BIOCH300D Bioinformatics Project 2012

Dr. Silvia Vidal
McGill Life Sciences Complex room 367
Phone: 398-2362
E-mail: silvia.vidal@mcgill.ca

http://teodorolab.mcgill.ca/300D/

OUTLINE INTRODUCTION

1- **WHAT IS BIOINFORMATICS?** definition, applications, goal

2- **WHERE DOES IT OCCUR?** Some important institutions

3- **HOW DOES IT HAPPEN?**

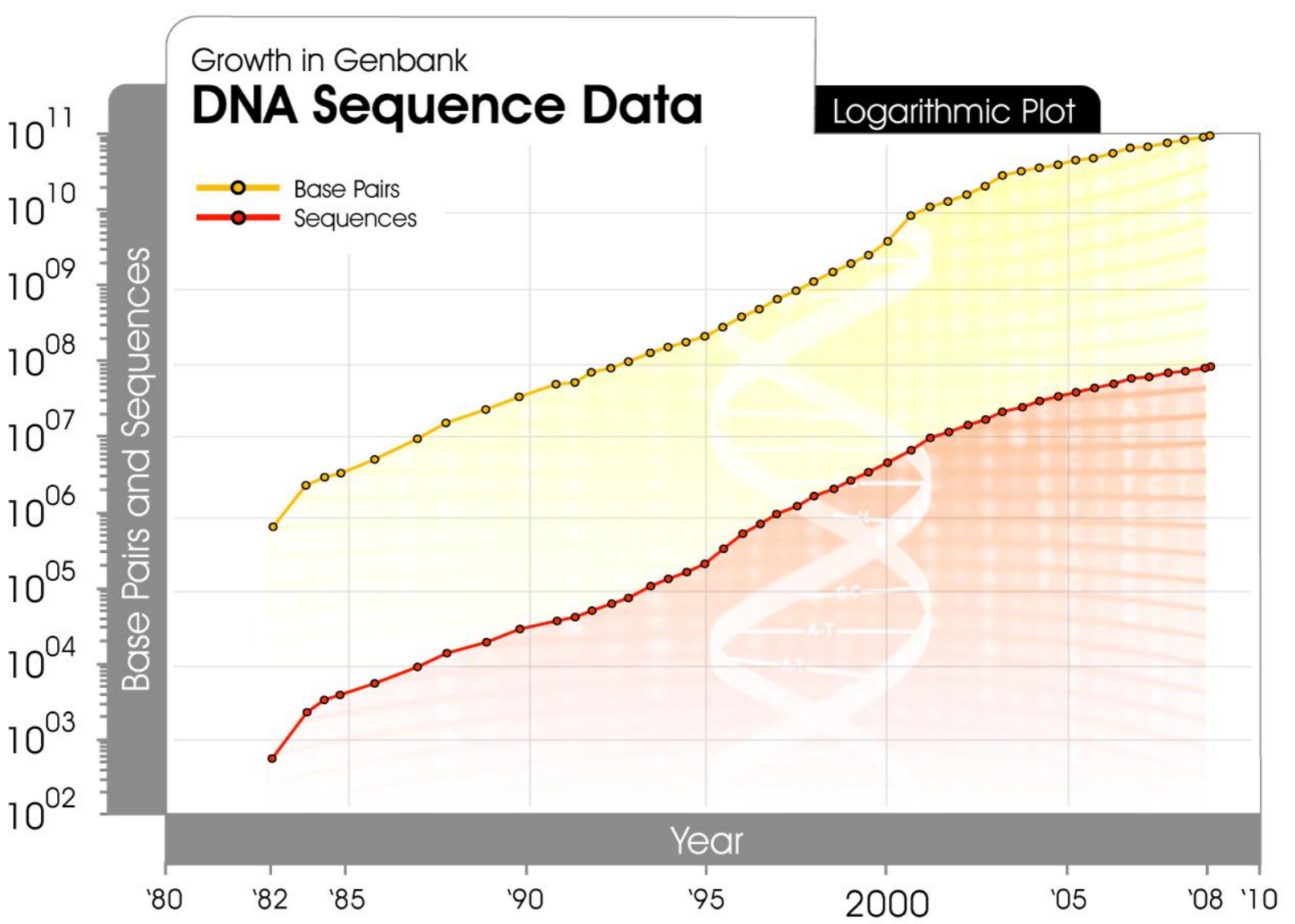
- **3.1 DATABASES** NCBI databases
- **3.2 TOOLS** for protein sequence analysis
  - 3.2.1 BLAST
  - 3.2.2 CLUSTAL W
  - 3.3.3 3-D VIEWER

EXERCISE

What is Bioinformatics?

Scientific discipline resulting from the merge of biology and information sciences

Access and Storage of information is important



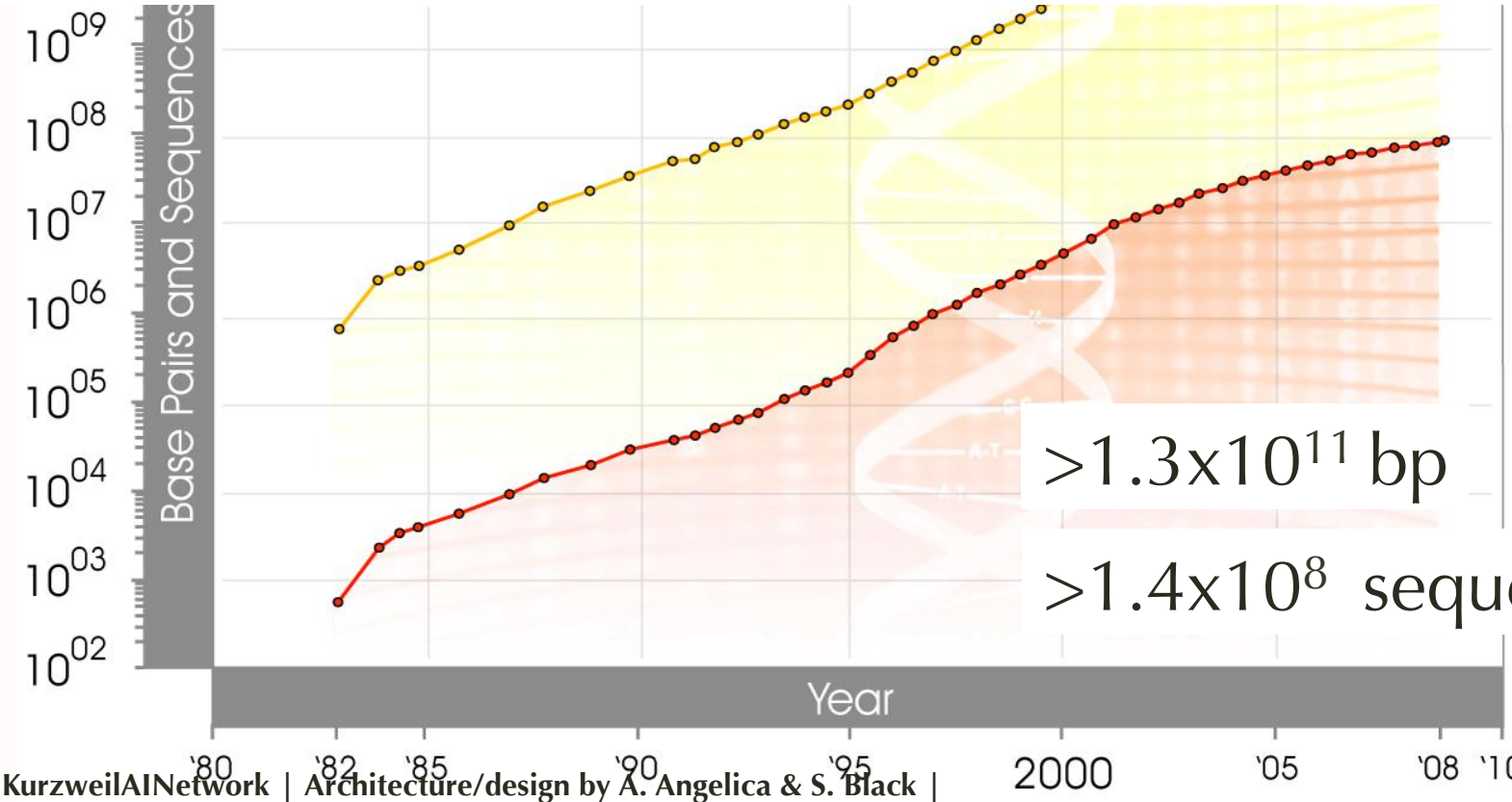
Access and Storage of information is important

GBREL.TXT Genetic Sequence Data Bank
February 15 2012

NCBI-GenBank Flat File Release 188.0

Distribution Release Notes

149819246 loci, 137384889783 bases, from 149819246 reported sequences



$>1.3x10^{11}$ bp

$>1.4x10^{8}$ sequences

Bioinformatics comprises two main activities:

1- Computerized annotation of genomic and biological information and data: **DATABASES**

2- Transformation and manipulation of the data: **SOFTWARE TOOLS**

The general goal of bioinformatics is:

1- Providing **testable predictions** from data and/or their manipulation

2- Needs to support data with **statistics** and probability calculations

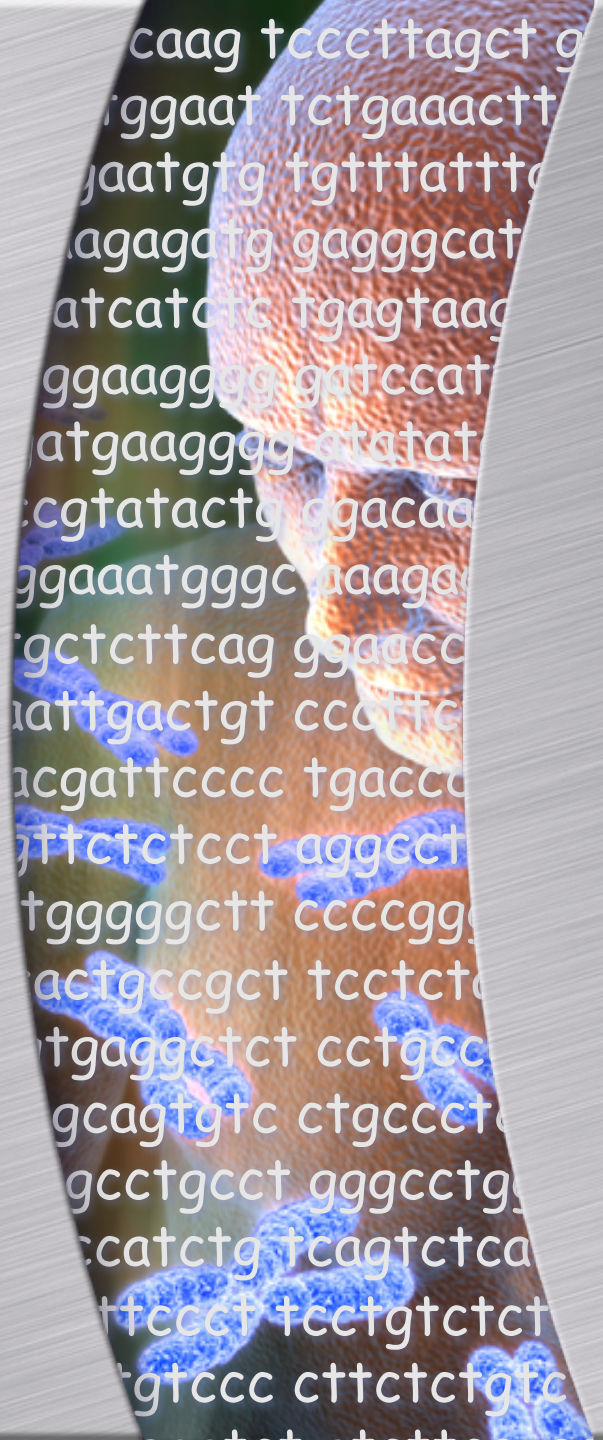
Bioinformatics is presently used in...

- Medical sciences
- Genomics (humans, plants, animals, pathogens)
- Ecology / population genetics
- Functional genomics
- Systems biology
- Pharmaceutical development
- Molecular biology

Bioinformatics and ... gene/protein sequence analysis

- ✓ identify new gene/protein
- ✓ predict protein function
- ✓ identify active sites, functional domains
- ✓ evolution
- ✓ 3D structure
- ✓ protein design
- ✓ identify inhibitors of protein function

Enter Bioinfo



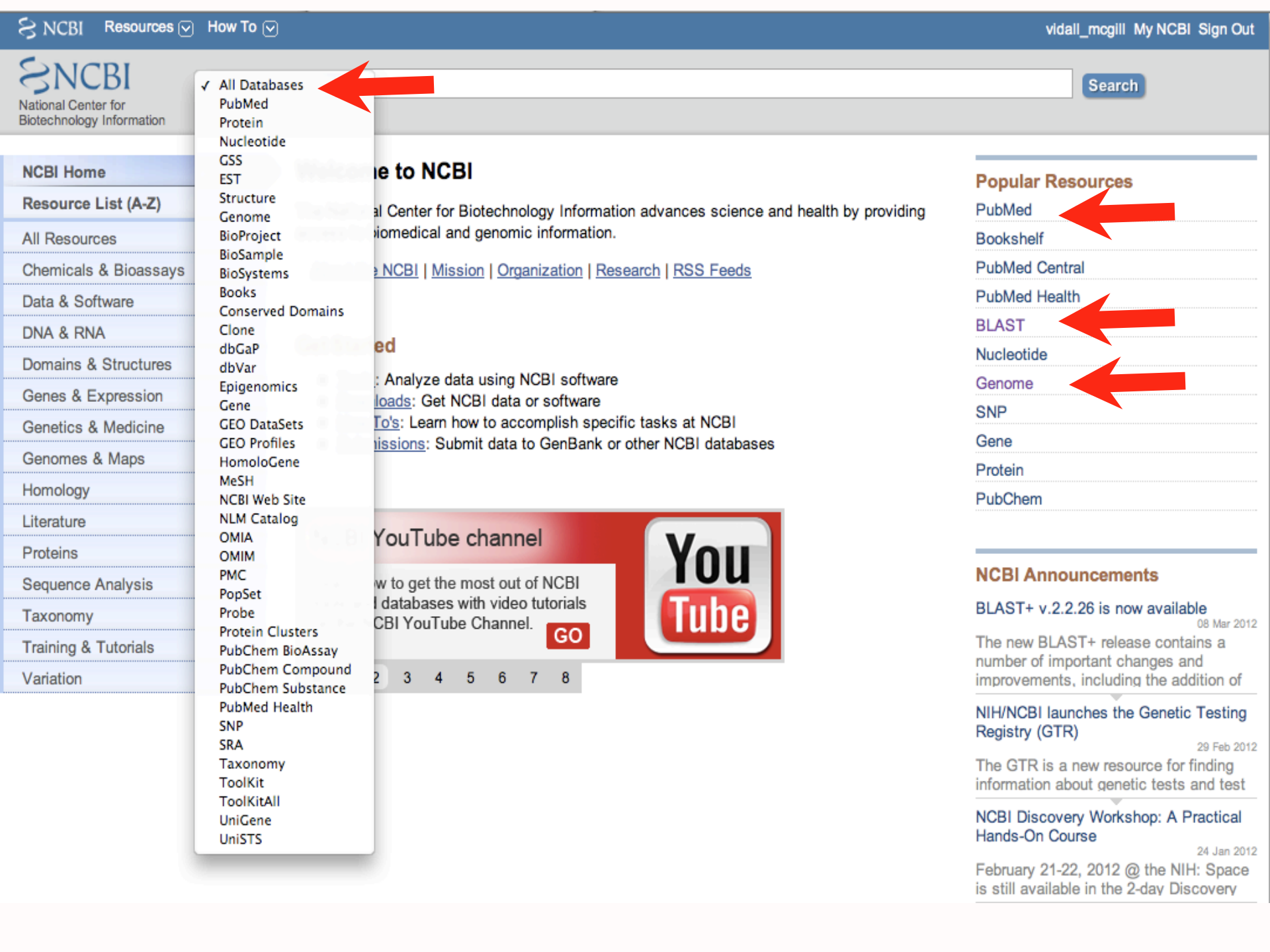
International institutions centralize bioinformatic information, tools and processes

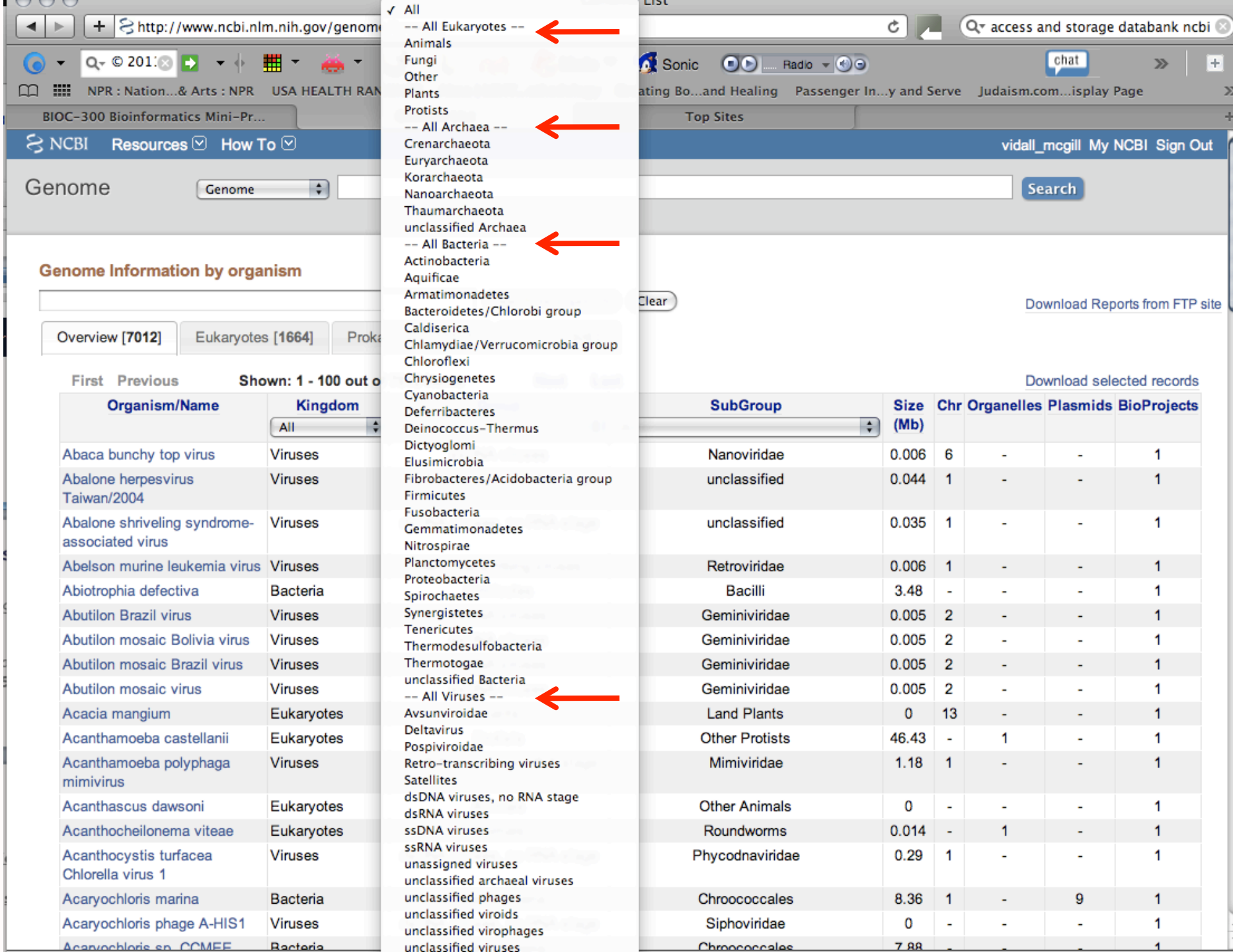
NCBI: National Center for Biotechnology Information, National library of Medicine and National Institutes of Health

EBI: European Bioinformatics Institute

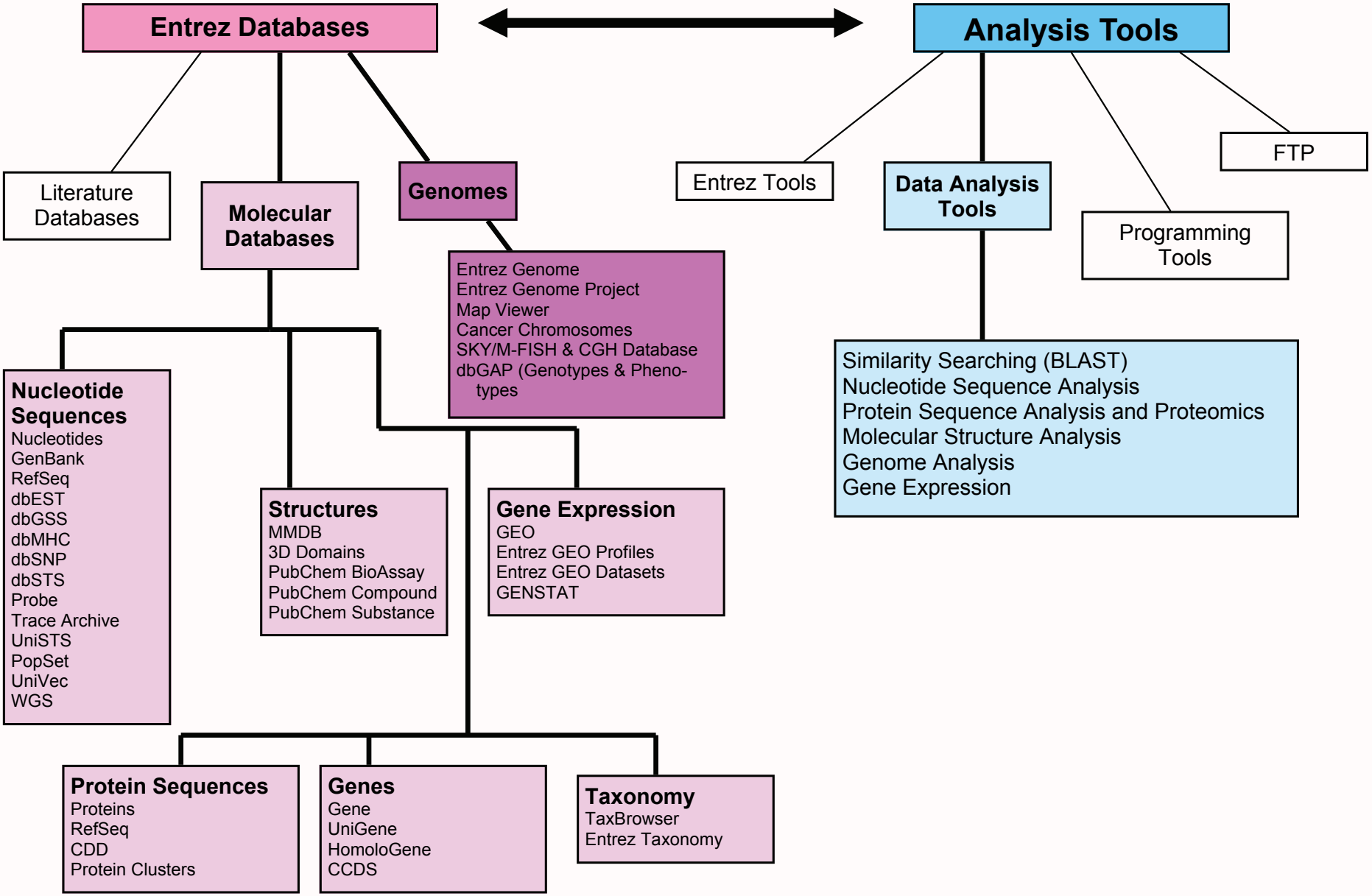
PDB: Protein Data Bank.

The NCBI Homepage


NCBI Resources How To
vidall_mcgill My NCBI Sign Out
NCBI
National Center for Biotechnology Information
All Databases
PubMed
Protein
Nucleotide
GSS
EST
Structure
Genome
BioProject
BioSample
BioSystems
Books
Conserved Domains
Clone
dbGaP
dbVar
Epigenomics
Gene
GEO DataSets
GEO Profiles
HomoloGene
MeSH
NCBI Web Site
NLM Catalog
OMIA
OMIM
PMC
PopSet
Probe
Protein Clusters
PubChem BioAssay
PubChem Compound
PubChem Substance
PubMed Health
SNP
SRA
Taxonomy
ToolKit
ToolKitAll
UniGene
UniSTS
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
e to NCBI
l Center for Biotechnology Information advances science and health by providing
iomedical and genomic information.
e NCBI | Mission | Organization | Research | RSS Feeds
ed
: Analyze data using NCBI software
loads: Get NCBI data or software
To's: Learn how to accomplish specific tasks at NCBI
issions: Submit data to GenBank or other NCBI databases
YouTube channel
w to get the most out of NCBI
databases with video tutorials
CBI YouTube Channel.
GO
You Tube
2 3 4 5 6 7 8
Popular Resources
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI Announcements
BLAST+ v.2.2.26 is now available
08 Mar 2012
The new BLAST+ release contains a number of important changes and improvements, including the addition of
NIH/NCBI launches the Genetic Testing Registry (GTR)
29 Feb 2012
The GTR is a new resource for finding information about genetic tests and test
NCBI Discovery Workshop: A Practical Hands-On Course
24 Jan 2012
February 21-22, 2012 @ the NIH: Space is still available in the 2-day Discovery

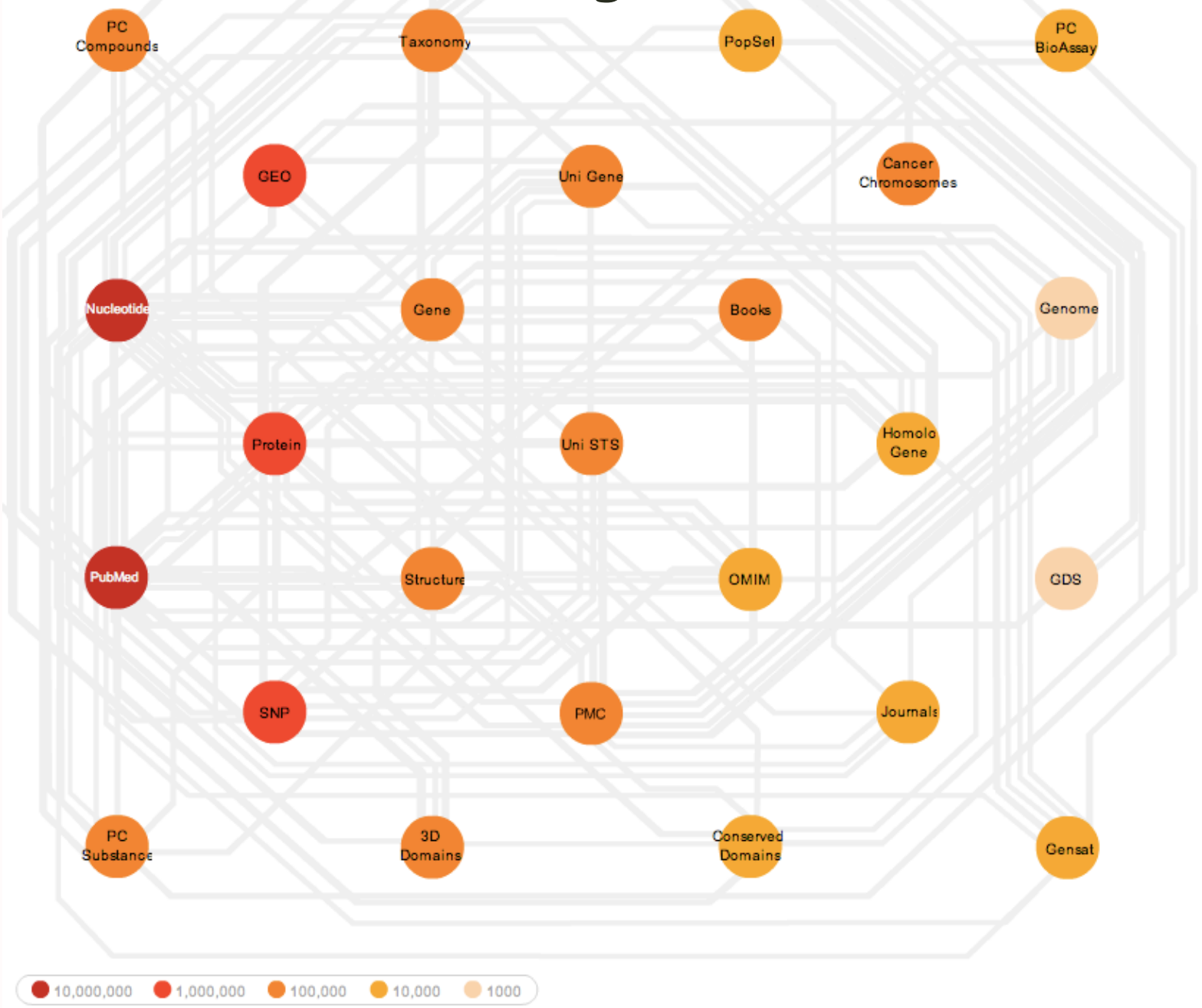


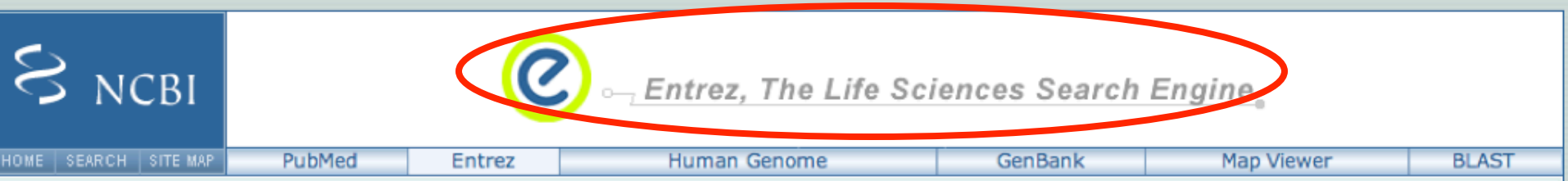
Organism/Name	Kingdom	SubGroup	Size (Mb)	Chr	Organelles	Plasmids	BioProjects
Abaca bunchy top virus	Viruses	Nanoviridae	0.006	6	-	-	1
Abalone herpesvirus Taiwan/2004	Viruses	unclassified	0.044	1	-	-	1
Abalone shriveling syndrome-associated virus	Viruses	unclassified	0.035	1	-	-	1
Abelson murine leukemia virus	Viruses	Retroviridae	0.006	1	-	-	1
Abiotrophia defectiva	Bacteria	Bacilli	3.48	-	-	-	1
Abutilon Brazil virus	Viruses	Geminiviridae	0.005	2	-	-	1
Abutilon mosaic Bolivia virus	Viruses	Geminiviridae	0.005	2	-	-	1
Abutilon mosaic Brazil virus	Viruses	Geminiviridae	0.005	2	-	-	1
Abutilon mosaic virus	Viruses	Geminiviridae	0.005	2	-	-	1
Acacia mangium	Eukaryotes	Land Plants	0	13	-	-	1
Acanthamoeba castellanii	Eukaryotes	Other Protists	46.43	-	1	-	1
Acanthamoeba polyphaga mimivirus	Viruses	Mimiviridae	1.18	1	-	-	1
Acanthascus dawsoni	Eukaryotes	Other Animals	0	-	-	-	1
Acanthocheilonema viteae	Eukaryotes	Roundworms	0.014	-	1	-	1
Acanthocystis turfacea Chlorella virus 1	Viruses	Phycodnaviridae	0.29	1	-	-	1
Acaryochloris marina	Bacteria	Chroococcales	8.36	1	-	9	1
Acaryochloris phage A-HIS1	Viruses	Siphoviridae	0	-	-	-	1
Acaryochloris sp. CCMEE	Bacteria	Chroococcales	7.88	-	-	-	1



Created by Lyra Eugenio-Thomson.
For more complete information on NCBI Databases and Tools, go to:
http://www.ncbi.nlm.nih.gov/Sitemap/index.html

NCBI databases are integrated





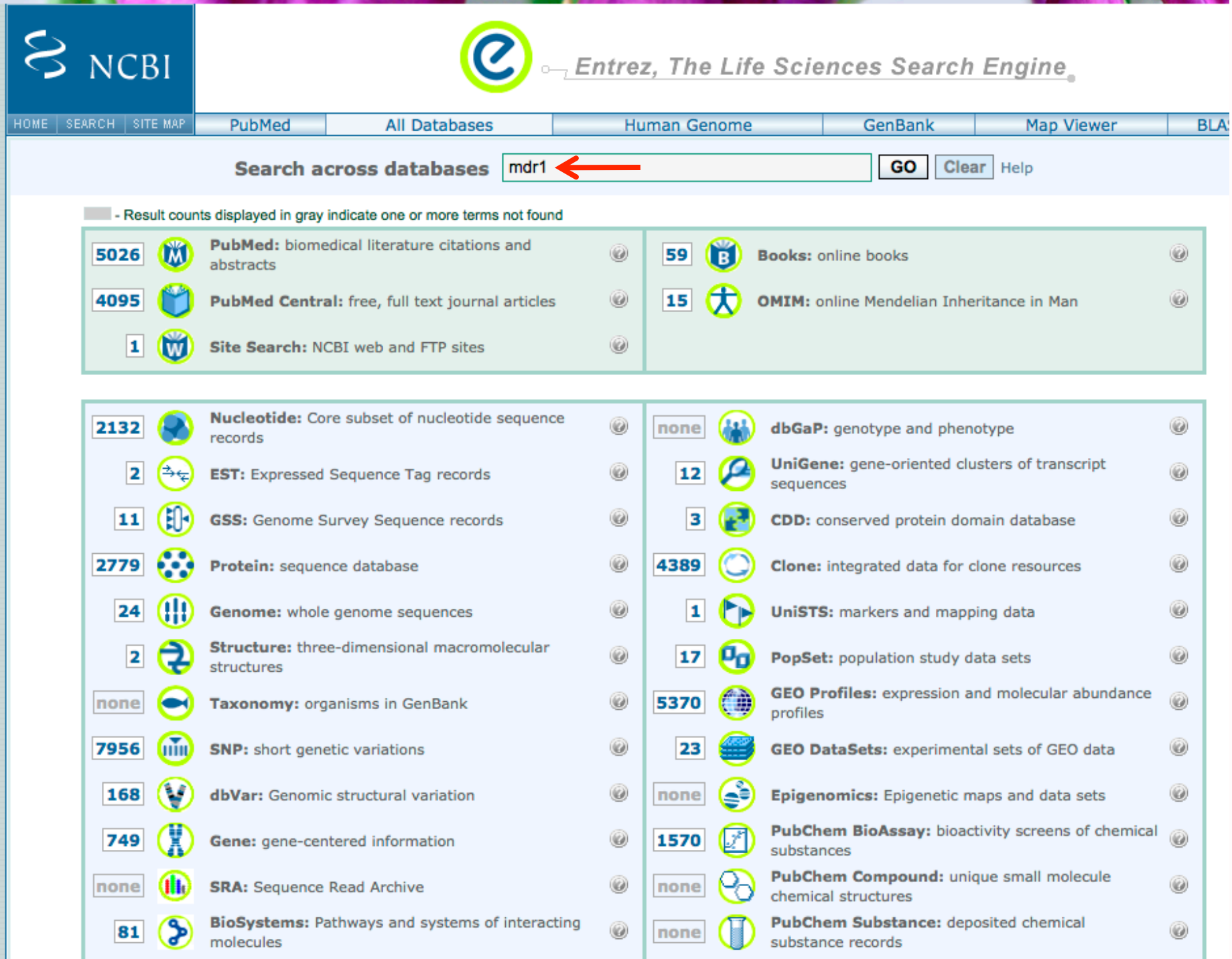
Search across databases GO CLEAR Help

Welcome to the new Entrez cross-database search page

- **PubMed:** biomedical literature citations and abstracts
- **PubMed Central:** free, full text journal articles
- **Books:** online books
- **OMIM:** online Mendelian Inheritance in Man
- **Site Search:** NCBI web and FTP sites

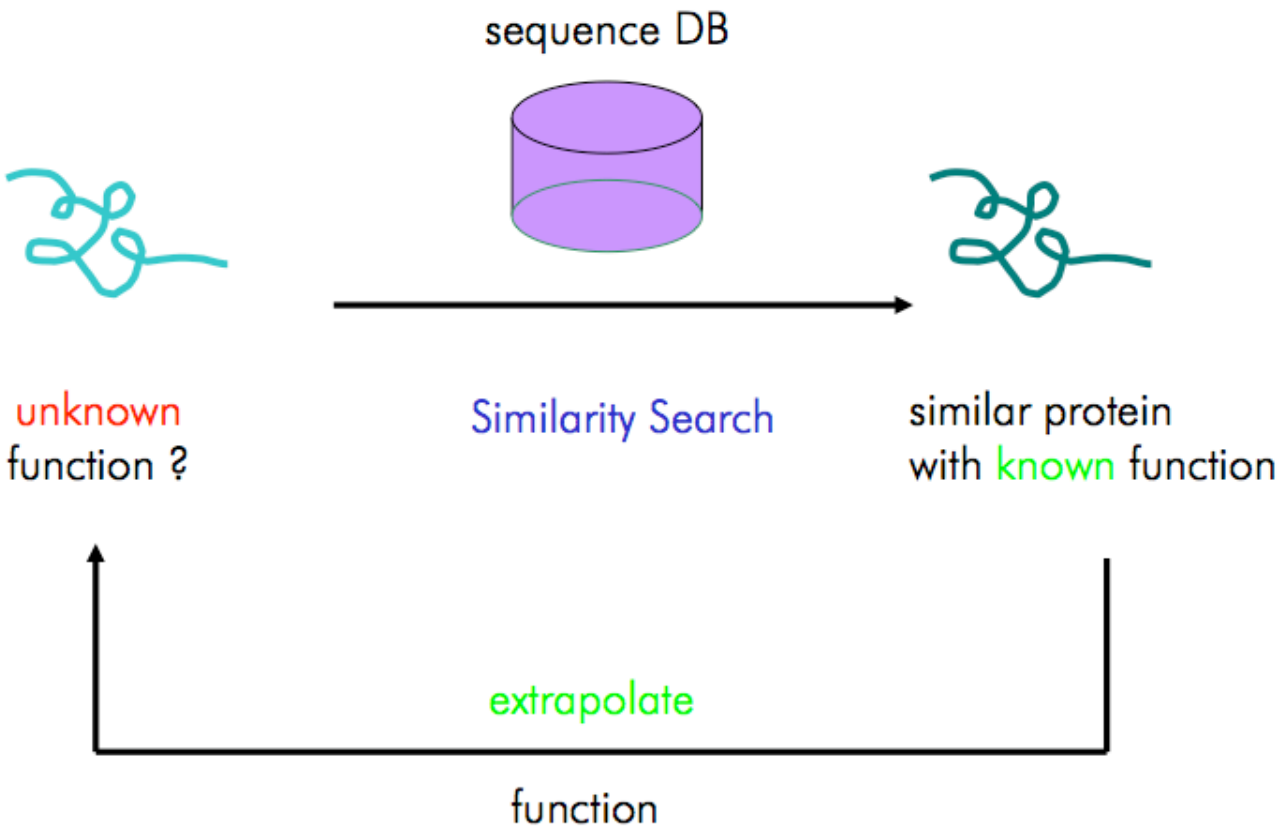
- **Nucleotide:** sequence database (GenBank)
- **Protein:** sequence database
- **Genome:** whole genome sequences
- **Structure:** three-dimensional macromolecular structures
- **Taxonomy:** organisms in GenBank
- **SNP:** single nucleotide polymorphism
- **Gene:** gene-centered information
- **HomoloGene:** eukaryotic homology groups
- **PubChem Compound:** small molecule chemical structures
- **PubChem Substance:** chemical substances screened for bioactivity
- **Genome Project:** genome project information
- **UniGene:** gene-oriented clusters of transcript sequences
- **CDD:** conserved protein domain database
- **3D Domains:** domains from Entrez Structure
- **UniSTS:** markers and mapping data
- **PopSet:** population study data sets
- **GEO Profiles:** expression and molecular abundance profiles
- **GEO DataSets:** experimental sets of GEO data
- **Cancer Chromosomes:** cytogenetic databases
- **PubChem BioAssay:** bioactivity screens of chemical substances
- **GENSAT:** gene expression atlas of mouse central nervous system

- **Journals:** detailed information *about* the journals indexed
- **MeSH:** detailed information about NLM's controlled

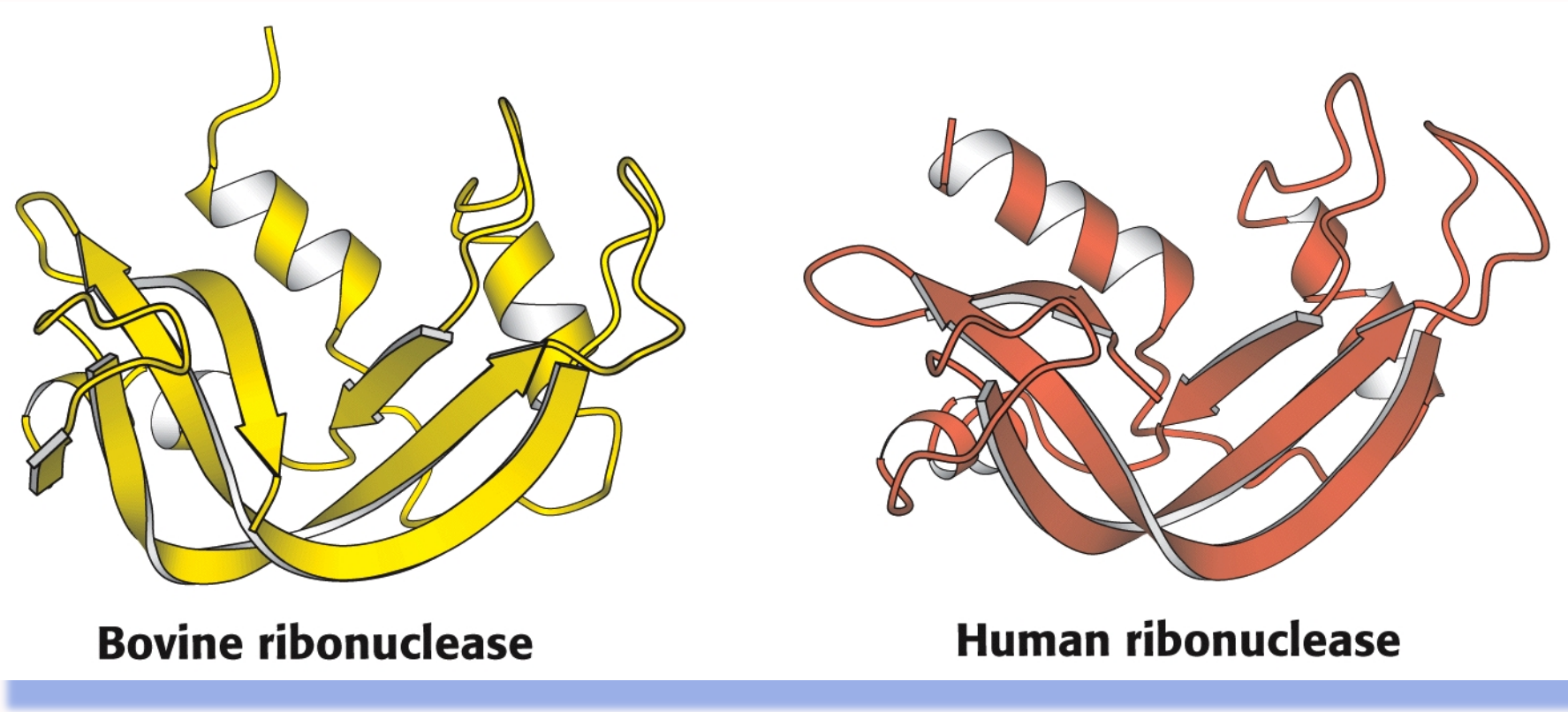

NCBI
Entrez, The Life Sciences Search Engine
HOME
SEARCH
SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLA
Search across databases
mdr1
GO
Clear
Help
- Result counts displayed in gray indicate one or more terms not found
5026 PubMed: biomedical literature citations and abstracts
4095 PubMed Central: free, full text journal articles
1 Site Search: NCBI web and FTP sites
59 Books: online books
15 OMIM: online Mendelian Inheritance in Man
2132 Nucleotide: Core subset of nucleotide sequence records
2 EST: Expressed Sequence Tag records
11 GSS: Genome Survey Sequence records
2779 Protein: sequence database
24 Genome: whole genome sequences
2 Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
7956 SNP: short genetic variations
168 dbVar: Genomic structural variation
749 Gene: gene-centered information
none SRA: Sequence Read Archive
81 BioSystems: Pathways and systems of interacting molecules
none dbGaP: genotype and phenotype
12 UniGene: gene-oriented clusters of transcript sequences
3 CDD: conserved protein domain database
4389 Clone: integrated data for clone resources
1 UniSTS: markers and mapping data
17 PopSet: population study data sets
5370 GEO Profiles: expression and molecular abundance profiles
23 GEO DataSets: experimental sets of GEO data
none Epigenomics: Epigenetic maps and data sets
1570 PubChem BioAssay: bioactivity screens of chemical substances
none PubChem Compound: unique small molecule chemical structures
none PubChem Substance: deposited chemical substance records

1. **BLAST**: similarity search through sequence alignment

Value of sequence alignment



sequence/structure/function homology in proteins from a **common ancestor** (homologous proteins)



Structures of Ribonucleases from Cows and Human Beings. Structural similarity often follows functional similarity.

HOMOLOGY: The presence of a similar feature because of descent from a common ancestor

SIMILARITY: quantitative measurement of identities, gaps, conservative substitutions between protein sequences.

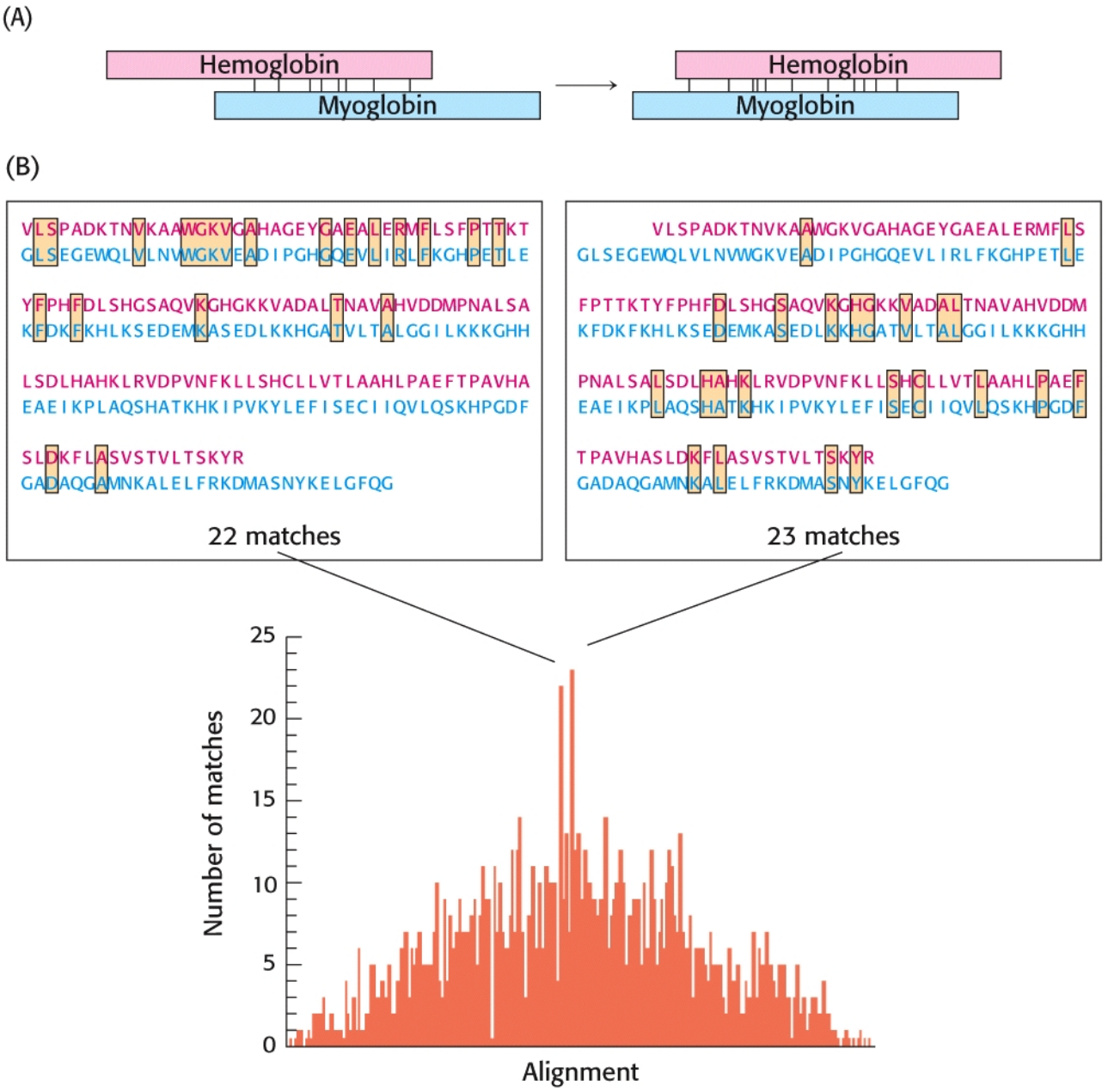
Human hemoglobin (α chain)

VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHG
SAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLS
HCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

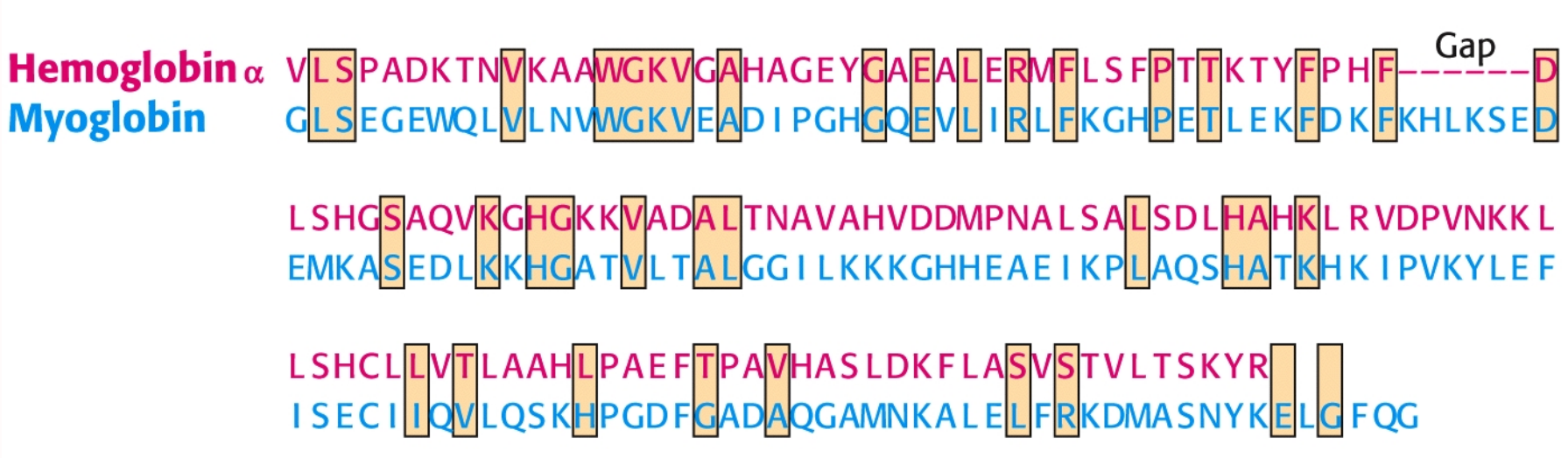
Human myoglobin

GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKS
EDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVK
YLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG

Molecular Cell Biology, Sixth Edition



Molecular Cell Biology, Sixth Edition



Alignment with gap insertion

Alignment with conservative substitutions

BLAST

A method to ascertain sequence **similarity**

1- The program takes a ***query*** sequence and searches it against the **database** selected by user.

2- It aligns a ***query*** sequence against each of all ***subject*** sequence in the database.

3- The results are reported in a form of a ranked list followed by a series of individual sequence alignments, plus various **statistics and scores**.

Sequence Analysis:

Basic Local Alignment Search Tool

```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || |  | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
```

```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                    ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

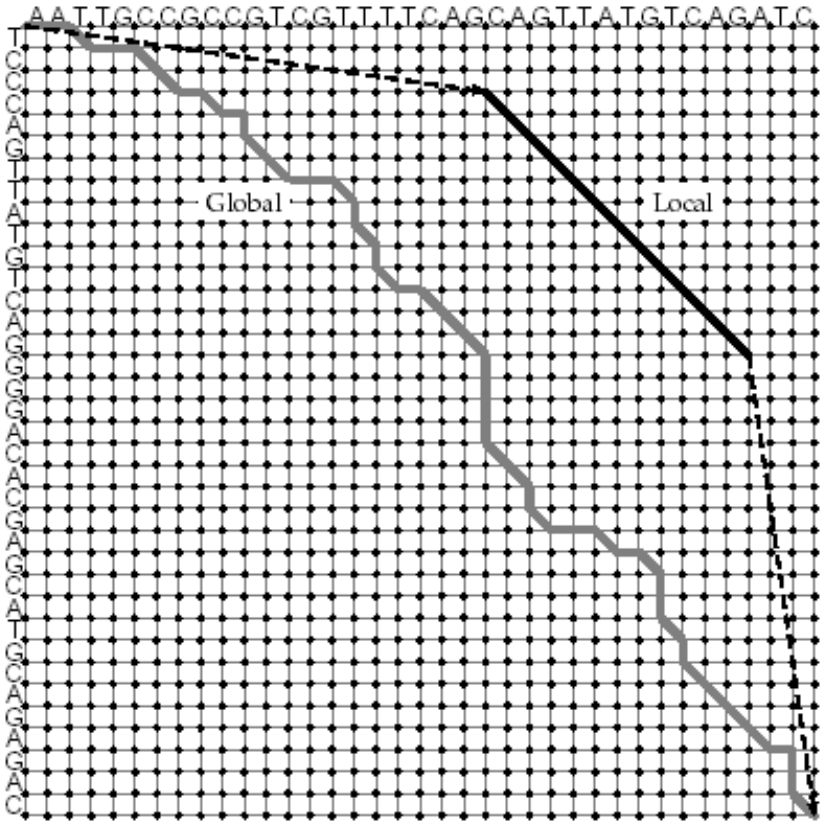
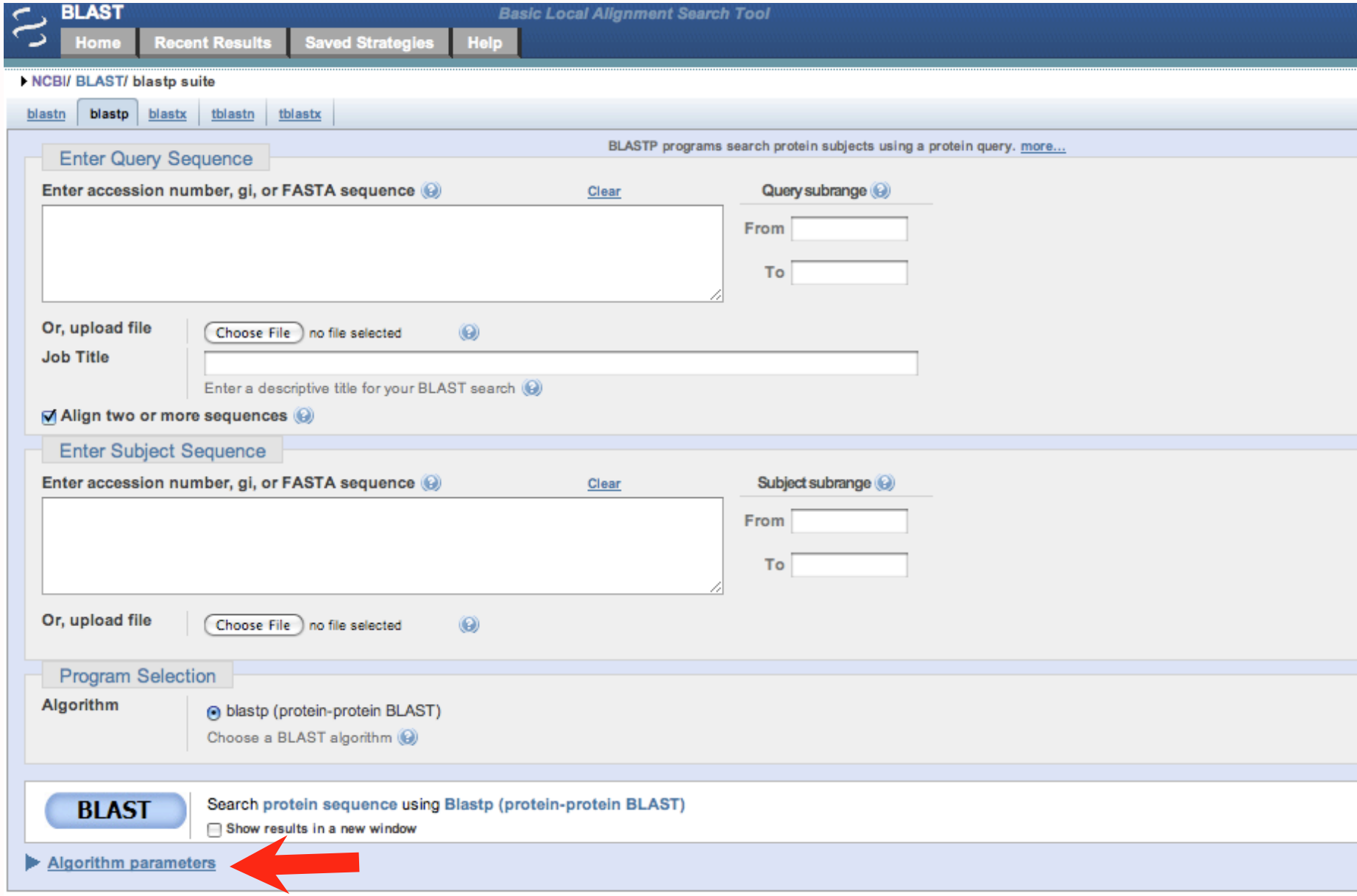


Figure 6.16 (a) Global and (b) local alignments of two hypothetical genes that each have a conserved domain. The local alignment has a much worse score according to the global scoring scheme, but it correctly locates the conserved domain.


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein subjects using a protein query. more...
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Query subrange
From
To
Or, upload file
Choose File
no file selected
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Enter Subject Sequence
Enter accession number, gi, or FASTA sequence
Clear
Subject subrange
From
To
Or, upload file
Choose File
no file selected
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters

BLAST Search protein sequence using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Assigns a score for aligning pairs of residues, and determines overall alignment score. more...

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only
Mask lower case letters

BLOSUM-62

BLAST substitution matrices

A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues. The theory of amino acid substitution matrices is described in [1], and applied to DNA sequence comparison in [2]. In general, different substitution matrices are tailored to detecting similarities among sequences that are diverged by differing degrees [1-3]. A single matrix may nevertheless be reasonably efficient over a relatively broad range of evolutionary change [1-3]. Experimentation has shown that the BLOSUM-62 matrix [4] is among the best for detecting most weak protein similarities. For particularly long and weak alignments, the BLOSUM-45 matrix may prove superior. A detailed statistical theory for gapped alignments has not been developed, and the best gap costs to use with a given substitution matrix are determined empirically. Short alignments need to be relatively strong (i.e. have a higher percentage of matching residues) to rise above background noise. Such short but strong alignments are more easily detected using a matrix with a higher "relative entropy" [1] than that of BLOSUM-62. In particular, short query sequences can only produce short alignments, and therefore database searches with short queries should use an appropriately tailored matrix. The BLOSUM series does not include any matrices with relative entropies suitable for the shortest queries, so the older PAM matrices [5,6] may be used instead. For proteins, a provisional table of recommended substitution matrices and gap costs for various query lengths is:

Query Length	Substitution Matrix	Gap Costs
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
85	BLOSUM-62	(10,1)

Gap Costs

NCBI

results of BLAST

BLASTP 2.2.10 [Oct-19-2004]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

RID: 1111270526-12062-108720234474.BLASTQ2

Query=
(153 letters)

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
2,367,365 sequences; 802,797,248 total letters

If you have any problems or questions with the results of this search
please refer to the **BLAST FAQs**

Taxonomy reports

Sequences producing significant alignments:

		Score (Bits)	E Value	
gi|58802449|gb|AAW82430.1|	ATP-binding cassette, sub-family B..	2365	0.0	G
gi|42741659|ref|NP_000918.2|	ATP-binding cassette sub-family ..	2364	0.0	G
gi|307180|gb|AAA59575.1|	P-glycoprotein [Homo sapiens]	2362	0.0	G
gi|2353264|gb|AAB69423.1|	P-glycoprotein [Homo sapiens]	2358	0.0	G
gi|60326712|gb|AAX18881.1|	P-glycoprotein [Cercopithecus aethiop	2285	0.0	
gi|31442763|gb|AAN07780.2|	multidrug resistance p-glycoprotein [	2285	0.0	
gi|74136329|ref|NP_001028059.1|	multidrug resistance p-glycop...	2280	0.0	G
gi|46394984|gb|AAS91648.1|	multidrug resistance protein; P-glyco	2279	0.0	G
gi|46394982|gb|AAS91647.1|	multidrug resistance protein 1; P-...	2175	0.0	G
gi|67462127|gb|AAY67840.1|	multidrug resistance protein 1 [Canis	2174	0.0	G

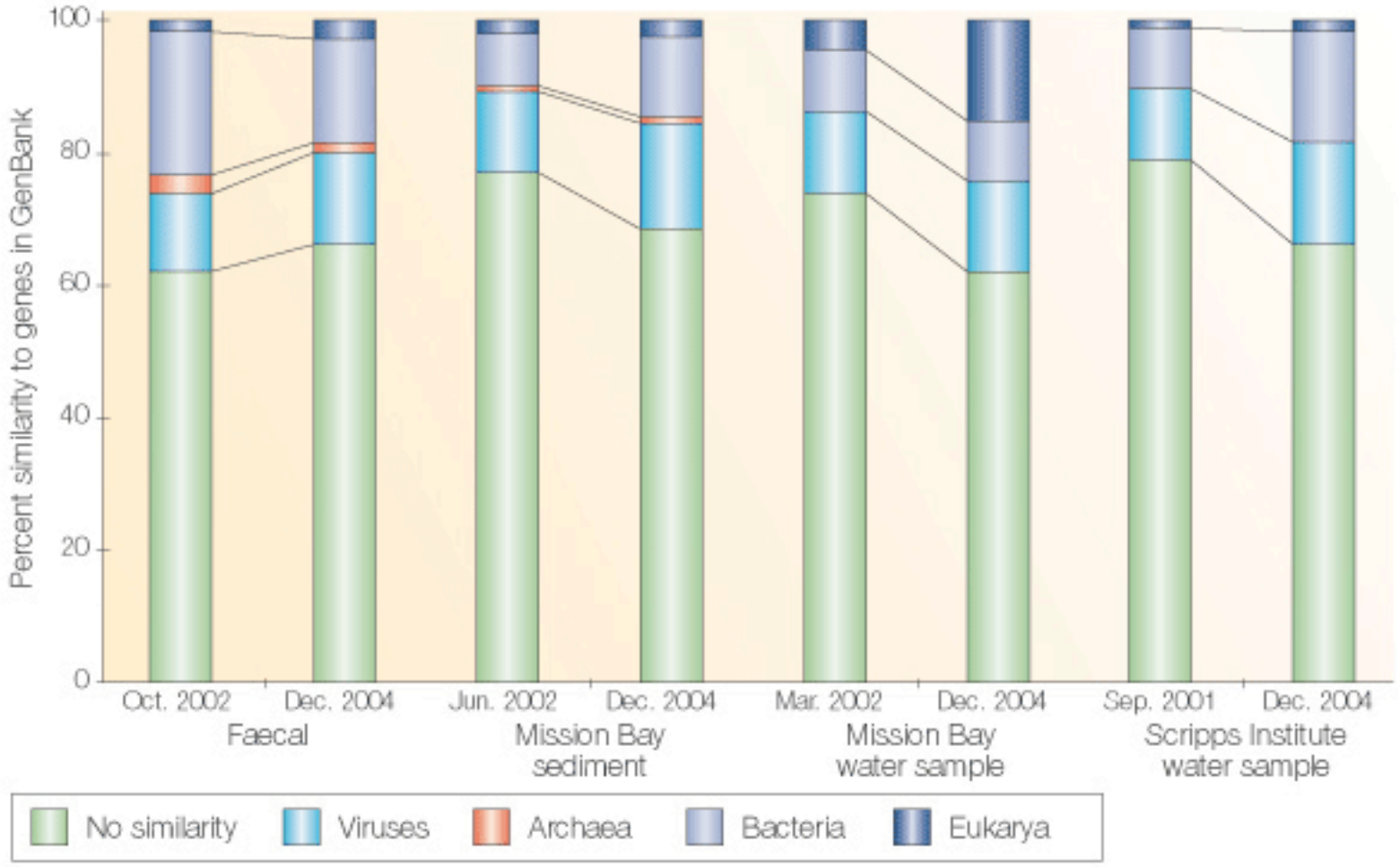
Identity Score: it is the extent of similarity between subject and query sequences. Takes into account similar and identical residues, as well as gaps introduced. The higher the identity score the better the alignment.

The E-Value: expected number of random comparisons in the database scoring above a given threshold; the E value indicates statistical significance of a given alignment. The lower the E value, the most significant is the hit.

USES of similarity searches with BLAST

- ✓ to identify homology between a new and a known sequence
- ✓ to identify new genes in any organism
- ✓ to predict gene structure of a new sequence (exon boundaries, regulatory regions)
- ✓ to identify members of a gene or protein family
- ✓ to infer biochemical function of a new protein
- ✓ to infer domain architecture, secondary and tertiary structure of a protein
- ✓ to infer evolutionary history of a sequence

Viral metagenomics: most viral genomes are novel



Nature Reviews | Microbiology
Vol 3, page 504, 2005

Viral metagenomic sequences from human faeces10, a marine sediment sample9 and two seawater samples2 were compared to the GenBank non-redundant database at the date of publication and in December 2004. The percentage of each library that could be classified as Eukarya, Bacteria, Archaea, viruses or showed no similarities (E-value >0.001) is shown.

New hypotheses:

Were DNA viruses involved in the origin of the eukaryotic cell nucleus?

Does evolution tinker with the wide genetic diversity offered by virus sequences, for example, through gene transfer?

2. ClustalW2

multiple sequence alignment

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

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- ClustalW2 Help
- ClustalW2 FAQ
- Jalview Help
- Scores Table
- Alignment
- Guide Tree
- Colours

- Similar Applications
  - Align
  - Kalign
  - MAFFT
  - MUSCLE
  - T-Coffee

- ClustalW Programmatic Access

- www.clustal.org

Clustal Related Literature

Search for Clustal related literature in Medline...
more

EBI > Tools > Sequence Analysis > ClustalW2

ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.
New users, please read the FAQ.
>> Download Software

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	
	Sequence	interactive	full	
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	NO END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	yes	def	def

ITERATION: none
NUMITER: 1

OUTPUT

OUTPUT FORMAT	OUTPUT ORDER
aln w/numbers	aligned

PHYLOGENETIC TREE

TREE TYPE	CORRECT DIST.	IGNORE GAPS	CLUSTERING
none	off	off	NJ

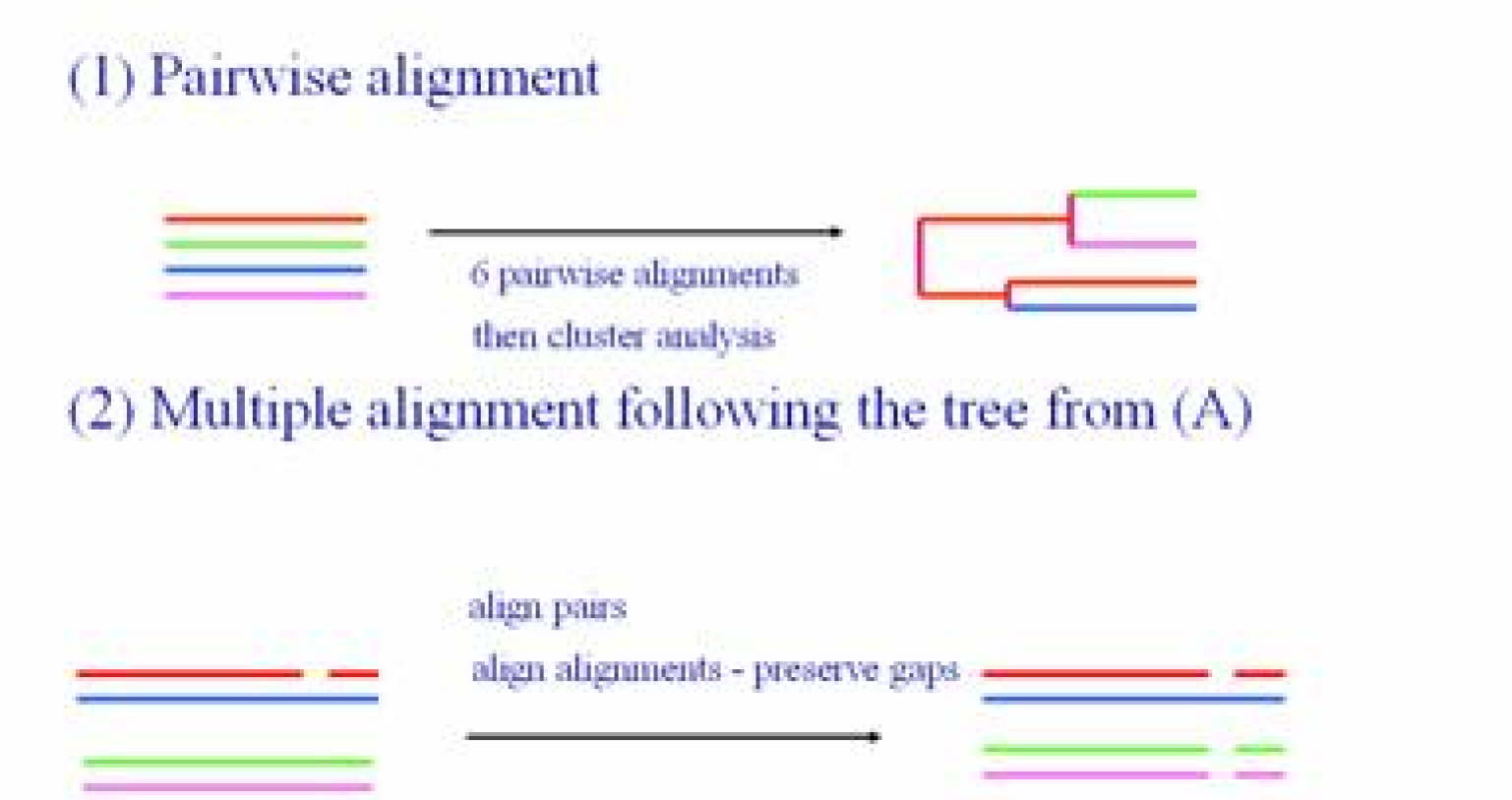
Enter or paste a set of sequences in any supported format: Help

Upload a file: Choose File no file selected Run Reset

ClustalW is a heuristic approach

- ✓It exploits the fact that homologous sequences are evolutionary related
- ✓Closely related sequences are aligned first gradually adding in the more distant ones.
- ✓Penalties are increased as more sequences are added
- ✓Different weight matrices (PAM and BLOSUM series) are used as the alignment proceeds because different matrices are used for highly-related sequences or more divergent sequences

Steps in a multiple alignment



USES of multiple sequence alignments

- ✓to find diagnostic patterns of protein families
- ✓to detect homology between new sequences and existing families of sequences
- ✓to help predict secondary and tertiary strucuture on new sequences
- ✓to determine evolutionary relationships: phylogenetics

alignment file

```
MDR1_HUMAN          YPSRKEVKILKGLNLKVQSGQTVALVGNSGCGKSTTVQLMQRLYDPTEGMVSVDGQDIRT 460
MDR3_HUMAN          YPSRANVKILKGLNLKVQSGQTVALVGSSGCGKSTTVQLIQRLYDPDEGTINIDGQDIRN 462
1B0U_A|PDBID        YGG---HEVLKGVSLQARAGDVISIIGSSGSGKSTFLRCINFLEKPSEGAIIVNGQNINL 72
STE6_YEAST          YPSRPSEAVLKNVSLNFSAGQFTFIVGKSGSGKSTLSNLLLRFYDGYNGSISINGHNIQT 425
TAP1_HUMAN          YPNRPDVLVLQGLTFTLRPGEVTALVGPNGSGKSTVAALLQNLYQPTGGQLLLDGKPLPQ 571
CFTR_HUMAN          NFSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMMIMGELEPSEGKIKHSG----- 486
                      .     :*:.:.:    *:   : * .*.**::    :    .   * :  .*
```

"*" means that the residues or nucleotides in that column are identical in all sequences in the alignment.

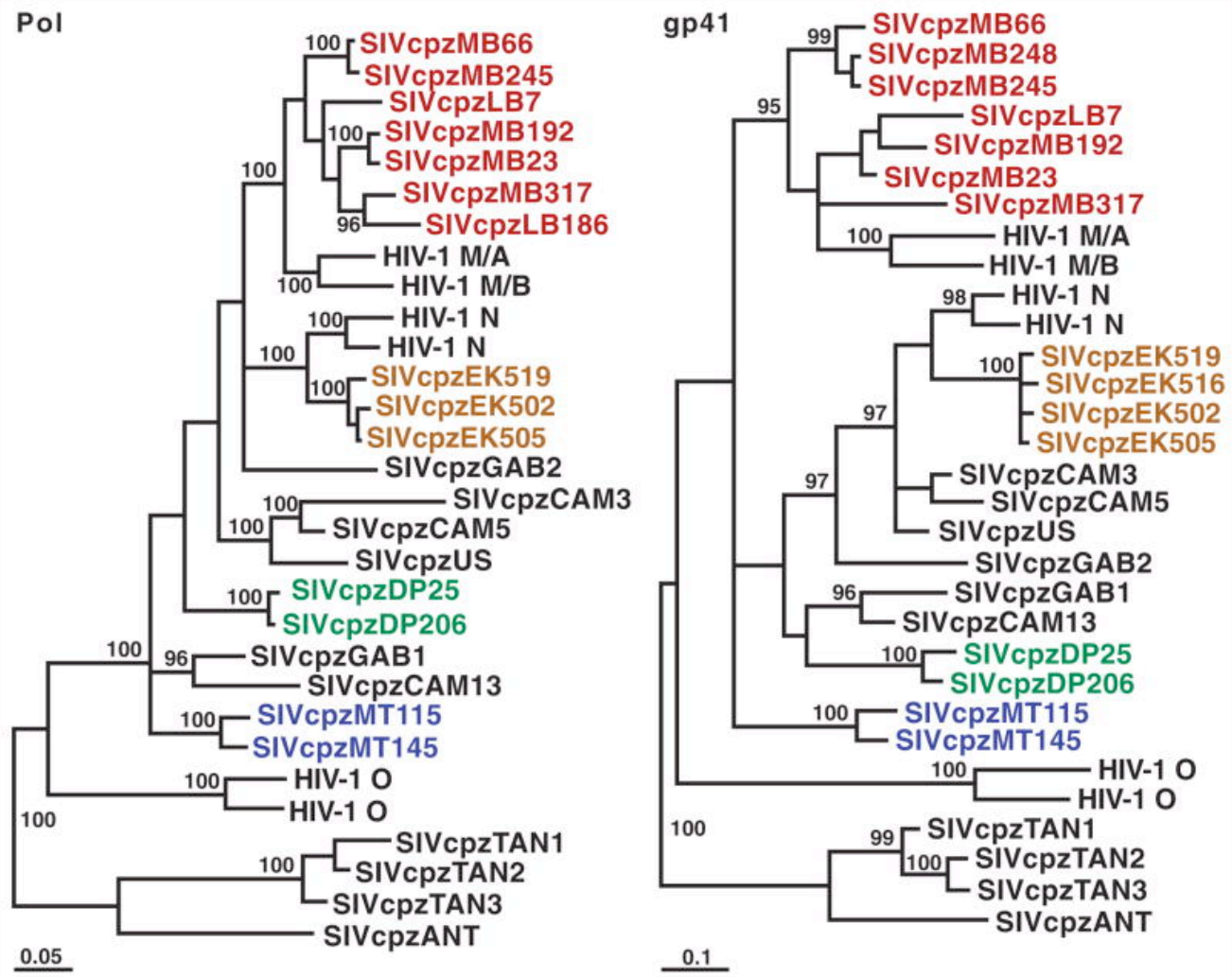
":" means that conserved substitutions have been observed, according to the CPLOR table at http://www.ebi.ac.uk/clustalw.

"." means that semi-conserved substitutions are observed.

PRESENCE OF STRICTLY CONSERVED AMINO ACID POSITIONS AMONG RELATED PROTEINS INDICATE THE LOCALIZATION OF KEY RESIDUES, WICH MIGHT BE IMPORTANT FOR FUNCTION

Published in final edited form as:
Science. 2006 July 28; 313(5786): 523–526.

Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1



3. Viewer 3D-protein structure

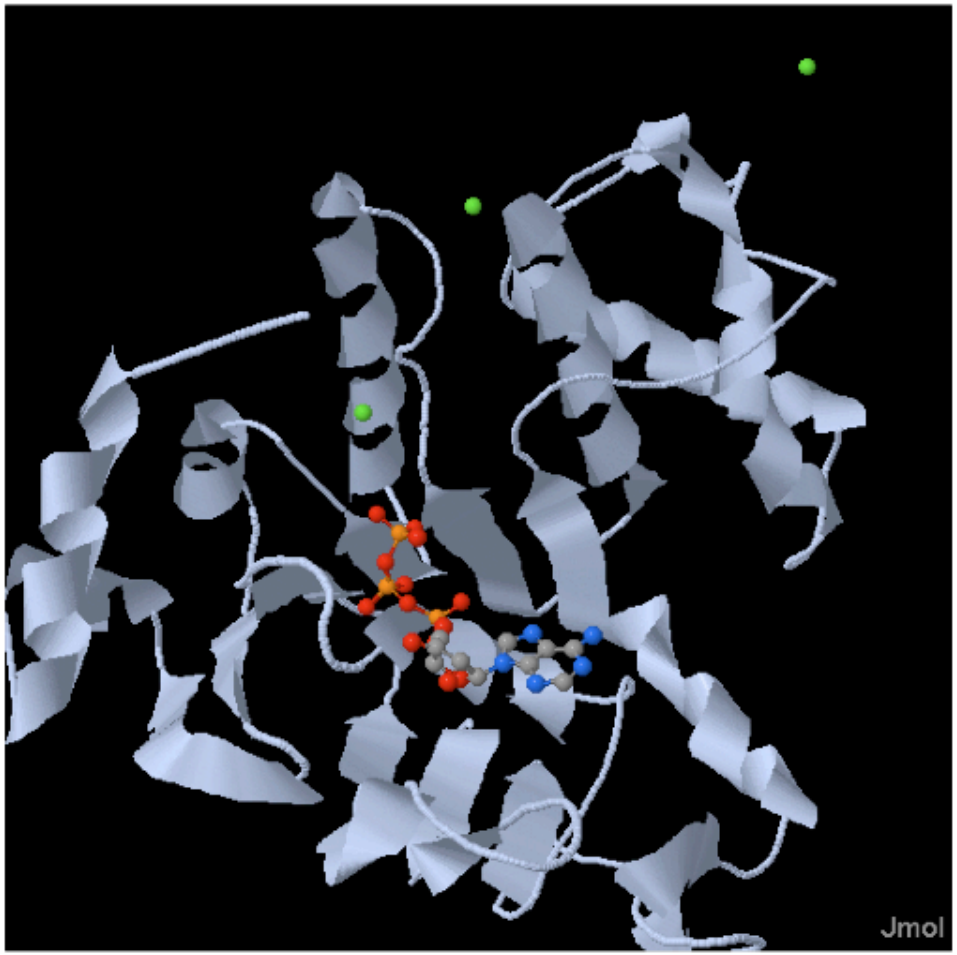
Protein Database

Jmol

For help select one of the options below:

- Help interacting with JMol
- Simple Interaction Guide (requires flash)
- Advanced JMol Help

1B0U



USES of 3D-viewers

- ✓ to annotate 3D-structure
- ✓ to predict functional consequences of mutations
- ✓ to identify drug targets
- ✓ to reveal evolutionary relationships not detected by sequence analysis

Unexpected similarities in coat protein topology suggest a common origin for viruses that infect bacteria and eukarya).

Curr Opin Struct Biol. 2005 Dec;15(6):655-63. Epub 2005 Nov 3

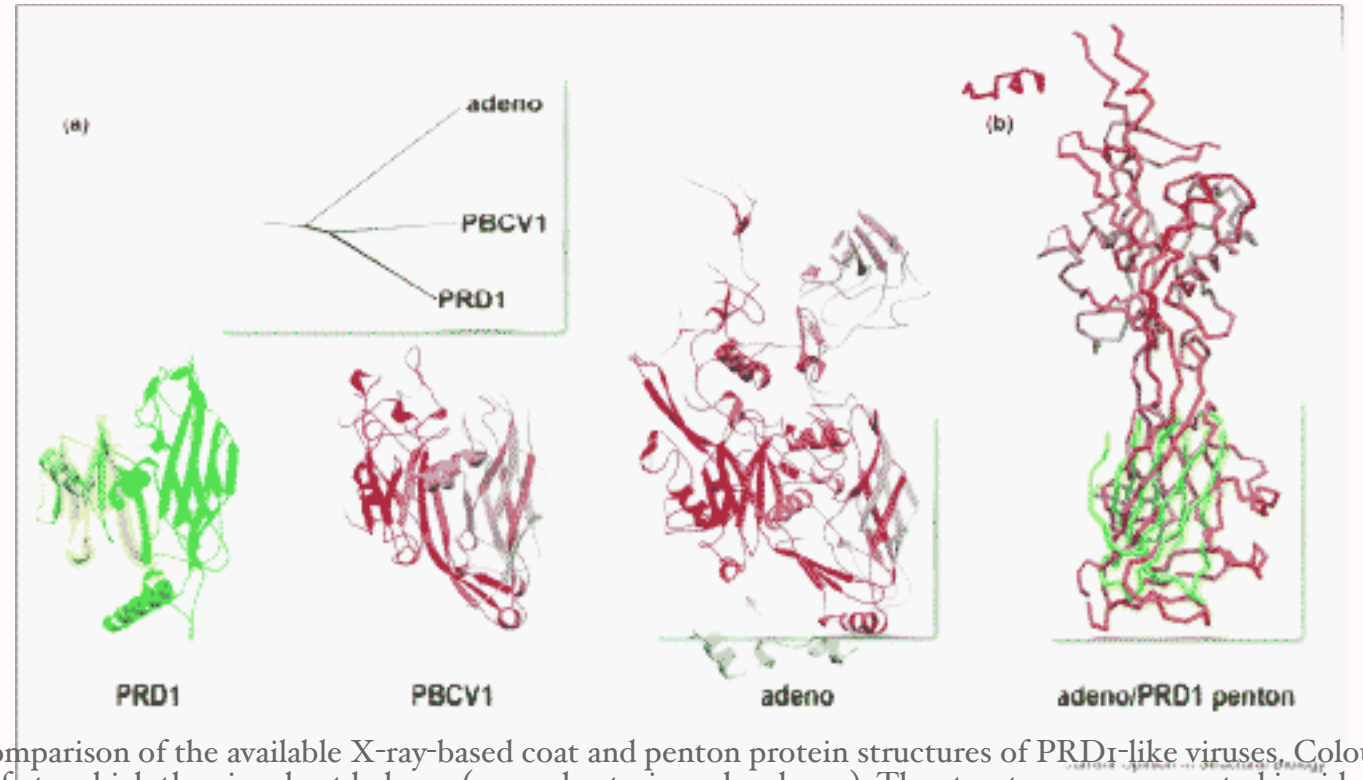


Figure 2. Comparison of the available X-ray-based coat and penton protein structures of PRD1-like viruses. Colouring is according to the domain of life to which the virus host belongs (green, bacteria; red, eukarya). The structures are presented in side view, so that the lower towards the virus interior. (a) The hexon-like proteins of PRD1, PBCV-1 and adenovirus. For adenovirus, the portion of the molecule comprising the two jelly-rolls is boxed (in performing the structure comparison, 728 of the 887 residues of adenovirus hexon were used, th deemed to comprise the jelly-rolls). Inset above is the relevant portion of the structure-based phylogenetic tree, which was calculated bas these structures. Note that the PBCV-1 protein, although somewhat closer to the equivalent protein of PRD1 than of adenovirus, also ha some structural features (e.g. the lower β extensions) that are closer to adenovirus. (b) The adenovirus penton compared to its stripped-d counterpart from PRD1. Note that the two β sheets that form the jelly-roll are packed at different angles in the two molecules; neverthe the superposition matches 87 out of <100 residues of the PRD1 jelly-roll.

The Exercise ...

go to the website

http://teodorolab.mcgill.ca/300D/

McGill

BIOC-300 Bioinformatics Mini-Project (BIMP)

Monday April 2nd, 2012

Contact

Coordinator

Dr Silvia Vidal
silvia.vidal@mcgill.ca [url]
514-398-2362

Meta-Instructions

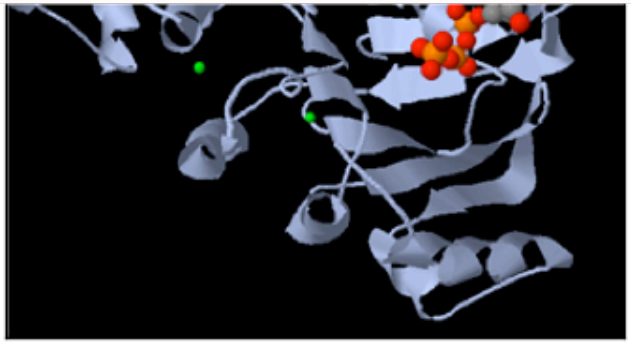
- **Write a flow sheet.** Not required, but you will find that it is much easier to perform the exercises when you organize yourself - just like in the wet lab.
- **Use the references/tutorials to guide you.** After attending the lecture, you will want to consolidate what you have seen by looking at the step-by-step tutorials written on each tool used in these exercise. The tutorials are in Chapter 3 of the "Introduction to Bioinformatics", a manual originally written for the Bioinformatics Project in the MicroImm lab course.
- **Write the report.** A good report should be concise and to the point. Show what you have understood, and present it in an orderly fashion, as suggested below.

Instructions

This exercise intends to show you how to use some of the most basic tools used in applied bioinformatics. This year, we have chosen to use the ATP-binding cassette (ABC) transporter superfamily to illustrate the use of these tools. The NCBI has a free online book on ABC transporters on their website.

Part One: BLAST

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. It is routinely used by millions of people to find sequences in extremely large databases. Typically, we say that we "BLAST" a sequence to say we "search" a sequence, without specifying where we look for it. The pool of sequences to search from is generally the NCBI sequence database, Genbank (which contained in 2005 about 60 billion bp, for 55 million sequence entries!).



The ABC transporter Histidine permease with ATP, seen with Jmol.

Your task is to use the human MDR1 sequence and perform a BLAST to discover similar sequences stored in public databases. We want you to be able to read the BLAST output and parse information from it.

Download: [Human MDR1 sequence]

Part Two: ClustalW

ClustalW is a multiple sequence alignment program. On top of aligning sequences, it may also be used to infer phylogenetic trees (while other programs are usually preferred).

Using ClustalW, we compare different ATB-binding cassette (ABC) transporters and try to find conserved motifs. ABC transporters are known to contain sequences that are involved in the binding of the ATP and the catalysis of the ATP hydrolysis reaction, which are present in all members of the superfamily. (What are ABC transporters? See the References section at the bottom of this page.)

Part Three: PDB

The Protein Data Bank is a repository of structural data, consisting overwhelmingly of protein structures. Much of the data can be appreciated best using standard molecular visualization tools. Classically, RasMol was used, but several new tools have been developed recently, many of which are web-based and just require Java to run.

Follow the instructions to:

1) identify human MDR1 protein homologs;

2) provide multiple sequence alignment of the NBD loop in ABC transporters;

3) visualize the structural localization of the ATP binding site in Histidine permease: highlight walker A, linker LSGGQ and walker B domains.

The Report

✓Once you have done the sequence manipulation in each of 3 parts, observe the results. Try to answer question for each point of the exercise. You may need to look at your results again. **Your written answer is required for all questions** in the exercise, either in point form, or in short-assay form.

✓Probably running the programs has taken you 15 min. Observation and reporting of the results might take you much longer but it is absolutely worth it.

✓The report should be at least 2 pages. However, if you are inspired go ahead. You may add figures, tables, references as you need. Please, no BLAST printouts out of the web.


Thank you and Good Luck!