

# CG020 Genomika

## 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 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**
  - Struktura genomů



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Schéma předmětu

- **Kapitola 09**
  - Evoluce genomů
  
- **Kapitola 10**
  - Genomika a systémová biologie
  
- **Kapitola 11**
  - Praktické aspekty funkční genomiky
  - Modelové organismy
  - PCR
  - Zásady navrhování primerů



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Literatura

- Literární zdroje pro kapitulu 01:
  - **Bioinformatics and Functional Genomics**, 3rd Edition, Jonathan Pevsner, Wiley-Blackwell, 2015  
<http://www.bioinfbook.org/php/?q=book3>
  - **Úvod do praktické bioinformatiky**, Fatima Cvrčková, 2006, Academia, Praha
  - **Plant Functional Genomics**, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání  
pro konkurenceschopnost



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?

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

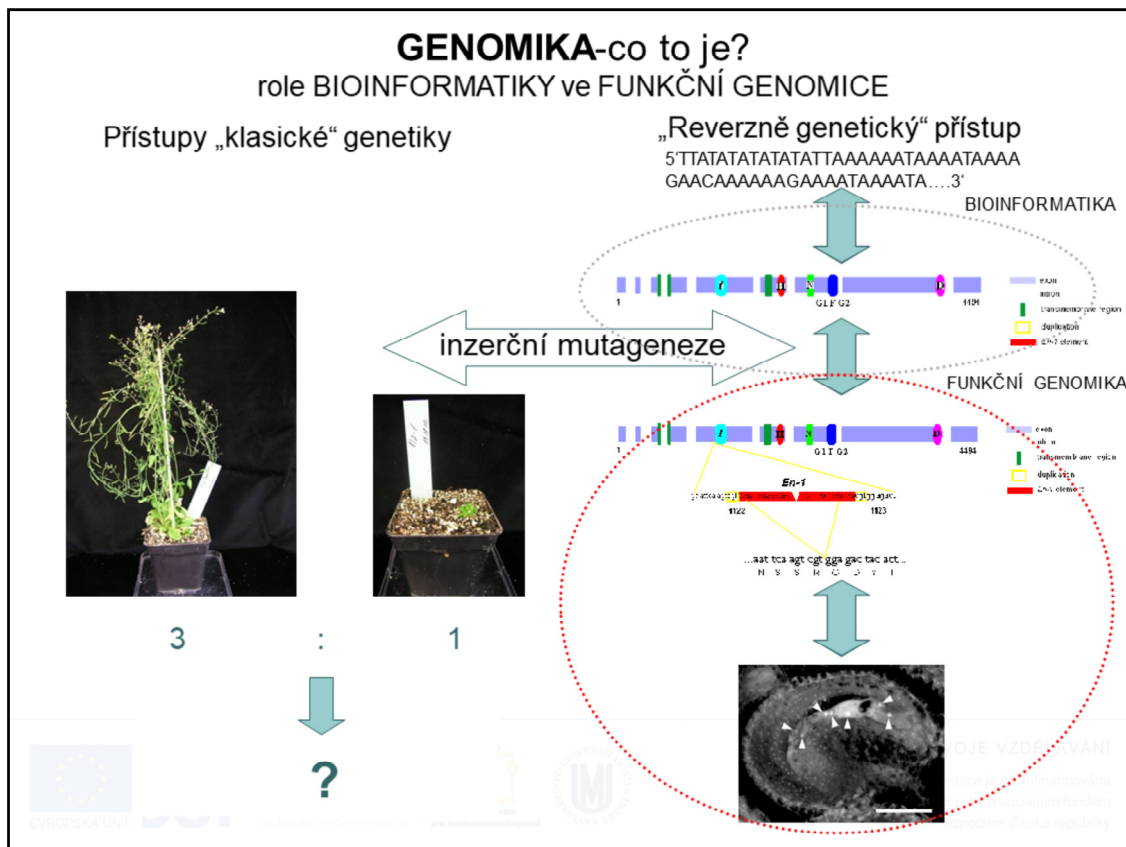
Tato prezentace je spolufinancována  
Evropským sociálním fondem  
a státním rozpočtem České republiky

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

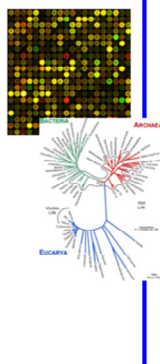
- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY



## 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.

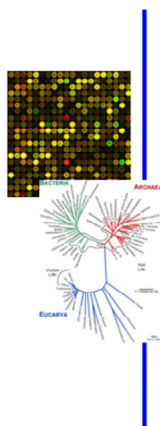
J. Pevsner,  
<http://www.bioinfbook.org/index.php>



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

<b>EMBL National Nodes</b>		
Vicente Blanco	Austria	<a href="http://www.at.emblnet.org/">http://www.at.emblnet.org/</a>
BEU	Belgium	<a href="http://www.be.emblnet.org/">http://www.be.emblnet.org/</a>
BioBase	Denmark	<a href="http://biobase.dk/">http://biobase.dk/</a>
CSC	Finland	<a href="http://www.fi.emblnet.org/">http://www.fi.emblnet.org/</a>
INFORMAGEN	France	<a href="http://www.infololgen.fr/">http://www.infololgen.fr/</a>
CRISOLINET	Germany	<a href="http://genome.sfb-hirchberg.de/bioinfo/">http://genome.sfb-hirchberg.de/bioinfo/</a>
IMBB	Greece	<a href="http://www.imbb.forth.gr/">http://www.imbb.forth.gr/</a>
HUN	Hungary	<a href="http://www.hu.emblnet.org/">http://www.hu.emblnet.org/</a>
INCEI	Ireland	<a href="http://www.incei.tcd.ie/">http://www.incei.tcd.ie/</a>
JIN	Israel	<a href="http://dapsil.wellman.ac.il/bcd/fin.html">http://dapsil.wellman.ac.il/bcd/fin.html</a>
JIN-ADN	Italy	<a href="http://fin-www.ba.cnr.it/8000/BioWWW/Bio-WWW.htm">http://fin-www.ba.cnr.it/8000/BioWWW/Bio-WWW.htm</a>
CAS/CQAMN	Netherlands	<a href="http://www.cas.kun.nl/">http://www.cas.kun.nl/</a>
IBO	Norway	<a href="http://www.no.emblnet.org/">http://www.no.emblnet.org/</a>
IBB	Poland	<a href="http://www.ibb.wzpa.pl/">http://www.ibb.wzpa.pl/</a>
ISC	Portugal	<a href="http://www.ig.gulbenkian.pt/">http://www.ig.gulbenkian.pt/</a>
GeneBee	Russia	<a href="http://www.genebee.msu.ru/">http://www.genebee.msu.ru/</a>
CNB-CSC	Spain	<a href="http://www.es.emblnet.org/">http://www.es.emblnet.org/</a>
BNC	Sweden	<a href="http://www.se.emblnet.org/">http://www.se.emblnet.org/</a>
SIB	Switzerland	<a href="http://www.ch.emblnet.org/">http://www.ch.emblnet.org/</a>
SIGNET	UK	<a href="http://www.signet.dl.ac.uk/">http://www.signet.dl.ac.uk/</a>
<b>EMBL Specialist Nodes</b>		
MPS	Germany	<a href="http://www.mips.biochem.mpg.de/">http://www.mips.biochem.mpg.de/</a>
ICGB	Italy	<a href="http://www.icgb.internic.it/">http://www.icgb.internic.it/</a>
Pharmacia Uppsala	Sweden	<a href="http://www.gnu.com/">http://www.gnu.com/</a>
FaH/FaHwv-La Roche	Switzerland	<a href="http://www.roche.com/">http://www.roche.com/</a>
EBI	UK	<a href="http://www.ebi.ac.uk/">http://www.ebi.ac.uk/</a>
HGMP-BC	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>
EMBL	UK	<a href="http://www.embl.ac.uk/">http://www.embl.ac.uk/</a>
<b>EMBL Associate Nodes</b>		
IBBH	Argentina	<a href="http://iui.biot.unip.edu.ar/emblnet">http://iui.biot.unip.edu.ar/emblnet</a>
ANIGS	Australia	<a href="http://www.anigs.usc.edu.au/">http://www.anigs.usc.edu.au/</a>
CEI	China	<a href="http://www.cei.cbi.cas.edu.cn/">http://www.cei.cbi.cas.edu.cn/</a>
CISB	Cuba	<a href="http://ibc.cigb.edu.cu/">http://ibc.cigb.edu.cu/</a>
CFDQ	India	<a href="http://falarjung.emblnet.org.in/">http://falarjung.emblnet.org.in/</a>
SANBE	South Africa	<a href="http://www.sanbi.ac.za">http://www.sanbi.ac.za</a>
<b>USA Information Providers</b>		
NCBI	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



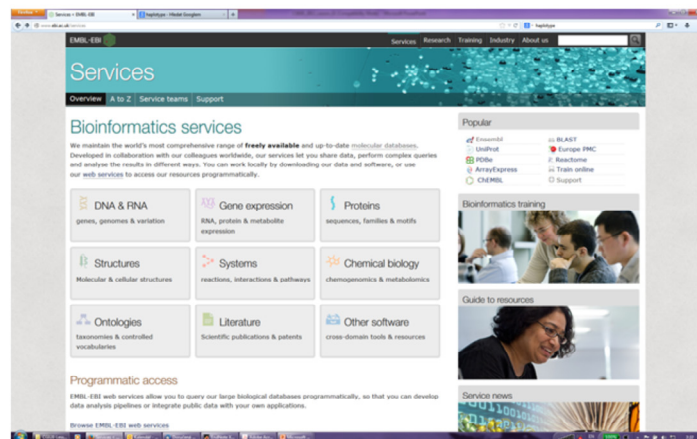
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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/>

NCBI Home  
Resource List (A-Z)  
All Resources  
Chemicals & Bioassays  
Data & Software  
DNA & RNA  
Domains & Structures  
Genes & Expression  
Genetics & Medicine  
Genomes & Maps  
Homology  
Literature  
Proteins  
Sequence Analysis  
Taxonomy  
Training & Tutorials  
Variation

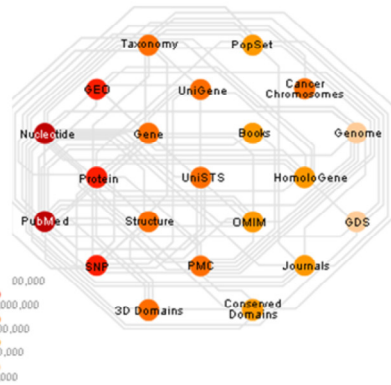
**Welcome to NCBI**  
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.  
About the NCBI | Mission | Organization | Research | RSS Feeds

**Get Started**

- **Tools:** Analyze data using NCBI software
- **Downloads:** Get NCBI data or software
- **HowTo's:** Learn how to accomplish specific tasks at NCBI
- **Submissions:** Submit data to GenBank or other NCBI databases

**NCBI YouTube channel!**  
Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel. [GO](#)

**NCBI Announcements**  
New version of GenBank available  
SHIP  
Gene  
Protein  
PubChem



i  
a  
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

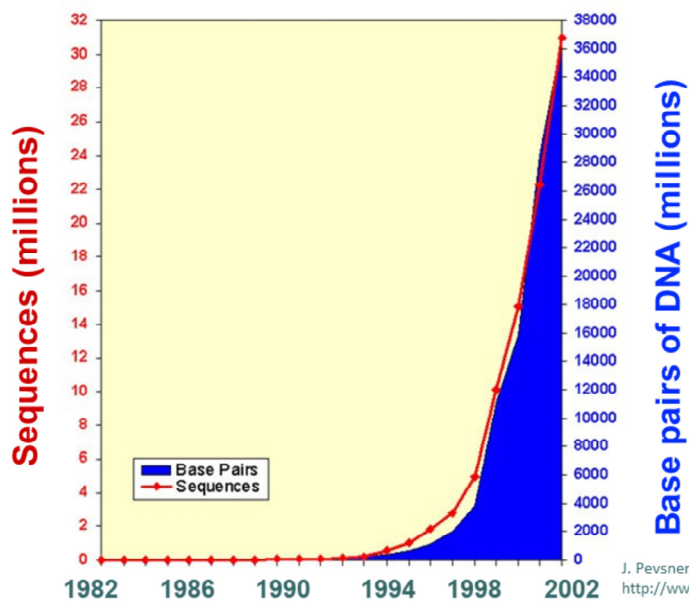
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - Sekvence v databázích tzv. „Velké trojky“:
    - EMBL
      - <http://www.ebi.ac.uk/embl/>
    - GenBank,
      - <https://www.ncbi.nlm.nih.gov/>
    - DDBJ,
      - <http://www.ddbj.nig.ac.jp>
  - denně vzájemná výměna a zálohování dat
  - velká datová náročnost (kapacita i software)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

## Growth of GenBank



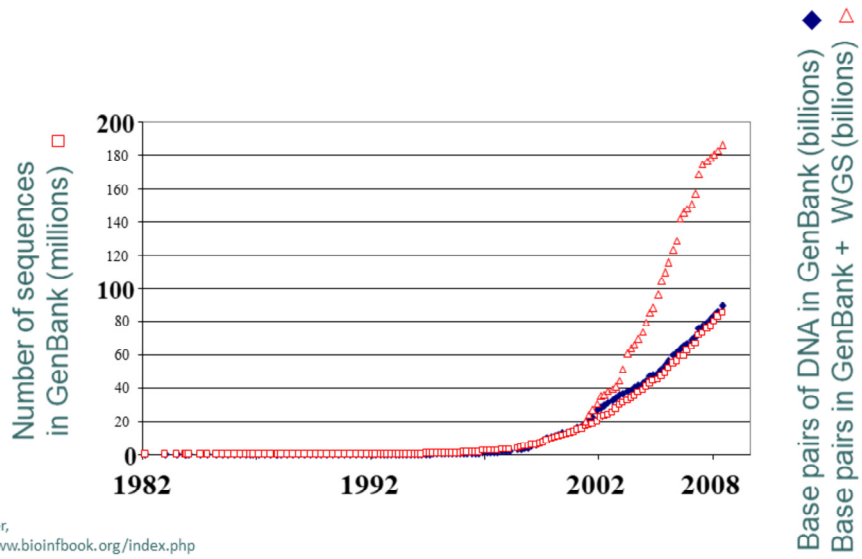
J. Pevsner,  
<http://www.bioinfbook.org/index.php>



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>

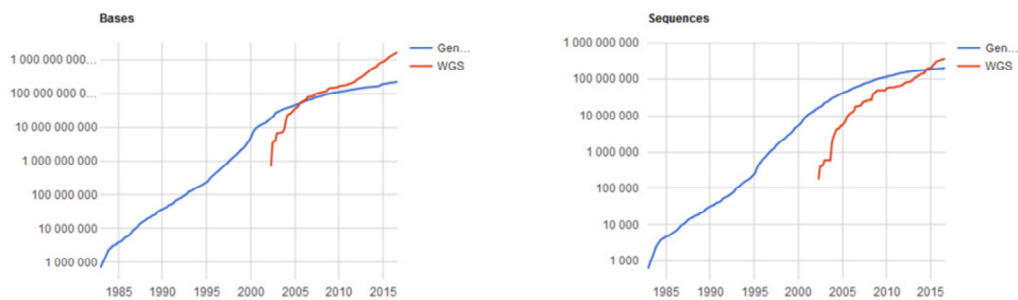


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Growth of GenBank

## Aug 2016



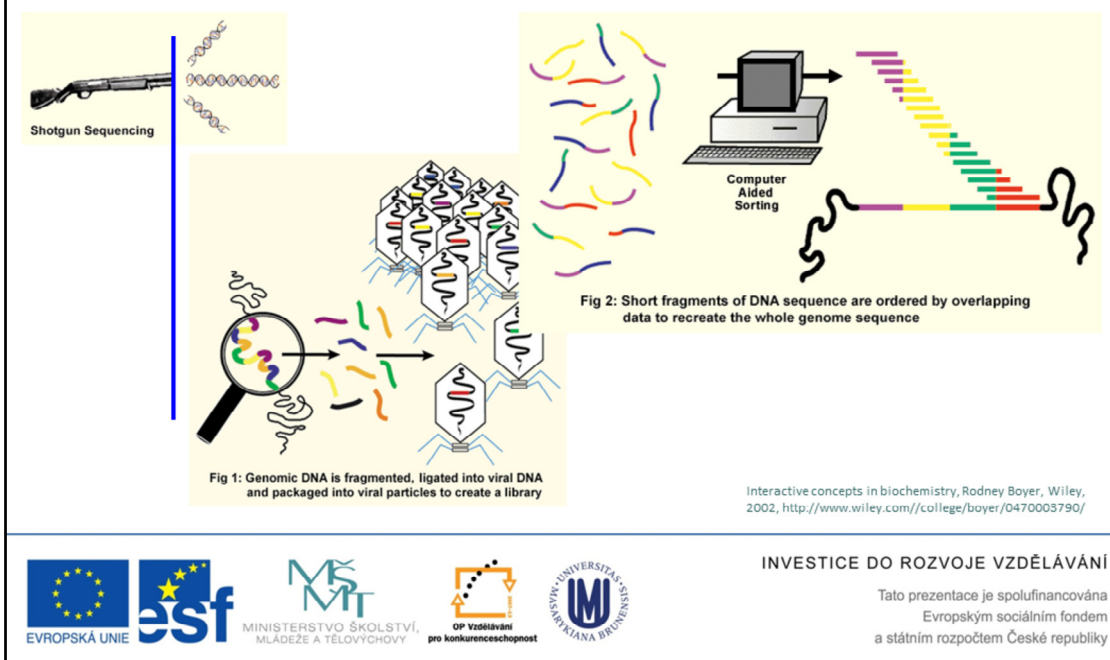
- Prosinec 1982 680 338 bp, 606 sekvencí
- Duben 2002  $19 \times 10^9$  bp,  $17 \times 10^6$  sekvencí + WGS  $692 \times 10^6$  bp, 172 768 sekvencí
- Srpen 2016  $218 \times 10^9$  bp,  $196 \times 10^6$  sekvencí + WGS  $1,6 \times 10^{12}$  bp,  $360 \times 10^6$  sekvencí



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# WGS

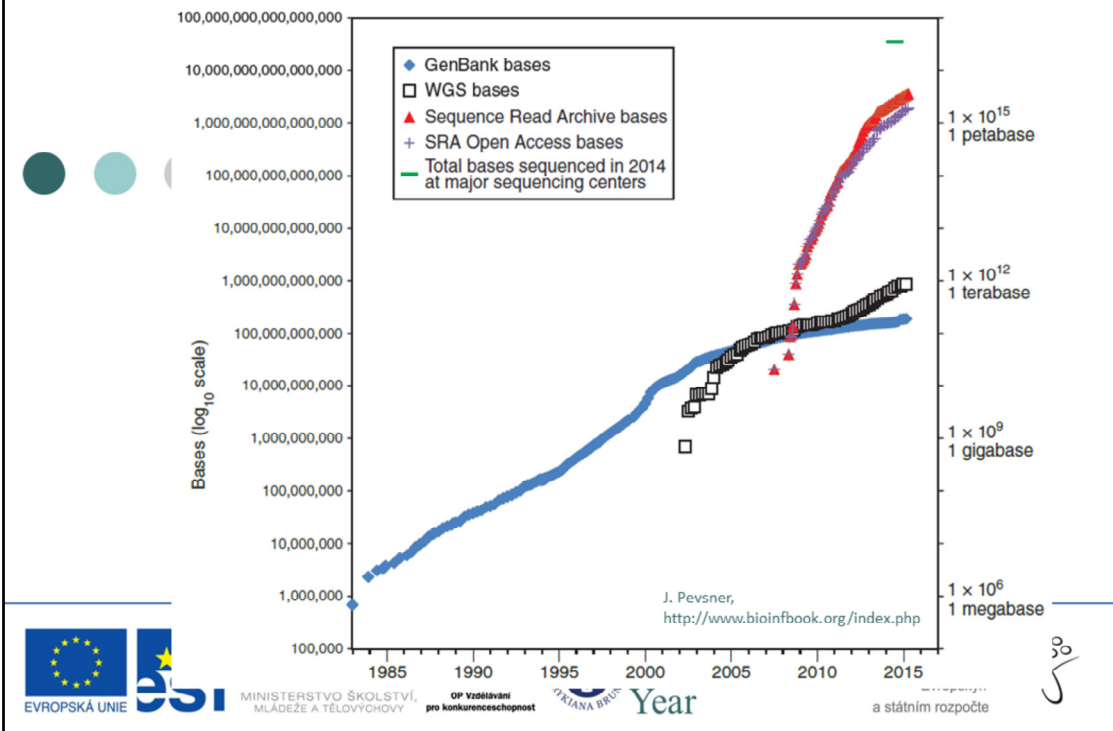


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic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

The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence. Placing the genomic DNA within the viral genome allows bacteria infected with the virus to faithfully replicate the genomic DNA pieces. Additionally, since a little bit of known sequence is needed to start the sequencing reaction, the reaction can be primed off the known flanking viral DNA.

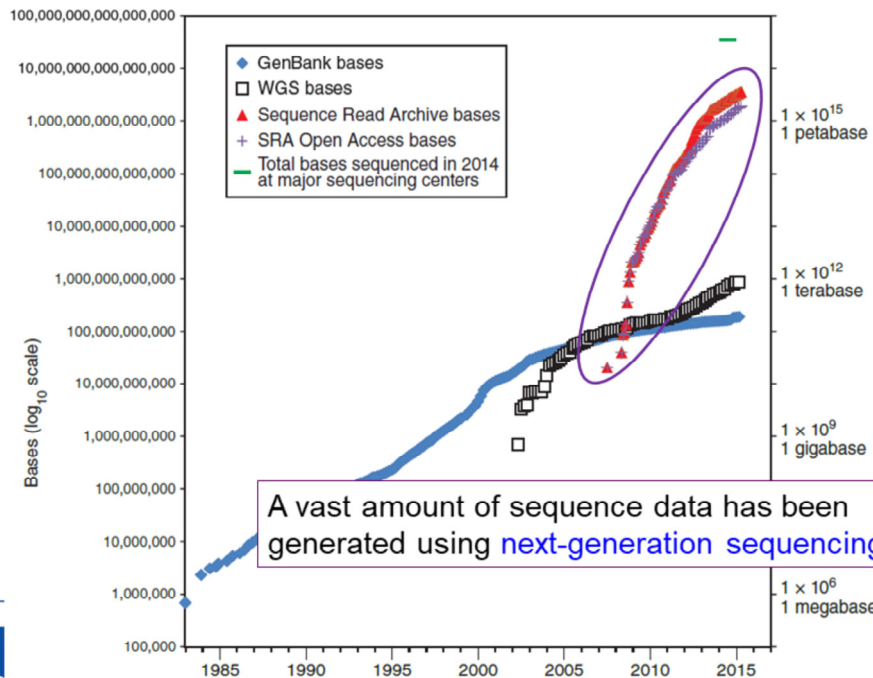
In order to read all the nucleotides of one organism, millions of individual clones are sequenced. The data is sorted by computer, which compares the sequences of all the small DNA pieces at once (in a "shotgun" approach) and places them in order by virtue of their overlapping sequences to generate the full-length sequence of the genome (Fig. 2). To statistically ensure that the whole genome sequence is acquired by this method, an amount of DNA equal to five to ten times the length of the genome must be sequenced. (Interactive concepts in biochemistry, Rodney Boyer, Wiley, 2002, <http://www.wiley.com/college/boyer/0470003790/>)

# Growth of DNA Sequence in Repositories





# Growth of DNA Sequence in Repositories



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY

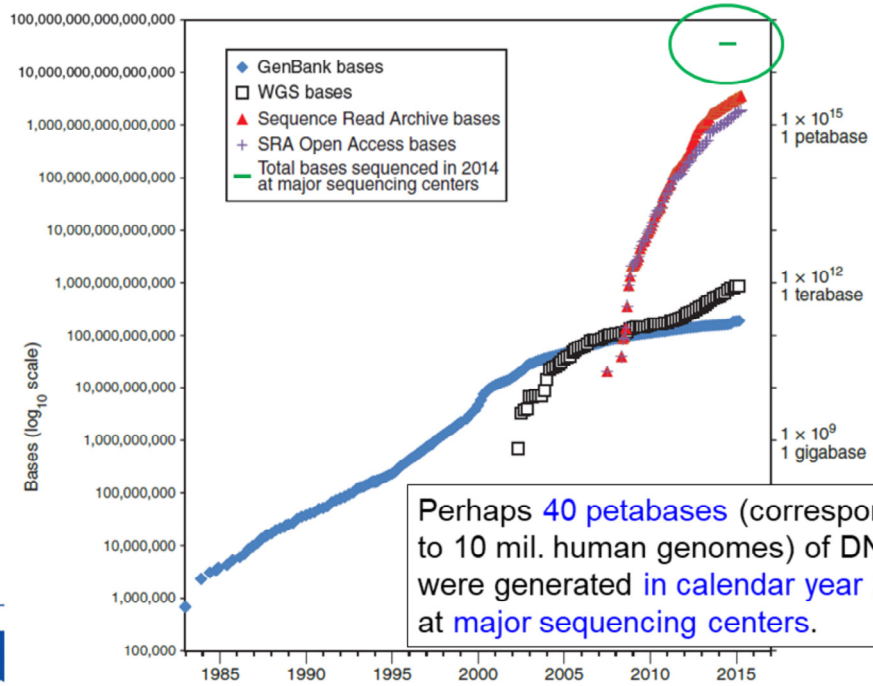
OP Vzdělávání  
pro konkurenceschopnost

EVROPSKÝ ROZVOJ  
Year

a státním rozpočte



# Growth of DNA Sequence in Repositories



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání  
pro konkurenceschopnost

EVROPSKÝ  
ROK  
2014

a státním rozpočte

# 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



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

# Primární databáze

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



The screenshot shows the NCBI homepage with a search bar at the top. The main content area is divided into three columns. The left column contains a navigation menu with categories like 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The middle column features a 'Welcome to NCBI' message, a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions', and a 'NCBI YouTube channel' section with a 'GO' button. The right column lists 'Popular Resources' such as 'PubMed', 'Bookshelf', 'PubMed Central', 'PubMed Health', 'BLAST', 'Nucleotide', 'Genome', 'SNP', 'Gene', 'Protein', and 'PubChem', along with an 'NCBI Announcer' section.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Primární databáze

The screenshot displays the NCBI Gene database entry for the *uak* gene. Key information includes:

- Gene symbol:** *uak*
- Gene description:** non-component VWA-like sensor kinase
- Location:** *uak\_129*
- Gene type:** protein coding
- RefSeq status:** PROVISIONAL
- Organism:** *Agrobacterium tumefaciens* subsp. *Agrobacterium tumefaciens*, subsp. *Rhizobium solanacei*
- Lineage:** Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiales; Rhizobium/Agrobacterium group; Agrobacterium; Agrobacterium tumefaciens complex

The 'Genomic context' section shows the gene's location on chromosome NC\_023277.1. The 'Genomic regions, transcripts, and products' section provides a detailed view of the gene structure and associated features. The 'Related articles' section, highlighted with a yellow circle, lists four scientific publications related to the gene.



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] [GO] [NUC] [PROTEIN] [GENE]

**Přístupový kód**

GeneBank Identifier

```
LOCUS BC_023777            2490 bp    DNA    linear    BCT 29-DEC-2003
DEFINITION  Open-reading frame for a transcription factor in plasmid T1, complete sequence.
ACCESSION   BC_023777
VERSION    BC_023777.1
KEYWORDS   map
SOURCE     Arabidopsis thaliana (Arabidopsis thaliana)
ORGANISM   Arabidopsis thaliana (Arabidopsis thaliana)
INSTRUMENT
AUTHOR     Farrand, S.K., Schramm, R., Kopyov, P.J., Farrand, S.K. and
            Schramm, R.
TITLE       Open-reading frame T1 plasmid sequence
JOURNAL    Submitted (07-MAR-2003) Microbiology, Cornell University, Wing
            Hall, Ithaca, NY 14853, USA
REFERENCE  Schramm, R., Kopyov, P.J., Farrand, S.K. and
            Farrand, S.K.
           2 (Issue 1 to 2490)
FEATURES   Location/Qualifiers
            source            1..2490
                    /organism="Arabidopsis thaliana"
                    /mol_type="genomic DNA"
                    /db_xref="taxon:3108"
                    /plasmid="T1"
                    /contig="transcriptome1"
            gene             1..2490
                    /gene="vial"
                    /db_xref="GeneID:1224316"
            CDS              1..2490
                    /gene="vial"
                    /note="Two-component regulator of vir regulon; Vial is a
                    transmembrane histidine kinase"
                    /coding_start=1
                    /trans_start=1
                    /product="vial"
                    /protein_id="NP_592951.1"
                    /db_xref="GI:11955143"
```



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	<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

Page 27

# RefSeq

two-component VIA-like sensor kinase

NCBI Reference Sequences (RefSeq)

Genome Annotation

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

Reference assembly

Genomic

1. **NC\_003065.3**

Range: 18031..18332  
Download: [GenBank](#), [FASTA](#), [Sequence Viewer](#), [Graphics](#)

mRNA and Protein(s)

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

UniProtKB/Swiss-Prot: [E18640](#)

Conserved Domains (3) [summary](#)

<a href="#">cd00075</a>	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.
<a href="#">cd00082</a>	HskA: Histidine Kinase A (dimerization/phosphoreceptor) domain: Histidine Kinase A dimers are formed through parallel association of 2 domains creating L-helix bundles; usually these domains contain a conserved His residue and are activated via ...
<a href="#">PRK13637</a>	PRK13637: two-component VIA-like sensor kinase. Provisional

Location:14 - 833  
Blast Score: 2944

Related Sequences



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

<u>Accession</u>	<u>Molecule</u>	<u>Method</u>	<u>Note</u>
AC_123456	Genomic	Mixed	Alternate complete genomic
AP_123456	Protein	Mixed	Protein products; alternate
NC_123456	Genomic	Mixed	Complete genomic molecules
NG_123456	Genomic	Mixed	Incomplete genomic regions
NM_123456	mRNA	Mixed	Transcript products; mRNA
NM_123456789	mRNA	Mixed	Transcript products; 9-digit
NP_123456	Protein	Mixed	Protein products;
NP_123456789	Protein	Curation	Protein products; 9-digit
NR_123456	RNA	Mixed	Non-coding transcripts
NT_123456	Genomic	Automated	Genomic assemblies
NW_123456	Genomic	Automated	Genomic assemblies
NZ_ABCD12345678	Genomic	Automated	Whole genome shotgun data
XM_123456	mRNA	Automated	Transcript products
XP_123456	Protein	Automated	Protein products
XR_123456	RNA	Automated	Transcript products
YP_123456	Protein	Auto. & Curated	Protein products
ZP_12345678	Protein	Automated	Protein products

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Primární databáze

The screenshot displays a genomic browser interface. At the top, the chromosome region is identified as NC\_002377.1: 145K..148K (2.9Kbp). A scale bar below shows coordinates from 145,400 to 147,600. A gene track shows a red bar representing the gene NP\_059797.1. A tooltip window is open over this gene, providing the following information:

- NP\_059797.1**
- NP\_059797.1: two-component VirA-like sensor kinase
- total range: NC\_002377.1 (145,694..148,183)
- total length: 2,490
- strand: plus
- protein product length: 829
- Links & Tools**
- GenBank View: [NC\\_002377.1 \(145,694..148,183\)](#), [NP\\_059797.1 \(145,694..148,183\)](#)
- 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](#)

Below the tooltip, there are sections for **Bibliography** and **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 GenBank entry for Agrobacterium tumefaciens plasmid Ti. The main content is the DNA sequence in FASTA format, starting with >110555014:14594-14813. The right-hand side features several interactive panels: 'Change region shown' (set to 'Selected region'), 'Customize view' (with options like 'Run BLAST'), 'Related information' (listing BioProject, Gene, etc.), and 'Recent activity' (showing recent searches for the plasmid).



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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





# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáváním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

```
>PDOC00001 PS00001 SULFATION Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].
171 - 585 skwsaatTetelaaa

>PDOC00004 PS00004 CAMP_PHOSPHO_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
744 - 747 RRVT
814 - 817 RRSE

>PDOC00005 PS00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
148 - 150 ESD
144 - 146 TSD
171 - 173 ESK
219 - 221 ESK
369 - 371 TSD
440 - 442 ESK
513 - 515 ESK
585 - 587 ESK
602 - 604 TSK
612 - 614 TSK
716 - 718 ESK
726 - 728 ESK
747 - 749 TSK
784 - 786 ESK
804 - 806 ESK
864 - 866 ESK
868 - 870 ESK
921 - 923 ESK
957 - 959 ESK
960 - 962 TSK
974 - 976 TSK
997 - 999 ESK
1009 - 1011 TSK
1018 - 1020 ESK
1031 - 1033 TSK
1119 - 1121 ESK
```



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáváním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

Hits for all PROSITE (release 2020\_05) motifs on sequence USERSEQ1 :

found 2 hits in 1 sequence

USERSEQ1 (1122 aa)

```
MRKQVTLVASSPIVWCVLAFLVWVFCINISDRITTEHVVWVAFPEKISTIVKRENTGK
FFAKMGLSTGLARVIGDITENDDSTFTEGQIAPLPAVSTIIGVQVYISBQKAPFIA
ESITSTAVANSSSSSSSDIYRITCTVQGLTQGLNDSIKQSLQVTHIDFCAAGNNTIAPY
DILGGEDRELLIQVYVLSRSGGLGSPFVGLTEVLSLGLSREELNHTKQDVLVREGSLM
SDFPSSSSICFQREISLNDKCFRCHSSQVSEVTEKPLSQAQVQVREYQVFAKTLIFPQGG
GATRIQAKAGATVQLYVHGFLQFQVNFVYHQAIRKEDD@ATLIQWQATQAKRQKQK
DGLFANNDYFSLKSNQGLLQVQDQVRFQVQVYVLLQVYVQVQVYVLLQVYVQVQVYV
NGLVEZDPLSKLLEVDIFPFLAGQVQVYVLLQVYVQVYVQVYVQVYVQVYVQVYVQVYV
RFTVQGLIARVAAQDFQSSSYVLAISIPQVSEPKQKQKQKQKQKQKQKQKQKQKQKQKQK
RYEYVQVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYV
DQVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYV
KFFVQVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYV
PFGQVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYV
FRSRLPFDMDGIDISGLQSSSIFRVAVLLVDAKTSDFRFLCQVQVQVQVQVQVQVQVQV
KLESSTVSEKQVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYV
NGLVEZDPLSKLLEVDIFPFLAGQVQVYVLLQVYVQVYVQVYVQVYVQVYVQVYVQVYV
VDCVFSRQVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYV
DQVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYV
DQVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYVQVYV
```

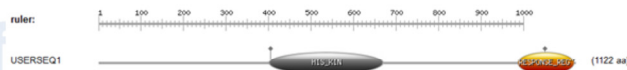
Legend:

□ disulfide bridge    ● active site    — other 'ranges'    ○ other sites

Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter. For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/mydomains/>

hits by profiles: [2 hits (by 2 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



DO ROZVOJE VZDĚLÁVÁNÍ

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

# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáváním primárních dat (sekvencí)
- PROSITE, <http://www.expasy.org/prosite/>

Hits for all PROSITE (release 2020\_05) motifs on sequence USERSEQ1 :

found 2 hits in 1 sequence

USERSEQ1 (1122 aa)

```
MMWVTKVLRSPVVFVLAFLVVFECINISNRITTEMLVVEVAFTEELATSLVSEINIQ
FTAKPLSTIGLARVIDITINDVDFTEIQIAPLLPVAISTILQVQVVISRDGLMFYIA
KSTFVAVFANSSSSSSSSVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
GTSLSGSESETLIQVQVLSVQVQVLSVQVQVLSVQVQVLSVQVQVLSVQVQVLSVQV
DSFFISDQICFRESNSLQGCIPENCSSQVVEIKGLRIGAFVIVSDFVFLTLIFPQNG
GALILHNSQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
SQAFAASDIPGALAGGLIDICRQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
NGLTEETFLSLELVITLTFSPVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
KFTVQSIARVAGQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
PVFVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
TCTFVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
PQESRQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
FQSELPFQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
WLESSTVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
KVEQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
VQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
DQKATRETRVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQVQV
```

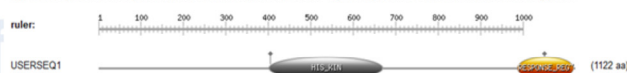
Legend:



Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter. For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/mydomains/>.

hits by profiles: 2 hits (by 2 distinct profiles) on 1 sequence

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



DO ROZVOJE VZDĚLÁVÁNÍ

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

# Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáváním 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 characterise a protein family; its diagnostic power is refined by iterative scanning of a PROTEIN/FAMILY/QUERY sequence. Usually the motifs do not overlap, but are scattered along a sequence, though they may be contiguous in 3D space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. [Reference](#)

#### New:

- [SPRINT](#) - Score & PRINTS-3 evolutionary PRINTS
- [comPRINTS](#) - Score & PRINTS automatic assignment
- [InterPro](#) - Search the integrated InterPro family database

#### Direct PRINTS access:

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

#### PRINTS search:

- Search PRINTS with **NEW FingerprintScan**
- [FPScan](#)
- [U.F.P.Scan](#)
- [MULScan](#)
- FingerprintScan binaries and source are available: [patrick.scofield@bioinf.man.ac.uk](mailto:patrick.scofield@bioinf.man.ac.uk)

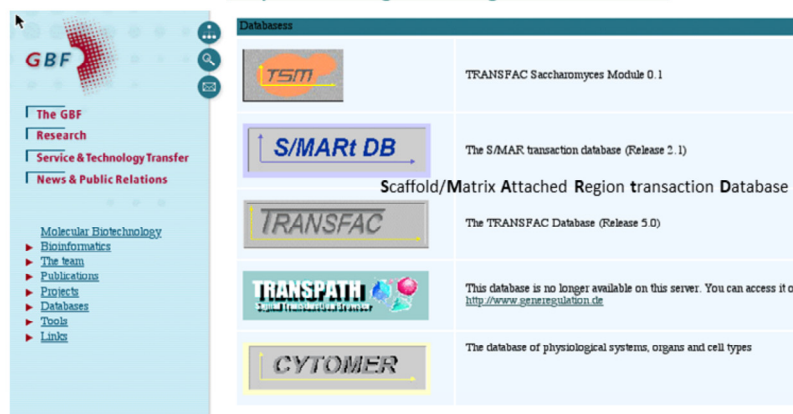


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 '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)
TRANSPATL	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 to be involved in the interaction of these elements with the nuclear matrix. <http://transfac.gbf.de/SMARTDB/index.html>

# Strukturální databáze

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

The screenshot shows the PDB website interface. At the top, it says "PROTEIN DATA BANK" with the RCSB logo and navigation links for Home, Contact, and Help. A welcome message states: "Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data." Below this are navigation links for "ABOUT PDB", "DATA UNIFORMITY", "RECENT FEATURES", "USER GUIDES", "FILE FORMATS", "EDUCATION", "STRUCTURAL GENOMICS", "PUBLICATIONS", and "SOFTWARE".

On the left side, there are links for "DEPOSIT data", "DOWNLOAD files", "Browse LINKS", "BETA TEST new features", and "BETA release files". Below these are "Current Holdings" statistics: "19623 Structures", "Last Update: 30-Dec-2002", and "PDB Statistics". A "Molecule of the Month" section features a 3D protein structure and the name "Cytochrome c".

The main content area includes a "Search the Archive" section with a search box, a "Find a structure" button, and checkboxes for "query by PDB id only", "match exact word", and "remove sequence homologues". There are also links for "SearchLite" and "Status Search".

On the right, there is a "PDB Mirrors" section listing various international mirrors such as "San Diego Supercomputer Center", "Rutgers University", "National Institute of Standards and Technology", "Cambridge Crystallographic Data Centre, UK", "National University of Singapore", "Osaka University, Japan", "Universidade Federal de Minas Gerais, Brazil", and "Max Delbrück Center for Molecular Medicine, Germany".

At the bottom of the page, there is a "News" section dated "23-Dec-2002" with a "Happy Holidays from the PDB!" message. There are also links for "Camelids News Newsletter" and "PDB Archive Subscribe".



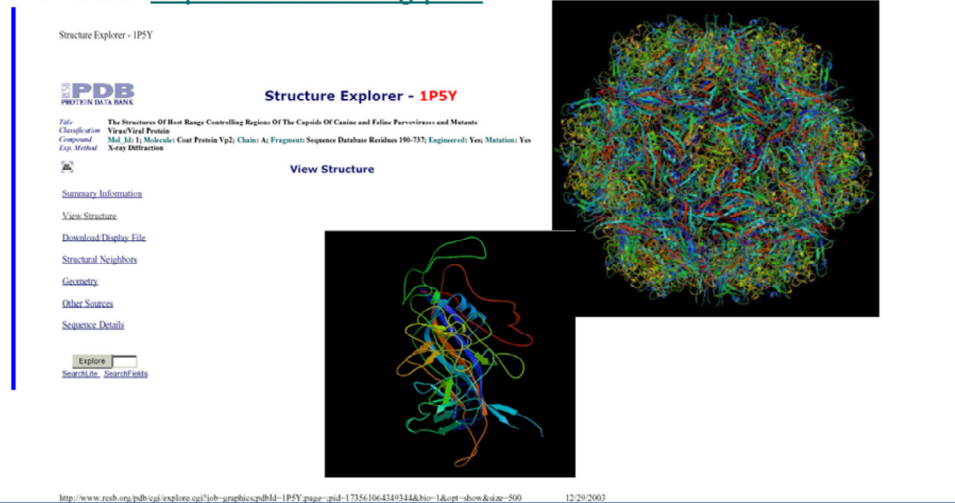
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Strukturální databáze

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

Structure Explorer - 1PSY



**PDB**  
PROTEIN DATA BANK

**Structure Explorer - 1PSY**

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

**View Structure**

**Summary Information**  
**View Structure**  
**Download Display File**  
**Structural Neighbors**  
**Geometry**  
**Other Sources**  
**Sequence Details**

**Explore**   
**Search**  **Search**

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

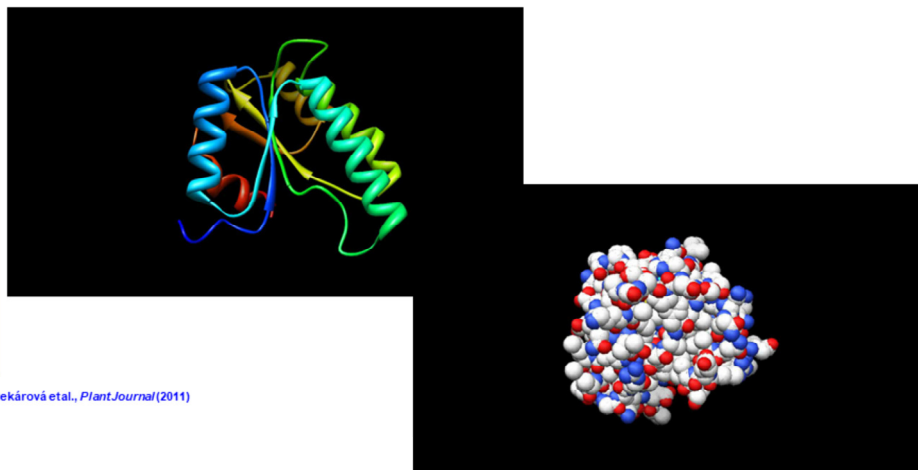


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Strukturální databáze

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



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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



# 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

□ **NCBI Genome Data Viewer** <https://www.ncbi.nlm.nih.gov/genome/gdv/>

## Genome Data Viewer

GDV is a genome browser supporting the exploration and analysis of more than 920 eukaryotic RefSeq genome assemblies. ⓘ

Select organism

Homo sapiens (human)



**Homo sapiens (human) genome**

Search in genome  
Location, gene or phenotype

Examples: TP53, chr17:7667000-7668000, rs334, DNA repair

Assembly  
GRCh38.p13

[Browse genome](#) [BLAST genome](#)

**Assembly details**

**Name** GRCh38.p13  
**RefSeq accession** GCF\_000001405.39  
**GenBank accession** GCA\_000001405.28  
**Download via FTP** RefSeq, GenBank  
**Submitter** Genome Reference Consortium  
**Level** Chromosome  
**Category** Reference genome

**Annotation details**

**Annotation Release** 109  
**Release date** 2020-08-17

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----



MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost



evropskými sociálními fondy a státním rozpočtem České republiky

# Genomové zdroje

Genome Browser Gateway <https://genome.ucsc.edu/>

The screenshot shows the UCSC Genome Browser Gateway interface. At the top, there is a search bar with fields for 'clade' (set to 'Human'), 'genome' (set to 'Feb. 2009 (GRCh37/hg18)'), and 'position' (set to 'chr21:33,031,597-33,041,570'). Below the search bar, there are links for 'Click here to reset the browser user interface settings to their defaults', 'track search', 'add custom tracks', 'track help', and 'configure tracks and display'. The main content area is titled 'Human Genome Browser - hg18 assembly (sequences)'. It includes a section for 'Sample position queries' and a table of 'Request' and 'Genome Browser Response' examples.

Request	Genome Browser Response
chr7	Displays all of chromosome 7
chr7p_g00212	Displays all of the unpaired contig p00212
20p13	Displays region for band p13 on chr 20
081.1:100000	Displays first million bases of chr 1, counting from p-arm telomere
chr3:100000-2000	Displays a region of chr3 that spans 2000 bases, starting with position 100000
RH1801:R80175 15q11:15q13 rs154252/rs1600376	Displays region between genome landmarks, such as the STS markers RH1801 and R80175, or chromosome bands 15q11 to 15q13, or SNPs rs154252 and rs1600376. This syntax may also be used for other range queries, such as between unparquet determined ESTs, mRNAs, refSeq, etc.
D18S3046	Displays region around STS marker D18S3046 from the Genethon/Manfield maps. Includes 100,000 bases on each side as well.
A020414	Displays region of EST with GenBank accession A020414 on BRCA1 cancer gene on chr 17
AC08101	Displays region of clone with GenBank accession AC08101
AF38211	Displays region of mRNA with GenBank accession number AF38211
FSNP	Displays region of genome with HSCO Gene Nomenclature Committee identifier FSNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_056160	Displays the region of genome with protein accession number NP_056160
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs
homovex cascade	Lists mRNAs for causal homeobox genes
zinc finger	Lists many zinc finger mRNAs
knapped zinc finger	Lists only knapped-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease
zeller	Lists mRNAs deposited by scientist named Zeller
Evans, J E	Lists mRNAs deposited by co-author J E. Evans



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# 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 top navigation bar includes 'Genome', 'Genome Browser', 'Tools', 'Downloads', 'My Data', and 'About Us'. The main content area is titled 'Human Gene HBB (uc001mae.1) Description and Page Index'. It contains a 'Description' section with a summary of the gene's function and clinical significance, a 'Page Index' with various links, and a 'Sequence and Links to Tools and Databases' section. A green arrow points to this section, which lists various databases and tools such as Genomic Sequence, Gene Name, CSAP, Ensembl, EMBL, Gene, ExonPrinter, GeneCards, GeneNetwork, Crisp, H-Inv, HNC, HNF, Jackson Lab, MIMED, OMIM, PubMed, Reaction, Standard SOURCE, TrEMBL, UniProtKB, and Wikipedia. Below this is the 'Comments and Description Text from UniProtKB' section, which provides detailed information about the protein's structure, function, and associated diseases.

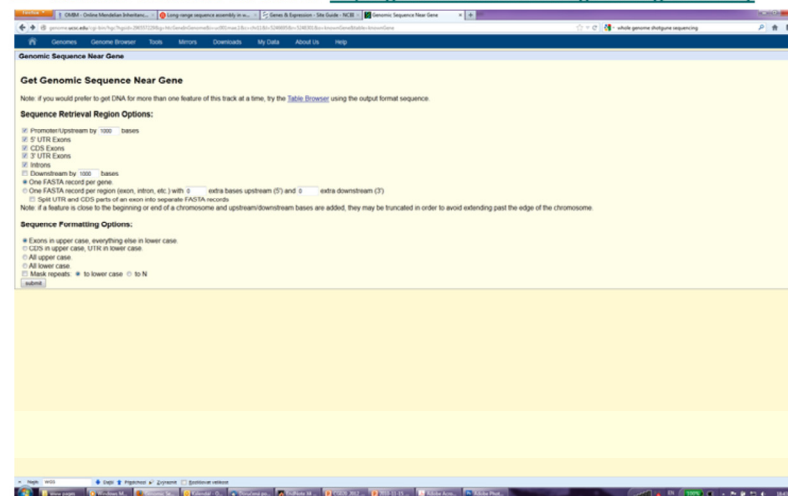


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Genomové zdroje

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

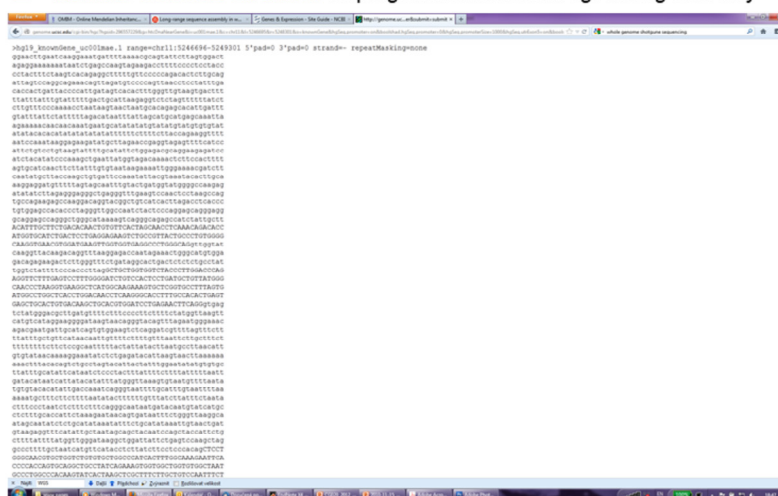


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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>



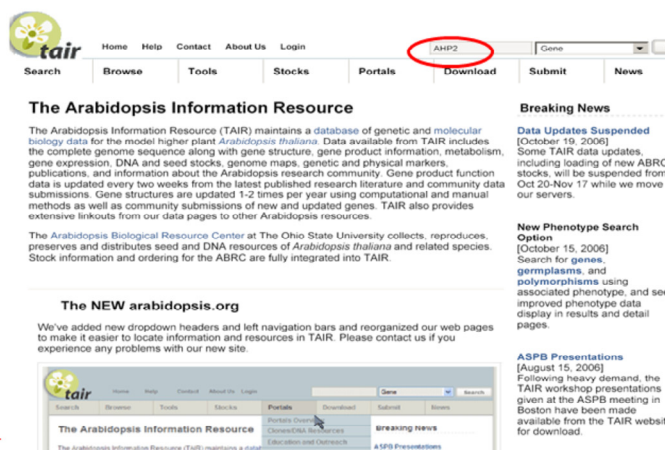
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



# Genomové zdroje

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



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

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

**The NEW arabidopsis.org**

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

**Breaking News**

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

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

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



## INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# 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  
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVE
```

### Lokální přiřazení

```
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVE  
-----NAPATNIKSECVRA-PIQNYRRVEHVRA-----
```

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

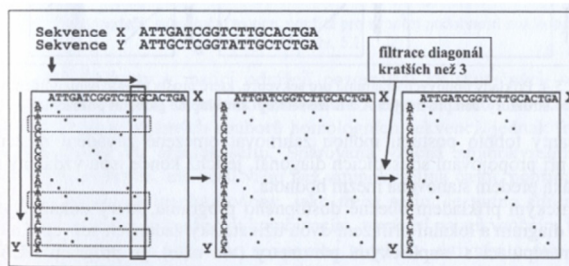


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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)

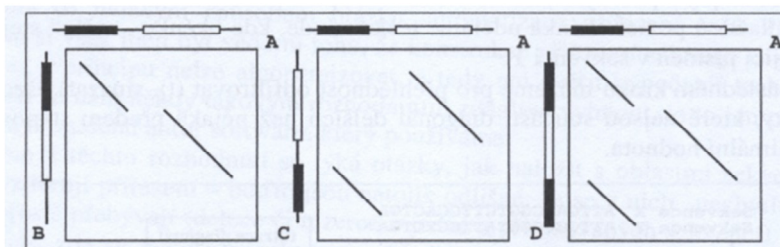


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 RID

[Search](#)

```
acaccatcgtg  
acaccatcattatcacc atcgcttttg ggcgatgttg tgggttcca  
gcytattaat  
ataattaatt tattccacat gagatgat atgatatact atgtattttt  
tgttttttt  
ttatttgtaa acotthaata taacaagaac tacaaaaaat gaaa
```

[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

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



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

# BLAST

## Basic Local Alignment Search Tool

actin, beta (ACTB), mRNA

Score = 1110 bits (560), Expect = 0.0  
Identities = 965/1100 (87%)  
Strand = Plus / Plus

Query: 156 gtgacaaaggctctggcatgtgcaaggcggattgccggagacgatgctccccggcc 215  
Sbjct: 101 gtgacaaaggctccggcatgtgcaaggcggcttcgggggacgatgccccgggcc 160

Query: 216 gtcttcccatcgattgtggacgtccccgtcaccagggtgtgatggtggcatggccag 275  
Sbjct: 161 gtcttccctccatcgtggggcggccaggcaccaggcgtgatggtggcatgggtcag 220

Query: 276 aaggactcgtacgtgggtgatggggcagagcaagcgtggtatcctcaccctgaagtac 335  
Sbjct: 221 aaggattcctatgtggggacgagggccagagcaagagaggcctcctcaccctgaagtac 280

Query: 336 ccattgagcacggatcgtgaccaactgggacgatggagaagatctggcaccacacc 395  
Sbjct: 281 ccctcggacacggatcgtcaccaactgggacgatggagaaatctggcaccacacc 340

E= expectancy value

ds..S=1213 E=0.0

>=200

250 1500

- „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

The screenshot shows a web browser displaying a GenBank entry for the gene NP\_059797.1. The browser address bar shows the URL [https://www.ncbi.nlm.nih.gov/nuccore/NC\\_002377.1](https://www.ncbi.nlm.nih.gov/nuccore/NC_002377.1). The main content area shows a genomic map with a red bar representing the gene NP\_059797.1. A tooltip window is open over the gene, displaying the following information:

- NP\_059797.1**
- NP\_059797.1: two-component VirA-like sensor kinase
- total range: NC\_002377.1 (145,694..148,183)
- total length: 2,490
- strand: plus
- protein product length: 829
- Links & Tools**
- GenBank View: [NC\\_002377.1 \(145,694..148,183\)](#), [NP\\_059797.1 \(145,694..148,183\)](#)
- 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](#)

Below the tooltip, there are sections for **Bibliography** and **Related articles in PubMed**.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

# BLAST

## Basic Local Alignment Search Tool

Pre-computed BLAST results for: [a16119781rvf/NP\\_396485.1](#) two component sensor kinase [Agrobacterium tumefaciens str. C58]

Matching gis: [15163423.20141874.1019660](#)

Total (score > 100) : 147086 hits in 146754 proteins in 6309 species

Selected: 147086 hits in 146754 proteins in 6309 species Filter: Min Score: 100 |

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

[Reset all filters](#)

Choose Display Options

1203 Archaea 138295 Bacteria 13 Metazoa 1349 Fungi 554 Plants 6 Viruses 5676 The Others [reset selection](#)

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

% hits	Score	Accession	Length	Protein Description
833 aa				
4166	AM99527	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]	
4166	P18548	833	ProName: Full-Wide host range virA protein Short-WDR virA	
4166	AAA79262	833	virA [Plasmid pTIC58]	
4159	NP_053300	833	hypothetical protein pT1-GAMMA_p142 [Agrobacterium tumefaciens]	
4159	AAA07765	833	tiorf140 [Agrobacterium tumefaciens]	
4153	AAA91590	833	virA [Plasmid Ti]	
4153	g11737127	833	virA protein	
4153	CAA34777	833	91.3 kDa protein [Agrobacterium tumefaciens]	
3800	CAA33380	829	virA [Agrobacterium rhizogenes]	
3718	g11227240	849	virA gene	
3348	AAA88643	829	virA [Plasmid Ti]	



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# BLAST

## 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 **datábázi proteinových sekvencí**
  - **BLASTN**
    - vyhledávání podobnosti k **nukleotidové sekvenci** v **datábázi 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 **datábázi 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 Initiated BLAST)**
    - 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  
VADAKGDSSESEDEDELEVVPVPSRFNRRVSVCAETYNPDEBBEDTDPRVIHPKTDEQRCRLQBEACKDILLF  
KNLDQEQLSQVLDAMFERIVKADHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA  
LMYNTPRAAITVATSEGSWGLDRVTFRRIIVKNNAKRRKMFESFIESVPLLSLEVSRMKIVDVIGEK  
IYKDBERIITQGEKADSFYIIBSGEVSILIRSRTKSNKDGNGQBEVEIARCHKGQYFGBLALVINKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYBQLVKMFSSVDLGNLGQ
```

```
[LIVMF] -G-E-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 **Biology Workbench** <http://workbench.sdsc.edu/>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Analytické nástroje

- o **Biology Workbench** <http://workbench.sdsc.edu/>

View  
View Nucleic Sequence(s)

Format: Fasta Case: Upper Change Format

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

[NEXT] [BOTTOM]

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

> 170248  
GAGCTCCCTTGGGGGGCAAGGGCAAAAACCTTTTGGCTAAATGGAAAAATATATACCAGTGTGTTGTAATA  
GTTACTCAATTTGAATTAACAAGGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCACAAAAC  
ATAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTCTTTAATTTATAGTTAGTT  
GACAAAACACTATCAAGATATCATTAATTAATAAATAACTTCAAGTCCATCATCTTAGCTGCTCCCTCA  
GTTAGAGCCCGAGTAAATAGACCGATCAATAAAGCCCGCATTAATAATAGAAATTTAGGACTCTC  
GATTTGGACGTAACTCCAAAACCTTCCAAATCTTTTCCCAAGCTTTGGGGCTCTGAGTCTGAGCTTC  
CAGATATGGGATATTTCTAAGTTTATCTCTTAATTTACATCTCAACTAATATTAAGAAATTAACAGGTA  
CAGCAATCATAAAATTTCTCTTAAGAAAGCAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG  
TCATGATCCCATATTCCTAAGGGGTCGTTTGGTACAAAGAAATAATAATAATTTGGGATAGAAATTT  
GAGATTCATTTATCTTTGTTTAAATTAAGATTTAGCTAATTCAGAAATTAATTTGCTTAAATATAG  
TAAATCACTTTCACATTTAGAAAGTGAATGGAAATGCTAATCCATAGCCACTCACTAGAAATTTCC  
TTAATTTATCTACATTTTACCAAATGATCGTTAGTCTTCAAGAGAAATCCAGTATCTCAATAAATGCA  
GTAGAAAGTTAGAAAATTTCTAATTAATCAATTCATATAATTTAAATAATTTAGATATGGAGCACTTAAG  
ATACAAATAAAGATGTACCGTTAATAATAAAGATAAGATAGATTTTAAATAGGAAAAAAAACGGTT  
CGAGACTCTTTATGGAAAGGGGTGCTTCAAGTAGATTTCTATCATTTGCTCTGGTGCATAGCAAAA  
TACACTTTGCTCTTAGATTCACCGAGCCACTTCAATCTCTTATTTATCTCAAAAGAAAGTTTAA  
GAGACTTTCAAACTTCAACTACTTTTAAAGGAAATTCAAAATACGACCAATTTATTTACTTACTTAC  
TTATAGTTAAATGATATGAATTTTAAATTTGAAATGAAAATATTAATTTACTTGTATTAATATA



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Analytické nástroje

- o **Biology Workbench** <http://workbench.sdsc.edu/>

**Regex pattern:**

ott. {1, 32}ott

0 sequences were searched

1 match was found

Matches are indicated in blue

```
>170248
GAGCTCCCTTGGGGGGCAAGGGCAAACCTTTGGCTAAATGGAAAAATATATACCAAGTGTGTAATA
GTATCTCAATTGAAATTAACAAAGGGCAAAATTTGACTATTTTGGCCCTTATATCTTTTGGTCACAAAAAC
ATAAATATCCCATCCGAATTCGAATGGTCCATATCGGCAAGTAGCTTTCTTTTAAATATAGTTAGTT
GACAAACCTATCTACAGATTCATTATTAATTAATTAATTTCAAGGTCCTTCTTTAGTCCCTCTCA
GTAGAGCCGCGATAAATAGACCGATCAANTRAAAGCCGCCATTAATAATAGAAATTTAGGACTCTC
GATGGCACGTAAAGTCCAAAACCTCTCCAACTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATCTAAGTTTATCTCCTAATTCATCTCAACTAATTAAGAAATTAACAGGTA
CAGCAAKTATAAATTTCTCTAAAGAGACAAATGAGTCCGTTACTATTCATGSCCTTTCTAGAG
TCTGATGCCATTTACTAAGGGGCTGTTGGTACAAGAAATATAATAATTTCCGGATAGAAATTT
TAAATCAACTATACATGTAGAGGTGGAATGGAATAGTAATCCATAGCCACTACATAGAAATCC
TATTATCTACTATTTTACAAATGATCGTTAGTCTTCAATGAAATCCAGTATCTCAATTAATGCA
GTAGAAATTTAGAAATTTTCAATTAATCAATTTCAATTTTAAATTTTAGATTTAGGCACTTAG
ATACAATAAAGATGACCGTAAATTAATAAAGATAGATAGAGTTTAAATAGGAAAAAAAACCGTT
CGAGACTTTTATGGAGGGCTTGTCTCAAGGTAGATCTCATTCATGCTCTGGTCAATAGCAAAA
TGACATTTACTCTTAGATACAGCGACTCTACAACTTCTATTTGTAATCAAAATGAAAGTTTAA
GAGAACTTCAAACTCTCACTACTTTTAGGGAAATCAAAATACGACCAATTTATTAATTTACTAC
TTATGTTAAATGATAGAAATTTATTTAAATTTGAAATGAAATTTAAATTTAGTTTAAATATAA
ACAATAGATATCGCTAAGTATTTACCAACAACATGGAGATACAGAAAGATTTATTTATTTAGCAT
GATTAAGCAGCTATTCATCTGGTTGTGAGGATGAAGAAAGTAACAGCTATTAATTTCTTTGTAAGT
```



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Analytické nástroje

- **Biology Workbench** <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 K L FAKWKNI I P S V C N S Y S I \* I N K G A N L T I L P L

E L P W G A R A K L F A K W K N I I P S  
1 g a g c t c c o t t g g g g g c a a g g g c a a a a c t t t t g c t a a 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  
S S L G Q Q N F L L N G K I L V Q V F I V T Q F E L T R G Q I \* L F C P

S S L G Q Q N F L L N G K I L Y Q V  
2 a g c t c c o t t g g g g g c a a g g g c a a a a c t t t t g c t a a t g g a a a a t a t t a t a c c a a g t 61  
F V I V T Q F E L T R 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 **Biology Workbench** <http://workbench.sdsc.edu/>

```
= Linear Map of Sequence:

      SbyI
      BsaJI
      CviJI
      AluI
      SacI
      EcoICRI
      Bsp1286I
      BsiHKAI
      BanII  BslI
      \ \ \ \ \
1 gagctcccttgggggcaagggaacttttgcataatgaaaaatattataccaagt 60
ctcgagggaacccccctccgctttgaaaaagattaccttttataataggttca
      * * * * *
1 E L P W G A R A K L F A K W K N I I P S
2 S S L G G Q G Q N F L L N G K I L Y Q V
3 A P L G G K G K T F C * M E K Y Y T K C
4 L E R P P C P C F K K S F F F I N Y W T
5 S S G Q P A L A F S K A L H F F I I G L
6 L A G K P P L P L V K Q * I S F Y * V L

      Tsp509I
      MaeIII Tsp509I  MseI
      \ \ \ \ \
61 gtttgaatggttactcaattgaattaacaaagggaactttgactattttgcoccta 120
caaacattatcaatgagttaaacttaattgtttccocgttaaacgtataaacggggaat
      * * * * *
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
```

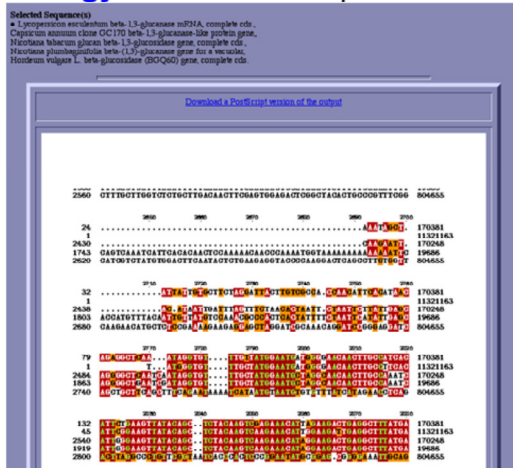


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Analytické nástroje

- **Biology Workbench** <http://workbench.sdsc.edu/>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Analytické nástroje

- o **Biology Workbench** <http://workbench.sdsc.edu/>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Analytické nástroje

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

SEARCH  ABOUT DOWNLOAD LINKS

VPCR 2.0 (WWW interface) - Please, enter nucleotide primer sequences (IUB codes allowed for degenerate primers). VPCR 2.0 searches the specified database for matches to the primers. If matches are found within 10000 bases, a PCR simulation model predicts amplification. Calculated PCR products are displayed within a minute.  
NOTE: Abilities of VPCR 2.0 are still limited by BLAST capabilities and settings, as well as instability of our current software to deal with more than a couple thousand matches per primer. For example, using primers shorter or roughly equal to our 11-base word size misses most matches. Primers with overrepresented sequences cause problems as well. We are now busy solving most of these problems, please, be patient. If you have a minute, please, let us know what kind of expectations you have for VPCR 2.0 etc. Currently, this address is for testing VPCR 2.0, stable features will be installed on [VPCR 2.0 Homepage](#).

Search using: BLAST in the database for: M. musculus

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

Annealing temperature: 50

Do PCR! 



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



# Analytické nástroje

- Virtual PCR (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 homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst...
  - Další [www genomové nástroje](#)



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

# Další WWW zdroje

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

PHACTR4 phosphatase and actin regulator 4 [Homo sapiens] - Gene - NCBI - Medline Plus

PHACTR4 phosphatase and actin reg... PHACTR4 phosphatase and actin reg... PHACTR4 phosphatase and actin reg... PHACTR4 phosphatase and actin reg...

NCBI Gene

PHACTR4 phosphatase and actin regulator 4 [Homo sapiens]

Gene ID: 65978, updated on 27-Aug-2011

Summary

Official Symbol: PHACTR4 provided by HUGO

Official Full Name: phosphatase and actin regulator 4 provided by HUGO

Primary source: HUGO (2012)

Location tag: BP11-442N24\_A.1

See related: Ensembl (ENSG00000204813), RefSeq (NM\_020725)

Gene type: protein coding

RefSeq status: REVIEWED

Organism: HOMO SAPIENS

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhina; Catarrhini; Hominoidea; Homo

Also known as: FLJ13171, MGC25818, MGC34186, DAF-2a68L7.205, BP11-442N24\_A.1

Summary: This gene encodes a member of the phosphatase and actin regulator (PHACTR) family. Other PHACTR family members have been shown to inhibit protein phosphatase 1 (PP1) activity, and the homolog of this gene in the mouse has been shown to interact with actin and PP1. Multiple transcript variants encoding different isoforms have been found for this gene. (provided by RefSeq, Jul 2008)

Genomic context

Location: 1:1235

Sequence: Chromosome 1, NC\_000011.10 (3894993-3895891)

Chromosome 1 - NC\_000011.10

Genomic regions, transcripts, and products

Genomic Sequence: NC\_000011 chromosome 1 reference GRCh37 p5 Primary Assembly



MINISTERSTVO  
MLÁDEŽE

JE VZDĚLÁVÁNÍ  
je spolufinancována  
kým sociálním fondem  
České republiky

# Další WWW zdroje

- **Online Mendelian Inheritance in Man (OMIM)** <http://www.omim.org/>



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