CG020 Genomika

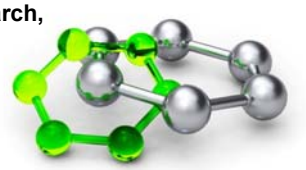
Lesson 1

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

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

CEITEC

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

CEITEC

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

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

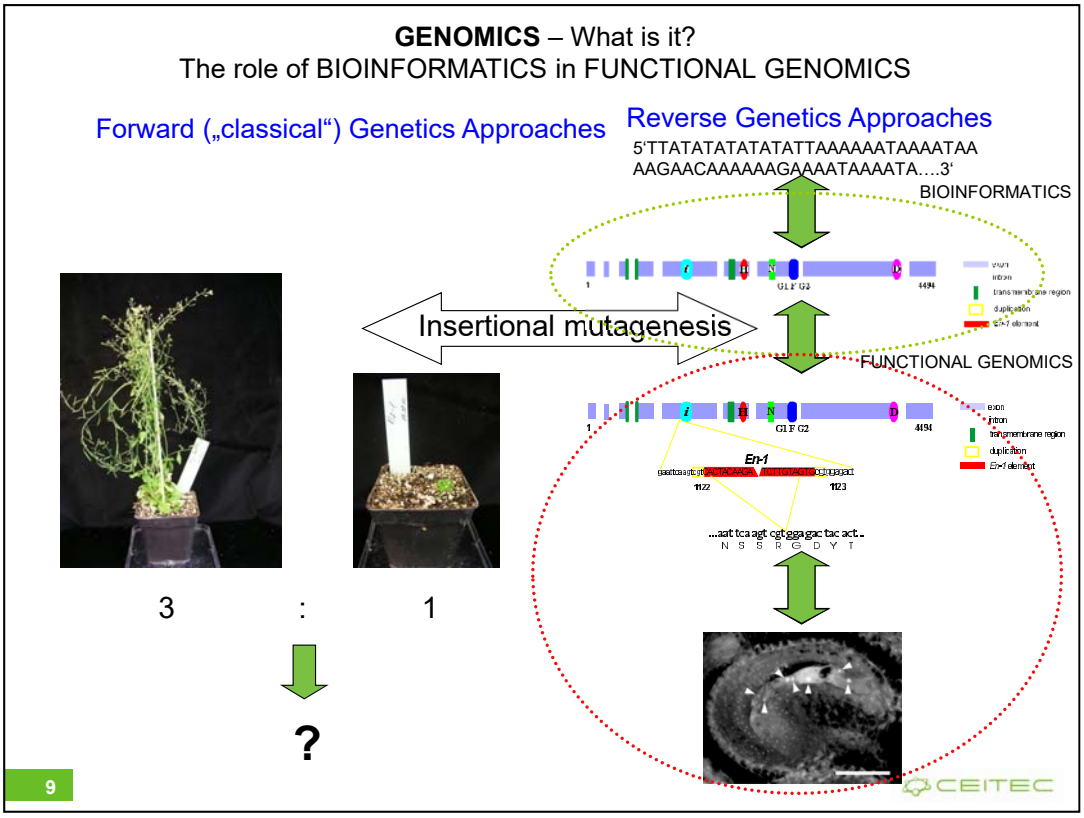
8

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.

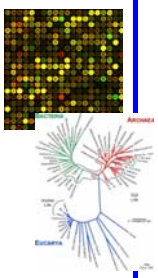
The necessity of having phenotype alterations in the forward genomics approach introduces important difference between those two approaches. Thus, the gene is no longer understood as a factor (*trait*) determining *phenotype,* but rather as a piece of DNA characterized by the unique *string of nucleotides*.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

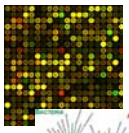
- **Interface** between the **biology** and **computers**
- **Analysis** of **proteins, genes** and **genomes** using **computer algorithms** and **databases**
- **Genomics** is the **analysis** of **genomes**.

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/
BIN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENDUSnet	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/
INN	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.bioi.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://salarjung.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/

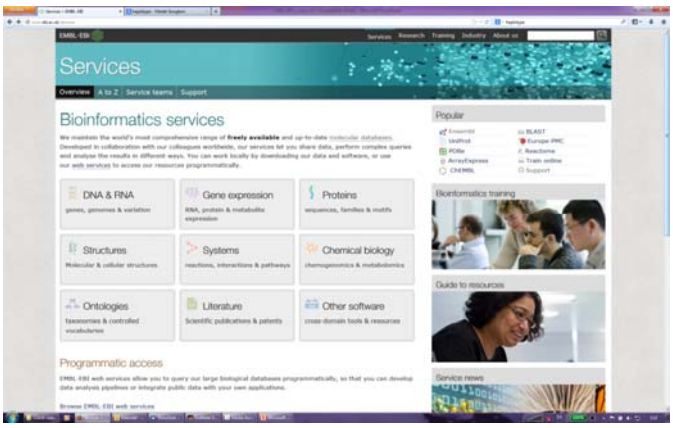
15

CEITEC

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

Spectre of On Line Resources

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



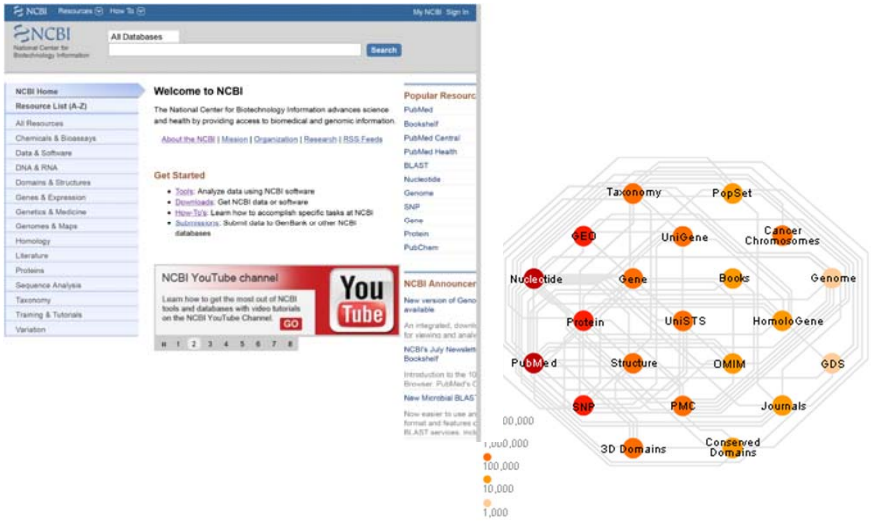
16

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

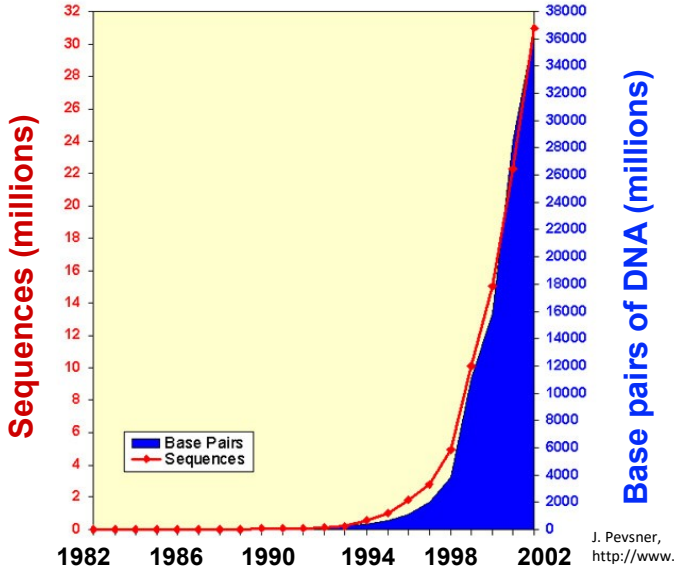
CEITEC

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

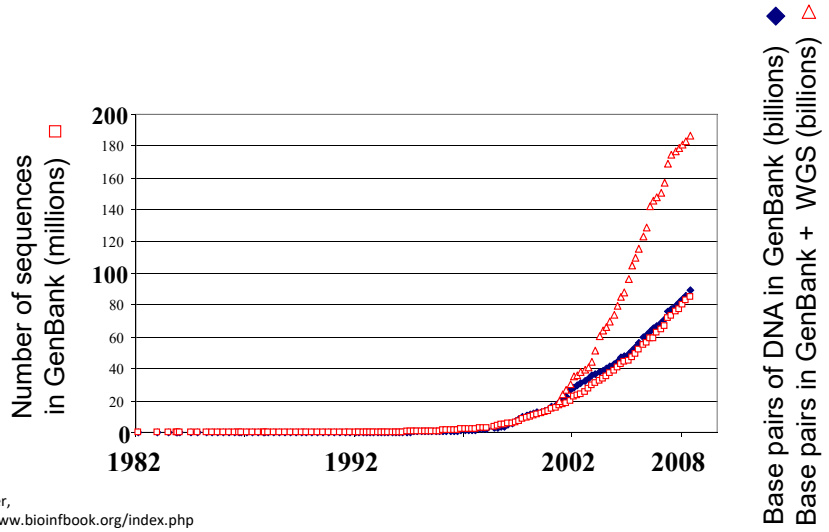
Growth of GenBank



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

CEITEC

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

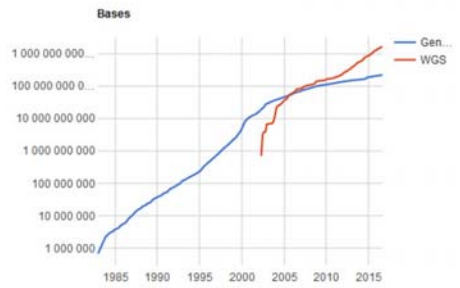


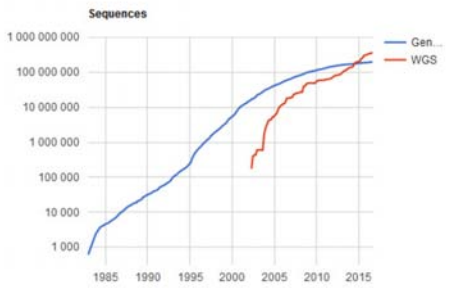
J. Pevsner,
http://www.bioinfbook.org/index.php

CEITEC

Growth of GenBank

Aug 2016

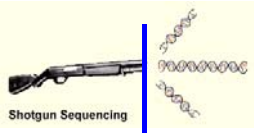




- Dec **1982** 680 338 bp, 606 sequences
- Apr **2002** 19 x 10^9 bp, 17 x 10^6 sequences + WGS 692 x 10^6 bp, 172 768 sequences
- Aug **2016** 218 x 10^9 bp, 196 x 10^6 sequences + WGS 1,6 10^{12} bp, 360 x 10^6 sequences

CEITEC

WGS



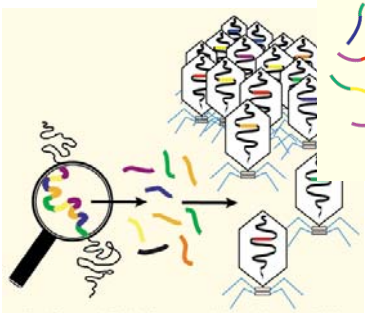


Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

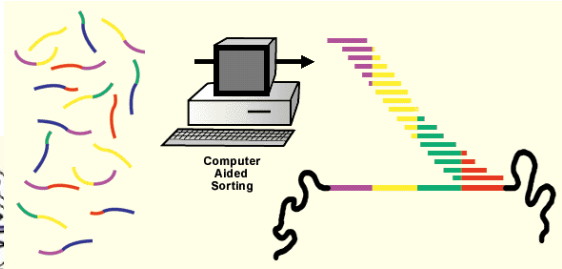


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

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

23

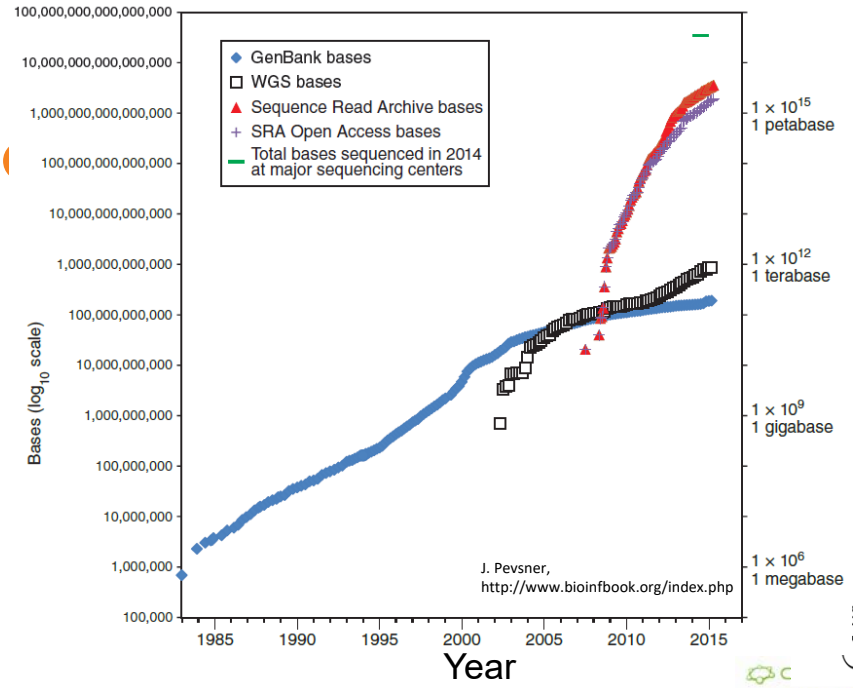
CEITEC

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).

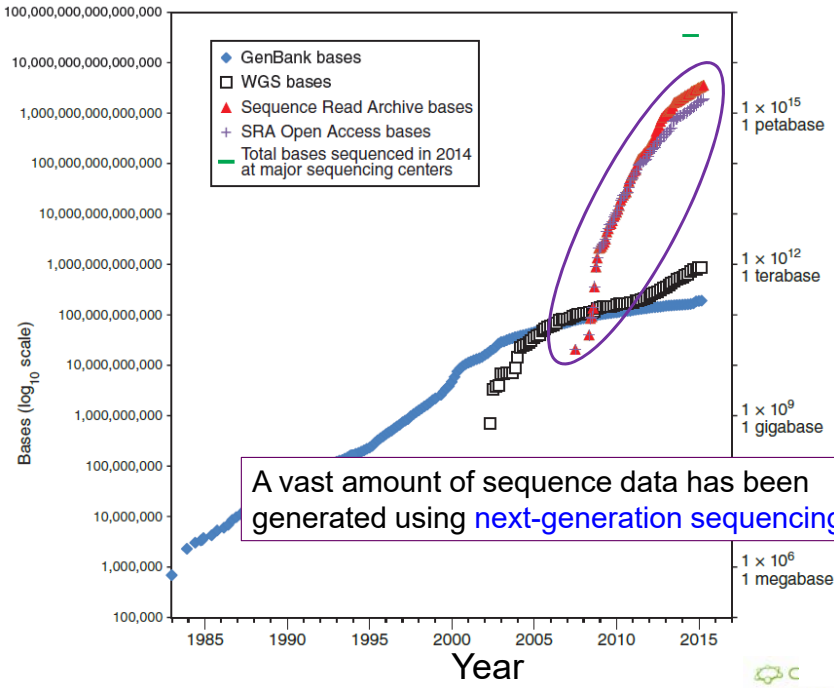
The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence.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

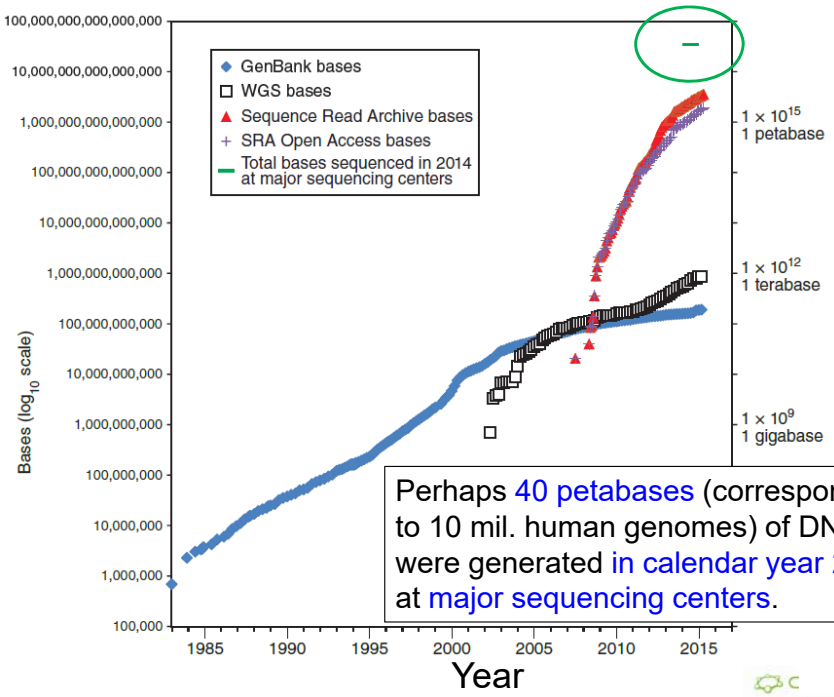


Growth of DNA Sequence in Repositories



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)

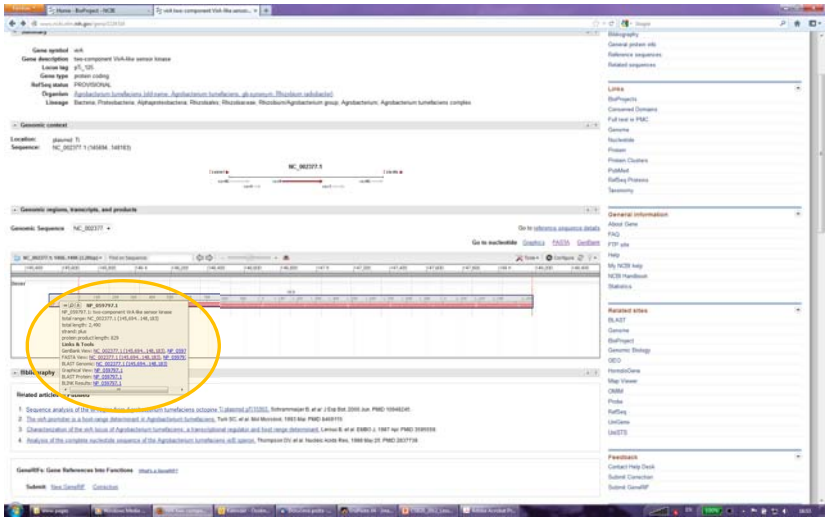
CEITEC

Primary Databases

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

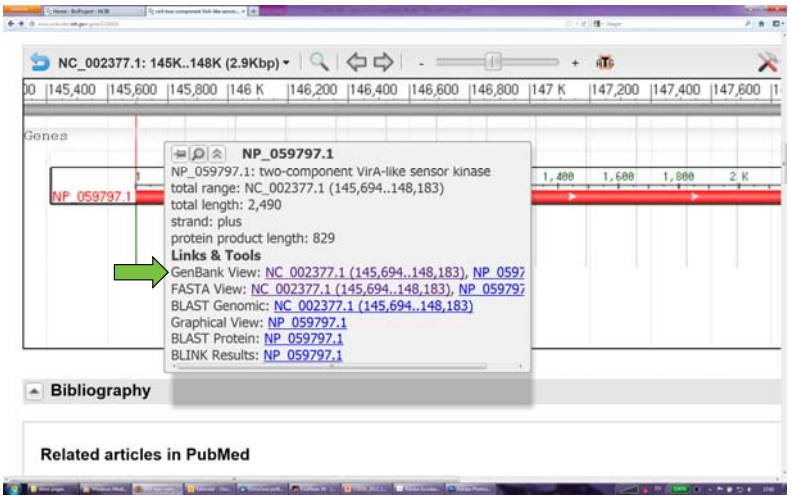
CEITEC

Primary Databases



CEITEC

Primary Databases



CEITEC

Primary Databases

CEITEC

Primary Databases

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.

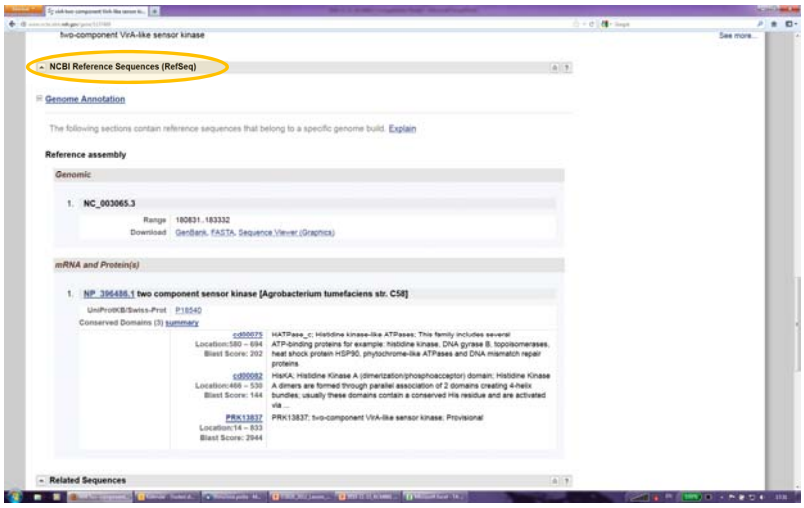
RefSeq identifiers include the following formats:

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

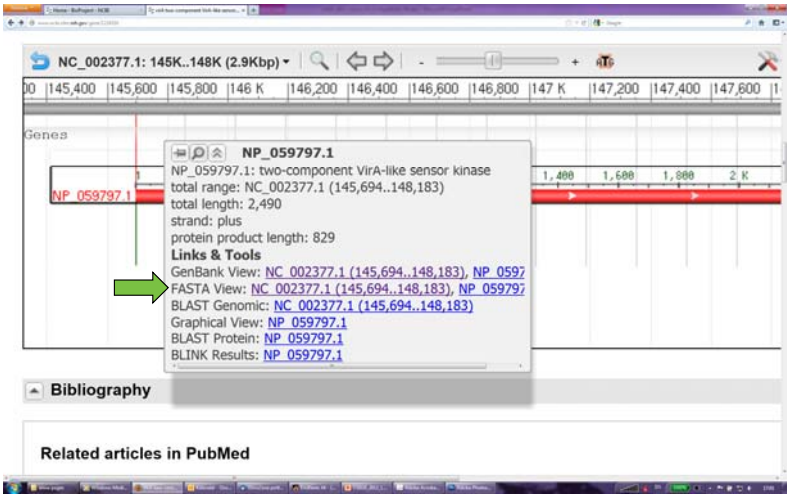
NCBI's RefSeq project: many accession number formats for genomic, mRNA, protein sequences

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

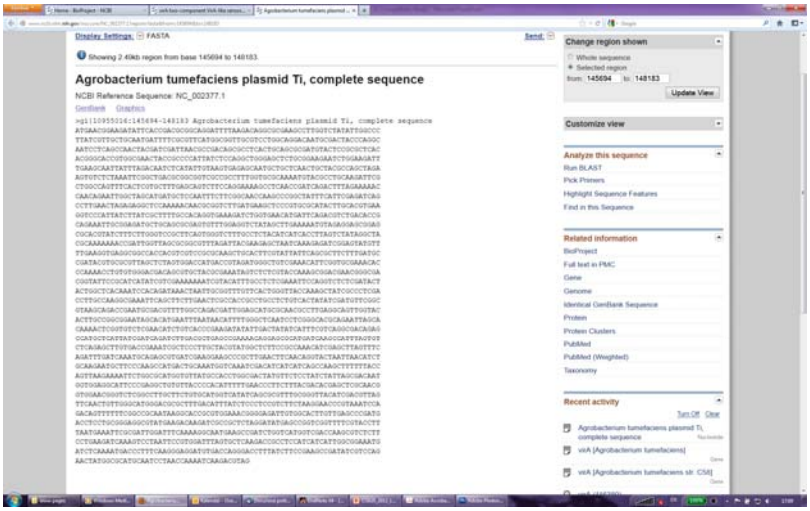
CEITEC

Primary Databases



CEITEC

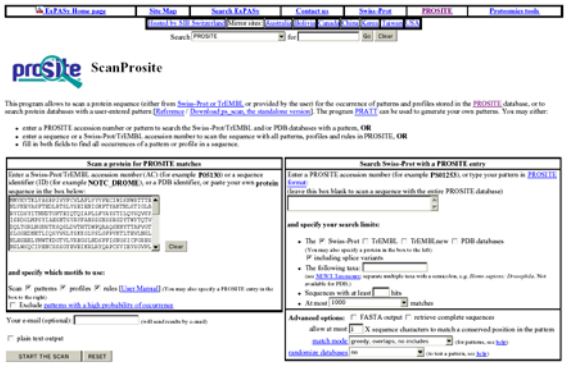
Primary Databases



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/

>PDOC00001 PS00001 **SULFATION** Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].

171 - 185 nkeeaatYeniaaa

>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 KRsS

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

149 - 151 SaR
164 - 166 TyR
171 - 173 SlR
219 - 221 SkR
369 - 371 TvR
460 - 462 SyR
513 - 515 SgR
585 - 587 SlR
402 - 404 TyR
412 - 414 TdK
716 - 718 SpR
726 - 728 SpR
747 - 749 TrR
794 - 796 SaR
854 - 856 SvK
864 - 866 SlR
868 - 870 SaR
921 - 923 SyR
957 - 959 SVR
960 - 962 TyR
974 - 976 TwR
997 - 999 SvK
1002 - 1004 TyR
1018 - 1020 SgR
1031 - 1033 TyR
1119 - 1121 SWR

41

CEITEC

Secondary Databases

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

CEITEC

Secondary Databases

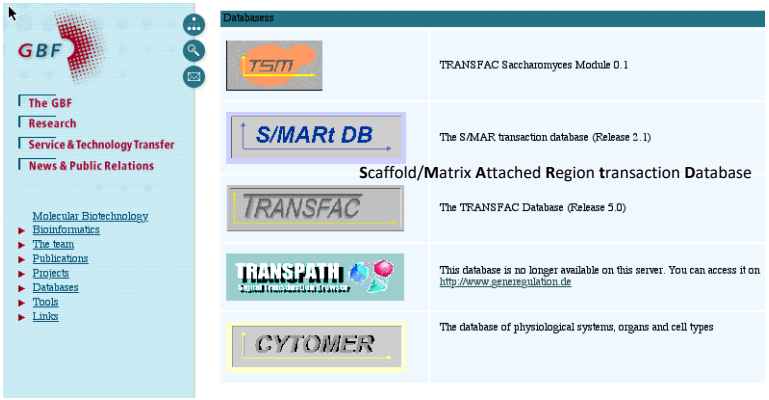
- Databases of **functional** or **structural** ***motifs,*** acquired by primary data (sequences) comparison

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

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)

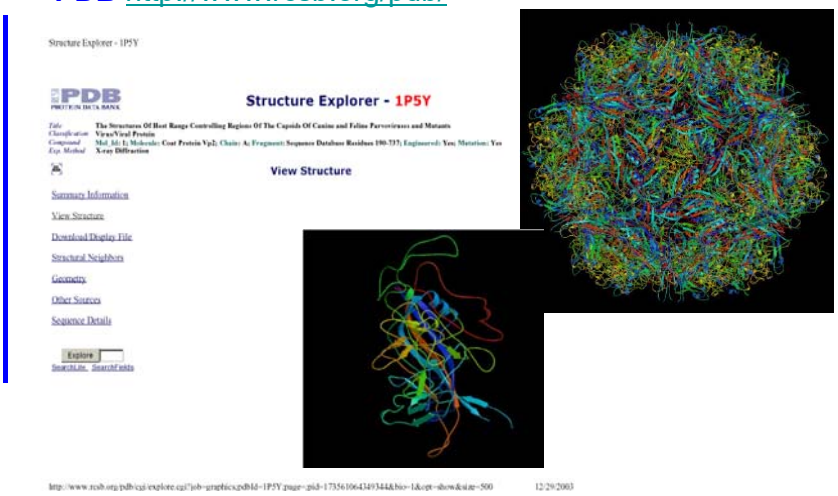
Structural Databases

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

CEITEC

Structural Databases

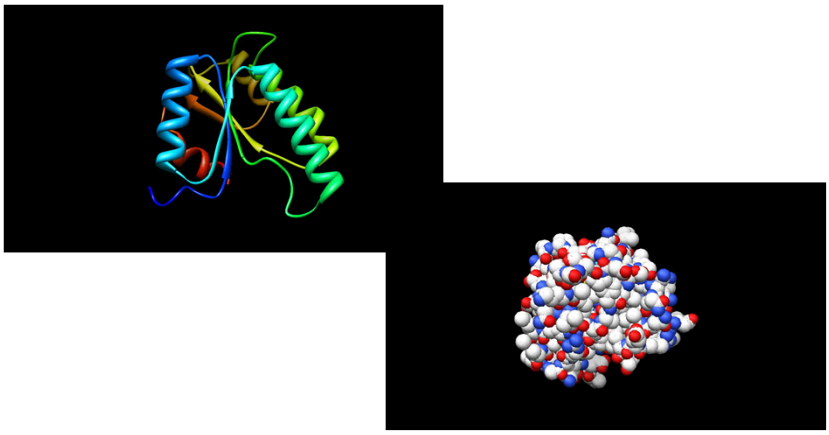
- **PDB** http://www.rcsb.org/pdb/



CEITEC

Structural Databases

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



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

CEITEC

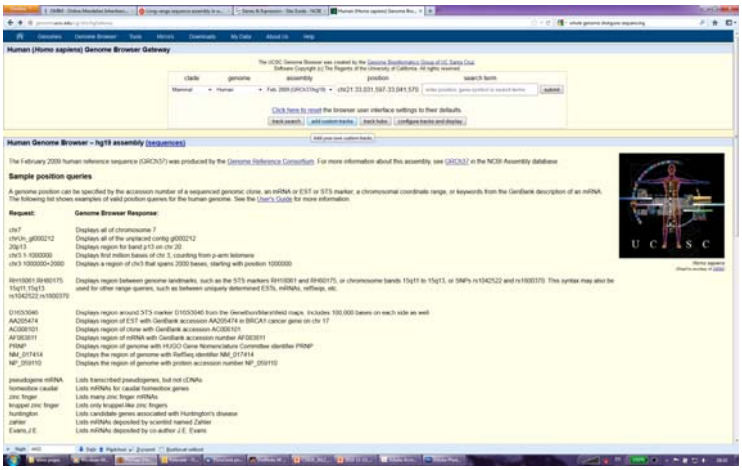
Outline

- 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

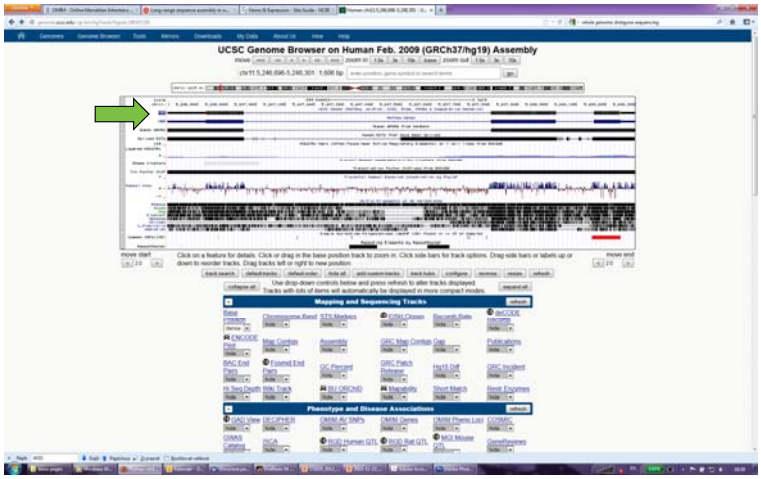
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genome Resources

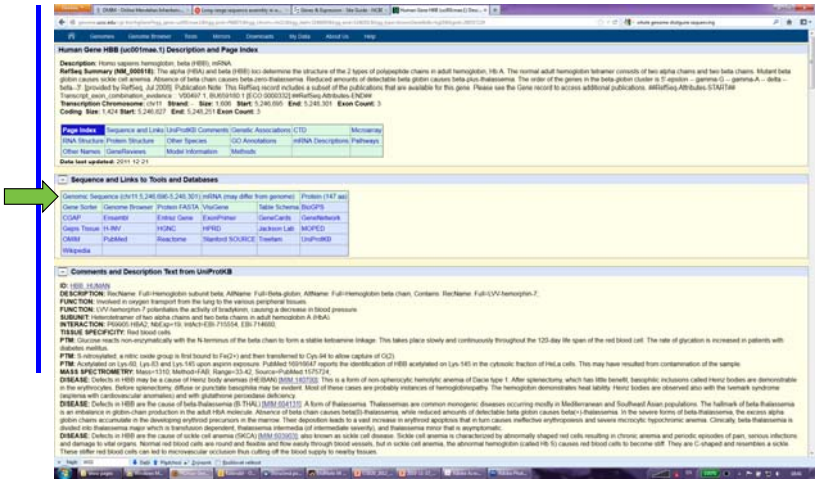
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genome Resources

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



CEITEC

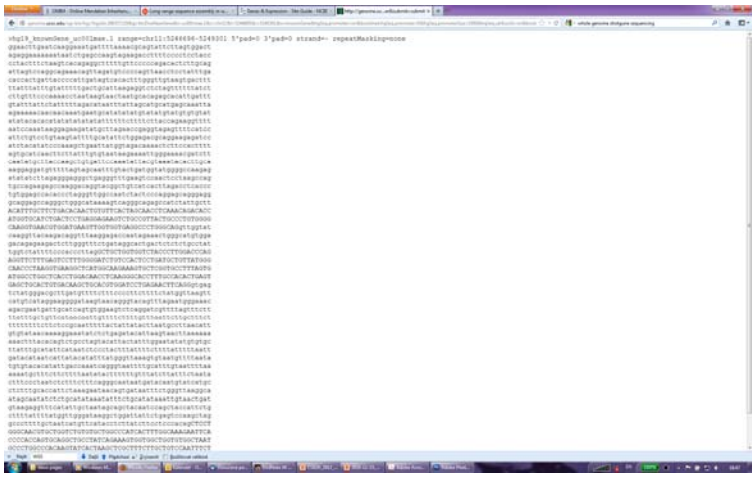
Genome Resources

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

CEITEC

Genome Resources

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



CEITEC

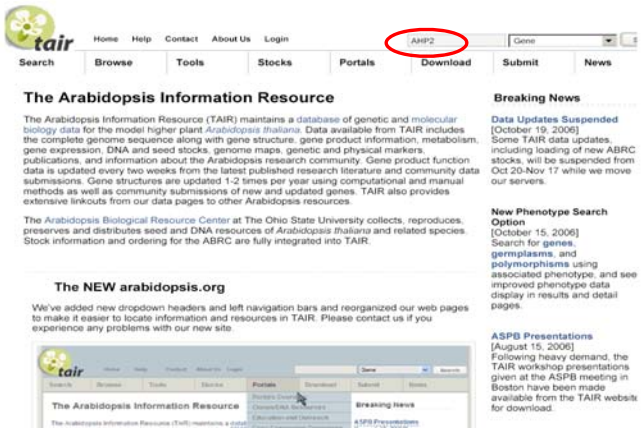
Genome Resources

- 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-------------
```

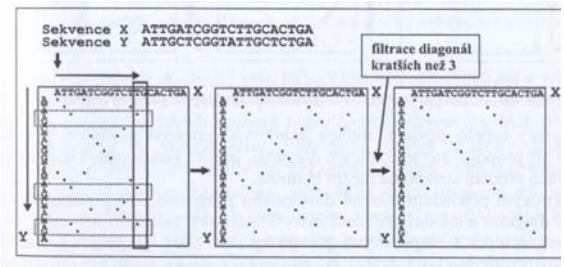
Cvrčková, Úvod do praktické bioinformatiky

- 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



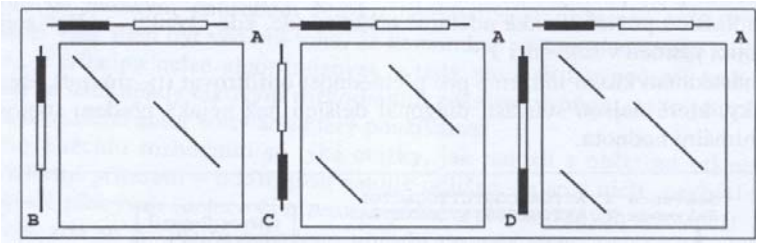
Cvrčková, Úvod do praktické bioinformatiky

  - 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



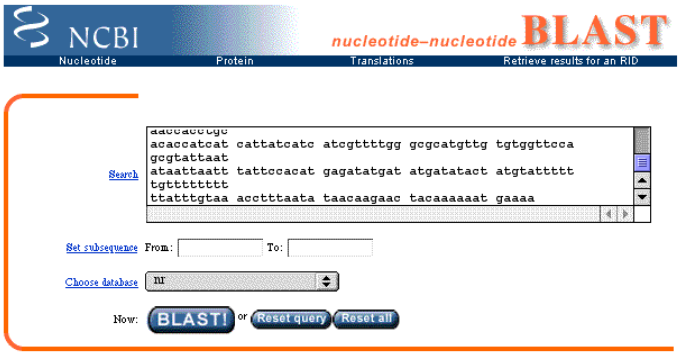
Cvrčková, Úvod do praktické bioinformatiky

- **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/

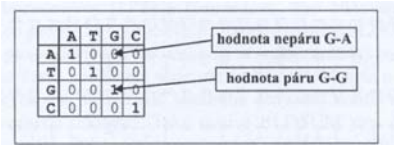


CEITEC

BLAST

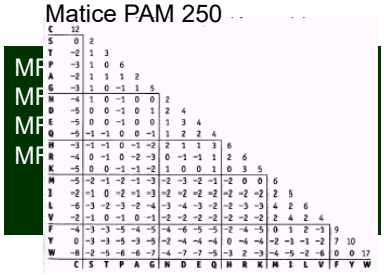
Basic **L**ocal **A**lignment **S**earch **T**ool

- 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



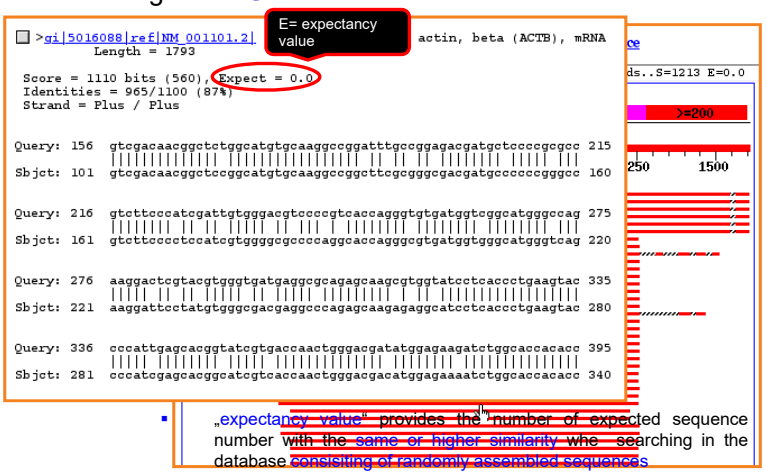
Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250



BLAST

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

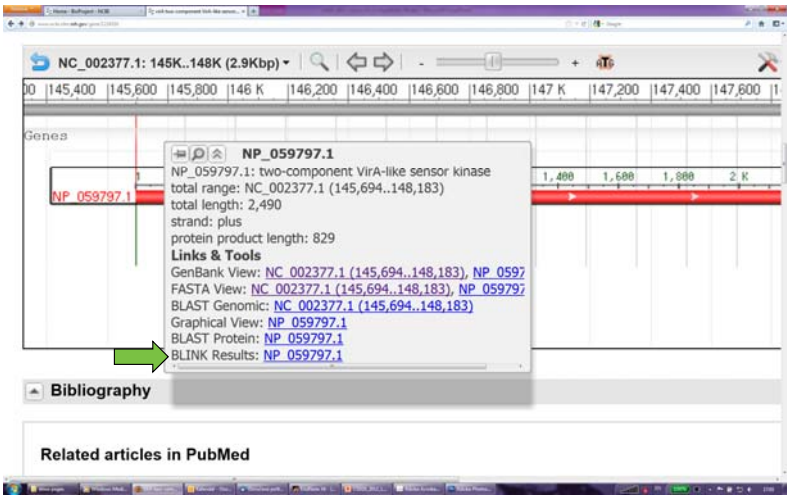


- „expectancy value" provides the number of expected sequence number with the same or higher similarity whe searching in the database consisting of randomly assembled sequences

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

CEITEC

Primary Databases



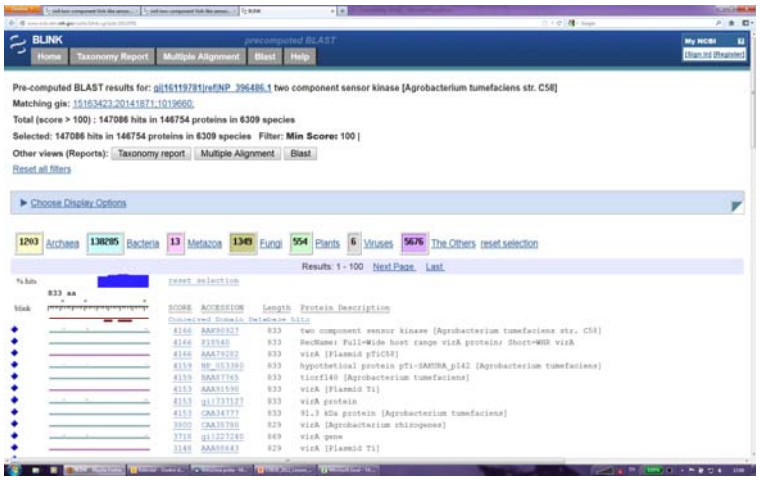
63

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

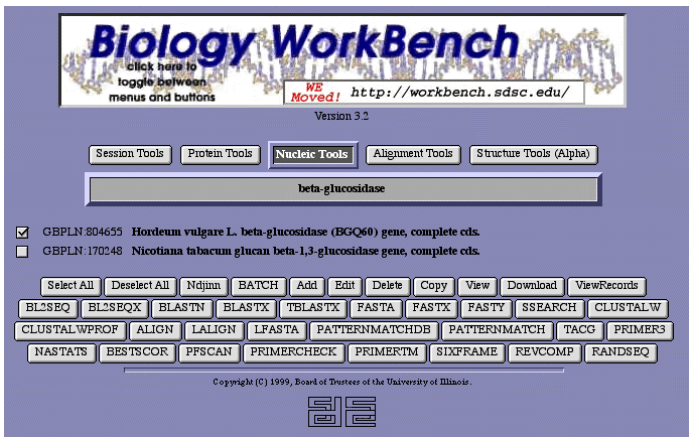
Outline

- 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

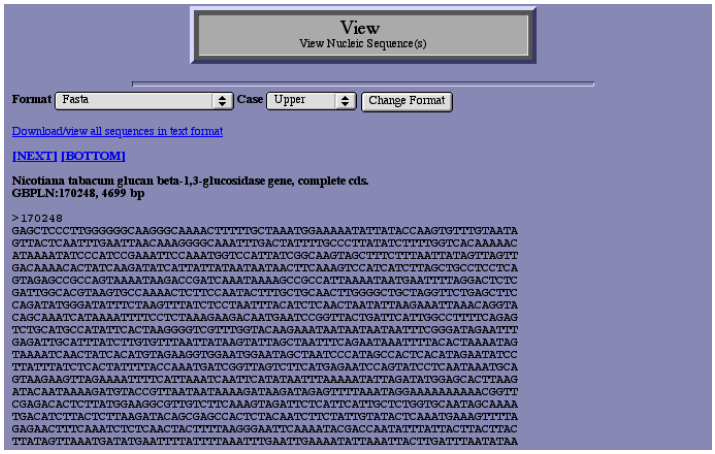
- http://workbench.sdsc.edu/



CEITEC

Analytical Tools

- http://workbench.sdsc.edu/



CEITEC

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
GAGCTCCCTTGGGGGGCAAGGGCAAAACTTTTTTGCTAAATGGAAAAATATTATACCAAGTGTTTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAATTTGACTATTTTGCCCTTATATCTTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTTCTTTAATTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAATAATAACTTCAAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAATAAGACCGATCAAATAAAAGCCGCCATTAAAATAATGAATTTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACTCTTCCAATACTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTAAACAGGTA
CAGCAAATCATAAAATTTTTCCTCTTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCATGCCATATTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAATAATTTCGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTAATTATAAGTATTAGCTAATTTCAGAATAAATTTTACACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTTACCAAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAATTTTCATTAAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTTATGGAAGGCGTTTCTTCTTCAAAGTAGATTCTCATTTCATTGCTCTGGTGCAATAGCAAAA
TGACATCTTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTTA
GAGAACTTTTCAAATCTCTCAACTACTTTTTAAGGGAATTCAAAATACGACCAATATTTATTACTTACTTAC
TTTATAGTTAAATGATATGAATTTTTATTTTTAAATTTTGAATTGAAAATATTAAATTACTTGATTTAATATAA
ACAATAGATATCGCTAAGTATTTACCACAAACATGGAGATACTACAGAAGATTTTTATTATTTTGTAACGAT
GATTAAGCAGCTATTCATCTGGTTGTGCAGGATGAAAGAAAGTAACTAGCTATAATTTCTTTTGTAAAGT
```


CEITEC

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
ELPWGARAKLFAKWKNIIPSVCNSYSI*INKGANLTILPL

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

Frame 2, 1 stop codon

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

```
>170248 Translated - Frame 2
SSLGGQGQNFLLNGKILYQVFVIVTQFELTKGQI*LFCP

    S  S  L  G  G  Q  G  Q  N  F  L  L  N  G  K  I  L  Y  Q  V
2   agctcccttgggggcaagggcaaaacttttgctaaatggaaaaatattataccaagtg  61
    F  V  I  V  T  Q  F  E  L  T  K  G  Q  I  *  L  F  C  P
62  tttgtaatagttactcaatttgaattaacaaaggggcaaatttgactattttgccctta  120
```


CEITEC

Analytical Tools

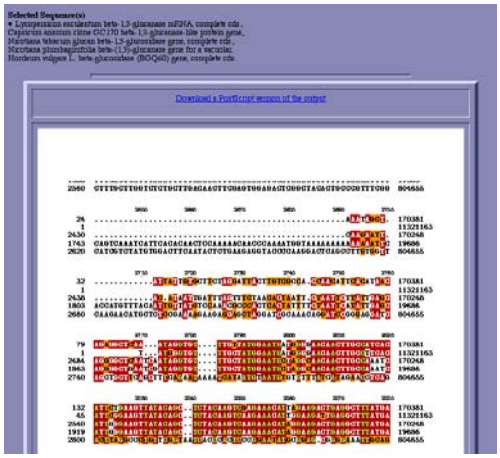
- http://workbench.sdsc.edu/

```
== Linear Map of Sequence:

                StyI
                BsaJI
            CviJI
            AluI
              SacI
            EcoICRI
              Bsp1286I
              BsiHKAI
              BanII  BslI                                SspI
            \ \ \    \                                   \
     1    gagctcccttgggggcaagggcaaaactttttgctaaatggaaaaatattataccaagt    60
          ctcgagggaacccccgttcccgttttgaaaaacgatttaccttttttataatatggttca
            ^    *    ^    *    ^    *    ^    *    ^    *    ^    *
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    gtttgtaatagttactcaatttgaattaacaaaggggcaaatttgactattttgccctta    120
          caaacattatcaatgagttaaacttaattgtttccccgtttaaactgataaaacgggaat
            ^    *    ^    *    ^    *    ^    *    ^    *    ^    *
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
```


CEITEC

Analytical Tools

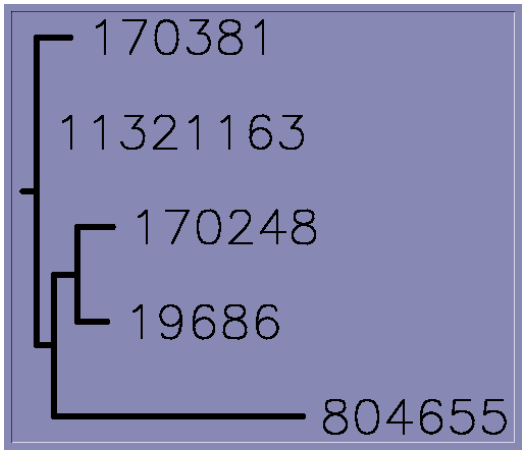
- http://workbench.sdsc.edu/



CEITEC

Analytical Tools

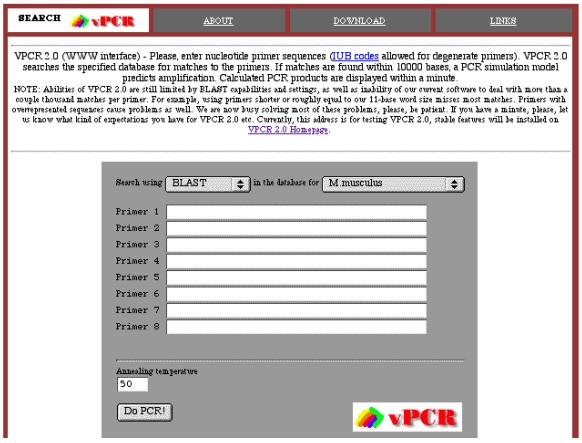
- http://workbench.sdsc.edu/



CEITEC

Analytical Tools

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



CEITEC

Analytical Tools

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

CEITEC

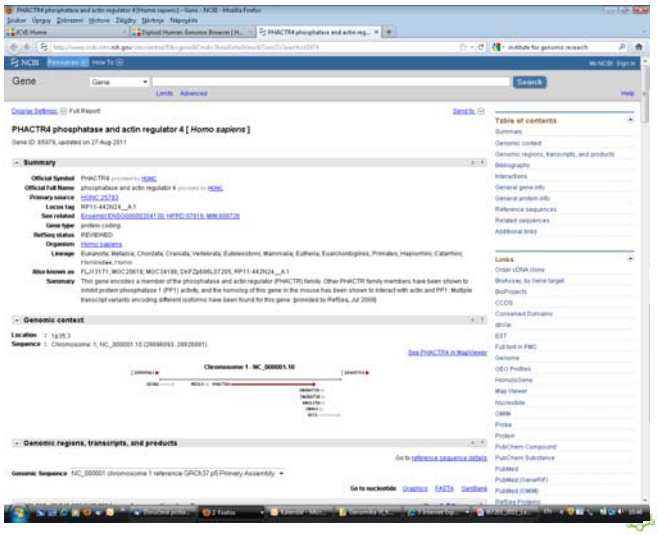
Outline

- 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…
  - 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



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