

CG020 Genomika

Lesson 1

Introduction into Bioinformatics

Jan Hejátko

Functional Genomics and Proteomics of Plants,
Mendel Centre for Plant Genomics and Proteomics,
CEITEC - Central European Institute of Technology
and

National Centre for Biomolecular Research,
Faculty of Science,

Masaryk University, Brno

hejatko@sci.muni.cz, www.ceitec.eu

MUNI
SCI



Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY and STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...
 - Other On-line Genome Tools

Course Syllabus

- **Chapter 01**
 - Introduction into Bioinformatics

- **Chapter 02**
 - Identification of Genes

- **Chapter 03**
 - Reverse Genetics Approaches

- **Chapter 04**
 - Forward Genetics Approaches

Course Syllabus

- **Chapter 05**
 - Functional Genomics Approaches
- **Chapter 06**
 - Protein-Protein Interactions And Their Analysis
- **Chapter 07**
 - Current Methods of DNA Sequencing
- **Chapter 08**
 - Structure of Genomes

Course Syllabus

- **Chapter 09**
 - Genome evolution

- **Chapter 10**
 - Genomics and Systems Biology

- **Chapter 11**
 - Practical Aspects Of Functional Genomics
 - Model Organisms,
 - PCR and Primer Design

Literature

- Literature resources for **Chapter 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

Outline

- Syllabus of the course
- Definition of Genomics

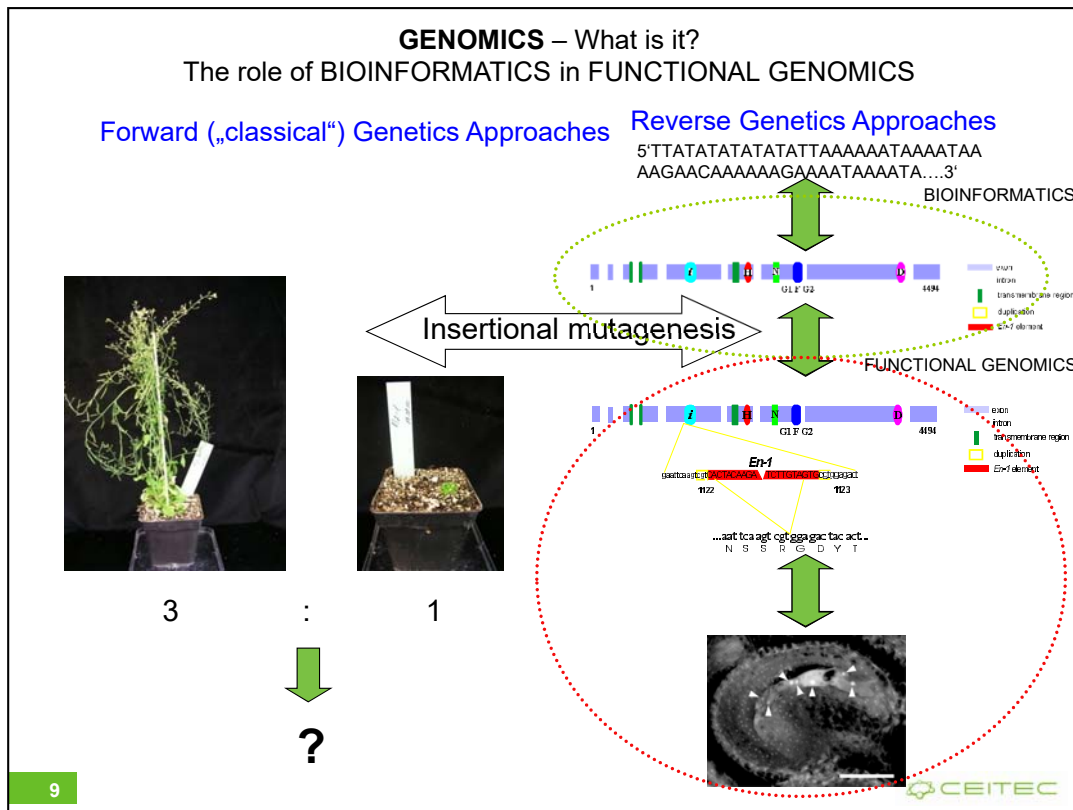
GENOMICS – What is it?

- *Sensu lato* (in the broad sense) – it is interested in **STRUCTURE and FUNCTION** of genomes
 - Necessary prerequisite: knowledge of the genome (sequence) – work with databases
- *Sensu stricto* (in the narrow sense) – it is interested in **FUNCTION** of **INDIVIDUAL GENES** – **FUNCTIONAL GENOMICS**
 - It uses mainly the reverse genetics approaches

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).

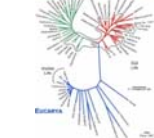
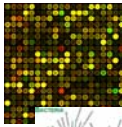
Reverse genetics uses a spectrum of approaches that will be described in the Lesson 03 that allow isolation of sequence-specific mutants and thus their phenotype analysis.

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**.

Outline

- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS

Bioinformatics



- **Definition of Bioinformatics** (according to NIH Biomedical Information Science and Technology Initiative Consortium)

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.

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.

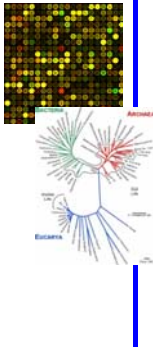
What is bioinformatics?

- **Interface** between the **biology** and **computers**
- **Analysis** of **proteins, genes** and **genomes** using **computer algorithms** and **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>

Bioinformatics



- **Bioinformatics in functional genomics**

- **Processing and analysis of sequencing data**

- Identification of reference sequences
 - Identification of genes
 - Identification of homologues, orthologues and paralogues
 - Correlative analysis of genomes and phenotypes (incl. human)

- **Processing and analysis of transcriptional data**

- Transcriptional profiling using DNA chips or next-gen sequencing

- **Evaluation of experimental data and prediction of new regulations in systems biology approaches**

- Mathematical modelling of gene regulatory networks

Outline

- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS
- Databases
 - Spectre of „on-line“ resources

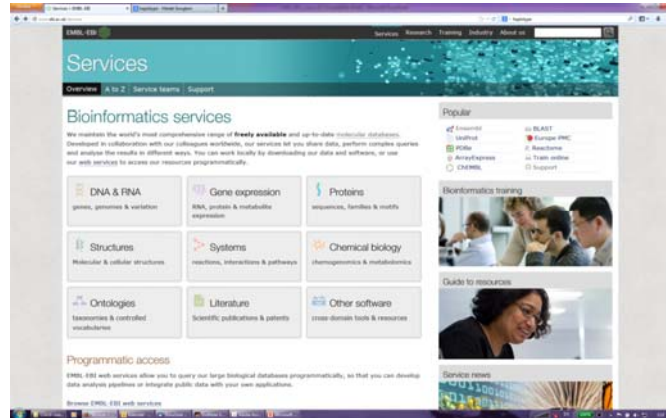
Spectre of On Line Resources

EMBLnet National Nodes		
Venna BioCenter	Austria	http://www.at.emblnet.org/
BIB	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://genoma.dlr-babelberg.de/biounit/
IMBB	Greece	http://www.imbb.forth.gr/
HEH	Hungary	http://www.hu.emblnet.org/
INEBI	Ireland	http://icet.gen.tcd.ie/
IMI	Israel	http://kspgal.wellmann.ac.il/foef/imm.htm
ITSAADR	Italy	http://bio-www.bio.cnr.it/2000/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.gdiberkian.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/
ICLISB	Italy	http://www.iclgb.bioche.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/
EMBL-EBI	UK	http://www.bioinf.embl.ac.uk/ebi/browser
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.

Spectre of On Line Resources

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



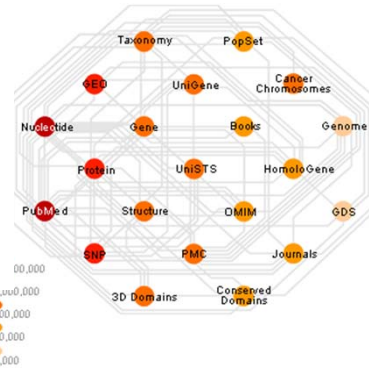
16

CEITEC

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 (

Spectre of On Line Resources

□ NCBI <http://www.ncbi.nlm.nih.gov/>



17

CEITEC

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

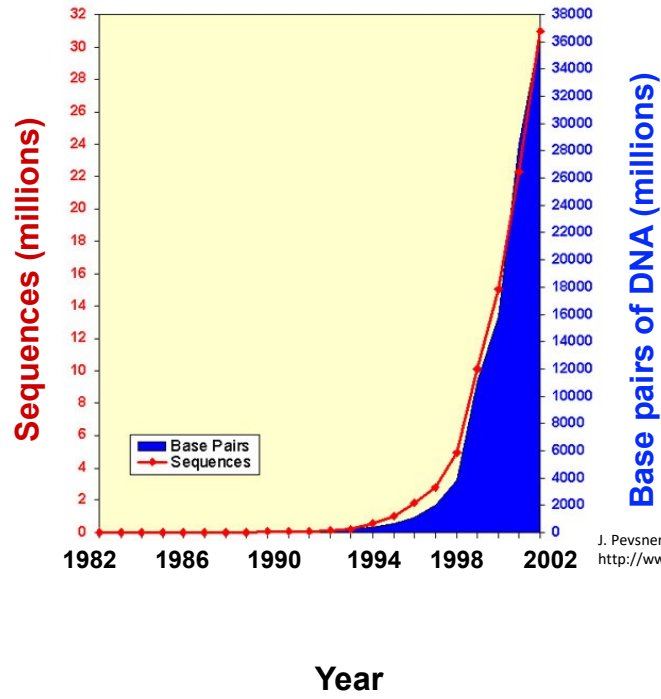
Outline

- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS
- Databases
 - Spectre of „on-line“ resources
 - PRIMARY, SECONDARY and STRUCURAL databases

Primary Databases

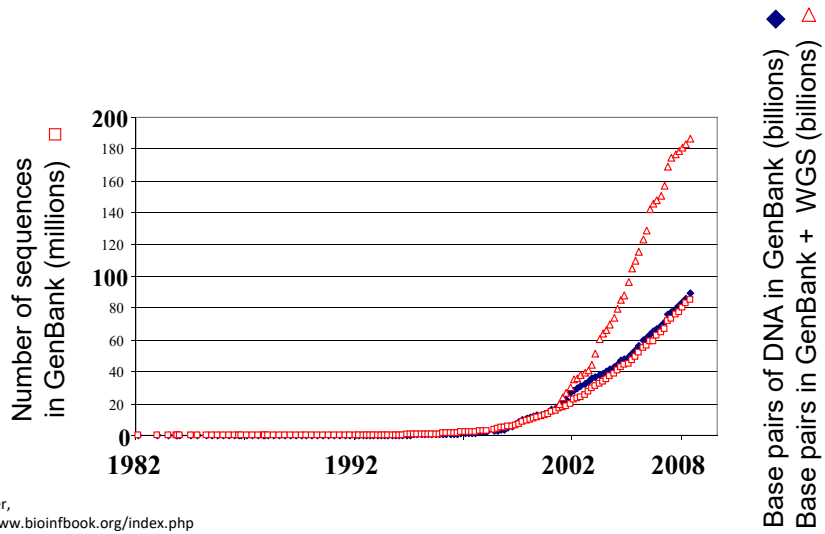
- Include primary datasets – DNA and Protein sequences
 - Sequences in databases of „The Big Three“:
 - **EMBL**
 - <http://www.ebi.ac.uk/embl/>
 - **GenBank**
 - <http://www.ncbi.nih.gov/Genbank/GenbankSearch.html>
 - **DDBJ**
 - <http://www.ddbj.nig.ac.jp>
 - Daily mutual exchange and backup of data
 - Works with large amount of data (capacity and software requirements)
 - September 2003 27,2 x 10⁶ entries (approx. 33 x 10⁹ bp)
 - August 2005 100 x 10⁹ bp from 165.000 organisms

Growth of GenBank

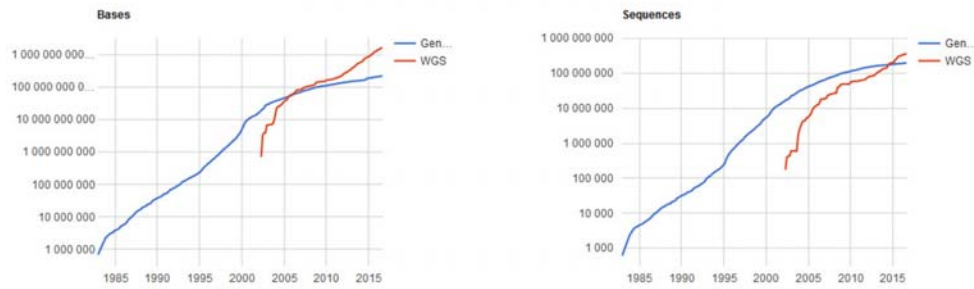


J. Pevsner,
<http://www.bioinfbook.org/index.php>

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

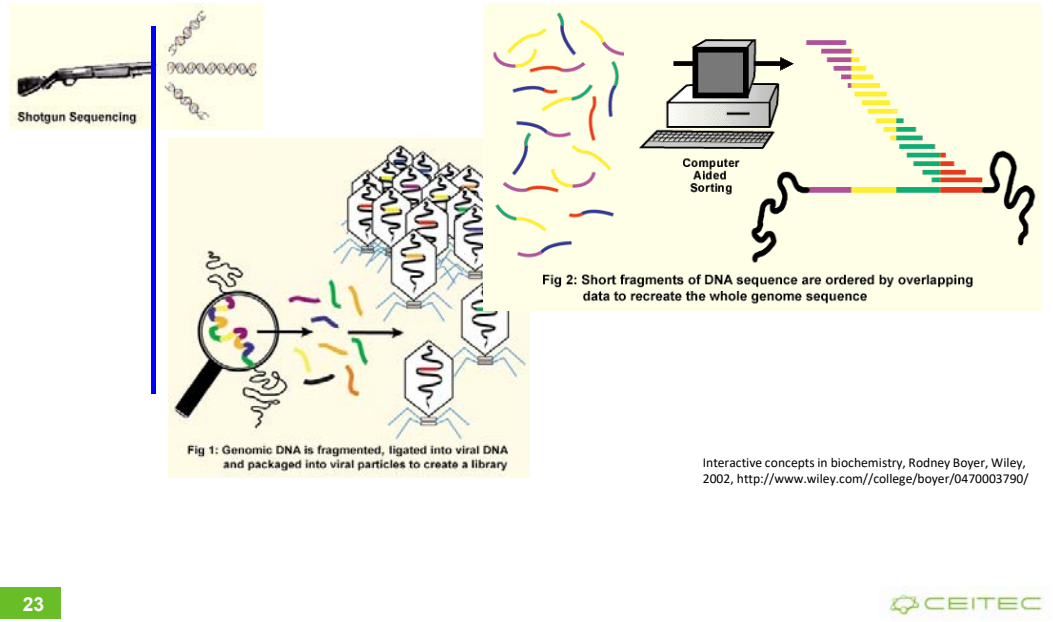


Growth of GenBank Aug 2016



- Dec **1982** 680 338 bp, 606 sequences
- Apr **2002** 19×10^9 bp, 17×10^6 sequences + WGS 692×10^6 bp, 172 768 sequences
- Aug **2016** 218×10^9 bp, 196×10^6 sequences + WGS $1,6 \times 10^{12}$ bp, 360×10^6 sequences

WGS

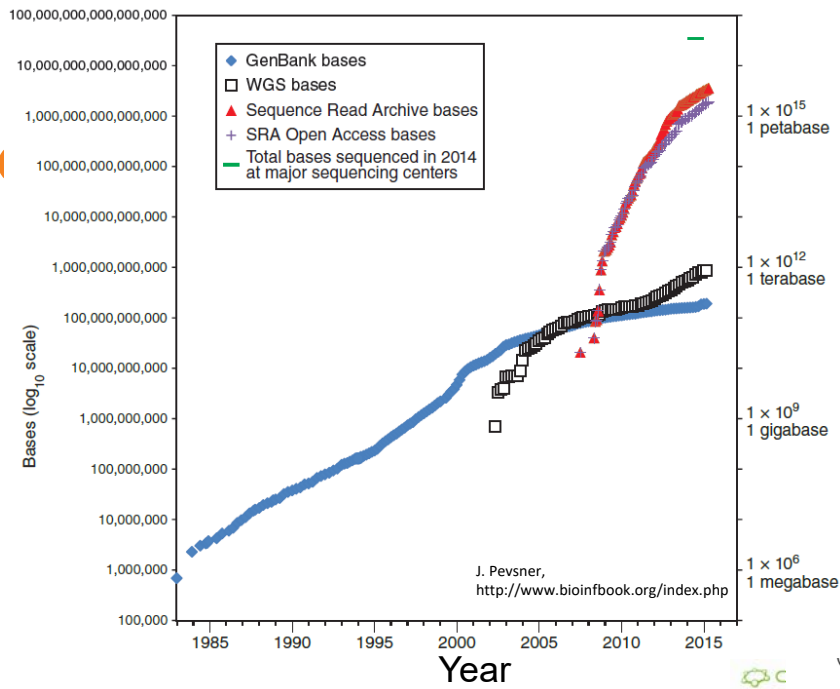


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic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

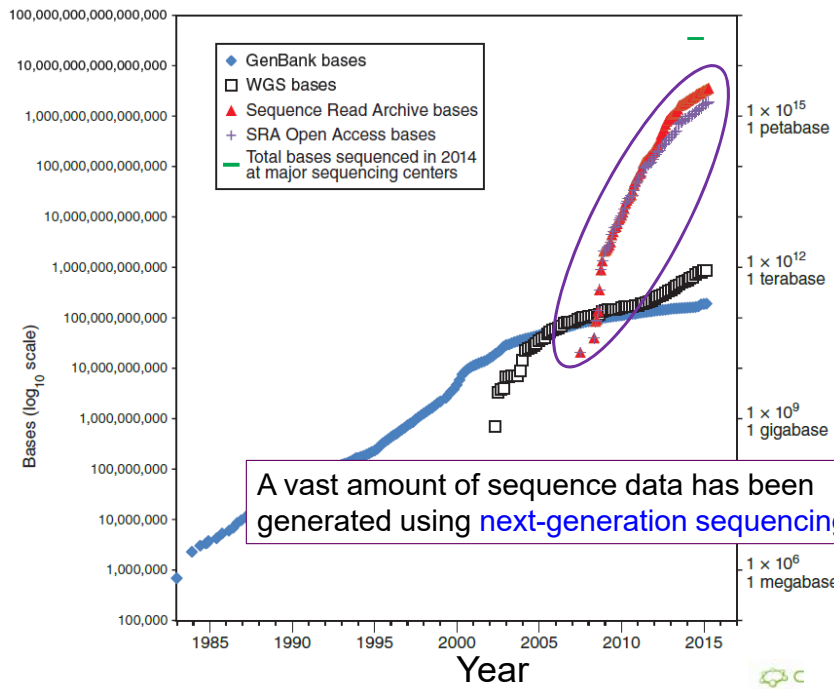
The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence.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

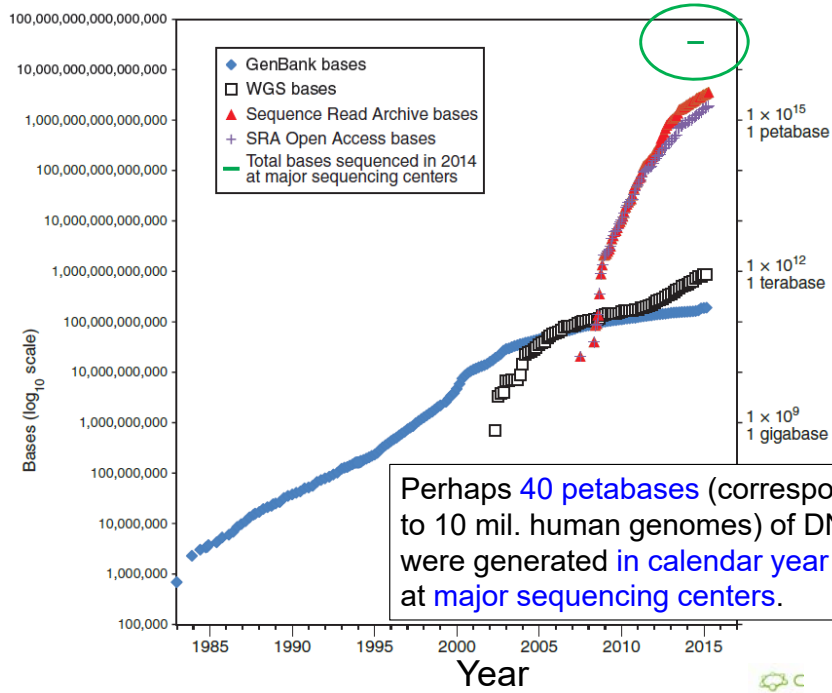
Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



B&FG 3e
Fig. 2-3
22



800

Primary Databases

- They include sets of primary data – [DNA](#) and [Protein](#) sequences
 - Protein sequences:
 - **PIR**, <http://pir.georgetown.edu/>
 - **MIPS**, <http://www.mips.biochem.mpg.de>
 - **SWISS-PROT**, <http://www.expasy.org/sprot/>

Primary Databases

- Types of sequences in primary databases
 - **Standard nucleotide sequences** acquired by high quality sequencing
 - **ESTs** (**E**xpressed **S**equences **T**ags)
 - **HGTS** (**H**igh **T**hroughput **G**enome **S**equencing)
 - Results of sequencing projects without annotation
 - **Reference Sequences** of annotated genomes
 - **TPAs** (**T**hird **P**arty **A**nnotation)
 - sequences annotated by third party (by someone else, not the original authors)

Primary Databases

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

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

Primary Databases

The screenshot displays the NCBI Gene database entry for gene NC_020771.1. The interface is organized into several sections:

- Gene symbol:** *actA*
- Gene description:** actA component of the actin filament
- Gene type:** protein coding
- RefSeq status:** PROVISIONAL
- Organism:** *Agrobacterium tumefaciens* (strain: Agrobacterium tumefaciens subsp. tumefaciens) [Rickettsiales: Rhizobium/Agrobacterium group: Agrobacterium: Agrobacterium tumefaciens complex]
- Genomic context:** Location: plasmid p1; Sequence: NC_020771.1 (24534..14915)
- Genomic regions, transcripts, and products:** A diagram shows the gene's location on the plasmid p1.
- Genomic Sequence:** A sequence viewer for NC_020771.1 (24534..14915) is shown. A yellow circle highlights the 'Links & Tools' section, which includes links to the gene's page, the sequence, and other resources.
- Related articles:** A list of four related articles is provided, including 'Sequence analysis of the actA gene of Agrobacterium tumefaciens subsp. tumefaciens' and 'The actA gene as a host-range determinant of Agrobacterium tumefaciens'.
- General Information:** A sidebar on the right contains various links and tools, including 'General information', 'About Gene', 'FAQ', 'FTP site', 'Help', 'My NCBI help', 'NCBI Handbook', 'Statistics', 'Related sites', 'BLAST', 'Gene', 'BioProject', 'Genomic Mapping', 'GEO', 'HomoloGene', 'Map Viewer', 'OMIM', 'PubMed', 'UniProt', 'UniSTS', and 'Feeds/atom'.

Primary Databases

The screenshot displays a web browser window showing a GenBank record for NC_002377.1 (2.9Kbp). A gene entry for NP_059797.1 is highlighted, and a tooltip provides 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**.

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 titled "two-component VIA-like sensor kinase" and includes a search bar. A yellow circle highlights the "NCBI Reference Sequences (RefSeq)" link in the top navigation bar. Below the search bar, the "Genome Annotation" section is visible, followed by a description: "The following sections contain reference sequences that belong to a specific genome build. Explain". The "Reference assembly" section lists genomic data for **NC_003065.3**, including its range (10061-10332) and download options (GenBank, FASTA, Sequence View, GI/NCBI). The "mRNA and Protein(s)" section lists the protein **NP_396486.1 two component sensor kinase [Agrobacterium tumefaciens str. C58]** with UniProtKB/Swiss-Prot ID P18542. It also displays conserved domains: **cd00075** (HATPase_c: Histidine kinase-like ATPases) and **cd00082** (HAKA: Histidine Kinase A). The HATPase_c domain description includes: "Location: 180 - 694; Blast Score: 202. 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." The HAKA domain description includes: "Location: 698 - 520; Blast Score: 144. A dimer is formed through parallel association of 2 domains creating 4-helix bundles, usually these domains contain a conserved His residue and are activated via ...". The PRK13837 domain description includes: "Location: 14 - 633; Blast Score: 2944. PRK13837: two-component VIA-like sensor kinase. Provisional". The "Related Sequences" section is partially visible at the bottom.

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>

Primary Databases

The screenshot displays a web browser window showing a GenBank record 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**, which is annotated as a **two-component VirA-like sensor kinase**. A green arrow points to the **Links & Tools** section of the entry, which contains the following information:

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

Below the **Links & Tools** section, there is a **Bibliography** section and a **Related articles in PubMed** section.

Primary Databases

The screenshot shows the NCBI GenBank entry for the Agrobacterium tumefaciens plasmid Ti, complete sequence. The entry is identified as NC_002377.1. The sequence is displayed in FASTA format, starting with 'xpi110955014:14584-14813 Agrobacterium tumefaciens plasmid Ti, complete sequence'. The sequence is a long string of nucleotide bases (A, T, C, G). On the right side of the page, 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, and Highlight Sequence Features), 'Find in this sequence', 'Related information' (with links for RefProject, Full text in PMC, Gene, Genome, Metadata, GenBank Sequence, Protein, Protein Clusters, PubMed, and Taxonomy), and 'Recent activity' (listing recent searches for 'Agrobacterium tumefaciens plasmid Ti, complete sequence' and 'vIA [Agrobacterium tumefaciens]').

Secondary Databases

- Databases of **functional** or **structural motifs**, acquired by **primary data** (sequences) **comparison**
- **PROSITE**, <http://www.expasy.org/prosite/>

```
>PDOC0001 PS0001 SULFATION Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].
171 - 185 nbsaaTxxxxxx

>PDOC0004 PS0004 CAMP_FOSPHO_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
744 - 745 SSST
814 - 817 PStC

>PDOC0005 PS0005 PKC_FOSPHO_SITE Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].
148 - 150 GAK
161 - 162 TGP
171 - 173 GSK
219 - 221 SAK
269 - 271 TTK
440 - 442 SGP
513 - 515 PGP
585 - 587 SPS
632 - 634 TGP
672 - 674 TSK
716 - 718 SPS
726 - 728 SPS
747 - 749 SPS
794 - 796 SAK
816 - 818 SPS
844 - 846 SSK
868 - 870 SPS
921 - 923 SPS
957 - 959 SPS
960 - 962 TGP
974 - 976 TSK
997 - 999 SPS
1062 - 1064 TGP
1018 - 1020 SGP
1031 - 1033 TGP
1119 - 1121 SAK
```


Secondary Databases

- Databases of **functional** or **structural motifs**, acquired by **primary data** (sequences) **comparison**
- **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 a protein family; its diagnostic power is defined by iterative scanning of a PROTEIN/GENOME 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, but diagnostic potency deriving from the central content 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:

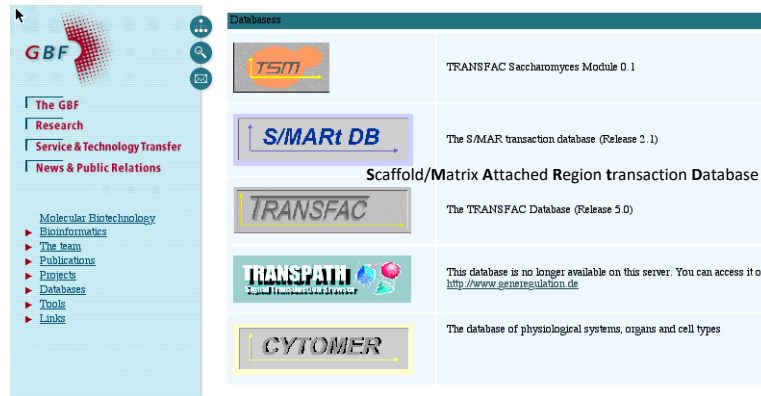
- [By accession number](#)
- [By PRINTS code](#)
- [By domain code](#)
- [By motif](#)
- [By sequence](#)
- [By motif of motifs](#)
- [By motif](#)
- [By access information](#)

PRINTS search:

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

Secondary Databases

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



The screenshot shows the TRANSFAC website interface. On the left is a navigation menu for GBF (Gene Bioinformatics Facility) with 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>)

Structural Databases

- **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](#)



[Molecule 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



Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

[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

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

News

[Complete News](#) [pdb1 Archive](#)
[Newsletter](#) [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](#)

[Bologna University](#)

[National Institute of Standards and Technology](#)

[Cambridge Crystallographic Data Centre, UK](#)

[National University of Singapore](#)

[Osaka University, Japan](#)

[Universidade Federal de Minas Gerais, Brazil](#)

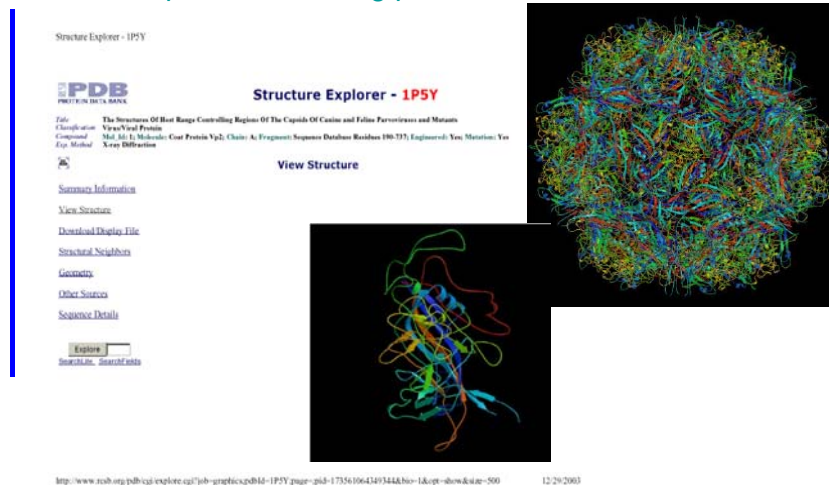
[Max Delbrück Center for Molecular Medicine, Germany](#)

[OTHER SITES](#)

Structural Databases

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

Structure Explorer - 1PSY



PDB
PROTEIN DATA BANK

Structure Explorer - 1PSY

Title: The Structure Of Host Range Controlling Region Of The Capsid Of Canine and Feline Parvoviruses and Mutants
Classification: Virus/Viral Protein
Compound: 304; 301; 3; Molecular: Coat Protein Vp2; Chain: A; Fragment: Sequence Database Residues 180-373; Engineered: Yes; Mutation: Yes
Exp. Method: X-ray Diffraction

View Structure

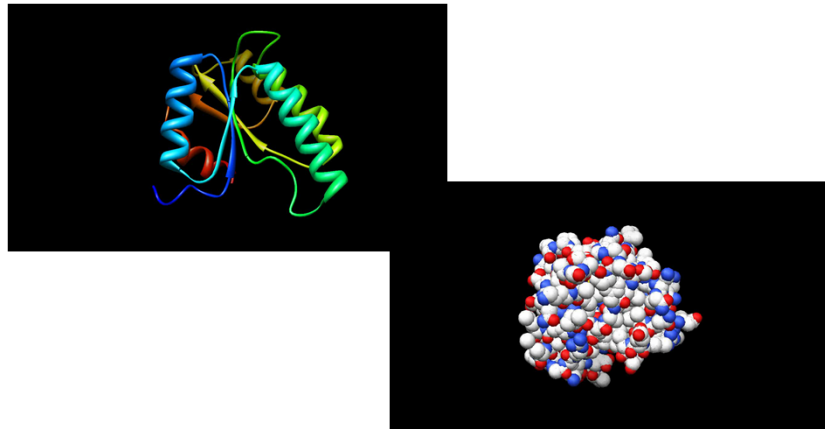
[Summary Information](#)
[View Structure](#)
[Download Display File](#)
[Structural Neighbors](#)
[Geometry](#)
[Other Sources](#)
[Science Details](#)

[Search for... Search for...](#)

<http://www.rcsb.org/pdb/cgi/explorer.cgi?pdb=1psycsdm&1-1PSY.znuc-pdb-173361064340344&bio-1&opt-show&size=500> 12/29/2003

Structural Databases

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



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

Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre of „on-line“ Resources
 - PRIMARY, SECONDARY And STRUCURAL Databases
 - **GENOME Resources**

Genome Resources

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

The screenshot shows the Human Genome Browser (hg19 assembly) interface. The search bar contains the coordinates 'chr1:100,000,000-100,000,000'. Below the search bar, there is a table with columns for 'chr1', 'genome', 'assembly', and 'position'. The main content area displays a list of 'Sample position queries' with their corresponding descriptions. A UCSC logo is visible on the right side of the page.

Request:	Genome Browser Response:
chr1	Displays all of chromosome 1
chr1:g000212	Displays all of the unplaced contig g000212
25k17	Displays region for band 25k on chr 20
chr3:1,000,000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000-2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
rs100000000	Displays region between genome landmarks, such as the STS markers D10S1081 and D10S171, or chromosome bands 15q11 to 15q11.1, or SNPs rs104252 and rs1000000. This syntax may also be used for other range queries, such as between uniquely determined ESTs, miRNAs, refSeqs, etc.
rs104252:rs1000000	
D10S1081	Displays region around STS marker D10S1081 from the Genethon/Whitehead map. 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
F08P	Displays region of genome with HGST Gene Nomenclature Committee identifier F08P
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	Lists identified phenotypes, list ref IDs
ncsdbase ncdbase	Lists miRNAs for causal transitive genes
zinc finger	Lists many zinc finger miRNAs
knockout zinc finger	Lists only knockout zinc finger genes
huntington	Lists candidate genes associated with Huntington's disease
zinc	Lists miRNAs deposited by scientist named Zinke
Evans J.E.	Lists miRNAs deposited by co-author J.E. Evans

Genome Resources

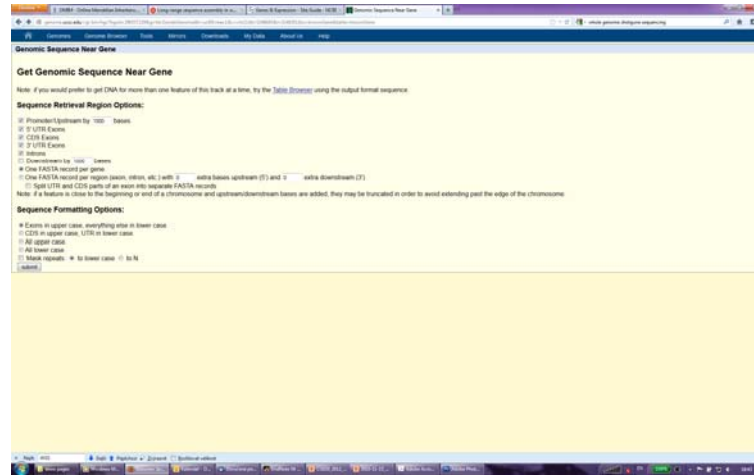
Human Genome Browser <http://genome.ucsc.edu/cgi-bin/hgGateway>

The screenshot displays the Human Genome Browser interface for the HBB gene. The page is titled "Human Gene HBB (uc0010aa.1) Description and Page Index". It includes a description of the gene, a list of publications, and a table of sequence and links to tools and databases. A green arrow points to the "Sequence and Links to Tools and Databases" section, which contains a table with columns for Gene Symbol, Genome Browser, Protein FASTA, UniProt, RefSeq, and others. The table lists various tools and databases such as Ensembl, Eukaryotic Genome, UniProt, RefSeq, and others.

Gene Symbol	Genome Browser	Protein FASTA	UniProt	RefSeq	Ensembl	Eukaryotic Genome	UniProt	RefSeq	Ensembl
HBB	UCSC	FASTA	UniProt	RefSeq	Ensembl	Eukaryotic Genome	UniProt	RefSeq	Ensembl

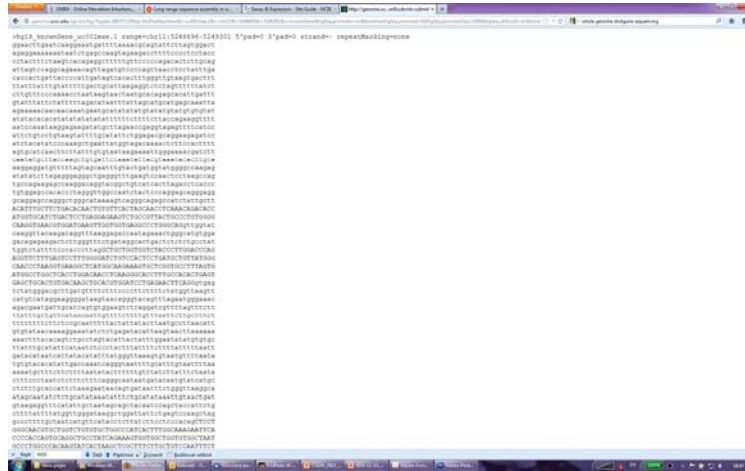
Genome Resources

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



Genome Resources

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



Genome Resources

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



Genome Resources

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

The screenshot shows the TAIR website homepage. At the top left is the TAIR logo. A navigation bar contains links for Home, Help, Contact, About Us, Login, and a search box with 'Gene' entered. Below this is a secondary navigation bar with links for Search, Browse, Tools, Stocks, Portals, Download (circled in red), 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 interface is shown at the bottom of the page.

Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY And STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homology Searching

Analytical Tools

□ Global versus Local alignment

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

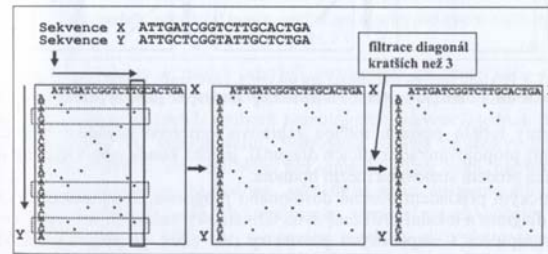
Lokální přiřazení
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVE
-----NAPATNIKSECVRA-PIQNYRRVEHVRA-----
```

Cvrčková, Úvod do praktické bioinformatiky

- **Global Alignment:** only for sequences, which are **similar** and of a **similar length** (BUT can insert spaces into one or both sequences)
- **Global Alignment** is used mainly in case of **multiple alignment** (CLUSTALW, further in the presentation)
- **Local Alignment** provides identification and comparison even in case of alignment of **regions of sequences with high similarity**, e.g. even in case of **change of order of protein domains** during evolution

Analytical Tools

- Choosing the right type of alignment using dotplot

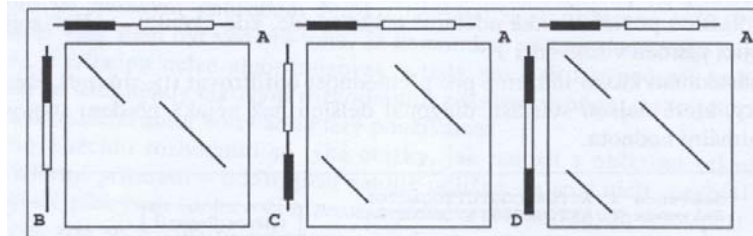


Cvrčková, Úvod do praktické bioinformatiky

- Plotting the sequences against each other (x and y axis)
- Identification of identity in „dot“ of specific size (e.g. 2 bp)
- Filtering the diagonals of lengths lower than a threshold

Analytical Tools

□ Examples of sequence alignment using dotplot



Cvrčková, Úvod do praktické bioinformatiky

- **Global Alignment:** possible *only* for sequences A and B
- The rest of the sequences underwent change of order of protein domains and therefore it is necessary to do a local alignment
- Dotplot can be obtained using [BLAST2](#) (see further in the presentation)

Analytical Tools

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

The screenshot shows the NCBI BLAST interface for a nucleotide-nucleotide search. 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 search area contains a text box with the following sequence:
aacacccggc
aacacacat cattatcacc atcgctttgg ggcgatggtg tctggttcca
gggtattaat
ataattaatt tattccacat gagatgatg atgatatact atgtattttt
tgcttttttt
ttatttgtaa acctttaata taacaagaac tacaaaaaat gaaaa

Below the text box is a "Search" link. Underneath, there are fields for "Set subsequence" with "From:" and "To:" input boxes. A "Choose database" dropdown menu is set to "nr". At the bottom, there are three buttons: "BLAST!" (highlighted in blue), "Reset query", and "Reset all".

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 gtcgacaacggctctggcatgtgcaaggccggatttgcgggagacgatgctccccggccc 215
Sbjct: 101 gtcgacaacggctctggcatgtgcaaggccggcttcgctgggagacgatgccccggggcc 160

Query: 216 gctctcccatcgatttgggaagtcctccgtaaccagggtgtgatggcctggccag 275
Sbjct: 161 gctctcccatcgatttgggaagtcctccgtaaccagggtgtgatggcctggccag 220

Query: 276 aaggactcgtaacgtgggtgatgagggcagagcaagcgtggtatcctcacctgaagtac 335
Sbjct: 221 aaggatcctctatgtgggacagggccagagcaagagagggatcctcacctgaagtac 280

Query: 336 cccattgagcaaggatctgtgaccaactgggacgatggagaagatctggcaccacacc 395
Sbjct: 281 cccatgagcaaggatctgtgaccaactgggacgatggagaaaatctggcaccacacc 340

ds..S=1213 E=0.0
>=200
250 1500

- „expectancy value“ provides the number of expected sequence number with the same or higher similarity when searching in the database consisting of randomly assembled sequences
- the results shows fraction of identical and in case of proteins also similar sequence positions and/or inserted spaces

Primary Databases

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 gene information, there is a "Bibliography" section and a "Related articles in PubMed" section.

63

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

BLINK pre-computed BLAST

Home Taxonomy Report Multiple Alignment Blast Help My NCBI

Pre-computed BLAST results for: [gi|16119781|ref|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

1263 Archaea 13825 Bacteria 13 Metazoa 1349 Fungi 554 Plants 6 Viruses 5676 The Others [reset selection](#)

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

% identity	823 aa	query selection	score	accession	length	protein description
				Conserved Domain Database hits		
100			823	AF093322	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]
100			823	F15345	833	Declume1 Full-Wide host range virA protein/ Short-WDR virA
100			823	AA073232	833	virA [Flamnid pTIC8]
100			823	WP_013390	833	Hypothetical protein pT1-SAMBA_g142 [Agrobacterium tumefaciens]
100			823	NM017065	833	virA140 [Agrobacterium tumefaciens]
100			823	AA015596	833	virA [Flamnid T1]
100			823	gi 1731317	833	virA protein
100			823	CA043773	833	virA kinase protein [Agrobacterium tumefaciens]
100			823	CA033760	829	virA [Agrobacterium rhizogenes]
100			829	gi 12272265	829	virA gene
100			829	AA016883	829	virA [Flamnid T1]

BLAST

Specialized Versions

- Currently there exists a lot of specialized versions of **BLAST**
 - Searching according to source (organism) of sequences, e.g. known genomes of **microorganisms**
 - **BLASTP**
 - Given the **protein query**, it returns the most similar protein sequences from the **protein database**.
 - **BLASTN**
 - Given the **DNA query**, it returns the most similar DNA sequences from the **DNA database**.
 - Other variants, e.g. **MEGABLAST**, for identification of identical or **very similar sequences** (searches **long similar regions** of nucleotide sequences)
 - **BLASTX**
 - Compares the all possible **six-frame translation products** of a **nucleotide query sequence** (both strands) against a **protein sequence database**.

BLAST

Specialized Versions

- Currently there exists a lot of specialized versions of BLAST
 - **TBLASTN**
 - Compares a **protein query** against the **all six reading frames** of a **nucleotide sequence database**.
 - **TBLASTX**
 - **Translates** the **query nucleotide sequence** in **all six possible frames** and compares it against the **six-frame translations** of a **nucleotide sequence database**.

BLAST

Specialized Versions

- Currently there exist a lot of specialized versions of BLAST
 - **PSI-BLAST** (Position-Specific Iterated Blast)
 - **First step:** standard BLAST, during which PSI-BLAST identifies a list of similar sequences with E value better than minimal value (standard = 0,005)
 - For every alignment, PSI-BLAST creates so-called PSSM (Position Specific Substitution Matrix)
 - PSSM takes into account relative frequency of specific aminoacid residue in a specific position within sequences identified as similar in first step, which can mean functional conservation.

BLAST

Specialized Versions

- Currently there exists a lot of specialized versions of BLAST
 - **PHI-BLAST** (Pattern-Hit Initiated BLAST)
 - For identification of **specific sequence**, e.g. motif (pattern) in sequence of similar protein sequences
 - Sequence of motif must be inserted using **special syntax**:
 - [LVIMF] means either Leu, Val, Ile, Met or Phe
 - - is spacer (means nothing)
 - x(5) means 5 positions in which any residue is allowed
 - x(3, 5) means 3 to 5 positions where any residue is allowed

BLAST

Specialized Versions

□ Example of search by PHI-BLAST

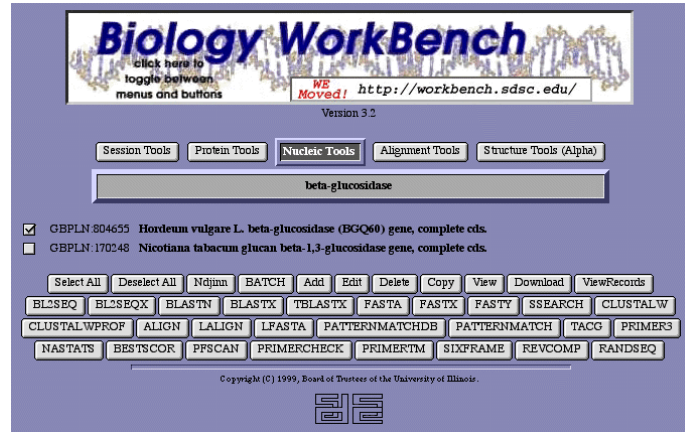
```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase  
MSHIQIPPGLETLLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAAATPRQSLGHPPPEPGPDR  
VADAKGDSSEEEDEDLEVPVPSRFNRRVSVCAATYNPDEEBEEDTDPRIHPKTDEQRCRLQEBACKDILLF  
KNLDQEQLSQVLDAMFERIVKADEHVIDQDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGRLA  
LMYNTPRAATIVA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

Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY And STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...

Analytical Tools

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



Analytical Tools

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

View
View Nucleic Sequence(s)

Format: Fasta Case: Upper Change Format

[Download/view all sequences in text format](#)
[\[NEXT\]](#) [\[BOTTOM\]](#)

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

>170248
GAGCTCCCTTGGGGGGCAAGGGCAAACTTTTGGCTAAATGGAAAAATTTATACCANGTGTGTAATA
GTTACTCAATTTTGAATTAACAAGGGGCAATTTTACTATTTTGGCCCTTATCTTTTGGTCAAAAAAC
ATAAAATATCCCATCCGAAATCCAAATGGTCCAAATATCGGCAAGTAGCTTTCMTTAAATATAGTTAGTT
GACAAAACCTATCAAGATATCATTAATTAATAATAACTTCAAAATGTCATCATCTTAGCTGCCCTCTCA
GTAGAGCCGCGAGTAAATTAAGCCGATCAAAATAAAGCCGCGCAATTAATAATGAATTTTAGGACTCTC
GATTGGCACGTAAAGTCCAAAACCTTCCAAATCTTTGCTCAACTTGGGGGCTCTAGGTTCTAGCTTC
CGATATGGGATTTTCTTAGGTTTATCTCTAATTTTACATCTCACTAATTTTAAAGAAATTAAGCGGTA
CAGCAAACTATAAAATTTTCTCTAAAGAAAGACAAATGATCCGGTTACTGATTCATTTGGCTTTTCAGAG
TCTGCATGCCATATCTACTAAGGGTCTGTTGGTACAAAGAAATAATAATAATAATTTGGGATAGAAATTT
GAGATTGCATTTATCTTGTGTTTAAATTAAGTATTAGCTAATTTTCAGAAATAAATTTTACACTAAATAG
TAAATCACTTCTACATTTGAGGTTGAAATGAAATGCTTAAATCCATGCGCTCACTAGAAATATCC
TTAATTTATCTACTAATTTTCCAAATGATCGGTTAGTCTTCAAGAGATCCAGTATCTCAATAAATGCA
GTAAAGGTTAGAAAATTTTCAATTAATCAATTCATATAAATTAATAATATAGATATGGAGCACTTAAG
ATCAATAAAGATGTACCGTTAATAATAAAGATAGATAGAGTTTAAATAGGAAAAAACAACGGTT
CGAGCACCTTTATGGAAAGCGGTTTCTCAAGATAGATTTCTCAATTAATGCTTGGTCAATAGCAAAA
TCACTCTTACTTTTAGATACAGCGACCCACTTCAATCTTCTATTTGATCTCAAAATGAAATTTTA
GGAACTTCAAACTCTCAACTACTTTTAAAGGAAATCAAAATACGACCAATATTTTACTTACTTAC
TTATAGTTAAATGATATGAATTTTATTTTAAATTTGAAATGAAAAATTTAAATACCTGATTTAATATAA

Analytical Tools

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

Regex pattern:

ctt. {1, 32}ctt

0 sequences were searched

1 match was found

Matches are indicated in blue

> 170248

```
GAGCTCCTTGGGGGGCAAGGGCAAAACTTTTTCCTAAATGGAAAAATATATACCAAGTGTTTTGTAAAT
GTACTCAATTTGAATTAACAAGGGGCAAAATTTGACTATTTTGGCCTTATACTTTTTGGTACAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGTTTTCTTTAAATATAGTTAGTT
GACAAAACACTATCAGATATCAATTTTATAATAATAATTCAAAATTCCAATTTTAGTGGCTCCCTCA
GTAGGCCCCCGTAAATTAAGCGATCAAATAAGGGCCCAATRAAATAAGAAATTTTGGACTCTC
GATTTGGCAGTAAAGTCCAAAACTTTCCAATACTTTGCTCAACTTTGGGGCTGTAGGTCTTGAGCTTC
CAGATATGGGATATTTTAAGTTTATCTCCTAATTTTACATCTCAACTAAATTAAGAAATTAACAGGTA
CAGCAATCATAAAATTTTCCCTAAAGAGGACAATAATCCGGTTACTGATTCATTTGGCCTTTTCAGAG
TCTCCATGCCATTTCACTAGGGGTCTTTTGGTAAAGAAATTAATATAATTTTCCGGATAGAAATTT
GAGATTGCATTTATCTTGTGTTTAAATTAAGATTAAGTATAGCTAATTTCAGATAAATTTTACATAAATAG
TAAAATCAACTATCACATGTAGAGGTGGAATGGATAGCTAATCCATAGCCACTCACATAGAAATATCC
TTATTTATCTCACTATTTTACCAATGATCGGTTAGTCTTCATAGAAATCCAGTATCTCAATAAATGCA
GTAGAGTTTGAATAATTTCTTTTAAATCAATTCATTAATTTAAATAATTTTGGTATGGGCACTTAG
ATACAATAAAGATGTACCGTAAATAAAGATAGATAGAGTTTAAATAGGAAAAAAAAACCGTT
CGAGACTTTATGGAGGGGTGTCTTTCAAGTAGATTCATTCATTTGCTCTGGTGCATATGCAAAA
TGACATTTACTCTTAGATACAGCGAGCACTTCAATCTTCTATGATATACIAAATGAAGTTTTA
GGAGTTTTAAATGTTAGCTTTTTAGGGAATTCAAAATGACAAATTTTATTATTACTTAC
TTATAGTTAAATGATAGAATTTATTTTAAATTTGAATGAATATTAAATTACTTGATTTAATATAA
ACAATAGATATCGCTAGATTTTACCAAAAATGGAGATCACAGAAGATTTTATTTTGTAACGAT
GTTAAGAGCTATTTATCTGGTTTGGAGGATGAAGAAAGTAACTAGCTATTTTCTTTTTAAGT
```

Analytical Tools

- <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*INRGNLTILEL

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 gtttgaatagtactcaattgaattaacaaggggcaaatgactatttgcctta 120

Frame 2, 1 stop codon

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

>170248 Translated - Frame 2
SSLGGQGQNFLLNGKILYQV

S S L G G Q G Q N F L L N G K I L Y Q V
2 agtcacctgggggcaaggcaaaacttttgctaaatggaaaaatattataccaagt 61
F V I V T Q F E L T K G Q I * L F C P
62 tttgtaatagtactcaattgaattaacaaggggcaaatgactatttgcctta 120

Analytical Tools

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

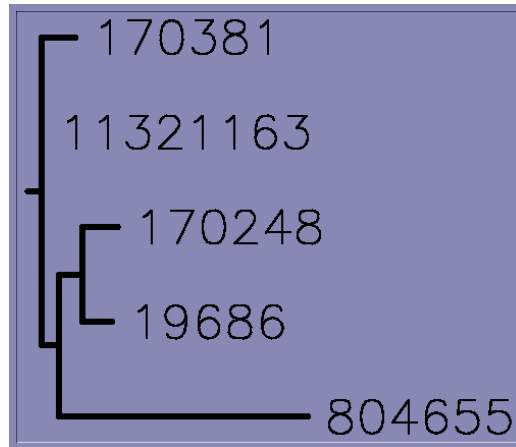
== Linear Map of Sequence:

```
StyI
BsaJI
CviJI
AluI
SacI
EcoICRI
Bsp1286I
BsiHKAI
BamII BslI
SspI
1 gagtcctctgggggcaaggcaaaaacttttgtaaatggaaaaatataccaagt 60
ctcgagggaaaccccggttcccgtttgaaaaacgatttaccctttataataggttca
^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^
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 Tsp509I
MaeIII Tsp509I MseI ApoI
61 gttgtaatggttactaacttgaatttaacaaagggcaaatgactatttgcotta 120
caaacattatcaatgagttaaacttaattgttcccggttaaacgtataaacgggaat
^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^ ^
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
```



Analytical Tools

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



Analytical Tools

- o 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

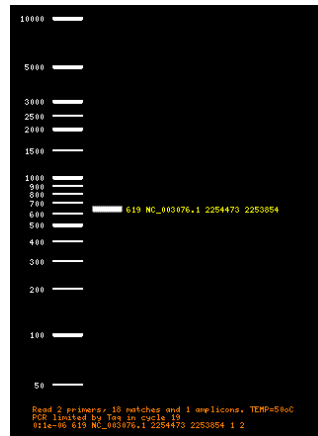
Primer 8

Annealing temperature



Analytical Tools

- o VPCR <http://grup.cribi.unipd.it/cgi-bin/mateo/vpccr2.cgi>



Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY And STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...
 - **Other On-line Genome Tools**

Other On-Line Genome Resources

- **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, followed by the gene name and a 'Table of contents' sidebar. The main content area is divided into sections: Summary, Official Symbol, Official Full Name, Primary Source, Locus tag, Size related, Gene type, RefSeq status, Organism, and Location. A 'Genomic context' section shows a map of the gene on Chromosome 1, with a scale bar and a 'See PHACTR4 in Map Viewer' link. Below the map is a 'Genomic regions, transcripts, and products' section with a 'Go to RefSeq' link. The bottom of the page features a 'Genomic sequence' section with a 'Go to RefSeq' link.

Other On-Line Genome Resources

- Online Mendelian Inheritance in Man (**OMIM**)



Summary

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY and STRUCURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...
 - Other On-line Genome Tools

Discussion