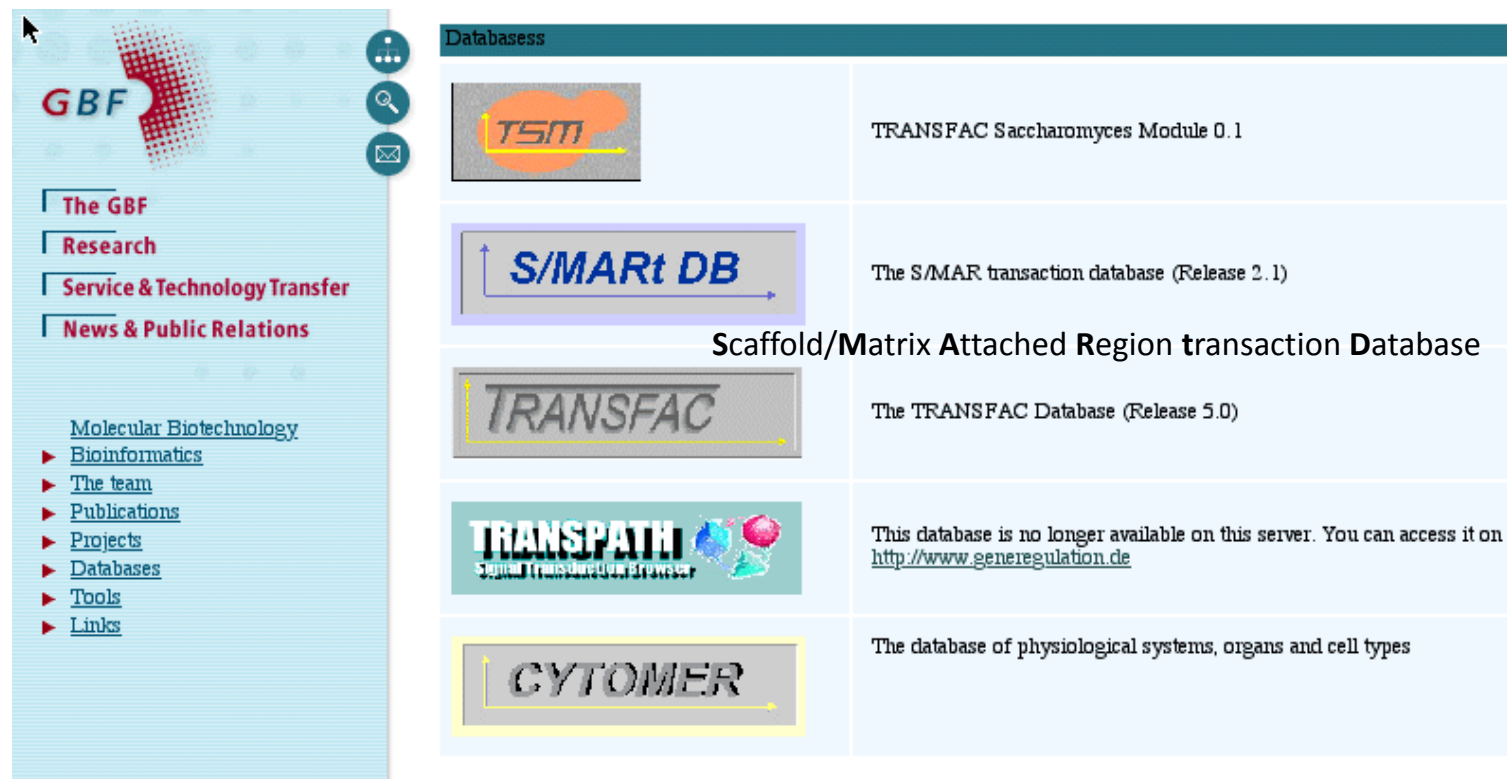
- [By accession number](#)
- [By PRINTS code](#)
- [By database code](#)
- [By text](#)
- [By sequence](#)
- [By title](#)
- [By number of motifs](#)
- [By author](#)
- [By query language](#)

PRINTS search:





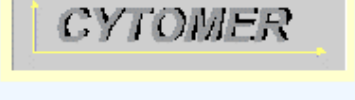
- [Search PRINTS with NEW FingerPRINTScan](#)
- [FPScan](#)
- [GRAPHScan](#)
- [MULScan](#)
- FingerPRINTScan binaries and source are available: contact.scordis@bioinf.man.ac.uk

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 entries.

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

PDB

PROTEIN DATA BANK



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](#)

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](#) [pdb4 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



Structure Explorer - 1P5Y

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



[View Structure](#)

[Summary Information](#)

[View Structure](#)

[Download/Display File](#)

[Structural Neighbors](#)

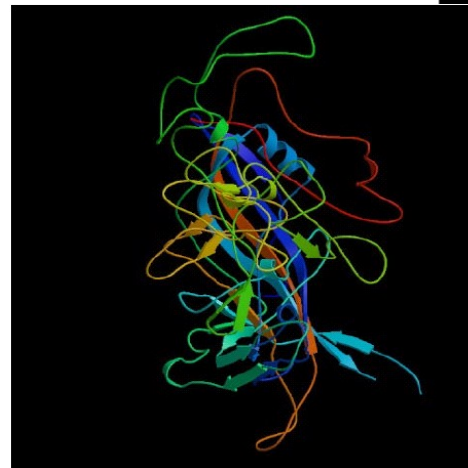
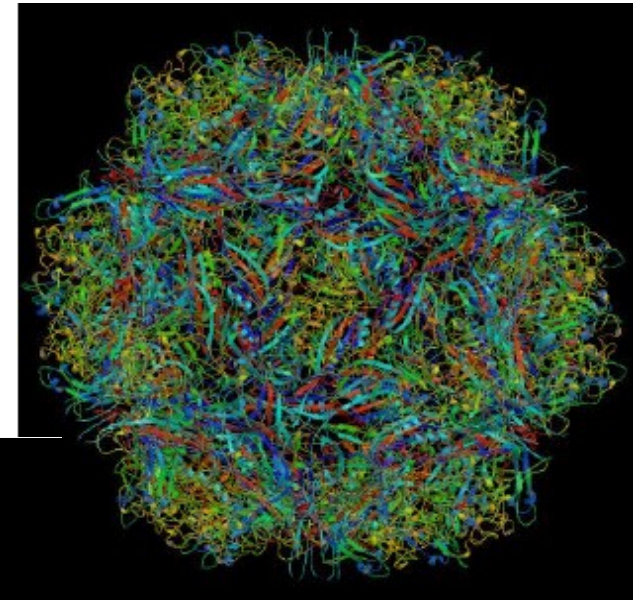
[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

Explore

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

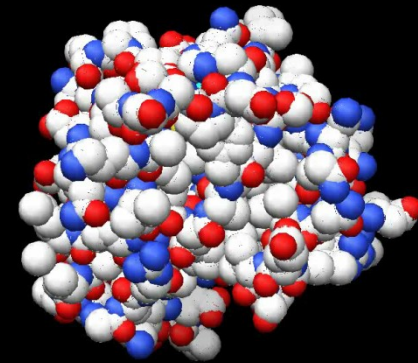


<http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=:pid=173561064349344&bio=1&opt=show&size=500>

12/29/2003

Strukturální databáze

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



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

Osnova

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

Genomové zdroje

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

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

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

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

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

[Add your own custom tracks](#)

Human Genome Browser - hg19 assembly (sequences)

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

Sample position queries

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

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gI000212	Displays all of the unplaced contig gi000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061,RH80175 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

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

Genomové zdroje

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

The screenshot displays the UCSC Genome Browser interface. The main content area is divided into three sections:

- UCSC Genes:** Lists genes such as HBB, HBD, HBE1, HBE2, SATB1, SATB2, KLF1, KLF11, GATA1, and MAFG, along with their genomic coordinates and descriptions.
- Non-Human RefSeq Genes:** Lists orthologous genes from other species, including LOC100174873, LOC100190885, LOC100136584, LOC100135791, LOC100135791, LOC100135791, and LOC573653.
- Human Aligned mRNA Search Results:** Lists mRNA sequences like A18171, AF349114, M25079, M12050, Y00497, Y00500, M14574, M11428, and AY195861.
- Human Unaligned mRNA Search Results:** Shows results for a deletion junction (S77349) and other unaligned sequences.
- Non-Human Aligned mRNA Search Results:** Lists orthologous mRNA sequences from other species like M19548, H0843793, H0843792, AY383035, AY775302, and AY329629.

The browser window title is "UCSC Genes" and the address bar shows the URL "genome.ucsc.edu/cgi-bin/hgTracks/hgHubConnect.destUrl=...". The taskbar at the bottom shows various open applications and system icons.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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

The screenshot shows the UCSC Genome Browser interface for Human Feb. 2009 (GRCh37/hg19) Assembly. The main track displays genomic data for chromosome 11 (chr11) at position 5,246,696-5,248,301 (1,606 bp). The tracks shown include RefSeq Genes, Human RefSeq, Spliced ESTs, H3K27Ac marks, DNase clusters, transcription factor ChIP-seq data, and various other genomic features. A green arrow points to the RefSeq Genes track. The interface includes navigation controls, track search, and expand/collapse options.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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

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 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>

The screenshot shows the UCSC Genome Browser interface. The browser window title is 'Genomic Sequence Near Gene'. The page content includes a navigation bar with links like 'Genomes', 'Genome Browser', 'Tools', 'Mirrors', 'Downloads', 'My Data', 'About Us', and 'Help'. The main heading is 'Genomic Sequence Near Gene'. Below this is a section titled 'Get Genomic Sequence Near Gene' with a note: '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.' The 'Sequence Retrieval Region Options:' section contains several checked checkboxes: 'Promoter/Upstream by 1000 bases', '5' UTR Exons', 'CDS Exons', '3' UTR Exons', and 'Introns'. There are also input fields for 'Downstream by 1000 bases', 'One FASTA record per gene.', and 'One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra downstream (3')'. A 'Split UTR and CDS parts of an exon into separate FASTA records' checkbox is unchecked. The 'Sequence Formatting Options:' section has radio buttons for 'Exons in upper case, everything else in lower case.', 'CDS in upper case, UTR in lower case.', 'All upper case.', and 'All lower case.'. There are also radio buttons for 'Mask repeats: to lower case' and 'to N'. A 'submit' button is located at the bottom of this section. The browser's address bar shows the URL 'genome.ucsc.edu/cgi-bin/hgGateway?hgId=296557229&g=htcGeneInGenome&u=uc001mae.1&c=chr11&l=5246695&r=5248301&o=knownGene&t=knownGene'. The taskbar at the bottom shows various open applications like 'Dělní', 'Přidchozí', 'Zvýraznit', 'Božislav velikost', 'www pages', 'Windows M...', 'Genomic Se...', 'Kalendář - O...', 'Doručená po...', 'EndNote X4...', 'CG020_2012_...', '2010-11-15_...', 'Adobe Acro...', and 'Adobe Phot...'. The system tray shows the date '18:43' and other icons.

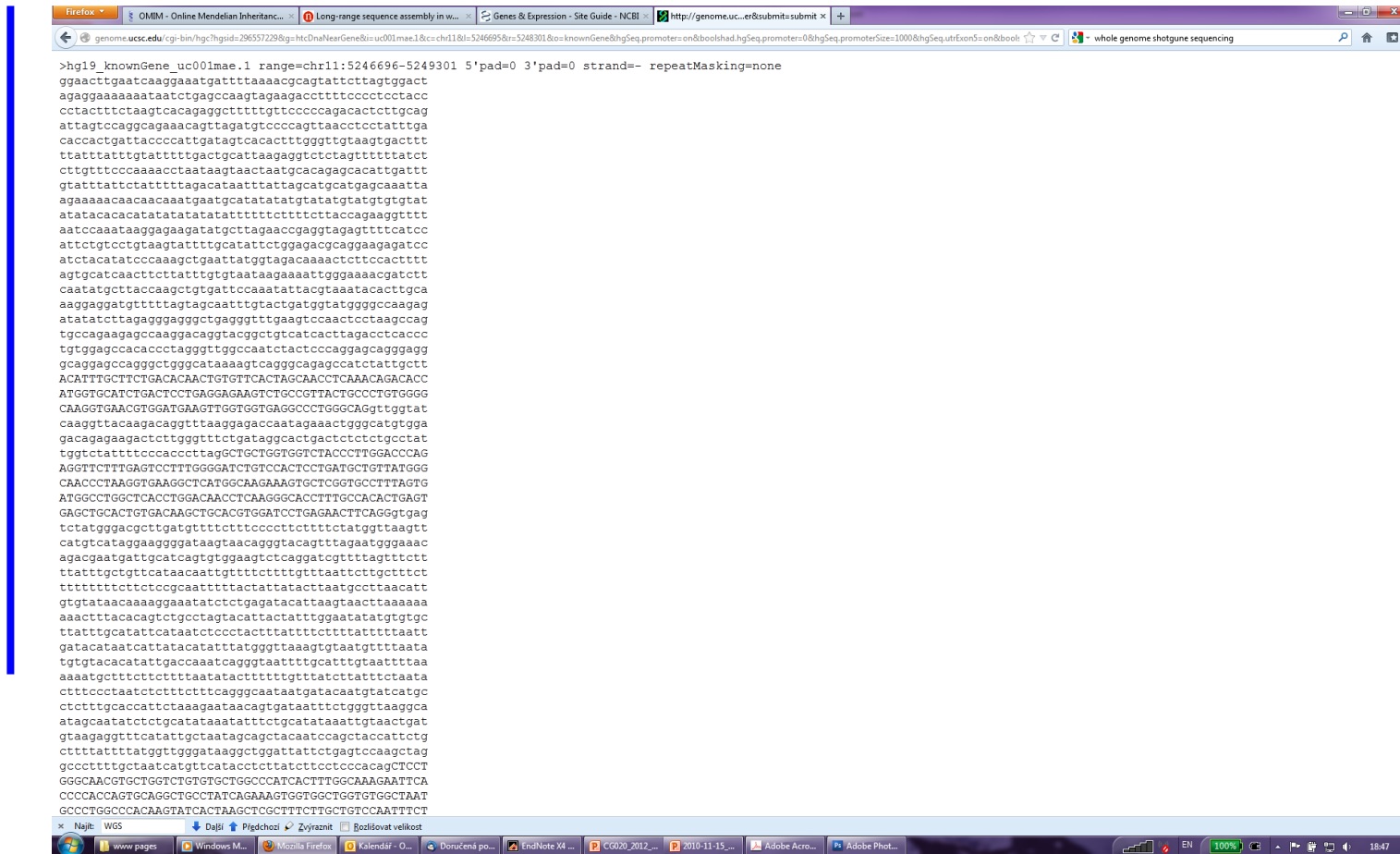


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

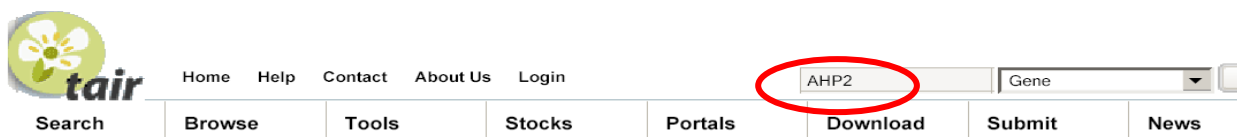
Genomové zdroje

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

The screenshot shows the TAIR website homepage. The browser window title is "TAIR - Home Page". The website features a search bar at the top right and a navigation menu with options like Home, Help, Contact, About Us, and Login/Register. Below the navigation, there are tabs for Search, Browse, Tools, Portals, Download, Submit, News, and ABRC Stocks. The main content area is titled "The Arabidopsis Information Resource" and contains several sections: "Breaking News" with links to subscribe to a news feed, follow on Twitter, and join a Facebook group; "2012 MASC Report Now Available" dated July 11, 2012; "New Protein Chip and Cell Cultures at ABRC" dated May 9, 2012; "Share Your Education Resources" dated February 1, 2012; and "GO Annotations At TAIR" dated January 25, 2012. A central banner promotes a new online submission form, with a "Click here" button and text describing the form's purpose: "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". The banner also includes a "SUBMIT PAPER" button and a "SUBMIT DATA" button. The website footer includes the Carnegie Institution for Science logo and text: "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."

Genomové zdroje

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



The Arabidopsis Information Resource

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

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

Breaking News

Data Updates Suspended

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

New Phenotype Search Option

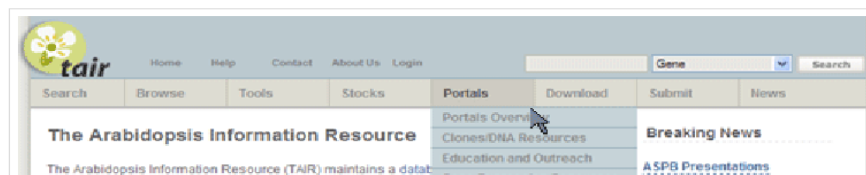
[October 15, 2006]
Search for [genes](#), [germplasms](#), and [polymorphisms](#) using associated phenotype, and see improved phenotype data display in results and detail pages.

ASPB Presentations

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

The NEW arabidopsis.org

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

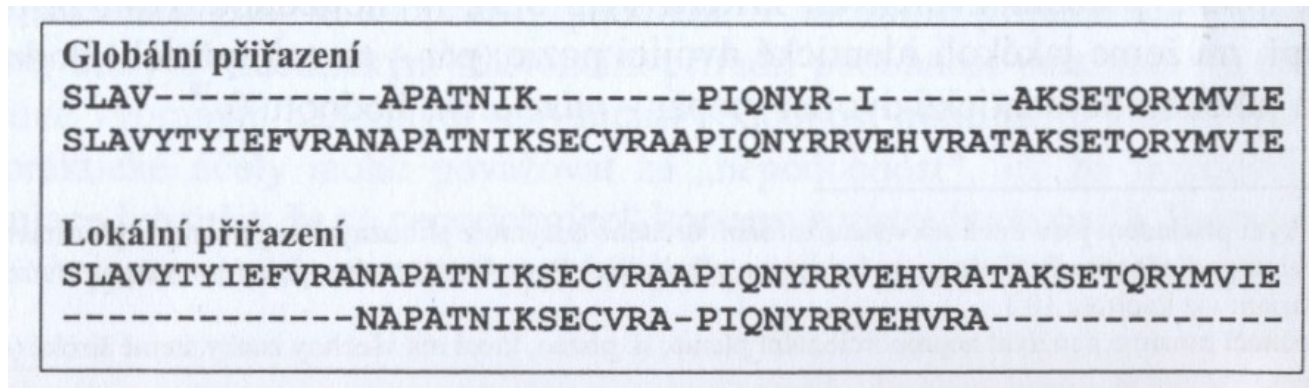


Osnova

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

Analytické nástroje

□ Globální vs. lokální přiřazení

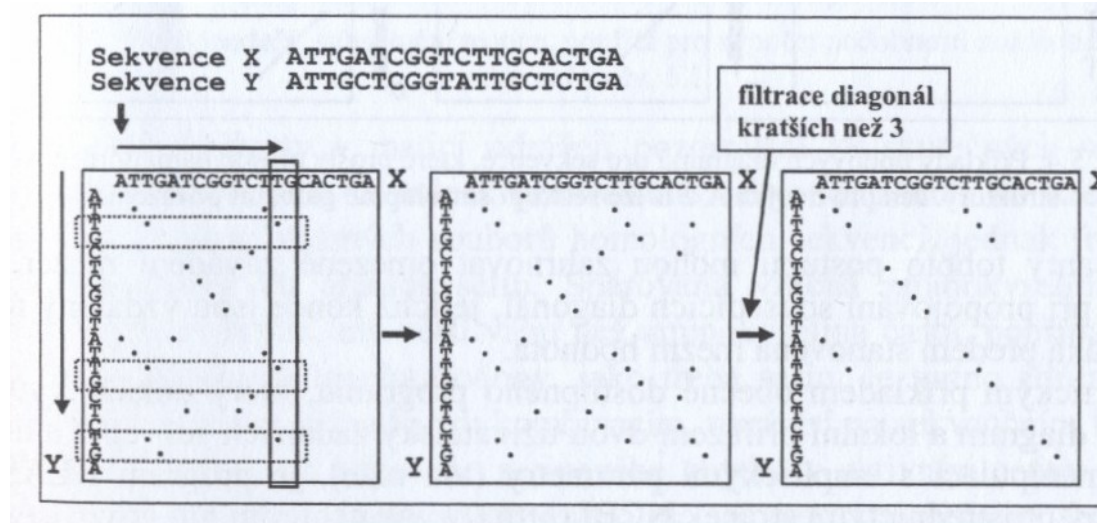


Cvrčková, Úvod do praktické bioinformatiky

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

Analytické nástroje

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)

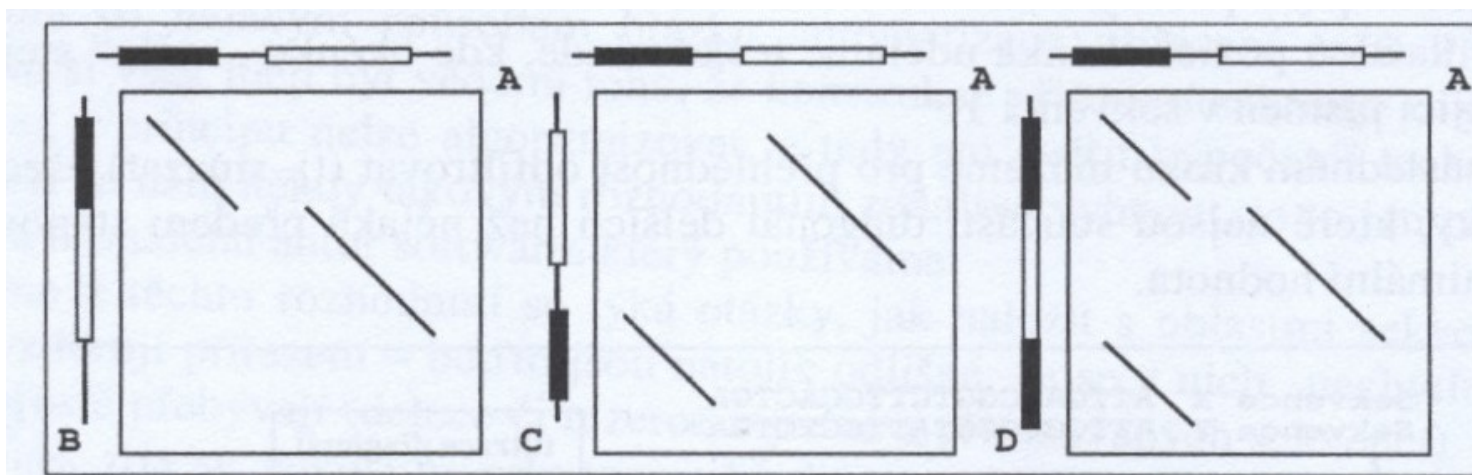


Cvrčková, Úvod do praktické bioinformatiky

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

Analytické nástroje

- příklady srovnání sekvencí pomocí bodového diagramu



Cvrčková, Úvod do praktické bioinformatiky

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

Analytické nástroje

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

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

[Search](#)

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

[Set subsequence](#) From: To:

[Choose database](#)

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

BLAST

Basic Local Alignment Search Tool

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

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

hodnota nepáru G-A

hodnota páru G-G

Cvrčková, Úvod do praktické bioinformatiky

Matrice PAM 250

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



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

Primární databáze

NC_002377.1: 145K..148K (2.9Kbp)

Genes

NP_059797.1

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

Links & Tools

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

Bibliography

Related articles in PubMed

BLAST

Basic Local Alignment Search Tool

BLINK precomputed BLAST

Home Taxonomy Report Multiple Alignment Blast Help

My NCBI [Sign In] [Register]

Pre-computed BLAST results for: [gi|16119781|ref|NP_396486.1](#) two component sensor kinase [Agrobacterium tumefaciens str. C58]
Matching gis: [15163423;20141871;1019660](#)
Total (score > 100) : 147086 hits in 146754 proteins in 6309 species
Selected: 147086 hits in 146754 proteins in 6309 species Filter: **Min Score: 100** |
Other views (Reports): [Taxonomy report](#) [Multiple Alignment](#) [Blast](#)
[Reset all filters](#)

Choose Display Options

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

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

% hits 833 aa [reset selection](#)

blink	SCORE	ACCESSION	Length	Protein Description
				Conserved Domain Database hits
◆	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 řada specializovaných verzí programu BLAST
 - vyhledávání podle zdroje (organismu) sekvencí, např. známých genomů mikroorganismů
 - **BLASTP**
 - vyhledávání podobnosti k proteinu v databázi proteinových sekvencí
 - **BLASTN**
 - vyhledávání podobnosti k nukleotidové sekvenci v databázi nukleotidových sekvencí
 - další varianty jako např. MEGABLAST pro identifikaci totožných nebo velice podobných sekvencí (vyhledává dlouhé podobné úseky nukl. sekvencí)
 - **BLASTX**
 - vyhledávání podobnosti k proteinu v databázi nukleotidových sekvencí přeložených do sekvence aa



BLAST

Specializované verze

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

BLAST

Specializované verze

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

BLAST

Specializované verze

□ Příklad vyhledávání pomocí PHI-BLAST

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase  
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPEPGPDR  
VADAKGDSESEEDLLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF  
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGS  
TSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIgek  
IYKDGERIITQGEKADSFYIESGEVSIILRSRTKSNKDGGNQEVVEIARCHKGQYFGELALVTNKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYEEQLVKMFGSSVDLGNLQ
```

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

Osnova

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

Analytické nástroje

- <http://workbench.sdsc.edu/>

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

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

beta-glucosidase

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

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

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

Analytické nástroje

- <http://workbench.sdsc.edu/>

View
View Nucleic Sequence(s)

Format Case

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

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

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

>170248
GAGCTCCCTTGGGGGGCAAGGGCAAAAACTTTTGCTAAATGGAAAAATATTATACCAAGTGTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAAATTTGACTATTTTGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCCAAGTAGCTTTCTTTAATTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAAATAAATTCAAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAATAAAGACCGATCAAAATAAAGCCGCCATTAAAAAATGAATTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACCTTTCCAATACTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTTATCTCTAATTTACATCTCAACTAATATTAAGAAAATTAACAGGTA
CAGCAAAATCATAAAAATTTCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCTTTTTCAGAG
TCTGCATGCCATATTCCTAAGGGGTCGTTTGGTACAAGAAAATAAATAAATAAATTTCCGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTTAATTATAAGTATTAGCTAATTTTCAGAAATAAATTTTACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACCAAATGATCGGTTAGTCTTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAAGTTAGAAAATTTTCAITTAATCAATTCATATAAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAAATAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAACCGTT
CGAGACTCTTTATGGAAGGCGTTTCTTCAAAATGAGATTCTCATTCAATTGCTCTGGTGC AATAGCAAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA
GAGAATTTCAAATCTCTCAACTACTTTTTAAGGGAATTCAAAATACGACC AATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTTATTTAAATTTGAATTGAAAATATTAATTTACTTGTATTAATATAA

Analytické nástroje

- <http://workbench.sdsc.edu/>

Regex pattern:

ett. {1,32}ett

0 sequences were searched

1 match was found

Matches are indicated in blue

>170248

```
GAGCTCCCTTTGGGGGGCAAGGGCAAAACTTTTGTCTAAATGGAAAAATATTATACCAAGTGTGTTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCCAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTTCTTTAAATATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAATAATAAACTTCAAGTCCATCATCTTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAAATAAGACCAGATCAAATAAAAAGCCGCCATTAAAAATAATGAATTTTAGGACTCTC
GATTTGGCAGGTAAGTGGCAAAACTCTTCCAATACTTTTGTGCAACTTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTA AACAGGTA
CAGCAAATCATAAAATTTTCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCATGCCATATTTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAAATTTTCGGGATAGAATTT
GAGATTGCATTTATCTTTGTTTTAATTATAAGTATTAGCTAATTTTCAAGAATAAATTTTACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACCAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAAATTTTCAATTAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAGATGTACCGTTAATAATAAAGATAAGATAGAGTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTGTCTTCAAAGTAGATTTCTCATTCATTGCTCTGGTGCATAGCAAAA
TGACATCTTACTCTTAAGATACAGGAGCCACTCTACAATCTTCTATTGTATACTCAAAATGAAAGTTTTA
GAGAACTTTTCAAACTCTCAACTACTTTTAAAGGGAATTCAAAAATACGACCAATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTAAATTTGAAATTTGAAATATTAAATTTACTTGAATTAATATAA
ACAAATAGATATCGCTAAGTATTTACCACAAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT
GATTAAGCAGCTATTCATCTGGTTTGTGCAGGATGAAAGAAAGTAACTAGCTATAATTTCTTTTGTAAAGT
```

Analytické nástroje

- <http://workbench.sdsc.edu/>

Frame 1, 1 stop codon

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

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

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

Frame 2, 1 stop codon

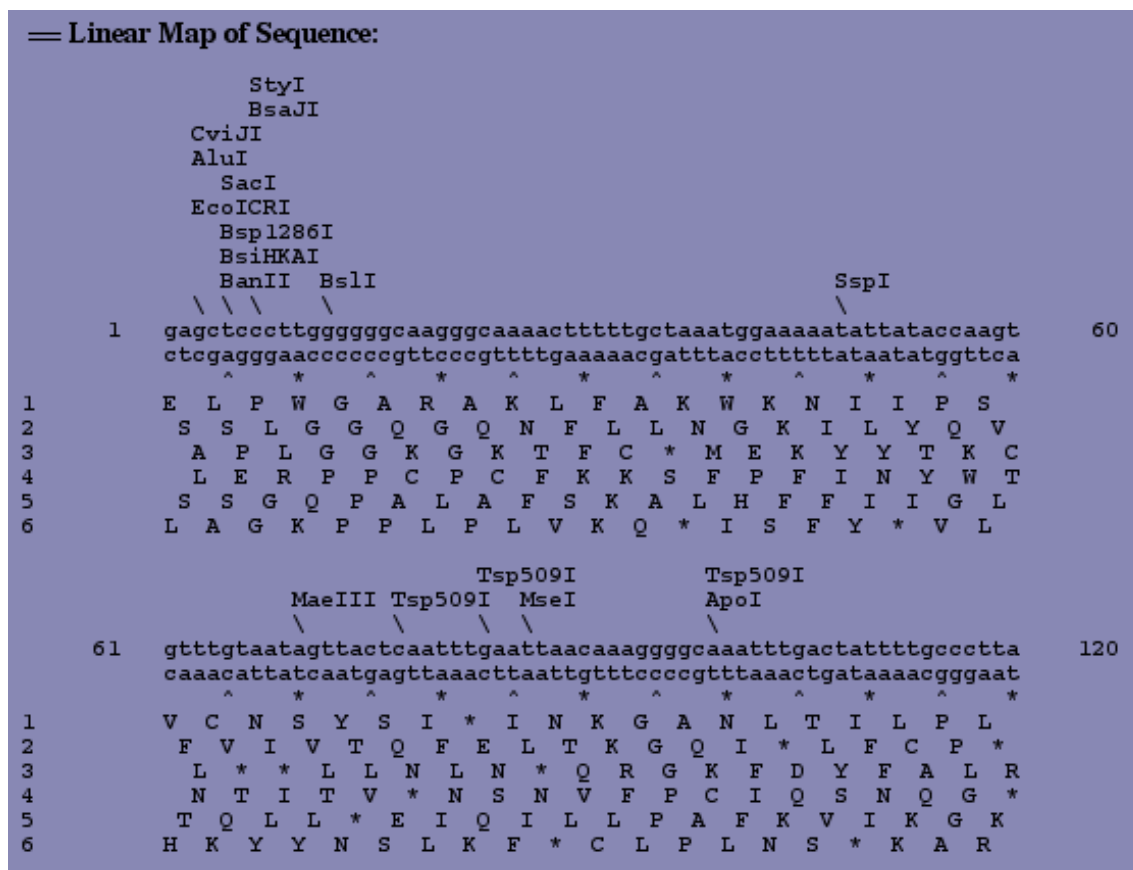
Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. Tran

```
>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    agtcccttggggggcaagggcaaaactttttgctaaatggaaaaatattataccaagt 61  
      F V I V T Q F E L T K G Q I * L F C P  
62   tttgtaatagttactcaatttgaattaacaaaggggcaatttgactattttgcctta 120
```

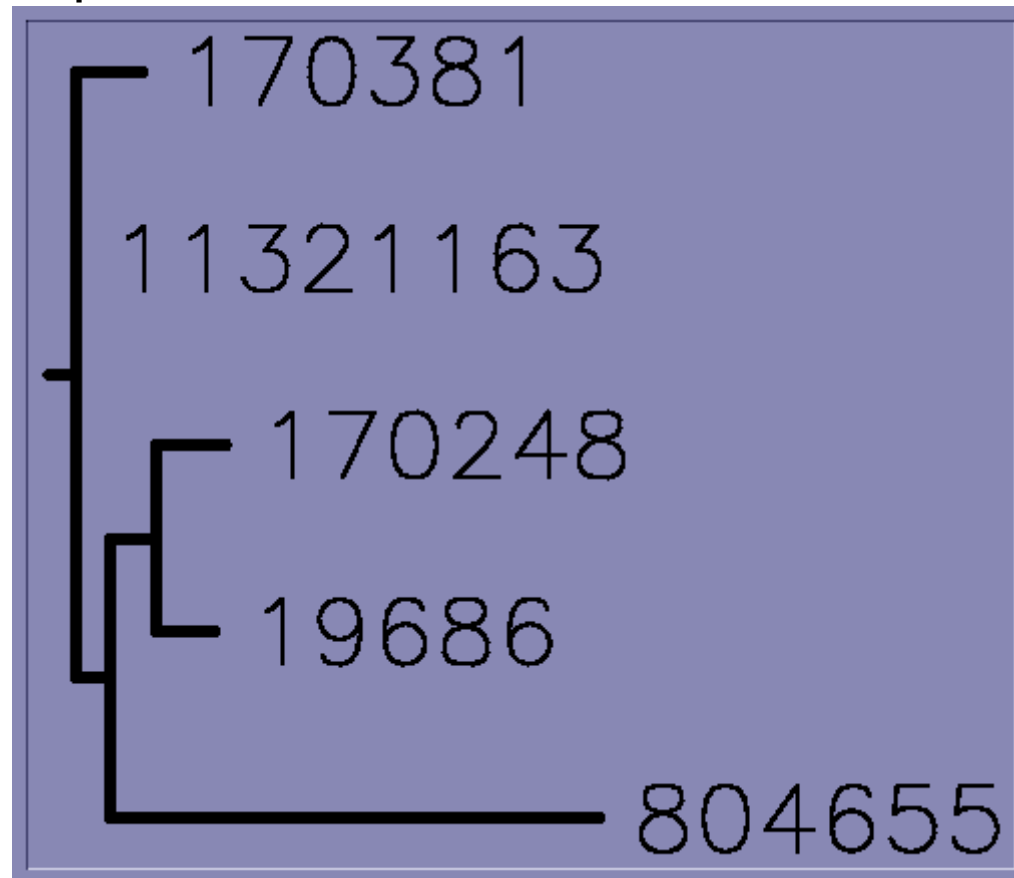
Analytické nástroje

- <http://workbench.sdsc.edu/>




Analytické nástroje

- <http://workbench.sdsc.edu/>



Analytické nástroje

- 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 ([UB 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

- VPCR <http://grup.cribi.unipd.it/cgi-bin/mateo/vpccr2.cgi>



Osnova

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

Další WWW zdroje

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

PHACTR4 phosphatase and actin regulator 4 [Homo sapiens]
Gene ID: 65979, updated on 27-Aug-2011

Summary

Official Symbol PHACTR4 provided by [HGNC](#)
Official Full Name phosphatase and actin regulator 4 provided by [HGNC](#)
Primary source [HGNC:25793](#)
Locus tag RP11-442N24_A.1
See related [Ensembl:ENSG00000204138](#); [HPRD: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 (28896093..28826881)

Genomic regions, transcripts, and products

Genomic Sequence NC_000001 chromosome 1 reference GRCh37.p5 Primary Assembly

Links

- Order cDNA clone
- BioAssay, by Gene target
- BioProjects
- CCDS
- Conserved Domains
- dbVar
- EST
- Full text in PMC
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Probe
- Protein
- PubChem Compound
- PubChem Substance
- PubMed
- PubMed (GeneRIF)
- PubMed (OMIM)
- RefSeq Proteins

See PHACTR4 in MapViewer

Go to [reference sequence details](#)

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

Další WWW zdroje




▪ Online Mendelian Inheritance in Man (OMIM)

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

Search OMIM [Sample Searches](#)

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

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

OMIM[®] 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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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