# CG020 Genomika
# Bi7201 Základy genomiky

## Přednáška 1

Úvod do bioinformatiky

Jan Hejátko

**Funkční genomika a proteomika rostlin**,
Mendelovo centrum genomiky a proteomiky rostlin,
Středoevropský technologický institut (CEITEC), Masarykova univerzita, 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

# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst….
  - Další www genomové nástroje

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

# Schéma předmětu

- **Kapitola 01** (CG020 , Bi7201)
  - Úvod do bioionformatiky
- **Kapitola 02** (CG020 , Bi7201)
  - Identifikace genů
- **Kapitola 03** (CG020 , Bi7201)
  - Přístupy reverzní genetiky
- **Kapitola 04** (CG020 , Bi7201)
  - Přístupy genetiky přímé

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

# Schéma předmětu

- **Kapitola 05** (CG020 , Bi7201)
  - Přístupy funkční genomiky
- **Kapitola 06** (CG020 , Bi7201)
  - Protein-protein interakce a jejich analýza
- **Kapitola 07** (CG020)
  - Moderní postupy funkční genomiky
- **Kapitola 08** (CG020)
  - Strukturní genomika

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

# Schéma předmětu

- **Kapitola 09** (CG020)
  - Lokalizace genů a genových produktů v buňce
- **Kapitola 10** (CG020)
  - Genomika a systémová biologie
- **Kapitola 11** (CG020)
  - Praktické aspekty funkční genomiky
- **Kapitola 12** (CG020)
  - Lokalizace genů a genových produktů v buňce

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

# Literatura

- Zdrojová literatura ke kapitole I:
  - **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

# Osnova

- Schéma předmětu
- Definice

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

# GENOMIKA-co to je?

- V širším pojetí-zkoumá **STRUKTURU** a **FUNKCI** genomů
  - Předpokladem je znalost genomu (sekvencí)- práce s databázemi
- V užším pojetí zkoumá FUNKCI jednotlivých genů - **FUNKČNÍ GENOMIKA**
  - používá zejména přístupy REVERZNÍ GENETIKY

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

# **GENOMIKA**-co to je?

## role BIOINFORMATIKY ve FUNKČNÍ GENOMICE

Přístupy „klasické“ genetiky

„Reverzně genetický“ přístup

5‘TTATATATATATATTAAAAAATAAAATAAAA
GAACAAAAAAGAAAATAAAATA….3‘

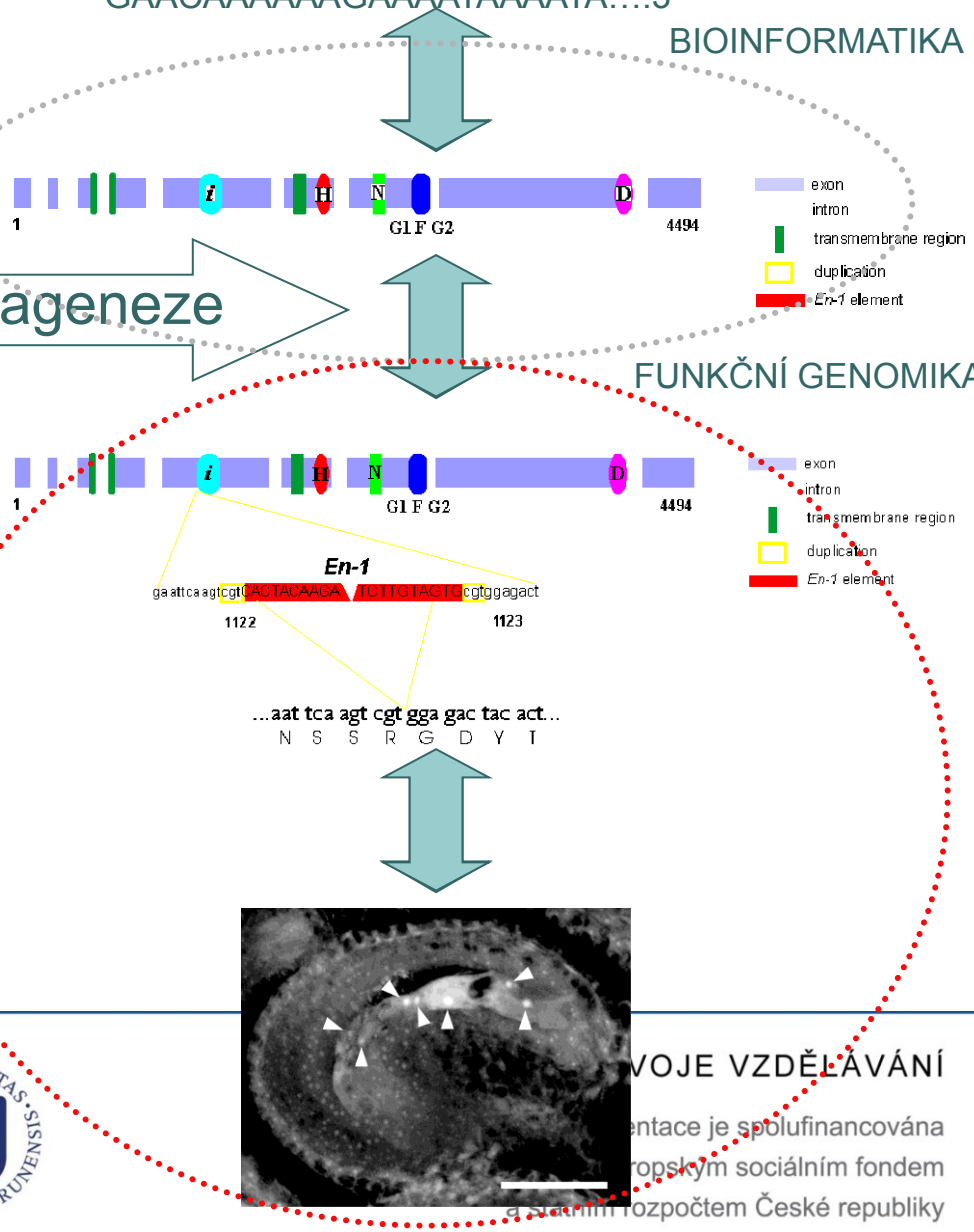
3 : 1

?

BIOINFORMATIKA

inzerční mutageneze

FUNKČNÍ GENOMIKA

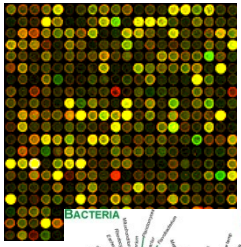


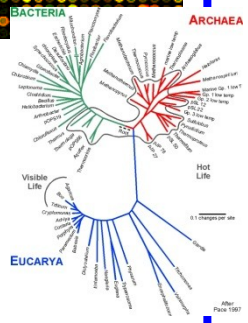
# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY

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

# Bioinformatika






- **Definice bioinformatiky** (podle NIH vědeckého a technologického konsorcia pro biomedicínské informace)

**Výzkum, vývoj nebo aplikace výpočetních nástrojů a přístupů za účelem zvyšování rozvoje využití biologických, lékařských, dat o chování nebo zdraví, včetně těch, které umožňují taková data získávat, ukládat, organizovat, archivovat, analyzovat nebo vizualizovat.**

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

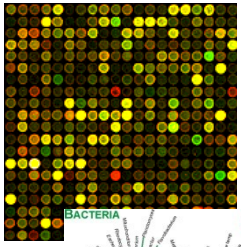
# What is bioinformatics?

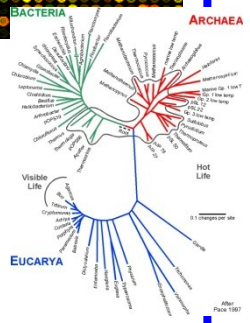
- Interface of biology and computers
- Analysis of proteins, genes and genomes using computer algorithms and computer databases
- Genomics is the analysis of genomes. The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.

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

EVROPSKÁ UNIE
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

# Bioinformatika






- **Bioinformatika** ve **funkční genomice**
  - **Zpracování a analýza sekvenačních dat**
    - Identifikace referenčních sekvencí
    - Identifikace genů
    - Identifikace homologů, ortologů a paralogů
    - Korelační analýzy mezi sekvencemi a fenotypy (včetně člověka)
  - **Zpracování a analýza transkripčních dat**
    - Transkripční profilování pomocí DNA čipů nebo next-gen sekvenování
  - **Vyhodnocování experimentálních dat a predikce nových regulací v přístupech systémové biologie**
    - Matematické modelování genových regulačních sítí

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

# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů

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

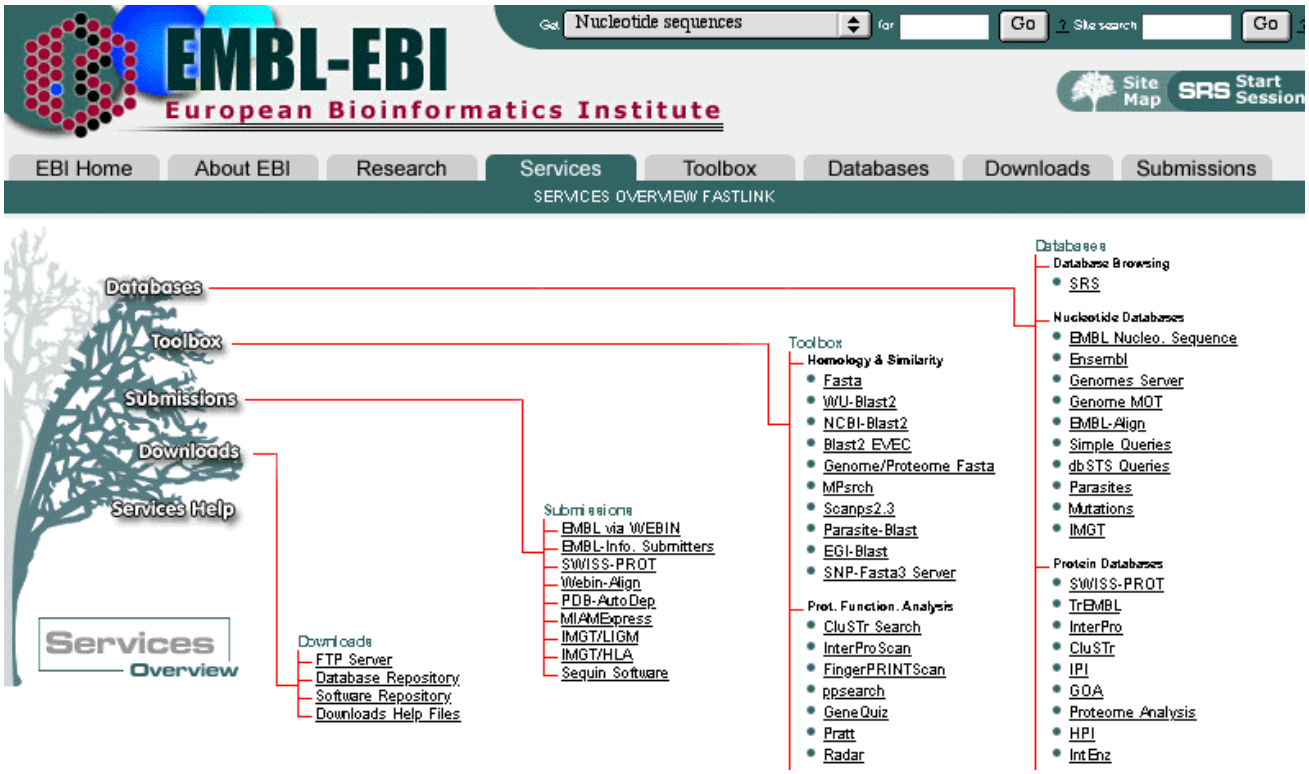
# Spektrum on-line zdrojů

| | | |
|---|---|---|
| ***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/ |

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

# Spektrum on-line zdrojů

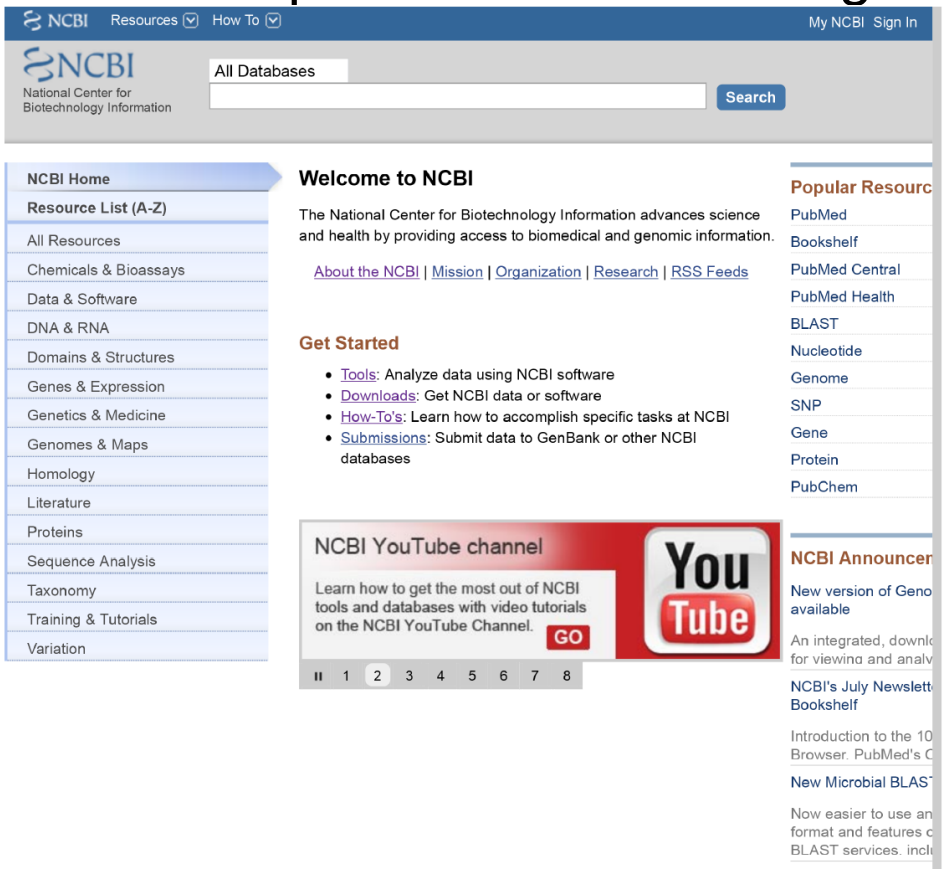
- EBI http://www.ebi.ac.uk/services

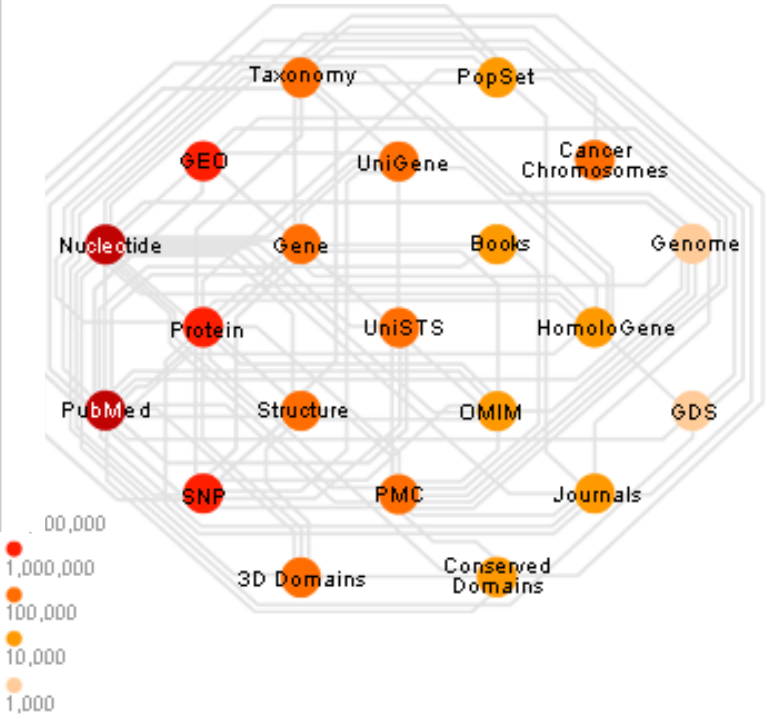



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

# Spektrum on-line zdrojů

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




EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
a
Evropským sociálním fondem
a státním rozpočtem České republiky

# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze

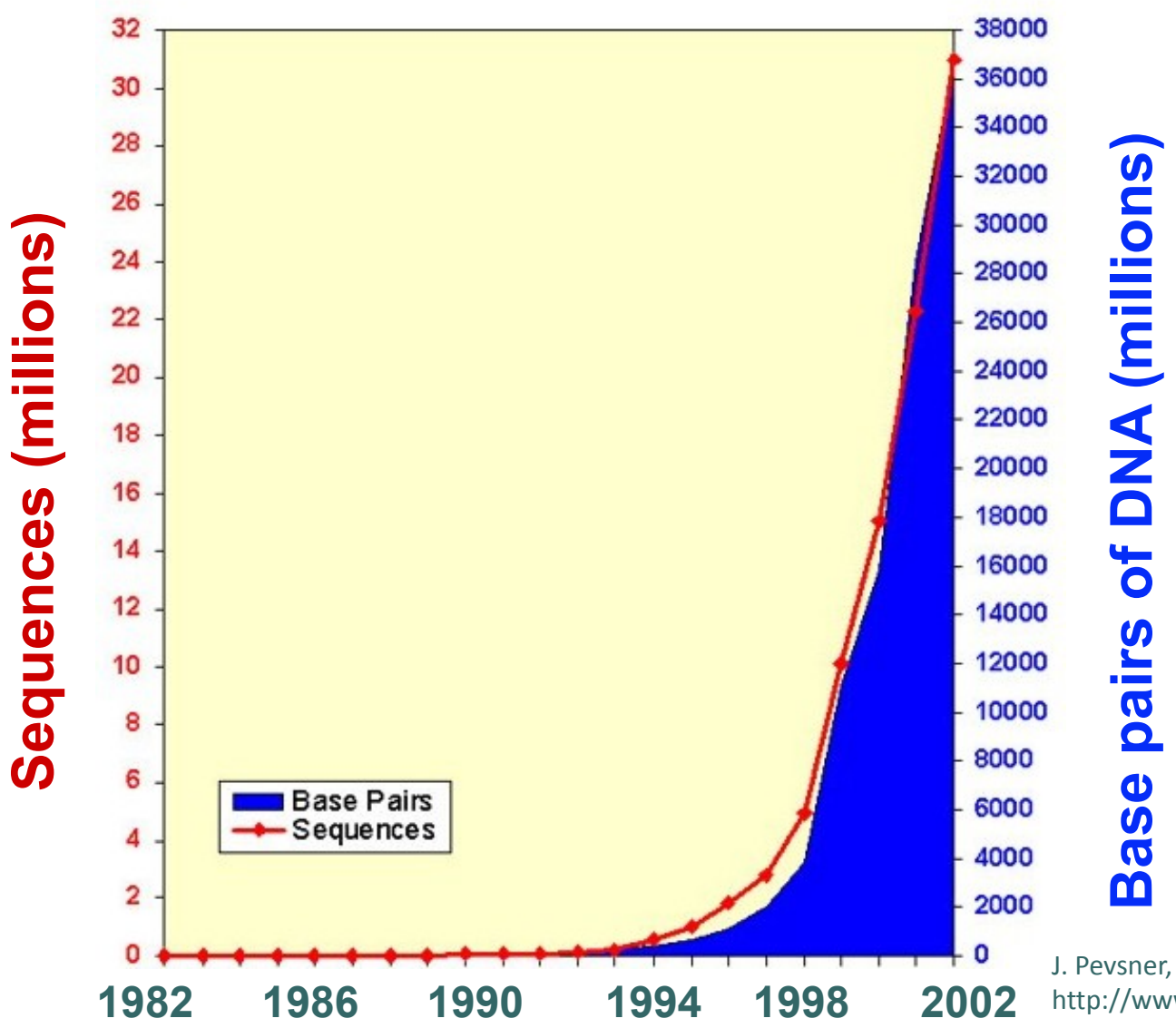
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

# Primární databáze

- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - Sekvence v databázích tzv. „Velké trojky“:
    - EMBL, http://www.ebi.ac.uk/embl/
    - GenBank,http://www.ncbi.nih.gov/Genbank/GenbankSearch.html
    - DDBJ, http://www.ddbj.nig.ac.jp
  - denně vzájemná výměna a zálohování dat
  - velká datová náročnost (kapacita i software)
  - září 2003 27,2 x $10^6$ záznamů o zhruba 33 x $10^9$ bp
  - srpen 2005 100 x $10^9$ bp ze 165.000 organizmů

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

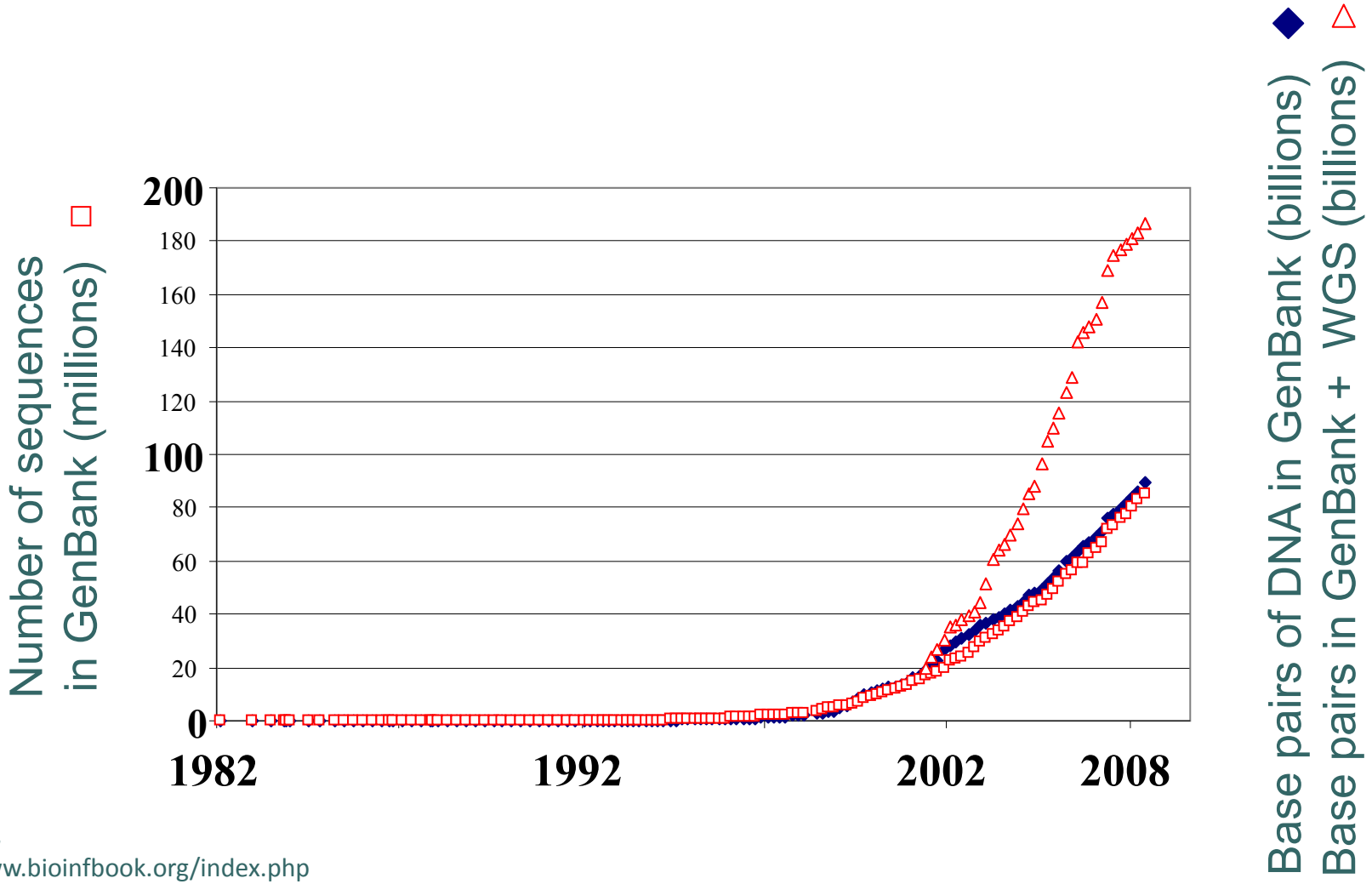
# Growth of GenBank




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


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

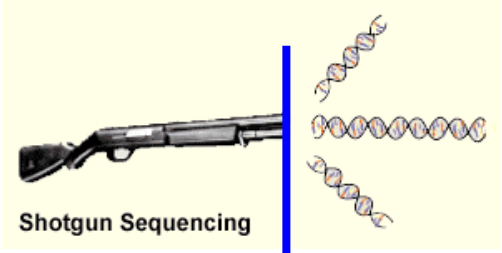
# Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases




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

# WGS



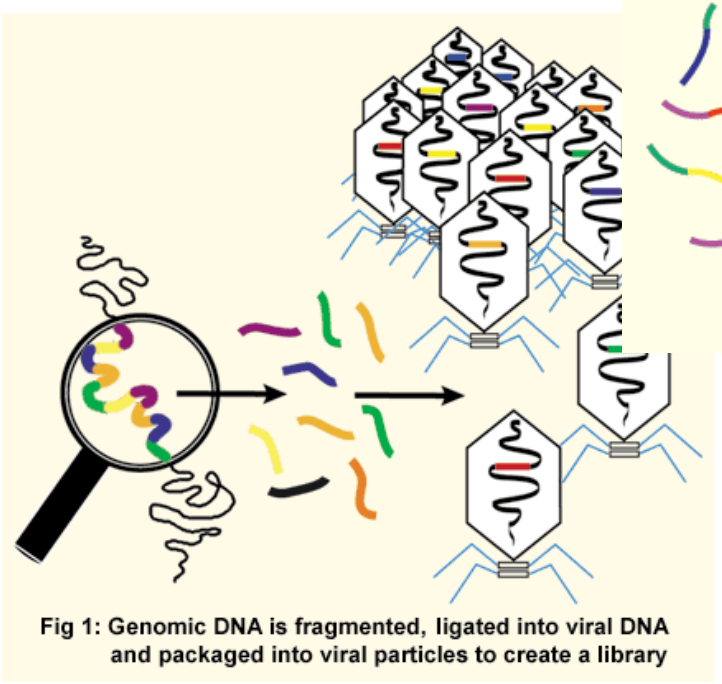



Fig 1: Genomic DNA is fragmented, ligated into viral DNA and packaged into viral particles to create a library

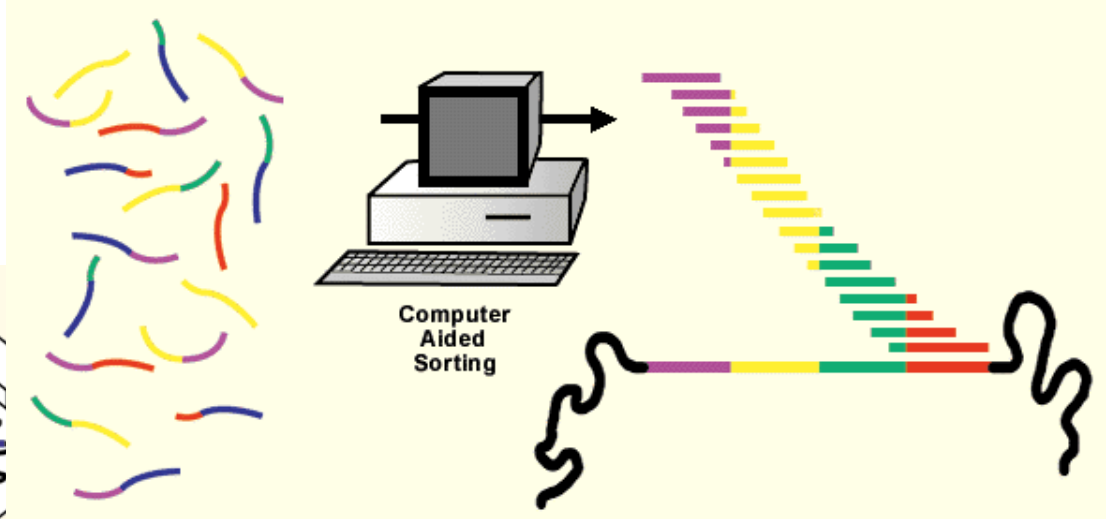
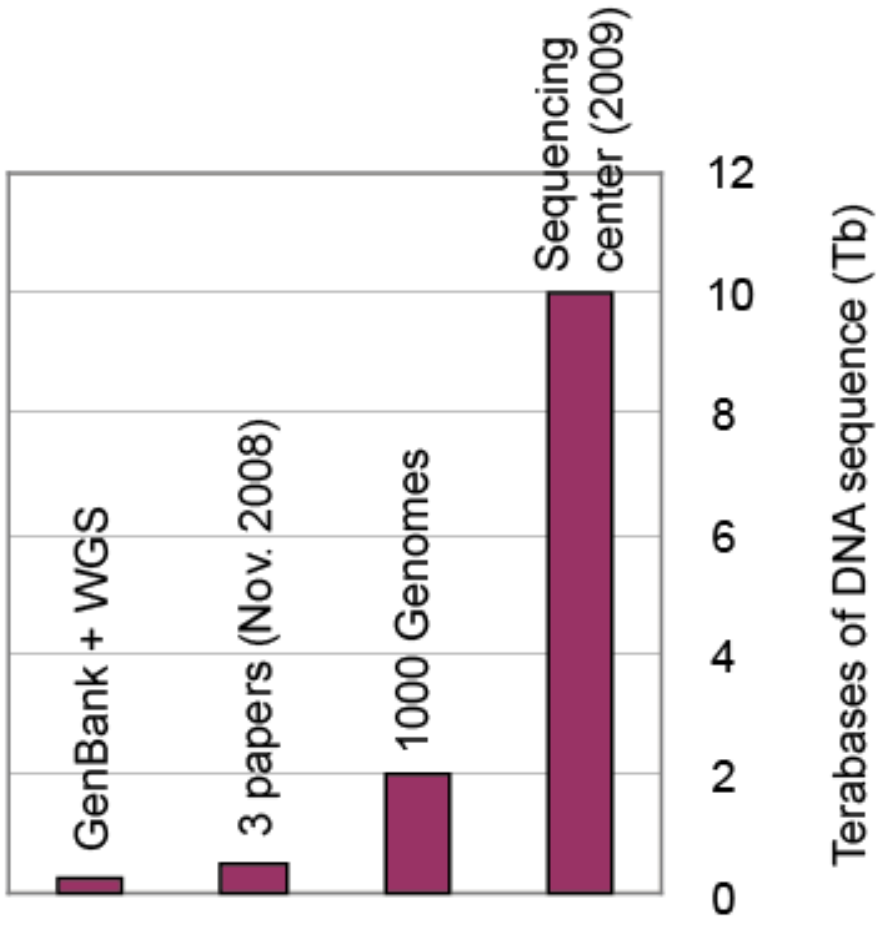



Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

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

# 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

# Primární databáze

- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - Proteinové sekvence:
    - PIR, http://pir.georgetown.edu/
    - MIPS, http://www.mips.biochem.mpg.de
    - SWISS-PROT, http://www.expasy.org/sprot/

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

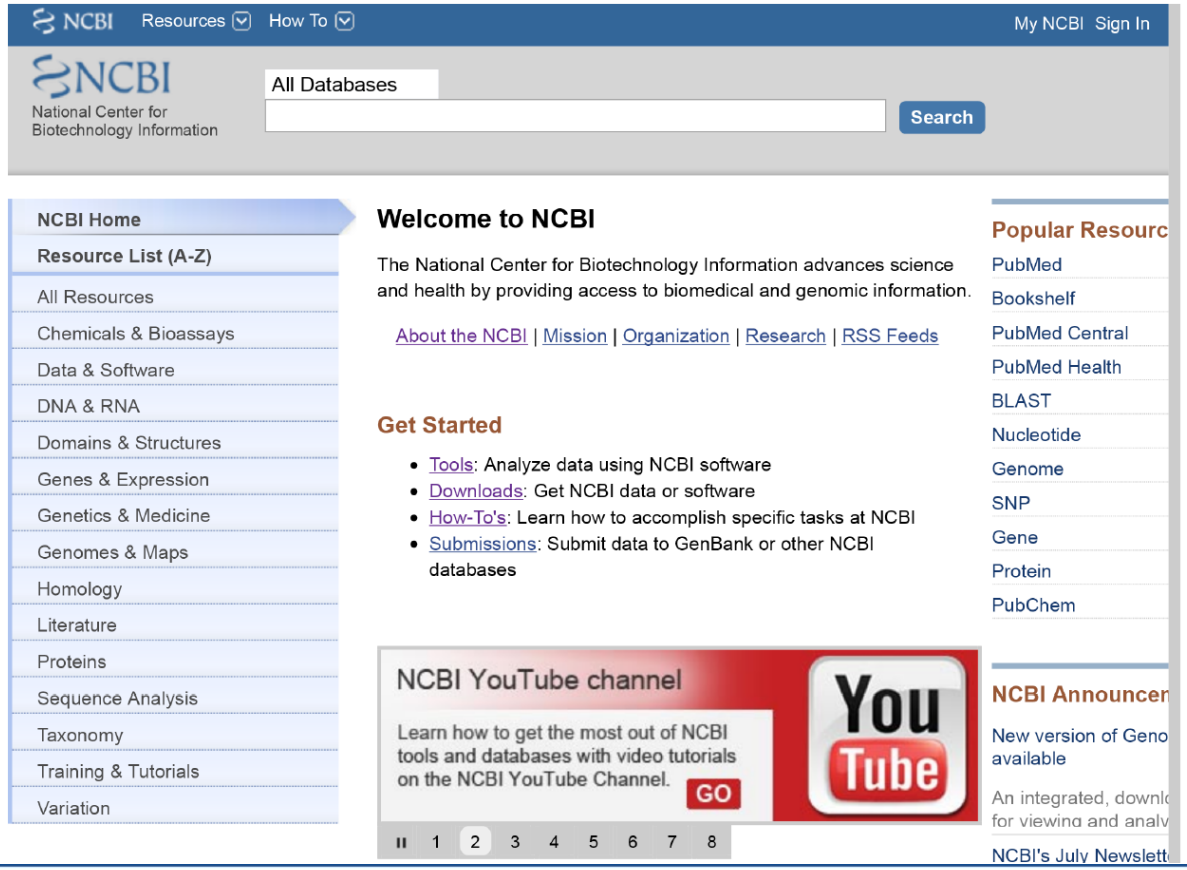
# Primární databáze

- Typy sekvencí v primárních databázích
  - standardní nukleotidové sekvence získané kvalitním sekvencováním
  - **EST**s (**E**xpressed **S**equence **T**ags)
  - **HGTS** (**H**igh **T**hroughput **G**enome **S**equencing)
    - neanotované „surové“ výsledky sekvenačních projektů
  - referenční sekvence anotovaných genomů
  - **TPA**s (**T**hird **P**arty **A**nnotation)
    - sekvence anotované jinými než původními autory

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

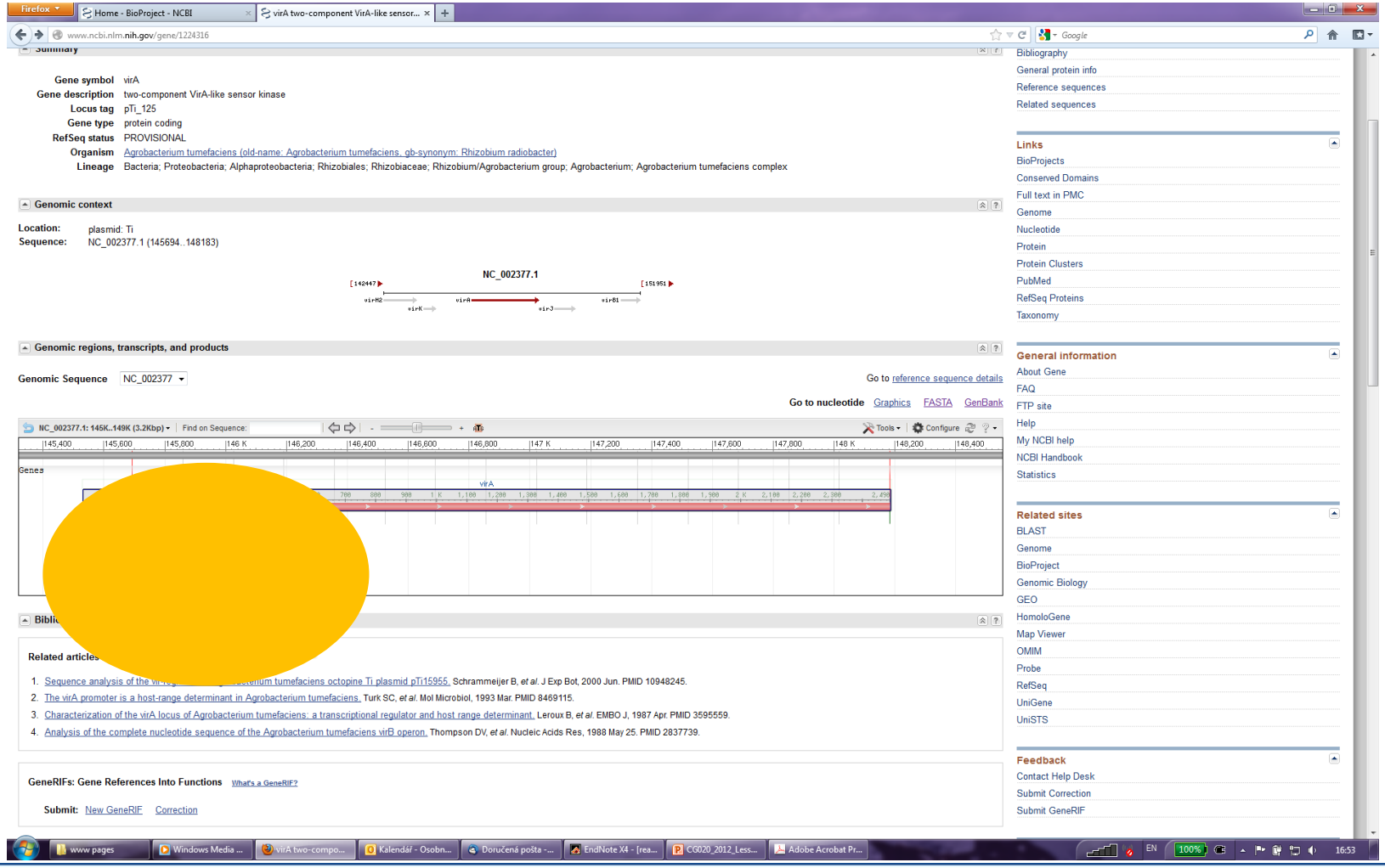
# Primární databáze

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




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

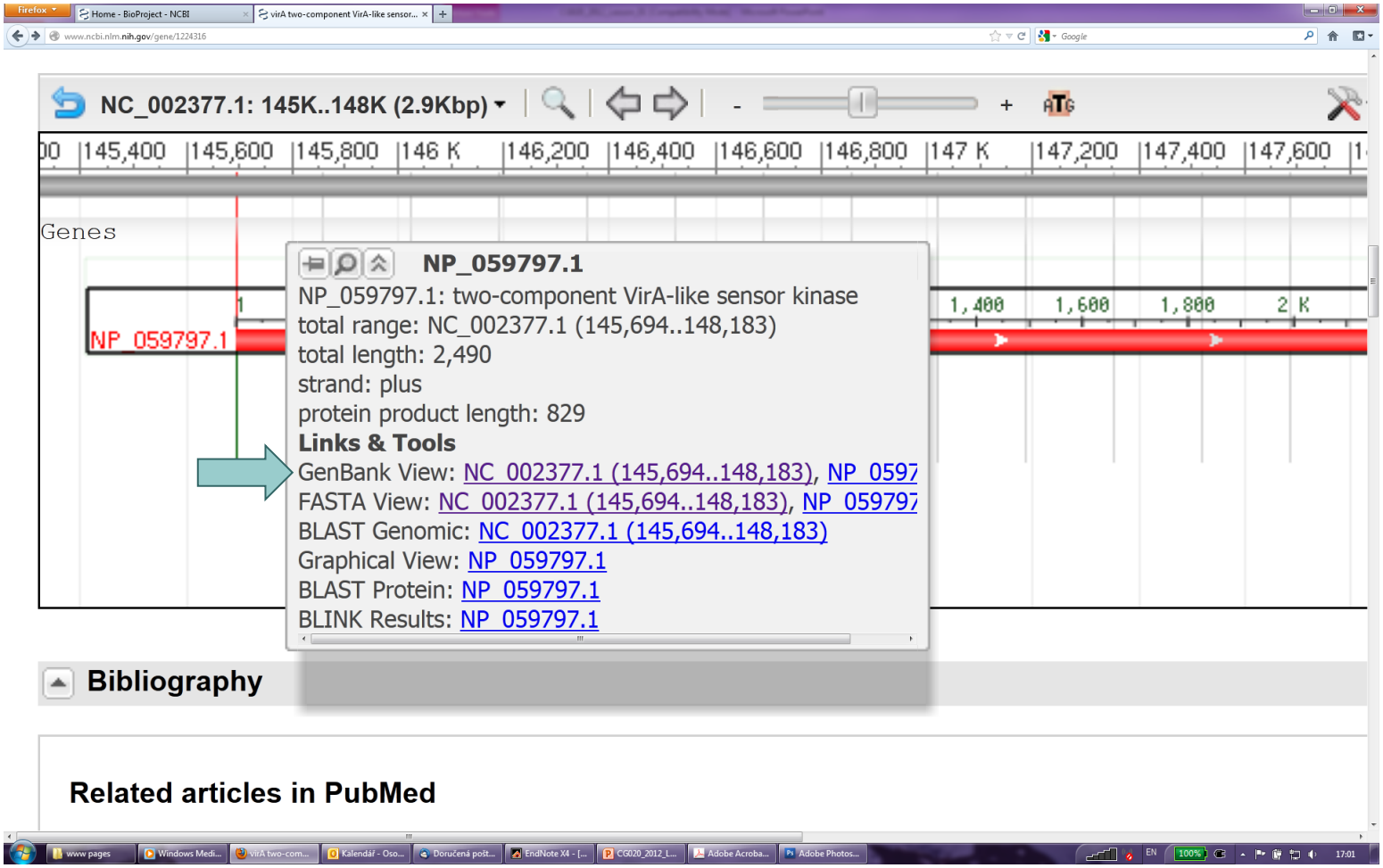
# Primární databáze





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

# Primární databáze




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

# Primární databáze

# Primární databáze

```
                     /translation="MNGRYSPTRQDFKTGAKPWSILALIVAAMIFAFMAVASWQDNAT
                     TQAILSQLRSINADSASLQRDVLRAHTGTVANYRPIISRLGALRKNLEDLKQLPRQSH
                     IVSENAAQLLRQLEVSLNSADAAVAAPGAQNVRLQDSLASPTRALSSLPGKASTDQT
                     LEKPTELASMMLQFLRQPSPAISFEISLELERLQKQRGLDEAPVRILAREGPIILSLL
                     PQVKDLVNMIQTSDTAEIAEMLQRECLEVYSLKNVEERSARIFLGSASVGLCLYIITL
                     VYRLRKKTDWLARRLDYEELIKEIGVCFPGHAATTSSAQAALRIIQRPFDADTCALAL
                     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 tcggcogcaa 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
```

# 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.

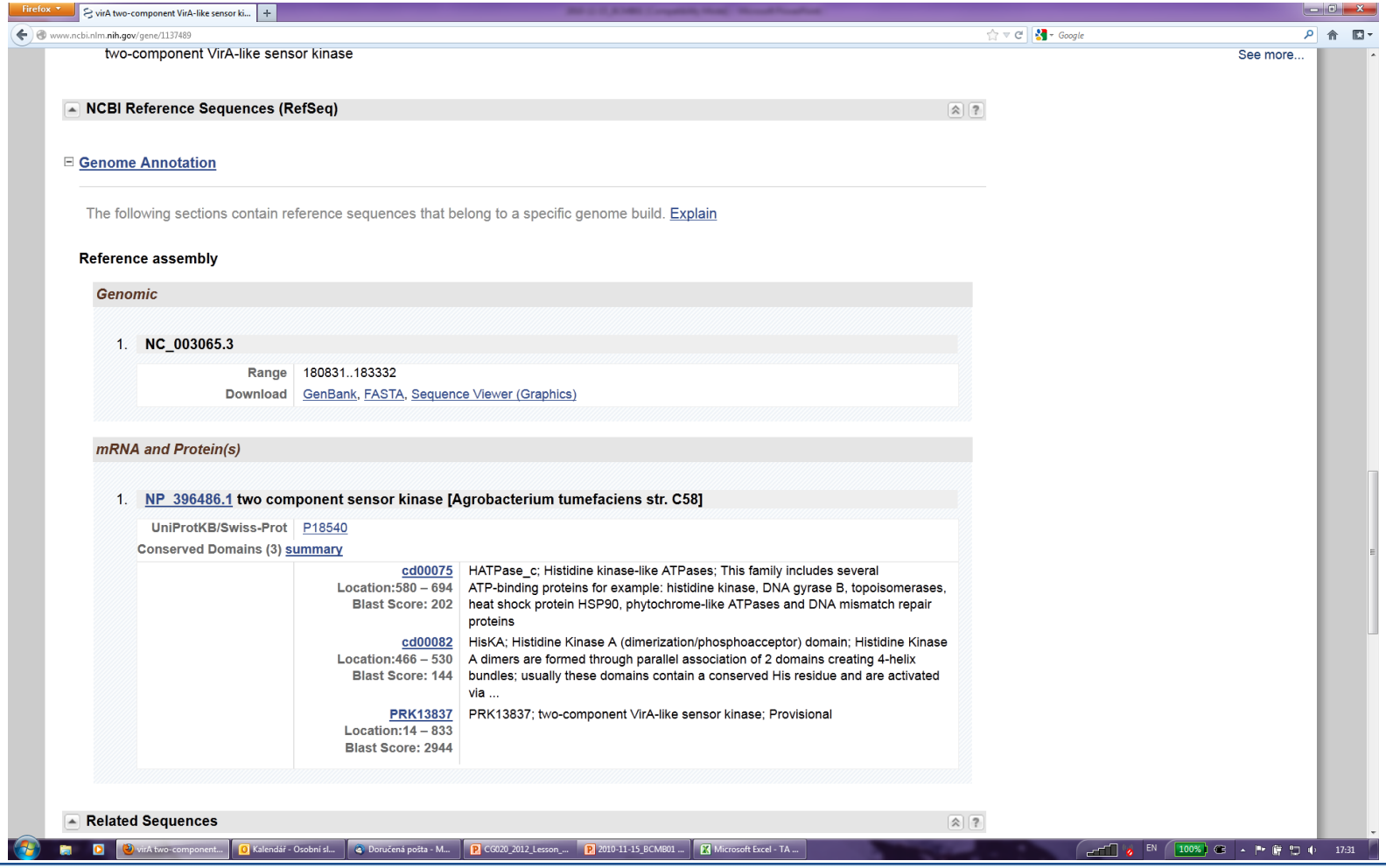
RefSeq identifiers include the following formats:

| | |
|---|---|
| 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

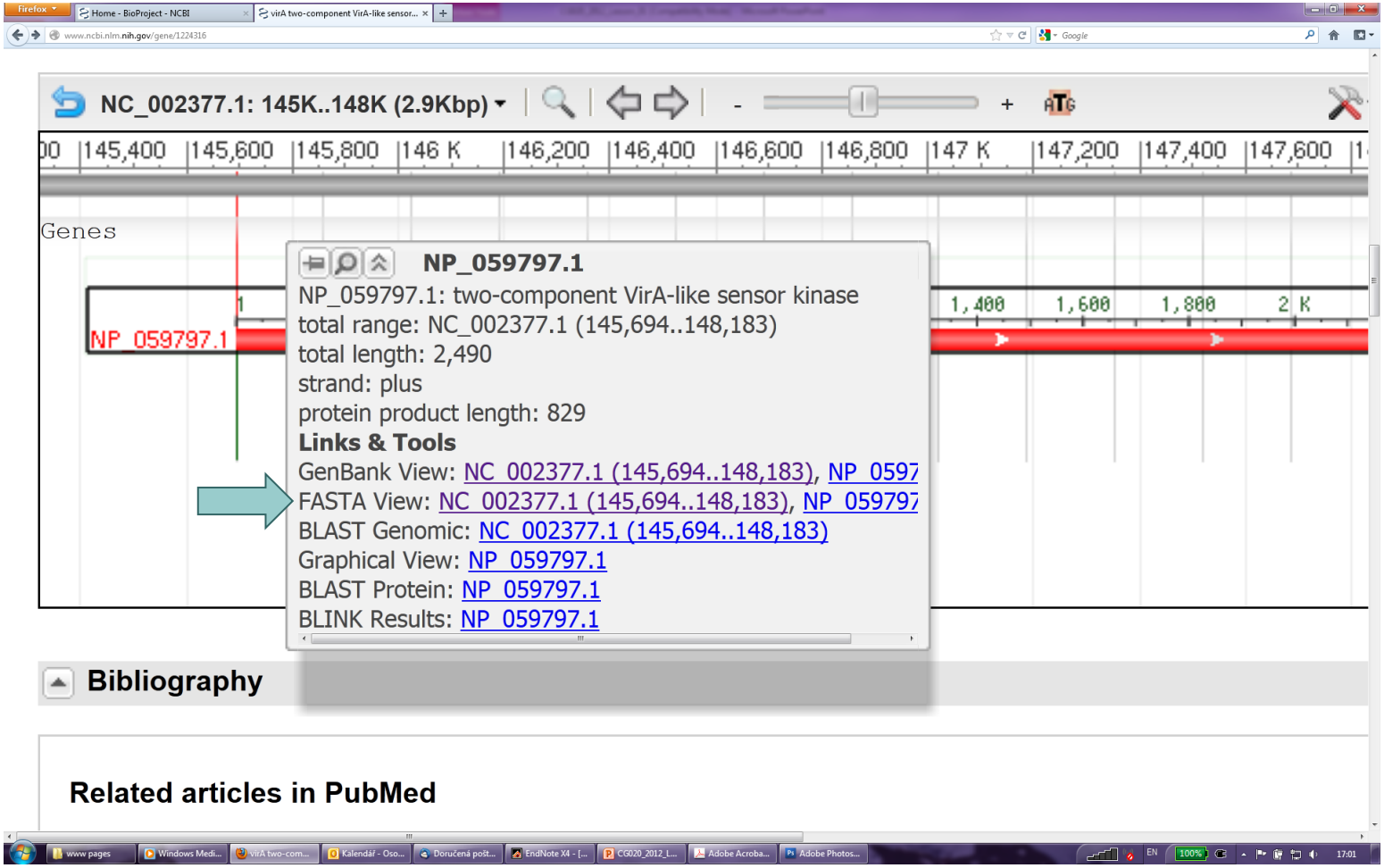
# NCBI's RefSeq project: many accession number formats for genomic, mRNA, protein sequences

| 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

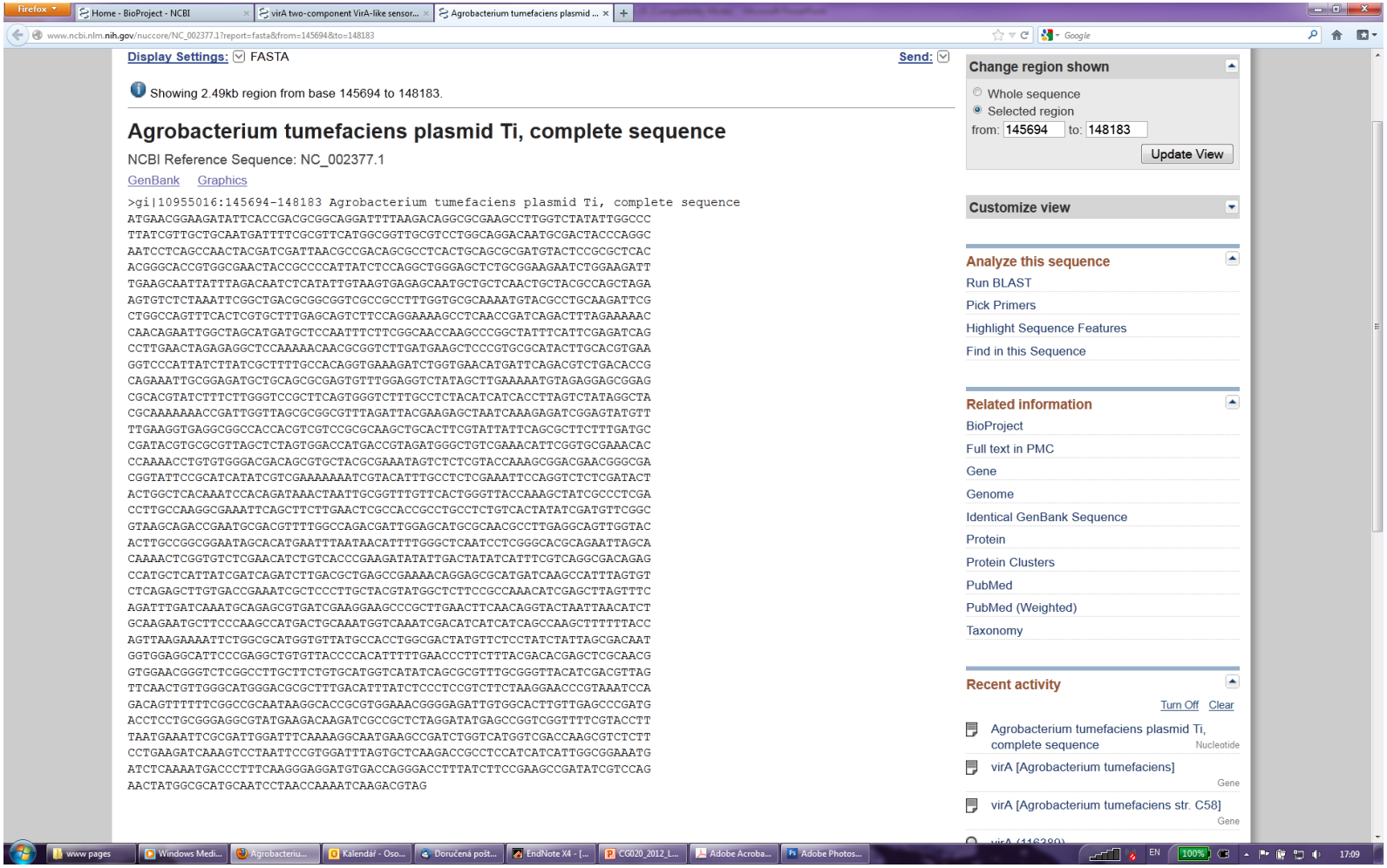
EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

# Primární databáze




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

# Primární databáze





EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

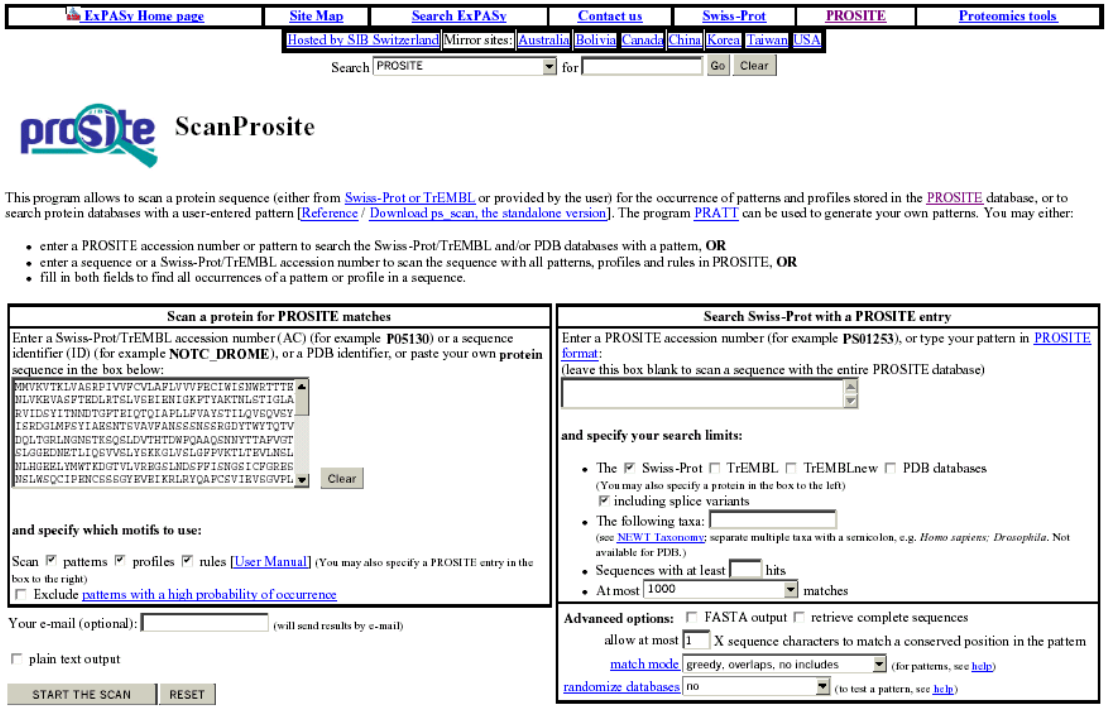
OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

# Sekundární databáze

- databáze funkčních nebo strukturálních ***motivů*** získaných srovnáváním primárních dat (sekvencí)
  - PROSITE, http://www.expasy.org/prosite/



# Sekundární databáze

- databáze funkčních nebo strukturálních ***motivů*** získaných srovnáváním primárních dat (sekvencí)
  - 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

# Sekundární databáze

- databáze funkčních nebo strukturálních ***motivů*** získaných srovnáváním primárních dat (sekvencí)
  - PROSITE, http://www.expasy.org/prosite/

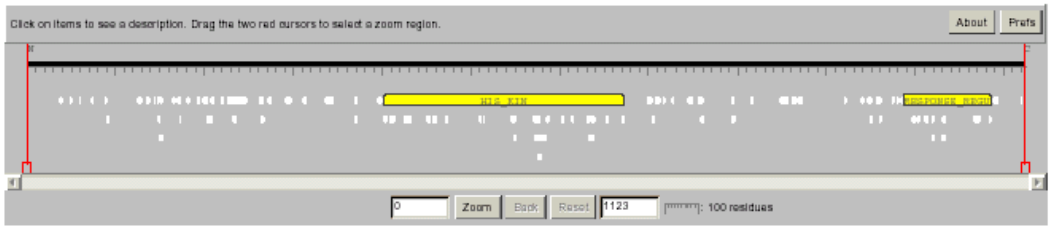
>PDOC50109 PS50109 **HIS_KIN** Histidine kinase domain [profile].

```
  402 - 671  NASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVNVCAKDLVALLNSVLDMSKIRSG
             KMQLVREDFNLSKLLKDVIDFYHPVAMKKGVDVVLDPHDgsvfKPSNVRGDSGRLKQILN
             NLVSNAVKFTVD--GHIAVRAWAQrpgsnsavvlaaypkgvakfvksmfcknkeesstye
             teisnsirnnanTMEFVFEVDDTGKGIPMEMRKSVFENYVQVRRtAQGHQGTGLGLGIVQ
             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 |
|---|---|---|---|---|---|---|


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

# Sekundární databáze

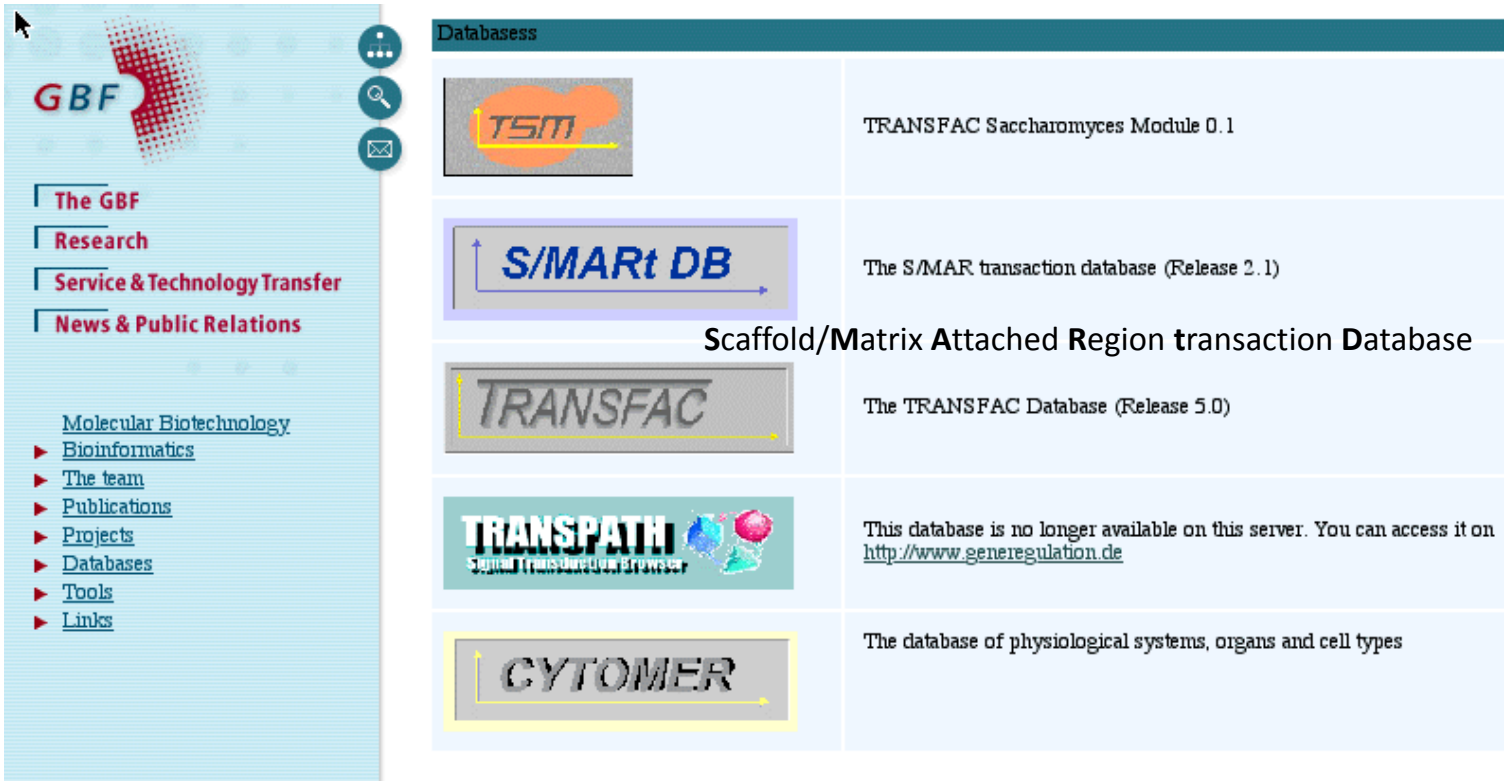
- databáze funkčních nebo strukturálních ***motivů*** získaných srovnáváním primárních dat (sekvencí)

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


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

# Sekundární databáze

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




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

# Strukturální databáze

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


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

# Strukturální databáze

- 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

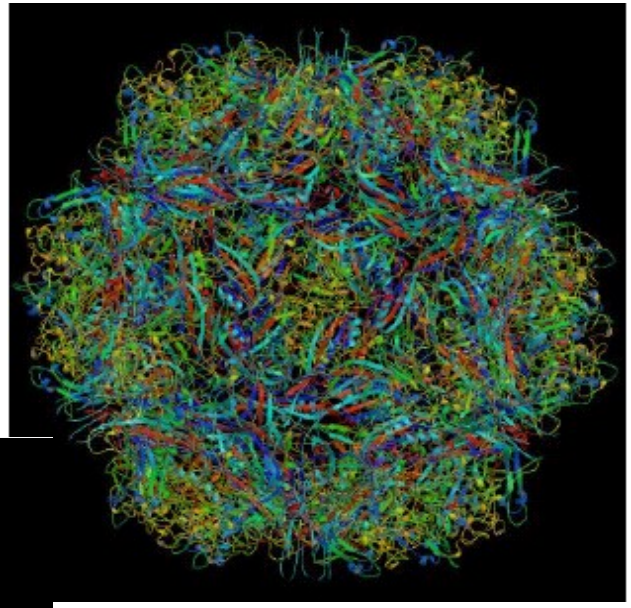
Geometry

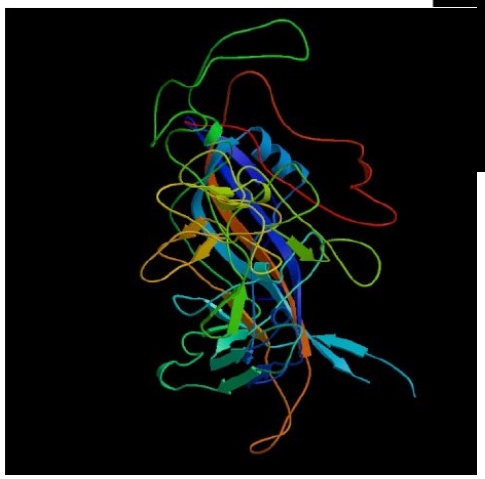
Other Sources

Sequence Details

Explore

SearchLite SearchFields

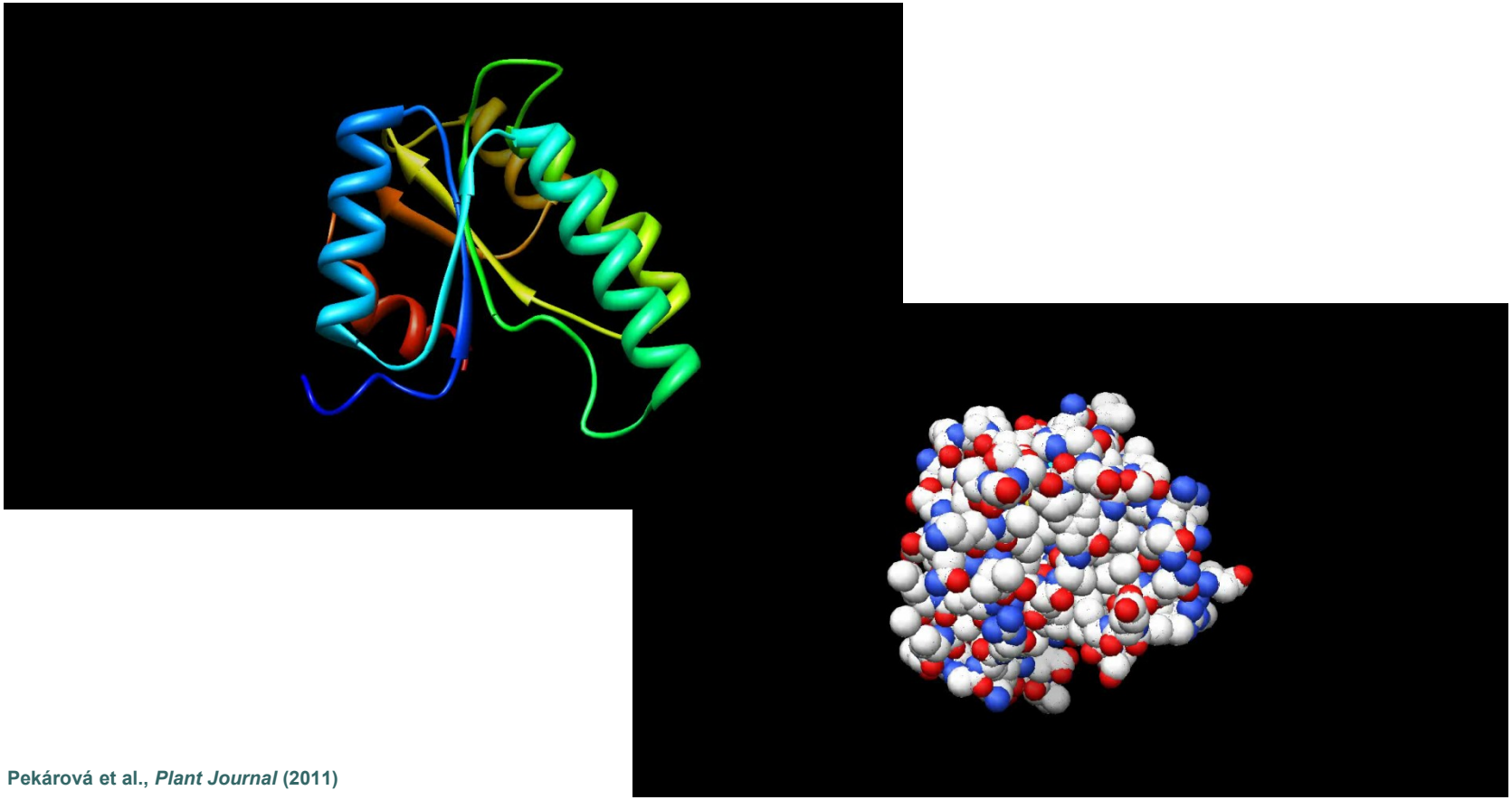




http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=;pid=1735610643493 44&bio=1&opt=show&size=500 12/29/2003

# Strukturální databáze

- 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

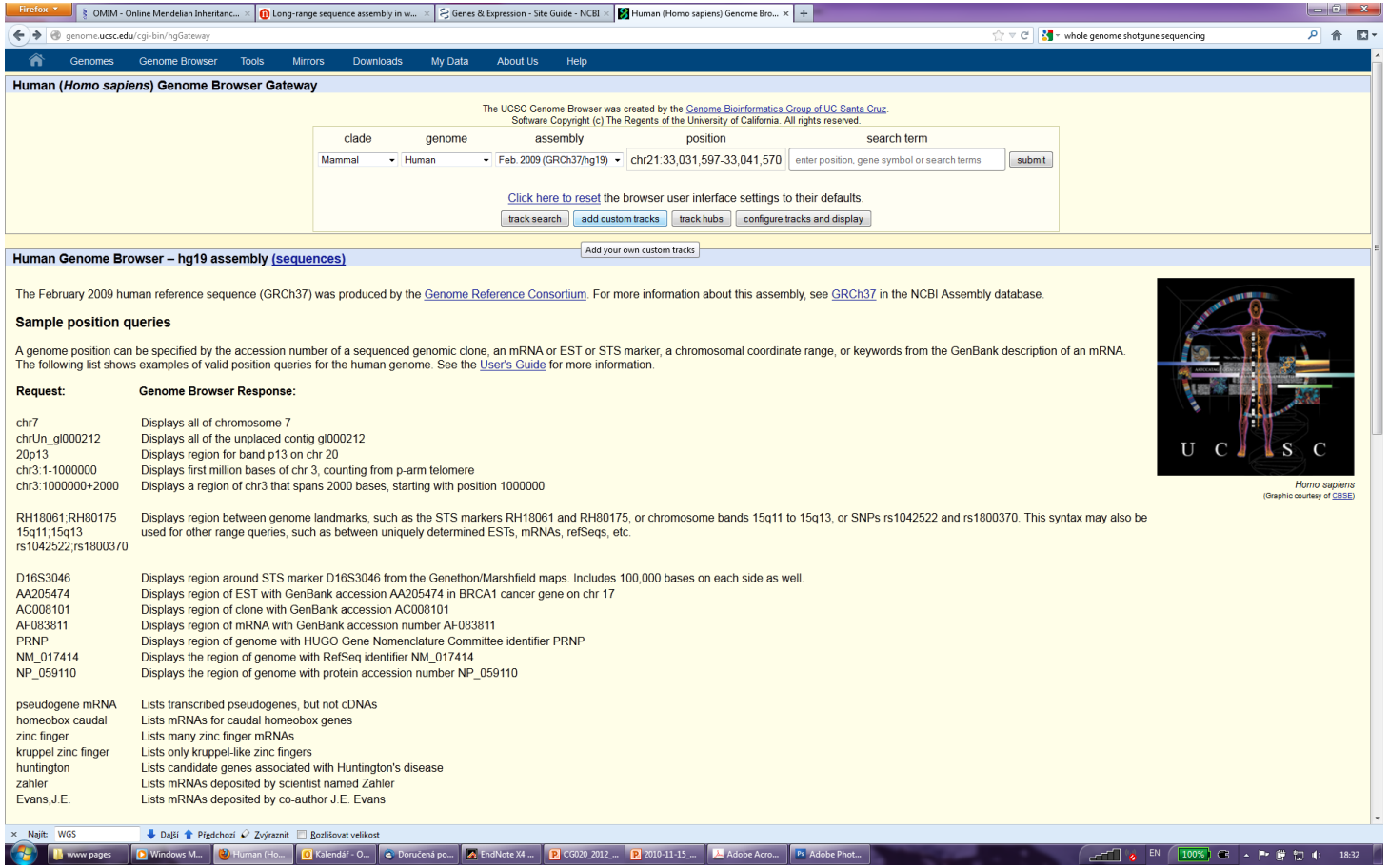
# Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje

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

# Genomové zdroje

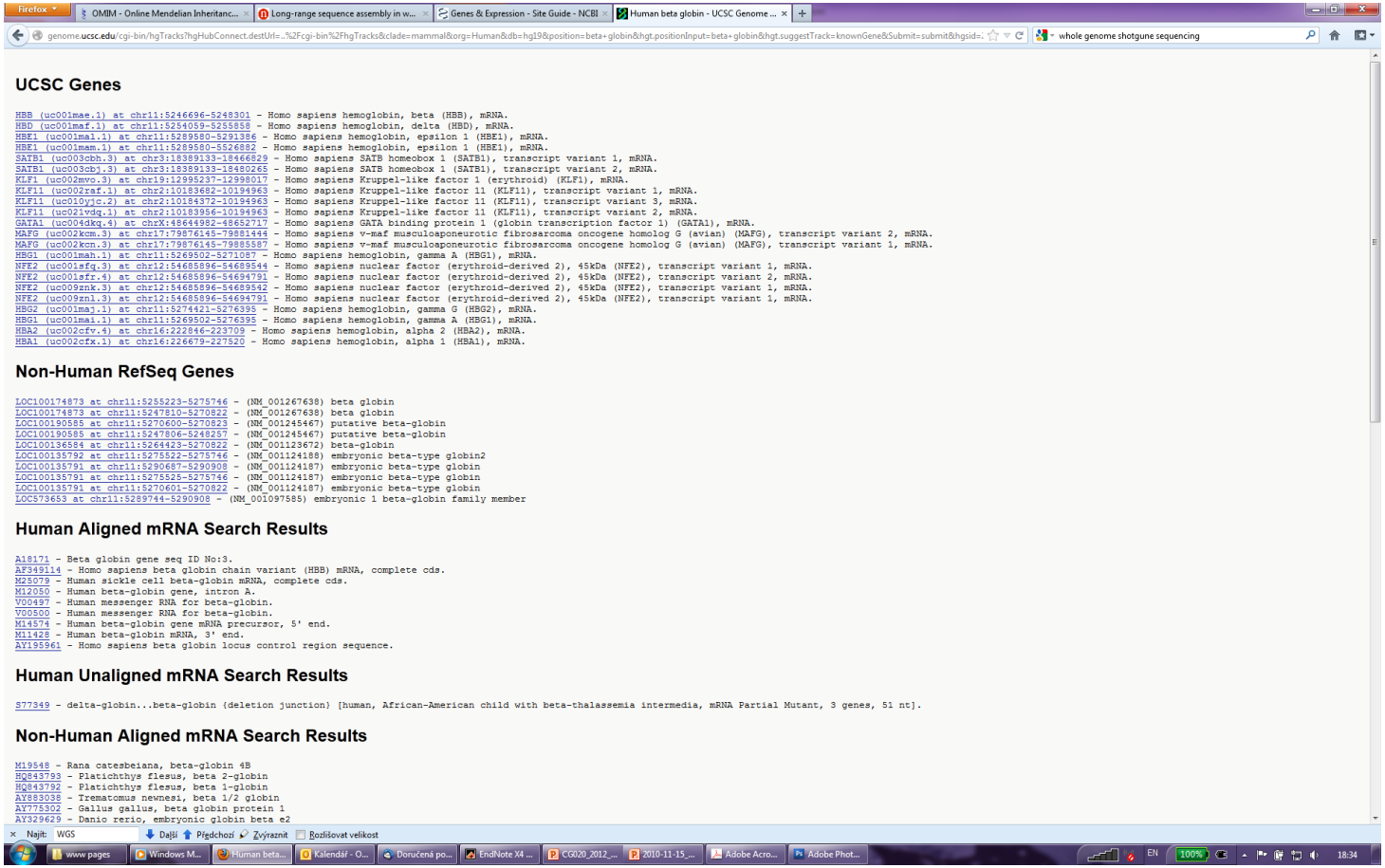
- Human Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway




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

# Genomové zdroje

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



**UCSC Genes**

HBB (uc001mae.1) at chr11:5246696-5248301 - Homo sapiens hemoglobin, beta (HBB), mRNA.
HBD (uc001maf.1) at chr11:5254059-5255858 - Homo sapiens hemoglobin, delta (HBD), mRNA.
HBE1 (uc001mal.1) at chr11:5289580-5291386 - Homo sapiens hemoglobin, epsilon 1 (HBE1), mRNA.
HBE1 (uc001mam.1) at chr11:5289580-5526882 - Homo sapiens hemoglobin, epsilon 1 (HBE1), mRNA.
SATB1 (uc003cbh.3) at chr3:18389133-18466829 - Homo sapiens SATB homeobox 1 (SATB1), transcript variant 1, mRNA.
SATB1 (uc003cbj.3) at chr3:18389133-18480265 - Homo sapiens SATB homeobox 1 (SATB1), transcript variant 2, mRNA.
KLF1 (uc002mvo.3) at chr19:12995237-12998017 - Homo sapiens Kruppel-like factor 1 (erythroid) (KLF1), mRNA.
KLF11 (uc002raf.1) at chr2:10183682-10194963 - Homo sapiens Kruppel-like factor 11 (KLF11), transcript variant 1, mRNA.
KLF11 (uc010yjc.2) at chr2:10184372-10194963 - Homo sapiens Kruppel-like factor 11 (KLF11), transcript variant 3, mRNA.
KLF11 (uc021vdq.1) at chr2:10183956-10194963 - Homo sapiens Kruppel-like factor 11 (KLF11), transcript variant 2, mRNA.
GATA1 (uc004dkq.4) at chrX:48644982-48652717 - Homo sapiens GATA binding protein 1 (globin transcription factor 1) (GATA1), mRNA.
MAFG (uc002kcm.3) at chr17:79876145-79881444 - Homo sapiens v-maf musculoaponeurotic fibrosarcoma oncogene homolog G (avian) (MAFG), transcript variant 2, mRNA.
MAFG (uc002kcn.3) at chr17:79876145-79885587 - Homo sapiens v-maf musculoaponeurotic fibrosarcoma oncogene homolog G (avian) (MAFG), transcript variant 1, mRNA.
HBG1 (uc001mah.1) at chr11:5269502-5271087 - Homo sapiens hemoglobin, gamma A (HBG1), mRNA.
NFE2 (uc001sfq.3) at chr12:54685896-54689544 - Homo sapiens nuclear factor (erythroid-derived 2), 45kDa (NFE2), transcript variant 1, mRNA.
NFE2 (uc001sfr.4) at chr12:54685896-54694791 - Homo sapiens nuclear factor (erythroid-derived 2), 45kDa (NFE2), transcript variant 2, mRNA.
NFE2 (uc009znk.3) at chr12:54685896-54689542 - Homo sapiens nuclear factor (erythroid-derived 2), 45kDa (NFE2), transcript variant 1, mRNA.
NFE2 (uc009znl.3) at chr12:54685896-54694791 - Homo sapiens nuclear factor (erythroid-derived 2), 45kDa (NFE2), transcript variant 1, mRNA.
HBG2 (uc001maj.1) at chr11:5274421-5276395 - Homo sapiens hemoglobin, gamma G (HBG2), mRNA.
HBG1 (uc001mai.1) at chr11:5269502-5276395 - Homo sapiens hemoglobin, gamma A (HBG1), mRNA.
HBA2 (uc002cfv.4) at chr16:222846-223709 - Homo sapiens hemoglobin, alpha 2 (HBA2), mRNA.
HBA1 (uc002cfx.1) at chr16:226679-227520 - Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA.

**Non-Human RefSeq Genes**

LOC100174873 at chr11:5255223-5275746 - (NM_001267638) beta globin
LOC100174873 at chr11:5247810-5270822 - (NM_001267638) beta globin
LOC100190585 at chr11:5270600-5270823 - (NM_001245467) putative beta-globin
LOC100190585 at chr11:5247806-5248257 - (NM_001245467) putative beta-globin
LOC100136584 at chr11:5264423-5270822 - (NM_001123672) beta-globin
LOC100135792 at chr11:5275522-5275746 - (NM_001124188) embryonic beta-type globin2
LOC100135791 at chr11:5290687-5290908 - (NM_001124187) embryonic beta-type globin
LOC100135791 at chr11:5275525-5275746 - (NM_001124187) embryonic beta-type globin
LOC100135791 at chr11:5270601-5270822 - (NM_001124187) embryonic beta-type globin
LOC573653 at chr11:5289744-5290908 - (NM_001097585) embryonic 1 beta-globin family member

**Human Aligned mRNA Search Results**

A18171 - Beta globin gene seq ID No:3.
AF349114 - Homo sapiens beta globin chain variant (HBB) mRNA, complete cds.
M25079 - Human sickle cell beta-globin mRNA, complete cds.
M12050 - Human beta-globin gene, intron A.
V00497 - Human messenger RNA for beta-globin.
V00500 - Human messenger RNA for beta-globin.
M14574 - Human beta-globin gene mRNA precursor, 5' end.
M11428 - Human beta-globin mRNA, 3' end.
AY195961 - Homo sapiens beta globin locus control region sequence.

**Human Unaligned mRNA Search Results**

S77349 - delta-globin...beta-globin {deletion junction} [human, African-American child with beta-thalassemia intermedia, mRNA Partial Mutant, 3 genes, 51 nt].

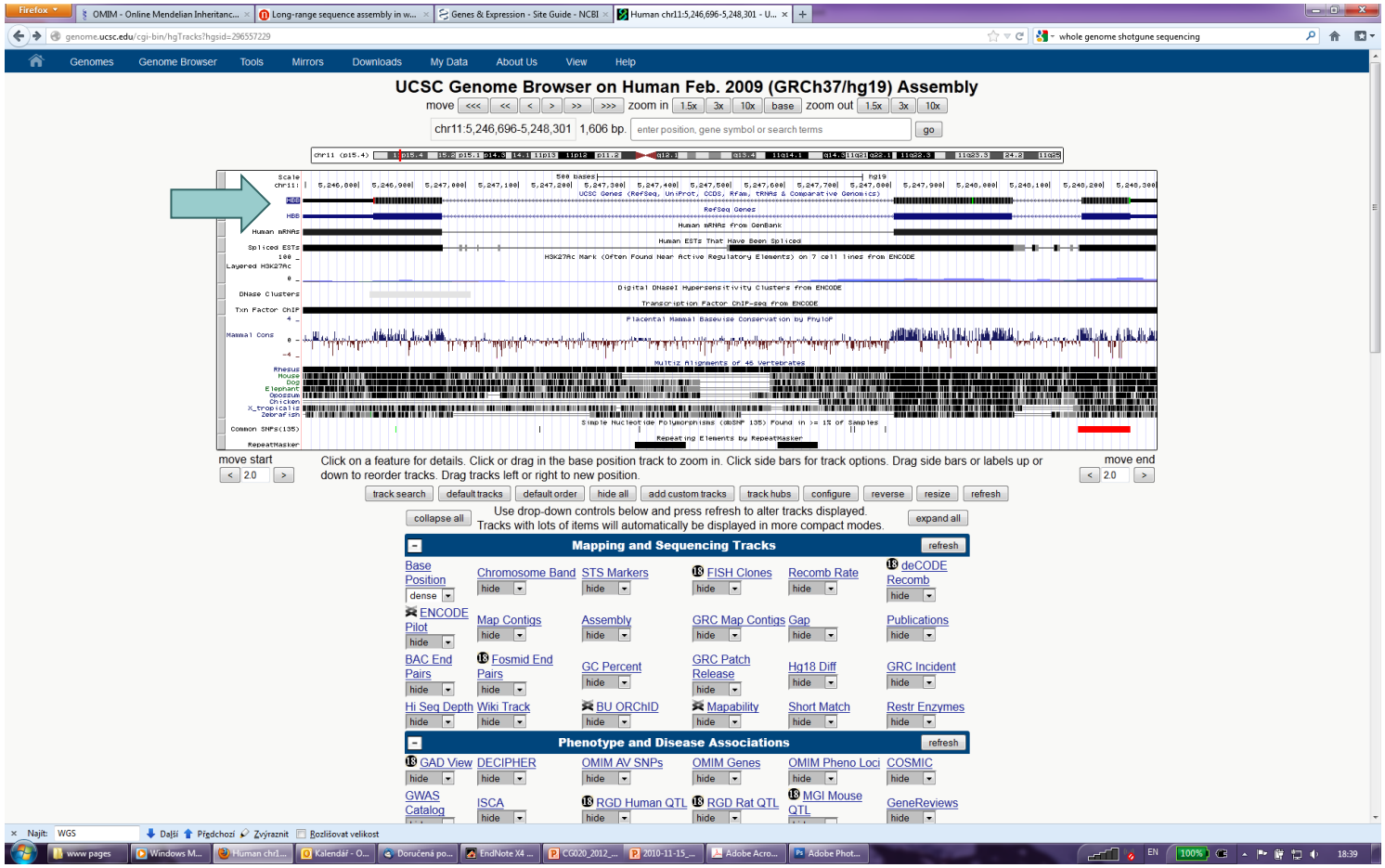
**Non-Human Aligned mRNA Search Results**

M19548 - Rana catesbeiana, beta-globin 4B
HQ843793 - Platichthys flesus, beta 2-globin
HQ843792 - Platichthys flesus, beta 1-globin
AY883038 - Trematomus newnesi, beta 1/2 globin
AY775302 - Gallus gallus, beta globin protein 1
AY329629 - Danio rerio, embryonic globin beta e2

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

# Genomové zdroje

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

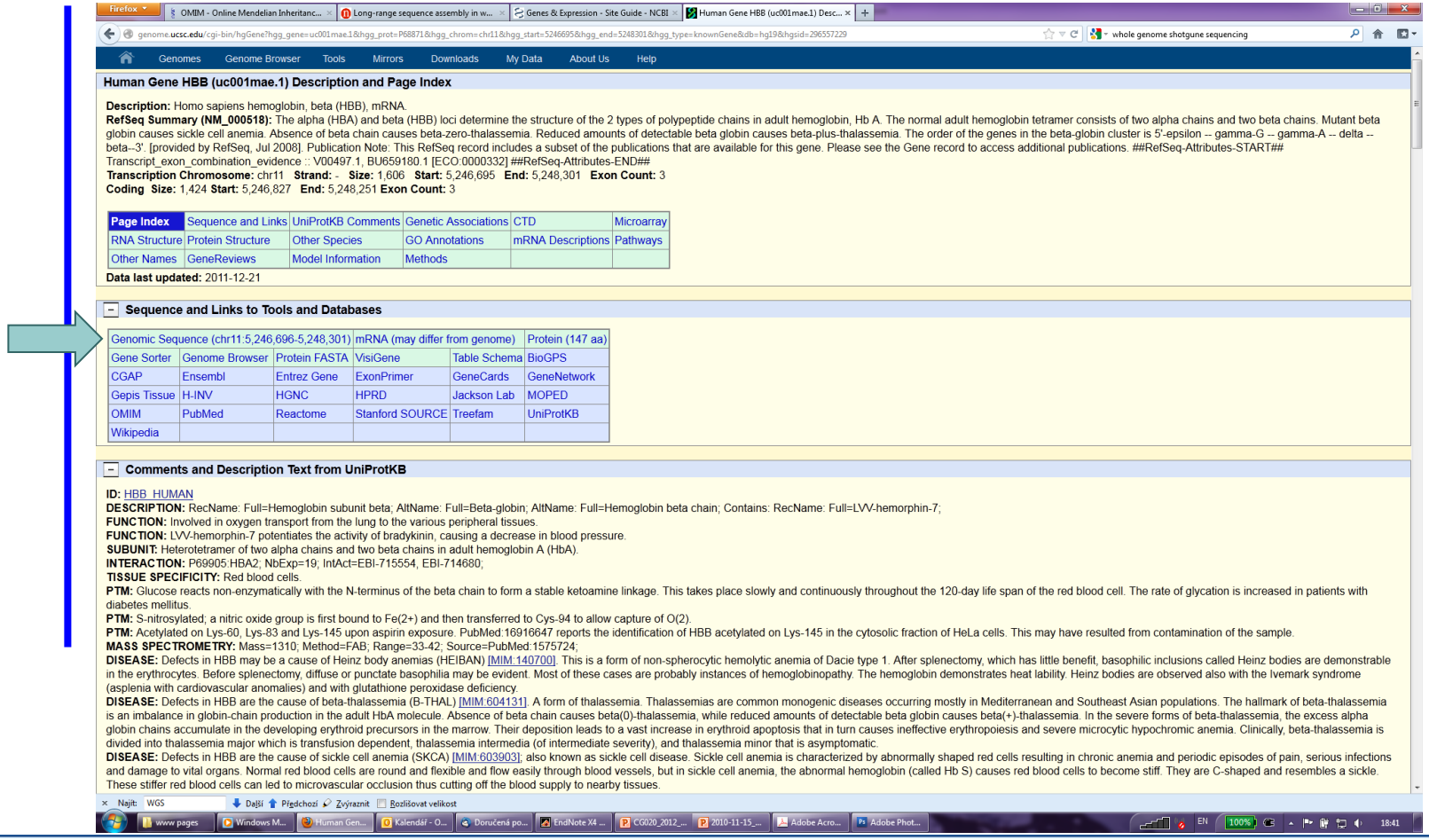




EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

# Genomové zdroje

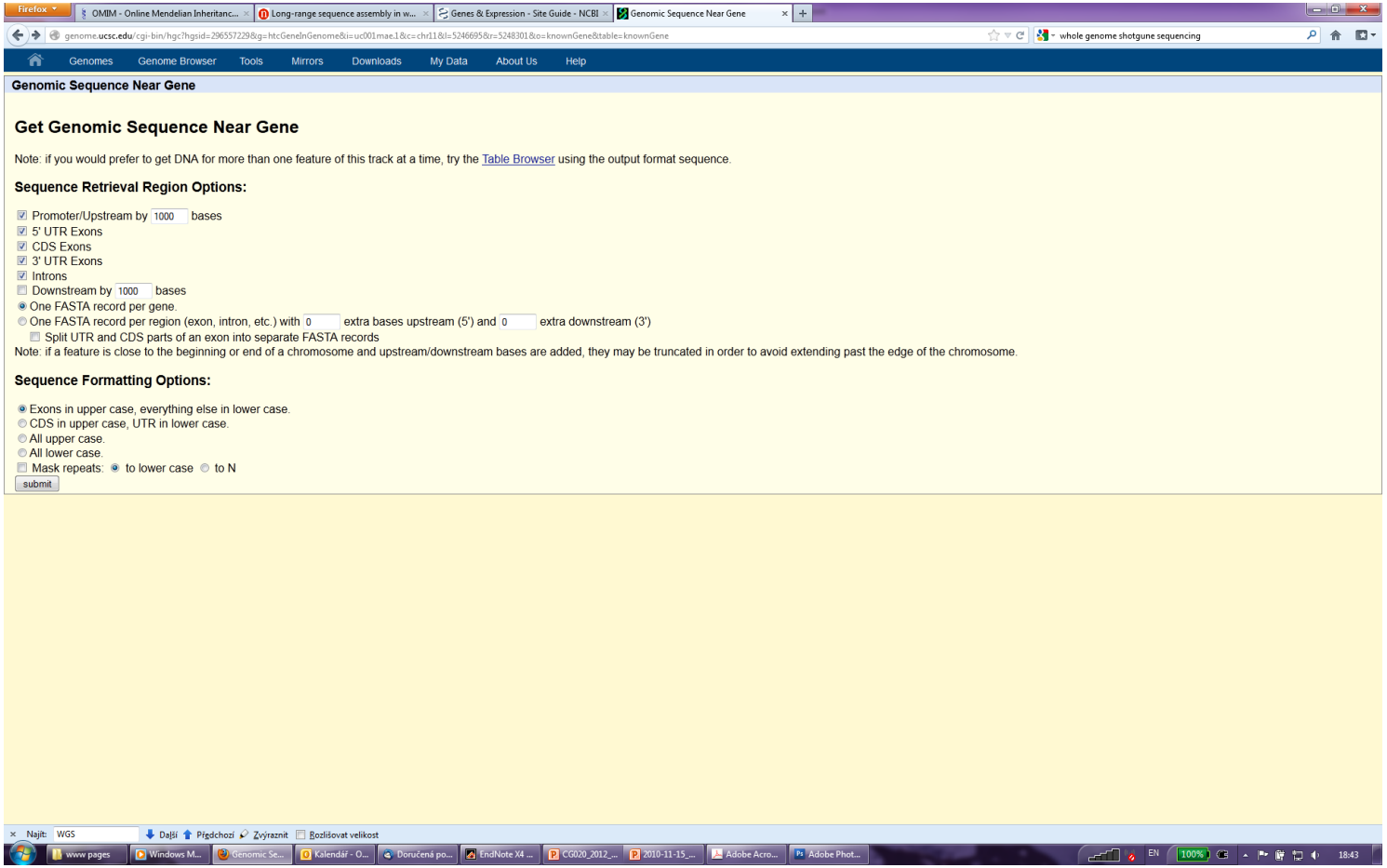
- Human Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway




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

# Genomové zdroje

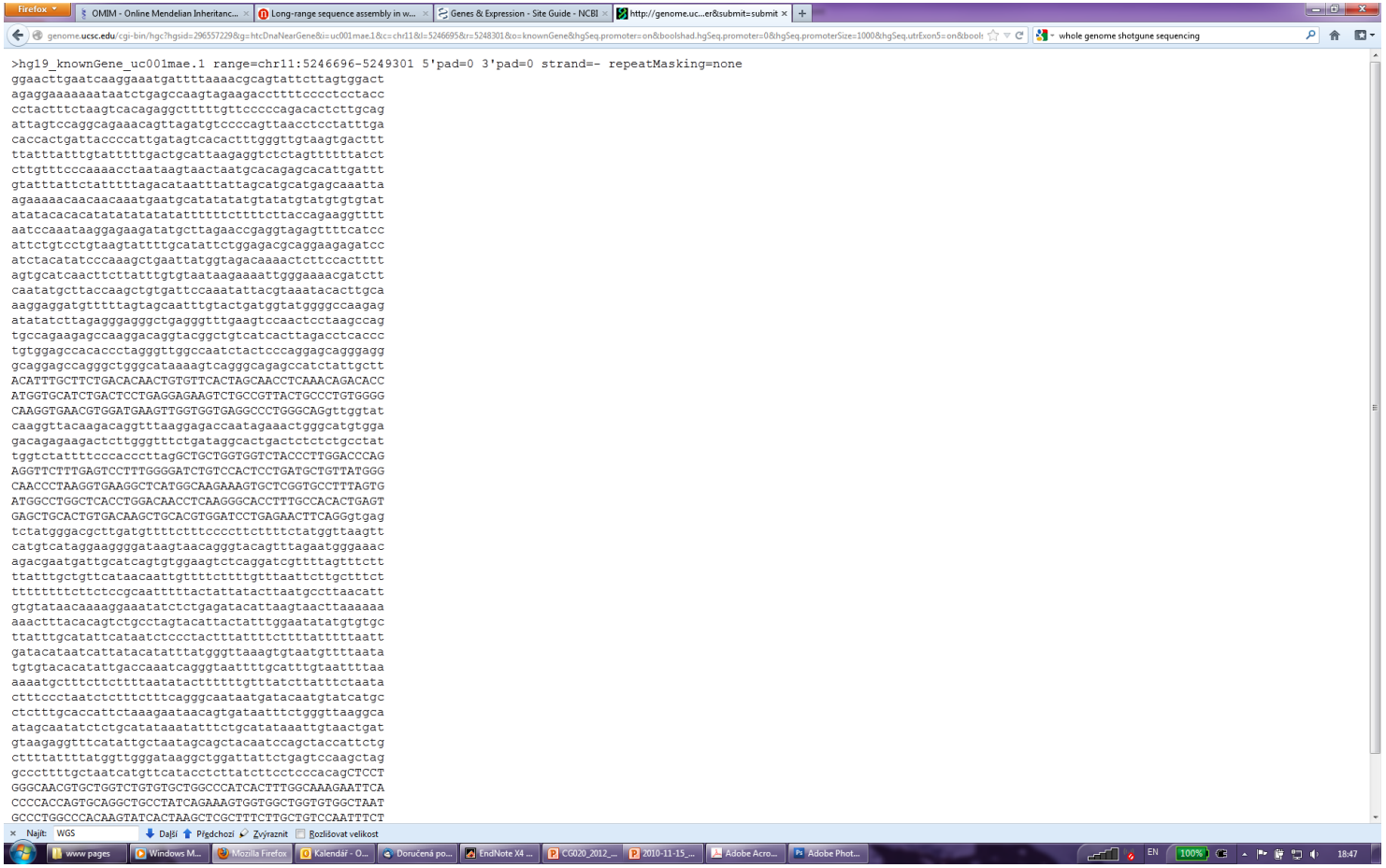
- Human Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway




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

# Genomové zdroje

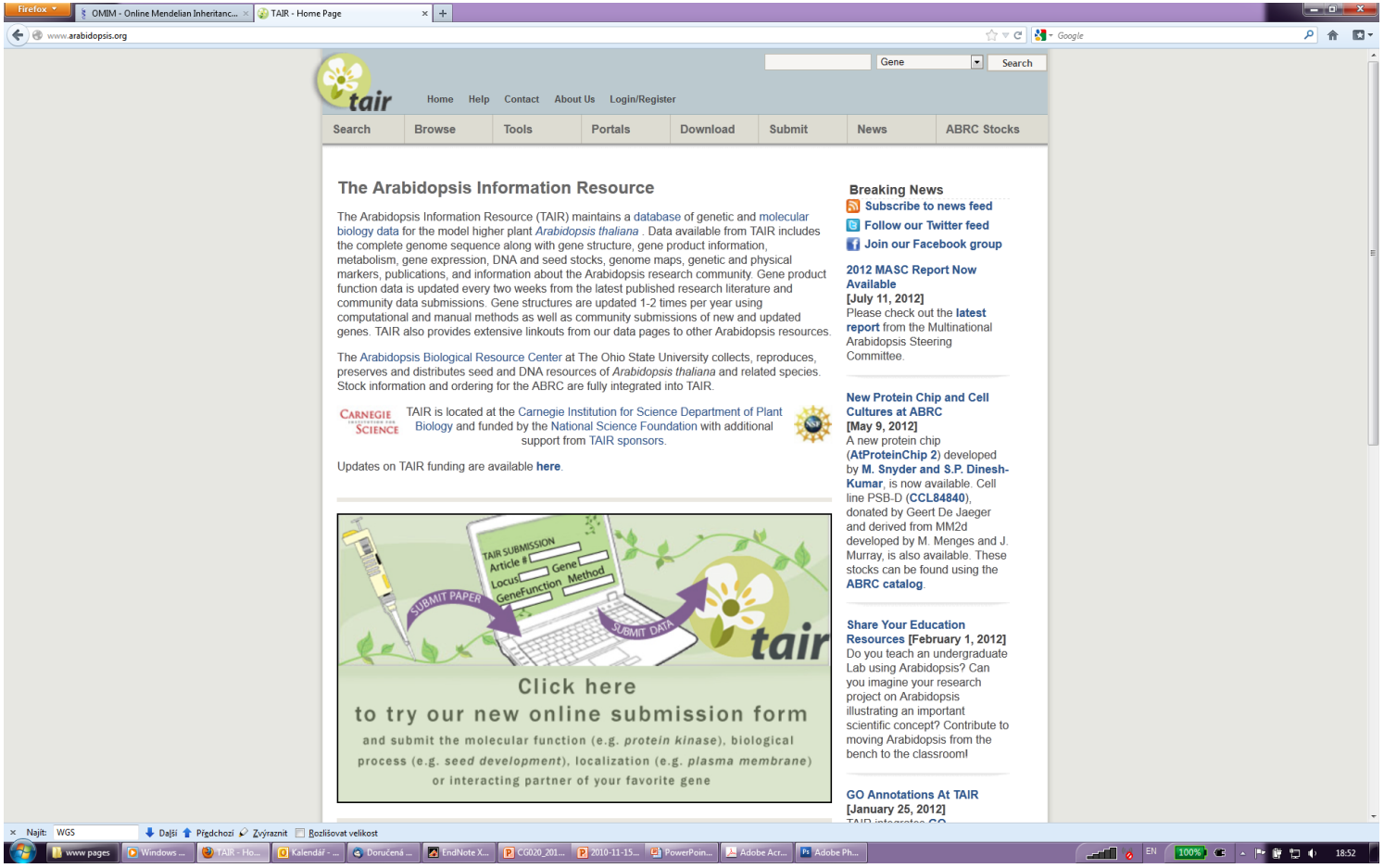
- Human Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway




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

# Genomové zdroje

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

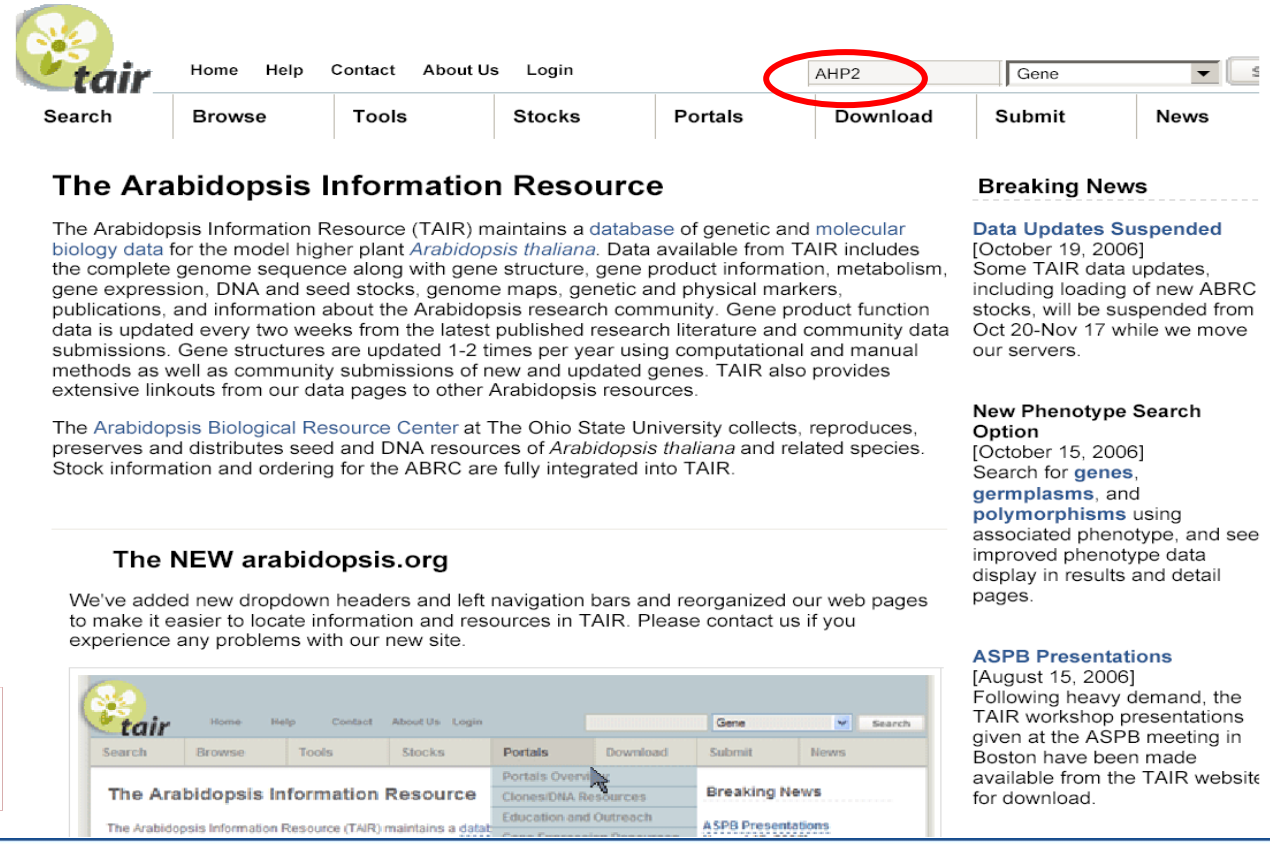




EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

# Genomové zdroje

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




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

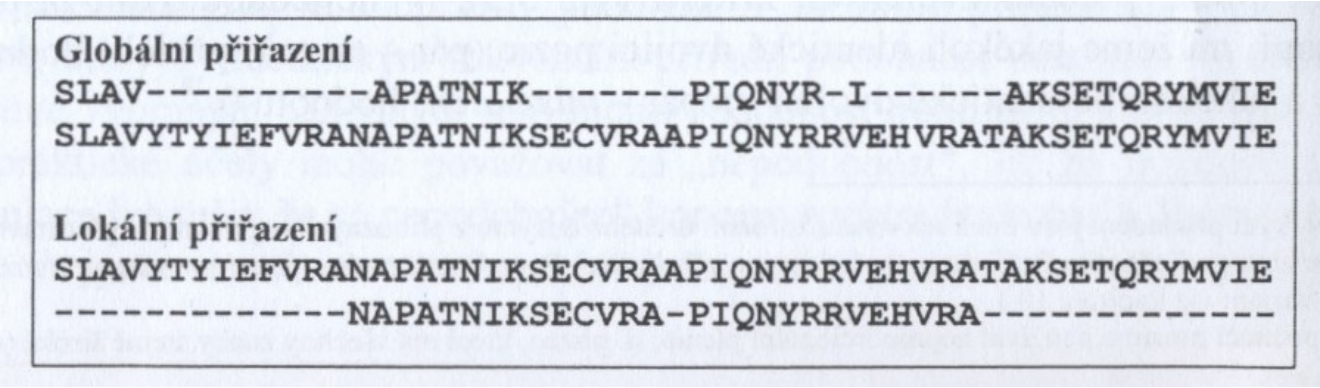
# Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií

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

# Analytické nástroje

- Globální vs. lokální přiřazení



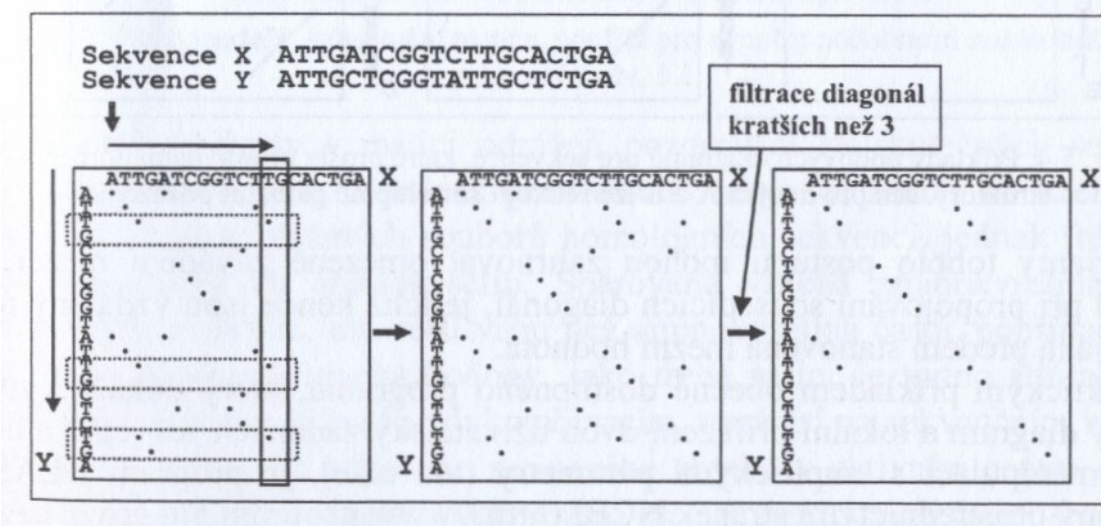

Cvrčková, Úvod do praktické bioinformatiky

  - globální přiřazení pouze u sekvencí, které jsou si podobné (za cenu vnášení mezer do jedné nebo obou sekvencí)
  - globální přiřazení se používá především v případě mnohačetného přiřazování (CLUSTALW, viz dále)
  - lokální přiřazení umožní identifikaci a srovnání i v případě porovnávání pouze **úseků sekvencí** s významnou mírou podobnosti, např. i při záměně pořadí proteinových domén během evoluce

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

# Analytické nástroje

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)



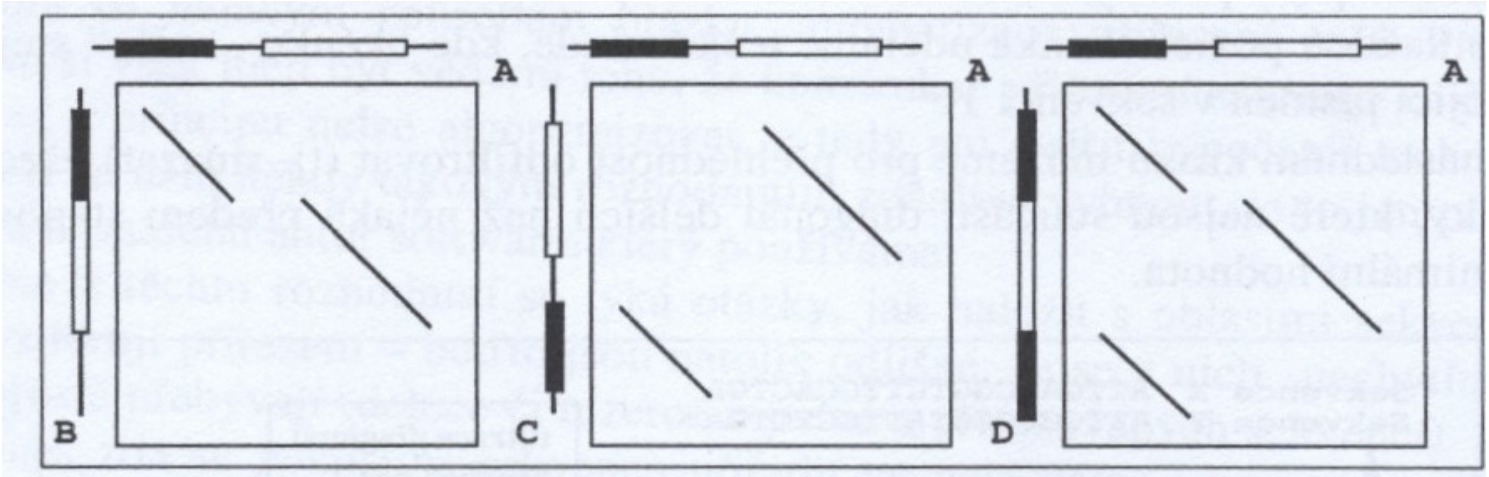

Cvrčková, Úvod do praktické bioinformatiky

  - vynesení sekvencí proti sobě
  - identifikace shody v okně o dané velikosti (např. 2 bp)
  - „odfiltrování“ diagonál o délce menší než je mezní hodnota (threshold)


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

# Analytické nástroje

- příklady srovnání sekvencí pomocí bodového diagramu



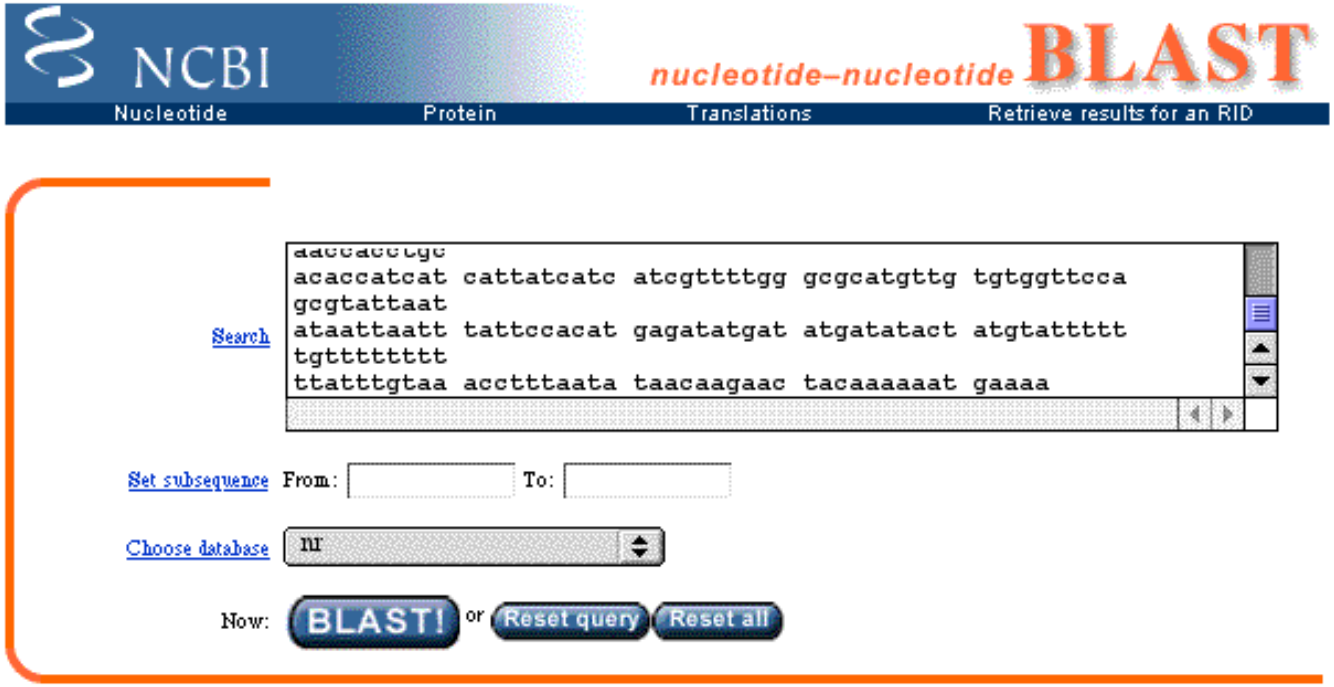

Cvrčková, Úvod do praktické bioinformatiky

- globálně lze srovnávat pouze sekvence A, B
- ostatní sekvence prošly během evoluce záměnou domén a je nutné je porovnávat lokálně
- bodový diagram lze získat pomocí srovnávání programem BLAST2 (viz dále)

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

# Analytické nástroje

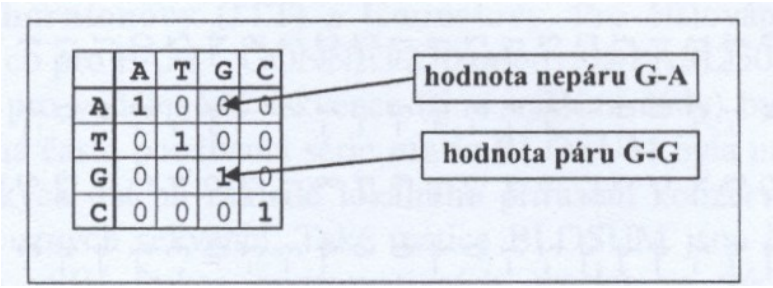
- BLAST http://ncbi.nlm.nih.gov/BLAST/



# BLAST

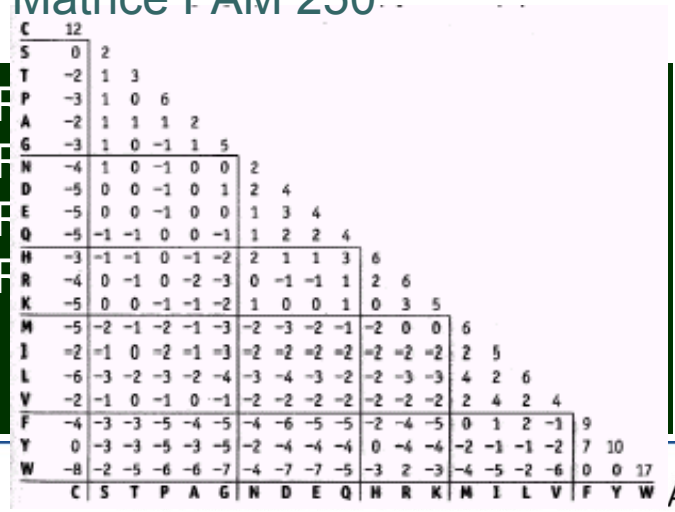
**B**asic **L**ocal **A**lignment **S**earch **T**ool

- Velikost vyhledávacího slova (word size): 10-11 bp, resp. 2-3 aa
  - Primární podobnosti (seed matches)
  - Rozšiřování oblasti homologie doprava i doleva
- Hodnocení homologie pomocí matice PAM (Point Accepted Mutation) nebo BLOSUM (BLOcks Substitution Matrix)
- Zobrazení výsledků




Cvrčková, Úvod do praktické bioinformatiky

Matrice PAM 250




EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

UNIVERSITAS MASARYKIANA BRUNENSIS

ÁVÁNÍ

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

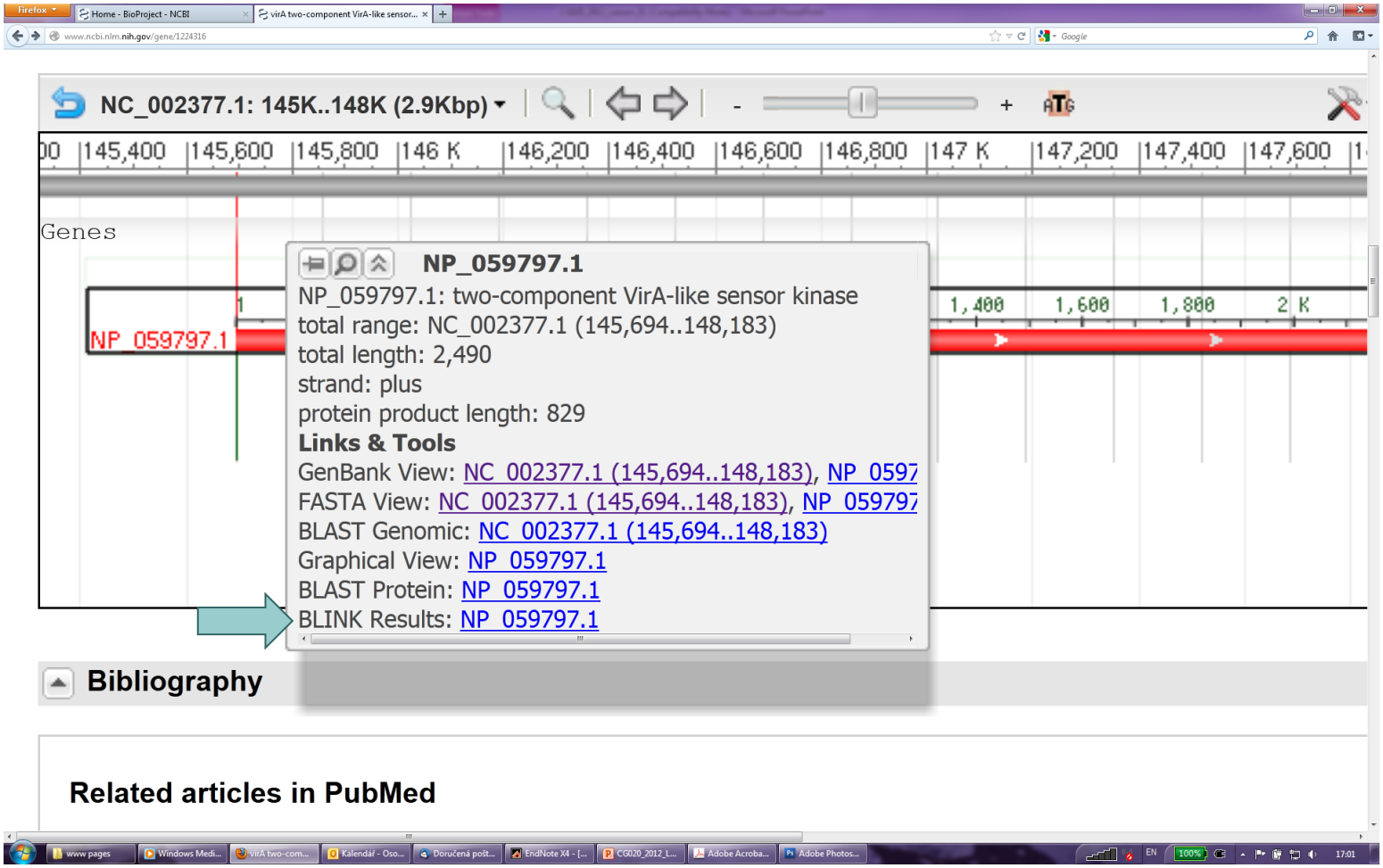
# BLAST

**B**asic **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

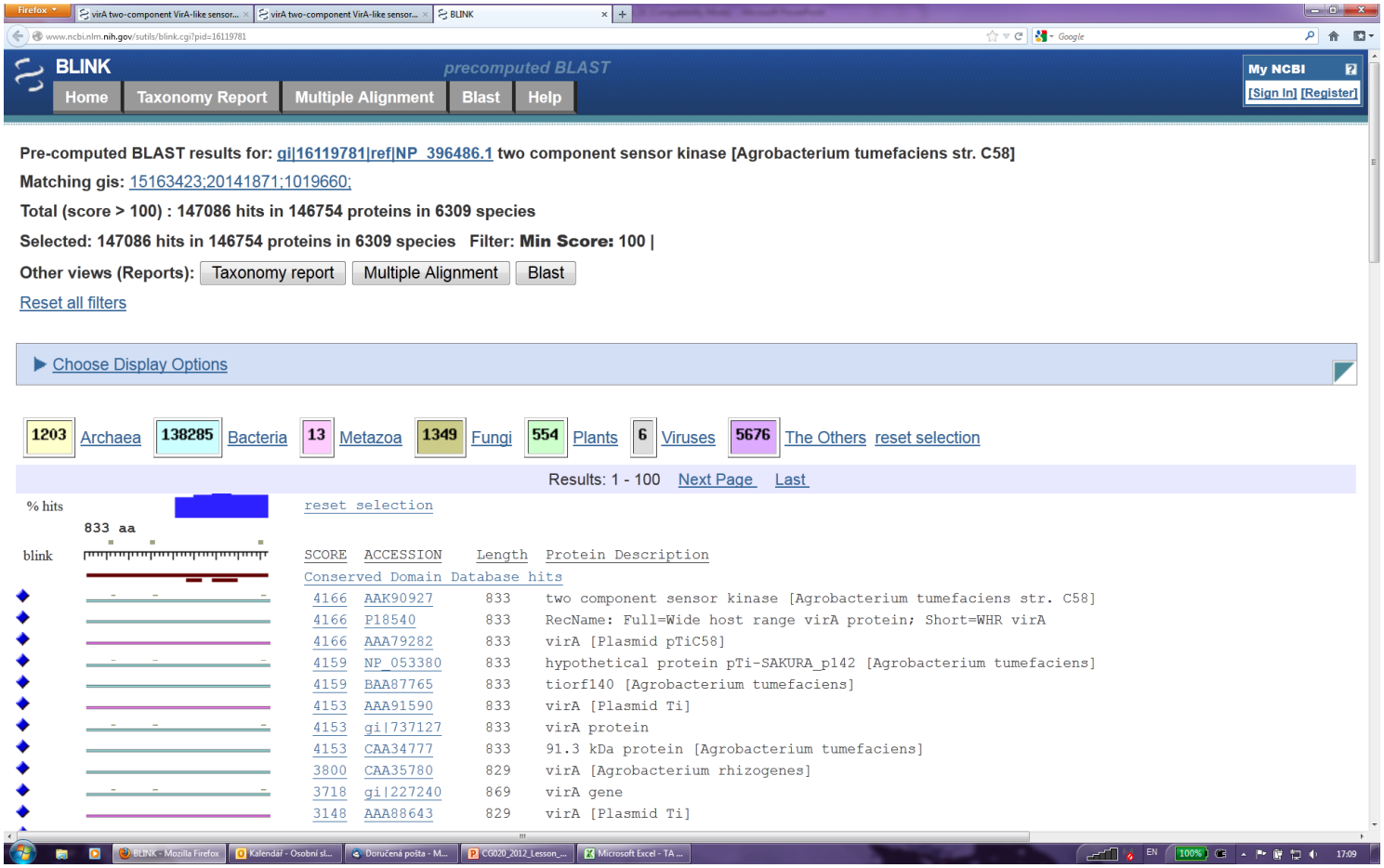
# Primární databáze




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

## **B**asic **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

## Specializované verze

- V současnosti existuje celá řad specializovaných verzí programu BLAST
  - vyhledávání podle zdroje (organismu) sekvencí, např. známých genomů mikroorganismů
  - **BLASTP**
    - vyhledávání podobnosti k proteinu v databázi proteinových sekvencí
  - **BLASTN**
    - vyhledávání podobnosti k nukleotidové sekvenci v databázi nukleotidových sekvencí
    - další varianty jako např. MEGABLAST pro identifikaci totožných nebo velice podobných sekvencí (vyhledává dlouhé podobné úseky nukl. sekvencí)
  - **BLASTX**
    - vyhledávání podobnosti k proteinu v databázi nukleotidových sekvencí přeložených do sekvence aa

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

## Specializované verze

- V současnosti existuje celá řad specializovaných verzí programu BLAST

  - **TBLASTN**
    - vyhledávání k sekvenci nukleotidů přeložené do sekvence aa v databázi proteinů

  - **TBLASTX**
    - vyhledávání k sekvenci nukleotidů přeložené do sekvence aa v databázi nukleotidových sekvencí přeložených do sekvence aa

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

## Specializované verze

- V současnosti existuje celá řad specializovaných verzí programu BLAST

  - **PSI-BLAST** (**P**osition-**S**pecific **I**terated Blast)
    - Prvním krokem je standardní BLAST, při kterém PSI-BLAST identifikuje skupinu podobných sekvencí s E hodnotou lepší než minimální hodnota (standardně 0,005)
    - PSI-BLAST vytváří pro každé přiřazení tzv. PSSM (position specific substitution matrix)
    - PSSM matice zohledňuje výskyt jedné aminokyseliny ve stejné pozici se zvýšenou frekvencí u sekvencí identifikovaných jako podobné v prvním kole pomocí BLAST, což může znamenat funkční konzervovanost

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

## Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST

  - **PHI-BLAST** (**P**attern-**H**it **I**nitiatedBlast)
    - Určen k identifikaci specifické sekvence, např. motivu (pattern) v sekvenci podobných proteinových sekvencí
    - Sekvenci motivu je třeba vložit pomocí speciálního syntaxu
      - [LVIMF] znamená buď Leu, Val, Ile, Met nebo Phe
      - - je oddělovník (neznamená nic)
      - x(5) znamená 5 jakýchkoliv aminokyselin
      - x(3, 5) znamená 3 až 5 jakýchkoliv aminokyselin

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

## Specializované verze

- Příklad vyhledávání pomocí 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].
```

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

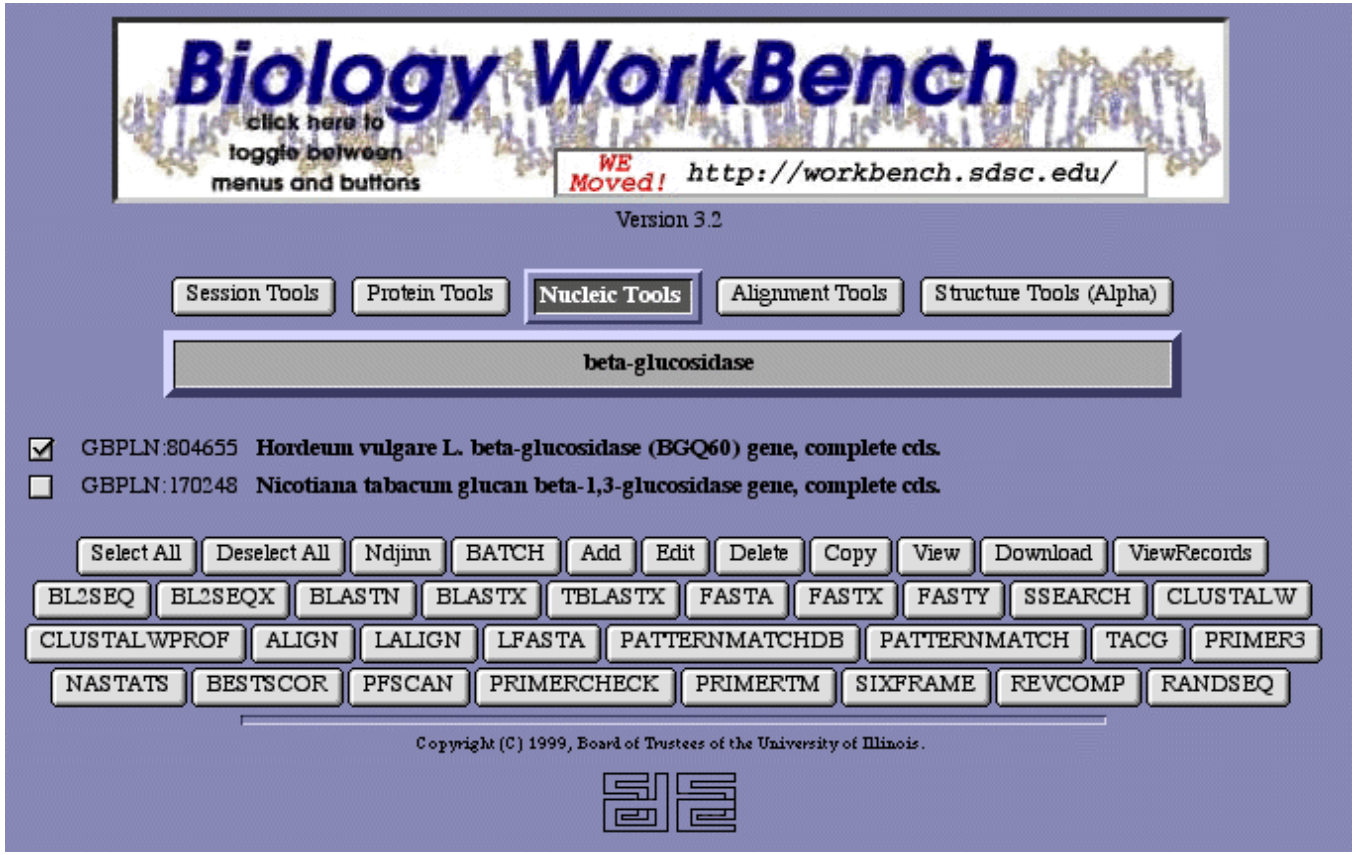
# Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst….

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

# Analytické nástroje

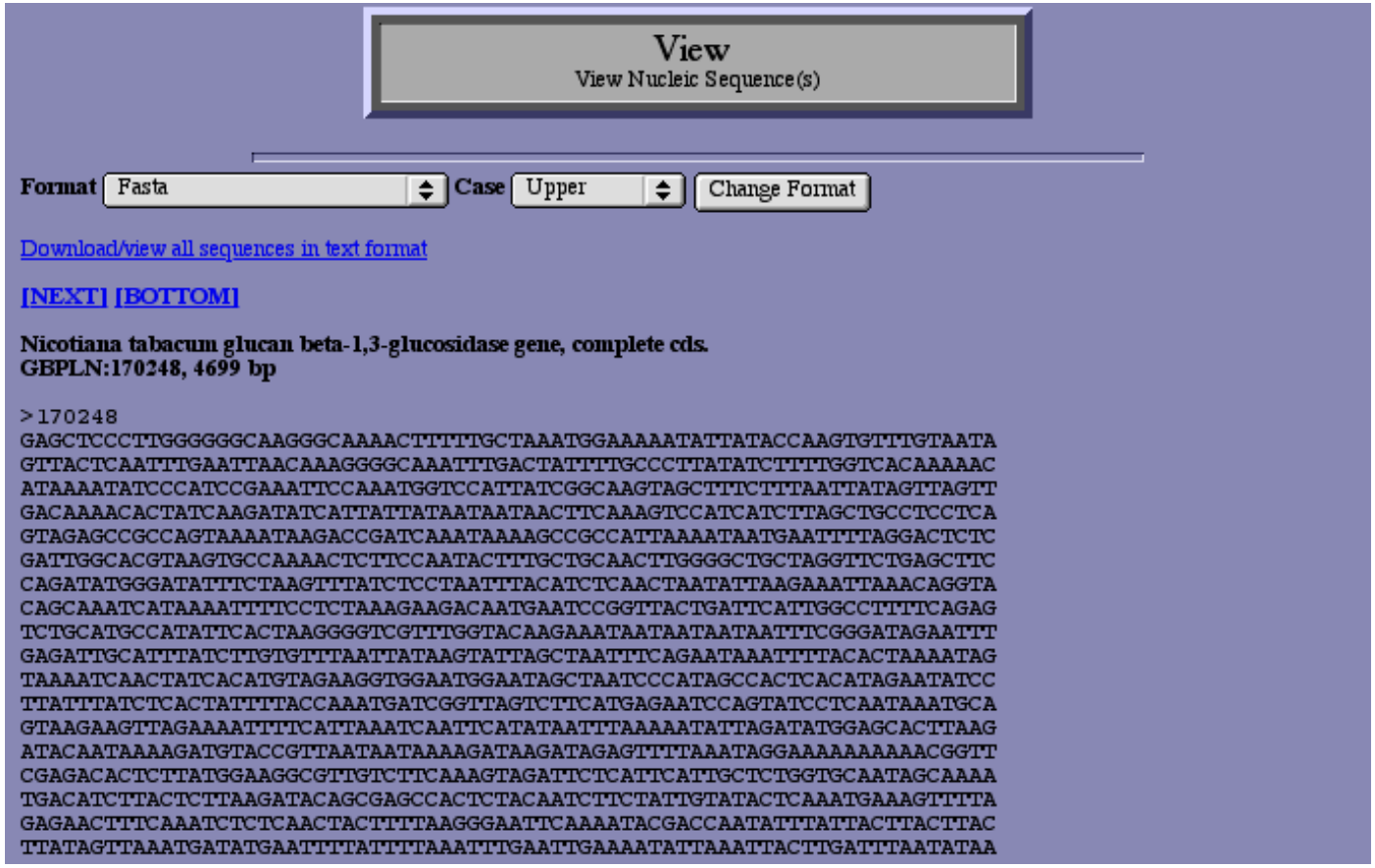
- http://workbench.sdsc.edu/




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

# Analytické nástroje

- http://workbench.sdsc.edu/



# Analytické nástroje

- http://workbench.sdsc.edu/

**Regex pattern:**

ctt.{1,32}ctt

**0 sequences were searched**

**1 match was found**

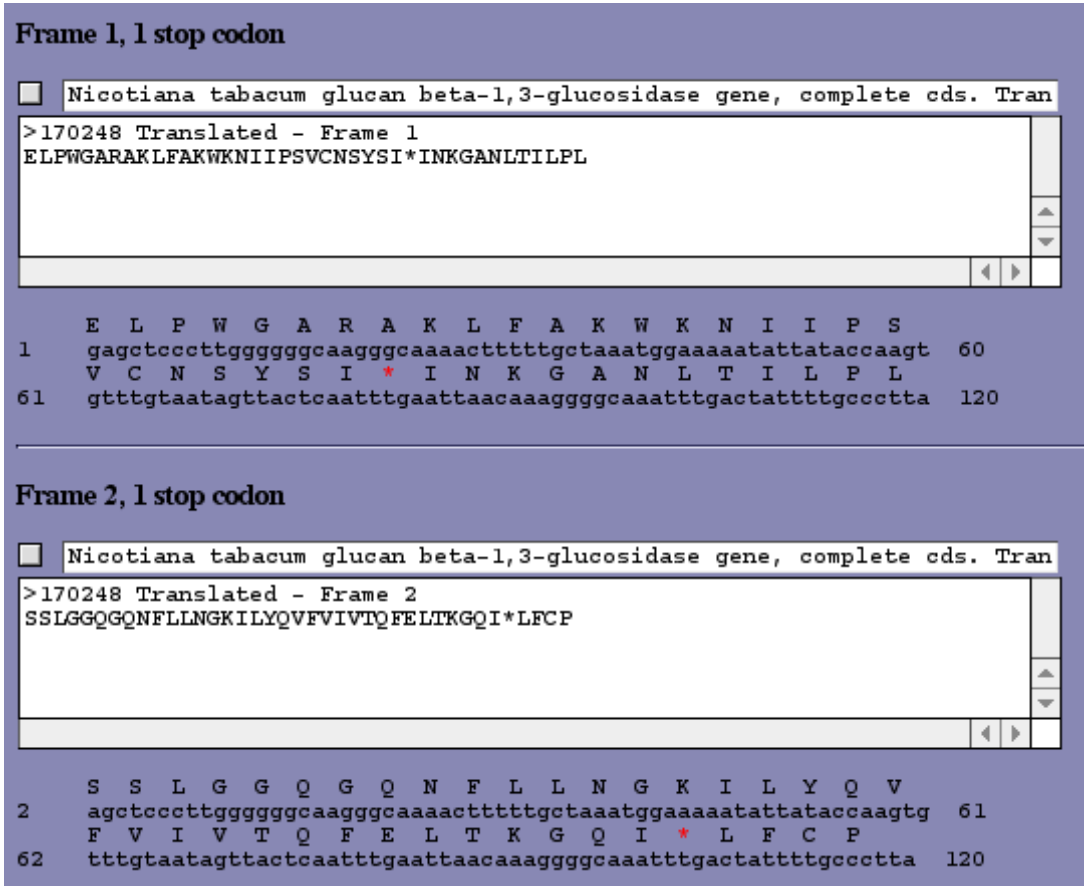
**Matches are indicated in blue**

---

>170248
GAGCTCCCTTGGGGGGCAAGGGCAAAACTTTTTGCTAAATGGAAAAATATTATACCAAGTGTTTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAATTTGACTATTTTGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTTCTTTAATTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAATAATAACTTCAAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAATAAGACCGATCAAATAAAAGCCGCCATTAAAATAATGAATTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACTCTTCCAATACTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTAAACAGGTA
CAGCAAATCATAAAATTTTCCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCATGCCATATTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAATAATTTCGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTAATTATAAGTATTAGCTAATTTCAGAATAAATTTTACACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACCAAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAATTTTCATTAAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTGTCTTCAAAGTAGATTCTCATTCATTGCTCTGGTGCAATAGCAAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA
GAGAACTTTCAAATCTCTCAACTACTTTTAAGGGAATTCAAAATACGACCAATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTATTTTAAATTTGAATTGAAAATATTAAATTACTTGATTTAATATAA
ACAATAGATATCGCTAAGTATTTACCACAAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT
GATTAAGCAGCTATTCATCTGGTTGTGCAGGATGAAAGAAAGTAACTAGCTATAATTTCTTTTGTAAAGT

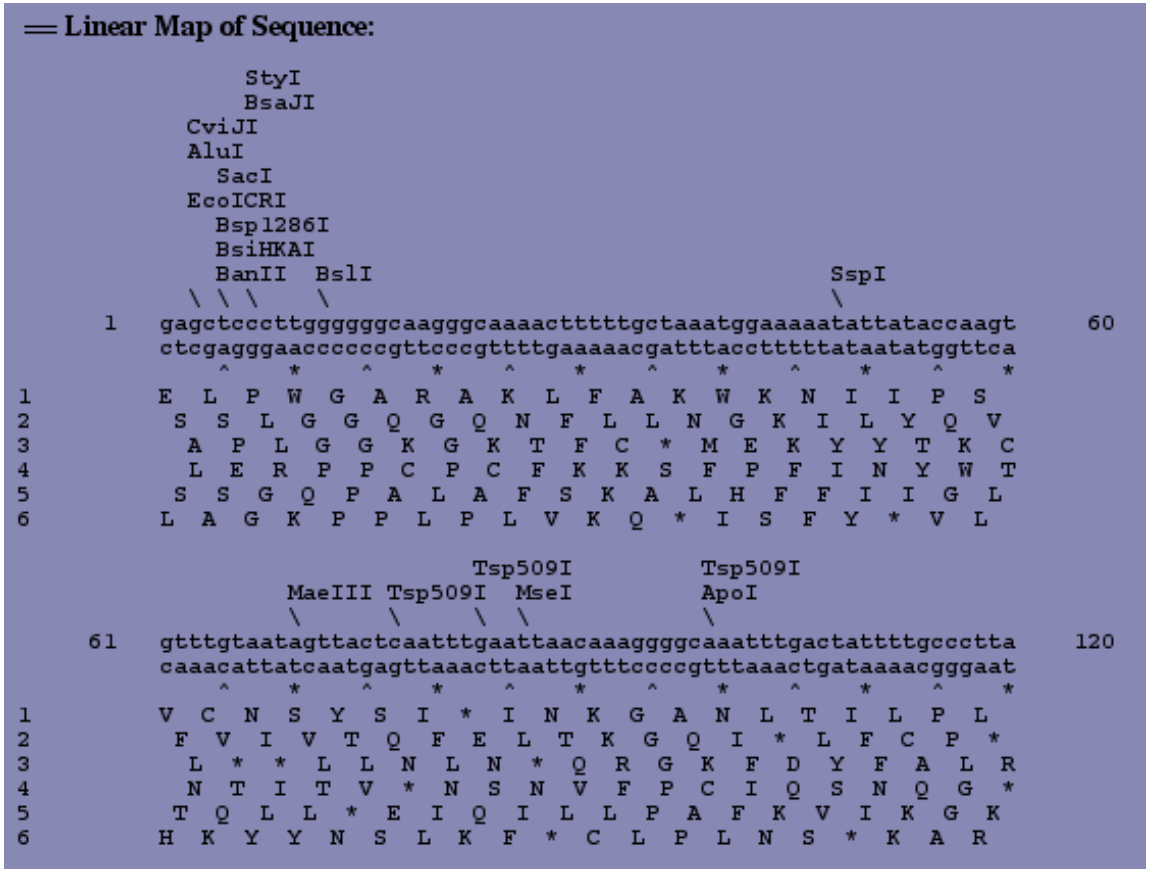
# Analytické nástroje

- http://workbench.sdsc.edu/



# Analytické nástroje

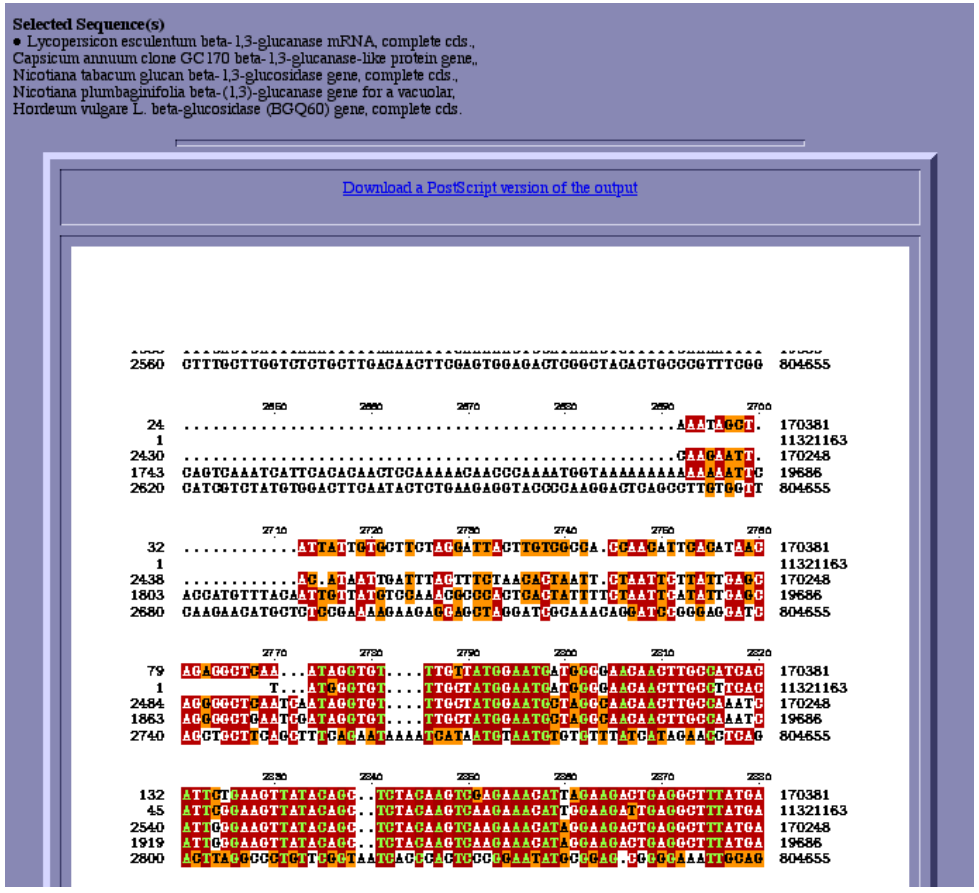
- http://workbench.sdsc.edu/




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

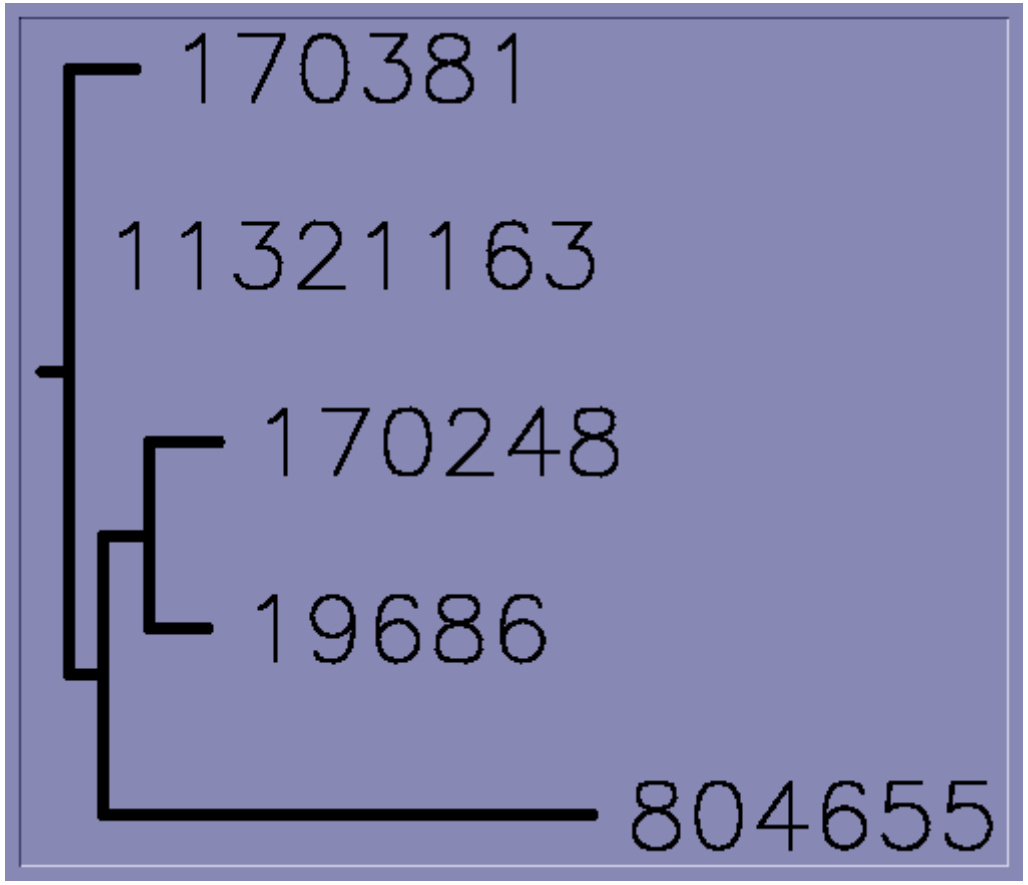
# Analytické nástroje

- http://workbench.sdsc.edu/



# Analytické nástroje

- http://workbench.sdsc.edu/



# Analytické nástroje

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


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

# Analytické nástroje

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

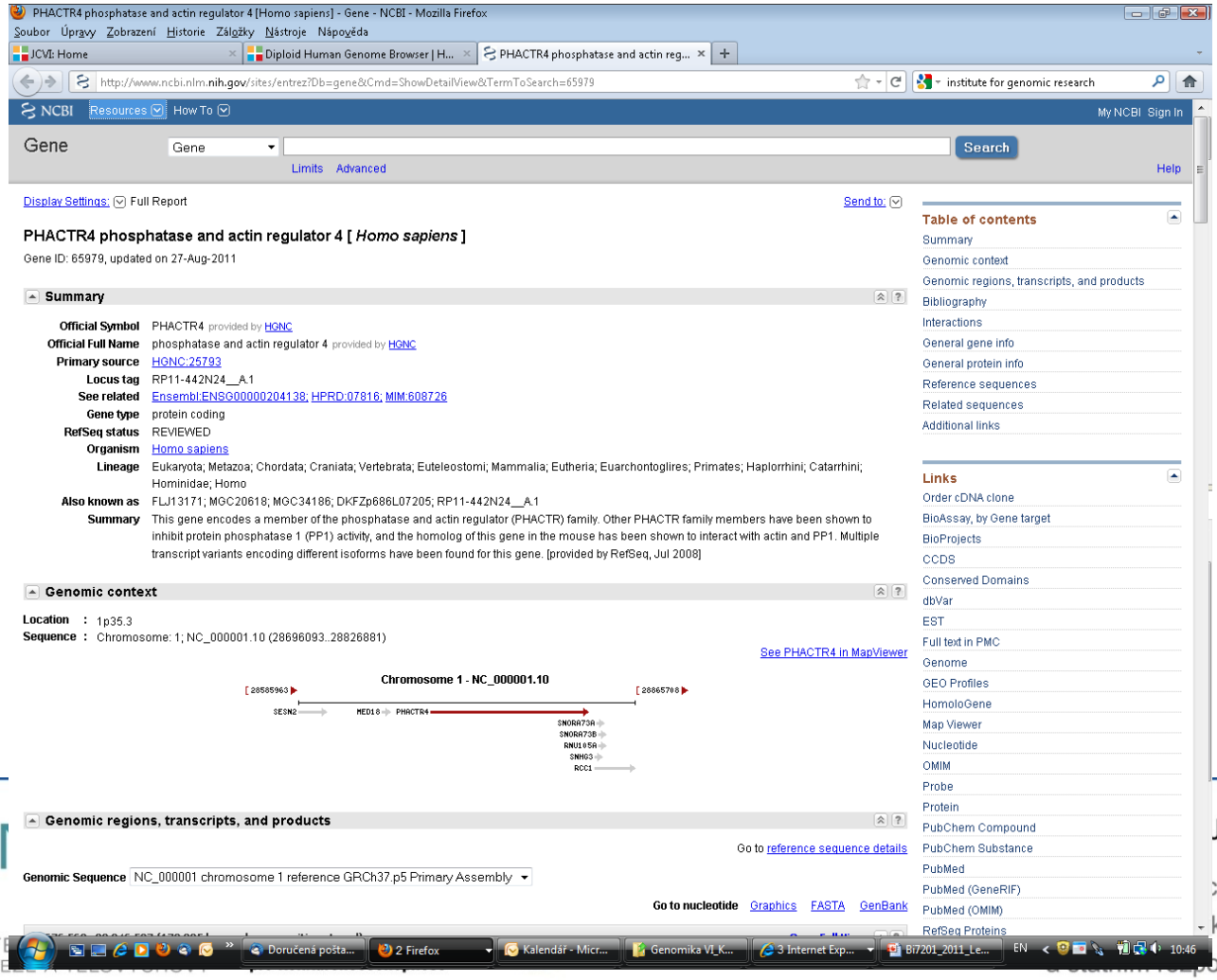
# Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst….
  - Další www genomové nástroje

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

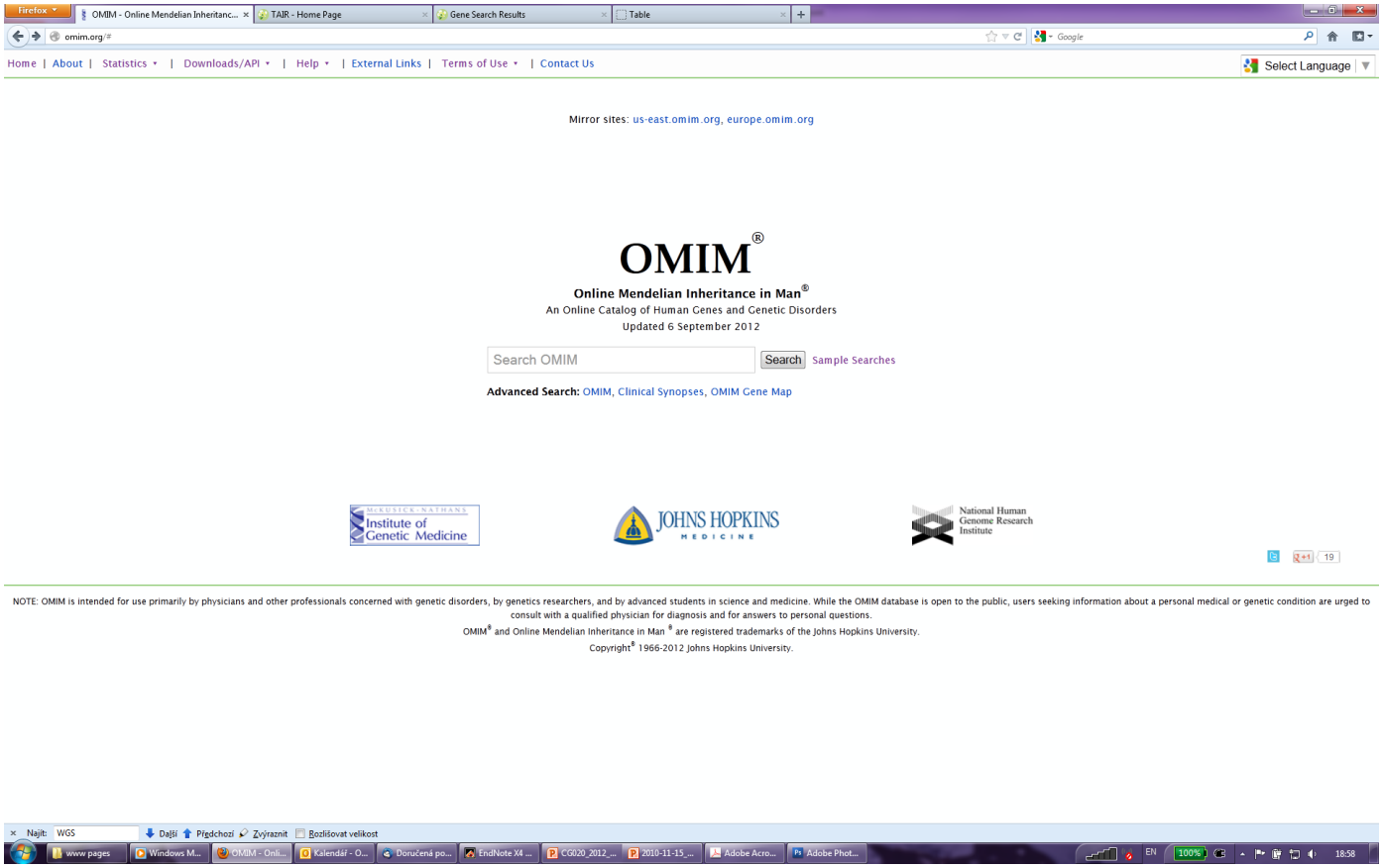
# Další WWW zdroje

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



# Další WWW zdroje

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




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

# Shrnutí

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst….
  - Další www genomové nástroje

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

# Diskuse

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