

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



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



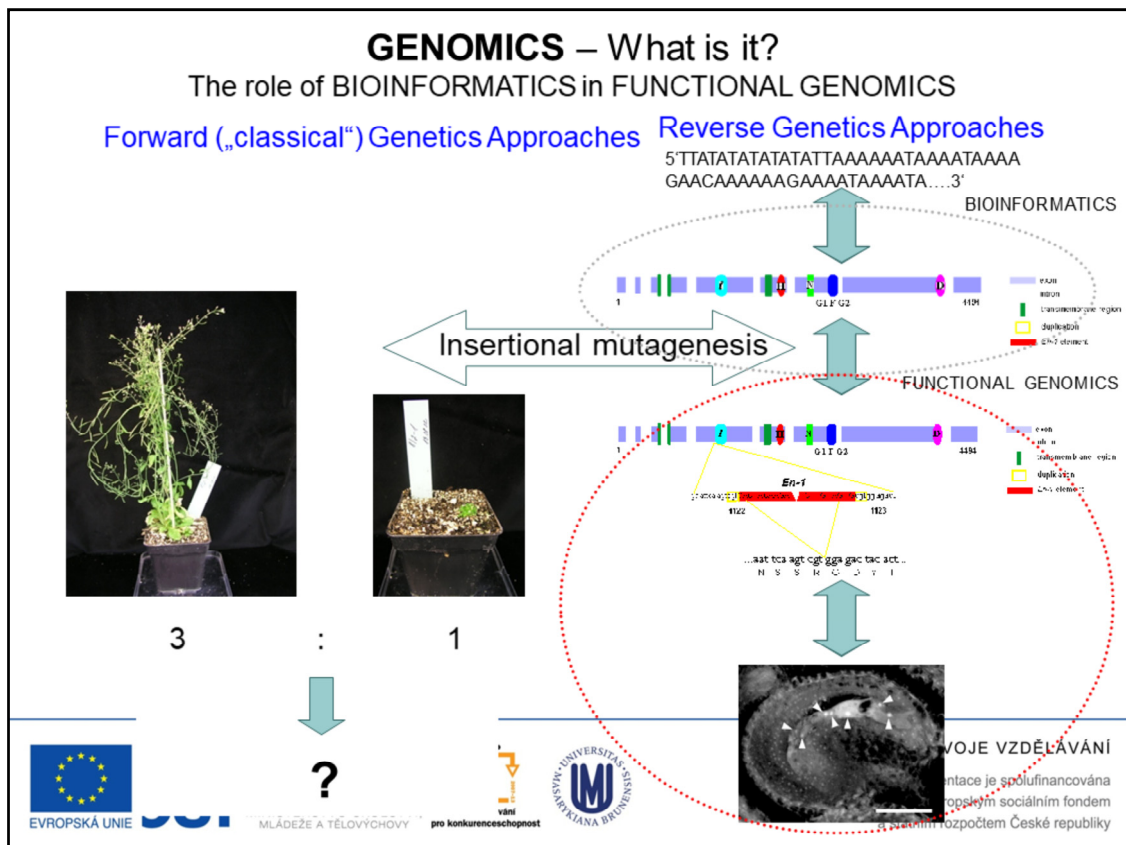
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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 eukaryotic 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.



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

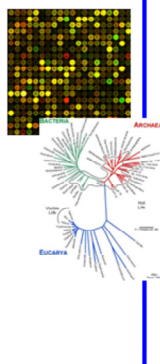
- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS



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.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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. 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.

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.

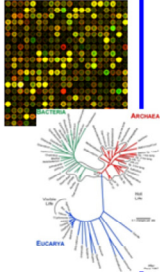
J. Pevsner,
<http://www.bioinfbook.org/index.php>



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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

EMBLnet National Nodes		
Vienne BioCenter	Austria	http://www.at.emblnet.org/
BBN	Belgium	http://www.be.emblnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.emblnet.org/
INFORMAGEN	France	http://www.infoblogem.fr/
CRISOLINET	Germany	http://genome.zibb-helmholtz.de/biocom/
IMBB	Greece	http://www.imbb.forth.gr/
HIN	Hungary	http://www.hu.emblnet.org/
INCEI	Ireland	http://www.gen.tcd.ie/
JNN	Israel	http://dapsil.wellman.ac.il/bcd/inn.html
JIB-ADN	Italy	http://jib-www.ba.cnr.it/8000/BioWWW/Bio-WWW.htm
CAS/C/CAIN	Netherlands	http://www.cas.kun.nl/
IBO	Norway	http://www.no.emblnet.org/
IBB	Poland	http://www.ibb.wzpa.pl/
ISC	Portugal	http://www.lgc.gulbenkian.pt/
GeneBee	Russia	http://www.genebee.msu.ru/
CNB-CSC	Spain	http://www.es.emblnet.org/
BNC	Sweden	http://www.se.emblnet.org/
SIB	Switzerland	http://www.ch.emblnet.org/
SIGNET	UK	http://www.signet.dl.ac.uk/
EMBLnet Specialist Nodes		
MPS	Germany	http://www.mips.biochem.mpg.de/
ICGB	Italy	http://www.icgb.internic.it/
Pharmacia Uppsala	Sweden	http://www.gnu.com/
FaH/Faasac-La Roche	Switzerland	http://www.roche.com/
EBI	UK	http://www.ebi.ac.uk/
HGMP-BC	UK	http://www.hgmp.mrc.ac.uk/
Sanger	UK	http://www.sanger.ac.uk/
EMBL-EB	UK	http://www.embl.ac.uk/
EMBLnet Associate Nodes		
IBBH	Argentina	http://iui.biot.unip.edu.ar/emblnet
ANGS	Australia	http://www.angis.usc.edu.au/
CEI	China	http://www.cei.cbi.cas.edu.cn/
CISB	Cuba	http://ibc.cigb.edu.cu/
CFDQ	India	http://falarjung.emblnet.org.in/
SANBE	South Africa	http://www.sanbi.ac.za
USA Information Providers		
NCBI	USA	http://www.ncbi.nlm.nih.gov/
HLM	USA	http://www.nlm.nih.gov/
NDH	USA	http://www.nih.gov/



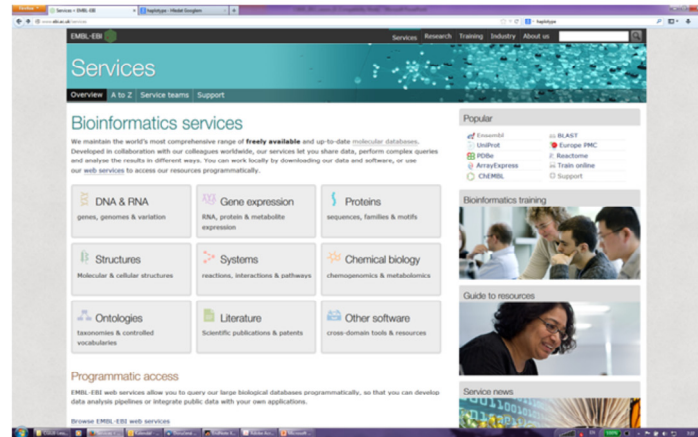
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

Spectre of on-line Resources

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

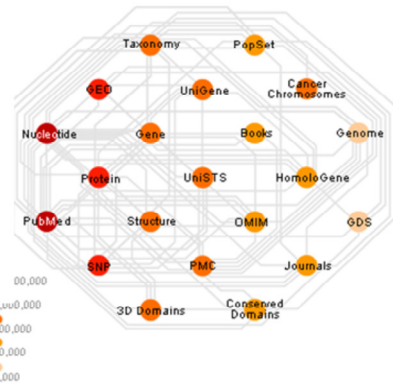
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Nowadays, the resources are interconnected and could be accessed via dedicated web pages. Among the best and mostly used www resources integrating plenty of database resources belong www portal of European Bioinformatics Institute (EBI) in Europe (Germany) and National Center of Biotechnology Information (NCBI) in the USA (

Spectre of on-line Resources

□ NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI homepage with a search bar at the top. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to Tools, Downloads, and Submissions, and a 'NCBI YouTube channel' section with a 'GO' button. A sidebar on the left lists various resources like 'All Databases', 'Chemicals & Bioassays', and 'Genetics & Medicine'. A 'Popular Resources' list is on the right, including PubMed, Bookshelf, and BLAST.



EVROPSKÝM SOCIÁLNÍM FONDEM
a státním rozpočtem České republiky

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**



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 x 10⁶ entries (approx. 33 x 10⁹ bp)
 - August 2005 100 x 10⁹ bp from 165.000 organisms



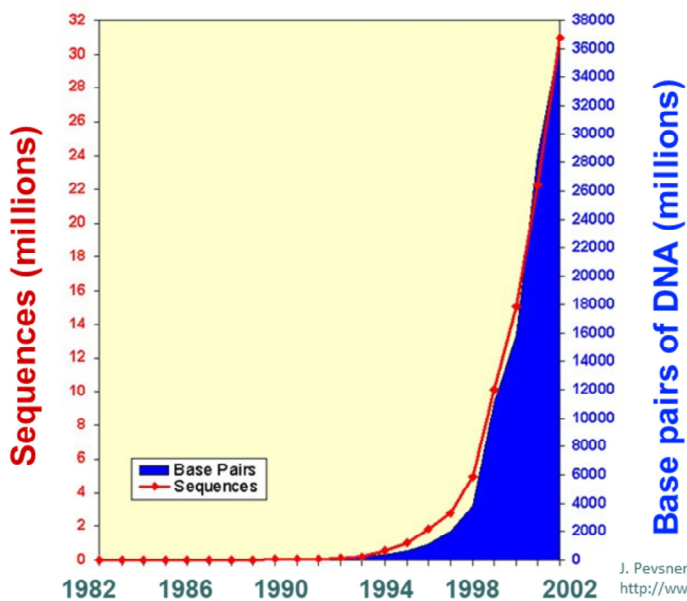
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Growth of GenBank



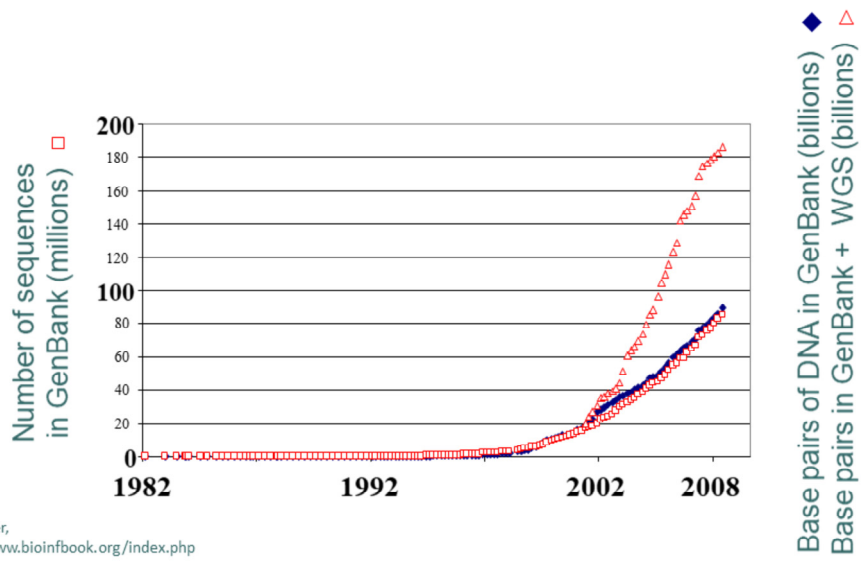
J. Pevsner,
<http://www.bioinfbook.org/index.php>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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



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

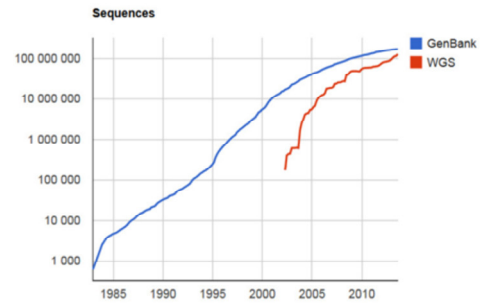
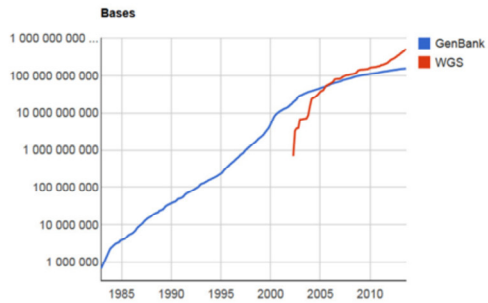


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Growth of GenBank

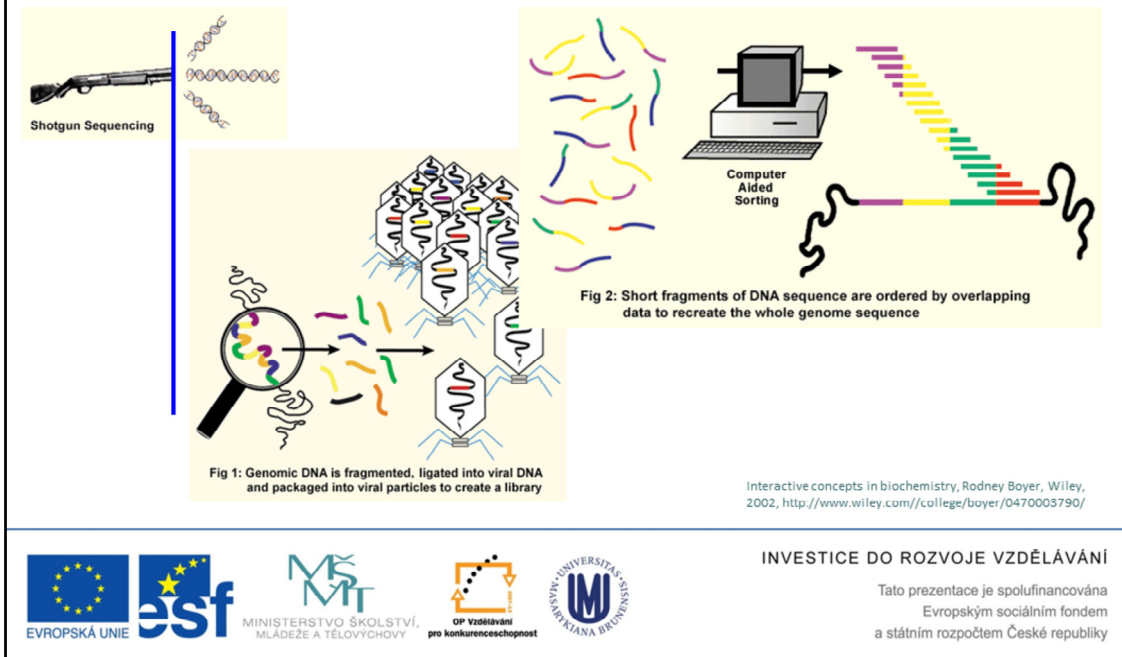
Feb 15 2013



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

WGS

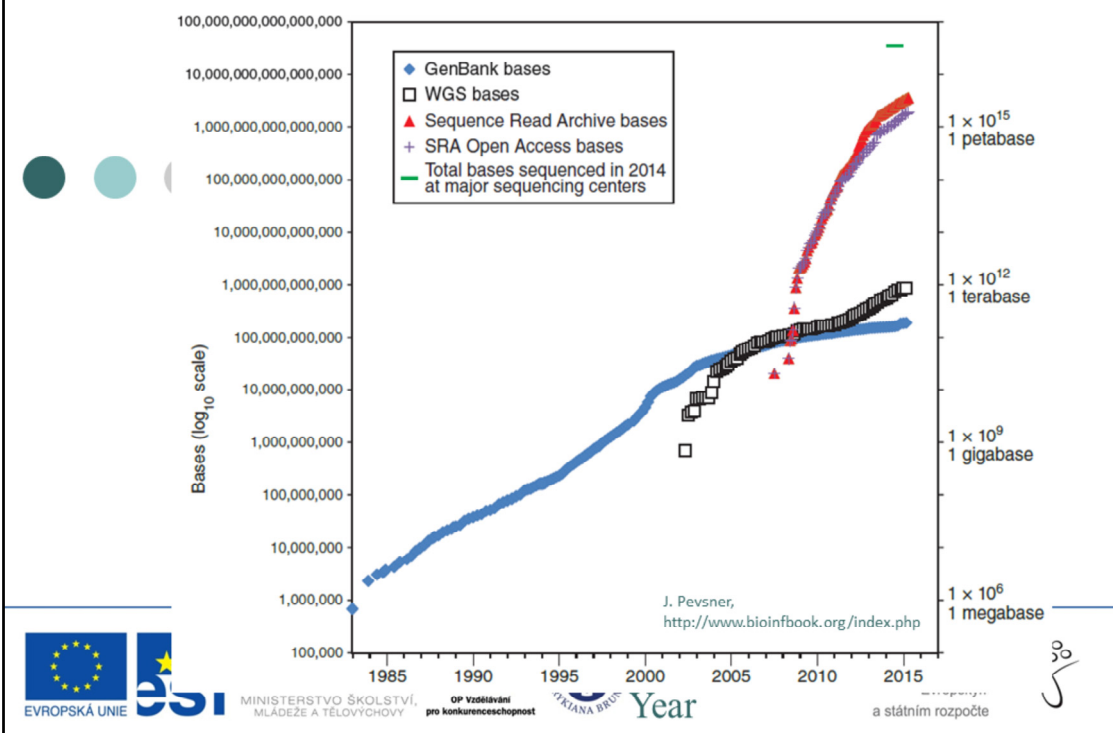


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic 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).

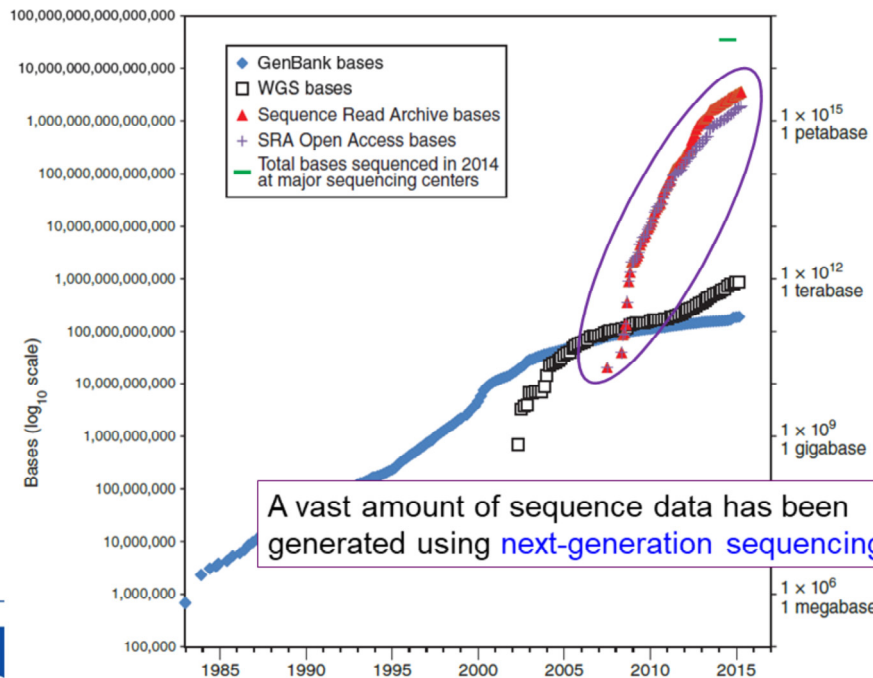
The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence.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

Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

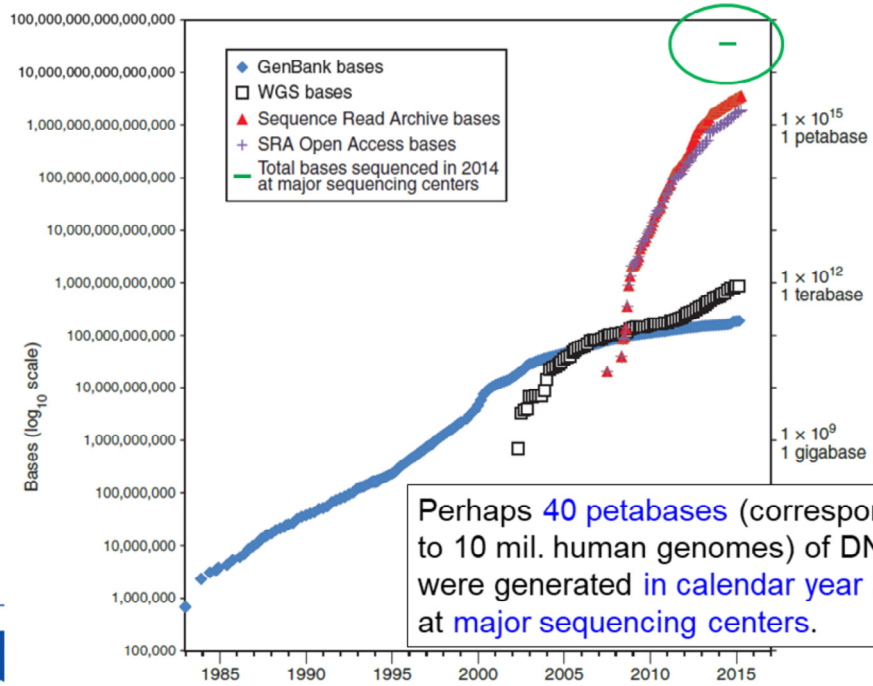
OP Vzdělávání
pro konkurenceschopnost

EVROPSKÝ
ROK
2014

a státním rozpočte



Growth of DNA Sequence in Repositories



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání
pro konkurenceschopnost

EVROPSKÝ
ROK

a státním rozpočte

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/>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primary Databases

- Types of sequences in primary databases
 - Standard nucleotide sequences acquired by high quality sequencing
 - **ESTs** (**E**xpressed **S**equences **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)

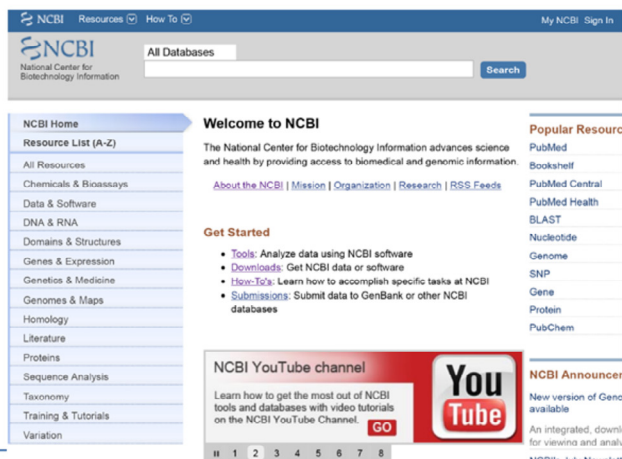


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primary Databases

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



The screenshot shows the NCBI homepage with a search bar at the top. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to Tools, Downloads, How-To's, and Submissions, and a 'Popular Resources' list on the right. A 'NCBI YouTube channel' banner is also visible.



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 entry for the 'suk' gene. Key information includes:

- Gene symbol:** suk
- Gene description:** non-component VWA-like sensor kinase
- Location:** plasmid 01
- Sequence:** NC_023771 (146584..146783)
- Genomic context:** A linear map showing the gene's position on the plasmid.
- Genomic regions, transcripts, and products:** A detailed view of the gene structure with exons and introns.
- Related articles:** A list of four scientific papers related to the gene, with the first article highlighted by a yellow circle.
 1. Sequences *ankA* of *Agrobacterium tumefaciens* and *Agrobacterium tumefaciens* *Ti* plasmid *pTi3355*, Schrammeyer B. et al. J Euk Bot 2000 Jun; PMID 10940245
 2. The *suk* gene from *Agrobacterium tumefaciens* *Agrobacterium tumefaciens*, Turb NC. et al. Mol Microbiol 1993 Mar; PMID 8491915
 3. Characterization of the *suk* locus of *Agrobacterium tumefaciens*, a transcriptional regulator and host range determinant, Lewis R. et al. EMBO J 1987 Jul; PMID 3098978
 4. Analysis of the complete nucleotide sequence of the *Agrobacterium tumefaciens* *suk* locus, Thompson D.V. et al. Nucleic Acids Res 1988 May 25; PMID 2837738



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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primary Databases

The screenshot shows the NCBI GenBank database interface. At the top, there is a search bar with 'Nucleotide' selected. Below the search bar, the accession number 'F022377.1' is highlighted with a red circle. The text 'Accession number' is written in red above the circle. Below the accession number, the text 'GeneBank Identifier' is written in red, with a red circle around the accession number itself. The main content area displays the following information:

```

LOCUS       F022377.1                2490 bp    DNA    linear    BCT 29-DEC-2003
DEFINITION  Opuntiacoccus tumefaciens carotachrom plasmid T1, complete sequence.
ACCESSION   F022377.1
VERSION    F022377.1
KEYWORDS   Agrobacterium
SOURCE     Agrobacterium tumefaciens (Rhizobium radiobacter)
           strain: Rhizobiales;
           type: Agrobacterium.
           Organism: Agrobacterium tumefaciens;
           Project: Schrammeyer, E., Mooykang, P.J. and
           Farrand, S.K.
TITLE       Octopine-type T1 plasmid sequence
JOURNAL     Unpublished
REFERENCE   2 (bases 1 to 2490)
AUTHORS    Zhu, J., Oger, P.M., Schrammeyer, E., Mooykang, P.J., Farrand, S.K. and
           Wisniew, S.C.
TITLE       Direct Submission
JOURNAL     Submitted (07-MAR-2003) Microbiology, Cornell University, Ithaca,
           New York, NY 14853, USA
COMMENT    PROVISIONAL accession. This record has not yet been subject to final
           NCBI review. The reference sequence was derived from EU041113.
FEATURES             Location/Qualifiers
     source           1..2490
                     /organism="Agrobacterium tumefaciens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:258"
                     /plasmid="T1"
                     /contig="contachromosomal"
     octopine-type    1..2490
                     /gene="viraA"
                     /db_xref="GeneID:1224316"
                     /gene="viraA"
     CDS              1..2490
                     /gene="viraA"
                     /note="Two-component regulator of vir regulon; ViraA is a
                     transmembrane histidine kinase"
                     /coding_start=1
                     /transl_start=11
                     /product="viraA"
                     /protein_id="F022377.1"
                     /db_xref="GI:11955343"
  
```



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 ViA-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: 18031..18332
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: [E18640](#)

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	HskA: Histidine Kinase A (dimerization/phosphoreceptor) 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 ...
PRK13637	PRK13637: two-component ViA-like sensor kinase. Provisional

Location: 14 - 833
Blast Score: 2944

Related Sequences



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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



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/>

```
>PDOC00003 PS00003 SULFATION Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].
573 - 585 skneaatTet.e1aae

>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 RRzG

>PDOC00005 PS00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
148 - 150 EGR
164 - 166 TGR
172 - 173 EKK
219 - 221 EKK
369 - 371 TGR
460 - 462 EGR
513 - 516 EGR
585 - 587 ELD
602 - 604 TGR
652 - 654 TGR
716 - 718 RGR
726 - 728 EGR
747 - 749 TGR
794 - 796 EGR
854 - 856 EGR
884 - 886 EGR
888 - 890 EGR
921 - 923 RGR
957 - 959 EGR
960 - 962 TGR
974 - 976 TGR
997 - 999 EGR
1002 - 1004 TGR
1018 - 1020 EGR
1031 - 1033 TGR
1139 - 1141 EGR
```



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/>

```
>PDOC-50100 PS0100 HIS_KIN Histidine kinase domain [profile]
402 - 671  SAEKEDVSGALADWHLIDICDQVTKPQDQVDTLNAVNVCAKGLVALLRSLVLMKIEIDGG
DQGLTEEDPRLLELLELVLPVDFVANKKQVLLQSDGKPKFFPTDGGDDELEQILR
RLTVNDVTFPTD-----GKLVAGKIKYVGRKQKQVYIYKQYKPVKAFKQKQKQKQKQKQKQK
PGLKQKILKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQK
KLVTEKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQK
-----
>PDOC-50110 PS0110 RESPONSE_REGULATORY Response regulatory domain [profile]
887 - 1085  PVLVTVDFKLSRVAATQSLKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQK
LPGKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQK
-----
```

Graphical summary of hits (*java applet*)



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**
- PRINTS, <http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>



PRINTS is a compilation 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 PRINTS/FPRINT/SMART composite. Usually the motifs do not overlap, but are scattered 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. [Reference](#)

New:

- [SPRINT](#) - Search PRINTS-3 evolutionary PRINTS
- [comPRINTS](#) - Search PRINTS automatic map/cluster
- [PRINTS](#) - Search the integrated InterPro family database

Direct PRINTS access:

- [By accession number](#)
- [By PRINTS code](#)
- [By FASTA code](#)
- [By ID](#)
- [By name](#)
- [By number of motifs](#)
- [By domain](#)
- [By query language](#)

PRINTS search:

- Search PRINTS with **NEW FingerprintScan**
- [FPrint](#)
- [U.FPRINTScan](#)
- [MULTISeq](#)
- FingerprintScan binaries and source are available: patrick.accedia@bioinf.man.ac.uk

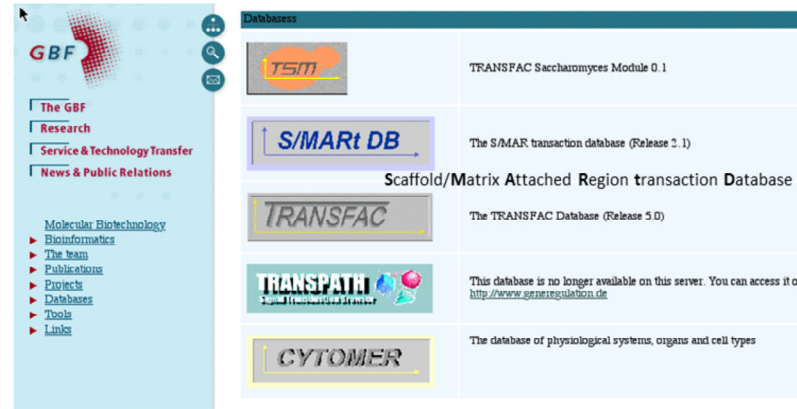


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Secondary Databases

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



The screenshot shows the TRANSFAC website interface. On the left is a navigation menu for GBF (German Biotechnology Foundation) with categories like Research, Service & Technology Transfer, and News & Public Relations. 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) Scaffold/Matrix Attached Region transaction Database
TRANSFAC	The TRANSFAC Database (Release 5.0)
TRANSPATI	This database is no longer available on this server. You can access it on http://www.gene-regulation.de
CYTOMER	The database of physiological systems, organs and cell types



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

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

The screenshot shows the PDB website interface. At the top, it says 'PROTEIN DATA BANK' with the RSCB logo and navigation links for 'Home' and 'Us'. A welcome message states: 'Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.' Below this are navigation links for 'ABOUT PDB', 'DATA UNIFORMITY', 'RECENT FEATURES', 'USER GUIDES', 'FILE FORMATS', 'EDUCATION', 'STRUCTURAL GENOMICS', 'PUBLICATIONS', and 'SOFTWARE'. The main content area is divided into several sections: 'DEPOSIT data', 'DOWNLOAD files', 'Browse LINKS', 'BETA TEST new features', and 'BETA release files'. A 'Current Holdings' section reports '19623 Structures' and 'Last Update: 30-Dec-2002'. The 'Molecule of the Month' is 'Cytochrome c'. A search section titled 'Search the Archive' includes a search box and options for 'query by PDB id only', 'match exact word', and 'remove sequence homologues'. A 'PDB Mirrors' section lists various international mirror sites. A 'News' section features a holiday message dated '23-Dec-2002'.




INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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


Structural Databases

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

Structure Explorer - 1PSY

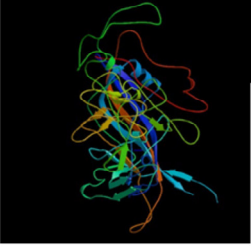
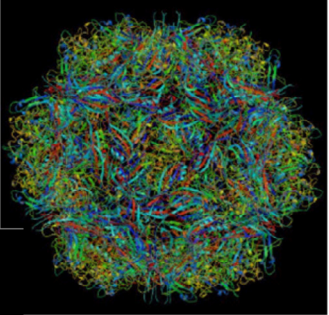
 **Structure Explorer - 1PSY**

Title: The Structure Of Host Range Controlling Regions Of The Capsids Of Canine and Feline Parvoviruses and Mutants
Classification: Virus/Viral Protein
Compound: Mol. Wt. 11, Molecular: Coat Protein Yp2, Chain: A; Fragment: Sequence Database Residues 190-231, Engineering: Yes, Mutation: Yes
Exp. Method: X-ray Diffraction

 **View Structure**

Summary Information
[View Structure](#)
[Download Display File](#)
[Structural Neighbors](#)
[Geometry](#)
[Other Sources](#)
[Sequence Details](#)

[Search by...](#) [Search by...](#)



<http://www.rcsb.org/pdb/cgi/structure.cgi?job=graphics&pdb=1PSY&page=pdb-173561064329344&bio=1&opt=show&size=500> 12/20/2003

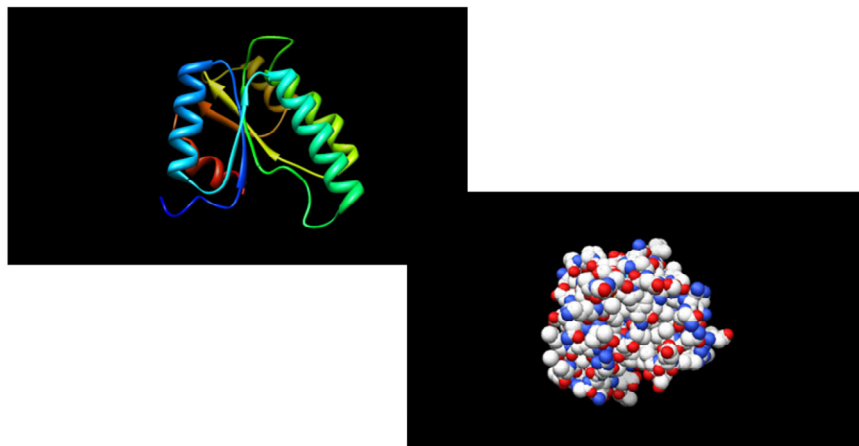


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Structural Databases

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



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



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**



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. At the top, there's a search bar with fields for 'clade' (set to 'Human'), 'genome', 'assembly', and 'position'. Below this, there's a section titled 'Human Genome Browser - hg19 assembly (sequences)'. It includes a 'Sample position queries' section with a table of 'Request' and 'Genome Browser Response'.

Request	Genome Browser Response
chr7	Displays all of chromosome 7
chr7p_g000212	Displays all of the unpaired contig g000212
20p13	Displays region for band p13 on chr 20
08S1.1000000	Displays first million bases of chr 1, counting from p-arm telomere
chr3:100000-2000	Displays a region of chr3 that spans 2000 bases, starting with position 100000
RH1801:R80175 15q11-15q13 rs154252/rs1600376	Displays region between genome landmarks, such as the STS markers RH1801 and R80175, or chromosome bands 15q11 to 15q13, or SNPs rs154252 and rs1600376. This syntax may also be used for other range queries, such as between unpaired contigs, ESTs, mRNAs, refSeq, etc.
D18S3046	Displays region around STS marker D18S3046 from the Genethon/Manfield maps. Includes 100,000 bases on each side as well.
AK20414	Displays region of EST with GenBank accession AK20414 on BRCA1 cancer gene on chr 17
AC089101	Displays region of clone with GenBank accession AC089101
AF382811	Displays region of mRNA with GenBank accession number AF382811
FSNP	Displays region of genome with HSCO Gene Nomenclature Committee identifier FSNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_056160	Displays the region of genome with protein accession number NP_056160
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs
homocysteine oxidase	Lists mRNAs for causal/homocysteine genes
zinc finger	Lists many zinc finger mRNAs
knapped zinc finger	Lists only knapped-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease
zeller	Lists mRNAs deposited by scientist named Zeller
Evans, J E	Lists mRNAs deposited by co-author J E. Evans

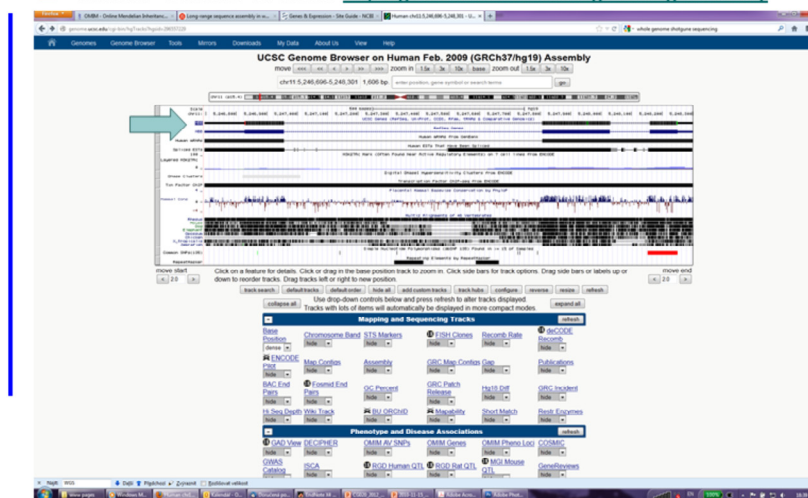


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genome Resources

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



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>

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, HbA. 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 A - delta - beta 2 - (provided by RefSeq, Jul 2005). **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. **Entrez Attributes STATED Transcribed, exon, Coordinates, evidence, Y060497.1, BC009180.1 [CCDC300002], RefSeq Attributes STATED**

Transcription Changelog: chr11:5246185-5246185. **Size:** 1555. **Start:** 5246185. **End:** 5246301. **Exon Count:** 3. **Coding Size:** 1,011. **Start:** 5246187. **End:** 5246251. **Exon Count:** 3.

Page Index: Sequence and Links, UniProtKB, Comments, Gene, Associations, CTD, Microarray, RNA Structure, Protein Structure, Other Species, GO Annotations, miRNA Descriptions, Pathways, Other Names, GeneReviews, Model Information, Methods.

Sequence and Links to Tools and Databases

Genomic Sequence (chr11:5,246,185-5,246,301) mRNA (may differ from genome)	(Protein 147 aa)
Gene Name	Genome Browser
Protein FASTA	Location
Table	Sequence Alignment
CSAP	Ensembl
Ensembl	Ensembl
Gene	Ensembl
GeneCards	GeneNetwork
Craps Tissue (H.MV)	HGNC
HPRED	Jackson Lab
MDPRED	NCBI
OMIM	PubMed
Reaction	Standard SOURCE
TrEMBL	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-LV-hemophorin.7.

FUNCTION: Involved in oxygen transport from the lung to the various peripheral tissues.

FUNCTION: LVV hemophorin 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: P10605:HBA2_NBE019; HBA3:EBI-715554; EBI-714690.

TISSUE SPECIFICITY: Red blood cells.

PTM: Globin chains form covalently with the N-terminus of the beta chain to form a stable ketamine linkage. This takes place slowly and continuously throughout the 120-day life span of the red blood cell. The rate of glycosylation is increased in patients with sickle cell anemia.

PTM: N-glycosylated, a nitric oxide group is first bound to Fe(2+) and then transferred to Cys-64 to allow capture of O(2).

PTM: N-glycosylated on Lys-60, Lys-61 and Lys-145 upon oxygen exposure. PubMed|Hirose et al. 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 (110) Method (AS) Range (33-42) Source (PubMed) 1537274.

DISEASE: Defects in HBB may be a cause of Hereditary hemochromatosis (HESAH) (OMIM 161033). This is a form of non-spherocytic hemolytic anemia of Dacie type 1. After splenectomy, which has little benefit, spherocytic inclusions called Heinz bodies are demonstrable in the erythrocytes. Before splenectomy, splenic or purpura purpurina may be evident. Most of these cases are probably instances of hemoglobinopathy. The hemoglobin demonstrates: (a) instability; Heinz bodies are observed also with the hemoglobin (hemoglobin peroxidase deficiency).

DISEASE: Defects in HBB are the cause of beta-thalassemia (B-THAL) (OMIM 101113). 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 (SCA) (OMIM 603202), 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 stiff red blood cells can lead to microvascular occlusions thus cutting off the blood supply to nearby tissues.

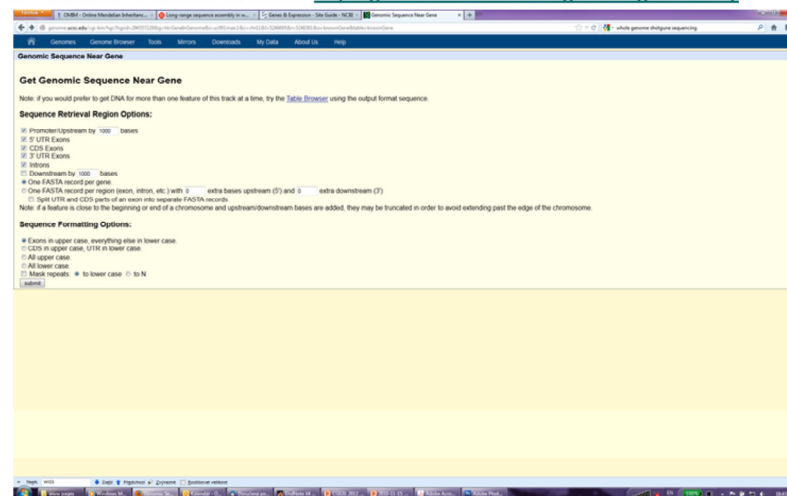


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genome Resources

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



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>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genome Resources

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

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

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

The NEW arabidopsis.org

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

Breaking News

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

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

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



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
 - Homology Searching



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

□ Global versus Local alignment

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

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

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

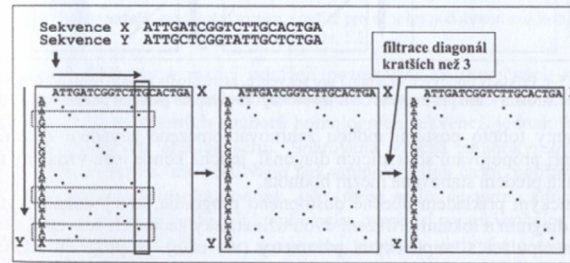


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

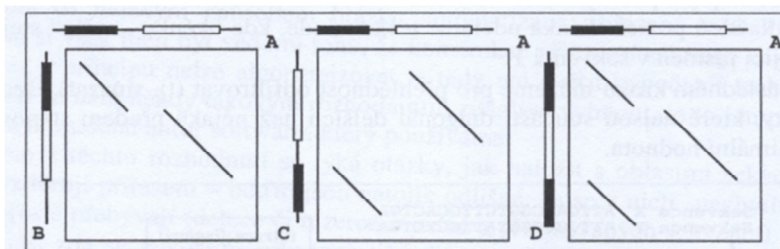


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

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

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

[Search](#)

```
aacccaccc ugu  
acaccatcat cattatcacc atcgttttgg ggcgatgttg tgggttcca  
gcytattaat  
ataattaatt tattccacat gagatgat atgatatact atgtattttt  
tttttttttt  
ttatttgtaa acotttaata taacaagaac tacaaaaaat gaaa
```

[Set subsequence](#) From: To:

[Choose database](#)

Now: **BLAST!** or [Reset query](#) [Reset all](#)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

Maticice PAM 250

	E	S	D	T	P	A	G	C	M	F	H	R	K	N	I	L	V	F	Y	W	
E	12	4	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
S	4	12	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
D	3	5	12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
T	1	1	1	12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
P	1	1	1	1	12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
A	1	1	1	1	1	12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
G	1	1	1	1	1	1	12	1	1	1	1	1	1	1	1	1	1	1	1	1	
C	1	1	1	1	1	1	1	12	1	1	1	1	1	1	1	1	1	1	1	1	
M	1	1	1	1	1	1	1	1	12	1	1	1	1	1	1	1	1	1	1	1	
F	1	1	1	1	1	1	1	1	1	12	1	1	1	1	1	1	1	1	1	1	
H	1	1	1	1	1	1	1	1	1	1	12	1	1	1	1	1	1	1	1	1	
R	1	1	1	1	1	1	1	1	1	1	1	12	1	1	1	1	1	1	1	1	
K	1	1	1	1	1	1	1	1	1	1	1	1	12	1	1	1	1	1	1	1	
N	1	1	1	1	1	1	1	1	1	1	1	1	1	12	1	1	1	1	1	1	
I	1	1	1	1	1	1	1	1	1	1	1	1	1	1	12	1	1	1	1	1	
L	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	12	1	1	1	1	
V	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	12	1	1	1	
F	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	12	1	1	
Y	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	12	
W	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	12



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

BLAST

Basic Local Alignment Search Tool



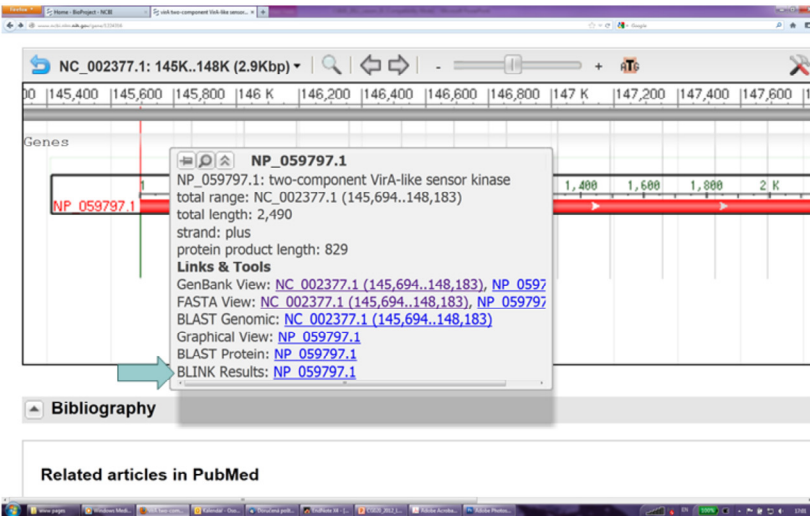
- „expectancy value“ provides the number of expected sequence number with the same or higher similarity when searching in the database consisting of randomly assembled sequences
- the results shows fraction of identical and in case of proteins also similar sequence positions and/or inserted spaces



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primary Databases



The screenshot shows a web browser displaying a GenBank record for NP_059797.1. The record is titled "NP_059797.1" and is described as a "two-component VirA-like sensor kinase". Key details include: total range: NC_002377.1 (145,694..148,183), total length: 2,490, strand: plus, and protein product length: 829. Below the description, there is a "Links & Tools" section with several hyperlinks: "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", and "BLINK Results: NP_059797.1". A green arrow points from the "BLINK Results" link to the "Bibliography" section below. The "Bibliography" section is currently collapsed. Below the bibliography, there is a section for "Related articles in PubMed".



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

Pre-computed BLAST results for: [a16119781rvf/NP_396485.1](#) two component sensor kinase [Agrobacterium tumefaciens str. C58]

Matching gis: [15163423.20141874-1019660](#)

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 138295 Bacteria 13 Metazoa 1349 Fungi 554 Plants 6 Viruses 5676 The Others [reset selection](#)

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

% hits	Score	Accession	Length	Protein Description
833 aa				
4166	AM99527	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]	
4166	P18548	833	RecName: Full=Wide host range virA protein; Short=WRB virA	
4166	AAA79262	833	virA [Plasmid pTIC58]	
4159	NP_053300	833	hypothetical protein pT1-GAMMA_p142 [Agrobacterium tumefaciens]	
4159	AAA07765	833	tiorf140 [Agrobacterium tumefaciens]	
4153	AAA91590	833	virA [Plasmid Ti]	
4153	g1173727	833	virA protein	
4153	CAA34777	833	91.3 kDa protein [Agrobacterium tumefaciens]	
3800	CAA33380	829	virA [Agrobacterium rhizogenes]	
3718	g11227240	849	virA gene	
3148	AAA88643	829	virA [Plasmid Ti]	



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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**.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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**.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

BLAST

Specialized Versions

□ Example of search by PHI-BLAST

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase  
MSHIQIPPLTELLQGYTVEVLRQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPPEPGPDR  
VADAKGDSSESEDEDELEVVPVPSRFNRRVSVCAETYNPDEEBEDTDPRVIHPKTDEQRCLQEACKDILLF  
KNLDQEQLSQVLDAMFERIVKADHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA  
LMYNTPRAAITVATSEGLWGLDRVTFRRIIVKNNAKRRKMFESFIESVPLLSLEVSRMKIVDVIGEK  
IYKDGERRITQGEKADSFYIIBSGEVSILIRSRTKSNKDGNGQEBE IARCHKGQYFGBLALVINKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYBQLVKMFGSSVDLGNLGQ
```

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



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...



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

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

Biology WorkBench
click here to toggle between menus and buttons
NE 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

BLSEQ BLSEQX BLASTN BLASTX TBLASTX FASTA FASTX FASTY SSEARCH CLUSTALW
CLUSTALWPROF ALIGN LALIGN LFASTA PATTERNMATCHDB PATTERNMATCH TACG PRIMER3
NASTATS BESTSCOR PFSCAN PRIMERCHECK PRIMER3M SIXFRAME REVCOMP RANDSEQ

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

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

The screenshot shows the 'View' section of the NCBI Workbench. It includes a 'View Nucleic Sequence(s)' header, a 'Format' dropdown menu set to 'Fasta', a 'Case' dropdown menu set to 'Upper', and a 'Change Format' button. Below the menu is a link to 'Download/View all sequences in text format'. The main content area displays the following information:

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

> 170248
GAGCTCCCTTGGGGGGCAAGGGCAAAAACCTTTTGGCTAAATGGAAAAATATATACC AAGTGTGTGTATA
GTTACTCAATTTGAATTAACAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCACAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTTATCGGCAAGTAGCTTCTTTAATTTAGTTAGTT
GACAAAACACTATCAAGATATCATTATTAATAATAAATCTCAAGTCCATCATCTTAGCTGCTCCTCA
GTTAGAGCCCGAGTAAATAGACCGATCAATAAAGCCCGCATTAATAATGAATTTTAGGACTCTC
GATTTGGACGTAGTCCAAAACCTTTCCAAATCTTTTCCGCAAGCTTTGGGCTCCGAGTCTTAGCTTC
CAGATATGGGATATTTCTAGTTTATCTCTTAATTTACATCTCAACTAATTTAAGAAATTAACAGGTA
CAGCAATCATAAAATTTCCCTTAAGGAAGCAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCAATGCCATATTCCTAAGGGGTCGTTTGGTACAGAAATTAATAATAATTTGGGATAGAAATTT
GAGATTCGATTTATCTTTGTTTAAATTAAGATTTAGCTAATTCAGATAAATTTTGCCTAATAATAG
TAAATCACTTTCACATGTAGAAAGTGAATGGATAGCTAATCCATAGCCACTCACTAGATATTC
TTATTTATCTACATTTTACC AAATGATCGTTAGTCTTCATAGAGATCCAGTATCTCAATAAATGCA
GTAGAAAGTTAGAAAATTTCTAATTAATCAATTCATATAATTTAAAATATTAGATATGGAGCTTAG
ATACATAAAGATGTACCGTTAATAATAAAGATAAGATAGATTTTAAATAGGAAAAAAAACGGTT
CGAGACTCTTTATGGGAAGGGGTGCTCTCAAGTAGATTCATTCATTTGCTCTGGTGCATAGCAAAA
TACACTTTAGCTCTTAGATTCAGCCGAGCCACTTCGATCTCTTATTTATCTCAAGTGAAGTTTA
GGAACCTTCAAACTTCAACTACTTTTAGGGAAATTCAAAATACGACCAATGTTATTTCTACTTAC
TTATAGTTAAATGATATGAATTTTAAATTTGAATTTGAAAATTTAAATTTACTTGATTTAATAATA



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

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

Regex pattern:

ott. {1, 32}ott

0 sequences were searched

1 match was found

Matches are indicated in blue

```
>170248
GAGCTCCCTTGGGGGGCAAGGGCAAACCTTTTGGCTAAATGGAAAAATATATACCAAGTGTITGTAATA
GTIATCTAATTGAAATTAACAAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAATTCGAATGGTCCATATCGCCAGGTAGCTTTCTTTTAAITTAGTITAGTT
GACAAACCTATCTACAGATATCTATTATTAATTAATTAATTTCAAGGGTCTTCTTTAGTCCCTCTCA
GTAGAGCCGCCAGTAAATAGACCGATCAANTRAAGCCGCCATTAATAATGAATTTTAGGACTCTC
GATGGCACGTAAAGTCCAAAACCTCTCCAAATCTTGGCTGCAACTTGGGGCTGTAGGTTCTGAGCTTC
CAGATATGGGATATCTAAGTTTATCTCCTAATTTACATCTCAACTAATTAAGAAATTAACAGGTA
CAGCAAKTATAAAATTTCTCTAAAGAGACAAATCCGGTTACTATTCATGSSCTTTCTAGAG
TCTGATGCCATATTCACDAAGGGGTCGTTGGTACAAGAAATAATAATAATTTCCGGATAGAAATTT
TAAATCAACTATACATGTAGAGGTGGAATGGAATAGTAATCCCATAGCCACTCACATAGAAATCC
TATTATCTACTATTTTACCAAATGATCGGTTAGTCTTCATGAAATCCAGTATCCCAATTAATGCA
GTAGAAATTTAGAAATTTTCAATTAATCAATTCATTAATTTTAAAAATTTAGATTTGGACACTTAG
ATACAATAAAGATGTACCGTAAATAAAGATAGATAGAGTTTAAATAGGAAAAAAAACCGGTT
CGAGACACTTTATGGAGGGCTTGTCTCAAGGTAGATCTCATTCATTTGCTCTGGTCAATAGCAAAA
TGACATTTACTCTTAGATACAGCGACCTCTACAACTTCTATTTGTACTTAAATGAAAGTTTAA
GAGAACTTAAATCTTCACTACTTTTAGGGAAATCAAAATACGACCAATTTATTAATTTACTAC
TTATGTTAAATGATAGAAATTTATTTAAATTTGAAATGAAATTTAAATTTAGATTTAATATAA
ACAATAGATATCGCTAAGTATTTACCACAACATGGAGATACACAGAAGATTTATTTATTTAGCAT
GATTAAGCAGCTATTCATCTGGTTGTGACAGGATGAAGAAAGTAACAGCTATTAATTTCTTTGTAAGT
```



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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
ELPFGARAKLFAKWKNIIPVCNSYSI*INKGANLTILPL

E L P W G A R A K L F A K W K N I I P S
1 g a g t c c c o t t g g g g g c a a g g g c a a a a c t t t t g c t a a a t g g a a a a t a t t a t a c c a a g t 60
V C N S Y S I * I N K G A N L T I L P L
61 g t t t g t a a t a g t t a c t c a a t t t g a a t t a a c a a a g g g g c a a a t t g a c t a t t t t g c c o t t a 120

Frame 2, 1 stop codon

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

>170248 Translated - Frame 2
SSLGGQQNPLNGKILVQVFIVTQFELTKGQI*LFCP

S S L G G Q Q N P L N G K I L V Q V F I V T Q F E L T K G Q I * L F C P
2 a g t c c c o t t g g g g g c a a g g g c a a a a c t t t t g c t a a a t g g a a a a t a t t a t a c c a a g t g 61
F V I V T Q F E L T K G O I * L F C P
62 t t t g t a a t a g t t a c t c a a t t t g a a t t a a c a a a g g g g c a a a t t g a c t a t t t t g c c o t t a 120



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

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

```
== Linear Map of Sequence:
      SbyI
      BsaJI
      CviJI
      AluI
      SacI
      EcoICRI
      Bsp1286I
      BsiHKAI
      BanII  BslI
      \ \ \ \ \
1 gagctcccttgggggcaagggaacaaacttttgcataatgaaaaatattataccaagt 60
ctcgagggaacccccctcccgtttgaaaaacgattaccttttataataggttca
      * * * * *
1 E L P W G A R A K L F A K W K N I I P S
2 S S L G G Q G Q N F L L N G K I L Y Q V
3 A P L G G K G K T F C * M E K Y Y T K C
4 L E R P P C P C F K K S F F F I N Y W T
5 S S G Q P A L A F S K A L H F F I I G L
6 L A G K P P L P L V K Q * I S F Y * V L
      \ \ \ \ \
      Tsp509I      Tsp509I
MaeIII Tsp509I  MseI      ApoI
      \ \ \ \ \
61 gtttgaatgattactcaattgaattaacaaagggaacaaattgactattttgcoccta 120
caaacattatcaatgagttaaacttaattgtttccccgttaaacgtataaacggggaat
      * * * * *
1 V C N S Y S I * I N K G A N L T I L P L
2 F V I V T Q F E L T K G Q I * L F C P *
3 L * * L L N L N * Q R G K F D Y F A L R
4 N T I T V * N S N V F P C I Q S N Q G *
5 T Q L L * E I Q I L L P A F K V I K G K
6 H K Y Y N S L K F * C L P L N S * K A R
```

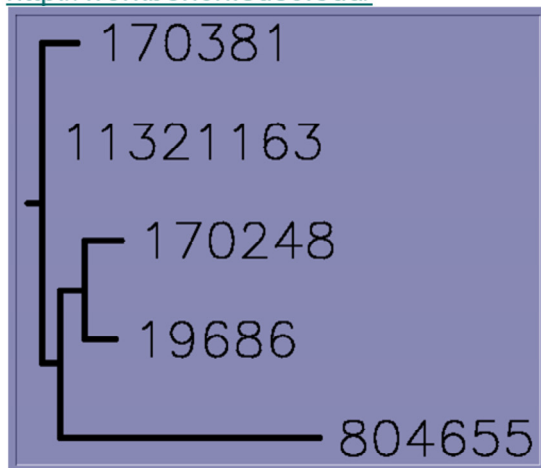


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

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



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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 instability 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: BLAST in the database for: M. musculus

Primer 1
Primer 2
Primer 3
Primer 4
Primer 5
Primer 6
Primer 7
Primer 8

Annealing temperature: 50

Do PCR! 



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytical Tools

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



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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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 | NCBI - Mozilla Firefox

PHACTR4 phosphatase and actin reg... | Digital Human Genome Browser | NCBI

NCBI | My NCBI | Sign In

Gene | Search

PHACTR4 phosphatase and actin regulator 4 [Homo sapiens]
Gene ID: 65978, updated on 27-Aug-2011

Summary

Official Symbol: PHACTR4 (provided by HUGO)
Official Full Name: phosphatase and actin regulator 4 (provided by HUGO)
Primary source: NC_000011.10
Locus tag: PF11_442124_A.1
See also: Ensembl: ENSEMBL00000204130, RefSeq: NM_020725
Gene type: protein_coding
RefSeq status: REVIEWED
Organism: HOMO SAPIENS
Lineage: Eukaryota; Metazoa; Chordata; Craniota; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhina; Catarrhini; Hominoidea; Homo

Also known as: FLJ13171, MGC25818, MGC34186, DAF-2a68L7.205, PF11_442124_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: 1:10,351
Sequence: Chromosome 1: NC_000011.10 (3894993..3895981)

Chromosome 1 - NC_000011.10

Genomic regions, transcripts, and products

Genomic Sequence: NC_000011 chromosome 1 reference GRCh37 p5 Primary Assembly

Go to nucleotide | Graphics | FASTA | Downloads

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Interactions
- General gene info
- General protein info
- Reference sequences
- Published sequences
- Additional links
- Links
- Order CDNA clones
- BioAssay, by Gene target
- BioProjects
- CCDS
- Conserved Domains
- dbMOP
- EST
- Full set in TrEMBL
- Genome
- Gene Ontology
- Homology
- Map Viewer
- Nucleotide
- ORF
- Protein
- PubChem Compound
- PubChem Substance
- PubMed
- PubMed (GenRef)
- PubMed (OMIM)



MINISTERSTVO
MLÁDEŽE

JE VZDĚLÁVÁNÍ
je spolufinancována
kým sociálním fondem
Česka a Evropské unie
řídí jej Mládež a tělesná výchova
Česka a Evropské unie

Other On-Line Genome Resources

- Online Mendelian Inheritance in Man (OMIM)



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



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Discussion



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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