

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
hejatk@sci.muni.cz, www.ceitec.muni.cz



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Osnova

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Schéma předmětu

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



EVROPSKÁ UNIE

esf



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



MUNI
MASARYKŮVA UNIVERZITA BRNO

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**
 - Strukturální genomika



EVROPSKÁ UNIE

esf



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



MUNI
MASARYKIANA BRUNENSIS

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, Wiley-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 (sekvencí)-práce s databázemi
- V užším pojetí zkoumá FUNKCI jednotlivých genů - **FUNKČNÍ GENOMIKA**
 - používá zejména přístupy REVERZNÍ GENETIKY



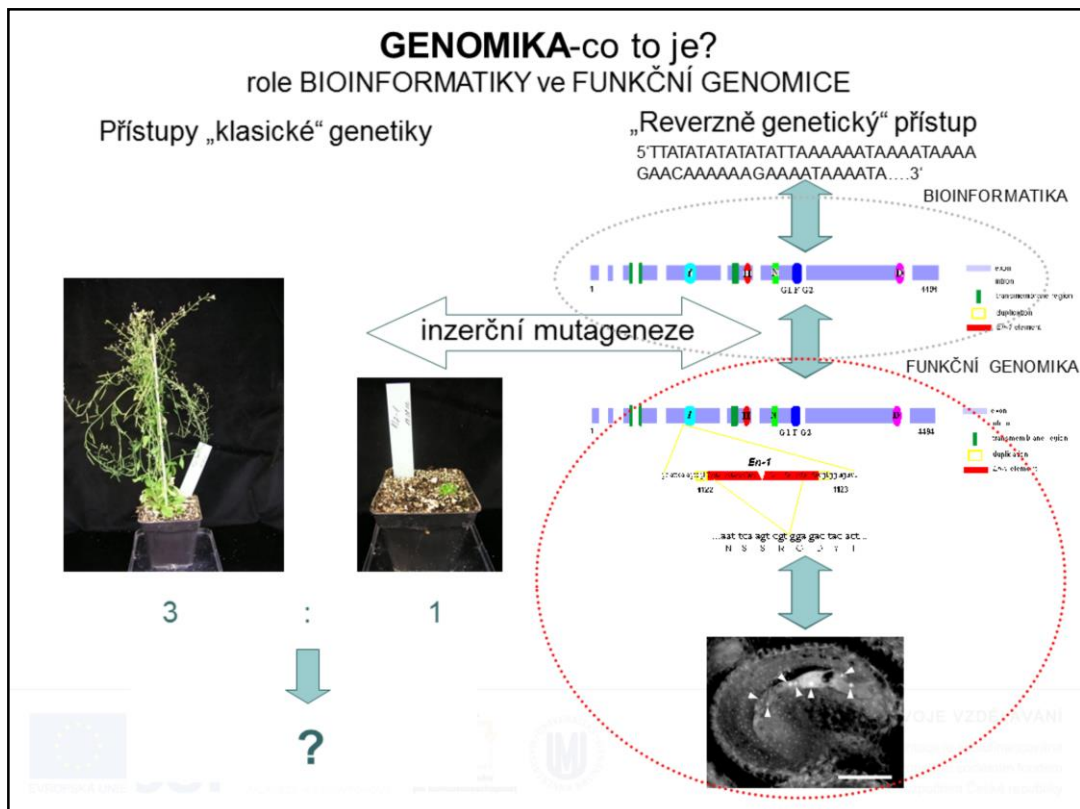
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



EVROPSKÁ UNIE

esf



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost

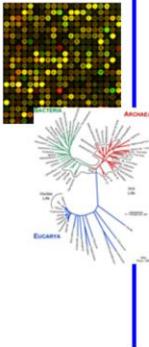


MUNI
MASARYKŮVA UNIVERZITA
BRNO

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.

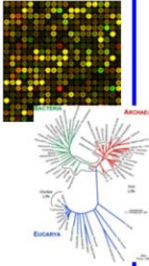
J. Pevsner,
<http://www.bioinfbook.org/index.php>



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



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



MASARYKŮVA
UNIVERSITA
BRNO

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Spektrum on-line zdrojů

EMNet National Nodes		
Vienna BioCenter	Austria	http://www.at.embnet.org/
EMN	Belgium	http://www.be.embnet.org/
EMN	Denmark	http://dnk.embnet.org/
EMN	Finland	http://www.fi.embnet.org/
EMN	France	http://www.fr.embnet.org/
EMN	Germany	http://www.de.embnet.org/
EMN	Greece	http://www.gr.embnet.org/
EMN	Hungary	http://www.hu.embnet.org/
EMN	Ireland	http://www.ie.embnet.org/
EMN	Italy	http://www.it.embnet.org/
EMN	Netherlands	http://www.nl.embnet.org/
EMN	Norway	http://www.no.embnet.org/
EMN	Poland	http://www.pl.embnet.org/
EMN	Portugal	http://www.pt.embnet.org/
EMN	Russia	http://www.ru.embnet.org/
EMN	Spain	http://www.es.embnet.org/
EMN	Sweden	http://www.se.embnet.org/
EMN	Switzerland	http://www.ch.embnet.org/
EMN	UK	http://www.uk.embnet.org/
EMNet Specialist Nodes		
EMN	Germany	http://www.emn.de.embnet.org/
EMN	Italy	http://www.emn.it.embnet.org/
EMN	Sweden	http://www.emn.se.embnet.org/
EMN	Switzerland	http://www.emn.ch.embnet.org/
EMN	UK	http://www.emn.uk.embnet.org/
EMN	UK	http://www.emn.uk.embnet.org/
EMN	UK	http://www.emn.uk.embnet.org/
EMNet Associate Nodes		
EMN	Argentina	http://www.emn.ar.embnet.org/
EMN	Australia	http://www.emn.au.embnet.org/
EMN	China	http://www.emn.cn.embnet.org/
EMN	Cuba	http://www.emn.cu.embnet.org/
EMN	India	http://www.emn.in.embnet.org/
EMN	South Africa	http://www.emn.za.embnet.org/
EMNet Information Providers		
EMN	USA	http://www.emn.us.embnet.org/
EMN	USA	http://www.emn.us.embnet.org/
EMN	USA	http://www.emn.us.embnet.org/



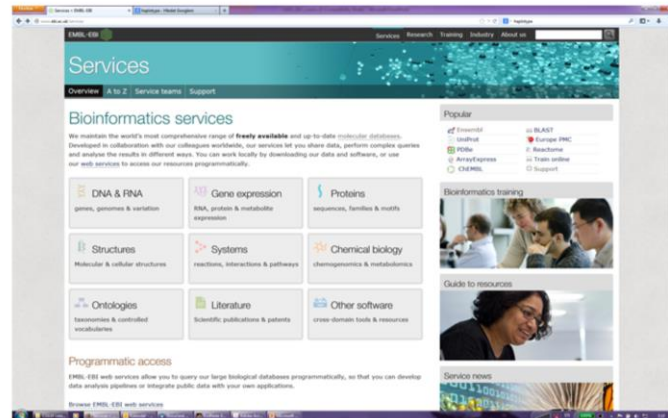
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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>



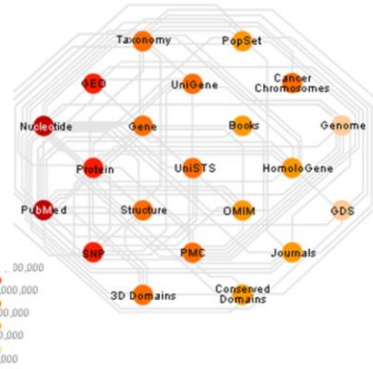
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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**



EVROPSKÁ UNIE

esf



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



MASARYKŮVA UNIVERZITA
BRNO

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)



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost

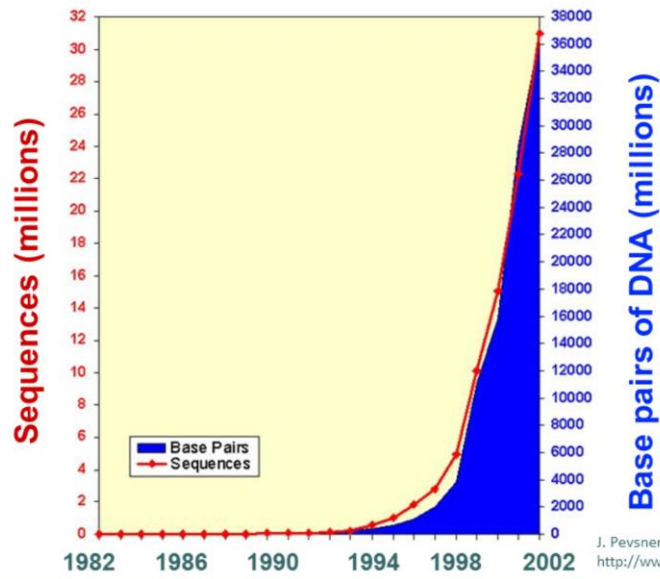


MASARYKŮVA
UNIVERSITA
BRNO

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

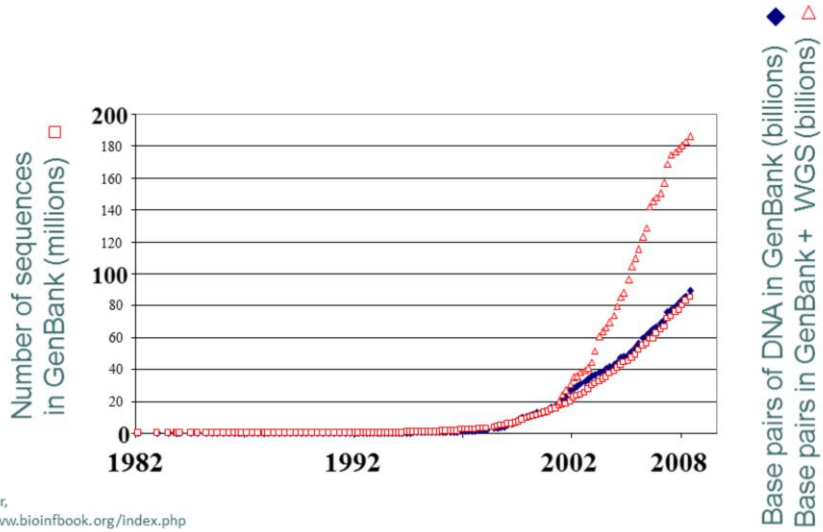
Growth of GenBank



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>

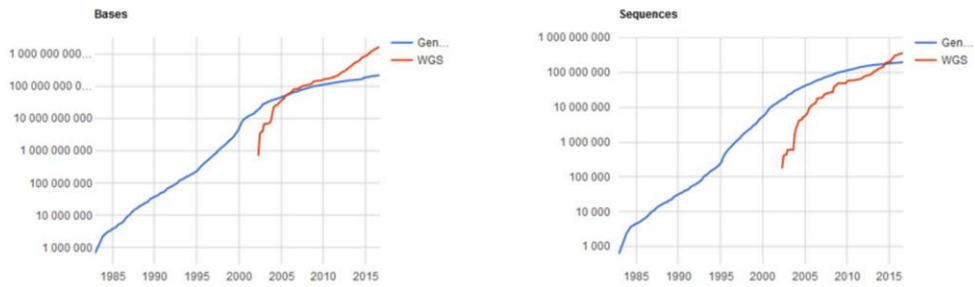


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Growth of GenBank

Aug 2016



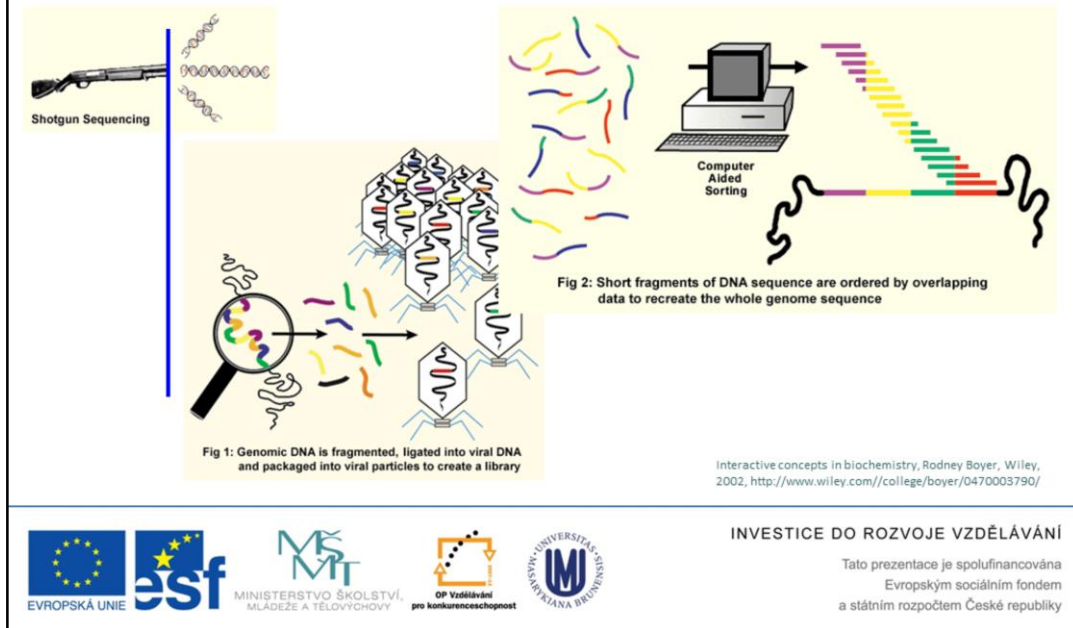
- Prosinec 1982 680 338 bp, 606 sekvencí
- Duben 2002 19×10^9 bp, 17×10^6 sekvencí + WGS 692×10^6 bp, 172 768 sekvencí
- Srpen 2016 218×10^9 bp, 196×10^6 sekvencí + WGS $1,6 \times 10^{12}$ bp, 360×10^6 sekvencí



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

WGS

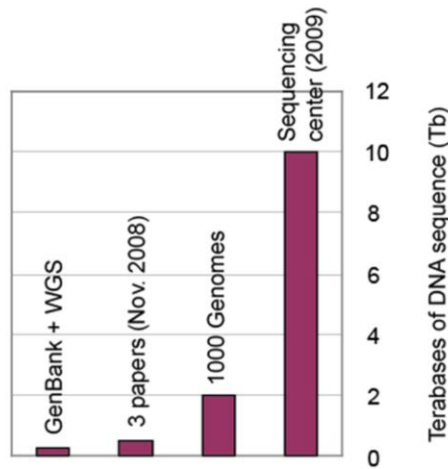


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic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

Arrival of **next-generation sequencing**:
 In **two years** we have gone from **0.2 terabases** to **71 terabases** (71,000 gigabases) (November 2010)



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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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

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

Primární databáze

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



EVROPSKÁ UNIE

esf



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



MASARYKŮVA
UNIVERZITA
BRNO

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** (**E**xpressed **S**equences **T**ags)
 - **HGTS** (**H**igh **T**hroughput **G**enome **S**equencing)
 - neanotované „surové“ výsledky sekvenačních projektů
 - Referenční sekvence anotovaných genomů
 - **TPAs** (**T**hird **P**arty **A**notation)
 - sekvence anotované jinými než původními autory

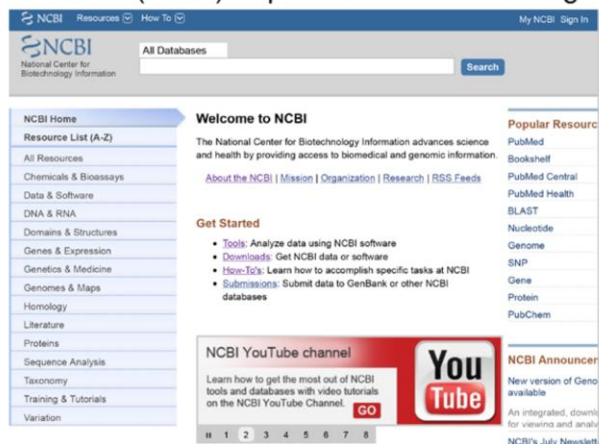


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze

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



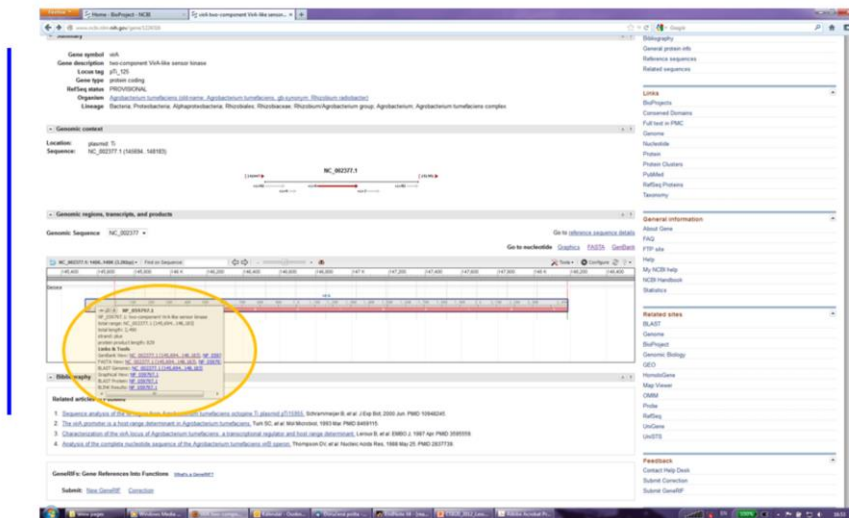
The screenshot shows the NCBI homepage with a search bar at the top. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to tools, downloads, and submissions, and a 'Popular Resources' list on the right. A 'NCBI YouTube channel' banner is also visible.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

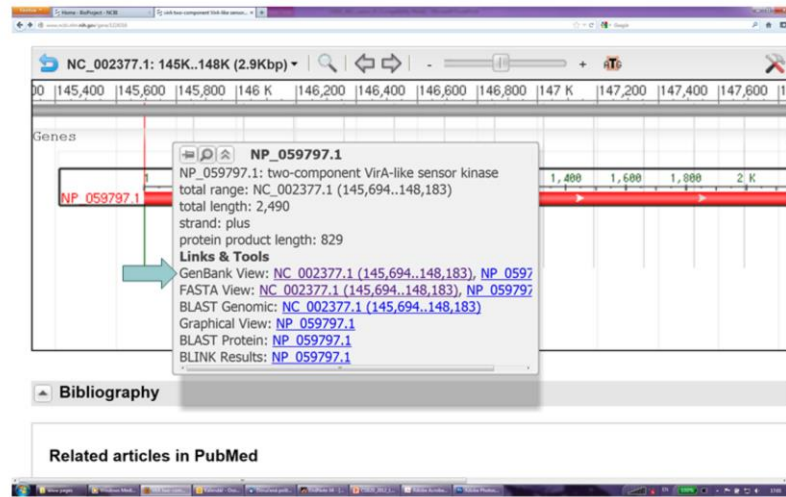
Primární databáze



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 record for the gene NP_059797.1. The record is titled "NC_002377.1: 145K..148K (2.9Kbp)". The gene is described as a "two-component VirA-like sensor kinase" with a total range of NC_002377.1 (145,694..148,183) and a total length of 2,490. The strand is plus, and the protein product length is 829. A "Links & Tools" section provides various links: GenBank View, FASTA View, BLAST Genomic, Graphical View, BLAST Protein, and BLINK Results. Below the links, there are sections for "Bibliography" and "Related articles in PubMed".

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] for [virk] [Go] [Clear] [History]

NC_024277.1

LOCUS NC_024277 2490 bp DNA linear BCT 19-DEC-2003

DEFINITION *Agrobacterium tumefaciens* strain C58 plasmid Ti, complete sequence.

ACCESSION [NC_024277.1](#) ([GI:15955014](#))

VERSION NC_024277.1 ([GI:15955014](#))

KEYWORDS

SOURCE *Agrobacterium* (Rhizobium radiobacter)

ORGANISM *Agrobacterium* (Rhizobium radiobacter);

REPLICATION ORGAN *Agrobacterium*;

TITLE Ferrand, E.E.

JOURNAL Orthopox-type Ti plasmid sequence

DEPOSITED Deposited

REFERENCE 2 (bases 1 to 2490)

AUTHORS Dm, J., Oper, P.R., Schramm, J.W., R., Hookey, P.J., Ferrand, E.E. and Winans, S.C.

TITLE Direct Subsequence

JOURNAL Submitted (07-MAR-2003) Microbiology, Cornell University, Ithaca, NY, USA

COMMENT **PROTEIN** [virk](#). This record has not yet been subject to final NCBI review. The reference sequence was derived from [EF141261](#).

FEATURES

Location/Qualifiers

1..2490

source /organism="Agrobacterium tumefaciens" /mol_type="genomic DNA" /db_xref="GenBank:15955014" /plasmid="Ti" /taxid="1224314" /feature="retrotransposon"

gene orthopox-type" 1..2490 /gene="virk" /db_xref="GeneID:1224314" /gene="virk" 1..2490 /note="Two-component regulator of vir regulon, VirK is a transmembrane histidine kinase" /coding_start=1 /trans_start=1 /product="virk" /protein_id="NP_052795.1" /db_xref="GI:15955111"



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

What is an accession number?

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

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

X02775	GenBank genomic DNA sequence	DNA
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	RNA
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	Protein
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Page 27

NCBI's important RefSeq project: best representative sequences

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

RefSeq identifiers include the following formats:

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

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Page 27

RefSeq

The screenshot shows the NCBI RefSeq page for the gene 'two-component VWA-like sensor kinase'. The page is organized into several sections:

- NCBI Reference Sequences (RefSeq):** The top section, indicating the source of the reference sequences.
- Genome Annotation:** A section with a link to 'Explain' and a note that the following sections contain reference sequences for a specific genome build.
- Reference assembly:**
 - Genomic:** Lists assembly NC_003065.3 with a range of 18031-18332 and download options for GenBank, FASTA, Sequence, and Vector.
 - mRNA and Protein(s):** Lists transcript NP_295166.1 for the protein 'two component sensor kinase [Agrobacterium tumefaciens str. C58]'. It includes UniProtKB/Swiss-Prot ID P18540 and a 'Summary' section.
- Conserved Domains (3):** A table of domains with their locations and descriptions:

Domain ID	Description
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 ...
PRK13037	PRK13037: two-component VWA-like sensor kinase, Provisional
- Related Sequences:** A section at the bottom of the page.



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

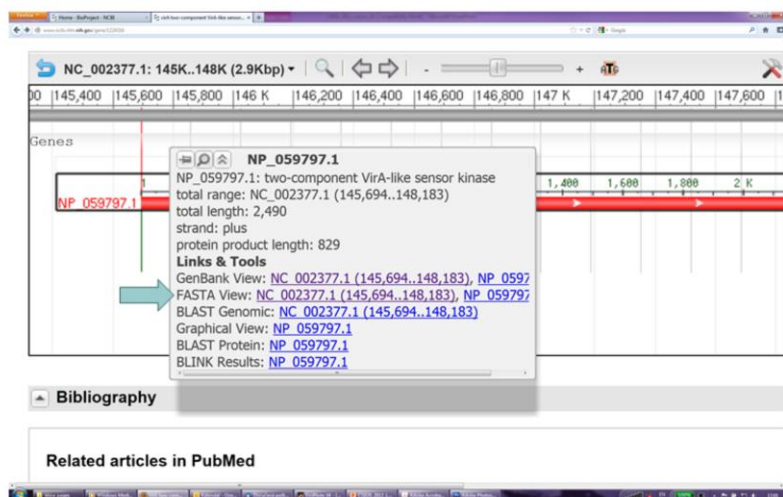
J. Pevsner,
<http://www.bioinfbook.org/index.php>



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze



The screenshot shows a web browser displaying a GenBank record for NP_059797.1. The record is titled "NC_002377.1: 145K..148K (2.9Kbp)". 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, 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](#)
- FASTA View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- BLAST Genomic: [NC_002377.1 \(145,694..148,183\)](#)
- Graphical View: [NP_059797.1](#)
- BLAST Protein: [NP_059797.1](#)
- BLINK Results: [NP_059797.1](#)

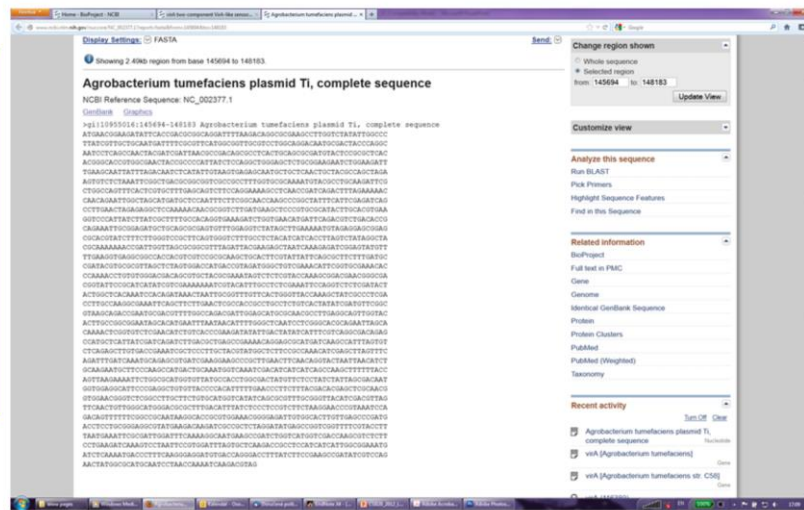
Below the tooltip, there are sections for "Bibliography" and "Related articles in PubMed".



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Primární databáze



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/>



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/>

>PDOC0001 PS0001 SULFATION Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].

171 - 185 sdsesstTetsttss

>PDOC0004 PS0004 CAMP_FIOPHOSPHO_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

744 - 747 DRVT

814 - 817 RRvG

>PDOC0005 PS0005 PKC_FIOPHOSPHO_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

148 - 150 QxR

184 - 186 TgR

172 - 173 QxR

219 - 221 QxR

368 - 370 TgR

440 - 442 QgR

513 - 515 QgR

585 - 587 QxR

602 - 604 TgR

632 - 634 TgR

714 - 716 QgR

726 - 728 QgR

787 - 789 TgR

798 - 800 QxR

804 - 806 QxR

884 - 886 QxR

868 - 870 QxR

922 - 924 QgR

957 - 959 QxR

1000 - 1002 TgR

974 - 976 TgR

997 - 999 QxR

1002 - 1004 TgR

1018 - 1020 QgR

1031 - 1033 TgR

1119 - 1121 QxR



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



PRINTS is a comparison 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 PROSITE-PRODOML1L1L2L3L4L5L6L7L8L9L10L11L12L13L14L15L16L17L18L19L20L21L22L23L24L25L26L27L28L29L30L31L32L33L34L35L36L37L38L39L40L41L42L43L44L45L46L47L48L49L50L51L52L53L54L55L56L57L58L59L60L61L62L63L64L65L66L67L68L69L70L71L72L73L74L75L76L77L78L79L80L81L82L83L84L85L86L87L88L89L90L91L92L93L94L95L96L97L98L99L100L101L102L103L104L105L106L107L108L109L110L111L112L113L114L115L116L117L118L119L120L121L122L123L124L125L126L127L128L129L130L131L132L133L134L135L136L137L138L139L140L141L142L143L144L145L146L147L148L149L150L151L152L153L154L155L156L157L158L159L160L161L162L163L164L165L166L167L168L169L170L171L172L173L174L175L176L177L178L179L180L181L182L183L184L185L186L187L188L189L190L191L192L193L194L195L196L197L198L199L200L201L202L203L204L205L206L207L208L209L210L211L212L213L214L215L216L217L218L219L220L221L222L223L224L225L226L227L228L229L230L231L232L233L234L235L236L237L238L239L240L241L242L243L244L245L246L247L248L249L250L251L252L253L254L255L256L257L258L259L260L261L262L263L264L265L266L267L268L269L270L271L272L273L274L275L276L277L278L279L280L281L282L283L284L285L286L287L288L289L290L291L292L293L294L295L296L297L298L299L300L301L302L303L304L305L306L307L308L309L310L311L312L313L314L315L316L317L318L319L320L321L322L323L324L325L326L327L328L329L330L331L332L333L334L335L336L337L338L339L340L341L342L343L344L345L346L347L348L349L350L351L352L353L354L355L356L357L358L359L360L361L362L363L364L365L366L367L368L369L370L371L372L373L374L375L376L377L378L379L380L381L382L383L384L385L386L387L388L389L390L391L392L393L394L395L396L397L398L399L400L401L402L403L404L405L406L407L408L409L410L411L412L413L414L415L416L417L418L419L420L421L422L423L424L425L426L427L428L429L430L431L432L433L434L435L436L437L438L439L440L441L442L443L444L445L446L447L448L449L450L451L452L453L454L455L456L457L458L459L460L461L462L463L464L465L466L467L468L469L470L471L472L473L474L475L476L477L478L479L480L481L482L483L484L485L486L487L488L489L490L491L492L493L494L495L496L497L498L499L500L501L502L503L504L505L506L507L508L509L510L511L512L513L514L515L516L517L518L519L520L521L522L523L524L525L526L527L528L529L530L531L532L533L534L535L536L537L538L539L540L541L542L543L544L545L546L547L548L549L550L551L552L553L554L555L556L557L558L559L560L561L562L563L564L565L566L567L568L569L570L571L572L573L574L575L576L577L578L579L580L581L582L583L584L585L586L587L588L589L590L591L592L593L594L595L596L597L598L599L600L601L602L603L604L605L606L607L608L609L610L611L612L613L614L615L616L617L618L619L620L621L622L623L624L625L626L627L628L629L630L631L632L633L634L635L636L637L638L639L640L641L642L643L644L645L646L647L648L649L650L651L652L653L654L655L656L657L658L659L660L661L662L663L664L665L666L667L668L669L670L671L672L673L674L675L676L677L678L679L680L681L682L683L684L685L686L687L688L689L690L691L692L693L694L695L696L697L698L699L700L701L702L703L704L705L706L707L708L709L710L711L712L713L714L715L716L717L718L719L720L721L722L723L724L725L726L727L728L729L730L731L732L733L734L735L736L737L738L739L740L741L742L743L744L745L746L747L748L749L750L751L752L753L754L755L756L757L758L759L760L761L762L763L764L765L766L767L768L769L770L771L772L773L774L775L776L777L778L779L780L781L782L783L784L785L786L787L788L789L790L791L792L793L794L795L796L797L798L799L800L801L802L803L804L805L806L807L808L809L810L811L812L813L814L815L816L817L818L819L820L821L822L823L824L825L826L827L828L829L830L831L832L833L834L835L836L837L838L839L840L841L842L843L844L845L846L847L848L849L850L851L852L853L854L855L856L857L858L859L860L861L862L863L864L865L866L867L868L869L870L871L872L873L874L875L876L877L878L879L880L881L882L883L884L885L886L887L888L889L890L891L892L893L894L895L896L897L898L899L900L901L902L903L904L905L906L907L908L909L910L911L912L913L914L915L916L917L918L919L920L921L922L923L924L925L926L927L928L929L930L931L932L933L934L935L936L937L938L939L940L941L942L943L944L945L946L947L948L949L950L951L952L953L954L955L956L957L958L959L960L961L962L963L964L965L966L967L968L969L970L971L972L973L974L975L976L977L978L979L980L981L982L983L984L985L986L987L988L989L990L991L992L993L994L995L996L997L998L999L1000

News:
[New!](#) - Search PRINTS, extended PRINTS
[New!](#) - Search PRINTS, automatic registration
[New!](#) - Search the integrated InterPro family database

Direct PRINTS access:

[By accession number](#)
[By PRINTS code](#)
[By domain code](#)
[By motif](#)
[By sequence](#)
[By title](#)
[By number of motifs](#)
[By author](#)
[By email address](#)

PRINTS search:

[Search PRINTS with NEW EprintPRINTS](#)
[EprintPRINTS](#)
[EprintPRINTS](#)
[EprintPRINTS](#)
EprintPRINTS binaries and source are available print@bioinf.man.ac.uk

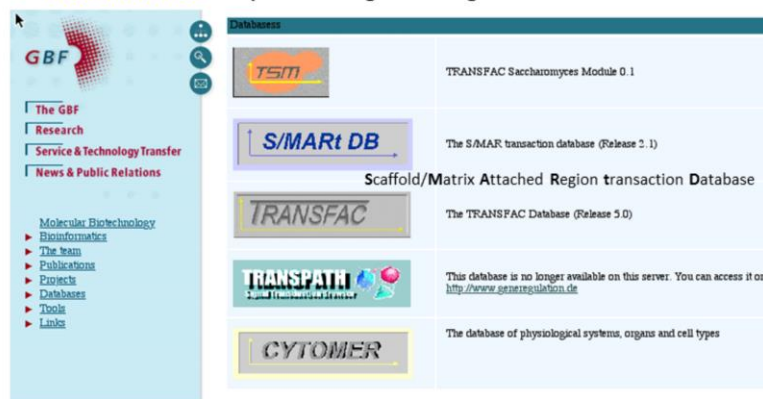


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Sekundární databáze

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



The screenshot shows the TRANSFAC website interface. On the left is a navigation menu for GBF (Gene Bioinformatics Foundation) with categories like 'The GBF', 'Research', 'Service & Technology Transfer', and 'News & Public Relations'. The main content area is titled 'Databases' and lists several databases:

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

S/MARt DB (saffold/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/>

[DEPOSIT data](#)
[DOWNLOAD files](#)
[Browse LINKS](#)
[BETA TEST new features](#)
[BETA mmCIF files](#)

Current Holdings

19623 Structures
Last Update: 30-Dec-2002
[PDB Statistics](#)



[Molecules of the Month](#)
[Cytochrome c](#)

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego, and the National Institute of Standards and Technology — three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the

PROTEIN DATA BANK

[RCSB Home](#) [Contact Us](#) [Help](#)

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data. [Did you find what you wanted?](#)

[ABOUT PDB](#) | [DATA UNIFORMITY](#) | [RECENT FEATURES](#) | [USER GUIDES](#) | [FILE FORMATS](#) | [EDUCATION](#) | [STRUCTURAL GENOMICS](#) | [PUBLICATIONS](#) | [SOFTWARE](#)

Search the Archive

Enter a **PDB ID** or **keyword**

[Query Tutorial](#)

query by PDB id only match exact word
 remove sequence homologues

[Searchable](#) keyword search form with examples
[Searchable](#) customizable search form
[Status Search](#) find entries awaiting release

News

[Complete Name](#) [pdb.J Archive](#)
[How to enter](#) [Subscribe](#)

23-Dec-2002

Happy Holidays from the PDB! The PDB staff wish to extend our **best wishes** to the community for a happy holiday season and a wonderful new year!



PDB Mirrors

"Please bookmark a mirror site!"

[San Diego Supercomputer Center](#)
[Rutgers University](#)
[National Institute of Standards and Technology](#)
[Cambridge Crystallographic Data Centre, UK](#)
[National University of Singapore](#)
[Osaka University, Japan](#)
[Universitäts-Federal-IdMuseum, Graz, Austria](#)
[Max Delbrück Center for Molecular Medicine, Germany](#)

[OTHER SITES](#)



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Strukturální databáze

○ **PDB** <http://www.rcsb.org/pdb/>

Structure Explorer - 1PSY

PDB
PROTEIN DATA BANK

Structure Explorer - 1PSY

Title: The Structures Of Hot Spots Controlling Regions Of The Capsids Of Canine and Feline Parvoviruses and Mutants
Classification: Virus/Viral Protein
Compound: Mol. Wt.: 13.366kDa; Cont. Protein Vp2; Chain: A; Fragment: Sequence Database Residues 180-735; Engineered: Yes; Mutation: Yes
Exp. Method: X-ray Diffraction

[View Structure](#)

[Summary Information](#)

[View Structure](#)

[Download Display File](#)

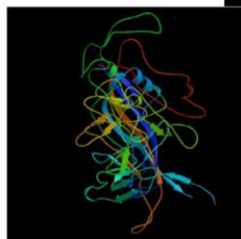
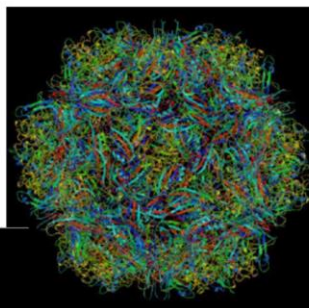
[Structural Neighbors](#)

[Geometry](#)

[Other Sources](#)

[Sequence Details](#)

[Search file](#) [Search ref seqs](#)



<http://www.rcsb.org/pdb/cgi/explorer.cgi?job=graphics&pdbid=1PSY&page=job-173561064249344&bio=1&opt=show&size=500> 12/25/2003

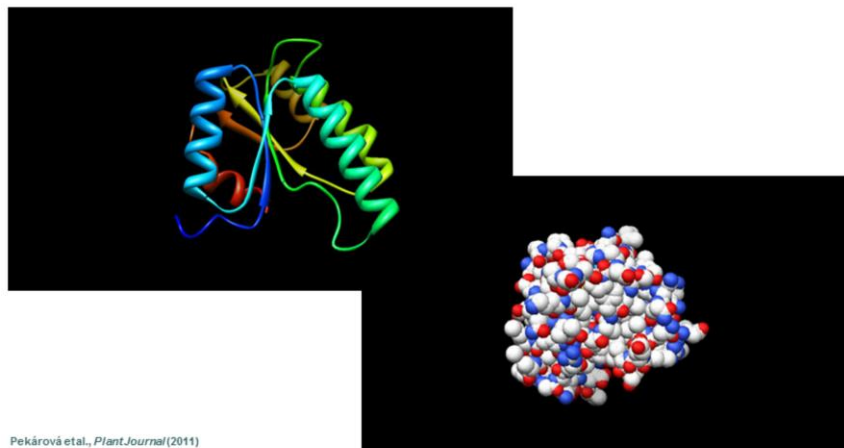


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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)



EVROPSKÁ UNIE



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



EVROPSKÁ UNIE

esf



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



MASARYKŮVA UNIVERZITA
BRNO

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 Human Genome Browser interface. At the top, there are navigation tabs for 'Genomes', 'Genome Browser', 'Tools', 'Metrics', 'Downloads', 'My Tools', and 'About Us'. Below this is a search bar with fields for 'genome', 'assembly', and 'position'. A search button is located to the right of the 'position' field. Below the search bar, there is a section titled 'Human Genome Browser - hg19 assembly [hg19/ncbi]' with a link to 'Get your own custom tracks...'. The main content area is titled 'Sample position queries' and contains a list of queries with their corresponding descriptions. A small image of a human figure with a genome map is visible on the right side of the page.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosome coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the [FAQ](#) for more information.

Request:

Request	Genome Browser Response:
chr7	Displays all of chromosome 7
chr7L_g000012	Displays all of the unpaired contig g000012
20q11	Displays region for band p11 on chr 20
chr7:1-1000000	Displays first million bases of chr 7, counting from p-arm telomere
chr7:1000000-2000	Displays a region of chr7 that spans 2000 bases, starting with position 1000000
8040215-8040215 15q11.1q11.1 nt104022-nt104022	Displays region between genome landmarks, such as the STS markers 8040215 and 8040215, or chromosome bands 15q11 to 15q11, or SNPs nt104022 and nt104022. This syntax may also be used for other range queries, such as between unpaired contigs, ESTs, mRNAs, ncRNAs, etc.
D10S2046	Displays region around STS marker D10S2046 from the GeneRef/MapInfo map. Includes 100,000 bases on each side as well.
A020474	Displays region of EST with GenBank accession A020474 as BRCA1 cancer gene on chr 17
A020474	Displays region of clone with GenBank accession A020474
A020474	Displays region of mRNA with GenBank accession number A020474
F0081	Displays region of genome with F0081 Gene Name/Label/Committee identifier F0081
MM_017414	Displays the region of genome with RefSeq identifier MM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110
penicillamine-mRNA	Links to nucleotide/peptide/genes, full and cDNA
hormonase (cauda)	Links mRNAs for caudal hormone genes
zinc finger	Links many zinc finger mRNAs
trapped zinc finger	Links only trapped zinc finger
huntingtin	Links candidate genes associated with Huntington's disease
zab1	Links mRNAs discovered by covalent named Zuber
Evans_LJE	Links mRNAs discovered by co-author J.E. Evans

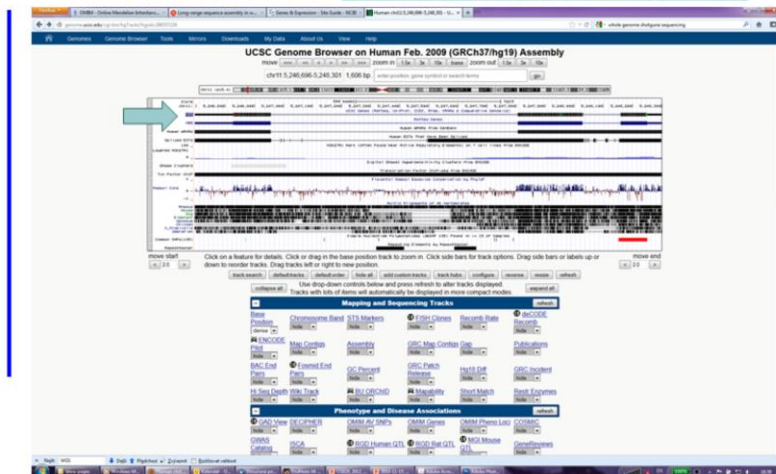


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 page includes a description of the gene, a table of sequence and links to various databases, and a detailed description of the protein. A green arrow points to the 'Sequence and Links to Tools and Databases' section.

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

Description: Homo sapiens hemoglobin, beta (HBB), mRNA. **Hereditary hemochromatosis (HFE):** 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. Mutated beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon - gamma 2 - gamma 1 - delta - beta - 3' (provided by HuRef, Jul 2005). Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. **RefSeq Attributes:** ENSEMBL: Transcription_start_coordination_evidence: V00487.1, B000491.1, J000000332, RefSeqKey: Attributes: ENSEMBL

Description: Chromosome: 11q11.1, Strand: -, Size: 1,008, Start: 5,245,695, End: 5,246,301, Exon Count: 3, Coding Size: 1,414 Start: 5,246,037, End: 5,246,211, Exon Count: 3

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

Data last updated: 2011-12-21

Sequence and Links to Tools and Databases

Genomic Sequence (chr11:5,245,695-5,246,301) mRNAs (may differ from genome): (Protein: 147 aa)

Gene Sorter	Gene Browser	Protein FASTA	MapGene	Table Schema	RefSeq
CGMAP	Consensus	Ensembl Gene	Ensembl	GeneCards	GeneHank
Depth Tissue	NCBI	NCIC	NCBI	NCBI	NCBI
OMIM	PUBMED	Reactome	Standard SOURCE	TranScrip	UniProtKB

UniProtKB

Comments and Description Text from UniProtKB

ID: HBB_HUMAN

DESCRIPTION: Hemoglobin subunit beta. A1Name: Full-beta-globin. A2Name: Full-hemoglobin beta chain. Contains: A1Name: Full-LVY-hemoglobin.7.

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

FUNCTION: LVY-hemoglobin 7 potentiates the activity of thrombin, causing a decrease in blood pressure.

SUBUNIT: Heterodimer of two alpha chains and two beta chains in adult hemoglobin A (HbA).

INTERACTION: Interacts with HFE2, HFE3, HFE4, HFE5, HFE6, HFE7, HFE8, HFE9, HFE10, HFE11, HFE12, HFE13, HFE14, HFE15, HFE16, HFE17, HFE18, HFE19, HFE20, HFE21, HFE22, HFE23, HFE24, HFE25, HFE26, HFE27, HFE28, HFE29, HFE30, HFE31, HFE32, HFE33, HFE34, HFE35, HFE36, HFE37, HFE38, HFE39, HFE40, HFE41, HFE42, HFE43, HFE44, HFE45, HFE46, HFE47, HFE48, HFE49, HFE50, HFE51, HFE52, HFE53, HFE54, HFE55, HFE56, HFE57, HFE58, HFE59, HFE60, HFE61, HFE62, HFE63, HFE64, HFE65, HFE66, HFE67, HFE68, HFE69, HFE70, HFE71, HFE72, HFE73, HFE74, HFE75, HFE76, HFE77, HFE78, HFE79, HFE80, HFE81, HFE82, HFE83, HFE84, HFE85, HFE86, HFE87, HFE88, HFE89, HFE90, HFE91, HFE92, HFE93, HFE94, HFE95, HFE96, HFE97, HFE98, HFE99, HFE100.

TISSUE SPECIFICITY: Red blood cells.

PTM: Glycosylated, but not enzymatically with the N-terminus of the beta chain to form a stable tetrasaccharide 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 diabetes mellitus.

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

PTM: Acetylated on Lys-60, Lys-61 and Lys-141 upon aspirin exposure. PubMed 19996667 reports the identification of HBB acetylated on Lys-141 in the cytosolic fraction of H6a cells. This may have resulted from contamination of the sample.

MASS SPECIOMETRY: Mass: 13122, Method: APL, Range: 0-42, Source: PubMed 157724.

DISEASE: Defects in HBB may be a cause of hereditary hemochromatosis (HFE) [MIM:163100]. This is a form of iron-siderophore hemochromatosis of the type 1. After splenectomy, which has little benefit, therapeutic phlebotomy is used to reduce iron stores in the erythrocytes. Before splenectomy, effluent or peritoneal hemochromatosis may be evident. Most of these cases are probably instances of hemochromatosis. The hemoglobin demonstrates high affinity. Hemoglobin is observed also with the hemochromatosis syndrome (HFE) [MIM:163100] and with glutathione S-transferase deficiency (GSTM1) [MIM:146250].

DISEASE: Defects in HBB are the cause of beta-thalassemia (BT) [MIM:604202]. 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. These descriptions apply to a wide range of clinical severity that is from clinical ineffective erythropoiesis and severe microcytic hypochromic anemia. Clinically, beta-thalassemia is divided into thalassemia major which is transfusion dependent, thalassemia intermedia of intermediate severity, and thalassemia minor that is asymptomatic.

DISEASE: Defects in HBB are the cause of sickle cell anemia (SCA) [MIM:603902], 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. Sickle cell disease cells are rigid and flexible and flow slowly through blood vessels. 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 occlusion thus cutting off the blood supply to nearby tissues.

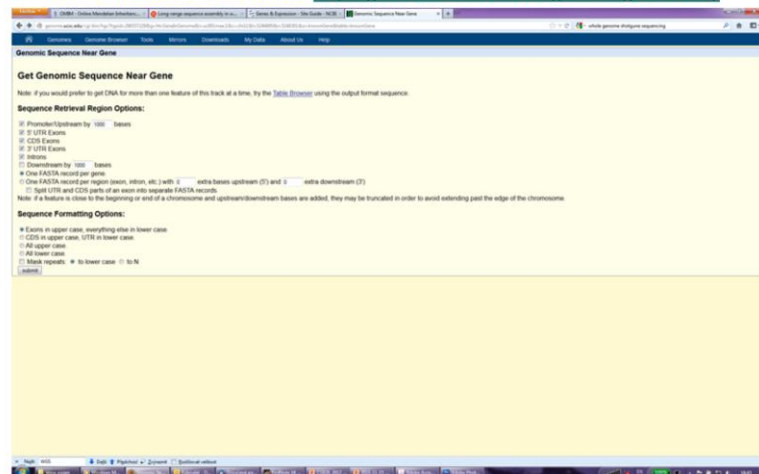


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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

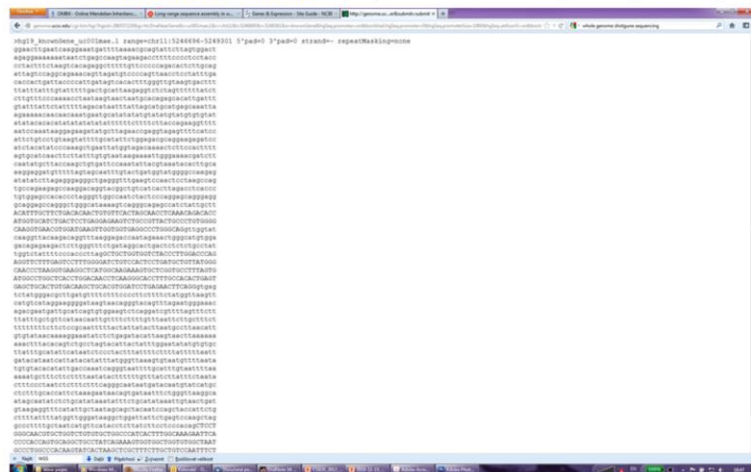


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Genomové zdroje

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

The Arabidopsis Information Resource

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

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

Breaking News

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

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

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

The NEW arabidopsis.org

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



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



EVROPSKÁ UNIE

esf



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



MASARYKOVA
UNIVERSITA
BRNO

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-----AKSETQRYMVIE  
SLAVTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE
```

Lokální přiřazení

```
SLAVTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE  
-----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

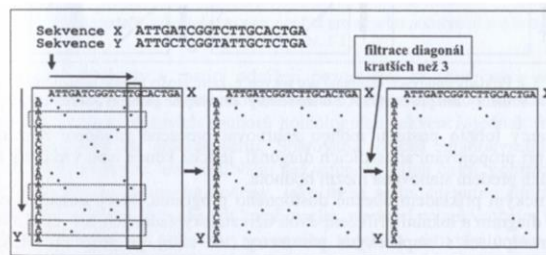


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)



Čurčková, 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)

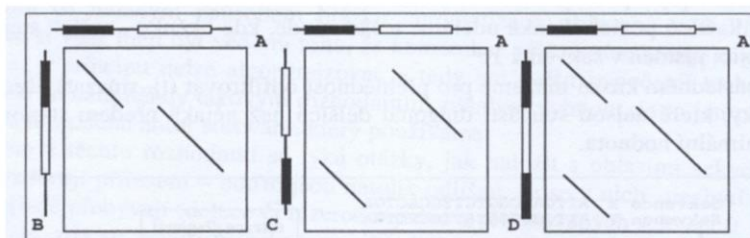


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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 RFD

[Search](#)

```
aaUUCUUUgU cATTATcATc ATGtTtTtTg GcGcATGtTg tGtGtTcCA  
gcGtATtAAT  
AtAAtTtAAt tAttcAcAT gAgAtATgAT AtGAtATcAT AtGtATtTtT  
tGtTtTtTtT  
tTAttTgTAA AcTtTtAAtA tAcAAGaAc tAcAAAAAt gAAAA
```

[Set subsequence](#) From: To:

[Choose database](#)

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



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

BLAST

Basic Local Alignment Search Tool

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

AVÁNÍ



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

BLAST

Basic Local Alignment Search Tool

>gi|5016088|ref|NM_001101.2| actin, beta (ACTB), mRNA
Length = 1793

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

Query: 156 gtcgacaacggctctggcattgtgcaaggccggattgcccggagacgatgccccgcgc 215
Sbjct: 101 gtcgacaacggctccgcatgtgcaaggccggctccggggcagcagatgccccgggccc 160

Query: 216 gtcttcccatcgatttgggacgtccccgtcaccagggtgtgatggtcggcatggccag 275
Sbjct: 161 gtcttcccatcctcctggtggggcggccaggaaccaggcgtgatggtggcattgggtcag 220

Query: 276 aaggactcgtacgtgggtgatgagggcagagcaagcgtggtatcctcaccctgaagtac 335
Sbjct: 221 aaggattcctatgtggggcagcagggccagagcaagagaggcattcctcaccctgaagtac 280

Query: 336 cccattgagcaggtatcgtgaccaactgggacgatatggagaagatctggcaccacacc 395
Sbjct: 281 cccatgagcagggcctcgtcaccaactgggacgatggagaaaatctggcaccacacc 340

„expectancy value – měra pravděpodobnosti, že se stejnou nebo lepší podobností či vztahováním 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 displays a GenBank record for the gene **NP_059797.1**. The record includes the following information:

- Gene:** NP_059797.1
- Description:** 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

The **Links & Tools** popup menu provides the following links:

- GenBank View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- FASTA View: [NC_002377.1 \(145,694..148,183\)](#), [NP_059797.1](#)
- BLAST Genomic: [NC_002377.1 \(145,694..148,183\)](#)
- Graphical View: [NP_059797.1](#)
- BLAST Protein: [NP_059797.1](#)
- BLINK Results: [NP_059797.1](#)

Below the popup menu, 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: [gi16119781ref|NP_396486.1](#) two component sensor kinase [Agrobacterium tumefaciens str. C58]

Matching gi: [15163423.20141871.1019660](#)

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

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

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

[Reset all filters](#)

Choose Display Options

1203 Archaea 136295 Bacteria 13 Metazoa 1349 Fungi 554 Plants 6 Viruses 5676 The Others [reset selection](#)

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

SCORE	ACCESSION	Length	Protein Description
4146	AAK0562	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]
4146	E15540	833	RecName: Full=Wide host range virA protein; Short=WRB virA
4146	AAA79282	833	virA [Plasmid pTIC58]
4150	NP_053280	833	Hypothetical protein pT1-SM08A_p142 [Agrobacterium tumefaciens]
4150	AAA17645	833	VirA140 [Agrobacterium tumefaciens]
4153	AAA11390	833	virA [Plasmid T1]
4153	gi172727	833	virA protein
4153	CAA34772	833	VirA kinase protein [Agrobacterium tumefaciens]
3820	CAA35780	829	virA [Agrobacterium rhizogenes]
3718	gi1227240	849	virA gene
3142	AAA10442	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 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  
MSHIQIPPGLTBELLQGYTVEVLRQQPPDLVBFVAVFYFTRLREARAPASVLPAAATPRQSLGHPPPPEPGPDR  
VADAKGDSSESEDEDELEVVPVPSRFNRRVSVCAETYNPDEEEDTDPRVIHPKTDEQRCRLQEACKDILLF  
KNLDQBQLSQVLDAMFERIVKADHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSRGRLA  
LMYNTPEAATIVATSEGLWGLDRVTFRRIIVKNNAKKRKMFESFIBSVPLKSLVSERMKIVDVIGEK  
IYKDGRIITQGEKADSFYIIESEVSIILRSRTKSNKGGNQEVEIARCHKGQYFGELALVTNKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYBEQLVKMPGSSVDLGNLGG
```

```
[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 homologii
 - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restričních míst....

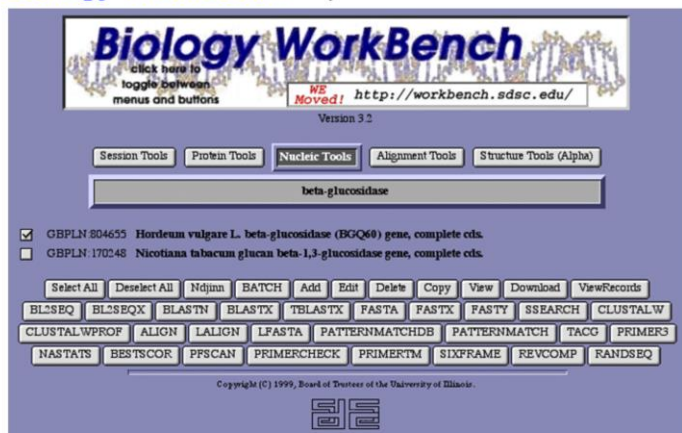


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- **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/>

The screenshot shows the 'View' interface of the Biology Workbench. At the top, there is a title bar 'View' and a subtitle '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. Below the dropdowns, there are two links: 'Download/view all sequences in text format' and '[NEXT] [BOTTOM]'. The main content area displays the following text:

Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds.
GenBank accession number: U0248, 4699 bp

>170248
GAGCTCCCTTGGGGGCAAGGGCAAACTTTTGGTAAATGGAAAATATTATACCAAGTGTGGTAAATA
GTACTCAATTTGAATTAACAAAAGGGGCAAACTTTGACATATTTGGCCCTTATCTTTTGGTCAAAAAC
ATAAAATATCCATCCGAAATCCAAAATGGTCCATTATCGGCAAGTASCTTCTTAAATATAGTDSGT
GACAAAACACTATCAAGATATCATTTATTAATAAATAACTTCAAAATCCATCTTAGCTGCTCCCTCA
GTAGAGCCGCCAGTAAATAAGACCGATCAATAAAAGCCGCCATTAATAATAGAAATTTAGGACTCTC
GATTNGCAGGTAAATGACCAAACTCTTCCAACTTTTGTGCACTTGGGGCTGCTAGGTCTTAGCTTC
CAGKMTGSGAATTTCTAGGTTATCTCTAAATTCAGATVCACTAATTAAGAAATTAAGAGTA
CAGCAAAATATAAAATTTTCTTAAAGAGACAAATGATCCGGTTACTGATTCATTGGCCCTTTCAGAG
TCGCAATGCATATCTCAAGGGGTCTTGGTCAAGAAATAATAAATTTCCGGATAGAAATTT
GAGTTGCAATTTATCTTGTGTTTAAATTAAGTATAGCTAATTTTCAATAAATTTTACTTAAATAG
TAAATCACTATCAATGAGAGGTGAAATGAAATGCTAATCCGATGSCACTTCAATGAAATGCTC
GTATTTATCTACATTTTACCAAATGATCGTTAGTCTTCAATGAAATCCAGTATCTTCAATAAATGCA
GTAAAGATTTGAAAATTTTCAATAAATCAATTCATAAATTTAAAAATATGATATGAGACCTTAAG
ATACAAATAAAGATGTCCGTTAAATAAAGATAGATAGATTTTAAATAGGAAAATAAAACGGTT
CGAGCACTCTTATGAAAGCCGTGCTTCAAGTATAGTCTTCAATCTAATCTGCTGCTCAATAGCAAAA
TGACATTTTACTCTTAAATGACAGGAGCACTTCAATGTTTCTATGTAATCTCAATGAAATTTA
GAGACTTTCAAATCTCAACTACTTTTAAAGGGAATCAAAATACGACCAATTTTATCTACTTACT
TTATAGTTAAATGATAGAAATTTTAAATTTGAAATGAAAATATTAAATCTTGAATTTAATATAA



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:
ett. {1,32}ett
0 sequences were searched
1 match was found

Matches are indicated in blue

> 170248
GAGCTCCCTTGGGGGCCAAGGCCAAAAGCTTTTGGCTAAATGAAAAATATATACCAGTGTGTTGTAATA
GTTACTAATTTGAAATTAACAAGGGCCAAATTTGACTATTTTGGCCCTTATATCTTTTGGTACAAAAC
ATAAAATATCCCAATCCGAAMTCCAAATGGTCCATTAACGGCAAGTAGCTTTCTTTAAATATAGTAGTT
GACAAACACTATCAAGATATCATATATATATATATATCTCAAGATCCATATCTTGGCTGCTCTCA
GTAGAGCCCCAGTAAATTAAGCCCAATCAATTAAGCCGCCATTAATATATATATTTTGGGCTCTC
GATTGGCACGTAAATGCCAAAACCTTCCAAATACTTTGCTGCAACTTGGGGCTCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAAATTAATCTCAACTTAATATTAAGAAATTAACAGGTA
CAGCAATCATAAATTTTCTCTTAAGAGACAAATGAAATCCGGTACTGATTCATTTGGCCTTTTAAAG
CTTCTATGCCAATTTACTTAAGGGGTCTTTTGGTACAGAAATATATATATATATTTCTGGGATGATTT
GAGATGCAATTTATCTTGTGTTAACTTAAGTATTAAGCTAATTTAGATAAAATTTTACTAAATAG
TAAATCAACTATACATGTAGAAGTGGAAATGAAATAGCTAATCCATAGCCACTCACATGAAATATCC
TTATTTATCTACTATTTTACCAATGATCGTTTGTCTTATGAAATCCAGTATCTCAATAAATGCA
GTAGAAATTTGAAATTTTCTTAAATCAATTCATTAATTTTAAATTTAGATGAGTGGGCTTAAAG
ATCAATAAAGATGATACCTTAAATAAAGATGAAATGAGTGTAAATAGGAAAAAAAACGGTT
CGAGACTCTTATGGAAGGCTTGTCTCAAGTAGATCTCAATCTATGCTTGGTCAATAGCAAAA
TGACATTTTCTCTTAAGATCAGCGAGCCCTCTCAATCTCTATTTGATATCAAAATGAAATTTTA
GAGAACTTTTAAATCTCACTACTCTTTAAGGAAATCAAAATCGCCCAATATTTTACTTCTTCTAC
TTATAGTAAATGATAGATTTTACCAAAACATGGAGATACAGAAATTTTATTTTGTATACGAT
ACAAATGATATCGTAAGTATTTACCAAAACATGGAGATACAGAAATTTTATTTTGTATACGAT
GATTAAGGACTATCTCTGTTTGGAGGATGAAAGAAAGTAACTAGCTATAATTTCTTTGTAAGT
```



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Analytické nástroje

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

```
== Linear Map of Sequence:

      StyI
      BsaJI
      CviJI
      AluI
      SacI
      EcoICRI
      BspI286I
      BsiHRAI
      BaniI BslI
      SspI
      1 gagctcccttgggggcaagggcaaaaacttttgtaaatgaaaaatattataccaagt 60
      ctgagggaacccccgttcccgtttgaaaaacgattacctttttataatggttca
      * * * * *
      1 E L P H G A R A K L F A K H 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 P 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 Tsp509I
      ApoI
      61 gtttgaatagttactcaattgaattaacaaggggcaaatgtgacttttgcctta 120
      caacattatcaatgagtttaacttaattgttcccgtttaactgataaaacgggaat
      * * * * *
      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 F V I K G K
      6 H K Y Y N S L K F * C L P L N S * K A R
```

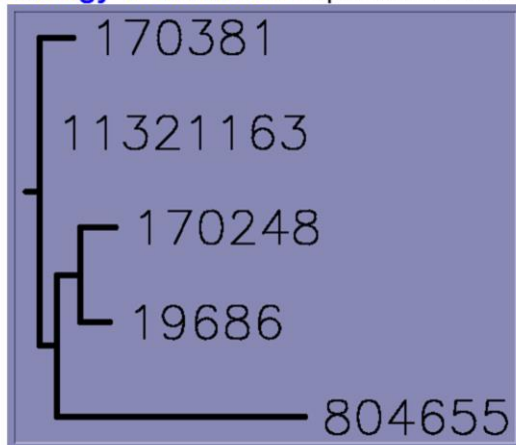


INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- o **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/vpccr2.cgi>

SEARCH ABOUT DOWNLOAD LINKS

VPCR 2.0 (WWW interface) - Please, enter nucleotide primer sequences (IUPAC codes allowed for degenerate primers). VPCR 2.0 searches the specified database for matches to the primers. If matches are found within 10000 bases, a PCR simulation model predicts amplification. Calculated PCR products are displayed within a minute.
NOTE: Abilities of VPCR 2.0 are still limited by BLAST capabilities and settings, as well as stability of our 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 Beta0001](#).

Search using: in the database for:

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

Annealing temperature:



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

Analytické nástroje

- o **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](#)



EVROPSKÁ UNIE

esf



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



MASARYKŮVA UNIVERZITA
BRNO

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 at the top, a navigation menu, and a table of contents on the right. The main content area is divided into several sections: Summary, Official full name, Primary source, Locus tag, Gene type, RefSeq status, Organism, Location, Genomic context, and Genomic regions, transcripts, and products. The Summary section provides a brief overview of the gene, including its official name and primary source. The Genomic context section shows the gene's location on chromosome 1, NC_000001.10. The Genomic regions, transcripts, and products section provides a detailed view of the gene's structure, including exons and introns. The table of contents on the right lists various sections such as Summary, Genomic context, Genomic regions, transcripts, and products, Interactions, General gene info, General protein info, Reference sequences, Repeat sequences, and Additional links.



MINISTERSTVO
MLÁDEŽI

JE VZDĚLÁVÁNÍ
je spolufinancována
kým sociálním fondem
řádem České republiky

Další WWW zdroje

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



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



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