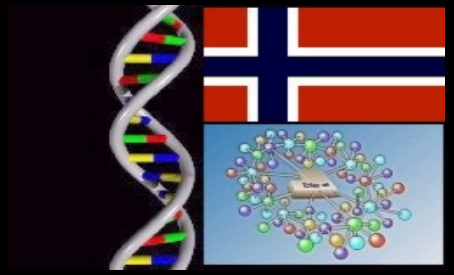


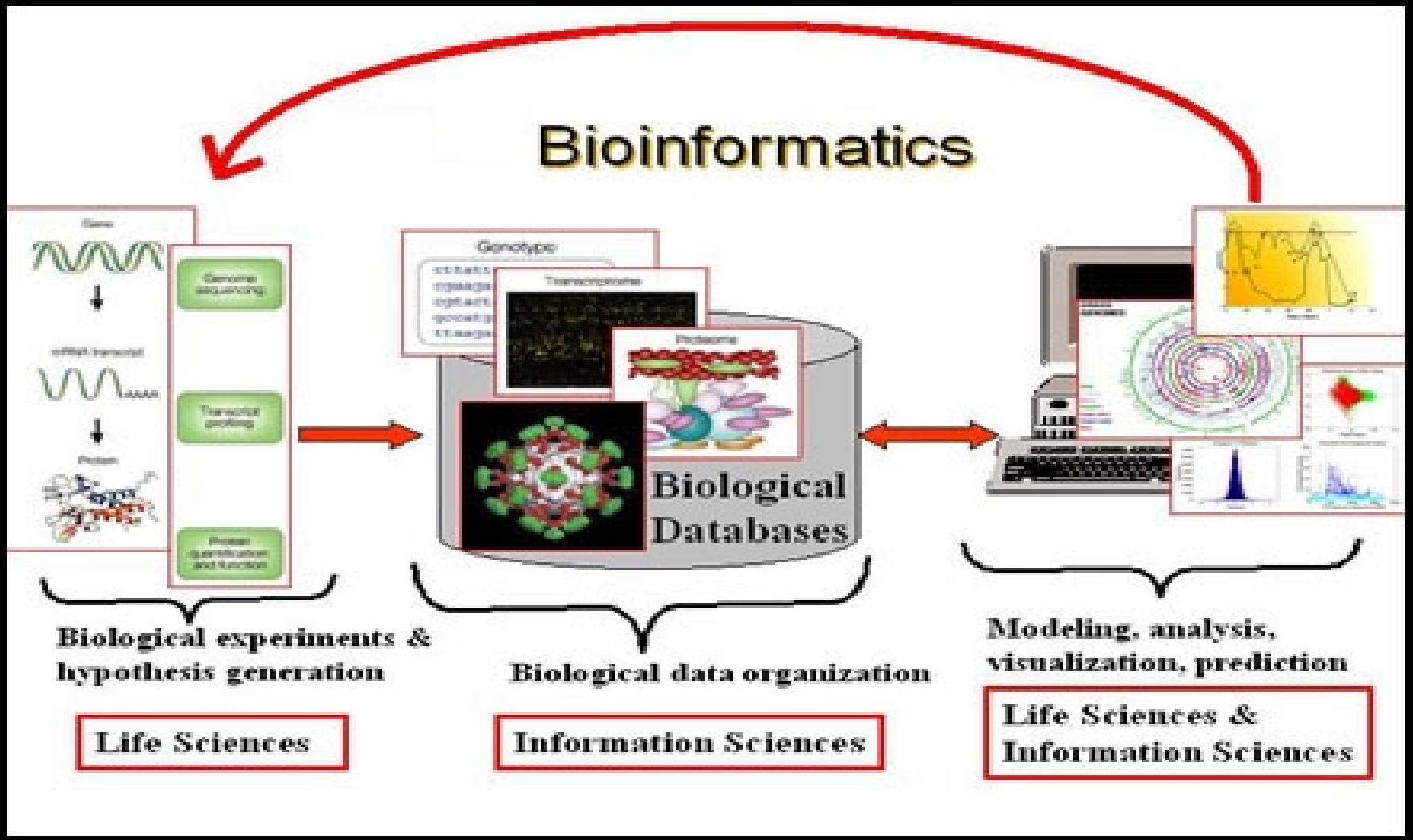
Biological sequence mining with MRS and EMBOSS

George Magklaras
EMBnet Norway

http://www.no.embnet.org
admmast@biotek.uio.no

EMBnet

Why should I bother?



Suggested bibliography/notes:

- ✓**EMBNET UNIX Quick Guide:**
  http://www.no.embnet.org/quickguides/UNIX03.pdf
- ✓**UNIX command line lecture notes:**
  http://www.no.embnet.org/lectures/UNIXforTIGR2006.ppt
- ✓**Introduction to Bioinformatics by Lesk**
  http://tinyurl.com/yzrxncq

Lecture Agenda

- ✓ Basic information about sequence databanks (flat file databases) (*slides 5-10*)
- ✓ Major sequence databases (slides 11-20)
- ✓ Other notable databases (slides 21-23)
- ✓ Sequence formats (slides 24-28)
- ✓ The Web/GUI interface versus the command line pipeline construction paradigm (slides 29-34)
- ✓ The EMBOSS suite (slides 35-47)
- ✓ Sequence retrieval systems (slides 48-57)
- ✓ MRS (slides 58-70)
- ✓ Case studies/Questions and Answers

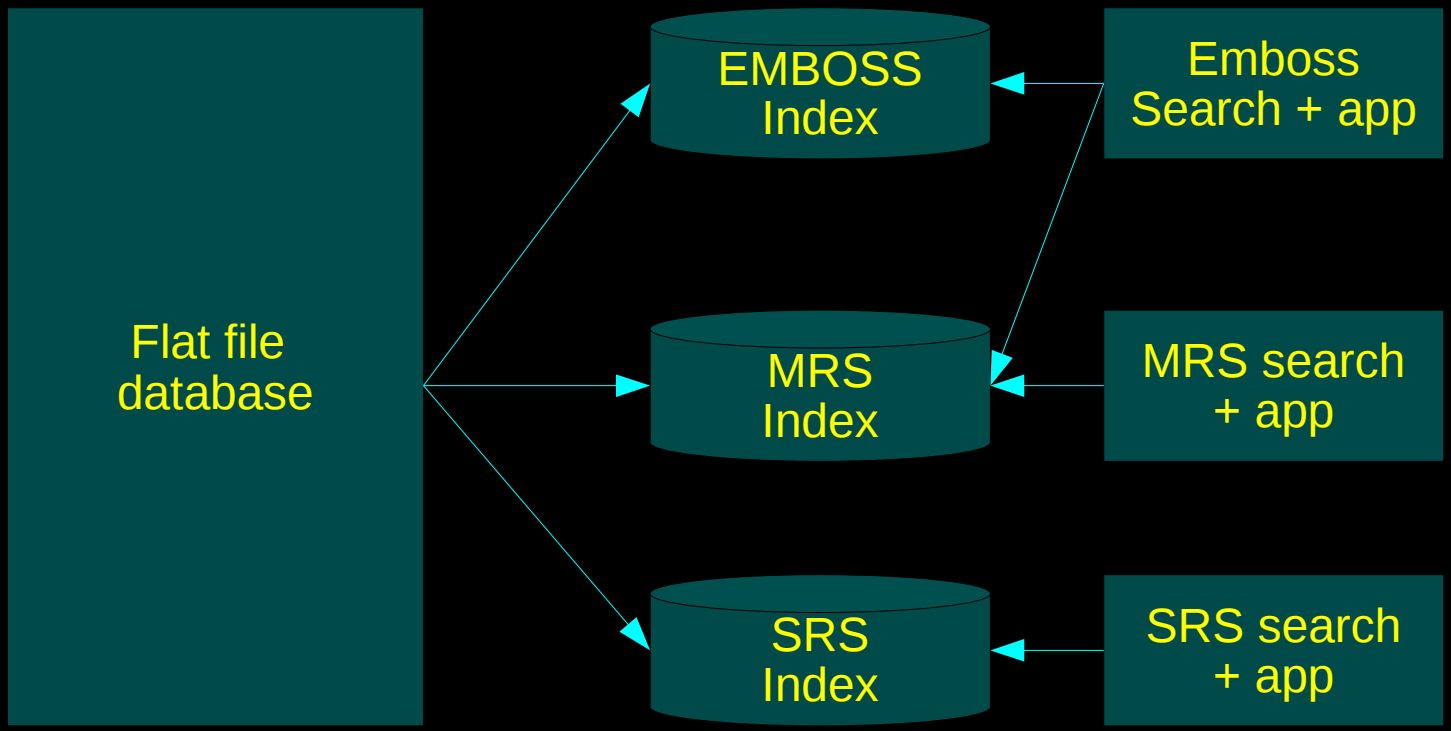
Flat file databases

- ✓ The computer world has many different types of databases:
  - *Relational*
  - *Object oriented*
- ✓ Biological sequences are often organized in flat file databases. This means that the source files that contain the sequences are simple human readable files, as opposed to unreadable binary files.

Flat file databases (2)

- ✓ One way to access the information is to obtain the flat files and look in them with the help of a text editor (Notepad, Wordpad, vi, pico, emacs, etc).
- ✓ Some of the files are too big to open on text editors (fancy reading 100000 entries until you find what you want?): flat files are indexed by various application toolkits (EMBOSS, GCG, SRS, MRS).
- ✓ An index is a set of pointers to information in the database. It speeds up the process of sequence searching/retrieval.

Flat file databases (3)



EMBOSS and MRS indexes are interoperable.

Flat file databases (4)

- ✓ A database index allows us to:
  - ***retrieve info fast****: out of 2 million entries find the one with accession number X04049 and give me the sequence.*
  - ***filter out (mine) sequences of interest****:*
    - Give me all the hemoglobin sequences of the EMBL database.
    - Give me only the human hemoglobin sequences of the EMBL database.
    - Give me the only the human hemoglobin sequences that also contain the word 'dehydrogenase' AND not the word 'anion'.

Flat file databases (5)

- ✓ All flat file databases follow a consistent pattern of displaying data:

header_of_sequence1

sequence1

header_of_sequence2

sequence2

.

.

.

header_of_sequencen

sequencen

  - The header contains info about the sequence that help us to identify it. Header fields are often indexed to help us find the sequence easily

```
ID   Q06S78_9INFA            Unreviewed;       757 AA.
AC   Q06S78;
DT   31-OCT-2006, integrated into UniProtKB/TrEMBL.
DT   31-OCT-2006, sequence version 1.
DT   01-SEP-2009, entry version 14.
DE   RecName: Full=RNA-directed RNA polymerase catalytic subunit;
DE            EC=2.7.7.48;
OS   Influenza A virus (A/cat/Germany/606/2006(H5N1)).
OC   Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
OC   Influenzavirus A.
...
KW   Nucleotide-binding; Nucleotidyltransferase; RNA replication;
KW   RNA-directed RNA polymerase; Transferase.
SQ   SEQUENCE   757 AA;  86462 MW;  273457664D64BC0D CRC64;
     MDVNPTLLFL KVPVQNAIST TFPYTGDPPY SHGTGTGYTM DTVNRTHQYS EKGKWTTNTE
     TGAPQLNPID GPLPEDNEPS GYAQTDCVLE AMAFLEESHP GIFENSCLET MEIVQQTRVD
     KLTQGRQTYD WTLNRNQPAA TALANTIEIF RSNGLTANES GRLIDFLKDV MESMDKEEME
     ITTHFQRKRR VRDNMTKKMV TQRTIGKKKQ RLNKKSYLIR ALTLNTMTKD AERGKLKRRA
     IATPGMQIRG FVYFVETLAR SICEKLEQSG LPVGGNEKKA KLANVVRKMM TNSQDTELSF
     TITGDNTKWN ENQNPRMFLA MITYITRNQP EWFRNVLSIA PIMFSNKMAR LGRGYMFESK
     SMKLRTQIPA EMLANIDLKY FNELTKKKIE KIRPLLIDGT ASLSPGMMMG MFNMLSTVLG
     VSILNLGQKR YTKTTYWWDG LQSSDDFALI VNAPNHEGIQ AGVDRFYRTC KLVGINMSKK
     KSYINRTGTF EFTSFFYRYG FVANFSMELP SFGVSGINES ADMSIGSTVI RNNMINNDLG
     PATAQMALQL FIKDYRYTYR CHRGDTQIQT RRSFELKKLW EQTRSKAGLL VSDGGPNLYN
     IRNLHIPEVC LKWELMDEDY QGRLCNPLNP FVSHKEIESV NNAVVMPAHG PAKGMEYDAV
     ATTHSWIPKR NRSILNTSQR GILEDEQMYQ KCCNLFEKFF PSSSYRRPVG ISSMVEAMVS
     RARIDARIDF ESGRIKKEEF AEIMKICSTI EELRRPK
//
```

Major sequence databases

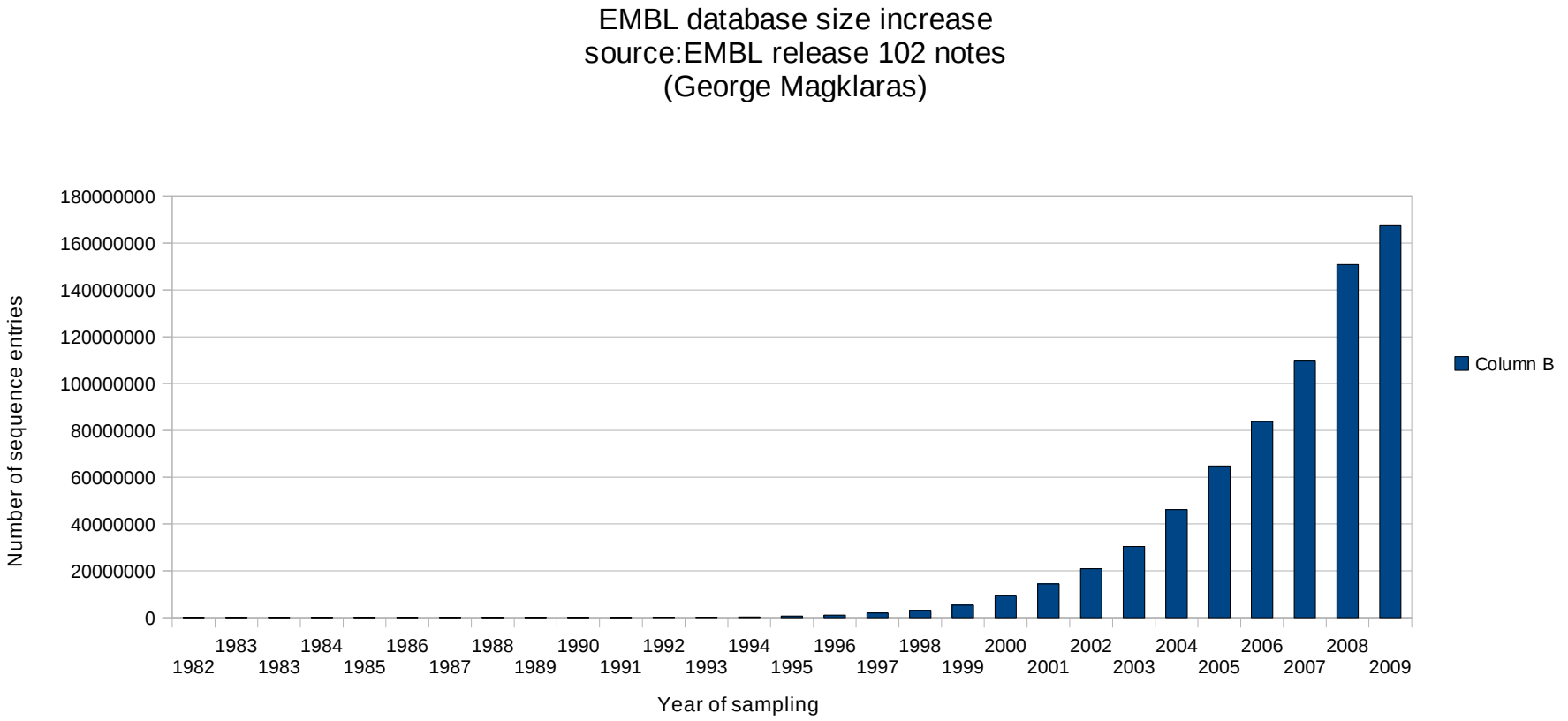
- ✓ Who makes them and why?

*'The steady conversion of new techniques into purchasable kits and the accumulation of nucleotide sequence data in the electronic data banks leads one practitioner to cry, "**Molecular biology is dead - Long live molecular biology!**"'*

"Towards a paradigm shift in biology".

W. Gilbert NATURE 349:99 1991.

Major sequence databases (2)



Major sequence databases (3)

- ✓ There are three major nucleotide sequence databases.
  - **EMBL** (European Molecular Biology Laboratory)
  - **NCBI's Genbank** (the U.S. National Center for Biotechnology Information)
  - **DDBJ** (the DNA Data Bank of Japan).
- ✓ Each of these databases convey the same info (all of the known nucleic acid (DNA/RNA) sequences) and they are often in sync with each other.
- ✓ However, each database has its own flat file format.
- ✓ The plethora of flat file formats creates problems in information exchange and pipeline construction.

Example of an EMBL record for an H5N1 polymerase acidic sequence

```
ID   AF046087; SV 1; linear; genomic RNA; STD; VRL; 2151 BP.
XX
AC   AF046087;
XX
DT   23-JUL-1998 (Rel. 56, Created)
DT   19-MAY-2005 (Rel. 83, Last updated, Version 4)
XX
DE   Influenza A virus (A/Chicken/Hong Kong/220/97 (H5N1)) polymerase acidic
DE   protein (PA) gene, partial cds.
XX
OS   Influenza A virus (A/Chicken/Hong Kong/220/97 (H5N1))
OC   Viruses; ssRNA negative-strand viruses; Orthomyxoviridae; Influenzavirus A.
XX
SQ   Sequence 2151 BP; 715 A; 435 C; 517 G; 484 T; 0 other;
     aaaatggaag actttgtgcg acaatgcttc aatccaatga ttgtcgagct tgcggaaaag        60
     acaatgaagg agtacgggga agatccgaaa attgaaacaa acaagttcgc tgcaatatgc       120
     acacacttag aagtctgctt catgtattca gacttccatt tcattgacga acgaggcgaa       180
     tcaataattg tggaatctgg tgatccgaat gcattgttga aacaccgatt tgaaataatt       240
     gaaggaagag accgagcaat ggcctggaca gtggtgaata gcatctgcaa caccacagga       300
.....
     accttcgatc ttgaagggct atatggagca attgaggagt gcctgattaa tgatccctgg      2100
     gttttgctta atgcatcttg gttcaactcc ttcctcacac atgcactaag a               2151
//
```

Example of a Genbank record for the same H5N1 polymerase acidic sequence

LOCUS AF046087 2151 bp RNA linear VRL 17-MAY-2005

DEFINITION Influenza A virus (A/Chicken/Hong Kong/220/97 (H5N1)) polymerase acidic protein (PA) gene, partial cds.

ACCESSION AF046087

VERSION AF046087.1 GI:3335418

KEYWORDS .

SOURCE Influenza A virus (A/Chicken/Hong Kong/220/97 (H5N1))

ORGANISM Influenza A virus (A/Chicken/Hong Kong/220/97 (H5N1))
Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
Influenzavirus A.

REFERENCE 1 (bases 1 to 2151)

...

ORIGIN

```
       1 aaaatggaag actttgtgcg acaatgcttc aatccaatga ttgtcgagct tgcggaaaag
      61 acaatgaagg agtacgggga agatccgaaa attgaaacaa acaagttcgc tgcaatatgc
     121 acacacttag aagtctgctt catgtattca gacttccatt tcattgacga acgaggcgaa
     181 tcaataattg tggaatctgg tgatccgaat gcattgttga aacaccgatt tgaaataatt
     241 gaaggaagag accgagcaat ggcctggaca gtggtgaata gcatctgcaa caccacagga
     301 gtcgataaac ccaaatttct tccggatcta tacgactaca aggaaaaccg attcactgaa
....
1981 gctgaatcga gaaaactact actcattgtt caagcactta gggacaacct ggaacctgga
2041 accttcgatc ttgaagggct atatggagca attgaggagt gcctgattaa tgatccctgg
2101 gttttgctta atgcatcttg gttcaactcc ttcctcacac atgcactaag a
```

Major sequence databases (4)

- ✓ Most known nucleotide database have Feature Table (FT) in their header. An FT is an annotation component that may note sequence regions that:
  - Perform or affect a function (like coding regions CDS)
  - interact with other molecules
  - affect replication
  - are involved in recombination
  - exhibit secondary or tertiary structure
  - are revised or corrected.

Example of an EMBL FT for the H5N1 polymerase acidic sequence

```
FT   source          1..2151
FT                   /organism="Influenza A virus (A/Chicken/Hong Kong/220/97
FT                   (H5N1))"
FT                   /strain="A/Chicken/Hong Kong/220/97"
FT                   /mol_type="genomic RNA"
FT                   /note="RT-PCR amplified and sequenced from virus grown in
FT                   embryonated chicken eggs"
FT                   /db_xref="taxon:100834"
FT   gene            4..>2151
FT                   /gene="PA"
FT   CDS             4..>2151
FT                   /codon_start=1
FT                   /gene="PA"
FT                   /product="polymerase acidic protein"
FT                   /db_xref="GOA:O89750"
FT                   /db_xref="InterPro:IPR001009"
FT                   /db_xref="UniProtKB/Swiss-Prot:O89750"
FT                   /protein_id="AAC32087.1"
FT                   /translation="MEDFVRQCFNPMIVELAEKTMKEYGEDPKIETNKFAAICTHLEVC
FT                   FMYSDFHFIDERGESIIVESGDPNALLKHRFEIIEGRDRAMAWTVVNSICNTTGVDKPK
FT                   FLPDLYDYKENRFTEIGVTRREIHIYYLEKANKIKSEKTHIHIFSFTGEEMATKADYTL
FT                   DEESRARIKTRLFTIRQEMASRGLWDSFRQSERGEETIEERFEITGTMRRLADQSLPPN.....
```

Example of a Genbank FT for the H5N1 polymerase acidic sequence

```
FEATURES             Location/Qualifiers
     source          1..2151
                     /organism="Influenza A virus (A/Chicken/Hong Kong/220/97
                     (H5N1))"
                     /mol_type="genomic RNA"
                     /strain="A/Chicken/Hong Kong/220/97"
                     /db_xref="taxon:100834"
                     /note="RT-PCR amplified and sequenced from virus grown in
                     embryonated chicken eggs"
     gene            4..>2151
                     /gene="PA"
     CDS             4..>2151
                     /gene="PA"
                     /codon_start=1
                     /product="polymerase acidic protein"
                     /protein_id="AAC32087.1"
                     /db_xref="GI:3335419"
                     /translation="MEDFVRQCFNPMIVELAEKTMKEYGEDPKIETNKFAAICTHLEV
                     CFMYSDFHFIDERGESIIVESGDPNALLKHRFEIIEGRDRAMAWTVVNSICNTTGVDK
                     PKFLPDLYDYKENRFTEIGVTRREIHIYYLEKANKIKSEKTHIHIFSFTGEEMATKAD
                     YTLDEESRARIKTRLFTIRQEMASRGLWDSFRQSERGEETIEERFEITGTMRRLADQS
                     LPPNFSSLENFRAYVDGFKPNGCIEGKLSQMSKEVNARIEPFLKTTPRPLRLPDGPPC
```

Major sequence databases (5)

- ✓ Observe similarities and differences in the header fields.
- ✓ Notable differences
  - EMBL has more structured format than Genbank which makes parsing easier.
  - A number of different fields and/or summary is expressed.
- ✓ Notable similarities
  - The accession number remains the same across these 3 databases.
  - They all convey more or less the same amount of information.

Major sequence databases (6)

- ✓ Notable protein sequences include:
  - **UniProt (Universal Protein Resource)**: It is the absolute reference that unifies the following databases under one umbrella:
    - **Swiss-Prot**: Fully annotated source of protein sequence information.
    - **TrEMBL (Translated EMBL)**: Take the EMBL nucleotide database, do a translation and have all the records as protein sequences.
    - **PIR (The Protein Information Resource)**: Also fully annotated with good references to x-ray crystallography and active site determination data.

http://www.uniprot.org/

Other notable databases

- ✓ Specialized **sequence related** databases include:
  - **IntePro**: A database of predictive protein signatures with annotation. It classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites.

    http://www.ebi.ac.uk/interpro/
  - **GOA**: Gene Ontology Annotated contains high quality Gene Ontology (GO) annotation records for proteins found in the UniProt database.

    http://www.ebi.ac.uk/GOA/

Other notable databases (2)

- **OMIM**: Are you a physician or genetics scientist concerned with genetic disorders? Online Mendelian Inheritance in Man (OMIM) contains on all known Mendelian disorders for more than 12000 genes.

http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim

- **Taxonomy**: Gene Ontology Annotated contains high quality Gene Ontology (GO) annotation records for proteins found in the UniProt database.

http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

Other notable databases (3)

- **IrefIndex**: Do you wish to have all known Protein-to-Protein interactions with data being consolidated from various interaction databases?

http://irefindex.uio.no/wiki/iRefIndex

✓ Note that some databases act as '*constellations*' of sequence related information (Uniprot combines SwissProt-TrEMBL-PIR, IrefIndex combines protein interaction databases such as BIND, BIOGRID, IntAct). Using these databases makes sense instead of looking into each of the consolidated databases separately.

Sequence formats:

- ✓ Biological sequences are encoded in specific formats
- ✓ Various utilities/applications understand one or more sequence formats.
- ✓ Translating sequences from one format to another is a necessity in the field of bioinformatics.
- ✓ The EMBOSS application 'seqret' (see latter slides) is the Swiss Army knife of sequence conversion.
- ✓ I present below some common sequence formats. The list is by no means exhaustive. For more info see:

http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

The FASTA (Pearson) format

>BRCA2_FELCA Q864S8 Breast cancer type 2 susceptibility protein homolog (Fanconi anemia group D1 protein homolog)

MPIGCKERPTFFEIFRTRCNKADLGPISLNWFEELCLEAPPYNSEPTEESGYKISYE
PNLFKTPQRKPCHQLASTPIIFKEQGLIPPIYQQSPLKELGKDITNSKHRSCCTMKS
KMDQTNDVTSPPLNSCLSESPLLRSTHVTPQREKSVVCGSLFHTPKLTKGQTPK
RISESLGAEVDPDMSWSSSLATPPTLSSTVLIVRDEEASAAVFPNDTTAIFKSYFC
NHDESLKKNDRFIPSGPDSENKSQREAKSQGLGKMVGNSCDKVNSCKDPFGN
STLNVLEDGVRERVADVSEEDSFP...

- ✓ Very commonly used in many tools.
- ✓ The simplest type of format.
- ✓ Contains always a single header line and the sequence data.
- ✓ Pioneered with the FASTA sequence alignment suite of programs

The SWISS format

ID BRCA2_FELCA Reviewed; 3372 AA.
AC Q864S8;
DT 07-MAR-2006, integrated into UniProtKB/Swiss-Prot.
DT 04-JAN-2005, sequence version 2.
DT 24-NOV-2009, entry version 36.
DE RecName: Full=Breast cancer type 2 susceptibility protein homolog;
...
SQ SEQUENCE 3372 AA; 377346 MW; 37F23DA23CA94665 CRC64;
MPIGCKERPT FFEIFRTRCN KADLGPISLN WFEELCLEAP PYNSEPTEES
GYKISYEPNL FKTPQRKPCH QLASTPIIFK EQGLIPPIYQ QSPLKELGKD
ITNSKHRSCC TMKSKMDQTN DVTSPPLNSC LSESPLLRST HVTPQREKSV
VCGSLFHTPK LTKGQTPKRI SESLGAEVDP DMSWSSSLAT PPTLSSTVLI
VRDEEASAAV FPNDTTAIFK SYFCNHDESL KKNDRFIPSG PDSENKSQRE
AKSQGLGKMV GNSCDKVNSC KDPFGNSTLN...

The ASN.1 format

seq {

id { local id 1 },

descr { title "Breast cancer type 2 susceptibility protein homolog (Fanconi anemia group D1 protein homolog)" },

inst {

repr raw, mol aa, length 3372, topology linear,

{

seq-data

Iupacaa"MPIGCKERPTFFEIFRTRCNKADLGPISLNWFEELCLEAPPYNSEPTEESGYKISYE
PNLFKTPQRKPCHQLASTPIIFKEQGLIPPIYQQSPLKELGKDITNSKHRSCCTMKSKMDQTN
DVTSPPLNSCLSESPLLRSTHVTPQREKSVVCGSLFHTPKLTKGQTPKRISESLGAEVDPDM
SWSSSLATPPTLSSTVLIVRDEEASAAVFPNDTTAIFKSYFCNHDESLKKNDRFIPSGPDSEN
KSQREAKSQGLGKMVGNSCDKVNSCKDPFGNSTLNV...”

} } ,

The MSF format

```
!!AA_MULTIPLE_ALIGNMENT 1.0

  brcas.msf MSF:  3372 Type: P 16/03/10 CompCheck:  694 ..

  Name: BRCA2_FELCA Len: 3372  Check: 7262 Weight: 1.00
  Name: BRCA2_RAT   Len: 3372  Check: 3432 Weight: 1.00
//
         1                                        50
BRCA2_FELCA MPIGCKERPTFFEIFRTRCNKADLGPISLNWFEELCLEAPPYNSEPTEES
BRCA2_RAT   MTVEYKRRPTFWEIFKARCSTADLGPISLNWFEELFSEAPPYNTEHPEES
         51                                       100
BRCA2_FELCA GYKISYEPNLFKTPQRKPCHQLASTPIIFKEQGLIPPIYQQSPLKELGKD
BRCA2_RAT   EYKPQGHEPQLFKTPQRNPSYHQFASTPIMFKEQSQTLPLDQSPFKELGN
         101                                      150
BRCA2_FELCA ITNSKHRSCCTMKSKMDQTNDVTSPPLNSCLSESPLLRSTHVTPQREKSV
BRCA2_RAT   VVANSKRKHHSKKKARKDPVVDVASLPLKACPSESPCTPRCTQVAPQRRK
```

Web interfaces

- ✓ Many times it is desirable to search for sequences in a web interface, and have the data presented for you.
- ✓ There are several services that offer this, we will look at two – SRS and MRS
- ✓ Intuitive, easy to use
- ✓ However, what if you want to repeat the process many times AND/OR connect the process to other programs?

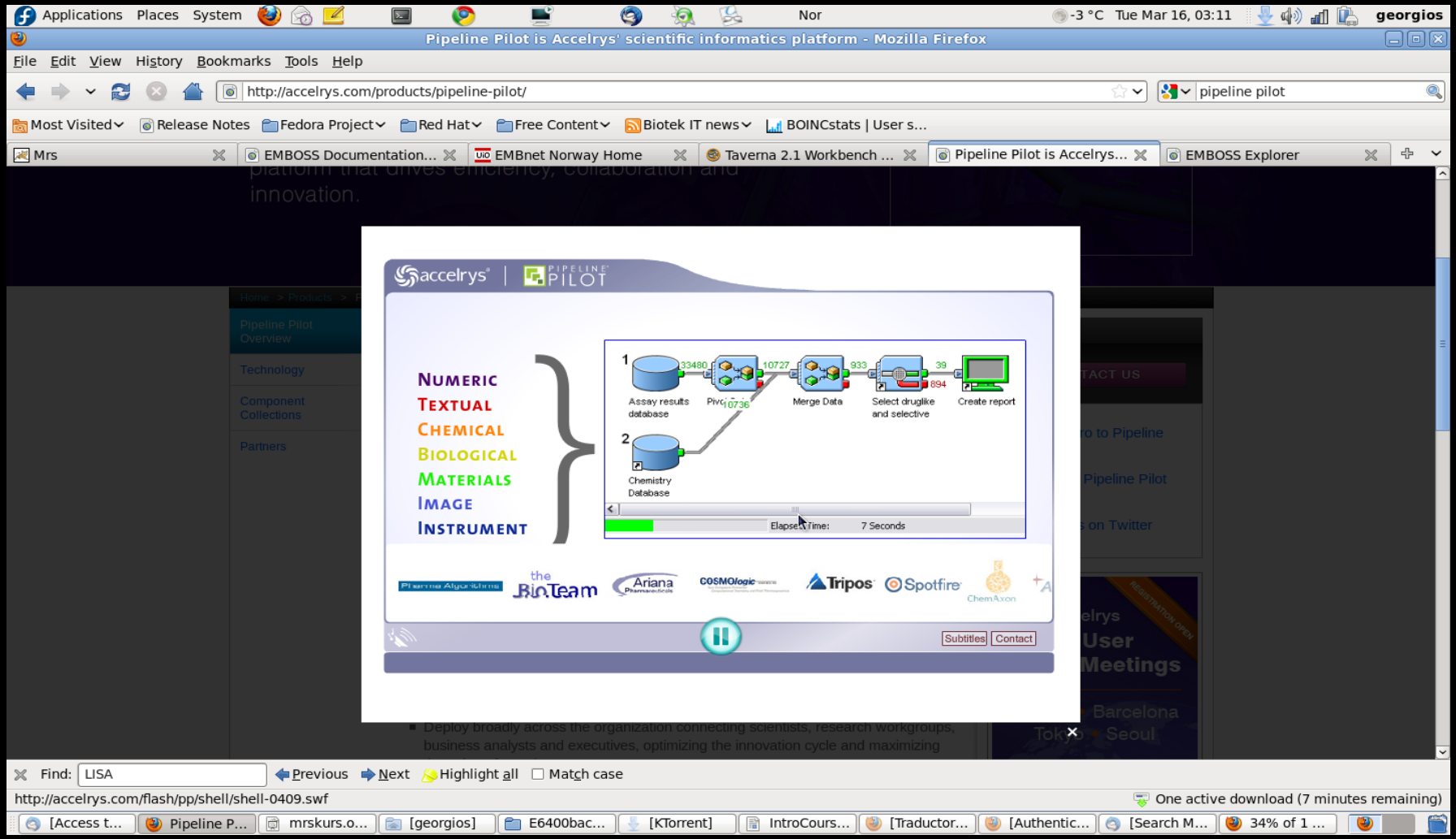
Command Line Interfaces

- ✓ For power-users, the best way of using any system is via a command line interface.
- ✓ Steep learning curve but once mastered, a command line interface is a powerful tool.
- ✓ It allows you to script and repeat processes and/or access data sets programmatically
- ✓ Both MRS and EMBOSS contrain powerful command line features that facilitate the contruction of **pipelines** (some people use the term workflow):
  - *<input>program<output>program2<output2>*

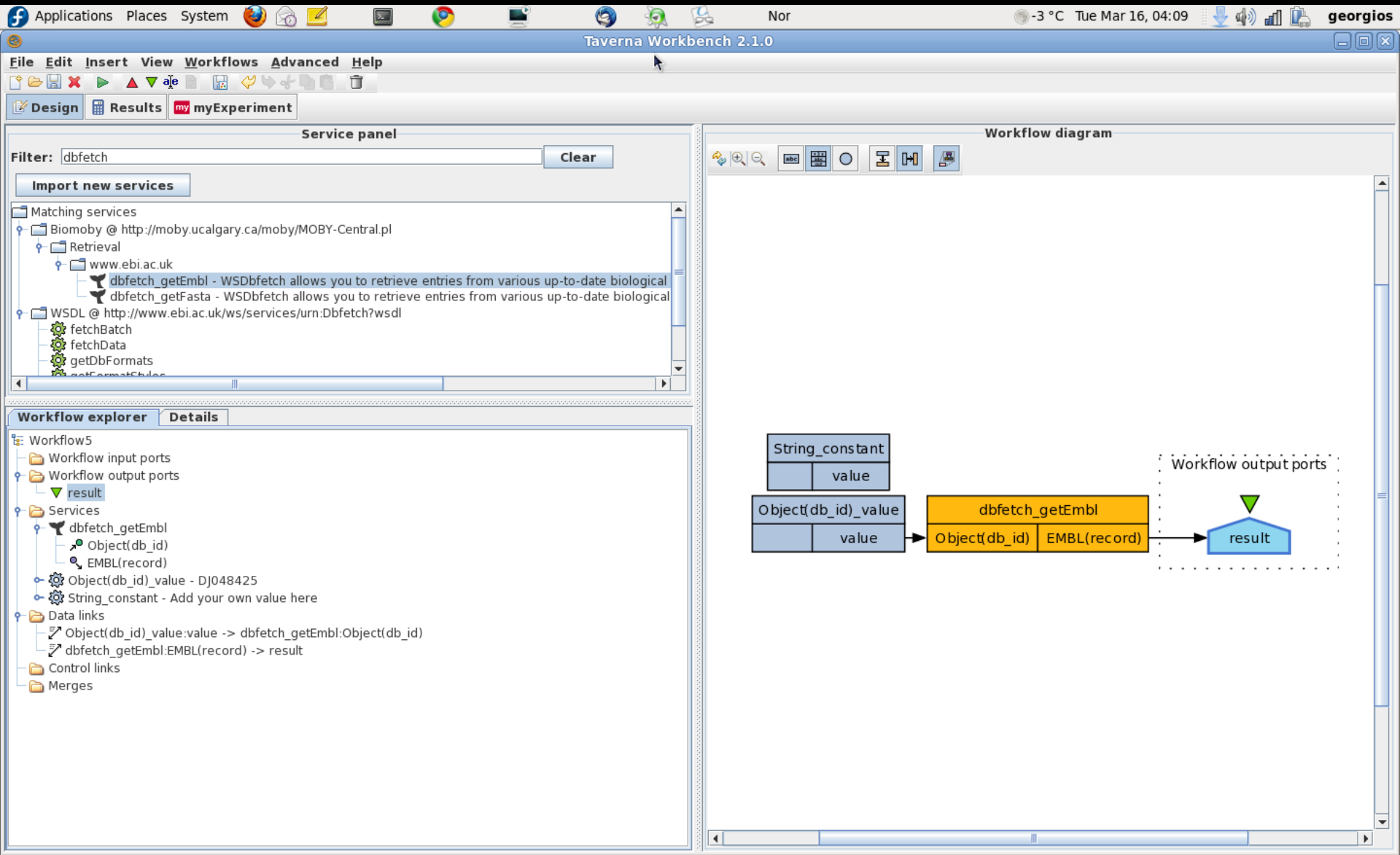
Alternatives to command-line pipeline construction:

- ✓ Can I build pipelines (workflows) graphically without having to resort to command line wizardry?

Accelrys Pipe-line pilot
Image (C) Accelrys corporation



Taverna interface



Graphical workflow systems

- ✓ **Are in principle a nice idea because:**
  - *they conceptualize a workflow graphically making it easier to grasp.*
  - *They include verified (ready-made) components and allow you to plug and play applications with ease.*
- ✓ **However, they have disadvantages:**
  - *They are computing intensive (Taverna requires at least 1.6 Gigs of RAM just to run), making it difficult to run large datasets.*
  - *Commercial versions are expensive (thousands of dollars for license fees per user).*

EMBOSS

- ✓ **European Molecular Biology Open Software Suite**
- ✓ The latest release (6.2) contains hundreds of applications for manipulating and analyzing biological sequences. These applications are often referenced in different categories:

http://emboss.sourceforge.net/apps/release/6.2/emboss/apps/groups.html

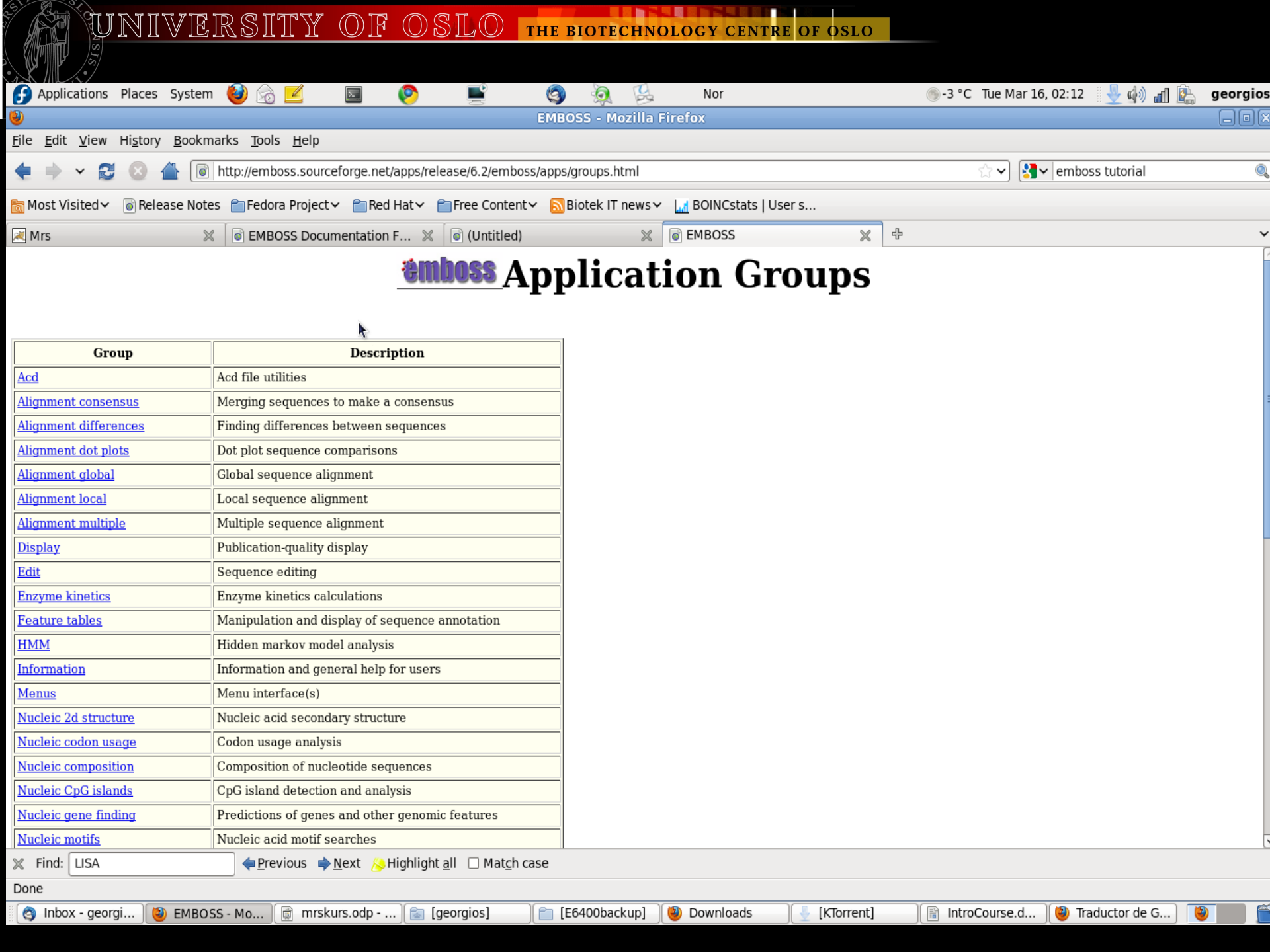
EMBOSS (2)

- ✓ Initially command-line driven.
- ✓ There are also graphical user interfaces made for it:
  - *JEMBOSS: Java driven, maintained by the EMBOSS team, but resource intensive and a bit inflexible.*

    http://emboss.sourceforge.net/Jemboss/
  - *wEMBOSS: Based on web-browser standards, more flexible and comprehensive.*

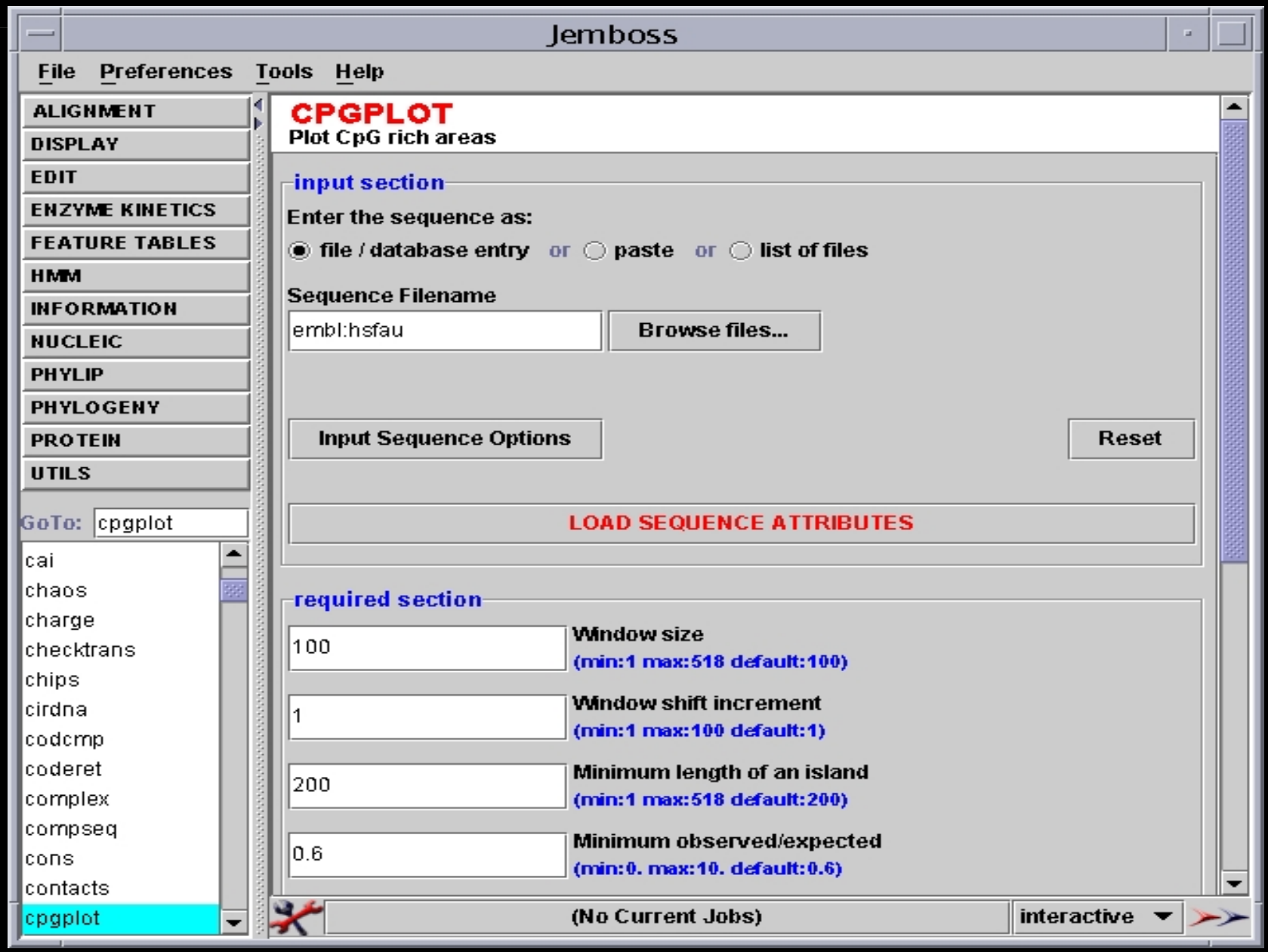
    http://wemboss.sourceforge.net/
  - *EMBOSS EXPLORER: Simple interface, requiring no authentication.*

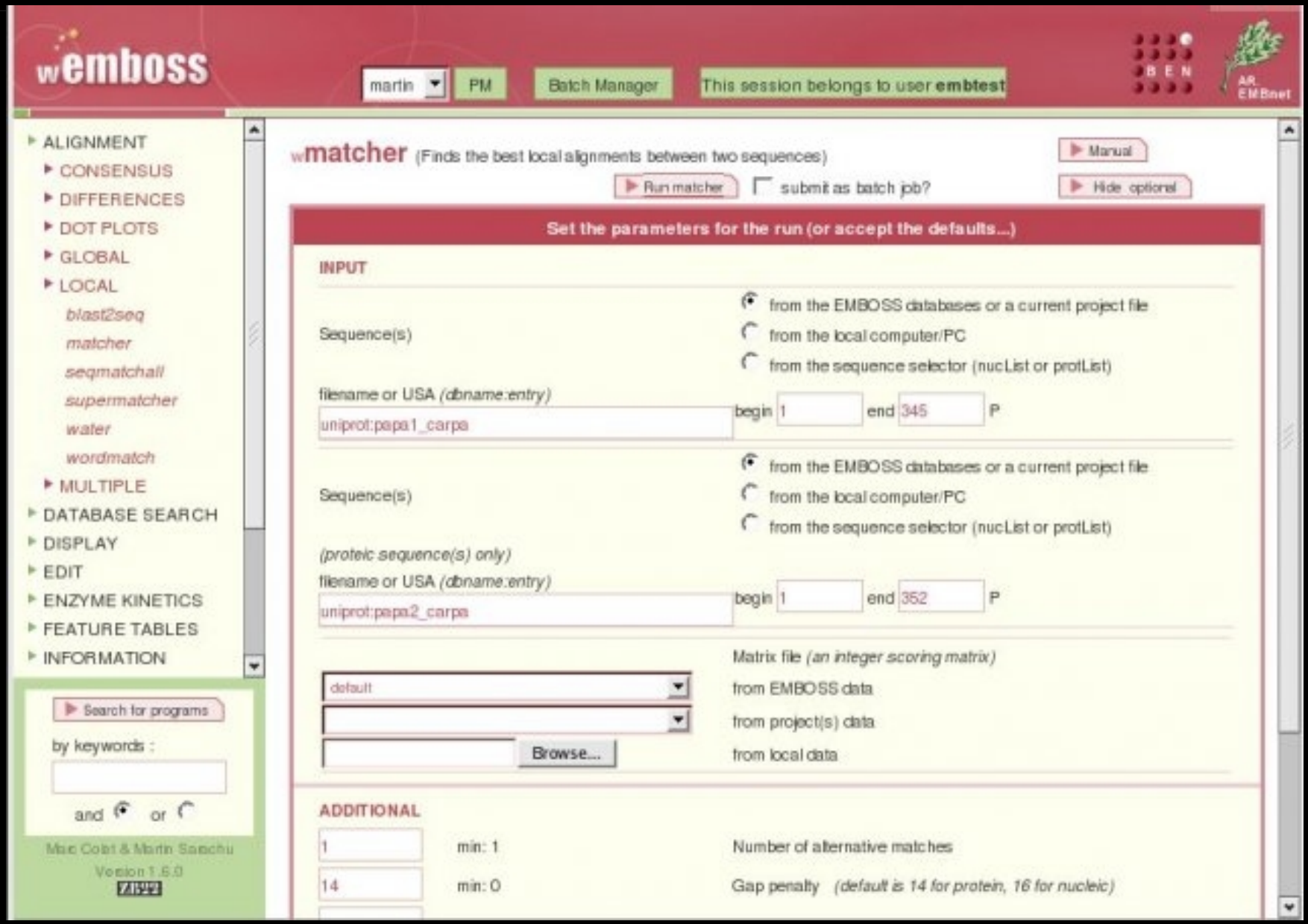
    http://cnkeeper.uio.no/

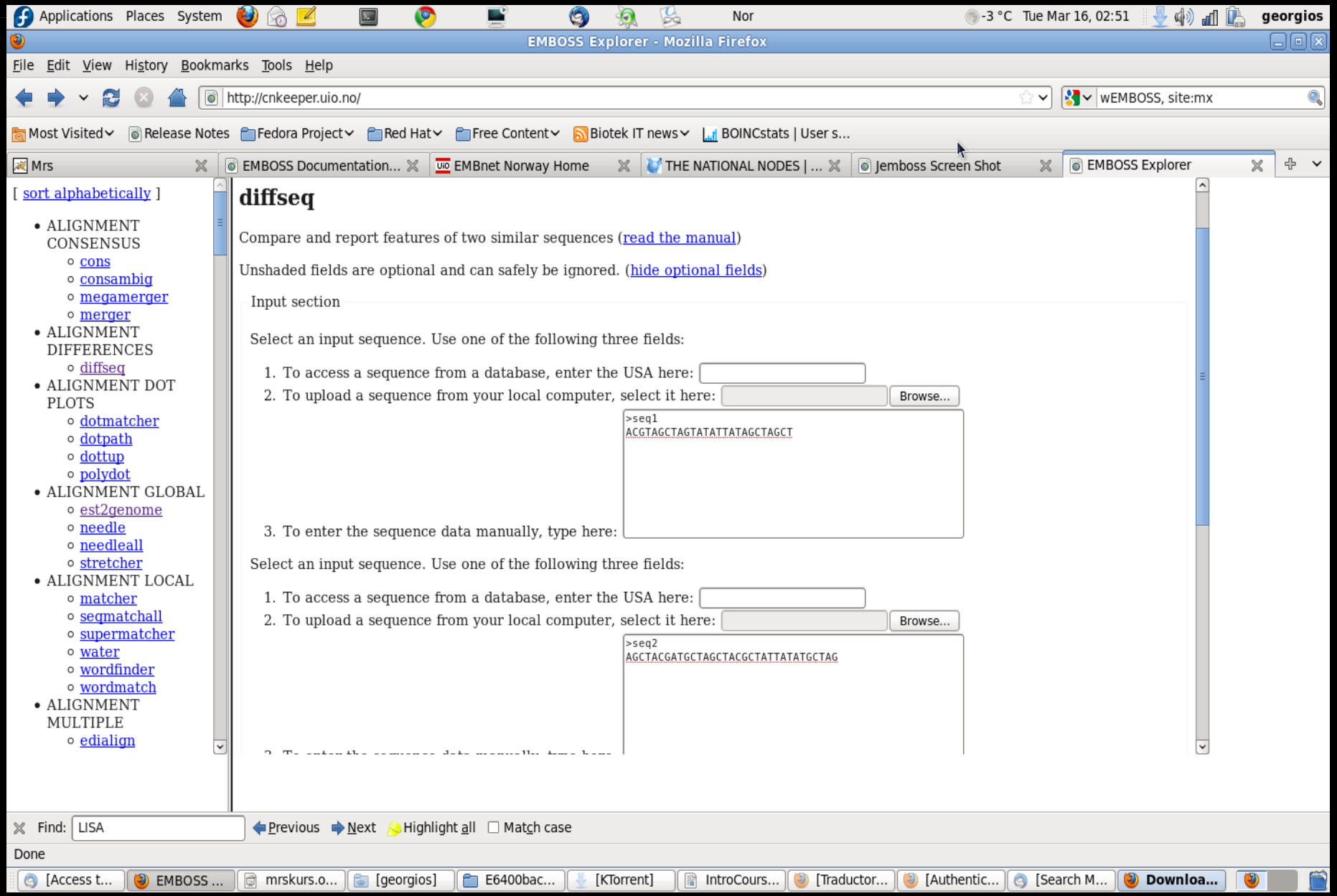


emboss Application Groups

Group	Description
Acd	Acd file utilities
Alignment consensus	Merging sequences to make a consensus
Alignment differences	Finding differences between sequences
Alignment dot plots	Dot plot sequence comparisons
Alignment global	Global sequence alignment
Alignment local	Local sequence alignment
Alignment multiple	Multiple sequence alignment
Display	Publication-quality display
Edit	Sequence editing
Enzyme kinetics	Enzyme kinetics calculations
Feature tables	Manipulation and display of sequence annotation
HMM	Hidden markov model analysis
Information	Information and general help for users
Menus	Menu interface(s)
Nucleic 2d structure	Nucleic acid secondary structure
Nucleic codon usage	Codon usage analysis
Nucleic composition	Composition of nucleotide sequences
Nucleic CpG islands	CpG island detection and analysis
Nucleic gene finding	Predictions of genes and other genomic features
Nucleic motifs	Nucleic acid motif searches


Jemboss
File Preferences Tools Help
ALIGNMENT
DISPLAY
EDIT
ENZYME KINETICS
FEATURE TABLES
HMM
INFORMATION
NUCLEIC
PHYLIP
PHYLOGENY
PROTEIN
UTILS
GoTo: cpgplot
cai
chaos
charge
checktrans
chips
cirdna
codcmp
coderet
complex
compseq
cons
contacts
cpgplot
CPGPLOT
Plot CpG rich areas
input section
Enter the sequence as:
file / database entry or paste or list of files
Sequence Filename
embl:hsfau
Browse files...
Input Sequence Options
Reset
LOAD SEQUENCE ATTRIBUTES
required section
100
Window size
(min:1 max:518 default:100)
1
Window shift increment
(min:1 max:100 default:1)
200
Minimum length of an island
(min:1 max:518 default:200)
0.6
Minimum observed/expected
(min:0. max:10. default:0.6)
(No Current Jobs)
interactive


wemboss
martin
PM
Batch Manager
This session belongs to user embtest
ALIGNMENT
CONSENSUS
DIFFERENCES
DOT PLOTS
GLOBAL
LOCAL
blast2seq
matcher
seqmatchall
supermatcher
water
wordmatch
MULTIPLE
DATABASE SEARCH
DISPLAY
EDIT
ENZYME KINETICS
FEATURE TABLES
INFORMATION
Search for programs
by keywords :
and
or
Marc Colet & Martin Sanchu
Version 1.6.0
wmatcher (Finds the best local alignments between two sequences)
Manual
Run matcher
submit as batch job?
Hide optional
Set the parameters for the run (or accept the defaults...)
INPUT
Sequence(s)
from the EMBOSS databases or a current project file
from the local computer/PC
from the sequence selector (nucList or protList)
filename or USA (dbname:entry)
uniprot:papa1_carpa
begin 1
end 345
P
Sequence(s)
from the EMBOSS databases or a current project file
from the local computer/PC
from the sequence selector (nucList or protList)
(proteic sequence(s) only)
filename or USA (dbname:entry)
uniprot:papa2_carpa
begin 1
end 352
P
Matrix file (an integer scoring matrix)
default
from EMBOSS data
from project(s) data
Browse...
from local data
ADDITIONAL
1
min: 1
Number of alternative matches
14
min: 0
Gap penalty (default is 14 for protein, 16 for nucleic)


Applications Places System
-3 °C Tue Mar 16, 02:51
georgios
EMBOSS Explorer - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://cnkeeper.uio.no/
wEMBOSS, site:mx
Most Visited Release Notes Fedora Project Red Hat Free Content Biotek IT news BOINCstats | User s...
Mrs
EMBOSS Documentation...
EMBnet Norway Home
THE NATIONAL NODES | ...
Jemboss Screen Shot
EMBOSS Explorer
[sort alphabetically]
ALIGNMENT CONSENSUS
cons
consambig
megamerger
merger
ALIGNMENT DIFFERENCES
diffseq
ALIGNMENT DOT PLOTS
dotmatcher
dotpath
dottup
polydot
ALIGNMENT GLOBAL
est2genome
needle
needleall
stretcher
ALIGNMENT LOCAL
matcher
seqmatchall
supermatcher
water
wordfinder
wordmatch
ALIGNMENT MULTIPLE
edialign
diffseq
Compare and report features of two similar sequences (read the manual)
Unshaded fields are optional and can safely be ignored. (hide optional fields)
Input section
Select an input sequence. Use one of the following three fields:
1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Browse...
>seq1
ACGTAGCTAGTATATTATAGCTAGCT
3. To enter the sequence data manually, type here:
Select an input sequence. Use one of the following three fields:
1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Browse...
>seq2
AGCTACGATGCTAGCTACGCTATTATATGCTAG
Find: LISA
Previous Next Highlight all Match case
Done
[Access t... EMBOSS ... mrskurs.o... [georgios] E6400bac... [KTorrent] IntroCours... [Traductor... [Authentic... [Search M... Downloa...

EMBOSS (3)

- ✓ **How do I get information about a particular emboss application?**
  - *tfm application_name*
    - Example: **tfm seqret**
- ✓ **What if I do not know the application name, but I know more or less what I want to do with that application?**
  - *wossname (single_keyword)*
    - Example: **wossname translate** (if I wanted to find the names of apllications that perform sequence translation)

EMBOSS: The USA format

- ✓ **The Uniform Sequence Address USA format**
  - *It is an EMBOSS convention to specify sequences and their formats.*
  - *Many (if not all) emboss applications support the same sequence (input, output) and feature formats.*
  - *Following the USA convention makes sure that we can use all/most emboss applications in a consistent manner.*

EMBOSS: The USA format (2)

USA type	Command example	Explanation
“file”	seqret brcas.fasta	Reference the file brcas.fasta (as input)
“format::file”	seqret brcas.fasta swiss::brcas.txt	Read the file brcas.fasta and convert the fasta format to SWISS format in the file brcas.txt
"format::file:entry"	infoseq fasta::brcas.fasta:BRCA2_RAT	From the file brcas.fasta which is in fasta format, give me sequence info about the entry with ID BRCA2_RAT
"dbname:entry"	seqret sprot:BRCA2_RAT	From the database sprot retrieve the entry with ID BRCA2_RAT
"@listfile"	infoseq @myfileids.txt	Retrieve info about the sequences specified in the file myfileids.txt

The EMBOSS index

- ✓ **EMBOSS indexes six different fields:**
  - ***id****: the sequence ID*
  - ***acc****: the sequence accession number*
  - ***sv****: The sequence version*
  - ***key****: The keyword(s) of the sequence*
  - ***des****: Word(s) in the description field*
  - ***org****: Taxonomy/organism species info*
- ✓ **All of these fields are part of the sequence header.**
- ✓ **Most core EMBOSS applications understand the USA format and these index names.**

EMBOSS – 'seqret'

- ✓ Retrieves sequences from a specified data bank, matching one specified index at a time.
- ✓ Output from this will then have to be searched for further matches.
- ✓ The main application from mining sequences in EMBOSS.

Using the index in USA expressions

- ✓ Allows us to form meaningful queries.
- ✓ Forms the basis of sequence mining in EMBOSS.

Expression	Explanation
seqret viral**-des**:'H1N1'	Give me all sequences of the 'viral' database that have the string 'H1N1' in the description field of the sequence header
infoseq prokaryotic**-key**:'phosphorylation'	Give me info on all sequences of the 'prokaryotic' database that have the word 'phosphorylation' in the keyword sequence header field
backtranseq trembl**-des**:globin embl::globins.embl	1.Select all sequences from 'trembl' that have the word 'globin' in their description 2. Convert them back to nucleotide sequences. 3.Put the results in the 'globins.embl' file in EMBL format.

seqret example

Terminal

File Edit View Terminal Tabs Help

```
lingeek@cnkeeper /biotek/cnkeeper/storage/databases/release102 $ seqret viral-des:H1N1 ~/h1n1.fasta
Reads and writes (returns) sequences
lingeek@cnkeeper /biotek/cnkeeper/storage/databases/release102 $ cat ~/h1n1.fasta | grep ^">" | grep -i "human" > ~/h1n1human.f
asta
lingeek@cnkeeper /biotek/cnkeeper/storage/databases/release102 $ head -25 ~/h1n1human.fasta
>EU032585 EU032585.1 Influenza A virus (A/human/Pandharpur/20055210/2005(H1N1)) hemagglutinin (HA1) gene, partial cds.
>EU032584 EU032584.1 Influenza A virus (A/human/Pandharpur/20055200/2005(H1N1)) hemagglutinin (HA1) gene, partial cds.
>EU032583 EU032583.1 Influenza A virus (A/human/Pandharpur/20055194/2005(H1N1)) hemagglutinin (HA1) gene, partial cds.
>EU032582 EU032582.1 Influenza A virus (A/human/Pandharpur/20055182/2005(H1N1)) hemagglutinin (HA1) gene, partial cds.
>EU032581 EU032581.1 Influenza A virus (A/human/Pandharpur/20055180/2005(H1N1)) hemagglutinin (HA1) gene, partial cds.
>EU032580 EU032580.1 Influenza A virus (A/human/Pandharpur/20055169/2005(H1N1)) hemagglutinin (HA1) gene, partial cds.
>AF362803 AF362803.1 Influenza A virus (A/human/Taiwan/0012/00(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362802 AF362802.1 Influenza A virus (A/human/Taiwan/5063/99(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362801 AF362801.1 Influenza A virus (A/human/Taiwan/4943/99(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362800 AF362800.1 Influenza A virus (A/human/Taiwan/4845/99(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362799 AF362799.1 Influenza A virus (A/human/Taiwan/4415/99(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362798 AF362798.1 Influenza A virus (A/human/Taiwan/4360/99(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362797 AF362797.1 Influenza A virus (A/human/Taiwan/3825/00(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362796 AF362796.1 Influenza A virus (A/human/Taiwan/3355/97(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362795 AF362795.1 Influenza A virus (A/human/Taiwan/2200/95(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362794 AF362794.1 Influenza A virus (A/human/Taiwan/1190/95(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362793 AF362793.1 Influenza A virus (A/human/Taiwan/1184/99(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362792 AF362792.1 Influenza A virus (A/human/Taiwan/0892/99(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362791 AF362791.1 Influenza A virus (A/human/Taiwan/0657/95(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362790 AF362790.1 Influenza A virus (A/human/Taiwan/0563/95(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362789 AF362789.1 Influenza A virus (A/human/Taiwan/0562/95(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362788 AF362788.1 Influenza A virus (A/human/Taiwan/0464/99(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362787 AF362787.1 Influenza A virus (A/human/Taiwan/0342/96(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362786 AF362786.1 Influenza A virus (A/human/Taiwan/0337/96(H1N1)) hemagglutinin (HA) mRNA, partial cds.
>AF362785 AF362785.1 Influenza A virus (A/human/Taiwan/0255/96(H1N1)) hemagglutinin (HA) mRNA, partial cds.
lingeek@cnkeeper /biotek/cnkeeper/storage/databases/release102 $
```

Limitations of EMBOSS sequence mining

- ✓ It does not allow complex 'relational ' style queries:
  - *Give me all the -des:'phosphorylation' sequences that have -org:human.*
- ✓ It can be slow.
- ✓ Nevetheless useful for creating filtered datasets/databases.
- ✓ That's why we will look seriously at MRS in latter slides.

How to make your own EMBOSS databases:

We will look at an exercise in the tutorial, but the procedure in plain English is given below:

- ✓ Find a representative name for your database/dataset.
- ✓ Filter out the sequences that you wish with 'seqret'.
- ✓ Modify the .embossrc file in your home area to write the definition of your database.
- ✓ Invoke the 'dbxflat' utilities to produce the index files.
- ✓ Verify with 'showdb' that your database is there and start using it.

Sequence retrieval system

- ✓ EMBOSS 'seqret' is not a proper sequence retrieval system. In contrast, the following are:
  - *SRS (Sequence Retrieval System)* *http://srs.ebi.ac.uk/*
  - *Entrez cross-database reference* http://www.ncbi.nlm.nih.gov/sites/gquery
  - MRS (Maarten's Retrieval System) http://cnkeeper.uio.no:8080/mrs-web/

Sequence Retrieval Systems (2)

- ✓ **What's a 'proper' sequence retrieval system?**
  - *One that allows the usage of relational/boolean queries:*

    Example: 'of all the H1N1 sequences, give me only those that belong to the human species and do not include the keyword 'neuradaminase' in the description OR keyword fields'

$$\sigma_{a\theta b}(R) = \{\ t : t \in R,\ t(a)\ \theta\ t(b)\ \}$$

Sequence Retrieval Systems (3)

- ✓ **What's a 'proper' sequence retrieval system?**
  - *One that allows to examine a plethora of datasets including all major biological sequence databanks.*
  - *One that returns results in reasonable time.*
  - *One that has a simple and intuitive interface to construct complex queries and transfer the results in different sequence formats.*
  - *One that has a command line client to facilitate query and result storage in batch mode (pipeline construction)*

Sequence Retrieval Systems (4)

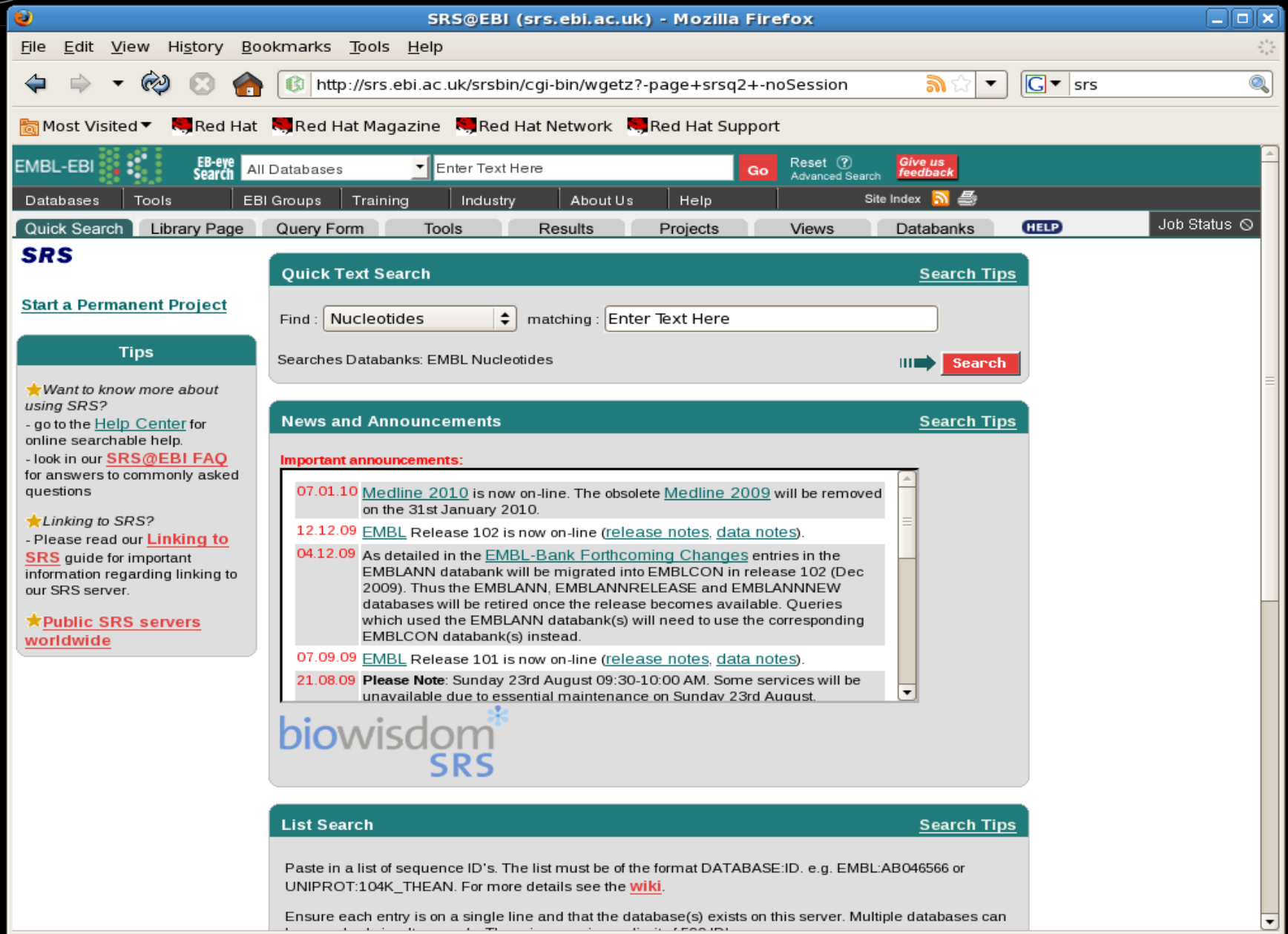
System	Boolean-Relational Queries	Major Databases	result production times	Command Line client*	Easy/intuitive interface
SRS	Yes	Yes	medium/high	yes	no
Entrez	Yes	Yes	low	no	yes
MRS	Yes	Yes	low	**yes**	yes

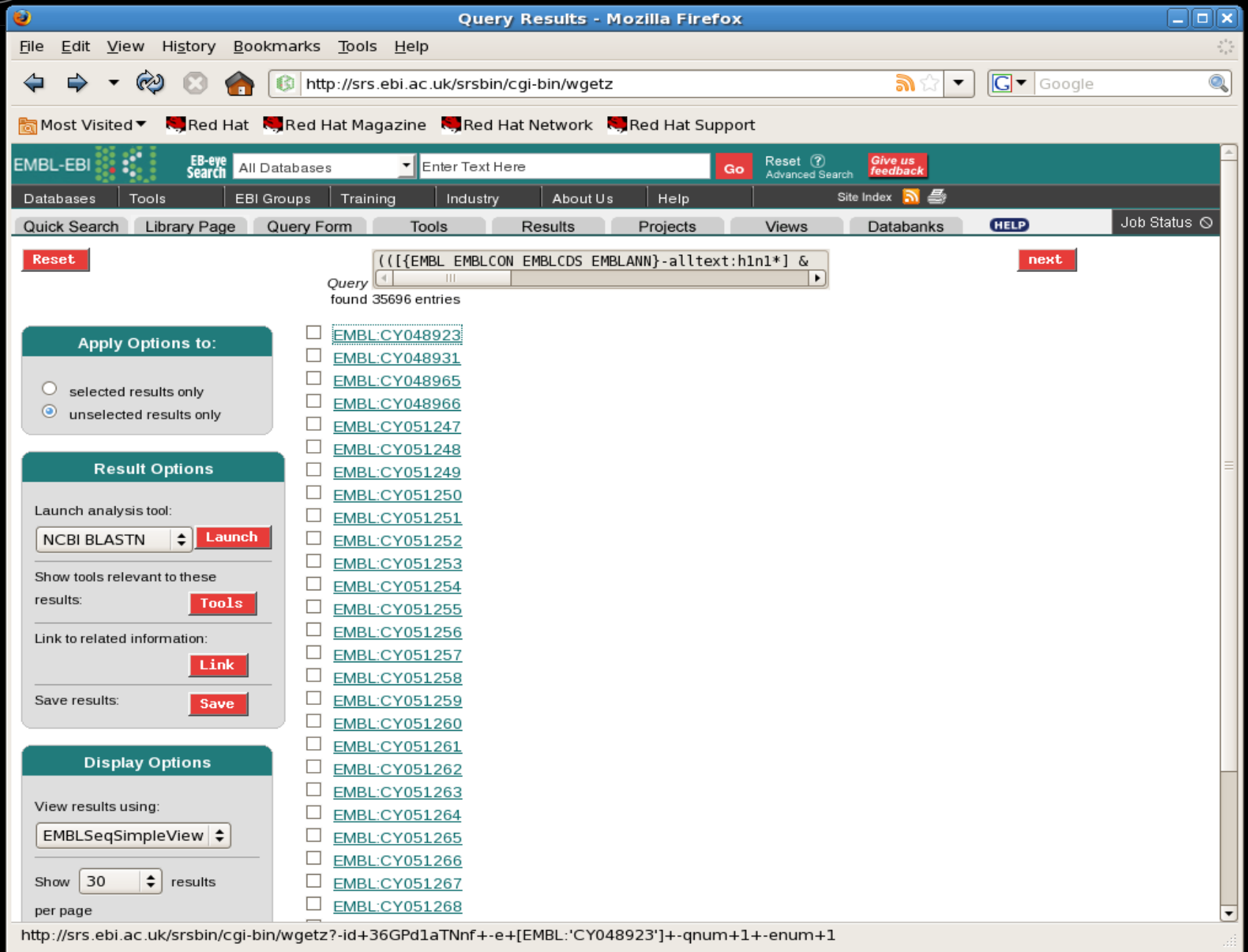
In addition, MRS is free and open source.

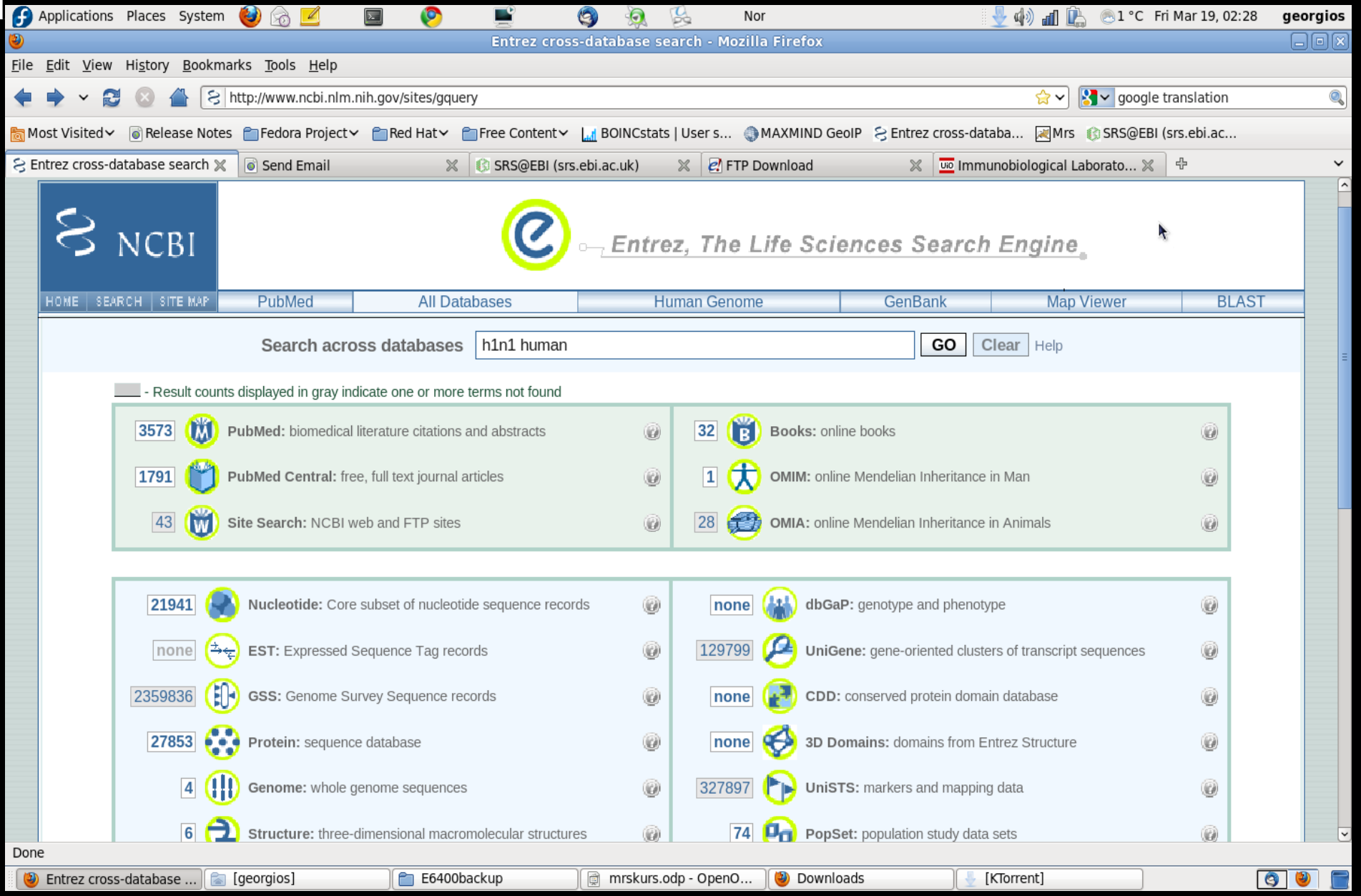
*The command line client does not examine software based on SOAP/REST web services.

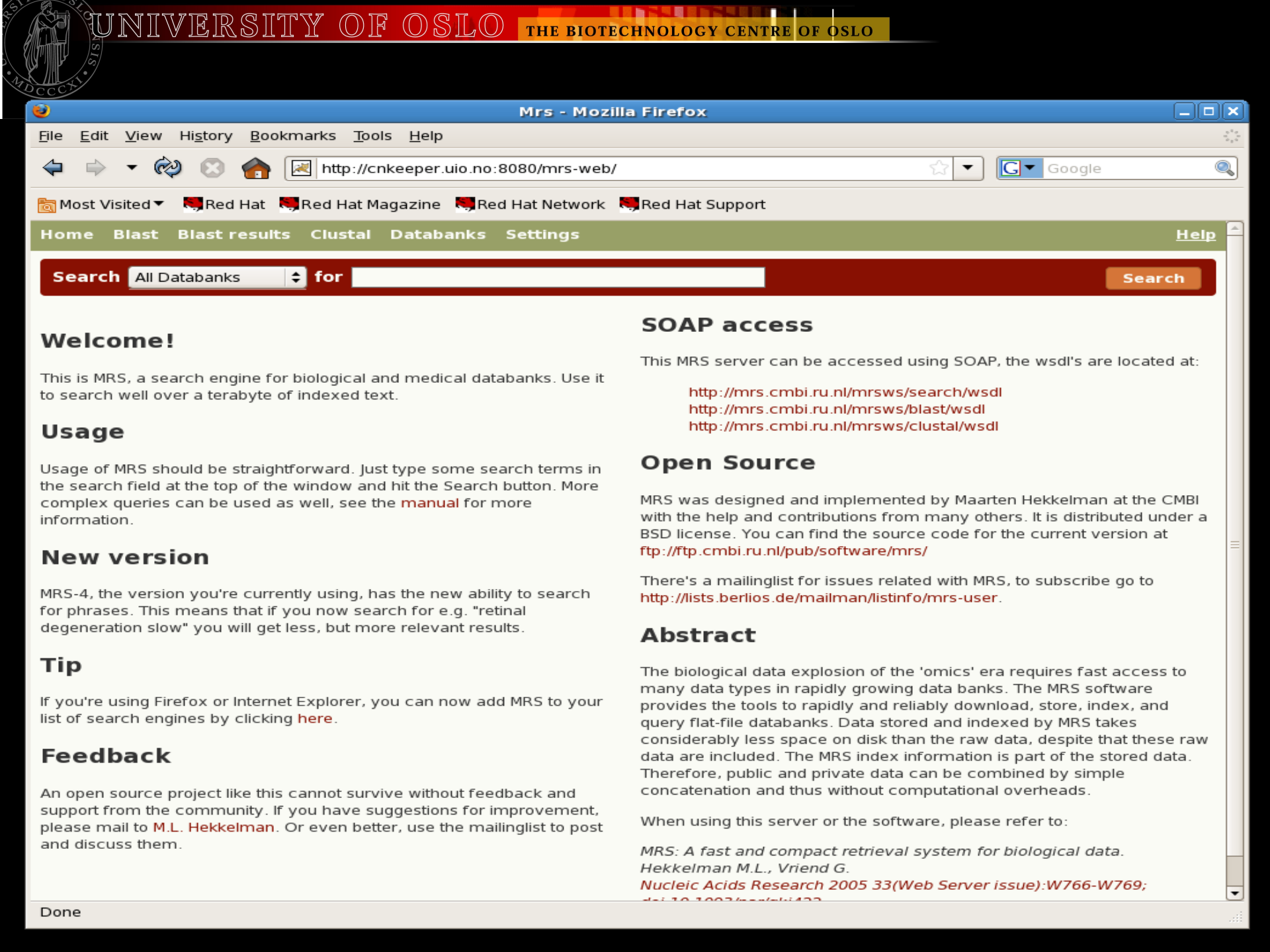
SRS - Sequence Retrieval System

- ✓ Web interface for retrieving sequences that match a given pattern
- ✓ One of the most used: http://srs.ebi.ac.uk/
- ✓ Typical time to retrieve the examples listed: 1 minutes 30 seconds


SRS@EBI (srs.ebi.ac.uk) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession
srs
Most Visited Red Hat Red Hat Magazine Red Hat Network Red Hat Support
EMBL-EBI
EB-eye Search
All Databases
Enter Text Here
Go
Reset
Advanced Search
Give us feedback
Databases Tools EBI Groups Training Industry About Us Help
Site Index
Quick Search Library Page Query Form Tools Results Projects Views Databanks HELP
Job Status
SRS
Start a Permanent Project
Tips
Want to know more about using SRS?
- go to the Help Center for online searchable help.
- look in our SRS@EBI FAQ for answers to commonly asked questions
Linking to SRS?
- Please read our Linking to SRS guide for important information regarding linking to our SRS server.
Public SRS servers worldwide
Quick Text Search
Search Tips
Find : Nucleotides
matching : Enter Text Here
Searches Databanks: EMBL Nucleotides
Search
News and Announcements
Search Tips
Important announcements:
07.01.10 Medline 2010 is now on-line. The obsolete Medline 2009 will be removed on the 31st January 2010.
12.12.09 EMBL Release 102 is now on-line (release notes, data notes).
04.12.09 As detailed in the EMBL-Bank Forthcoming Changes entries in the EMBLANN databank will be migrated into EMBLCON in release 102 (Dec 2009). Thus the EMBLANN, EMBLANNRELEASE and EMBLANNNEW databases will be retired once the release becomes available. Queries which used the EMBLANN databank(s) will need to use the corresponding EMBLCON databank(s) instead.
07.09.09 EMBL Release 101 is now on-line (release notes, data notes).
21.08.09 Please Note: Sunday 23rd August 09:30-10:00 AM. Some services will be unavailable due to essential maintenance on Sunday 23rd August.
biowisdom SRS
List Search
Search Tips
Paste in a list of sequence ID's. The list must be of the format DATABASE:ID. e.g. EMBL:AB046566 or UNIPROT:104K_THEAN. For more details see the wiki.
Ensure each entry is on a single line and that the database(s) exists on this server. Multiple databases can
Done


Query Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz
Google
Most Visited Red Hat Red Hat Magazine Red Hat Network Red Hat Support
EMBL-EBI
EB-eye Search
All Databases
Enter Text Here
Go
Reset
Advanced Search
Give us feedback
Databases Tools EBI Groups Training Industry About Us Help
Site Index
Quick Search Library Page Query Form Tools Results Projects Views Databanks HELP Job Status
Reset
(([{EMBL EMBLCON EMBLCDS EMBLANN}-alltext:h1n1*] &
next
Query
found 35696 entries
Apply Options to:
selected results only
unselected results only
Result Options
Launch analysis tool:
NCBI BLASTN
Launch
Show tools relevant to these results:
Tools
Link to related information:
Link
Save results:
Save
Display Options
View results using:
EMBLSeqSimpleView
Show 30 results per page
EMBL:CY048923
EMBL:CY048931
EMBL:CY048965
EMBL:CY048966
EMBL:CY051247
EMBL:CY051248
EMBL:CY051249
EMBL:CY051250
EMBL:CY051251
EMBL:CY051252
EMBL:CY051253
EMBL:CY051254
EMBL:CY051255
EMBL:CY051256
EMBL:CY051257
EMBL:CY051258
EMBL:CY051259
EMBL:CY051260
EMBL:CY051261
EMBL:CY051262
EMBL:CY051263
EMBL:CY051264
EMBL:CY051265
EMBL:CY051266
EMBL:CY051267
EMBL:CY051268
http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+36GPd1aTNnf+-e+[EMBL:'CY048923']+-qnum+1+-enum+1


Applications Places System
Nor
1 °C Fri Mar 19, 02:28 georgios
Entrez cross-database search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/sites/gquery
google translation
Most Visited Release Notes Fedora Project Red Hat Free Content BOINCstats | User s... MAXMIND GeoIP Entrez cross-databa... Mrs SRS@EBI (srs.ebi.ac...
Entrez cross-database search
Send Email
SRS@EBI (srs.ebi.ac.uk)
FTP Download
Immunobiological Laborato...
NCBI
Entrez, The Life Sciences Search Engine
HOME SEARCH SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
h1n1 human
GO
Clear
Help
- Result counts displayed in gray indicate one or more terms not found
3573 PubMed: biomedical literature citations and abstracts
1791 PubMed Central: free, full text journal articles
43 Site Search: NCBI web and FTP sites
32 Books: online books
1 OMIM: online Mendelian Inheritance in Man
28 OMIA: online Mendelian Inheritance in Animals
21941 Nucleotide: Core subset of nucleotide sequence records
none EST: Expressed Sequence Tag records
2359836 GSS: Genome Survey Sequence records
27853 Protein: sequence database
4 Genome: whole genome sequences
6 Structure: three-dimensional macromolecular structures
none dbGaP: genotype and phenotype
129799 UniGene: gene-oriented clusters of transcript sequences
none CDD: conserved protein domain database
none 3D Domains: domains from Entrez Structure
327897 UniSTS: markers and mapping data
74 PopSet: population study data sets
Done
Entrez cross-database ...
[georgios]
E6400backup
mrskurs.odp - OpenO...
Downloads
[KTorrent]


Mrs - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://cnkeeper.uio.no:8080/mrs-web/
Google
Most Visited Red Hat Red Hat Magazine Red Hat Network Red Hat Support
Home Blast Blast results Clustal Databanks Settings Help
Search All Databanks for Search
Welcome!
This is MRS, a search engine for biological and medical databanks. Use it to search well over a terabyte of indexed text.
Usage
Usage of MRS should be straightforward. Just type some search terms in the search field at the top of the window and hit the Search button. More complex queries can be used as well, see the manual for more information.
New version
MRS-4, the version you're currently using, has the new ability to search for phrases. This means that if you now search for e.g. "retinal degeneration slow" you will get less, but more relevant results.
Tip
If you're using Firefox or Internet Explorer, you can now add MRS to your list of search engines by clicking here.
Feedback
An open source project like this cannot survive without feedback and support from the community. If you have suggestions for improvement, please mail to M.L. Hekkelman. Or even better, use the mailinglist to post and discuss them.
SOAP access
This MRS server can be accessed using SOAP, the wsdl's are located at:
http://mrs.cmbi.ru.nl/mrsws/search/wsdl
http://mrs.cmbi.ru.nl/mrsws/blast/wsdl
http://mrs.cmbi.ru.nl/mrsws/clustal/wsdl
Open Source
MRS was designed and implemented by Maarten Hekkelman at the CMBI with the help and contributions from many others. It is distributed under a BSD license. You can find the source code for the current version at ftp://ftp.cmbi.ru.nl/pub/software/mrs/
There's a mailinglist for issues related with MRS, to subscribe go to http://lists.berlios.de/mailman/listinfo/mrs-user.
Abstract
The biological data explosion of the 'omics' era requires fast access to many data types in rapidly growing data banks. The MRS software provides the tools to rapidly and reliably download, store, index, and query flat-file databanks. Data stored and indexed by MRS takes considerably less space on disk than the raw data, despite that these raw data are included. The MRS index information is part of the stored data. Therefore, public and private data can be combined by simple concatenation and thus without computational overheads.
When using this server or the software, please refer to:
MRS: A fast and compact retrieval system for biological data.
Hekkelman M.L., Vriend G.
Nucleic Acids Research 2005 33(Web Server issue):W766-W769;
Done

Mrs - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://cnkeeper.uio.no:8080/mrs-web/query.do;jsessionid=93EA585F6DC

Most Visited Red Hat Red Hat Magazine Red Hat Network Red Hat Support

Home Blast Blast results Clustal Databanks Settings Help

Search All Databanks for h1n1 human Search

Order by databank or relevance Show 1 2 3 4 5 hits per databank

Results for query "h1n1 human"

Databank	Hits	ID	Relevance	Title
PDB	7	3gbn		hemagglutinin; hemagglutinin; fab heavy chain; fab lambda light chain; (VIRAL PROTEIN/IMMUNE SYSTEM); crystal structure of fab cr6261 in complex with the 1918 h1n1 influenza virus hemagglutinin
		2zko		non-structural protein 1; rna (5'- (RNA BINDING PROTEIN/RNA); structural basis for dsrna recognition by ns1 protein of human influenza virus a
		1ruy		hemagglutinin; hemagglutinin; (VIRAL PROTEIN); swine h1 hemagglutinin
		more hits		
PDBFinder2	1	3gbn		viral protein/immune system
Swiss-Prot	355	pa_i33a0		Polymerase acidic protein;
		pa_i82a2		Polymerase acidic protein;
		ncap_i76a7		Nucleoprotein;
		more hits		
Taxonomy	44	405099		Influenza A virus (A/Kentucky/2/2006(H1N1))
		405098		Influenza A virus (A/Virginia/20/2003(H1N1))
		162539		Influenza A virus (A/human/Taiwan/5779/98(H1N1))
		more hits		
TrEMBL	6,035	c4rth4_9infa		RNA-directed RNA polymerase catalytic subunit;
		c5mqp8_9infa		Neuraminidase;
		c4rth3_9infa		Polymerase PA;
		more hits		

Alternative spelling suggestions

h1n1: h10n1 h11n1 h12n1 h1n1v 1h1n

human: chuman ghuman hauman hhuman huhman

Done

MRS

- ✓ Created as an alternative to SRS: A faster and more compact system. Can search all databases or particular ones with a single query.
- ✓ Several servers available to public, including ours at: http://cnkeeper.uio.no:8080/mrs-web/
- ✓ Typical time to retrieve the examples listed: <10 seconds on a moderately loaded server (20-30 user sessions)
- ✓ Think of it as the “Google of biological sequence databases”.

MRS (2)

- ✓ Like EMBOSS had its index query list (id,acc,sv,des,key,org), so does MRS.
- ✓ In fact, every MRS database has its own indices.
- ✓ Some of the index names remain the same across different databases (for example os (organism species) across EMBL, UniProt)
- ✓ **Remembering the names of indices is key to getting the most out of the power of MRS**

MRS (3)

Indices for embl

id	description	kind	count
__ALL_TEXT__	__ALL_TEXT__	FullText	386,368,586
ac	Accession number	FullText	169,888,166
as	as	FullText	78,707
cc	Comments and Notes	FullText	12,153,853
cd_date	cd_date	Date	7,823
co	Contigs	FullText	18,803,020
dc	Data class	FullText	19
de	Description	FullText	122,413,122
dr	Database cross-reference	FullText	16,399,644
ft	Feature table data	FullText	90,147,670
id	Identification	Unique	169,850,727
kw	Keywords	FullText	74,119
length	Sequence length	Number	309,576
oc	Organism classification	FullText	76,464
og	Organelle	FullText	5,056
os	Organism species	FullText	307,042
pr	Project	FullText	2,458
ref	Any reference field	FullText	977,604
sv	Sequence version	Number	106
td	Taxonomic division	FullText	30
topology	Topology (circular or linear)	FullText	4
up_date	up_date	Date	5,693
xref	xref	FullText	17,604,133

Done

Indices for uniprot

id	description	kind	count
__ALL_TEXT__	__ALL_TEXT__	FullText	71,477,743
ac	Accession number	FullText	10,750,272
cc	Comments and Notes	FullText	213,787
crc64	The CRC64 checksum for the sequence	FullText	8,746,681
de	Description	FullText	847,259
dr	Database cross-reference	FullText	36,765,488
dt	Date	Date	431
ft	Feature table data	FullText	1,264,692
gn	Gene name	FullText	8,971,223
id	Identification	Unique	10,547,998
kw	Keywords	FullText	1,468
length	The length of the sequence	Number	9,240
mw	Molecular weight	Number	320,079
oc	Organism classification	FullText	57,068
og	Organelle	FullText	4,897
oh	oh	FullText	4,424
os	Organism species	FullText	179,557
ox	Taxonomy cross-reference	FullText	233,637
pe	pe	FullText	24
ref	Any reference field	FullText	1,589,486

MRS (4)

- ✓ The role of Boolean/Relational operators
  - | *OR*
  - & *AND*
  - ! *NOT*

(use the textual representation (<u>in capitals</u>), <u>not</u> the symbols on the command-line)

- *lysosyme | lysozyme* *lysosyme OR lysozyme*
- *lysozyme & !kw:est* *lysozyme AND NOT(kw:est)*
- *(lysosyme | lysozyme) & !kw:est* *(lysosyme OR lysozyme) AND NOT(kw:est)*

MRS (5)

- ✓ 'mrs-query' is the command line tool of MRS.
- ✓ Very useful for the construction of pipelines.

Mrs-query search mode

mrs-query -d databank -q term1 -n no_of_results -o file

OR

mrs-query -d databank -f “bool_expr” -n no_of_results -o file

- *If you do not include the -n switch only the first 10 results will be shown.*
- *-q may be repeated and in that case, MRS will attempt to combine the text search terms (implicit AND boolean switch)*
- *This mode retrieves identifiers but not sequences*

MRS (6)

Mrs-query sequence retrieval mode

mrs-query -d databank -e seq_id

- *This mode retrieves actual sequences*
- *It could be used in batch mode to retrieve a list of sequence ids with a bit of scripting (see tutorial Day 3)*

✓ Pipeline:

mrs-query(search mode)->results->mrs-query(sequence_retrieval_mode)

MRS (7)

Fixing the glue between the mrs-query mode

```
#!/site/perl510/bin/perl -w
use strict;

my $databank="embl";

#loose the \n character
chomp(my @ids=<STDIN>);

for my $id (@ids) {
        #Now loose the \t whitespace character
        chomp($id);
        my $result=`mrs-query -d $databank -e $id`;
        print $result;
}
```

Mrs-query search mode example:

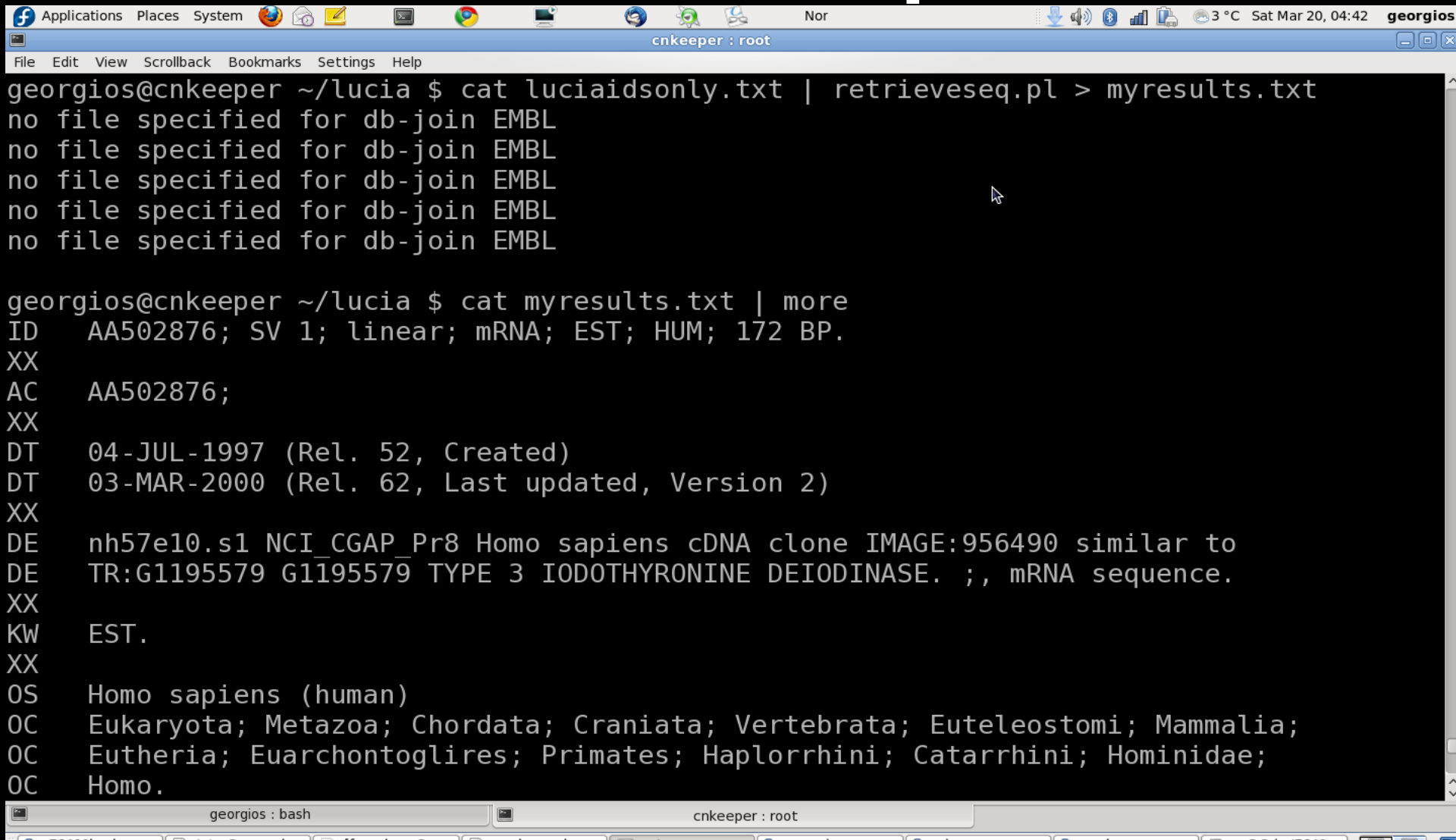
Terminal

File Edit View Terminal Tabs Help

```
lingeek@cnkeeper /biotek/cnkeeper/storage/rdbms/bin $ mrs-query -d embl -q h1n1 -q human -n 25 -o ~/h1n1human.txt
no file specified for db-join EMBL
lingeek@cnkeeper /biotek/cnkeeper/storage/rdbms/bin $ cat ~/h1n1human.txt
Found 19426 hits, displaying the first 25
gq221696        55.5    Influenza A virus (A/GuangzhouSB/01/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds.
gq221694        55.4    Influenza A virus (A/GuangzhouSB/01/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds.
gq221693        55.2    Influenza A virus (A/GuangzhouSB/01/2009(H1N1)) segment 3 polymerase PA (PA) gene, complete cds.
cy053484        55.2    Influenza A virus (A/Taiwan/156/2009(H1N1)) segment 6 sequence.
cy053483        55.1    Influenza A virus (A/Taiwan/156/2009(H1N1)) segment 5 sequence.
gq221695        55      Influenza A virus (A/GuangzhouSB/01/2009(H1N1)) segment 5 nucleocapsid protein (NP) gene, complete cds.
gq225359        54.9    Influenza A virus (A/Shanghai/1/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds.
cy053481        54.8    Influenza A virus (A/Taiwan/156/2009(H1N1)) segment 3 sequence.
cy053500        54.7    Influenza A virus (A/Taiwan/177/2009(H1N1)) segment 6 sequence.
gq200292        54.7    Influenza A virus (A/Shandong/1/2009(H1N1)) segment 1 polymerase PB2 gene, complete cds.
cy053476        54.7    Influenza A virus (A/Taiwan/143/2009(H1N1)) segment 6 sequence.
cy053499        54.6    Influenza A virus (A/Taiwan/177/2009(H1N1)) segment 5 sequence.
gq225357        54.6    Influenza A virus (A/Shanghai/1/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds.
gq232095        54.6    Influenza A virus (A/Beijing/4/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds.
cy053492        54.6    Influenza A virus (A/Taiwan/167/2009(H1N1)) segment 6 sequence.
cy053475        54.6    Influenza A virus (A/Taiwan/143/2009(H1N1)) segment 5 sequence.
cy053508        54.6    Influenza A virus (A/Taiwan/206/2009(H1N1)) segment 6 sequence.
gq225356        54.5    Influenza A virus (A/Shanghai/1/2009(H1N1)) segment 3 polymerase PA (PA) gene, complete cds.
cy053507        54.4    Influenza A virus (A/Taiwan/206/2009(H1N1)) segment 5 sequence.
cy053491        54.4    Influenza A virus (A/Taiwan/167/2009(H1N1)) segment 5 sequence.
cy053497        54.4    Influenza A virus (A/Taiwan/177/2009(H1N1)) segment 3 sequence.
gq232093        54.3    Influenza A virus (A/Beijing/4/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds.
cy053473        54.3    Influenza A virus (A/Taiwan/143/2009(H1N1)) segment 3 sequence.
gq225383        54.3    Influenza A virus (A/Beijing/3/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds.
gq225375        54.3    Influenza A virus (A/Beijing/01/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds.

lingeek@cnkeeper /biotek/cnkeeper/storage/rdbms/bin $
```

Mrs-query sequence retrieval mode example:



MRS::Client PERL module

- ✓ A better way to query MRS databanks on the command line
- ✓ Based on Martin Senger's CPAN MRS::Client module

  http://search.cpan.org/~tulsoft/MRS-Client-0.53/lib/MRS/Client.pod
- ✓ SOAP-based
- ✓ Contains functionality to access the BLAST and CLUSTALW MRS facilities

MRS::Client PERL module (2)

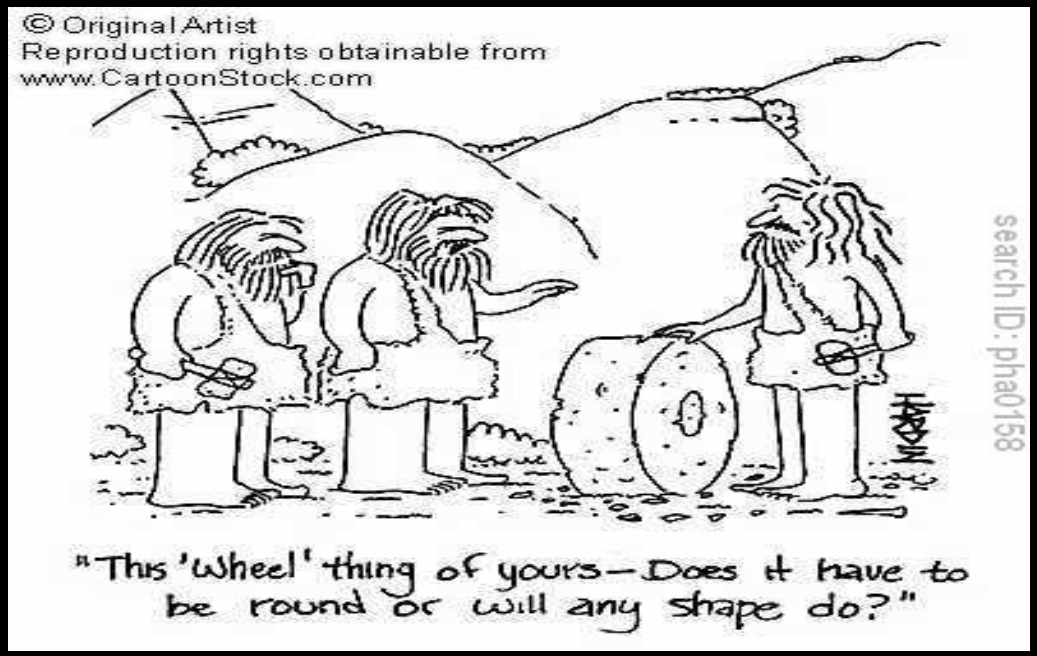
```
#!/site/perl510/bin/perl -w
use MRS::Client;

# Create a new MRS Client object and point to the right server
my $client = MRS::Client->new ( search_url  => 'http://localhost:18081/', blast_url   => '
http://localhost:18082/',  clustal_url => 'http://localhost:18083/');

#Receive an approximate count of the results by calling the count method
print $client->db ('embl')->find ('(kw:transporter OR de:transporter) AND oc:"Viridiplantae" AND
de:complete')->count;
#form the query and store the result.
my $query = $client->db ('embl')->find ('(kw:transporter OR de:transporter) AND
oc:"Viridiplantae" AND de:complete');

 #Now obtain the results by examining the data structure and call the next method
  while (my $record = $query->next) {
          print $record . "\n";
          }
```

Questions/Case studies



georgios@biotek.uio.no