

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

M U N I
S C I



Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY and STRUCTURAL 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 – What is it?

The role of BIOINFORMATICS in FUNCTIONAL GENOMICS

Forward („classical“) Genetics Approaches



3

:

1

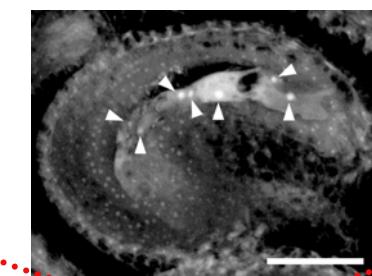
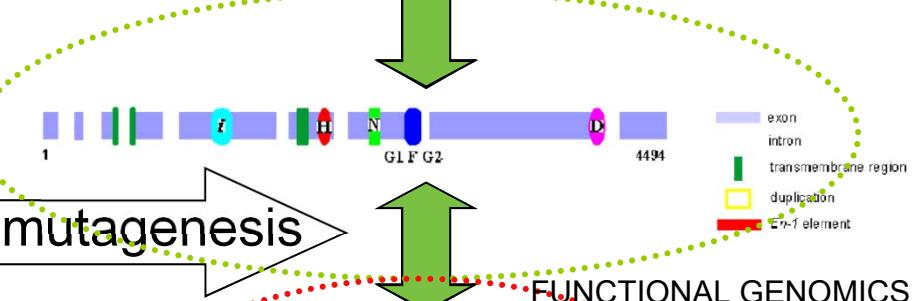


?

Reverse Genetics Approaches

5'TTATATATATATTAAAAAAATAAAATA...3'
AAGAACAAAAAAAGAAAATAAAATA...3'

BIOINFORMATICS

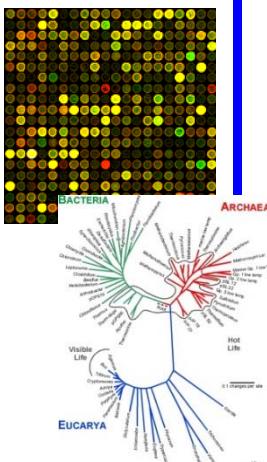


CEITEC

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**.

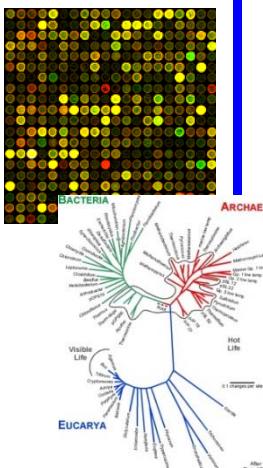
What is bioinformatics?

- 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

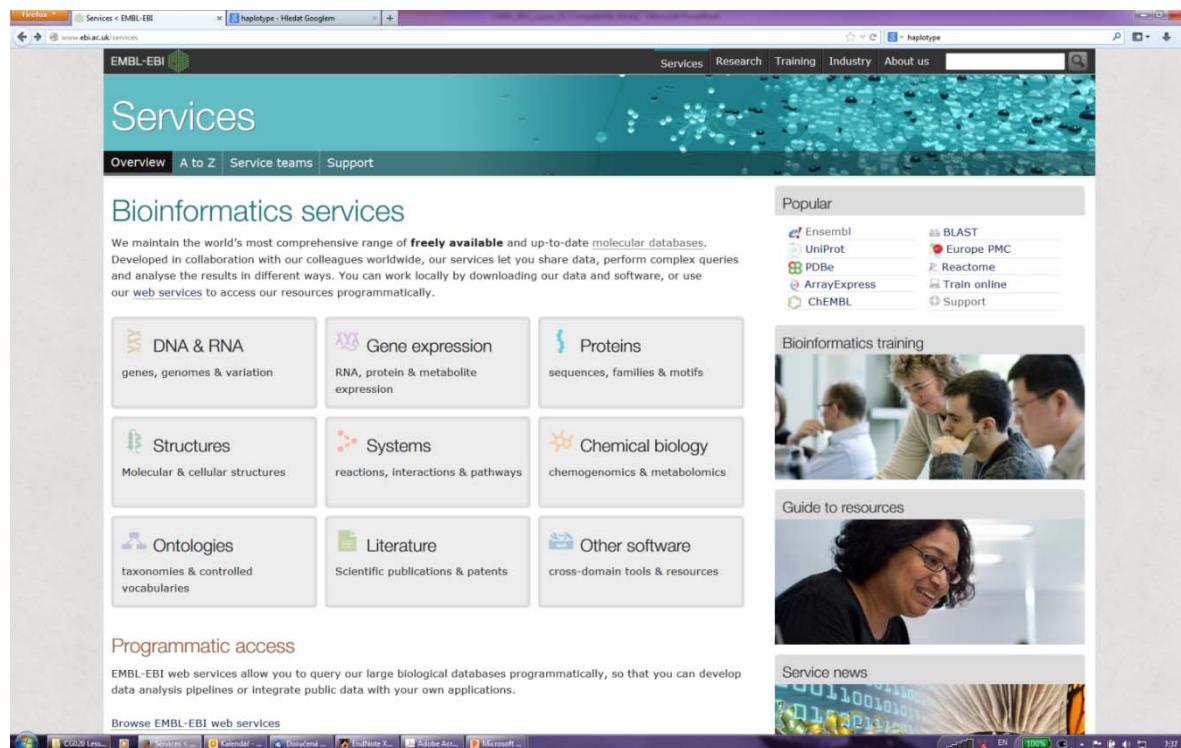
- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS
- Databases
 - Spectre of „on-line“ resources

Spectre of On Line Resources

<i>EMBnet National Nodes</i>		
Vienna Biocenter	Austria	http://www.at.embnet.org/
BEN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://blobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENIUSnet	Germany	http://genome.dkfz-heidelberg.de/bionet/
IMBB	Greece	http://www.imbb.forth.gr/
HEN	Hungary	http://www.hu.embnet.org/
INCBI	Ireland	http://acer.gen.tcd.ie/
INN	Israel	http://dapsas.weizmann.ac.il/bcd/inn.html
IEN-ADR	Italy	http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm
CADS/CAMM	Netherlands	http://www.caos.kun.nl/
Bio	Norway	http://www.no.embnet.org/
IBB	Poland	http://www.ibb.waw.pl/
IGC	Portugal	http://www.igc.gulbenkian.pt/
GeneBee	Russia	http://www.genabee.msu.su/
CNB-CSIC	Spain	http://www.es.embnet.org/
BMC	Sweden	http://www.embnet.se/
SIB	Switzerland	http://www.ch.embnet.org/
SEQNET	UK	http://www.seqnet.dLac.uk/
<i>EMBnet Specialist Nodes</i>		
MIPS	Germany	http://www.mips.biochem.mpg.de/
ICGEB	Italy	http://www.icgeb.trieste.it/
Pharmacia Upjohn	Sweden	http://www.pnu.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.man.ac.uk/dbbrowser
<i>EMBnet Associate Nodes</i>		
IBBM	Argentina	http://sol.biol.unlp.edu.ar/embnet
ANGIS	Australia	http://www.angis.su.oz.au/
CBI	China	http://www.cbi.pku.edu.cn/
CIGB	Cuba	http://bio.cigb.edu.cu/
CDPD	India	http://salajung.embnet.org.in/
SANBI	South Africa	http://www.sanbi.ac.za
<i>USA Information Providers</i>		
NCBI	USA	http://www.ncbi.nlm.nih.gov/
NLM	USA	http://www.nlm.nih.gov/
NIH	USA	http://www.nih.gov/

Spectre of On Line Resources

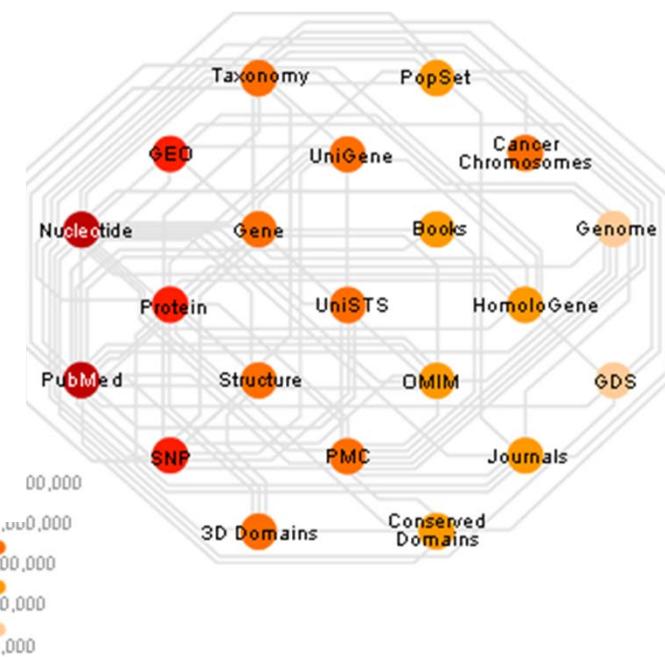
- EBI <http://www.ebi.ac.uk/services>



Spectre of On Line Resources

- NCBI <http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI homepage. At the top is a search bar with the placeholder "All Databases". Below the search bar is a sidebar with links to various resources: 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, and Variation. The main content area features a "Welcome to NCBI" section with a brief introduction and links to About the NCBI, Mission, Organization, Research, and RSS Feeds. Below this is a "Get Started" section with a bulleted list of links to Tools, Downloads, How-To's, and Submissions. To the right is a "Popular Resource" sidebar listing PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. At the bottom left is a "NCBI YouTube channel" section with a link to the channel and a "GO" button. On the right is a "NCBI Announcer" section with news items about the new version of Geno, NCBI's July Newsletter, and the Bookshelf.



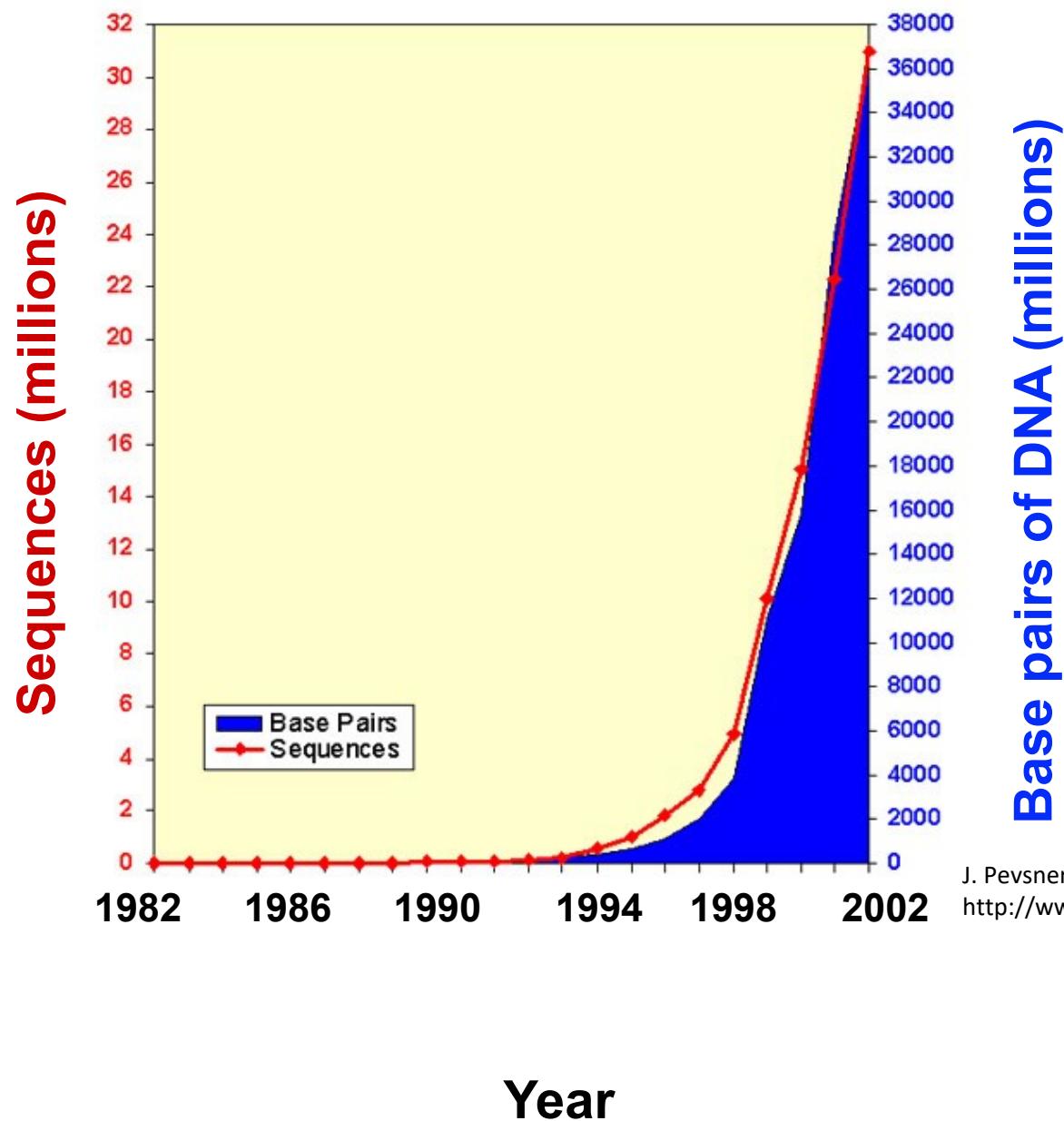
Outline

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

Primary Databases

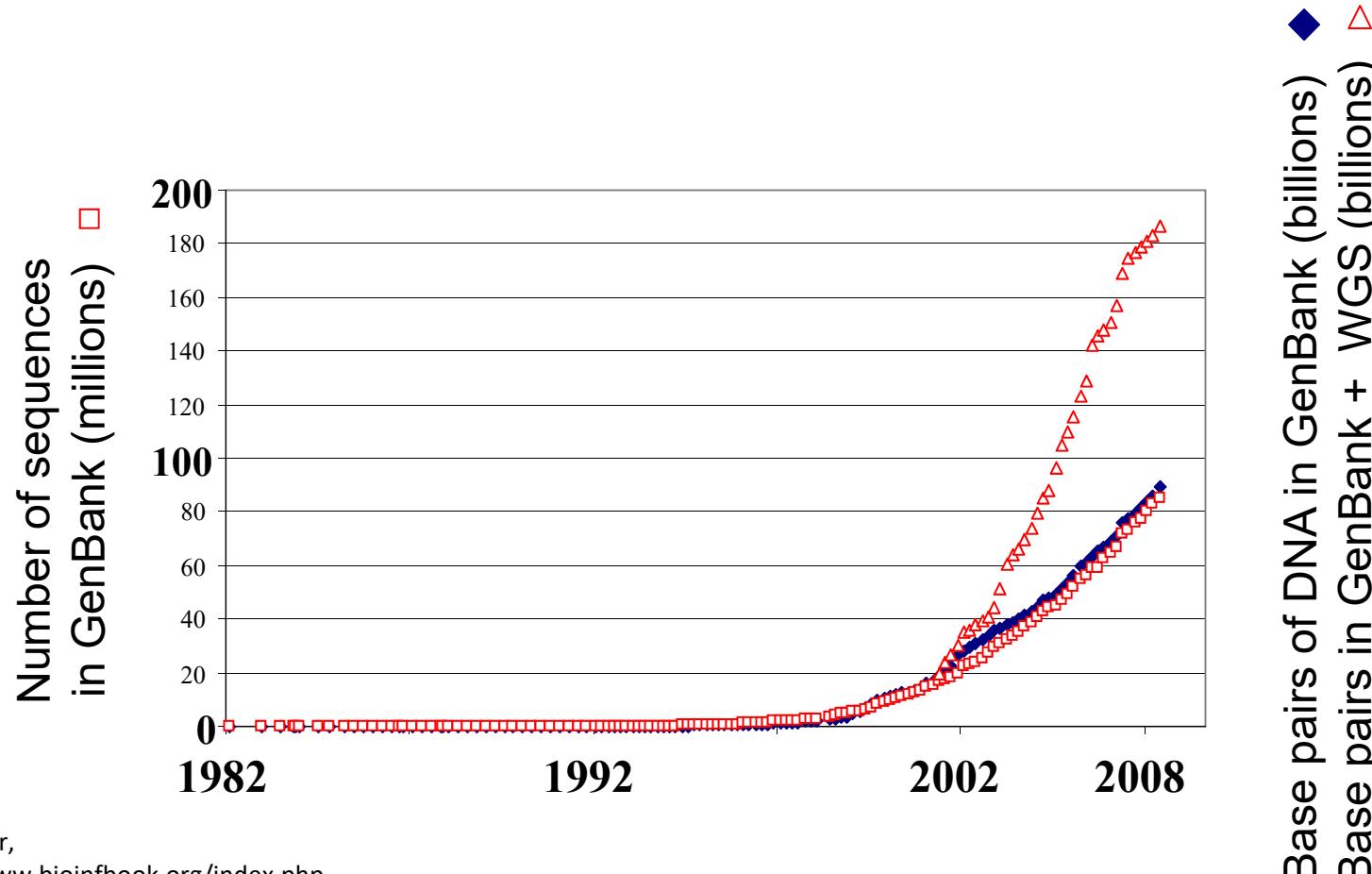
- Include primary datasets – DNA and Protein sequences
 - Sequences in databases of „The Big Three“:
 - **EMBL**
 - <http://www.ebi.ac.uk/embl/>
 - **GenBank**
 - <http://www.ncbi.nih.gov/Genbank/GenbankSearch.html>
 - **DDBJ**
 - <http://www.ddbj.nig.ac.jp>
 - Daily mutual exchange and backup of data
 - Works with large amount of data (capacity and software requirements)
 - September 2003 $27,2 \times 10^6$ entries (approx. 33×10^9 bp)
 - August 2005 100×10^9 bp from 165.000 organisms

Growth of GenBank



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

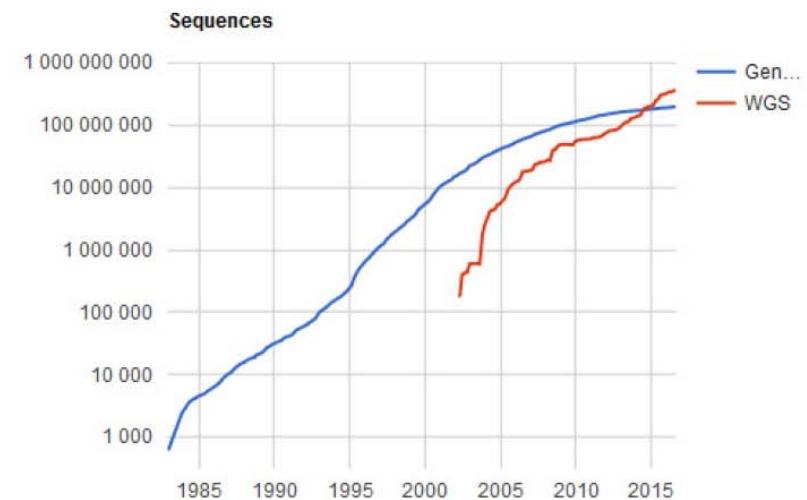
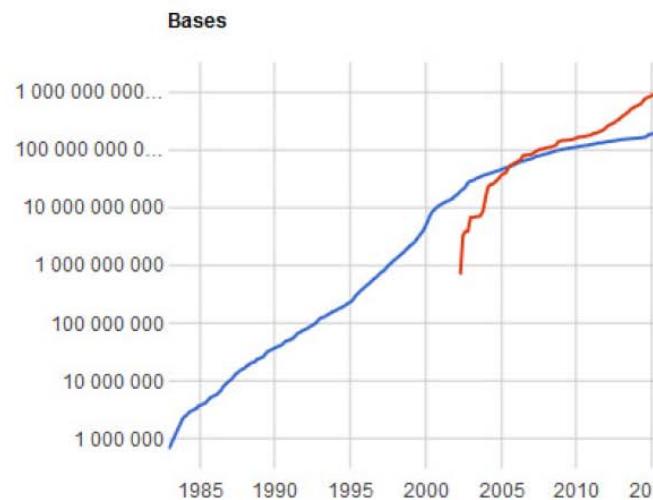
Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached **0.2 terabases**



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

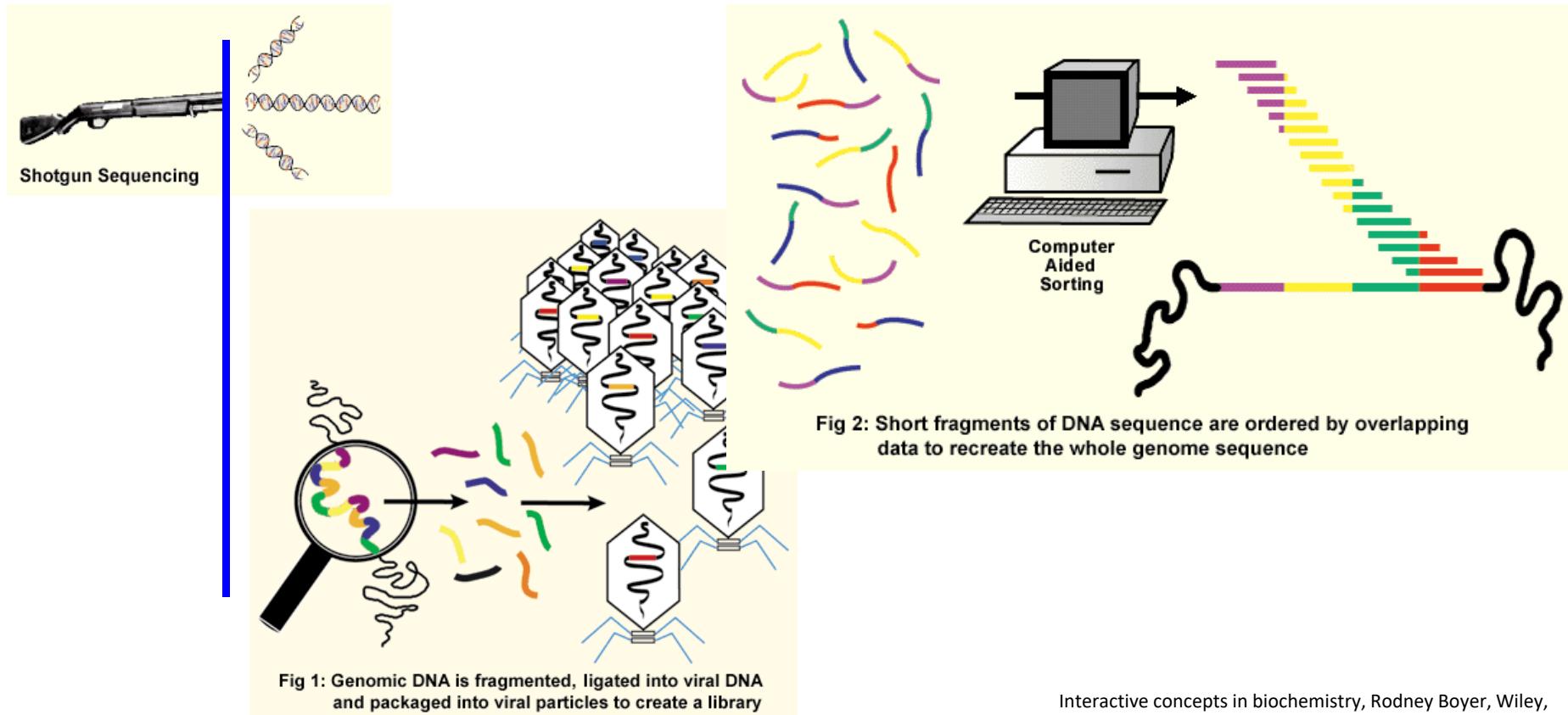
Growth of GenBank

Aug 2016



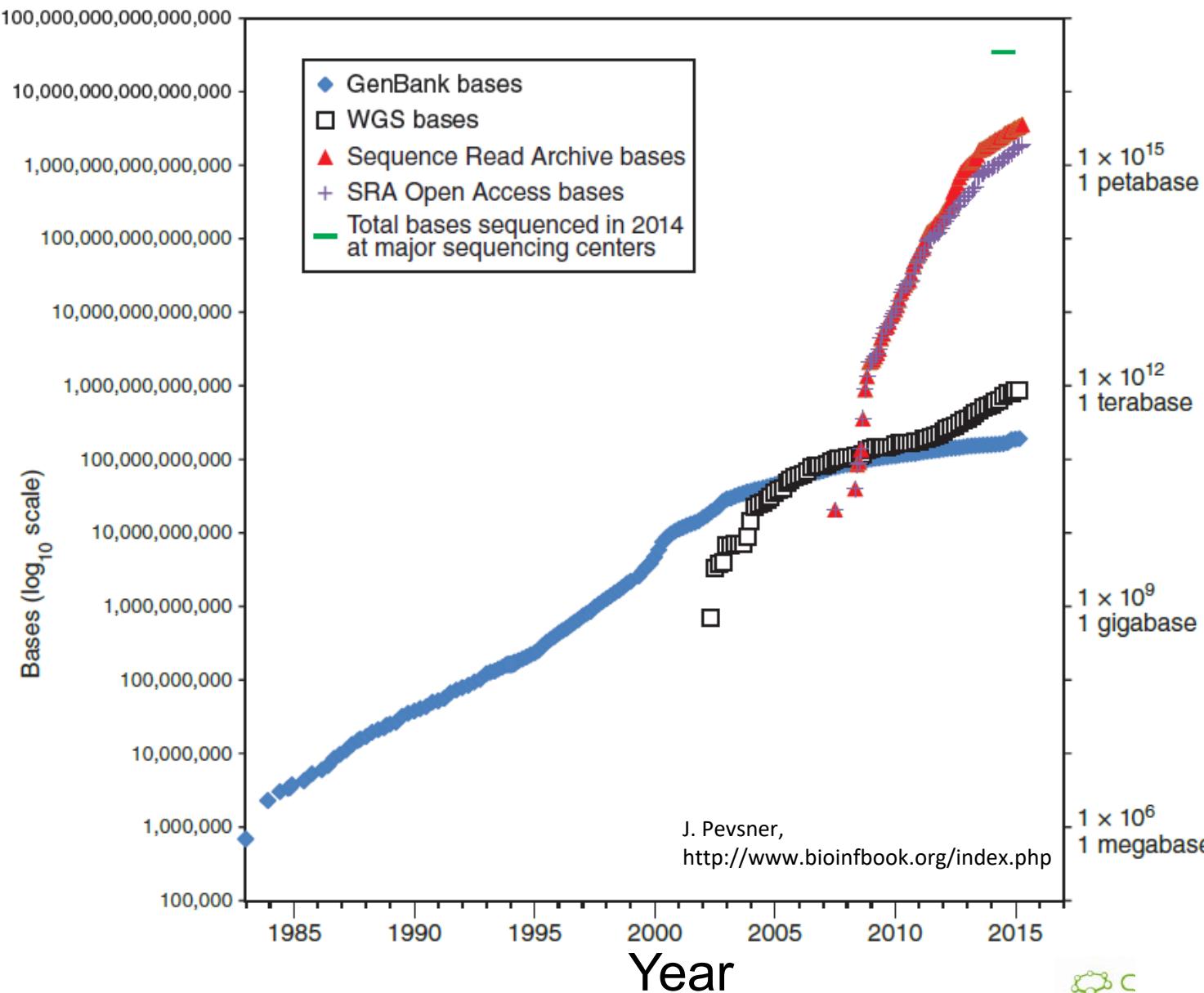
- Dec **1982** $680\,338$ bp, 606 sequences
- Apr **2002** 19×10^9 bp, 17×10^6 sequences + WGS 692×10^6 bp, $172\,768$ sequences
- Aug **2016** 218×10^9 bp, 196×10^6 sequences + WGS $1,6 \times 10^{12}$ bp, 360×10^6 sequences

WGS

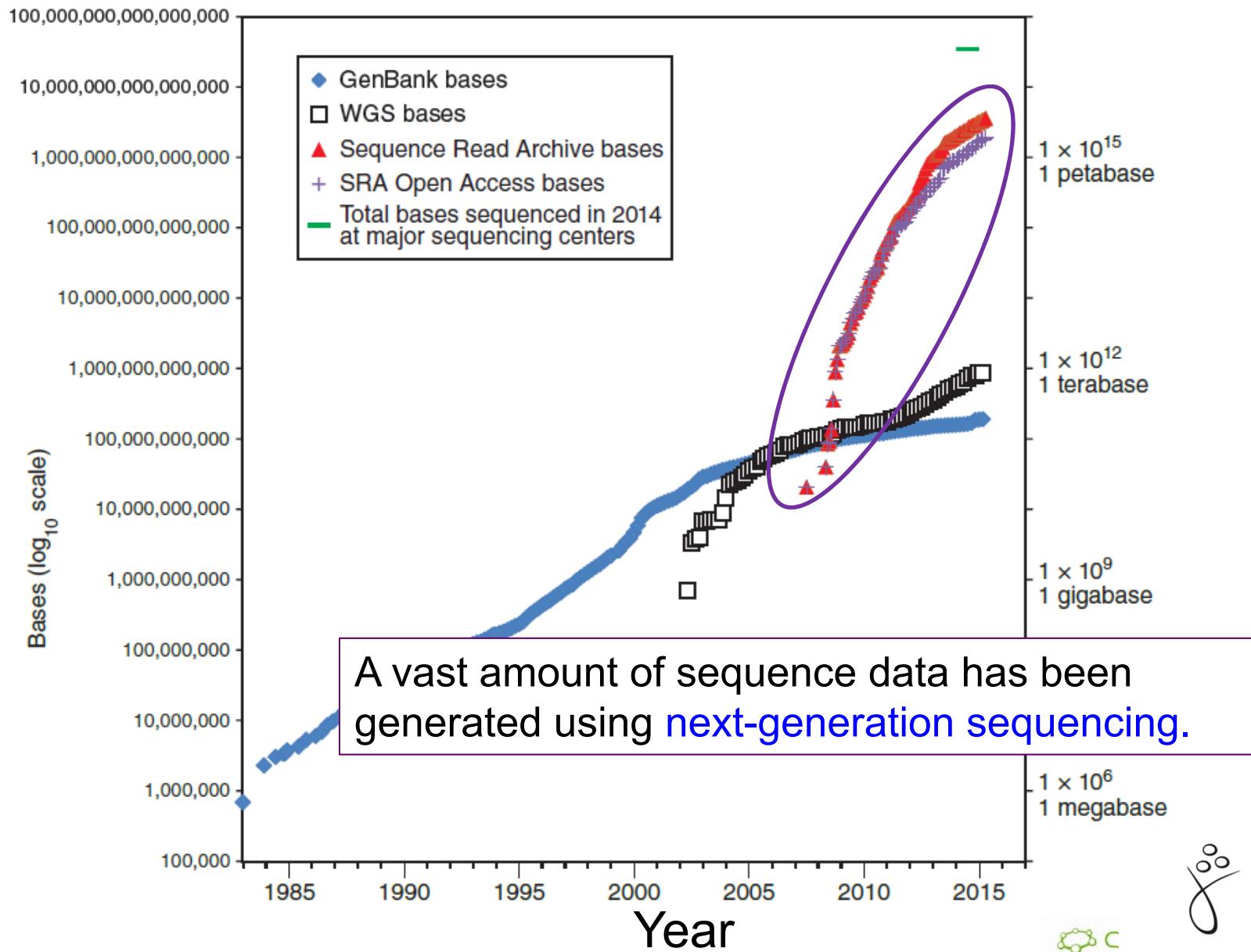


Interactive concepts in biochemistry, Rodney Boyer, Wiley, 2002, <http://www.wiley.com/college/boyer/0470003790/>

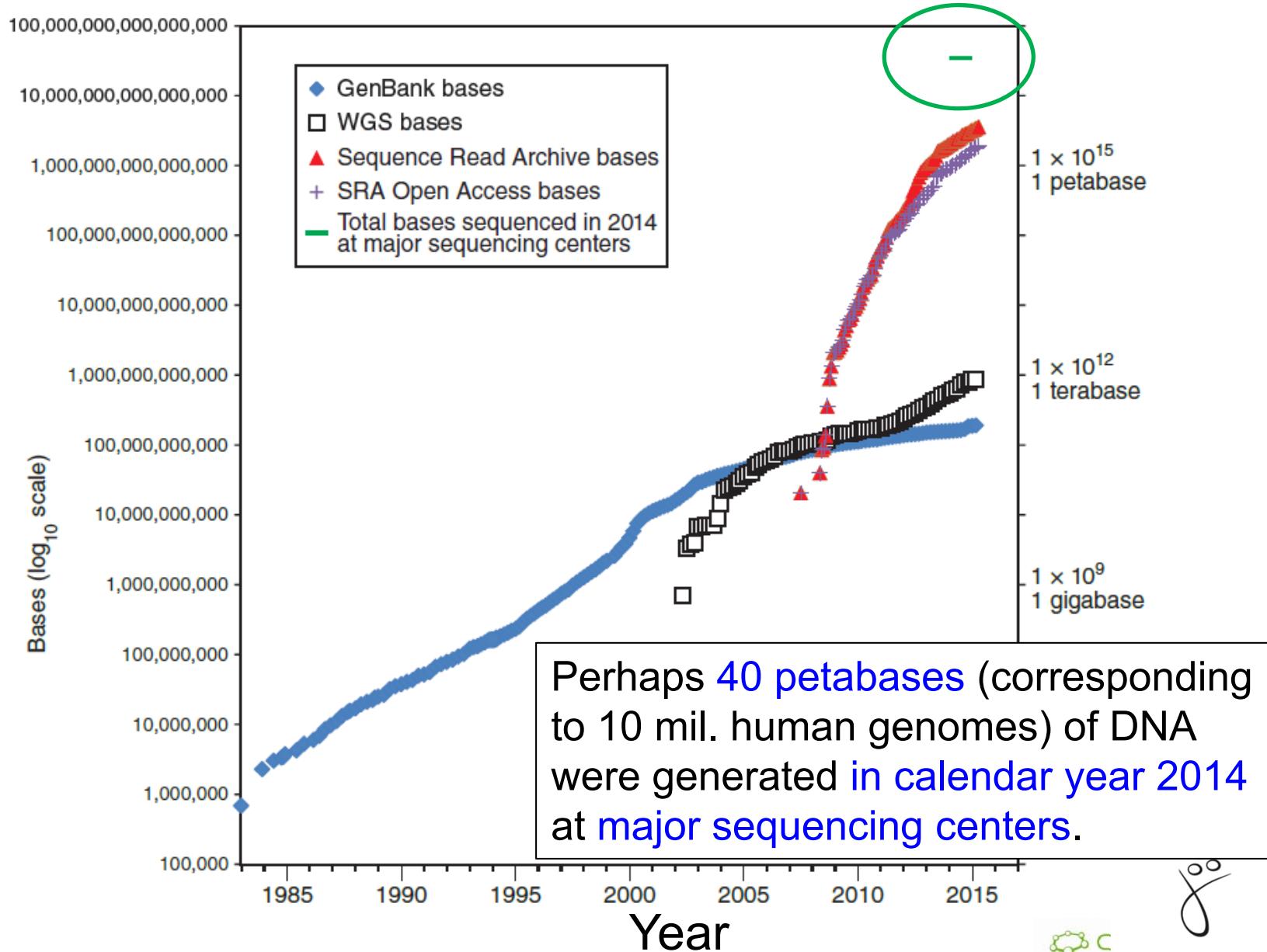
Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



Growth of DNA Sequence in Repositories



B&FG 3e

Fig. 2-3

Page 22



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 (Expressed Sequence Tags)
 - HGTS (High Throughput Genome Sequencing)
 - Results of sequencing projects without annotation
 - Reference Sequences of annotated genomes
 - TPA_s (Third Party Annotation)
 - sequences annotated by third party (by someone else, not the original authors)

Primary Databases

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

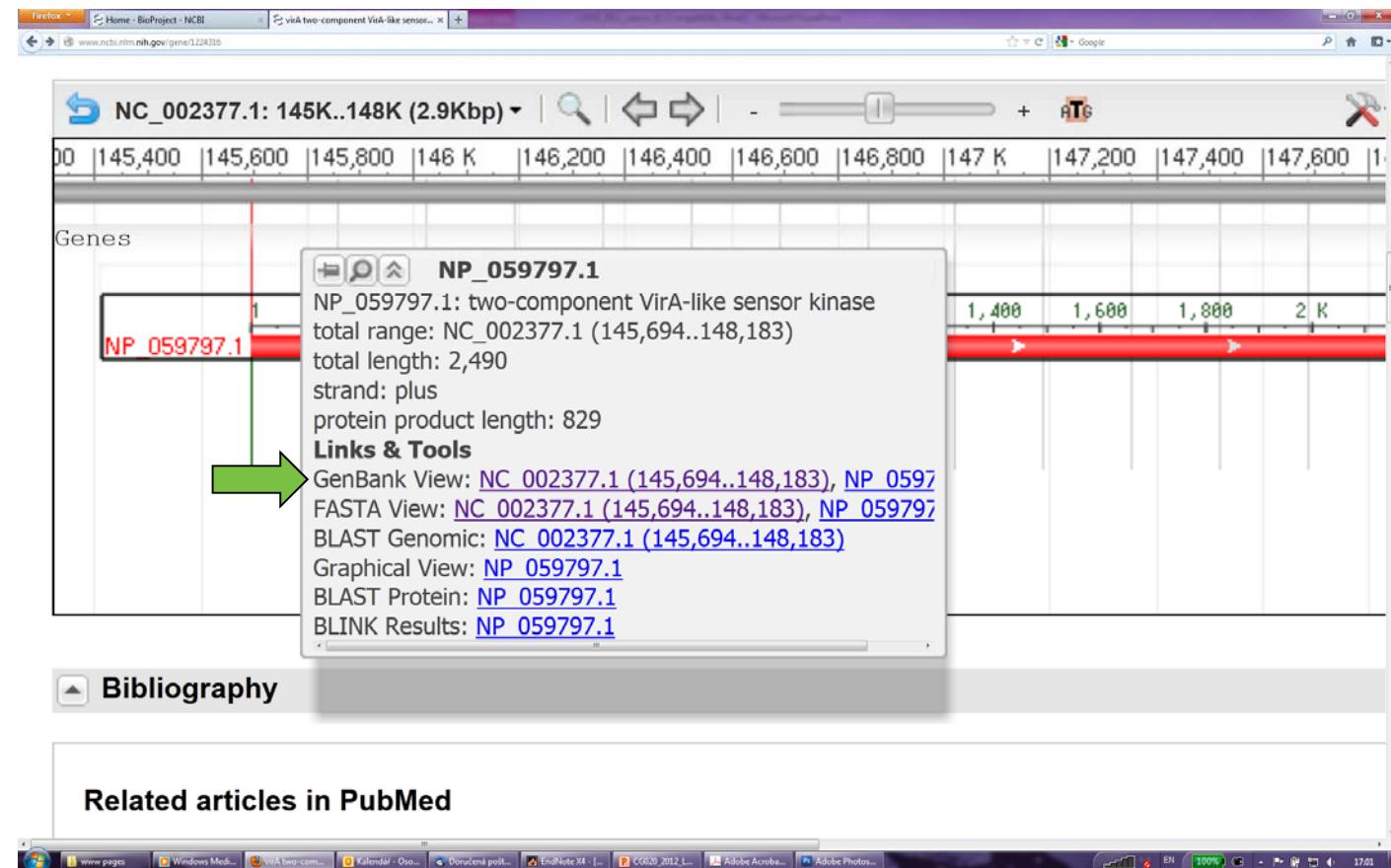
The screenshot shows the NCBI homepage. At the top, there's a blue header bar with the NCBI logo, a "Resources" dropdown, a "How To" dropdown, and "My NCBI Sign In" links. Below the header is the NCBI logo and the text "National Center for Biotechnology Information". A search bar with the placeholder "All Databases" and a "Search" button is positioned above the main content area. On the left, a sidebar titled "NCBI Home" lists various categories: 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, and Variation. The main content area features a "Welcome to NCBI" section with a brief description of the center's mission and links to About the NCBI, Mission, Organization, Research, and RSS Feeds. Below this is a "Get Started" section with a bulleted list of links: Tools, Downloads, How-To's, and Submissions. To the right, there's a "Popular Resource" sidebar listing PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. At the bottom, there's a "NCBI YouTube channel" section with a link to the channel and a "GO" button, followed by a navigation bar with links 1 through 8.

Primary Databases

The screenshot shows a Firefox browser window displaying the NCBI BioProject page for the gene *virA*. The URL is www.ncbi.nlm.nih.gov/gene/3224216. The page includes the following sections:

- Summary**:
 - Gene symbol: *virA*
 - Gene description: two-component VirA-like sensor kinase
 - Locus tag: pTi_125
 - Gene type: protein coding
 - RefSeq status: PROVISIONAL
 - Organism: *Agrobacterium tumefaciens* (old name: *Agrobacterium tumefaciens*; gb-synonym: *Rhizobium radiobacter*)
 - Lineage: Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiaceae; Rhizobium/Agrobacterium group; Agrobacterium; *Agrobacterium tumefaciens* complex
- Genomic context**: Shows the location of the gene on plasmid Ti, sequence NC_002377.1 (145694..148183).
- NC_002377.1**: A genomic map showing the *virA* gene (red arrow) and other genes (blue arrows) on a plasmid.
- Genomic regions, transcripts, and products**:
 - Genomic Sequence: NC_002377
 - Sequence View: Shows the nucleotide sequence from 145,400 to 148,400. A yellow circle highlights the *virA* gene region.
 - Links & Tools: Provides links to GenBank View, FASTA View, BLAST Genomic, Graphical View, BLAST Protein, BLINK Results, and UniProt.
- Bibliography**:
 - 1. Sequence analysis of the *virA* locus from *Agrobacterium tumefaciens* octopine Ti plasmid pTi15955. Schrammeijer B, et al. J Exp Bot. 2000 Jun. PMID: 10948245.
 - 2. The *virA* promoter is a host-range determinant in *Agrobacterium tumefaciens*. Turk SC, et al. Mol Microbiol. 1993 Mar. PMID: 8499115.
 - 3. Characterization of the *virA* locus of *Agrobacterium tumefaciens*: a transcriptional regulator and host range determinant. Leroux B, et al. EMBO J. 1987 Apr. PMID: 3595559.
 - 4. Analysis of the complete nucleotide sequence of the *Agrobacterium tumefaciens* *virB* operon. Thompson DV, et al. Nucleic Acids Res. 1988 May 25. PMID: 2837739.
- GeneRIFs: Gene References Into Functions**:
 - Submit: New GeneRIF, Correction
- Links**: Includes BioProjects, Conserved Domains, Full text in PMC, Genomes, Nucleotide, Protein, Protein Clusters, PubMed, RefSeq Proteins, and Taxonomy.
- General information**: About Gene, FAQ, FTP site, Help, My NCBI help, NCBI Handbook, Statistics.
- Related sites**: BLAST, Genome, BioProject, Genomic Biology, GEO, HomoloGene, Map Viewer, OMIM, Probe, RefSeq, UniGene, UniSTS.
- Feedback**: Contact Help Desk, Submit Correction, Submit GeneRIF.

Primary Databases



Primary Databases

NCBI

Search [Nucleotide] for [Go] [Clear]

Display: [List] [Get Subsequence] [Features]

1: NC_002377 [View details]

LOCUS NC_002377 2490 bp DNA linear BCT 29-DEC-2003

DEFINITION Agrobacterium tumefaciens extrachromosomal plasmid Ti, complete sequence.

ACCESSION NC_002377 REGION: 145694..148183

VERSION NC_002377.1 (GI:109551016)

KEYWORDS .

SOURCE Agrobacterium tumefaciens (Rhizobium radiobacter)

GeneBank Identifier

Farrand, S.K., Oger, P.M., Schrammeijer, B., Hooykaas, P.J. and Winans, S.C.

TITLE Octopine-type Ti plasmid sequence

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 2490)

AUTHORS Zhu, J., Oger, P.M., Schrammeijer, B., Hooykaas, P.J., Farrand, S.K. and Winans, S.C.

TITLE Direct Submission

JOURNAL Submitted (07-MAR-2000) Microbiology, Cornell University, Wing Hall, Ithaca, NY 14853, USA

COMMENT PROVISIONAL REFEREE: This record has not yet been subject to final NCBI review. The reference sequence was derived from AF242801.

FEATURES Location/Qualifiers

source 1..2490 /organism="Agrobacterium tumefaciens" /mol_type="genomic DNA" /db_xref="taxon:358" /plasmid="Ti" /note="extrachromosomal octopine-type"

gene 1..2490 /gene="virA" /db_xref="GeneID:1224316"

CDS 1..2490 /gene="virA" /note="two-component regulator of vir regulon; VirA is a transmembrane histidine kinase" /codon_start=1 /transl_table=11 /product="virA" /protein_id="NP_059797.1" /db_xref="GI:109551016"

Primary Databases

```
translation="MNGRYSPTRQDFKTGAKPWEILALIVAMIAFMAWASWQDNAT
TQAILEQLEISINADASLQRDVIRAHHTGVANTYPIISRLIGALKRNKLEDLQLPROSH
IVSEHESNAQILRQELESELNSADAIAVAAFGAQNVELQDLSLSPTRALSLPKASTDQT
LEKPTELASMMUQFLEQQSPAISFISLRLERLQKRGULDEAPVILAREBGPILSLL
PQVEDLVNNIQTSEDTEAIAABMLQRECLEVYSLNVERVERSARPLISASVYGLYIITL
VTRLEKTTWLAERLDYELIKEIIVCFEGEAATTSAQUALIIICOPFDADTCALAL
VDHDERWAVETPGAKHFKPVWDDSVLKEIVSERTKADEHATVPRVISSKVHLPLBIP
GLSILLAHKSTDKLIAVCVSLGYOSYRFPFCOGHQLLBLELATACLCHYIDVERRKETCSD
VLARELHQAQRLLEAVGTLAGIATHEFNWLGEILGHARLQANSVERTSVTERYIDVII
SSGRANMLIIDILTLRKQEMIKPPSVELVTBIAPILEMALPFIIRLSPRDMMQ
SVIENGSPLEIQLQVLINICKHAQCMANTANGQIDIIISQAPLEVVKRILAHQVMPGDYVL
LSIEDNNGQIIPBAVLPHIFFFFPTTARRNMGTGLGLASVHGHISAPAGYIDVSTVGH
GTRPDIVLPPSKEPVNNDSFVGRNKAAPHGNHEIVALVBPUDLREAYEDKIAALGYE
PUGFRTPNHIIRDWISKGHEADLWMDQASLPEDQSPNEVDLVLTASIIIGGNDLKMNT
LSREDVTREDVLYLPKPIISERTMABAIIITKIKT"
ORIGIN
  1 atgaaacggaa gatattcaac gacgcggcag gattttaaga caggcgccaa gccttggct
  61 atattggccc ttatgttgc tgaatgatt ttccggctca tgccgggttc gtctcgccag
121 gacaatgcga ctaccggcc aatccatcaga caactaegat cggataaegc cgacacggcc
181 tcaatcggcc gggatgttcc cggcgccatc acggggccacg tgccgcaacta cggcccatc
241 attcggccgg tggggatgtt cggggaaatg ttggaaatgtt tgaaagcaat atttagacaa
301 totatatttg taatggatg caatgttgtt ccgttgttcc gcaatgttca agtgcgttca
361 aatccggctg acggcgccgg cggccgttcc gttggccaaa atgcgcgtt ggaaatgtt
421 ctggccatgtt tcaatcgatg ttttggatgtt ccgttccggaa aacgttccaa cgtatcgact
481 tttagaaaaat caacaaatgtt ggttgcgttcc atgttccaaat tttttggccca accaaacccg
541 gttatccatc tccatgttcc ctttggatgtt ccgttccaaatggggatgttcc ttttgcgttcat
601 gaatgttccgg tggccatgtt tggccatgtt ccgttccatgtt ttttgcgttcc ttttgcgttcat
661 gttaaaatgtt tggttgcgttcc gtttgcgttcc ttttgcgttcc ttttgcgttcat
721 cggccggatgtt tttttggatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
781 cggccggatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
841 cggaaaaaaa cccatgttcc agcggccgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
901 ggatgttcc ttggaaatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
961 attcgcgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1021 tggccgttcc aaacattccgg tggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1081 cggaaatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1141 tcgaaaaaaa tccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1201 aaatccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1261 ccttgcggaa gggaaatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1321 gatgttccgg gtaacccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1381 cggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1441 ggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1501 cggatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1561 atccgttccgg tggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1621 gtgaccggaa tggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1681 agatttgcgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1741 attaacatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1801 atccgcggaa ctttttttcc agttaaaatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1861 gactatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1921 cacattttttcc aaccctttcc tacggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
1981 gttttttcc atggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
2041 gggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
2101 gacatgttcc tggccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
2161 gagccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
2221 cccgttccgg tttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
2281 gatccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt
2341 ttatgttccaa agatccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt ccgttccatgtt"
```

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

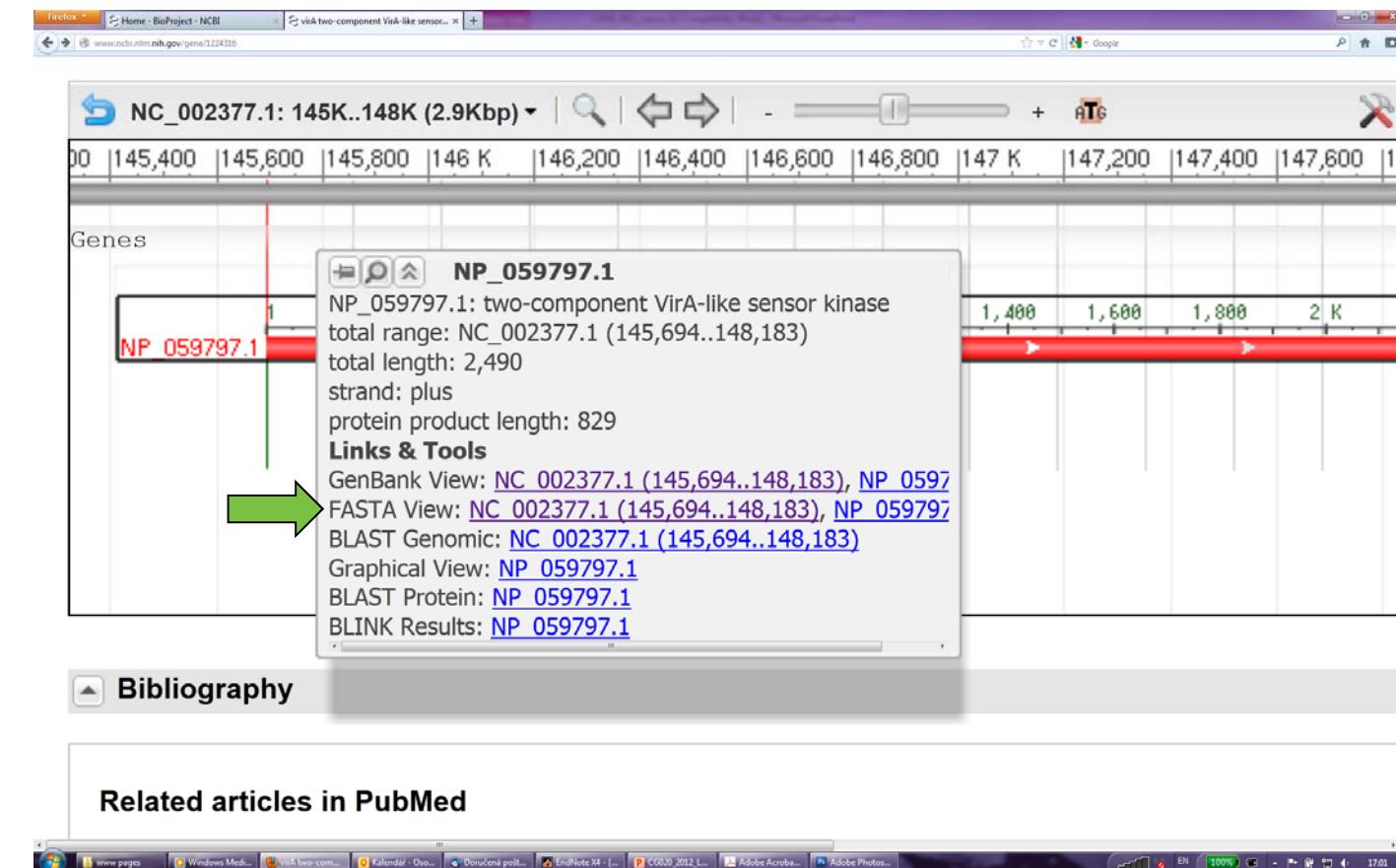
The screenshot shows a Firefox browser window displaying the NCBI Reference Sequences (RefSeq) page for the gene NP_396486.1. The URL in the address bar is www.ncbi.nlm.nih.gov/gene/1137489. The page title is "two-component VirA-like sensor kinase". A yellow circle highlights the "NCBI Reference Sequences (RefSeq)" section in the header. Below it, the "Genome Annotation" section is visible, containing a note about reference sequences belonging to a specific genome build and a link to "Explain". The "Reference assembly" section shows the "Genomic" assembly with entry NC_003065.3, spanning from position 180831..183332, with links to "GenBank", "FASTA", and "Sequence Viewer (Graphics)". The "mRNA and Protein(s)" section lists the protein NP_396486.1, which is a two-component sensor kinase from Agrobacterium tumefaciens str. C58. It includes UniProtKB/Swiss-Prot entry P18540 and a summary of conserved domains. The domain cd00075 is described as HATPase_c: Histidine kinase-like ATPases, located at positions 580–694 with a Blast Score of 202. The domain cd00082 is described as HisKA: Histidine Kinase A (dimerization/phosphoacceptor) domain; Histidine Kinase, located at positions 466–530 with a Blast Score of 144. The domain PRK13837 is described as PRK13837: two-component VirA-like sensor kinase; Provisional, located at positions 14–833 with a Blast Score of 2944. At the bottom, the "Related Sequences" section is partially visible. The taskbar at the bottom of the screen shows various open applications.

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



Primary Databases

The screenshot shows a web browser window with the URL www.ncbi.nlm.nih.gov/nuccore/NC_002377.1?report=fasta&from=145694&to=148183. The page displays the complete sequence of *Agrobacterium tumefaciens* plasmid Ti, reference sequence NC_002377.1. The sequence is presented in FASTA format, showing a 2.49 kb region from base 145694 to 148183. The browser interface includes various tools and links for sequence analysis, such as BLAST, Primers, and GenBank.

Secondary Databases

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

[ExPASy Home page](#) [Site Map](#) [Search ExPASy](#) [Contact us](#) [Swiss-Prot](#) [PROSITE](#) [Proteomics tools](#)
Hosted by SIB Switzerland | Mirror sites: [Australia](#) [Bolivia](#) [Canada](#) [China](#) [Korea](#) [Taiwan](#) [USA](#)

Search PROSITE for

proSite ScanProsite

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

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

Scan a protein for PROSITE matches
Enter a Swiss-Prot/TrEMBL accession number (AC) (for example P05130) or a sequence identifier (ID) (for example NOTC_DROME), or a PDB identifier, or paste your own protein sequence in the box below:
`MSVKVTKLVALEPIVIVPCVLEPDPVPTPECMISNGPPTR
MVKVTKLVALEPIVIVPCVLEPDPVPTPECMISNGPPTR
EVQDITPEPTEPPTGTTTAPLLPVAASTLQVSGVVE
LSRDQIMPEVYIAHNTGUNVANSSNSNSRSDYTVKWTOTV
DQLTGELNLGNSTSKSQSLDVHTHTWPAQSNNYTTAIPVGT
SLGEHDNTTICGSVELYLKEGLVSLQDFPVKLTIFVLNLSL
MLHEELIYMWTIDGFTVLRERBESLMDSFPIFLNSICFGREES
MSLWSQCIPENCSSEGGVEVIEKLRLYQAFCSVIRVEGVPL`

and specify which motifs to use:
Scan patterns profiles rules [User Manual] (You may also specify a PROSITE entry in the box to the right)
 Exclude patterns with a high probability of occurrence

Your e-mail (optional): (will send results by e-mail)
 plain text output

Search Swiss-Prot with a PROSITE entry
Enter a PROSITE accession number (for example PS01253), or type your pattern in [PROSITE format](#):
(leave this box blank to scan a sequence with the entire PROSITE database)

and specify your search limits:

- The Swiss-Prot TrEMBL TrEMBLnew PDB databases
(You may also specify a protein in the box to the left)
 including splice variants
- The following taxa:
(see [NEW TAXONOMY](#); separate multiple taxa with a semicolon, e.g. *Homo sapiens; Drosophila*. Not available for PDBs.)
- Sequences with at least hits
- At most 1000 matches

Advanced options: FASTA output retrieve complete sequences
allow at most X sequence characters to match a conserved position in the pattern
 match mode greedy, overlaps, no includes (for patterns, see [help](#))
 randomize databases (to test a pattern, see [help](#))

Secondary Databases

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

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

571 - 585 nkeesstYeteians

>[PDOC00004 PS00004](#) **CAMP_PHOSPHO_SITE** cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

744 - 747 RrvT
814 - 817 KRrS

>[PDOC00005 PS00005](#) **PKC_PHOSPHO_SITE** Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

148 - 150 SsR
164 - 166 TgR
171 - 173 SsK
219 - 221 SKK
369 - 371 TrR
460 - 462 SqK
511 - 515 SqR
585 - 587 SsR
602 - 604 TgK
652 - 654 TdK
716 - 718 SpR
726 - 728 SpK
747 - 749 TeK
794 - 796 SeR
854 - 856 ScK
864 - 866 SrR
868 - 870 SeR
921 - 923 SpK
957 - 959 SvR
960 - 962 TgR
974 - 976 TaK
997 - 999 SsK
1002 - 1004 TgK
1018 - 1020 SqK
1031 - 1033 TgR
1119 - 1121 SkR

Secondary Databases

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

```
>PDOC50100 PS50109 HIS_KIN Histidine kinase domain [profile].  
402 - 671 NASHDIRGALAGMEGLIDICRDGVKPGSDVDTTLNQVNVAEAKDLVALLNSVLIMSKIRSG  
KQQLVEEDPNLSKLLRDVLDPYHPVAMKKGVNVLDHDgavEKPSNTRGDSGRLKQILN  
NLVSNAVKPTVD--GHIAVRAMAQrpgeasvvlasypkgvafkvamfccknkeesatye  
teiensirnnantTMEPVPEVIDTGKGIPMEMRKSVFPAVTVQVRBtAQGHQGTGLGLGIVQ  
SLVRLMGGIEIRITDKAMdekgGTCPOFWLITT
```

```
>PDOC50110 PS50110 RESPONSE_REGULATORY Response regulatory domain [profile].  
987 - 1085 RVLVVDDNPISRKVATGKLKKMGVSeVEQCDSGKRALRLVTBGLtqreeggsvdklpFDY  
IPMDQQMPEMDGYBATREIRkvekSYGVRTPIIAVSGHD-----
```

Graphical summary of hits (*java applet*)



98 hits with 12 PROSITE entries

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 compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a SWISS-PROT/TrEMBL composite. 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. [References](#)

New:

- [PRINTS - Search PRINTS-S \(relational PRINTS\)](#)
- [prePRINTS - Search PRINTS' automatic supplement](#)
- [InterPro - Search the integrated InterPro family database](#)

Direct PRINTS access:

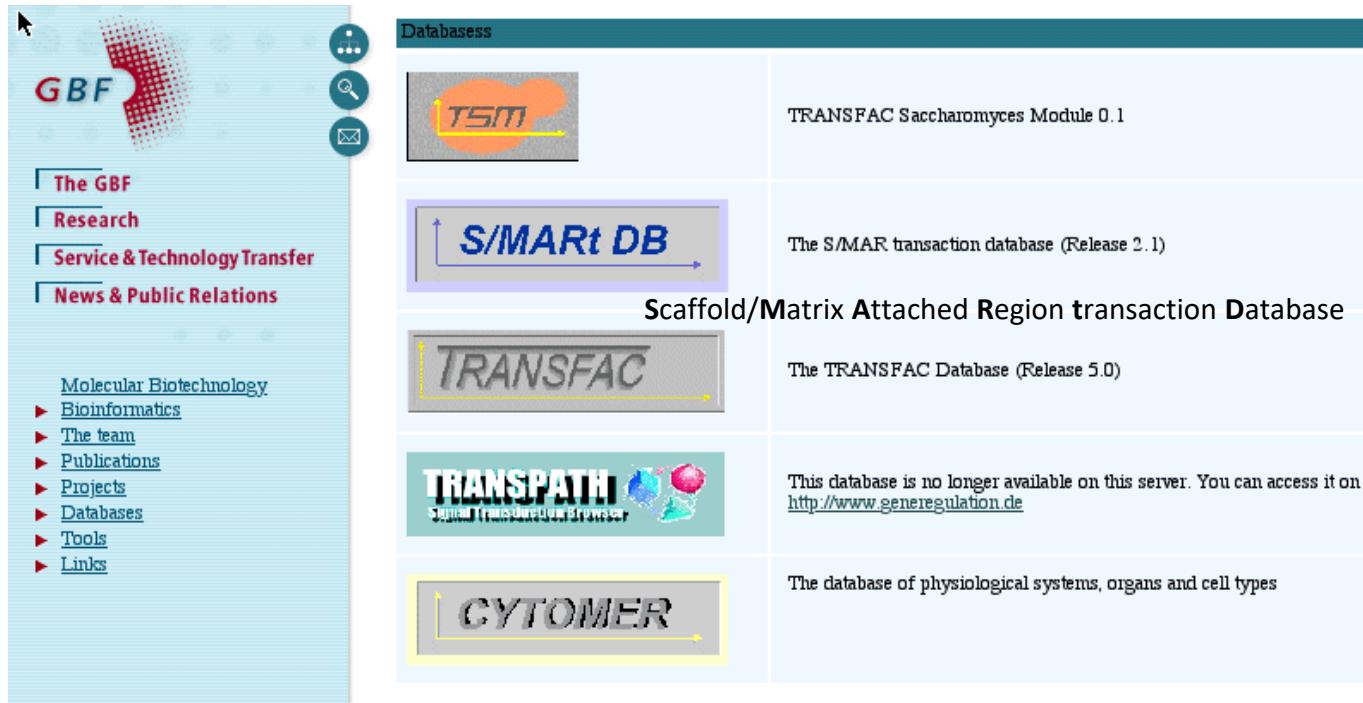
- [By accession number](#)
- [By PRINTS code](#)
- [By database code](#)
- [By text](#)
- [By sequence](#)
- [By title](#)
- [By number of motifs](#)
- [By author](#)
- [By query language](#)

PRINTS search:

- [Search PRINTS with NEW FingerPRINTScan](#)
- [FPScan](#)
- [GRAPHScan](#)
- [MULScan](#)
- [FingerPRINTScan binaries and source are available: contact scordis@bioinf.man.ac.uk](#)

Secondary Databases

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



The screenshot shows the homepage of the German Biotechnology Federation (GBF) with a sidebar and a main content area. The sidebar includes links for 'The GBF', 'Research', 'Service & Technology Transfer', 'News & Public Relations', 'Molecular Biotechnology' (with sub-links for Bioinformatics, The team, Publications, Projects, Databases, Tools, and Links), and a search bar.

The main content area features a 'Databases' section with five entries:

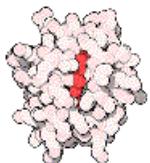
Databases	Description
	TRANSFAC Saccharomyces Module 0.1
	The S/MAR transaction database (Release 2.1)
	Scaffold/Matrix Attached Region transaction Database
	The TRANSFAC Database (Release 5.0)
	This database is no longer available on this server. You can access it on http://www.generegulation.de
	The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

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

P R O T E I N D A T A B A N K

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](#)

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

[Did you find what you wanted?](#)

Search the Archive [?](#)
Enter a [PDB ID](#) or keyword [Query Tutorial](#)
 [Find a structure](#)
 query by PDB id only match exact word
 remove sequence homologues

SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

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

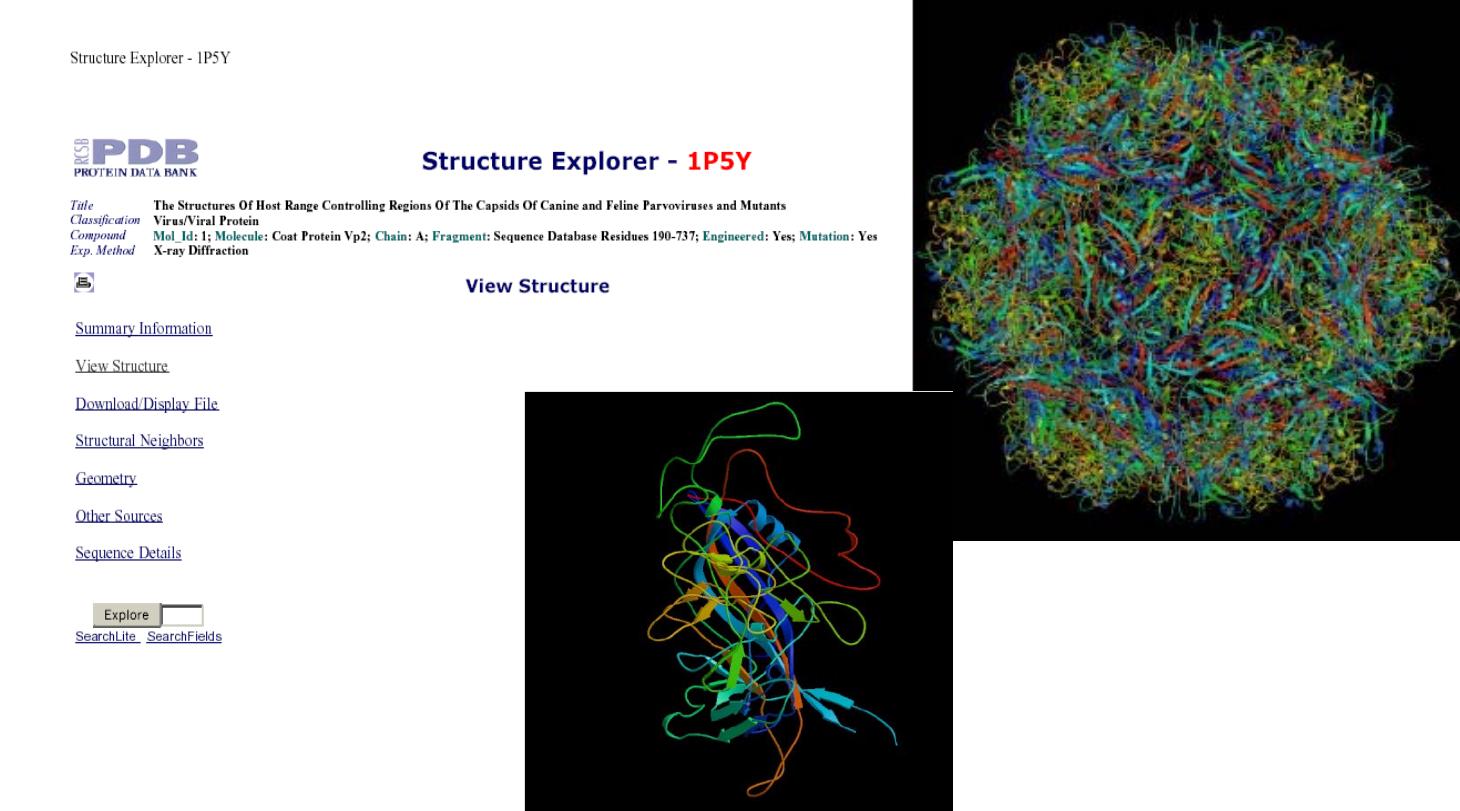
23-Dec-2002
Happy Holidays from the PDB! The PDB staff wish to extend our [best wishes](#) to the community for a happy holiday season and a wonderful new year! 

PDB Mirrors
Please bookmark a mirror site
[San Diego Supercomputer Center*](#)
[Rutgers University*](#)
[National Institute of Standards and Technology*](#)
[Cambridge Crystallographic Data Centre, UK](#)
[National University of Singapore](#)
[Osaka University, Japan](#)
[Universidade Federal de Minas Gerais, Brazil](#)
[Max Delbrück Center for Molecular Medicine, Germany](#)

OTHER SITES

Structural Databases

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

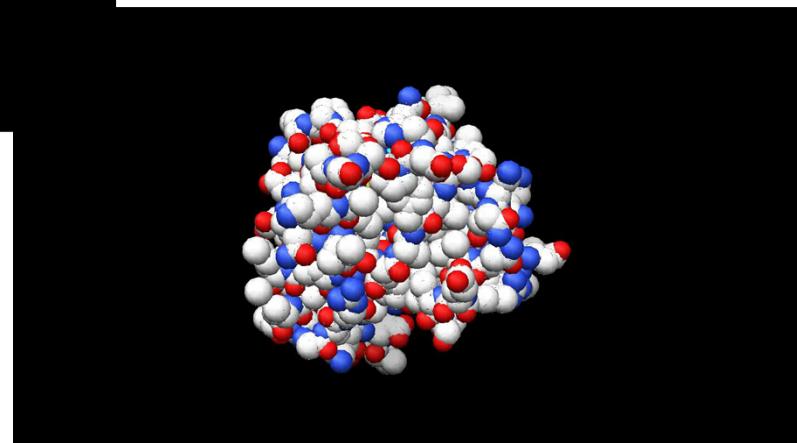
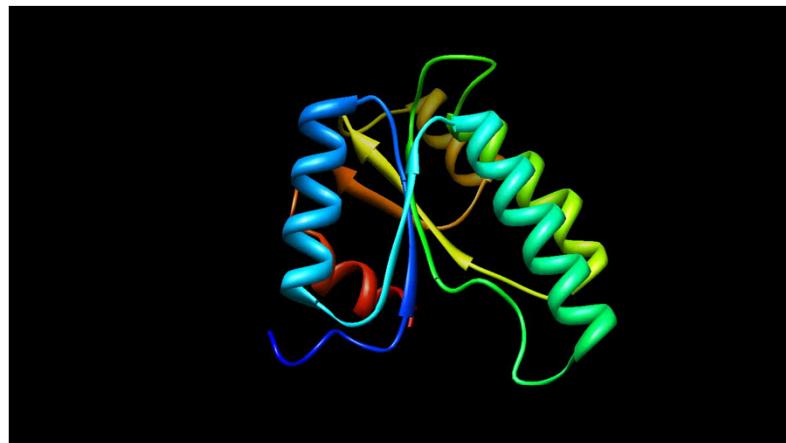


<http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics; pdbId=1P5Y; page=; pid=173561064349344&bio=1&opt=show&size=500>

12/29/2003

Structural Databases

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



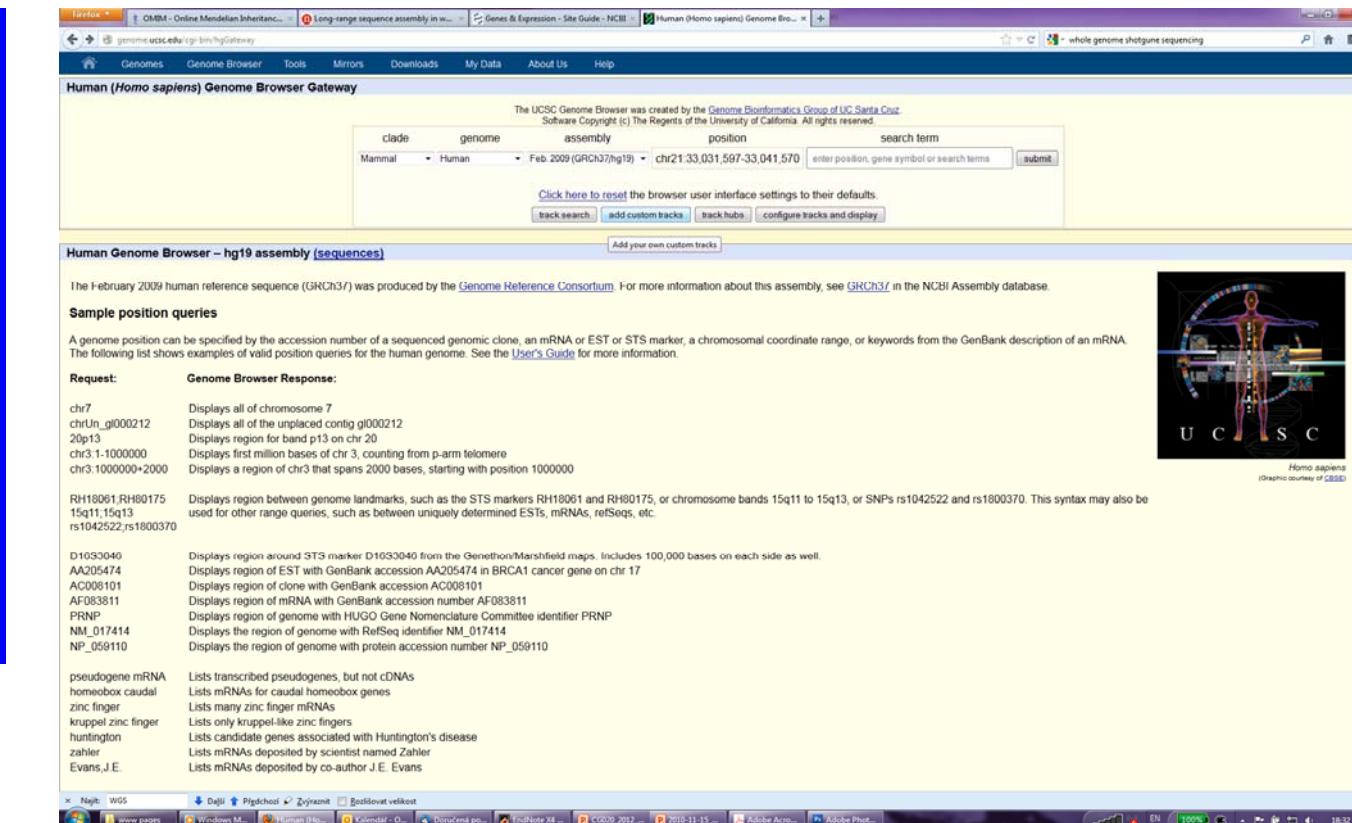
Pekárová et al., *Plant Journal* (2011)

Outline

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

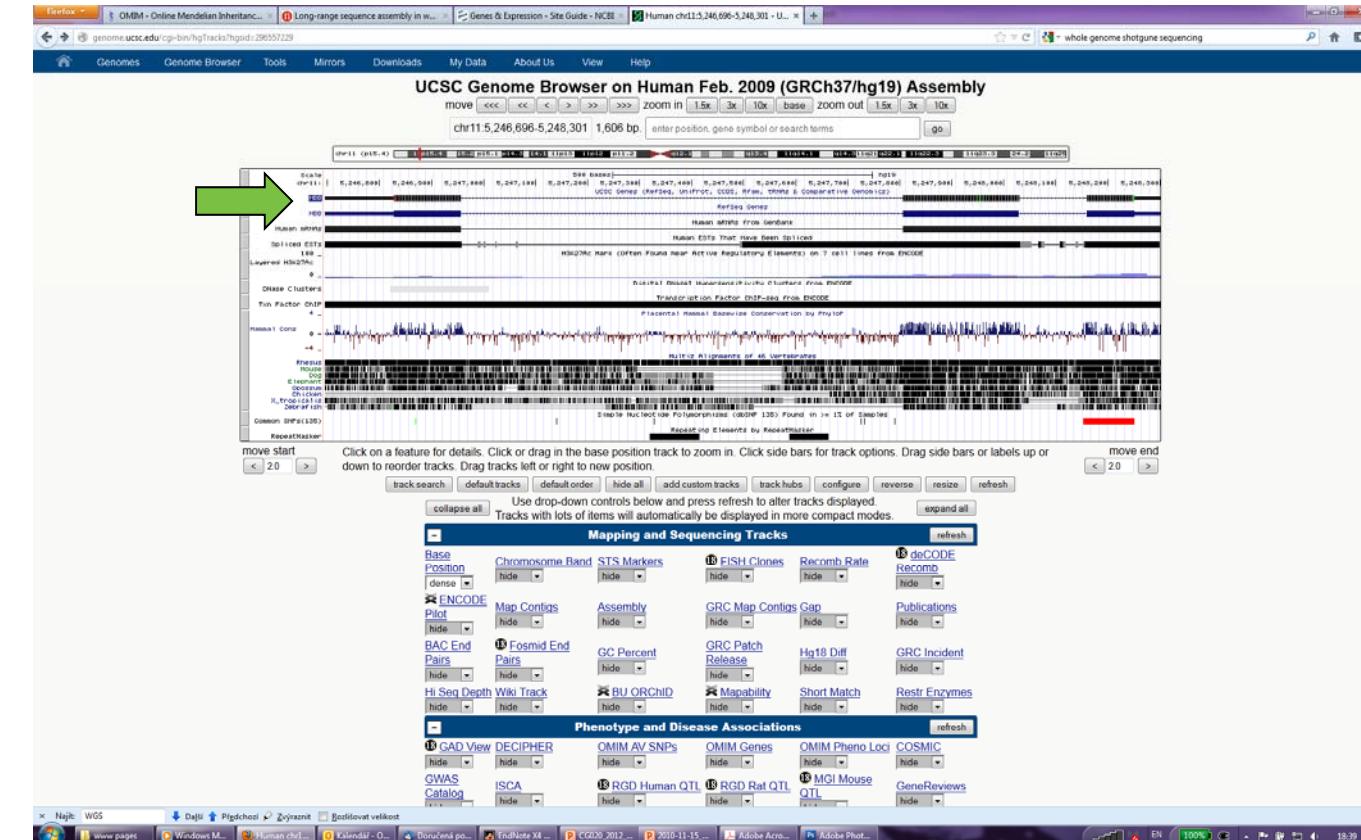
Genome Resources

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



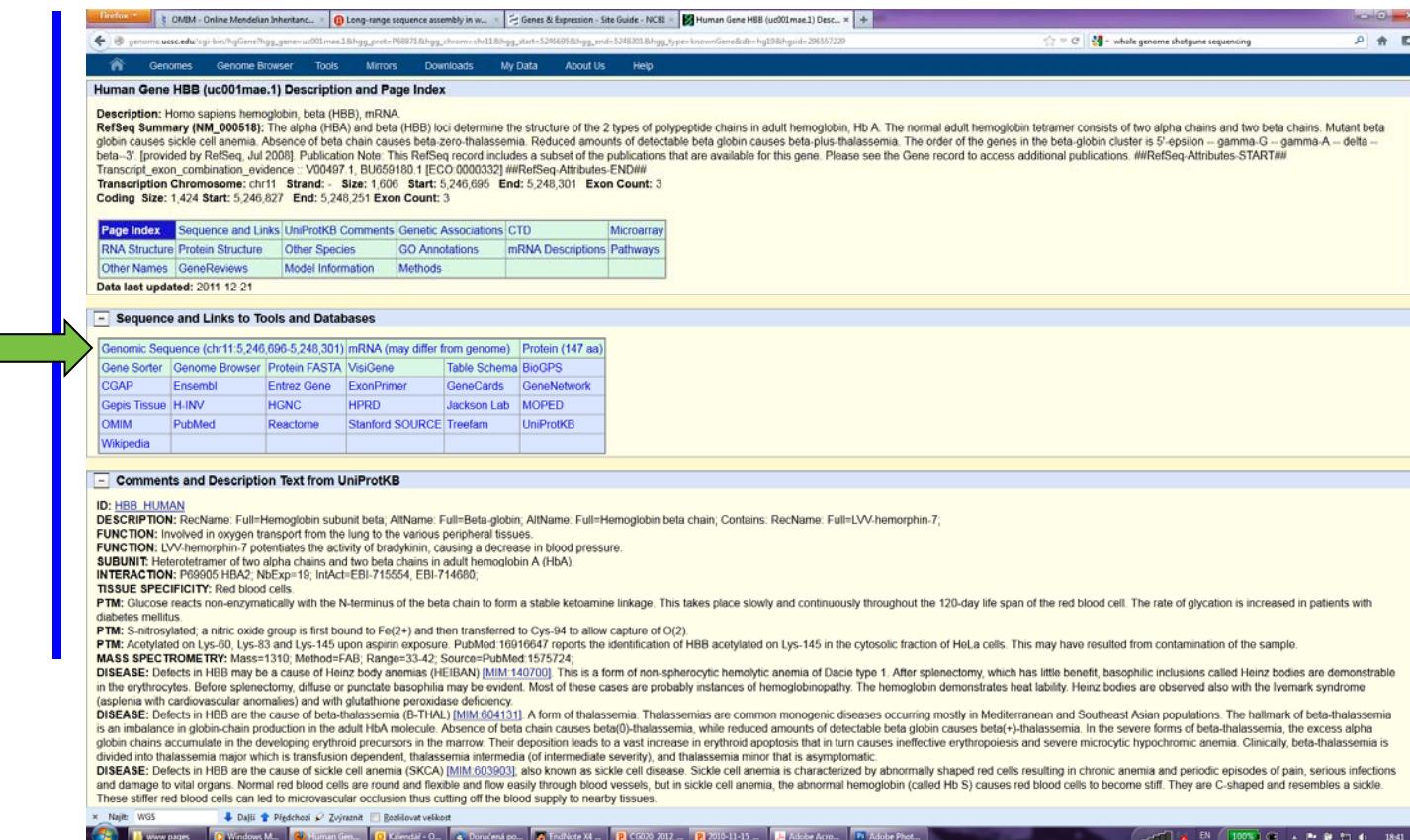
Genome Resources

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



Genome Resources

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



The screenshot shows the UCSC Genome Browser interface for the HBB gene. The main content area displays the gene's description, RefSeq summary, and various links to databases like UniProtKB, GenBank, and Ensembl. A green arrow points to the 'Sequence and Links' section, which contains links to genomic sequence, protein structure, and other databases.

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

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

RefSeq Summary (NM_000518): The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -> gamma-G -> gamma-A -> delta -> beta-3' [provided by RefSeq, Jul 2008]. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications. ##RefSeq-Attributes-START##

Transcript_Chrosome: chr11 Strand: - Size: 1,606 Start: 5,246,696 End: 5,248,301 Exon Count: 3

Coding Size: 1,424 Start: 5,246,627 End: 5,246,251 Exon Count: 3

Data last updated: 2011-12-21

Sequence and Links to Tools and Databases

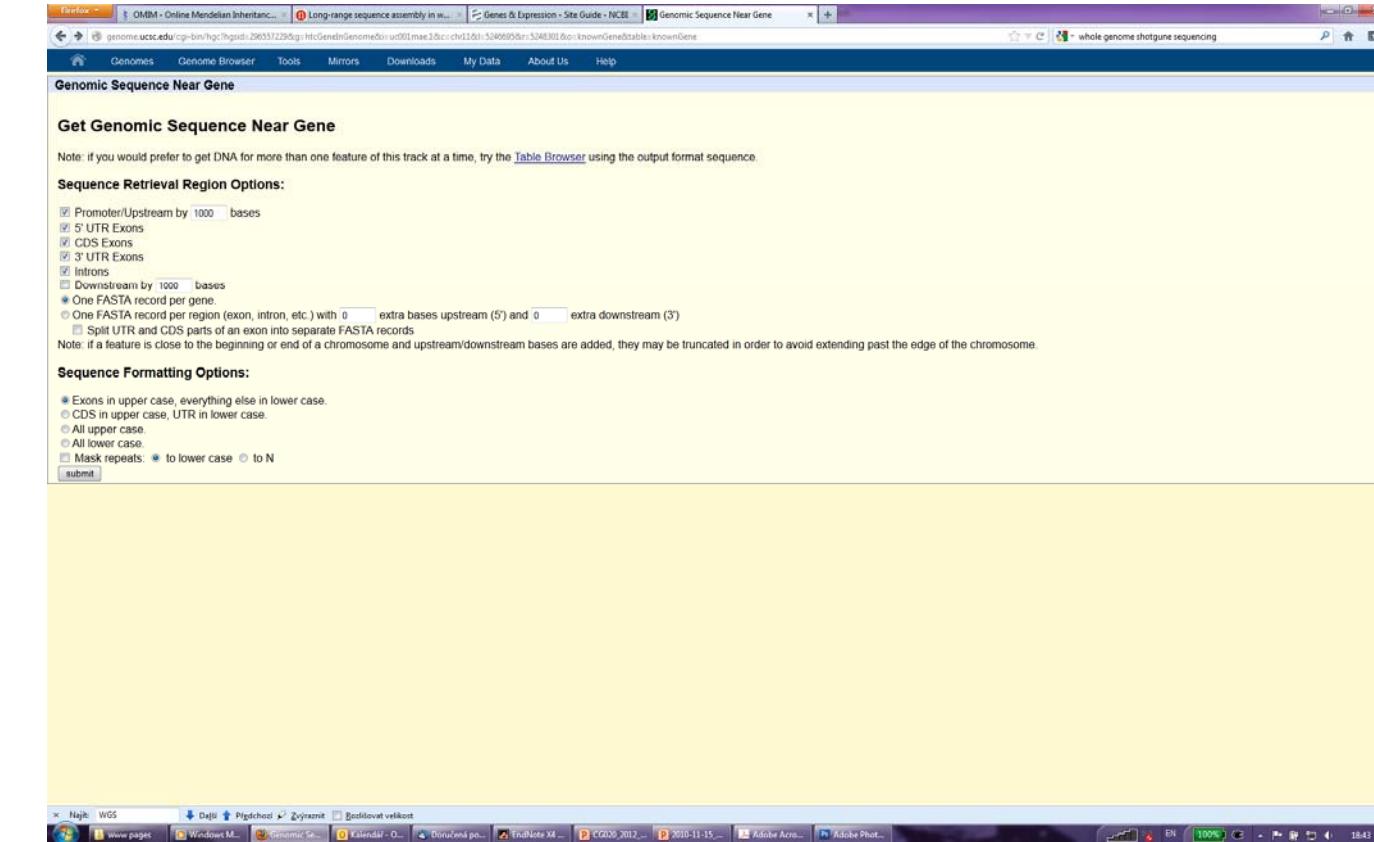
Genomic Sequence (chr11:5,246,696-5,248,301)	mRNA (may differ from genome)	Protein (147 aa)
Gene Sorter	Genome Browser	Protein FASTA
CGAP	Ensembl	Entrez Gene
Geps Tissue	H-INV	HGNC
OMIM	PubMed	Reactome
Wikipedia		

Comments and Description Text from UniProtKB

ID: HBB_HUMAN
DESCRIPTION: RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain; Contains: RecName: Full=LVV-hemorphin-7;
FUNCTION: Involved in oxygen transport from the lung to the various peripheral tissues.
FUNCTION: LVV-hemorphin-7 potentiates the activity of bradykinin, causing a decrease in blood pressure.
SUBUNIT: Heterotrimer of two alpha chains and two beta chains in adult hemoglobin A (HbA).
INTERACTION: P06905 HBA2; NbExp=19; IntAct:EBI-715554; EBI-714680;
TISSUE SPECIFICITY: Red blood cells.
PTM: Glucose reacts non-enzymatically with the N-terminus of the beta chain to form a stable ketoamine linkage. This takes place slowly and continuously throughout the 120-day life span of the red blood cell. The rate of glycation is increased in patients with diabetes mellitus.
PTM: S-nitrosylated; a nitric oxide group is first bound to Fe(2+) and then transferred to Cys-94 to allow capture of O(2).
PTM: Acetylated on Lys-60, Lys-83 and Lys-145 upon aspinin exposure. PubMed:16916647 reports the identification of HBB acetylated on Lys-145 in the cytosolic fraction of HeLa cells. This may have resulted from contamination of the sample.
MASS SPECTROMETRY: Mass=1310; Method=FAB; Range=33-42; Source=PubMed:1575724;
DISEASE: Defects in HBB may be a cause of Heinz body anemias (HEIBAN) [MIM:140700]. This is a form of non-spherocytic hemolytic anemia of Dacie type 1. After splenectomy, which has little benefit, basophilic inclusions called Heinz bodies are demonstrable in the erythrocytes. Before splenectomy, diffuse or punctate basophilia may be evident. Most of these cases are probably instances of hemoglobinopathy. The hemoglobin demonstrates heat lability. Heinz bodies are observed also with the Ivemark syndrome (asplenia with cardiovascular anomalies) and with glutathione peroxidase deficiency.
DISEASE: Defects in HBB are the cause of beta-thalassemia (B-THAL) [MIM:604131]. A form of thalassemias. Thalassemias are common monogenic diseases occurring mostly in Mediterranean and Southeast Asian populations. The hallmark of beta-thalassemia is an imbalance in globin-chain production in the adult HbA molecule. Absence of beta chain causes beta(0)-thalassemia, while reduced amounts of detectable beta globin causes beta(+)-thalassemia. In the severe forms of beta-thalassemia, the excess alpha globin chains accumulate in the developing erythroid precursors in the marrow. Their deposition leads to a vast increase in erythroid apoptosis that in turn causes ineffective erythropoiesis and severe microcytic hypochromic anemia. Clinically, beta-thalassemia is divided into thalassemia major which is transfusion dependent, thalassemia intermedia (of intermediate severity), and thalassemia minor that is asymptomatic.
DISEASE: Defects in HBB are the cause of sickle cell anemia (SKCA) [MIM:603903], also known as sickle cell disease. Sickling cell anemia is characterized by abnormally shaped red cells resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resembles a sickle. These stiffer red blood cells can lead to microvascular occlusion thus cutting off the blood supply to nearby tissues.

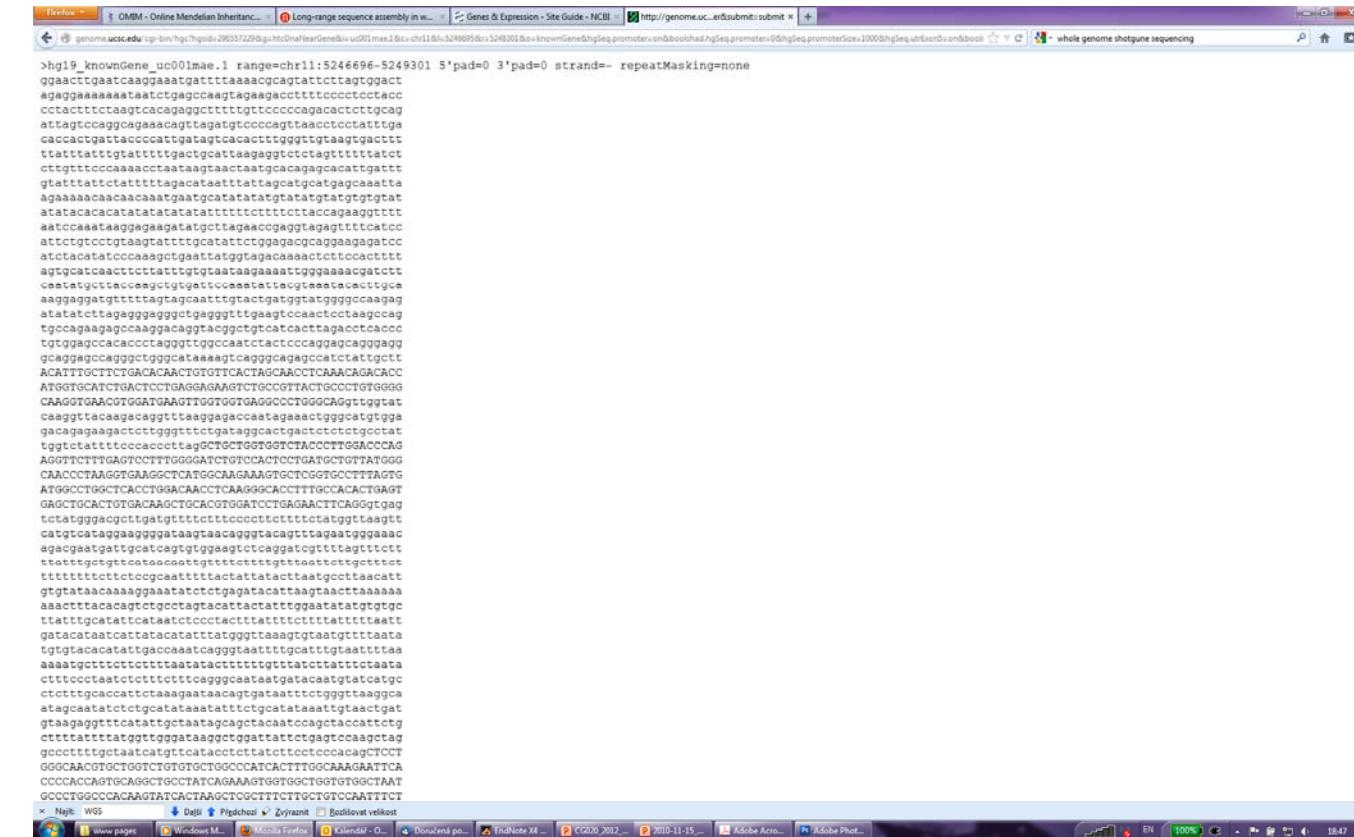
Genome Resources

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



Genome Resources

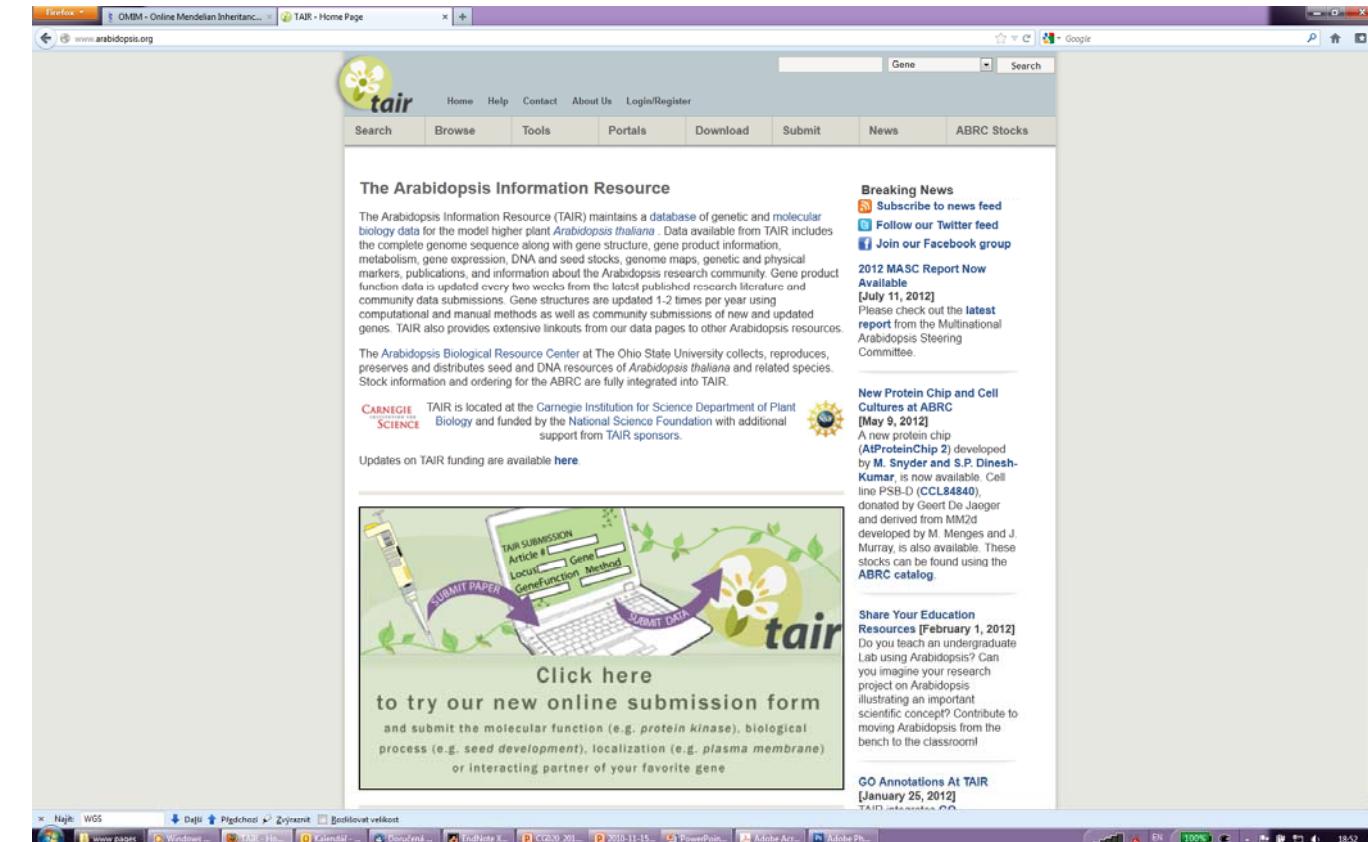
- Human Genome Browser <http://genome.ucsc.edu/cgi-bin/hgGateway>



The screenshot shows a web browser window displaying a DNA sequence from the UCSC Genome Browser. The sequence is presented in a monospaced font, showing a long stretch of nucleotide bases. The browser interface includes a top navigation bar with tabs for 'Home', 'OMIM - Online Mendelian Inheritance in Man', 'Long range sequence assembly in ...', 'Genes & Expression - Site Guide - NCBI', 'http://genome.uc...erdssubmit: submit', and 'whole genome shotgun sequencing'. Below the tabs, there are several toolbars and status bars. The main content area contains the DNA sequence, which is extremely long and repetitive, consisting mostly of 'A' and 'T' nucleotides.

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 main navigation bar of the TAIR website. The top bar includes links for Home, Help, Contact, About Us, Login, and a search bar labeled "Gene". Below the main menu, there are several tabs: Search, Browse, Tools, Stocks, Portals, Download, Submit, and News. The "Portals" tab is currently selected. A red circle highlights the "AHP2" search bar, which is part of the "Gene" dropdown menu. The main content area features a section titled "The Arabidopsis Information Resource" with a brief description of the database's purpose and a note about data updates being suspended. Below this is another section titled "The NEW arabidopsis.org" with a message about the site's reorganization. On the right side, there is a "Breaking News" sidebar with three entries: "Data Updates Suspended" [October 19, 2006], "New Phenotype Search Option" [October 15, 2006], and "ASPB Presentations" [August 15, 2006].

Outline

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

Analytical Tools

□ Global versus Local alignment

The diagram illustrates two types of sequence alignment:

Globální přiřazení
SLAV-----APATNIK-----PIQNYR-I-----AKSETQRYSMIE
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYSMIE

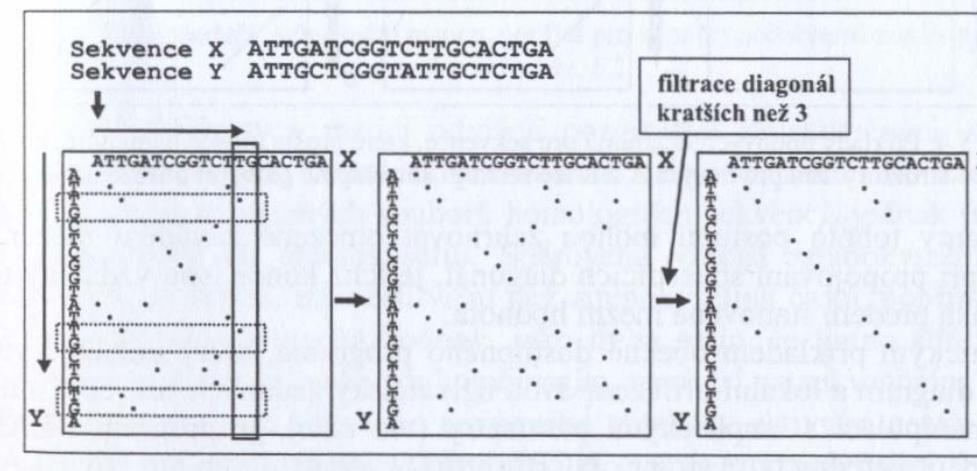
Lokální přiřazení
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYSMIE
-----NAPATNIKSECVRA-PIQNYRRVEHVR-----

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**

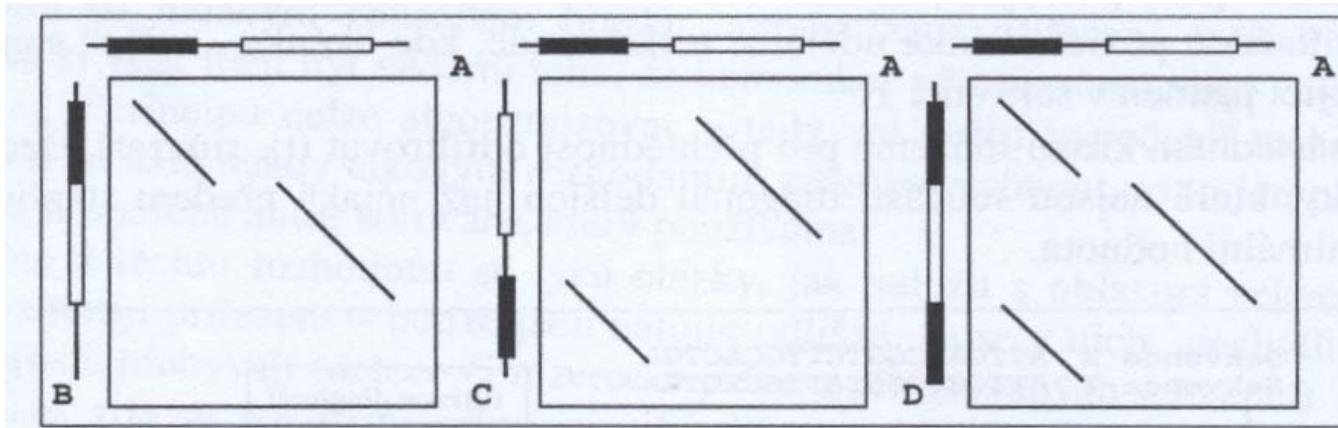


Cvrčková, Úvod do praktické bioinformatiky

- 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

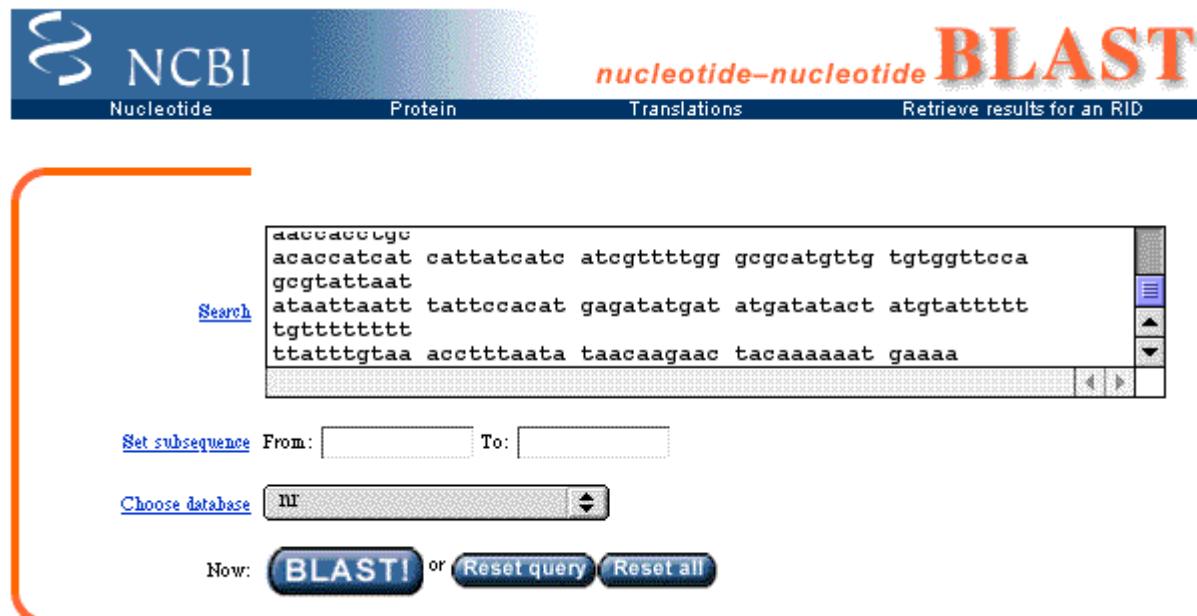


Cvrčková, Úvod do praktické bioinformatiky

- **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/>



BLAST

Basic Local Alignment Search Tool

- Word size: 10-11 bp or 2-3 aa
 - Primary similarities (seed matches)
 - Expanding the homology regions to the left and to the right
- Scoring the homology with matrices PAM (Point Accepted Mutation) or BLOSUM (BLOcks Substitution Matrix)
- Showing the results

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

Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250																			
C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
0	2																		
T	-2	1	3																
P	-3	1	0	6															
A	-2	1	1	1	2														
G	-3	1	0	-1	1	5													
H	-4	1	0	-1	0	0	2												
D	-5	0	0	-1	0	1	2	4											
E	-5	0	0	-1	0	0	1	3	4										
Q	-5	-1	-1	0	0	-1	1	2	2	4									
H	-3	-1	-1	0	-1	-2	2	1	1	3	6								
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6							
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6			
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0
																			17

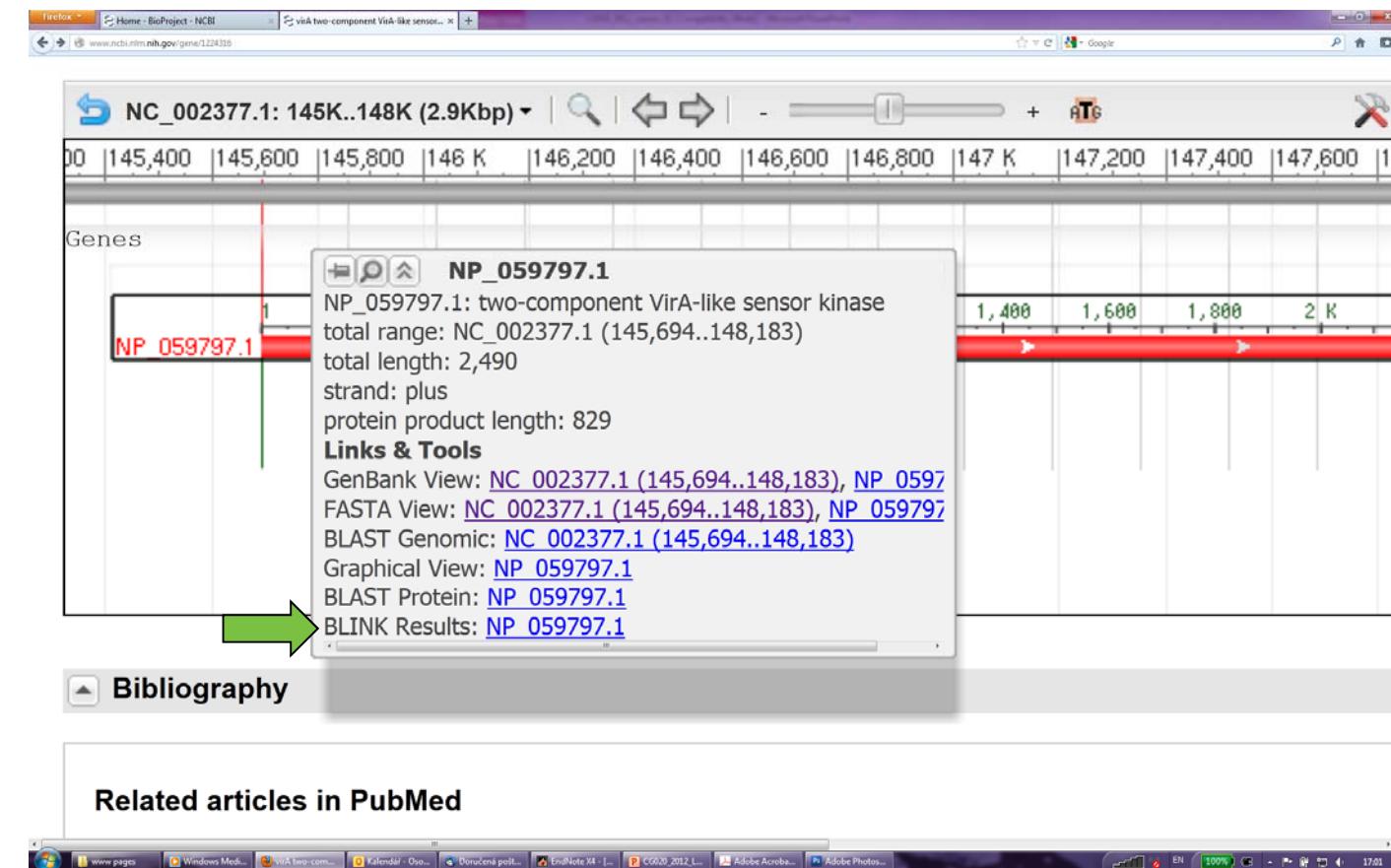
BLAST

Basic Local Alignment Search Tool



- the results shows fraction of identical and in case of proteins also similar sequence positions and/or inserted spaces

Primary Databases



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

1203 Archaea 138285 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
4166	AAK90927	833	two component sensor kinase [Agrobacterium tumefaciens str. C58]
4166	P18540	833	RecName: Full=Wide host range virA protein; Short=WHR virA
4166	AAA79282	833	virA [Plasmid pTiC58]
4159	NP_053380	833	hypothetical protein pTi-SAKURA_p142 [Agrobacterium tumefaciens]
4159	BAA87765	833	tiorf140 [Agrobacterium tumefaciens]
4153	AAA91590	833	virA [Plasmid Ti]
4153	gi 737127	833	virA protein
4153	CAA34777	833	91.3 kDa protein [Agrobacterium tumefaciens]
3800	CAA35780	829	virA [Agrobacterium rhizogenes]
3718	gi 227240	869	virA gene
3148	AAA88643	829	virA [Plasmid Ti]

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** (**P**osition-**S**pecific **I**terated **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** (**P**osition **S**pecific **S**ubstitution **M**atrix)
 - 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

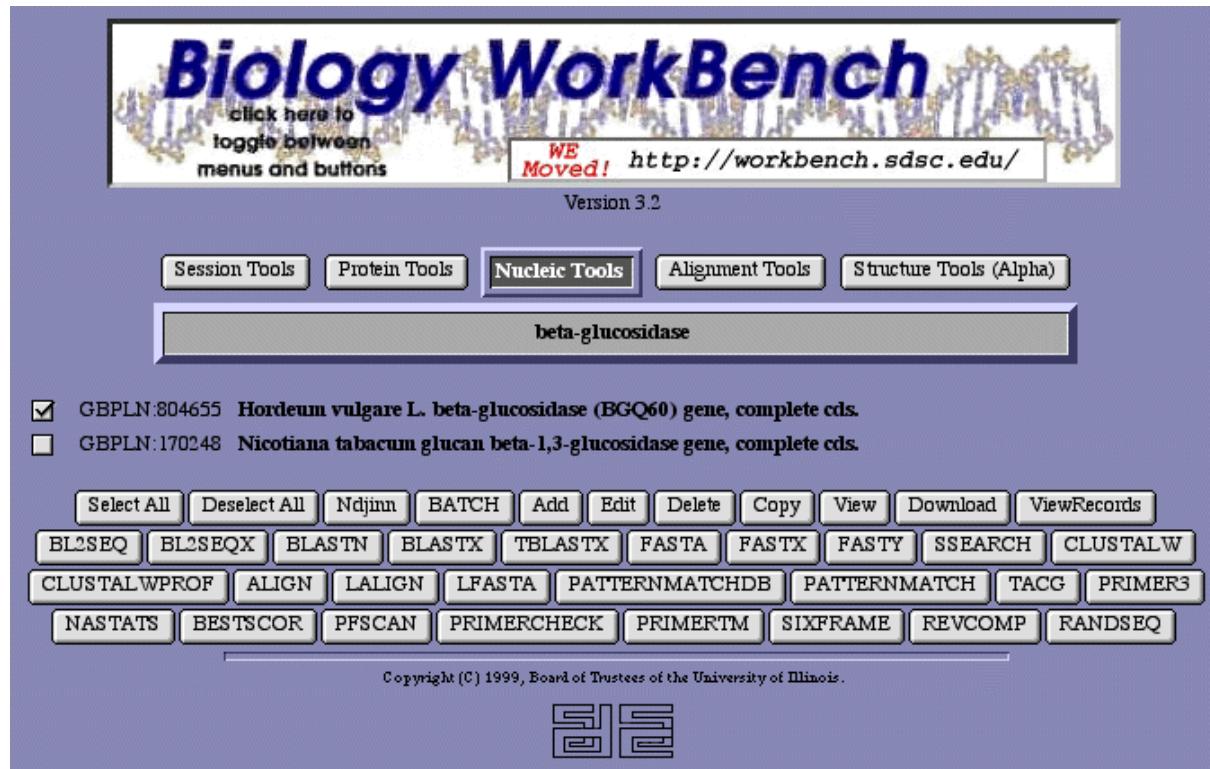
```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase  
MSHIQIPPGLELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAATPRQLGHPPPEPGPDR  
VADAKGDSESEDEDLEVPPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCLQEACKDILLF  
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA  
LMLYNTPRAATIVATSEGSLWGLDRVTFRRRIVKNNNAKKRKMFESESFIESTVPLKSLEVSERMKIVDVIGEK  
IYKDGERIITQGEKADSFYIIIESGEVSILIRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS  
AYAVGDVKCLVMVDVQAFLRLGPCMDIMKRNISHYEEQLVKMFGSSVLDGNLQQ  
[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 STRUCTURAL 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/>

The screenshot shows a web-based application for viewing nucleic acid sequences. At the top, there is a button labeled "View Nucleic Sequence(s)". Below it, there are two dropdown menus: "Format" set to "Fasta" and "Case" set to "Upper". A link "Download/view all sequences in text format" is also present. There are navigation links "[NEXT]" and "[BOTTOM]". The main content area displays the sequence information for "Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds." with GBPLN:170248, 4699 bp. The sequence itself is a long string of DNA bases (A, T, C, G) starting with >170248.

```
>170248
GAGCTCCCTGGGGGCAAGGGCAAAACTTTTGCTAAATGGAAAATTATACCAAGTGTGTATAA
GTTACTCAATTGAATTAAACAAGGGCAATTGACTATTTCGCCTTATATCCTTTGGTCACAAAAAC
ATAAAAATATCCCATCGGAAATTCGAATTCGGTCATTATCGGAAGTAGCTTCTTTAATTATAGTTAGTT
GACAAAAACTATCAAGATATCATTATAATAACTTCAGGCACTCATCTTAGCTGCCTCTCA
GTAGAGGCCAGTAAATAAGACCGATCAAAATAAGCCGCATTAAAATAATGAATTITAGGACTCTC
GATTTGGCACGTAAGTGCCTAAACTCTTCCAATACCTTGTGCAACTTGGGCTAGGTCTGAGCTTC
CAGATATGGGATATTCTAAGTTATCTCTAAATTACATCTCAACTAATATAAGAAATTAAACAGGTA
CAGCAAAATCTAAATAAAATTTCCTCTAAAGAACATGAATCCGGTTACTGATTCAATTGGCTTTTCAGAG
TCTGCATGCCATTCTCAAGGGGCGTTGGTACAAGAAATAATAATAATTTCGGGATAGAATTIT
GAGATTGCAATTATCTTGTGTTAAATTATAAGTATTAGCTAATTTCAGAATAAAATTTCACACTAAAATAG
TAAAATCAACTATCACATGAGAAGGTGGAATGGAATAGCTAACCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTAATTACCAAATGATCGGTAGTCCTCATGAGAATCCAGTATCTCAATAAATGCA
GTAAGAAGTTAGAAAATTTTCAATTAAATCAATTCTATATAATTAAAAATAATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAAATAATAAAAGATAAGAGTAGAGTTAAATTAGGAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCCGTTGCTTCAGTCAAGTAGATTCTCATTCATTGCTCTGCAATAGCAAAA
TGACATCTTACTCTTAAGATACAGCAGGCCACTCTACAATCTCTATTGTATACTCAAAATGAAAGTTTA
GAGAACTTCAAAATCTCTCAACTACTTTAAAGGAATTCAAAATACGACCAATAATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTAAATTGAAATGAAATATAAAATTACTTGATTTAATATAA
```

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
GAGCTCCCTTGGGGGGCAAGGGCAAAACTTTTTGCTAAATGGAAAAAATTATAACCAAGTGTTTGTAATA
GTACTCAATTGAATTAAACAAGGGGCAAATTGACTATTGCGCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTCCAATGGTCAATTATCGGCAAGTAGCTTTCTTTAATTATAGTTAGTT
GACAAAACACTATCAAGGATATCATTATAATAAACTTCAAAAGCCATCATCTTAGCTGCCCTCTCA
GTAGAGGCCAGTAAAATAAGACCGATCAAATAAAAGGCCAATTAAAATAATGAATTITAGGACTCTC
GATTGGCAGTAAAGTGCCAAAACTTCTGCAACTTGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTCTCAAGTTATCTCAATTACATCTCAACTAATTAAAGAAATTAAACAGGTA
CAGCAATCATAAATTCTCTAAAGAAGACATGAAATCGGGTTACTGATTCAATTGCCCTTTCAGAG
TCTGCATGCCATATTCAACTAAGGGTCGTTGGTCAAGAAAATAATAAATTCGGGATAGAATIT
GAGATTGCAATTATCTTGTTTAATTATAAGTATTAGCTAACTGAGAATAAAATTACACTAAAATAG
TAAATCAACTTACATGTAGAAGGGATGGAATAGCTAACATCCATAGCCACTCACATAGAAATATCC
TTATTATCTCACTATTACCAATGATCGGGTTAGCTTCATGAGAACTCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAATTCTCAATTAACTCAATTCAATAATTAAAATATTAGATATGGAGCACTTAA
ATACAATAAAAAGATGTACCGTTAATAATAAAAAGATAAGATAGAGTTAAATAGGAAAAAAACGGTT
CGAGACACTCTTATGGAAAGGCGTTGCTCTCAAAGTAGATTCTCATTCATIGCTGGTGCAATAGCAAAA
TGACATCTTACCTCTAAAGATAACAGCGAGCCACTCTAACATCTTCTATTGTTACTCAAAATGAAAGTTTA
GAGAACCTTCAATCTCAACTACTTTTAAGGGAATTCAAAATACGACCAATAATTATTACITACTTAC
TTATAGTTAAATGATATGAATTTTAAATTGAAATTGAAAATTAAATTACTTGATTAAATATAAA
ACAATAGATAICGCTAACTTACCAACAAACATGGAGACTACTACAGAAGATTTATTATTGAAACGAT
GATTAAGCAGCTATTCTCTGGTTGTGAGGATGAAAGAAAGTAACTAGCTATAATTCTTTGTAAAGT
```

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
ELPWNGARAKLFAWKNIIPSVVCNSYSI*INKGANLTILPL

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

Frame 2, 1 stop codon

Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds. Tran
>170248 Translated - Frame 2
SSLGGQGQNFLNGKILYQVFVIVIQFELTKGQI*LFCP

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

Analytical Tools

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

— Linear Map of Sequence:

```
StyI          BsaJI
CviJI          AluI
SacI          EcoICRI
Bsp1286I      BsiHKAI
BanII        BsII          SspI
\ \ \ \ \
1  gagctccttggggcaagggcaaaacttttgcataatggaaaaattataccaaatgttcac 60
    ^ * ^ * ^ * ^ * ^ * ^ * ^ * ^ * ^ *
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          Tsp509I
MaeIII Tsp509I  MseI          ApoI
\ \ \ \ \
61  gtttgaatagttaactcaatttgaattaacaaggggcaatttgactatggccctta 120
    ^ * ^ * ^ * ^ * ^ * ^ * ^ * ^ *
1  V C N S Y S I * I N K G A N L T I L P L
2  F V I V T Q F E L T K G Q I * L F C P *
3  L * * L L N L N * Q R G K F D Y F A L R
4  N T I T V * N S N V F P C I Q S N Q G *
5  T Q L L * E I Q I L L P A F K V I K G K
6  H K Y Y N S L K F * C L P L N S * K A R
```

Analytical Tools

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

Selected Sequence(s)

- Lycopersicon esculentum beta-1,3-glucanase mRNA, complete cds.
- Capsicum annuum clone GC170 beta-1,3-glucanase-like protein gene.
- Nicotiana tabacum glucan beta-1,3-glucosidase gene complete cds.
- Nicotiana plumbaginifolia beta-(1,3)-glucanase gene for a vacuolar.
- Hordeum vulgare L. beta-glucosidase (BGQ60) gene, complete cds.

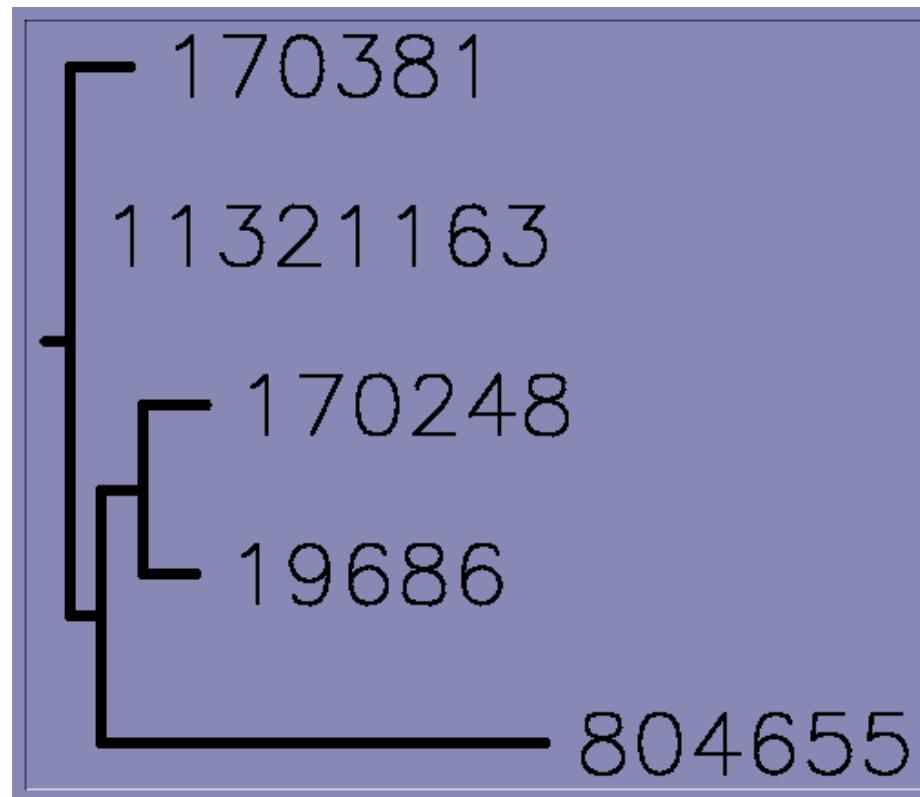
Download a PostScript version of the output

The screenshot shows a sequence alignment interface with three panels. The top panel displays the selected sequences: Lycopersicon esculentum beta-1,3-glucanase mRNA, Capsicum annuum clone GC170 beta-1,3-glucanase-like protein gene, Nicotiana tabacum glucan beta-1,3-glucosidase gene complete cds, Nicotiana plumbaginifolia beta-(1,3)-glucanase gene for a vacuolar, and Hordeum vulgare L. beta-glucosidase (BGQ60) gene, complete cds. Below this is a toolbar with various icons. The main panel contains sequence data with columns for position, sequence, and identifiers. The bottom panel shows a detailed view of a specific sequence segment with color-coded regions and annotations.

Position	Sequence	Identifier	
2560	GTTTGTTGGTCCTGCTGCTGACAACCTCGAGTGGAGACTGGGCTACACTGGGGTTTCGGG	804655	
24	AATGGC 170381	
1	CAAAGATT 11321163	
2430	CAAGAATT 170248	
1743	GAGTAAAATGATGAGACAGCTGGAAAAAAAGAACGGAAATAATGGTAAAGAAAAAA	AAATTG 19686	
2620	GATCGCTATGTTGAGCTTCATAACTCTGAGAGGTAGCCCAAAGGACTCAGGGCTTGCGG	804655	
32ATTATTTATGTTGAGATTACTGGGGCA	CAAGAATTGATAAC 170381	
1TGATGTTGATGTTGTTAACCTTATT	CAATTTCATTCAC 11321163	
2438AGCATGTTTAAATGCTTATGCGAA	GGCCCACTCGCTATTTTAATTGATATTGAC 19686	
1803	AGGATGTTTAAATGCTTATGCGAA	GGCCCACTCGCTATTTTAATTGATATTGAC 804655	
2680	GAAGAACATGGCTGGGA	AGAAAGAGAGCTAGGATGGAAACAGGATCGGGAGACATG	
79	AQGGCTAAATAGGTTG	TGCTATGGAAATGATGGCA	GAACCTGGGATCAC 170381
1TATGGCTGTGCTATGGAAATGATGGCA	GAACCTGGGATCAC 11321163
2484	AGGGCTAAATGCTAGGTTGTGCTATGGAAATGATGGCA	GAACCTGGGATCAC 170248
1863	AGGGCTAAATGCTAGGTTGTGCTATGGAAATGATGGCA	GAACCTGGGAAATG 19686
2740	AGCTGGCTGCGCTTGAGAATGATGTTGTTGTTTGCTAGAACCTG	804655
132	ATTCGAAAGTTATACAGC	CTTACAGTCAGAACACATGAGAGCTGGTTATGA	CTGGAGCTTATGA 170381
45	ATTCGAAAGTTATACAGC	CTTACAGTCAGAACACATGAGAGCTGGTTATGA	CTGGAGCTTATGA 11321163
2540	ATTCGAAAGTTATACAGC	CTTACAGTCAGAACACATGAGAGCTGGTTATGA	CTGGAGCTTATGA 170248
1919	ATTCGAAAGTTATACAGC	CTTACAGTCAGAACACATGAGAGCTGGTTATGA	CTGGAGCTTATGA 19686
2860	ACTTCGGCCCTGGCTGGAAATGAGCTGGCGAAATGCGAG	CTGGAGCTTATGA	CTGGAGCTTATGA 804655

Analytical Tools

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



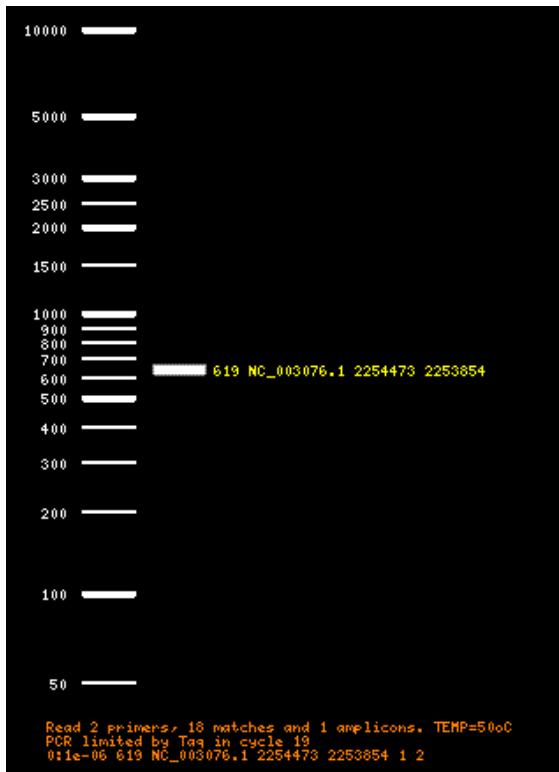
Analytical Tools

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

The screenshot shows the VPCR 2.0 web interface. At the top, there is a navigation bar with links for SEARCH, ABOUT, DOWNLOAD, and LINKS. The main content area contains a note about the software's capabilities and limitations, mentioning BLAST search and support for IUB codes. Below this, there is a search form where users can specify the search method (BLAST) and database (M. musculus). There are eight input fields for Primer 1 through Primer 8. Underneath these, there is a field for Annealing temperature set to 50, and a "Do PCR!" button. The VPCR logo is located at the bottom right of the form.

Analytical Tools

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



Outline

- Syllabus Of The Course
- Definition Of Genomics
- Role Of Bioinformatics In Functional Genomics
- Databases
 - Spectre Of „On-line“ Resources
 - PRIMARY, SECONDARY And STRUCTURAL 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 page for the PHACTR4 gene in *Homo sapiens*. The page includes a summary of the gene's characteristics, its genomic context on Chromosome 1, and links to various databases and resources.

Summary:

- Official Symbol: PHACTR4
- Official Full Name: phosphatase and actin regulator 4
- Primary source: HGNC:25793
- Locus tag: RP11-442N24_A1
- See related: Ensembl:ENSG00000204138, HPRD:07816, MIM:608726
- Gene type: protein coding
- RefSeq status: REVIEWED
- Organism: *Homo sapiens*
- Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrini; Hominoidea; Homo
- Also known as: FLJ13171, MGC20618, MGC34186, DKFZp686L07205, RP11-442N24_A1
- Summary: This gene encodes a member of the phosphatase and actin regulator (PHACTR) family. Other PHACTR family members have been shown to inhibit protein phosphatase 1 (PP1) activity, and the homolog of this gene in the mouse has been shown to interact with actin and PP1. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]

Genomic context:

Location: 1p35.3
Sequence: Chromosome 1: NC_000001.10 (28696093..28826881)

Chromosome 1 - NC_000001.10

28685943 → MEO1A → PHACTR4 → SNORD7B → SNORD7B → RNU15A → SNORD5 → RNU1

Genomic regions, transcripts, and products:

Genomic Sequence: NC_000001 chromosome 1 reference GRCh37.p5 Primary Assembly

Go to reference sequence details

Go to nucleotide Graphics Fasta GenBank

PubChem Compound PubChem Substance PubMed PubMed (GeneRIF) PubMed (OMIM) RefSeq Proteins

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 STRUCTURAL Databases
 - GENOME Resources
- Analytical Tools
 - Homologies Searching
 - Searching Of Sequence Motifs, Open Reading Frames, Restriction Sites...
 - Other On-line Genome Tools

Discussion