

CG920 Genomics

Lesson 1

Introduction into Bioinformatics

Jan Hejátko

Functional Genomics and Proteomics of Plants,
Mendel Centre for Plant Genomics and Proteomics,
Central European Institute of Technology (CEITEC), Masaryk University, 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

Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY and STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...
 - Other On-line Genome Tools

Course Syllabus

- **Chapter 01**
 - Introduction into Bioinformatics

- **Chapter 02**
 - Identification of Genes

- **Chapter 03**
 - Reverse Genetics Approaches

- **Chapter 04**
 - Forward Genetics Approaches



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Course Syllabus

- **Chapter 05**
 - Functional Genomics Approaches

- **Chapter 06**
 - Protein-Protein Interactions And Their Analysis

- **Chapter 07**
 - Current Methods of DNA Sequencing

- **Chapter 08**
 - Structure of genomes



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Course Syllabus

- **Chapter 09**
 - Genome evolution

- **Chapter 10**
 - Genomics and Systems Biology

- **Chapter 11**
 - Practical Aspects Of Functional Genomics
 - Model Organisms,
 - PCR and Primer Design



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Literature

- Literature resources for **Chapter 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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Outline

- Syllabus of the course
- Definition of Genomics



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

GENOMICS – What is it?

- *Sensu lato* (in the broad sense) – it is interested in **STRUCTURE and FUNCTION** of genomes
 - Necessary prerequisite: knowledge of the genome (sequence) – work with databases
- *Sensu stricto* (in the narrow sense) – it is interested in **FUNCTION** of **INDIVIDUAL GENES** – **FUNCTIONAL GENOMICS**
 - It uses mainly the reverse genetics approaches

GENOMICS – What is it?

The role of BIOINFORMATICS in FUNCTIONAL GENOMICS

Forward („classical“) Genetics Approaches

Reverse Genetics Approaches

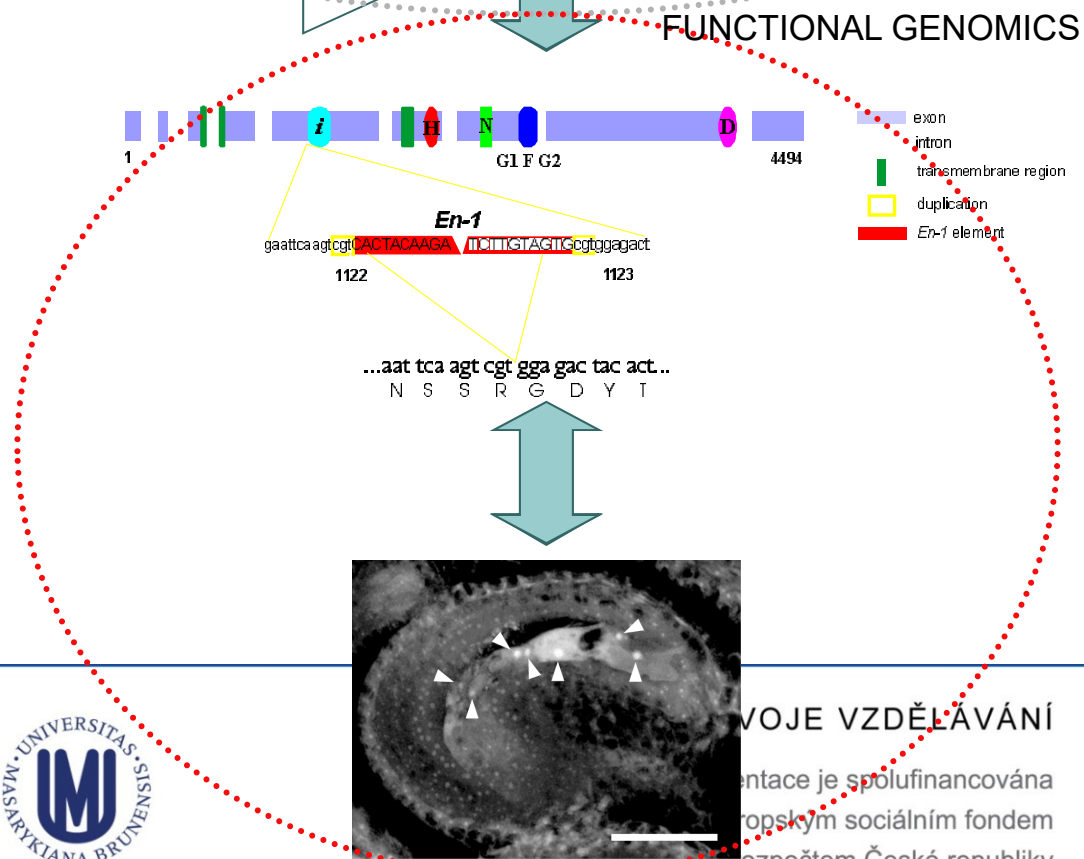
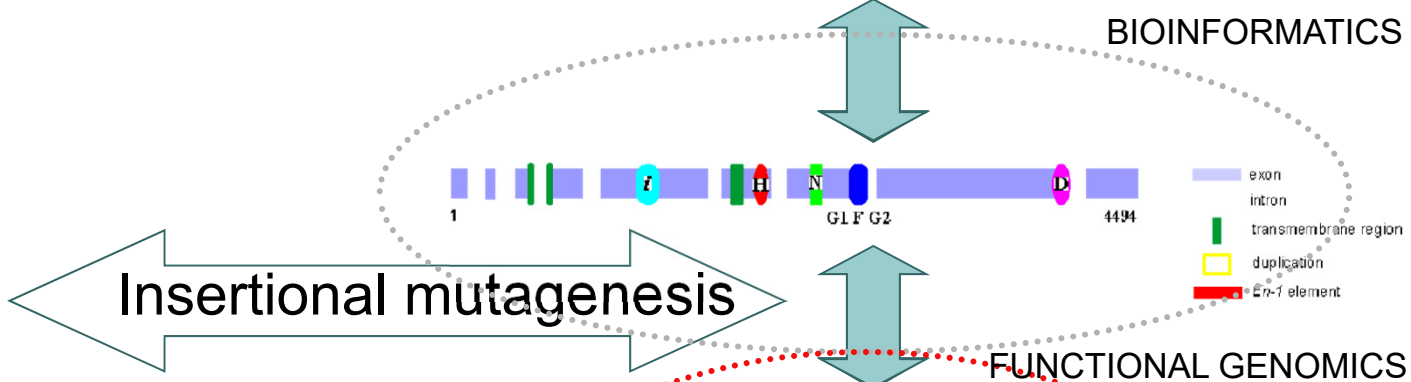
5'TTATATATATATATTAATAAAATAAAATAAAA
GAACAAAAAGAAAATAAATA....3'



3

:

1



?

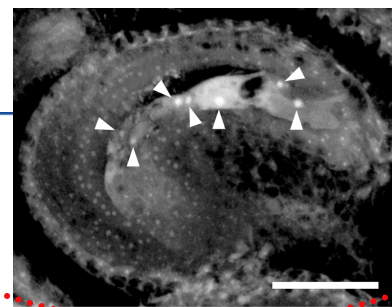


EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

pro konkurenceschopnost



VOJE VZDĚLÁVÁNÍ

entace je spolufinancována
ropským sociálním fondem
a státním rozpočtem České republiky

Outline

- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS

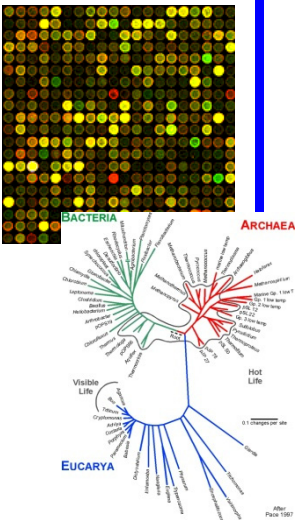


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Bioinformatics

- **Definition of Bioinformatics** (according to NIH Biomedical Information Science and Technology Initiative Consortium)



Research, development, or application of computational tools and approaches for expanding the **use of biological, medical, behavioral or health data**, including those to **acquire, store, organize, archive, analyze, or visualize such data.**

What is bioinformatics?

- **Interface** between the **biology** and **computers**
- **Analysis** of **proteins, genes** and **genomes** using **computer algorithms** and **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.

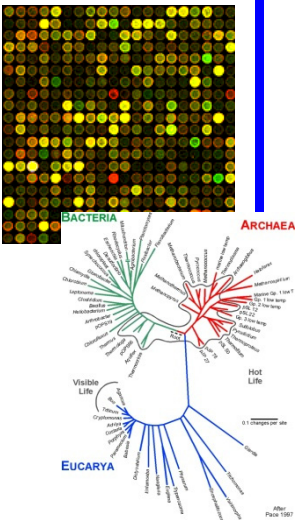
J. Pevsner,
<http://www.bioinfbook.org/index.php>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Bioinformatics



- **Bioinformatics in functional genomics**
 - **Processing and analysis of sequencing data**
 - Identification of reference sequences
 - Identification of genes
 - Identification of homologues, orthologues and paralogues
 - Correlative analysis of genomes and phenotypes (incl. human)
 - **Processing and analysis of transcriptional data**
 - Transcriptional profiling using DNA chips or next-gen sequencing
 - **Evaluation of experimental data and prediction of new regulations in systems biology approaches**
 - Mathematical modelling of gene regulatory networks

Outline

- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS
- **Databases**
 - Spectre of „on-line“ resources



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

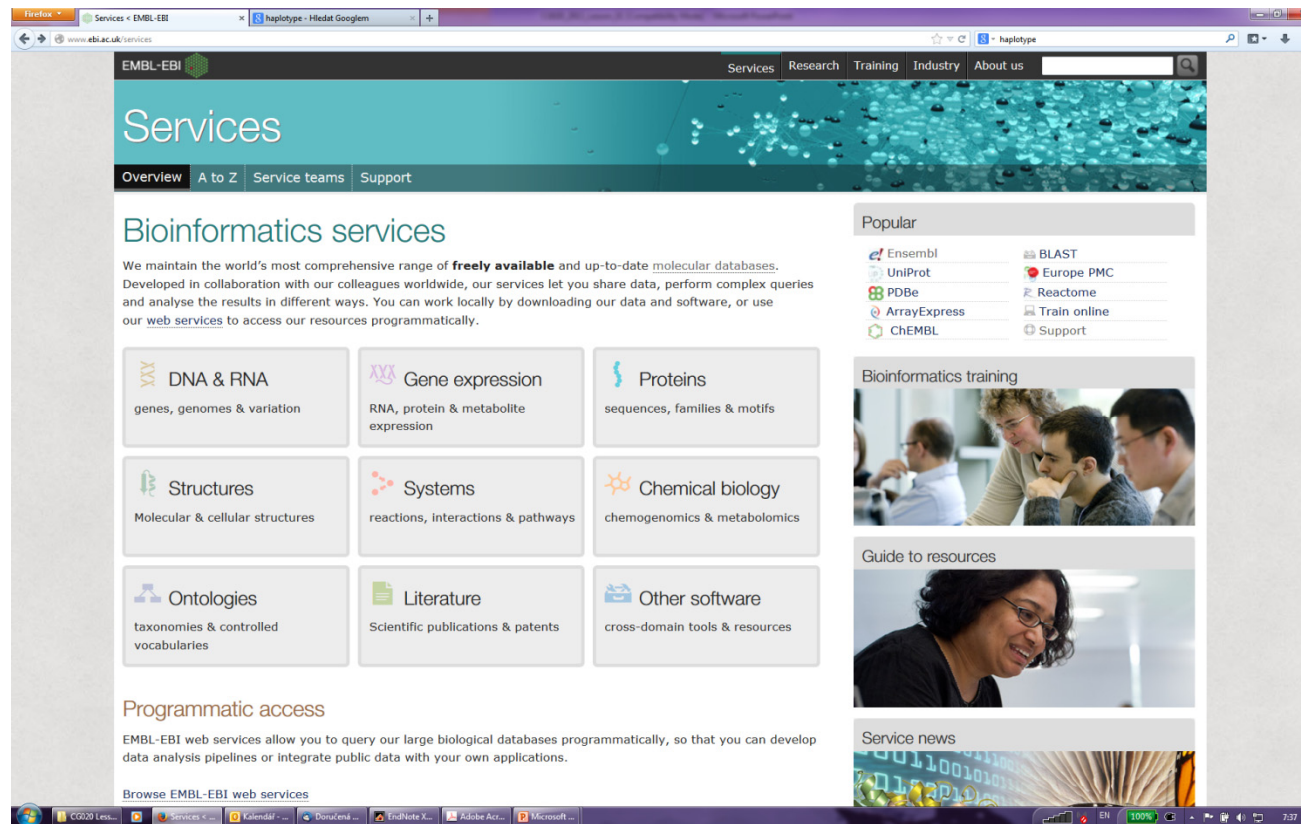
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Spectre of on-line Resources

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/
INFOTIAGEN	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/
INCEBI	Ireland	http://acer.gen.tcd.ie/
INN	Israel	http://dapsas.weizmann.ac.il/bcd/inn.html
IEN-ADR	Italy	http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm
CAOS/CAMM	Netherlands	http://www.caos.kun.nl/
Bio	Norway	http://www.no.embnet.org/
IBB	Poland	http://www.ibb.waw.pl/
IGC	Portugal	http://www.igc.gulbenkian.pt/
GeneBee	Russia	http://www.genebee.msu.su/
CNB-CSIC	Spain	http://www.es.embnet.org/
BMC	Sweden	http://www.embnet.se/
SIB	Switzerland	http://www.ch.embnet.org/
SEQNET	UK	http://www.seqnet.dl.ac.uk/
EMBNet Specialist Nodes		
MIPS	Germany	http://www.mips.biochem.mpg.de/
ICGEB	Italy	http://www.icgeb.trieste.it/
Pharmacia Upjohn	Sweden	http://www.pnu.com/
F.Hoffmann-La Roche	Switzerland	http://www.roche.com/
EBI	UK	http://www.ebi.ac.uk/
HGMP-RC	UK	http://www.hgmp.mrc.ac.uk/
Sanger	UK	http://www.sanger.ac.uk/
UMBER	UK	http://www.bioinf.man.ac.uk/dbbrowser
EMBNet Associate Nodes		
IBBM	Argentina	http://sol.biol.unlp.edu.ar/embnet
ANGIS	Australia	http://www.angis.su.oz.au/
CBI	China	http://www.cbi.pku.edu.cn/
CIGB	Cuba	http://bio.cigb.edu.cu/
CDFD	India	http://salarjung.embnet.org.in/
SANBI	South Africa	http://www.sanbi.ac.za
USA Information Providers		
NCBI	USA	http://www.ncbi.nlm.nih.gov/
NLM	USA	http://www.nlm.nih.gov/
NIH	USA	http://www.nih.gov/

Spectre of on-line Resources

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



Spectre of on-line Resources

□ NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI homepage with the following elements:

- Search Bar:** "All Databases" search field with a "Search" button.
- Navigation Menu (Left):**
 - NCBI Home
 - Resource List (A-Z)
 - All Resources
 - Chemicals & Bioassays
 - Data & Software
 - DNA & RNA
 - Domains & Structures
 - Genes & Expression
 - Genetics & Medicine
 - Genomes & Maps
 - Homology
 - Literature
 - Proteins
 - Sequence Analysis
 - Taxonomy
 - Training & Tutorials
 - Variation
- Welcome to NCBI:**

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

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)
- Get Started:**
 - Tools:** Analyze data using NCBI software
 - Downloads:** Get NCBI data or software
 - How-To's:** Learn how to accomplish specific tasks at NCBI
 - Submissions:** Submit data to GenBank or other NCBI databases
- NCBI YouTube channel:**

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel. [GO](#)
- Popular Resources:**
 - PubMed
 - Bookshelf
 - PubMed Central
 - PubMed Health
 - BLAST
 - Nucleotide
 - Genome
 - SNP
 - Gene
 - Protein
 - PubChem
- NCBI Announcer:**

New version of Gen... available

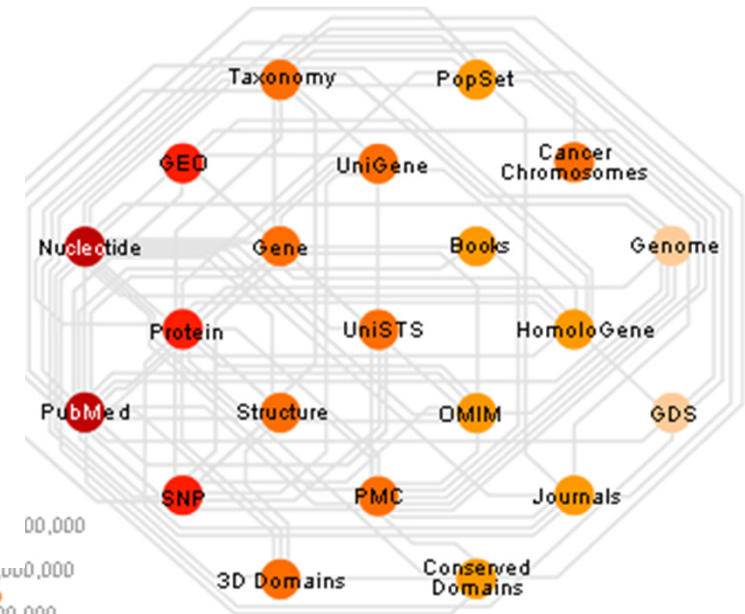
An integrated, downlo... for viewing and analy...

NCBI's July Newslett... Bookshelf

Introduction to the 10... Browser. PubMed's C...

New Microbial BLAS...

Now easier to use an... format and features c... BLAST services. inclu...



Outline

- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS
- Databases
 - Spectre of „on-line“ resources
 - PRIMARY, SECONDARY and STRUCURAL databases



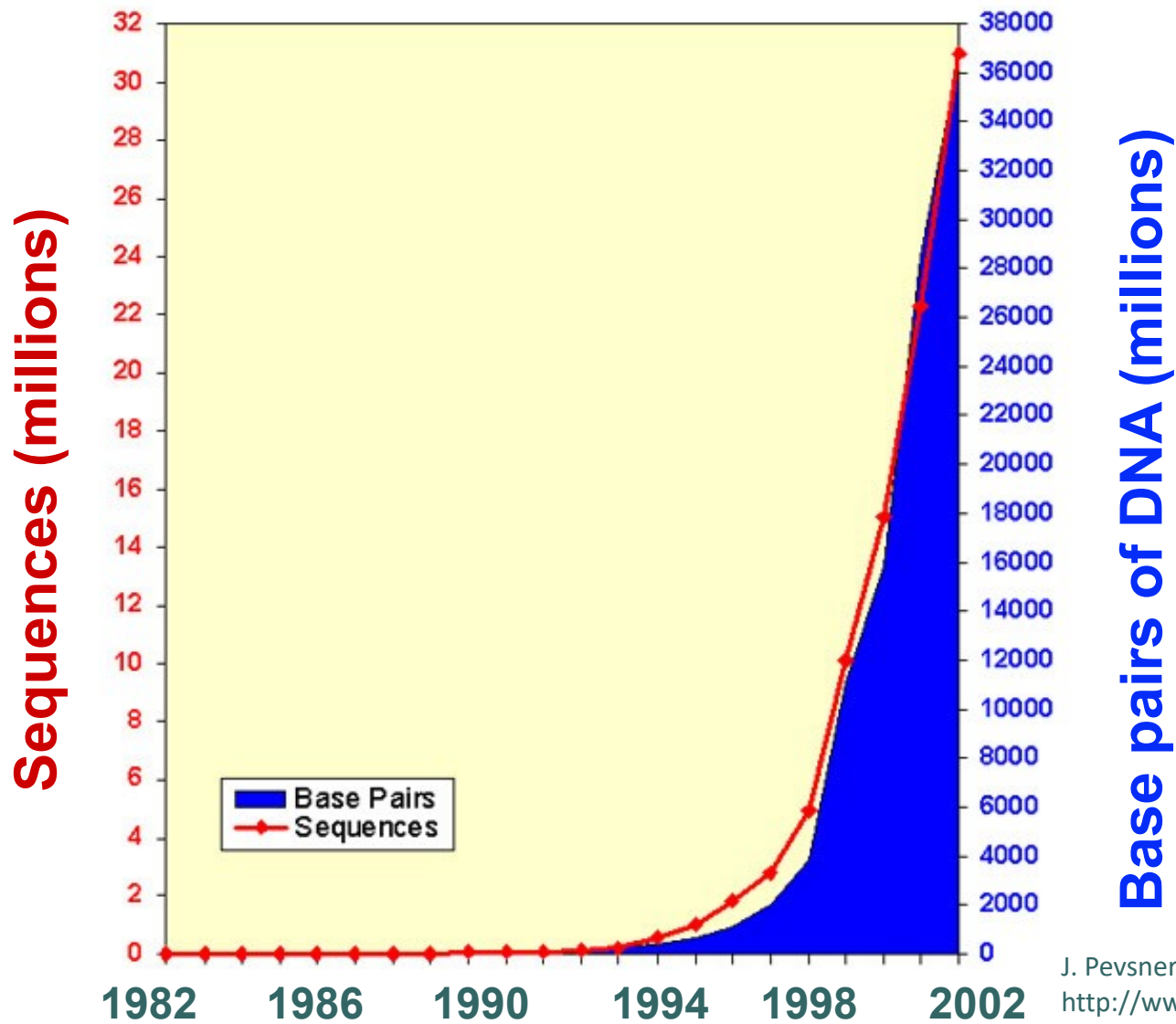
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primary Databases

- Include primary datasets – DNA and Protein sequences
 - Sequences in databases of „The Big Three“:
 - EMBL
 - <http://www.ebi.ac.uk/embl/>
 - GenBank,
 - <http://www.ncbi.nih.gov/Genbank/GenbankSearch.html>
 - DDBJ,
 - <http://www.ddbj.nig.ac.jp>
 - Daily mutual exchange and backup of data
 - Works with large amount of data (capacity and software requirements)
 - September 2003 $27,2 \times 10^6$ entries (approx. 33×10^9 bp)
 - August 2005 100×10^9 bp from 165.000 organisms

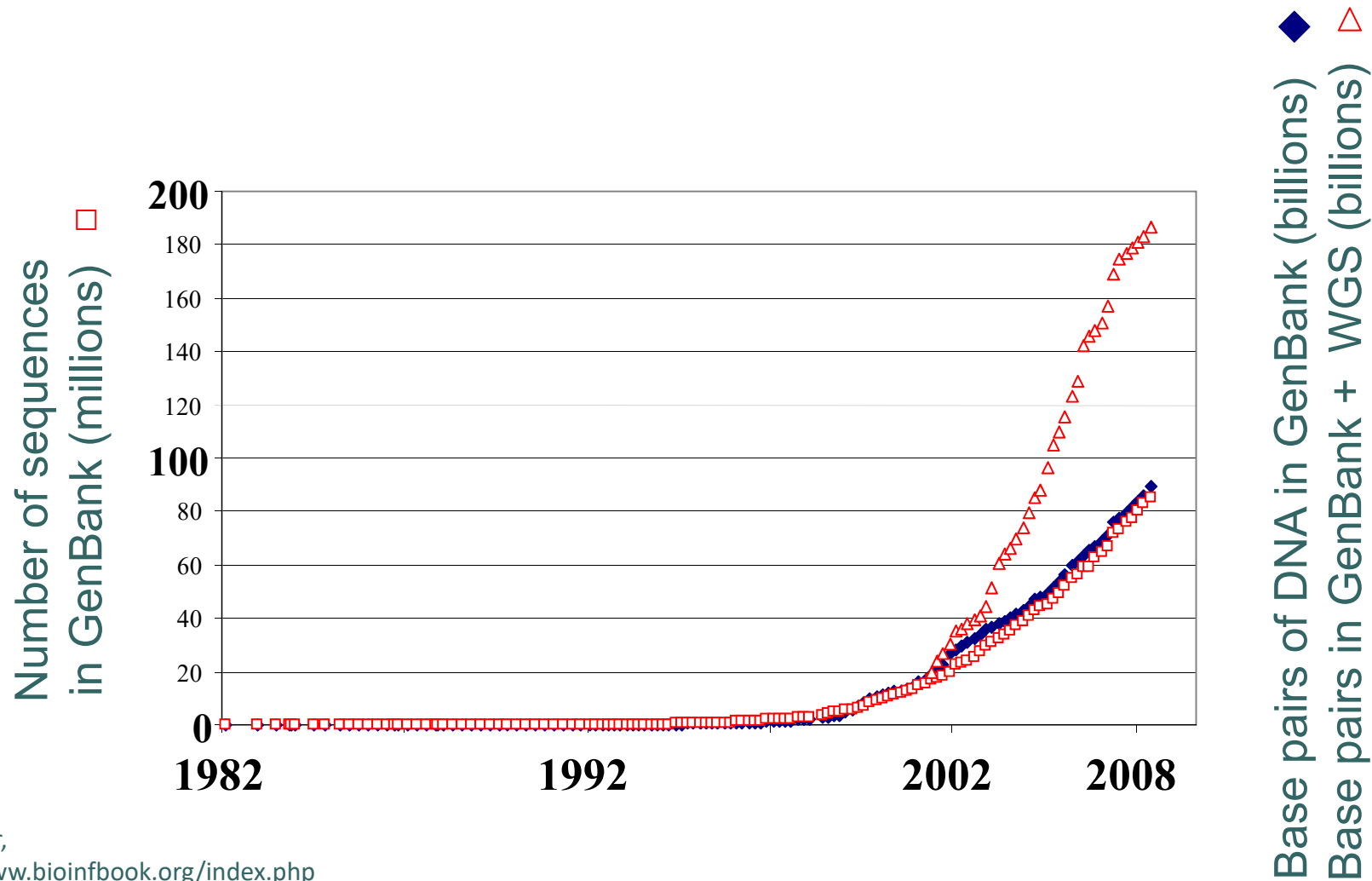
Growth of GenBank



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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



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

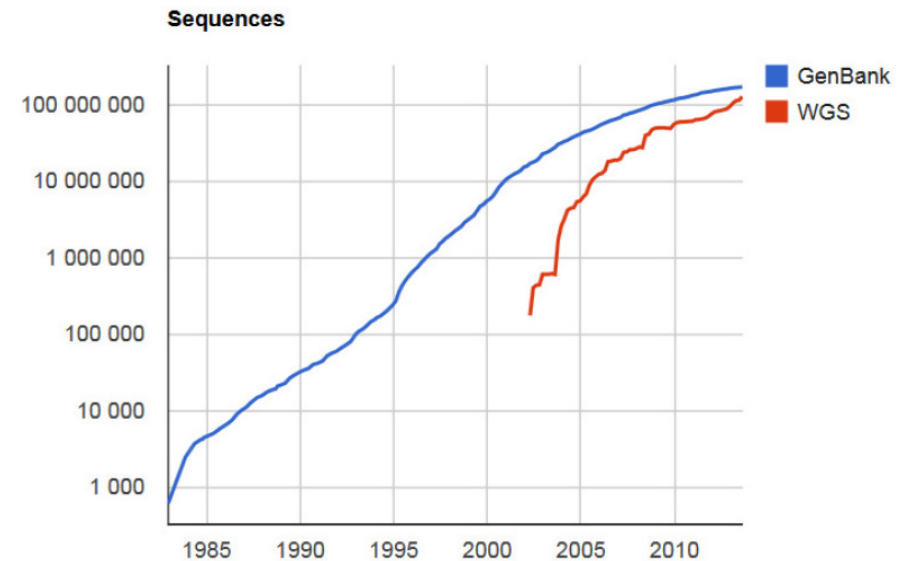
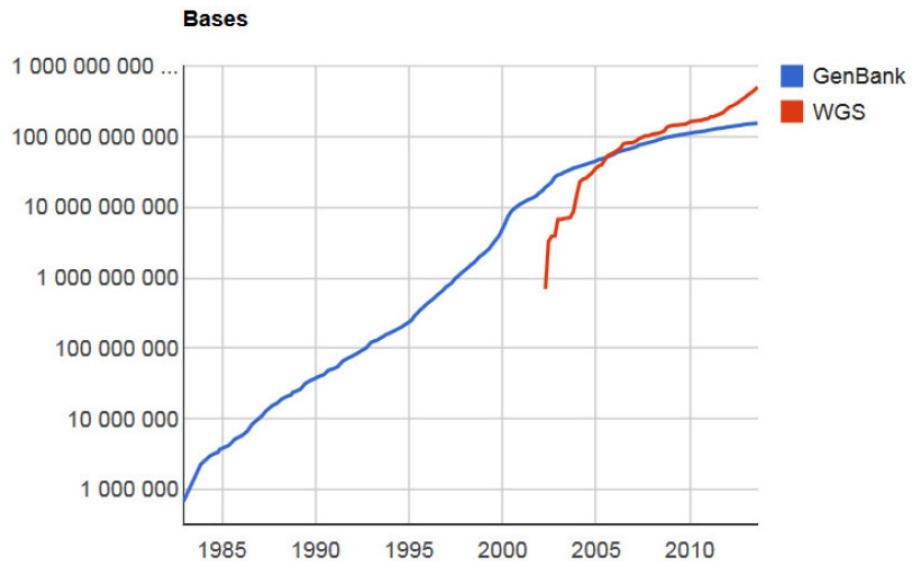


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

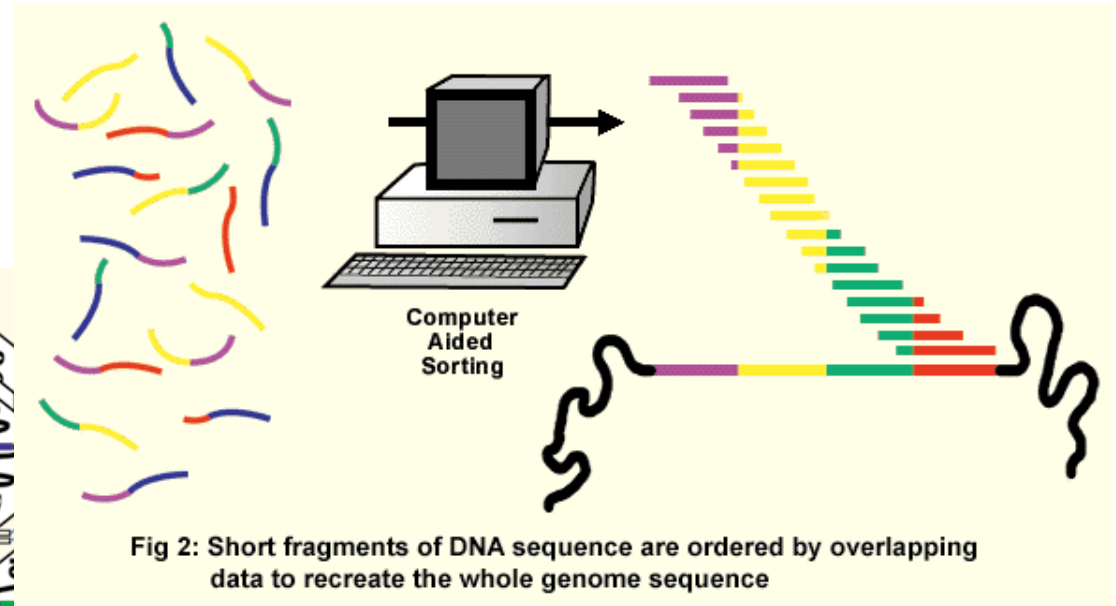
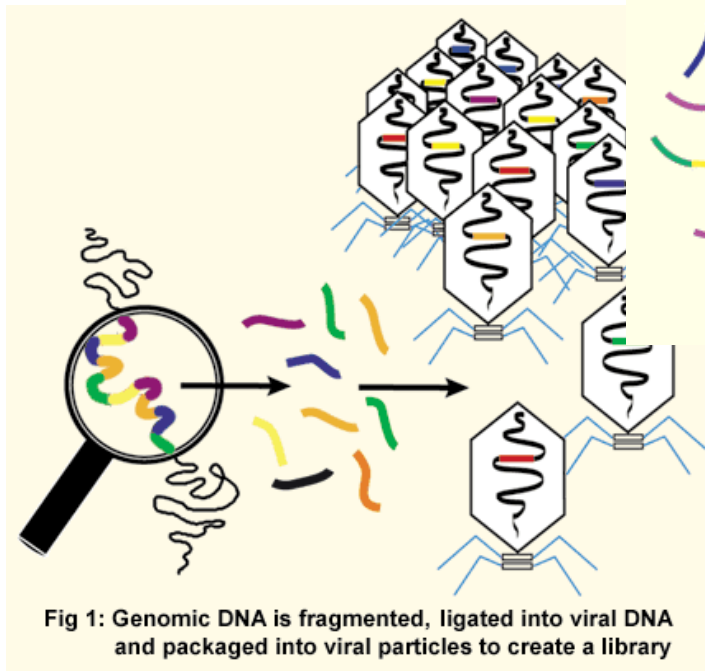
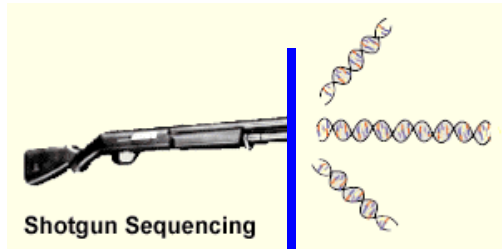
Tato prezentace je spolufinancována
 Evropským sociálním fondem
 a státním rozpočtem České republiky

Growth of GenBank

Feb 15 2013

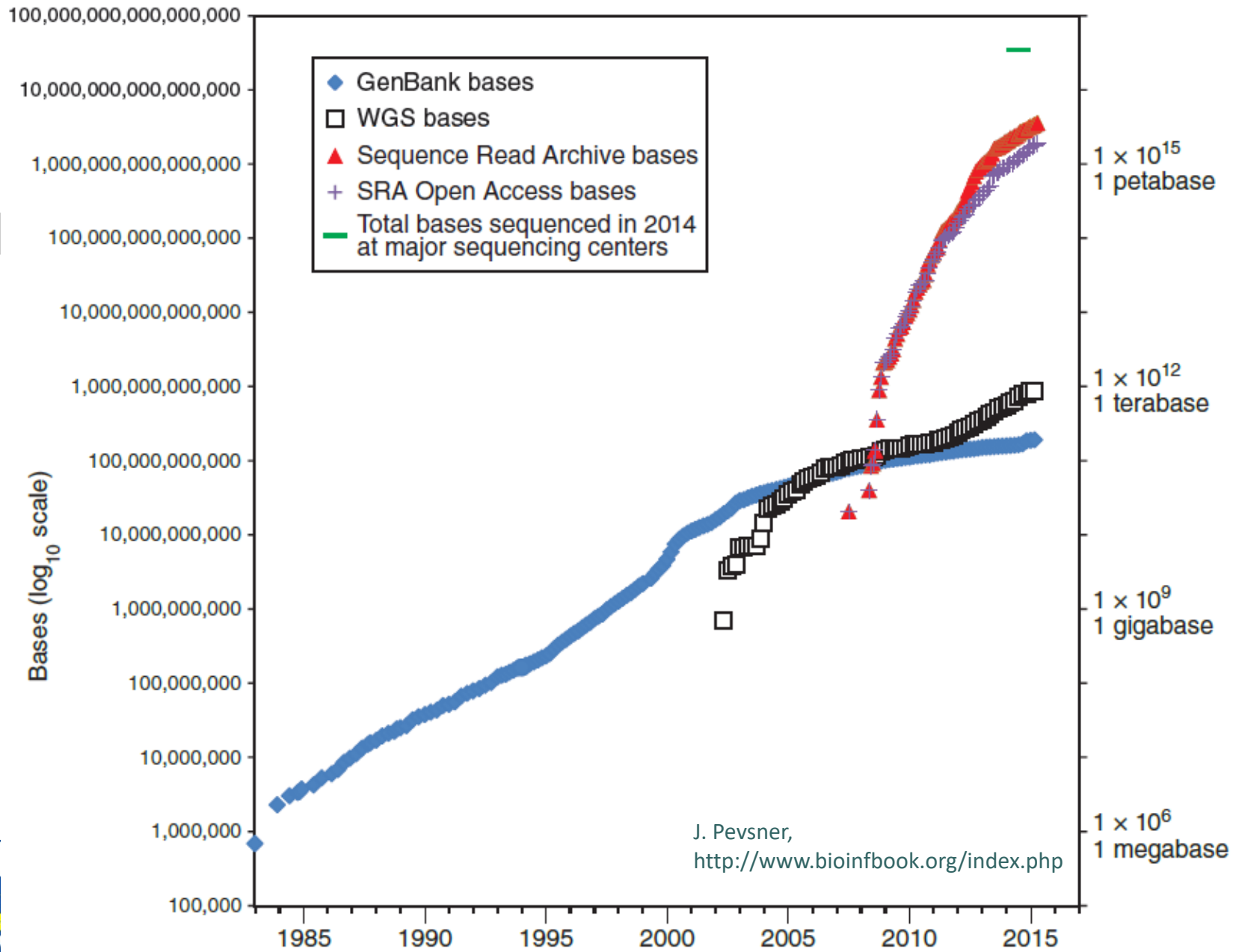


WGS



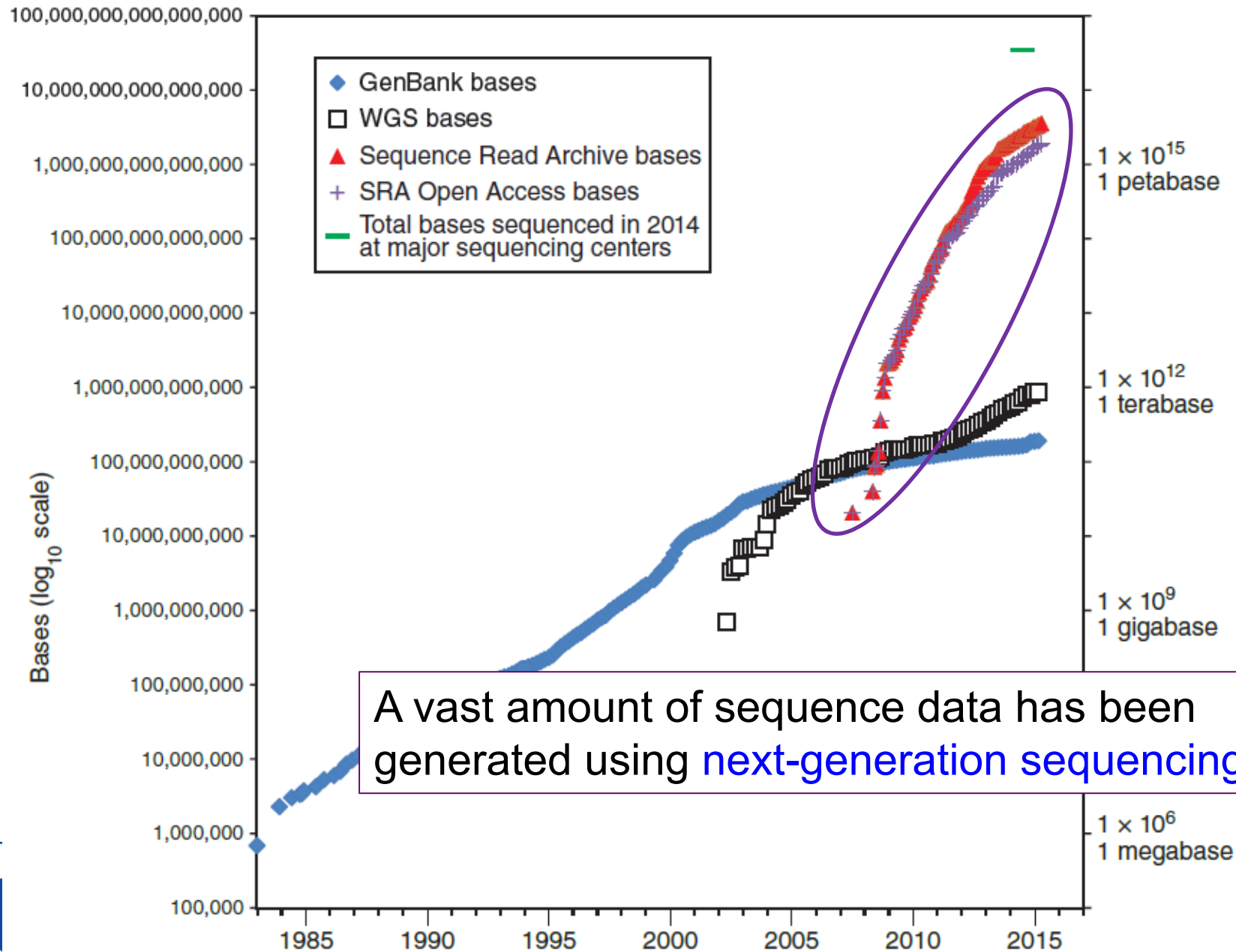
Interactive concepts in biochemistry, Rodney Boyer, Wiley, 2002, <http://www.wiley.com//college/boyer/0470003790/>

Growth of DNA Sequence in Repositories

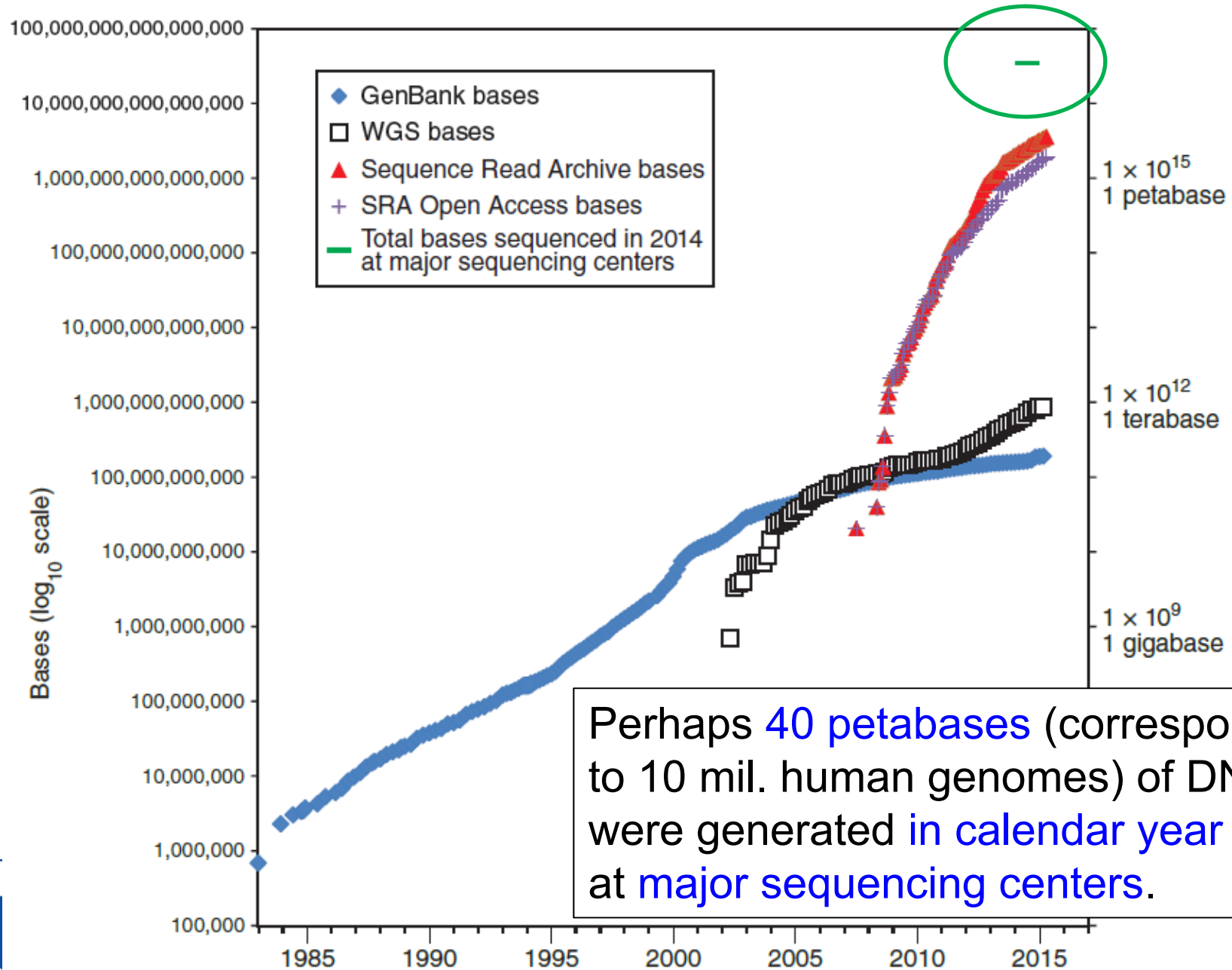


J. Pevsner,
<http://www.bioinfbook.org/index.php>

Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



Primary Databases

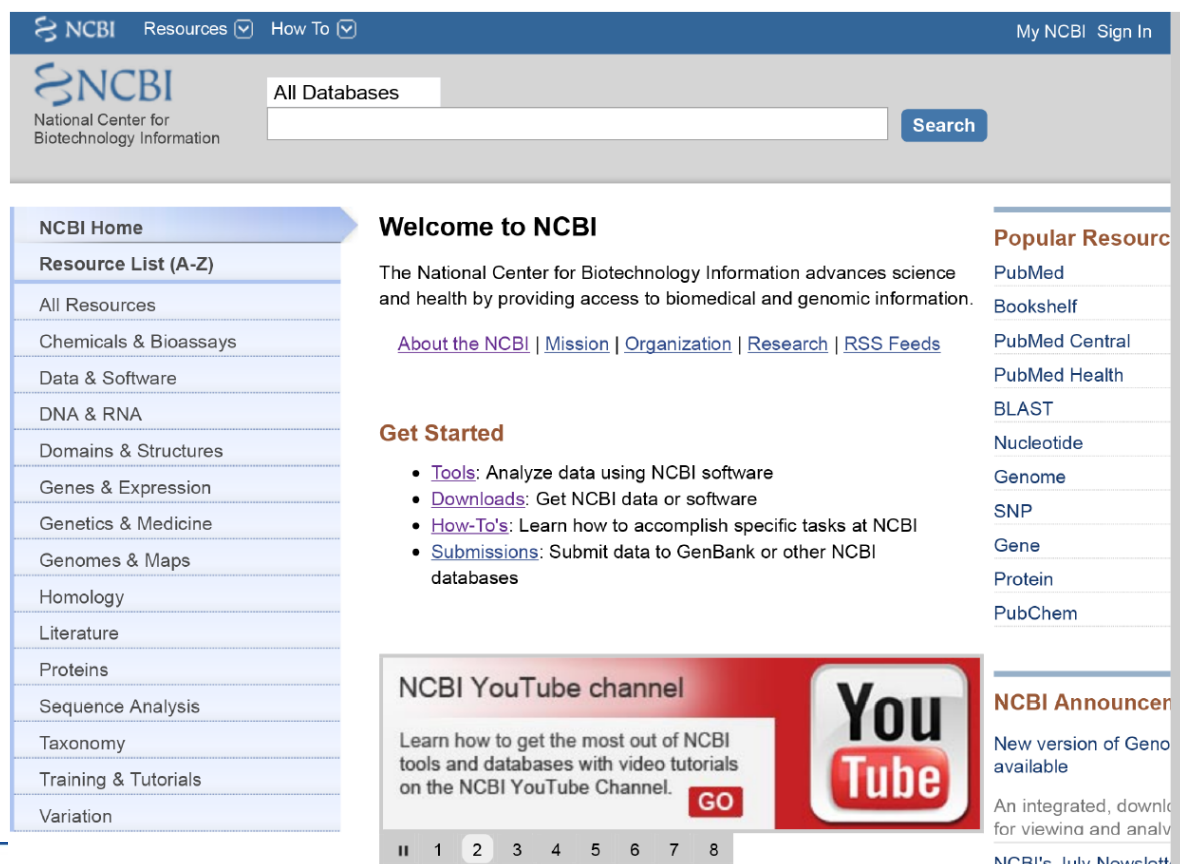
- They include sets of primary data – [DNA](#) and [Protein](#) sequences
 - Protein sequences:
 - PIR, <http://pir.georgetown.edu/>
 - MIPS, <http://www.mips.biochem.mpg.de>
 - SWISS-PROT, <http://www.expasy.org/sprot/>

Primary Databases

- Types of sequences in primary databases
 - Standard nucleotide sequences acquired by high quality sequencing
 - **ESTs** (**E**xpressed **S**equence **T**ags)
 - **HGTS** (**H**igh **T**hroughput **G**enome **S**equencing)
 - Results of sequencing projects without annotation
 - **Reference Sequences** of annotated genomes
 - **TPAs** (**T**hird **P**arty **A**nnotation)
 - sequences annotated by third party (by someone else, not the original authors)

Primary Databases

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', 'Popular Resources', and 'NCBI Announcements'. The 'Get Started' section lists links for 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 tool.

Primary Databases

Gene symbol virA
Gene description two-component VirA-like sensor kinase
Locus tag pTl_125
Gene type protein coding
RefSeq status PROVISIONAL
Organism *Agrobacterium tumefaciens* (old-name: *Agrobacterium tumefaciens*, qb-synonym: *Rhizobium radiobacter*)
Lineage Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiaceae; Rhizobium/Agrobacterium group; Agrobacterium; Agrobacterium tumefaciens complex

Genomic context
Location: plasmid: Ti
Sequence: NC_002377.1 (145694..148183)

Genomic regions, transcripts, and products
Genomic Sequence: NC_002377

Sequence Viewer: NC_002377.1:145K-148K (3.2Kbp) | Find on Sequence: | Go to nucleotide | Graphics | FASTA | GenBank

Gene Details (circled in yellow):
NP_059797.1
NP_059797.1: two-component VirA-like sensor kinase
total range: NC_002377.1 (145,694..148,183)
total length: 2,490
strand: plus
protein product length: 829
Links & Tools
GenBank View: NC_002377.1 (145,694..148,183); NP_059797.1
FASTA View: NC_002377.1 (145,694..148,183); NP_059797.1
BLAST Genomic: NC_002377.1 (145,694..148,183)
Graphical View: NP_059797.1
BLAST Protein: NP_059797.1
BLINK Results: NP_059797.1

Bibliography
Related articles
1. [Sequence analysis of the virA locus from Agrobacterium tumefaciens octopine Ti plasmid pTl15955](#), Schrammeijer B, et al. J Exp Bot. 2000 Jun. PMID 10948245.
2. [The virA promoter is a host-range determinant in Agrobacterium tumefaciens](#), Turk SC, et al. Mol Microbiol. 1993 Mar. PMID 8469115.
3. [Characterization of the virA locus of Agrobacterium tumefaciens: a transcriptional regulator and host range determinant](#), Leroux B, et al. EMBO J. 1987 Apr. PMID 3595559.
4. [Analysis of the complete nucleotide sequence of the Agrobacterium tumefaciens 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](#)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primary Databases

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

Genes

NP_059797.1

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

Links & Tools

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

Bibliography

Related articles in PubMed

Primary Databases

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

TITLE Octopine-type Ti plasmid sequence

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 2490)

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

TITLE Direct Submission

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

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

FEATURES

Location/Qualifiers

source

1..2490

/organism="Agrobacterium tumefaciens"

/mol_type="genomic DNA"

/db_xref="taxon:358"

/plasmid="Ti"

/note="extrachromosomal octopine-type"

gene

1..2490

/gene="virA"

/db_xref="GeneID:1224316"

CDS

1..2490

/gene="virA"

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

/codon_start=1

/transl_table=11

/product="virA"

/protein_id="NP_059797.1"

/db_xref="GI:10955141"

Primary Databases

```
/translation="MNGRYSPTRODFKTKGAKPWSILALIYAAMI FAFMAVASWQDNMT
TQAILSQLRSINADSASLQRDVLRAHTCTVANYRPI I SRLGALRKNLEDLKQLFRQSH
IVSEENRQQLRQLEVSLSADAAVAAPGQNVRLQDSIASPTRALSSLPKASTDQT
LEKPTELASMMQLQFLRQSPASISPHISLELEELKQRGLDEAFVILAREGPIILSLL
PQVKDLVNMQTSDBTAEIEMLRQRCLEVYSLKNVEERSARIPLSSASVGLCLYIITL
VYLRKKTDWLARRLDYHEL I KEI GVCFBGRATTTSSQAALRI IQRPFDADTCALAL
VDHDERWAVETFGAKHFKPVWDSVLRRIVSRKADBRATVFRILSSKKIVHLFLHIP
GLSILLAHKSTDKLIAVCSLGYQSYRFPFCQGEIQLLELATACLCHYIDVRRKQTECD
VLARRLEHAQRLAVGTLAGGIAHFFNNILGSLGHAEALQNSVSRTEVTRRYIDYII
SSGDRAMLIIDQILTLRKRQEMIKPPSVSELVTEIAFLRMLPFPNIELSFRPDQMC
SVI EGSPLRLQQLINICKNASQAMTANQIDII IIGQAPLPVKKILAHGVMPFGDYVL
LSISDNQGGIPRAVLPHIPEPFFPTRARNGGTGLGLASVHGHSAPAGYIDVSTVGH
GTRFDIYLPSPSKKFPVNPDSFFGRNKA PRGNGHI VALVPPDDLREAYRDKI AALGYE
PVGFRTPNKRDIWISKGNEADLVMDQASLPEDQSPNSVDLVKTA SIIIGGNDLKM T
LSREDVT RDLYLPKPISSRTMAHALTKIKT"
```

ORIGIN

```
1 atgaacggaa gatattcacc gaecggcgag gattttaaga caggcgcgaa gccctggctt
61 atattggccc ttatcgttgc tgaatgatt ttocggttca tggcggttgc gtcctggcag
121 gacaatgcca ctaccocagc aatcctcagc caactacgat cgat taacgc cgcacagccc
181 tcactgcagc gogatgact cgcgcctcac acgggcaacgc tggcgaacta ccgcccacat
241 atctccaggc tgggagctct gcggagaagt ctggaagatt tgaagcaatt atttagacaa
301 tctcatattg taagtgcagc caatgctgct caactgctac gccagctaga agtgtctcta
361 aatctggctg acgocggctg cgcgcctctt ggtgcgcaaa atgtacgctt gcaagattcg
421 ctggcagctt tcactcgtgc tttgagcagt ctccagcaga aagcctcaac cgatcagact
481 ttgaaaaaac caacagaatt ggttagcagt atgctccaat ttctcggca accaagcccg
541 gctatttcat toagatcag ccttgaacta gagaggctcc aaaaacaacg cggctttgat
601 gaagctcccg tgcgcaact tgcacgtgaa ggtccaccata tcttctcgtt tttgccacag
661 gtgaaagatc tggtaaacat gatcagcagc tctgacacgc cagaatctgc gtagatgctg
721 cagcgcgagt gtttgagggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc
781 ttctctgggt cgcctcagc gggctcttgc ctctacatca tcacctagc ctataggcta
841 cgcacacaaa cagattggtt agcgcggcgt ttagattcag aagagctaat caaagagatc
901 ggagtagttt ttgaagtgta ggcggccacc acgtcgtcgc cgcacagctc actctgtatt
961 atcagcgcct tcttgatgc cgtacgtgce gcttagctc tagtggacca tgacgtaga
1021 tgggctgtgc aaacattcgg tgcgaaacac caaaacctg tctgggacga cagcgtgcta
1081 cgcgaaatag tctctcgtac caaagcggac gaacgggcca cggtatccg catcatatcg
1141 tgcacacaaa tctacattt gccctctcga atccagctc tctcgatact actggtctcgc
1201 aaatccacag ataaactaat tgcggtttgt tcaactgggtt accaaagcta tgcgcctcga
1261 ccttgccaag gcaaaatca gctcttggaa ctgcacacgc cctgcctctg tcatatatac
1321 gatgttcggc gtaagcagac cgaatgcgac gttttggcca gacgatgga gcatgcgcaa
1381 cgccttgagg cagttggtac acttgcgggc ggaatgacac atgaatttaa taacattttg
1441 gctcaaatcc tgggcaacgc agaattagca caaaactcgg tctctcgaac atctgtcacc
1501 cgaagatata ttgactatat cattctgcga ggcgacagag ccatgctcat tctcagcag
1561 atcttgacgc tgagccgaaa acaggagcgc atgatcaagc catttagtgt ctcagagctt
1621 gtgaccgaaa tgcctcctt gctacgtatg gctcttcgcg caaacatcga gcttagtttc
1681 agatttgatc aaatgcagag cgtgatcgaa ggaagccgcg ttgaacttca acaggtacta
1741 ataacatct gcaagaatgc tcccaagcc atgacgcaa atggtcaaat cgcacatcct
1801 atcagccaag cttttttacc agttaagaaa attctggcgc atggttttat gccocctggc
1861 gactatgttc tctatctat tagcgaacat ggtggaggca tcccgaggc tgttttacc
1921 cacatttttg aacctctct taagcagca gctgcacagc gttgaaacgg tctcggcctt
1981 gctctgtgce atggtcatt cagcgcgctt gcgggttaca tcgacgttag ttoactgtt
2041 gggcatggga cgcgcttga catttatctc cctcgtctt ctaaagaaec cgtaaatcna
2101 gacagttttt bccggccgaa taaggccacc cgtgaaacgc gggagattgt ggcactgtt
2161 gacccgatg acctcctgag gtagcgtat gaagacaaga tgcgcgctc aggatgatg
2221 ccggtcgggt ttctgactct taatgaattt cgcgatggga tttcaaaagg caatgaagcc
2281 gatctggtca tggctcagca agcgtctctt cctgaagatc aaagtcttaa tctcgtggat
2341 ttagtctca agacgcctc catcatcatt ggcggaatg atctcaaat gccoccttca
```



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

What is an **Accession Number**?

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

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

X02775	GenBank genomic DNA sequence	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>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

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

RefSeq identifiers include the following formats:

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

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

RefSeq

two-component VirA-like sensor kinase

NCBI Reference Sequences (RefSeq)

Genome Annotation

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

Reference assembly

Genomic

- NC_003065.3**
Range: 180831..183332
Download: [GenBank](#), [FASTA](#), [Sequence Viewer \(Graphics\)](#)

mRNA and Protein(s)

- NP_396486.1 two component sensor kinase [Agrobacterium tumefaciens str. C58]**
UniProtKB/Swiss-Prot: [P18540](#)
Conserved Domains (3) [summary](#)

cd00075 Location:580 – 694 Blast Score: 202	HATPase_c; Histidine kinase-like ATPases; This family includes several ATP-binding proteins for example: histidine kinase, DNA gyrase B, topoisomerases, heat shock protein HSP90, phytochrome-like ATPases and DNA mismatch repair proteins
cd00082 Location:466 – 530 Blast Score: 144	HisKA; Histidine Kinase A (dimerization/phosphoacceptor) domain; Histidine Kinase A dimers are formed through parallel association of 2 domains creating 4-helix bundles; usually these domains contain a conserved His residue and are activated via ...
PRK13837 Location:14 – 833 Blast Score: 2944	PRK13837; two-component VirA-like sensor kinase; Provisional

Related Sequences

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

<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

Primary Databases

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, and a green arrow points to its start. A detailed popup window provides the following information:

- NP_059797.1**
- NP_059797.1: two-component VirA-like sensor kinase
- total range: NC_002377.1 (145,694..148,183)
- total length: 2,490
- strand: plus
- protein product length: 829
- Links & Tools**
- GenBank View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- FASTA View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- BLAST Genomic: [NC_002377.1 \(145,694..148,183\)](#)
- Graphical View: [NP_059797.1](#)
- BLAST Protein: [NP_059797.1](#)
- BLINK Results: [NP_059797.1](#)

Below the popup, there are sections for **Bibliography** and **Related articles in PubMed**. The browser's taskbar at the bottom shows various open applications like Firefox, Windows Media Center, and EndNote.

Primary Databases

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
ATGAACGGAAGATATTCACCGACGGCGCAGGATTTAAGACAGGCGCGAAGCCTTGGTCTATATTGGCCC
TTATCGTTGCTGCAATGATTTTCGCGTTTCATGGCGGTTGCGTCTGGCAGGACAAATGCGACTACCCAGGC
AATCTCAGCCAACACGATCGATTAACCGCGACAGCCCTCACTGACAGCGGATGACTCCGCGCTCAC
ACGGCACCGTGGCGAATACCGCCCATTTATCTCCAGGCTGGAGCTCTGCGGAAGAAATCGAAAGATT
TGAAGCAATTTAGCAATCTCATATTTGAAGTGAAGCAATGCTGCTCAACTGCTACGCGAGCTAGA
AGTGTCTTAAATTCGGCTGACGCGCGGCTCGCCGCTTTGGTGGCAAAATGTACGCTCAAAAGATTG
CTGGCCAGTTTCACTCGTCTTTGAGCAGTCTCCGGAAGGCTCAACCGATCAGACTTTAGAAAAAC
CAACAGAATTGGTAGCATGATGCTCCAATTTCTTCGGCAACCAAGCCGGCTATTTCAATTCGAGATCAG
CCTTGAAGTGAAGAGGCTCCAAAAACAACCGCGCTTGTATGAAGTCCCGTGGCATACTTGACCTGAA
GGTCCCATTTATCTTATCGCTTTTGGCCAGGTTGAAAGATCTGGTGAACATGATTCAGACGCTGACACCC
CAGAAATTCGGGATGCTGACGCGGAGTGTGGAGGTCTATAGCTTGAATAATGTAGAGGAGCGGAG
CGCAGTATCTTTCTGGTCCGCTTCAGTGGGTCTTGGCTCTACATCATCACTTGTCTATAGGCTA
CGCAAAAAACCGATTGGTTAGCGCGGCTTAGATTACGAAGAGCTAATCAAGAGATCGGAGTAGTGT
TTGAAGTGAAGCGGCCACCACTGCTCGCGCAAGCTGCATTCGTATTATTCAGCGCTTTTGGATGC
CGATACGTCGCGCTTAGCTTAGTGGACCATGACCGTAGAGGGCTGTGCAAAACATTCGTTGCGAAACAC
CCAAAACTGTGGGACGACAGCGTGTACGGCAATAGTCTCTGTACCAAGCGGACGACGCGGCGA
CGGTATTCGCAATCATGCTGCAAAAAAATCGTACATTTGCCTCTCGAAATCCAGGCTCTCTCGATCT
ACTGGCTCAAAATCCACAGATAAACTAATGCGGCTTTGTTCACTGGTATCCAAAGCTATCGCCCTCGA
CCTTGCCAGGCGAAATTCAGCTTCTTGAATCGCCACCGCTGCTCTGCTACTATATCGATGTTGCGG
GTAAGCAGACCGAATGCGACGTTTGGCCAGACGATTGGAGCATGCGCAACGCTTGAGGCACTTGGTAC
ACTTCCGCGGGAATAGCACATGAATTAATAACATTTTGGGCTCAATCCTCGGGCAGCAGAAATAGCA
CAAACTCGGTCTCGAACATCTGTACCCGAAATATATGACTATATCATTTCTGTCAGGCGCAGAG
CCATGCTCATATCGATCAGATCTTGAAGCTGAGCGGAAACAGGAGCGCATGATCAAGCCATTTAGTGT
CTCAGAGTGTGACCGAAATCGTCCCTTGTCTAGCTATGGCTTCCGCAAAACATCGAGCTTAGTTTC
AGATTTGATCAAAATGACAGCGGTGATCGAAGGAAGCCCGCTTGAACCTCAACAGGTAATTAACATCT
GCAAGATGCTTCCAGCCATGACTGCAATGGTCAATCGACATCATCATAGCCAAAGCTTTTTTACC
AGTTAAGAAATTTCTGGCGCATGGTGTATGCCACCTGGCGACTATGTTCTCCTATCTATTAGCGCAAT
GGTGGAGGCAATCCCGAGGCTGTGTACCCACATTTTGAACCTTTTACGACACGAGCTCGCAACG
GTGGAACGGGCTCTGGCCCTGCTCTGTGTCATGTTGATATCAGCGGCTTTCGGGTTACATCGAGCTTAG
TTCAACTGTTGGCATGGACGCGCTTGCATTTATCTCCCTCGCTTCTAAGGAACCCGTAATTCGA
GACAGTTTTTTCGGCCGCAATAAGGCACCGCTGGAAACGGGAGATTGTGGCACTTTTGGAGCCGATG
ACCTCCTGGGGAGGCGTATGAAACAAGATCGCCGCTTAGGATATGAGCGGTCGGTTTTTCTGATCCTT
TAATGAAATTCGCGATTGGATTTCAAAAGCAATGAAGCCGATCTGGTCAATGTTGCAACCAAGCGCTCT
CCTGAAGATCAAACTCCTAATTCGGTGGATTTAGTGTCAAGACCGGCTCCATCATCATTTGGCGAAATG
ATCTCAAAATGACCCCTTCAAGGGAGGATGTGACCGGAGCTTTATCTCCGAAAGCGGATATCGTCCAG
AATATGGCGCATGCAATCTCAACAAATCAAGACGATG
```

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

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Secondary Databases

- Databases of **functional** or **structural motifs**, acquired by primary data (sequences) **comparison**
- 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 Bolivia 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
<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>MMVKVTKLYASPTVTPCVLAPLVVPECTWISNMTTTE DLVKEVASFTEDLRLSLVSEIENIGKPTVAKTHLSTGLA RVIDEYITNNDTQPTFIQTQIALPLLFVAYSTILQVQVSY ISRDGIMPSYIARNTSVAVFASSSSNSRGGDTTYTQTV DQLTGRLRNGNSTRSQSLDVTHTWQQAQSHNYTTPVGT ELGGEDMETLIQSVVSLYSRGLVSLGFPFRTITVNLGL NLHRELIYMTEDVLYVRESLNDSPFISGSIQFGRRE NSLWQCPENCSSSGYEVKRLRYQAPCSYIRVSGVPL</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 NEWT 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 <p>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"/> greedy, overlaps, no includes (for patterns, see help) randomize databases: <input type="text"/> no (to test a pattern, see help)</p>

Secondary Databases

- Databases of **functional** or **structural motifs**, acquired by **primary data** (sequences) **comparison**
- PROSITE, <http://www.expasy.org/prosite/>

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

571 - 585 nkeesstYeteians

>[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 SsR
164 - 166 TgR
171 - 173 SsK
219 - 221 SsK
369 - 371 TrR
460 - 462 SgK
513 - 515 SgR
585 - 587 SiR
602 - 604 TgK
652 - 654 TgK
716 - 718 SpR
726 - 728 SpK
747 - 749 TeK
794 - 796 SsR
854 - 856 SsK
864 - 866 SsR
868 - 870 SsR
921 - 923 SpK
957 - 959 SvR
960 - 962 TgR
974 - 976 TsK
997 - 999 SsK
1002 - 1004 TgK
1018 - 1020 SgK
1031 - 1033 TgR
1119 - 1121 SsR

Secondary Databases

- Databases of **functional** or **structural motifs**, acquired by primary data (sequences) **comparison**
- PROSITE, <http://www.expasy.org/prosite/>

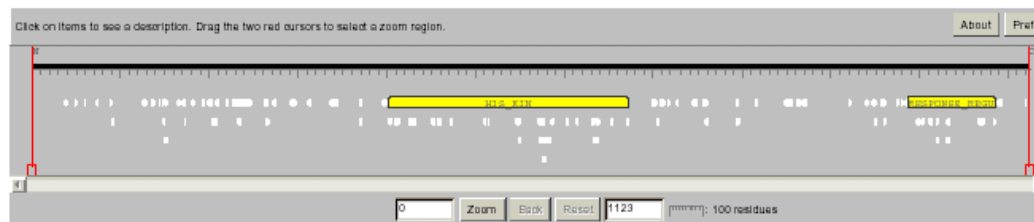
>[PDOC50109 PS50109 HIS_KIN](#) Histidine kinase domain [profile].

```
402 - 671 NASHDIRGALAGMEGLIDICRDGVKPGSDVDTTINQVMVCAKDLVALLNSVLEMSKIESG
KMQLVRHDFNLSKLLLEDVIDFHPVAMKKGVVLDPHDgavEKPSNVRGDSGRLKQILN
NLVSNARVFTVD--GHIAVRAWAQrpgensavvlasyppgvskfvkcmfcnkkeaatye
teianairnnaTMEFVFEVDITGKGIHMEMRKSVPENYVQVREtAQGHQGTGLGLGIVQ
SLVRLMG3EIRITDKAMGeKGTCPQPNVLLTT
```

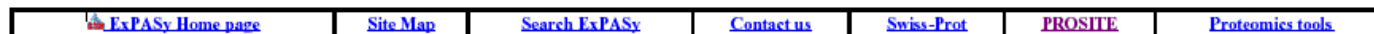
>[PDOC50110 PS50110 RESPONSE_REGULATORY](#) Response regulatory domain [profile].

```
987 - 1085 RVLVVDNPFISRRKVTGKLLKMGVSeVEQCDSGKEALRLVTEGLtqreeggvdklpFDY
IFMDQMPEMDGYRATREIRkvekSYGVRTPITAVSGHD-----
```

Graphical summary of hits (*java applet*)



98 hits with 12 PROSITE entries



Secondary Databases

- Databases of **functional** or **structural motifs**, acquired by primary data (sequences) comparison
- 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:

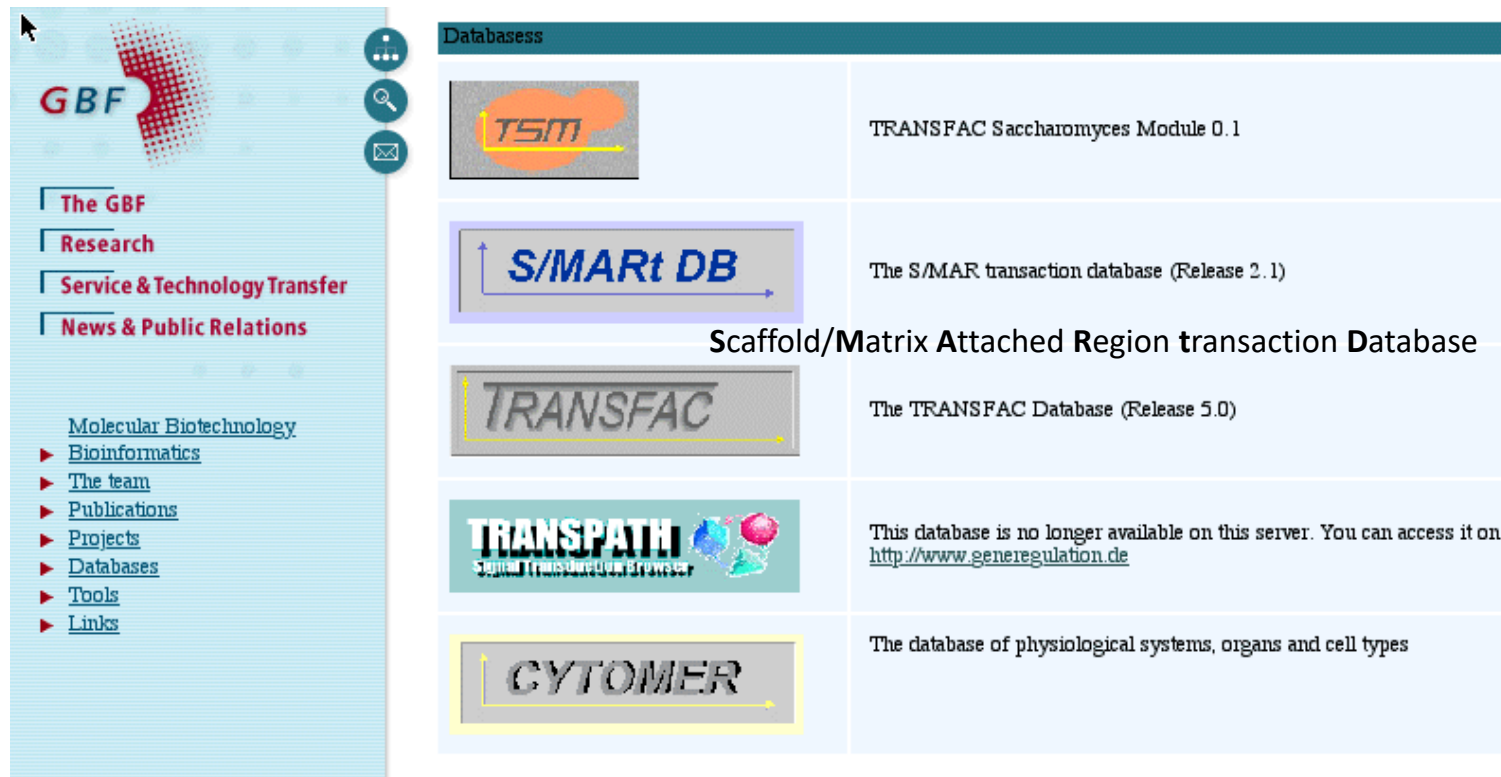
- [By accession number](#)
- [By PRINTS code](#)
- [By database code](#)
- [By text](#)
- [By sequence](#)
- [By title](#)
- [By number of motifs](#)
- [By author](#)
- [By query language](#)

PRINTS search:





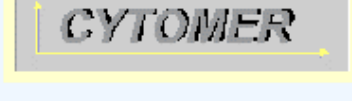
- [Search PRINTS with NEW FingerPRINTScan](#)
- [FPScan](#)
- [GRAPHScan](#)
- [MULScan](#)
- FingerPRINTScan binaries and source are available: contact.scordis@bioinf.man.ac.uk

Secondary Databases

- 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


Structural Databases

- 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.

[RCSB Home](#) [Contact Us](#) [Help](#)

[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](#)

Structural Databases

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

Structure Explorer - 1P5Y

RCSB
PDB
PROTEIN DATA BANK

Structure Explorer - 1P5Y

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



[View Structure](#)

[Summary Information](#)

[View Structure](#)

[Download/Display File](#)

[Structural Neighbors](#)

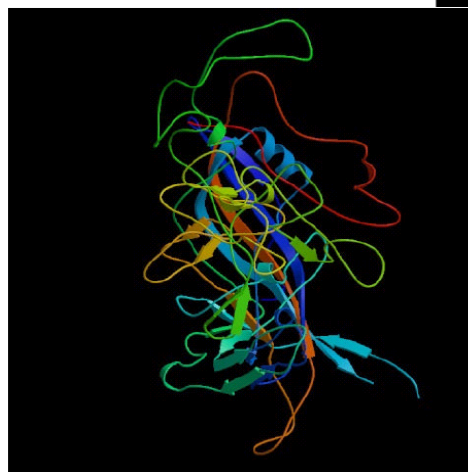
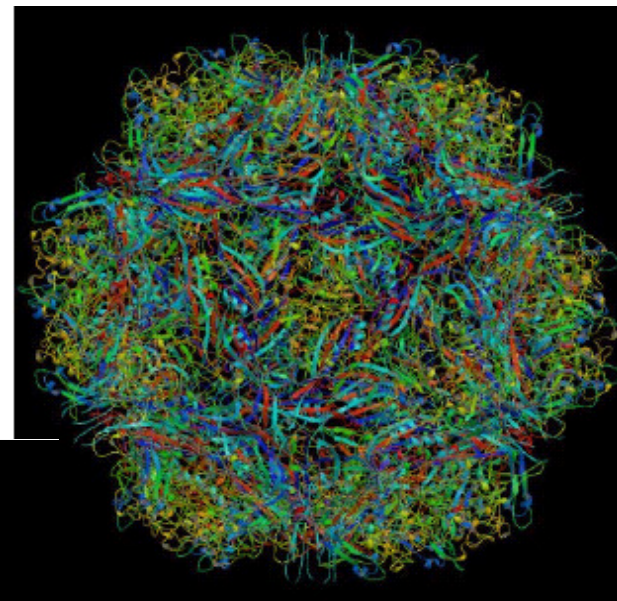
[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

Explore

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

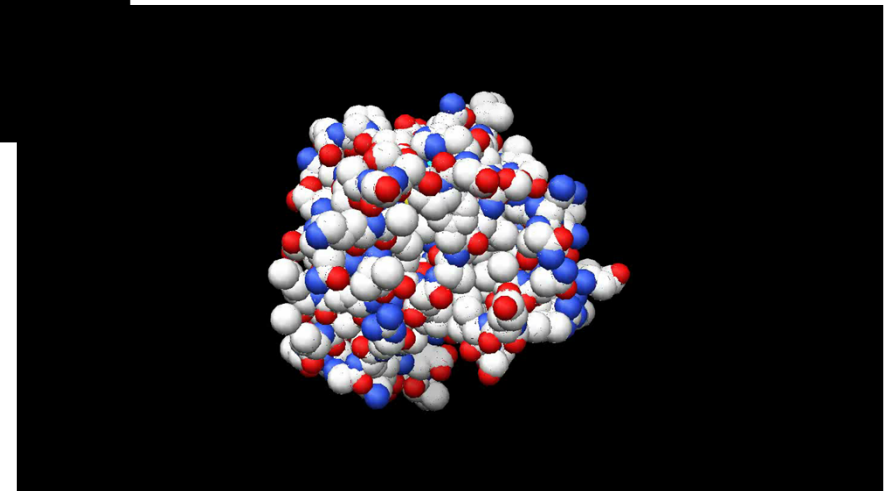


<http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=;pid=173561064349344&bio=1&opt=show&size=500>

12/29/2003

Structural Databases

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



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

Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre of „on-line“ Resources
 - PRIMARY, SECONDARY And STRUCURAL Databases
 - GENOME Resources



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genome Resources

- 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 [CBSE](#))

Genome Resources

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

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

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

Scale: chr11: [5,246,000 5,246,900 5,247,800 5,247,700 5,247,600 5,247,500 5,247,400 5,247,300 5,247,200 5,247,100 5,247,000 5,246,900 5,246,800 5,246,700 5,246,600 5,246,500]

RefSeq Genes

Human mRNAs

ESTs

HDK27hc Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

Digital DNase Hypersensitivity Clusters from ENCODE

Transcription Factor ChIP-seq from ENCODE

Phylo-P

Conservation

Multiple Alignments of 46 Vertebrates

Repeating Elements by RepeatMasker

track search | default tracks | default order | hide all | add custom tracks | track hubs | configure | reverse | resize | refresh

collapse all | expand all

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

Mapping and Sequencing Tracks					
Base Position	Chromosome Band	STS Markers	FISH Clones	Recomb Rate	deCODE Recomb
dense	hide	hide	hide	hide	hide
ENCODE Pilot	Map Contigs	Assembly	GRC Map Contigs	Gap	Publications
hide	hide	hide	hide	hide	hide
BAC End Pairs	Fosmid End Pairs	GC Percent	GRC Patch Release	Hq18 Diff	GRC Incident
hide	hide	hide	hide	hide	hide
Hi Seq Depth	Wiki Track	BU_ORChID	Mapability	Short Match	Restr Enzymes
hide	hide	hide	hide	hide	hide

Phenotype and Disease Associations					
GAD View	DECIPHER	OMIM AV SNPs	OMIM Genes	OMIM Pheno Loci	COSMIC
hide	hide	hide	hide	hide	hide
GWAS Catalog	ISCA	RGD Human QTL	RGD Rat QTL	MGI Mouse QTL	GeneReviews
hide	hide	hide	hide	hide	hide

Genome Resources

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

The screenshot displays the UCSC Genome Browser interface for the Human Gene HBB (uc001mae.1). The page is titled "Human Gene HBB (uc001mae.1) Description and Page Index". It provides a detailed description of the gene, including its function in hemoglobin synthesis and its association with various diseases like sickle cell anemia and thalassemia. A table of "Sequence and Links to Tools and Databases" is highlighted with a green arrow, listing various resources such as Genomic Sequence, Gene Sorter, Ensembl, and UniProtKB. Below this, the "Comments and Description Text from UniProtKB" section is visible, providing further details on the protein's structure and function.

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

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

TISSUE SPECIFICITY: Red blood cells.

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

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

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

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

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

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

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

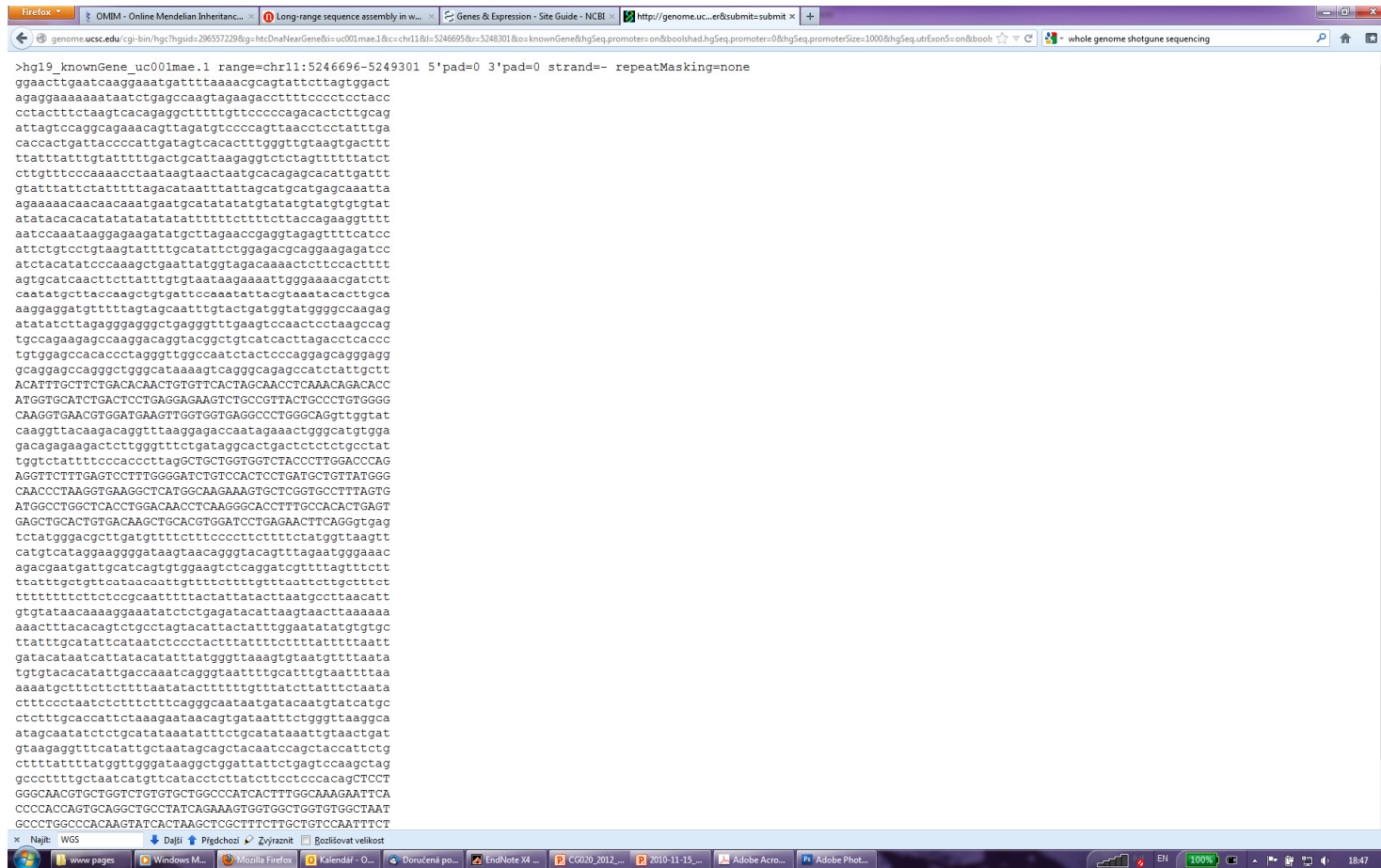
Genome Resources

- 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 URL in the address bar is 'genome.ucsc.edu/cgi-bin/hgGateway?hgsid=296557229&g=htcGeneInGenome&i=uc001.mae.1&cc=chr11&l=5246695&r=5248301&o=knownGene&table=knownGene'. The page content includes a navigation menu (Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, About Us, Help) and a main heading 'Genomic Sequence Near Gene'. Below this is a sub-heading 'Get Genomic Sequence Near Gene' and 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 includes 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 is also a 'Mask repeats:' section with radio buttons for 'to lower case' and 'to N'. A 'submit' button is located at the bottom of the options section. The browser's taskbar at the bottom shows various open applications and the system tray with the time 18:43.

Genome Resources

- 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

Genome Resources

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

The screenshot shows the TAIR website homepage. The browser window title is "TAIR - Home Page". The website features a search bar at the top right and a navigation menu with options like Home, Help, Contact, About Us, and Login/Register. Below the navigation, there are tabs for Search, Browse, Tools, Portals, Download, Submit, News, and ABRC Stocks. The main content area is titled "The Arabidopsis Information Resource" and includes a detailed description of the resource, a "Breaking News" section with links to subscribe to a news feed, follow on Twitter, and join a Facebook group, and a "2012 MASC Report Now Available" section. There is also a "New Protein Chip and Cell Cultures at ABRC" section and a "Share Your Education Resources" section. A large green banner at the bottom of the main content area promotes a new online submission form, with a "Click here" link and a "SUBMIT PAPER" button. The banner also mentions submitting molecular function, biological process, localization, or interacting partner of a favorite gene. The website footer includes "GO Annotations At TAIR" and "TAIR Introduction".

Genome Resources

- 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

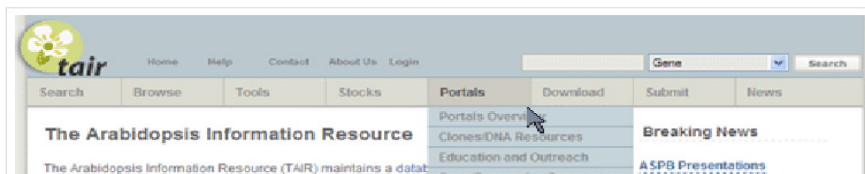
[October 15, 2006]
Search for [genes](#), [germplasms](#), and [polymorphisms](#) using associated phenotype, and see improved phenotype data display in results and detail pages.

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.



Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY And STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homology Searching

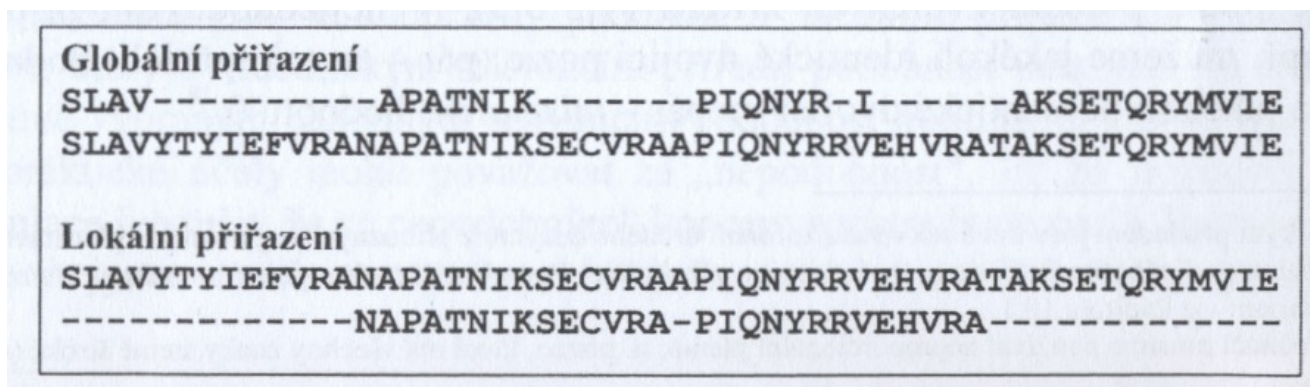


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

□ Global versus Local alignment

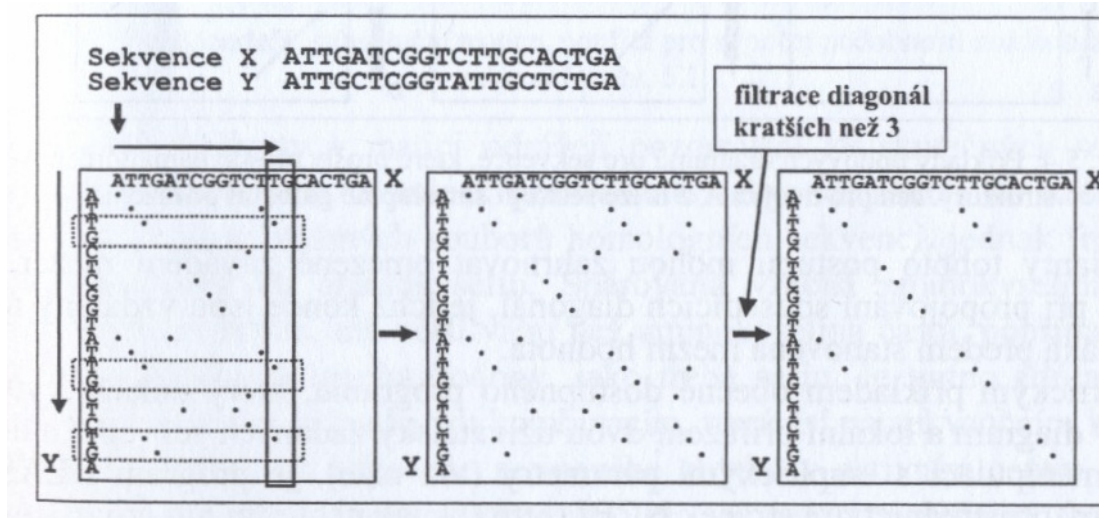


Cvrčková, Úvod do praktické bioinformatiky

- **Global Alignment:** only for sequences, which are **similar** and of a **similar length** (BUT can insert spaces into one or both sequences)
- **Global Alignment** is used mainly in case of **multiple alignment** (CLUSTALW, further in the presentation)
- **Local Alignment** provides identification and comparison even in case of alignment of **regions of sequences with high similarity**, e.g. even in case of **change of order of protein domains** during evolution

Analytical Tools

- Choosing the right type of alignment using [dotplot](#)

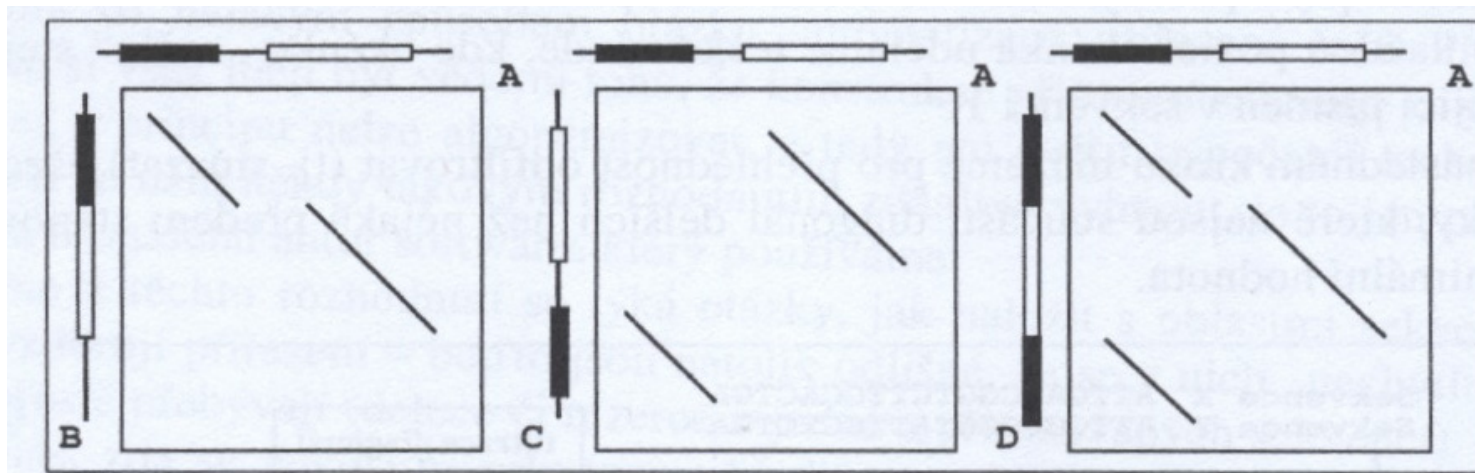


Cvrčková, Úvod do praktické bioinformatiky

- Plotting the sequences against each other (x and y axis)
- Identification of identity in „dot“ of specific size (e.g. 2 bp)
- Filtering the diagonals of lengths lower than a threshold

Analytical Tools

- Examples of sequence alignment using dotplot



Cvrčková, Úvod do praktické bioinformatiky

- **Global Alignment:** possible **only** for **sequences A and B**
- The rest of the sequences underwent change of order of protein domains and therefore it is necessary to do a local alignment
- Dotplot can be obtained using **BLAST2** (see further in the presentation)

Analytical Tools

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

NCBI *nucleotide-nucleotide* **BLAST**

Nucleotide Protein Translations Retrieve results for an RID

[Search](#)

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

[Set subsequence](#) From: To:

[Choose database](#) nr

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

BLAST

Basic Local Alignment Search Tool

- Word size: 10-11 bp or 2-3 aa
 - Primary similarities (seed matches)
 - Expanding the homology regions to the left and to the right
- Scoring the homology with matrices PAM (Point Accepted Mutation) or BLOSUM (BLOcks Substitution Matrix)
- Showing the results

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

hodnota nepáru G-A

hodnota páru G-G

Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250

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

BLAST

Basic Local Alignment Search Tool



- „**expectancy value**“ provides the number of expected sequence number with the **same or higher similarity** when searching in the database **consisting of randomly assembled sequences**
- the results shows **fraction of identical** and in case of proteins also **similar sequence positions** and/or **inserted spaces**

Primary Databases

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 gene, NP_059797.1, is highlighted in red. A detailed view of this gene is shown in a pop-up window:

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**.

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
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

Specialized Versions

- Currently there exists a lot of specialized versions of BLAST
 - Searching according to source (organism) of sequences, e.g. known genomes of microorganisms
 - **BLASTP**
 - Given the **protein query**, it returns the most similar protein sequences from the **protein database**.
 - **BLASTN**
 - Given the **DNA query**, it returns the most similar DNA sequences from the **DNA database**.
 - Other variants, e.g. **MEGABLAST**, for identification of identical or **very similar sequences** (searches **long similar regions** of nucleotide sequences)
 - **BLASTX**
 - Compares the all possible **six-frame translation products** of a **nucleotide query sequence** (both strands) against a **protein sequence database**.



BLAST

Specialized Versions

- Currently there exists a lot of specialized versions of BLAST
 - **TBLASTN**
 - Compares a **protein query** against the **all six reading frames** of a **nucleotide sequence database**.
 - **TBLASTX**
 - **Translates** the **query nucleotide sequence** in **all six possible frames** and **compares** it against the **six-frame translations** of a **nucleotide sequence database**.

BLAST

Specialized Versions

- Currently there exist a lot of **specialized versions** of BLAST
 - **PSI-BLAST** (**P**osition-**S**pecific **I**terated **B**last)
 - **First step: standard BLAST**, during which PSI-BLAST identifies a **list of similar sequences** with **E value better than minimal value** (standard = 0,005)
 - For every alignment, PSI-BLAST creates so-called **PSSM** (**P**osition **S**pecific **S**ubstitution **M**atrix)
 - **PSSM** takes into account **relative frequency of specific aminoacid residue in a specific position** within sequences identified as similar in first step, which can mean functional conservation.



BLAST

Specialized Versions

- Currently there exists a lot of specialized versions of BLAST
 - **PHI-BLAST** (Pattern-Hit Initiated BLAST)
 - For identification of **specific sequence**, e.g. motif (pattern) in sequence of similar protein sequences
 - Sequence of motif must be inserted using **special syntax**:
 - [LVIMF] means either Leu, Val, Ile, Met or Phe
 - - is spacer (means nothing)
 - x(5) means 5 positions in which any residue is allowed
 - x(3, 5) means 3 to 5 positions where any residue is allowed

BLAST

Specialized Versions

□ Example of search by PHI-BLAST

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

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

Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY And STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...

Analytical Tools

- <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.

Analytical Tools

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

View
View Nucleic Sequence(s)

Format Case

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

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

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

>170248
GAGCTCCCTTGGGGGGCAAGGGCAAAAACTTTTGCTAAATGGAAAAATATTATACCAAGTGTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAAATTTGACTATTTTGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTC AAATGGTCCATTATCGGCCAAGTAGCTTTCTTTAATTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATATAATAATAAATTC AAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAAAAGACCGATCAAAATAAAGCCGCCATTAAAAAATGAATTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACCTTTCCAATACCTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTTATCTCTAATTTACATCTCAACTAATATTAAGAAATTAACAGGTA
CAGCAAATCATAAAATTTCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCATGCCATATTC ACTAAGGGGTCGTTTGGTAC AAGAAATAATAATAAATTTTCGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTTAATTATAAGTATTAGCTAATTT CAGAATAAAATTTTACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACC AAATGATCGGTTAGTCTTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAATTTTCATTAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTTGTCTTCAAAGTAGATTCTCATTCAATTGCTCTGGTGC AATAGCAAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA
GAGAATTTCAAATCTCTCAACTACTTTTAAGGGAATTC AAAATACGACC AATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTATTTTAAATTTGAATTGAAAAATTTAAATTTACTTTGATTTAATATAA

Analytical Tools

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

Regex pattern:

ctt. {1, 32}ctt

0 sequences were searched

1 match was found

Matches are indicated in blue

>170248

```
GAGCTCCCTTGGGGGGCAGGGCAAAACCTTTTGGCTAAATGGAAAAATATTATACCAAGTGTGTTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTC AAAATGGTCCATTATCGGCAAGTAGCTTTCTTTAAATATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAATAATAA CTTCAAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAAAAGACCAGATCAAAATAAAGCCGCCATTAAAAATAATGAATTTTAGGACTCTC
GATTTGGCAGGTAAGTGCCAAAACCTTTCCAATACTTTTGTGCAACTTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTA AACAGGTA
CAGCAAATCATAAAATTTTCTCTAAGAAAGACAATGAATCCGGTFACTGATTCATTGGCCTTTTAGAG
TCTGCATGCCATATTTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAAATTTTCGGGATAGAATTT
GAGATTGCATTTATCTTTGTTTTAATTATAAGTATTAGCTAATTTTACAATAAAATTTTACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACC AAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAAATTTTCATTAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAGATGTACCGTTAATAATAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTGTCTTCAAAGTAGATTTCTCATTCATTGCTCTGGTGC AATAGCAAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTTACAACTTCTATTGTATACTCAAAATGAAAGTTTTA
GAGAACTTTCAAATCTCTCAACTCTTTAAGGGAATTCAAAATACGACCAATATTTATTA CTTACTTAC
TTATAGTTAAATGATATGAATTTTTAATTTGAAATTTGAAAATATTAATTTACTTGTATTAATATAA
ACAATAGATATCGCTAAGTATTTACCACAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT
GATTAAGCAGCTATTCATCTGGTTTGTGCAGGATGAAAGAAAGTAACTAGCTATAATTTCTMMTGTAAAGT
```

Analytical Tools

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

Frame 1, 1 stop codon

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

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

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

Frame 2, 1 stop codon

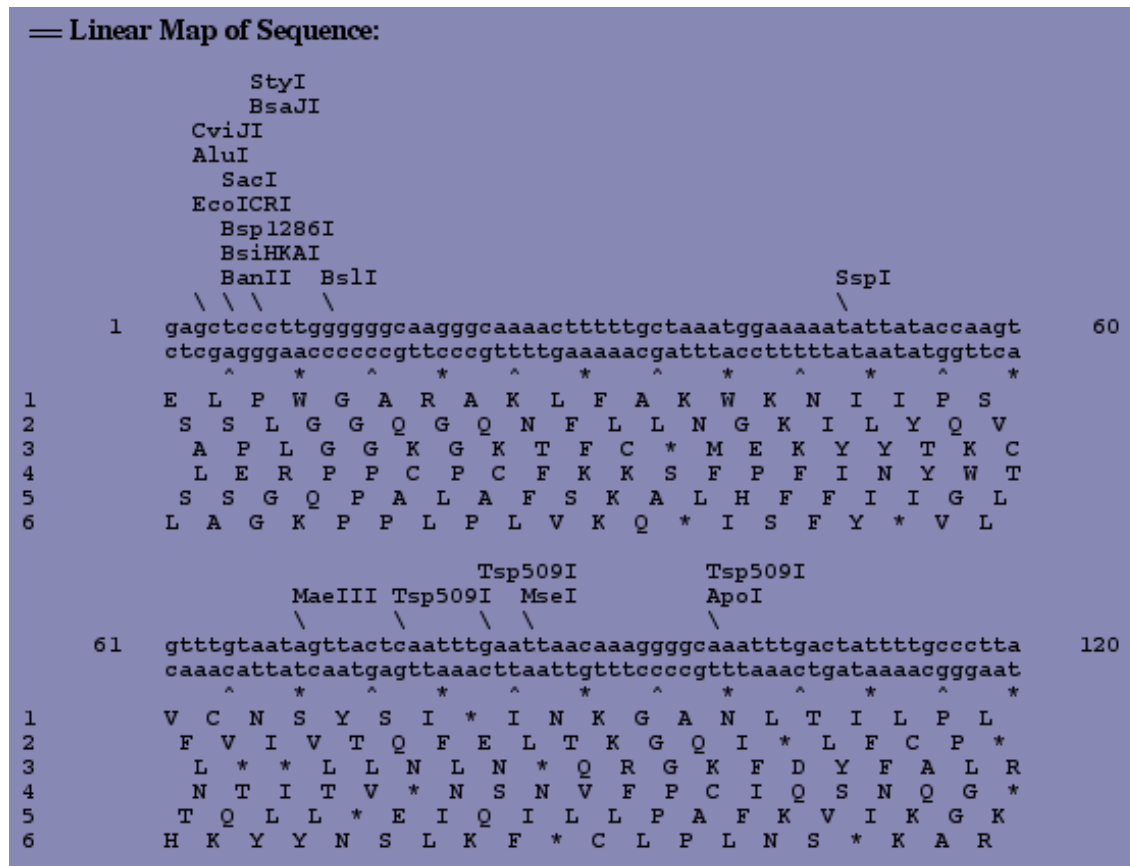
Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. Tran

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

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

Analytical Tools

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



Analytical Tools

- o <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](#)

```
.....
2560 GTTTGGTTGGTGTCTGGTTGAGAAGCTTGGAGTGGAGAGTCGGGTAGAGTGGGGTTTGGG 804855

          2850      2860      2870      2880      2890      2700
24 ..... A A A T G G C T . 170381
1 ..... 11321163
2430 ..... C A A G A A T T . 170248
1743 GAGTGAAATGATTGACAGAACTGCCAAAAACAAGCCAAAAATGGTAAAAAAA A A A A T T C 19686
2520 CATCGTCTATGTGGACTTCAATACTGTGAAGAGGTAGCCCAAGGACTGAGCGTTTGGCT 804855

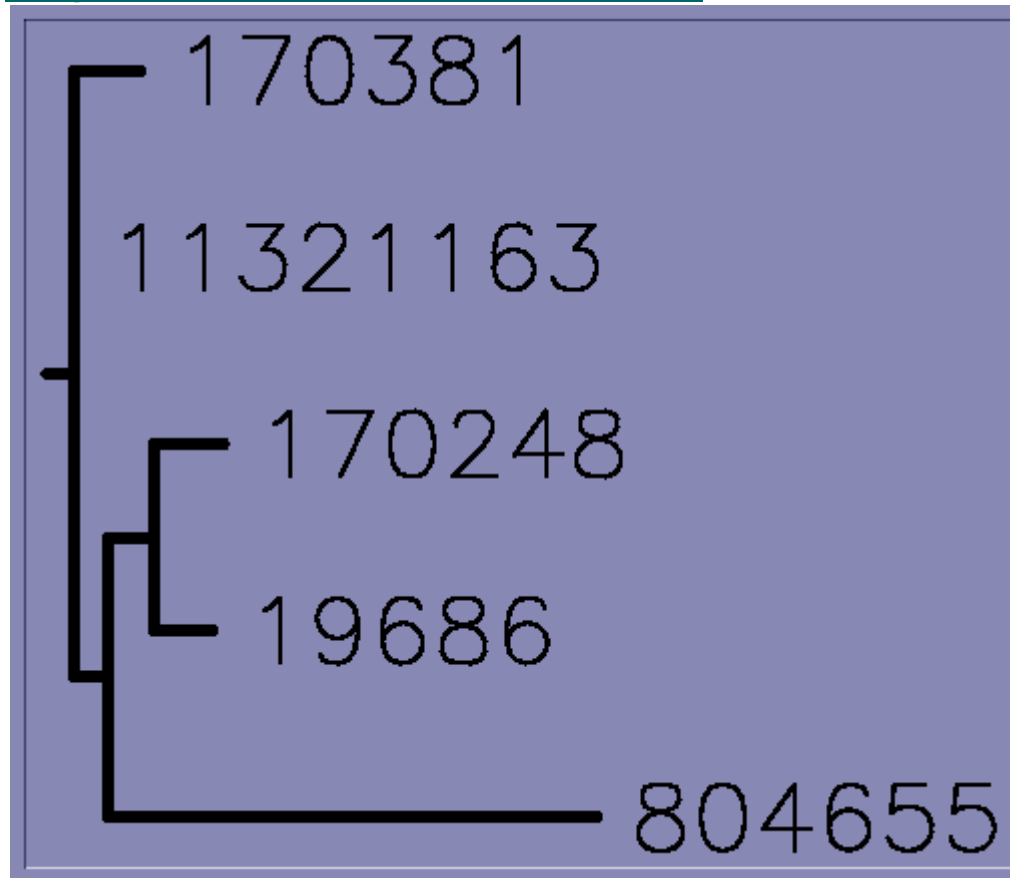
          2710      2720      2730      2740      2750      2760
32 ..... A T T A T T G T G C T T C T A C G A T T C T T G T G G C C A . C C A A C A T T C A G A T A G 170381
1 ..... 11321163
2438 ..... A G . A T A A T G A T T T A C T T T G T A A G A C T A A T T . C T A A T T C T T A T T G A G G 170248
1803 AGCATGTTTACA A T T G T T A T G T C C A A A C G C C G A C T G A C T A T T T T C A A T T C A A T T G A G G 19686
2580 CAAGAACATGCTCTCGAA A A G A A G A G C A G C T A G G A T C C A A A C A G G A T C G G G A G G A T C 804855

          2770      2780      2790      2800      2810      2820
79 A G A G G T F A A . . . . A T A G G T G T . . . . T T G T T A T G C A A T C A T G C C A A C A A C T T G C C A T G A C 170381
1 . . . . A T G G G T G T . . . . T T G G T A T G C A A T C A T G C C A A C A A C T T G C C A T T G A C 11321163
2484 A C C G G T F A A T C A A T A G G T G T . . . . T T G G T A T G C A A T C T A G C C A A C A A C T T G C C A A A T C 170248
1863 A C C G G T C A A T G C A T A G G T G T . . . . T T G G T A T G C A A T C T A G C C A A C A A C T T G C C A A A T C 19686
2740 A C T G G T T C A G C T T C A C A A A A A A A A G A T A T G T A A T C T T T T A T C T A G A A A C T G A G 804855

          2830      2840      2850      2860      2870      2880
132 A T T G T A A A C T T A T A C A G C . . . . T C T A C A A G T G G A G A A A C A T T A C A A G A C T G A G C T T T A T G A 170381
45 A T T C G A A C T T A T A C A G C . . . . T C T A C A A G T C A A G A A A C A T T C C A A G A T G A G C T T T A T G A 11321163
2540 A T T C G A A C T T A T A C A G C . . . . T C T A C A A G T C A A G A A A C A T T C C A A G A C T G A G C T T T A T G A 170248
1919 A T T C G A A C T T A T A C A G C . . . . T C T A C A A G T C A A G A A A C A T T C C A A G A C T G A G C T T T A T G A 19686
2800 A C T T A G C G C C T C T T G C C T A A A G A G C A C T G C C A A T A T G C C A G . . . . C C G A A A T T G C A G 804855
```


Analytical Tools

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



Analytical Tools

- 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



Analytical Tools

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



Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY And STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...
 - Other On-line Genome Tools

Other On-Line Genome Resources

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

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

Summary

Official Symbol PHACTR4 provided by HGNC
Official Full Name phosphatase and actin regulator 4 provided by HGNC
Primary source [HGNC:25793](#)
Locus tag RP11-442N24_A.1
See related [Ensembl:ENSG00000204138](#); [HPRD:07818](#); [MIM:608726](#)
Gene type protein coding
RefSeq status REVIEWED
Organism [Homo sapiens](#)
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Iliomniidae; Homo
Also known as FLJ13171; MGC20618; MGC34186; DKFZp686L07205; RP11-442N24__A.1
Summary This gene encodes a member of the phosphatase and actin regulator (PHACTR) family. Other PHACTR family members have been shown to inhibit protein phosphatase 1 (PP1) activity, and the homolog of this gene in the mouse has been shown to interact with actin and PP1. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]

Genomic context

Location : 1p35.3
Sequence : Chromosome 1; NC_000001.10 (288696093..28826881)

[See PHACTR4 in MapViewer](#)

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

Other On-Line Genome Resources

- Online Mendelian Inheritance in Man (OMIM)

The screenshot shows the OMIM website in a Firefox browser window. The address bar shows 'omim.org/#'. The page content includes the OMIM logo, the text 'Online Mendelian Inheritance in Man', and 'An Online Catalog of Human Genes and Genetic Disorders Updated 6 September 2012'. There is a search bar with the text 'Search OMIM' and a 'Search' button. Below the search bar, there are logos for the Institute of Genetic Medicine, Johns Hopkins Medicine, and the National Human Genome Research Institute. At the bottom of the page, there is 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.'



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Summary

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY and STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...
 - Other On-line Genome Tools

Discussion



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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