

CG020 Genomika

Přednáška 1

Úvod do bioinformatiky

Jan Hejátko

Funkční genomika a proteomika rostlin,
Středoevropský technologický institut (CEITEC)
a

Národní centrum pro výzkum biomolekul,
Přírodovědecká fakulta,

Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.eu

MUNI
SCI



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](#)

Schéma předmětu

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

Schéma předmětu

- **Kapitola 05**
 - RNA interference a genomové editování
- **Kapitola 06**
 - Genová exprese a chemická genetika
- **Kapitola 07**
 - Protein-proteinové interakce a jejich analýza
- **Kapitola 08**
 - Současné metody sekvenování DNA

Schéma předmětu

- **Kapitola 09**
 - **Struktura genomů**

- **Kapitola 10**
 - **Evoluce genomů**

- **Kapitola 11**
 - **Genomika a systémová biologie**

- **Kapitola 12**
 - **Praktické aspekty funkční genomiky**
 - **Modelové organismy**
 - **PCR**

Literatura

- Literární zdroje pro kapitolu 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

Osnova

- Schéma předmětu
- Definice

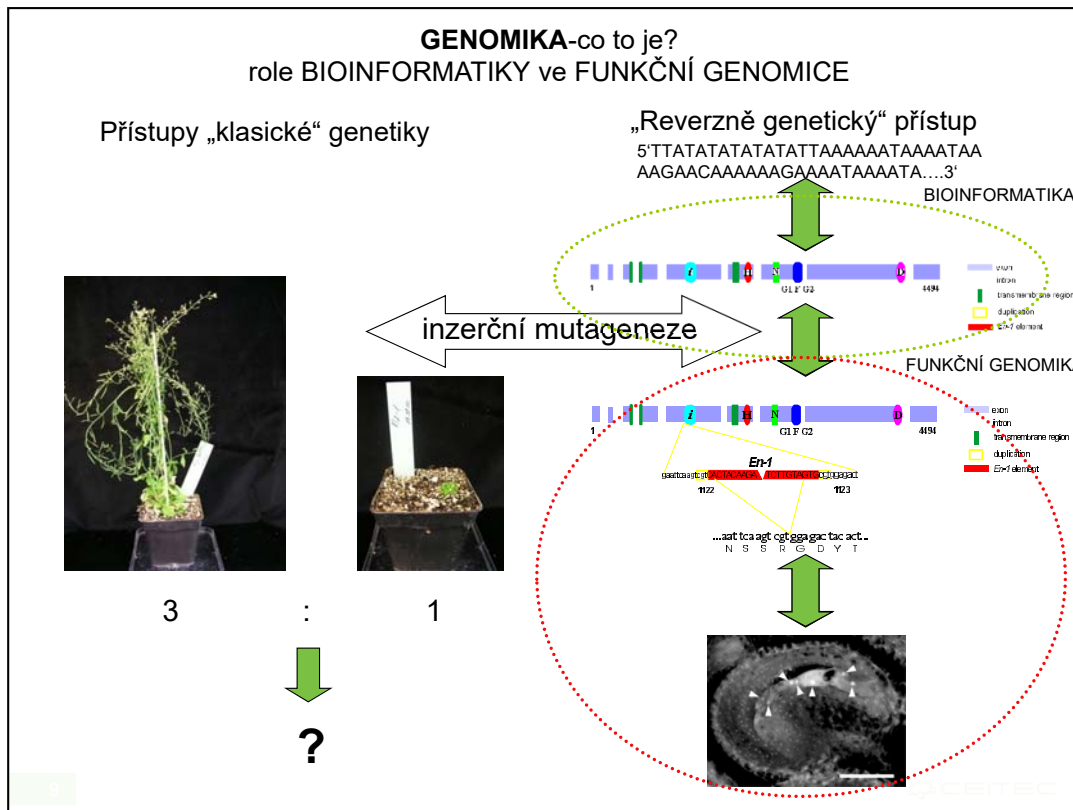
GENOMIKA-co to je?

- *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**

Genomics is a science discipline that is interested in the analysis of genomes. Genome of each organism is a complex of all genes of the respective organism. The genes could be located in cytoplasm (prokaryots) nucleus (in most eukaryotic organisms), mitochondria or chloroplasts (in plants).

The critical prerequisite of genomics is the knowledge of gene sequences.

Functional genomics is interested in function of individual genes.



With the knowledge of gene sequences (or the knowledge of the gene files in the individual organisms, i.e. the knowledge of genomes), **Reverse Genetics** appears that allows study their function.

In comparison to "classical" or **Forward Genetics**, starting with the phenotype, the reverse genetics starts with the sequence identified as a gene in the sequenced genome. The gene identification using approaches of **Bioinformatics** will be described later (see Lesson 02).

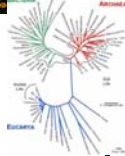
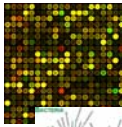
Reverse genetics uses a spectrum of approaches that will be described in the Lesson 03 that allow isolation of sequence-specific mutants and thus their phenotype analysis.

The necessity of having phenotype alterations in the forward genomics approach introduces important difference between those two approaches. Thus, the gene is no longer understood as a factor (*trait*) determining *phenotype*, but rather as a piece of DNA characterized by the unique *string of nucleotides*. i.e. **physical DNA molecule**.

Osnova

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

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.

11

CEITEC

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.

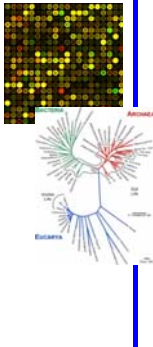
Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

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>

Bioinformatika



- **Bioinformatika** ve **funkční genomice**

- **Zpracování a analýza sekvenačních 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

Osnova

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

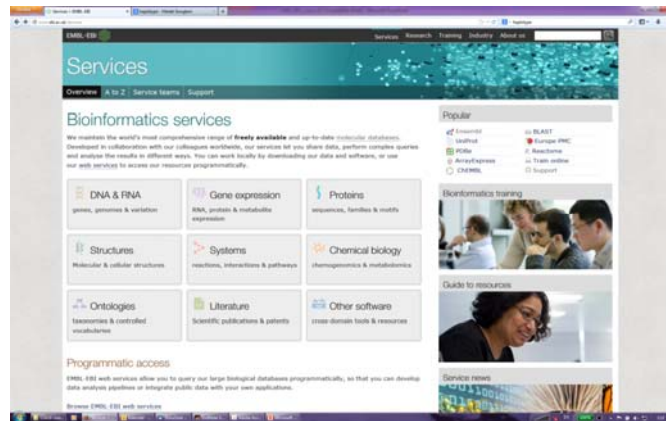
Spektrum on-line zdrojů

EMBLnet National Nodes		
Venna BioCenter	Austria	http://www.at.emblnet.org/
BIH	Belgium	http://www.be.emblnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.emblnet.org/
INCBIS/GEN	France	http://www.infobloges.fr/
GENUSnet	Germany	http://genomax.dlr-bioscience.de/bio-unit/
IMBB	Greece	http://www.imbb.forth.gr/
HEH	Hungary	http://www.hu.emblnet.org/
INEBI	Ireland	http://icet.gen.tcd.ie/
IMI	Israel	http://kippat.wellmann.ac.il/foef/imm.html
ITB-ADP	Italy	http://bio-www.bio.cnr.it/2004/BIOWWW/Bio-WWW.htm
CAOS/CAMH	Netherlands	http://www.caos.kun.nl/
Bio	Norway	http://www.no.emblnet.org/
IBB	Poland	http://www.ibb.waw.pl/
IGC	Portugal	http://www.igc.gulbenkian.pt/
GenWeb	Russia	http://www.genweb.msk.ac/
CNB-CSC	Spain	http://www.es.emblnet.org/
BMC	Sweden	http://www.emblnet.se/
SIB	Switzerland	http://www.ch.emblnet.org/
SEQNET	UK	http://www.seqnet.dl.ac.uk/
EMBLnet Specialist Nodes		
MPS	Germany	http://www.mips.biochem.mpg.de/
IGIB	Italy	http://www.igib.treche.it/
Pharmacia Upjohn	Sweden	http://www.upjohn.com/
F Hoffmann-La Roche	Switzerland	http://www.roche.com/
EBI	UK	http://www.ebi.ac.uk/
HGMP-RC	UK	http://www.hgmp.mrc.ac.uk/
Sanger	UK	http://www.sanger.ac.uk/
UMBER	UK	http://www.bioinf.mrc.ac.uk/1/bbrserver
EMBLnet Associate Nodes		
IBBM	Argentina	http://sol.biol.unlp.edu.ar/emblnet
ANIGS	Australia	http://www.angis.us.es.au/
CEB	China	http://www.cbi.pku.edu.cn/
CEB	Cuba	http://bio.cigb.edu.cu/
CPD	India	http://bala@ang.emblnet.org.in/
SABRE	South Africa	http://www.sabre.ac.za
USA Information Providers		
NCBI	USA	http://www.ncbi.nlm.nih.gov/
NLM	USA	http://www.nlm.nih.gov/
NH	USA	http://www.nih.gov/

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

Spektrum on-line zdrojů

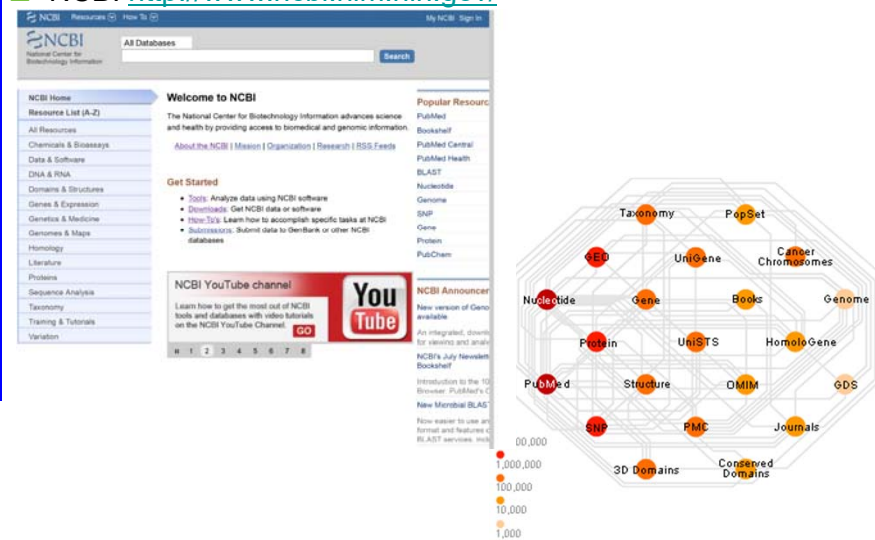
- EBI <http://www.ebi.ac.uk/services>



Nowadays, the resources are interconnected and could be accessed via dedicated web pages. Among the best and mostluy 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/>



17

CEITEC

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

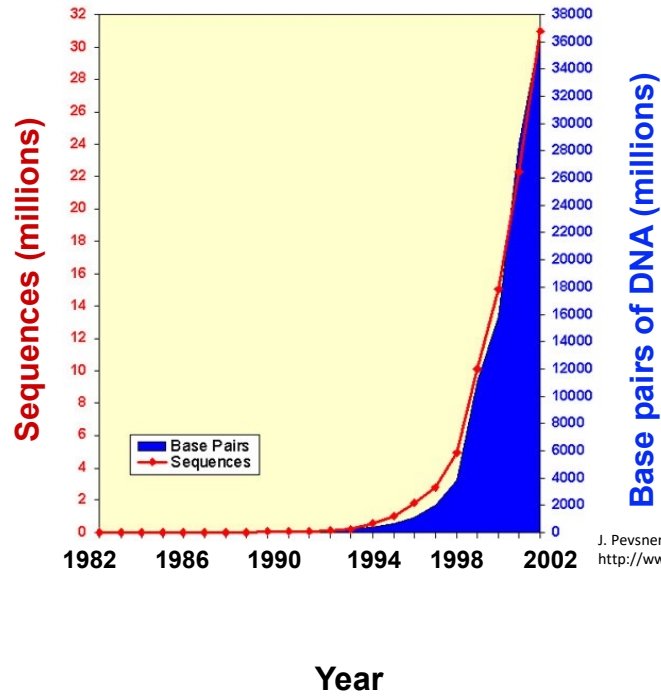
Osnova

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

Primární databáze

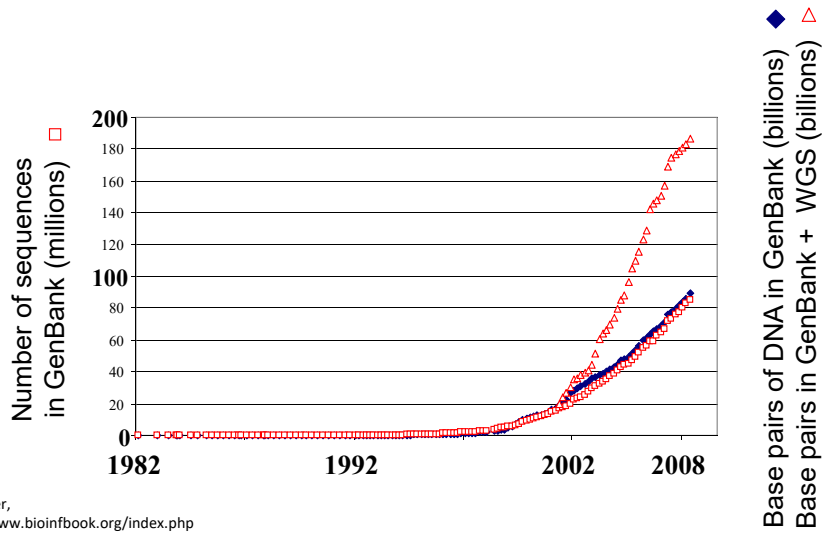
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
 - Sekvence v databázích tzv. „Velké trojky“:
 - **EMBL**
 - <http://www.ebi.ac.uk/embl/>
 - **GenBank,**
 - <https://www.ncbi.nlm.nih.gov/>
 - **DDBJ,**
 - <http://www.ddbj.nig.ac.jp>
 - denně vzájemná výměna a zálohování dat
 - velká datová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

Growth of GenBank

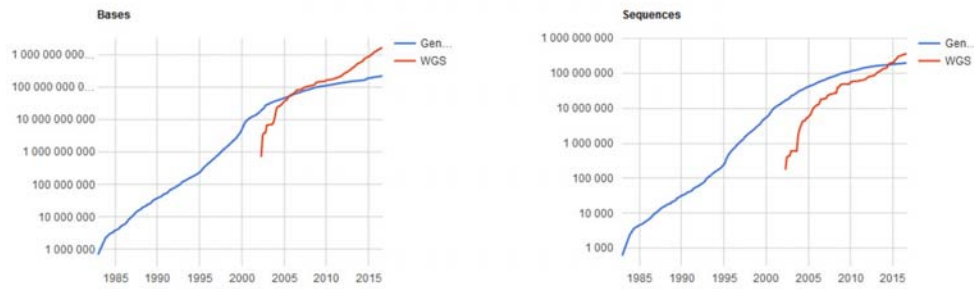


J. Pevsner,
<http://www.bioinfbook.org/index.php>

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

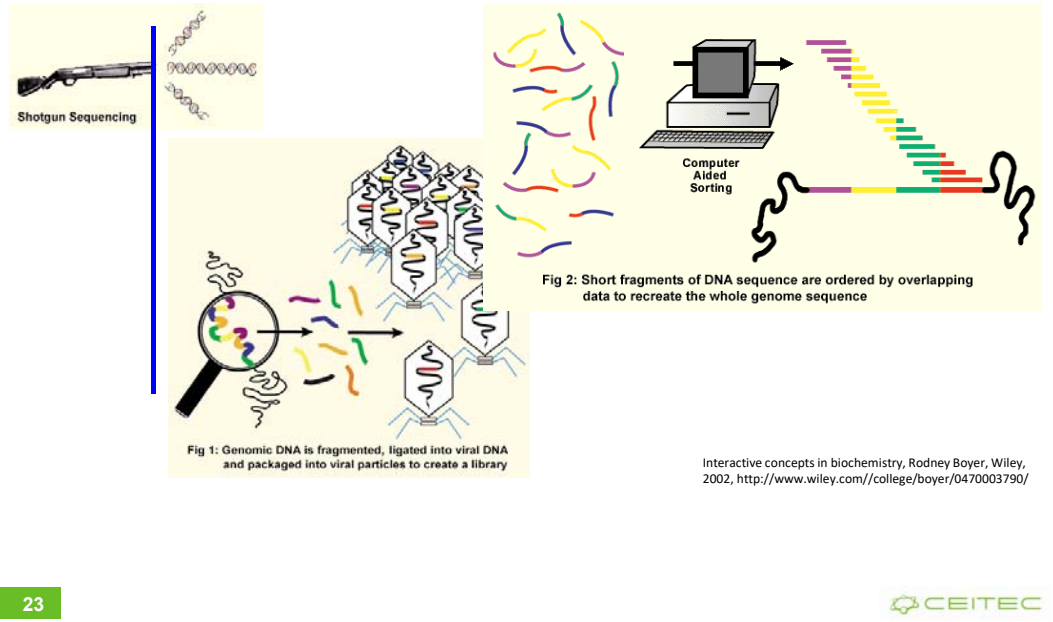


Growth of GenBank Aug 2016



- 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

WGS

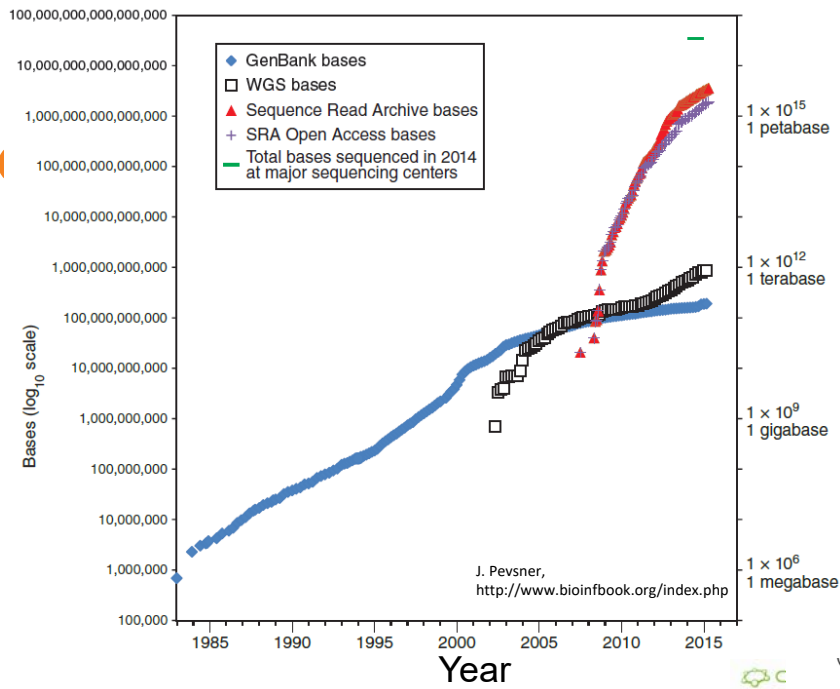


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic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

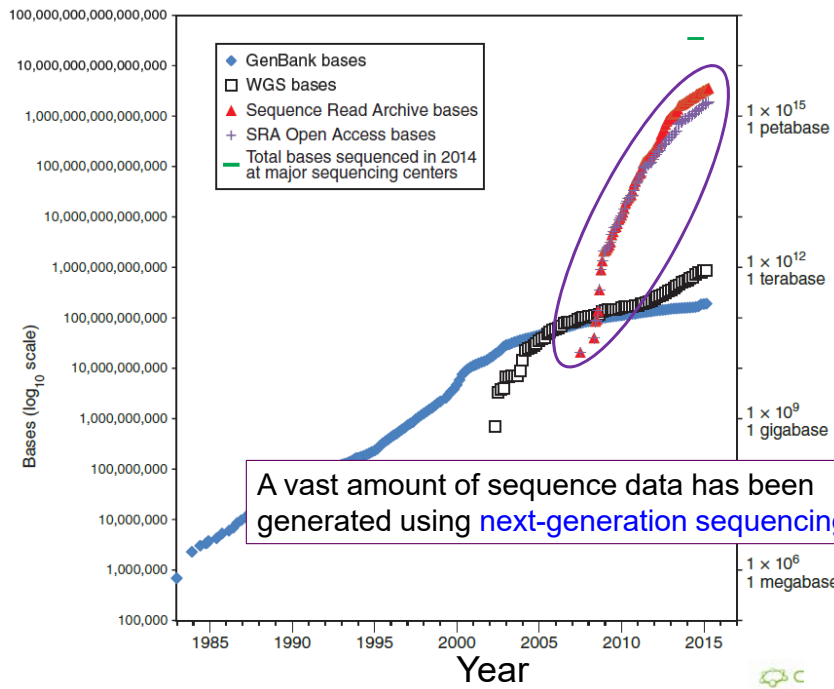
The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence.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

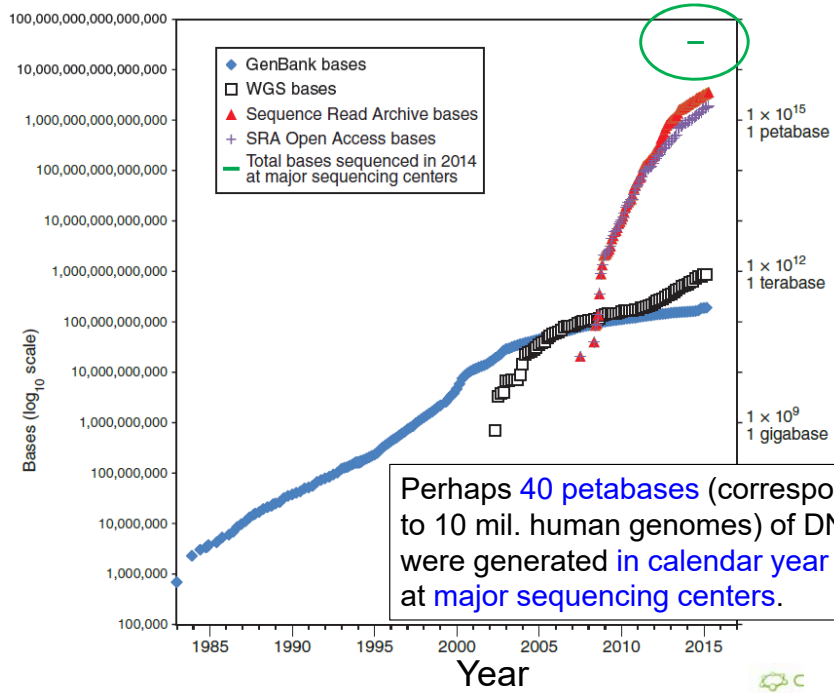
Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



B&FG 3e
Fig. 2-3
22



80

Primární databáze

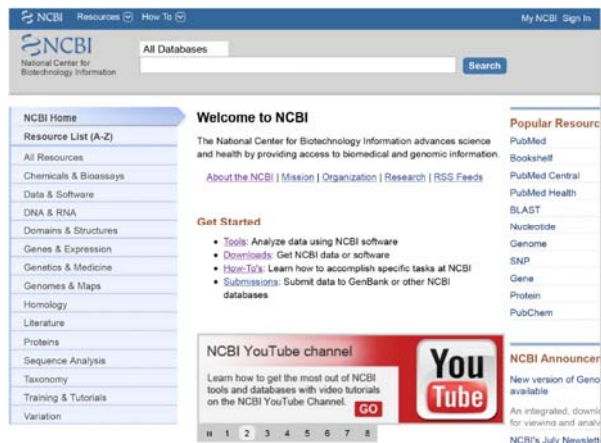
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
 - **Proteinové sekvence:**
 - **PIR**, <http://pir.georgetown.edu/>
 - **MIPS**, <http://www.mips.biochem.mpg.de>
 - **SWISS-PROT**, <http://www.expasy.org/sprot/>

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

Primární databáze

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



The screenshot shows the NCBI homepage with a navigation menu on the left, a search bar at the top, and a main content area with sections for 'Welcome to NCBI', 'Get Started', 'Popular Resources', and 'NCBI YouTube channel'.

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
- **How-To's:** Learn how to accomplish specific tasks at NCBI
- **Submissions:** Submit data to GenBank or other NCBI databases

Popular Resources
PubMed
Booksshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

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 Announcer
New version of Gen... available
An integrated, downl... for windows and mac...
NCBI's July Newsletter

Primární databáze

The screenshot displays a detailed view of a gene entry in a bioinformatics database. The main content area is divided into several sections:

- Gene symbol:** *actA*
- Gene description:** actA component of the actin filament
- Gene type:** protein coding
- RefSeq status:** PROVISIONAL
- Organism:** *Agrobacterium tumefaciens* subsp. *rubroaureum* [Strain: 3446] (NC_020771.1)
- Genomic context:** Location: plasmid 71; Sequence: NC_020771.1 (45834..46182)
- Genomic regions, transcripts, and products:** A diagram showing the gene's location on a plasmid map.
- Genomic sequence:** A sequence viewer showing the DNA sequence with a yellow circle highlighting a specific region.
- Links & Tools:** A section containing links to NCBI, EMBL, and other databases, along with tools for sequence analysis.
- Related articles:** A list of articles related to the gene, including "Sequence analysis of the actA gene of Agrobacterium tumefaciens subsp. rubroaureum 3446" and "The actA gene as a host range determinant of Agrobacterium tumefaciens".

The right sidebar contains various navigation and information links, including "General information", "About this page", "FAQ", "FTP site", "Help", "My NCBI", "NCBI Handbook", "Statistics", "Related sites", "BLAST", "Gene", "BioProject", "Genomic mapping", "GEO", "HomoloGene", "Map Viewer", "OMIM", "PDB", "RefSeq", "UniProt", "UniSTS", and "Feeds".

Primární databáze

The screenshot displays a web browser window showing a GenBank entry for NC_002377.1 (2.9Kbp). The main content area shows a genomic map with a red bar representing the gene NP_059797.1. A green arrow points to a popup menu titled 'NP_059797.1' which contains the following information:

- 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 popup menu, there are sections for 'Bibliography' and 'Related articles in PubMed'.

Primární databáze

NCBI Nucleotide

Search/Features/Map/Align/Download/Clipboard

Accession: **F03417** (Přístupový kód)

GeneBank Identifier: **U11955.1**

LOCUS F03417 1490 bp DNA linear BCT 18-DEC-2010

DEFINITION Agrobacterium tumefaciens coccolithum plasmid T1, complete sequence.

ACCESSION F03417

VERSION 1.0

KEYWORDS

SOURCE Agrobacterium tumefaciens (Strain: coccolithum)

ORGANISM Agrobacterium tumefaciens

ISOLATED FROM Agrobacterium tumefaciens

DEPOSITED ON Agrobacterium tumefaciens

RESEARCHER Schramm, R., Hoyer, P.J., Farrant, E.E. and

LABORATORY

TITLE Coccolithum T1 plasmid sequence

COMMENT

REFERENCE 1. Schramm, R., Hoyer, P.J., Farrant, E.E. and Hoyer, P.J. 2010. Microbiology, Cornell University, Ithaca, NY, USA. DOI: 10.1101/000000

FEATURES

Location/Qualifiers

1..1490

source

1..1490

/organism="Agrobacterium tumefaciens"

/mol_type="genomic DNA"

/db_xref="taxon:12101"

/plasmid="T1"

/subset="coccolithum"

organism_type

1..1490

/gene="tisk"

/db_xref="GeneID:122435"

1..1490

/gene="tisk"

/subset="two-component regulator of vir regulon; tisk is a transmembrane histidine kinase"

/rdbx_start=1

/rdbx_end=1490

/rdbx_subset=1

/rdbx_subset="tisk"

/gene_id="122435.1"

/db_xref="GI:122435.1"

/gb_key="122435.1"

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>

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>

RefSeq

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

- NCBI Reference Sequences (RefSeq)**: This section is highlighted with a yellow oval and contains a dropdown menu.
- Genome Annotation**: This section contains the text: "The following sections contain reference sequences that belong to a specific genome build. [Explain](#)".
- Reference assembly**: This section includes a table with the following information:

Genomic
1. NC_003065.3
Range: 10061 - 10332
Download: GenBank , FASTA , Sequence Viewer (GenBank)
- mRNA and Protein(s)**: This section lists the following entry:

mRNA and Protein(s)
1. NP_396486.1 two component sensor kinase [<i>Agrobacterium tumefaciens</i> str. C58]
UniProtKB/Swiss-Prot: P18542
Conserved Domains (3) summary
cd00075 HATPase_c: Histidine kinase-like ATPases. This family includes several ATP-binding proteins for example: histidine kinase, DNA gyrase B, topoisomerases, heat shock protein HSP90, phytochrome-like ATPases and DNA mismatch repair proteins. Location: 180 - 694 Blast Score: 202
cd00082 HAKA: Histidine Kinase A (dimerization/phosphoreceptor) domain; Histidine Kinase A dimers are formed through parallel association of 2 domains creating 4-helix bundles, usually these domains contain a conserved His residue and are activated via ... Location: 610 - 520 Blast Score: 144
PRK13837 PRK13837: two-component VIA-like sensor kinase. Provisional Location: 14 - 633 Blast Score: 2944
- Related Sequences**: This section is partially visible at the bottom of the page.

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>

Primární databáze

The screenshot displays a web browser window showing 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. The main content area shows a genomic map with a scale from 145,400 to 147,600. A red bar represents the gene **NP_059797.1**, with a green arrow pointing to its details. The details include:

- 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 details, there is a **Bibliography** section and a **Related articles in PubMed** section.

Primární databáze

The screenshot shows the NCBI GenBank web interface for the Agrobacterium tumefaciens plasmid Ti, complete sequence. The main content area displays the FASTA format sequence, which is a long string of nucleotide bases (A, T, C, G) starting with 'xpl10955014:14584-14813 Agrobacterium tumefaciens plasmid Ti, complete sequence'. To the right of the sequence, there are several interactive panels: 'Change region shown' (set to 'Selected region' from 14584 to 14813), 'Customize view', 'Analyze this sequence' (with options for Run BLAST, Pick Primers, Highlight Sequence Features, Find in this Sequence), 'Related information' (with links for Full Report, Full text in PMC, Gene, Genome, Metadata, Credits, Sequences, Protein, Protein Clusters, PubMed, PubMed (highlighted), and Taxonomy), and 'Recent activity' (showing recent searches for 'Agrobacterium tumefaciens plasmid Ti, complete sequence' and 'vIA [Agrobacterium tumefaciens]'). The browser's address bar shows the URL 'www.ncbi.nlm.nih.gov/genbank/...'. The Windows taskbar is visible at the bottom of the screenshot.

Sekundární databáze

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

Navigation: Home page | Site Map | Search EAPAs | Contact us | Swiss-Prot | PROSITE | ExPASy form

Search PROSITE [input] [button]

ScanProsite

This program allows to scan a protein sequence (either from [Swiss-Prot](#) or [TIMBL](#)), or provided by the user for the occurrence of patterns and profiles used in the [PROSITE](#) database, or to search protein databases with a user-defined pattern ([Pattern](#), [Description](#), [Accession](#)). The program [PROSITE](#) can be used to generate your own patterns. You may either:

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

Scan a protein for PROSITE matches	Search Swiss-Prot with a PROSITE entry
Enter a Swiss-Prot/TIMBL accession number (AC) (for example P00510) or a sequence identifier (ID) (for example NOFY_DROME1), or a PDB identifier, or paste your own protein sequence in the box below: [input] [button]	Enter a PROSITE accession number (for example P00225), or type your pattern in PROSITE format: [input] [button]
and specify which motifs to use: Scan <input type="checkbox"/> patterns <input type="checkbox"/> profiles <input type="checkbox"/> rules [User Manual] (You may also specify a PROSITE entry in the box to the right) <input type="checkbox"/> Exclude entries with a high probability of occurrence	and specify your search limits: <ul style="list-style-type: none">• The <input type="checkbox"/> Swiss-Prot <input type="checkbox"/> TIMBL <input type="checkbox"/> TIMBLacc <input type="checkbox"/> PDB databases (You may also specify a protein in the box to the left) <input type="checkbox"/> including other variants• The following lists [input]• Use SCN2 database to scan multiple times with a connection, e.g. filter against <i>Streptococcus</i>. Not available in PROSITE• Sequence with at least [input] hits• At most 1000 [input] matches
Your e-mail (optional): [input] (optional matches to e-mail) <input type="checkbox"/> plain text output [button] [button]	Advanced options: <input type="checkbox"/> FASTA output <input type="checkbox"/> retrieve complete sequences allow at most [input] X sequence characters to match a conserved position in the pattern match mode [input] priority, settings, no restriction [input] (in pattern, use help) pattern database [input] (to list a pattern, use help)

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].
      371 - 585  sbsnsatYrcsksaa
      371 - 585  sbsnsatYrcsksaa

>PDOC00004 PS00004 CAMP_PHOSPHO_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
      744 - 743  ssv*
      814 - 817  KDaG

>PDOC00005 PS00005 PKC_PHOSPHO_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
      148 - 150  ssk
      144 - 146  Tsk
      172 - 173  ssk
      219 - 221  ssk
      369 - 370  Tsk
      400 - 402  ssk
      513 - 515  Tsk
      585 - 587  ssk
      602 - 604  Tsk
      612 - 614  ssk
      716 - 718  ssk
      758 - 759  Tsk
      747 - 749  Tsk
      790 - 790  ssk
      804 - 806  ssk
      864 - 866  ssk
      868 - 870  ssk
      921 - 923  ssk
      957 - 959  ssk
      960 - 962  Tsk
      974 - 976  Tsk
      997 - 999  ssk
      1050 - 1054  Tsk
      1018 - 1020  Tsk
      1031 - 1033  Tsk
      1118 - 1120  ssk
```


Sekundární databáze

- Databáze funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí)
- PRINTS, <http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>



PRINTS is a compilation of protein fingerprints. A fingerprint is a group of conserved motifs characteristic of a protein family; its diagnostic power is defined by iterative scanning of a PROSITE/PROSITE-LIKE sequence. Usually the motifs do not overlap, but are separated 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. [Background](#)

New:

- [PRINTS](#) - Search PRINTS or related PRINTS
- [PRINTS](#) - Search PRINTS automatic registration
- [Prints](#) - Search the integrated InterPro family database

Direct PRINTS access:

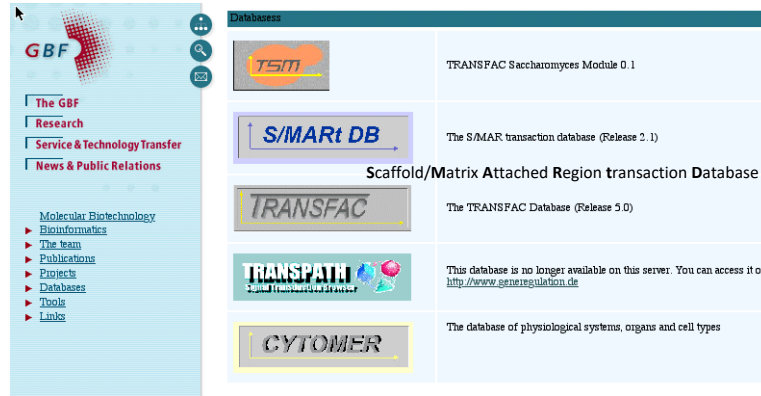
- [By accession number](#)
- [By PRINTS code](#)
- [By domain code](#)
- [By motif](#)
- [By sequence](#)
- [By motif of motifs](#)
- [By motif](#)
- [By motif domains](#)

PRINTS search:

- Search PRINTS with [NEW EprintPRINTScom](#)
- [EPrints](#)
- [EPRINTScom](#)
- [EPrints](#)
- [EprintPRINTScom](#) literature and access are available: contact@ceitec.man.ac.uk

Sekundární databáze

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



The screenshot shows the website interface for TRANSFAC. On the left is a navigation menu with the GBF logo and links for '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)
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

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/>

The screenshot shows the PDB website homepage. At the top, there is a navigation bar with links for 'DEPOSIT data', 'DOWNLOAD files', 'Browse LINKS', 'BETA TEST new features', and 'BETA mmCIF files'. Below this, there are sections for 'Current Holdings' (19623 Structures, Last Update: 30 Dec 2002, PDB Statistics) and 'Molecule of the Month: Cytochrome c'. The main content area features a search bar for 'Search the Archive' with a 'Find a structure' button and options for 'Query by PDB id only' and 'match exact word'. There is also a 'News' section with a date of '23-Dec-2002' and a 'Happy Holidays from the PDB!' message. On the right side, there is a 'PDB Mirrors' section listing various international mirrors.


Strukturální databáze

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

Structure Explorer - 1PSY

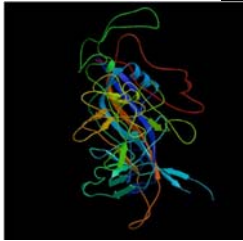
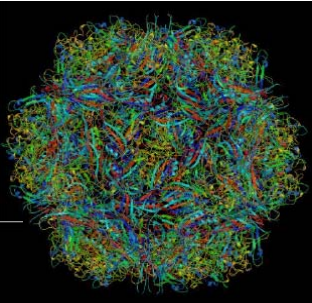
 **Structure Explorer - 1PSY**

Title: The Structure Of Hot Range Controlling Region Of The Capsid Of Canine and Feline Parvoviruses and Mutants
Classification: Virus/Viral Protein
Compound: 104; 10; 10; Molecular: Coat Protein Vp2; Chain: A; Fragment: Sequence Database Residues 180-373; Engineered: Yes; Mutation: Yes
Exp. Method: X-ray Diffraction

 **View Structure**

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

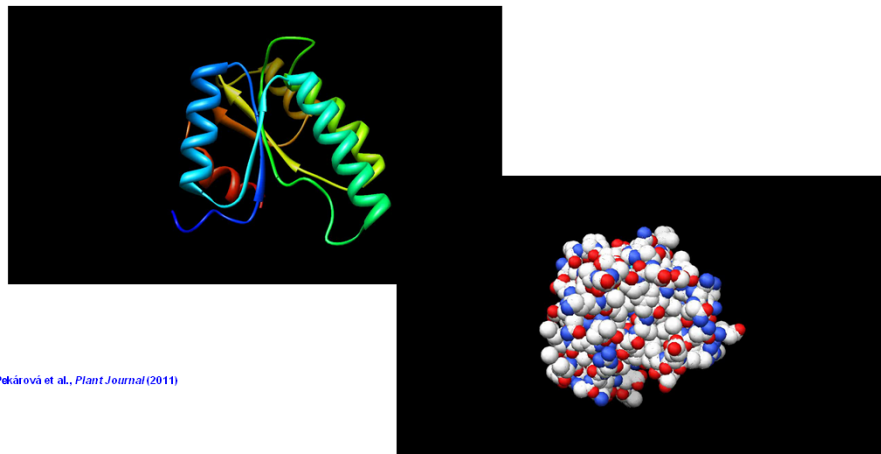
[Search](#), [Search/Info](#)



<http://www.rcsb.org/pdb/cgi/explorer.cgi?pdb=1psy&nr=173561064340344&bio=1&opt=show&size=500> 12/29/2003

Strukturální databáze

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



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

Osnova

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

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, chr17:7867000-7886000, rs334, DNA repair

Assembly
GRCh38.p13

Assembly details

Name GRCh38.p13
RefSeq accession GCF_000001405.20
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 2020-09-17

1	CHROM1
2	CHROM2
3	CHROM3
4	CHROM4
5	CHROM5
6	CHROM6
7	CHROM7
8	CHROM8
9	CHROM9
10	CHROM10
11	CHROM11
12	CHROM12
13	CHROM13
14	CHROM14
15	CHROM15
16	CHROM16
17	CHROM17
18	CHROM18
19	CHROM19
20	CHROM20
21	CHROM21
22	CHROM22
23	CHROM23
24	CHROM24
25	CHROM25
26	CHROM26
27	CHROM27
28	CHROM28
29	CHROM29
30	CHROM30
31	CHROM31
32	CHROM32
33	CHROM33
34	CHROM34
35	CHROM35
36	CHROM36
37	CHROM37
38	CHROM38
39	CHROM39
40	CHROM40
41	CHROM41
42	CHROM42
43	CHROM43
44	CHROM44
45	CHROM45
46	CHROM46
47	CHROM47
48	CHROM48
49	CHROM49
50	CHROM50
51	CHROM51
52	CHROM52
53	CHROM53
54	CHROM54
55	CHROM55
56	CHROM56
57	CHROM57
58	CHROM58
59	CHROM59
60	CHROM60
61	CHROM61
62	CHROM62
63	CHROM63
64	CHROM64
65	CHROM65
66	CHROM66
67	CHROM67
68	CHROM68
69	CHROM69
70	CHROM70
71	CHROM71
72	CHROM72
73	CHROM73
74	CHROM74
75	CHROM75
76	CHROM76
77	CHROM77
78	CHROM78
79	CHROM79
80	CHROM80
81	CHROM81
82	CHROM82
83	CHROM83
84	CHROM84
85	CHROM85
86	CHROM86
87	CHROM87
88	CHROM88
89	CHROM89
90	CHROM90
91	CHROM91
92	CHROM92
93	CHROM93
94	CHROM94
95	CHROM95
96	CHROM96
97	CHROM97
98	CHROM98
99	CHROM99
100	CHROM100

Genomové zdroje

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

The screenshot displays the UCSC Genome Browser Gateway interface. At the top, there are navigation tabs for 'Genome Browser', 'Tools', 'Network', 'Downloads', 'My Data', and 'About Us'. Below the navigation is a search bar with fields for 'class', 'genome', 'assembly', and 'position'. The 'class' dropdown is set to 'genome', 'genome' is set to 'Human', and 'assembly' is set to 'Feb. 2009 (GRCh37/hg19)'. 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 (accessions)'. It includes a paragraph about the February 2009 human reference sequence (GRCh37) and a section for 'Sample position queries'. This section explains that a genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. It then lists several example queries with their corresponding descriptions:

Request:	Genome Browser Response:
19q7	Displays all of chromosome 7
U937_p000212	Displays all of the unpaired contig p000212
25q17	Displays region for band 17q11 on chr 20
hg13 1 1000000	Displays first million bases of chr 1, counting from p-arm telomere
hg13 1 1000000-2000	Displays a region of hg13 that spans 2000 bases, starting with position 1000000
1000000-10400175 15q11 15q13	Displays region between genome landmarks, such as the STS markers D15K091 and D15K071, or chromosome bands 15q11 to 15q11, or SBPs rs154252 and rs180370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, RefSeqs, etc.
rs154252-rs180370	
D15K091	Displays region around STS marker D15K091 from the GenBank/RefSeq/Ensembl track. Includes 100,000 bases on each side as well.
AA20474	Displays region of EST with GenBank accession AA20474 in BRCA1 cancer gene on chr 17
AC005051	Displays region of clone with GenBank accession AC005051
AF003811	Displays region of mRNA with GenBank accession number AF003811
FBP9	Displays region of genome with HUGO Gene Nomenclature Committee identifier FBP9
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
phenotype mRNA	Lets filtered phenotypes, list and cDNA
transcript model	Lets mRNA for causal transcription genes
zinc finger	Lets many zinc finger mRNAs
knockout zinc finger	Lets only knockout zinc finger genes
huntington	Lets candidate genes associated with Huntington's disease
zinc	Lets mRNAs dependent by keyword named zinc
Ewins J.E.	Lets mRNAs dependent by co-author J.E. Ewins

Genomové zdroje

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

The screenshot displays the UCSC Genome Browser interface for the Human Feb. 2009 (GRCh37/hg19) Assembly. The main view shows a genomic region with various tracks including the RefSeq gene model, RepeatMasker, and a signal track. A green arrow points to the RefSeq gene track. Below the tracks is a list of available tracks categorized into 'Mapping and Sequencing Tracks' and 'Phenotype and Disease Associations'. The 'Mapping and Sequencing Tracks' list includes tracks like ChIP-chip, CNV, and GC content. The 'Phenotype and Disease Associations' list includes tracks like ClinVar, OMIM, and HUGO Gene Nomenclature Committee.

Genomové zdroje

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

Human Gene HBB (uc003fma.1) Description and Page Index

Description: Homo sapiens hemoglobin, beta (HBB), mRNA.

RefSeq Summary (NM_000518): The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta zero thalassemia. Reduced amounts of detectable beta globin causes beta plus thalassemia. The order of the genes in the beta globin cluster is 5' alpha 1 - gamma 1 - gamma 2 - beta 1 - 3' (located by DeFries, AJ (2002) Publication Note: The beta globin record includes a subset of the publications that are available for this gene. Please use the Gene record to access additional publications. RefSeqEntryAttributes:5184736).

Transcript, exon, combination, junction - VSIGAF 1 (B) (CC) 3000332 RefSeqSeq Attributes (NCBI)

Transcript: Chromosome: 11 (p15.5) - Span: 1,938,587 - 1,942,501 - Exon Count: 3

Coding: Size: 1,424 Start: 1,248,627 - End: 1,248,211 Exon Count: 3

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

Sequence and Links to Tools and Databases

Genomic Sequence (NM_000518.5,246,201) mRNA (may differ from genome) Protein (147 aa)

Gene Sorter	Genome Browser	Protein FASTA	VistaGene	Table Schema	BioGPS
CSAP	Ensembl	Ensembl Gene	Ensembl Trans	GeneCards	GeneNetwork
CGAP Tissue (1, 8M)	NCBI	NCBI	NCBI	NCBI	NCBI
CCMdb	PubMed	PubMed	PubMed	PubMed	PubMed
UniProt	UniProt	UniProt	UniProt	UniProt	UniProt

Comments and Description Text from UniProtKB

ID: HBB_HUMAN

DESCRIPTION: UniProt Name: Full-Hemoglobin subunit beta; UniProt Name: Full-Hemoglobin subunit beta; UniProt Name: Full-Hemoglobin subunit beta; UniProt Name: Full-Hemoglobin subunit beta; UniProt Name: Full-Hemoglobin subunit beta.

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

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

INTERACTION: Interacts with HBA2; HBA2 (HbA2) is a non-symplectic hemoglobin variant of HbA1. After separation, which has little benefit, samples sometimes called fetal hemoglobin are demonstrable in the erythrocytes before apoptosis, diffuse or parvane hemoglobin may be evident. Most of these cases are probably mixtures of hemoglobinopathy. The hemoglobin domains with high affinity HbA1 are observed also with the fetal hemoglobin syndrome with autosomal recessive and with quantitative parvane hemoglobin.

PTM: S-nitrosylated, a nitric oxide group is first bound to Fe(2+) and then transferred to Cys 94 to allow capture of CO.

PTM: S-nitrosylated, a nitric oxide group is first bound to Fe(2+) and then transferred to Cys 94 to allow capture of CO.

MASS SPECTROMETRY: Mass=1110; Molecule=AB; Range=23-42; Source=PubMed; 1157274

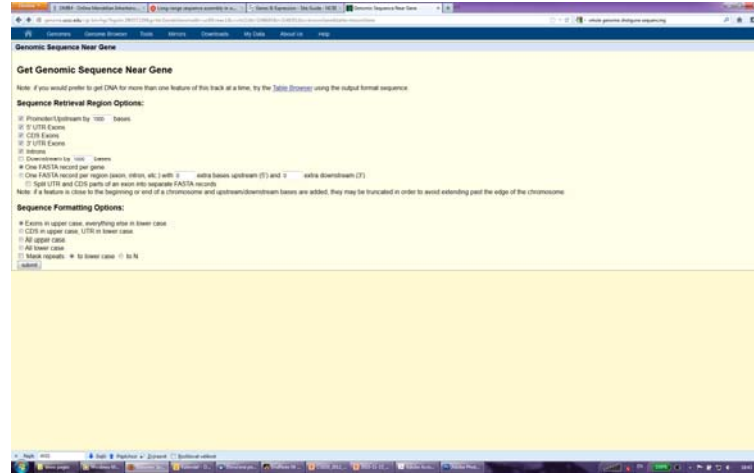
DISEASE: Defects in HBB may be a cause of hemolytic anemia (HBA) (OMIM 261300). This is a form of non-symplectic hemoglobinopathy. After separation, which has little benefit, samples sometimes called fetal hemoglobin are demonstrable in the erythrocytes before apoptosis, diffuse or parvane hemoglobin may be evident. Most of these cases are probably mixtures of hemoglobinopathy. The hemoglobin domains with high affinity HbA1 are observed also with the fetal hemoglobin syndrome with autosomal recessive and with quantitative parvane hemoglobin.

DISEASE: Defects in HBB are the cause of beta-thalassemia (BTHAL) (OMIM 260501). A form of beta-thalassemia. Thalassemias are common monogenic diseases occurring mostly in Mediterranean and Southeast Asian populations. The hallmark of beta-thalassemia is an increase in globin-chain production in the adult HbA molecule. Absence of beta chains causes beta-zero thalassemia, with reduced amounts of detectable beta globin causing beta-plus thalassemia. In the severe forms of beta-thalassemia, the excess alpha globin chains accumulate in the developing erythroid precursors in the marrow. Their deposition leads to a vast increase in erythroid apoptosis that in turn causes ineffective erythropoiesis and severe osteopenia, hepatosplenomegaly, anemia. Chronic beta-thalassemia is usually not transfusion-free which is transfusion-dependent, transfusion-independent (of intermediate severity), and beta-thalassemia intermedia (of intermediate severity).

DISEASE: Defects in HBB are the cause of sickle cell anemia (SCA) (MIM 603903), also known as sickle cell disease. Sickle cell anemia is characterized by abnormally shaped red cells resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resemble a sickle. These other red blood cells can lead to microvascular occlusion that cutting off the blood supply to nearby tissues.

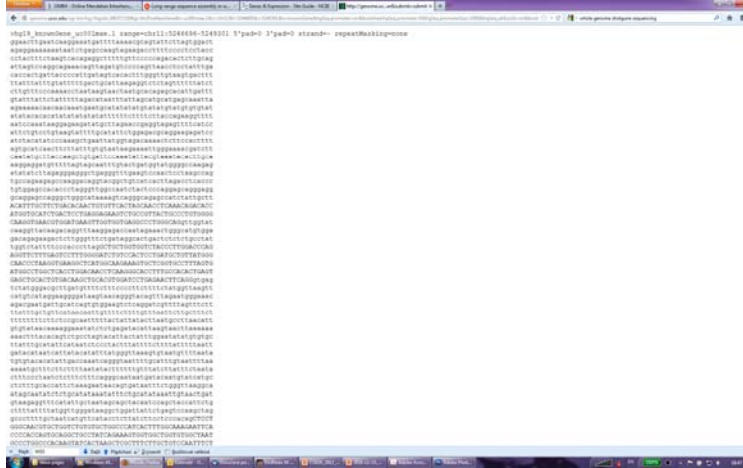
Genomové zdroje

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



Genomové zdroje

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



Genomové zdroje

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



Genomové zdroje

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

The screenshot shows the TAIR website interface. At the top, there is a navigation bar with links: Home, Help, Contact, About Us, Login, and AHP2 (circled in red). Below this is a search bar and a menu with options: Search, Browse, Tools, Stocks, Portals, Download, Submit, and News. The main content area is divided into several sections: 'The Arabidopsis Information Resource' with a detailed description of the database; 'Breaking News' with a notice about data updates; 'New Phenotype Search Option' with information about improved search capabilities; and 'ASPB Presentations' with a notice about workshop presentations. A smaller version of the website is shown at the bottom of the page.

AHP2@TAIR

Osnova

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

Analytické nástroje

□ Globální vs. lokální přiřazení

```
Globální přiřazení
SLAV-----APATNIK-----PIQNYR-I-----AKSETQRYMVE
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVE

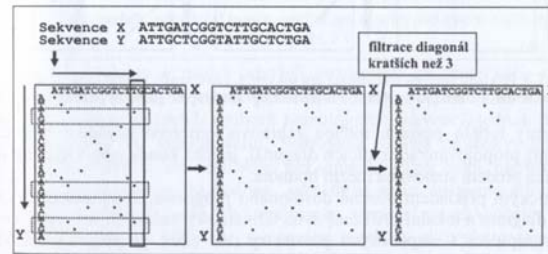
Lokální přiřazení
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVE
-----NAPATNIKSECVRA-PIQNYRRVEHVRA-----
```

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

Analytické nástroje

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)

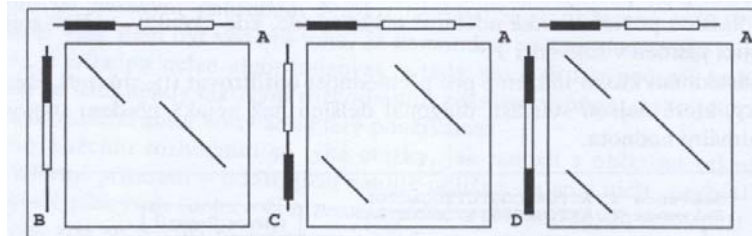


Cvrčková, Úvod do praktické bioinformatiky

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

Analytické nástroje

- příklady srovnání sekvencí pomocí bodového diagramu

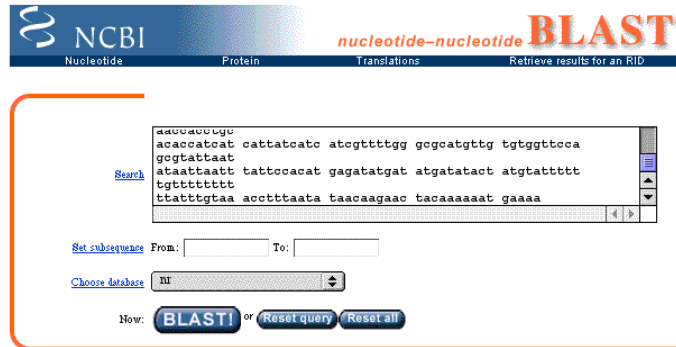


Cvrčková, Úvod do praktické bioinformatiky

- **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)

Analytické nástroje

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



The screenshot displays the NCBI BLAST web interface. At the top, the NCBI logo is on the left, and the text "nucleotide-nucleotide BLAST" is on the right. Below this, there are tabs for "Nucleotide", "Protein", and "Translations", with "Nucleotide" selected. A sub-header reads "Retrieve results for an RID". The main input area contains a text box with the following nucleotide sequence: `aacacccggc cabbatcacc atcgctttgg ggcgatggtg tgtggtcca
gggtattaat
ataattaatt tattccacat gagatgatg atgatatact atgtattttt
tgcttttttt
ttatttgtaa acotbtaata taacaagaac tacaaaaaat gaaaa`. Below the text box are fields for "Set subsequence" (From: and To:), a "Choose database" dropdown menu set to "nr", and buttons for "BLAST!", "Reset query", and "Reset all".

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

Diagram showing a 4x4 substitution matrix for nucleotides A, T, G, C. The diagonal elements are 1, representing matches. The off-diagonal elements are 0, representing mismatches. Two callouts point to the G-A and G-G cells, with labels 'hodnota nepáru G-A' and 'hodnota páru G-G' respectively.

Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250

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

BLAST

Basic Local Alignment Search Tool

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

E= expectancy value

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

Query: 156 gtcgacaacggctctgcatgtgcaaggccggatttgcggagacgatgctccccggccc 215
Sbjct: 101 gtcgacaacggctctgcatgtgcaaggccggcttcgctgggcaacgatgccccccggccc 160

Query: 216 gctctcccatcgatttgggaagtcctccgtaaccagggtgtgatggctggcatggccag 275
Sbjct: 161 gctctcccatcctatctgtggggcggcccaaggcaaccagggtgtgatggctggcatgggtcag 220

Query: 276 aaggactcgtaacgtgggtgatgagggcagagcaagcgtggtatcctcacctgaagtac 335
Sbjct: 221 aaggatcctctatgtgggcaagggccagagcaagagagggatcctcacctgaagtac 280

Query: 336 cccattgagcaaggatctgtgaccaactgggaagatggagaagatctggcaccacacc 395
Sbjct: 281 cccatgagcaaggcatctgtcaccactgggaagacatggagaaaatctggcaccacacc 340

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

Primární databáze

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

65

CEITEC

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: [gi|16119781|ref|NP_396486.1](#) two component sensor kinase [Agrobacterium tumefaciens str. C58]
Matching gis: [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

1263 Archaea 13825 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
3166	AA093322	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]
3164	F15345	833	Dectone1: Full-Wide host range virA protein/ Short-WDR virA
3164	AA073232	833	virA [Plasmid pTIC81]
3159	WP_013380	833	Hypothetical protein pT1-SAMBA_g142 [Agrobacterium tumefaciens]
3158	AA017065	833	virA140 [Agrobacterium tumefaciens]
3153	AA015590	833	virA [Plasmid T1]
3153	gi 1731317	833	virA protein
3153	CA043773	833	virA kinase protein [Agrobacterium tumefaciens]
3100	CA033760	829	virA [Agrobacterium rhizogenes]
3118	gi 1227245	849	virA gene
3149	AA016843	829	virA [Plasmid T1]

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 nukleotidové sekvence** přeložené do sekvence **aa** v **proteinové databázi**

BLAST

Specializované verze

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

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

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 (neznámá nic)
 - x(5) znamená 5 jakýchkoliv aminokyselin
 - x(3, 5) znamená 3 až 5 jakýchkoliv aminokyselin

BLAST

Specializované verze

□ Příklad vyhledávání pomocí PHI-BLAST

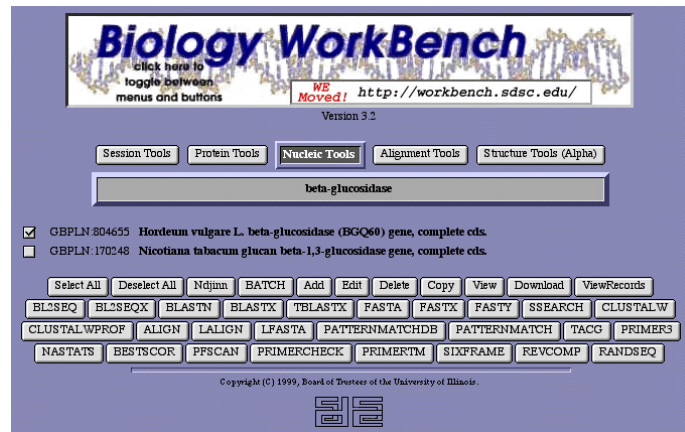
```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase  
MSHIQIPPGLELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPEPGPDR  
VADAKGDSSEEDLEVPVPSRFNRRVSVCAATYNPDEBEEEDTPRVIHPKTDEQRCRLQEACKDILLF  
KNLDQEQLSQVLDAMFERIVKADEHVIDQDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGRLA  
LMYNTPRAATIVA TSEGSLWGLDRVTFRRI IVKNNAKRRKMFESFIESVPLLKSLEVSRMKIVDVIGEK  
IYKDGERRITQGEKADSFYIIESGSEVSLIRSRTKSNKDGNGQVEIARCHKQYFGBLALVTNKPRAAS  
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNI SHYBEQLVKMFGSSVDLGNLQ  
  
[LIVMF] -G-E-x- [GAS] - [LIVM] -x(5,11) -R- [STAQ] -A-x- [LIVMA] -x- [STACV] .
```

Osnova

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

Analytické nástroje

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



Analytické nástroje

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



The screenshot shows the 'View' interface of 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. 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.  
GBFLN:170248, 4699 bp  
  
>170248  
GAGCTCCCTTGGGGGGCAAGGGCAAAACCTTTTGGCTAAATGGAAAAATTTATACCANGTGTGTAATA  
GTTACTCAATTTTGAATACAAAGGGCAAACTTTGACTATTTGGCCCTTATCTTTTGGTCAAAAAAC  
ATAAAATATCCCATCCGAAATCCAAATGGTCCAAATATCGGCAAGTAGCTTTCCTTAATATAGTTAGTT  
GACAAAACCTATCAAGATATCATTAATTAATTAATAACTTCAAAATGTCATCATCTTAGCTGCCCTCA  
GTAGAGCCCGAGTAAATTAAGCCGATCAAAATAAAGCCCGCAATTAATAATGAATTTTAGGACTCTC  
GATTGGCACGTAAAGTCCAAAACCTTCCAAATCTTTGCTCAACTTTGGGGGCTCTAGGTTCTAGCTTC  
CGATATGGGATTTTCTTAGGTTTATCTCCCTAATTTTACATCTCACTAATTTAAGAAATTAAGCGGTA  
CAGCAAACTATAAAATTTTCTCTAAAGAAAGACAATGAATCCGGTTACTGATCTATTTGGCTTTTCAGAG  
TCTGCATGCCATATCTACTAAGGGTCTGTTGGTACAAAGAAATAATAATAAATTTGGGATAGAATTT  
GAGATTGCATTTATCTTGTGTTTAATTAAGTATTAGCTAATTTCAAGATAAATTTTACACTAAAATAG  
TAAATCACTTCTACATTTGAGGTTGATGGAATGCTAATCCATCCATCCATCCATCCATCCATCCATCC  
TTAATTTATCTACTAATTTTCCAAATGATCGGTTAGTCTTCAATGAAATCCAGTATCTCAATAAATGCA  
GTAAGAAGTTAGAAAATTTTCAATTAATCAATCATATAAATTAATAATATAGATATGGAGCACTTAAG  
ATCAATAAAGATGTACCGTTAATAATAAAGATAGATAGAGTTTAAATAGGAAAAAAAACGGTT  
CGAGCACCTTTATGGAAAGCGGTTTCTCAAGATAGATTTCTCAATCAATTTGCTTGGTCAATAGCAAAA  
TCACTCTTACTTTAAGATACAGCGACCCACTTCAATCTTCTATTTGATCTCAATGAAAGTTTA  
GGAACCTTCAAACTCTCAACTACTTTTAAAGGAAATCAAAATACGACCAATATTTTACTTACTTAC  
TTATAGTTAAATGATATGAATTTTATTTAAATTTGAAATGAAATATTAATAATCTGATTTAATATAA
```

Analytické nástroje

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

```
Regex pattern:
ctt. {1,32}ctt
0 sequences were searched
1 match was found

Matches are indicated in blue

> 170248
GAGCTCCTTGGGGGGCAAGGGCAAAACTTTTCCTAAATGGAAAAATATATACCAAGTGTTTTGTAAAT
GTACTCAATTTGAATTAACAAGGGGCAAAATTTGACTATTTTGGCCTTATACTTTTGTGTACACAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGTTTCTTTTTAATATAGTTAGTT
GACAAAACACTATCAGATATCATTTTATATAATAACTTTCAAAGTCCATCATCTTAGCTGGCTCCCTCA
GTAGGCCCCCGTAATAATAGCCGATCAATTAAGGGCCCGCTTAATAATAGGAAATTTTGGACTCTC
GATTTGGCACGTAAGTCCAAAACTTCCAACTACTTTCCTCAACTTGGGGCTGTAGGTTCAGACTTC
CAGATATGGGATATTTTAAGTTTATCTCCTAATTTTACATCTCAACTAATATTAAGAAATTAACAGGTA
CAGCAATCATAAAATTTTCCCTTAAGAGAGCAATGAATCCGGTTACTGATTCATTTGGCCTTTTCAGAG
TCTCCATGCCATTTTCACTAGGGGTCTTTTGTCAAGAAATTAATATAATTTTCCGGATAGAAATTT
GAGATTTGCAATTTATCTTGTGTTTAAATTAAGATATAGCTAATTTCCAGATAAATTTTACATAAATAG
TAAAATCAACTATCACATGTAGAGGTGGAAATGGAAATAGCTAATCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTAATTTTACCAAATGATCGGTAGTCTTTCATGAGATCCAGTATCTCAATAAATGCA
GTAGAGTTTGGAAATTTCTTTAATCAATTTCACTTAAATTTAATAATTTTGGATTTGGGCACTTAG
ATACAATAAAGATGTACCCTTAATAATAAAGATAGATAGAGTTTAAATAGGAAAAAAAACCGTTT
CGAGACTTTTATGGAGGGGTGTCTTTCAAAGTAGATTCATTCATTTGCTCTGGTGCATATGCAAAA
TGACATTTACTCTTAGATACAGCGAGCCACTCTCAATCTTCTATGTATACIAAATGAAGTPTTA
GGAGTTTTCAAATGTTTACCTTTTTAGGGAAATCAAAATGACCAATTTTATTATTACTTTC
TTATAGTTAAATGATAGAAATTTTAAATTTGAATGAAATATTAAAATTTAGTTGATTTAATATAA
ACAATAGATATCGCTAAGTATTTTACCACAAACATGGAGATCACTACAGAAATTTTATTTTGTAAACGAT
GATTAAGCAGCTATTTATCTGGTTTGTGAGGATGAAGAAAGTAACTAGCTATAATTTCTTTTGAAGT
```

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
ELFWGARAKLFAKWKNIIPSVCHNSYSI*INRGNLTIPL

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

Frame 2, 1 stop codon
 Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. Tran
>170248 Translated - Frame 2
SSLGGQGQNFLLNGKILYQV
SSLGGQGQNFLLNGKILYQV
2 agtcacctgggggcaaggcaaaacttttgctaaatggaaaaatattataccaagt 61
F V I V T Q F E L T K G Q I * L F C P
62 tttgtaatagtactcaattgaattaacaaggggcaaatgactatttgcotta 120
```

Analytické nástroje

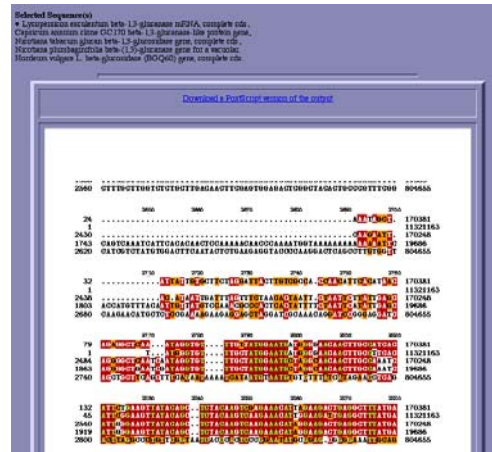
- **Biology Workbench** <http://workbench.sdsc.edu/>

```
== Linear Map of Sequence:
      StyI
      BsaJI
      CviJI
      AluI
      SacI
      EcoICRI
      Bsp1286I
      BsiHKAI
      BaniI BslI
      SspI
      \ \ \ \ \
1  gagtcctctgggggcaagggaacaaacttttggtaaatggaaaaatattataccaagt 60
ctcgagggaaacccccgttcccgtttggaaaaacgatttaccctttataataggttca
^ * * * * * ^ * * * * *
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 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 gttgtaatggttactcaatttgaattaaacaaagggcaaatgactatttgcotta 120
caaacattatcaatgagttaaacttaattgttcccgtttaactgataaacgggaat
^ * * * * * ^ * * * * *
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 R
6  H K Y Y N S L K F * C L P L N S * K A R
```

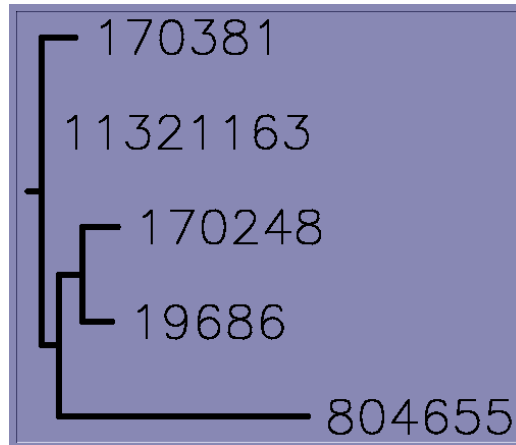
Analytické nástroje

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



Analytické nástroje

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



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 ([QUB 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 inability 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 mixes 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 Release](#).

Search using in the database for

Primer 1

Primer 2

Primer 3

Primer 4

Primer 5

Primer 6

Primer 7

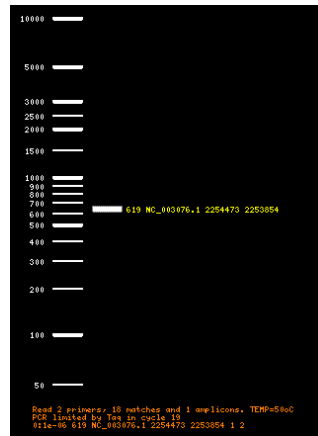
Primer 8

Assembling temperature



Analytické nástroje

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



Osnova

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

Další WWW zdroje

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

The screenshot shows the NCBI Gene database entry for PHACTR4 phosphatase and actin regulator 4 [Homo sapiens]. The page includes a search bar, a table of contents, and detailed gene information. The table of contents lists sections such as Summary, Genomic context, Genomic regions, transcripts, and products, Biography, Interactions, General gene info, General protein info, Reference sequences, Related sequences, and Additional info. The Summary section provides details on the official symbol, name, primary source, locus tag, size, gene type, and map location. The Genomic context section shows the gene's location on Chromosome 1, NC_000001.10, with a map of the region and a list of genomic regions, transcripts, and products. The Genomic region, transcripts, and products section provides a link to the RefSeq database for PHACTR4.

Další WWW zdroje

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



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

Diskuse