CG020 Genomika

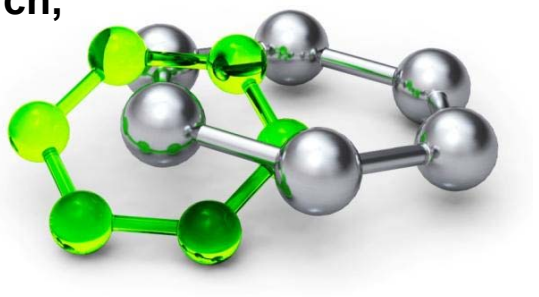
Lesson 1

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

Course Syllabus

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

Course Syllabus

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

Course Syllabus

- **Chapter 09**
  - Genome evolution
- **Chapter 10**
  - Genomics and Systems Biology
- **Chapter 11**
  - Practical Aspects Of Functional Genomics
  - Model Organisms,
  - PCR and Primer Design

Literature

- Literature resources for Chapter 01:
  - **Bioinformatics and Functional Genomics**, 3rd Edition, Jonathan Pevsner, Wiley-Blackwell, 2015 http://www.bioinfbook.org/php/?q=book3
  - **Úvod do praktické bioinformatiky**, Fatima Cvrčková, 2006, Academia, Praha
  - **Plant Functional Genomics**, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey

Outline

- Syllabus of thecourse
- Definition of Genomics

GENOMICS – What is it?

- *Sensu lato* (in the broad sense) – it is interested in **STRUCTURE and FUNCTION** of genomes
  - Necessary prerequisite: knowledge of the genome (sequence) – work with databases

- *Sensu stricto* (in the narrow sense) – it is interested in FUNCTION of INDIVIDUAL GENES – **FUNCTIONAL GENOMICS**
  - It uses mainly the reverse genetics approaches

GENOMICS – What is it?
The role of BIOINFORMATICS in FUNCTIONAL GENOMICS

Forward („classical“) Genetics Approaches

Reverse Genetics Approaches

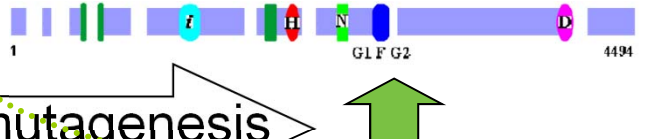
5‘TTATATATATATATTAAAAAATAAAATAA
AAGAACAAAAAAGAAAATAAAATA….3‘

3 : 1

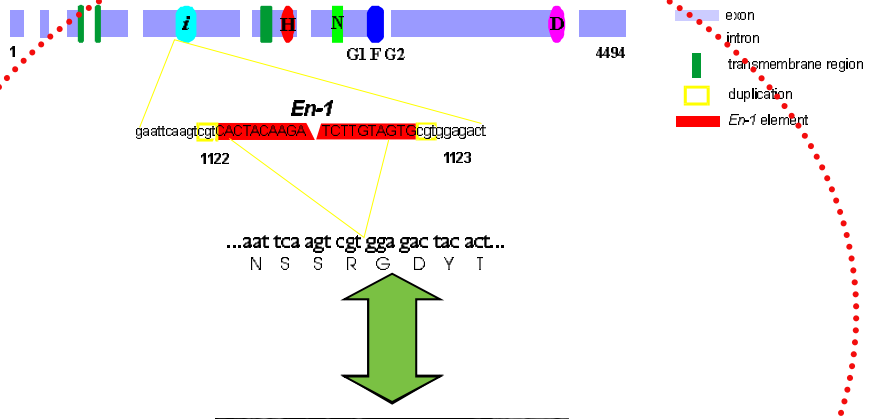
?

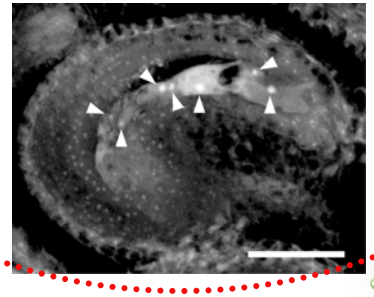
Insertional mutagenesis

BIOINFORMATICS



FUNCTIONAL GENOMICS

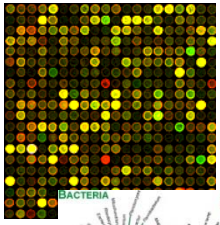


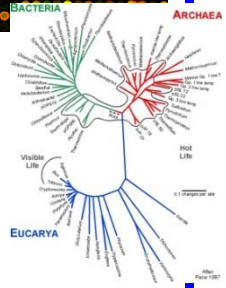


Outline

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

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.**

CEITEC

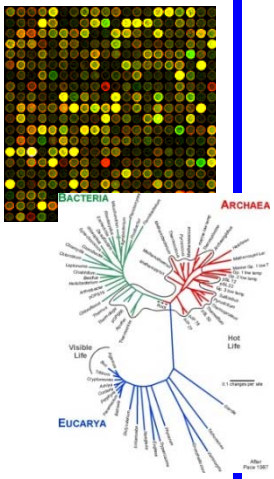
What is bioinformatics?

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

The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.

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

Bioinformatics



- **Bioinformatics** in **functional genomics**
  - **Processing and analysis of sequencing data**
    - Identification of reference sequences
    - Identification of genes
    - Identification of homologues, orthologues and paralogues
    - Correlative analysis of genomes and phenotypes (incl. human)
  - **Processing and analysis of transcriptional data**
    - Transcriptional profiling using DNA chips or next-gen sequencing
  - **Evaluation of experimental data and prediction of new regulations in systems biology approaches**
    - Mathematical modelling of gene regulatory networks

CEITEC

Outline

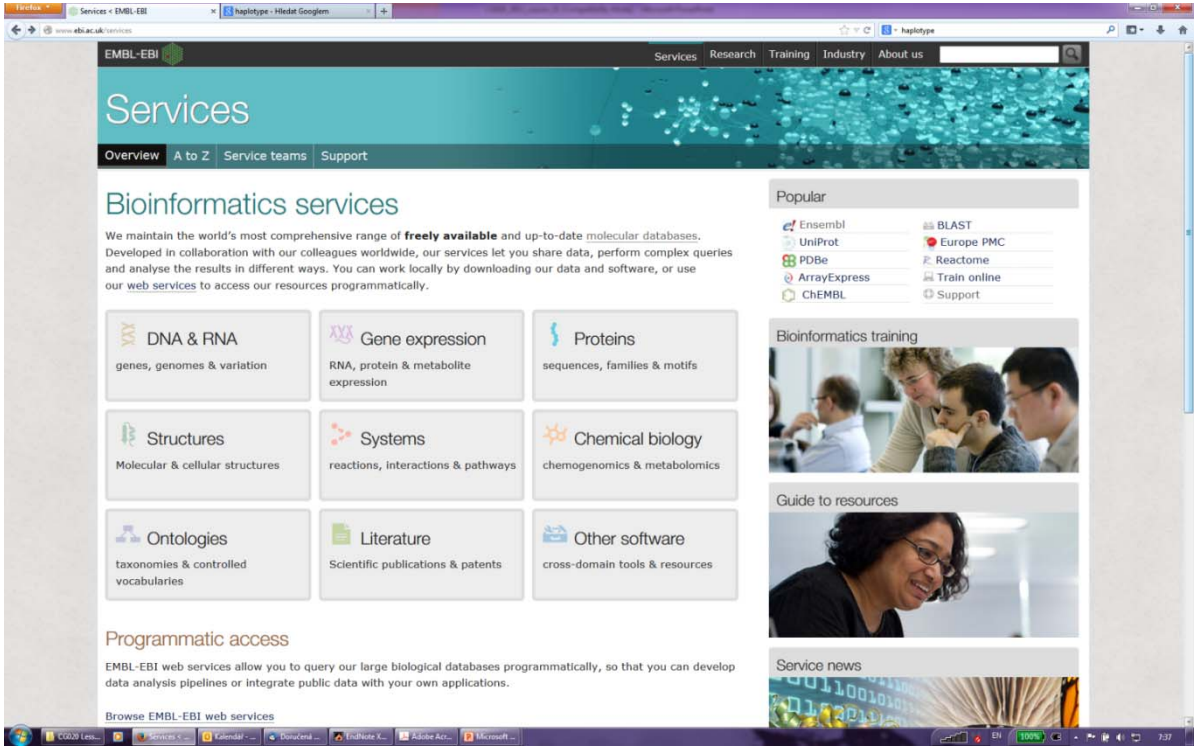
- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS
- Databases
  - Spectre of „on-line“ resources

Spectre of On Line Resources

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/
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.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://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/

Spectre of On Line Resources

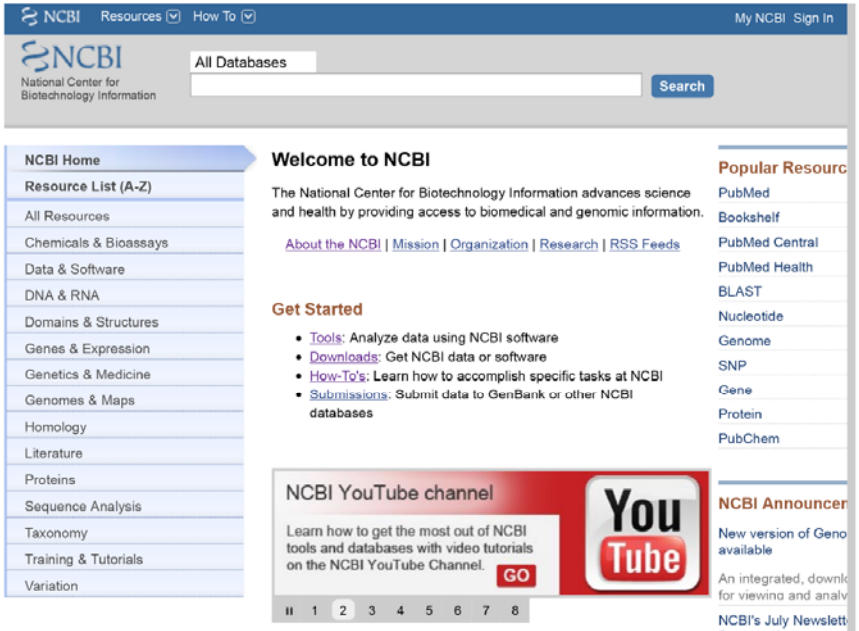
- EBI http://www.ebi.ac.uk/services

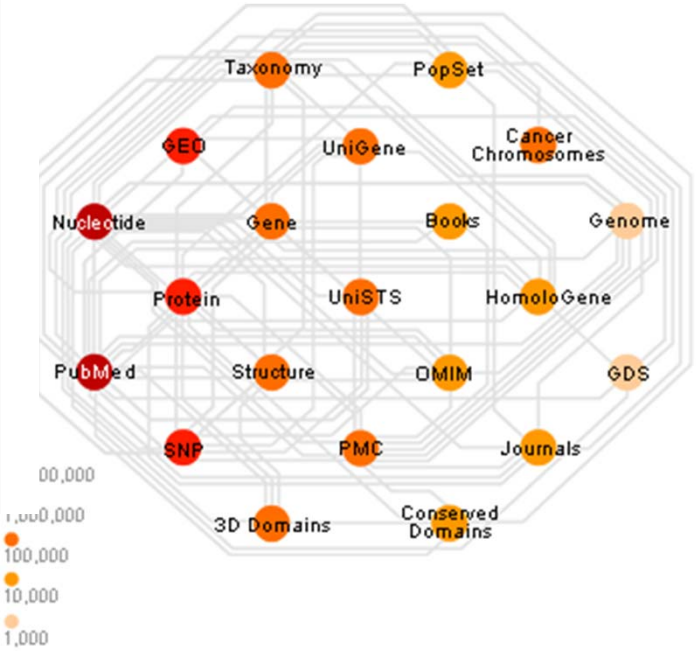


CEITEC

Spectre of On Line Resources

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





CEITEC

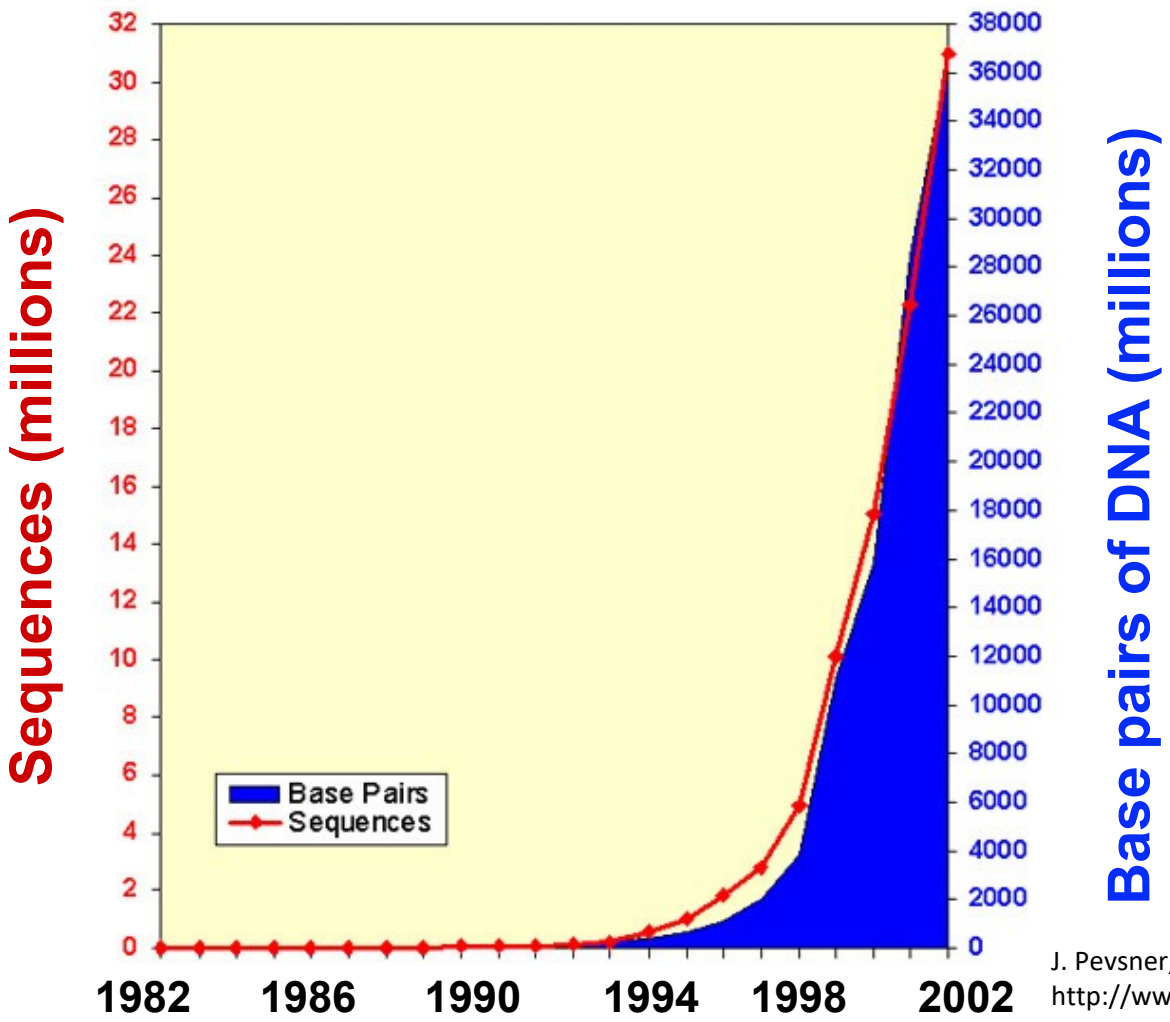
Outline

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

Primary Databases

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

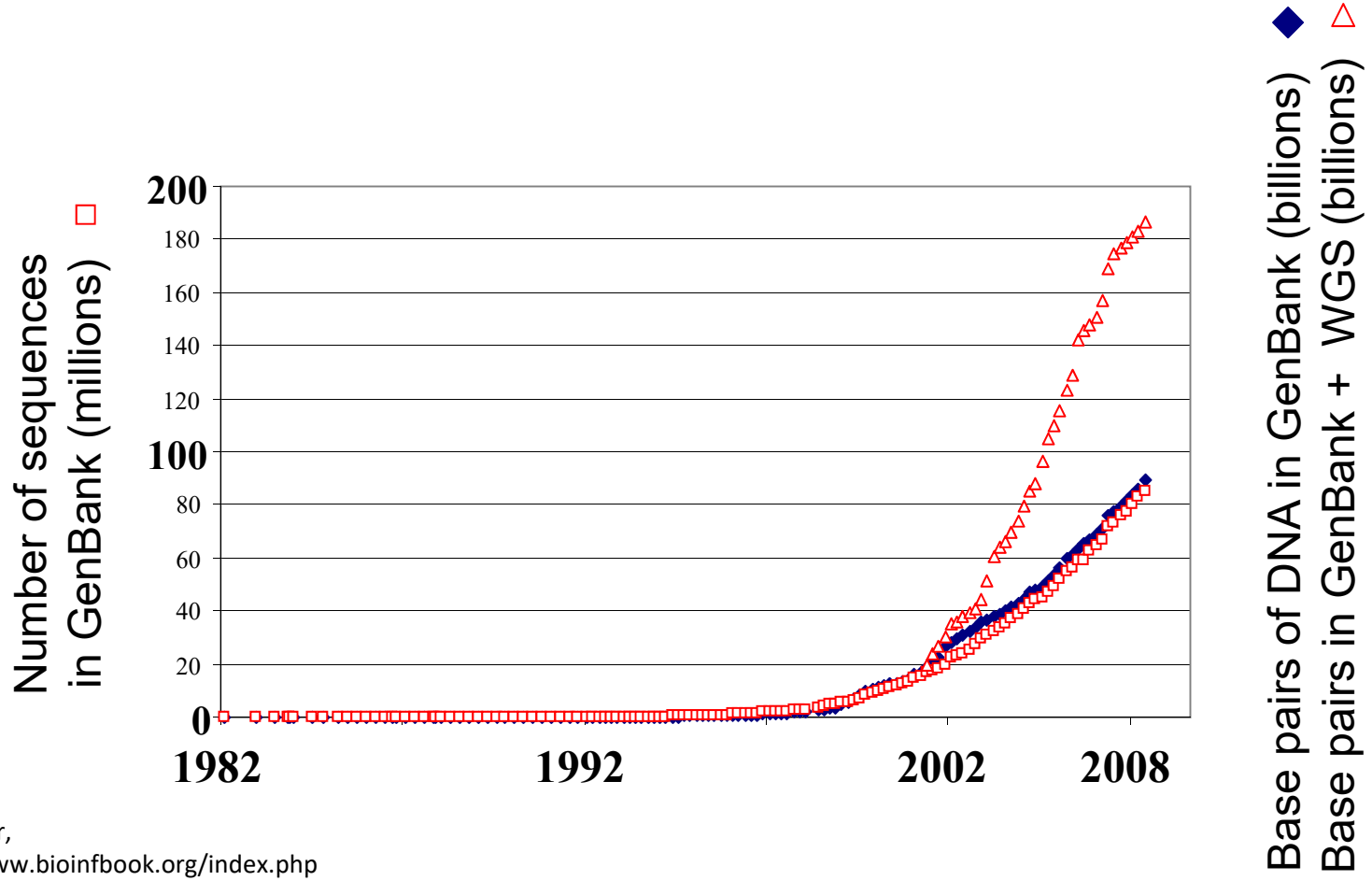
Growth of GenBank



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

CEITEC

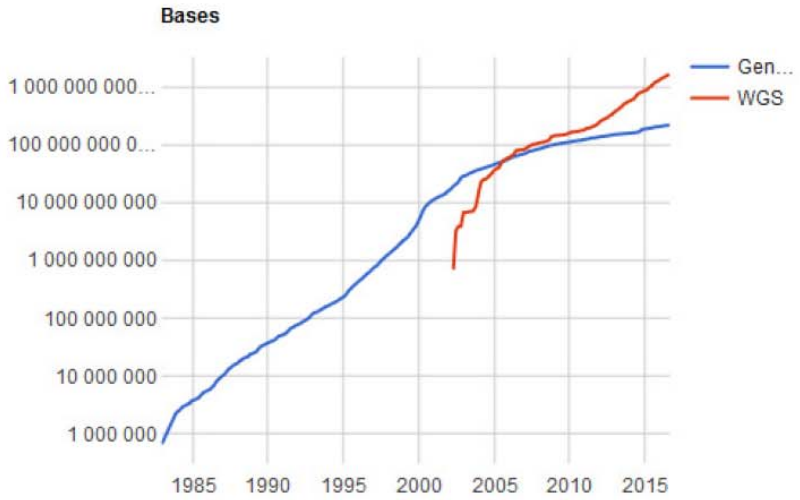
Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases

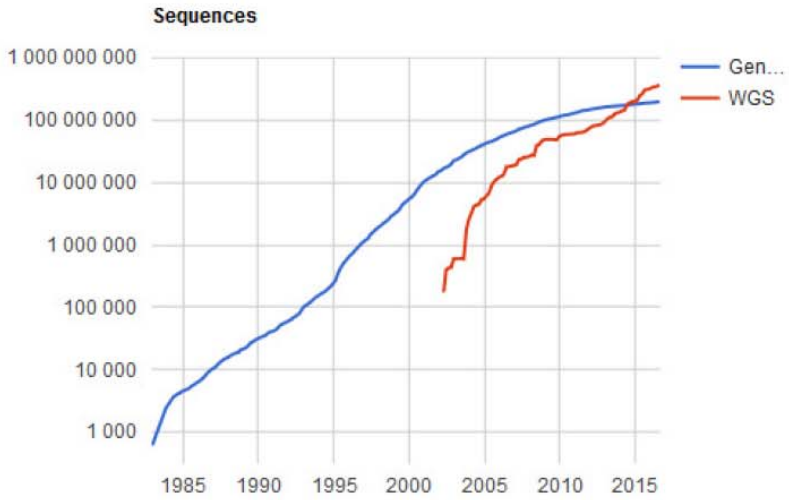


J. Pevsner,
http://www.bioinfbook.org/index.php

Growth of GenBank

Aug 2016

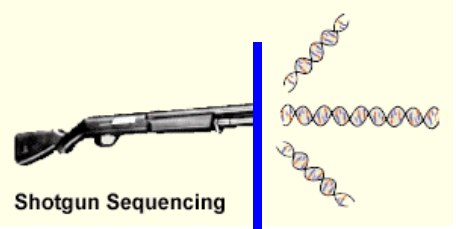




- Dec **1982** 680 338 bp, 606 sequences
- Apr **2002** 19 x 10^9 bp, 17 x 10^6 sequences + WGS 692 x 10^6 bp, 172 768 sequences
- Aug **2016** 218 x 10^9 bp, 196 x 10^6 sequences + WGS 1,6 10^{12} bp, 360 x 10^6 sequences

CEITEC

WGS



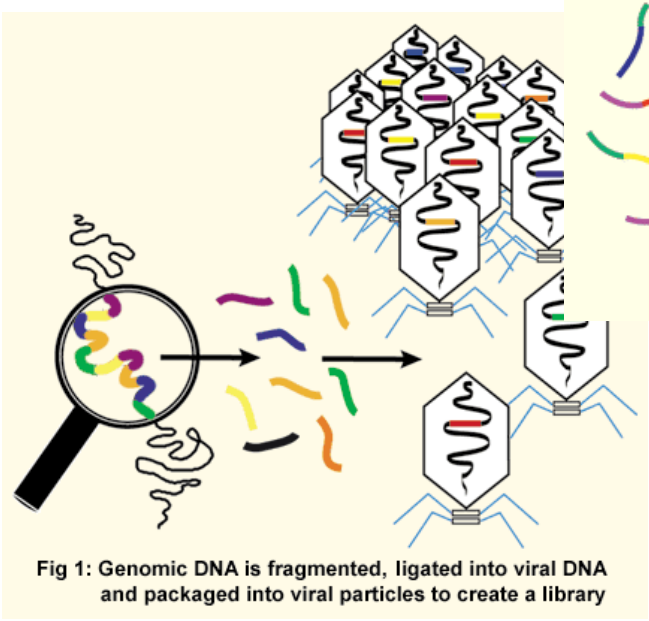


Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

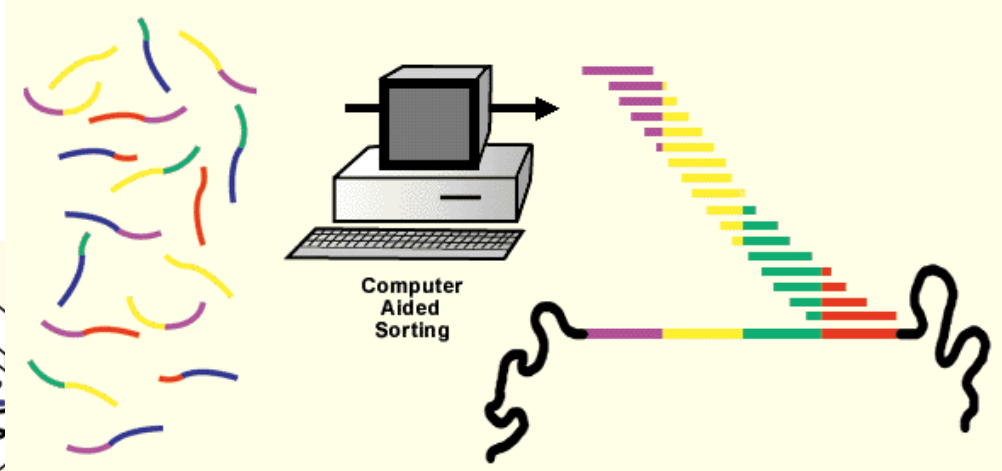
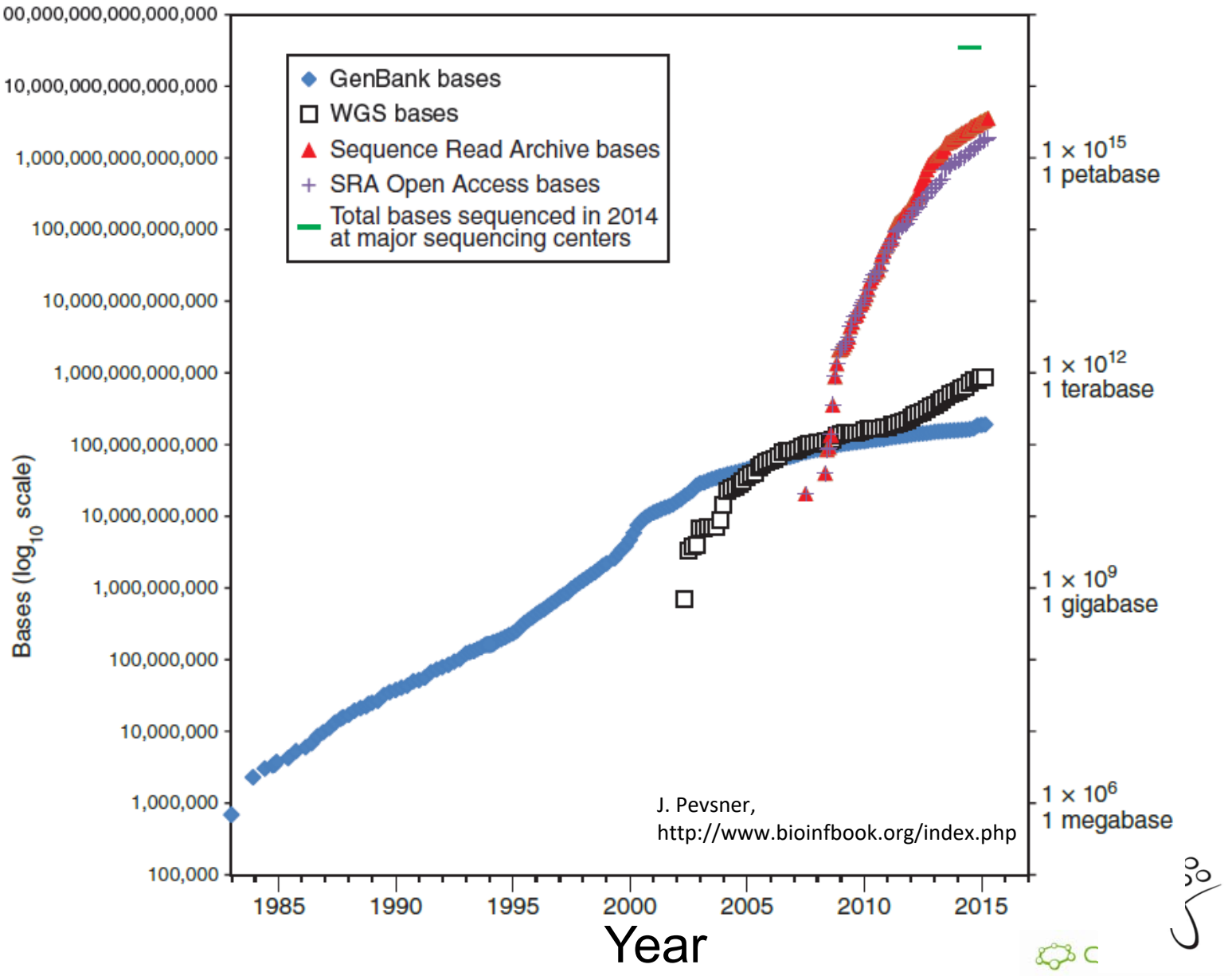


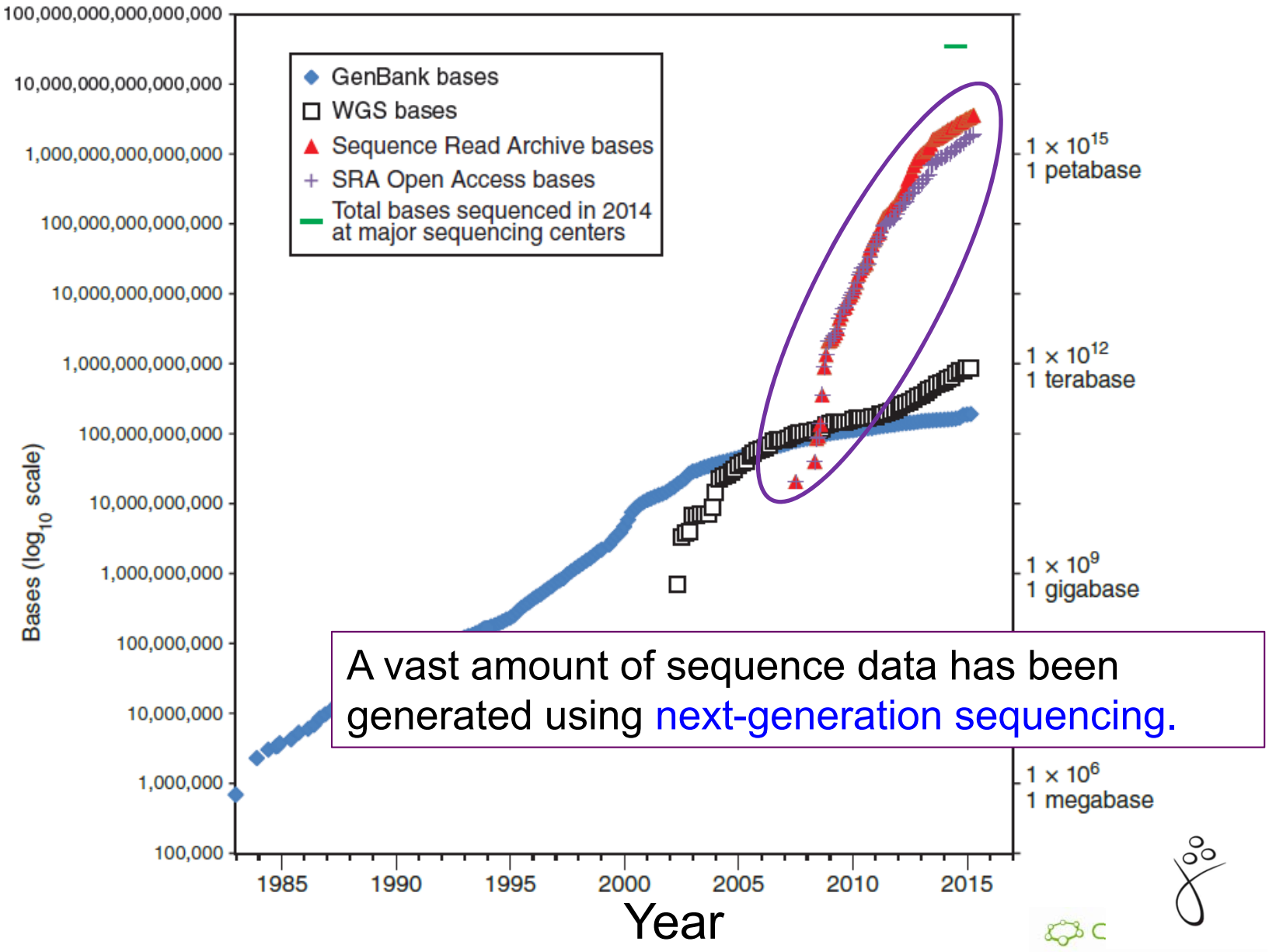
Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

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

Growth of DNA Sequence in Repositories

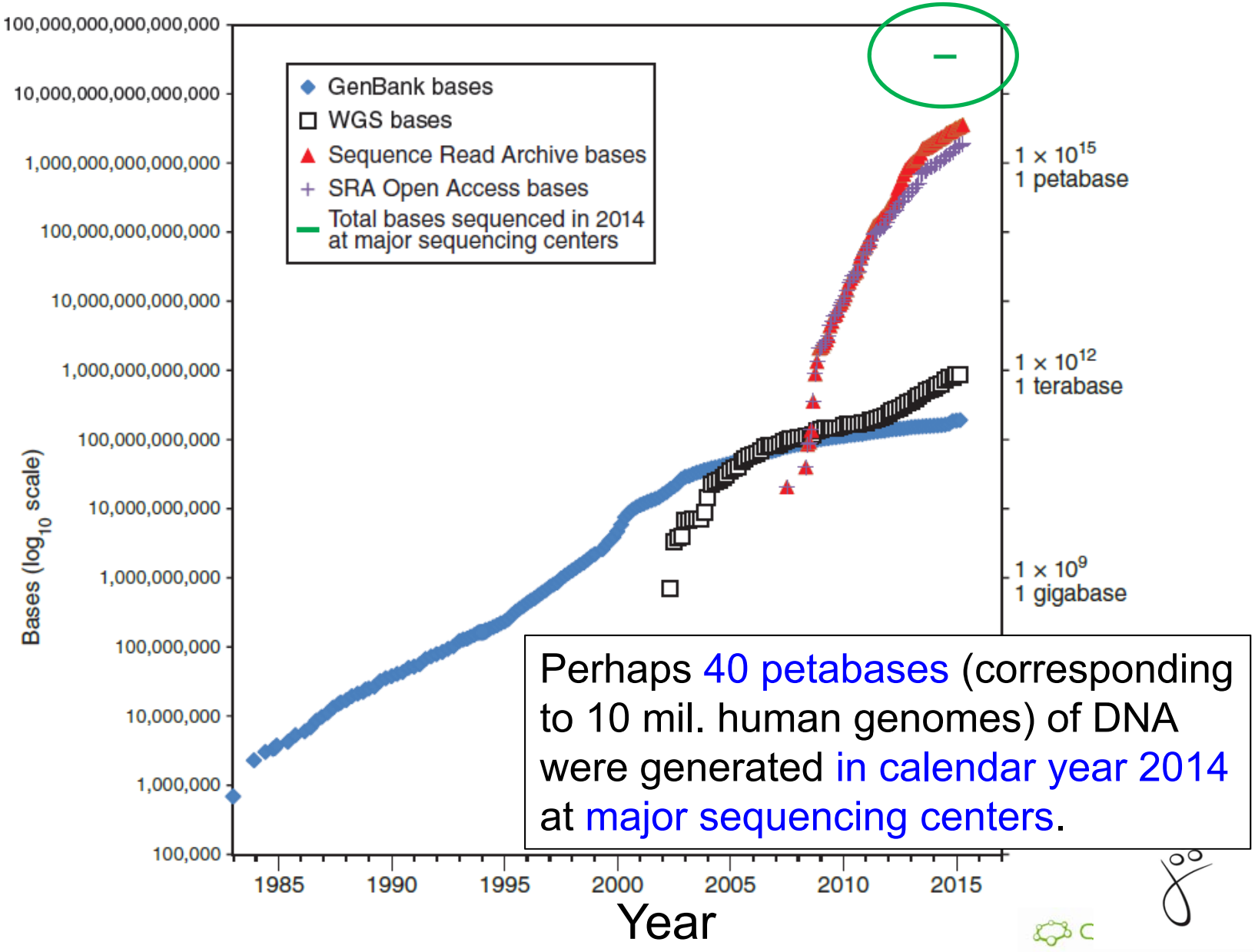


Growth of DNA Sequence in Repositories



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

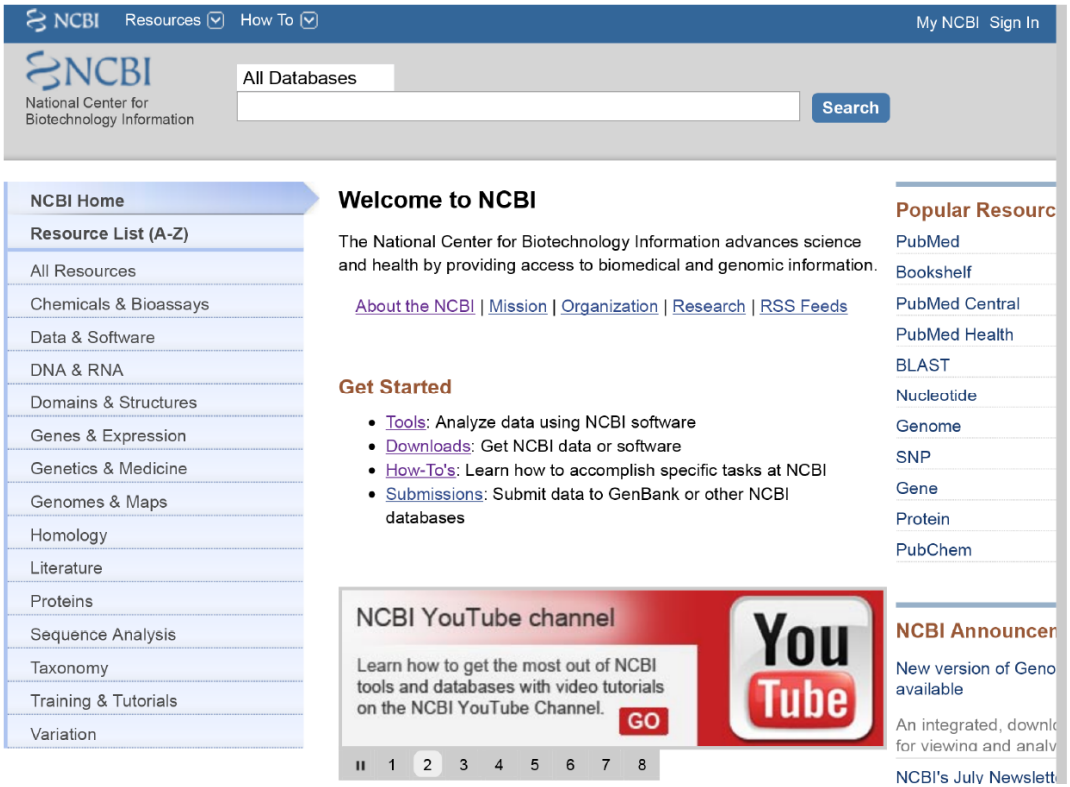
- They include sets of primary data – DNA and Protein sequences
  - Protein sequences:
    - **PIR**, http://pir.georgetown.edu/
    - **MIPS**, http://www.mips.biochem.mpg.de
    - **SWISS-PROT**, http://www.expasy.org/sprot/

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)

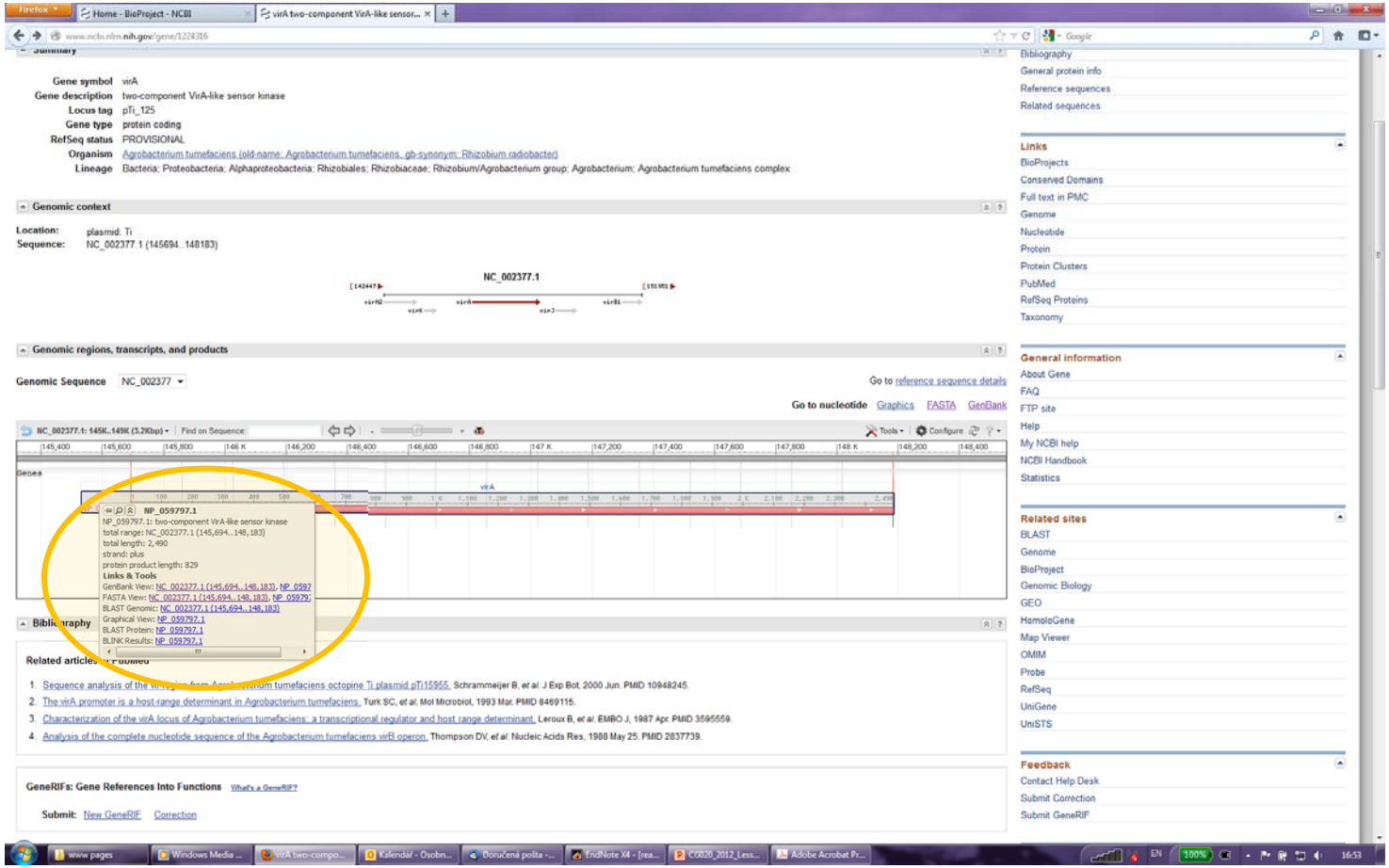
Primary Databases

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



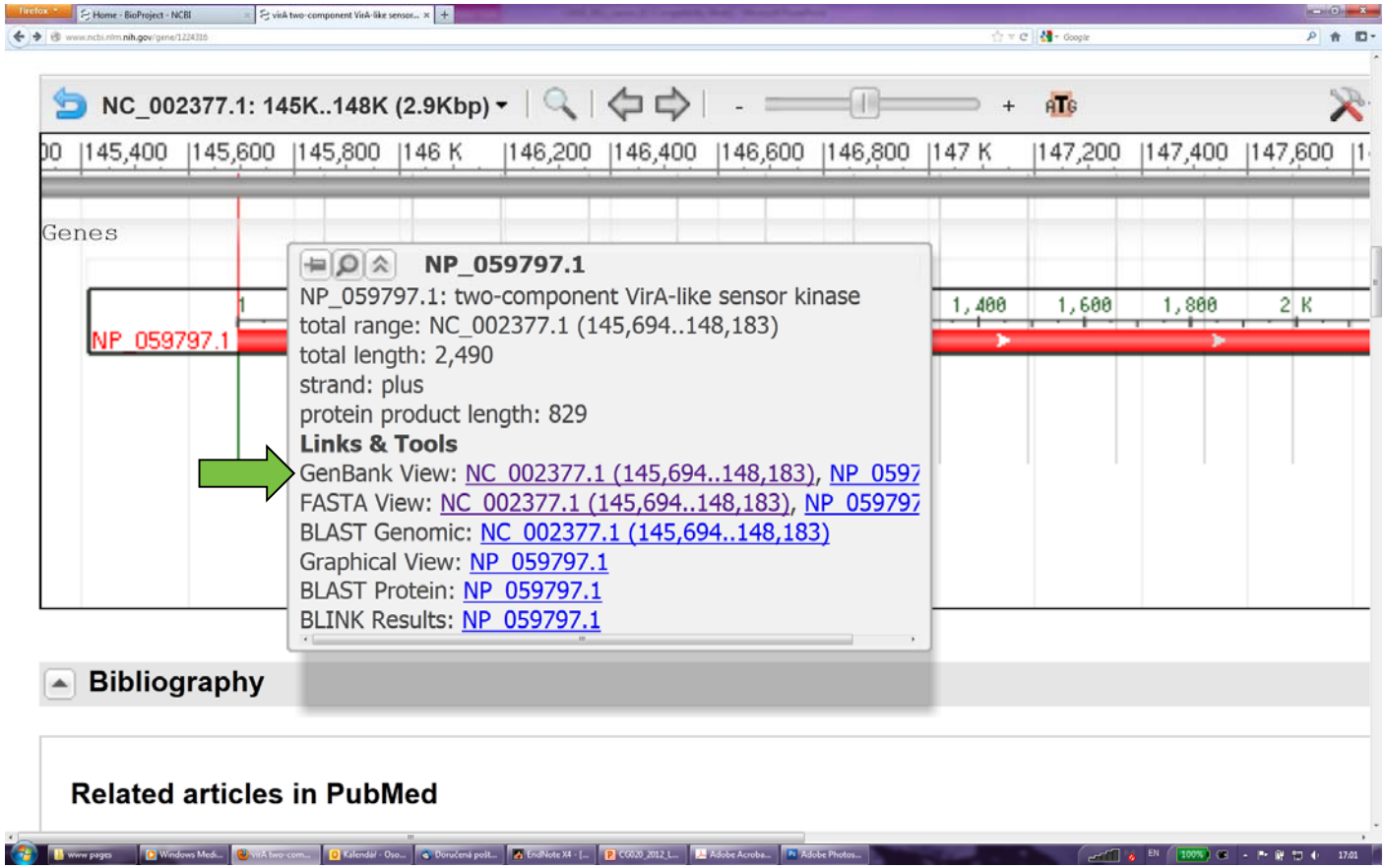
CEITEC

Primary Databases



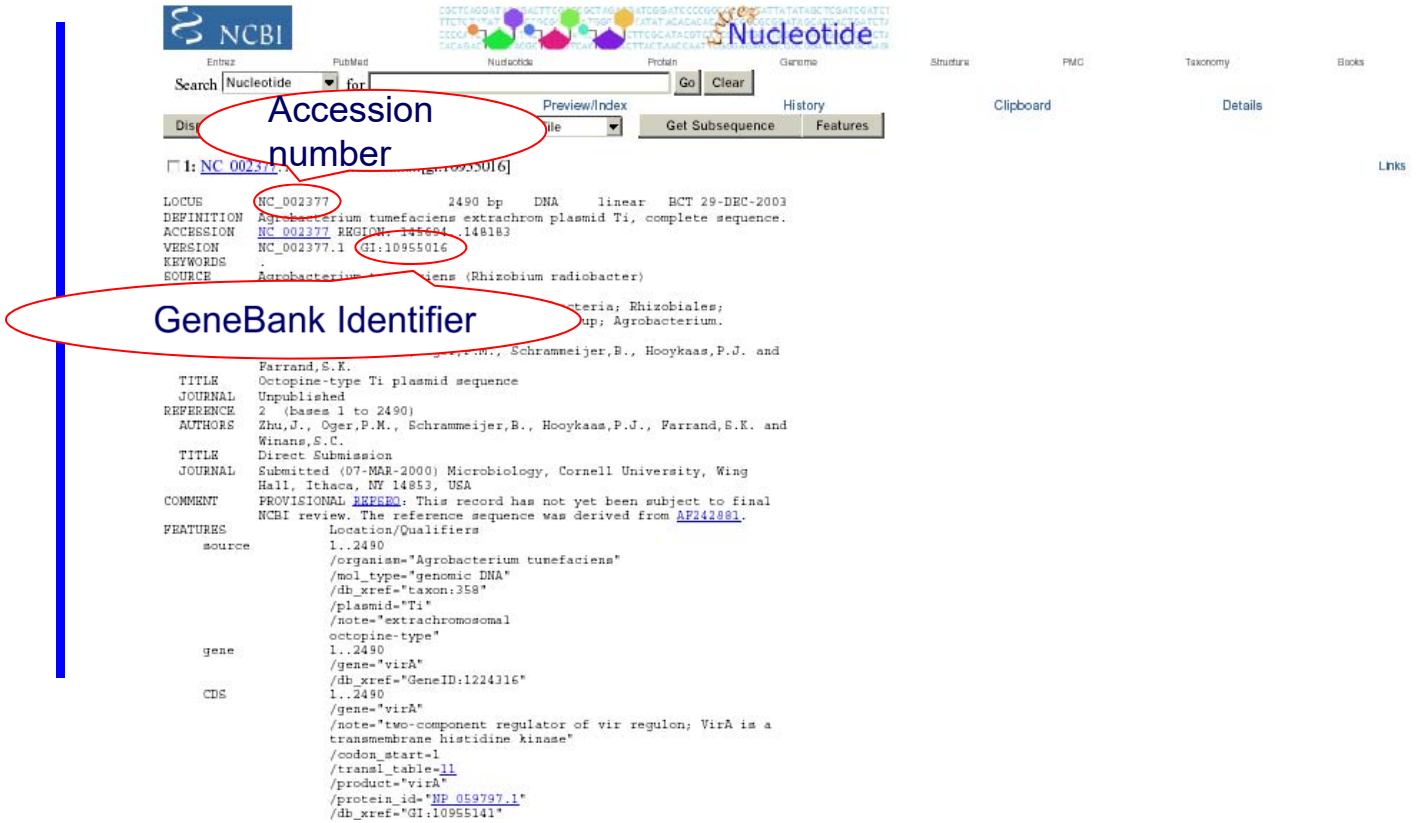
CEITEC

Primary Databases



CEITEC

Primary Databases



CEITEC

Primary Databases

```
                     /translation="MNGRYSPTRQDFKTGAKPWSILALIVAAMIFAFMAVASWQDNAT
                     TQAILSQLRSINADSASLQRDVLRAHTGTVANYRPIISRLGALRKNLEDLKQLPRQSH
                     IVSESNAAQLLRQLEVSLNSADAAVAAFGAQNVRLQDSLASPTRALSSLPGKASTDQT
                     LEKPTELASMMLQFLRQPSPAISFEISLELERLQKQRGLDEAPVRILAREGPIILSLL
                     PQVKDLVNNMIQTSDTAEIAEMLQRECLEVYSLKNVERRSARIFLGSASVGLCLYIITL
                     VYRLRKKTDWLARRLDYEELIKEIGVCFEGEAATTSSAQAALRIIQRPFDADTCALAL
                     VDHDRRWAVETPGAKHPKPVWDDSVLREIVSRTKADERATVFRIISSKKIVHLPLBIP
                     GLSILLAHKSTDKLIAVCSLGYQSYRPRPCQGEIQLLELATACLCHYIDVRRKQTECD
                     VLARRLEHAQRLEAVGTLAGGIAHEFNNILGSILGHAELAQNSVSRTSVTRRYIDYII
                     SSGDRAMLIIDQILTLSRKQERMIKPPSVSELVTEIAPLLRMALPPNIELSPRFDQMQ
                     SVIEGSPLELQQVLINICKNASQAMTANGQIDIIISQAFLPVKKILAHGVMPPGDYVL
                     LSISDNGGGIPEAVLPHIFEPFFTTRARNGGTGLGLASVHGHISAPAGYIDVSSTVGH
                     GTRFDIYLPPSSKEPVNPDSPFGRNKAPRGNGEIVALVEPDDLLREAYEDKIAALGYE
                     PVGPRTFNKIRDWISKGMEADLVMVDQASLPEDQSPNSVDLVLKTASIIIGGNDLKMT
                     LSREDVTRDLYLPKPISSRTMAHAILTKIKT"
ORIGIN
        1 atgaacggaa gatattcacc gacgcggcag gattttaaga caggcgcgaa gccttggtct
       61 atattggccc ttatcgttgc tgcaatgatt ttcgcgttca tggcggttgc gtcctggcag
      121 gacaatgcga ctacccaggc aatcctcagc caactacgat cgattaacgc cgacagcgcc
      181 tcactgcagc gcgatgtact ccgcgctcac acgggcaccg tggcgaacta ccgccccatt
      241 atctccaggc tgggagctct gcggaagaat ctggaagatt tgaagcaatt atttagacaa
      301 tctcatattg taagtgagag caatgctgct caactgctac gccagctaga agtgtctcta
      361 aattcggctg acgcggcggt cgccgccttt ggtgcgcaaa atgtacgcct gcaagattcg
      421 ctggccagtt tcactcgtgc tttgagcagt cttccaggaa aagcctcaac cgatcagact
      481 ttagaaaaac caacagaatt ggctagcatg atgctccaat ttcttcggca accaagcccg
      541 gctatttcat tcgagatcag ccttgaacta gagaggctcc aaaaacaacg cggtcttgat
      601 gaagctcccg tgcgcatact tgcacgtgaa ggtcccatta tcttatcgct tttgccacag
      661 gtgaaagatc tggtgaacat gattcagacg tctgacaccg cagaaattgc ggagatgctg
      721 cagcgcgagt gtttggaggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc
      781 tttcttgggt ccgcttcagt gggtctttgc ctctacatca tcaccttagt ctataggcta
      841 cgcaaaaaaa ccgattggtt agcgcggcgt ttagattacg aagagctaat caaagagatc
      901 ggagtatgtt ttgaaggtga ggcggccacc acgtcgtccg cgcaagctgc acttcgtatt
      961 attcagcgct tctttgatgc cgatacgtgc gcgttagctc tagtggacca tgaccgtaga
     1021 tgggctgtcg aaacattcgg tgcgaaacac ccaaaacctg tgtgggacga cagcgtgcta
     1081 cgcgaaatag tctctcgtac caaagcggac gaacgggcga cggtattccg catcatatcg
     1141 tcgaaaaaaa tcgtacattt gcctctcgaa attccaggtc tctcgatact actggctcac
     1201 aaatccacag ataaactaat tgcggtttgt tcactgggtt accaaagcta tcgccctcga
     1261 ccttgccaag gcgaaattca gcttcttgaa ctcgccaccg cctgcctctg tcactatatc
     1321 gatgttcggc gtaagcagac cgaatgcgac gttttggcca gacgattgga gcatgcgcaa
     1381 cgccttgagg cagttggtac acttgccggc ggaatagcac atgaatttaa taacattttg
     1441 ggctcaatcc tcgggcacgc agaattagca caaaactcgg tgtctcgaac atctgtcacc
     1501 cgaagatata ttgactatat catttcgtca ggcgacagag ccatgctcat tatcgatcag
     1561 atcttgacgc tgagccgaaa acaggagcgc atgatcaagc catttagtgt ctcagagctt
     1621 gtgaccgaaa tcgctccctt gctacgtatg gctcttccgc caaacatcga gcttagtttc
     1681 agatttgatc aaatgcagag cgtgatcgaa ggaagcccgc ttgaacttca acaggtacta
     1741 attaacatct gcaagaatgc ttcccaagcc atgactgcaa atggtcaaat cgacatcatc
     1801 atcagccaag ctttttacc agttaagaaa attctggcgc atggtgttat gccacctggc
     1861 gactatgttc tcctatctat tagcgacaat ggtggaggca ttccgaggc tgtgttaccc
     1921 cacatttttg aacccttctt tacgacacga gctcgcaacg gtggaacggg tctcggcctt
     1981 gcttctgtgc atggtcatat cagcgcgttt gcgggttaca tcgacgttag ttcaactgtt
     2041 gggcatggga cgcgctttga catttatctc cctccgtctt ctaaggaacc cgtaaatcca
     2101 gacagttttt tcggccgcaa taaggcaccg cgtggaaacg gggagattgt ggcacttgtt
     2161 gagcccgatg acctcctgcg ggaggcgtat gaagacaaga tcgccgctct aggatatgag
     2221 ccggtcggtt ttcgtacctt taatgaaatt cgcgattgga tttcaaaagg caatgaagcc
     2281 gatctggtca tggtcgacca agcgtctctt cctgaagatc aaagtcctaa ttccgtggat
     2341 ttagtgctca agaccgcctc catcatcatt ggcggaaatg atctcaaaat gaccctttca
```

What is an Accession Number?

An accession number is label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	**DNA**
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	**RNA**
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	**Protein**
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

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

NCBI's important RefSeq project: best representative sequences

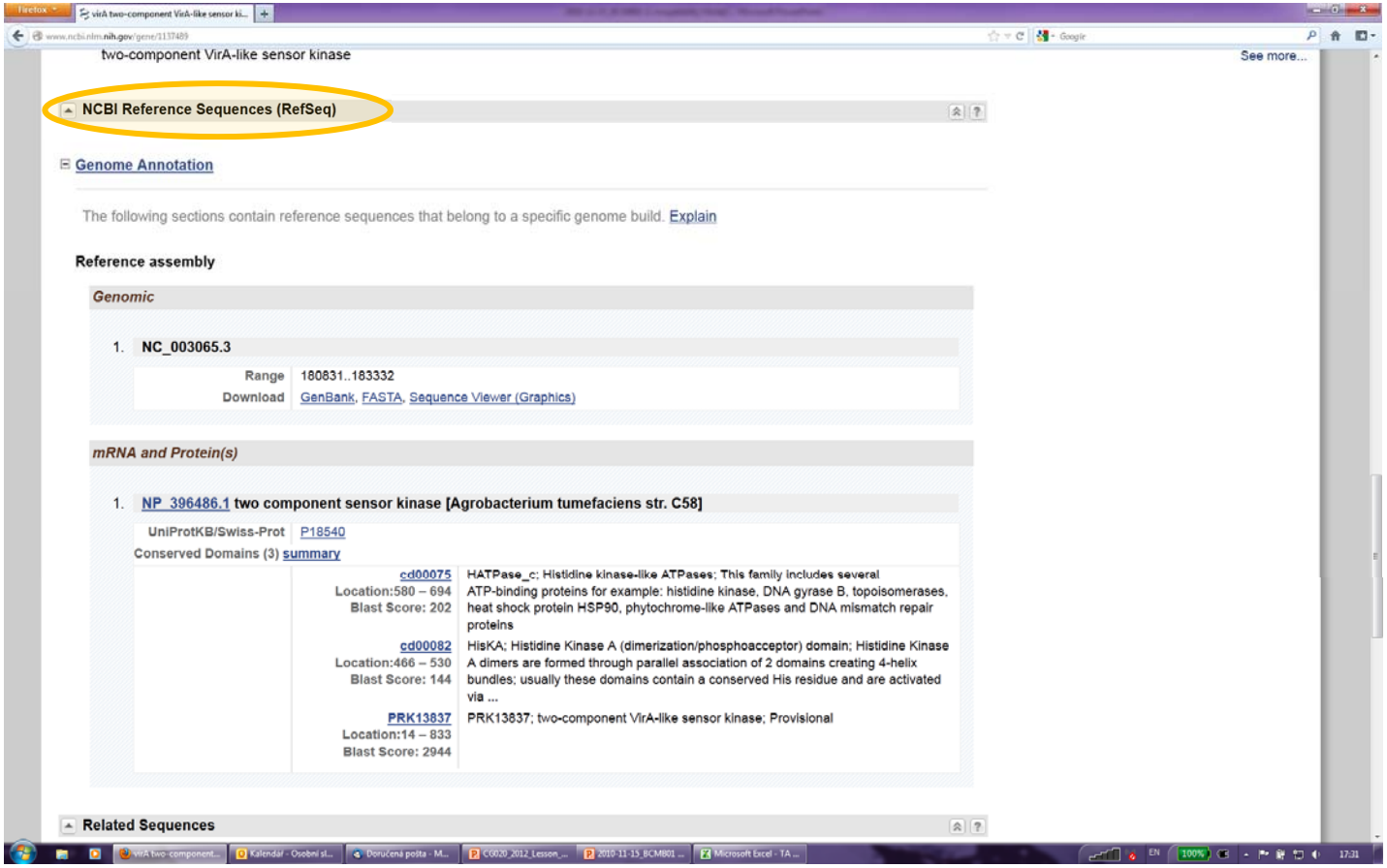
RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon "reference" version of a sequence.

RefSeq identifiers include the following formats:

Complete genome	NC_######
Complete chromosome	NC_######
Genomic contig	NT_######
mRNA (DNA format)	NM_###### e.g. NM_006744
Protein	NP_###### e.g. NP_006735

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

RefSeq



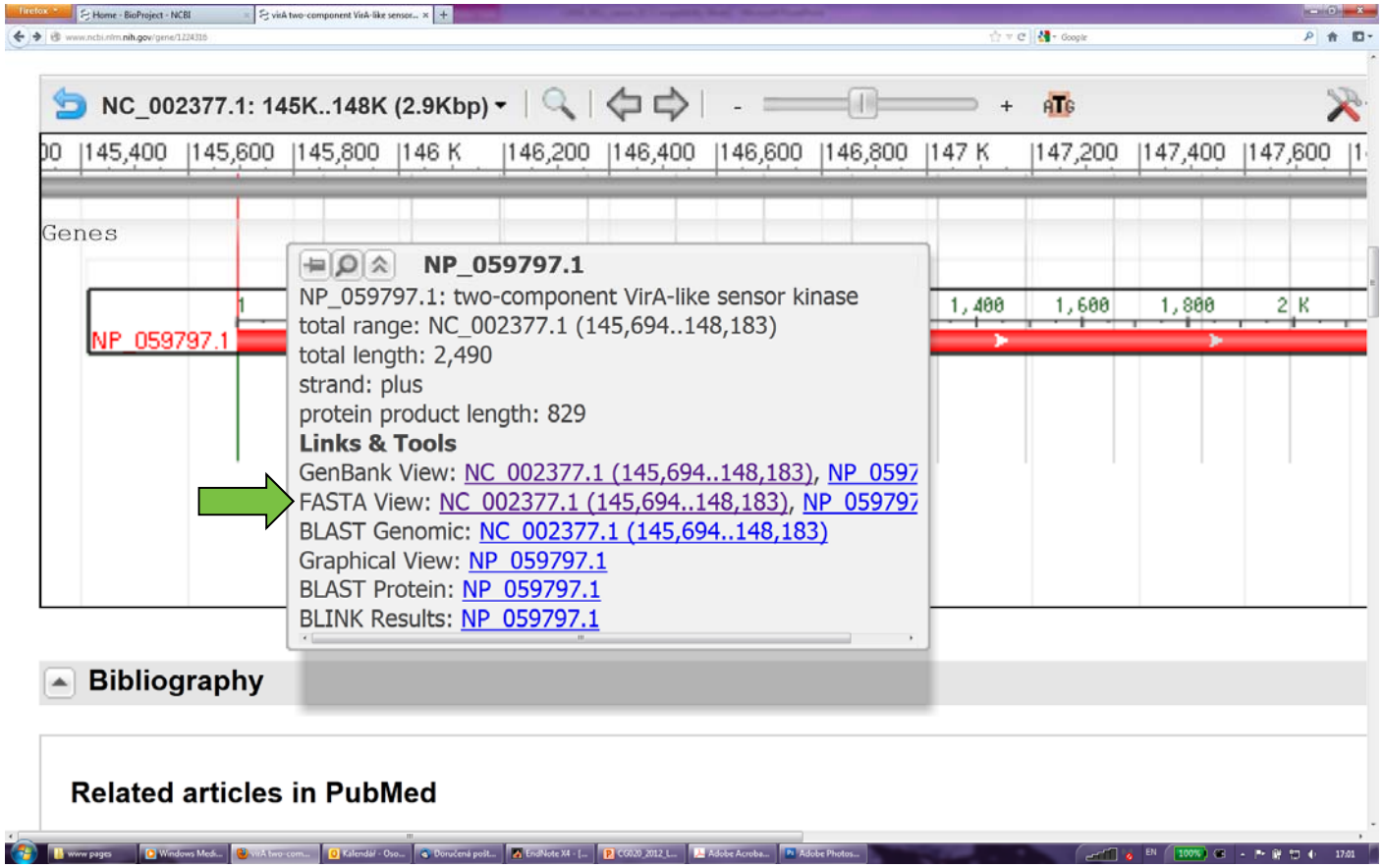
CEITEC

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

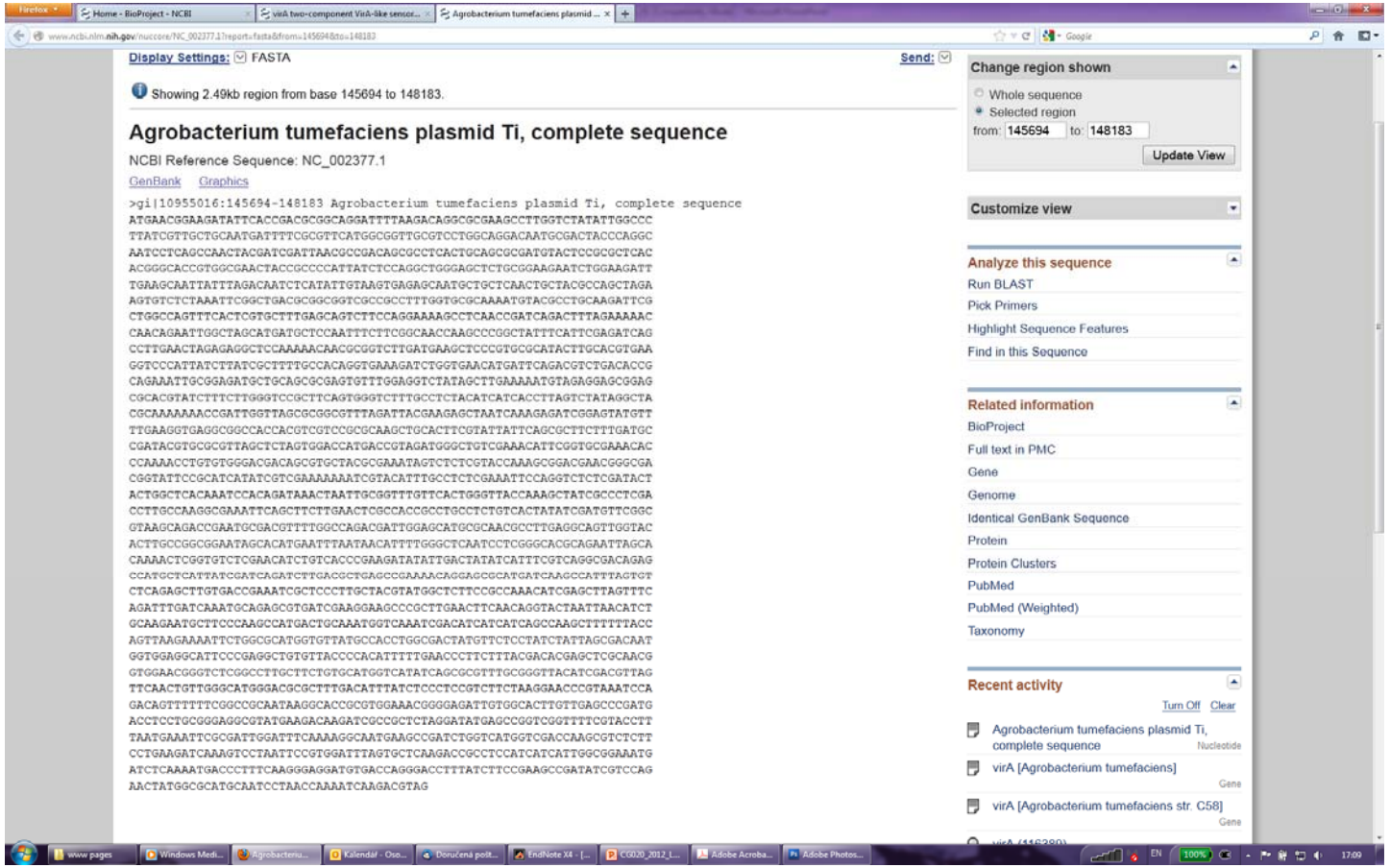
J. Pevsner,
http://www.bioinfbook.org/index.php

Primary Databases



CEITEC

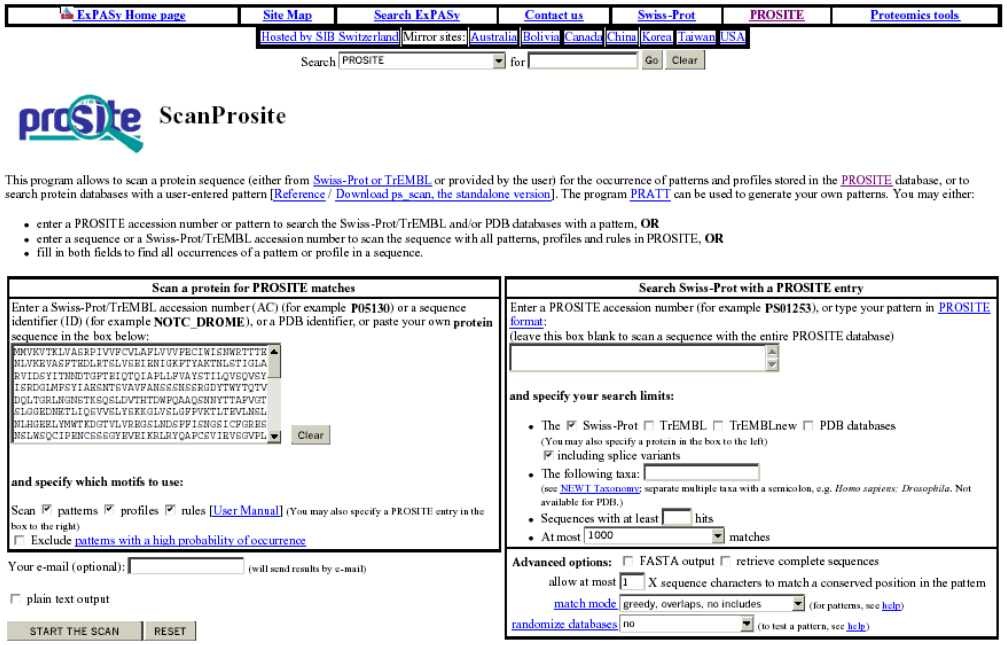
Primary Databases



CEITEC

Secondary Databases

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



Secondary Databases

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

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

571 - 585 nkeemstYeteisns

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

744 - 747 RRvT
814 - 817 KRrS

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

148 - 150 SmR
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 TeK
794 - 796 SsR
854 - 856 ScK
864 - 866 StR
868 - 870 SeR
921 - 923 SpK
957 - 959 SvR
960 - 962 TgR
974 - 976 TmK
997 - 999 SrK
1002 - 1004 TgK
1018 - 1020 SgK
1031 - 1033 TgR
1119 - 1121 SkR

Secondary Databases

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

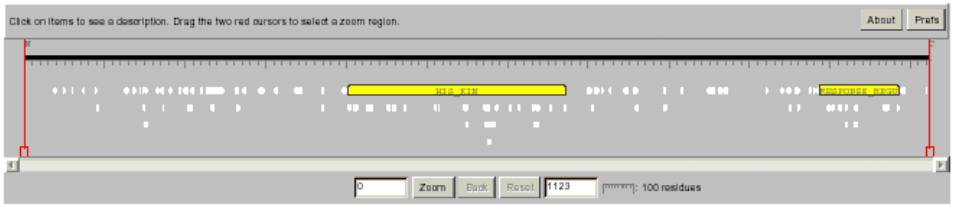
>PDOC50109 PS50109 **HIS_KIN** Histidine kinase domain [profile].

```
402 - 671  NASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVNVCAKDLVALLNSVLDMSKIESG
           KMQLVREDFNLSKLLKDVIDFYHPVAMKKGVDVVLDPHDgsvfKPSNVRGDSGRLKQILN
           NLVSNAVKFTVD--GHIAVRAWAQrpgsnsavvlsaypkgvskfvksmfcknkeesatye
           teisnsirnnanTMEFVPEVDDTGKGIPMEMRKSVPENYVQVREtAQGHQGTGLGLGIVQ
           SLVRLMGGEIRITDKAMGekGTCPQFNVLLTT
```

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

```
987 - 1085  RVLVVDDNPISRKVATGKLKKMGVSeVEQCDSGKEALRLVTEGLtqreeqgsvdklpFDY
            IFMDCQMPEMDGYEATREIRkvekSYGVRTPIIAVSGHD----------------------
            ------------
```

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

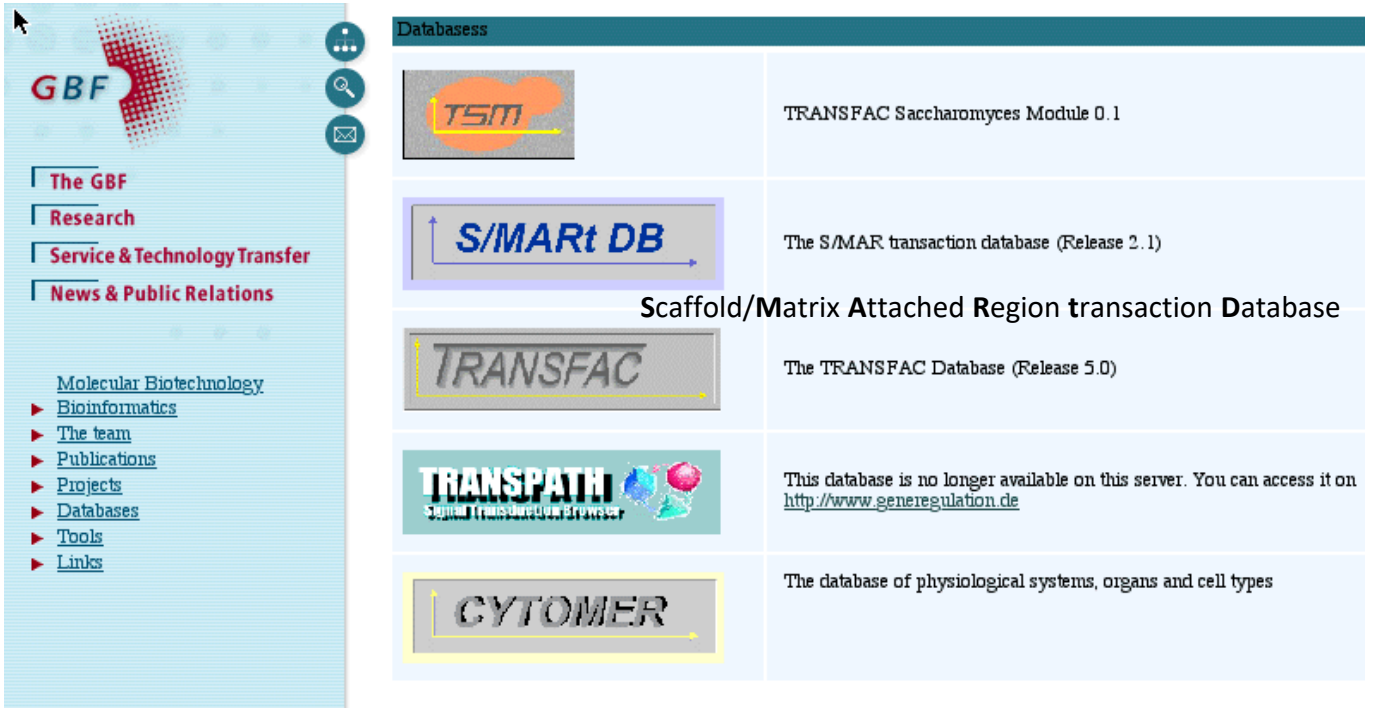
Secondary Databases

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

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

Secondary Databases

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



Scaffold/**M**atrix **A**ttached **R**egion **t**ransaction **D**atabase

CEITEC

Structural Databases

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

Structural Databases

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

Structure Explorer - 1P5Y

RCSB PDB PROTEIN DATA BANK

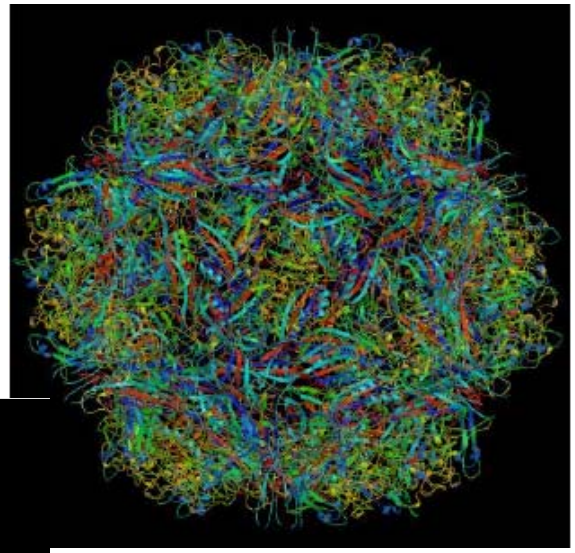
Structure Explorer - 1P5Y

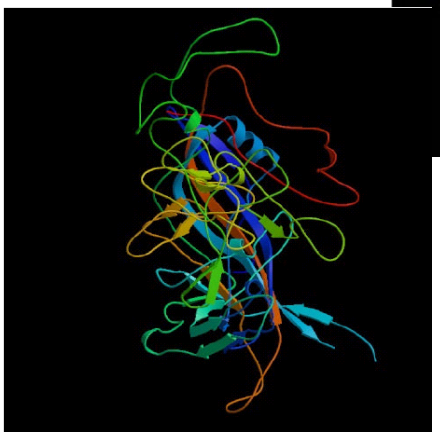
Title The Structures Of Host Range Controlling Regions Of The Capsids Of Canine and Feline Parvoviruses and Mutants
Classification Virus/Viral Protein
Compound Mol_Id: 1; Molecule: Coat Protein Vp2; Chain: A; Fragment: Sequence Database Residues 190-737; Engineered: Yes; Mutation: Yes
Exp. Method X-ray Diffraction

View Structure

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details

Explore
SearchLite SearchFields

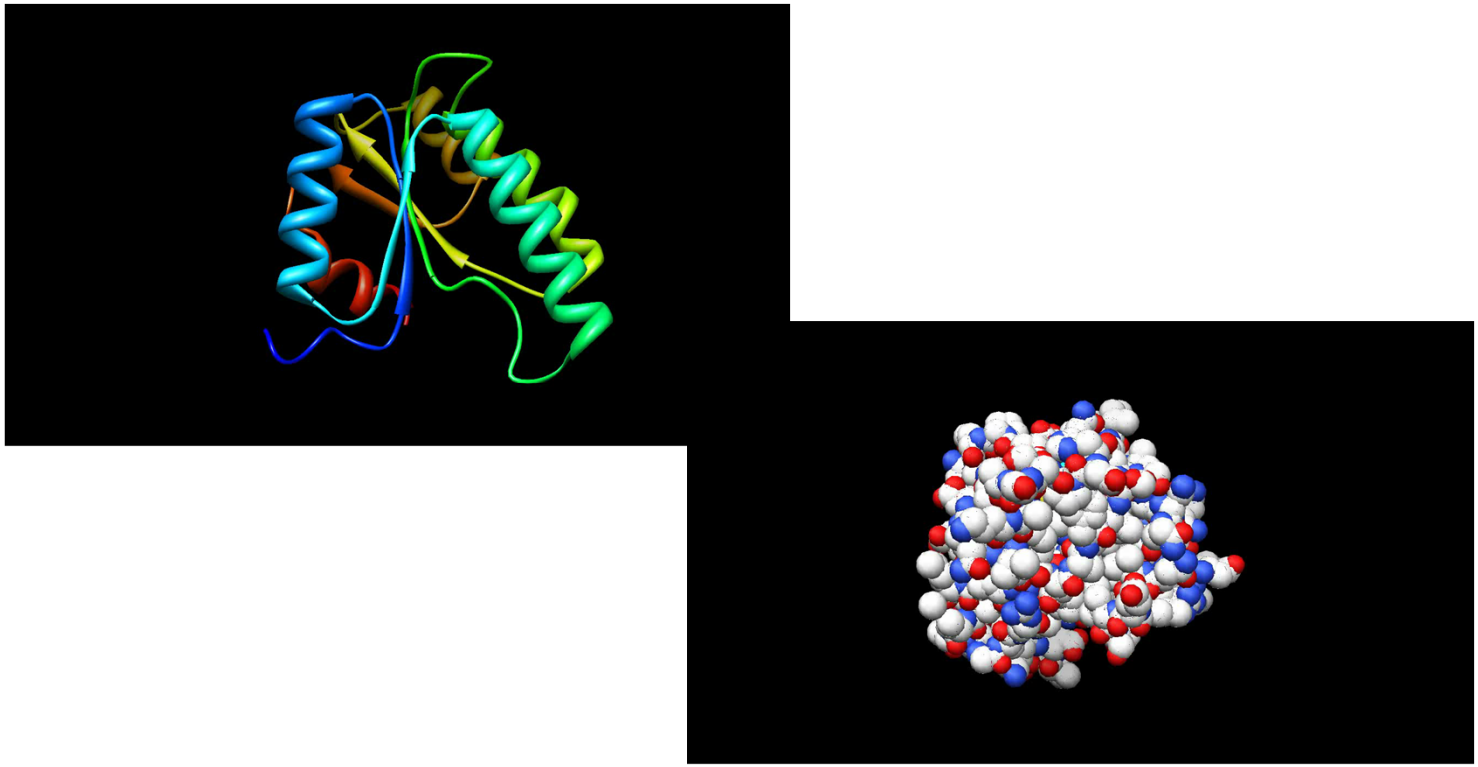




http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=;pid=1735610643493 44&bio=1&opt=show&size=500 12/29/2003

Structural Databases

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



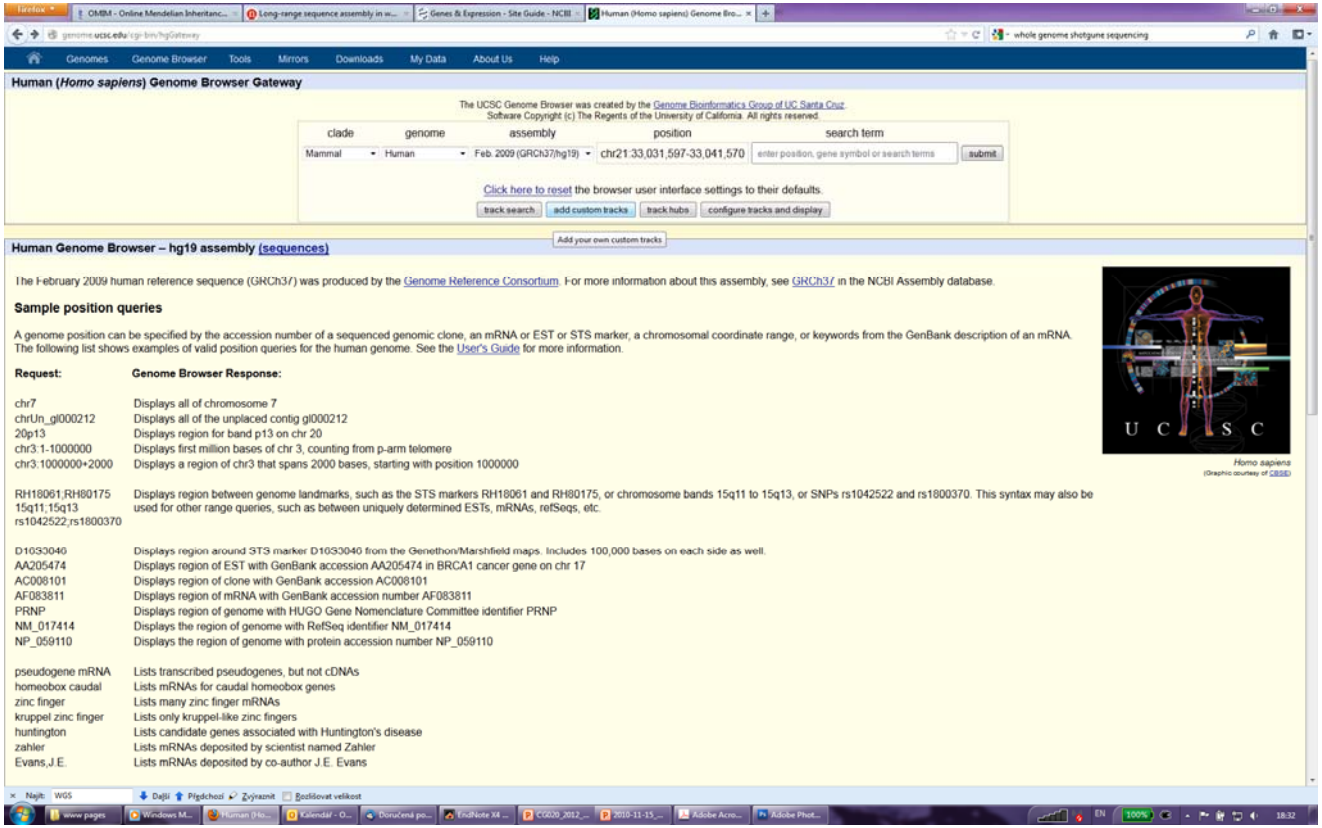
Pekárová et al., *Plant Journal* (2011)

Outline

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

Genome Resources

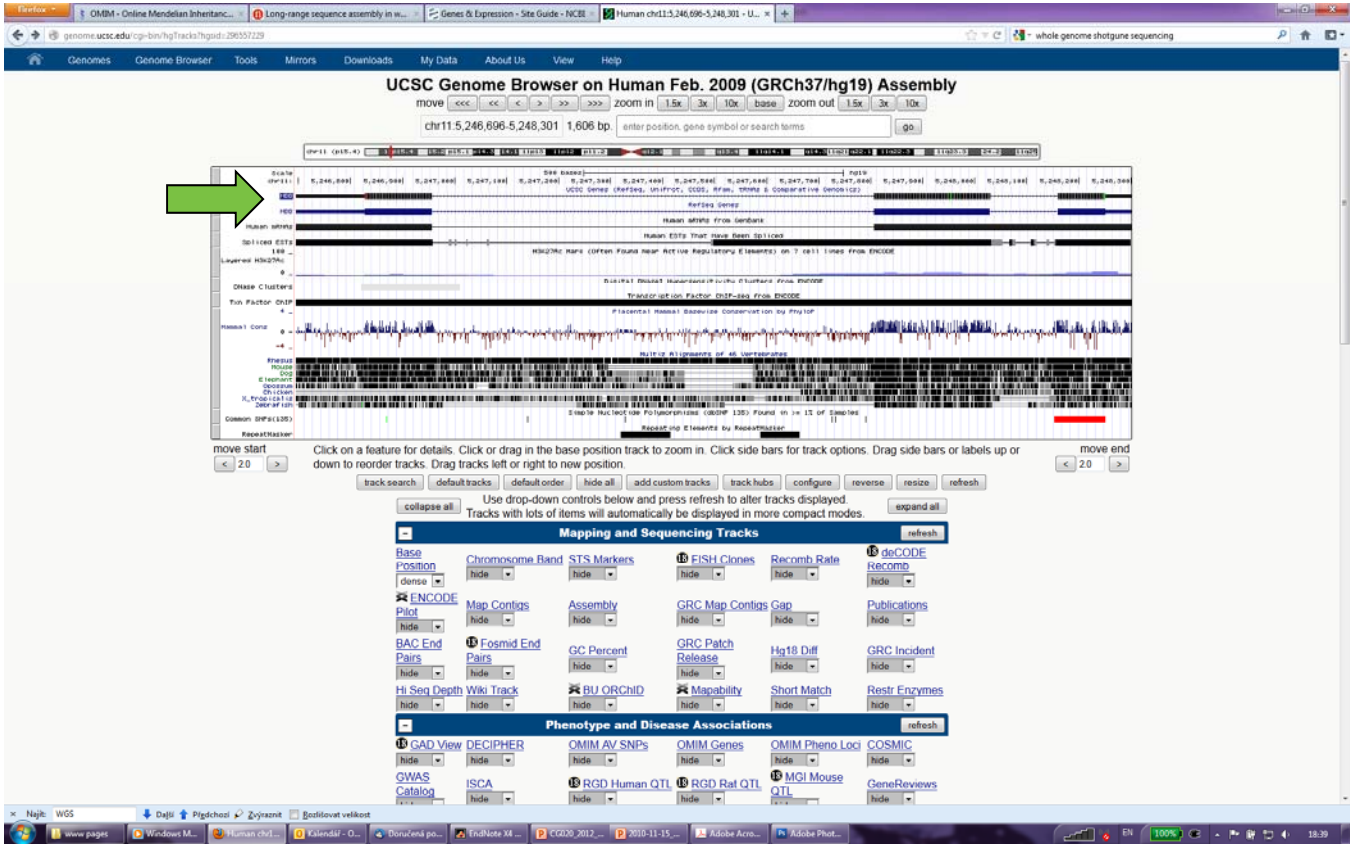
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genome Resources

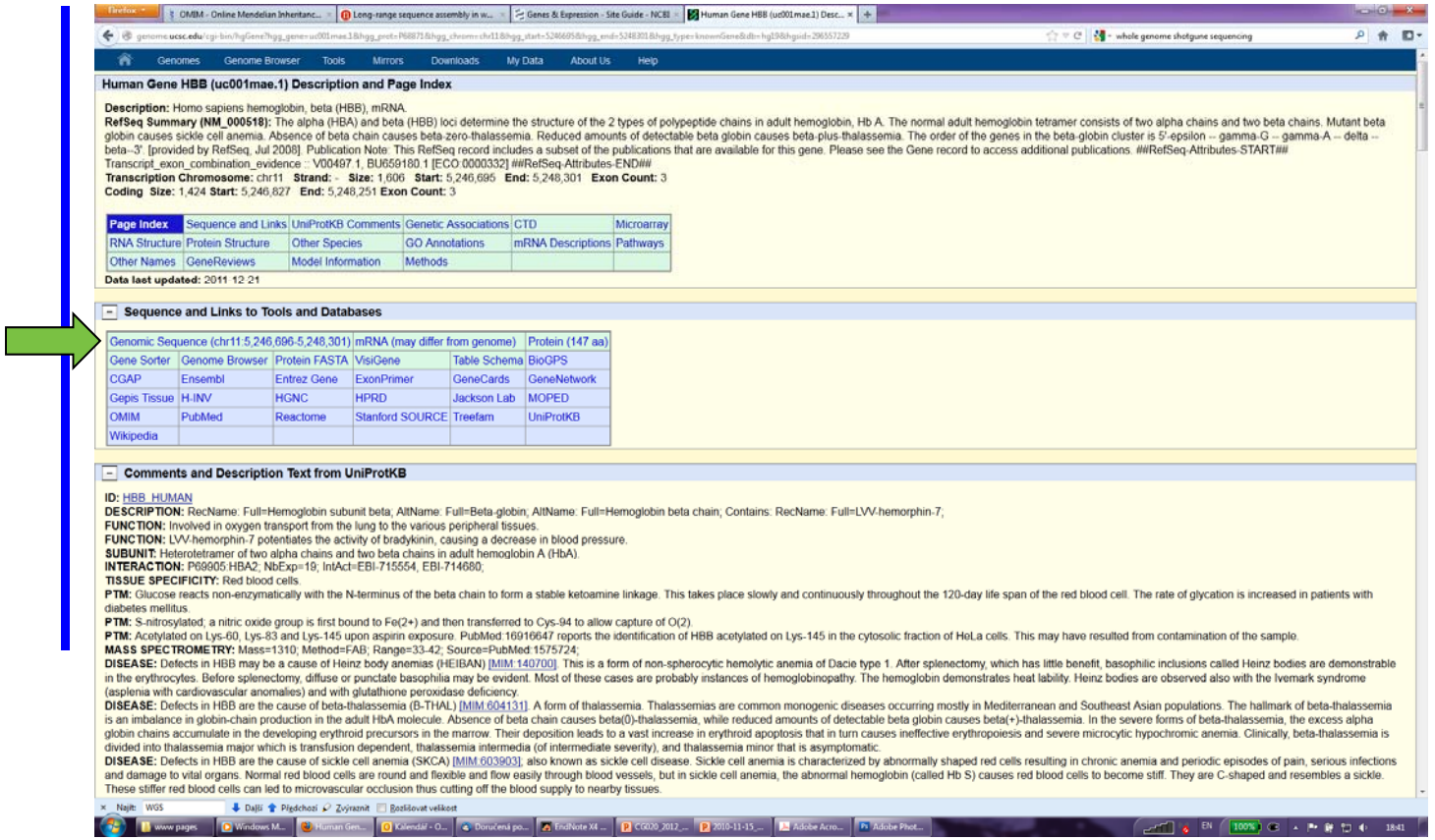
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

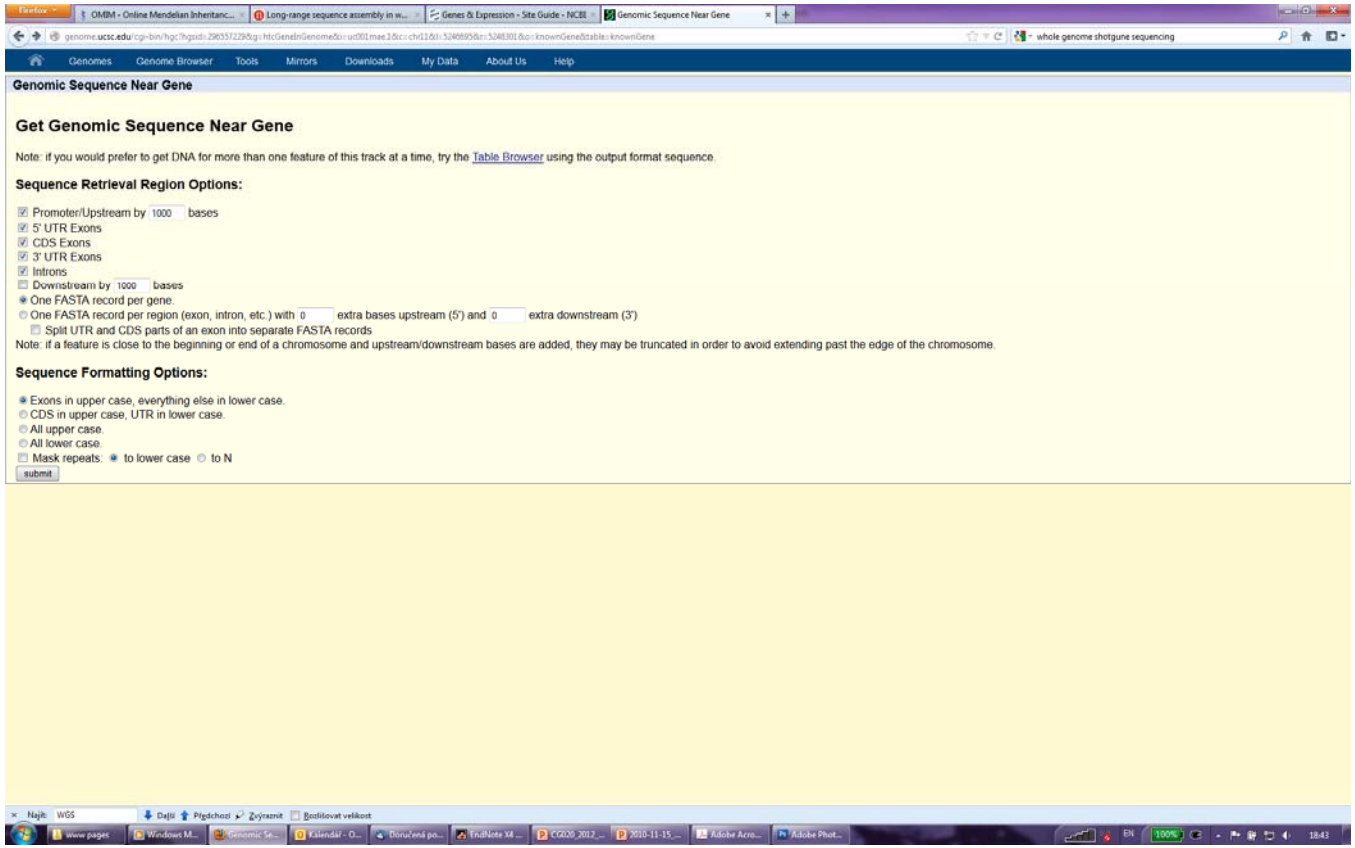
Genome Resources

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



Genome Resources

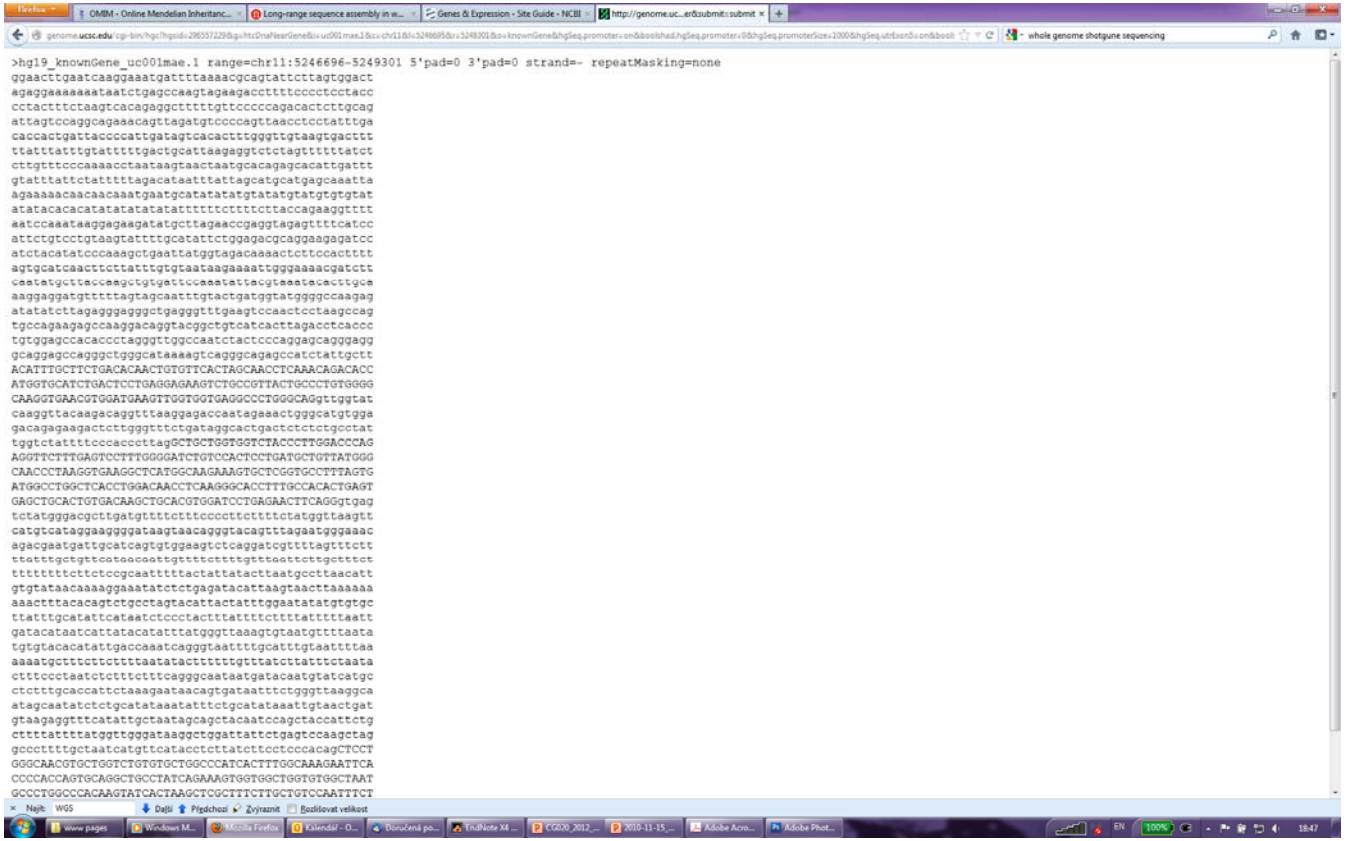
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genome Resources

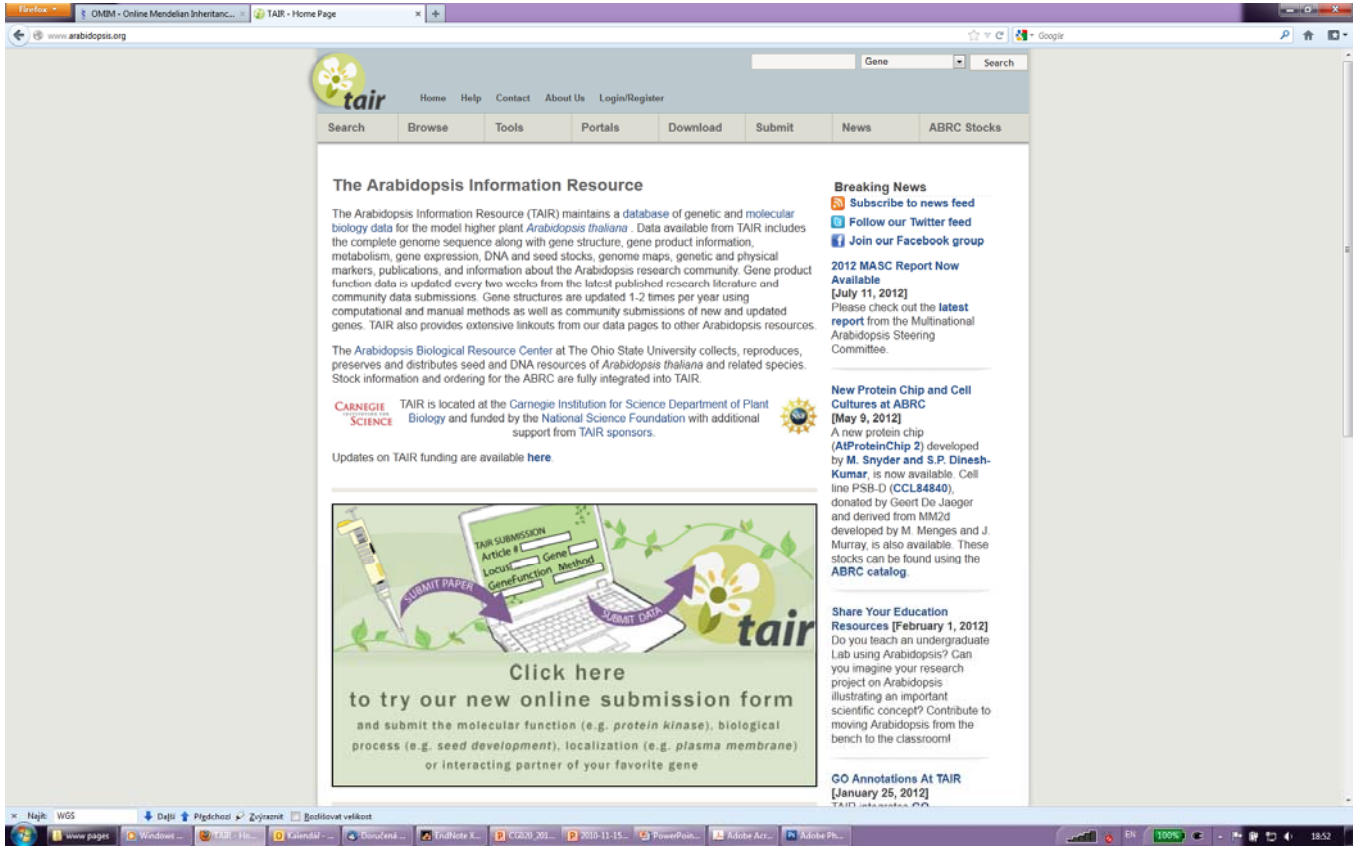
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

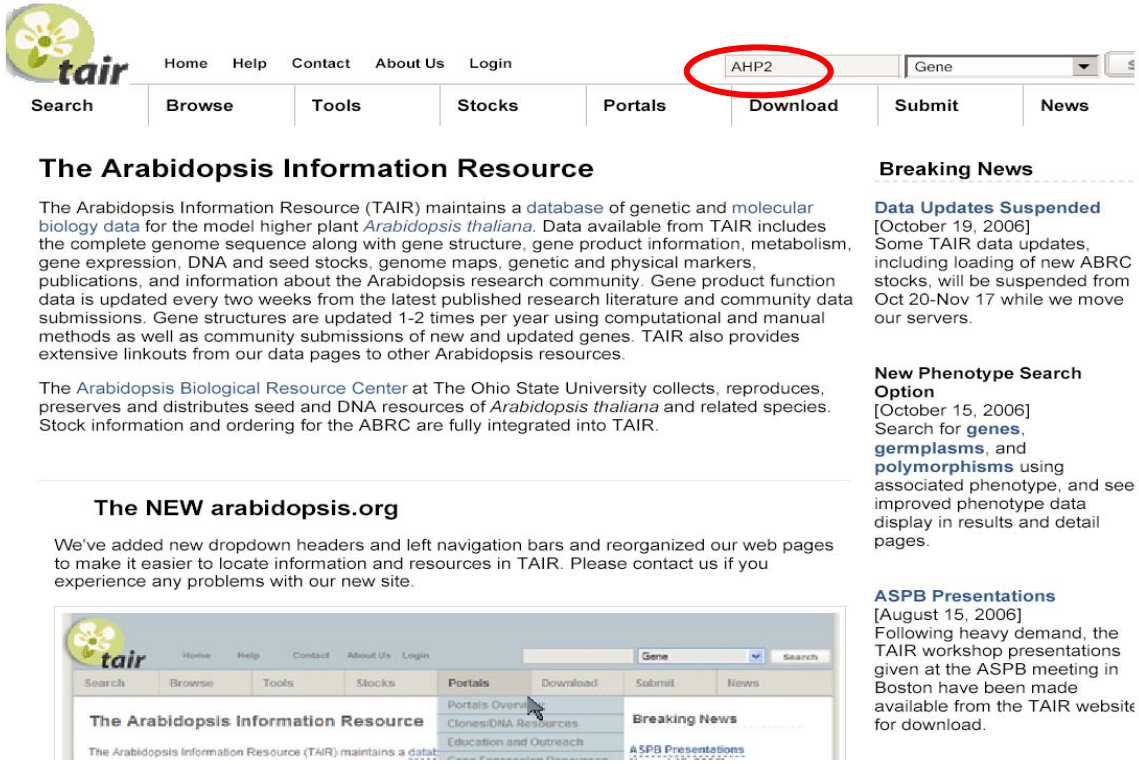
Genome Resources

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



Genome Resources

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



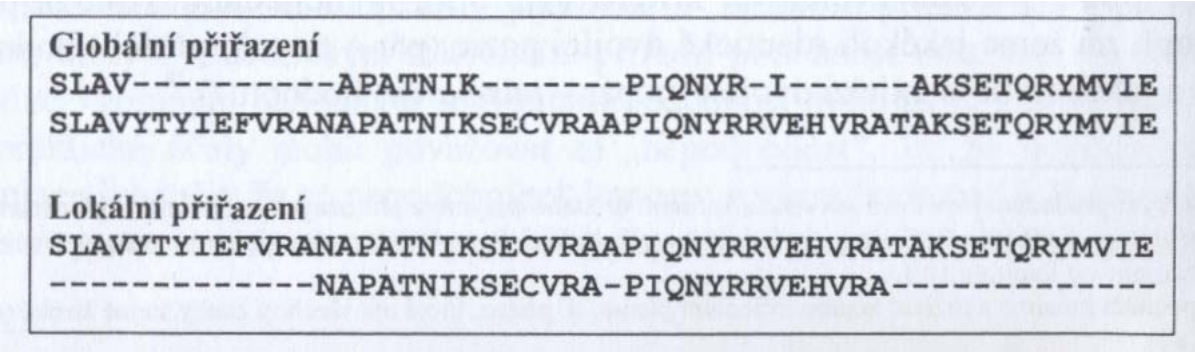
CEITEC

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

Analytical Tools

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

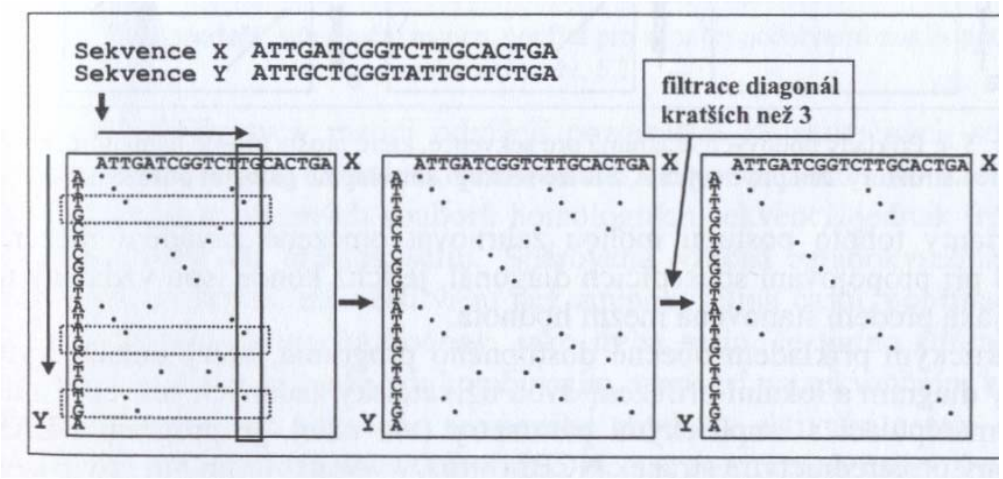


Cvrčková, Úvod do praktické bioinformatiky

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

Analytical Tools

- Choosing the right type of alignment using dotplot

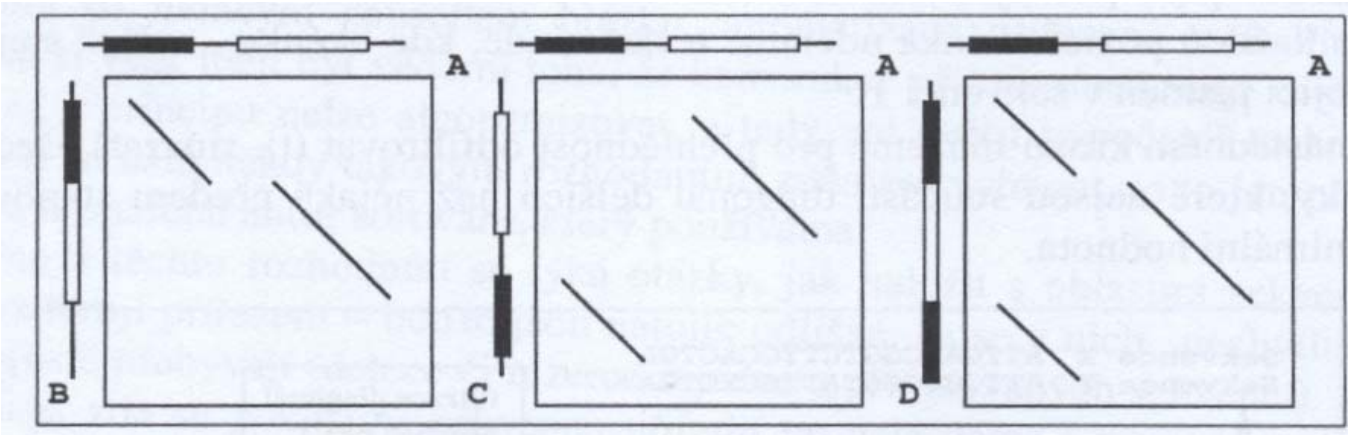


Cvrčková, Úvod do praktické bioinformatiky

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

Analytical Tools

- Examples of sequence alignment using dotplot

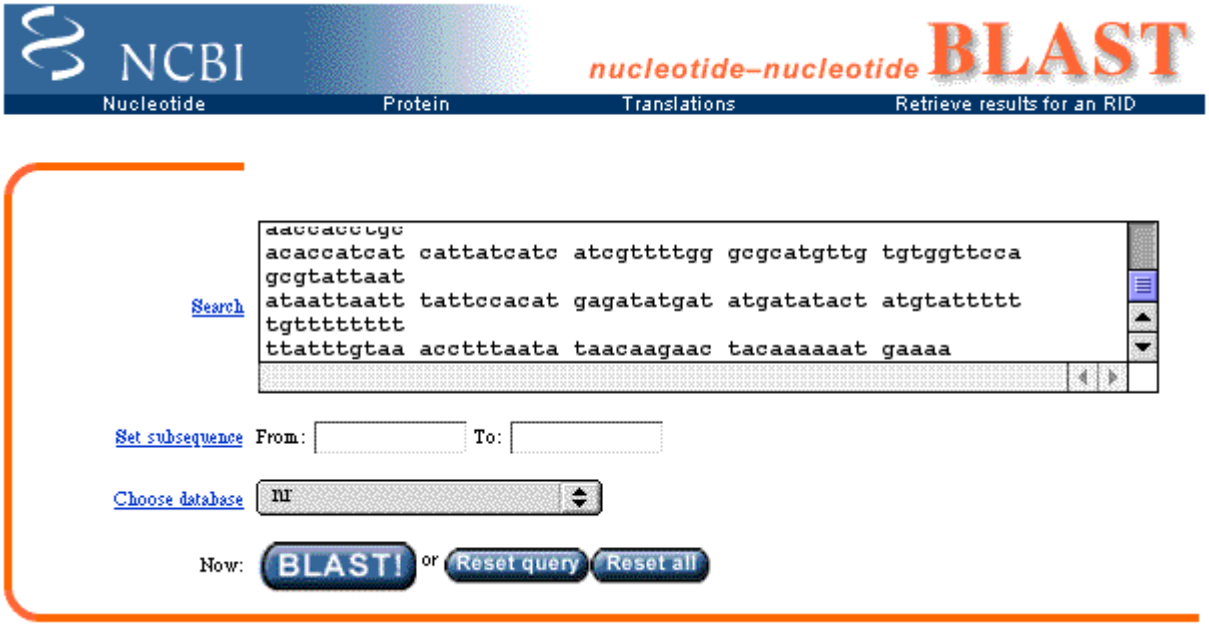


Cvrčková, Úvod do praktické bioinformatiky

- **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)

Analytical Tools

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

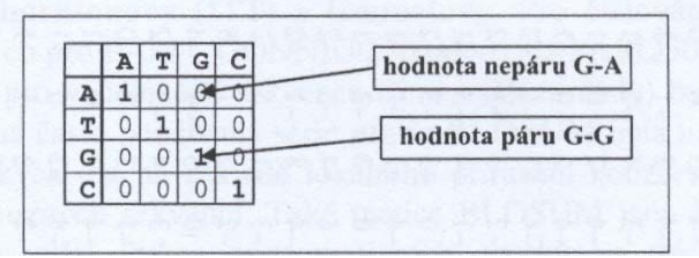


CEITEC

BLAST

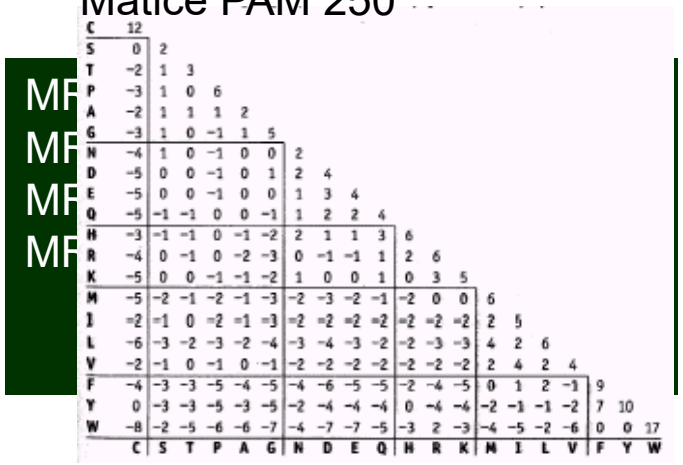
Basic **L**ocal **A**lignment **S**earch **T**ool

- 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



CEITEC

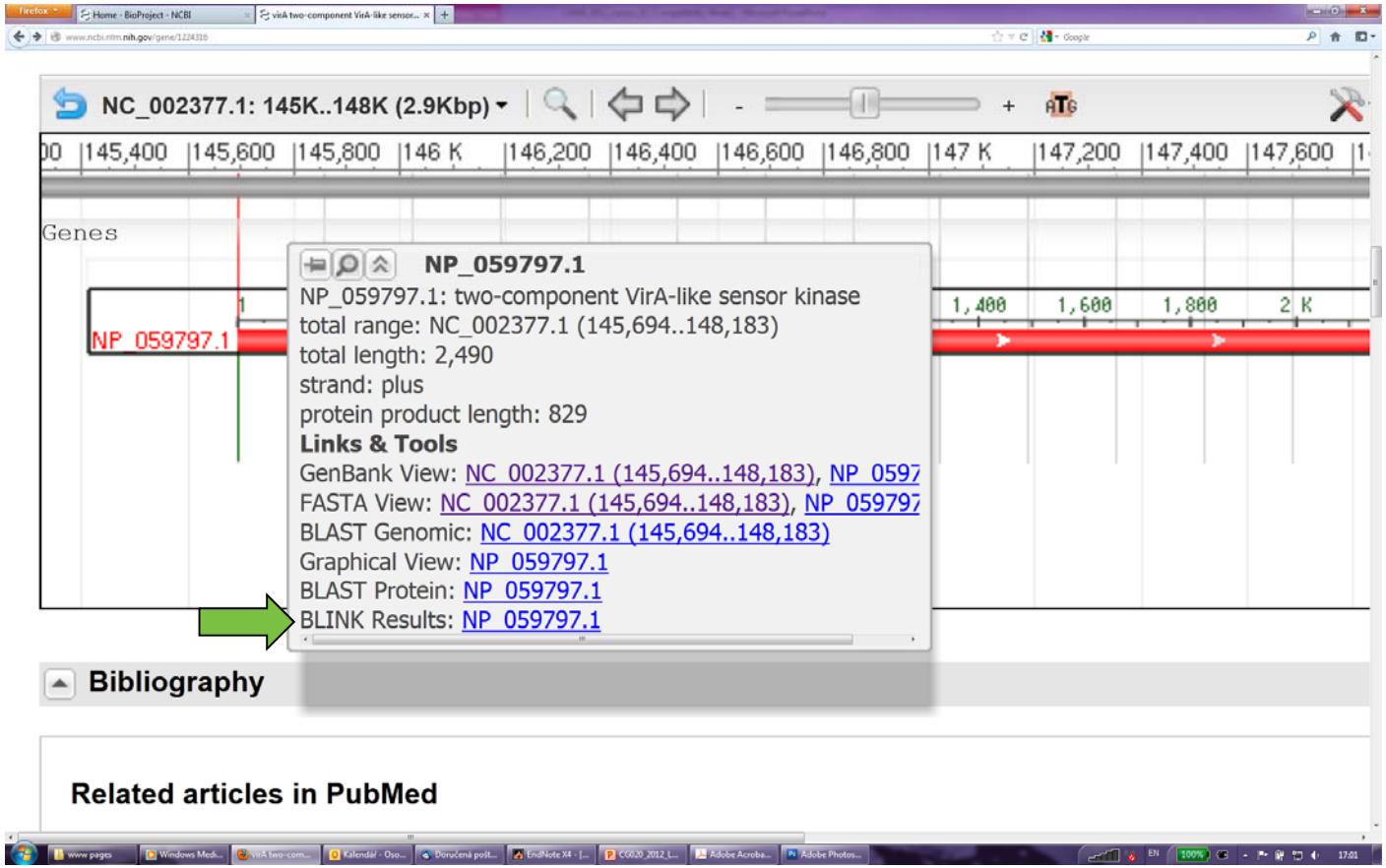
BLAST

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

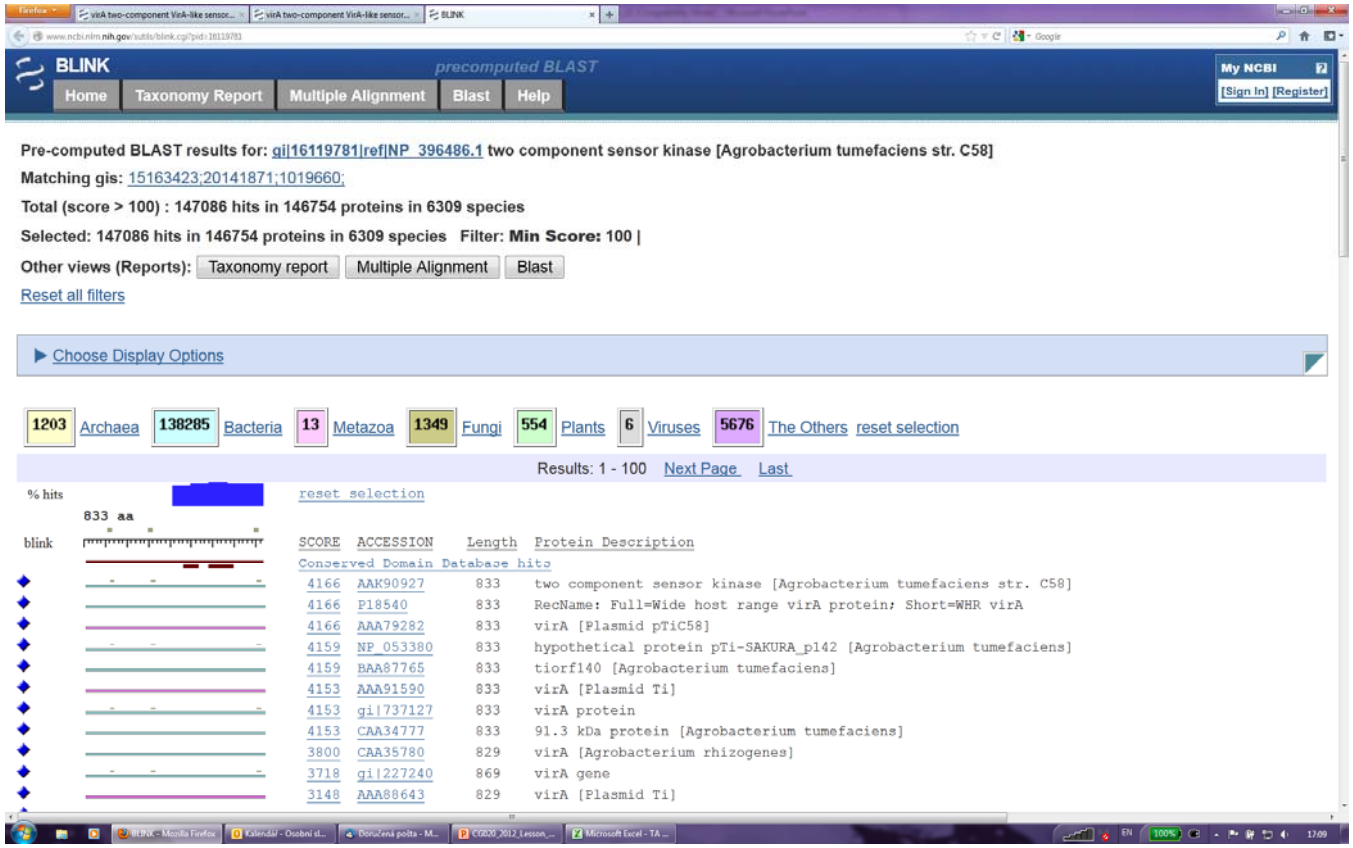
Primary Databases



CEITEC

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.

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.

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

BLAST

Specialized Versions

- Example of search by PHI-BLAST

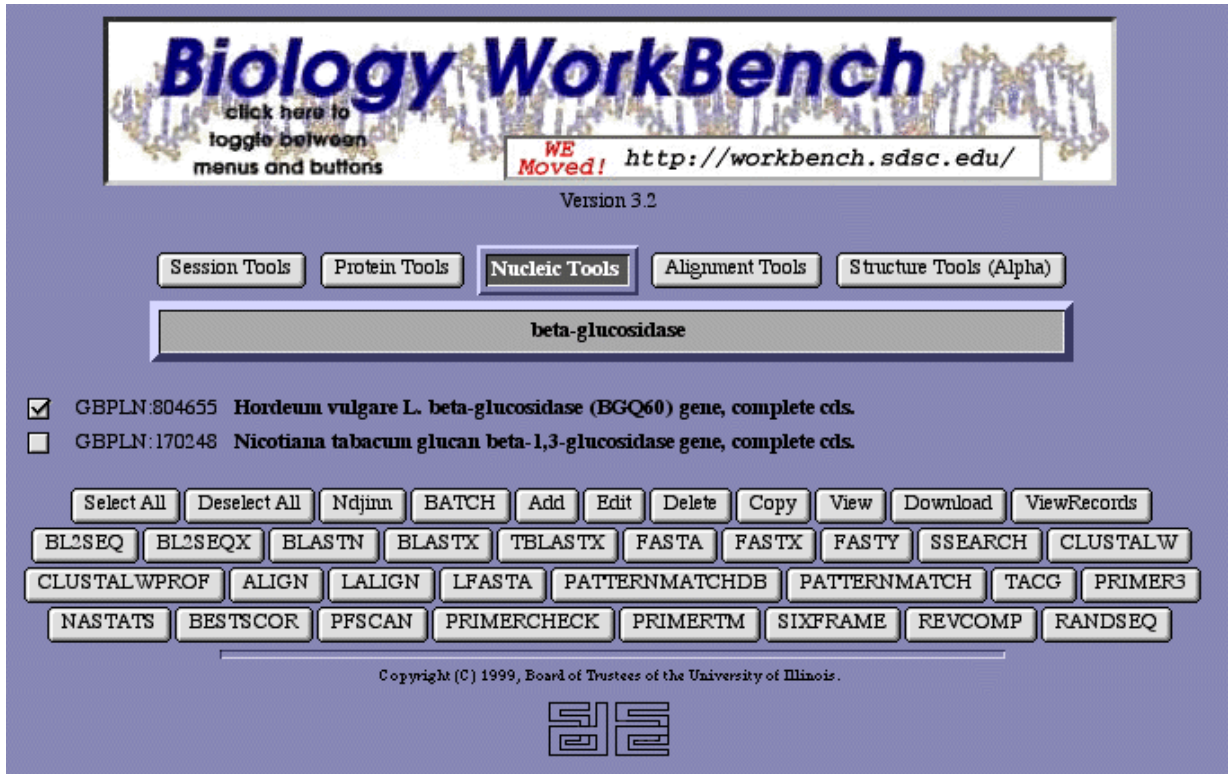
```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAATPRQSLGHPPPEPGPDR
VADAKGDSESEEDEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA
LMYNTPRAATIVATSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIGEK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

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

Outline

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

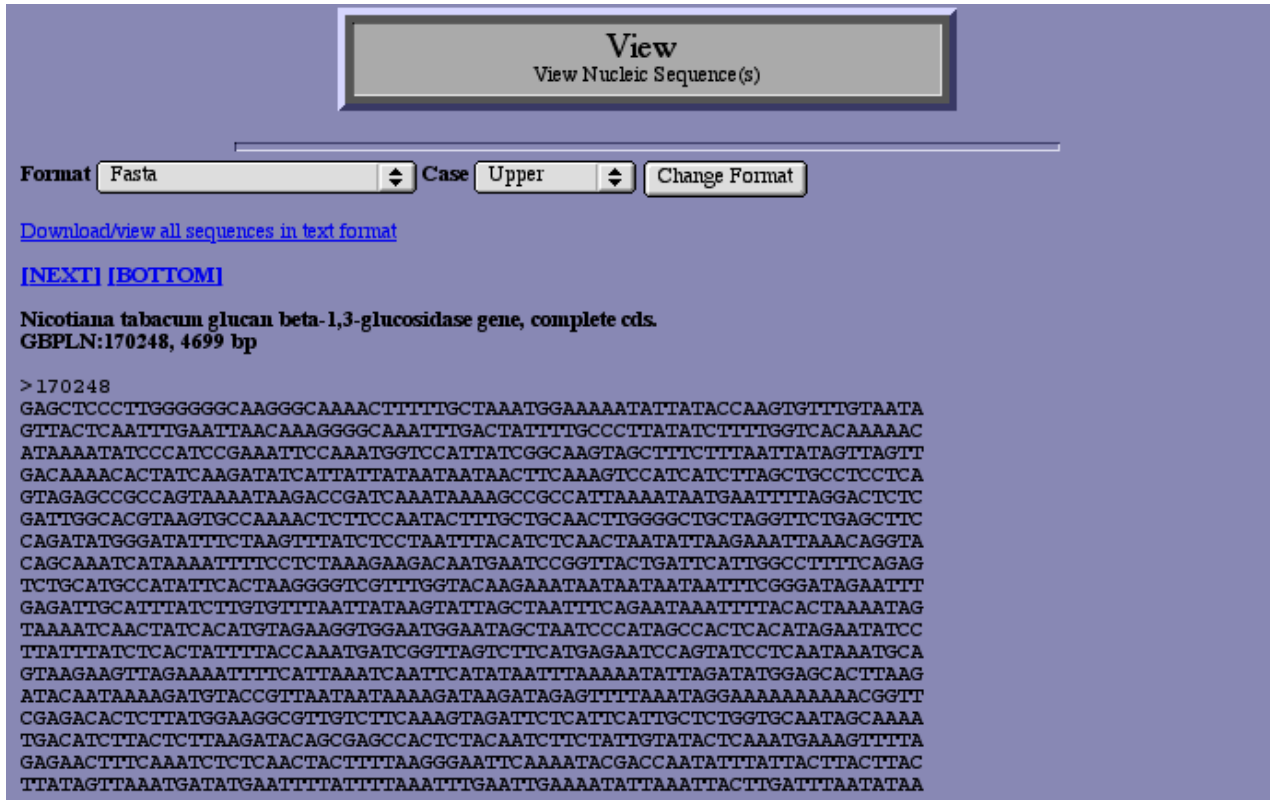
Analytical Tools

- http://workbench.sdsc.edu/



Analytical Tools

- http://workbench.sdsc.edu/



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
GAGCTCCCTTGGGGGGCAAGGGCAAAACTTTTTGCTAAATGGAAAAATATTATACCAAGTGTTTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAATTTGACTATTTTGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTTCTTTAATTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAATAATAACTTCAAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAATAAGACCGATCAAATAAAAGCCGCCATTAAAATAATGAATTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACTCTTCCAATACTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTAAACAGGTA
CAGCAAATCATAAAATTTTCCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCATGCCATATTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAATAATTTCGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTAATTATAAGTATTAGCTAATTTCAGAATAAATTTTACACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACCAAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAATTTTCATTAAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTGTCTTCAAAGTAGATTCTCATTCATTGCTCTGGTGCAATAGCAAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA
GAGAACTTTCAAATCTCTCAACTACTTTTAAGGGAATTCAAAATACGACCAATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTATTTTAAATTTGAATTGAAAATATTAAATTACTTGATTTAATATAA
ACAATAGATATCGCTAAGTATTTACCACAAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT
GATTAAGCAGCTATTCATCTGGTTGTGCAGGATGAAAGAAAGTAACTAGCTATAATTTCTTTTGTAAAGT

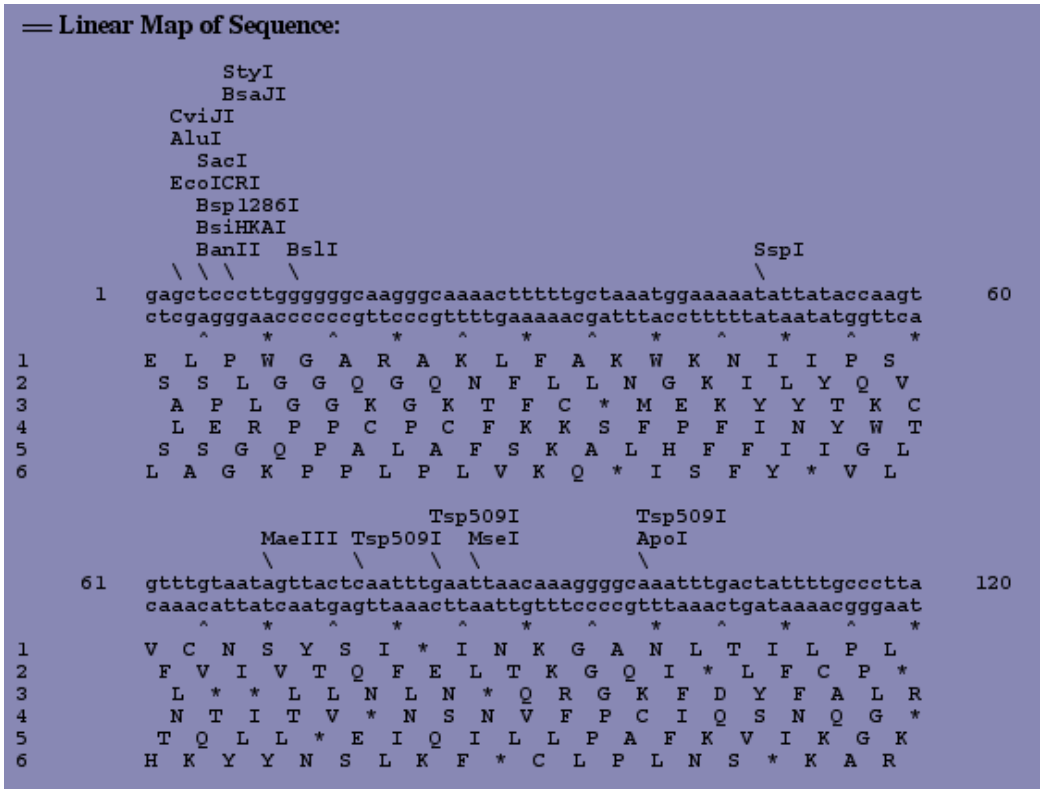
Analytical Tools

- http://workbench.sdsc.edu/

CEITEC

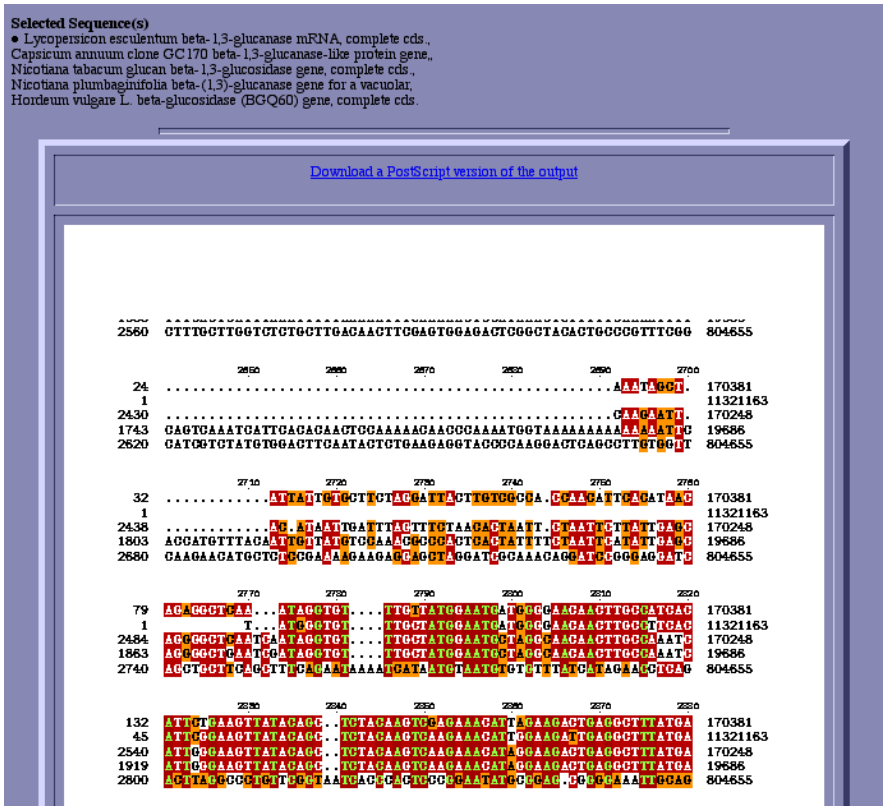
Analytical Tools

- http://workbench.sdsc.edu/



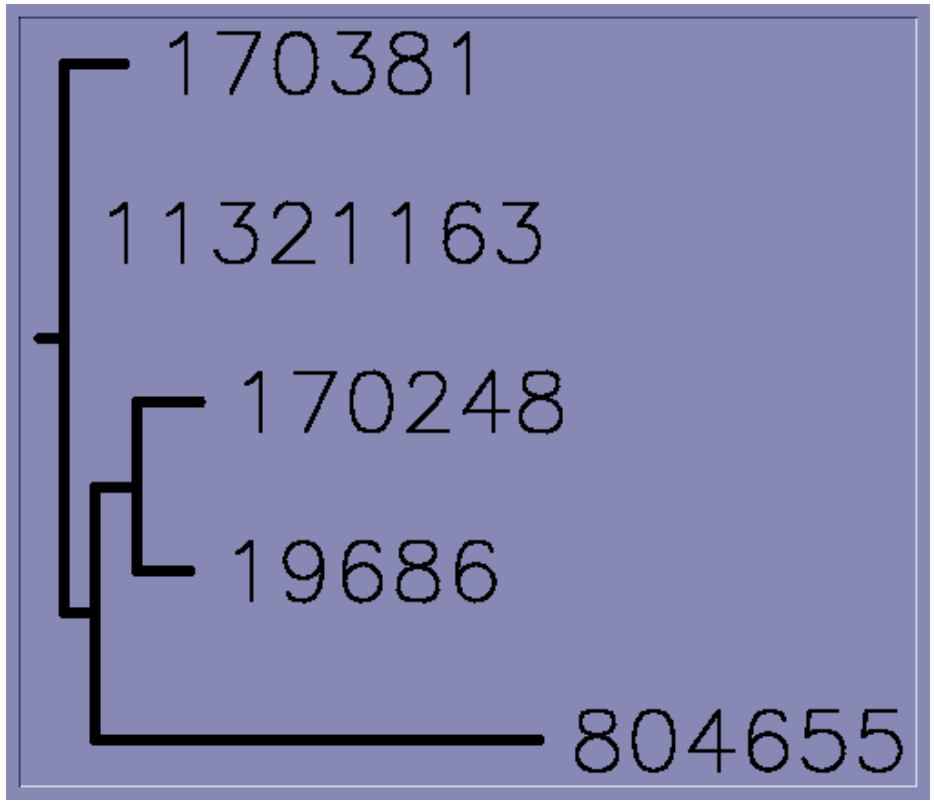
Analytical Tools

- http://workbench.sdsc.edu/



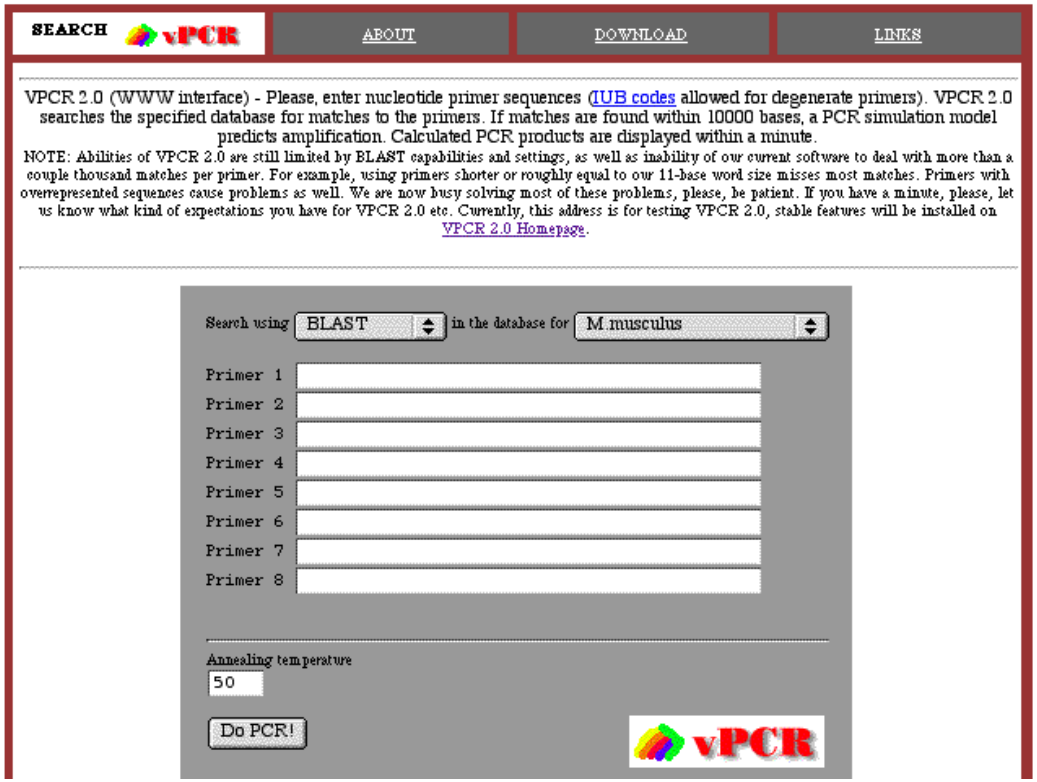
Analytical Tools

- http://workbench.sdsc.edu/



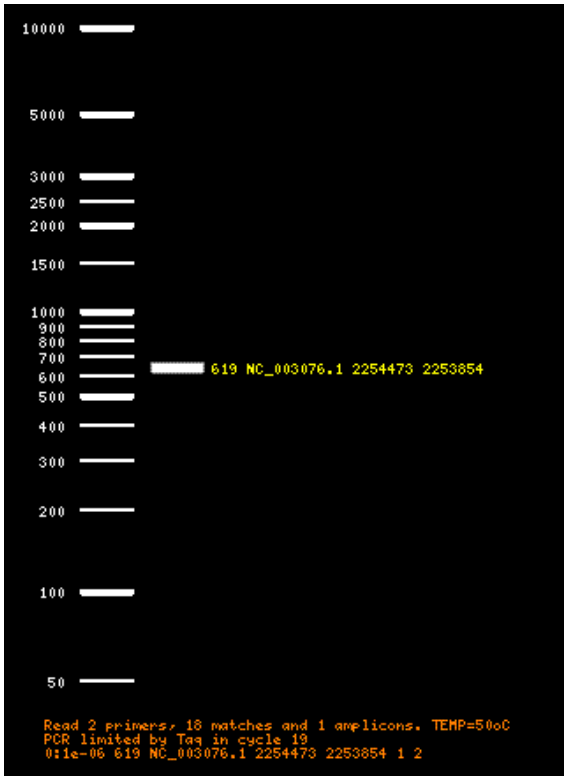
Analytical Tools

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



Analytical Tools

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



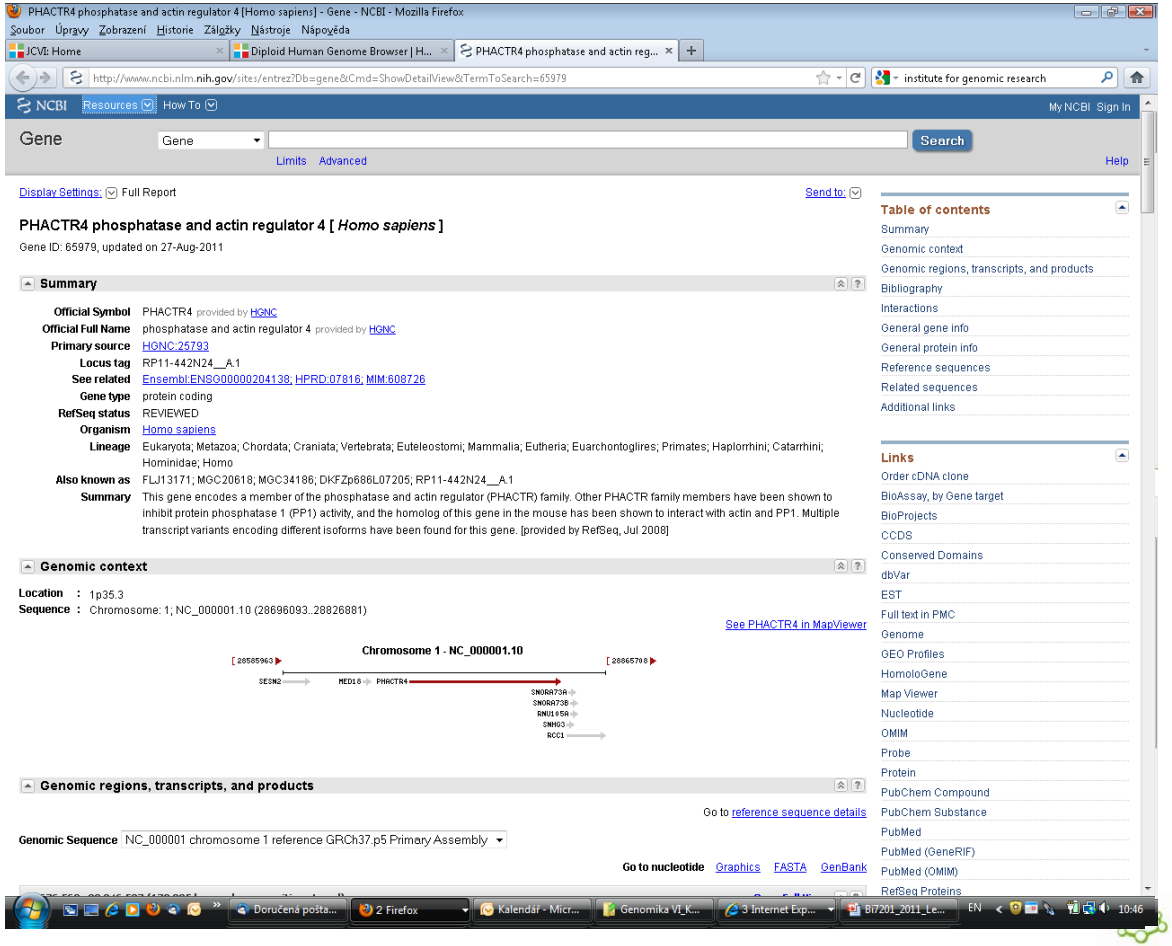
CEITEC

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

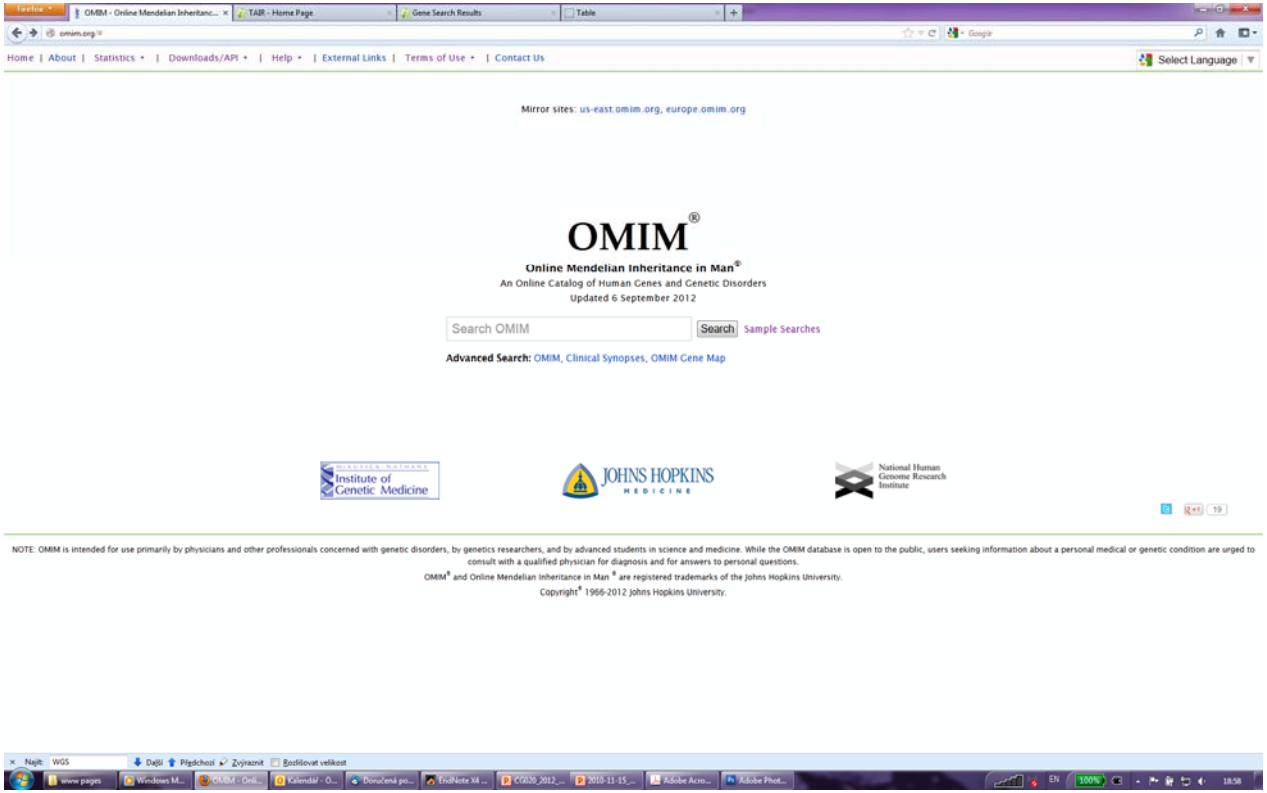
Other On-Line Genome Resources

- **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

Discussion