

CG020 Genomika Bi7201 Základy genomiky

Přednáška 1 Úvod do bioinformatiky

Jan Hejátko

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Osnova

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Schéma předmětu

- **Kapitola 01**
 - Úvod do bioinformatiky
- **Kapitola 02**
 - Identifikace genů
- **Kapitola 03**
 - Přístupy reverzní genetiky
- **Kapitola 04**
 - Přístupy genetiky přímé



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Schéma předmětu

- **Kapitola 05**
 - Přístupy funkční genomiky
- **Kapitola 06**
 - Protein-protein interakce a jejich analýza
- **Kapitola 07**
 - Současné metody sekvenování DNA
- **Kapitola 08**
 - Strukturní genomika



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Schéma předmětu

- **Kapitola 09**
 - Lokalizace genů a genových produktů v buňce

- **Kapitola 10**
 - Genomika a systémová biologie

- **Kapitola 11**
 - Praktické aspekty funkční genomiky

- **Kapitola 12**
 - Nástroje systémové biologie
 - Modelové organismy, PCR a zásady navrhování primerů



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Literatura

- Zdrojová literatura ke kapitole I:
 - **Bioinformatics and Functional Genomics**, 2009, Jonathan Pevsner, Willey-Blackwell, Hoboken, New Jersey
<http://www.bioinfbook.org/index.php>
 - **Úvod do praktické bioinformatiky**, Fatima Cvrčková, 2006, Academia, Praha
 - **Plant Functional Genomics**, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Osnova

- Schéma předmětu
- Definice



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

GENOMIKA-co to je?

- V širším pojetí-zkoumá **STRUKTURU** a **FUNKCI** genomů
 - Předpokladem je znalost genomu (sekvenci)-práce s databázemi
- V užším pojetí zkoumá FUNKCI jednotlivých genů - **FUNKČNÍ GENOMIKA**
 - používá zejména přístupy REVERZNÍ GENETIKY



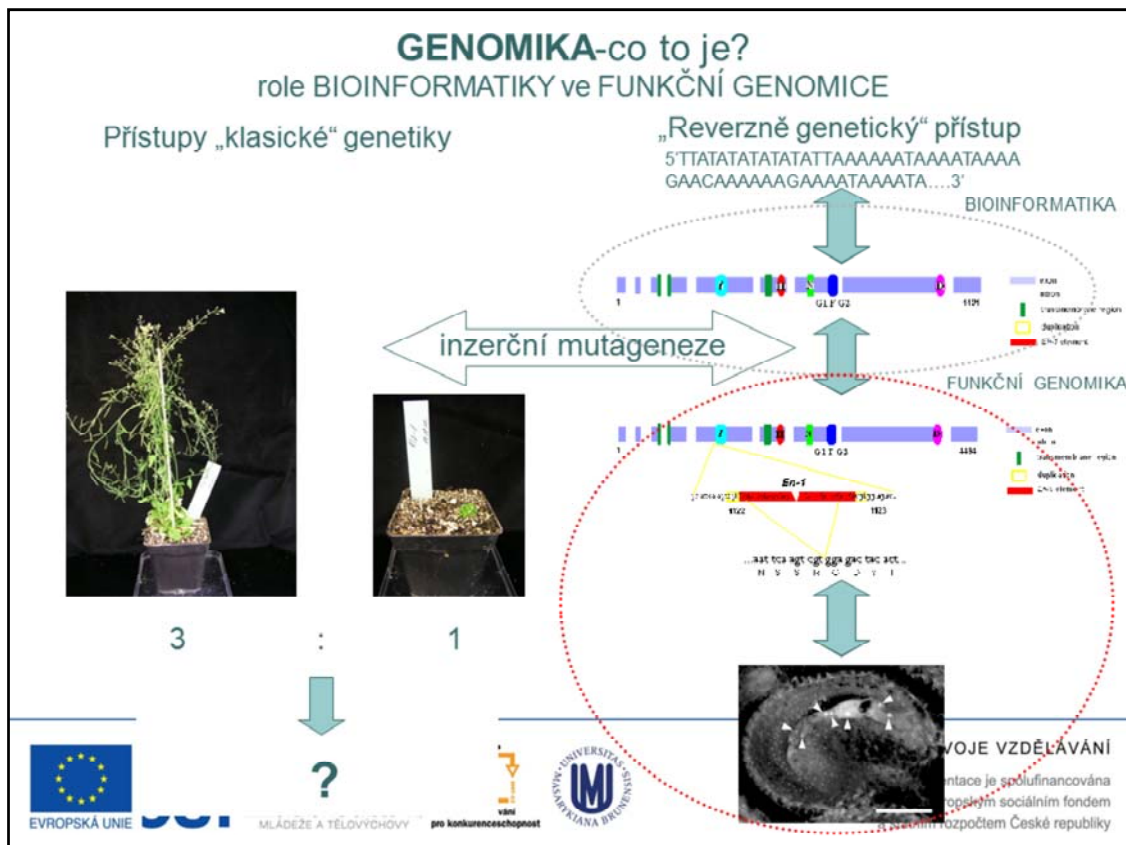
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 (prokaryotes) 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**.

Osnova

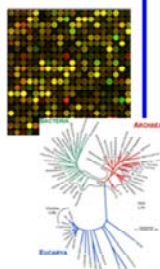
- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Bioinformatika



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

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



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

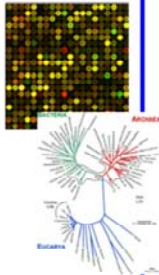
J. Pevsner,
<http://www.bioinfbook.org/index.php>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Bioinformatika



- **Bioinformatika ve funkční genomice**

- **Zpracování a analýza sekvenčních dat**

- Identifikace referenčních sekvencí
 - Identifikace genů
 - Identifikace homologů, ortologů a paralogů
 - Korelační analýzy mezi genomy a fenotypy (včetně člověka)

- **Zpracování a analýza transkripčních dat**

- Transkripční profilování pomocí DNA čipů nebo next-gen sekvenování

- **Vyhodnocování experimentálních dat a predikce nových regulací v přístupech systémové biologie**

- Matematické modelování genových regulačních sítí



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
 - Spektrum „on-line“ zdrojů



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Spektrum on-line zdrojů

EMBL National Nodes		
Vicente Blanco	Austria	http://www.at.emblnet.org/
SEN	Belgium	http://www.be.emblnet.org/
Biolinar	Denmark	http://biolinar.dk/
CSK	Finland	http://www.fi.emblnet.org/
INFODIGEN	France	http://www.infodigen.fr/
CSIC/CIEMAT	Germany	http://www.gemba.zemlin.uni-koeln.de/finwiki/
IMBB	Greece	http://www.imbb.forth.gr/
HON	Hungary	http://www.hu.emblnet.org/
IKM	Indonesia	http://www.gemba.ti.ac.id/
JIN	Israel	http://dipnet.wellmann.ac.il/fool/finn.html
JIN-ADN	Italy	http://finn-www.bio.cnr.it/8000/BIOWWW/finn-www.htm
CAJIC/CIAM	Netherlands	http://www.cajic.kun.nl/
IKU	Norway	http://www.no.emblnet.org/
ISB	Poland	http://www.plb.wzpi.pl/
SC	Portugal	http://www.lgp.gulbenkian.pt/
GemBio	Russia	http://www.gemba.msu.ru/
OND-CSC	Spain	http://www.es.emblnet.org/
BMC	Sweden	http://www.se.emblnet.org/
ISB	Switzerland	http://www.ch.emblnet.org/
SIGNET	UK	http://www.signet.ac.uk/
EMBL Specialist Nodes		
MPS	Germany	http://www.mps.biochem.mpg.de/
IGSB	Italy	http://www.igsb.infn.it/
Pharmacia Uppsala	Sweden	http://www.gnu.com/
F. Hoffmann-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	UK	http://www.embl.ac.uk/
EMBL Associate Nodes		
IBRA	Argentina	http://iut.ibt.cncp.edu.ar/emblnet
ANIGS	Australia	http://www.anigs.mq.edu.au/
CSI	China	http://www.csi.cbi.edu.cn/
CSIR	Cuba	http://ibc.cigb.edu.cu/
CSFO	India	http://halarjung.emblnet.org.in/
SARDC	South Africa	http://www.sarcsi.ac.za
EMBL Information Providers		
MSI	USA	http://www.ncbi.nlm.nih.gov/
HLH	USA	http://www.nih.nih.gov/
NIH	USA	http://www.nih.gov/



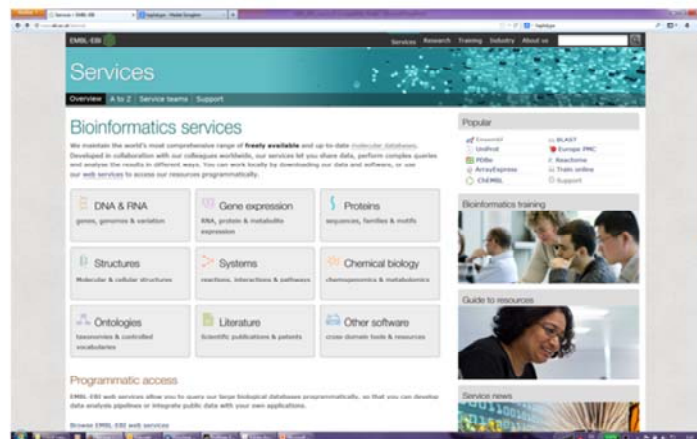
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

Spektrum on-line zdrojů

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




INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

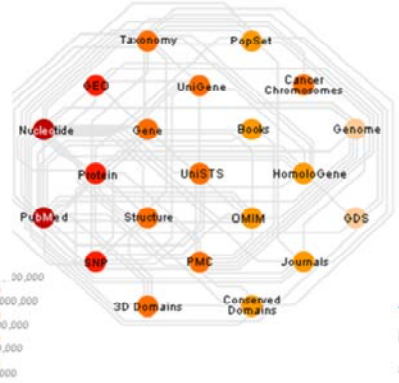
Nowadays, the resources are interconnected and could be accessed via dedicated web pages. Among the best and mostly used www resources integrating plenty of database resources belong www portal of European Bioinformatics Institute (EBI) in Europe (Germany) and National Center of Biotechnology Information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

Spektrum on-line zdrojů





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



The screenshot shows the NCBI homepage with a search bar at the top. On the left, there is a 'Resource List (A-Z)' menu with categories like 'All Resources', 'Chemicals & Bioassays', 'Data & Software', etc. The main content area includes a 'Welcome to NCBI' message, 'Get Started' instructions, and a 'Popular Resources' list. A 'YouTube' channel link is also visible.



The network diagram illustrates the interconnected nature of various biological databases and resources. Nodes are represented by colored circles of varying sizes, indicating their relative importance or volume. The nodes are interconnected by a dense web of lines, showing how these resources are linked together. The nodes include: Taxonomy, PopSet, UniGene, Cancer Chromosomes, Genome, Nucleotide, Gene, Books, HomoloGene, PubMed, Protein, UNISTs, OMM, ODS, Structure, PIR, Journals, SNP, 3D Domains, and Conserved Domains. A legend on the right indicates node sizes corresponding to values: 1,000, 10,000, 100,000, 1,000,000, and 10,000,000.

EVROPSKÁ UNIE    

evropským sociálním fondem
a státním rozpočtem České republiky

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

Osnova

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze

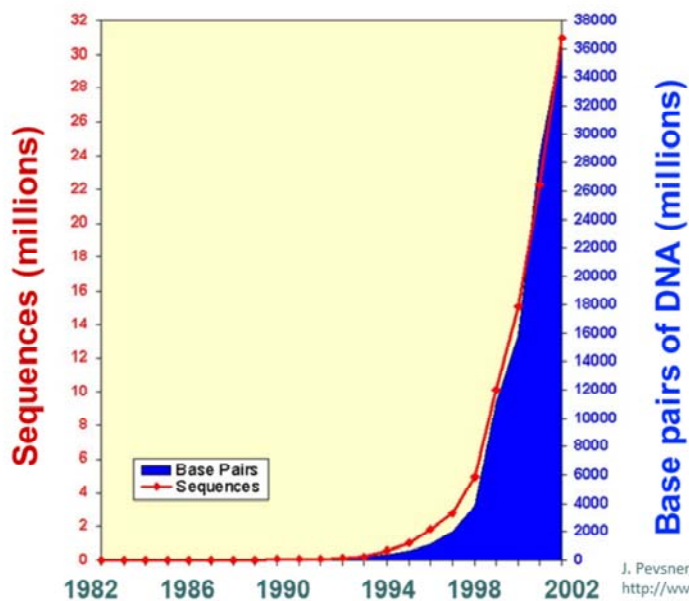
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
 - Sekvence v databázích tzv. „Velké trojky“:
 - EMBL
 - <http://www.ebi.ac.uk/embl/>
 - GenBank,
 - <http://www.ncbi.nih.gov/Genbank/GenbankSearch.html>
 - DDBJ,
 - <http://www.ddbj.nig.ac.jp>
 - denně vzájemná výměna a zálohování dat
 - velká datová náročnost (kapacita i software)
 - září 2003 $27,2 \times 10^6$ záznamů o zhruba 33×10^9 bp
 - srpen 2005 100×10^9 bp ze 165.000 organismů



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Growth of GenBank



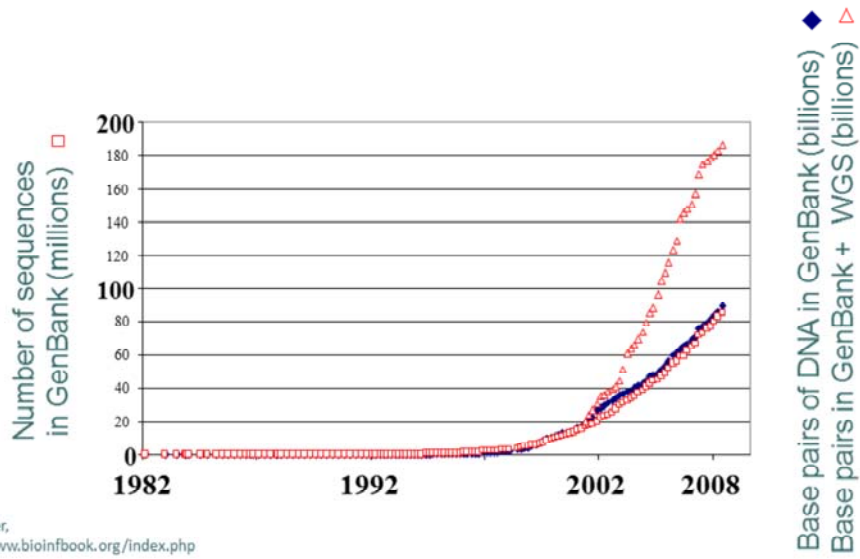
J. Pevsner,
<http://www.bioinfbook.org/index.php>



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>

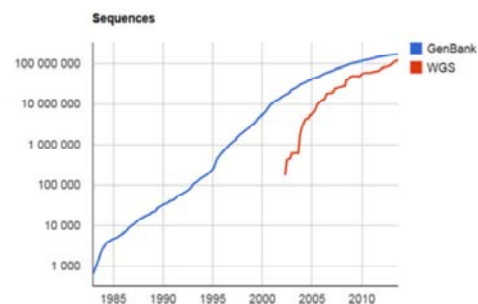
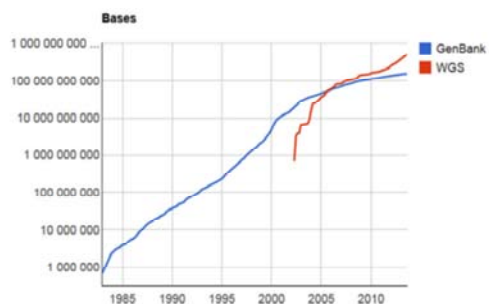


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Growth of GenBank

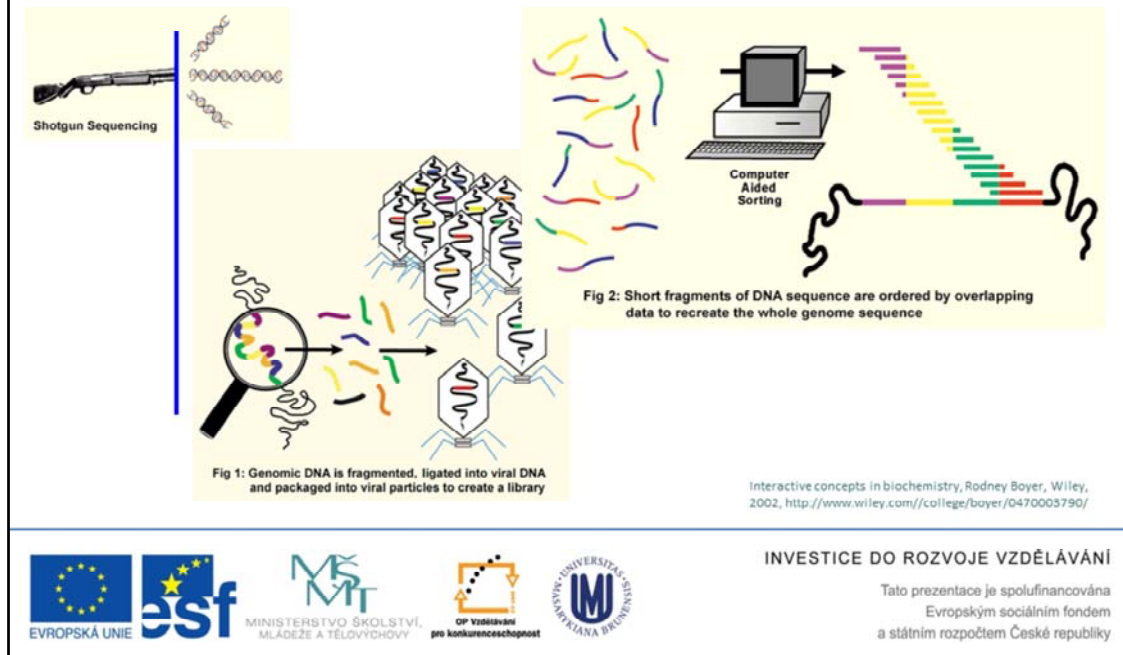
Feb 15 2013



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

WGS

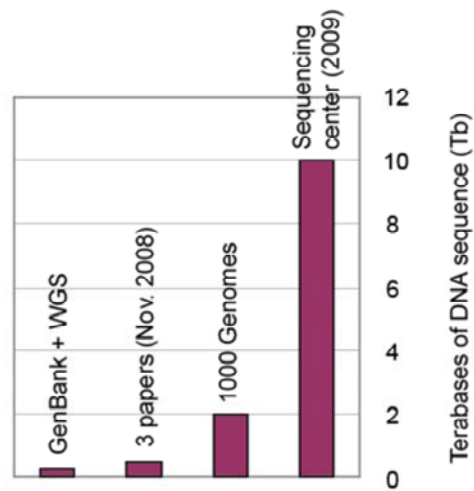


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic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

**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>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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.

Primární databáze

- zahrnují soubory primárních dat – sekvencí DNA a proteinů
 - Proteinové sekvence:
 - PIR, <http://pir.georgetown.edu/>
 - MIPS, <http://www.mips.biochem.mpg.de>
 - SWISS-PROT, <http://www.expasy.org/sprot/>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze

- Typy sekvencí v primárních databázích
 - standardní nukleotidové sekvence získané kvalitním sekvencováním
 - **ESTs (Expressed Sequence Tags)**
 - **HGTS (High Throughput Genome Sequencing)**
 - neanotované „surové“ výsledky sekvenačních projektů
 - referenční sekvence anotovaných genomů
 - **TPAs (Third Party Annotation)**
 - sekvence anotované jinými než původními autory



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze

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



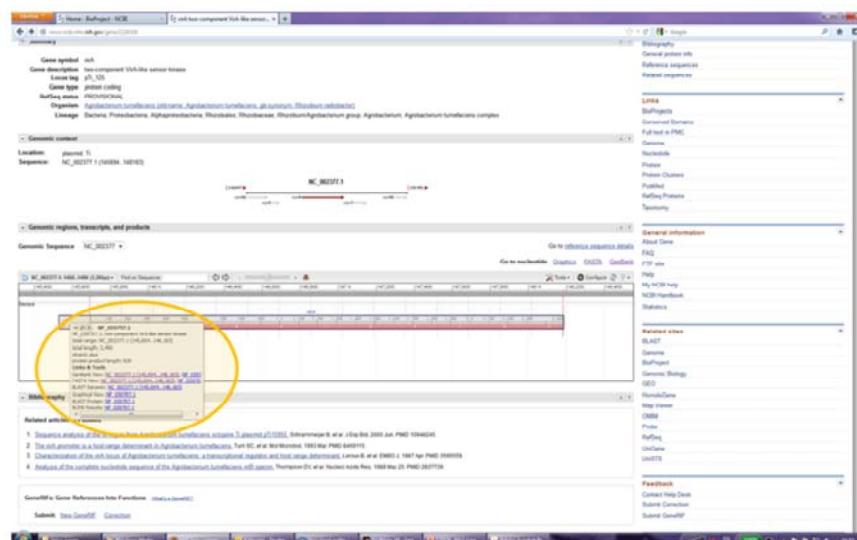
The screenshot shows the NCBI homepage with a navigation menu on the left, a search bar at the top, and a main content area with sections for 'Welcome to NCBI', 'Get Started', and 'NCBI YouTube channel'. The 'Get Started' section lists tools, downloads, how-to guides, and submissions. The 'NCBI YouTube channel' section promotes video tutorials. The footer of the screenshot includes a navigation bar with numbers 1-8.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze

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

Genes

NP_059797.1

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

Links & Tools

GenBank View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1 \(145,694..148,183\)](#)
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

Primární databáze

The screenshot shows the NCBI Nucleotide search interface. The search term 'vira' is entered in the search box. The results page displays the following information:

- Accession:** NC_022377.1
- GeneBank Identifier:** NC_022377.1
- Gene:** vira
- Organism:** *Agrobacterium tumefaciens*
- Accession:** NC_022377.1
- GeneBank Identifier:** NC_022377.1
- Gene:** vira
- Organism:** *Agrobacterium tumefaciens*

The screenshot also shows the NCBI logo, the search bar, and various navigation options like 'Clipboard' and 'Details'.



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

Page 27

NCBI's important RefSeq project: best representative sequences

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

RefSeq identifiers include the following formats:

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

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

RefSeq

The screenshot displays the NCBI RefSeq interface for the gene **two component sensor kinase** (NC_003065.3) in *Agrobacterium tumefaciens* str. C58. The page is divided into several sections:

- NCBI Reference Sequences (RefSeq):** Shows the genomic assembly with coordinates 18031..18332.
- mRNA and Protein(s):** Lists the mRNA sequence (U01700) and the protein product (PRK13027).
- Conserved Domains:** Identifies two domains: **HATPase_C** (Histidine kinase-like ATPase) and **HAKA** (Histidine Kinase A dimerization/phosphoreceptor domain).

The HATPase_C domain is described as a family that includes several ATP-binding proteins for enzymes like histidine kinase, GNA gyrase B, topoisomerases, and head shock protein HSP90. The HAKA domain is described as a dimer that forms through parallel association of 2 domains creating a-helical bundles, usually containing a conserved His residue and activated via ...



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

Primární databáze

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

Genes

NP_059797.1

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

Links & Tools

GenBank View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1 \(145,694..148,183\)](#)
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

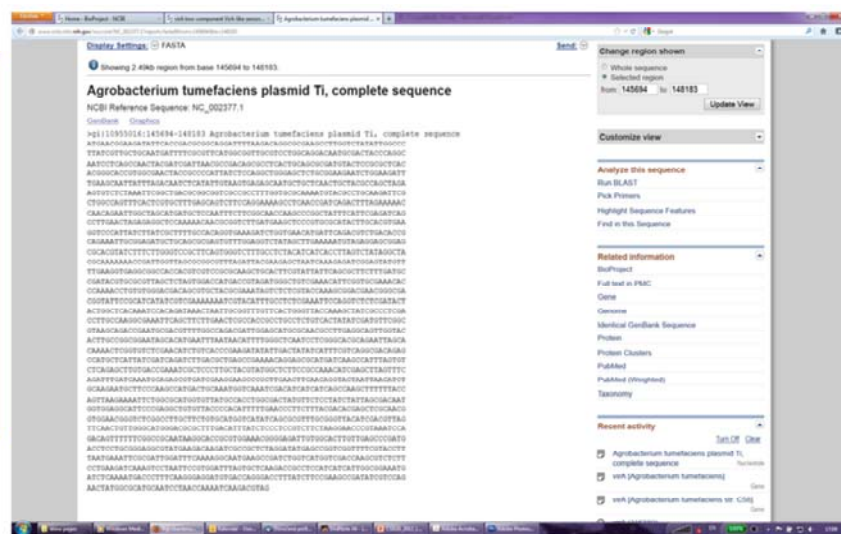
Related articles in PubMed



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Sekundární databáze

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

The screenshot shows the ScanProsite web interface. At the top, there is a navigation menu with links: 'Go to FASTA Basic page', 'Site Map', 'Search LAFAs', 'Contact us', 'Basic-Pro', 'PROSITE', and 'Evolutionary links'. Below the menu is a search bar with a 'Search' button. The main content area is titled 'ScanProsite' and contains two search forms. The left form is 'Scan a protein for PROSITE matches' and the right form is 'Search Basic-Pro with a PROSITE entry'. Both forms have 'SEARCH THE SCAN' and 'SEARCH' buttons.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Sekundární databáze

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

>PDOC00002 P00002 SULFATION Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].

171 - 185 `...S...T...S...`

>PDOC00004 P00004 CAMP_FIBOSPHO_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

744 - 747 `...D...T`

814 - 817 `...K...G`

>PDOC00005 P00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

149 - 150 `...K...R`

164 - 166 `...T...D`

171 - 173 `...K...R`

219 - 221 `...K...R`

349 - 371 `...T...D`

440 - 442 `...K...R`

513 - 515 `...K...R`

545 - 547 `...K...R`

692 - 694 `...T...D`

412 - 414 `...K...R`

716 - 718 `...K...R`

726 - 728 `...K...R`

747 - 749 `...K...R`

794 - 794 `...K...R`

814 - 816 `...K...R`

884 - 886 `...K...R`

849 - 870 `...K...R`

921 - 923 `...K...R`

957 - 959 `...K...R`

1000 - 1042 `...T...D`

974 - 974 `...K...R`

997 - 999 `...K...R`

1000 - 1004 `...T...D`

1019 - 1020 `...K...R`

1031 - 1033 `...T...D`

1119 - 1121 `...K...R`



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Sekundární databáze

- databáze funkčních nebo strukturálních *motivů* získaných srovnáním primárních dat (sekvencí)
- 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 characterize a protein family; its diagnostic power is refined by iterative scanning of a PRINTS-PROFIT database. Usually, the motifs do not overlap, but are conserved along a sequence. Though they may be conserved in 2D space, fingerprints can enable protein links and functionalisation more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. [Reference](#)

New:

- [PRINTS](#) - Search PRINTS 3-dimensional PRINTS
- [seqPRINTS](#) - Search PRINTS sequence alignments
- [PRINTS](#) - Search the integrated InterPro family database

Direct PRINTS access:

- [PRINTS](#) home page
- [PRINTS](#) search
- [PRINTS](#) help
- [PRINTS](#) manual
- [PRINTS](#) FAQ
- [PRINTS](#) contact
- [PRINTS](#) news
- [PRINTS](#) updates

PRINTS search:

- Search PRINTS with [NEW](#) [EuroPRINTS](#)
- [PRINTS](#)
- [seqPRINTS](#)
- [PRINTS](#) home page and source are available: print@bioinf.man.ac.uk

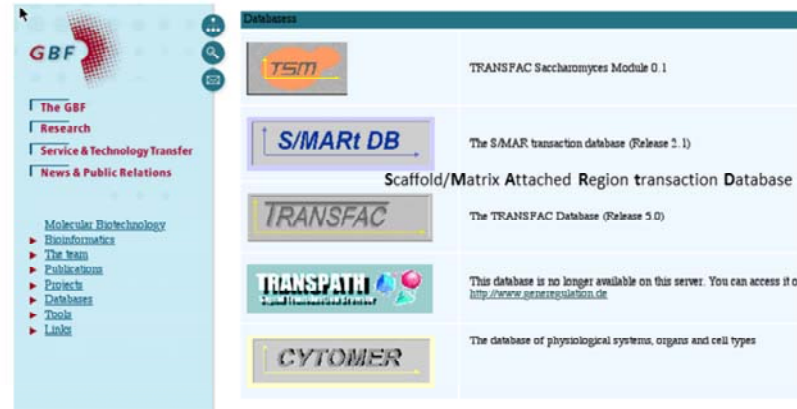


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Sekundární databáze

- 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)
TRANSFACt	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>)

Strukturální databáze

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

The screenshot shows the PDB website with the following sections:

- DEPOSIT data**, **DOWNLOAD files**, **Browse LINKS**, **BETA TEST new features**, **BETA RELEASE files**
- Current Holdings**: 19523 Structures, Last Update: 30-Dec-2002, PDB Statistics
- Molecule of the Month**: Cyclochrome c
- Search the Archive**: Search box with options for "query by PDB id only", "match exact word", and "remove sequence homologues". Includes a "Find a structure" button and a "Query Tutorial" link.
- PDB Mirrors**: List of mirror sites including San Diego Supercomputer Center, Rutgers University, National Institute of Standards and Technology, Cambridge Crystallographic Data Centre, etc.
- News**: Section dated 23-Dec-2002 with a "Happy Holidays from the PDB!" message.
- OTHER SITES**: A list of additional resources.



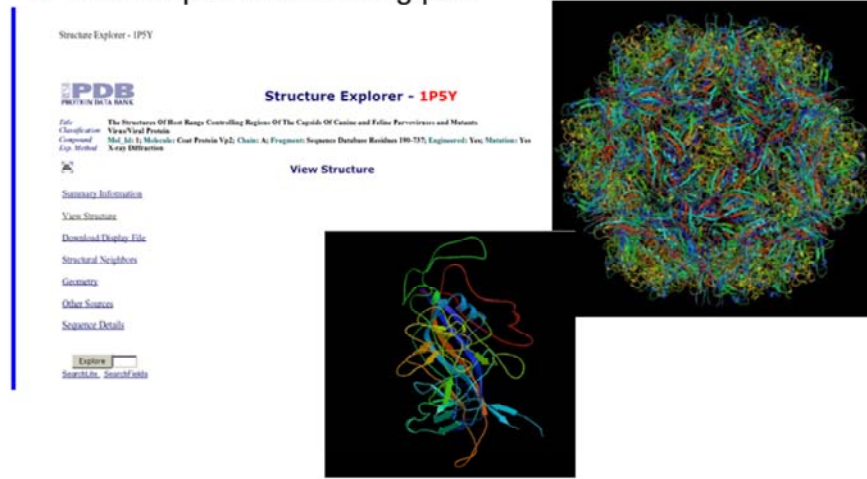
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Strukturální databáze

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

Structure Explorer - 1PSY



PDB
PROTEIN DATA BANK

Structure Explorer - 1PSY

Title: The Structures Of Hot Spots Controlling Regions Of The Capsids Of Coxsackie and Echo Parvoviruses and Mutants
Classification: Virus/Viral Protein
Compound: Mol. Wt. 1, Molecular: Coat Protein Yp2, Chain: A; Fragment: Sequence Database Residues 100-125, 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

Structure

Search/Find

<http://www.rcsb.org/pdb/cgi/structure.cgi?db=alphabeta&id=1PSY&page=pad-1734106432944&no=1&cut=show&size=800> 12/20/2003

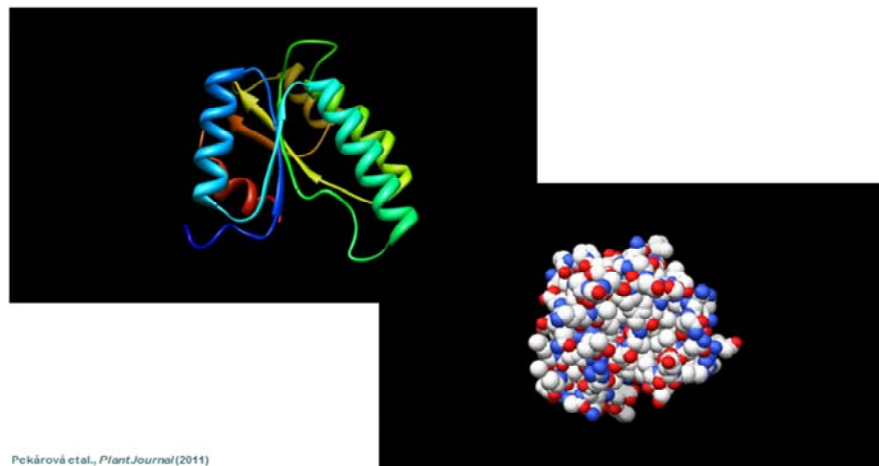


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Strukturální databáze

- 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

Osnova

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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

The screenshot displays the UCSC Genome Browser interface. At the top, there is a search bar with fields for 'chr', 'genome', 'assembly', and 'position'. Below this, the 'Human Genome Browser - hg19 assembly (GRCh37)' section is visible. It includes a 'Sample position queries' section with a list of genomic coordinates and their corresponding features. The list includes items like 'chr7', 'chr16_p11.21', '20p13', 'chr11_100000', 'chr3_100000-200', '841881-846171', '15q11-15q12', 'rs154322', 'D10S246', 'AKO2414', 'AC020101', 'AF032111', 'P19P', 'NM_017414', and 'NP_054440'. Each item has a brief description of the feature, such as 'Displays all of chromosome 7' or 'Displays the region of genome with RefSeq identifier NM_017414'. The interface also includes a 'UCSC' logo and a 'Home' link.

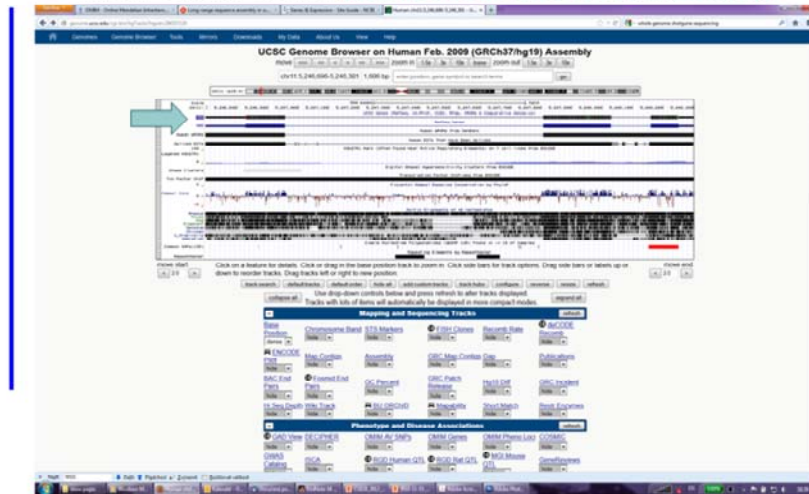


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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

The screenshot displays the Human Genome Browser interface for the HBB gene. The page includes a description of the gene, a table of sequence and links to tools and databases, and a section for comments and description text from UniProtKB. A green arrow points to the 'Sequence and Links to Tools and Databases' section.

Category	Tool/Database
Genomic Sequence	1,246,895-1,246,931 ntRNA (copy after from genome) Protein (147 aa)
Gene Name	Gene Browser, General Browser, Protein FASTA, Location, Taxonomy, UniProt
CGAP	Essays, Ethical Gene, Expression, GeneCards, GeneNetwork
Chain Tracer	Trinity, HMG, HMG, Jackson Lab, MGPED
Gene	Pubmed, BioSuite, Standard SOURCE, TrEMBL, UniProt
Mapview	

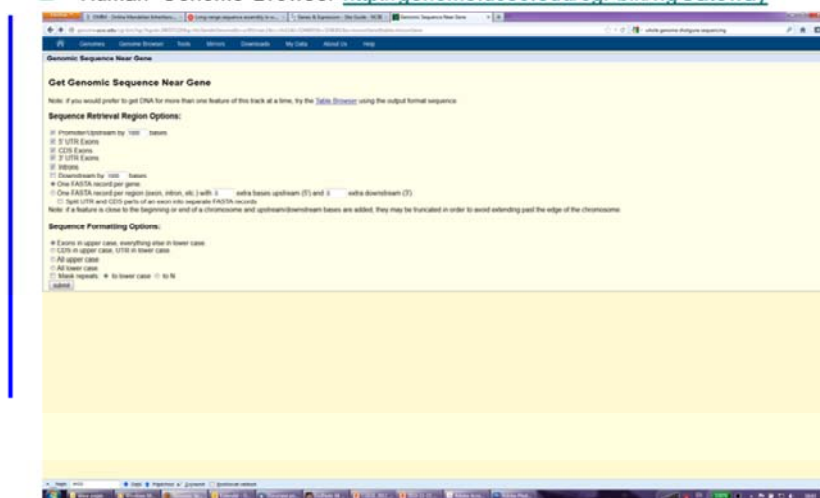


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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

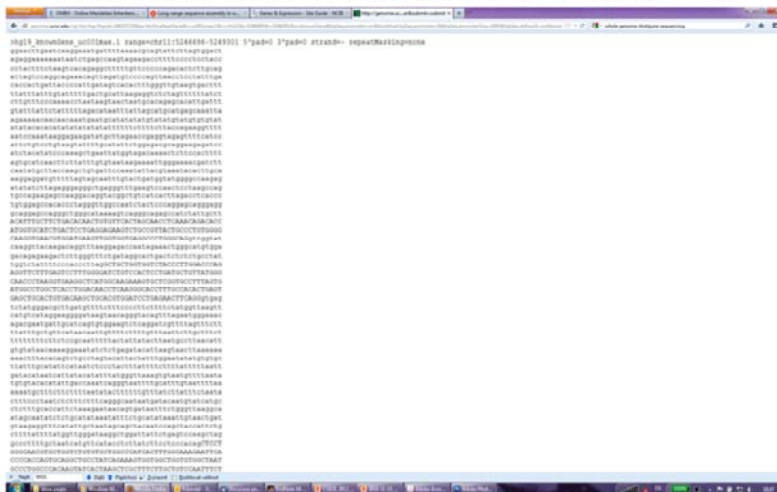


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

- 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, metabolite, 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, germplasm, 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

Osnova

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

□ Globální vs. lokální přiřazení

Globální přiřazení

```
SLAV-----APATNIK-----PIQNYR-I-----AKSETQRYMVE  
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRVEHVRATAKSETQRYMVE
```

Lokální přiřazení

```
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRVEHVRATAKSETQRYMVE  
-----NAPATNIKSECVRA-PIQNYRVEHVRA-----
```

Cvrčková, Úvod do praktické bioinformatiky

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

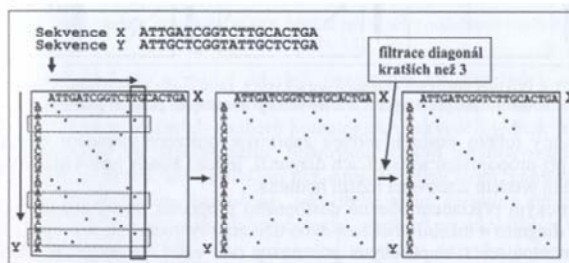


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)



Cvrčková, Úvod do praktické bioinformatiky

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

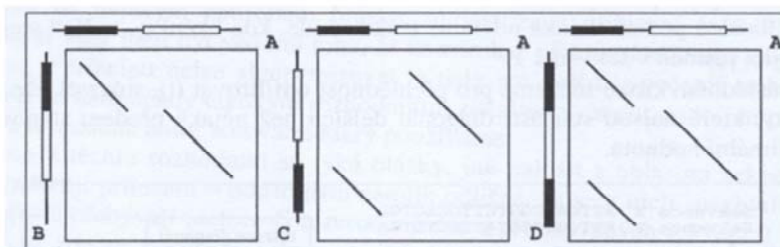


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- příklady srovnání sekvencí pomocí bodového diagramu



Cvrčková, Úvod do praktické bioinformatiky

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

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

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

[Search](#)

```
acaccatcatt cattatcaco atcgttttgg ggcgatgttg tctgggtcca  
gggtattaat  
ataattaatt tattccocat gagatgatg atgatatact atgtattttt  
tgtttttttt  
ttatttgtaa acotttaata taacaagaac taacaaaaat 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



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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze

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

Genes

NP_059797.1

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

Links & Tools

GenBank View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1 \(145,694..148,183\)](#)
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

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: `u16119781v0NP_396486.1` two component sensor kinase [Agrobacterium tumefaciens str. C58]

Matching gis: `15163423.20141874.1019662`

Total (score > 100) : 147086 hits in 148754 proteins in 6309 species

Selected: 147086 hits in 148754 proteins in 6309 species Filter: Min Score= 100 |

Other views (Reports): [Taxonomy report](#) | [Multiple Alignment](#) | [Blast](#) | [Reset all filters](#)

Choose Display Options

1293 Archaea
 138295 Bacteria
 13 Metazoa
 1340 Fungi
 954 Plants
 6 Viruses
 5676 The Others
 [reset selection](#)

Results: 1 - 100 Next Page Last

Hit	Score	Accession	Length	Protein Description
4164	833	AA099827	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]
4166	833	P18545	833	SecYemE: Full-wide hist range vWA protein Short-WID vWA
4166	833	AA179262	833	vWA (Flaximid pTIC58)
4153	833	NP_203200	833	hypothetical protein pT1-DMPWA_g142 [Agrobacterium tumefaciens]
4153	833	AA017368	833	tsorF140 [Agrobacterium tumefaciens]
4153	833	AA015590	833	vWA (Flaximid T1)
4153	833	u1173127	833	vWA protein
4153	833	AA141777	833	Yi-3 vWA protein [Agrobacterium tumefaciens]
3800	828	AA030390	828	vWA [Agrobacterium thiosphaera]
3718	849	u1227149	849	vWA gene
3163	828	AA036643	828	vWA (Flaximid T1)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

BLAST

Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
 - vyhledávání podle zdroje (organismu) sekvencí, např. známých genomů mikroorganismů
 - **BLASTP**
 - vyhledávání podobnosti k proteinu v databázi proteinových sekvencí
 - **BLASTN**
 - vyhledávání podobnosti k nukleotidové sekvenci v databázi nukleotidových sekvencí
 - další varianty jako např. MEGABLAST pro identifikaci totožných nebo velice podobných sekvencí (vyhledává dlouhé podobné úseky nukl. sekvencí)
 - **BLASTX**
 - vyhledávání podobnosti k proteinu v databázi nukleotidových sekvencí přeložených do sekvence aa



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

BLAST

Specializované verze

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

BLAST

Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST
 - **PSI-BLAST (Position-Specific Iterated Blast)**
 - Prvním krokem je standardní BLAST, při kterém PSI-BLAST identifikuje skupinu podobných sekvencí s E hodnotou lepší než minimální hodnota (standardně 0,005)
 - PSI-BLAST vytváří pro každé přiřazení tzv. PSSM (position specific substitution matrix)
 - PSSM matice zohledňuje výskyt jedné aminokyseliny ve stejné pozici se zvýšenou frekvencí u sekvencí identifikovaných jako podobné v prvním kole pomocí BLAST, což může znamenat funkční konzervovanost



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

BLAST

Specializované verze

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

BLAST

Specializované verze

□ Příklad vyhledávání pomocí PHI-BLAST

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase  
MSHIQIPPGLELLQGYTVEVLRQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPPEPGPDR  
VADAKGDSSESEDELDLEVPVPSRFNRRVSVCAETYNPDEBBEDTDPRVIHPKTDEQRCRLQBACKDILLF  
KNLDQEQLSQVLDAMFERIVKADHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA  
LMYNTPRRAATIVATSEGSWGLDRVTFRRIVKNNAKRRKMFESFIESVPLLSLEVSRMKIVDVIGBK  
IYKDGRIITQGEKADSFYIIBSSEVSLIRSRTKSNKDGNGQVEIARCHKGQYFGELALVINKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYEBQLVKMFGSSVDLGNLGG
```

```
[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

Osnova

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

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

The screenshot shows the NCBI Workbench interface. At the top, there is a 'View' button and a label 'View Nucleic Sequence(s)'. Below this, there are dropdown menus for 'Format' (set to 'Fasta') and 'Case' (set to 'Upper'), along with a 'Change Format' button. A link 'Download/View all sequences in text format' is visible. 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
GAGCTCCCTTGGGGGGCAAGGGCAAAAACCTTTTGGCTAAATGGAAAAATATATACCAAGTGTGTAATA
GTTACTCAATTTGAATTAACAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCACAAAAC
ATAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTCTTTAATTAAGTATAGTT
GACAAAACACTATCAAGATATCATTATATAATAAATACTTCAAAAGTCCATCATCTTAGCTGCCCTCA
GTTAGAGCCGCCAGTAAATAGACCAGATCAATAAAGCCGCCATTAAATATGAAATTTAGGACTCTC
GATTGGCACTGTAATGTCCTCAAACTTCTCAATCTTTGCTGCAAGCTTGGGGCTGCTAGGCTTCAAGCTC
CAGATATGGGATATTTCAAGTTATCTCCAAATTTACATCTCACTAATTAAGAAATTAACAGGTA
CAGCAAAATCAAAAATTTCCCTCAAGAAAGACAATGAATCCGGTACTGATTCATTGGCTTTTCAGAG
TCATGATCCCATATTTACTAAGGGGTCTGTTGGTACAGAAATTAATAATAATTTCCGGATAGAAATTT
GAGATGCAATTTATCTTGTGTTTAAATTAAGTATTAGCTAATTTCAAGATAAATTTGACTAATAATAG
TAAATCACTATCACATGTAAGGTTGAATGGAAATGCTAATCCCATGGCCACTCAATAGAAATATCTC
TTATTTATCTTCTCAATTTTACCAGAGGATCGTTAGTCTTCTGAGAAATCCAGTATCTCAATAAATCA
GTAGAAATTGAGAAATTTCTATTAATCAATTCATATAATTTAATAATTTAGATATGGAGCACTTAG
ATCAATAAAGATGTACCCTTAATAATAAAGATAAGATAGATTTTAAATAGGAAAAAAAACGGTT
CGAGACTCTTTATGGAGGCGTTGCTTCAAAAGTAGATTTCTATTCATGCTCTGGTGCATAGCAAAA
TACACTTTCCTTTAGATGTCAGCGAGCGACTTGCATCTCTATTTATATCTCAAGTAAAGTTTAA
GAGACTTCAAAATCTTCAACTCTTTTAGGGAAATCAAAATACGACCAAGTTTATCTTCACTTAC
TTATAGTTAATGATATGAAATTTATTTTAAATTTGAATGAAAATATAAATTAATCTGATTTAATATA



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- 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
GAGCTCCCTTGGGGGGCCAAAGGCCAAACTTTTGGCTAAATGGAAAAATATATACCAAGTGTGTTGTAATA
GTTACTCAATTGAAATTAACAAAGGGGCCAAATTTGACTATTTTGGCCCTTATATCTTTTGGTACAAAAAC
ATAAATATCCCATCCGAAATTCGAAATGGTCCATATTCGGCAGGTAGCCTTCTTTTAAITTAGITAGIT
GACAAACCTATCTAGAGATCTATTATTATTAATATAA-TTCAAAAGTTCATCTTACCTGGTCCCTCA
GTAGGCCGCCAGTAAATAGACCGATCAANTAAAGCCGCCATTAATAATGAAATTTAGGACTCTC
GATGGCACGTAAAGTCCAAAACCTCTCCAAACTTTGGCTCAACTTTGGGGCTGCTAGGTCTGAGCTTC
CAGATATGGGATATTCAGGTTTATCTCTAATTTACATCTCAACTAATATTAAGAAATTAACAGGTA
GAGCAAKCTATAAAATTTCTCTAAAGAGACAAATTAATCCGGTACTTAACTTACTGGCCCTTTCTAGAG
TCTGATGCCATATTCCTAAGGGGTCGTTTGGTCAAGAAATAATAATAAATTTCCGGATAGAAATTT
GAGATTCGACTTATCTTGGTGTAAATATAAGTATTAGCTAATTCAGATAAATTTTACACTAAATAG
TAAATCAACTATCACATOTAGAAAGTGGAAATGGATAGCTAATCCCATAGCCACTCACATAGAAATCTC
TATTATCTCTACTATTTTACCMAATGATCGTTAGTCTTCTATGAAATCCAGTATCTTCAATTAATGCA
GTAGAGATTGAAAAATTTTCTAATCAACTTCTATTATTTTAAAAATTTAGATTTGGGACCTTAG
ATCAATAAAGATGZACCGTAAATAAAGATAGATAGAGTTTAAATAGGAAAAAAAACCGTT
CGAGCACTTTATGGAAAGCCGTGTCTCAAGGTAGATCTCTATTCATPGCTCTGGTSCAATAGCAAAA
TGACATTTACTCTTAGATTCAGCGAGCCACTCTACAACTTCTATTGTATACTCAAAATGAAAGTTTAA
GAGAACTCTTAAATCTCTCACTACTTTTAAAGGAATTCAAAATCAGCCAAATTTTATTA-TTAC-TTAC
TATTGGTAAATGATGGAACTTATTTTAAATTTGAAATGAAATTTAAATCTTGAATTAATGAA
ACAATAGATATCGCTAAGTATTTACCACAACATGGAGATCTACAGAAATTTTATTATTGTAACGAT
GATTAAGCAGCTATTCATCTGGTGTGCGAGGATGAAAGAAATTAAGCTATAATTA-TTGTATAAGT
```



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- 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
ELPWGARA KLF A K W K N I I P S

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

Frame 2, 1 stop codon

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

>170248 Translated - Frame 2
SSLGGQQN F L L N G K I L M Q V F V I V T O F E L T W G Q I * L F C P

S S L G G Q Q N F L L N G K I L Y Q V
2 agtcccttgggggcaagggaacttttgcataatgaaaaatattataccaagt 61
F V I V T O F E L T K G O I * L F C P
62 tttgtaatagttactcaattgaaatacaaaagggcaaatgactatattggccotta 120



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

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

```
— Linear Map of Sequence:
      SbyI
      BsaJI
      CviJI
      AluI
      SacI
      EcoICRI
      Bsp1286I
      BsiHKAI
      BanII BslI
      SspI
1 gagctcccttgggggcaagggaacttttggtaaatggaaaatattataccaagt 60
ctcgaggggaacccccctcccgtttgaaaaagatttacccttttataataggttca
* * * * *
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 O 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 N 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
      MaeIII Tsp509I MseI
      Tsp509I
      ApoI
61 gtttgtaatgattactcaattgaattaacaagggaactttgactattttgcoccta 120
caaacattatcaatgagtttaacttaattggttccocgttaaacgtataaacggggaat
* * * * *
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 P R L F 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
```

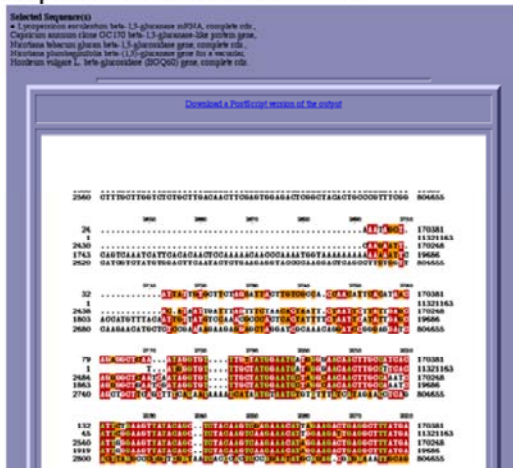


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

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

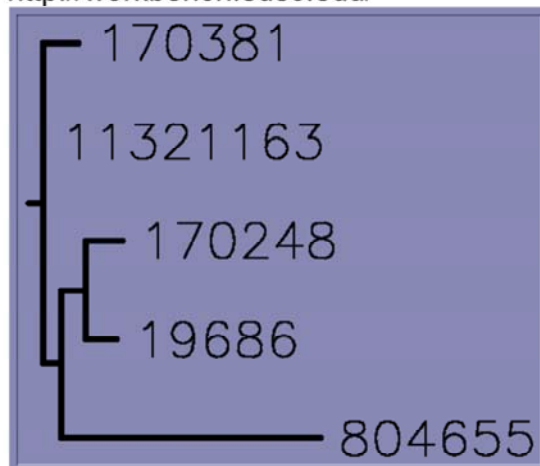


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

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

SEARCH ABOUT DOWNLOAD LINKS

VPCR 2.0 (WWW interface) - Please, enter nucleotide primer sequences (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 stability of our server 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 almost 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 operations you have for VPCR 2.0 site. Certainly, 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!



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- 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

Osnova

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Další WWW zdroje

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

The screenshot displays the NCBI Gene database entry for PHACTR4 phosphatase and actin regulator 4 (Homo sapiens). The page is organized into several sections:

- Summary:** Provides key information such as the Official Symbol (PHACTR4), Official Full Name (phosphatase and actin regulator 4), and Gene ID (65978).
- Genomic context:** Shows the gene's location on Chromosome 1, NC_000001.10, with a map of the region and coordinates.
- Genomic regions, transcripts, and products:** Lists various transcripts and protein products associated with the gene.



MINISTERSTVO
MLÁDEŽI

JE VZDĚLÁVÁNÍ
je spolufinancována
kým sociálním fondem
České republiky

Další WWW zdroje

▪ Online Mendelian Inheritance in Man (OMIM)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Shrnutí

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Diskuse



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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