CG920 Genomics

Lesson 1

Introduction into Bioinformatics

Jan Hejátko

Functional Genomics and Proteomics of Plants,
Mendel Centre for Plant Genomics and Proteomics,
Central European Institute of Technology (CEITEC), Masaryk University, Brno
hejatko@sci.muni.cz, www.ceitec.muni.cz

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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 genomic tools

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Course Syllabus

- **Chapter 01**
  - Introduction to bioinformatics
- **Chapter 02**
  - Identification of genes
- **Chapter 03**
  - Reverse genetics approaches
- **Chapter 04**
  - Forward genetics approaches

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Course Syllabus

- **Chapter 05**
  - Funcional genomics approaches
- **Chapter 06**
  - Protein-protein interactions and their analysis
- **Chapter 07**
  - Current DNA-sequencing methods
- **Chapter 08**
  - Structural genomics

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Course Syllabus

- **Chapter 09**
  - Localization of genes and gene products in the cell
- **Chapter 10**
  - Genomics and systems biology
- **Chapter 11**
  - Practical aspects of functional genomics
- **Chapter 12**
  - Tools of systems biology
  - Model organisms, PCR and PCR primer design

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Literature

- Literature sources for Chapter 01:
  - **Bioinformatics and Functional Genomics**, 2009, Jonathan Pevsner, Willey-Blackwell, Hobocken, New Jersey http://www.bioinfbook.org/index.php
  - **Úvod do praktické bioinformatiky**, Fatima Cvrčková, 2006, Academia, Praha
  - **Plant Functional Genomics**, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey

Outline

- Syllabus of this course
- Definition of genomics

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

GENOMICS – What is it?

- *Sensu lato* (in the broad sense) – it is interested in **STRUCTURE and FUNCTION** of genomes
  - Condition: knowing 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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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.

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

GENOMICS – What is it?

The role of BIOINFORMATICS in FUNCTIONAL GENOMICS

Forward („classical“) genetics approaches

Reverse genetics approaches

5‘TTATATATATATATTAAAAAATAAAATAAAA
GAACAAAAAAGAAAATAAAATA….3‘

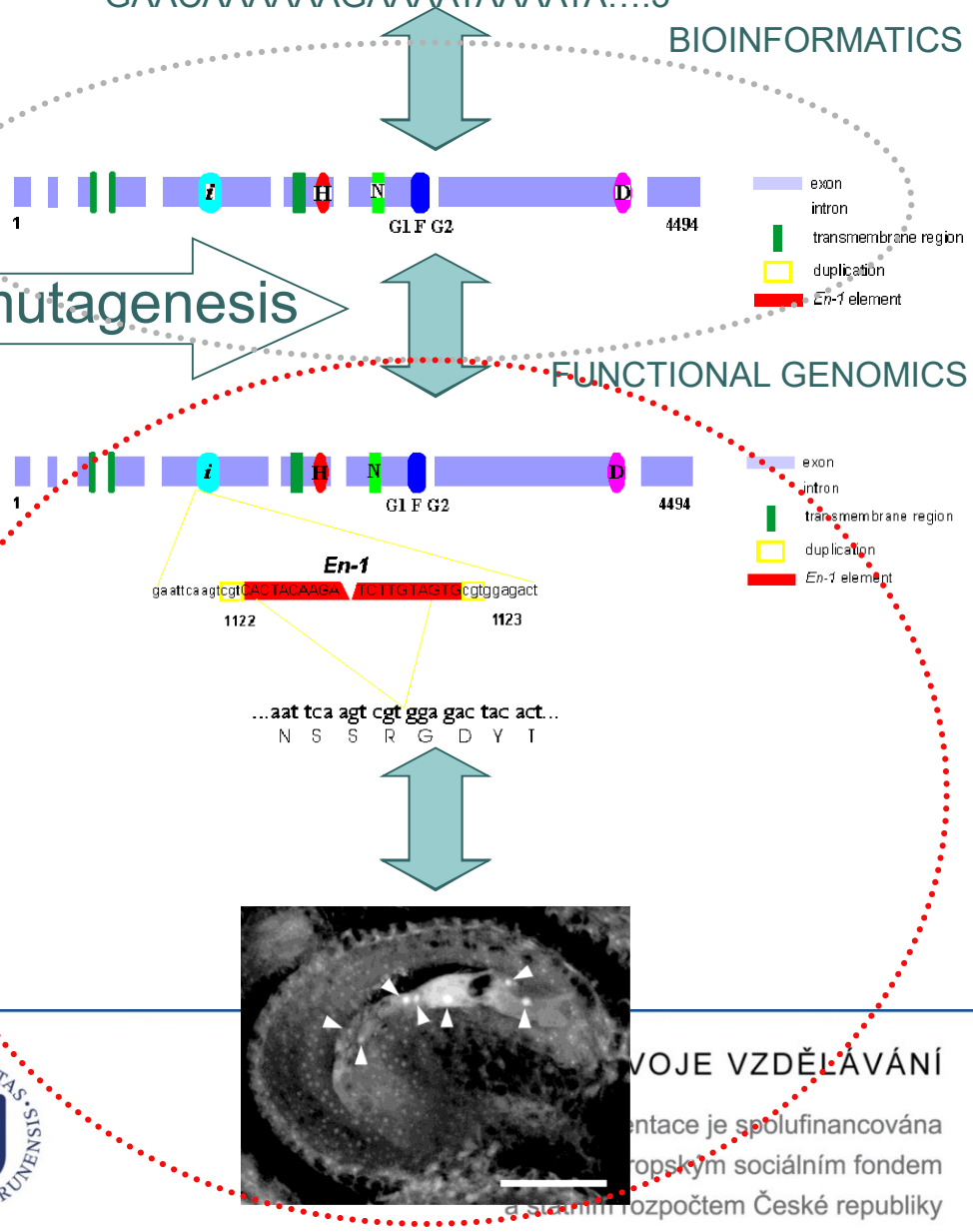
3 : 1

?

BIOINFORMATICS

Insertional mutagenesis

FUNCTIONAL GENOMICS



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**.

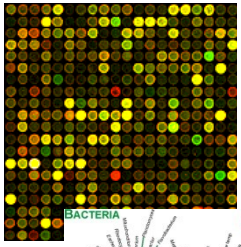
EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

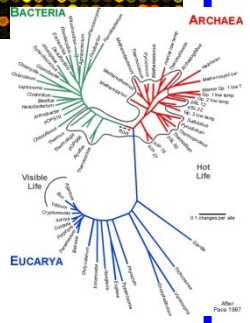
Outline

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

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

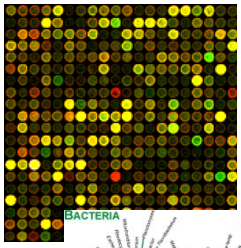
What is bioinformatics?

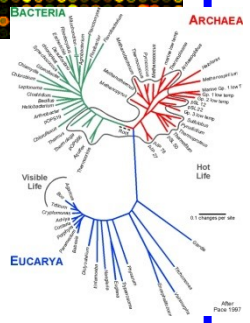
- Interface of biology and computers
- Analysis of proteins, genes and genomes using computer algorithms and computer databases
- Genomics is the analysis of genomes. The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.

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

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Bioinformatics





- **Bioinformatics** in **functional genomics**
  - **Processing and analysis of sequencing data**
    - Identification of reference sequences
    - Identification of genes
    - Identification of homologs, orthologs and paralogs
    - Correlation 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 regualtion networks

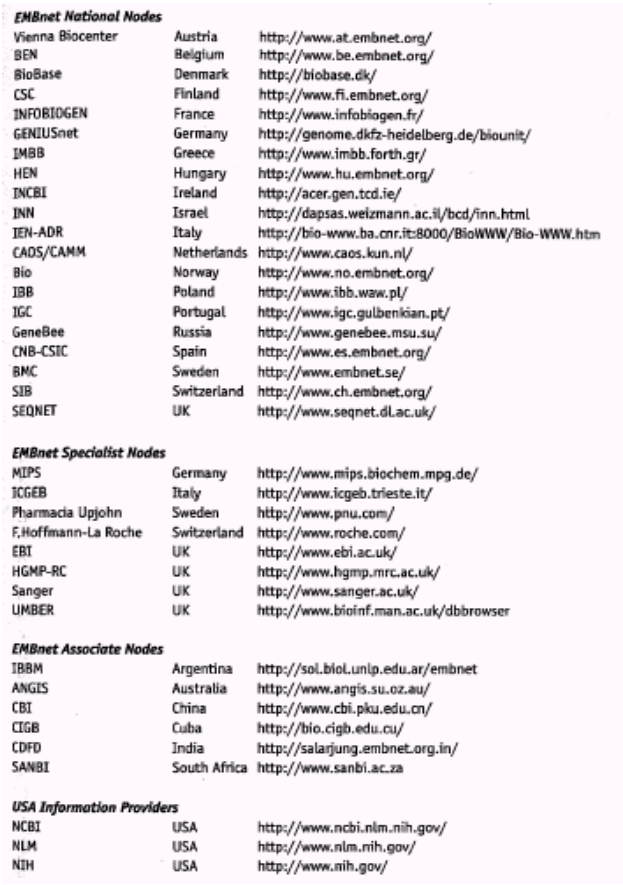
EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Outline

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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Spectre of on-line resources

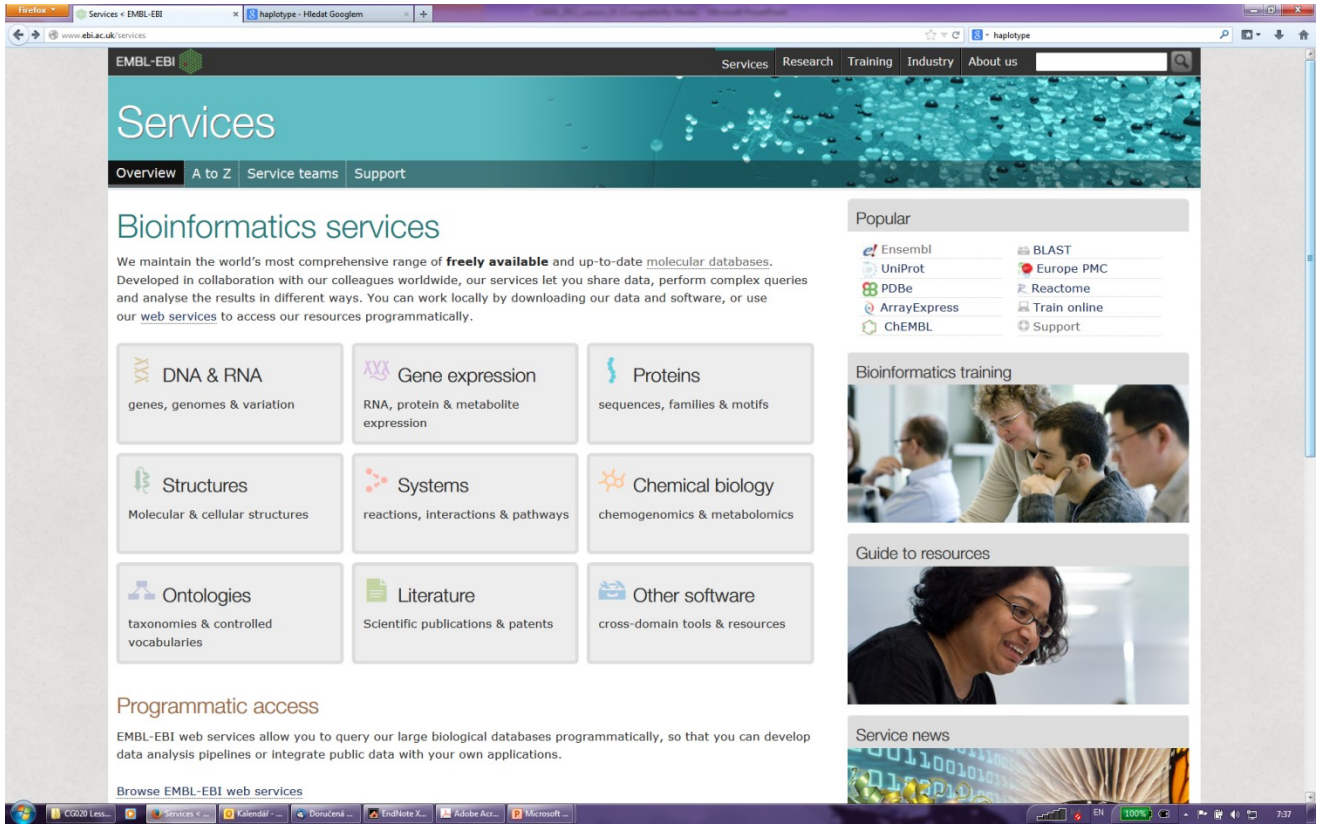


EMBnet National Nodes		
Vienna Biocenter	Austria	http://www.at.embnet.org/
BEN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENIUSnet	Germany	http://genome.dkfz-heidelberg.de/biounit/
IMBB	Greece	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/

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

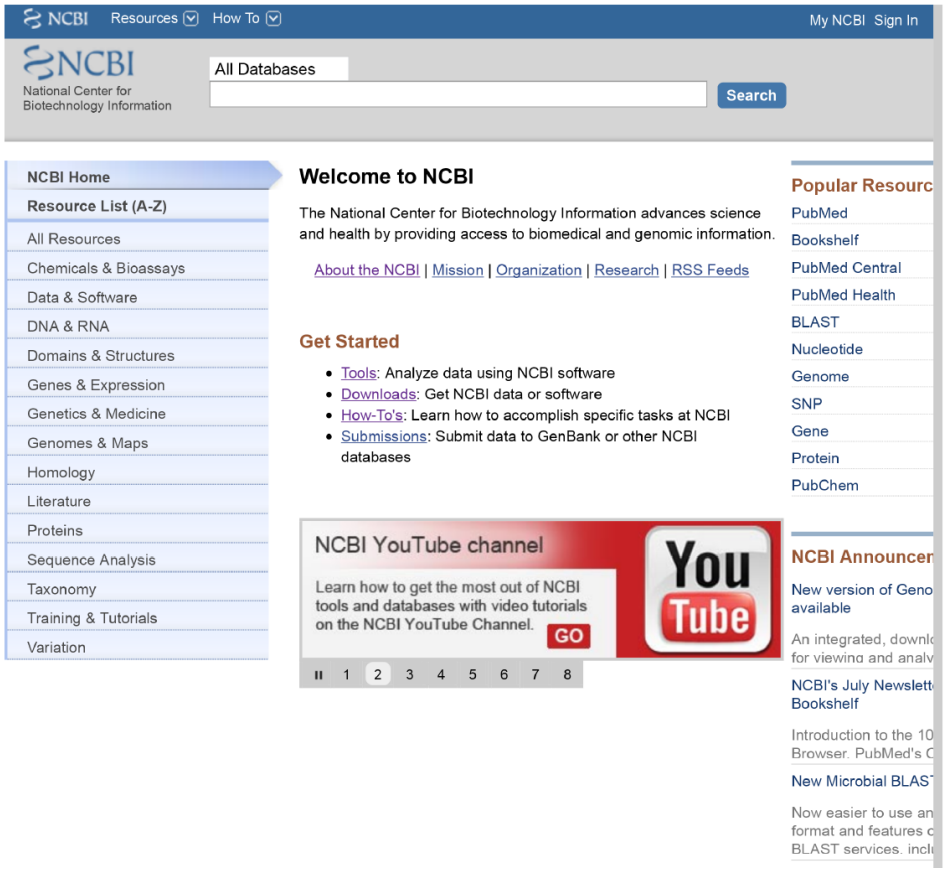
Spectre of on-line resources

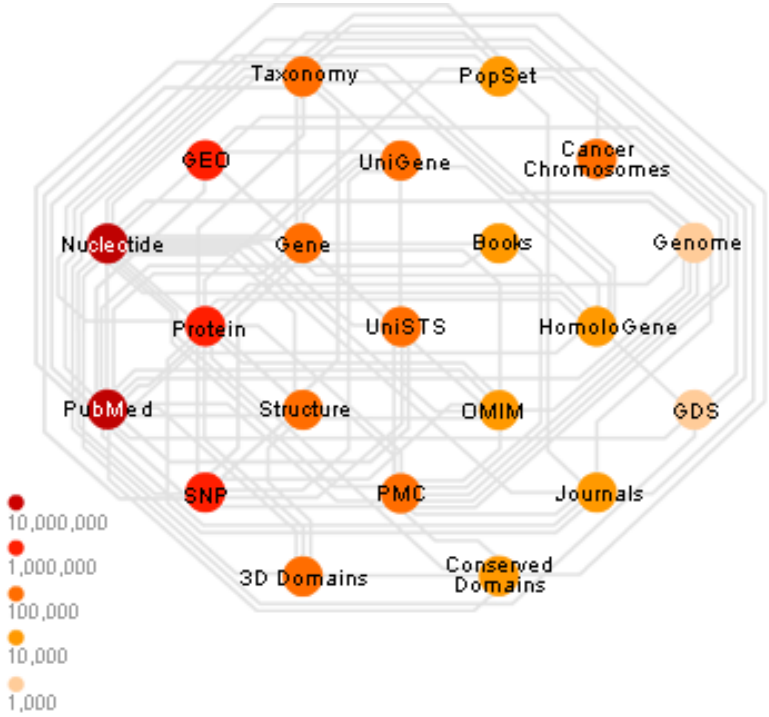
- EBI http://www.ebi.ac.uk/services



Spectre of on-line resources

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





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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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

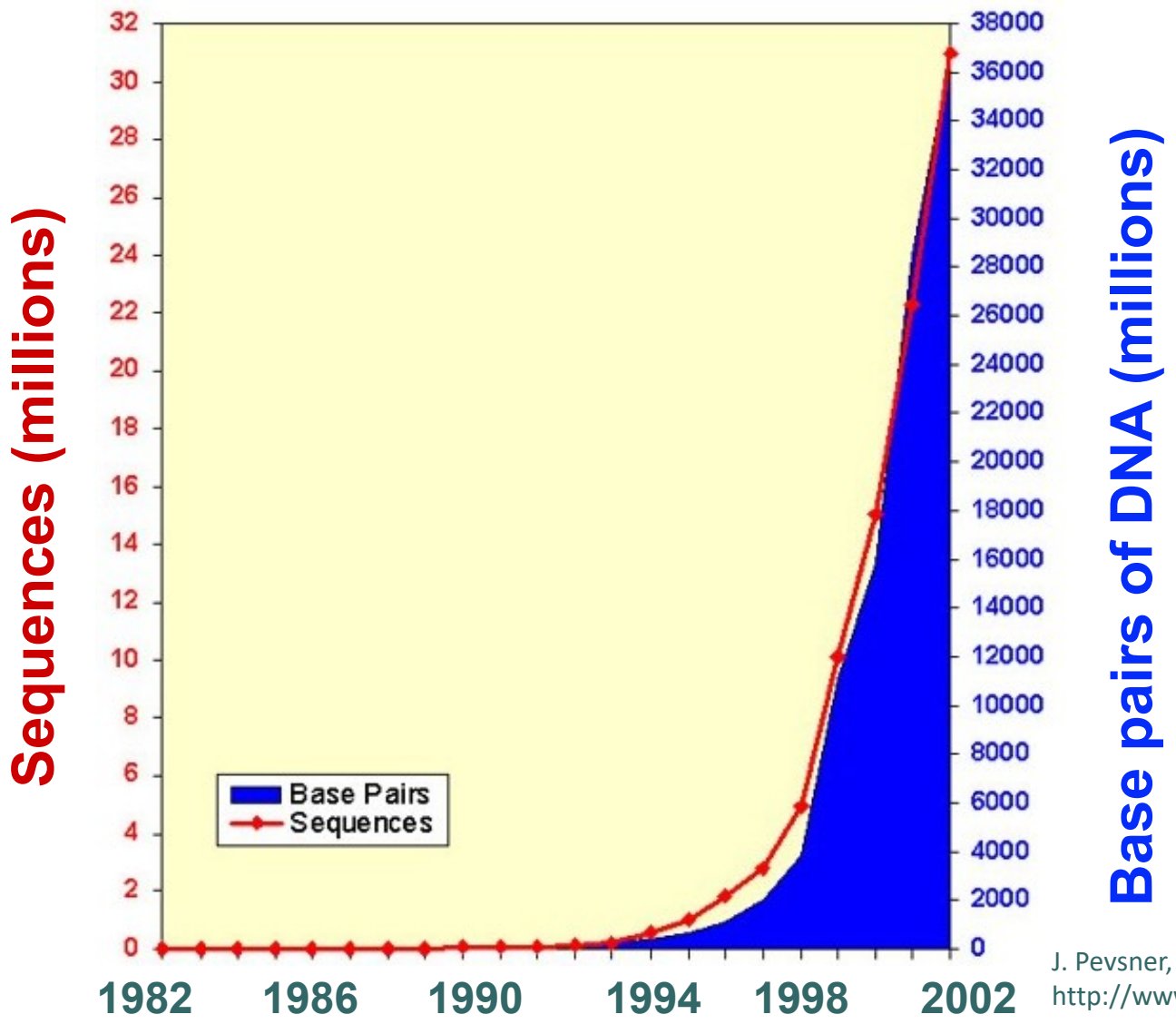
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Primary databases

- They include sets of primary data – 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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

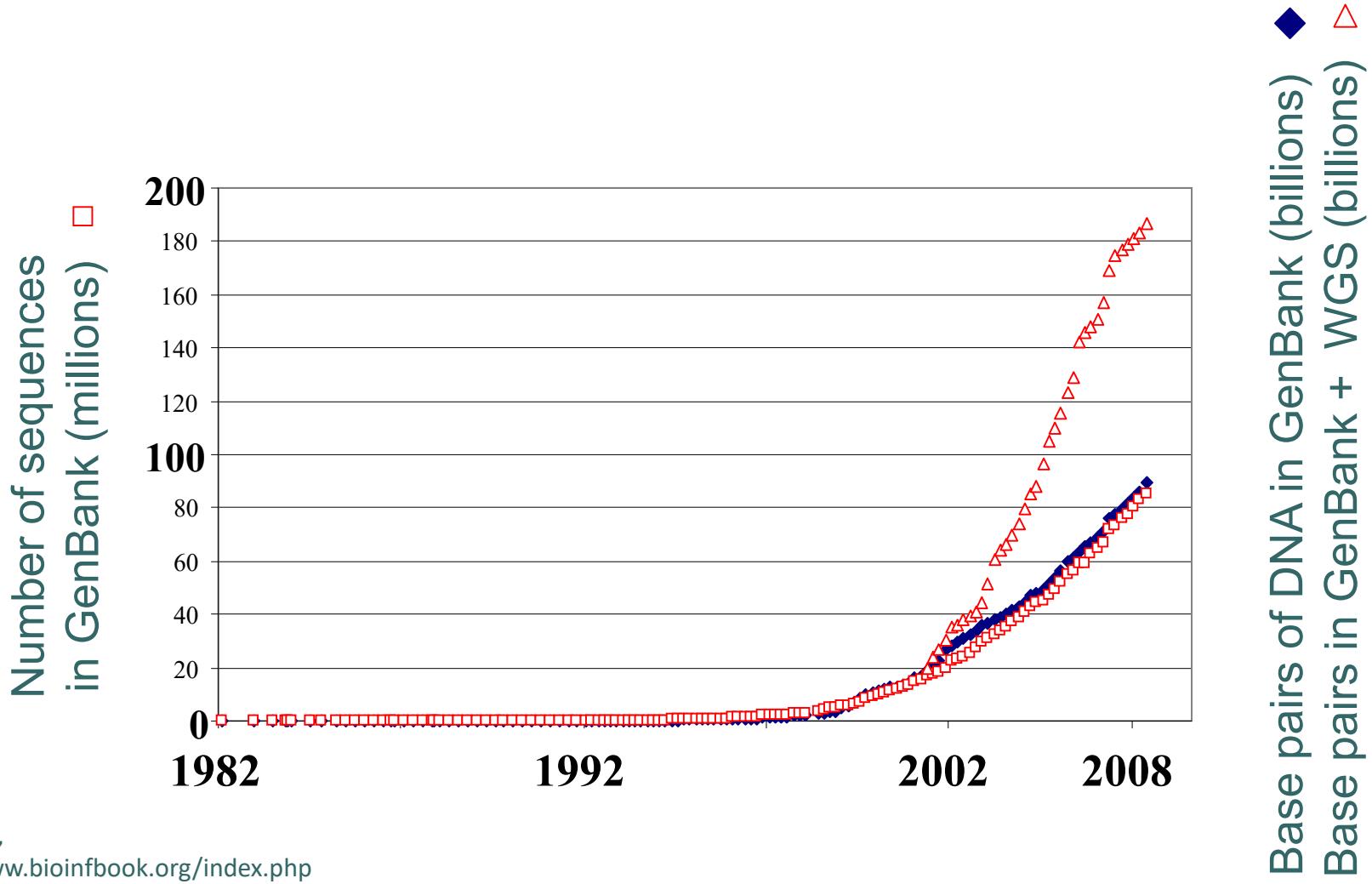
Growth of GenBank



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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
Year
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

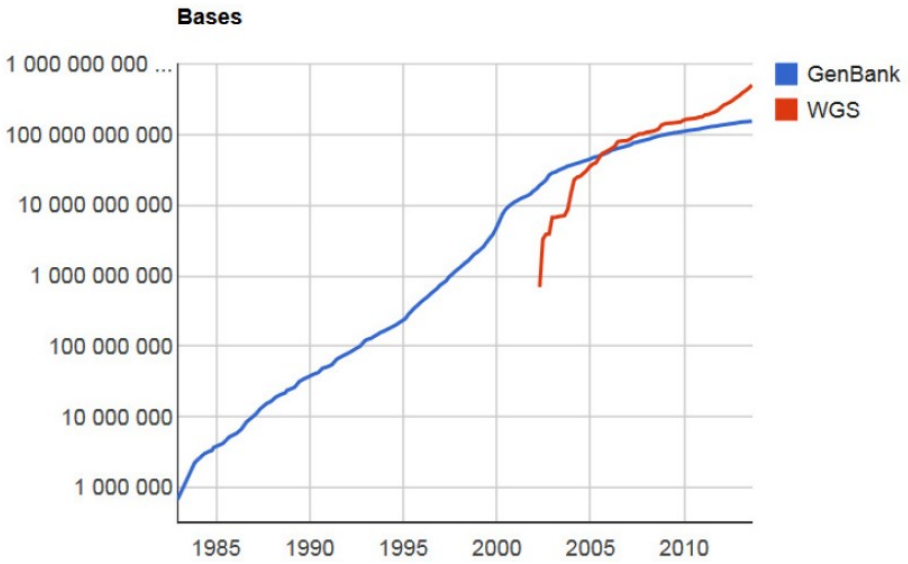
Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases

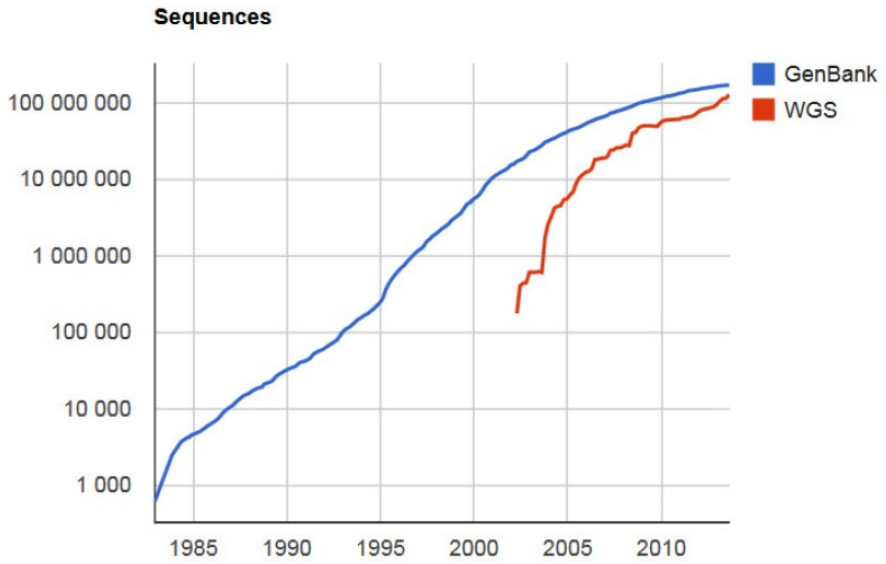


J. Pevsner,
http://www.bioinfbook.org/index.php

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

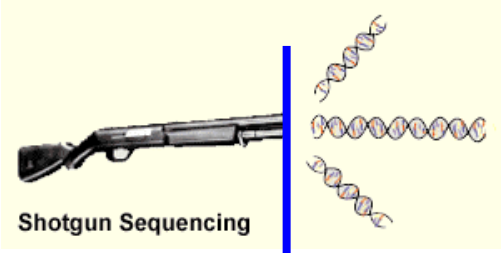
Growth of GenBank
Feb 15 2013





EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

WGS



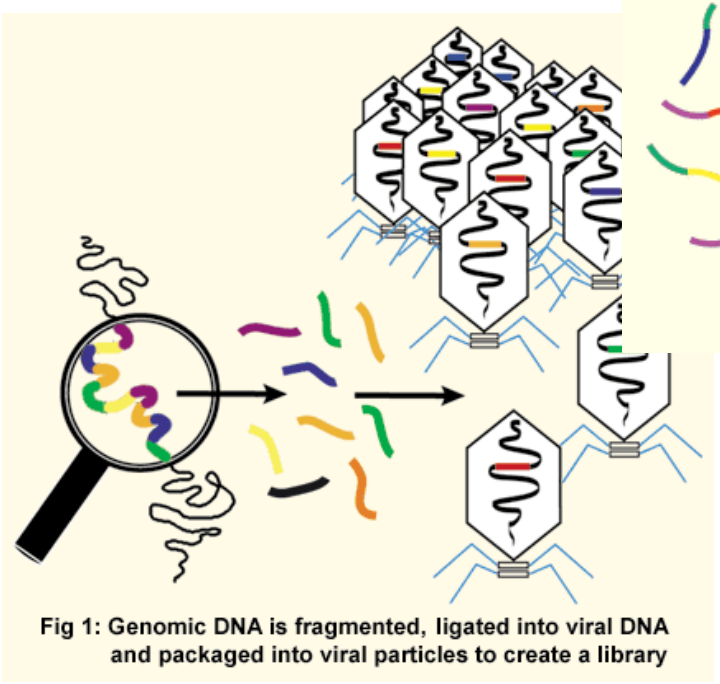


Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

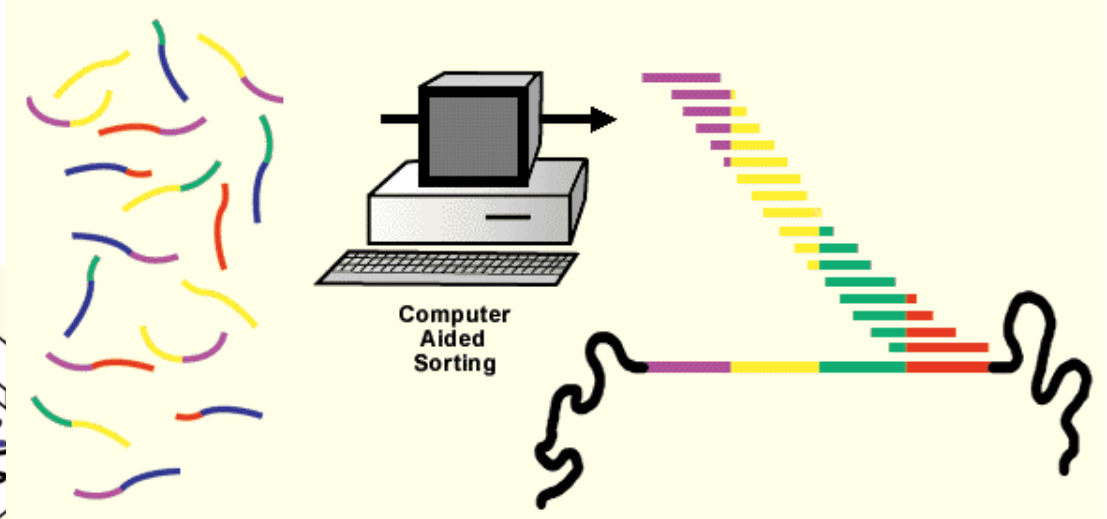


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

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

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

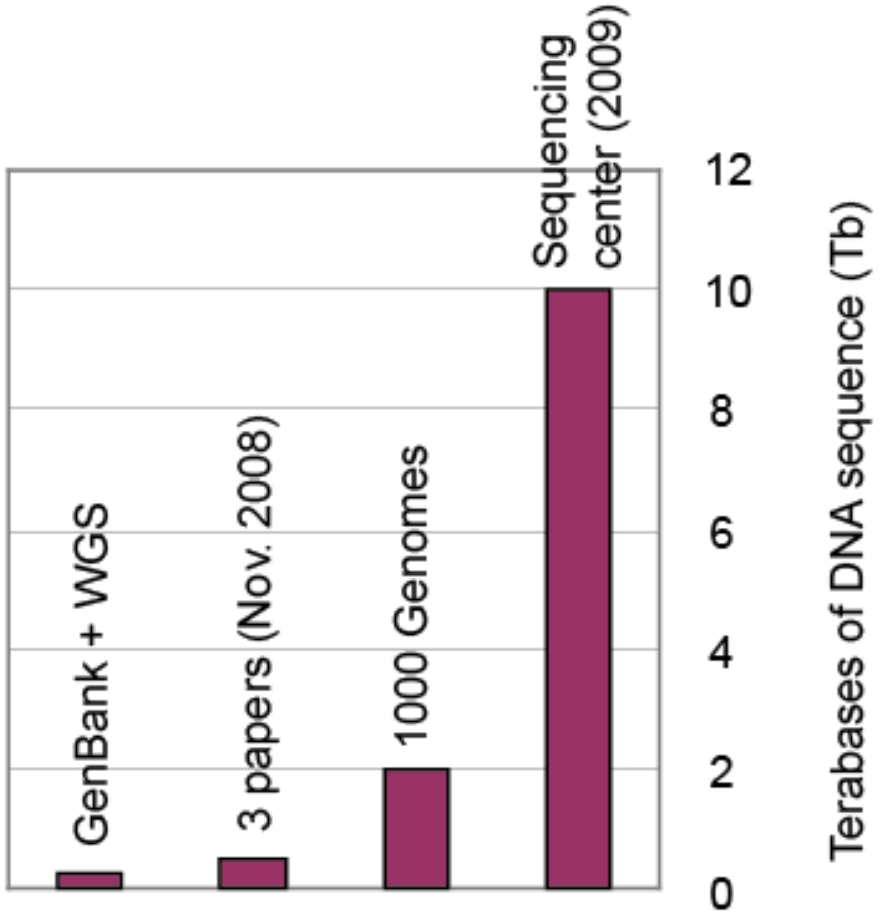
Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic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

Arrival of next-generation sequencing: In two years we have gone from 0.2 terabases to 71 terabases (71,000 gigabases) (November 2010)



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

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

DDBJ/EMBL/GenBank accepts both complete and incomplete genomes. Whole Genome Shotgun (WGS) sequencing projects are incomplete genomes or incomplete chromosomes that are being sequenced by a whole genome shotgun strategy. WGS projects may be annotated, but annotation is not required.

The pieces of a WGS project are the contigs (overlapping reads), and they do not include any gaps. An AGP file can be submitted to indicate how the contig sequences are assembled together into scaffolds (contig sequences separated by gaps) and/or chromosomes. We must have the contig sequences without gaps as the basic units for all WGS projects.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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/

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

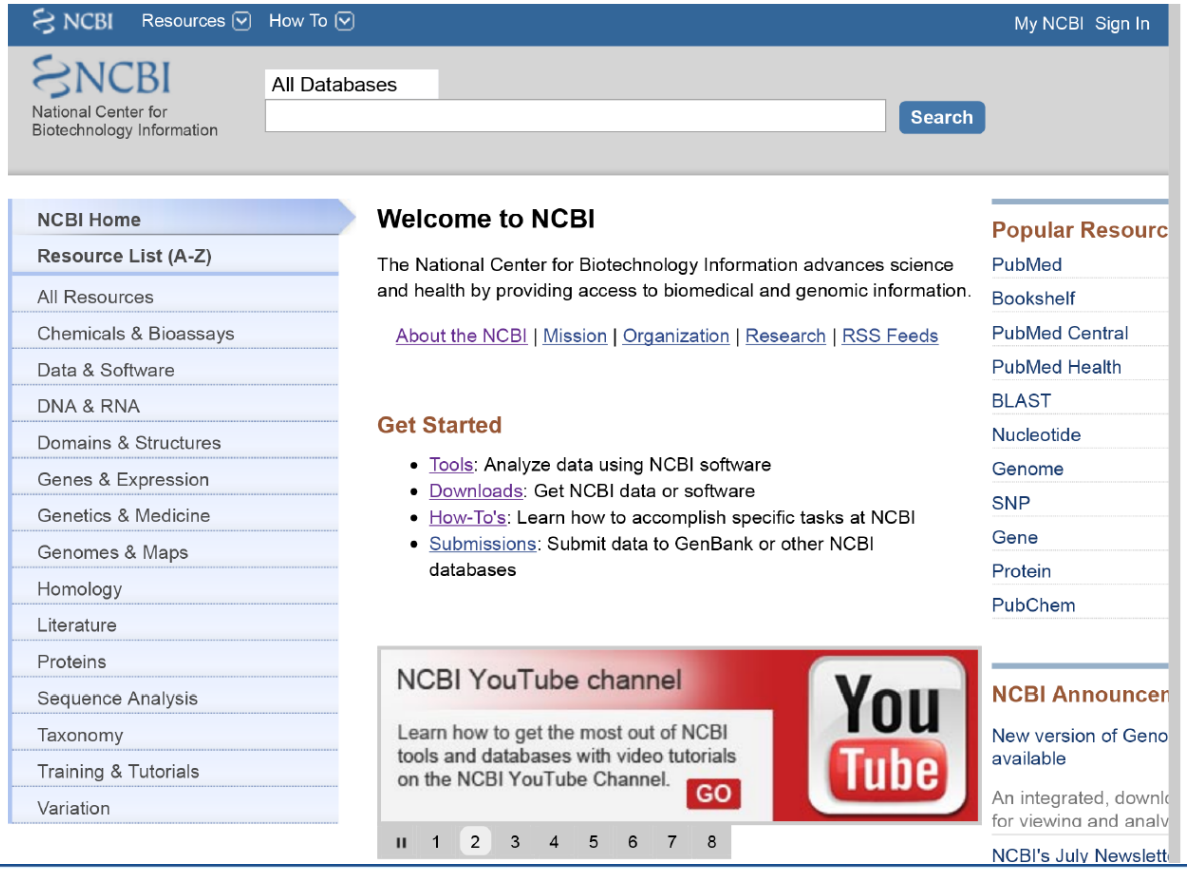
Primary databases

- 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)

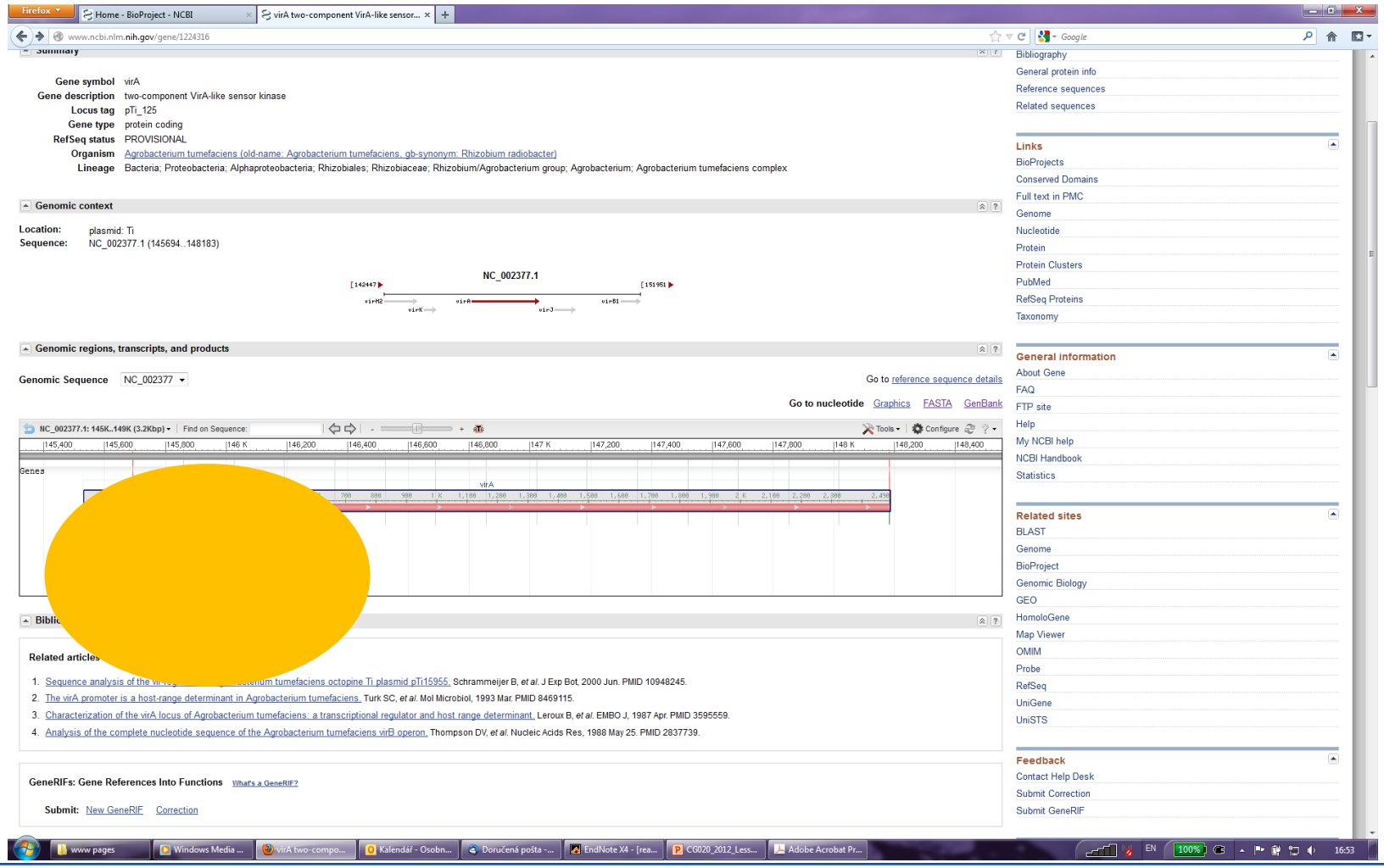
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Primary databases

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

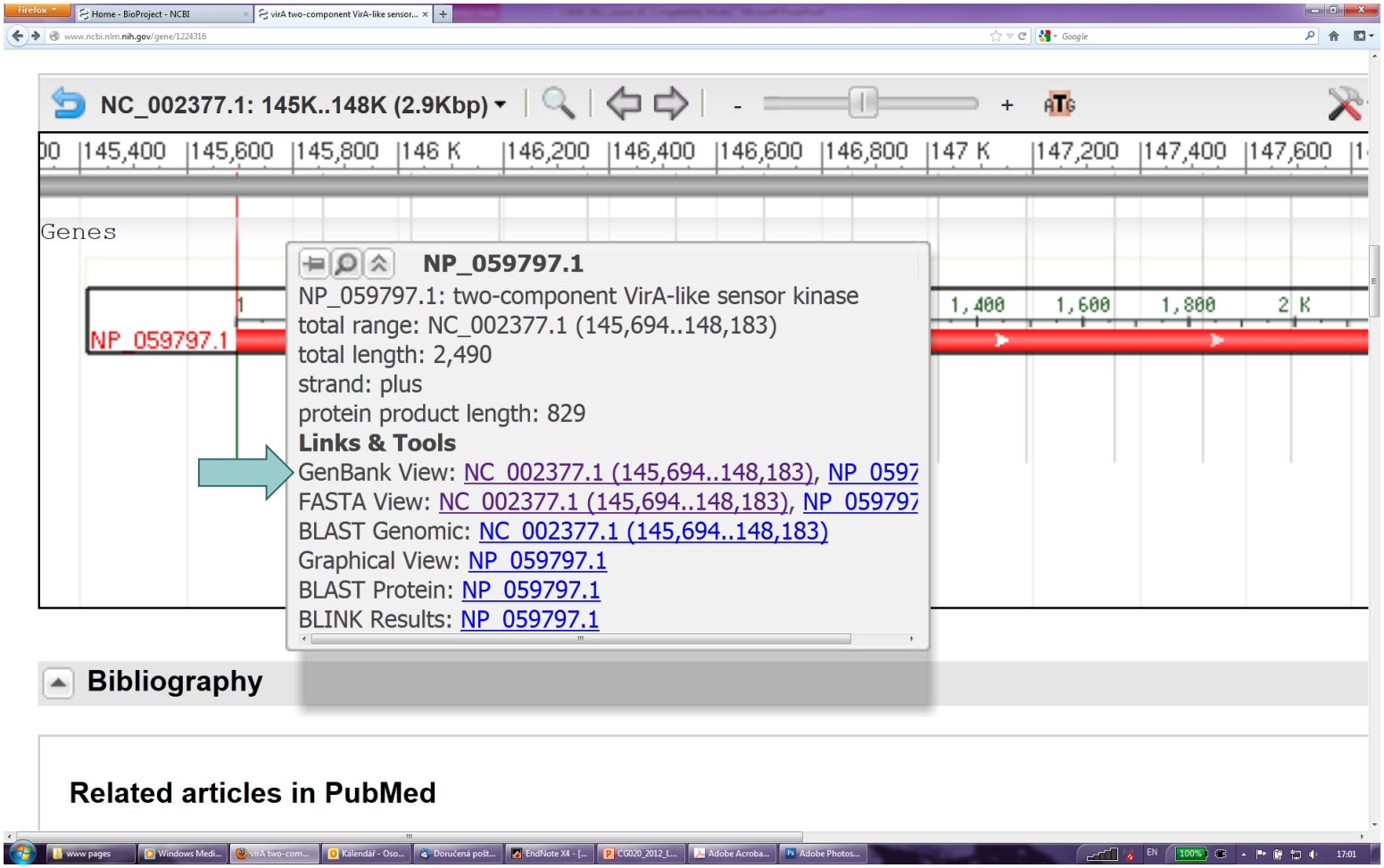


Primary databases

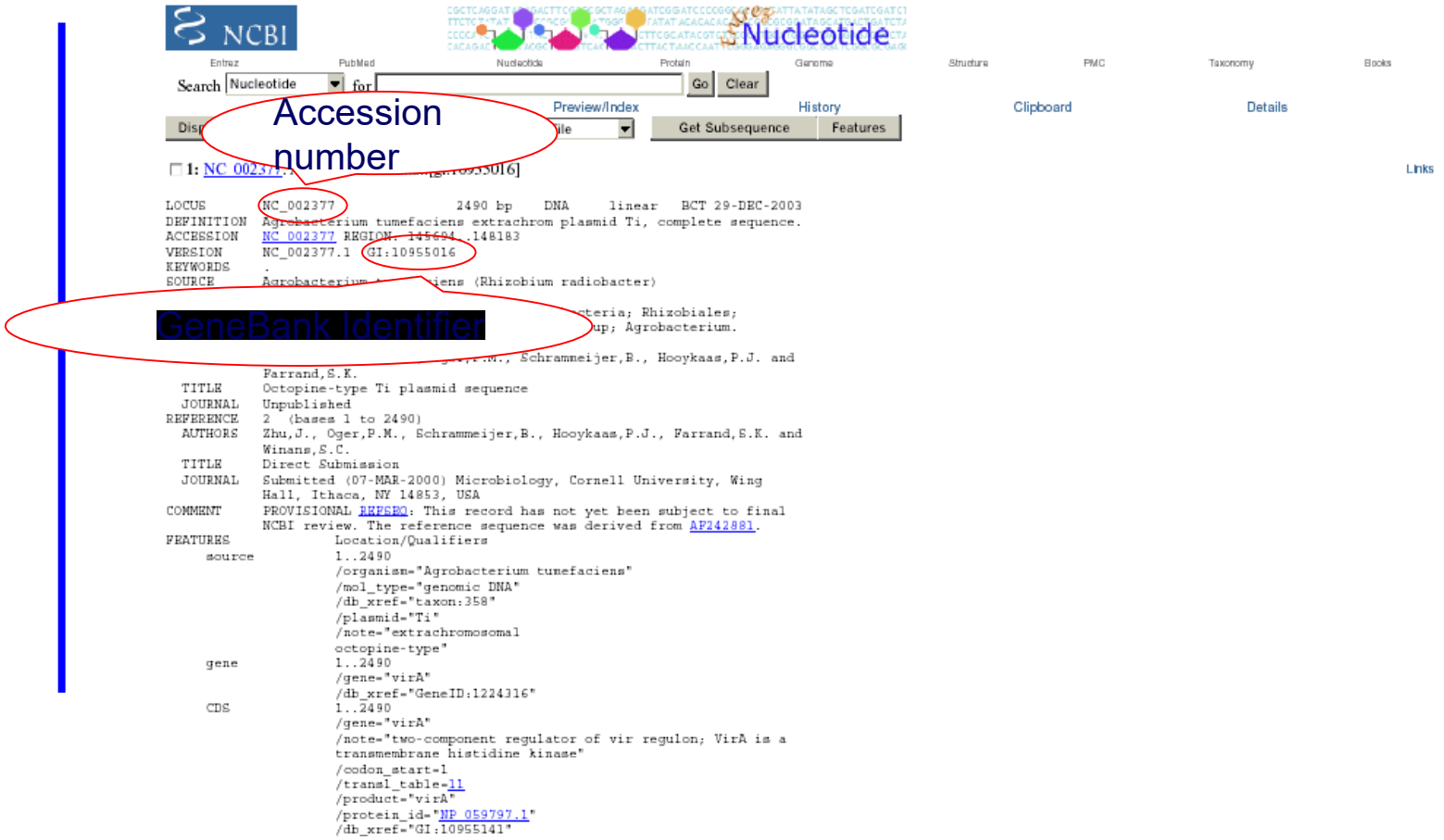


EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Primary databases



Primary databases



Primary databases

```
                     /translation="MNGRYSPTRQDFKTGAKPWSILALIVAAMIFAFMAVASWQDNAT
                     TQAILSQLRSINADSASLQRDVLRAHTGTVANYRPIISRLGALRKNLEDLKQLPRQSH
                     IVSENAAQLLRQLEVSLNSADAAVAAPGAQNVRLQDSLASPTRALSSLPGKASTDQT
                     LEKPTELASMMLQFLRQPSPAISFEISLELERLQKQRGLDEAPVRILAREGPIILSLL
                     PQVKDLVNMIQTSDTAEIAEMLQRECLEVYSLKNVEERSARIFLGSASVGLCLYIITL
                     VYRLRKKTDWLARRLDYEELIKEIGVCFPGHAATTSSAQAALRIIQRFFDADTCALAL
                     VDHDRRWAVETPGAKHPKPVWDDSVLREIVSRTKADERATVPRIISSKKIVHLPLEIP
                     GLSILLAHKSTDKLIAVCSLGYQSYRPRPCQGEIQLLELATACLCHYIDVRRKQTRCD
                     VLARRLEHAQRLEAVGTLAGGIAHEFNNILGSILGHAELAQNSVSRTSVTRRYIDYII
                     SSGDRAMLIIDQILTLSRKQERMIKPPSVSELVTEIAPLLRMALPPNIELSPRFDQMQ
                     SVIEGSPLELQQVLINICKNASQAMTANGQIDIIISQAFLPVKKILAHGVMPPGDYVL
                     LSISDNGGGIPEAVLPHIFEPFFTTRARNGGTGLGLASVHGHISAPAGYIDVSSTVGH
                     GTRFDIYLPPSSKEPVNPDSPFGRNKAPRGNGEIVALVEPDDLLREAYEDKIAALGYE
                     PVGPRTPNEIRDWISKGNEADLVMVDQASLPEDQSPNSVDLVLKTASIIIGGNDLKMT
                     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 ccgcccccatt
      241 atctccaggc tgggagctct gcggaagaat ctggaagatt tgaagcaatt atttagacaa
      301 tctcatattg taagtgagag caatgctgct caactgctac gccagctaga agtgtctcta
      361 aatttggctg acgcggcggt cgccgccttt ggtgcgcaaa atgtacgcct gcaagatttcg
      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 tctgacacog cagaaattgc ggagatgctg
      721 cagcgcgagt gtttggaggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc
      781 tttcttgggt ccgcttcagt gggtctttgc ctctacatca tcaccttagt ctataggcta
      841 cgcaaaaaaa cogattggtt 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 tcgcctcga
     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 tcgctcccct 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 ttcccgaggc tgtgttaccc
     1921 cacatttttg aacccttctt tacgacacga gctcgcaacg gtggaacggg tctcggcctt
     1981 gcttctgtgc atggtcatat cagcgcgttt gcgggttaca tcgacgttag ttcaactgtt
     2041 gggcatggga cgcgctttga catttatctc cctcogtctt 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 gacccttca
```

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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.

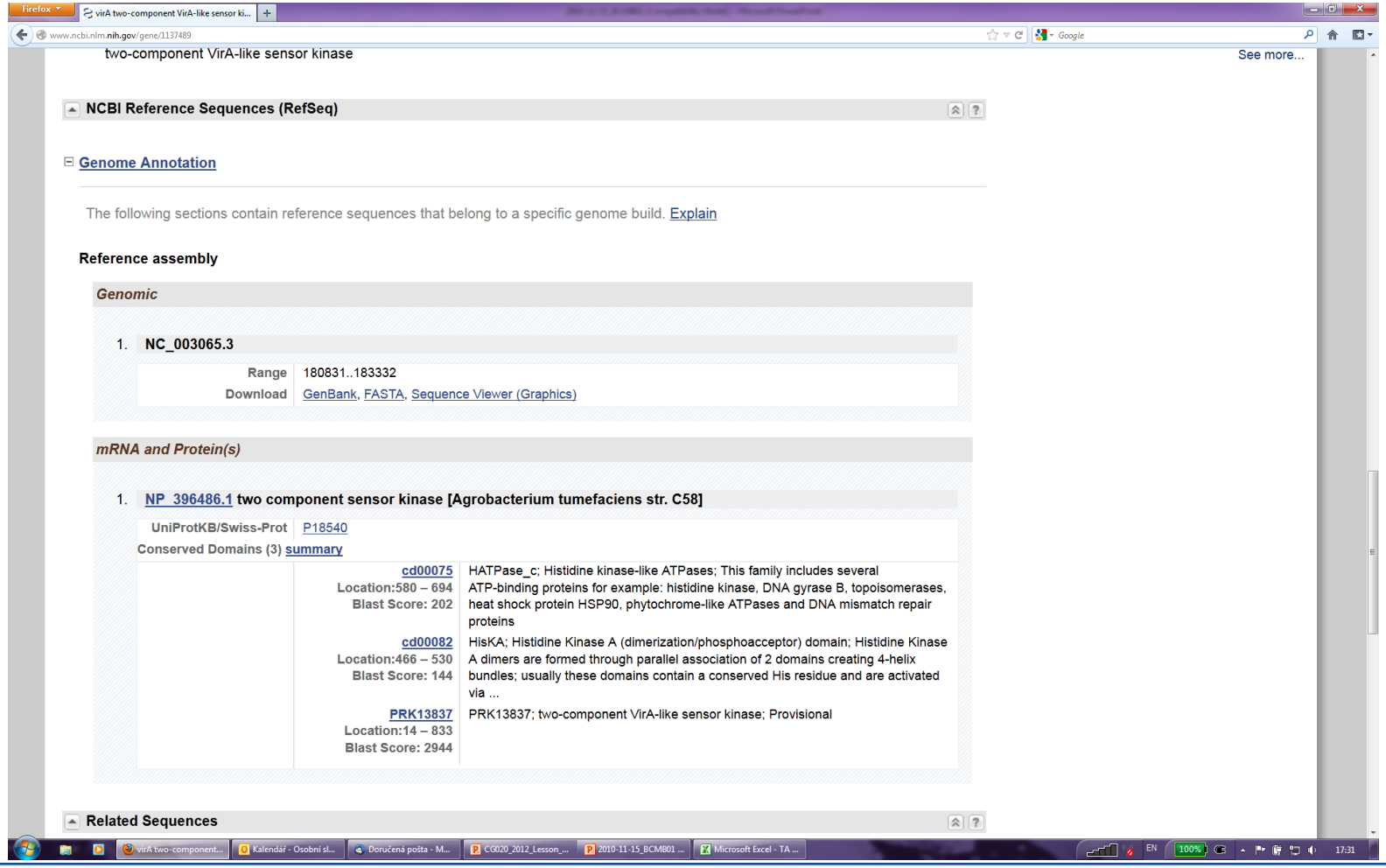
RefSeq identifiers include the following formats:

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

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

RefSeq



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

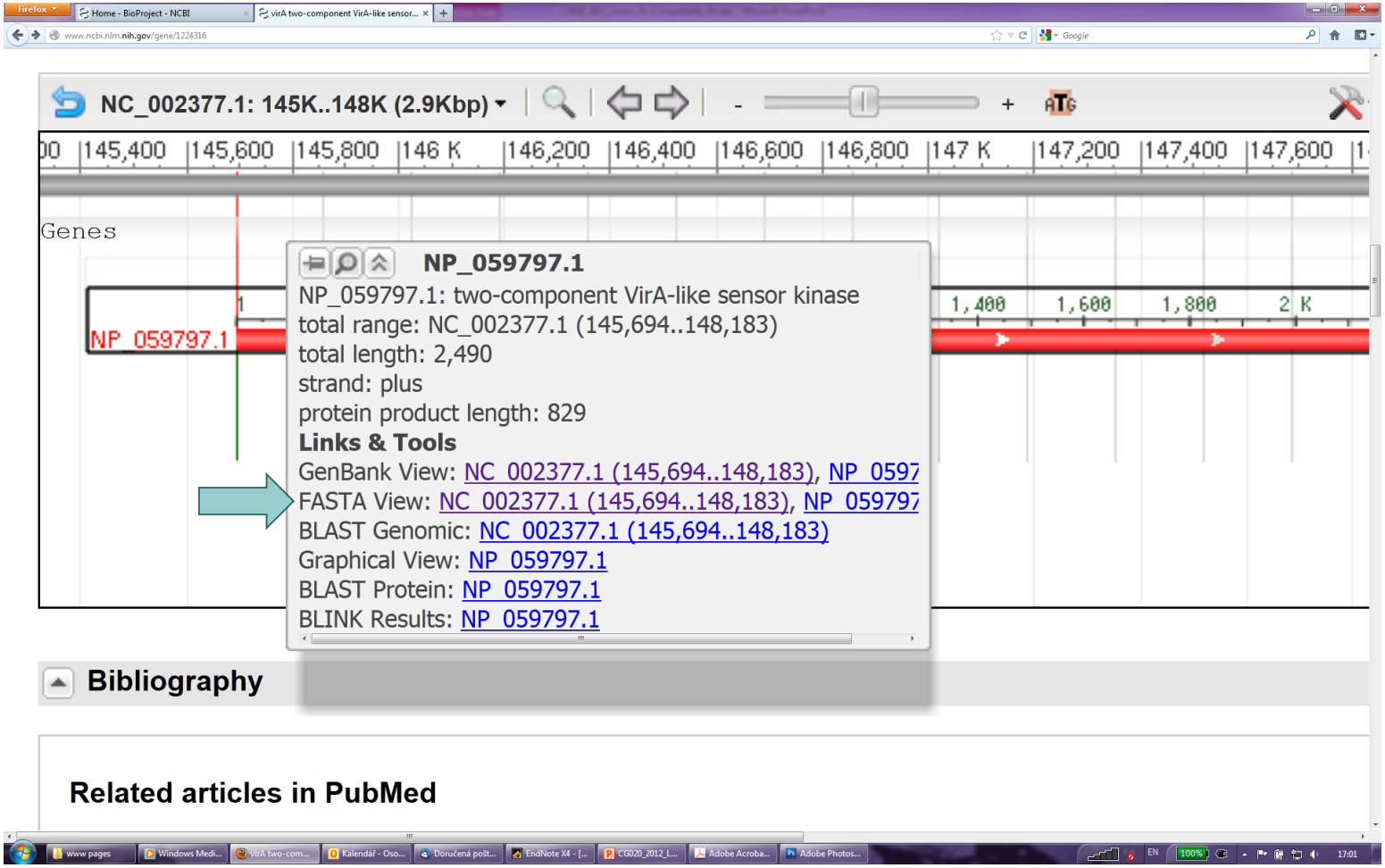
NCBI's RefSeq project: many accession number formats for genomic, mRNA, protein sequences

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

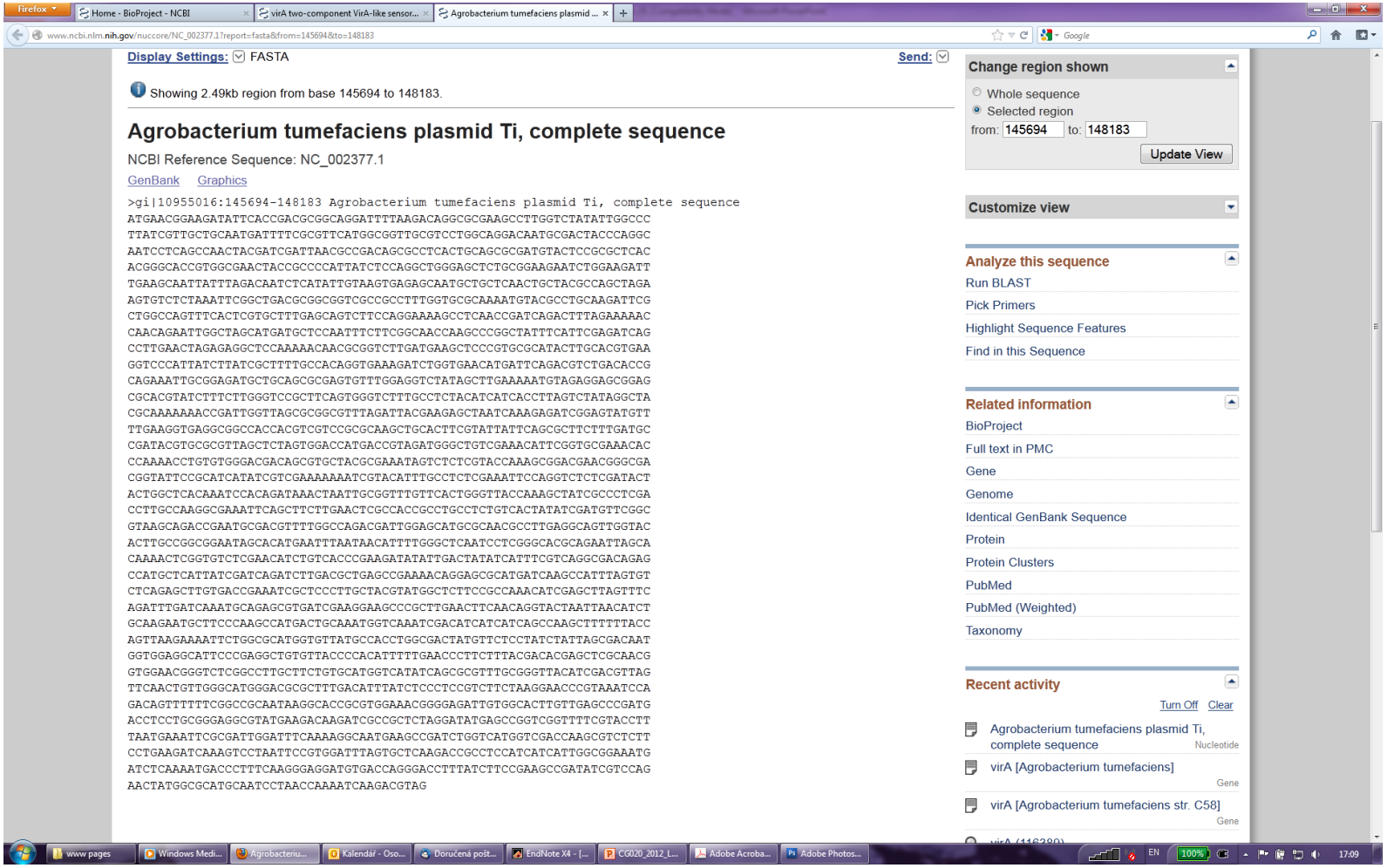
J. Pevsner,
http://www.bioinfbook.org/index.php

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Primary databases



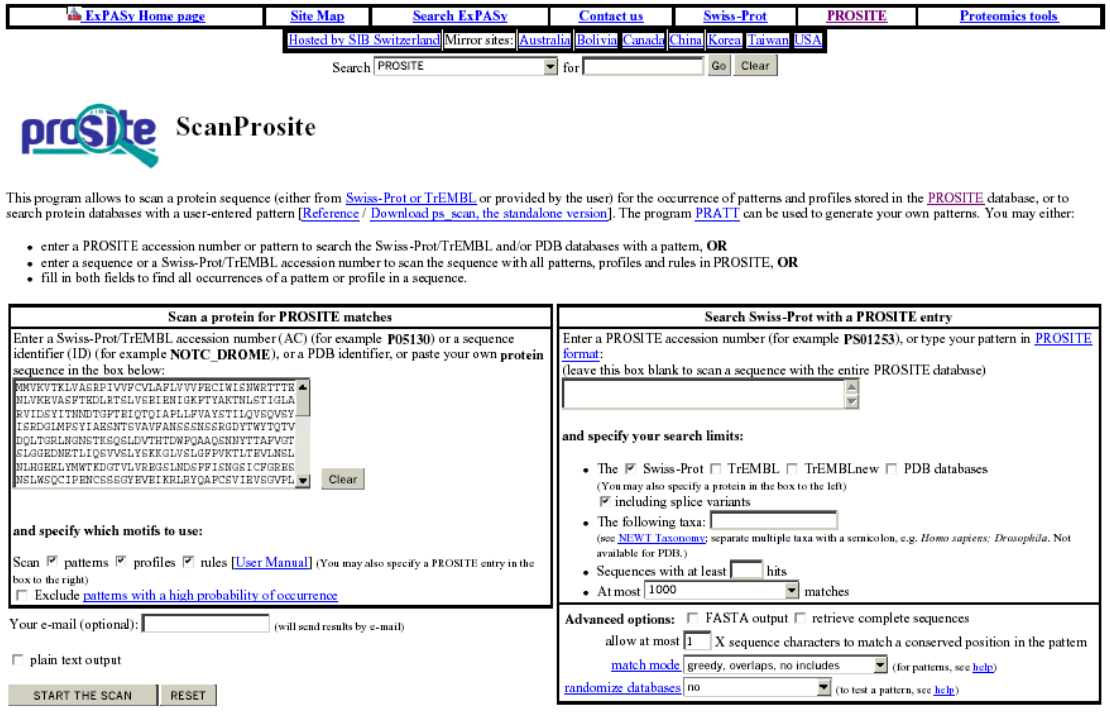
Primary databases



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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 nkeesstYeteisns

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

744 - 747 RRvT
814 - 817 KRrS

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

148 - 150 SsR
164 - 166 TgR
171 - 173 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 TsK
997 - 999 SrK
1002 - 1004 TgK
1018 - 1020 SgK
1031 - 1033 TqR
1119 - 1121 SkR

Secondary databases

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

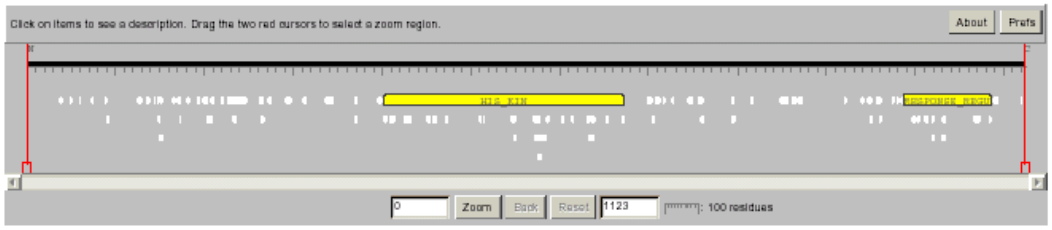
>PDOC50109 PS50109 **HIS_KIN** Histidine kinase domain [profile].

```
  402 - 671  NASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVNVCAKDLVALLNSVLDMSKIRSG
             KMQLVREDFNLSKLLKDVIDFYHPVAMKKGVDVVLDPHDgsvfKPSNVRGDSGRLKQILN
             NLVSNAVKFTVD--GHIAVRAWAQrpgsnsavvlaaypkgvakfvksmfcknkeesstye
             teisnsirnnanTMEFVPEVDDTGKGIPMEMRKSVPENYVQVRRtAQGHQGTGLGLGIVQ
             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

Secondary databases

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

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

PRINTS is a compendium of protein **fingerprints.** A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a *SWISS-PROT/TrEMBL* composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. References

New:

- 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

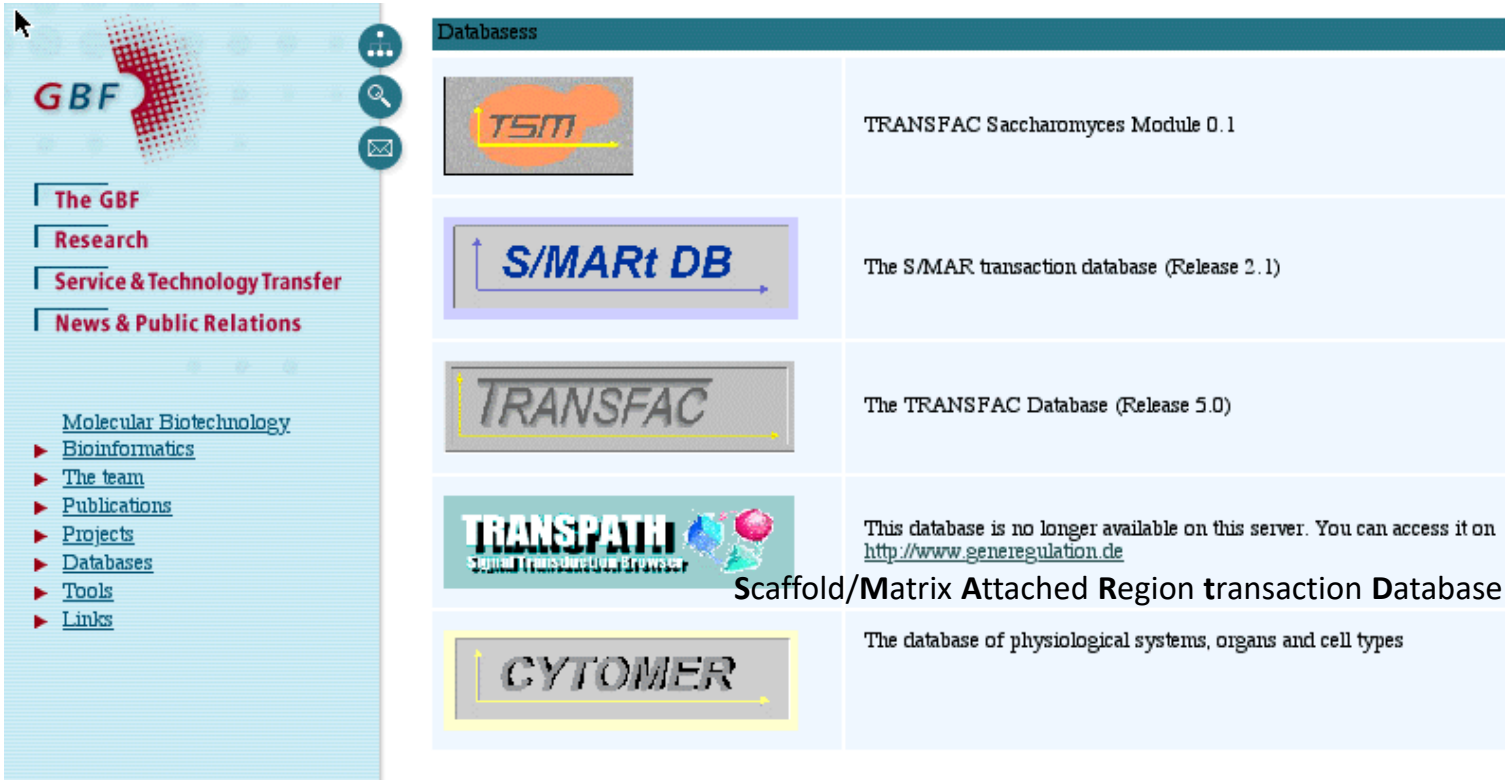
PRINTS search:

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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Secondary databases

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



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

S/MARt DB (scaffold/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)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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

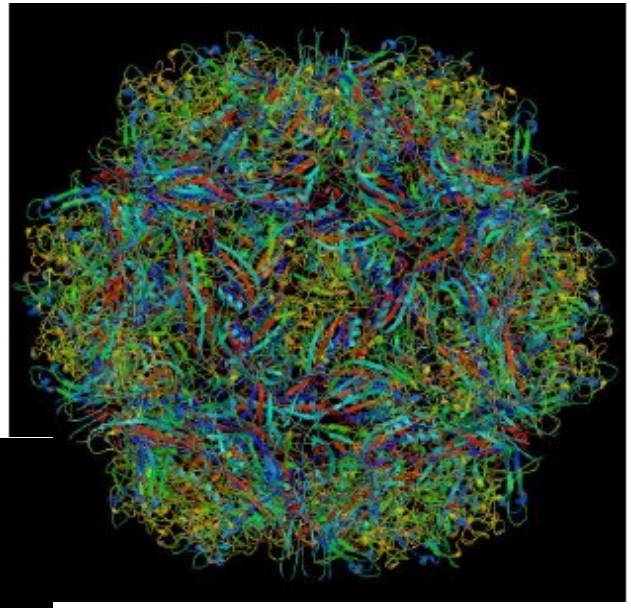
Geometry

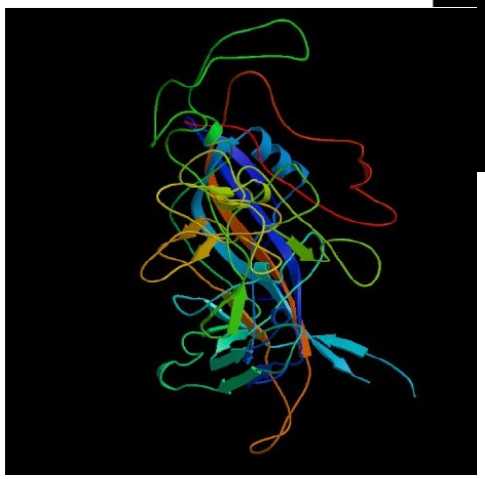
Other Sources

Sequence Details

Explore

SearchLite SearchFields

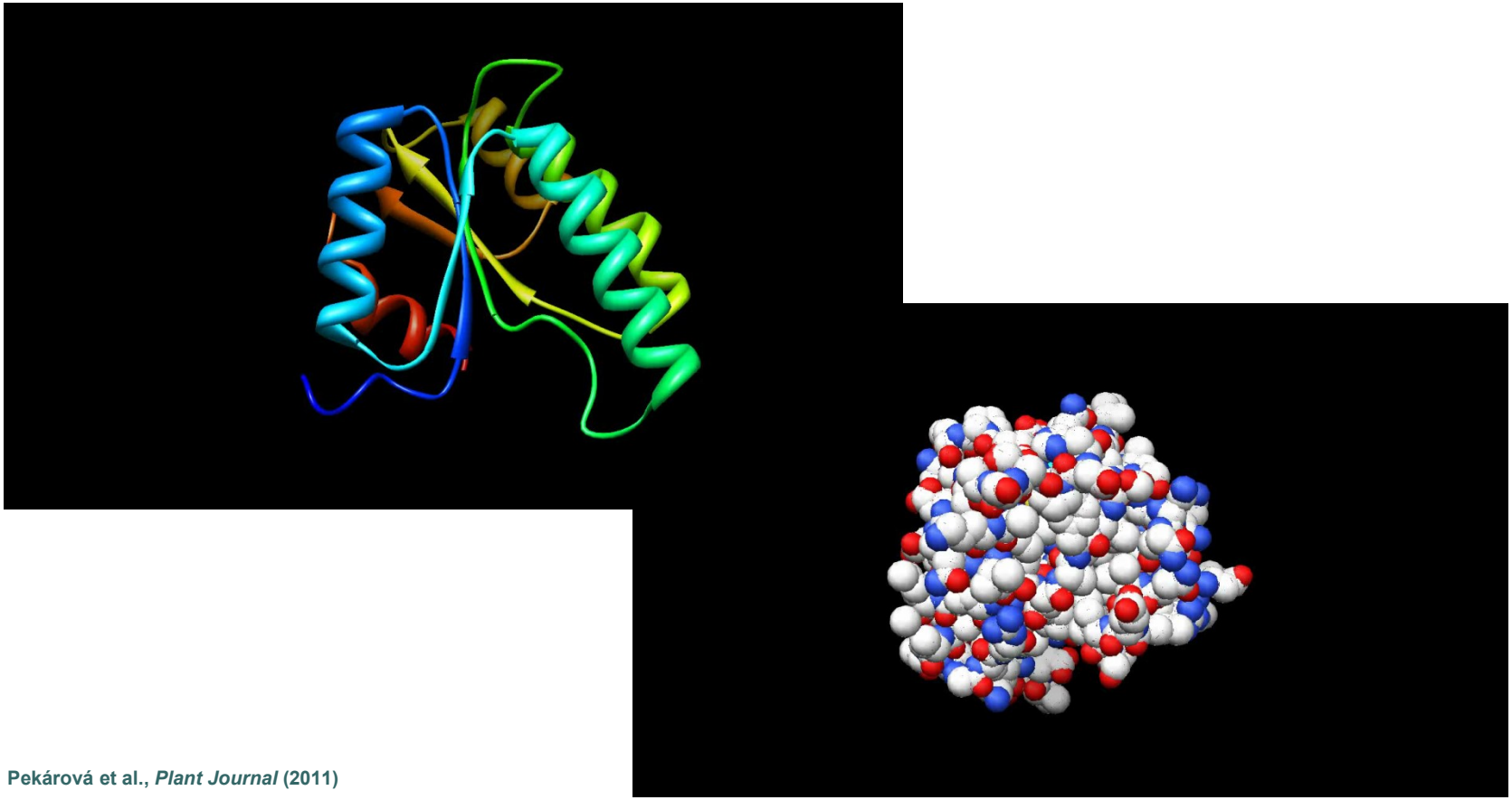




http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=;pid=1735610643493444&bio=1&opt=show&size=500 12/29/2003

Structural databases

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



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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

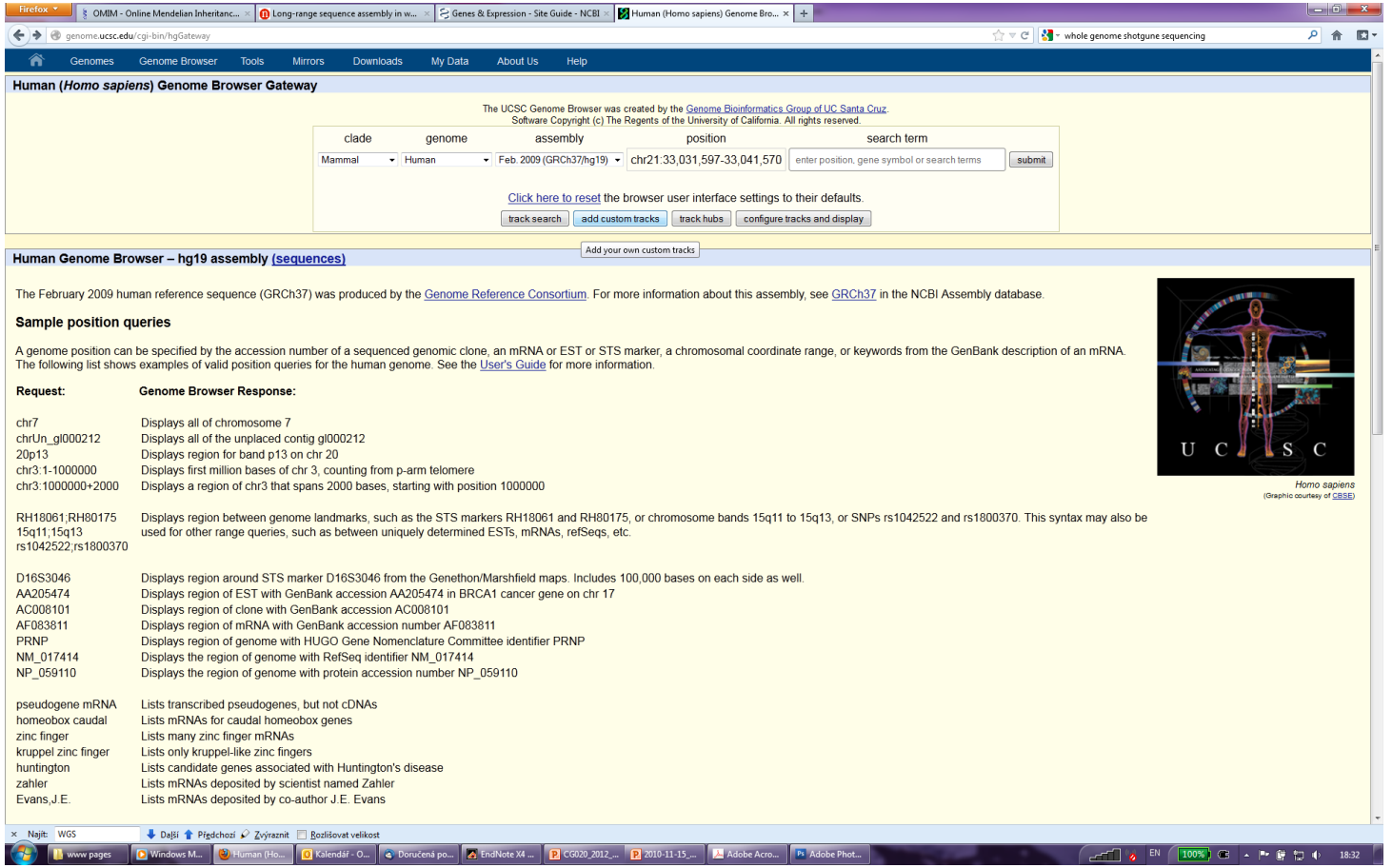
Outline

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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Genome resources

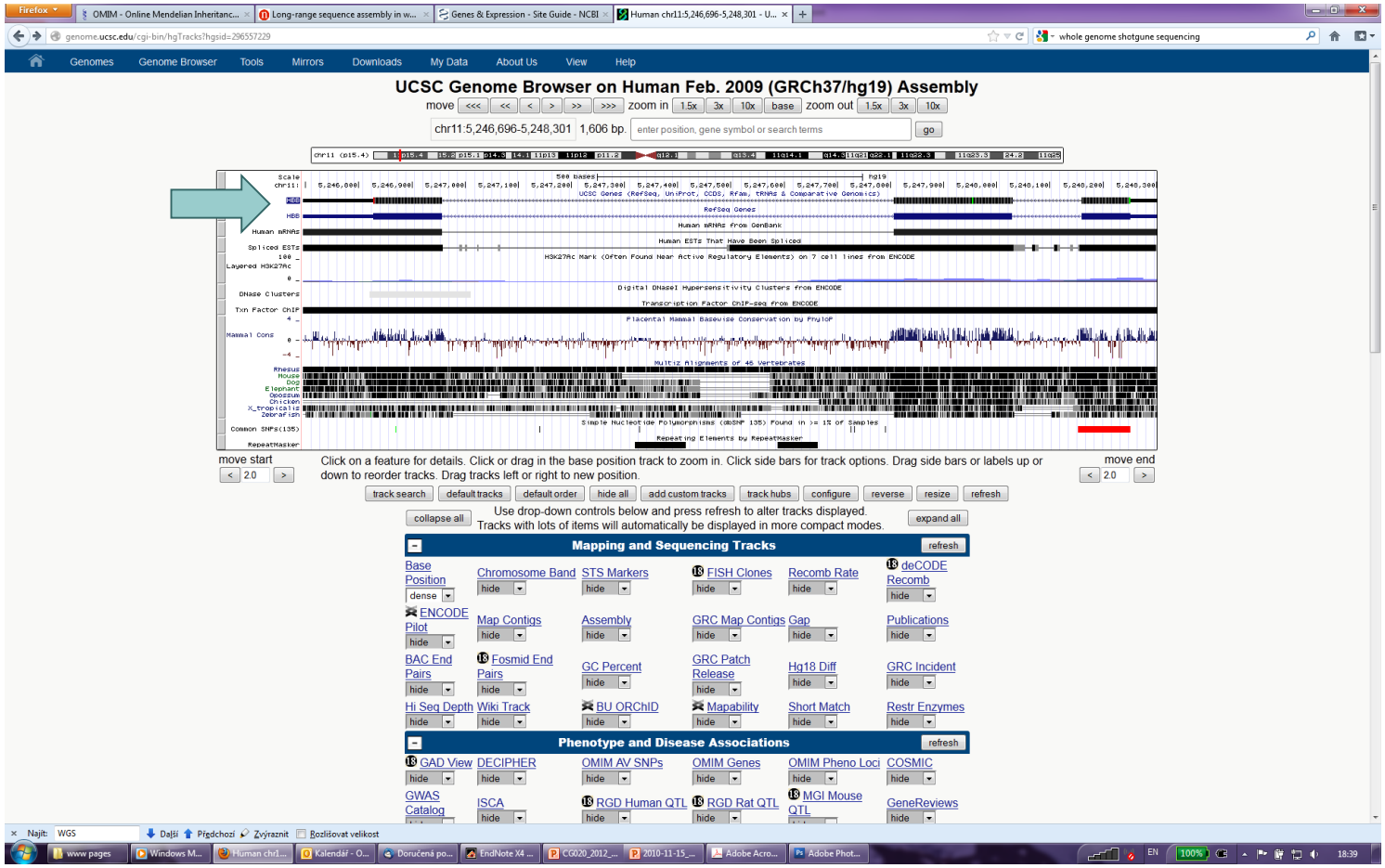
- Human Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

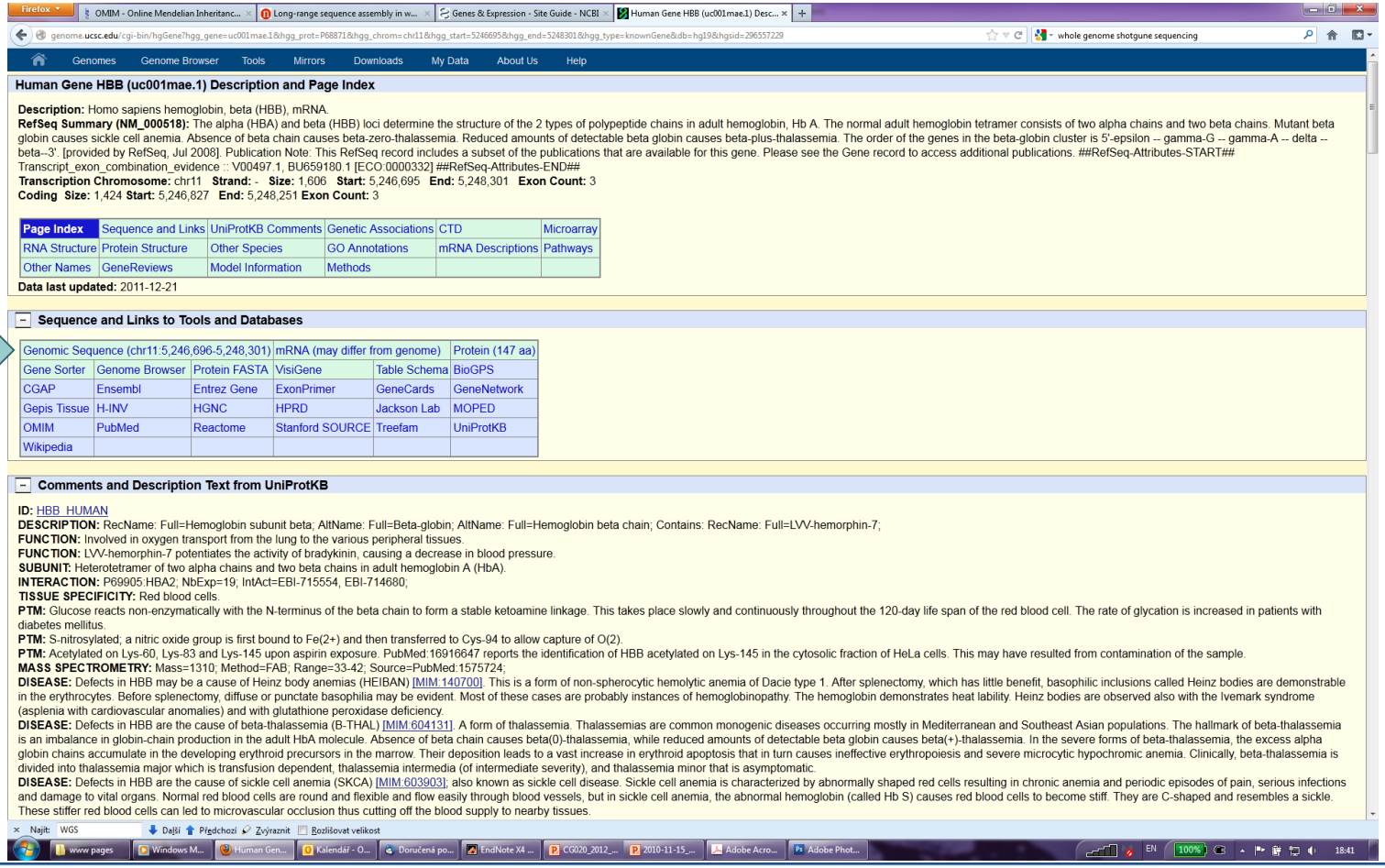
Genome resources

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



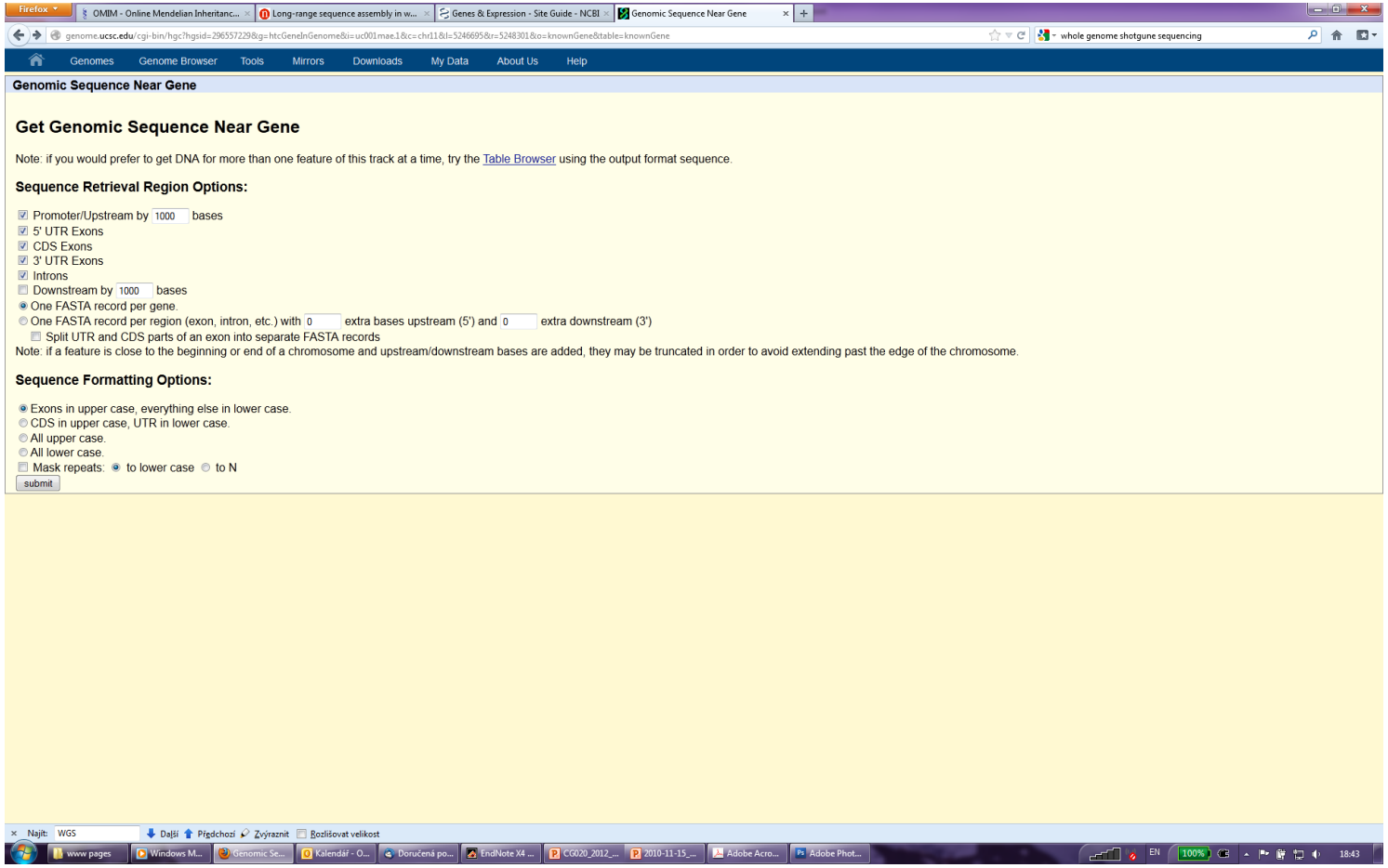
Genome resources

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



Genome resources

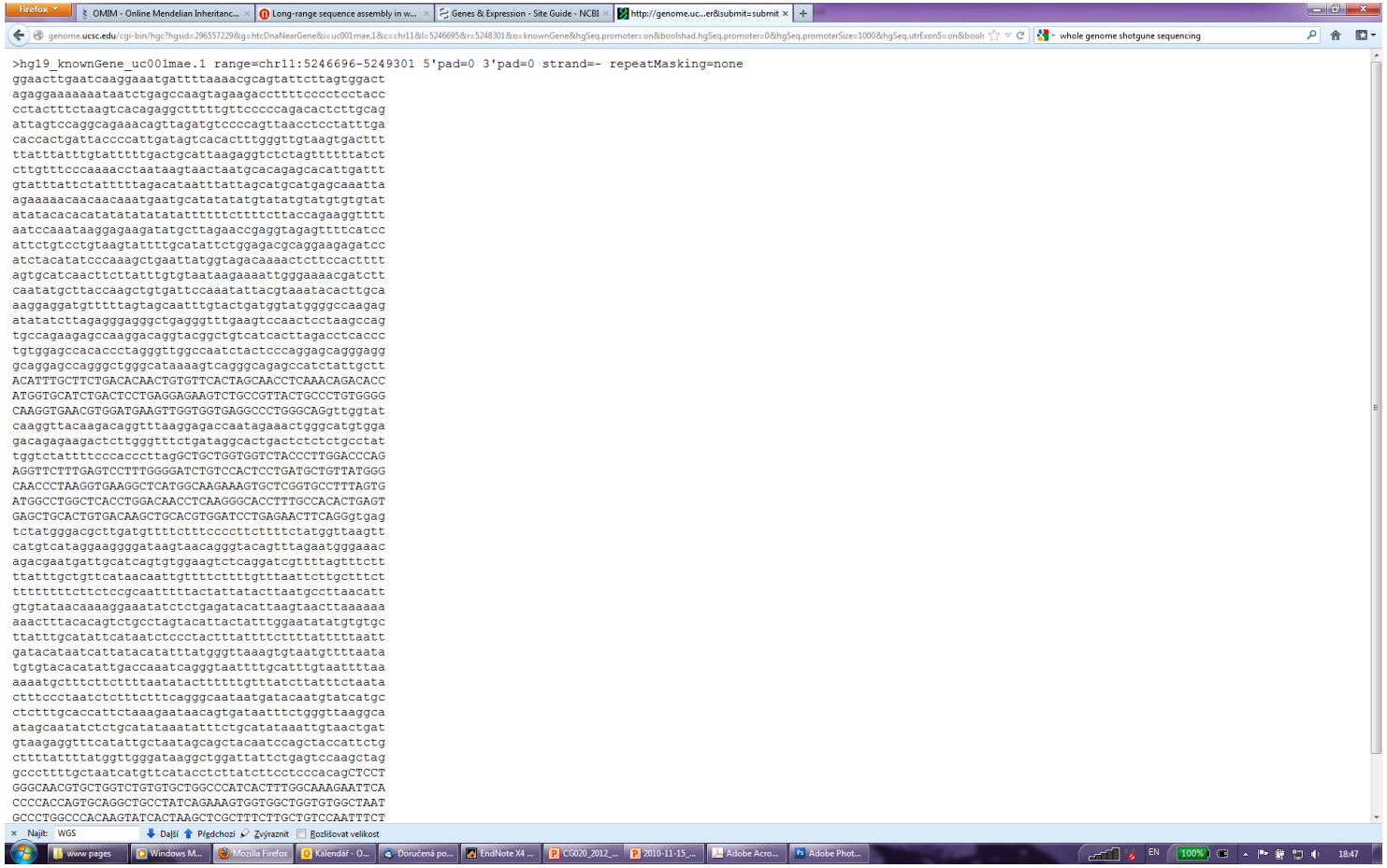
- Human Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Genome resources

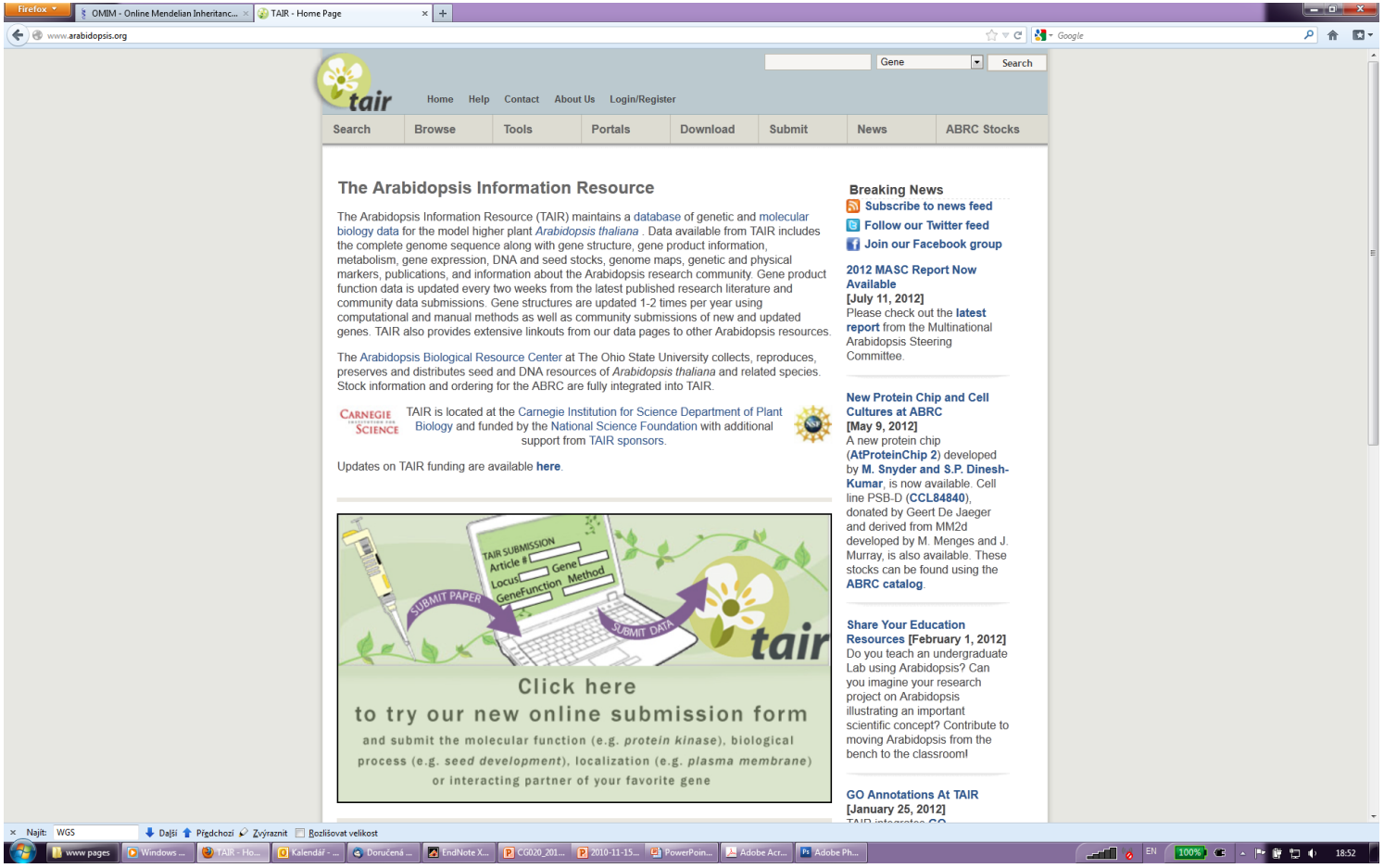
- Human Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

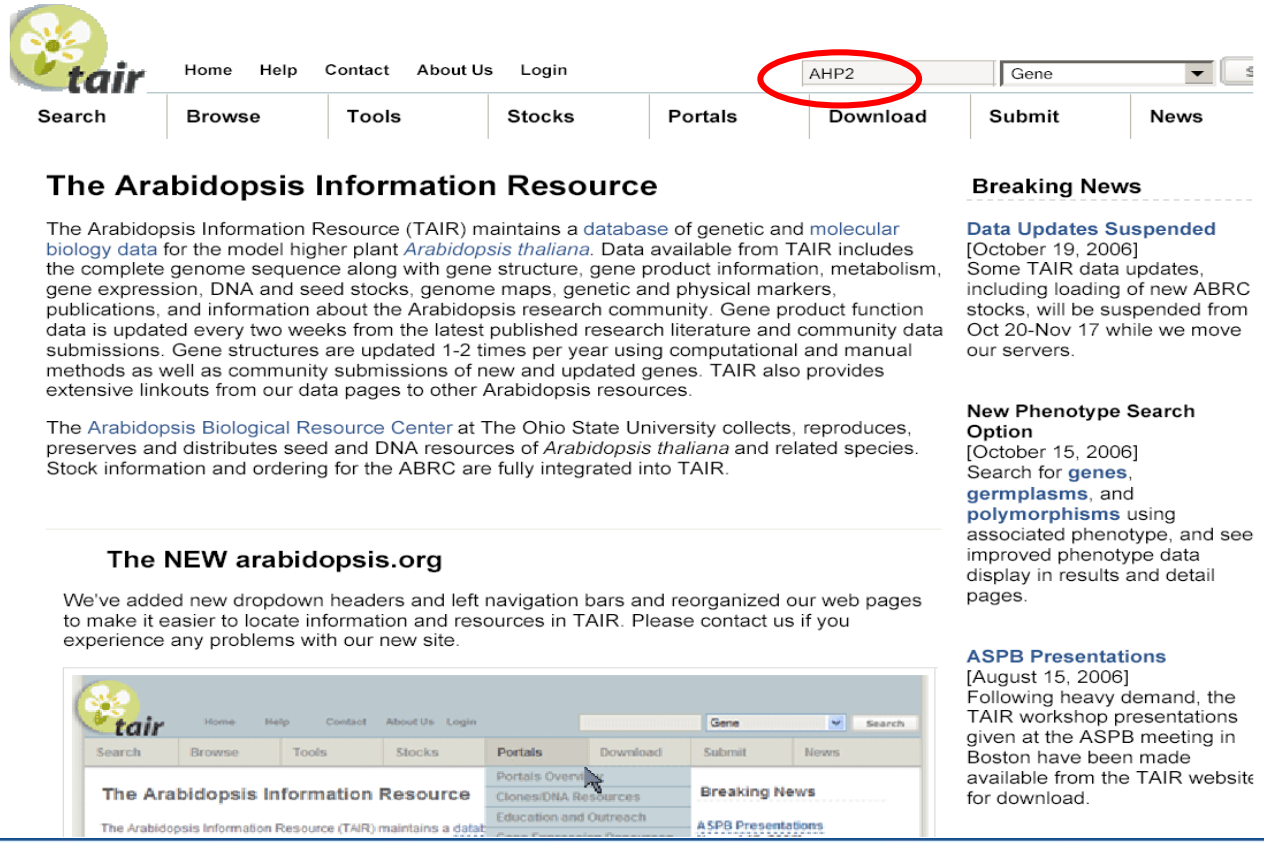
Genome resources

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



Genome resources

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



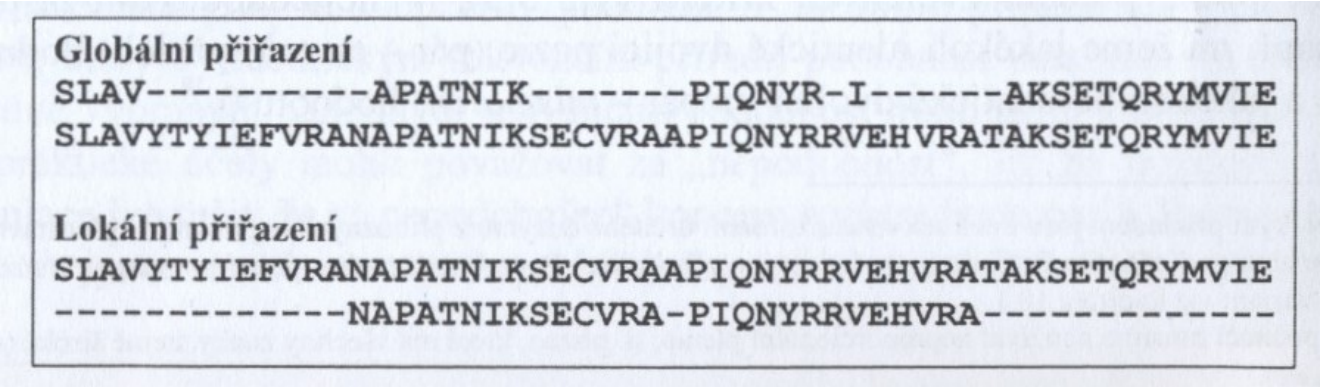
Outline

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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Analytical tools

- Global versus local alignment



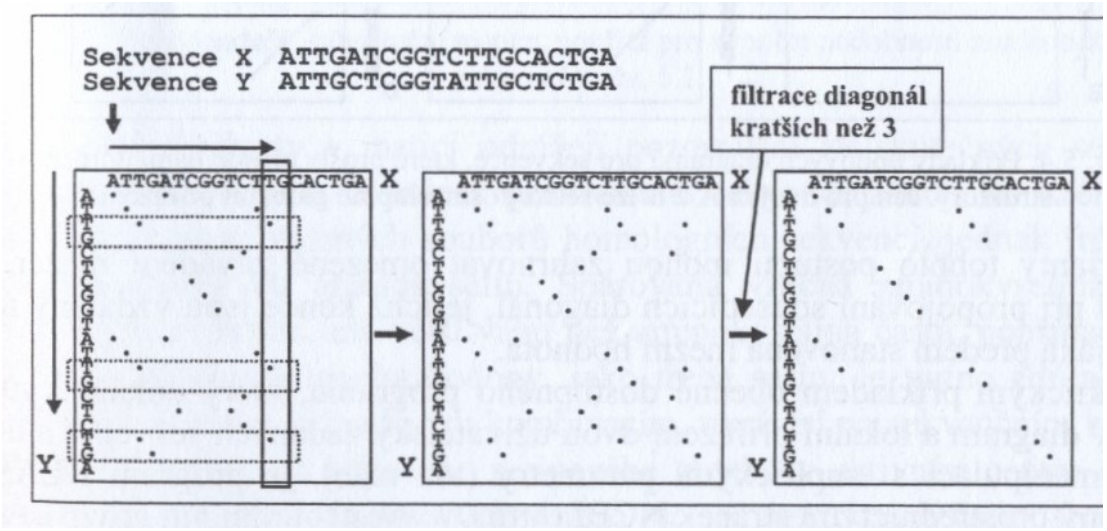
Cvrčková, Úvod do praktické bioinformatiky

  - 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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Analytical tools

- Choosing the right type of alignment using dotplot



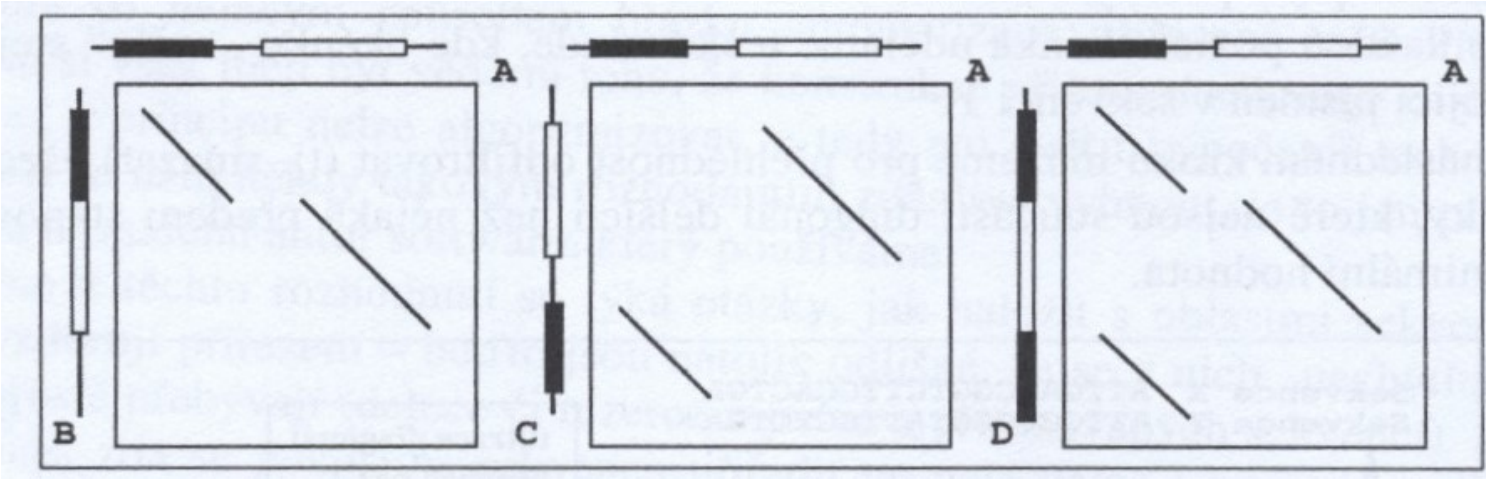
Cvrčková, Úvod do praktické bioinformatiky

  - Plotting the sequences (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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Analytical tools

- Examples of sequence alignment using dotplot



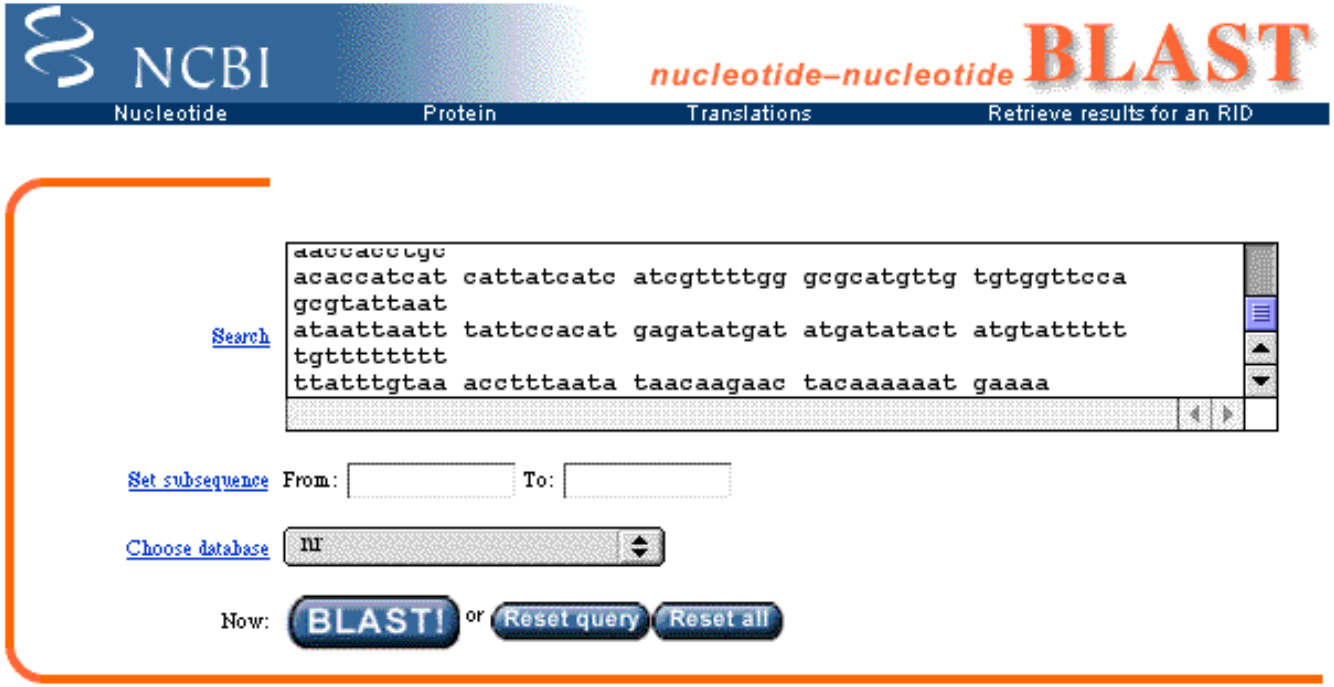
Cvrčková, Úvod do praktické bioinformatiky

- 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)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Analytical tools

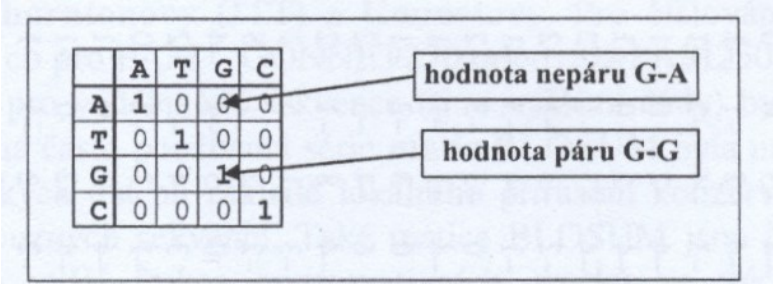
- BLAST http://ncbi.nlm.nih.gov/BLAST/



BLAST

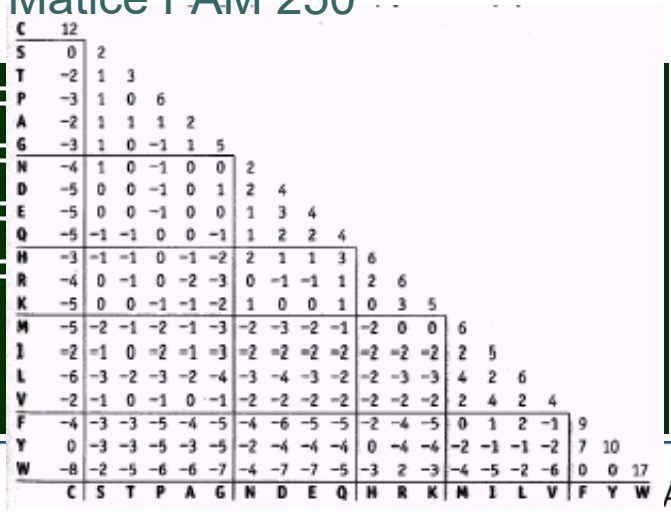
Basic **L**ocal **A**lignment **S**earch **T**ool

- 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



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

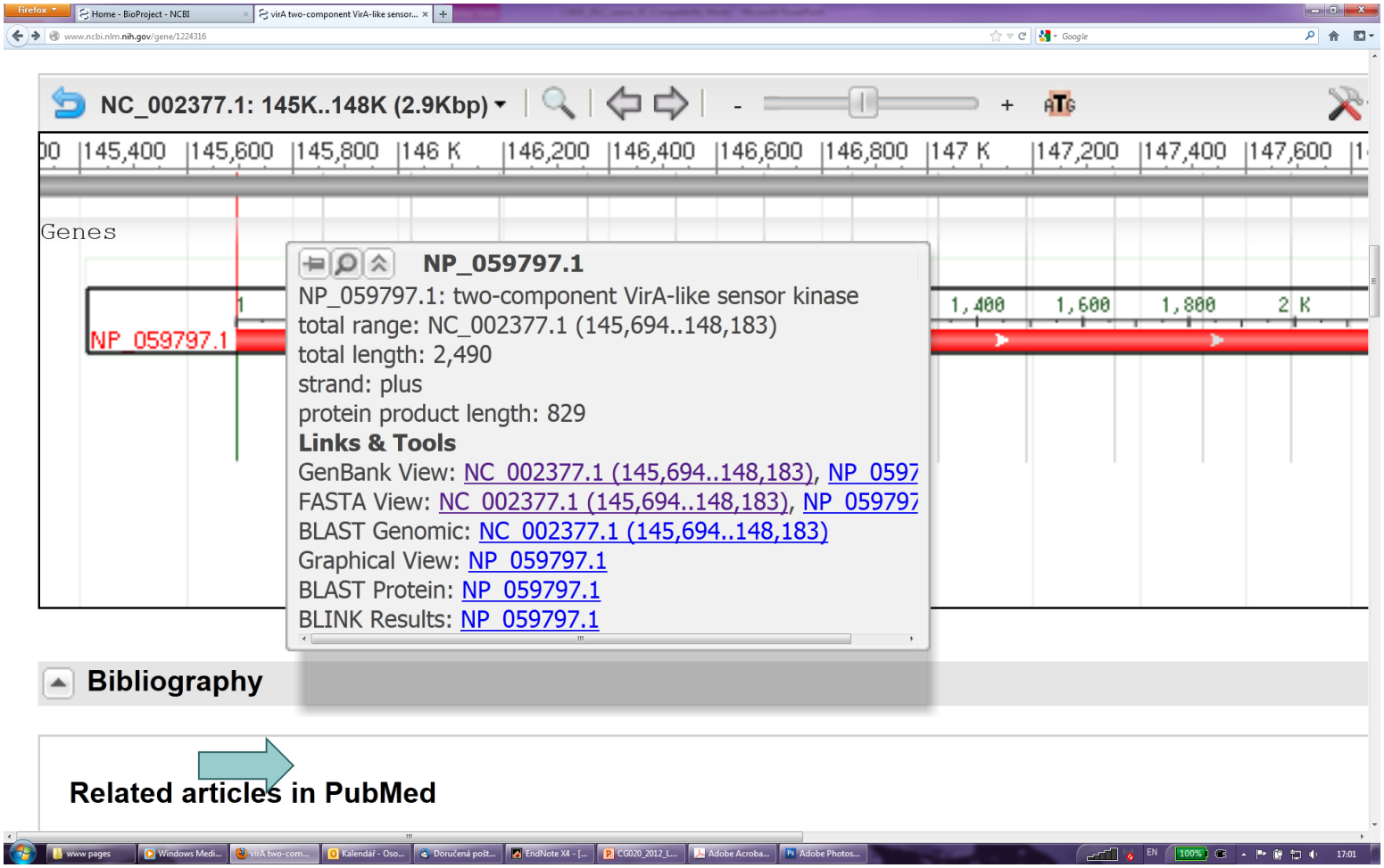
BLAST

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

- „expectancy value“ udává předpokládaný počet sekvencí se stejnou nebo lepší podobnosti při vyhledávání ve stejně velké databázi složené z náhodných sekvencí
- výsledek udává frakci totožných a u proteinů i podobných pozic, příp. počet vložených mezer

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

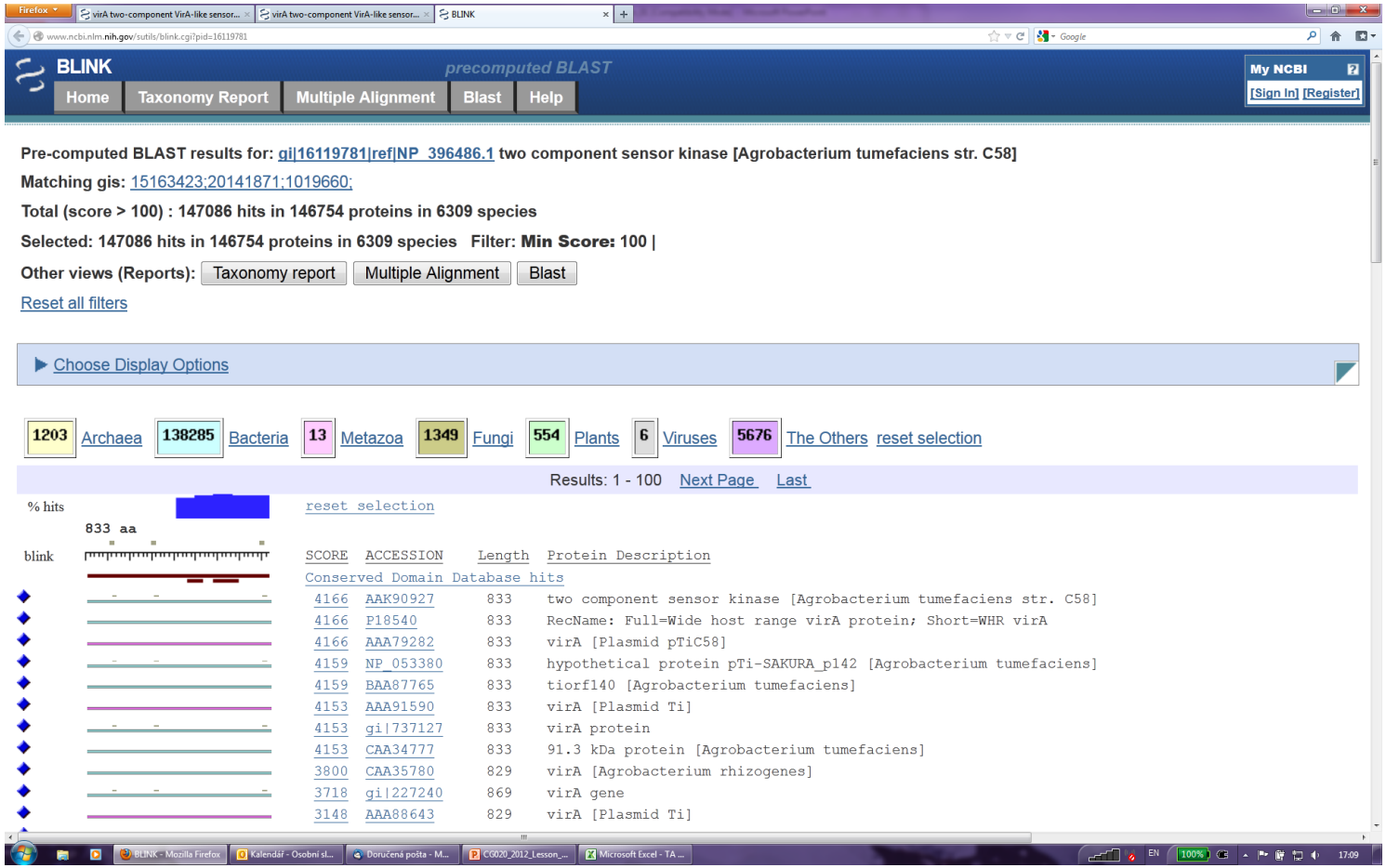
Primary databases



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



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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 six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

BLAST

Specialized versions

- Currently there exists 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.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

BLAST

Specialized versions

- Currently there exists a lot of specialized versions of BLAST

  - **PHI-BLAST** (**P**attern-**H**it **I**nitiatedBlast)
    - 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

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

BLAST

Specialized versions

- Example of search by PHI-BLAST

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

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


Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

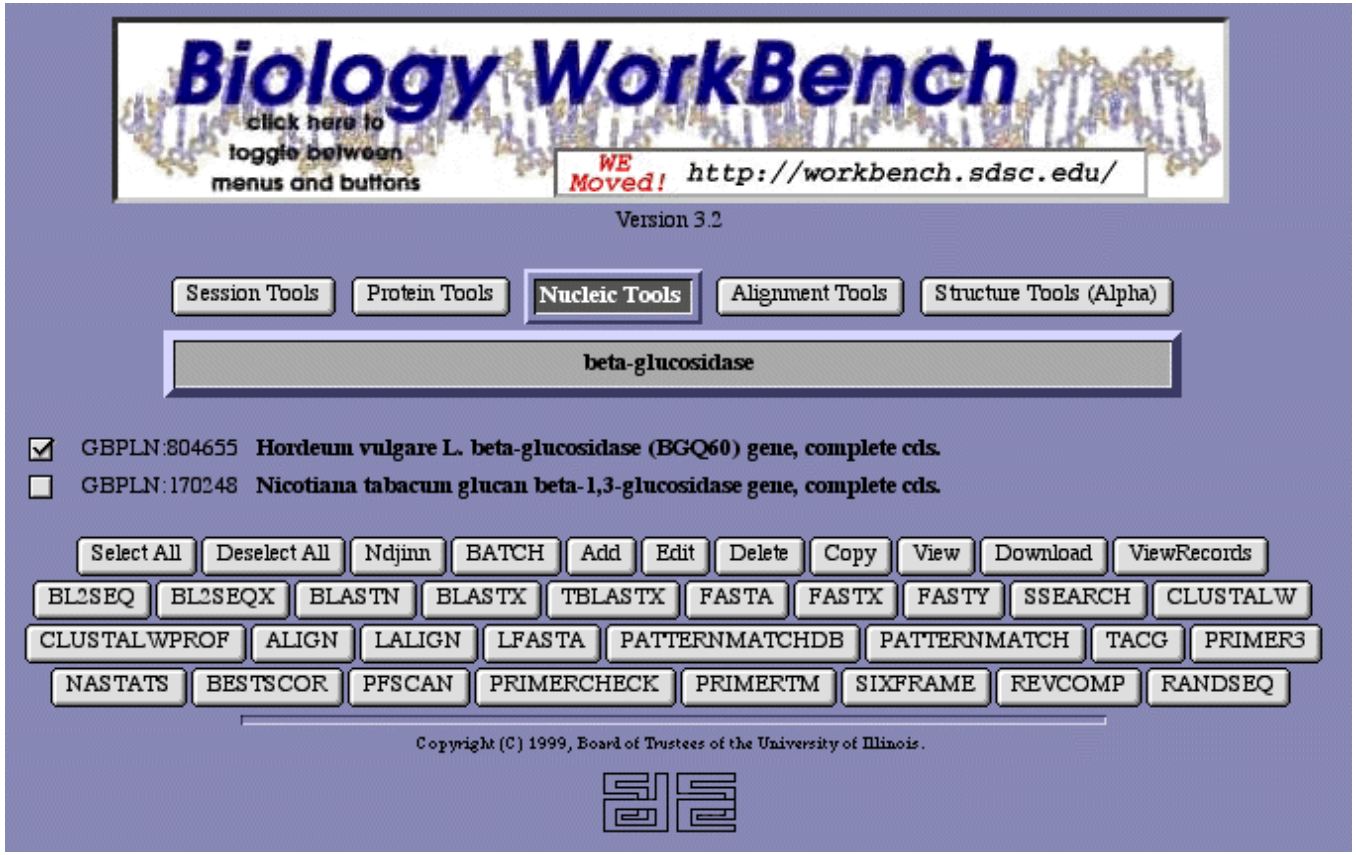
Outline

- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS
- Databases
  - Spectre of „on-line“ resources
  - PRIMARY, SECONDARY and STRUCURAL databases
  - GENOME resources
- Analytical tools
  - Homologies searching
  - Searching of sequence motifs, open reading frames, restriction sites…

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Analytical tools

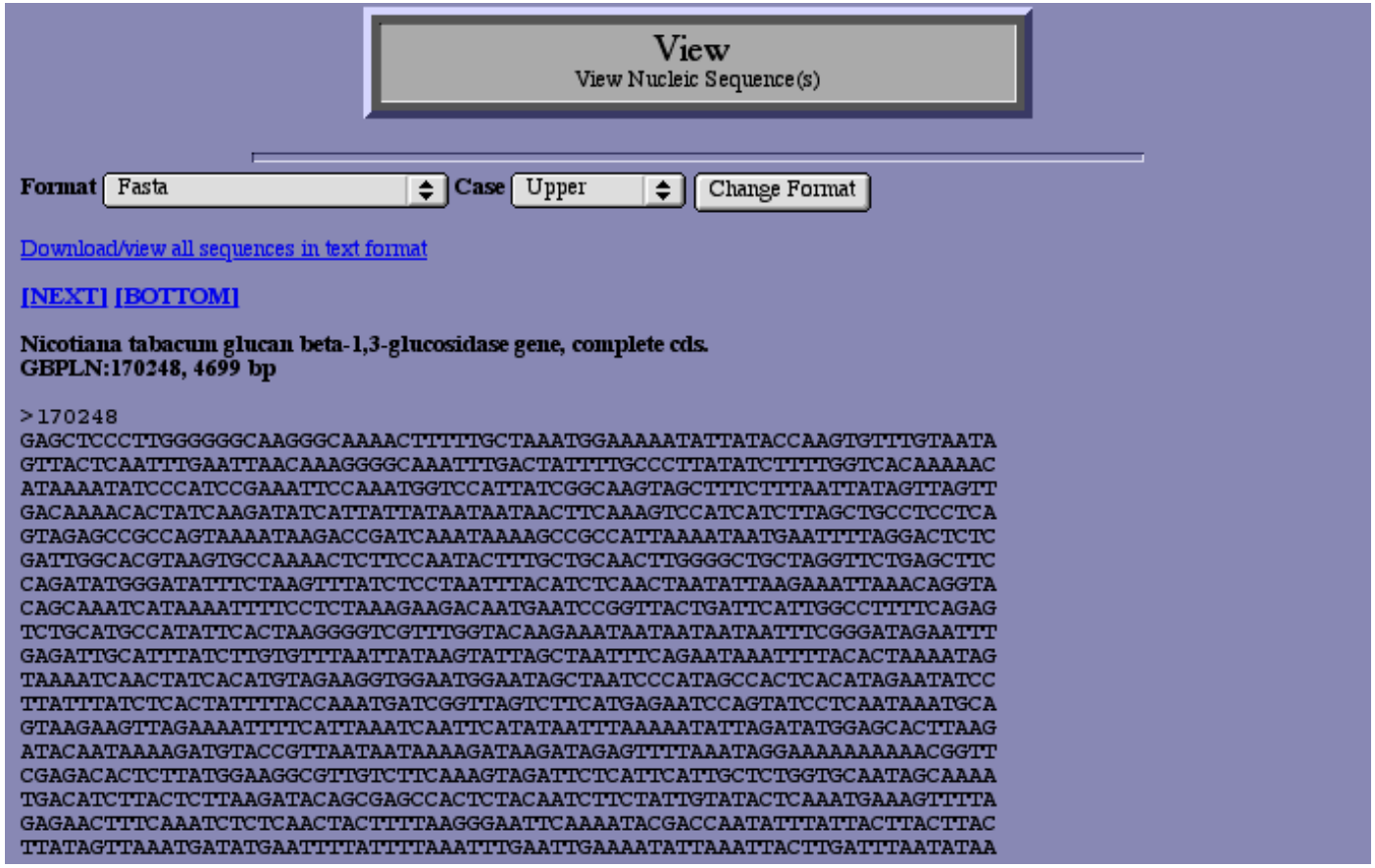
- http://workbench.sdsc.edu/



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Analytical tools

- http://workbench.sdsc.edu/



Analytical tools

- http://workbench.sdsc.edu/

Regex pattern:

ctt.{1,32}ctt

0 sequences were searched

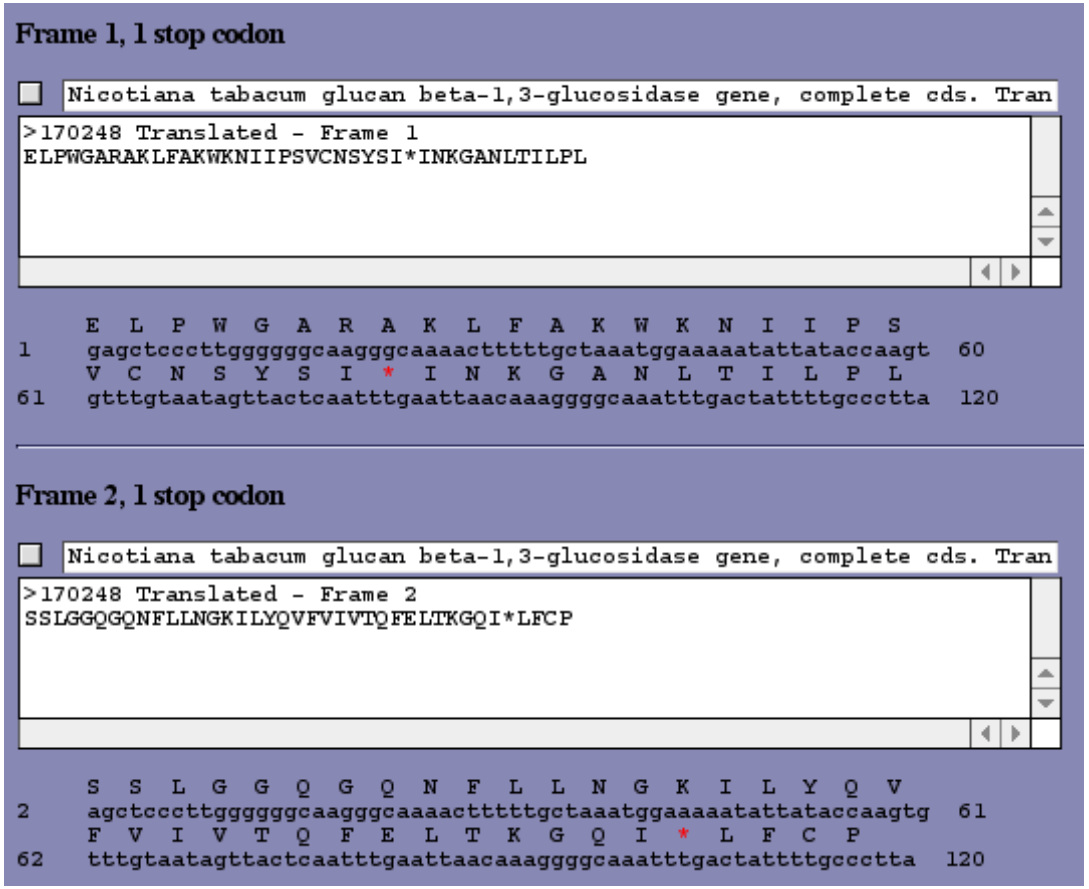
1 match was found

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

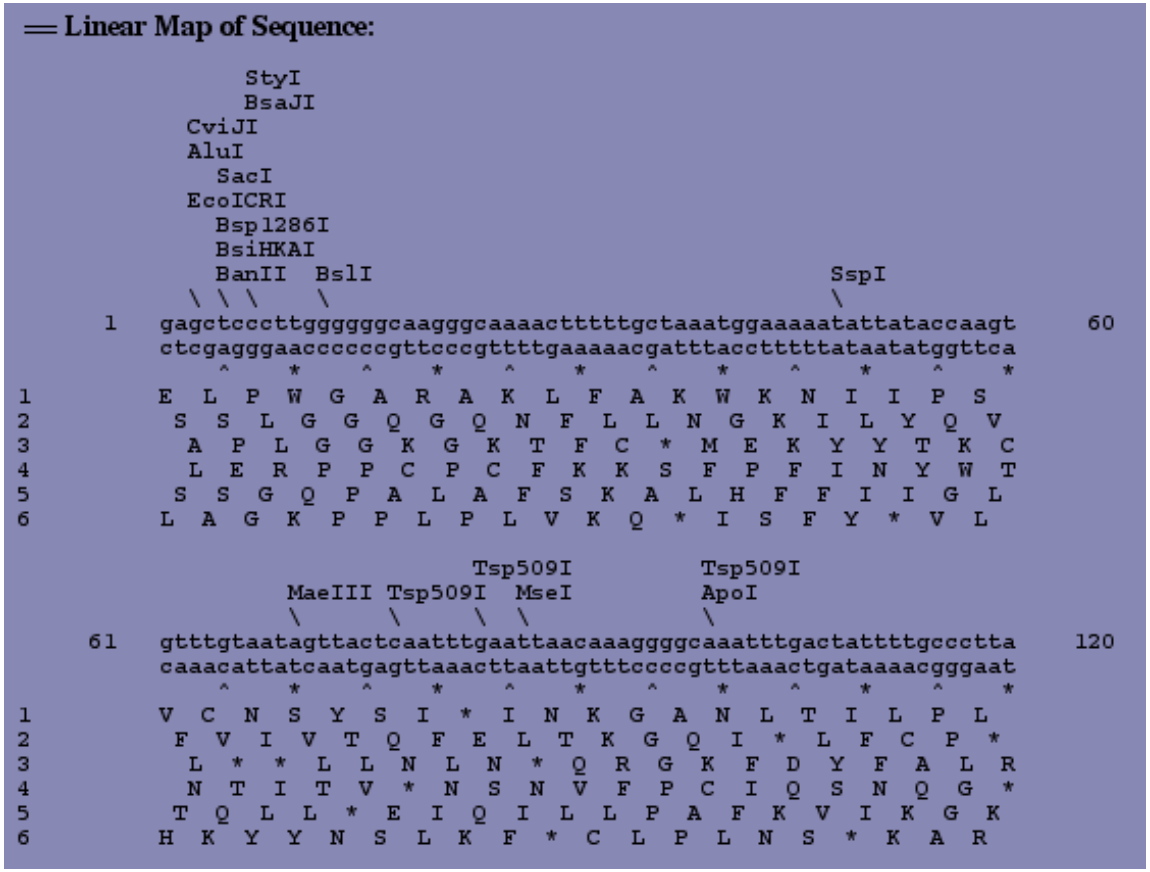
Analytical tools

- http://workbench.sdsc.edu/



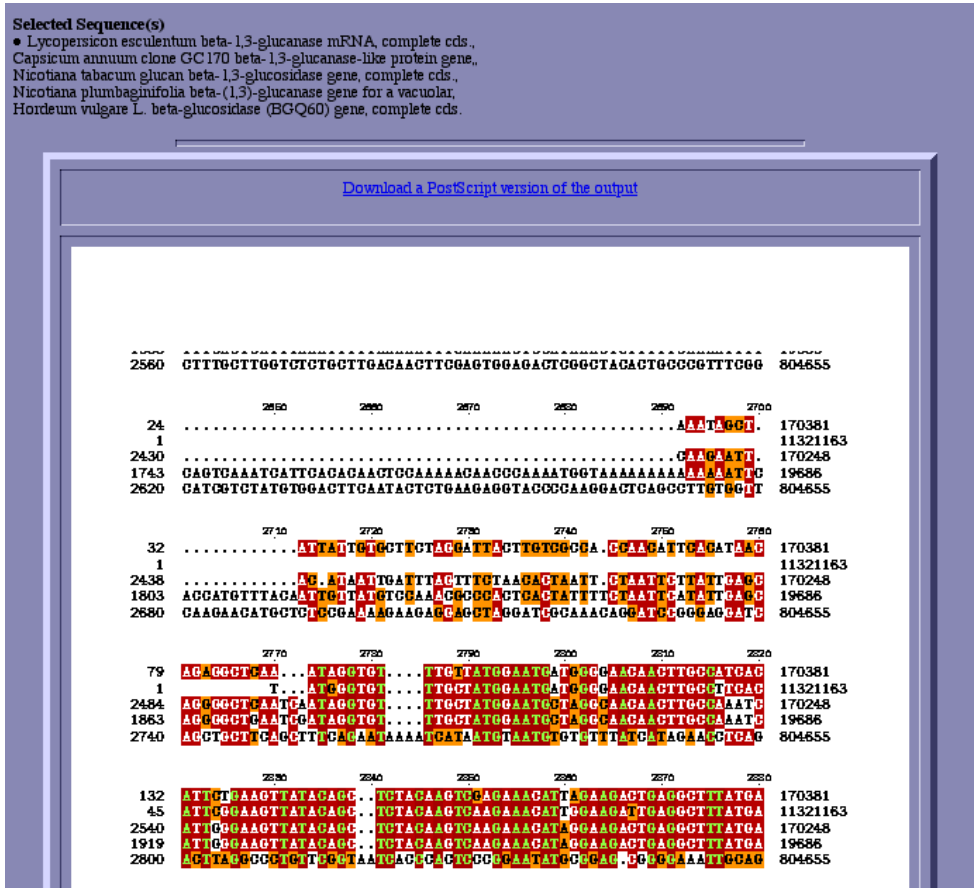
Analytical tools

- http://workbench.sdsc.edu/



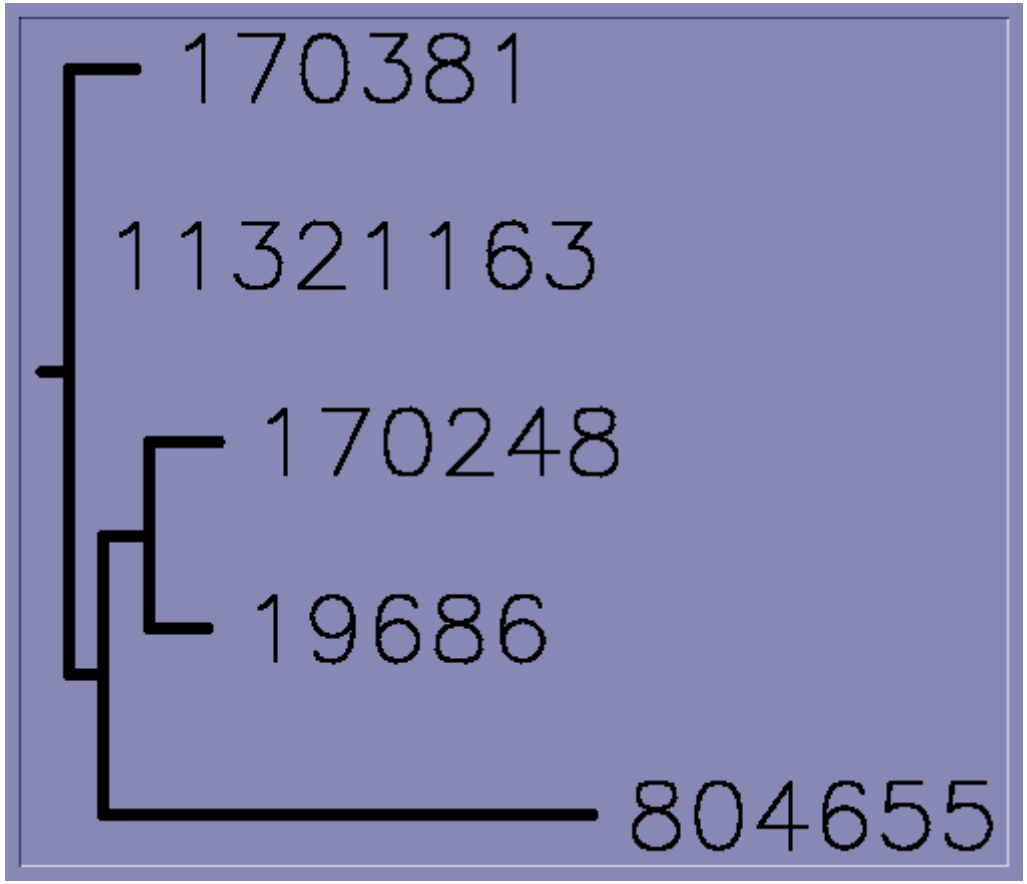
Analytical tools

- http://workbench.sdsc.edu/



Analytical tools

- http://workbench.sdsc.edu/



EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

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

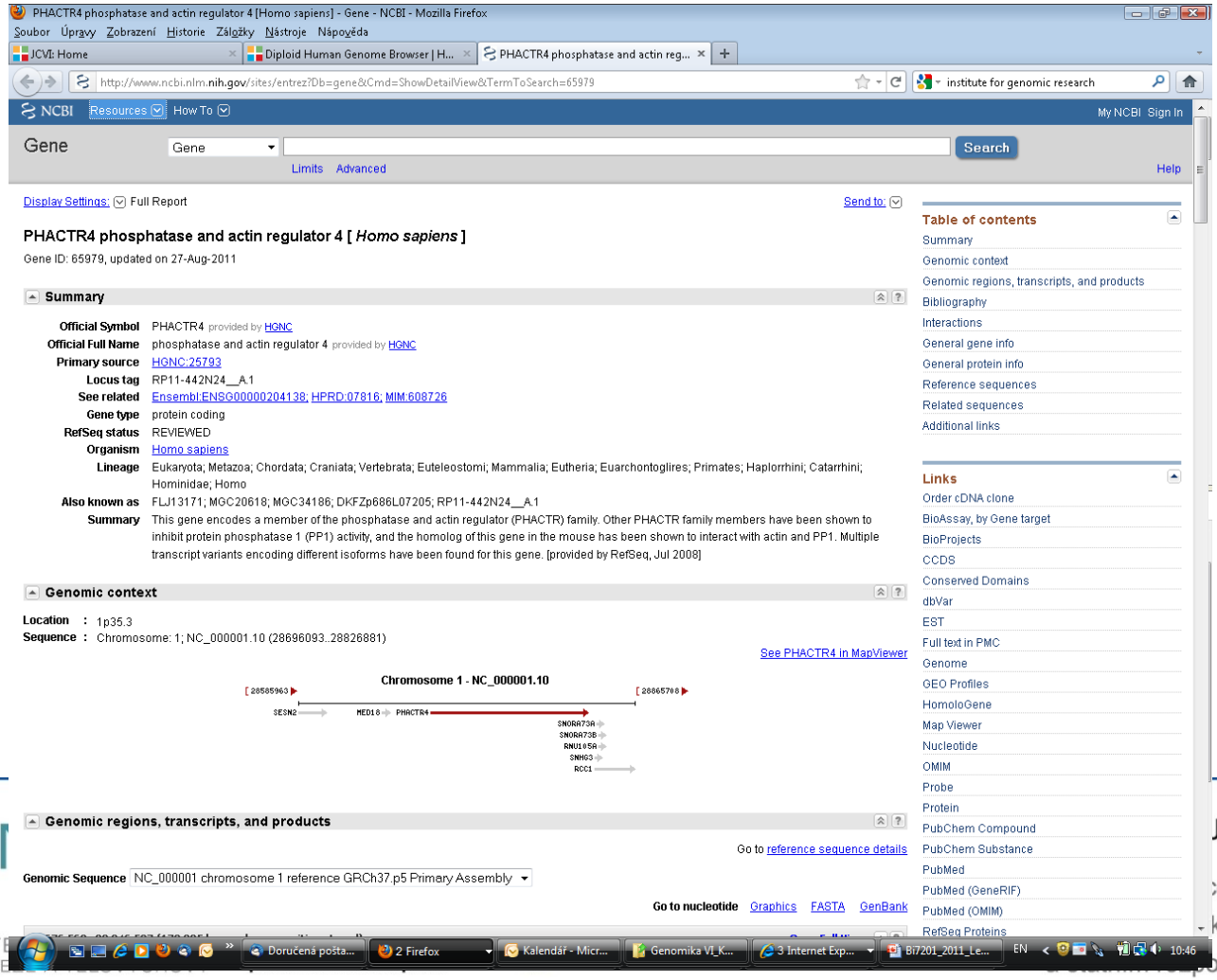
Outline

- Syllabus of this course
- Definition of genomics
- Role of BIOINFORMATICS in FUNCTIONAL GENOMICS
- Databases
  - Spectre of „on-line“ resources
  - PRIMARY, SECONDARY and STRUCURAL databases
  - GENOME resources
- Analytical tools
  - Homologies searching
  - Searching of sequence motifs, open reading frames, restriction sites…
  - Other on-line genome tools

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

Other online genome resources

- TIGR (The Institute for Genomic Research, http://www.tigr.org/software/)
  - Recently part of the J. Craig Venter Institute



Other online genome resources

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