

# 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](mailto:hejatko@sci.muni.cz), [www.ceitec.muni.cz](http://www.ceitec.muni.cz)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Osnova

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



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**
  - Moderní postupy funkční genomiky
- **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



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



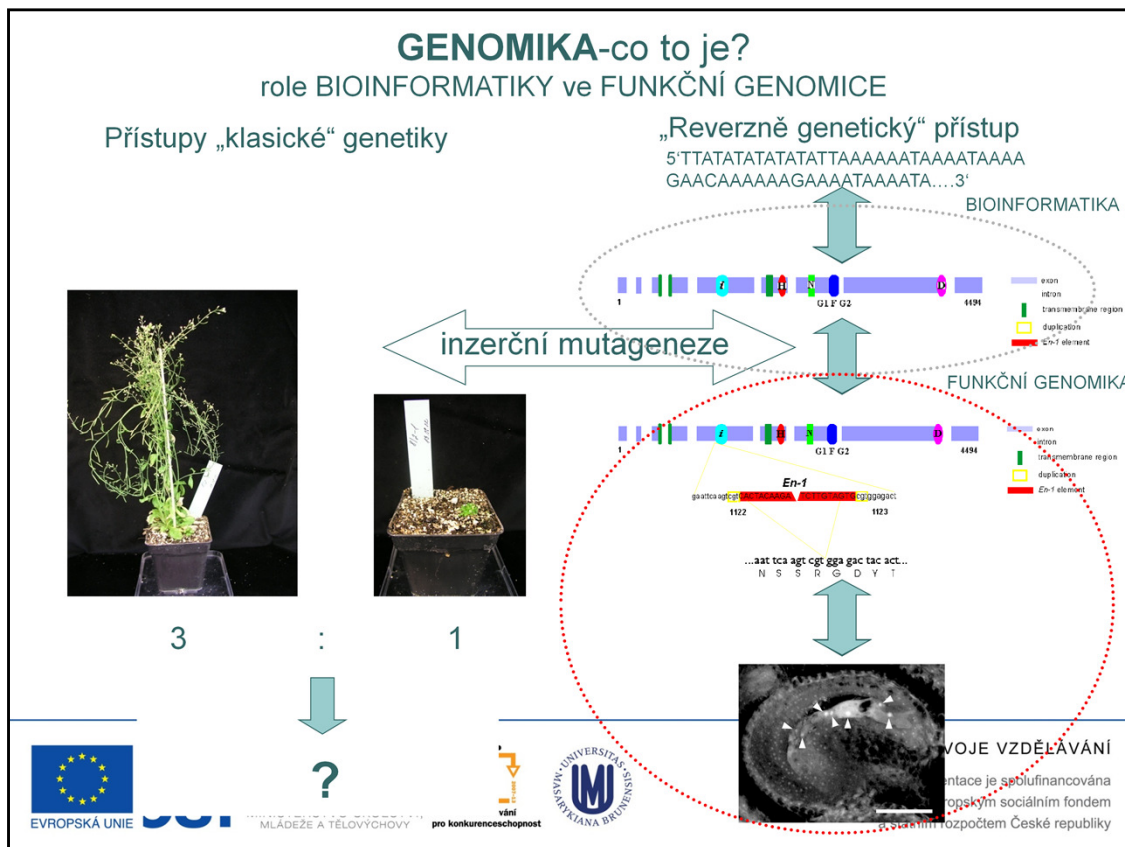
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

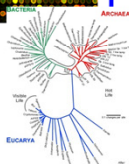
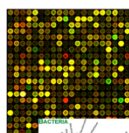
- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY



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>

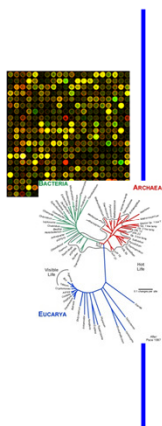


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



# Bioinformatika



- **Bioinformatika ve funkční genomice**
  - **Zpracování a analýza sekvenačních dat**
    - Identifikace referenčních sekvencí
    - Identifikace genů
    - Identifikace homologů, ortologů a paralogů
    - Korelační analýzy mezi sekvencemi a fenotypy (včetně člověka)
  - **Zpracování a analýza transkripčních dat**
    - Transkripční profilování pomocí DNA čipů nebo next-gen sekvenování
  - **Vyhodnocování experimentálních dat a predikce nových regulací v přístupech systémové biologie**
    - Matematické modelování genových regulačních sítí



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

<b>EMNET National Nodes</b>		
Vienna BioCenter	Austria	<a href="http://www.at.embnet.org/">http://www.at.embnet.org/</a>
BEIN	Belgium	<a href="http://www.be.embnet.org/">http://www.be.embnet.org/</a>
BioBase	Denmark	<a href="http://biobase.dk/">http://biobase.dk/</a>
CSC	Finland	<a href="http://www.fi.embnet.org/">http://www.fi.embnet.org/</a>
INFORMAGEN	France	<a href="http://www.infologgen.fr/">http://www.infologgen.fr/</a>
GENZUSnet	Germany	<a href="http://genzuse.citc-helldoberg.de/biounit/">http://genzuse.citc-helldoberg.de/biounit/</a>
DNBB	Greece	<a href="http://www.lmbb.forth.gr/">http://www.lmbb.forth.gr/</a>
HEN	Hungary	<a href="http://www.hu.embnet.org/">http://www.hu.embnet.org/</a>
INSEE	Ireland	<a href="http://www.gen.tcd.ie/">http://www.gen.tcd.ie/</a>
INN	Israel	<a href="http://dapsas.welzlmann.ac.il/bcd/finn.html">http://dapsas.welzlmann.ac.il/bcd/finn.html</a>
ISI-ADR	Italy	<a href="http://bio-www.bio.cnr.it:8000/BioWWW/Bio-WWW.htm">http://bio-www.bio.cnr.it:8000/BioWWW/Bio-WWW.htm</a>
CATC/CAMK	Netherlands	<a href="http://www.caco.kun.nl/">http://www.caco.kun.nl/</a>
Bio	Norway	<a href="http://www.no.embnet.org/">http://www.no.embnet.org/</a>
IBB	Poland	<a href="http://www.ibb.waw.pl/">http://www.ibb.waw.pl/</a>
ICC	Portugal	<a href="http://www.lgc.gilbenham.pt/">http://www.lgc.gilbenham.pt/</a>
GeneBee	Russia	<a href="http://www.genebee.msu.su/">http://www.genebee.msu.su/</a>
CHB-CSIC	Spain	<a href="http://www.es.embnet.org/">http://www.es.embnet.org/</a>
BMC	Sweden	<a href="http://www.embnet.se/">http://www.embnet.se/</a>
SIB	Switzerland	<a href="http://www.ch.embnet.org/">http://www.ch.embnet.org/</a>
SENET	UK	<a href="http://www.seenet.dl.ac.uk/">http://www.seenet.dl.ac.uk/</a>
<b>EMNET Specialist Nodes</b>		
MGPS	Germany	<a href="http://www.mips.biochem.mpg.de/">http://www.mips.biochem.mpg.de/</a>
ICSEB	Italy	<a href="http://www.icseb.biorche.it/">http://www.icseb.biorche.it/</a>
Pharmacia Upjohn	Sweden	<a href="http://www.gnu.com/">http://www.gnu.com/</a>
T.Hoffmann-La Roche	Switzerland	<a href="http://www.roche.com/">http://www.roche.com/</a>
EET	UK	<a href="http://www.ebi.ac.uk/">http://www.ebi.ac.uk/</a>
HGMP-RC	UK	<a href="http://www.hgmp.mrc.ac.uk/">http://www.hgmp.mrc.ac.uk/</a>
Sanger	UK	<a href="http://www.sanger.ac.uk/">http://www.sanger.ac.uk/</a>
UMBER	UK	<a href="http://www.biolif.msu.ac.uk/ibbrowser">http://www.biolif.msu.ac.uk/ibbrowser</a>
<b>EMNET Associate Nodes</b>		
IBBN	Argentina	<a href="http://sol.biol.unlp.edu.ar/embnet">http://sol.biol.unlp.edu.ar/embnet</a>
ANZIS	Australia	<a href="http://www.anzis.su.oz.au/">http://www.anzis.su.oz.au/</a>
CEI	China	<a href="http://www.cei.pku.edu.cn/">http://www.cei.pku.edu.cn/</a>
CIGB	Cuba	<a href="http://bio.cigb.edu.cu/">http://bio.cigb.edu.cu/</a>
CFRI	India	<a href="http://salarjung.embnet.org.in/">http://salarjung.embnet.org.in/</a>
SANBI	South Africa	<a href="http://www.sanbi.ac.za">http://www.sanbi.ac.za</a>
<b>USA Information Providers</b>		
NIH	USA	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>
NLM	USA	<a href="http://www.nlm.nih.gov/">http://www.nlm.nih.gov/</a>
NIH	USA	<a href="http://www.nih.gov/">http://www.nih.gov/</a>



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



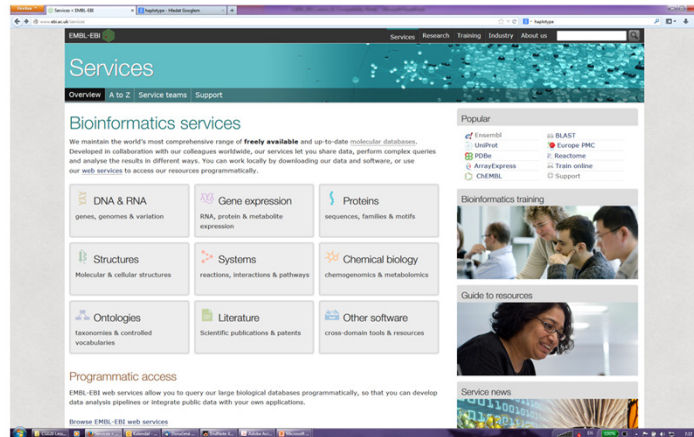
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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>



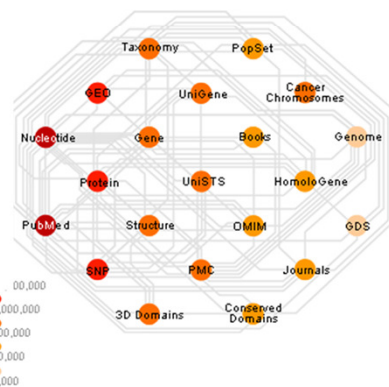
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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/>



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

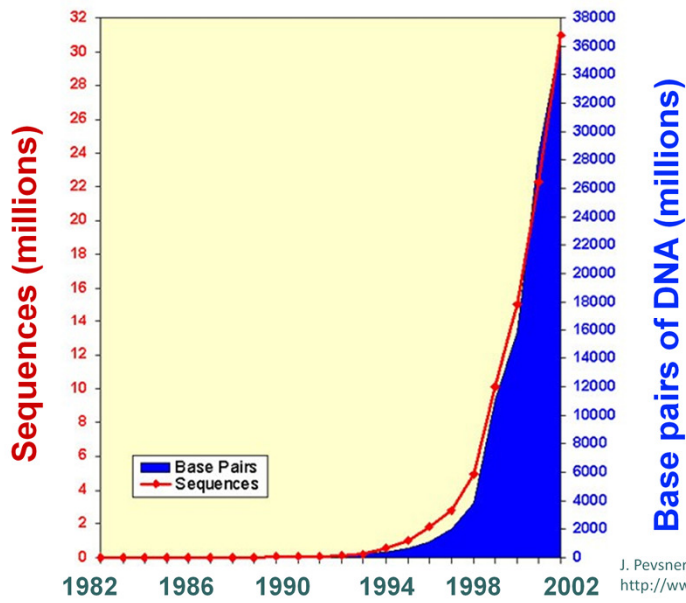
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - Sekvence v databázích tzv. „Velké trojky“:
    - EMBL
      - <http://www.ebi.ac.uk/embl/>
    - GenBank,
      - <http://www.ncbi.nih.gov/Genbank/GenbankSearch.html>
    - DDBJ,
      - <http://www.ddbj.nig.ac.jp>
  - denně vzájemná výměna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

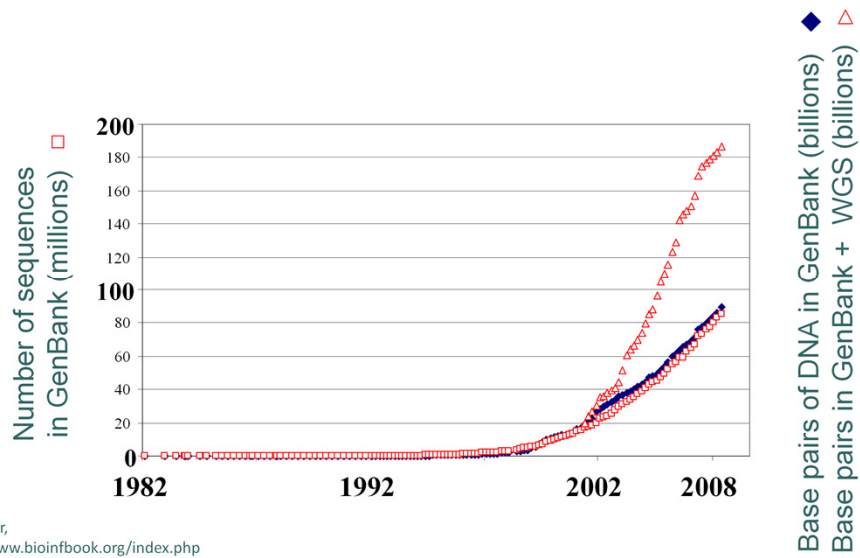


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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>

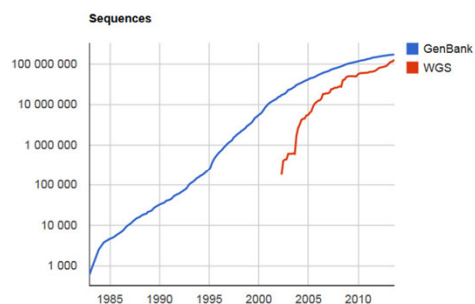
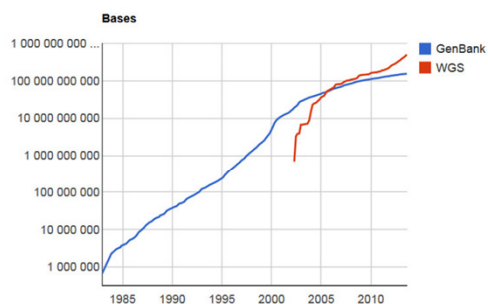


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Growth of GenBank

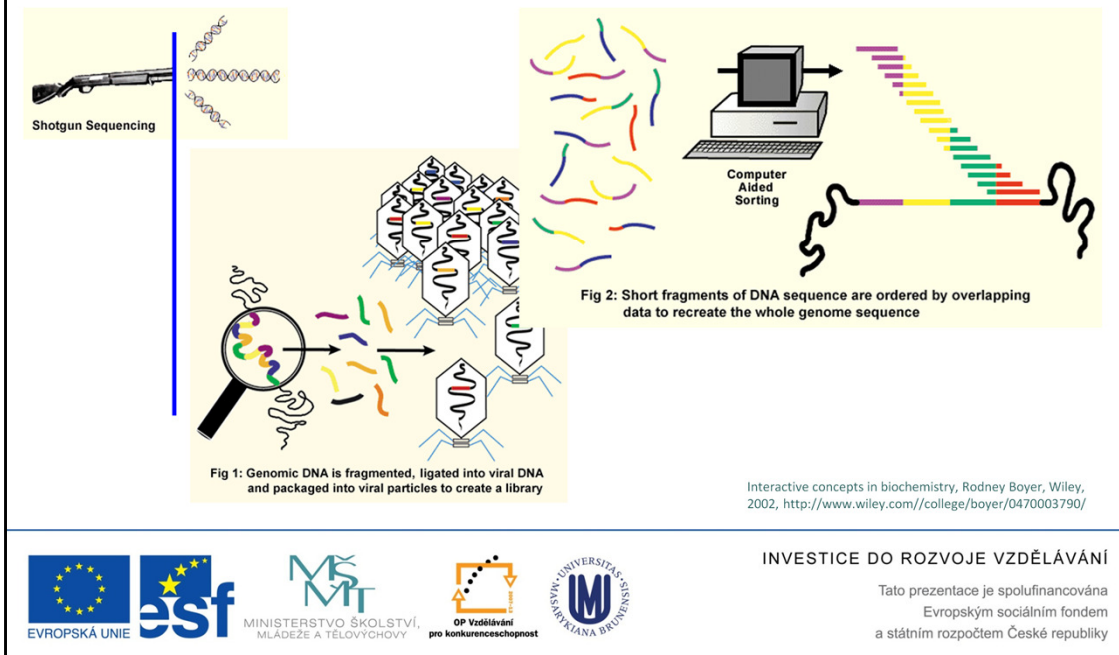
## Feb 15 2013



### INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# WGS

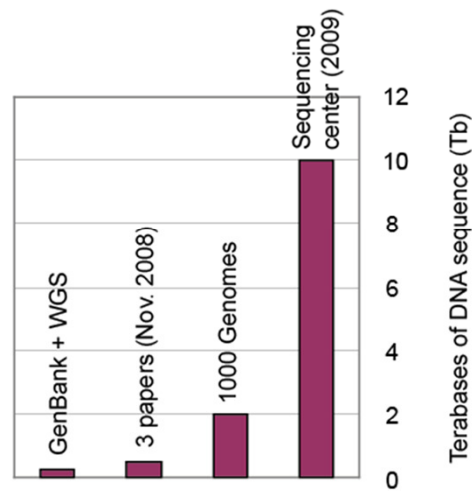


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic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



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



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 GenBank homepage. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' links, and a 'My NCBI Sign In' link. Below this is a search bar labeled 'All Databases' with a 'Search' button. The main content area is divided into several sections: 'Welcome to NCBI' with a brief description of the center's mission and links to 'About the NCBI', 'Mission', 'Organization', 'Research', and 'RSS Feeds'; 'Get Started' with a list of links for 'Tools', 'Downloads', 'How-To's', and 'Submissions'; 'Popular Resources' with links to 'PubMed', 'Bookshelf', 'PubMed Central', 'PubMed Health', 'BLAST', 'Nucleotide', 'Genome', 'SNP', 'Gene', 'Protein', and 'PubChem'; and 'NCBI YouTube channel' with a 'GO' button and a 'GO' button. There is also a 'NCBI Announcements' section with a link to 'New version of GenBank available' and a link to 'An integrated, downloadable tool for viewing and analyzing NCBI's July Newsletter'.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Primární databáze

Gene symbol: *isaA*  
Gene description: Non-component VanA-like sensor kinase  
Locus tag: *pf\_125*  
Gene type: protein coding  
RefSeq status: PROVISIONAL  
Organism: *Agrobacterium tumefaciens* subsp. *Agrobacterium tumefaciens* [strain: *Rhizobium radiobacter*]  
Lineage: Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiales; Rhizobium/Agrobacterium group; Agrobacterium; Agrobacterium tumefaciens complex

Location: plasmid: Ti  
Sequence: NC\_02377.1 (140594..148193)

Genomic context: NC\_02377.1

Genomic regions, transcripts, and products

Genomic Sequence: NC\_02377.1

Related articles

General information



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

NCBI Nucleotide

Search [Nucleotide]

Dir:

**Přístupový kód**

NC\_002377.1

**GeneBank Identifier**

NC\_002377.1

LOCUS NC\_002377 2490 bp DNA linear BCT 29-DEC-2003

DEFINITION *Agrobacterium tumefaciens* octachlorum plasmid T1, complete sequence.

ACCESSION NC\_002377.1

VERSION NC\_002377.1

KEYWORDS

SOURCE *Agrobacterium tumefaciens* (Rhizobium radiobacter);

Organism: Rhizobiales;  
Taxon: Agrobacterium.

Farrand, E.F., Schrammeyer, B., Hooykaas, P.J. and

TITLE *Agrobacterium tumefaciens* T1 plasmid sequence

JOURNAL *Agrobacterium*

REFERENCE 2 (bases 1 to 2490)

AUTHORS Zhu, J., Oper, P.M., Schrammeyer, B., Hooykaas, P.J., Farrand, E.F. and

Wainman, S.C.

TITLE Direct Submission

JOURNAL Submitted (07-DEC-2003) Microbiology, Cornell University, Wing

Hall, Ithaca, NY 14853, USA

COMMENT PROVISIONAL [BLAST](#). This record has not yet been subject to final

NCBI review. The reference sequence was derived from [BLAST](#).

FEATURES

source

1..2490

/organism="Agrobacterium tumefaciens"

/mol\_type="genomic DNA"

/db\_xref="taxon:358"

/plasmid="T1"

/note="extrachromosomal"

contigine-type

1..2490

/gene="vira"

/db\_xref="GeneID:1224314"

1..2490

/gene="vira"

/note="Two-component regulator of vir regulon; ViraA is a

transmembrane histidine kinase"

/codon\_start=1

/transl\_start=11

/product="vira"

/protein\_id="WP\_002377.1"

/db\_xref="GI:1095121"



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Primární databáze

```
/translation="KROUPEPTQDFETGAEWEILALIVAMIPAMAIAWQONAT
TQAILQJGRIINDEALIGREVLAMTUVAVPPIIIRGALMRELELGLPPOH
IFREBMAOLACQFPELACIADAVAFQWQWYFQGLQAFPTLALIPKATVOT
LREPTLANNQOFLEQWFAIFPELLELESQFELGZAPRILARQPIILGLL
POVELANQVOTLIRLHNGOQRELEVLQVREBQALIPKQKATVGLIITL
VLEBETKMEADLRELELKEIYVCPORATTECAALIIIGPPTAUTMALL
VREBMAQVOTVAFREPWQKETSREVMOTVREBQALIPKQKATVGLIITL
GLIILLAMETTELAVLCLQVTEFPEFCOHLLELALCLQVIVBSEQFBCD
VREBMAQVOTVAFREPWQKETSREVMOTVREBQALIPKQKATVGLIITL
SOTZAMLIIDQILTLASQWMIIEFPELLELESQFELGZAPRILARQPIIL
ETREBQALIPKQKATVGLIITLQVREBQALIPKQKATVGLIITLQVREBQ
LHICQWQIIPAVLREIIFRPPPTLADQVOTLGLACVHSHIAPAVIIVQVH
QTPPIIFPQKQKATVGLIITLQVREBQALIPKQKATVGLIITLQVREBQ
PQWPTREIENIKOMQALMOTQALPQKQKATVGLIITLQVREBQALIPK
LQREBQKATVGLIITLQVREBQALIPKQKATVGLIITLQVREBQALIPK
ORIGIN
1 atgaaagaa gataataaac gaagaggag gattitaaqa caggagcaga gctttggtt
41 aataagcct taaatcagg tgaagaaat tgcgcttca tggcgttgc gctctgag
121 gaaatgga ataacaggc aatctcagg caatagatc agatkaagc agaaagcgc
181 tcaatcagg gaaagaaat ccgctcacc aagaaagag tggcaataa ccgcaaat
241 atctcagg tggagctct gaggaaat cgaagaaat tgaagaaat attagaaa
301 tcaatcagg taagaaagc caatcgtc caatcagc gcaagaaat agcttcca
361 aatcaggc aagagaggc ccgctcct ggcagaaa atgagcct gaagattc
421 tggagctc taatcagg tgaagaaat ataaagaa agctcctc
481 tgaagaaat caagaaat gctagcag agctcctc tctctgca acaaagcgc
541 gataatcct tcaagcag cctgaaat gaggcctc aaagaaag cgtctcag
601 gaatcctc tgcgaaat tgaagaaat gctcctc tctcagc ttcgcaag
661 gaaagaaat tggaaatc gctcagc ttcgcaag caaatcgc ggaatcgc
721 aagagaaat gctcagc atagctc aaatcag agagagag agagctc
781 tctcagc agctcctc gctcagc cctcctc tcaatcct ctaagctc
841 gaaagaaat caatcagc aagagaaat tgaagaaat agctcctc caagaaat
901 gaaatcctc tgaagaaat ggcagaaat agctcagc agagaaat atctcagc
961 ataaagaa caatcagc ataaagaa gctcagc tgaagaaat tgaagaaat
1021 tggagctc aatcagc tgaagaaat caaaagc tggagaaat ccgctcctc
1081 ccgcaatc tctcagc caagaaat gaaagaaat ccgctcctc cctcagc
1141 tgaagaaat tgaagaaat gctcagc ataaagaa ataaagaa ataaagaa
1201 aatcagc ataaagaa ataaagaa ataaagaa ataaagaa ataaagaa
1261 cctcagc gaaagaaat gctcagc ataaagaa ataaagaa ataaagaa
1321 gctcagc gaaagaaat gctcagc ataaagaa ataaagaa ataaagaa
1381 gctcagc gaaagaaat gctcagc ataaagaa ataaagaa ataaagaa
1441 gctcagc gaaagaaat gctcagc ataaagaa ataaagaa ataaagaa
1501 gaaagaaat tgaagaaat ataaagaa gaaagaaat caatcagc ataaagaa
1561 ataaagaa gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
1621 gaaagaaat tgaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
1681 gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
1741 ataaagaa gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
1801 ataaagaa gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
1861 gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
1921 caatcagc aatcagc tgaagaaat gctcagc ataaagaa ataaagaa
2041 gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
2101 gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
2161 gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
2221 gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
2281 gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
2341 gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat gaaagaaat
```



## INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato realizace 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	<b>DNA</b>
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	<b>RNA</b>
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	<b>protein</b>
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

two-component VWA-like sensor kinase

NCBI Reference Sequences (RefSeq)

Genome Annotation

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

Reference assembly

Genomic

1. NC\_003065.3

Range 180831..183332  
Download [GenBank](#) [FASTA](#) [Sequence Viewer \(Graphics\)](#)

mRNA and Protein(s)

1. NP\_396486.1 two component sensor kinase [Agrobacterium tumefaciens str. C58]

UniProtKB/Swiss-Prot P18549  
Conserved Domains (2) [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/phosphoacceptor) domain: Histidine Kinase A dimers are formed through parallel association of 2 domains creating 4-helix bundles; usually these domains contain a conserved His residue and are activated via ...
PRK13837	PRK13837: two-component VWA-like sensor kinase; Provisional

Location:580 - 694  
Blast Score: 292

Location:452 - 550  
Blast Score: 144

Location:14 - 333  
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

# 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**

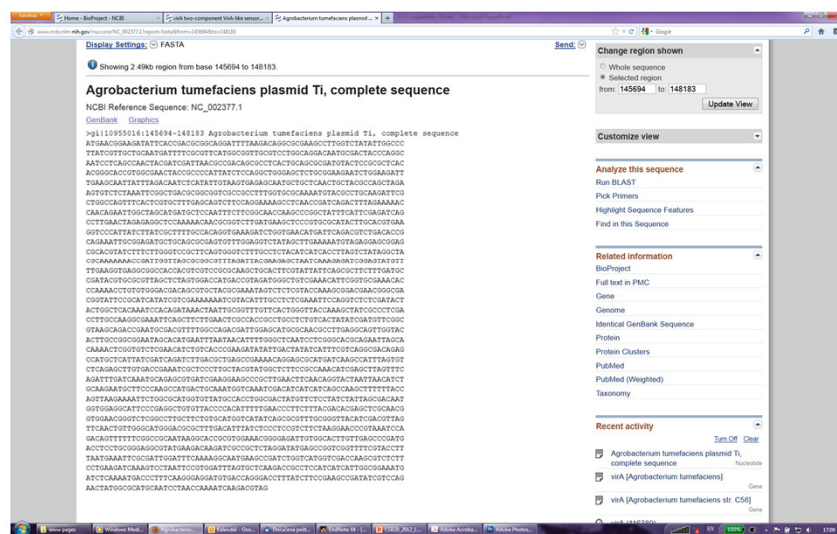


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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/>

Expasy Home page   Site Map   Search EAFAs   Contact us   Swiss-Prot   PROSITE   Extras/links

Home to SWISS-PROT   Search SWISS-PROT   Search PROSITE   Search TrEMBL   Search UniProt

Search PROSITE for:    Go   Clear



This program allows to scan a protein sequence (either from [Swiss-Prot](#), [TrEMBL](#), or provided by the user) for the occurrence of patterns and profiles stored in the [PROSITE](#) database, or to search protein databases with a user-entered pattern ([Reference](#) / [Download as scan, the standalone version](#)). The program [PROSITE](#) can be used to generate your own patterns. You may either:

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

<b>Scan a protein for PROSITE matches</b> Enter a Swiss-Prot/TrEMBL accession number (AC) (for example P08130) or a sequence identifier (ID) (for example NCIC_019061), or a PDB identifier, or paste your own protein sequence in the box below: <input type="text"/> and specify which motifs to use: Scan <input type="checkbox"/> patterns <input type="checkbox"/> profiles <input type="checkbox"/> rules [ <a href="#">User Manual</a> ] (You may also specify a PROSITE entry in the box to the right) <input type="checkbox"/> Exclude systems with a high probability of occurrence Your e-mail (optional): <input type="text"/> will send results by e-mail <input type="checkbox"/> plain text output <input type="button" value="START THE SCAN"/> <input type="button" value="RESET"/>	<b>Search Swiss-Prot with a PROSITE entry</b> Enter a PROSITE accession number (for example P081253), or type your pattern in <a href="#">PROSITE</a> format. Leave this box blank to scan a sequence with the entire PROSITE database: <input type="text"/> and specify your search limits: <ul style="list-style-type: none"><li>• The <input type="checkbox"/> Swiss-Prot <input type="checkbox"/> TrEMBL <input type="checkbox"/> TrEMBL/UniProt <input type="checkbox"/> PDB databases (You may also specify a protein in the box to the left)</li><li><input type="checkbox"/> including specific variants</li><li>• The following text: <input type="text"/> <a href="#">on NCBI TrEMBL</a> - reports multiple text with a connection, e.g. from sequences (Droghda, Not available for PDB)</li><li>• Sequences with at least <input type="text"/> hits</li><li>• At most 1000 <input type="text"/> matches</li></ul> Advanced options: <input type="checkbox"/> FASTA output <input type="checkbox"/> retrieve complete sequences allow at most <input type="text"/> X sequence characters to match a conserved position in the pattern <input type="checkbox"/> match inside <input type="checkbox"/> priority overlaps, no inclusion <input type="checkbox"/> the pattern, see <a href="#">help</a> <a href="#">Prosite database</a> (en) <input type="checkbox"/> see the pattern on <a href="#">help</a>
--	--



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/>

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

571 - 585 akessatTetelaaa

>PDOC00004 PS00004 CAMP\_PHOSPHO\_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

744 - 747 RDT

814 - 817 KRtS

>PDOC00005 PS00005 PKC\_PHOSPHO\_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

148 - 150 RKR

164 - 166 TGR

171 - 173 RKR

218 - 221 RKR

369 - 371 TGR

400 - 402 RGR

460 - 462 RGR

513 - 515 RGR

585 - 587 RGR

600 - 604 TGR

652 - 654 TGR

716 - 718 RGR

726 - 728 RGR

787 - 789 TGR

794 - 796 RGR

804 - 806 RGR

804 - 806 RGR

868 - 870 RGR

921 - 923 RGR

957 - 959 RGR

1000 - 1002 TGR

974 - 976 TGR

997 - 999 RGR

1002 - 1004 TGR

1018 - 1020 RGR

1031 - 1033 TGR

1119 - 1121 RGR



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



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/>

>PDOC50109 PS50109 HIS\_KIN Histidine kinase domain [profile].

```
402 - 471  SAERDQKSLAARWGLDIDCRGQYKPKQSDVDTLLNWNVCAKGLVALLAVLAKSEIIRGG  
RQGLVREKDFPLKELLRVTFIPFDFYANKKQVQVLLQSDYevEFKPYRGGGDLQKILIN  
MLTNDNTEPTD-----SLAYDARAKQYPRKQVYVLSKQYKRVKDFVQKDFQKQVAKVY  
teasawLrsaaTNEFFVFEVDTGRGIPRMEKCVPRNTVQVRELAQGRQDTLGLQTV  
KLVREKREKELTTEKAKDQVDTQVPELSTT
```

>PDOC50110 PS50110 RESPONSE\_REGULATORY Response regulatory domain [profile].

```
987 - 1045  RYFVYDGRNFCGRVATQKLNKQKVEVYRQCDGFEALALVTRGILqreegqevdklpPDT  
SPKQKQKPRKQTEKATKREKSRVWKAIVQVETFIIVQGGSD-----
```

Graphical summary of hits (*java applet*)



98 hits with 12 PROSITE entries

[Expasy Home page](#) [Site Map](#) [Search Expasy](#) [Contact us](#) [Swiss-Prot](#) [PROSITE](#) [Proteomics tools](#)



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

#### New:

- [SPRINT](#) - Search PRINTS's evolutionary PRINTS
- [enPRINTS](#) - Search PRINTS' automatic annotation
- [InterPro](#) - Search the integrated InterPro family database

#### Direct PRINTS access:

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

#### PRINTS search:

- Search PRINTS with **NEW FingerPRINTScan**
- [FPrint](#)
- [GRAPIScan](#)
- [MULScan](#)
- FingerPRINTScan binaries and source are available: [contact.bioinf@man.ac.uk](mailto:contact.bioinf@man.ac.uk)



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání  
pro konkurenceschopnost

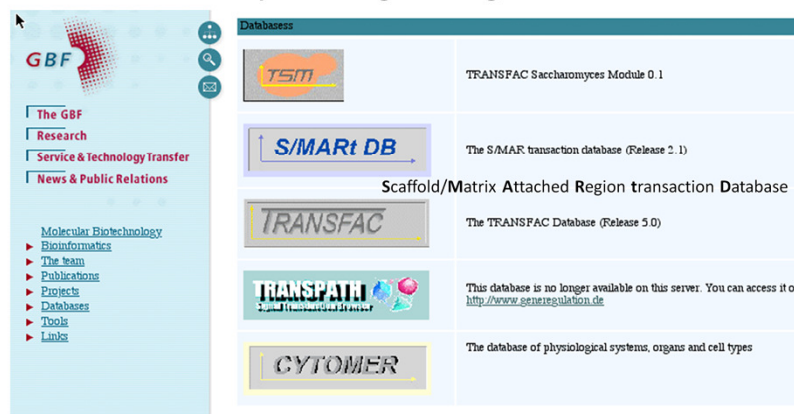


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Sekundární databáze

- 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 'The GBF', 'Research', 'Service & Technology Transfer', and 'News & Public Relations'. Below these are links for 'Molecular Biotechnology', 'Bioinformatics', 'The team', 'Publications', 'Projects', 'Databases', 'Tools', and 'Links'. The main content area is titled 'Databases' and lists several databases:

Database Name	Description
TSM	TRANSFAC Saccharomyces Module 0.1
S/MARt DB	The S/MAR transaction database (Release 2.1) 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 <a href="http://www.gene-regulation.de">http://www.gene-regulation.de</a>
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 interface. At the top left, there are links for 'DEPOSIT data', 'DOWNLOAD files', 'browse LINKS', 'BETA TEST new features', and 'BETA mmCIF files'. Below these are 'Current Holdings' and '19623 Structures' with a 'Last Update: 30-Dec-2002' and 'PDB Statistics' link. A 'Molecule of the Month: Cytochrome c' is featured with a 3D protein structure image. The main navigation bar includes 'ABOUT PDB | DATA UNIFORMITY | RECENT FEATURES | USER GUIDES | FILE FORMATS | EDUCATION | STRUCTURAL GENOMICS | PUBLICATIONS | SOFTWARE'. The central 'Search the Archive' section has a search box and options for 'query by PDB id only', 'match exact word', and 'remove sequence homologues'. To the right, 'PDB Mirrors' lists various international sites like 'San Diego Supercomputer Center', 'Rutgers University', etc. A 'News' section at the bottom left contains a holiday message dated '23-Dec-2002'.



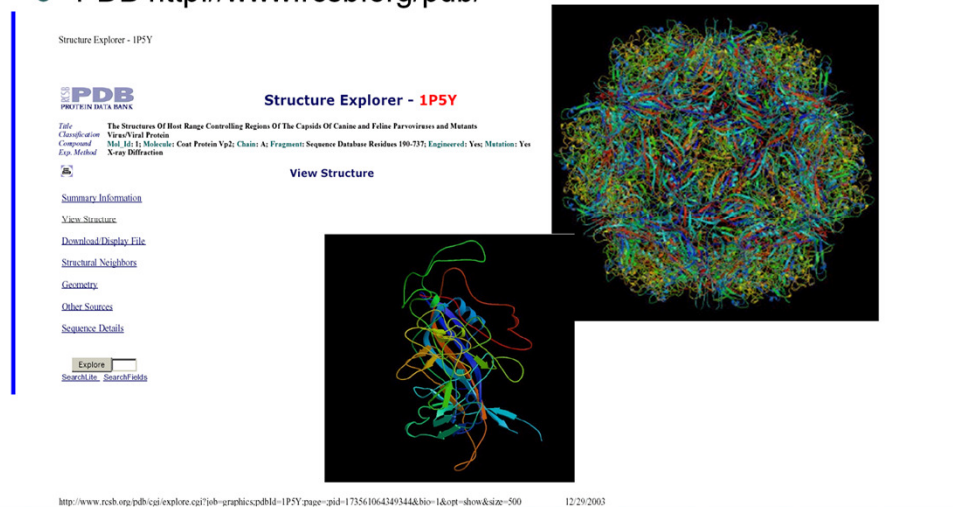
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 Host Range Controlling Regions Of The Capsids Of Canine And Feline Parvoviruses and Mutants  
**Classification:** Virus/Viral Protein  
**Compound:** M1\_H1\_Molecule: Coat Protein Vp2; Chain: A; Fragment: Sequence Database Residues 190-237; Engineered: Yes; Mutation: Yes  
**Exp. Method:** X-ray Diffraction

**View Structure**

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

[SearchFile](#) [SearchFields](#)

<http://www.rcsb.org/pdb/cgi/structure.cgi?job=graphics.pdbId=1PSY&page=pid=173561064349344&bio=1&opt=show&size=500> 12/29/2003



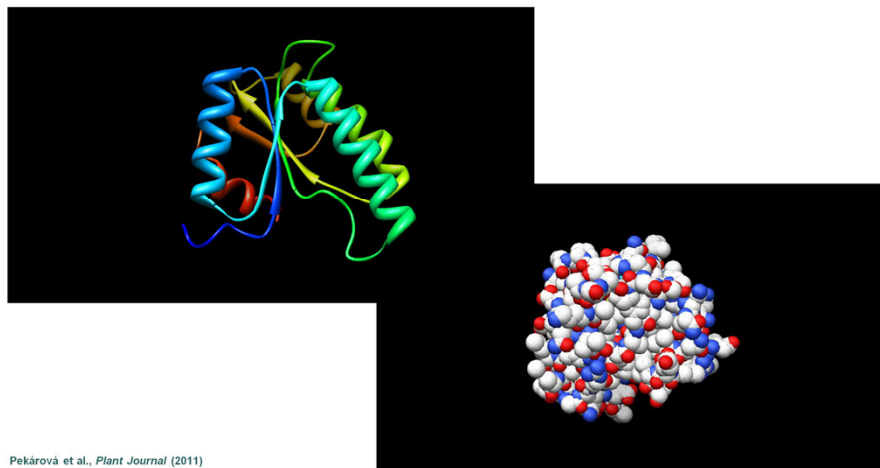
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 shows the UCSC Genome Browser homepage. At the top, there is a search bar with fields for 'clone', 'genome', 'assembly', and 'position'. Below this, there is a section titled 'Human Genome Browser - hg19 assembly (sequences)'. A table lists various queries and their descriptions:

Request	Genome Browser Response
chr7	Displays all of chromosome 7
chr11:q00212	Displays all of the unpaired coreg g000212
2p13	Displays region for band p13 on chr 2
chr2:1-1000000	Displays first million bases of chr 2, counting from p-arm telomere
chr3:100000-2000	Displays a region of chr3 that spans 2000 bases, starting with position 100000
RH1801, RH1815, 15q11.1-15q13, rs104252 and rs1102370	Displays region between genome landmarks, such as the STS markers RH1801 and RH1815, or chromosome bands 15q11.1 to 15q13, or SNPs rs104252 and rs1102370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
D15S3046	Displays region around STS marker D15S3046 from the Genethon/Mansfield maps. Includes 100,000 bases on each side as well.
A020474	Displays region of EST with GenBank accession A020474 in BRCA1 cancer gene on chr 17
AC020101	Displays region of clone with GenBank accession AC020101
AF030111	Displays region of mRNA with GenBank accession number AF030111
FSNP	Displays region of genome with HapMap Gene Nomenclature Committee identifier FSNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs
homocyst cascade	Lists mRNAs for dual homocyst genes
zinc finger	Lists many zinc finger mRNAs
knoppe zinc finger	Lists only knoppe-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

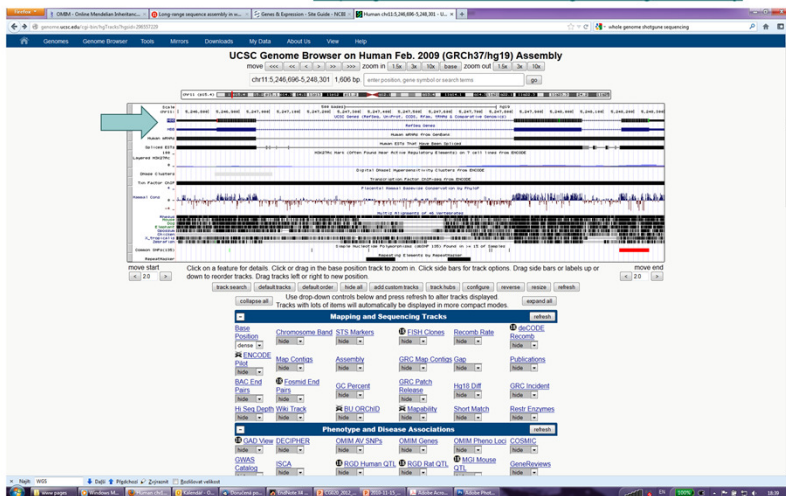


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# 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>

**Human Gene HBB (uc001msa.1) Description and Page Index**

**Description:** Homo sapiens hemoglobin, beta (HBB), mRNA.  
**RefSeq Summary (NM\_000518):** The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin. Hb A, the normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta zero-thalassemia. Reduced amounts of detectable beta globin causes beta plus-thalassemia. The order of the genes in the beta globin cluster is 5'-alpha 1 - gamma 2 - gamma 1 - beta 2 - beta 1 (provided by RefSeq, Jul 2009). Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. RefSeq Attributes: STAT(1) Transferrin\_receptor\_combination\_evidence\_1:V00471.1, B06050.0, E00000002, RefSeqAttributes:ENZYME

**Transcription Chromosome:** chr11 **Strand:** - **Size:** 1,600 **Start:** 5,245,695 **End:** 5,245,301 **Exon Count:** 3  
**Coding:** Size: 1,014 **Start:** 5,245,677 **End:** 5,245,253 **Exon Count:** 3

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

**Sequence and Links to Tools and Databases**

Genomic Sequence (chr11:5,245,695-5,245,301) mRNA (may differ from genomic) [Protein (147 aa)]  
Gene Source: Genome Browser, Protein (FASTA), VecSite, Table Schema (BioGPS)  
Gene Source: Ensembl, Ensembl Gene, Ensembl Transcript, GeneCards, GeneNetwork  
Clips Tissue: H-INV, HGNC, HPRD, Jackson Lab, MOPED  
OMIM: PubMed, Reactome, Standard SOURCE, TrEMBL, UniProtKB  
Wikipedia

**Comments and Description Text from UniProtKB**

**ID:** HBB\_HUMAN  
**DESCRIPTION:** UniProtKB: Full-Hemoglobin subunit beta, A1Name: Full-Beta-globin, A2Name: Full-Hemoglobin beta chain, Contains: UniProtKB: Full-LV-hemoglobin-7.  
**FUNCTION:** Involved in oxygen transport from the lung to the various peripheral tissues.  
**FUNCTION:** LVV-hemoglobin-7 potentiates the activity of bradykinin, causing a decrease in blood pressure.  
**SUBUNIT:** Heterodimer of one alpha chain and two beta chains in adult hemoglobin A (HbA).  
**INTERACTION:** P19695: HBA2, N6-Exp-19, HBA3: EBI-715554, EBI-714680.  
**ISSUE SPECIFICITY:** Red blood cells.  
**PTM:** Glucosylated non-enzymatically 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:** S-nitrosylated, a nitric oxide group is first bound to Fe2+ and then transferred to Cys-94 to allow capture of O2.  
**PTM:** Acetylated on Lys-60, Lys-61 and Lys-145 upon aspirin exposure. PubMed: 16919647 reports the identification of HBB acetylated on Lys-145 in the cytosolic fraction of HLs cells. This may have resulted from contamination of the sample.  
**MASS SPECITOMETRY:** Mass: 1310, Method: F4B, Range: 33, 42, Source: PubMed: 1579724.  
**DISEASE:** Defects in HBB may be a cause of Hereditary body anemias (HBBAN) (BMJ 181111) This is a form of non-spherocytic anemia of Diace type 1. After splenectomy, which has little benefit, basophilic inclusions called Heinz bodies are demonstrable in the erythrocytes. Before splenectomy, diffuse or punctate basophilia may be evident. Most of these cases are probably instances of hemoglobinopathy. The hemoglobin demonstrates heat stability. Heinz bodies are observed also with the hemlock syndrome (leptosis with cardiovascular anomalies) and with glutathione peroxidase deficiency.  
**DISEASE:** Defects in HBB are the cause of beta-thalassemia (B-THAL) (604133). 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. Classically, 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) (BMJ 610202), 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 that cutting off the blood supply to nearby tissues.



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>

**Get Genomic Sequence Near Gene**

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

**Sequence Retrieval Region Options:**

- Promoter Upstream by 1000 bases
- 5' UTR Exons
- CDS Exons
- 3' UTR Exons
- Introns
- Downstream by 1000 bases

• One FASTA record per gene.  
• One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (U) and 0 extra downstream (D).  
• Split UTR and CDS parts of an exon into separate FASTA records.  
Note: if features in close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

**Sequence Formatting Options:**

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

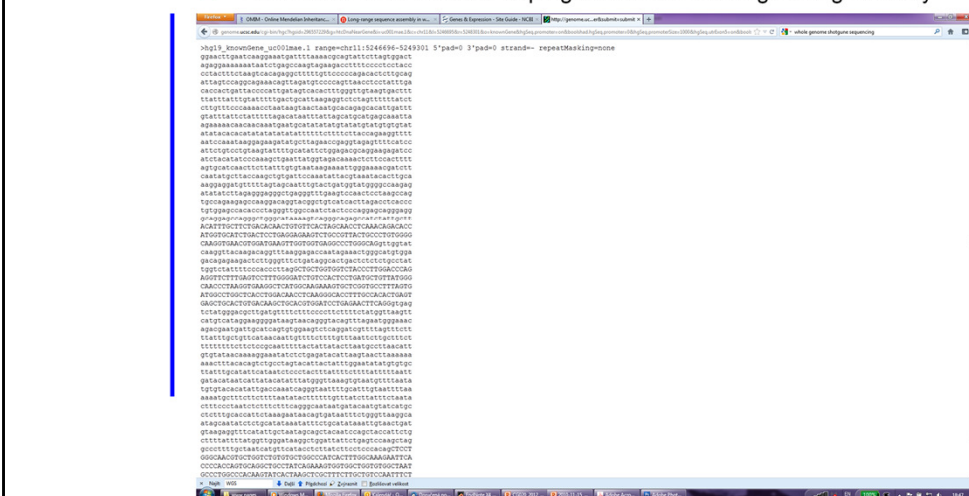


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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, 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, 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 homologíí



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 (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

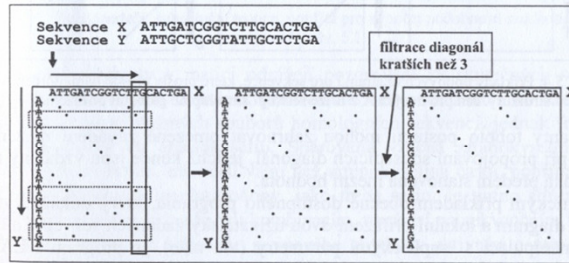


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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)

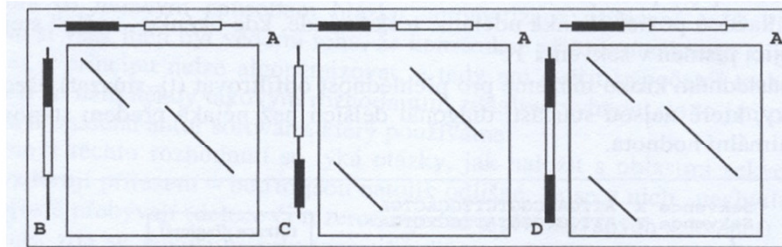


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 RID

[Search](#)

```
aacccacccgga  
acaccatcat cattatcacc atcggttttg gogcagttg tgggttcca  
gcytattaat  
aaaaataatt tattccacat gagatagat atgatatact atgtattttt  
tttttttttt  
ttatttgtaa acotttaata taacaagaac tacaaaaaat gaaaa
```

[Set subsequence](#) From:  To:

[Choose database](#)

Now: **BLAST!** or



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# BLAST

## Basic Local Alignment Search Tool

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

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

hodnota nepáru G-A  
 hodnota páru G-G

Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250

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



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

# BLAST

## Basic Local Alignment Search Tool

>gi|5016088|ref|NM\_001101.2| Length = 1793 E= expectancy value actin, beta (ACTB), mRNA ce  
 Score = 1110 bits (560), Expect = 0.0  
 Identities = 965/1100 (87%)  
 Strand = Plus / Plus

Query: 156 gtcgacaacggctctggcatgtgcaaggccggattgcccggagacgatgctccccggcc 215  
 Sbjct: 101 gtcgacaacggctccggcatgtgcaaggccggcttcgcccgggacgatgccccgggccc 160

Query: 216 gtcttcccatcgattgtgggacgtccccgtcaccagggtgtgatggtggcatggccag 275  
 Sbjct: 161 gtcttccctccatcgtggggcccaccaggccaggcgtgatggtggcatgggtoag 220

Query: 276 aaggactcgtacgtgggtgatgagggcagagcaagcgtggtatcctcaccctgaagtac 335  
 Sbjct: 221 aaggattcctatgtggggaogaggccagagcaagagaggcctcctcaccctgaagtac 280

Query: 336 occattgagcaggtatcgtgaccaactgggacgatggagaagatctggcaccacacc 395  
 Sbjct: 281 occatcgagcaggtatcgtcaccactgggacgatggagaaaatctggcaccacacc 340

- „expectancy value“ udává předpokládaný počet sekvencí se stejnou nebo lepší podobností při vyhledávání 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: [g|16119781|ref|NP\\_396486.1](#) two component sensor kinase [Agrobacterium tumefaciens str. C58]

Matching gis: [15163423-20141871-1019660](#)

Total (score > 100): 147088 hits in 146754 proteins in 6309 species

Selected: 147088 hits in 146754 proteins in 6309 species Filter: Min Score: 100 |

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

[Reset all filters](#)

Choose Display Options

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

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

% hits	SCOPE	ACCESSION	Length	Protein Description
833 aa	Conserved Domain Database hits			
4166	AAK9527	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]	
4166	P18540	833	RecName: Full=Wide host range virA protein; Short=WR virA	
4166	AAA79282	833	virA [Plasmid pTIC58]	
4159	NP_252390	833	hypothetical protein pTI-SANUKA_g142 [Agrobacterium tumefaciens]	
4159	AAA97765	833	tiorf140 [Agrobacterium tumefaciens]	
4153	AAA91590	833	virA [Plasmid T1]	
4153	g 1737127	833	virA protein	
4153	CAA34777	833	91.3 kDa protein [Agrobacterium tumefaciens]	
3800	CAA35790	829	virA [Agrobacterium rhizogenes]	
3718	g 1227240	829	virA gene	
3148	AAA8643	829	virA [Plasmid 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  
MSHIQIPPLGTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPEPGPDR  
VADAKGDSSEBEDELDLEVPVPSRFNRRVSVCAETYNPDEEBEDTDPRVIHPKTDEQRCRLQBACKDILLF  
KNLDQEQLSQVLDAMFERIVKADHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFCGLA  
LMYNTPRAATIVA TSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSRMKIVDVIGEK  
IYKDBERIITQGEKADSPYIIESEGVSLIRSRTKSNKDGNGQEVETARCHKQGYFGEALALVTKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNISHYEEQLVKMFGSSVDLGNLGG
```

```
[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....



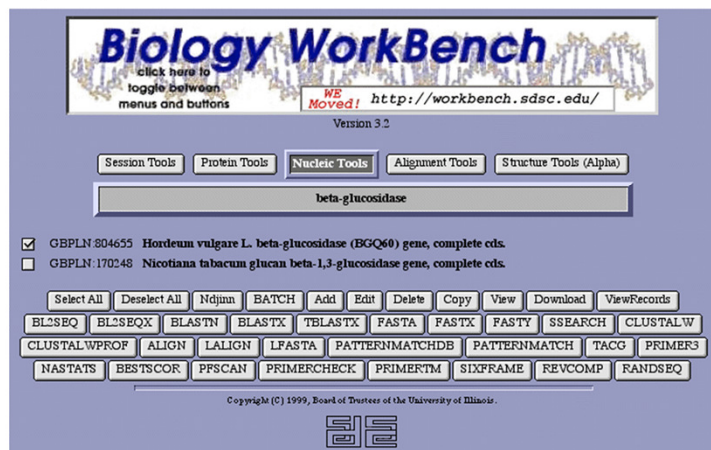
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



# Analytické nástroje

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



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



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 'View' interface for a nucleotide sequence. At the top, there is a 'View' button and the text 'View Nucleic Sequence(s)'. Below this, there are two dropdown menus: 'Format' set to 'Fasta' and 'Case' set to 'Upper', with a 'Change Format' button to the right. A link 'Download/View all sequences in text format' is present. Below the link are the words '[NEXT]' and '[BOTTOM]'. The main content is the sequence header: 'Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. GBPLN:170248, 4699 bp'. The sequence itself is a long string of nucleotide characters (A, T, C, G) starting with '> 170248'.



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
GAGCTCCCTTGGGGGCAAGGGCAAAACTTTTGGCTAAATGGAAAAATATTATACCAAGTGTITGTAATA
GTTACTCAATTTGAAITTAACRAAGGGGCAAAITTTGACTATTTTGGCCCTTATATCTTTGGTCACAAAAAC
ATAAATATCCCATCCGAAATTCGAAATGGTCCATTTATCGGCAAGTAGCTTTCTTTTATTTATAGTTAGTT
GCAAAACCTTCTCAGATTCATTTTATTAATTAATTTTCAAGGTCCTTCAAGGTCCTTGGTCTGCTCTCA
GTAGAGCCGCAATAAATAAGACCGATCAAAATAAAGCCGCCATTAATAATTAAGAAATTTAGGACTTC
GATGGCACGTAAGTGC AAAACTTTTCCAACTACTTGGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATTTTCTAAGTTTATCTCTAATTTACATCTCAACTAATATTAAGAAATTAACAGGTA
CAGCAATCATAAATTTTCTTAAAGAGGCAATTAATCCGGTACTGATTCATTTGGCTTTCTGAGG
TCTTCTATGSCADATTTACTAAGGGGCTCTTTTGGTACAGAAATAATAATAATTTTGGATAGAAATTT
GAGATGCAATTTCTTGGTTTAATTAAGTATTAGCTAATTTTCAGAAATAAATTTTACACTAAATAG
TAAATCAACTATCACATGTAGAAAGTGAATGGAATAGCTAATCCATAGCCACTCACATAGAAATCC
TTAATTTATCTACTATTTTACAAATGATCGGTTAGTCTTATGAGATCCAGTATCCCTAATAATGCA
GTAGAACTTGAATAATTTTCAATTAATCTATTTCTTAAATTTAAATTTTGAATTTGGGCACTTAG
ATCATAAAGATGACCGTTAATAATAAAGATAGATGAGTTTAAATAGGAAAAAAAACCGTT
CGAGACTCTTATGGAAGGGCTTGTCTTCAAAGTAGATTTCTATTCATTTGCTCTGGTGAATGCAAAA
TGACATTTACTCTTAGATACAGCGAGCCACTCTACAACTTTCTATTTGTTACTC AAATGAAGTTTITA
GAGAACTTTAAATCTCACTACTCTTAAAGGAAATCAAAADGAGGCAATTTATTTAATTAATTAATTTAG
TTATGTTAAAGATAGAAATTTTATTTAAATTTGAAATGAAATTTAATTTAATTTGATTTAATAA
ACATAGATATCGCTAAGTATTTACCACAACATGGAGATCTACAGAGATTTTATTTATTTTACGAT
GATTAAGCAGCTTATCTCTGTTTGGAGGATGAAAGAAAGTAACTAGCTATTAATTTTATTTAAGT
```



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  
ELPWGARAKLFAKWKNIIPSVCSYSI\*INKGANLTILEL

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 t t 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 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  
SSLGGQGNFLNGKILQVFIIVYFELTKGQI\*LFCP

S S L G G Q G N F L N G K I L Y Q V  
2 a g t c c c t t 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 a c c a a g t g 61  
F V I V T Q F E L T K G Q 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 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

# Analytické nástroje

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

```
= Linear Map of Sequence:

          StyI
          BsaJI
          CviJI
          AluI
          SacI
          EcoICRI
          Bsp1286I
          BsiHKA1
          BanII  BslI
          SspI

1  gagctcccttgggggcaaggcaaaacttttgcctaaatggaaaatattataccaagt 60
ctcgagggaacccccgtccogtttgaaaaagatttaacttttataaatggttca
  * * * * *
2  E L P W G A R A K L F A K W K N I I P S
3  S S L G G Q G Q N F L L N G K I L Y Q V
4  A P L G G K G K T F C * M E K Y Y T K C
5  L E R P P C F C F K K S F F F I N Y W T
6  S S G Q P A L A F S K A L H F F I I G L
L A G K P P L P L V K Q * I S F Y * V L

          Tsp509I
          MaeIII Tsp509I  MseI
          ApoI

61  gtttgaatgattactcaattgaaatacaaaagggcaaatgactattttgcoccta 120
caacattatcaatgagttaaacttaattgttccocgtttaaactgataaaoggggat
  * * * * *
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
```

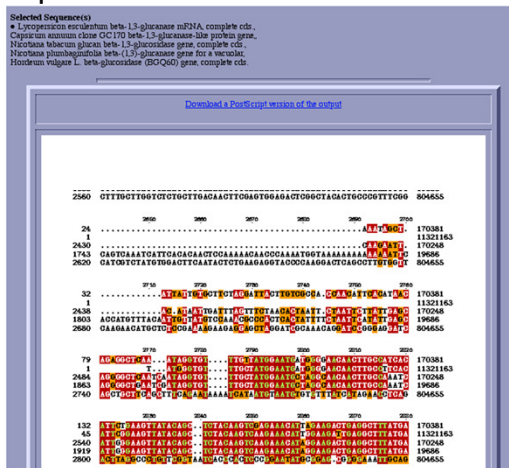


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Analytické nástroje

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

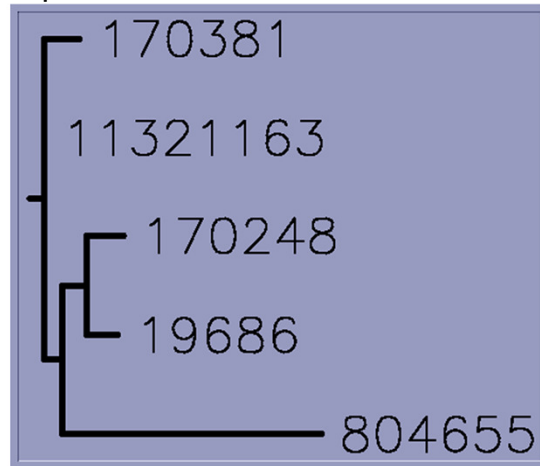


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 capability 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! 



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>



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 homologí
  - 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 shows the NCBI Gene database entry for PHACTR4 phosphatase and actin regulator 4 (Homo sapiens). The page includes a search bar, a table of contents, and detailed gene information. The 'Summary' section provides the official symbol (PHACTR4), full name (phosphatase and actin regulator 4), primary source (HGNC:2573), and location tag (PP11-442N24\_A1). It also lists related Ensembl IDs (ENSG00000204138, tEPHD:07816, MM:608748) and gene type (protein coding). The 'Genomic context' section shows the gene's location on Chromosome 1, NC\_000001.10 (2869693, 2892681) and a diagram of the gene structure with exons and introns. The 'Genomic regions, transcripts, and products' section provides a link to reference sequence details and the genomic sequence (NC\_000001 chromosome 1 reference GRCh37 p5 Primary Assembly).



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, restričních 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