Základy genomiky

I. Úvod do bioinformatiky

Jan Hejátko

Masarykova univerzita, Oddělení funkční genomiky a proteomiky
Laboratoř molekulární fyziologie rostlin

Základy genomiky I.

- Zdrojová literatura ke kapitole I:
  - Plant Functional Genomics, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey
  - Úvod do praktické bioinformatiky, Fatima Cvrčková, 2006, Academia, Praha

Základy genomiky I.

- 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

Základy genomiky I.

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

GENOMIKA-co to je?

role BIOINFORMATIKY ve FUNKČNÍ GENOMICE

Přístupy „klasické“ genetiky

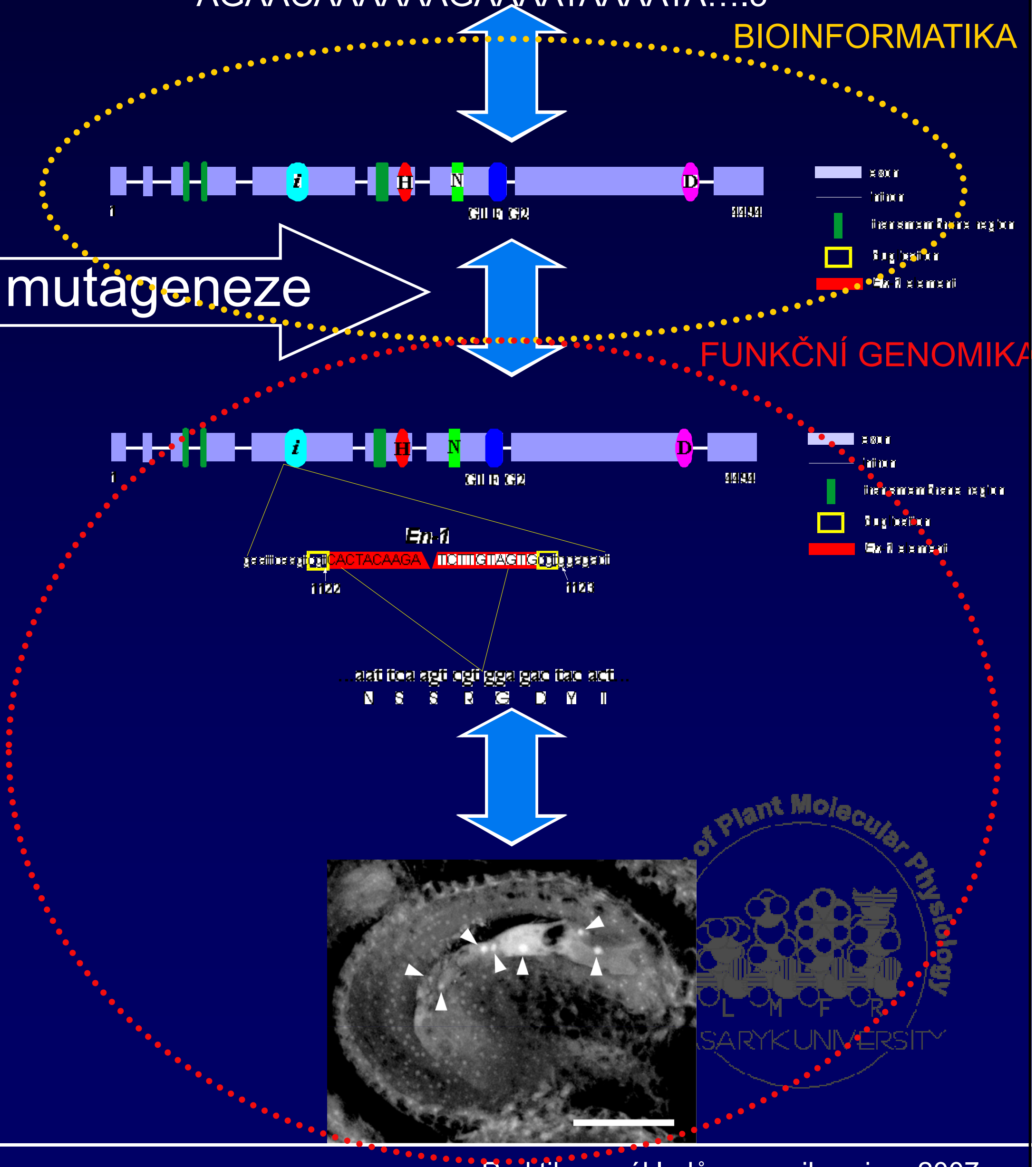
3 : 1

?

inzerční mutageneze

„Reverzně genetický“ přístup

5‘TTATATATATATATTAAAAAATAAAATAAA AGAACAAAAAAGAAAATAAAATA….3‘


BIOINFORMATIKA

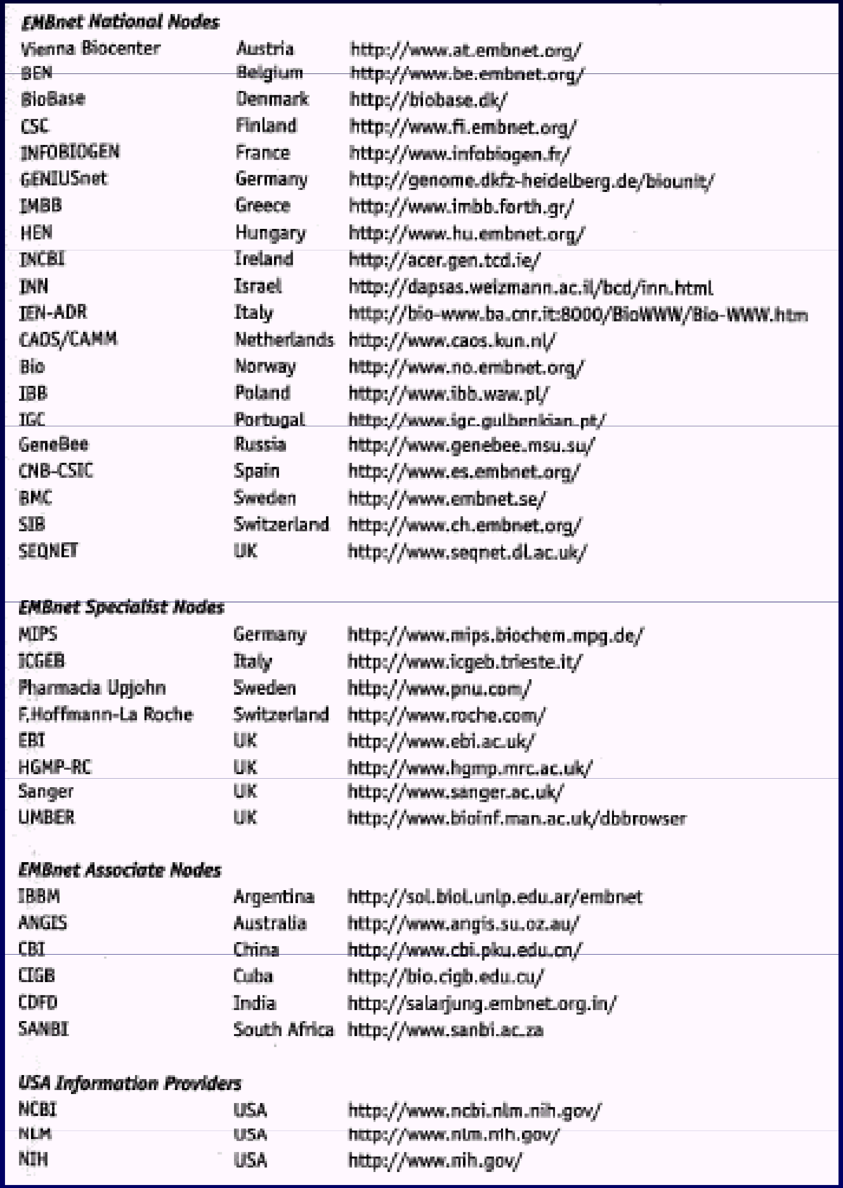
FUNKČNÍ GENOMIKA

Základy genomiky I.

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

Databáze

Spektrum on-line zdrojů

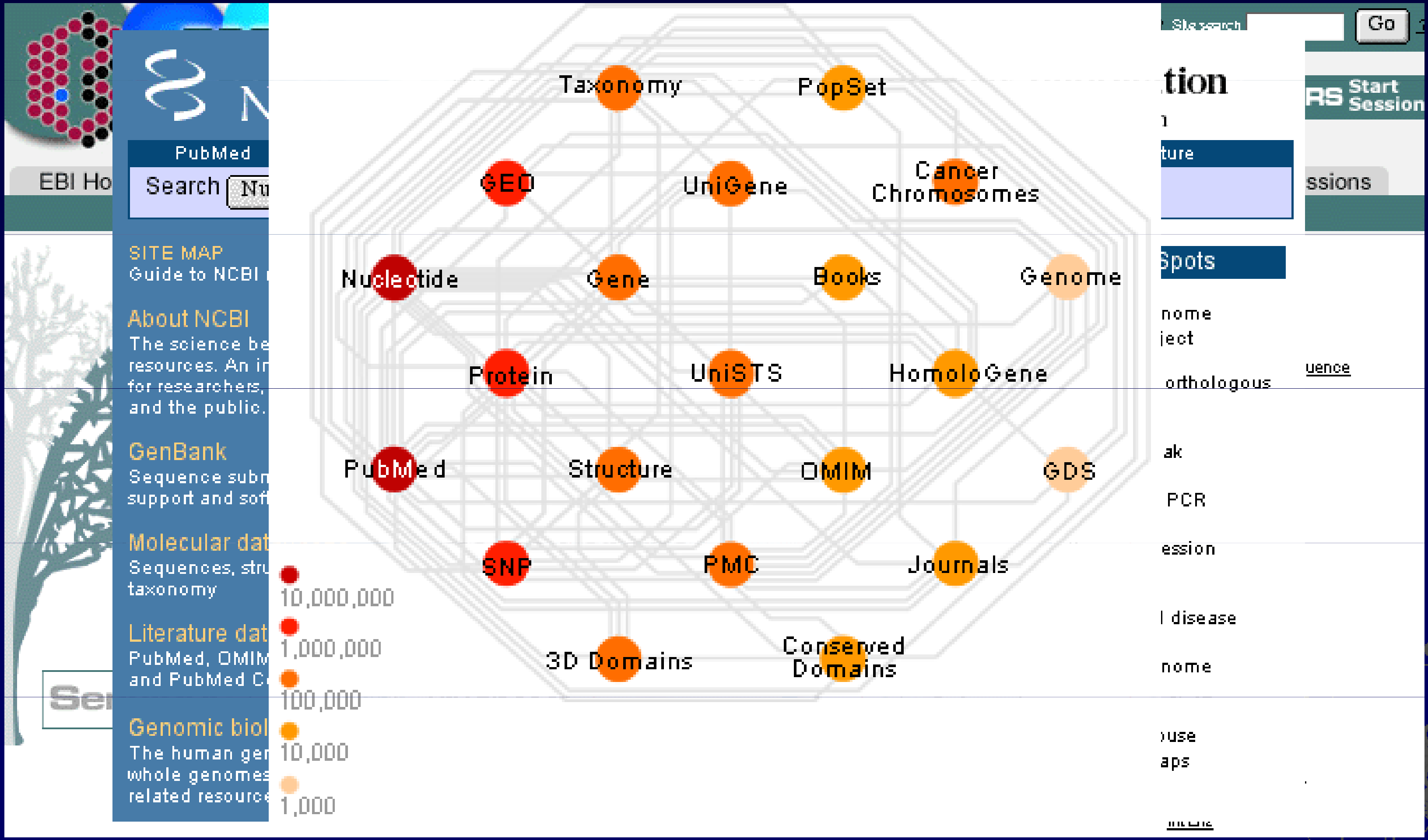


EMBnet National Nodes		
Vienna Biocenter	Austria	http://www.at.embnet.org/
BEN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENIUSnet	Germany	http://genome.dkfz-heidelberg.de/biounit/
IMBB	Greece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

Databáze

Spektrum on-line zdrojů

- EBI http://www.ebi.ac.uk/services
- NCBI http://www.ncbi.nlm.nih.gov/



Základy genomiky I.

- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY

- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze

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

Primární databáze

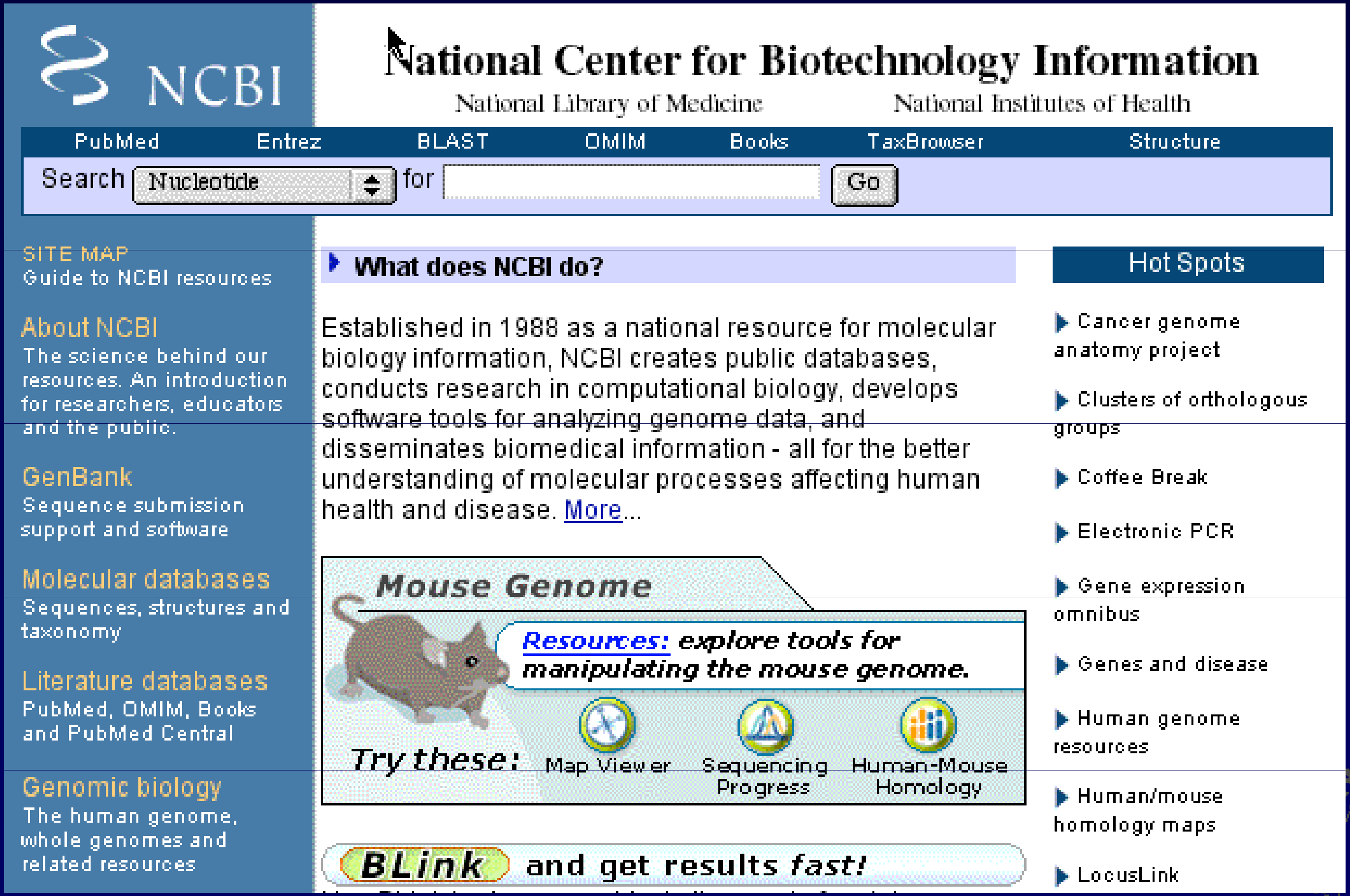
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - Proteinové sekvence:
    - PIR, http://pir.georgetown.edu/
    - MIPS, http://www.mips.biochem.mpg.de
    - SWISS-PROT, http://www.expasy.org/sprot/

Primární databáze

- Typy sekvencí v primárních databázích
  - standardní nukleotidové sekvence získané kvalitním sekvenováním
  - ESTs (Expressed Sequence Tags)
  - HGTS (High Throughput Genome Sequencing)
    - neanotované „surové“ výsledky sekvenačních projektů
  - referenční sekvence anotovaných genomů
  - TPAs (Third Party Annotation)
    - sekvence anotované jinými než původními autory

Primární databáze

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



NCBI

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

SITE MAP
Guide to NCBI resources

About NCBI
The science behind our resources. An introduction for researchers, educators and the public.

GenBank
Sequence submission support and software

Molecular databases
Sequences, structures and taxonomy

Literature databases
PubMed, OMIM, Books and PubMed Central

Genomic biology
The human genome, whole genomes and related resources

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Mouse Genome

Resources: explore tools for manipulating the mouse genome.

Try these: Map Viewer, Sequencing Progress, Human-Mouse Homology

BLink and get results *fast!*

Hot Spots

- Cancer genome anatomy project
- Clusters of orthologous groups
- Coffee Break
- Electronic PCR
- Gene expression omnibus
- Genes and disease
- Human genome resources
- Human/mouse homology maps
- LocusLink

Primární databáze

Přístupový kód

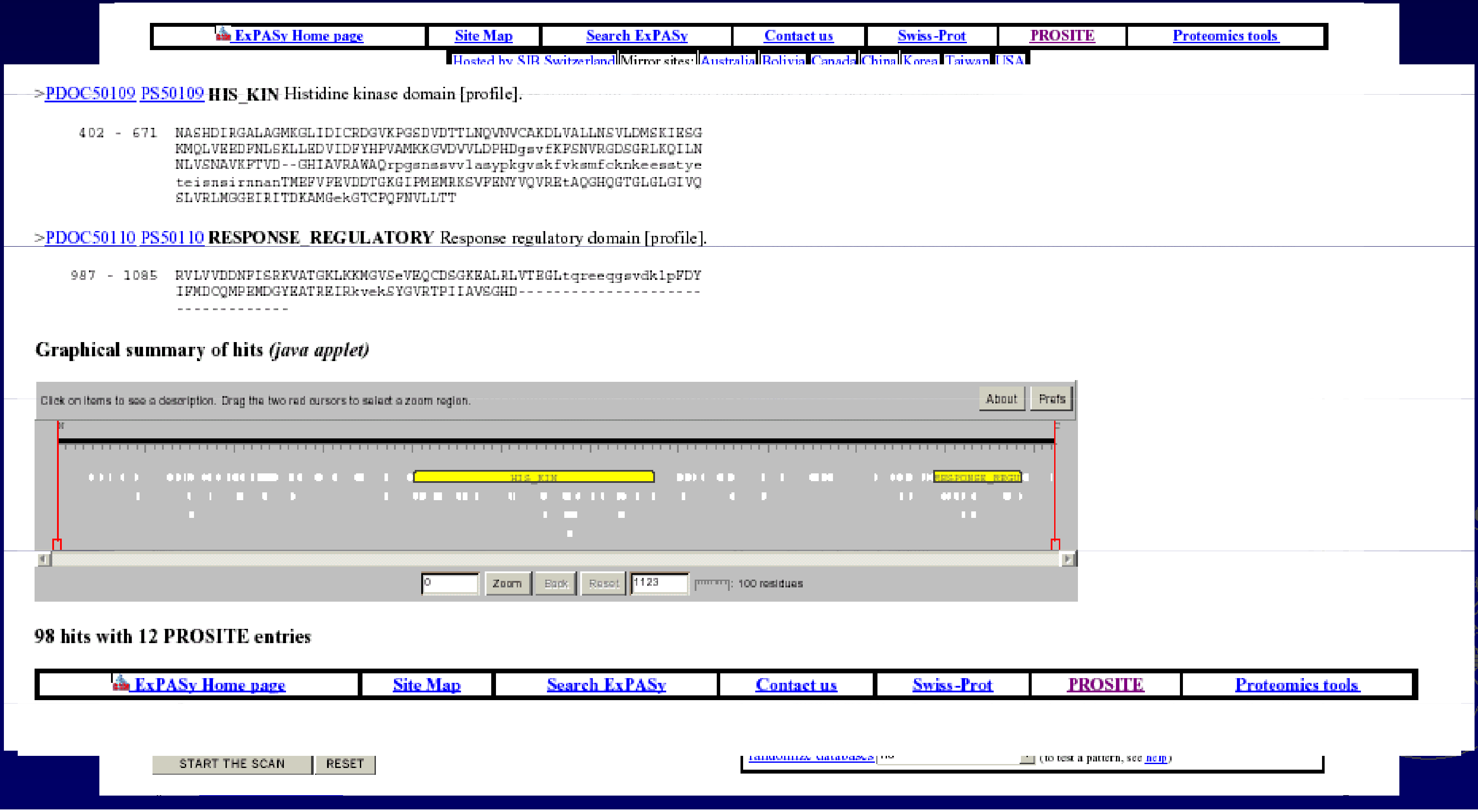
GeneBank Identifier

```
                     /translation="MNGRYSPTRQDFKTGAKPWSILALIVAAMIFAFMAVASWQDNAT
                     TQAILSQLRSINADSASLQRDVLRAHTGTVANYRPIISRLGALRKNLEDLKQLFRQSH
                     IVSESNAAQLLRQLEVSLNSADAAVAAFGAQNVRLQDSLASFTRALSSLPGKASTDQT
                     LEKPTELASMMLQFLRQPSPAISFEISLELERLQKQRGLDEAPVRILAREGPIILSLL
                     PQVKDLVNMIQTSDTAEIAEMLQRECLEVYSLKNVEERSARIFLGSASVGLCLYIITL
                     VYRLRKKTDWLARRLDYEELIKEIGVCFEGEAATTSSAQAALRIIQRFFDADTCALAL
                     [illegible] EIVSRTKADERATVFRIISSKKIVHLPLEIP
                     [illegible] EIQLLELATACLCHYIDVRRKQTECD
                     [illegible] LLGHAELAQNSVSRTSVTRRYIDYII
                     [illegible] ELVTEIAPLLRMALPPNIELSFRFDQMQ
                     [illegible] QAMTANGQIDIIISQAFLPVKKILAHGVMPPGDYVL
                     LSI[illegible]DNGGGIPEAVLPHIFEPFFTTRARNGGTGLGLASVHGHISAFAGYIDVSSTVGH
                     GTR[illegible]DIYLPPSSKEPVNPDSPFGRNKAPRGNGEIVALVEPDDLLREAYEDKIAALGYE
                     PVGFRTFNEIRDWISKGNEADLVMVDQASLPEDQSPNSVDLVLKTASIIIGGNDLKMT
                     LSREDVT[illegible]DLYLPKPIS[illegible]TMAHAILTKIKT"
ORIGIN
        1 atgaacggaa gatattcac[illegible] [illegible]ggcag gattttaaga caggcgcgaa gccttggtct
       61 [illegible] att ttcgcgttca tggcggttgc gtcctggcag
          [illegible] at cgattaacgc cgacagcgcc
          [illegible] ggcgaacta ccgccccatt
          [illegible] gaagcaatt atttagacaa
          [illegible] gctac gccagctaga agtgtctcta
      361 [illegible] cgccgccttt ggtgcgcaaa atgtacgcct gcaagattcg
      421 ctggccagtt tcactcgtgc tttgagcagt cttccaggaa aagcctcaac cgatcagact
      481 ttagaaaaac caacagaatt ggctagcatg atgctccaat ttcttcggca accaagcccg
      541 gctatttcat tcgagatcag ccttgaacta gagaggctcc aaaaacaacg cggtcttgat
      601 gaagctcccg tgcgcatact tgcacgtgaa ggtcccatta tcttatcgct tttgccacag
      661 gtgaaagatc tggtgaacat gattcagacg tctgacaccg cagaaattgc ggagatgctg
      721 cagcgcgagt gtttggaggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc
      781 tttcttgggt ccgcttcagt gggtctttgc ctctacatca tcaccttagt ctataggcta
      841 cgcaaaaaaa ccgattggtt agcgcggcgt ttagattacg aagagctaat caaagagatc
      901 ggagtatgtt ttgaaggtga ggcggccacc acgtcgtccg cgcaagctgc acttcgtatt
      961 attcagcgct tctttgatgc cgatacgtgc gcgttagctc tagtggacca tgaccgtaga
     1021 tgggctgtcg aaacattcgg tgcgaaacac ccaaaacctg tgtgggacga cagcgtgcta
     1081 cgcgaaatag tctctcgtac caaagcggac gaacgggcga cggtattccg catcatatcg
     1141 tcgaaaaaaa tcgtacattt gcctctcgaa attccaggtc tctcgatact actggctcac
     1201 aaatccacag ataaactaat tgcggtttgt tcactgggtt accaaagcta tcgccctcga
     1261 ccttgccaag gcgaaattca gcttcttgaa ctcgccaccg cctgcctctg tcactatatc
     1321 gatgttcggc gtaagcagac cgaatgcgac gttttggcca gacgattgga gcatgcgcaa
     1381 cgccttgagg cagttggtac acttgccggc ggaatagcac atgaatttaa taacattttg
     1441 ggctcaatcc tcgggcacgc agaattagca caaaactcgg tgtctcgaac atctgtcacc
     1501 cgaagatata ttgactatat catttcgtca ggcgacagag ccatgctcat tatcgatcag
     1561 atcttgacgc tgagccgaaa acaggagcgc atgatcaagc catttagtgt ctcagagctt
     1621 gtgaccgaaa tcgctccctt gctacgtatg gctcttccgc caaacatcga gcttagtttc
     1681 agatttgatc aaatgcagag cgtgatcgaa ggaagcccgc ttgaacttca acaggtacta
     1741 attaacatct gcaagaatgc ttcccaagcc atgactgcaa atggtcaaat cgacatcatc
     1801 atcagccaag ctttttttacc agttaagaaa attctggcgc atggtgttat gccacctggc
     1861 gactatgttc tcctatctat tagcgacaat ggtggaggca ttcccgaggc tgtgttaccc
     1921 cacattttttg aaccctcttt tacgacacga gctcgcaacg gtggaacggg tctcggcctt
     1981 gcttctgtgc atggtcatat cagcgcgttt gcgggttaca tcgacgttag ttcaactgtt
     2041 gggcatggga cgcgctttga catttatctc cctccgtctt ctaaggaacc cgtaaatcca
     2101 gacagttttt tcggccgcaa taaggcaccg cgtggaaacg gggagattgt ggcacttgtt
     2161 gagcccgatg acctcctgcg ggaggcgtat gaagacaaga tcgccgctct aggatatgag
     2221 ccggtcggtt ttcgtacctt taatgaaatt cgcgattgga tttcaaaagg caatgaagcc
     2281 gatctggtca tggtcgacca agcgtctctt cctgaagatc aaagtcctaa ttccgtggat
     2341 ttagtgctca agaccgcctc catcatcatt ggcggaaatg atctcaaaat gaccctttca
```

Proteinové sekundární databáze

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

- ☐ PROSITE, http://www.expasy.org/prosite/
- ☐ PRINTS, http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/



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

```
  402 - 671   NASHDIRGALAGMKGLIDICRDGVKPGSDVDTTLNQVNVCAKDLVALLNSVLDMSKIESG
              KMQLVEEDFNLSKLLEDVIDFYHPVAMKKGVDVVLDPHDgsvfKFSNVRGDSGRLKQILN
              NLVSNAVKFTVD--GHIAVRAWAQrpgsnssvvlasypkgvskfvksmfcknkeesstye
              teisnsirnnanTMEFVFEVDDTGKGIPMEMRKSVFENYVQVREtAQGHQGTGLGLGIVQ
              SLVRLMGGEIRITDKAMGekGTCFQFNVLLTT
```

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

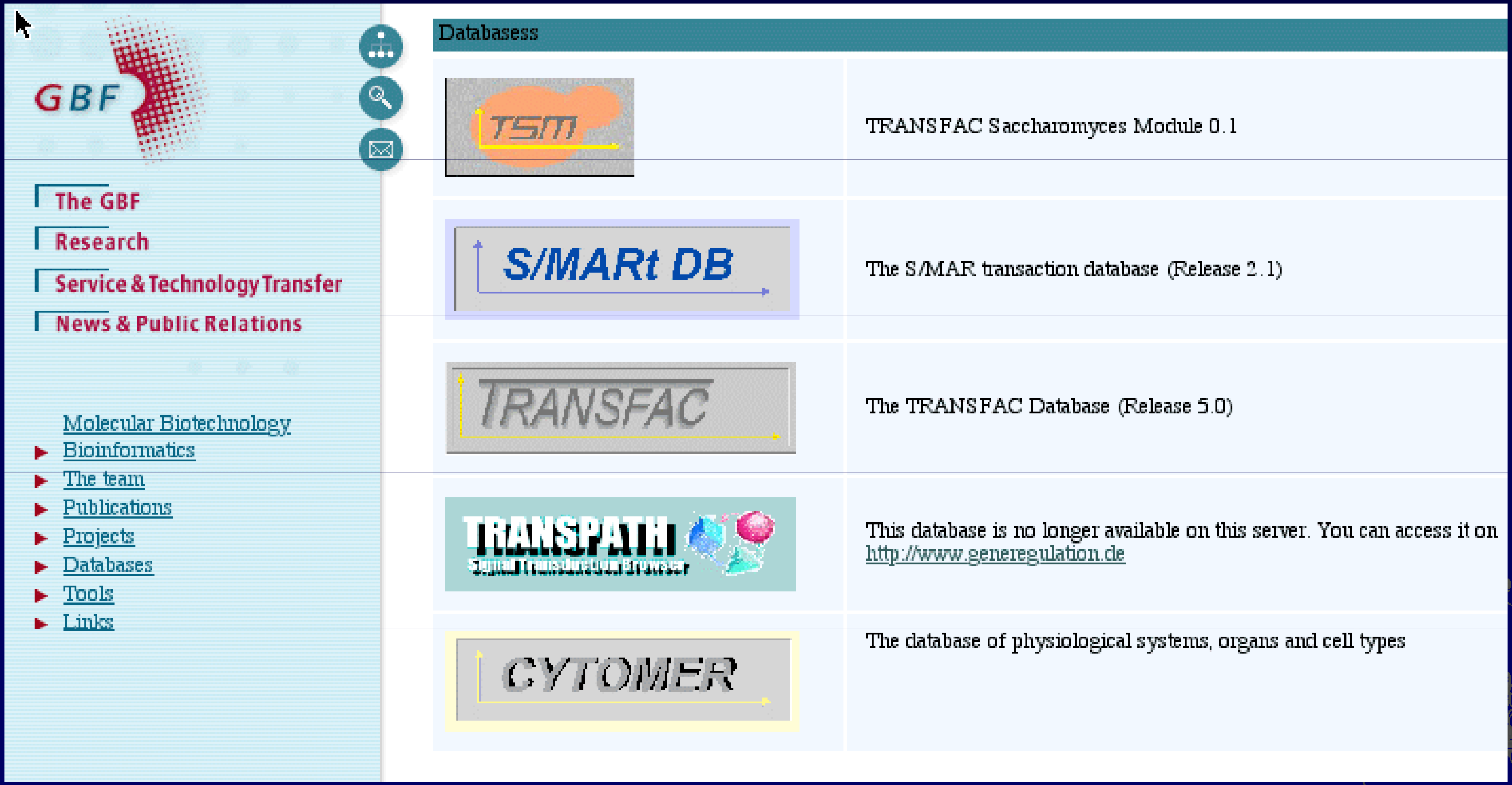
```
 987 - 1085   RVLVVDDNFISRKVATGKLKKMGVSeVEQCDSGKEALRLVTEGLtqreeqgsvdklpFDY
              IFMDCQMPEMDGYEATREIRkvekSYGVRTPIIAVSGHD---------------------
              -------------
```

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

98 hits with 12 PROSITE entries

Sekundární databáze DNA

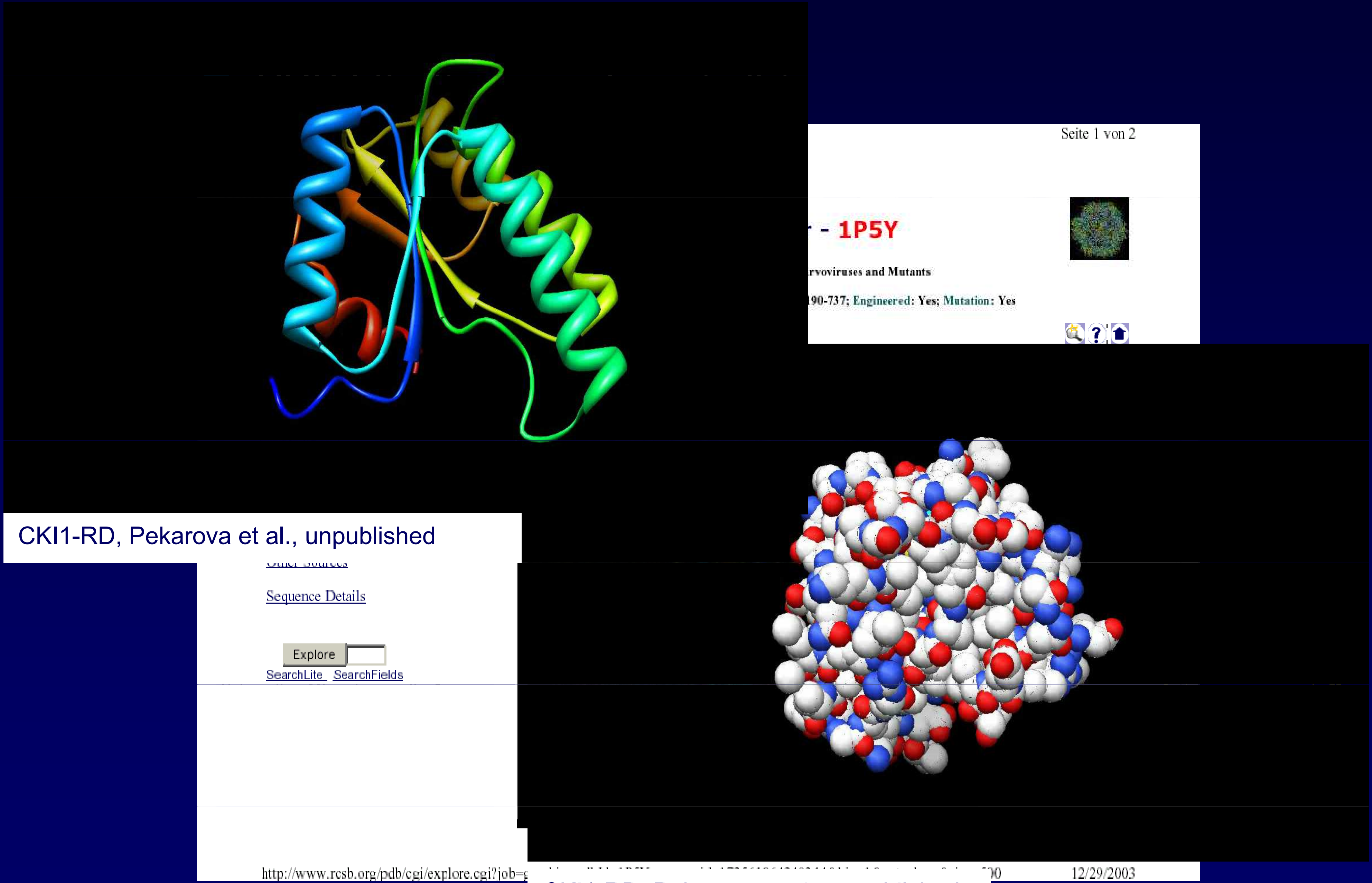
- TRANSFAC http://www.gene-regulation.com/



Databasess

TSM	TRANSFAC Saccharomyces Module 0.1
S/MARt DB	The S/MAR transaction database (Release 2.1)
TRANSFAC	The TRANSFAC Database (Release 5.0)
TRANSPATH	This database is no longer available on this server. You can access it on http://www.generegulation.de
CYTOMER	The database of physiological systems, organs and cell types

Strukturální databáze



Seite 1 von 2

- 1P5Y

rvoviruses and Mutants

190-737; Engineered: Yes; Mutation: Yes

http://www.rcsb.org/pdb/cgi/explore.cgi?job=g

12/29/2003

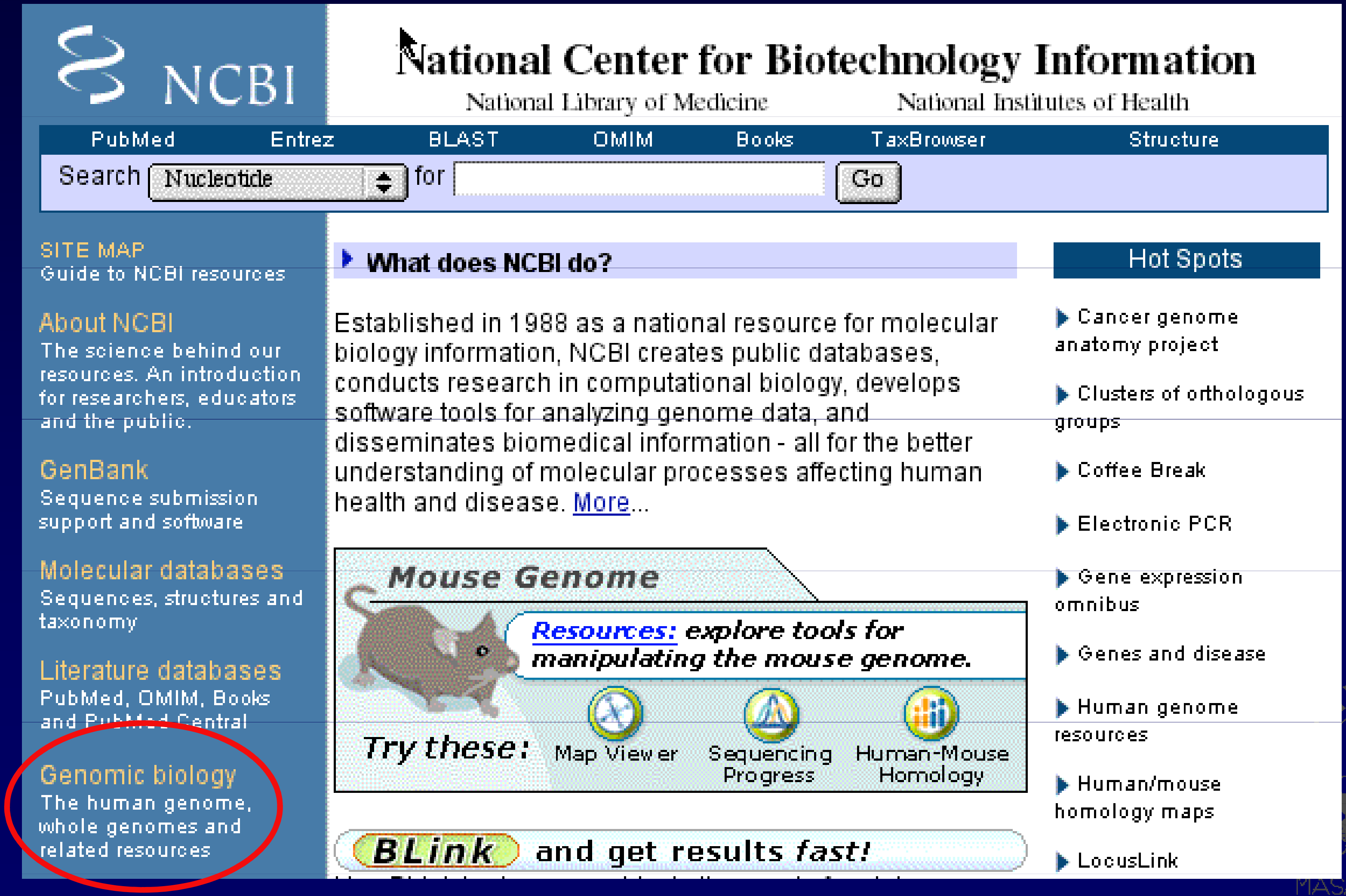
CKI1-RD, Pekarova et al., unpublished

CKI1-RD, Pekarova et al., unpublished

Základy genomiky I.

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

Primární data-genomové zdroje



National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

SITE MAP
Guide to NCBI resources

About NCBI
The science behind our resources. An introduction for researchers, educators and the public.

GenBank
Sequence submission support and software

Molecular databases
Sequences, structures and taxonomy

Literature databases
PubMed, OMIM, Books and PubMed Central

Genomic biology
The human genome, whole genomes and related resources

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Mouse Genome

Resources: explore tools for manipulating the mouse genome.

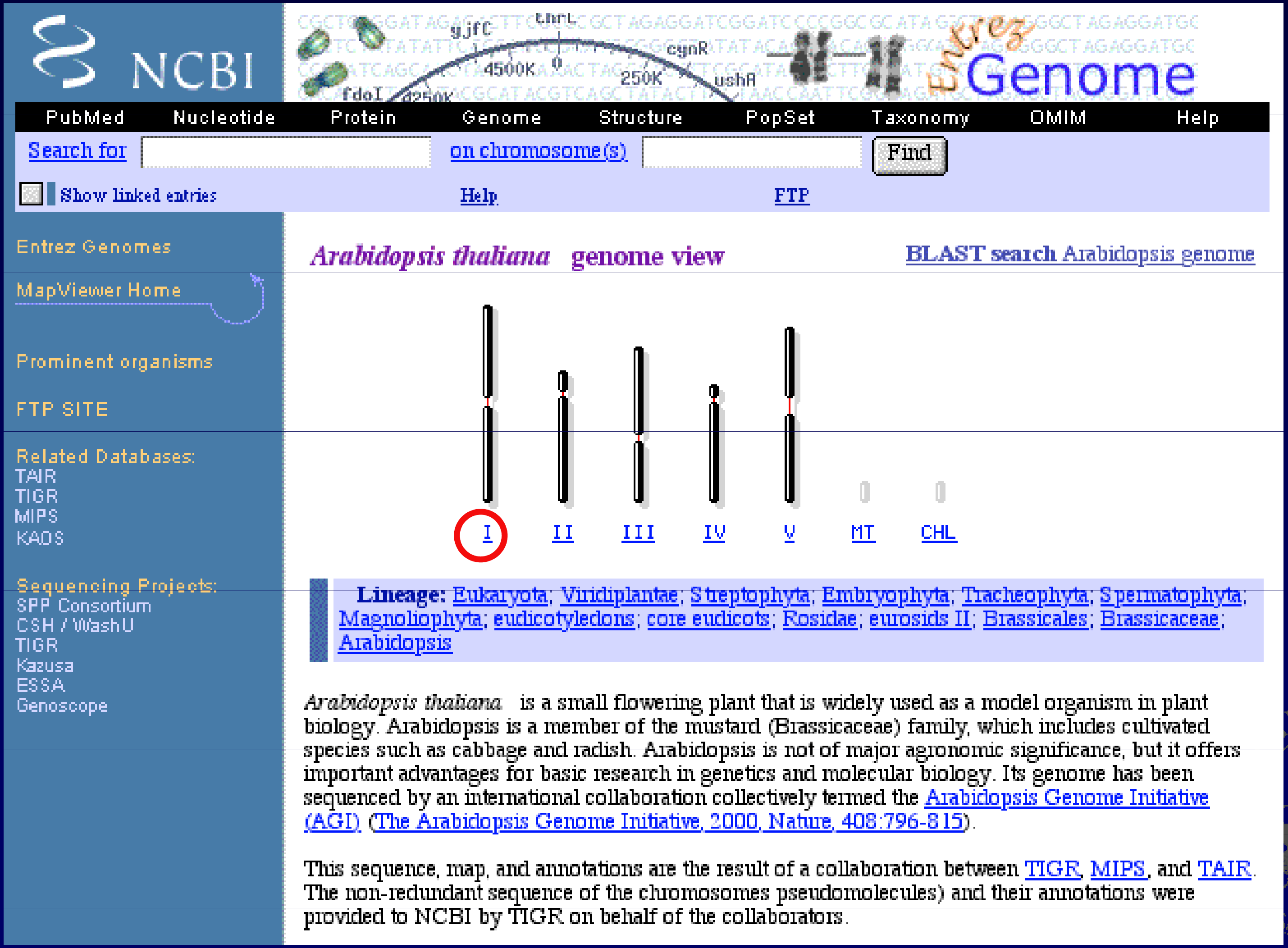
Try these: Map Viewer, Sequencing Progress, Human-Mouse Homology

BLink and get results *fast!*

Hot Spots

- Cancer genome anatomy project
- Clusters of orthologous groups
- Coffee Break
- Electronic PCR
- Gene expression omnibus
- Genes and disease
- Human genome resources
- Human/mouse homology maps
- LocusLink

Primární data-genomové zdroje



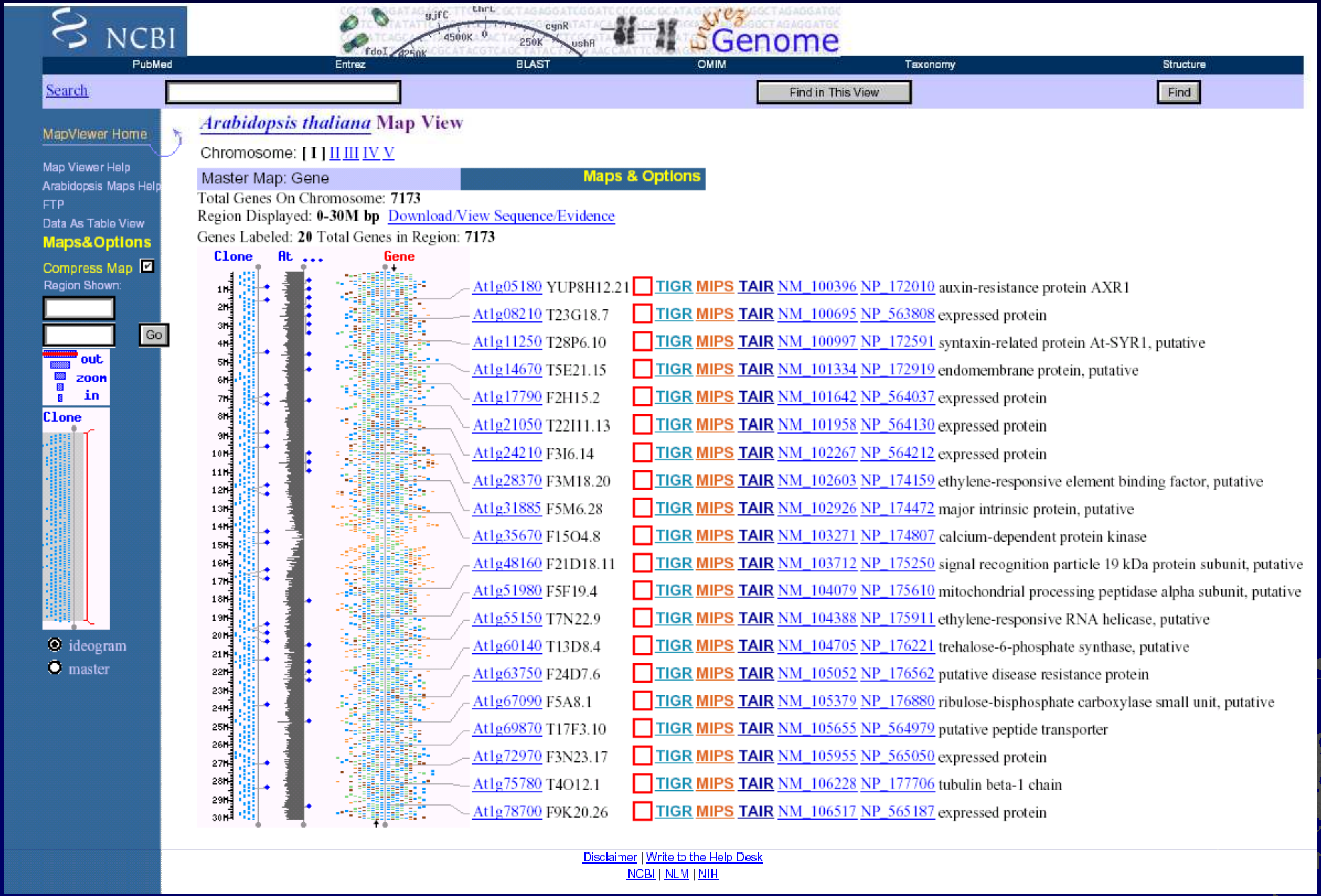
Arabidopsis thaliana genome view

Lineage: Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis

Arabidopsis thaliana is a small flowering plant that is widely used as a model organism in plant biology. Arabidopsis is a member of the mustard (Brassicaceae) family, which includes cultivated species such as cabbage and radish. Arabidopsis is not of major agronomic significance, but it offers important advantages for basic research in genetics and molecular biology. Its genome has been sequenced by an international collaboration collectively termed the Arabidopsis Genome Initiative (AGI) (The Arabidopsis Genome Initiative, 2000, Nature, 408:796-815).

This sequence, map, and annotations are the result of a collaboration between TIGR, MIPS, and TAIR. The non-redundant sequence of the chromosomes pseudomolecules) and their annotations were provided to NCBI by TIGR on behalf of the collaborators.

Primární data-genomové zdroje



***Arabidopsis thaliana* Map View**

Chromosome: [I] II III IV V

Master Map: Gene

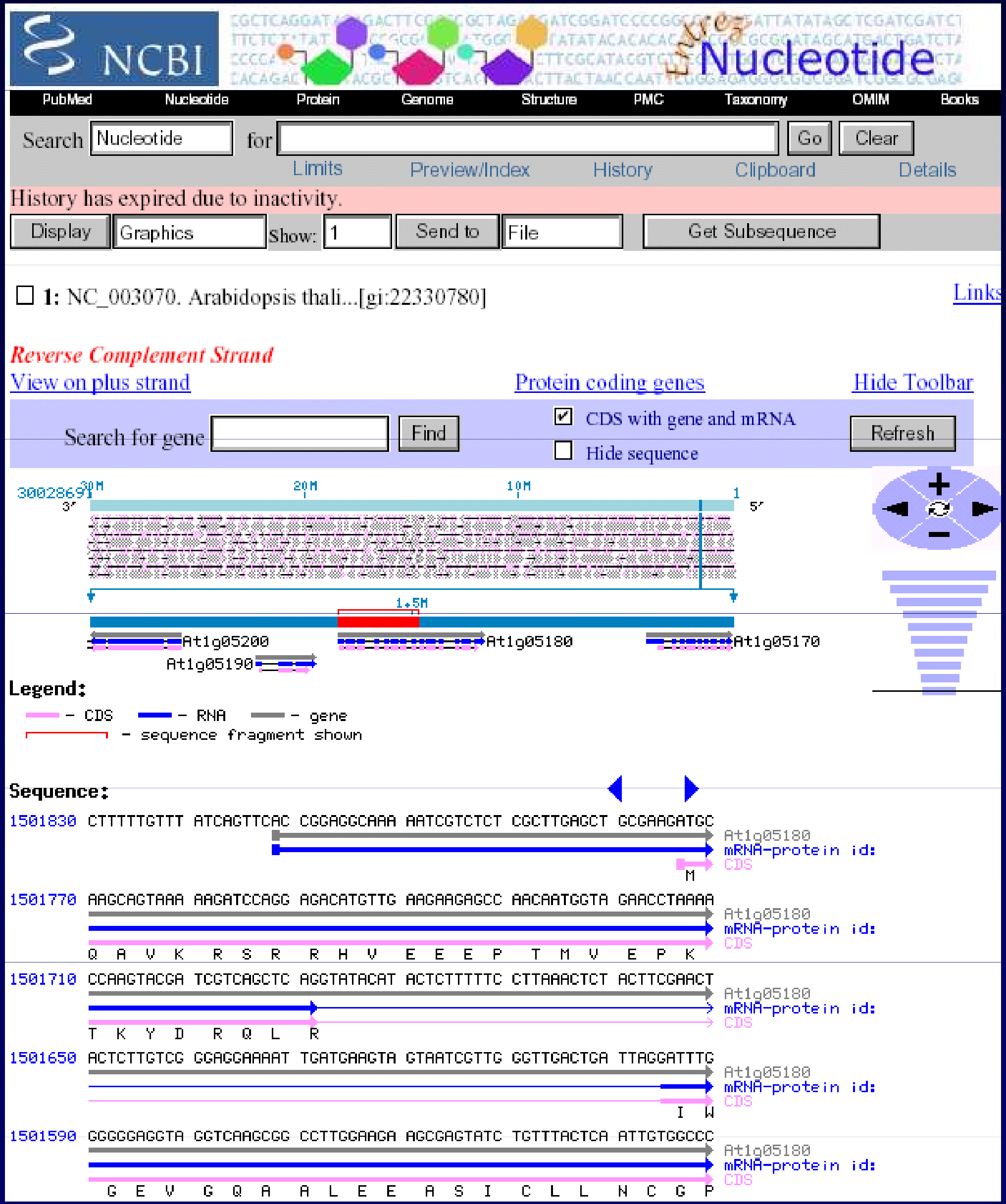
Total Genes On Chromosome: **7173**

Region Displayed: **0-30M bp** Download/View Sequence/Evidence

Genes Labeled: **20** Total Genes in Region: **7173**

At1g05180	YUP8H12.21	☐	TIGR	MIPS	TAIR	NM_100396	NP_172010	auxin-resistance protein AXR1
At1g08210	T23G18.7	☐	TIGR	MIPS	TAIR	NM_100695	NP_563808	expressed protein
At1g11250	T28P6.10	☐	TIGR	MIPS	TAIR	NM_100997	NP_172591	syntaxin-related protein At-SYR1, putative
At1g14670	T5E21.15	☐	TIGR	MIPS	TAIR	NM_101334	NP_172919	endomembrane protein, putative
At1g17790	F2H15.2	☐	TIGR	MIPS	TAIR	NM_101642	NP_564037	expressed protein
At1g21050	T22H1.13	☐	TIGR	MIPS	TAIR	NM_101958	NP_564130	expressed protein
At1g24210	F3I6.14	☐	TIGR	MIPS	TAIR	NM_102267	NP_564212	expressed protein
At1g28370	F3M18.20	☐	TIGR	MIPS	TAIR	NM_102603	NP_174159	ethylene-responsive element binding factor, putative
At1g31885	F5M6.28	☐	TIGR	MIPS	TAIR	NM_102926	NP_174472	major intrinsic protein, putative
At1g35670	F15O4.8	☐	TIGR	MIPS	TAIR	NM_103271	NP_174807	calcium-dependent protein kinase
At1g48160	F21D18.11	☐	TIGR	MIPS	TAIR	NM_103712	NP_175250	signal recognition particle 19 kDa protein subunit, putative
At1g51980	F5F19.4	☐	TIGR	MIPS	TAIR	NM_104079	NP_175610	mitochondrial processing peptidase alpha subunit, putative
At1g55150	T7N22.9	☐	TIGR	MIPS	TAIR	NM_104388	NP_175911	ethylene-responsive RNA helicase, putative
At1g60140	T13D8.4	☐	TIGR	MIPS	TAIR	NM_104705	NP_176221	trehalose-6-phosphate synthase, putative
At1g63750	F24D7.6	☐	TIGR	MIPS	TAIR	NM_105052	NP_176562	putative disease resistance protein
At1g67090	F5A8.1	☐	TIGR	MIPS	TAIR	NM_105379	NP_176880	ribulose-bisphosphate carboxylase small unit, putative
At1g69870	T17F3.10	☐	TIGR	MIPS	TAIR	NM_105655	NP_564979	putative peptide transporter
At1g72970	F3N23.17	☐	TIGR	MIPS	TAIR	NM_105955	NP_565050	expressed protein
At1g75780	T4O12.1	☐	TIGR	MIPS	TAIR	NM_106228	NP_177706	tubulin beta-1 chain
At1g78700	F9K20.26	☐	TIGR	MIPS	TAIR	NM_106517	NP_565187	expressed protein

Primární data-genomové zdroje



History has expired due to inactivity.

☐ **1:** NC_003070. Arabidopsis thali...[gi:22330780]

Reverse Complement Strand

View on plus strand

Protein coding genes

Hide Toolbar

Search for gene

☑ CDS with gene and mRNA

☐ Hide sequence

Legend:

▬ - CDS ▬ - RNA ▬ - gene

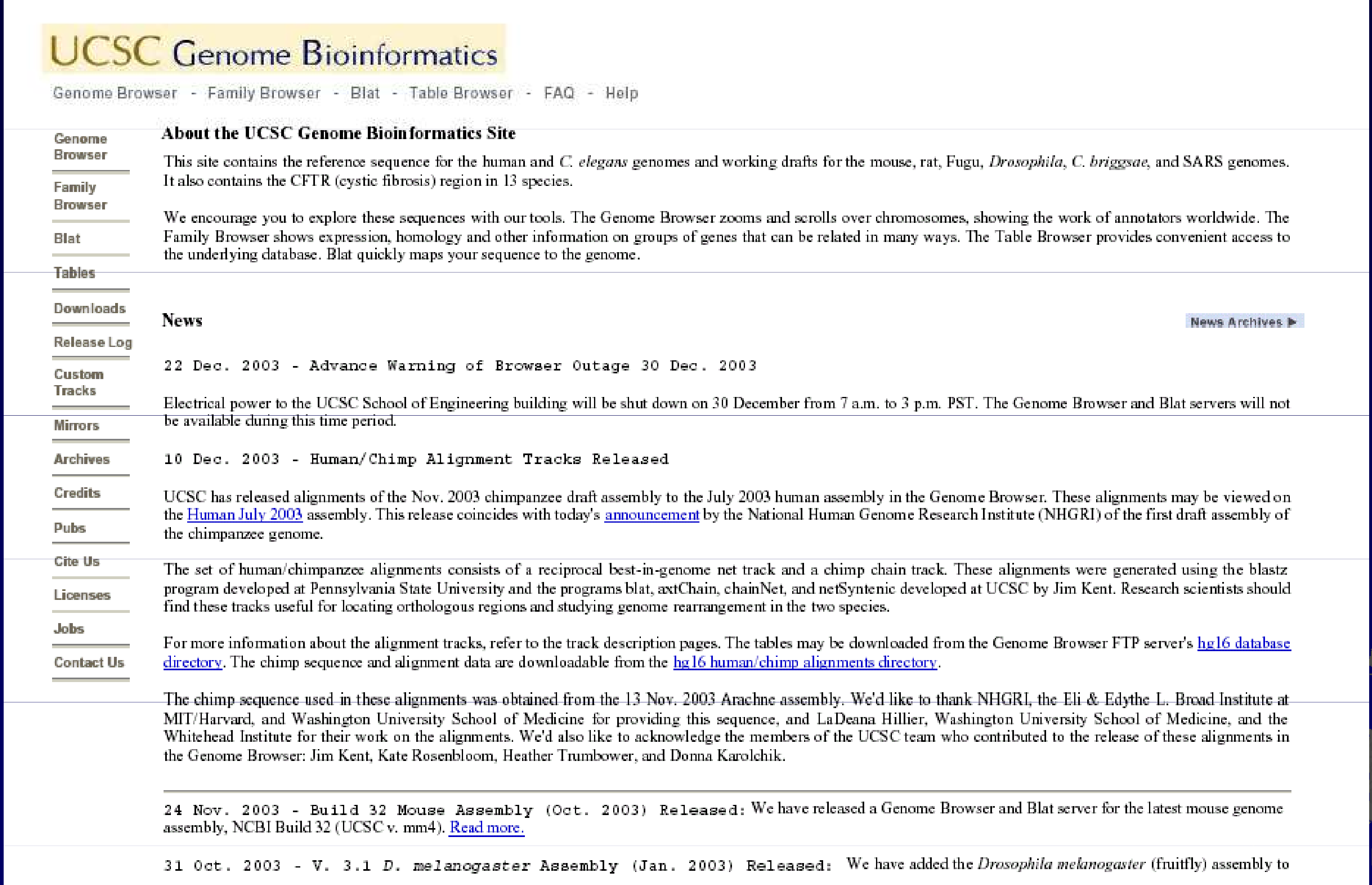
⌐¬ - sequence fragment shown

Sequence:

```
1501830 CTTTTTGTTT ATCAGTTCAC CGGAGGCAAA AATCGTCTCT CGCTTGAGCT GCGAAGATGC
                                                                   At1g05180
                                                                   mRNA-protein id:
                                                                   CDS
                                                                 M
1501770 AAGCAGTAAA AAGATCCAGG AGACATGTTG AAGAAGAGCC AACAATGGTA GAACCTAAAA
                                                                   At1g05180
                                                                   mRNA-protein id:
                                                                   CDS
        Q  A  V  K   R  S  R   R  H  V   E  E  E  P   T  M  V   E  P  K
1501710 CCAAGTACGA TCGTCAGCTC AGGTATACAT ACTCTTTTTC CTTAAACTCT ACTTCGAACT
                                                                   At1g05180
                                                                   mRNA-protein id:
                                                                   CDS
        T  K  Y  D   R  Q  L   R
1501650 ACTCTTGTCG GGAGGAAAAT TGATGAAGTA GTAATCGTTG GGTTGACTGA TTAGGATTTG
                                                                   At1g05180
                                                                   mRNA-protein id:
                                                                   CDS
                                                                I  W
1501590 GGGGGAGGTA GGTCAAGCGG CCTTGGAAGA AGCGAGTATC TGTTTACTCA ATTGTGGCCC
                                                                   At1g05180
                                                                   mRNA-protein id:
                                                                   CDS
          G  E  V   G  Q  A   A  L  E  E   A  S  I   C  L  L   N  C  G  P
```

Primární data-genomové zdroje

- Human Genome Browser http://genome.ucsc.edu/



About the UCSC Genome Bioinformatics Site

This site contains the reference sequence for the human and *C. elegans* genomes and working drafts for the mouse, rat, Fugu, *Drosophila*, *C. briggsae*, and SARS genomes. It also contains the CFTR (cystic fibrosis) region in 13 species.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Family Browser shows expression, homology and other information on groups of genes that can be related in many ways. The Table Browser provides convenient access to the underlying database. Blat quickly maps your sequence to the genome.

News

`22 Dec. 2003 - Advance Warning of Browser Outage 30 Dec. 2003`

Electrical power to the UCSC School of Engineering building will be shut down on 30 December from 7 a.m. to 3 p.m. PST. The Genome Browser and Blat servers will not be available during this time period.

`10 Dec. 2003 - Human/Chimp Alignment Tracks Released`

UCSC has released alignments of the Nov. 2003 chimpanzee draft assembly to the July 2003 human assembly in the Genome Browser. These alignments may be viewed on the Human July 2003 assembly. This release coincides with today's announcement by the National Human Genome Research Institute (NHGRI) of the first draft assembly of the chimpanzee genome.

The set of human/chimpanzee alignments consists of a reciprocal best-in-genome net track and a chimp chain track. These alignments were generated using the blastz program developed at Pennsylvania State University and the programs blat, axtChain, chainNet, and netSyntenic developed at UCSC by Jim Kent. Research scientists should find these tracks useful for locating orthologous regions and studying genome rearrangement in the two species.

For more information about the alignment tracks, refer to the track description pages. The tables may be downloaded from the Genome Browser FTP server's hg16 database directory. The chimp sequence and alignment data are downloadable from the hg16 human/chimp alignments directory.

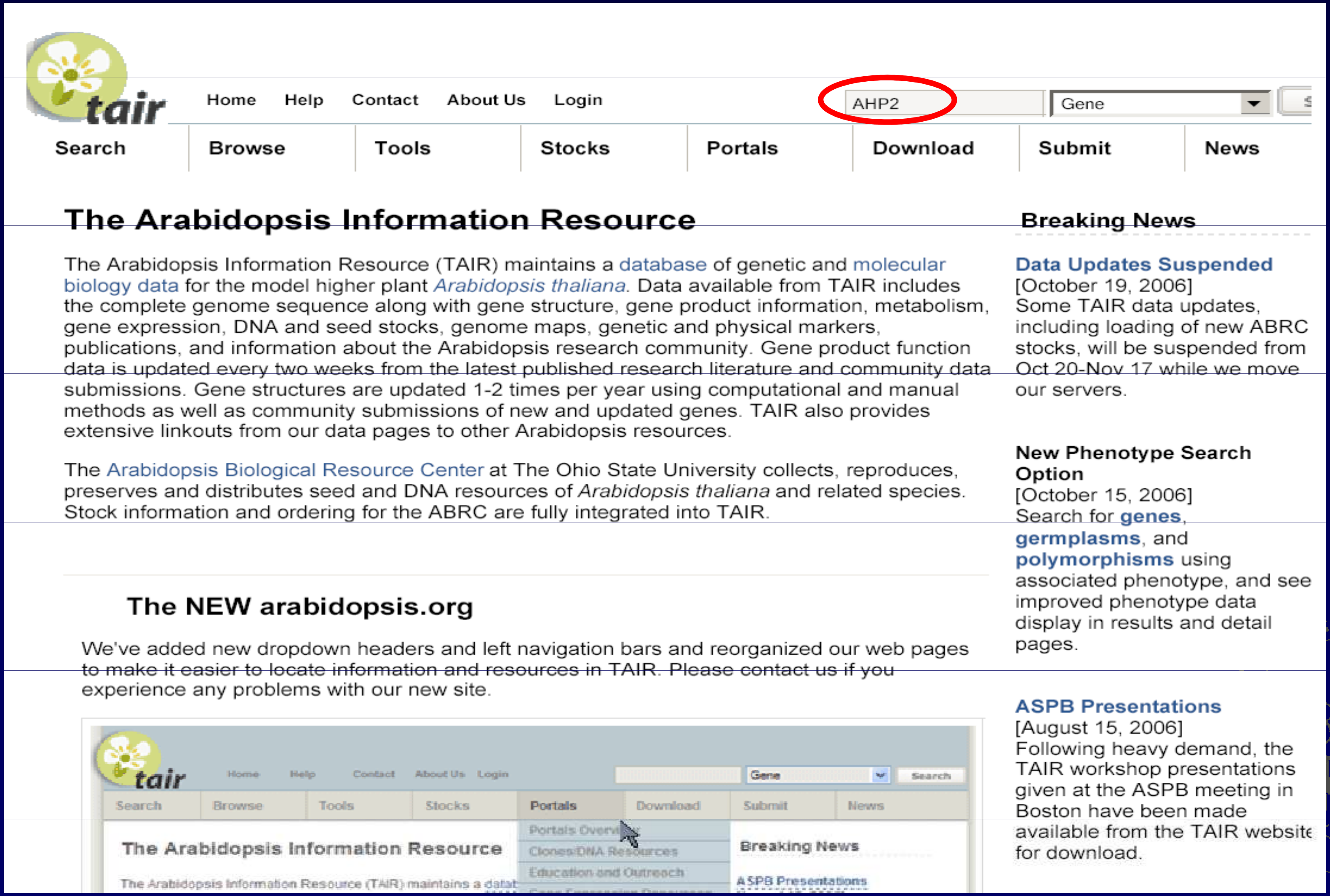
The chimp sequence used in these alignments was obtained from the 13 Nov. 2003 Arachne assembly. We'd like to thank NHGRI, the Eli & Edythe L. Broad Institute at MIT/Harvard, and Washington University School of Medicine for providing this sequence, and LaDeana Hillier, Washington University School of Medicine, and the Whitehead Institute for their work on the alignments. We'd also like to acknowledge the members of the UCSC team who contributed to the release of these alignments in the Genome Browser: Jim Kent, Kate Rosenbloom, Heather Trumbower, and Donna Karolchik.

`24 Nov. 2003 - Build 32 Mouse Assembly (Oct. 2003) Released:` We have released a Genome Browser and Blat server for the latest mouse genome assembly, NCBI Build 32 (UCSC v. mm4). Read more.

`31 Oct. 2003 - V. 3.1 D. melanogaster Assembly (Jan. 2003) Released:` We have added the *Drosophila melanogaster* (fruitfly) assembly to

Primární data-genomové zdroje

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



The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of *Arabidopsis thaliana* and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

The NEW arabidopsis.org

We've added new dropdown headers and left navigation bars and reorganized our web pages to make it easier to locate information and resources in TAIR. Please contact us if you experience any problems with our new site.

Breaking News

Data Updates Suspended
[October 19, 2006]
Some TAIR data updates, including loading of new ABRC stocks, will be suspended from Oct 20-Nov 17 while we move our servers.

New Phenotype Search Option
[October 15, 2006]
Search for **genes**, **germplasms**, and **polymorphisms** using associated phenotype, and see improved phenotype data display in results and detail pages.

ASPB Presentations
[August 15, 2006]
Following heavy demand, the TAIR workshop presentations given at the ASPB meeting in Boston have been made available from the TAIR website for download.

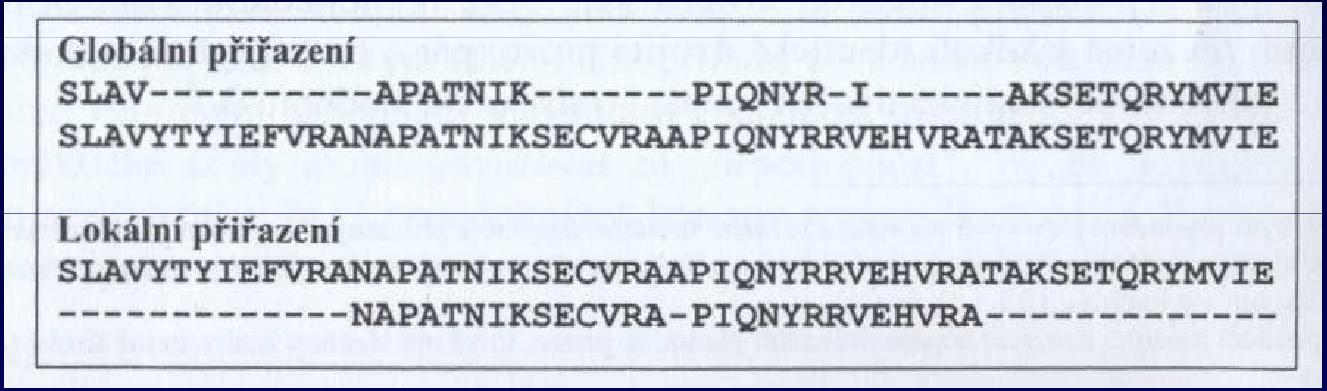
Základy genomiky I.

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

Analytické nástroje

porovnávání dvou sekvencí

- Globální vs. lokální přiřazení



Globální přiřazení

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

Lokální přiřazení

```
SLAVYTYIEFVRANAPATNIKSECVRAAPIQNYRRVEHVRATAKSETQRYMVIE
-------------NAPATNIKSECVRA-PIQNYRRVEHVRA-------------
```

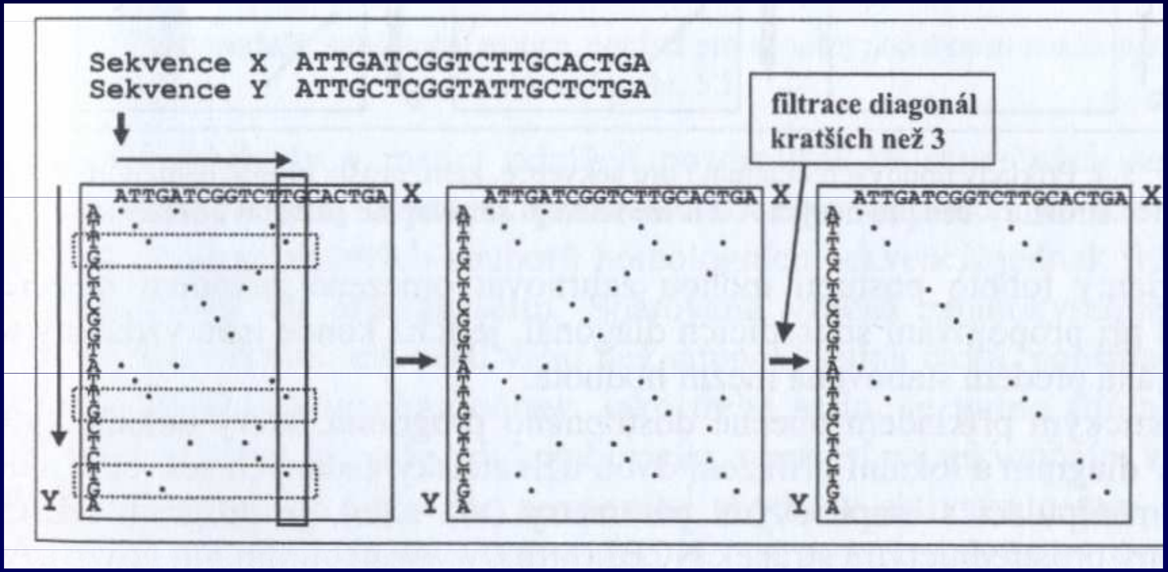
Cvrčková, Úvod do praktické bioinformatiky

- 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 pouze u úseků sekvencí s významnou mírou podobnosti, např. i při záměně pořadí proteinových domén během evoluce

Analytické nástroje

porovnávání dvou sekvencí

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)



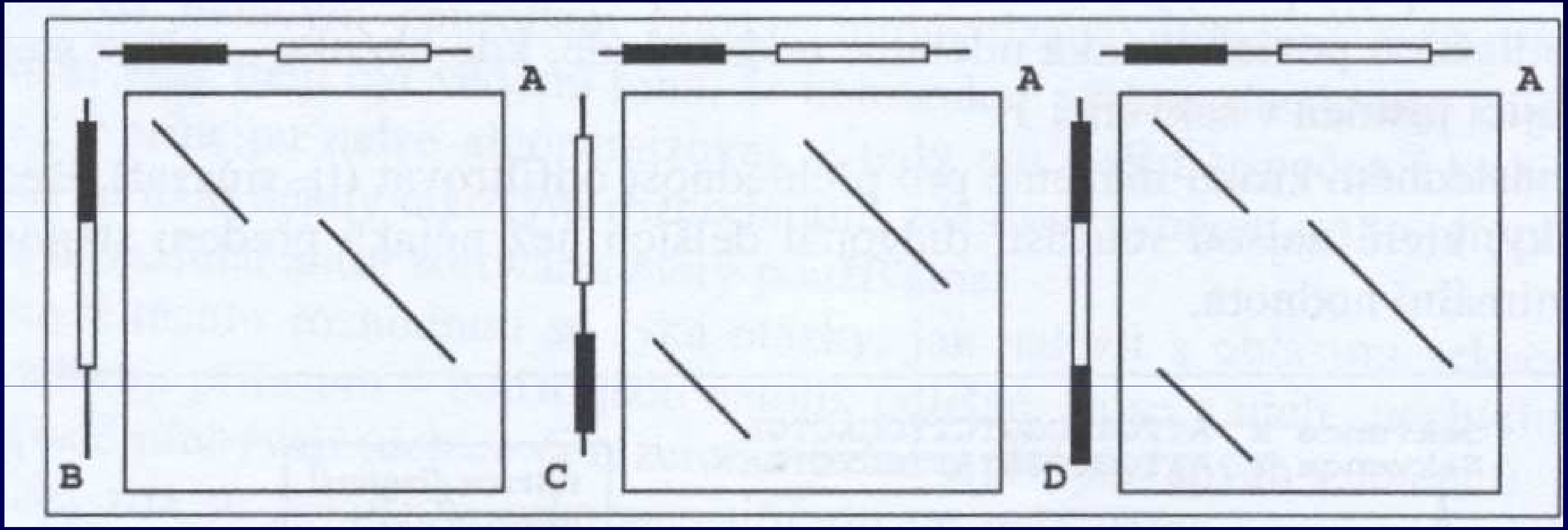
Cvrčková, Úvod do praktické bioinformatiky

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

Analytické nástroje

porovnávání dvou sekvencí

- příklady srovnání sekvencí pomocí bodového diagramu

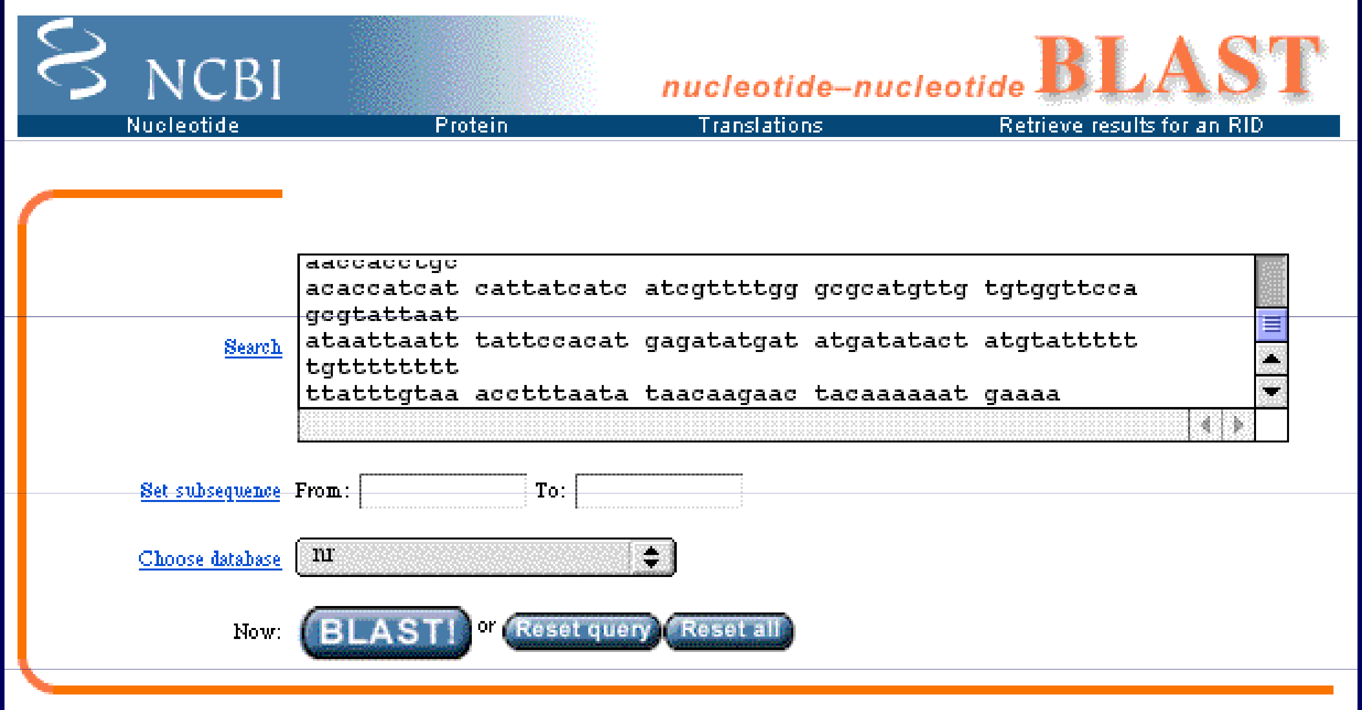


Cvrčková, Úvod do praktické bioinformatiky

- 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)

Analytické nástroje

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



```
acaccatcat cattatcatc atcgttttgg gcgcatgttg tgtggttcca
gcgtattaat
ataattaatt tattccacat gagatatgat atgatatact atgtattttt
tgtttttttt
ttatttgtaa acctttaata taacaagaac tacaaaaaat gaaaa
```

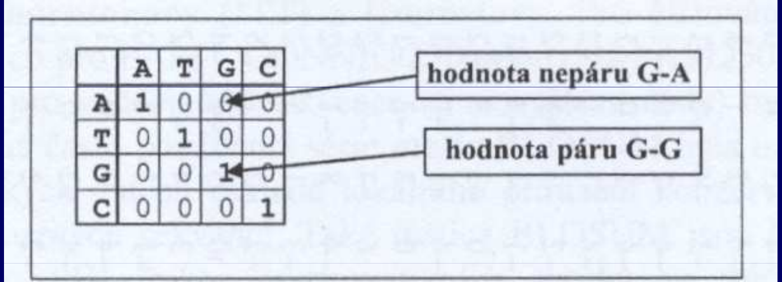
Set subsequence From: To:

Choose database: nr

Now: BLAST! or Reset query Reset all

Podstata algoritmu 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

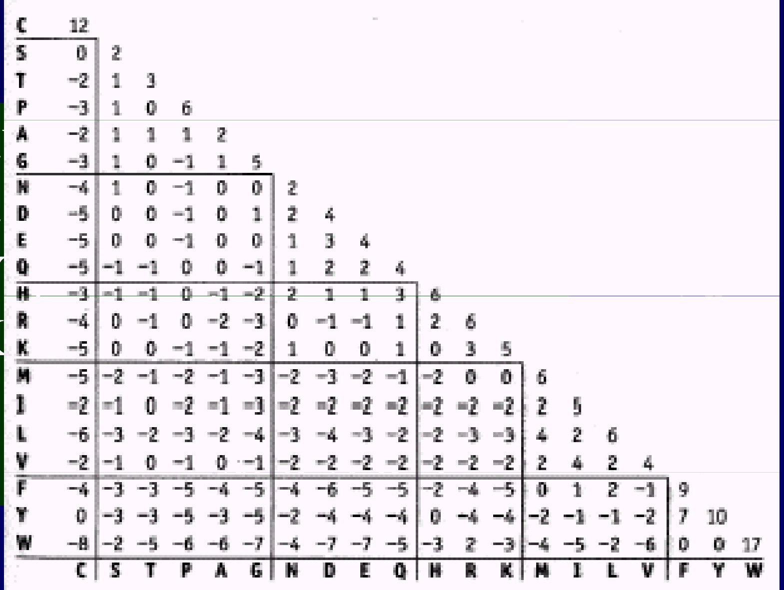


	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

Cvrčková, Úvod do praktické bioinformatiky

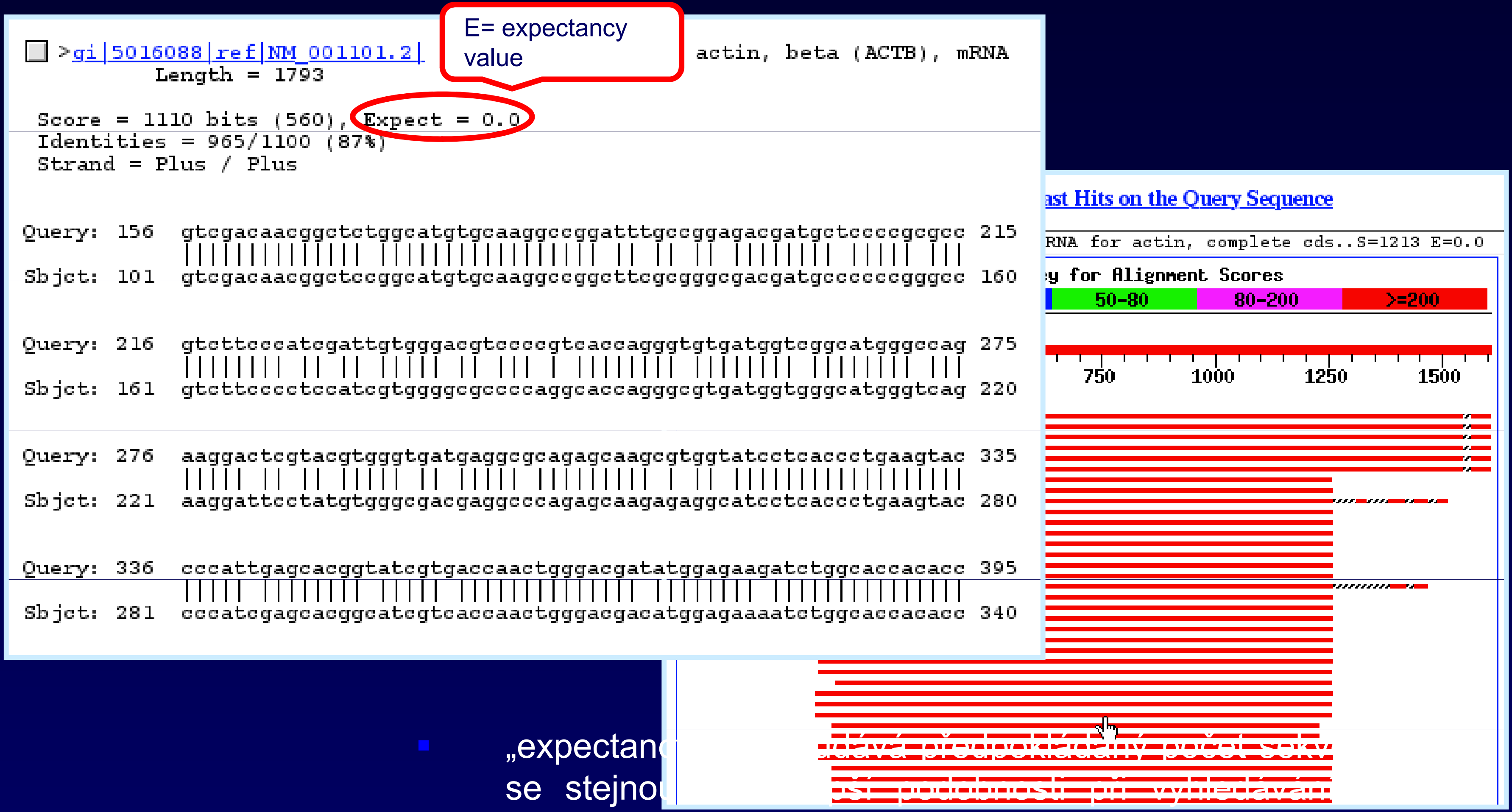
MRKE
MRKE
MRKY
MRAK

Matrice PAM 250



C	12																			
S	0	2																		
T	−2	1	3																	
P	−3	1	0	6																
A	−2	1	1	1	2															
G	−3	1	0	−1	1	5														
N	−4	1	0	−1	0	0	2													
D	−5	0	0	−1	0	1	2	4												
E	−5	0	0	−1	0	0	1	3	4											
Q	−5	−1	−1	0	0	−1	1	2	2	4										
H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Analytické nástroje



E= expectancy value

```
☐ >gi|5016088|ref|NM_001101.2|                actin, beta (ACTB), mRNA
          Length = 1793

 Score = 1110 bits (560), Expect = 0.0
 Identities = 965/1100 (87%)
 Strand = Plus / Plus


Query: 156   gtcgacaacggctctggcatgtgcaaggccggatttgccggagacgatgctccccgcgcc 215
             |||||||||||||| ||||||||||||||||| || || || |||||||| ||||| |||
Sbjct: 101   gtcgacaacggctccggcatgtgcaaggccggcttcgcgggcgacgatgccccccgggcc 160


Query: 216   gtcttcccatcgattgtgggacgtccccgtcaccagggtgtgatggtcggcatgggccag 275
             |||||||| || || ||||| || ||| | |||||||| |||||||| |||||||| |||
Sbjct: 161   gtcttcccctccatcgtggggcgccccaggcaccagggcgtgatggtgggcatgggtcag 220


Query: 276   aaggactcgtacgtgggtgatgaggcgcagagcaagcgtggtatcctcaccctgaagtac 335
             ||||| || || ||||| || ||||| ||||||||| | || ||||||||||||||||||
Sbjct: 221   aaggattcctatgtgggcgacgaggcccagagcaagagaggcatcctcaccctgaagtac 280


Query: 336   cccattgagcacggtatcgtgaccaactgggacgatatggagaagatctggcaccacacc 395
             ||||| |||||||| ||||| |||||||||||||| |||||||| |||||||||||||||
Sbjct: 281   cccatcgagcacggcatcgtcaccaactgggacgacatggagaaaatctggcaccacacc 340
```

- „expectanc udává předpokládaný počet sekv se stejnou ší podobnosti při vyhledávání stejně velké databázi složené z náhodných sekvencí
- výsledek udává frakci totožných a u proteinů i podobných pozic, příp. počet vložených mezer

Analytické nástroje

specializované typy BLASTu

☐ 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

Analytické nástroje

specializované typy BLASTu

- 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

Analytické nástroje

specializované typy BLASTu

- ☐ V současnosti existuje celá řad specializovaných verzí programu BLAST
  - PSI-BLAST (Position-Specific Iterated 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

Analytické nástroje

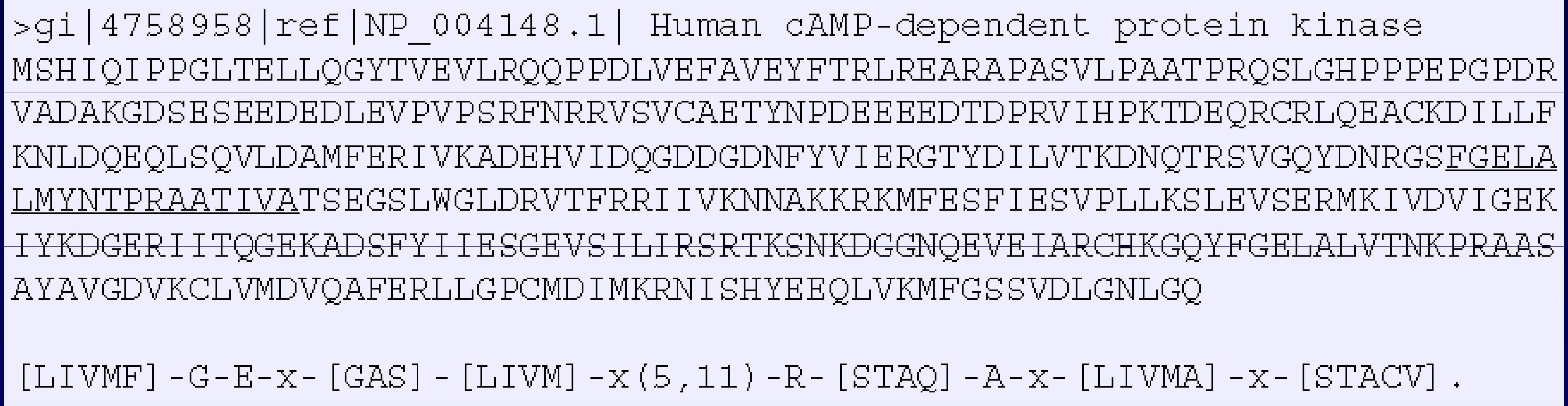
specializované typy BLASTu

- V současnosti existuje celá řad specializovaných verzí programu BLAST
  - PHI-BLAST (Pattern-Hit InitiatedBlast)
    - 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

Analytické nástroje

specializované typy BLASTu

- Příklad vyhledávání pomocí PHI-BLAST



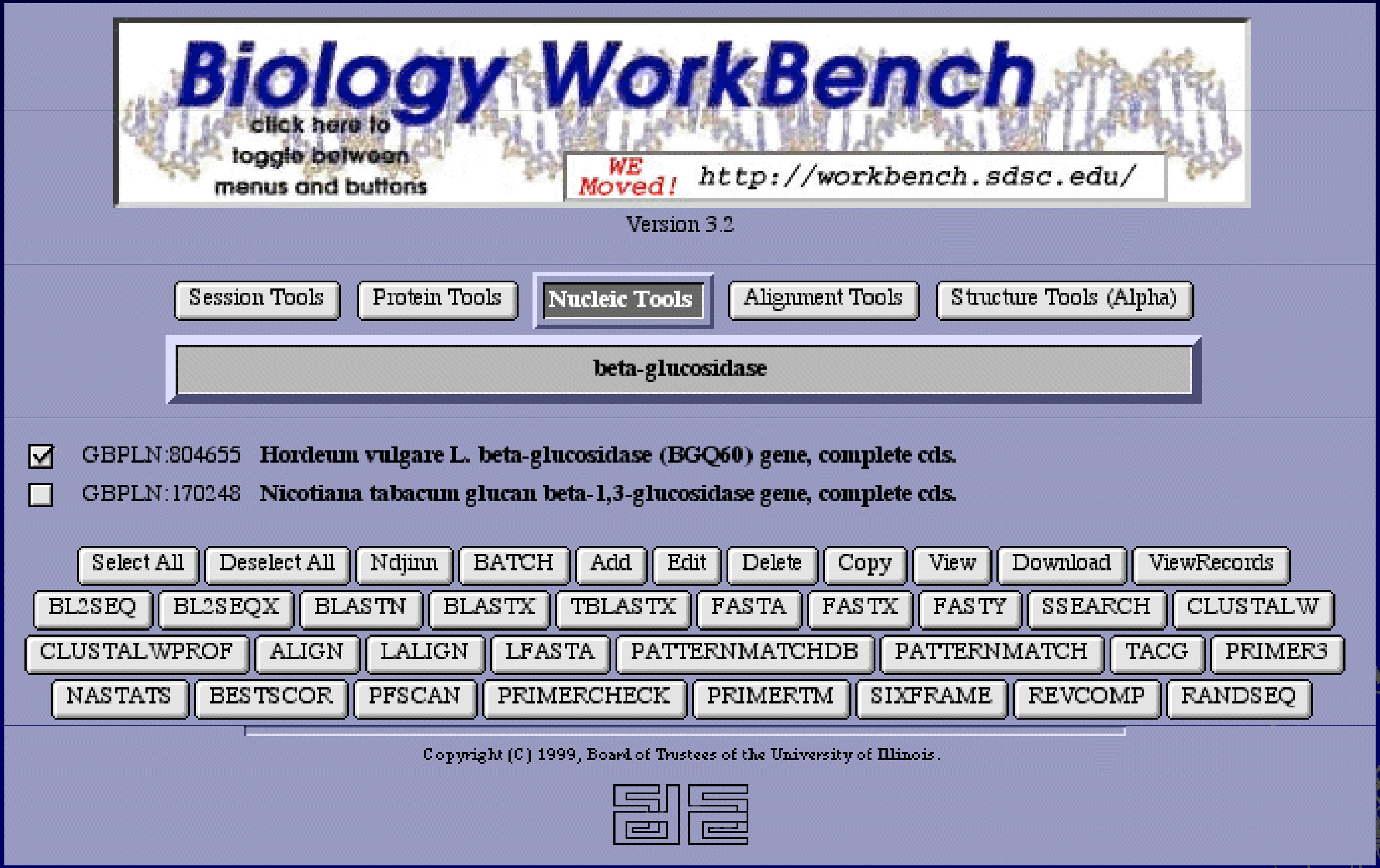
```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAATPRQSLGHPPPEPGPDR
VADAKGDSESEEDEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA
LMYNTPRAATIVATSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIGEK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

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

Základy genomiky I.

- 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….

Analytické nástroje

- http://workbench.sdsc.edu/



beta-glucosidase

☑ GBPLN:804655 **Hordeum vulgare L. beta-glucosidase (BGQ60) gene, complete cds.**

☐ GBPLN:170248 **Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds.**

Copyright (C) 1999, Board of Trustees of the University of Illinois.

Analytické nástroje

Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds.
GBPLN:170248, 4699 bp

```
>170248
GAGCTCCCTTGGGGGGCAAGGGCAAAACTTTTTGCTAAATGGAAAAATATTATACCAAGTGTTTGTAATA
GTTACTCAATTTGAATTAACAAAGGGGCAAATTTGACTATTTTGCCCTTATATCTTTTGGTCACAAAAAC
ATAAAATATCCCATCCGAAATTCCAAATGGTCCATTATCGGCAAGTAGCTTTCTTTAATTATAGTTAGTT
GACAAAACACTATCAAGATATCATTATTATAATAATAACTTCAAAGTCCATCATCTTAGCTGCCTCCTCA
GTAGAGCCGCCAGTAAAATAAGACCGATCAAATAAAAGCCGCCATTAAAATAATGAATTTTAGGACTCTC
GATTGGCACGTAAGTGCCAAAACTCTTCCAATACTTTGCTGCAACTTGGGGCTGCTAGGTTCTGAGCTTC
CAGATATGGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTAAACAGGTA
CAGCAAATCATAAAATTTTCCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCATGCCATATTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAATAATTTCGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTAATTATAAGTATTAGCTAATTTCAGAATAAATTTTACACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACCAAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAATTTTCATTAAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTGTCTTCAAAGTAGATTCTCATTCATTGCTCTGGTGCAATAGCAAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA
GAGAACTTTCAAATCTCTCAACTACTTTTAAGGGAATTCAAAATACGACCAATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTATTTTAAATTTGAATTGAAAATATTAAATTACTTGATTTAATATAA
```

Analytické nástroje

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
CAGATATGGATATTTCTAAGTTTATCTCCTAATTTACATCTCAACTAATATTAAGAAATTAAACAGGTA
CAGCAAATCATAAAATTTTCCTCTAAAGAAGACAATGAATCCGGTTACTGATTCATTGGCCTTTTCAGAG
TCTGCATGCCATATTCACTAAGGGGTCGTTTGGTACAAGAAATAATAATAATAATTTCGGGATAGAATTT
GAGATTGCATTTATCTTGTGTTTAATTATAAGTATTAGCTAATTTCAGAATAAATTTTACACTAAAATAG
TAAAATCAACTATCACATGTAGAAGGTGGAATGGAATAGCTAATCCCATAGCCACTCACATAGAATATCC
TTATTTATCTCACTATTTTACCAAATGATCGGTTAGTCTTCATGAGAATCCAGTATCCTCAATAAATGCA
GTAAGAAGTTAGAAAATTTTCATTAAATCAATTCATATAATTTAAAAATATTAGATATGGAGCACTTAAG
ATACAATAAAAGATGTACCGTTAATAATAAAAGATAAGATAGAGTTTTAAATAGGAAAAAAAAAACGGTT
CGAGACACTCTTATGGAAGGCGTTGTCTTCAAAGTAGATTCTCATTCATTGCTCTGGTGCAATAGCAAAA
TGACATCTTACTCTTAAGATACAGCGAGCCACTCTACAATCTTCTATTGTATACTCAAATGAAAGTTTTA
GAGAACTTTCAAATCTCTCAACTACTTTTAAGGGAATTCAAAATACGACCAATATTTATTACTTACTTAC
TTATAGTTAAATGATATGAATTTTATTTTAAATTTGAATTGAAAATATTAAATTACTTGATTTAATATAA
ACAATAGATATCGCTAAGTATTTACCACAAACATGGAGATACTACAGAAGATTTTATTATTTGTAACGAT
GATTAAGCAGCTATTCATCTGGTTGTGCAGGATGAAAGAAAGTAACTAGCTATAATTTCTTTTGTAAAGT
```

Analytické nástroje

Frame 1, 1 stop codon

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

```
>170248 Translated - Frame 1
ELPWGARAKLFAKWKNIIPSVCNSYSI*INKGANLTILPL
```

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

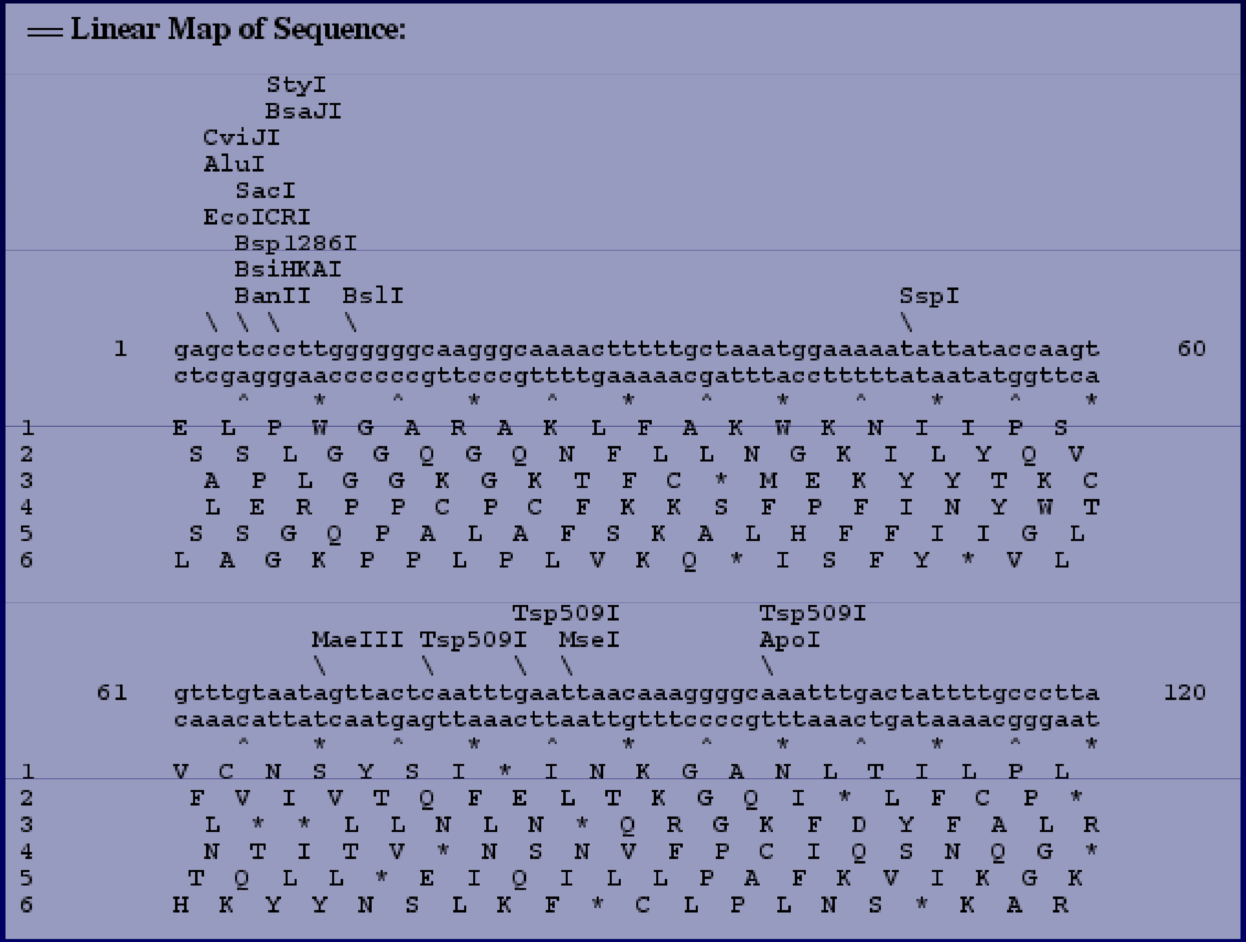
Frame 2, 1 stop codon

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

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

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

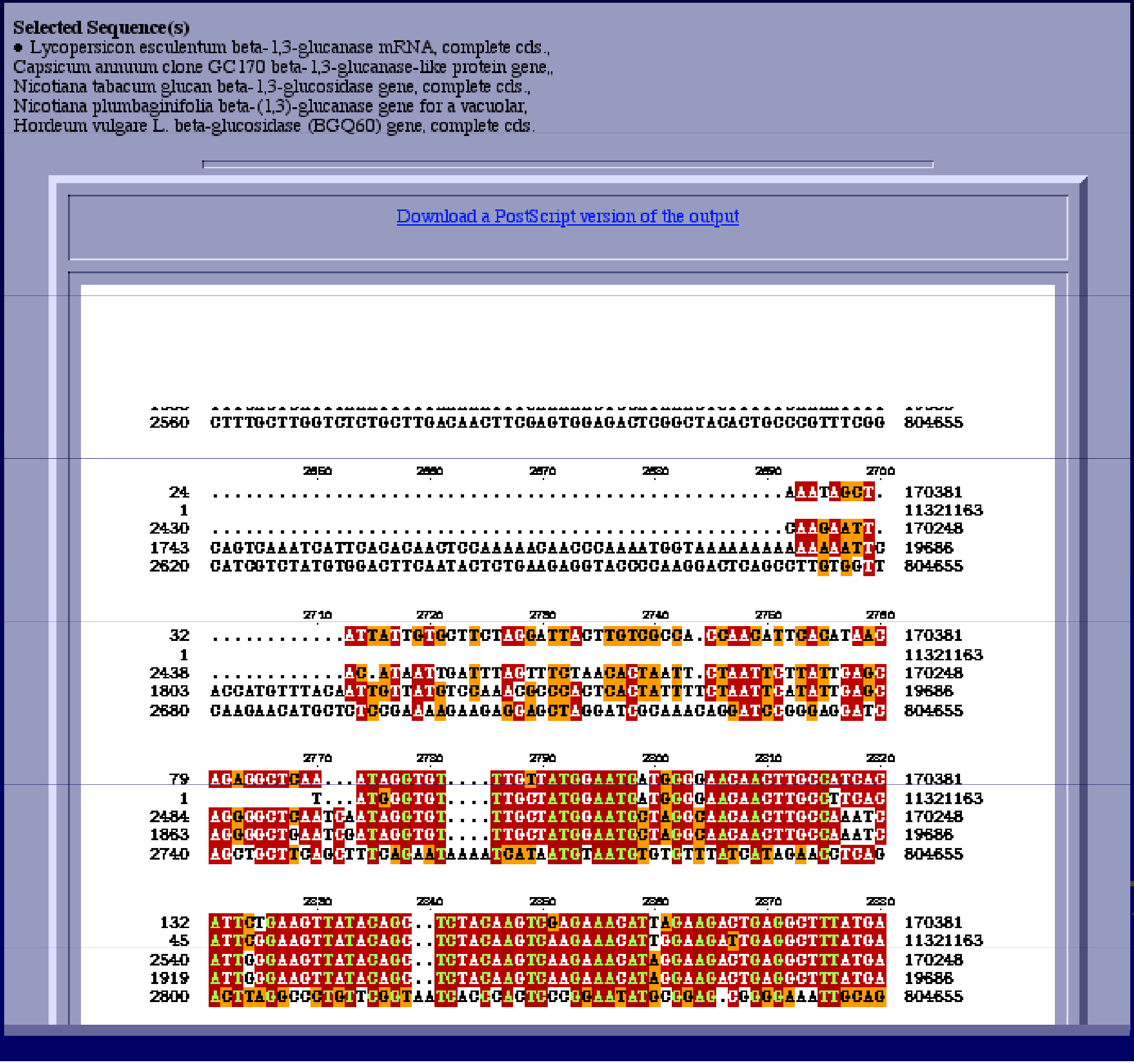
Analytické nástroje



═ Linear Map of Sequence:

```
                StyI
                BsaJI
            CviJI
            AluI
              SacI
            EcoICRI
              Bsp1286I
              BsiHKAI
              BanII  BslI                                SspI
            \ \ \    \                                   \
      1   gagctcccttggggggcaagggcaaaactttttgctaaatggaaaaatattataccaagt     60
          ctcgagggaaccccccgttcccgttttgaaaaacgatttacctttttataatatggttca
              ^    *    ^    *    ^    *    ^    *    ^    *    ^    *
1         E  L  P  W  G  A  R  A  K  L  F  A  K  W  K  N  I  I  P  S
2          S  S  L  G  G  Q  G  Q  N  F  L  L  N  G  K  I  L  Y  Q  V
3           A  P  L  G  G  K  G  K  T  F  C  *  M  E  K  Y  Y  T  K  C
4           L  E  R  P  P  C  P  C  F  K  K  S  F  P  F  I  N  Y  W  T
5          S  S  G  Q  P  A  L  A  F  S  K  A  L  H  F  F  I  I  G  L
6         L  A  G  K  P  P  L  P  L  V  K  Q  *  I  S  F  Y  *  V  L

                                Tsp509I         Tsp509I
                   MaeIII Tsp509I  MseI         ApoI
                   \      \     \  \            \
     61   gtttgtaatagttactcaatttgaattaacaaaggggcaaatttgactattttgccctta    120
          caaacattatcaatgagttaaacttaattgtttccccgtttaaactgataaaacgggaat
              ^    *    ^    *    ^    *    ^    *    ^    *    ^    *
1         V  C  N  S  Y  S  I  *  I  N  K  G  A  N  L  T  I  L  P  L
2          F  V  I  V  T  Q  F  E  L  T  K  G  Q  I  *  L  F  C  P  *
3           L  *  *  L  L  N  L  N  *  Q  R  G  K  F  D  Y  F  A  L  R
4           N  T  I  T  V  *  N  S  N  V  F  P  C  I  Q  S  N  Q  G  *
5          T  Q  L  L  *  E  I  Q  I  L  L  P  A  F  K  V  I  K  G  K
6         H  K  Y  Y  N  S  L  K  F  *  C  L  P  L  N  S  *  K  A  R
```

Analytické nástroje



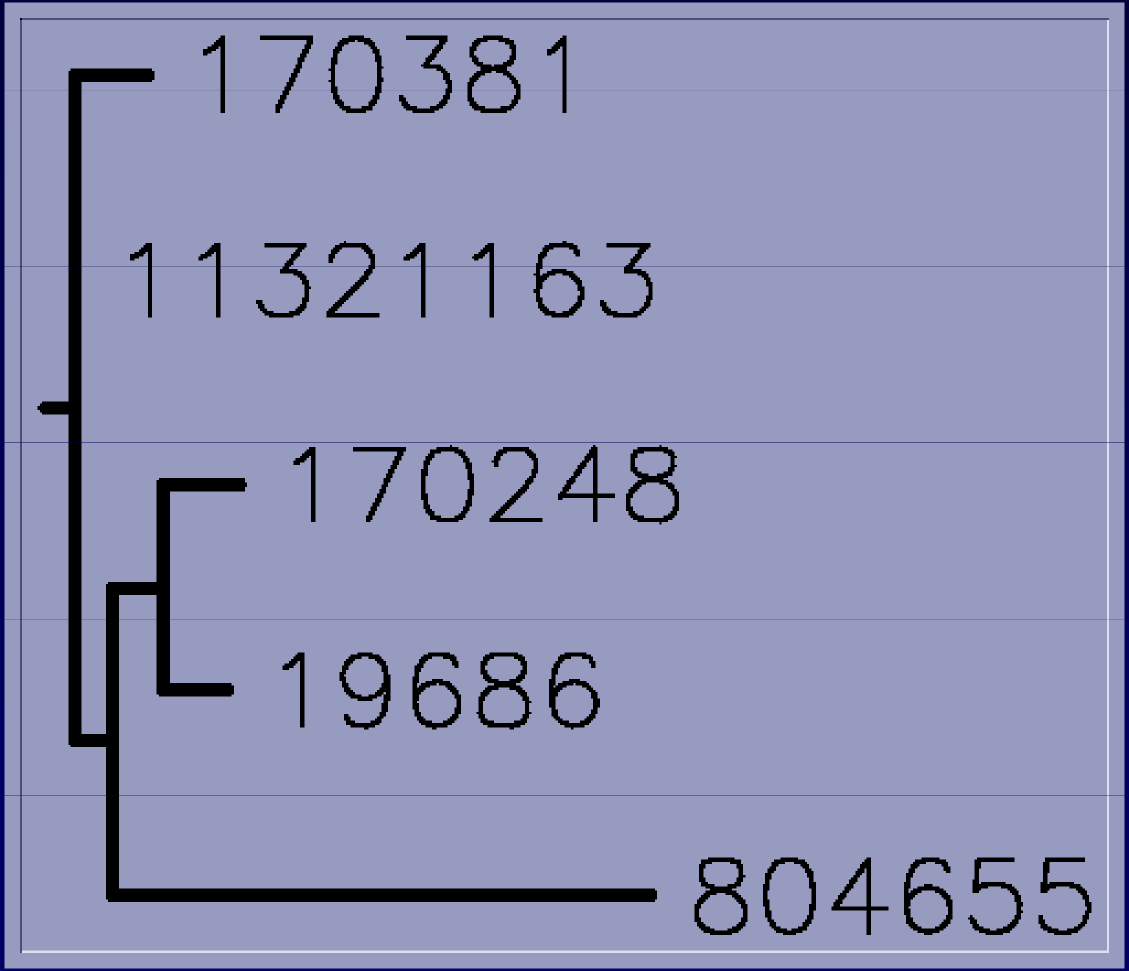
Selected Sequence(s)

- Lycopersicon esculentum beta-1,3-glucanase mRNA, complete cds., Capsicum annuum clone GC170 beta-1,3-glucanase-like protein gene,, Nicotiana tabacum glucan beta-1,3-glucosidase gene, complete cds., Nicotiana plumbaginifolia beta-(1,3)-glucanase gene for a vacuolar, Hordeum vulgare L. beta-glucosidase (BGQ60) gene, complete cds.

Download a PostScript version of the output

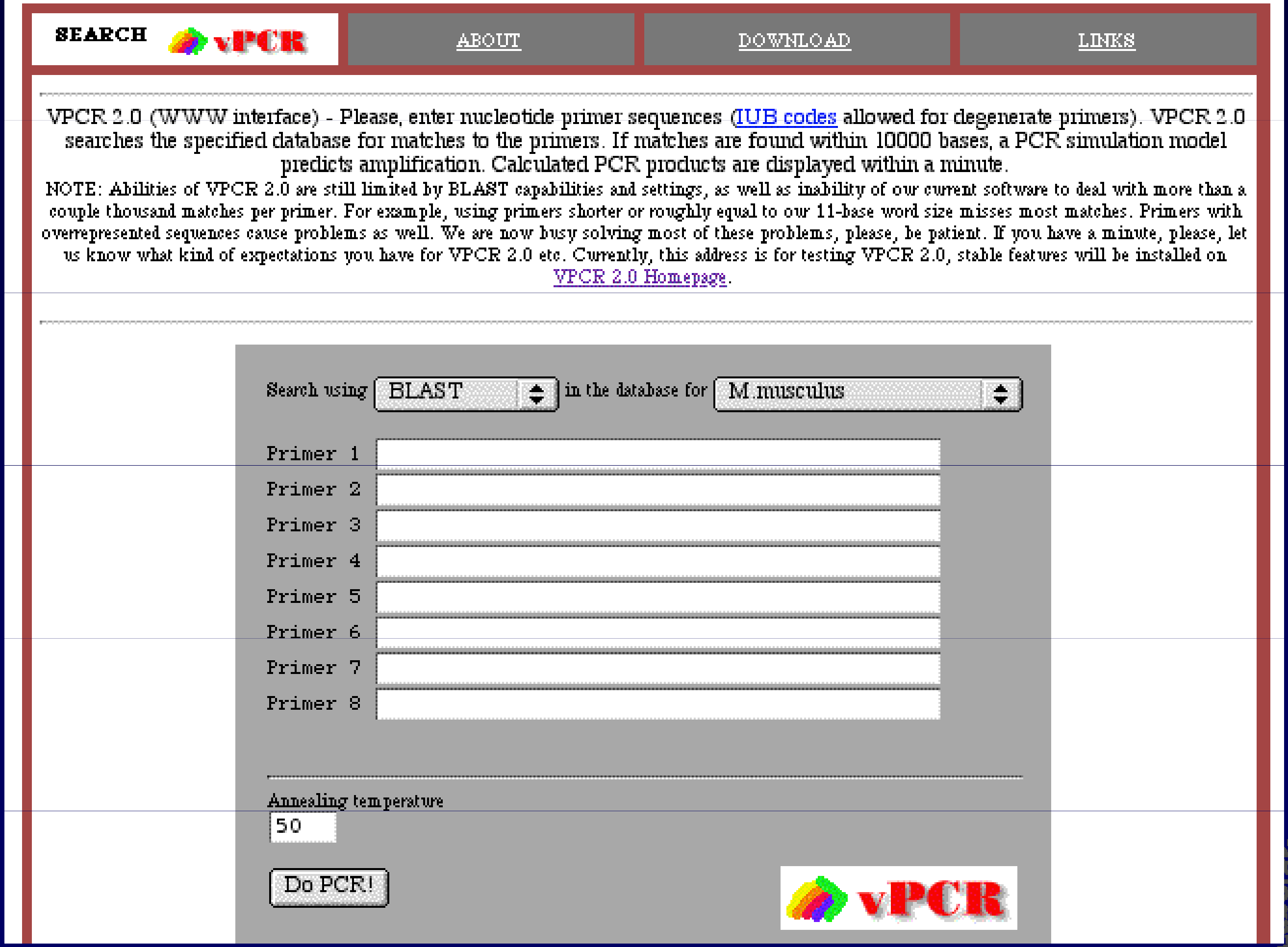
```
2560  CTTTGCTTGGTCTCTGCTTGACAACTTCGAGTGGAGACTCGGCTACACTGCCCGTTTCGG  804655

                2650      2660      2670      2680      2690      2700
  24  ...................................................AAATAGCT.  170381
   1                                                                11321163
2430  ...................................................CAAGAATT.  170248
1743  CAGTCAAATCATTCACACAACTCCAAAAACAACCCAAAATGGTAAAAAAAAAAAAAATTC  19686
2620  CATGGTCTATGTGGACTTCAATACTCTGAAGAGGTACCCCAAGGACTCAGCCTTGTGGTT  804655

                2710      2720      2730      2740      2750      2760
  32  ............ATTATTGTGCTTCTAGGATTACTTGTCGCCA.CCAAGATTCACATAAC  170381
   1                                                                11321163
2438  ............AC.ATAATTGATTTACTTTCTAACACTAATT.CTAATTCTTATTGAGC  170248
1803  ACCATGTTTACAATTGTTATGTCCAAACGCCCACTCACTATTTTCTAATTCATATTGAGC  19686
2680  CAAGAACATGCTCTCCGAAAAGAAGAGCAGCTAGGATCGCAAACAGGATCCGGGAGGATC  804655

                2770      2780      2790      2800      2810      2820
  79  AGAGGCTCAA...ATAGGTGT....TTGTTATGGAATGATGGCGAACAACTTGCCATCAC  170381
   1           T...ATGGGTGT....TTGCTATGGAATGATGGCGAACAACTTGCCTTCAC  11321163
2484  AGGGGCTCAATCAATAGGTGT....TTGCTATGGAATGCTAGGCAACAACTTGCCAAATC  170248
1863  AGGGGCTGAATCGATAGGTGT....TTGCTATGGAATGCTAGGCAACAACTTGCCAAATC  19686
2740  AGCTGCTTCAGCTTTCAGAATAAAATCATAATGTAATGTGTCTTTATCATAGAACCTCAG  804655

                2830      2840      2850      2860      2870      2880
 132  ATTCTGAAGTTATACAGC..TCTACAAGTCGAGAAACATTAGAAGACTGAGGCTTTATGA  170381
  45  ATTCGGAAGTTATACAGC..TCTACAAGTCAAGAAACATTGGAAGATTGAGGCTTTATGA  11321163
2540  ATTGGGAAGTTATACAGC..TCTACAAGTCAAGAAACATAGGAAGACTGAGGCTTTATGA  170248
1919  ATTGGGAAGTTATACAGC..TCTACAAGTCAAGAAACATAGGAAGACTGAGGCTTTATGA  19686
2800  ACTTAGGCCCTGTTCGGTAATCACCGACTCCCGGAATATGCCGAG.CGGGCAAATTGCAG  804655
```

Analytické nástroje



Analytické nástroje

- VPCR http://elanor.sci.muni.cz/LMFR/vpcr/



VPCR 2.0 (WWW interface) - Please, enter nucleotide primer sequences (IUB codes allowed for degenerate primers). VPCR 2.0 searches the specified database for matches to the primers. If matches are found within 10000 bases, a PCR simulation model predicts amplification. Calculated PCR products are displayed within a minute.

NOTE: Abilities of VPCR 2.0 are still limited by BLAST capabilities and settings, as well as inability of our current software to deal with more than a couple thousand matches per primer. For example, using primers shorter or roughly equal to our 11-base word size misses most matches. Primers with overrepresented sequences cause problems as well. We are now busy solving most of these problems, please, be patient. If you have a minute, please, let us know what kind of expectations you have for VPCR 2.0 etc. Currently, this address is for testing VPCR 2.0, stable features will be installed on VPCR 2.0 Homepage.

Search using BLAST in the database for M.musculus

Primer 1
Primer 2
Primer 3
Primer 4
Primer 5
Primer 6
Primer 7
Primer 8

Annealing temperature
50

Do PCR!

Analytické nástroje

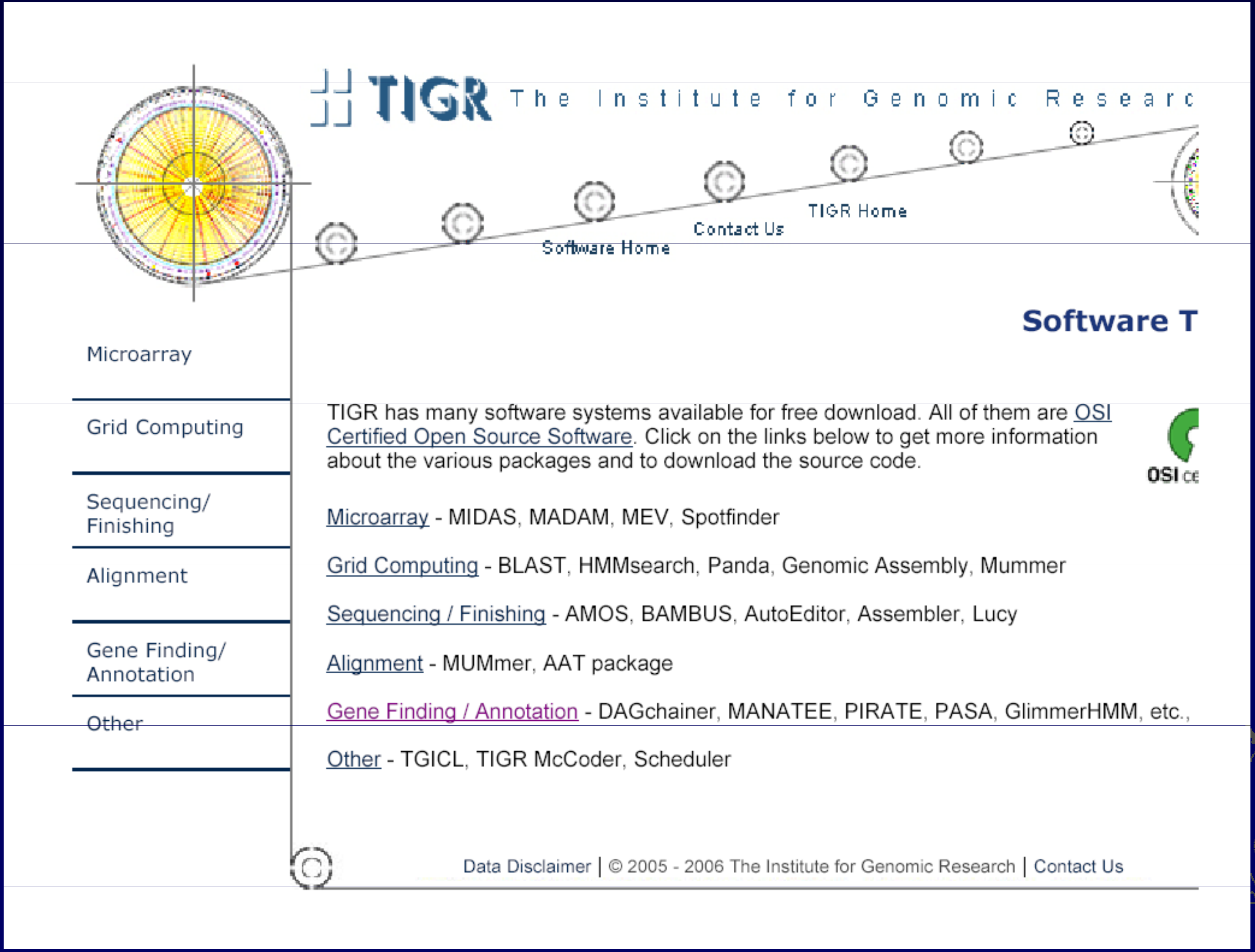
```
Read 2 primers, 18 matches and 1 amplicons. TEMP=50oC
PCR limited by Taq in cycle 19
0:1e-06 619 NC_003076.1 2254473 2253854 1 2
```

Základy genomiky I.

- 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

www analytické nástroje

- TIGR (The Institute for Genomic Research, http://www.tigr.org/software/)



Software T

TIGR has many software systems available for free download. All of them are OSI Certified Open Source Software. Click on the links below to get more information about the various packages and to download the source code.

Microarray - MIDAS, MADAM, MEV, Spotfinder

Grid Computing - BLAST, HMMsearch, Panda, Genomic Assembly, Mummer

Sequencing / Finishing - AMOS, BAMBUS, AutoEditor, Assembler, Lucy

Alignment - MUMmer, AAT package

Gene Finding / Annotation - DAGchainer, MANATEE, PIRATE, PASA, GlimmerHMM, etc.,

Other - TGICL, TIGR McCoder, Scheduler

Data Disclaimer | © 2005 - 2006 The Institute for Genomic Research | Contact Us

Základy genomiky I.

shrnutí

- 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

Základy genomiky I.

diskuse