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

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

CEITEC

Course Syllabus

- **Lesson 01**
  - Introduction into Bioinformatics
- **Lesson 02**
  - Identification of Genes
- **Lesson 03**
  - Reverse Genetics Approaches
- **Lesson 04**
  - Forward Genetics Approaches

CEITEC

Course Syllabus

- **Lesson 05**
  - RNA Interference and Genome Editing
- **Lesson 06**
  - Gene Expression and Chemical Genetics
- **Lesson 07**
  - Protein-Protein Interactions And Their Analysis
- **Lesson 08**
  - Recent Approaches in DNA Sequencing

CEITEC

Course Syllabus

- **Lesson 09**
  - Structure of Genomes
- **Lesson 10**
  - Genome evolution
- **Lesson 11**
  - Genomics and Systems Biology
- **Lesson 12**
  - Practical Aspects Of Functional Genomics
  - Model Organisms,
  - PCR

CEITEC

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

CEITEC

Outline

- Syllabus of thecourse
- Definition of Genomics

CEITEC

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

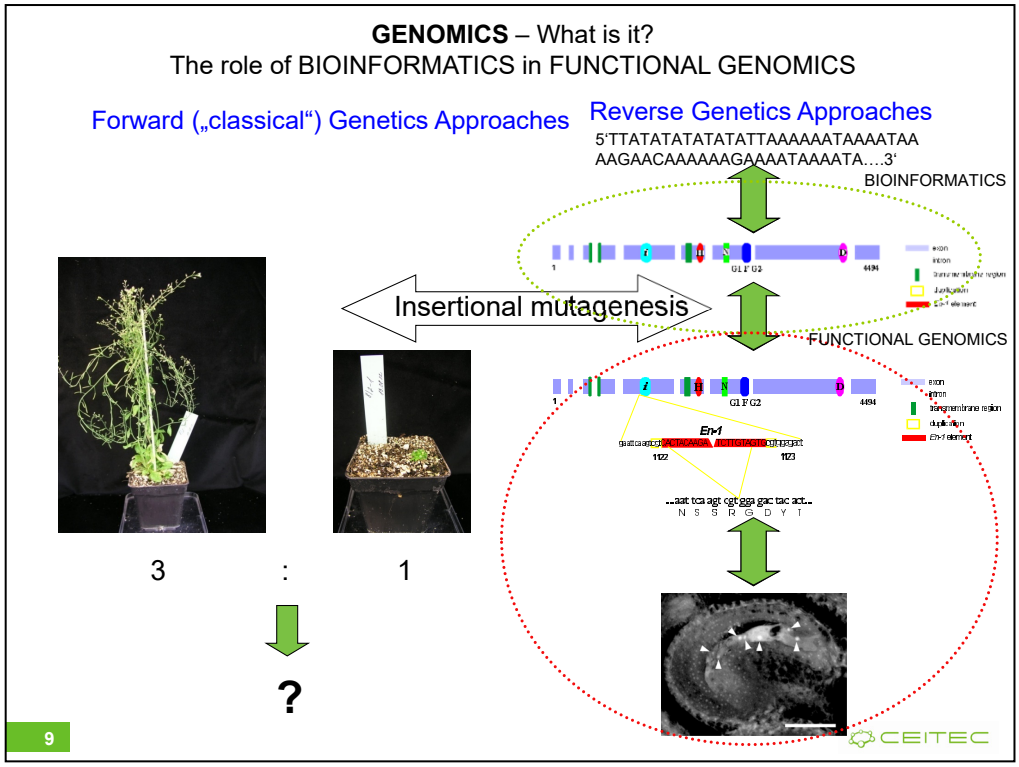
8

CEITEC

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 euckaryotic 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.

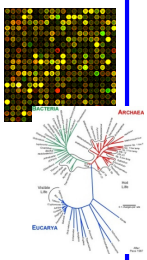
The necessity of having phenotype alterations in the forward genomics approach introduces important difference between those two approaches. Thus, the gene is no longer understood as a factor (*trait*) determining *phenotype,* but rather as a piece of DNA characterized by the unique *string of nucleotides*.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

Outline

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

CEITEC

Bioinformatics



- **Definiction 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?

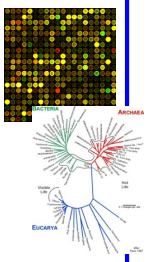
- **Interface** between the **biology** and **computers**
- **Analysis** of **proteins, genes** and **genomes** using **computer algorithms** and **databases**
- **Genomics** is the **analysis** of **genomes**.

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

CEITEC

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

CEITEC

Outline

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

CEITEC

Spectre of On Line Resources

EMBnet National Nodes		
Vienna Biocenter	Austria	http://www.at.embnet.org/
BEN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENIUSnet	Germany	http://genome.dkfz-heidelberg.de/biounit/
IMBB	Greece	http://www.imbb.forth.gr/
HEN	Hungary	http://www.hu.embnet.org/
INCBI	Ireland	http://acer.gen.tcd.ie/
INH	Israel	http://dapsas.weizmann.ac.il/bcd/inn.html
IEN-ADR	Italy	http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm
CAOS/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.genebee.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.dl.ac.uk/
EMBnet Specialist Nodes		
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
EMBnet Associate Nodes		
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/
CDFD	India	http://solarjung.embnet.org.in/
SANBI	South Africa	http://www.sanbi.ac.za
USA Information Providers		
NCBI	USA	http://www.ncbi.nlm.nih.gov/
NLM	USA	http://www.nlm.nih.gov/
NIH	USA	http://www.nih.gov/

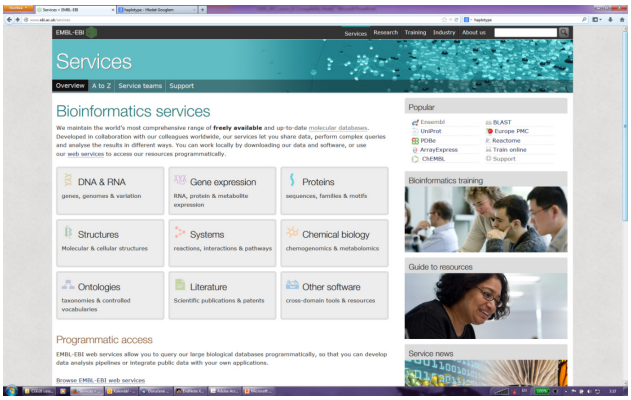
15

CEITEC

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

Spectre of On Line Resources

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



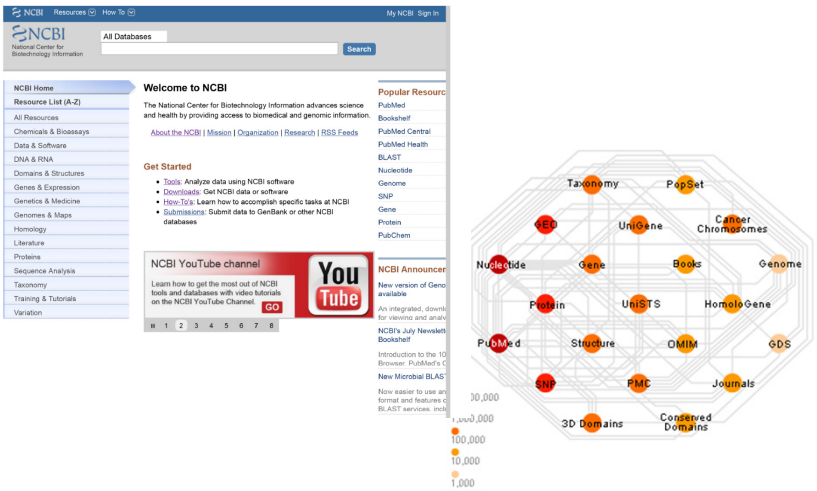
16

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

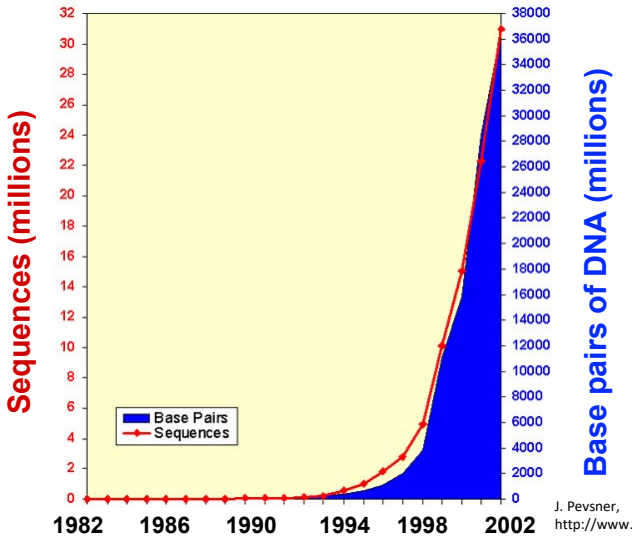
CEITEC

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 10^6 entries (approx. 33 x 10^9 bp)
  - August 2005 100 x 10^9 bp from 165.000 organisms

CEITEC

Growth of GenBank

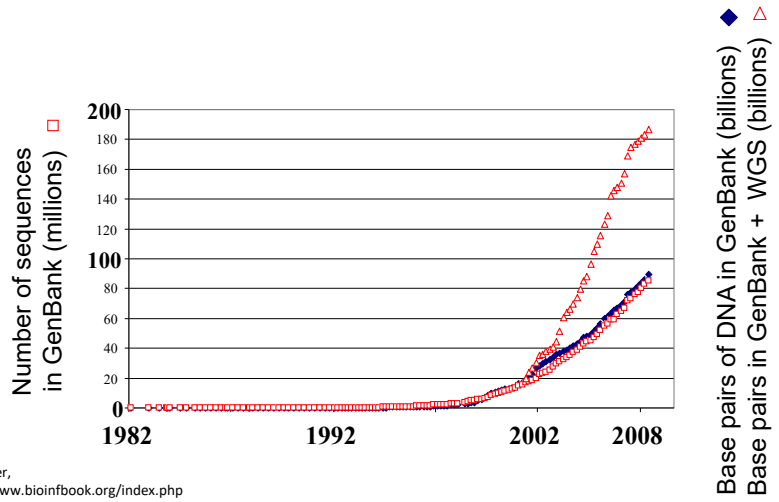


J. Pevsner,
http://www.bioinfbook.org/index.php

20

CEITEC

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

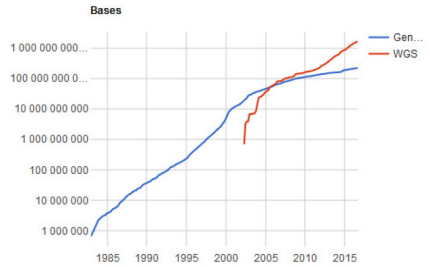


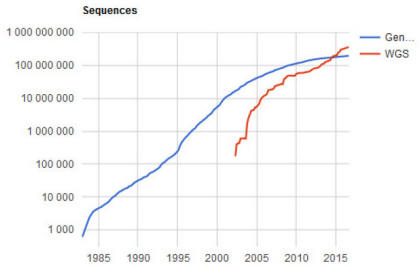
J. Pevsner,
http://www.bioinfbook.org/index.php

CEITEC

Growth of GenBank

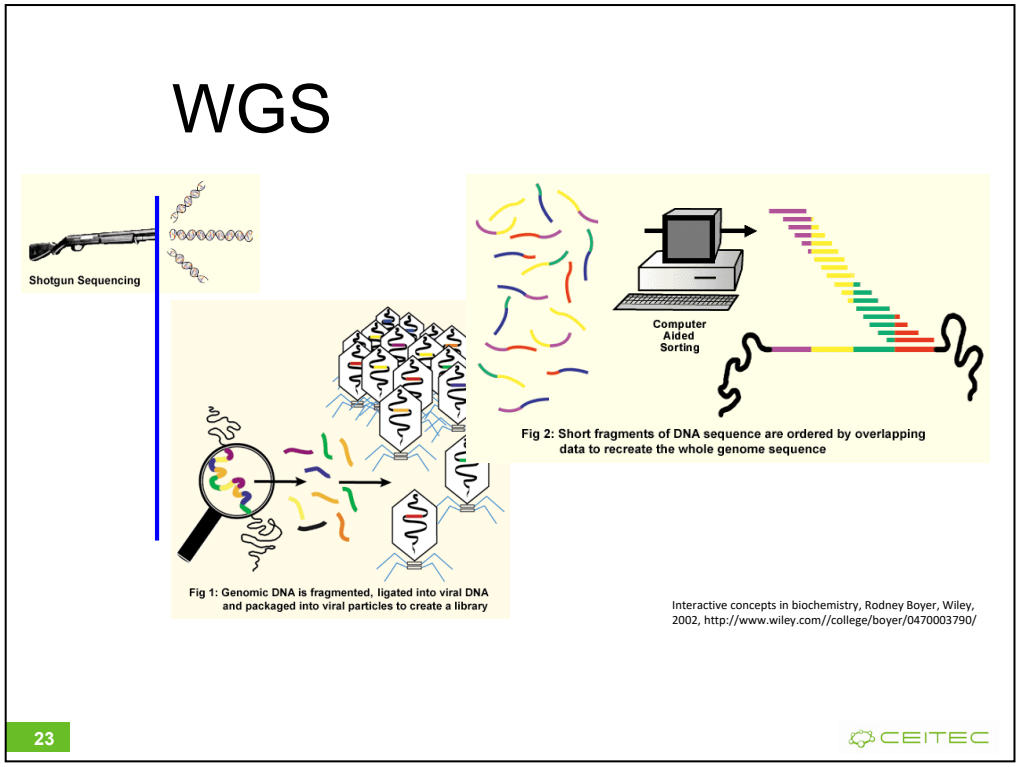
Aug 2016





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

CEITEC

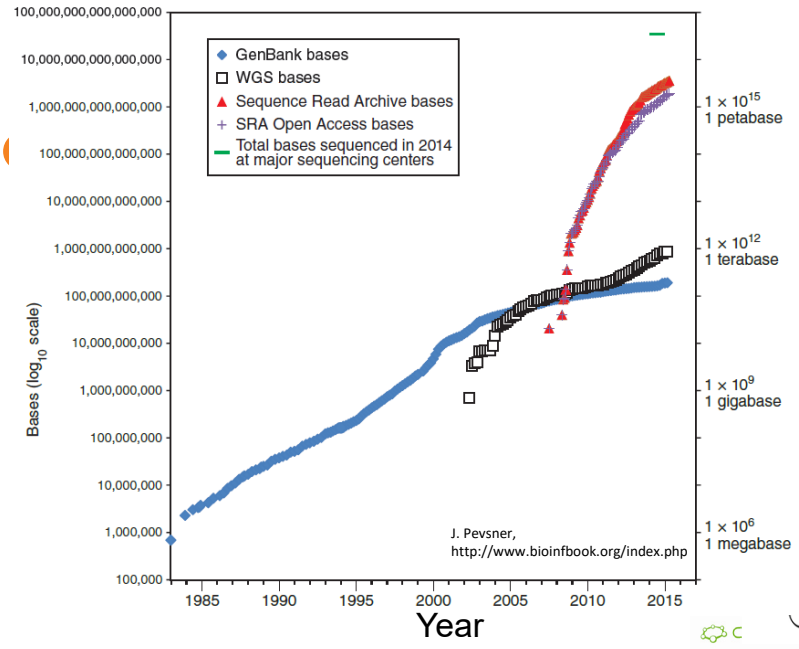


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic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

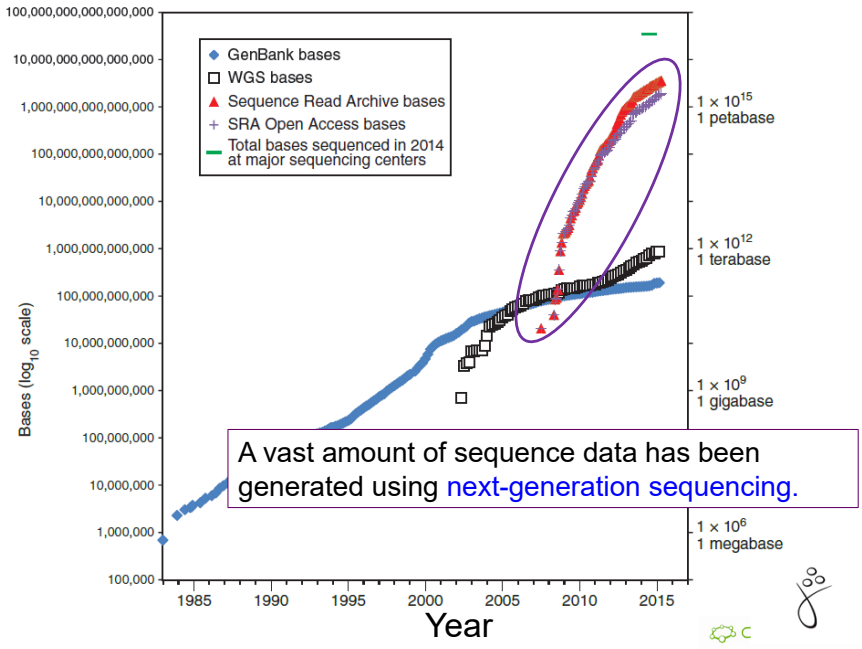
The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence.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

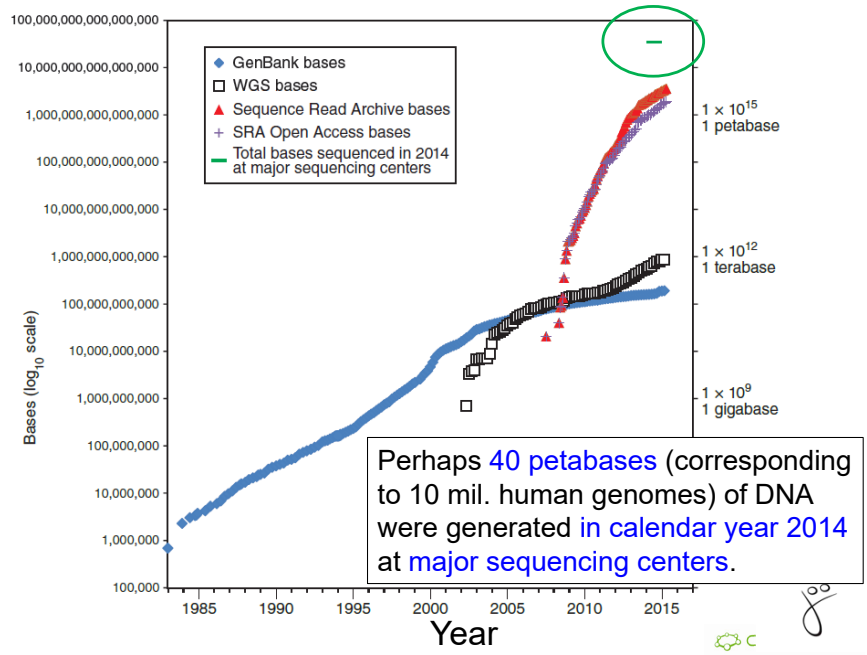


Growth of DNA Sequence in Repositories



A vast amount of sequence data has been generated using next-generation sequencing.

Growth of DNA Sequence in Repositories



Perhaps 40 petabases (corresponding to 10 mil. human genomes) of DNA were generated in calendar year 2014 at major sequencing centers.

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/

CEITEC

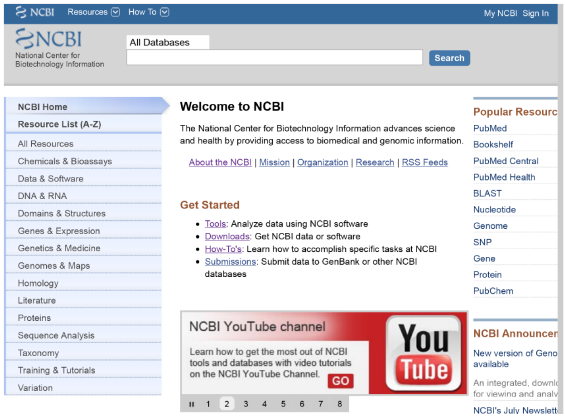
Primary Databases

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

CEITEC

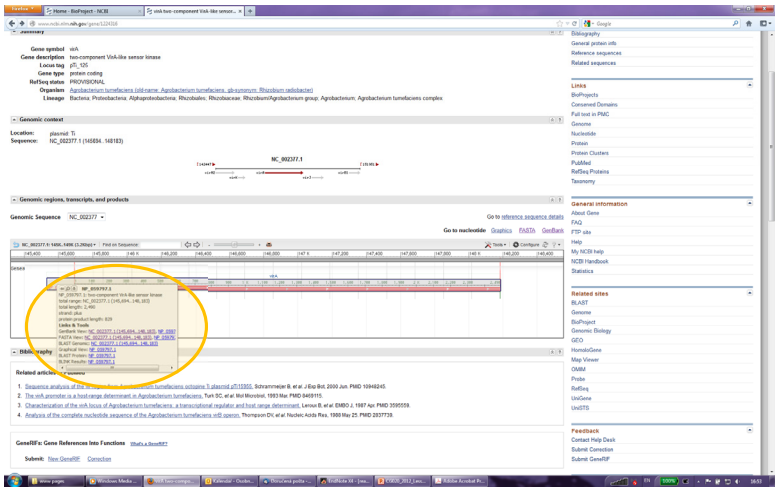
Primary Databases

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



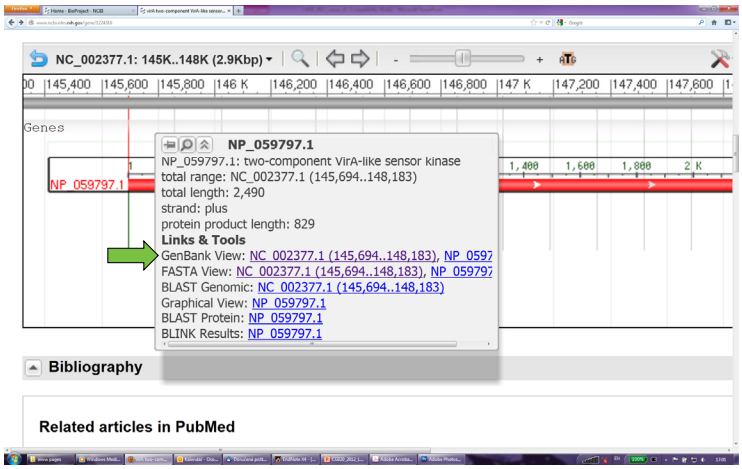
CEITEC

Primary Databases



CEITEC

Primary Databases



CEITEC

Primary Databases

CEITEC

Primary Databases

```
                     /translation="MNGRYGPTRQDPKTGAKPWGILALITAAMIFAFMAVAGWQNNAT
                     TQAILAQLSKINADSEAELQKEVLRANTDTVANYRPIIGRLGALEKKLRDLEQLFRQGH
                     IVRGNAQLLAQLAYYLANADAAYAAPGNQNVRLQDSLASFTRALGLLPGKASTEGT
                     LRKFTELASNMLQFLRQPGPAISFRISLELRSLQRGLLDSAPVRILARRGPILELL
                     PQVEDLWNMIQTYDTARIAEMLQRECLEYYSLKNVERHRARIFLGLASYGLCLTIITL
                     VYPLLKKTDWLARSLDYRKLIRKIVCVFRGRAATYHAAQAAIAIIQRFFRDKTCALAL
                     VDNLRRWAVRTFQARKRPKPVNDDSYLRRIVKSYTAADRRATVFRIISKKKIVRLPLRIP
                     GLSILARRSTCRLIAVCLGTQETRPRFQCRSQLELATACLCRYIDVRRRQTRTD
                     VLARRLRRNQRLEAVTTLAGGIAHRPRNTLDGTLDRAELAQMGYGTTYTRRYIDYII
                     GSDEGAMLIIDGILTLGRQGRMIRPPYVRLTTRIAPLGMALPPRIRLGPRPFCQWQ
                     SYIRGSPLRLQGYLINICKNASQAMTANRQIDIIIEGQAPLPYRKILARGYMPPGEIVL
                     LRIGRRXXIPRATLPRIPRPPFTTRARRRXTGLGLAFRRRRIEAPRAYIDYGETYRR
                     GTRPDIYLPPGEGEKPYRGGPPGRRPARPRGRRIVLYEPGDLLRKAYEDKIAALGYR
                     PYGFETYRLIRGWIKGRRALGWYDQAALRFQQSPNRYDLYLATASIIIGGMLKMT
                     LRRERVYRLYLPKFIGGKTKAHAILYTLRY"
ORIGIN
        1 atgaacggaa gatatttaac gacgcggcag gattttaagk cagggcgcgaa gccttggtct
       61 atattggccc ttatcgttgc tgcaatgatt ttcgcgttca tggcggttgc gtcctggcag
      121 gacaatcgcga ctacccaggc aatcctcagc caactacgat cgattaacgc cgacagcgcc
      181 tcactgcagc gcgatgtact ccgcgctcac acgggcaccg tggcgcaacta ccgcccccatt
      241 atctccaggc tgggagctct gcggaagaat ctgcaagatt tgaagcaatt atttagacaa
      301 tctcatattg taagtgagag caatgctgct caactgctac gccagctaga agtgtctcta
      361 aattcgctg acgcggcggt cgccgcctt ggtgcgcaaa atgtacctct gcaagattcg
      421 ctggccagtt tcactcgtgc tttgagcagt cttccagaaa aagcctcaac cgatcagact
      481 ttagaaaaac caacagaatt ggctagcatg atgctccaat ttcttcggca accaagccog
      541 gctatttcat tcgagatcag ccttgaacta gagaggctcc aaaacaacg cggtcttgat
      601 gaagctccog tgcgcatact tgcacgtgaa ggtcccatta tcttatcgct tttgccacag
      661 gtgaaagatc tgtgtaacat gattcagacg tctgacaccg cagaaattgc ggagatgctg
      721 cagcgcgagt gtttgtaggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc
      781 ttcctcggct cogctctagt ggtcctttgc ctctcatctca tcacctagc ctatagctca
      841 cgcaaaaaaa ccgattggtt agcgcggcgt ttagattacg aagagctaat caaagagatc
      901 ggagtatgtt ttgaagtgga ggcggccacc acgtcgtccg cgcaagctgc acttcgtatt
      961 attcagcgct tctttgatgc cgatacgtgc gcgtcagctc tagtggacca tgaccgtaga
     1021 tgggctgtcg aaacattcgg tgcgaaacac ccaaaacctg tgtgggacga cagcgtgcta
     1081 cgcgaaatag tctctcgtac caaagcggac gaacggggcg cggtatctcg catcatatcg
     1141 tcgaaaaaaa tcgtacattt gcctctcgaa attccaggtc tctcgatact actggctcac
     1201 aaatctacag ataactat tgcggttgt tcactggctt acctaaagcta tcgcctctga
     1261 cctgccaag gcgaaattca gcttcttgaa ctcgccaccg cctgcctctg tcactatatc
     1321 gatgtcccgc gtaagcagac cgaatgcgac gtcttgccca gacgattgga gcatgcgcaa
     1381 cgcctgaggg cagtctgtac acttgccggc ggaatagcac atgaatttaa taacatcttg
     1441 ggctcaatcc tcgggcacgc agaattagca caaaatcgg tgtctcgaac atctgtcacc
     1501 cgaagatata ttgatatat catttgctca ggcgacagag ccatgctcat tatcgatcag
     1561 atcttgacgc tgagccgaaa acaggagcgc atgatcaagc catttagtgt ctcagagctt
     1621 gtgacccgaaa ctcgctcctt gctacgtatg gctcttccgc caaacatcga gcttagttcc
     1681 agatttgatc aaatgcagag cgtgatcgaa ggaagcccgc ttgaacttca acaggtacta
     1741 attaccatct gcaagaatgc ttcccaagcc atgactgcaa atggtcaaat cgacatatcc
     1801 atcagccaag ctttttacc agtaagaaa attctggcgc atgtgtgttat gccacctggc
     1861 gactatgtc tcctatctat tagcgacaat gctgcgagcg ttccccgaggc tgtgtcaccc
     1921 cacatttttg aacccttctt tacgacacga gctcgcaacg gtggaacggg tctcggcctt
     1981 gcttctgtgc atggtcatat cagccgcgtt gcgggttaca tgcaccgttag ttcaactgct
     2041 gggcatggga cgcgctttga catttatctc cctcgtctt ctaaggaacc cgtaaatcca
     2101 gacagtttt tcggcgcaa taaggtaccg cgtggaaacg ggagagttgc ggcacttgct
     2161 gagcccgatg acctcctgcg ggaggcgtat gaagacaaga tcgccgctct aggatatgag
     2221 ccgtccggtt ttgtcacctt taatgaaatt cgcgattgga tttcaaaagg caatgaagcc
     2281 gatctggtca tggtcgacca agcgtctctt cctgaagatc aaagtcctaa ttccgtggat
     2341 ttagtgctca agacgcgctc catcatcatt ggcgggaatg atctcaaaat gaccctttca
```


CEITEC

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

CEITEC

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.

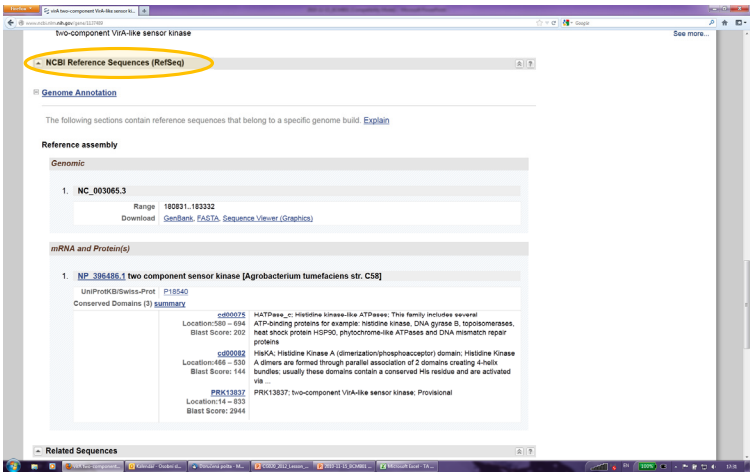
RefSeq identifiers include the following formats:

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

CEITEC

RefSeq



CEITEC

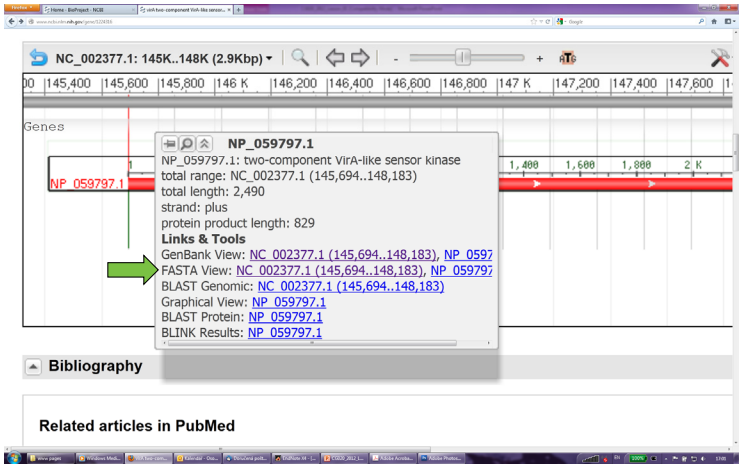
NCBI's RefSeq project: many accession number formats for genomic, mRNA, protein sequences

Accession	Molecule	Method	Note
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

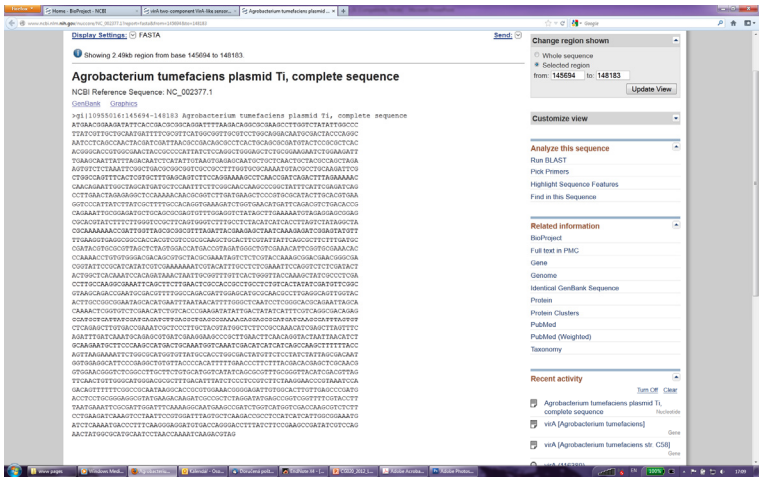
CEITEC

Primary Databases



CEITEC

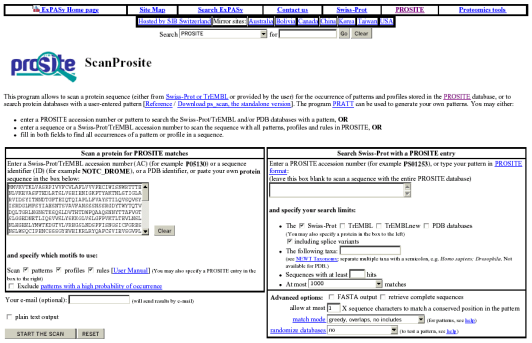
Primary Databases



CEITEC

Secondary Databases

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



CEITEC

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].

```
171 - 585  nkeesstYeteiasn
```

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

```
744 - 747  SRvT
414 - 417  KRrS
```

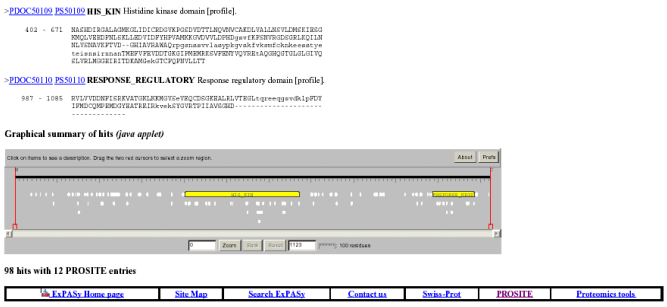
>PDOC00005 PS00005 **PKC_PHOSPHO_SITE** Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

```
148 - 150  SsR
164 - 166  TgR
171 - 173  StK
219 - 221  SkK
369 - 371  TrR
460 - 462  SgK
513 - 515  SgR
585 - 587  SiR
602 - 604  TgK
652 - 654  TdK
716 - 718  SpR
726 - 728  SpK
747 - 749  TsK
794 - 796  SsR
854 - 856  StK
864 - 866  StR
868 - 870  SsR
921 - 923  SpK
957 - 959  SvR
960 - 962  TgR
974 - 976  TsK
997 - 999  SrK
1002 - 1004  TgK
1018 - 1020  SgK
1031 - 1033  TgR
1119 - 1121  SkR
```


CEITEC

Secondary Databases

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


>PDOC50109 PS50109 **HIS_KIN** Histidine kinase domain [profile].

402 - 671 HAANDIRGALAGMEGLIDICRDGVKPGSDYDTTLNQVMVCAEDLVALLNSYLDMSKIEGG
KNQLVEEPNLSKLLEDVIDFYNPVAMKKGVDVVLDSNDgsvfKPSNVRGDSGRLEQILN
NLVENAVKFTVD--GHIAVRAMHQprgqnsasvvlsnyNgyvskfvksmfchqkseestys
teisnnirsnsnTMEFYFVDCTGKGIPMSMKESVFENYVQVRRSAQSNQHTGLGLGIVQ
SLVRLMGGEIRITDEANSekGTCFQFHVLLTT

>PDOC50110 PS50110 **RESPONSE_REGULATORY** Response regulatory domain [profile].

987 - 1085 RVLVVDDNPISRKVATGKLKNMGYGeVRQCDSGKEALRLVTRSLtqreqqsvdklpFDV
IFMDCQMPEMDGYEATRRIRkvekEYGVRTPIIAVGGED----------

Graphical summary of hits ***(java applet)***

98 hits with 12 PROSITE entries

ExPASy Home page	Site Map	Search ExPASy	Contact us	Swiss-Prot	PROSITE	Proteomics tools

CEITEC

Secondary Databases

- Databases of **functional** or **structural *motifs*,** acquired by primary data (sequences) comparison

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

PRINTS

Protein Fingerprint Database

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. Reference

New:

- SPRINT - *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

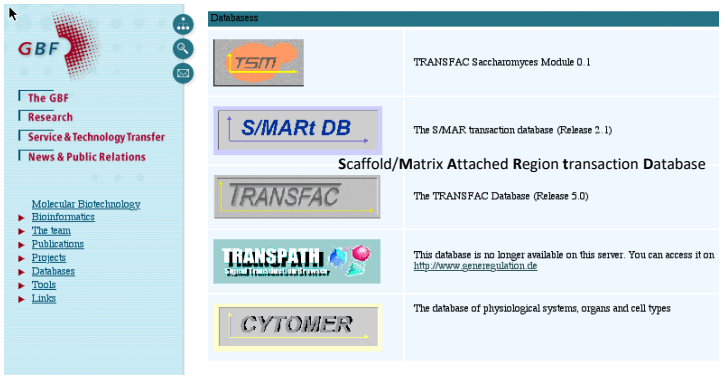
PRINTS search:

- Search PRINTS with NEW FingerPRINTScan
- FPScan
- GRAPHScan
- MULScan
- FingerPRINTScan binaries and source are available: contact scanfin@bioinf.man.ac.uk

CEITEC

Secondary Databases

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



CEITEC

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)

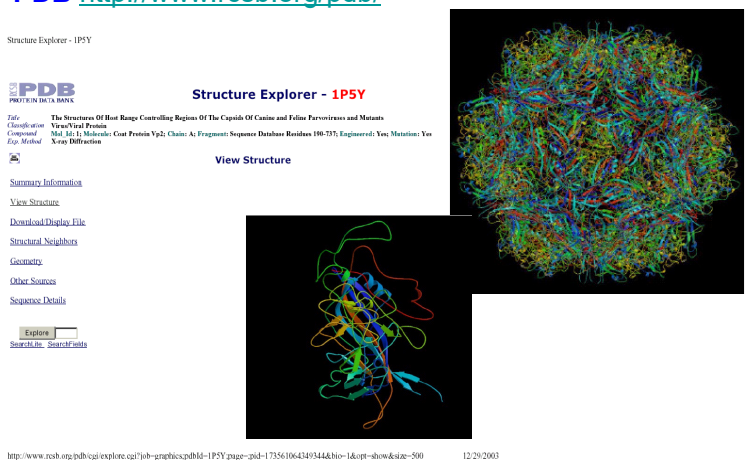
Structural Databases

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

CEITEC

Structural Databases

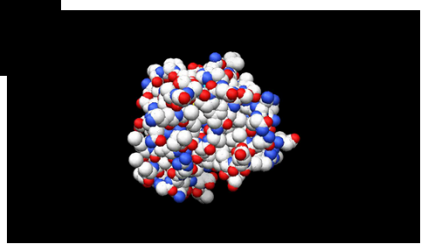
- **PDB** http://www.rcsb.org/pdb/



CEITEC

Structural Databases

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



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

CEITEC

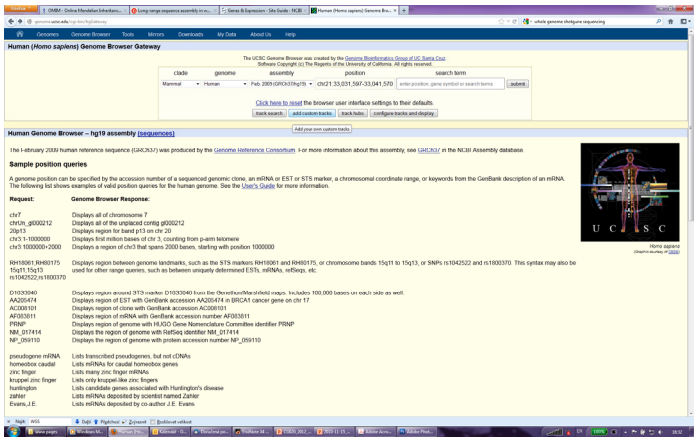
Outline

- 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

CEITEC

Genome Resources

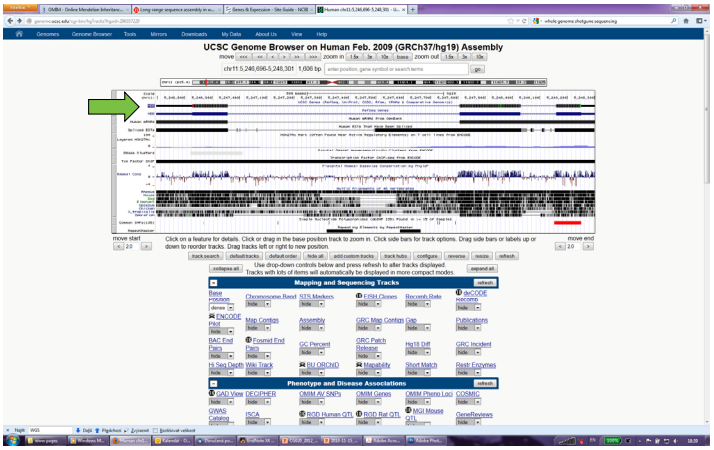
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genome Resources

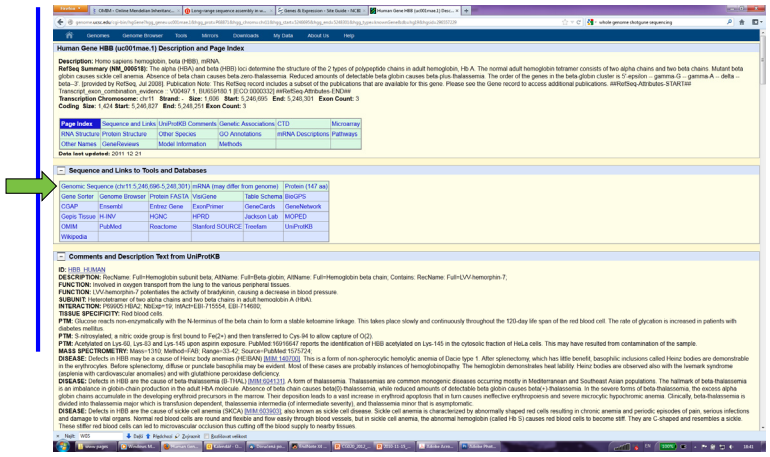
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genome Resources

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



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_exon_combination_evidence - V004617 1, [6,855/855] 1 [ECO:0000332] ##RefSeq-Attributes-END##
Transcription Chromosome: chr11 **Strand:** - **Size:** 1,636 **Start:** 5,246,695 **End:** 5,248,301 **Exon Count:** 3
Coding Size: 1,424 **Start:** 5,246,827 **End:** 5,248,251 **Exon Count:** 3

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

Data last updated: 2011 12 21

Sequence and Links to Tools and Databases

Genomic Sequence (chr11:5,246,695-5,248,301)		mRNA (may differ from genome)			Protein (147 aa)
Gene Sorter	Genome Browser	Protein FASTA	VisGene	Table Schema	BioGPS
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneNetwork
Gepis Tissue	H-INV	HGNC	HPRD	Jackson Lab	MOPED
OMIM	PubMed	Reactome	Stanford SOURCE	Treefam	UniProtKB
Wikipedia					

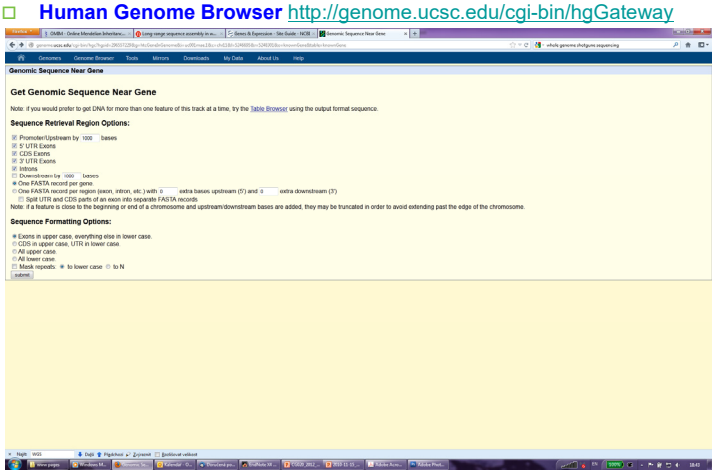
Comments and Description Text from UniProtKB

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: Heterotetramer of two alpha chains and two beta chains in adult hemoglobin A (HBA).
INTERACTION: P69905: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-93 to allow capture of O(2).
PTM: Acetylated on Lys-60, Lys-83 and Lys-145 upon aspirin 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=13192; Method=FAB; Range=33-42; Source=PubMed:15753724;
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 Iwemark 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 thalassemia. Thalassemias are common monogenic diseases occurring mostly in Mediterranean and Southeast Asian populations. The hallmark of beta-thalassemia is an imbalance in globin-chain production in the adult HbA molecule. Absence of beta chain causes beta(0)-thalassemia, while reduced amounts of detectable beta globin causes beta(+)-thalassemia. In the severe forms of beta-thalassemia, the excess alpha globin chains accumulate in the developing erythroid precursors in the marrow. 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 (SICA) [MIM:603903], also known as sickle cell disease. Sickle cell anemia is characterized by abnormally shaped red cells resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resembles a sickle. These stiffer red blood cells can lead to microvascular occlusion thus cutting off the blood supply to nearby tissues.

CEITEC

Genome Resources

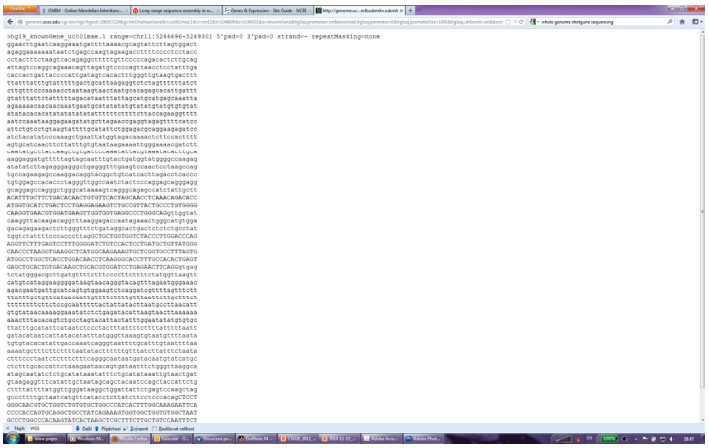
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genome Resources

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



CEITEC

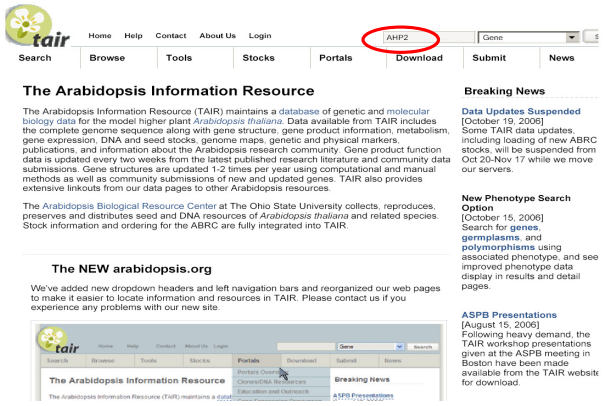
Genome Resources

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

CEITEC

Genome Resources

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



CEITEC

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

CEITEC

Analytical Tools

- **Global** versus **Local** alignment

```
Globální přiřazení
SLAV----------APATNIK-------PIQNYR-I------AKSETQRYMVIE
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE

Lokální přiřazení
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE
-------------NAPATNIKSECVRA-PIQNYRRVEHVRA-------------
```

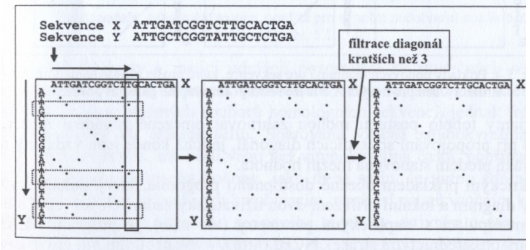
Cvrčková, Úvod do praktické bioinformatiky

- 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

CEITEC

Analytical Tools

- Choosing the right type of alignment using dotplot



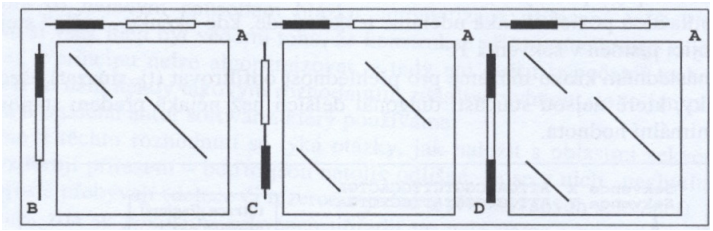
Cvrčková, Úvod do praktické bioinformatiky

  - 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 treshold

CEITEC

Analytical Tools

- Examples of sequence alignment using dotplot



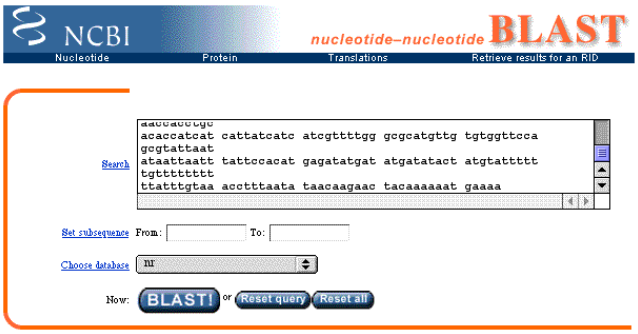
Cvrčková, Úvod do praktické bioinformatiky

- **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 neccessary to do a local alignment
- Dotplot can be obtained using BLAST2 (see further in the presentation)

CEITEC

Analytical Tools

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

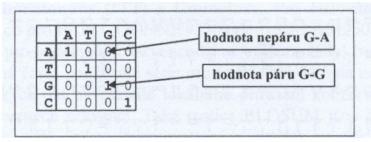


CEITEC

BLAST

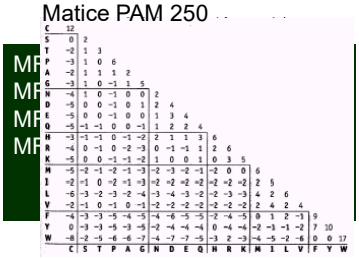
Basic **L**ocal **A**lignment **S**earch **T**ool

- 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



Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250

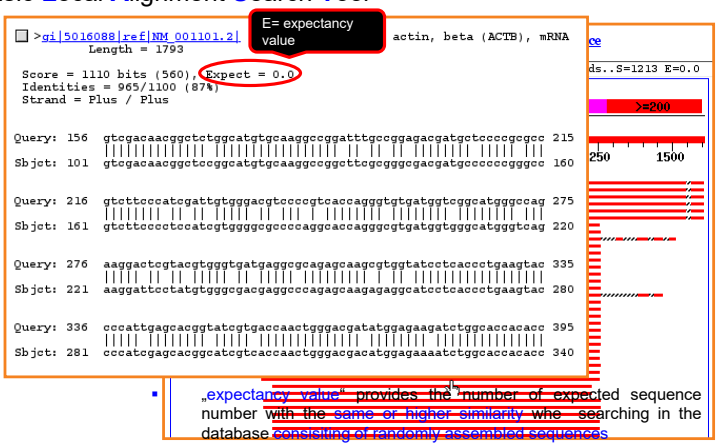


	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1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

CEITEC

BLAST

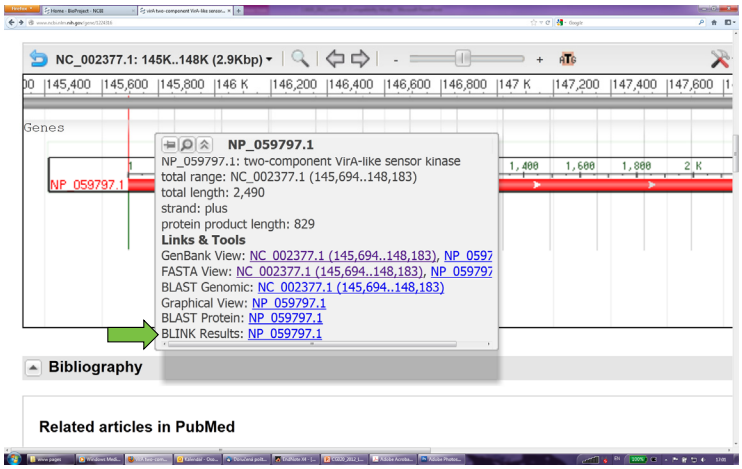
Basic **L**ocal **A**lignment **S**earch **T**ool



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

CEITEC

Primary Databases



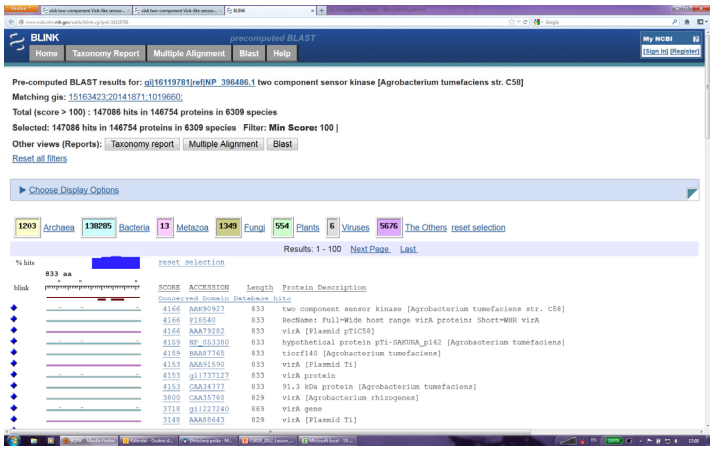
63

CEITEC

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

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool



CEITEC

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.

CEITEC

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 (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.

CEITEC

BLAST

Specialized Versions

- Currently there exists a lot of specialized versions of BLAST

  - **PHI-BLAST** (**P**attern-**H**it **I**nitiated 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

CEITEC

BLAST

Specialized Versions

- Example of search by PHI-BLAST

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAATPRQSLGHPPPEPGPDR
VADAKGDSESEEDEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA
LMYNTPRAATIVATSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIGEK
IYKDGERIITQGEKADSFYIIESGEVSILIRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNISHYEEQLVKMFGSSVDLGNLGQ

[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].
```


CEITEC

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…

CEITEC

Analytical Tools

https://blog.addgene.org/free-online-molecular-biology-tools

Early Career Researcher Toolbox: Free Online Molecular Biology Tools

By Beth Kenkel

Beth Kenkel
September 12, 2023

Share this article

Primer design. Plasmid mapping. DNA sequence analysis. We all have our favorite tools for tackling these particular tasks, but they tend to be scattered about the internet. To help you keep your virtual molecular biology toolbox organized, today's post features a list of free online molecular biology tools all in one place.

Plasmid mapping

These tools are for viewing, editing or making plasmid maps, but can also analyze and annotate any DNA sequence.

- SnapGene Viewer: The free SnapGene Viewer is great for looking at plasmid maps and viewing sequencing traces, while the paid version provides more tools for plasmid mapping and design (Figure 1).
- Benchling: While you might think of Benchling as an electronic lab notebook, it also has a suite of molecular biology tools and can make plasmid maps. Free for academic users.
- Serial Cloner: Free desktop-based software for plasmid design and mapping.
- ApE (A plasmid Editor): A free, donation-based plasmid analysis tool including editing, annotating, creating maps, and more. This tool is maintained by M. Wayne Davis from the University of Utah.

SnapGene

https://www.snapgene.com/snapgene-viewer/download

https://www.youtube.com/watch?v=0sQh2s182WQ

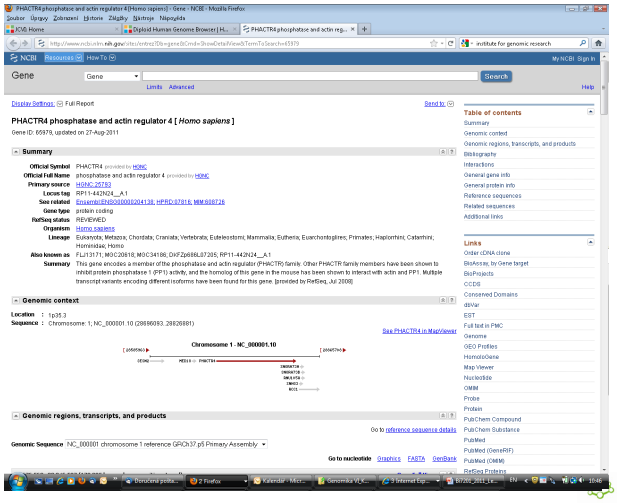
Outline

- 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

CEITEC

Other On-Line Genome Resources

- **TIGR** (**T**he **I**nstitute for **G**enomic **R**esearch, http://www.tigr.org/software/)
  - Recently part of the J. Craig Venter Institute



CEITEC

Other On-Line Genome Resources

- **O**nline **M**endelian **I**nheritance in **M**an (**OMIM**)

CEITEC

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

CEITEC

Discussion