

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 genomic tools

Course Syllabus

- **Chapter 01**
 - Introduction to bioinformatics

- **Chapter 02**
 - Identification of genes

- **Chapter 03**
 - Reverse genetics approaches

- **Chapter 04**
 - Forward genetics approaches

Course Syllabus

- **Chapter 05**
 - Funcional genomics approaches

- **Chapter 06**
 - Protein-protein interactions and their analysis

- **Chapter 07**
 - Current DNA-sequencing methods

- **Chapter 08**
 - Structural genomics

Course Syllabus

- **Chapter 09**
 - Localization of genes and gene products in the cell

- **Chapter 10**
 - Genomics and systems biology

- **Chapter 11**
 - Practical aspects of functional genomics

- **Chapter 12**
 - Tools of systems biology
 - Model organisms, PCR and PCR primer design

Literature

- Literature sources for Chapter 01:
 - **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

Outline

- Syllabus of this course
- Definition of genomics

GENOMICS – What is it?

- *Sensu lato* (in the broad sense) – it is interested in **STRUCTURE and FUNCTION** of genomes
 - Condition: knowing 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 is a science discipline that is interested in the analysis of genomes. Genome of each organism is a complex of all genes of the respective organism. The genes could be located in cytoplasm (prokaryots) nucleus (in most euckaryotic organisms), mitochondria or chloroplasts (in plants).

The critical prerequisite of genomics is the knowledge of gene sequences.

Functional genomics is interested in function of individual genes.

GENOMICS – What is it?

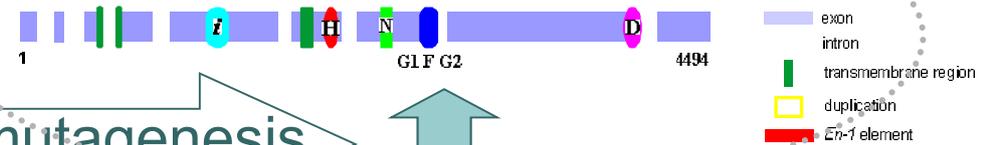
The role of BIOINFORMATICS in FUNCTIONAL GENOMICS

Forward („classical“) genetics approaches

Reverse genetics approaches

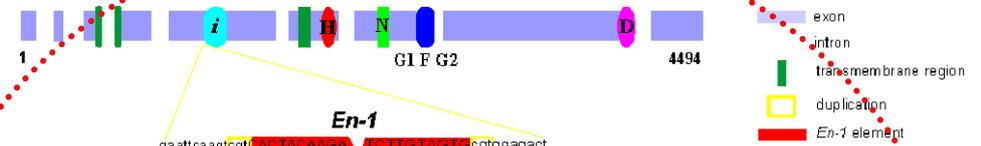
5'TTATATATATATATTAATAAAATAAAATAAAA
GAACAAAAAGAAAATAAAATA...3'

BIOINFORMATICS



Insertional mutagenesis

FUNCTIONAL GENOMICS



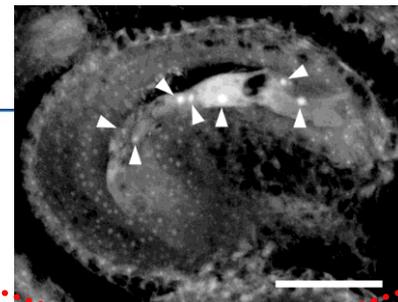
3

:

1



?



VOJE VZDĚLÁVÁNÍ

entace je spolufinancována
ropským sociálním fondem
a státním rozpočtem České republiky



EVROPSKÁ UNIE



MLÁDEŽE A TĚLOVÝCHOVY

pro konkurenceschopnost



UNIVERSITAS SILENSIS
MASARYKIANA BRUNENSIS

With the knowledge of gene sequences (or the knowledge of the gene files in the individual organisms, i.e. the knowledge of genomes), **Reverse Genetics** appears that allows study their function.

In comparison to "classical" or **Forward Genetics**, starting with the phenotype, the reverse genetics starts with the sequence identified as a gene in the sequenced genome. The gene identification using approaches of **Bioinformatics** will be described later (see Lesson 02).

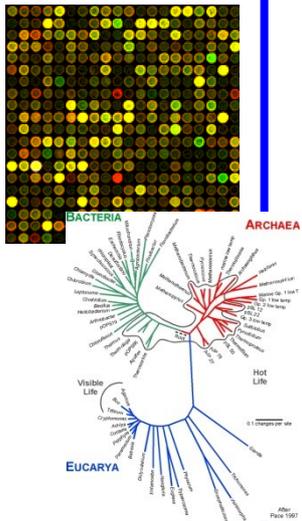
Reverse genetics uses a spectrum of approaches that will be described in the Lesson 03 that allow isolation of sequence-specific mutants and thus their phenotype analysis.

The necessity of having phenotype alterations in the forward genomics approach introduces important difference between those two approaches. Thus, the gene is no longer understood as a factor (*trait*) determining *phenotype*, but rather as a piece of DNA characterized by the unique *string of nucleotides*. i.e. **physical DNA molecule**.

Outline

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

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.

NIH WORKING DEFINITION OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY, July 17, 2000

The following working definition of bioinformatics and computational biology were developed by the BISTIC Definition Committee and released on July 17, 2000. The committee was chaired by Dr. Michael Huerta of the National Institute of Mental Health and consisted of the following members:

Bioinformatics Definition Committee BISTIC Members Expert Members

Michael Huerta (Chair) Gregory Downing

Florence Haseltine Belinda Seto

Yuan Liu

Preamble

Bioinformatics and computational biology are rooted in life sciences as well as computer and information sciences and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science. Bioinformatics and computational biology each maintain close interactions with life sciences to realize their full potential.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.

Definition

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: 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.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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.

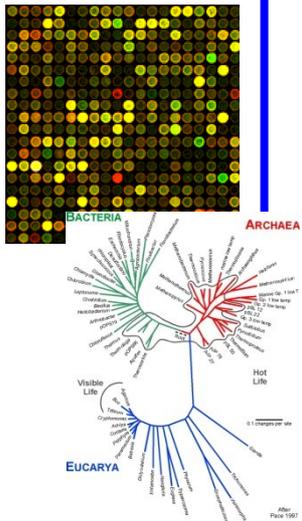
J. Pevsner,
<http://www.bioinfbook.org/index.php>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Bioinformatics



- **Bioinformatics in functional genomics**
 - **Processing and analysis of sequencing data**
 - Identification of reference sequences
 - Identification of genes
 - Identification of homologs, orthologs and paralogs
 - Correlation 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 regulation networks

Outline

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

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/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENIUSnet	Germany	http://genome.dkfz-heidelberg.de/biounit/
IMBB	Greece	http://www.imbb.forth.gr/
HEN	Hungary	http://www.hu.embnet.org/
INCBi	Ireland	http://acer.gen.tcd.ie/
INN	Israel	http://dapsas.weizmann.ac.il/bcd/inn.html
IEN-ADR	Italy	http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm
CAOS/CAMM	Netherlands	http://www.caos.kun.nl/
Bio	Norway	http://www.no.embnet.org/
IBB	Poland	http://www.ibb.waw.pl/
IGC	Portugal	http://www.igc.gulbenkian.pt/
GeneBee	Russia	http://www.genebee.msu.su/
CNB-CSIC	Spain	http://www.es.embnet.org/
BMC	Sweden	http://www.embnet.se/
SIB	Switzerland	http://www.ch.embnet.org/
SEQNET	UK	http://www.seqnet.dl.ac.uk/
EMBNet Specialist Nodes		
MIPS	Germany	http://www.mips.biochem.mpg.de/
ICGEB	Italy	http://www.icgeb.trieste.it/
Pharmacia Upjohn	Sweden	http://www.pnu.com/
F.Hoffmann-La Roche	Switzerland	http://www.roche.com/
EBI	UK	http://www.ebi.ac.uk/
HGMP-RC	UK	http://www.hgmp.mrc.ac.uk/
Sanger	UK	http://www.sanger.ac.uk/
UMBER	UK	http://www.bioinf.man.ac.uk/dbbrowser
EMBNet Associate Nodes		
IBBM	Argentina	http://sol.bio.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/

There are many of on-line resources that could be used.

Spectre of on-line resources

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

The screenshot displays the EBI Services website interface. The main heading is "Services" with sub-navigation for "Overview", "A to Z", "Service teams", and "Support". The primary section is "Bioinformatics services", which states: "We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically."

Services are categorized into several boxes:

- DNA & RNA**: genes, genomes & variation
- Gene expression**: RNA, protein & metabolite expression
- Proteins**: sequences, families & motifs
- Structures**: Molecular & cellular structures
- Systems**: reactions, interactions & pathways
- Chemical biology**: chemogenomics & metabolomics
- Ontologies**: taxonomies & controlled vocabularies
- Literature**: Scientific publications & patents
- Other software**: cross-domain tools & resources

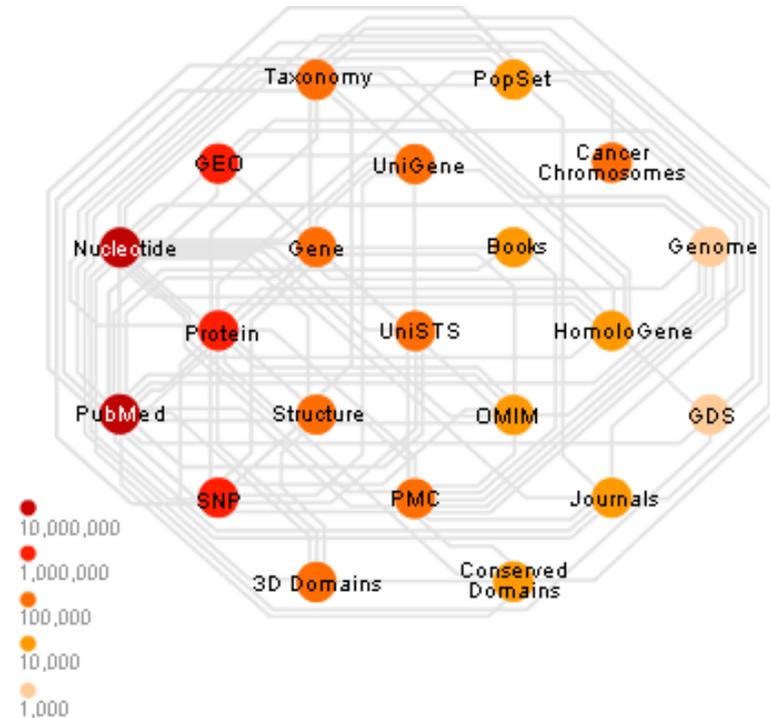
Additional sections include "Programmatic access" (EMBL-EBI web services allow you to query our large biological databases programmatically...), "Popular" (listing tools like Ensembl, UniProt, PDBe, ArrayExpress, ChEMBL, BLAST, Europe PMC, Reactome, Train online, and Support), "Bioinformatics training" (with a photo of people working), "Guide to resources" (with a photo of a woman), and "Service news" (with a photo of a book).

Spectre of on-line resources

□ NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' menus, and a 'My NCBI Sign In' link. Below this is a search bar with the text 'All Databases' and a 'Search' button. The main content area is divided into several sections:

- NCBI Home**: A blue header for the main navigation.
- Resource List (A-Z)**: A vertical list of categories including 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'.
- Welcome to NCBI**: A central text block stating the center's mission and providing links for 'About the NCBI', 'Mission', 'Organization', 'Research', and 'RSS Feeds'.
- Get Started**: A section with bullet points for 'Tools', 'Downloads', 'How-To's', and 'Submissions'.
- NCBI YouTube channel**: A red banner with a 'YouTube' logo and a 'GO' button, with a video player below it showing a progress bar from 1 to 8.
- Popular Resources**: A list of frequently accessed services like 'PubMed', 'Bookshelf', 'PubMed Central', 'BLAST', 'Nucleotide', 'Genome', 'SNP', 'Gene', 'Protein', and 'PubChem'.
- NCBI Announcer**: A section for news and updates, including 'New version of Genom', 'An integrated, downl', 'NCBI's July Newslett', 'Introduction to the 10', 'New Microbial BLAS', and 'Now easier to use an'.



Nowadays, the resources are interconnected and could be accessed via dedicated web pages.

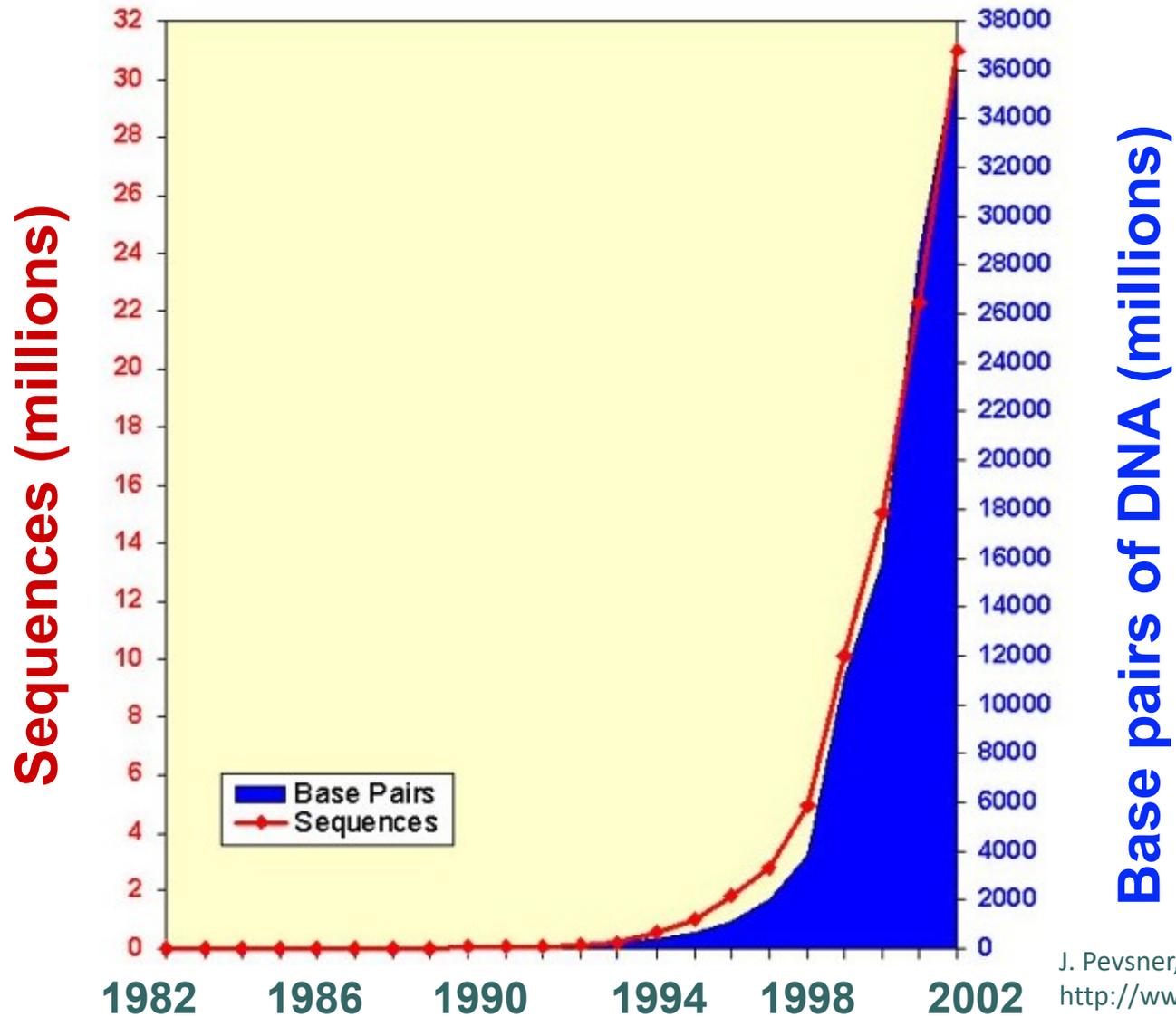
Outline

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

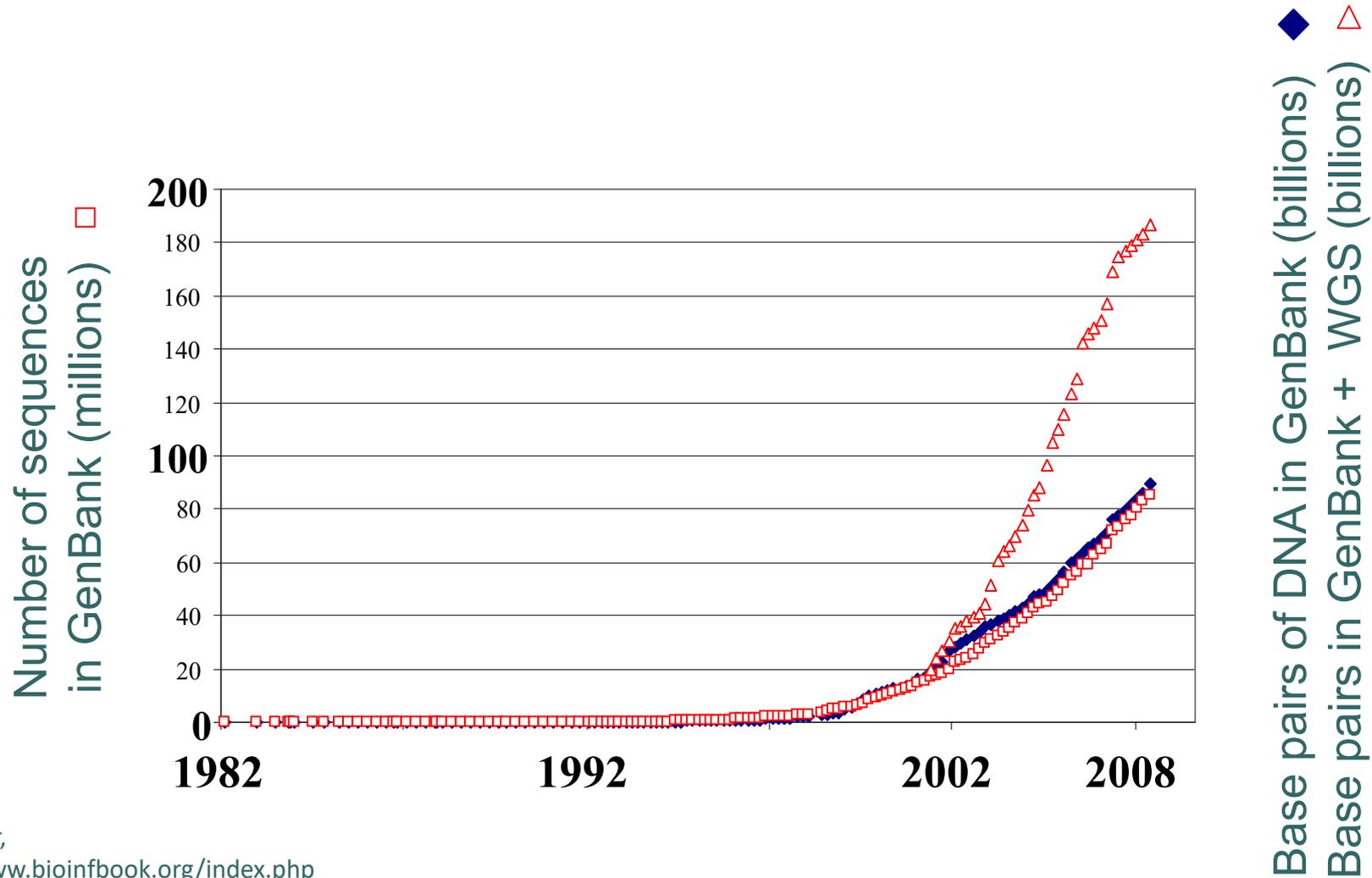
Primary databases

- They include sets of primary data – 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



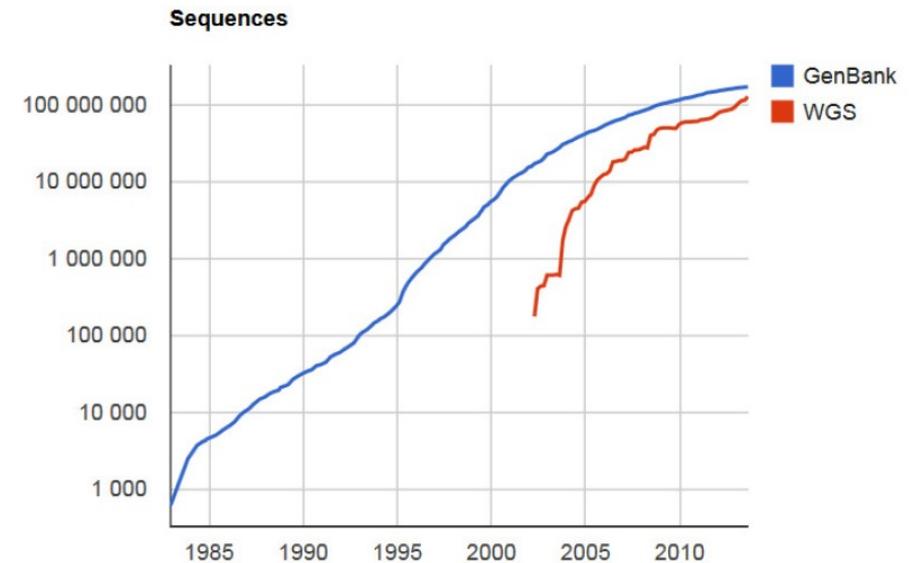
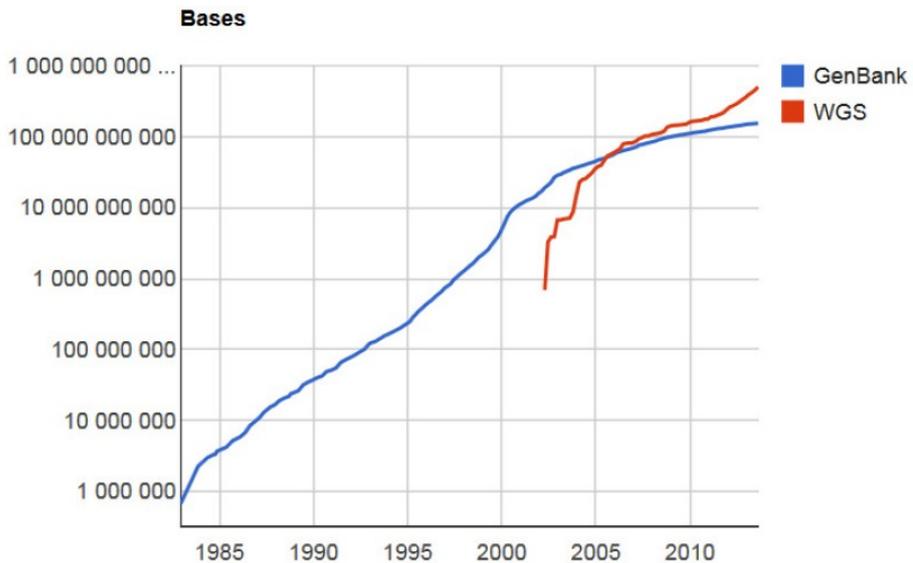
Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases



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

Growth of GenBank

Feb 15 2013



WGS

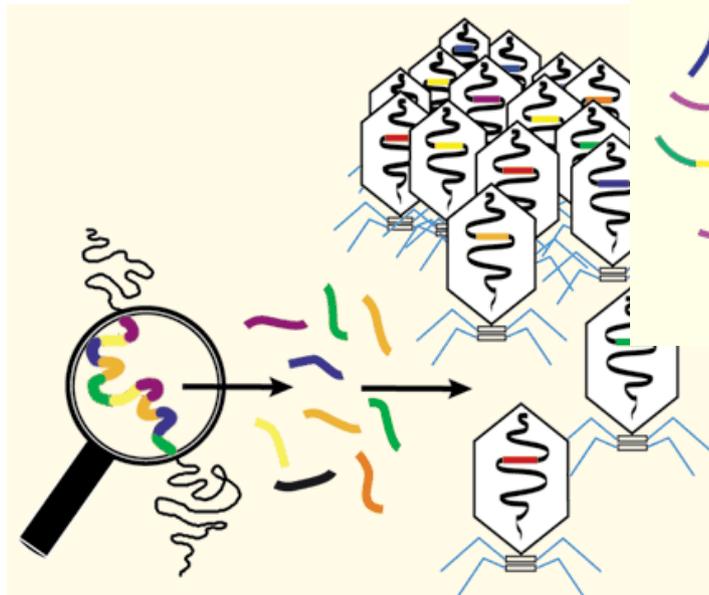
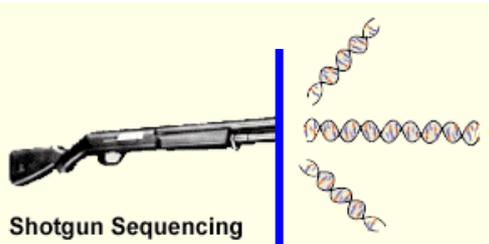


Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

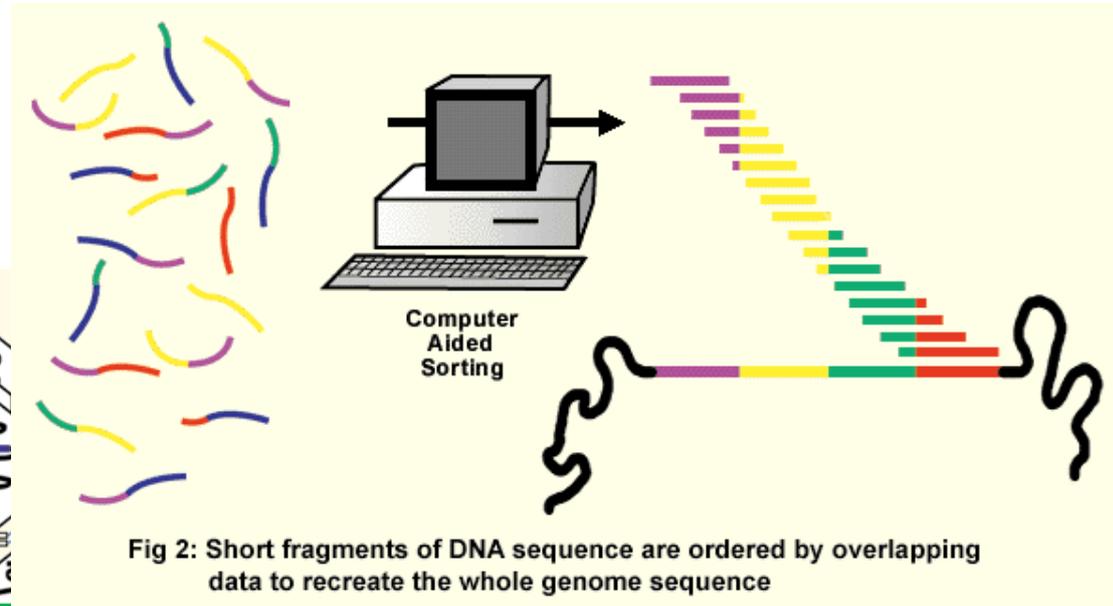


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

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

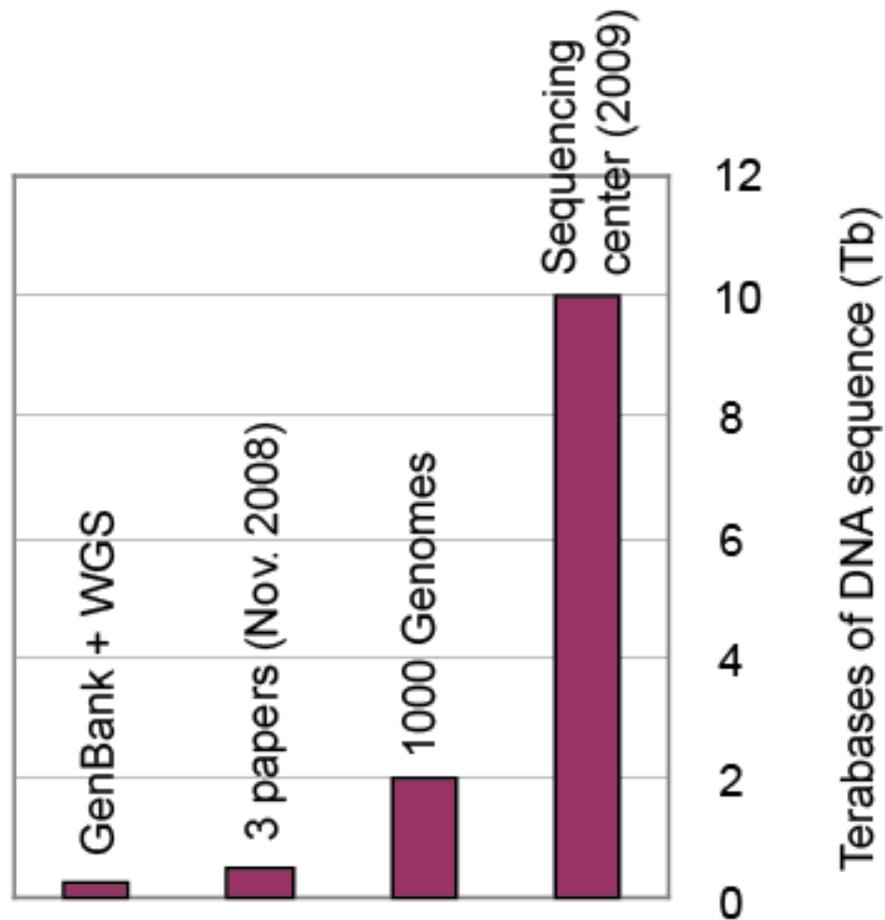
Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic DNA is usually first divided into its individual chromosomes. Each chromosome is then randomly broken into small strands of hundreds to several thousand base pairs, usually accomplished by mechanical shearing of the purified genetic material. Each of the short DNA pieces is then inserted into a DNA vector (a viral genome), resulting in a viral particle containing "cloned" genomic DNA (Fig. 1).

The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence.

Placing the genomic DNA within the viral genome allows bacteria infected with the virus to faithfully replicate the genomic DNA pieces. Additionally, since a little bit of known sequence is needed to start the sequencing reaction, the reaction can be primed off the known flanking viral DNA.

In order to read all the nucleotides of one organism, millions of individual clones are sequenced. The data is sorted by computer, which compares the sequences of all the small DNA pieces at once (in a "shotgun" approach) and places them in order by virtue of their overlapping sequences to generate the full-length sequence of the genome (Fig. 2). To statistically ensure that the whole genome sequence is acquired by this method, an amount of DNA equal to five to ten times the length of the genome must be sequenced. (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>

DDBJ/EMBL/GenBank accepts both complete and incomplete genomes. Whole Genome Shotgun (WGS) sequencing projects are incomplete genomes or incomplete chromosomes that are being sequenced by a whole genome shotgun strategy. WGS projects may be annotated, but annotation is not required.

The pieces of a WGS project are the contigs (overlapping reads), and they do not include any gaps. An [AGP file](#) can be submitted to indicate how the contig sequences are assembled together into scaffolds (contig sequences separated by gaps) and/or chromosomes. We must have the contig sequences without gaps as the basic units for all WGS projects.

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 (Expressed Sequence Tags)**
 - **HGTS (High Throughput Genome Sequencing)**
 - Results of sequencing projects without annotation
 - Reference sequences of annotated genomes
 - **TPAs (Third Party Annotation)**
 - sequences annotated by third party (by someone else, not the original authors)

Primary databases

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

The screenshot displays the NCBI website interface. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' dropdown menus, and 'My NCBI Sign In' links. Below this is a search bar labeled 'All Databases' with a 'Search' button. The main content area is divided into several sections:

- Left Sidebar:** A vertical menu with 'NCBI Home' at the top, followed by 'Resource List (A-Z)' and a list of categories including 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'.
- Center:** A 'Welcome to NCBI' section with a brief description of the center's mission and a list of links: 'About the NCBI', 'Mission', 'Organization', 'Research', and 'RSS Feeds'. Below this is a 'Get Started' section with a bulleted list:
 - Tools:** Analyze data using NCBI software
 - Downloads:** Get NCBI data or software
 - How-To's:** Learn how to accomplish specific tasks at NCBI
 - Submissions:** Submit data to GenBank or other NCBI databases
- Right Sidebar:** A 'Popular Resources' section listing various services: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Below this is an 'NCBI Announcements' section with a notice about a new version of GenBank and a link to an integrated download tool.
- Bottom Center:** A 'NCBI YouTube channel' banner featuring a 'GO' button and a YouTube logo, with text encouraging users to learn from video tutorials.

Primary databases

The screenshot displays the NCBI Gene database entry for the *virA* gene (NC_002377.1). The page is organized into several sections:

- Summary:**
 - Gene symbol:** *virA*
 - Gene description:** two-component VirA-like sensor kinase
 - Locus tag:** pTI_125
 - Gene type:** protein coding
 - RefSeq status:** PROVISIONAL
 - Organism:** *Agrobacterium tumefaciens* (old-name: *Agrobacterium tumefaciens*, gb-synonym: *Rhizobium radiobacter*)
 - Lineage:** Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiaceae; Rhizobium/Agrobacterium group; Agrobacterium; Agrobacterium tumefaciens complex
- Genomic context:**
 - Location:** plasmid: Ti
 - Sequence:** NC_002377.1 (145694..148183)
- Genomic regions, transcripts, and products:**
 - Genomic Sequence:** NC_002377
 - Sequence viewer:** Shows the genomic context of the *virA* gene (145,694-148,183 bp) with a yellow oval highlighting the gene's location. The viewer includes a scale bar and a search function.
- Related articles:**
 - Sequence analysis of the *virA* gene of *Agrobacterium tumefaciens* octopine Ti plasmid pTI15955. Schrammeyer B, et al. J Exp Bot. 2000 Jun. PMID 10948245.
 - The *virA* promoter is a host-range determinant in *Agrobacterium tumefaciens*. Turk SC, et al. Mol Microbiol. 1993 Mar. PMID 8469115.
 - 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.
 - 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](#)

The right sidebar contains various navigation and utility links:

- Bibliography:** General protein info, Reference sequences, Related sequences.
- Links:** BioProjects, Conserved Domains, Full text in PMC, Genome, Nucleotide, Protein, Protein Clusters, PubMed, RefSeq Proteins, Taxonomy.
- General information:** About Gene, FAQ, FTP site, Help, My NCBI help, NCBI Handbook, Statistics.
- Related sites:** BLAST, Genome, BioProject, Genomic Biology, GEO, HomoloGene, Map Viewer, OMIM, Probe, RefSeq, UniGene, UniSTS.
- Feedback:** Contact Help Desk, Submit Correction, Submit GeneRIF.

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](#)
 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

Primary databases

NCBI

Search Nucleotide for [] Go Clear

Accession number

NC_002377.1

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

DEFINITION Agrobacterium tumefaciens extrachrom plasmid Ti, complete sequence.

ACCESSION NC_002377 REGION: 148694..148813

VERSION NC_002377.1 GI:10955016

KEYWORDS

SOURCE Agrobacterium tumefaciens (Rhizobium radiobacter)

GeneBank identifier

Agrobacterium tumefaciens; Rhizobiales; Rhizobium; Agrobacterium.

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

TITLE Octopine-type Ti plasmid sequence

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 2490)

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

TITLE Direct Submission

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

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

FEATURES

Location/Qualifiers

source 1..2490

/organism="Agrobacterium tumefaciens"

/mol_type="genomic DNA"

/db_xref="taxon:358"

/plasmid="Ti"

/note="extrachromosomal octopine-type"

gene 1..2490

/gene="virA"

/db_xref="GeneID:1224316"

CDS 1..2490

/gene="virA"

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

/codon_start=1

/transl_table=11

/product="virA"

/protein_id="NP_059797.1"

/db_xref="GI:10955141"

Primary databases

```

/translation="MNGRYSPTRODFKTKGAKPWSILALIYAAMIYAFMAVASWQDNMT
TQAILSQLRSINADSASLQRDVLRHTTQVANYRPIISRLGALRNKLEDLKLFRQSH
IVSEBNAQQLRQLEVSLSADAAVAAPGQNVRLQDSLASPTRALSSSLFGKASTDQT
LEKPTELASMMQLQFLRQSPAISFPBISLELELELQKQRLDEAPVIRILAREGPIILSLL
PQVKDLVNMQISTDTAEIEMLQRCLEVYSLKNVEERSARIPLSASVGLCLYIITL
VYLRKKTDWLARRLDYELIKELGVCFBGRATTSQAALRIIQRPFDADTCALAL
VDHRRWAVETFGAKHFKPVWDSVLRRIVSRKADERATVFRILSSKKIVHLFLHIP
GLSILLAHKSTDKLIAVCSLGYQSYRPFPCQGETQLLELATACLCHYIDVRRKQTECD
VLARREHAQRLAVGTLAGGIAHFFNNILGSLGHAEALQNSVSRTEVTRRYIDYII
SSGDRAMLIIDQILTLRKRQEMIKPPSVSELVTEIAPLRLMALPPNIELESPFDQMC
SVI EGSPLRLQQLVINICKNASQAMTANQIDIIISQAPLPVKKILAHGVMPFGDYVL
LSISDNQGGIPEAVLPHIPEPFFTTARNGGTGLGLASVHGHSAPAGYIDVSTVGH
GTRFDIYLPFSSKPEVNPDSPPFRNKAPEGRNGEIVALVPPDDLREAYRDKIAALGYE
PVGPRTFNKRIDWISKGNEDLVMVDQASLPEDQSPNSVDLVLKTAIIIGGNDLKM
LSREDVTADLYLFPKPISSRTMAHAHLTKIKT"

ORIGIN
1 atgaacggaa gatattcacc gaecggcgag gattttaaga caggcgcgaa gcccttggct
61 atattggccc ttatcgttgc tgcaatgatt ttocogtcca tggcgggttc gtccttggcag
121 gacaatgcca ctaccacgga aatcctcage caactacgat cगततााााा cगकागगगग
181 tcaactgcaag cगतगगतत cगगगगगग cगगगगगग cगगगगगग cगगगगगग
241 atctccaggg tgggagctct ggggaagaat ctgggaagatt tgaagcaatt atttagacaa
301 tctcatttg taagtggag caatgctgct caactgctac gccagctaga agtgtctcta
361 aatcgggctg acgcgggctg cgcgccttt ggtgogcaaa atgtacgctt gcaagattcg
421 ctggcgagtt tcaactcgtc ttgagcagat ctccagagaa aagcctcaac cगतकगगत
481 ttagaaaaac caacagaatt gगतगगत atgctccaat ttcttggca accaagcccg
541 gctatttcat togagatcag ccttgaacta gagaggtctc aaaaacaag cगतtctgat
601 gaagctcccg tgcgcaact tgcacgtgaa ggtcccatta tcttatcgtt ttgcccacag
661 gtgaaagatc tgggtaacat gattcagagc tctgacacog cagaanaatgc gगतगगत
721 cagcgcgagt gtttggaggt ctatagcttg aaaaatgtag aggagcgag cgcagctac
781 ttcttgggtt cगतtctcag ggttcttctg ctctacatca tcaacttagt ctataggcta
841 cगकागगगग cगतtggtt agcgcgggct ttगतताग aagagctaat caaagagatc
901 gगतगगत ttgaaagtga ggcggccacc cगतtctcog cगकागगग cगतtctgat
961 atcagcgctt tcttctgct cगतtaactgc cगतtctcog tगतgagca tgcagctaga
1021 tgggctgctg aaacattcgg tgcgaaacac ccaaacctgt tctgagagca cगतgctga
1081 cगकागगगग tctctcgtac caaagcgagc gaacggggca cगतtctcog cगतcगत
1141 tगकागगगग tगतcगतt gctctcगग atctcगतt cगतगतat cगतgctcag
1201 aaatccacag ataaactaat tgcggttgt tcaactgggtt accaagcta tgcgctcga
1261 ccttgcacag gogaatata gcttctttaa ctgcacacog cगतcगतt tगतगतatc
1321 gatgttcggc gtaagcagac cगाatgcgac gttttggca cगतगतgga gगतgcgca
1381 cगतtgcagc cगतtggatc acttgcggc gगतगतcag atगाattaa taacatttg
1441 gctcgaatcc tgggcaogc agaatagca caaacctcgg tगतtcgaac atctgcacc
1501 cगागतata ttगतatata ctttctgca ggcgacagag cगतगतat tगतगतcag
1561 atctgacgc tगगcगगा acagggagcg atगतcaagc cगतtगतt cगतगतct
1621 gtgaccgaaa tगतcगतt gctcगतt gctctcगग caaacatga gगतगतt
1681 agatttgatc aaatcगगag cगतगतgaa gगाagccgc tगgaactta acगगतta
1741 attaacatct gcaagaatgc tcccaagcc atगacgca atगतcaaat cगतगतat
1801 atcagccaag cttttttacc agttaagaaa atctgगगc atगतगतt ggcacctgc
1861 gactatgttc tctatctat tagcgcaaat ggtgagggca tcccgaggc tगतगतacc
1921 cacatttttg aacctctct tacgacagca gctgcacag gगाagccgg tctgcgctt
1981 gcttctgctg atगतगतat cगगcगगgtt cगगगतta cगगगतt tcaactgtt
2041 ggcgatggga cगतcगतtga ctttatctc cctcगतt ctgaagacc cगगाatca
2101 gacagttttt cगगcगगca taagccacog cगतggaacog gगतगतt ggcctgtt
2161 gagccगतg actcगतg gगगगतat gaagacaaga tgcगतt agगतगतg
2221 cगगतगतt ttगतगतt taatgaatt cगतगतgga tttcaaaag caatgagcc
2281 gatctgctca tggtcgaca agcगतtctt cctgaagatc aaगतtca tctगतगत
2341 ttगतtca agacgcctc cगतcगतt ggcगाaatg atctcaaat gaccttca

```

What is an accession number?

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

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

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

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

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

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

NCBI's important RefSeq project: best representative sequences

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

RefSeq identifiers include the following formats:

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

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

RefSeq

two-component VirA-like sensor kinase

NCBI Reference Sequences (RefSeq)

Genome Annotation

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

Reference assembly

Genomic

1. **NC_003065.3**

Range 180831..183332

Download [GenBank](#), [FASTA](#), [Sequence Viewer \(Graphics\)](#)

mRNA and Protein(s)

1. **NP_396486.1 two component sensor kinase [Agrobacterium tumefaciens str. C58]**

UniProtKB/Swiss-Prot [P18540](#)

Conserved Domains (3) [summary](#)

cd00075	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	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	PRK13837; two-component VirA-like sensor kinase; Provisional

Location:580 – 694
Blast Score: 202

Location:466 – 530
Blast Score: 144

Location:14 – 833
Blast Score: 2944

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

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

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
ATGAACGGAAAGATATTCACCGACGCGCAGGATTTAAGACAGCGCGCAAGCCTTGGCTATATTTGGCC
TTATCGTTGCTGCAATGATTTTCGCGTTTCATGGCGTTTGGCTCCTGGCAGGACAAATGCGACTACCCAGGC
AATCCTCAGCCAACACGATCGATTACCGCCGACAGCCCTCACTGCAAGCCGATGACTCCGCGCTCAC
ACGGCACCGTGGCGAATACCGCCCATTTATCTCCAGGCTGGGAGCTCTGCGGAAGAAATCGAAAGATT
TGAAGCAATTTAGCAATCTCATATTTGTAAGTGAGAGCAATGCTGCTCAACTGCTACGCGCAGTAGA
AGTGTCTAAATTCGGCTGACGCGCGCTCGCCGCTTTGGTGGCAAAATGTACGCTCAAAAGATTG
CTGGCCAGTTTCACTCGTGTCTTGGAGCTTCCAGGAAAAGCCTCAACCGATCAGACTTTAGAAAAAC
CAACAGAATTGGCTAGCATGATGCTCCAATTTCTCGGCAACCAAGCCGGCTATTTCAATCGAGATCAG
CCTTGAAGTGAAGAGGCTCCAAAAACAACGCGCTTGTATGAAGCTCCCGTCCGCATCTTGACCGTAA
GGTCCCATTTATCGCTTTTGGCCAGGTAAGGATCTGGTGAACATGATTCAGACCTCTGACACCG
CAGAAATTCGGGATGCTGACGCGGAGTGTGGAGGCTATAGCTTGAATAATGAGAGGAGCGGAG
CGCACGTATCTTTCTGGTCCGCTTCAGTGGTCTTTGCCTCTACATCATCACTTAGTCTATAGGCTA
CGCAAAAAACCGATTGGTTAGCGCGGCTTTAGATTAGAAAGACTAATCAAGAGATCGGAGTATGTT
TTGAAGTGAGGCGGCCACCACTGCTCGCGCAAGCTGCATTTGATTTATTCAGCGCTTTTGGATGC
CGATACCTGCGCGTAGCTTAGTGGACCATGACCGTAGTGGGCTGTGAAAACATTCGTTGCGAAACAC
CCAAAACCTGTGGGACGACAGCGTCTACGCGAAATAGTCTCTGTAACAAAGCGGACGAAACGGGCGA
CGGATTCGCGATCATGCTGCAAAAAAATCGTACATTTGCCTCTCGAAATTCAGGCTCTCTGATACT
ACTGGCTCAAAATCCACAGATAAATTAATGCGGTTTGTTCCTGGGTTACCAAGCTATTCGCGCTCGA
CCTTGCCAGGCGAAATTCAGCTTCTTGAATCGCCACCGCTGCTCTGCTACTATATCGATGTTGGC
GTAAGCAGACCGAATGCGACGTTTGGCCAGACGATTGGAGCATGCGCAACGCTTGGAGCAGTTGGTAC
ACTTCCGCGCGAATAGCACATGAATTAATAACATTTTGGGCTCAATCCTCGGCAACGAGAAATAGCA
CAAACTCGGTCTCGAACATCTGTACCCGAAAGATATATGACTATATCATTTCTGTCAGGCGACAGAG
CCATGCTCATATCGATCAGATCTTGACGCTGAGCCGAAACAGGAGCGCATGATCAAGCCATTTAGTGT
CTCAGACTGTGACCGAAATCGCTCCCTTGTACGATGGCTTCCGCAAAACATCGAGCTTAGTTC
AGATTTGATCAAAATGCGAGCGTATCGAAGGAAGCCGCTTGAACCTCAACAGGTAATTAACATCT
GCAAGATGCTTCCCAAGCCATGACTGCAAAATGTTCAAAATCGACATCATCATAGCCAAAGCTTTTTTACC
AGTTAAGAAAATCTGGCGCATGGTGTATGCCACCTGGCGACTATGTTCTCCTATCTATAGCGCAAT
GGTGGAGGATTCGCGAGGCTGTGTAACCCACATTTTGAACCTTCTTACGACACGAGCTCGCAACG
GTGGAACGGGCTCGGCTGCTCTGTGATGTTGATGATGATGATGATGATGATGATGATGATGATGATGAT
TTCAACTGTGGCATGGGACCGCTTTGACATTTATCTCCCTCGCTTCTAAGGAACCGTAAATCCA
GACAGTTTTTTCGGCCGCAATAAGGCAACCGCTGGAAACGGGAGATTGTTGGCCTTTTGGACCCGATG
ACCTCCTCGGGGAGCGTATGAAAGCAAGATCGCCCTCTAGGATATGAGCCGCTCGTTTTTCTGATCCTT
TAATGAAATTCGCGATTGGATTTCAAAAGGCAATGAAGCCGATCTGGTATGTTGATGTTGATGTTGATG
CCTGAAGATCAAACTCCTAATTCCTGATTTAGTGTCAAGACCGCTCCATCATATTTGGCGAAATG
ATCTCAAAATGACCTTTCAAGGGAGGATGTGACCGGAGCTTTATCTCCGAAAGCGGATATGCTCCAG
AACTATGGCGCATGCAATCTAACAAAATCAAGACGTAG
```

Change region shown

Whole sequence
 Selected region

from: 145694 to: 148183

Update View

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Related information

BioProject

Full text in PMC

Gene

Genome

Identical GenBank Sequence

Protein

Protein Clusters

PubMed

PubMed (Weighted)

Taxonomy

Recent activity

Turn Off Clear

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

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
Enter a Swiss-Prot/TrEMBL accession number (AC) (for example P0130) or a sequence identifier (ID) (for example NOTC_DROME), or a PDB identifier, or paste your own protein sequence in the box below: <pre> SRVAVTKLYASRPTVFPVCLAPLVVPRCTWISNMTTTE NLVKEVASFTEDLRTSLVSRERINIGKPTVAKTHLSTIGLA RVIDSVITNNDTQPTIEIQQLAFLFLVAVSTILQVQVSY LSRDGLMPSYIARSNTEVAVAVANSSNSRQDVTWYQTV EDLTLGRLNGNSTRSQSLVHTDWPQAAQGNHTTAPVGT SLGGHNETLIQSVVSLYERKGLVSLQFPKTLTEVLNGL NLRIHELIMWTEDVTVLVEKSLMSFFIISGICQGRFE INLMSQCIPENCSSGVEVEIKRLRVAQPCSYIIEVGVPL </pre> <input type="button" value="Clear"/>	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) <input type="text"/>
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"/> greedy, overlaps, no includes (for patterns, see help) randomize databases: <input type="text"/> no (to test a pattern, see help)

Secondary databases

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

>[PDOC00003 PS00003](#) **SULFATION** Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].

571 - 585 nkeesstYeteisns

>[PDOC00004 PS00004](#) **CAMP_PHOSPHO_SITE** cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

744 - 747 RRvT
814 - 817 KRrS

>[PDOC00005 PS00005](#) **PKC_PHOSPHO_SITE** Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

148 - 150 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

Secondary databases

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

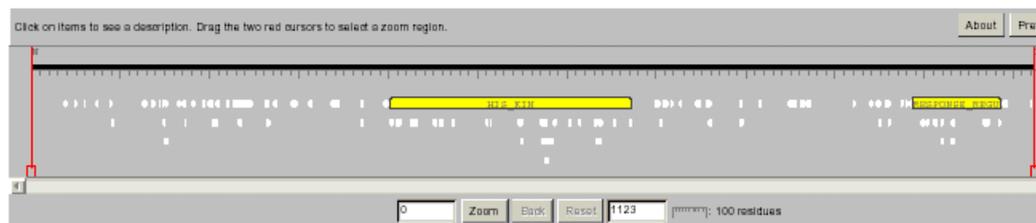
>[PDOC50109 PS50109 HIS_KIN](#) Histidine kinase domain [profile].

```
402 - 671 NASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVMVCAKDLVALLNSVLEMSKIESG
KMQLVREDFNLSKLLLEDVIDFYHFMKKGVDVLDPHDgveKPSNVRGDSGRKQILN
NLVSNVRFVFD - -GHIAVRAWAQrpgensavvlasyppgvkfvkkmfckkkesatye
teienairnnaTMEFVFEVDDTGKGIHMEMRKSVPBNYVQVREtAQGHQGTGLGLGIVQ
SLVRLMGGEIRITDKAMGeKGTCPQPNVLLTT
```

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

```
987 - 1085 RVLVDDNPISRKVTGKLNKMGVSeVEQCDSGKEALRLVTEGLtqreeggsvdklpPDY
IFMDQMPEMDGYRATREIRkvekSYGVRTPIIAVSGHD-----
```

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:

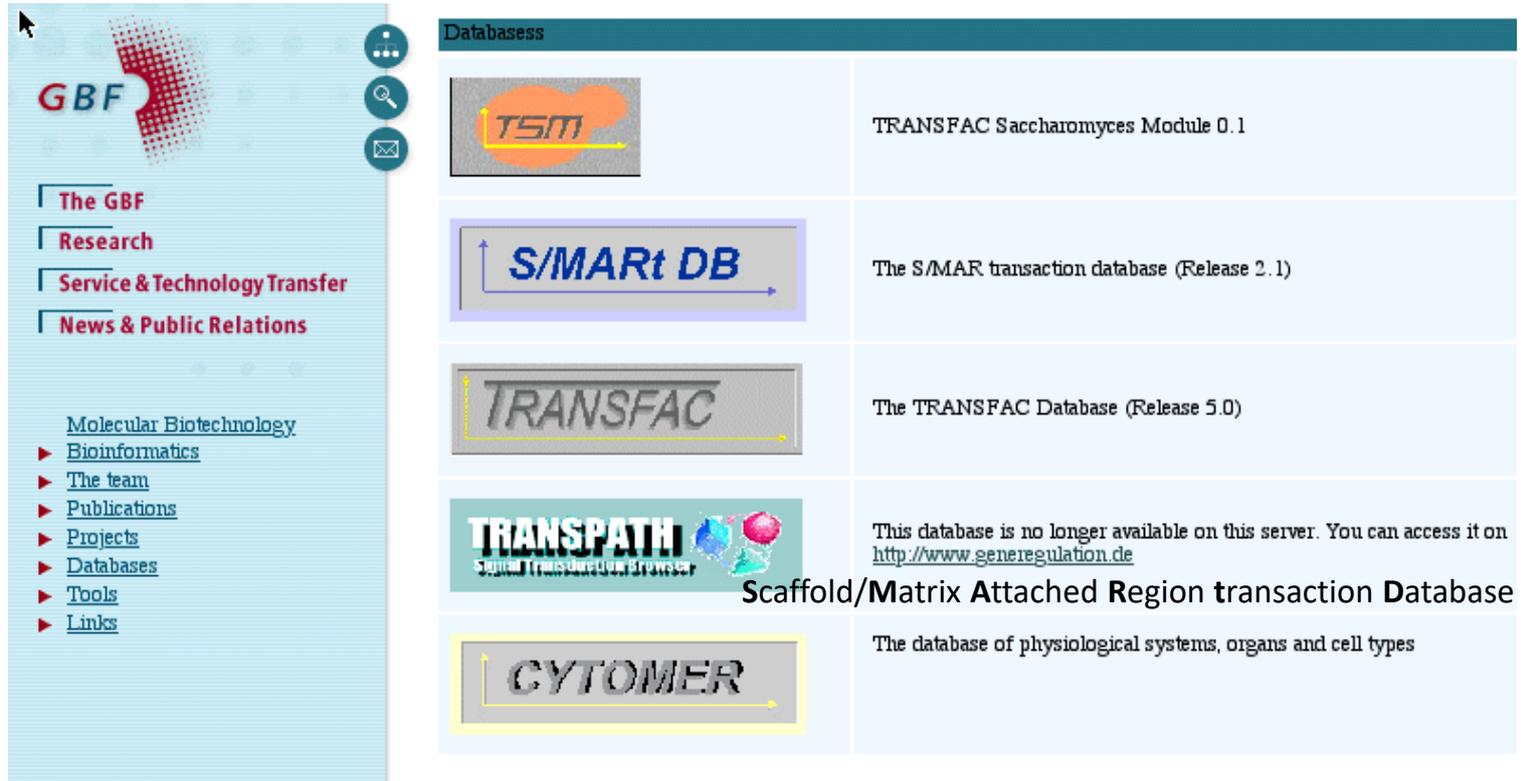
- [By accession number](#)
- [By PRINTS code](#)
- [By database code](#)
- [By text](#)
- [By sequence](#)
- [By title](#)
- [By number of motifs](#)
- [By author](#)
- [By query language](#)

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 website interface for TRANSFAC. On the left is a navigation menu with the GBF logo and links for 'The GBF', 'Research', 'Service & Technology Transfer', and 'News & Public Relations'. Below these are links for 'Molecular Biotechnology', 'Bioinformatics', 'The team', 'Publications', 'Projects', 'Databases', 'Tools', and 'Links'. The main content area is titled 'Databases' and lists several databases:

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

S/MARt DB (scaffold/matrix attached region transaction database). This database collects information about S/MARs and the nuclear matrix proteins that are supposed be involved in the interaction of these elements with the nuclear matrix.

<http://transfac.gbf.de/SMARTDB/index.html>

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

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](#)

Structural databases

- 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](#)

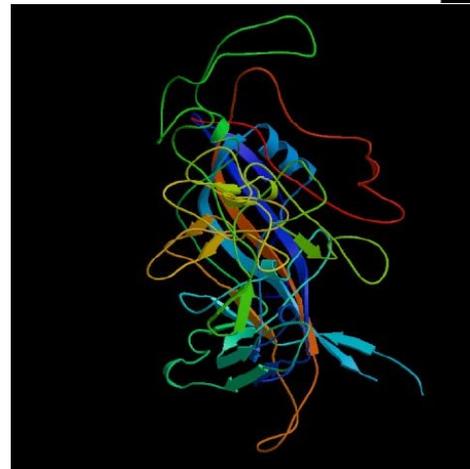
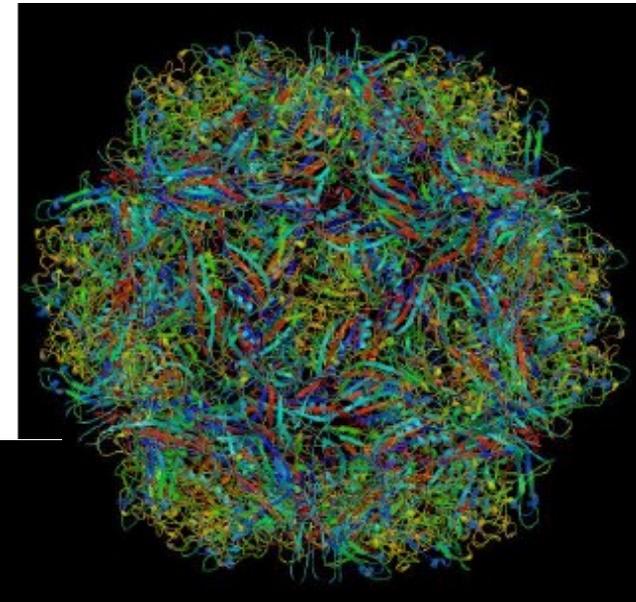
[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

Explore

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

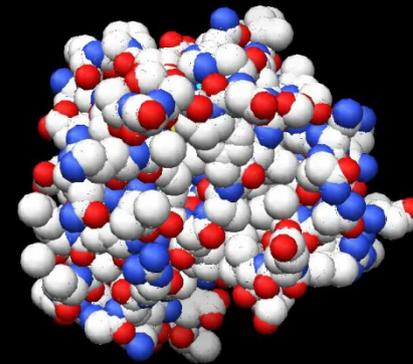
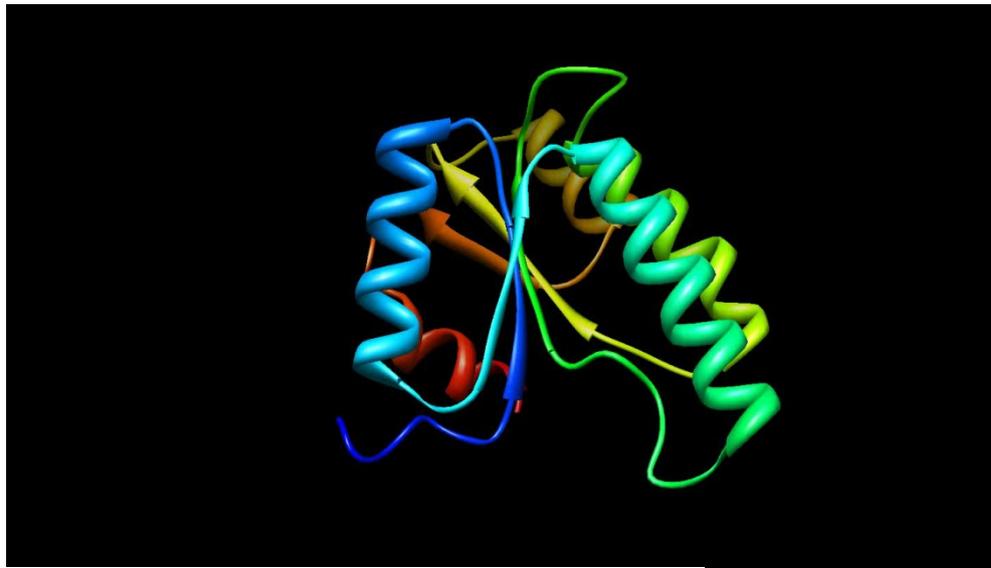


<http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=:pid=173561064349344&bio=1&opt=show&size=500>

12/29/2003

Structural databases

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



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

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
 - GENOME resources

Genome resources

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

The screenshot shows the UCSC Genome Browser interface. At the top, there's a navigation bar with links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, About Us, and Help. Below this is the main search area with a form containing fields for 'clade' (Mammal), 'genome' (Human), 'assembly' (Feb. 2009 (GRCh37/hg19)), 'position' (chr21:33,031,597-33,041,570), and 'search term'. There are also buttons for 'submit', 'track search', 'add custom tracks', 'track hubs', and 'configure tracks and display'. Below the search area, there's a section titled 'Human Genome Browser – hg19 assembly (sequences)' with a sub-section 'Sample position queries'. This section provides a list of queries and their corresponding responses, such as 'chr7' displaying all of chromosome 7, and 'RH18061,RH80175' displaying a region between genome landmarks.

The UCSC logo is visible on the right side of the page, featuring a stylized human figure with a DNA helix and the letters 'U C S C' below it.

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 (hg19.4) 500 bases

RefSeq Genes: HES

Human RefSeq: Human RefSeq from GenBank

Spliced ESTs: Human ESTs That Have Been Spliced

Layered HiChIP: HiChIP Marks (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

DNase ChIP: Digital DNase Hypersensitivity Clusters from ENCODE

Tbx Factor ChIP: Transcription Factor ChIP-seq from ENCODE

Mammal Cons: Placental Mammal Basepair Conservation by PhyloP

RepeatMasker: Multiple Alignments of 45 Vertebrates

Mapping and Sequencing Tracks

Phenotype and Disease Associations

Genome resources

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

Human Gene HBB (uc001mae.1) Description and Page Index

Description: Homo sapiens hemoglobin, beta (HBB), mRNA.

RefSeq Summary (NM_000518): The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta-3' [provided by RefSeq, Jul 2008]. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. ##RefSeq-Attributes-END##

Transcript_exon_combination_evidence :: V00497.1, BU659180.1 [ECO:0000332] ##RefSeq-Attributes-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
CGAP	Ensembl	Entrez Gene
Gepis Tissue	H-INV	HGNC
OMIM	PubMed	Reactome
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.

Genome resources

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

Genomic Sequence Near Gene

Get Genomic Sequence Near Gene

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

Sequence Retrieval Region Options:

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

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

Sequence Formatting Options:

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

Genome resources

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

```

>hg19_knownGene_uc001mae.1 range=chr11:5246696-5249301 5'pad=0 3'pad=0 strand=- repeatMasking=none
ggaaacttgaatcaaggaatgattttaaaacgcagattcttagtggaact
agaggaataaatactgagccaagtagaagacctttcccctcctacc
cctacttcttaagtcaacagggcttttgggtccccagacactctgag
atagtcagggcagaaaagcttagatgtcccagttaacctcctattga
caccactgatccccaattgagtcacactttgggttgaagtgaacttt
ttatttattgtattttgactgcaataagaggtcctagtttttatct
ctgtttccccaaacctaaagtaactaatgacagagcacattgattt
gtatttattctattttagacataattattagcatgcatgagcaaat
agaaaaaacacaataatgaaatgacataatgataatgataatgataat
ataatacacacataataataataattttttcttaccagaaggtttt
aatccaaataaggaagaatgcttagaacaggagtagagtttctacc
attctgctctgtaagtatttgcatactctggagagcaggaagagatcc
atctacataccccaaagctgaattagtagcaaaaactctccactttt
agtgcatacaacttctatttgtgtaataagaaaatgggaaaacgatctt
caatgcttaccagctgtgattccaaaatactgtaataacacttgca
aaggagatgtttttagtgaactttagtagtggtatggggccaagag
atatacttagagggagggctgaggggttgaagtccaactcctaagccag
tgccaagaagagccaaggaaggtacggctgtacacttagacctcaacc
tgtggagccaaccctagggttggccaactctaccaggagcagggagg
gcaaggagccaagggctgggcataaaagtcaagggcagagccatctattgctt
ACATTTGCTTCTGACACACTGTGTTCACTAGCAACTCAAAAGACACC
ATGGTGCACTGTGACTCTGAGGAGAGTCTGCCGTTACTGCCCTGTGGGG
CAAGGTGAAGCTGGATGAAGTTGGTGGTGAAGCCCTGGGCAGTtggat
caaggttaacaagcaaggtttaaaggagacaaatagaaaactgggcatgtgga
gacagagaagactctgggttctctgtaggcaactgactctctctgctat
tggctatatttcccaccccttagGCTGCTGGTGGTCTACCCCTGGACCCAG
AGGTTCTTTGAGTCTTTGGGGATCTGCTCCACTCTCTGATGCTGTATGGG
CRACCCTAAGGTGAAGGCTCATGGCAAGAAAGTCTCGGTGCTTTAGTG
ATGGCCCTGGCTCACTGGCAACCTCAAGGGCACCTTTGGCACACTGAGT
GAGCTCACTGTGACAGCTGCACTGAGTGGTCTGAGAACTTCAGGgtgag
tctatgggacgctgatttcttcccctcttctctatggttaagt
catgtcataaggaagggataagtaacagggtagagtttagaatgggaaac
agacgaatgattgcatcagttggaaggtctcagagctgttttagttctt
ttatttggcttcaatacaaatgtttcttggtttaattcttggcttctt
tttttttctctccgcaattttactattatacttaagtcttaacat
gtgtataacaaaaggaaatctctgagatacattaaagtaactaaaaaa
aaactttacacagctctgctagtaactactatttggaaataatgtgtgc
ttatttgcataatcaatactcccactctttttcttatttttaatt
gatacaataacattatacattttatgggttaaagtgtaatttttaata
tgytacaacataatgacaaaacagggtaatttgcattgtaatttttaa
aaaatgcttctctttaaataacttttggttattctatttctaata
cttcccctaaatctcttctcagggcaaatgatacaatgtatcatcgc
ctcttggaccattctaaagaataacagtgataattctgggttaaggca
atagcaatctctgcataataaatttctgcataaattgtaactgat
gtaagaggttcaatattgctaatagcagctcaaatccaagtaaccattctg
cttttattttaggttgggataaggtggatttctgagttccaagctag
gccccttggtaaatcatgttcatacctcttcttcccacagCTCCT
GGGCAAGCTGGTCTGTGTGCTGGCCCACTTTGGCAAGAAATTC
CCCCACAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGGCTAAT
GCCCTGGCCCAAGTATCACTAAGCTCGCTTTTGTGTGCCAATTTCT
  
```

Genome resources

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

The screenshot shows the TAIR website homepage. At the top, there is a search bar and navigation tabs: 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:** Includes links to "Subscribe to news feed", "Follow our Twitter feed", and "Join our Facebook group".
- 2012 MASC Report Now Available:** Dated July 11, 2012, it mentions the latest report from the Multinational Arabidopsis Steering Committee.
- New Protein Chip and Cell Cultures at ABRC:** Dated May 9, 2012, it describes a new protein chip (AtProteinChip 2) developed by M. Snyder and S.P. Dinesh-Kumar, and cell lines PSB-D (CCL84840) and MM2d.
- Share Your Education Resources:** Dated February 1, 2012, it encourages teachers to use Arabidopsis in their classrooms.
- GO Annotations At TAIR:** Dated January 25, 2012.

A central banner features a laptop with a "TAIR SUBMISSION" form and a syringe, with arrows labeled "SUBMIT PAPER" and "SUBMIT DATA". Below the banner, it says "Click here to try our new online submission form" and lists examples of data to submit: molecular function (e.g. protein kinase), biological process (e.g. seed development), localization (e.g. plasma membrane), or interacting partner of a gene.

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

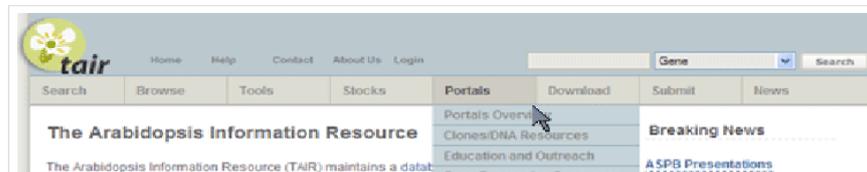
[October 15, 2006]
Search for [genes](#), [germplasms](#), and [polymorphisms](#) using associated phenotype, and see improved phenotype data display in results and detail pages.

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.

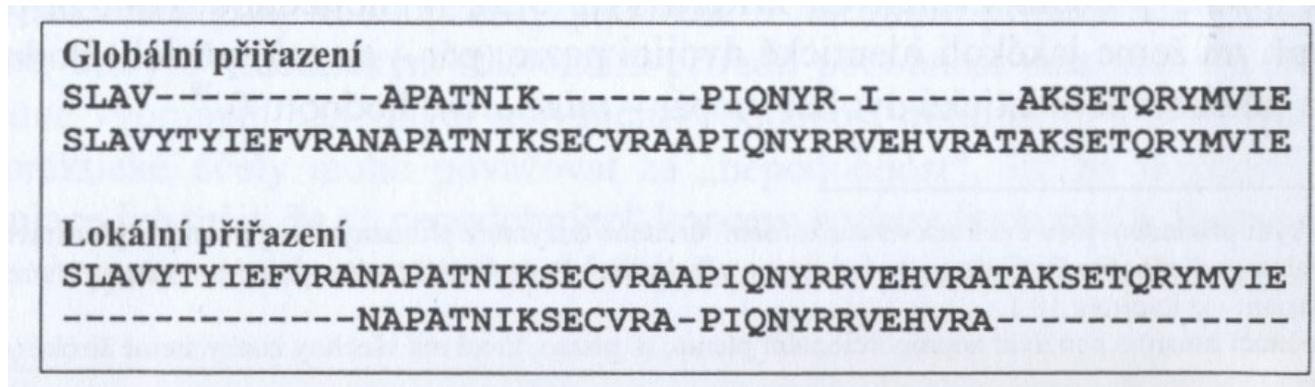


Outline

- Syllabus of this 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

Analytical tools

□ Global versus local alignment

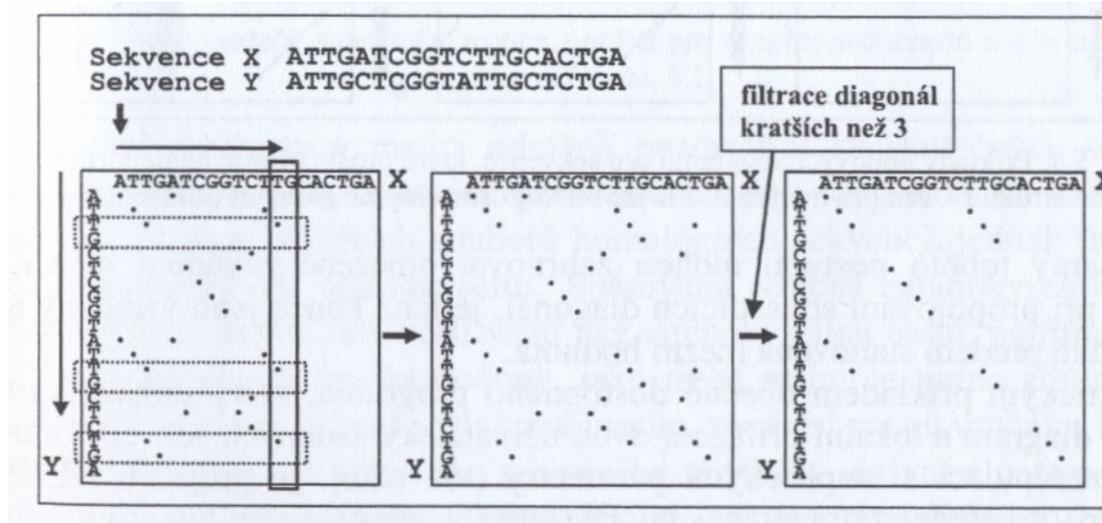


Cvrčková, Úvod do praktické bioinformatiky

- 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

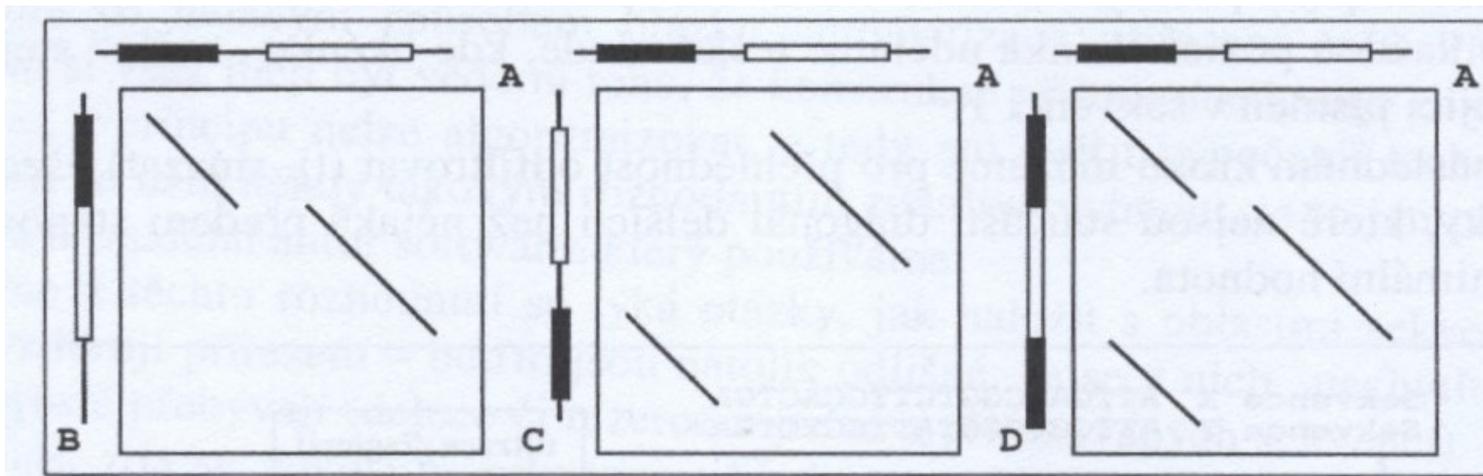


Cvrčková, Úvod do praktické bioinformatiky

- Plotting the sequences (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

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

NCBI *nucleotide-nucleotide* **BLAST**

Nucleotide Protein Translations Retrieve results for an RID

[Search](#)

```

aacccacg
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

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

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	2	-3	-2	-3	-4	-5	-5	-5	-4	-4	-5	-5	-2	-3	-3	-4	-3	-8
0	2	1	1	1	1	1	0	0	0	-1	-1	0	-2	-1	-1	-1	-3	-3	-2
-2	1	3	1	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	1
-2	1	1	1	2	1	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	1
-4	1	0	-1	0	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1
-5	0	0	-1	0	1	2	4	1	1	1	1	1	1	1	1	1	1	1	1
-5	0	0	-1	0	0	1	3	4	1	1	1	1	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	1
-3	-1	-1	0	-1	-2	2	1	1	3	6	1	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	1
-5	0	0	-1	-1	-2	1	0	0	1	0	3	5	1	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	1
-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5	1	1	1	1	1
-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6	1	1	1	1
-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4	1	1	1
-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	1	1
0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	1
-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

BLAST

Basic Local Alignment Search Tool

>gi|5016088|ref|NM_001101.2|
 Length = 1793
 Score = 1110 bits (560), **Expect = 0.0**
 Identities = 965/1100 (87%)
 Strand = Plus / Plus

Query: 156 gtcgacaacggctctgcatgtgcaaggccggatttgcgggagacgatgctccccgcgcc 215
 Sbjct: 101 gtcgacaacggctccggcatgtgcaaggccggcttcgcgggcgacgatgccccccgggcc 160

Query: 216 gtcttcccacatcgattgtgggacgtccccgtcaccaggggtgtgatggctggcatgggcccag 275
 Sbjct: 161 gtcttcccctccacatcggtggggcgccccagggcaccagggcggtgatgggcatggggtcag 220

Query: 276 aaggactcgtacgtgggtgatgagggcgcagagcaagcgtggtatcctcaccctgaagtac 335
 Sbjct: 221 aaggattcctatgtgggcgacgagggcccagagcaagagagggatcctcaccctgaagtac 280

Query: 336 cccattgagcacggatcgtgaccaactgggacgatatggagaagatctggcaccacacc 395
 Sbjct: 281 cccatcgagcacggcatcgtcaccaactgggacgacatggagaaaatctggcaccacacc 340

actin, beta (ACTB), mRNA
 ds..S=1213 E=0.0
 >=200

- „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

Primary databases

The screenshot displays the NCBI GenBank database interface for the gene **NP_059797.1**. The main view shows a genomic map with a red bar representing the gene. A detailed information popup is visible, providing the following details:

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

Below the information popup, there is a section titled **Bibliography** and a section titled **Related articles in PubMed** with a green arrow pointing to the right.

BLINK is a link to the pre-computed BLAST search results for the respective sequence (see the next slide).

BLAST

Basic Local Alignment Search Tool

BLINK *precomputed BLAST* My NCBI [Sign In] [Register]

Home Taxonomy Report Multiple Alignment Blast Help

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	qi 737127	833	virA protein
4153	CAA34777	833	91.3 kDa protein [Agrobacterium tumefaciens]
3800	CAA35780	829	virA [Agrobacterium rhizogenes]
3718	qi 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 six-frame conceptual 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 exists a lot of specialized versions of BLAST
 - **PSI-BLAST (Position-Specific Iterated Blast)**
 - 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 (position specific substitution matrix)
 - 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
VADAKGDSESEEDLLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQBACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGS
TSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIgek
IYKDGERIITQGEKADSFYIESGEVSIILRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYEEQLVKMFGSSVDLGNLQ
```

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

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
 - 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

- <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
GAGCTCCCTTGGGGGGCAAGGGC AAAACTTTTTGCTAAATGGAAAAATATTATACCAAGTGTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAATTTGACTATTTTGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTC AAATGGTCCATTATCGGCCAAGTAGCTTTCTTTAATTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAAATAAATTC AAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAATAAGACCGATCAAAATAAAGCCGCCATTAAAATAATGAATTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACCTTTCCAATACTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTTATCTCTAATTTACATCTCAACTAATATTAAGAAAATAAAACAGGTA
CAGCAAATCATAAAAATTTCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCTTTTTCAGAG
TCTGCATGCCATATTC ACTAAGGGGTCGTTTGGTACAAGAAAATAAATAAATAAATTTCTGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTTAATTATAAGTATTAGCTAATTTTCAGAATAAAATTTTACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACC AAATGATCGGTTAGTCTTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAAGTTAGAAAATTTTCAITTAATCAATTCATATAAATTTAAAATAATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAAATAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAAAACGGTT
CGAGACTCTTTATGGAAGGCGTTTCTTCAAAGTAGATTCTCATTCATTGCTCTGGTGC AATAGCAAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA
GAGAATTTCAAATCTCTCAACTACTTTTTAAGGGAATTC AAAATACGACC AATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTTATTTAAATTTGAATTGAAAATATTAATTTACTTGTATTAATATAA

Analytical tools

- <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
GTTACTCAATTTGAATTAACAAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTTCTTTAAATATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAATAATAACTTCAAAAGTCCATCATCTTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAAATAAGACCAGATCAAAATAAAAGCCGCCATTAAAAATAATGAATTTTAGGACTCTC
GATTTGGCAGGTAAGTGCCAAAACCTTCCAAACTTTTGTCTGCAACTTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTA AACAGGTA
CAGCAAATCATAAAATTTTCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTTCAGAG
TCTGCATGCCATATTTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAAATTTTCGGGATAGAATTT
GAGATTGCATTTATCTTTGTGTTTAAATTAAGTATTAGCTAATTTTCAAGAATAAATTTTACTAAAATAG
TAAAATCAACTATCACATGTAGAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACCAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAAATTTTCAATTAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAGATGTACCGTTAATAATAAAGATAAGATAGAGTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTGTCTTCAAAGTAGATTCTCATTCATTGCTCTGGTGCATAGCAAAA
TGACATCTTACTCTTAAGATACAGGAGCCACTCTACAATCTTCTATTGTATACTCAAAATGAAAGTTTTA
GAGAACTTTTCAAACTCTCAACTACTTTTAAAGGGAATTCAAAATACGACCAATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTAAATTTGAAATTTGAAATATTAAATTTACTTGAATTAATATAA
ACAAATAGATATCGCTAAGTATTTACCACAAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT
GATTAAGCAGCTATTCATCTGGTTTGTGCAGGATGAAAGAAAGTAACTAGCTATAATTTCTTTTGTAAAGT
```

Analytical tools

- <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
```

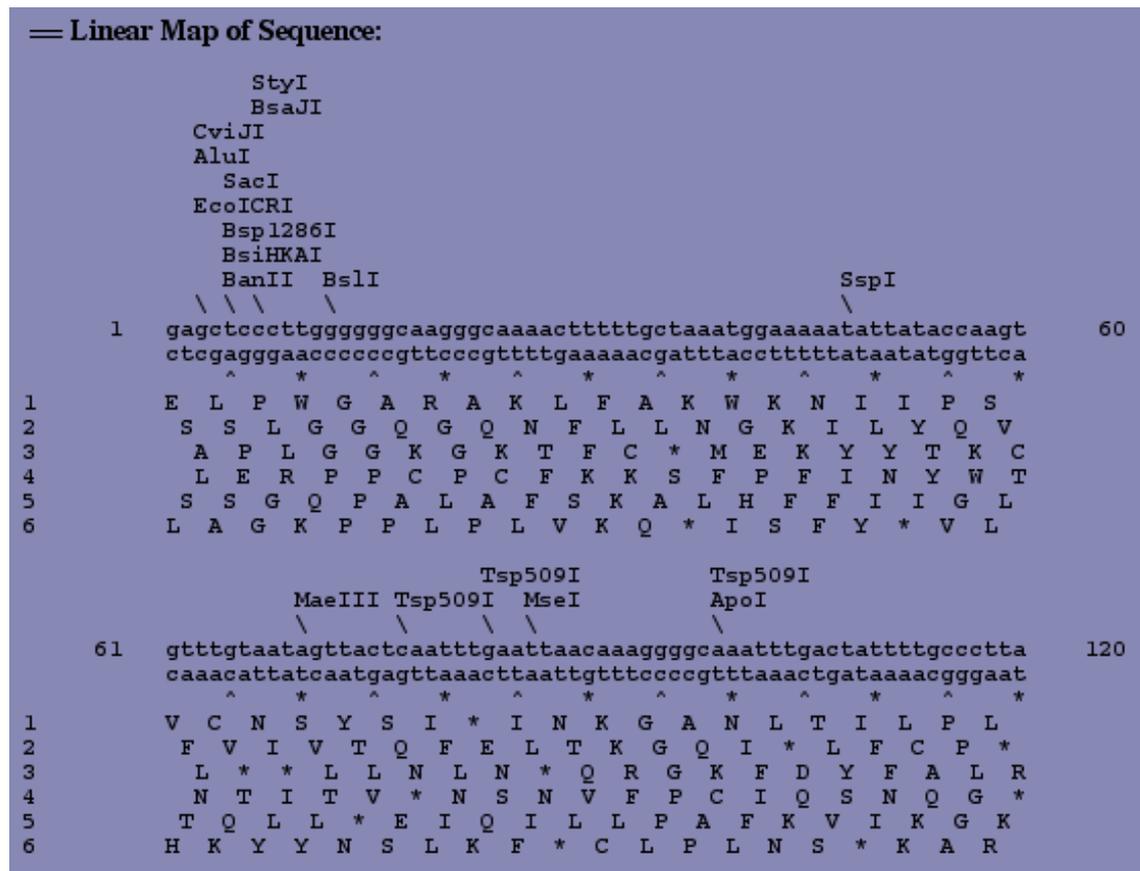
```

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

```

Analytical tools

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



Analytical tools

- <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  CTTTGGCTTGGTGTCTGGCTTGACAACTTGGAGTGGAGACTCGGGTAGACTGGCGGTTTGGG  804655

          2850      2860      2870      2880      2890      2700
24  ..... A A A T G G C T . 170381
1  ..... 11321163
2430 ..... C A A C A A T T . 170248
1743 CAGTGAAATGATTGACAGAACTGCCAAAAACAAGCCAAAATGGTAAAAAAAAAAAAAATTC 19686
2620 CATCGTCTATGTGGACTTCAATACTCTGAAGAGGTACGCCAAGGAGTCAAGGCTTGTGGT  804655

          2710      2720      2730      2740      2750      2760
32  ..... A T T A T G T G C T T C T A C G A T T A 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 C 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 ACGATGTTTACAATTGTTATGTCCAAACGCCGACTCACTATTTTCAATTCATATTGAGG 19686
2680 CAAGAACATGCTGTCCGAAAGAAAGCAGCTAGGATCGCAACAGGATCGGGAGGATC  804655

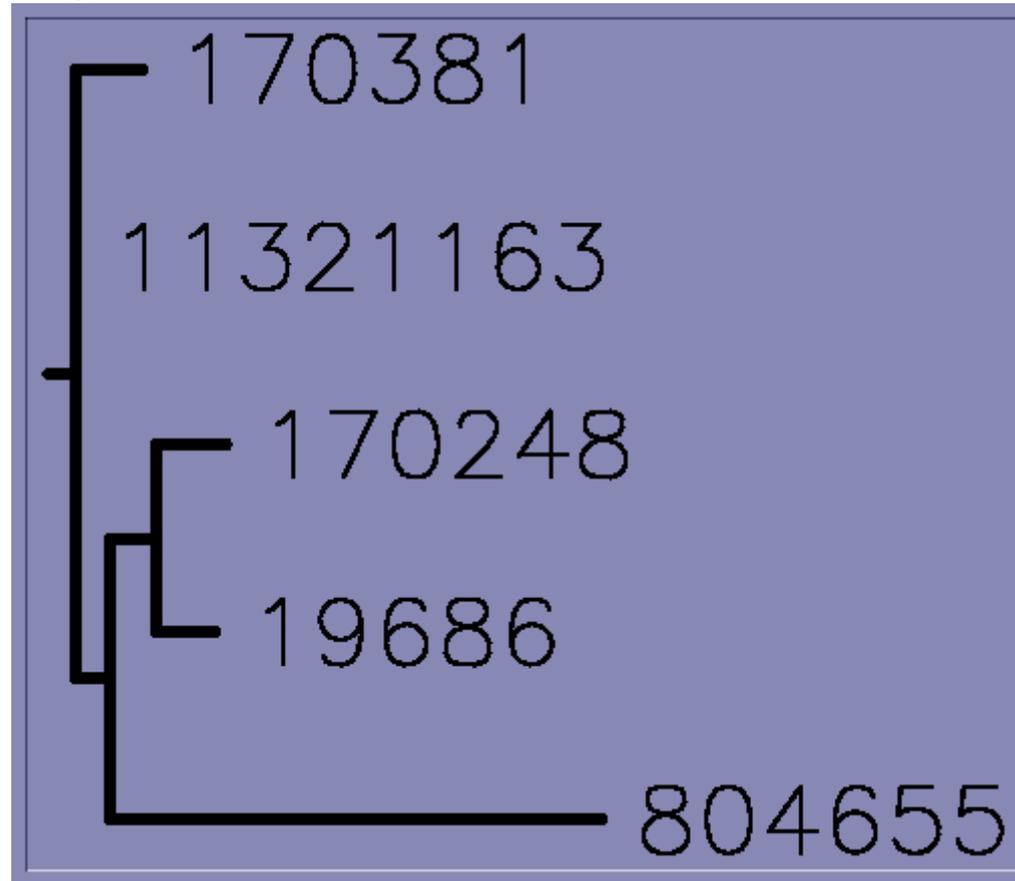
          2770      2780      2790      2800      2810      2820
79  A G A G G G T F A A . . . . A T A G G T G T . . . . T T G T A T G G A A T C A T G G C C A C A A A G T T G C C A T C A G 170381
1  T . . . . A T G G G T G T . . . . T T G C T A T G G A A T C A T G G C C A C A A A G T T G C C T T C A G 11321163
2484 A C C G G G T F A A T C A A T A G G T G T . . . . T T G C T A T G G A A T C T A G C C A C A A A G T T G C C A A A T C 170248
1863 A C C G G G T C A A T G G A T A G G T G T . . . . T T G T A T G G A A T C T A G C C A C A A A G T T G C C A A A T C 19686
2740 A C T G C T T C A G G 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 G T T T T A T C A T A G A A T G T C A G  804655

          2830      2840      2850      2860      2870      2880
132 A T T G T T A A G 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 G A T T A C A A A C T G A G G C T T T A T G A 170381
45  A T T C G G A A G 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 G A A A T T G A A G G C T T T A T G A 11321163
2540 A T T C G G A A G 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 G A A A C T G A G G C T T T A T G A 170248
1919 A T T C G G A A G 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 G A A A C T G A G G C T T T A T G A 19686
2800 A C T T A G G C C C T G T T G G T A A A A C A C C A C C C C A A T A T C C C A A . . . . C C G G A A A T T G C A G  804655

```

Analytical tools

- <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 ([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



Analytical tools

- VPCR <http://grup.cribi.unipd.it/cgi-bin/mateo/vpcr2.cgi>



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
 - GENOME resources
- Analytical tools
 - Homologies searching
 - Searching of sequence motifs, open reading frames, restriction sites...
 - Other on-line genome tools

Other online 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:07816](#); [MIM:608726](#)
Gene type protein coding
RefSeq status REVIEWED
Organism [Homo sapiens](#)
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo
Also known as FLJ13171; MGC20618; MGC34186; DKFZp686L07205; RP11-442N24__A.1
Summary This gene encodes a member of the phosphatase and actin regulator (PHACTR) family. Other PHACTR family members have been shown to inhibit protein phosphatase 1 (PP1) activity, and the homolog of this gene in the mouse has been shown to interact with actin and PP1. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]

Genomic context

Location : 1p35.3
Sequence : Chromosome 1; NC_000001.10 (28696093..28826881)

[See PHACTR4 in MapViewer](#)

Genomic regions, transcripts, and products

Genomic Sequence NC_000001 chromosome 1 reference GRCh37.p5 Primary Assembly

[Go to reference sequence details](#)

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

Table of contents

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

Links

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

Other online genome resources

- 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.