Hands-on lab with databases

- Homework #1
- Quiz #1
- Summary: Nucleotide and protein databases
- Sequence formats
- Lab exercises

Homework #1

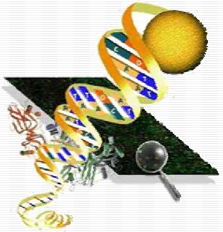
Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises

Genbank



Search and retrieval of sequences

Entrez

search and retrieval system

Entrez is a retrieval system for searching several linked databases. It provides access to: PubMed; Nucleotide; Protein; Structure; Genome; PopSet; OMIM; Taxonomy and more.

BLAST

BLAST® (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA.

Homework #1

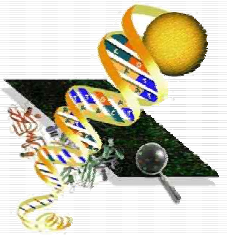
Quiz #1

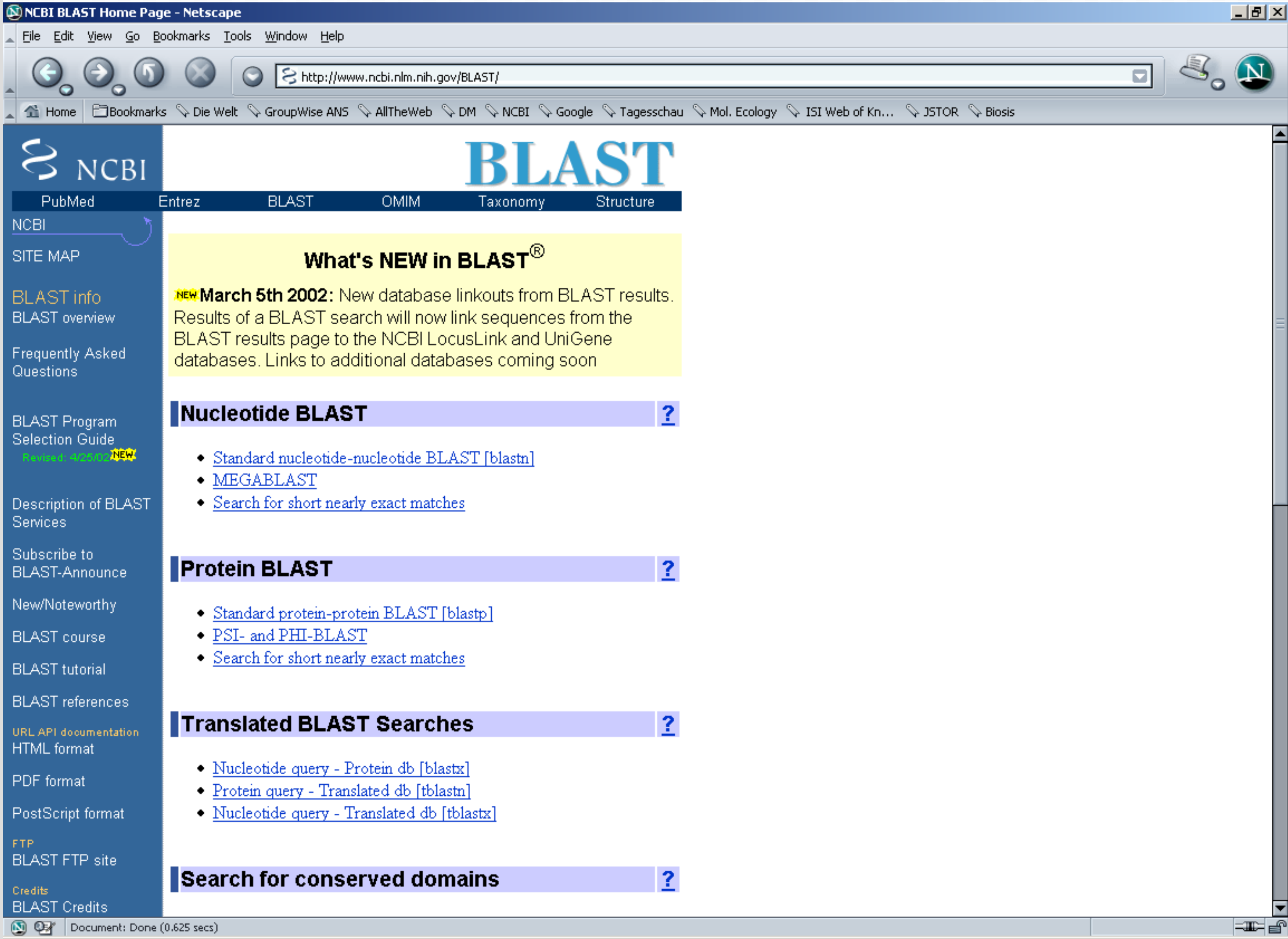
Summary: Nucleotide and protein databases

Sequence formats

Lab exercises

BLAST selections




NCBI BLAST Home Page - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/BLAST/
Home
Bookmarks
Die Welt
GroupWise ANS
AllTheWeb
DM
NCBI
Google
Tagesschau
Mol. Ecology
ISI Web of Kn...
JSTOR
Biosis
Document: Done (0.625 secs)

NCBI

BLAST

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NCBI

SITE MAP

BLAST info

BLAST overview

Frequently Asked Questions

BLAST Program Selection Guide

Revised: 4/25/02 NEW

Description of BLAST Services

Subscribe to BLAST-Announce

New/Noteworthy

BLAST course

BLAST tutorial

BLAST references

URL API documentation

HTML format

PDF format

PostScript format

FTP

BLAST FTP site

Credits

BLAST Credits

What's NEW in BLAST®

NEW **March 5th 2002:** New database linkouts from BLAST results. Results of a BLAST search will now link sequences from the BLAST results page to the NCBI LocusLink and UniGene databases. Links to additional databases coming soon

Nucleotide BLAST ?

- Standard nucleotide-nucleotide BLAST [blastn]
- MEGABLAST
- Search for short nearly exact matches

Protein BLAST ?

- Standard protein-protein BLAST [blastp]
- PSI- and PHI-BLAST
- Search for short nearly exact matches

Translated BLAST Searches ?

- Nucleotide query - Protein db [blastx]
- Protein query - Translated db [tblastn]
- Nucleotide query - Translated db [tblastx]

Search for conserved domains ?

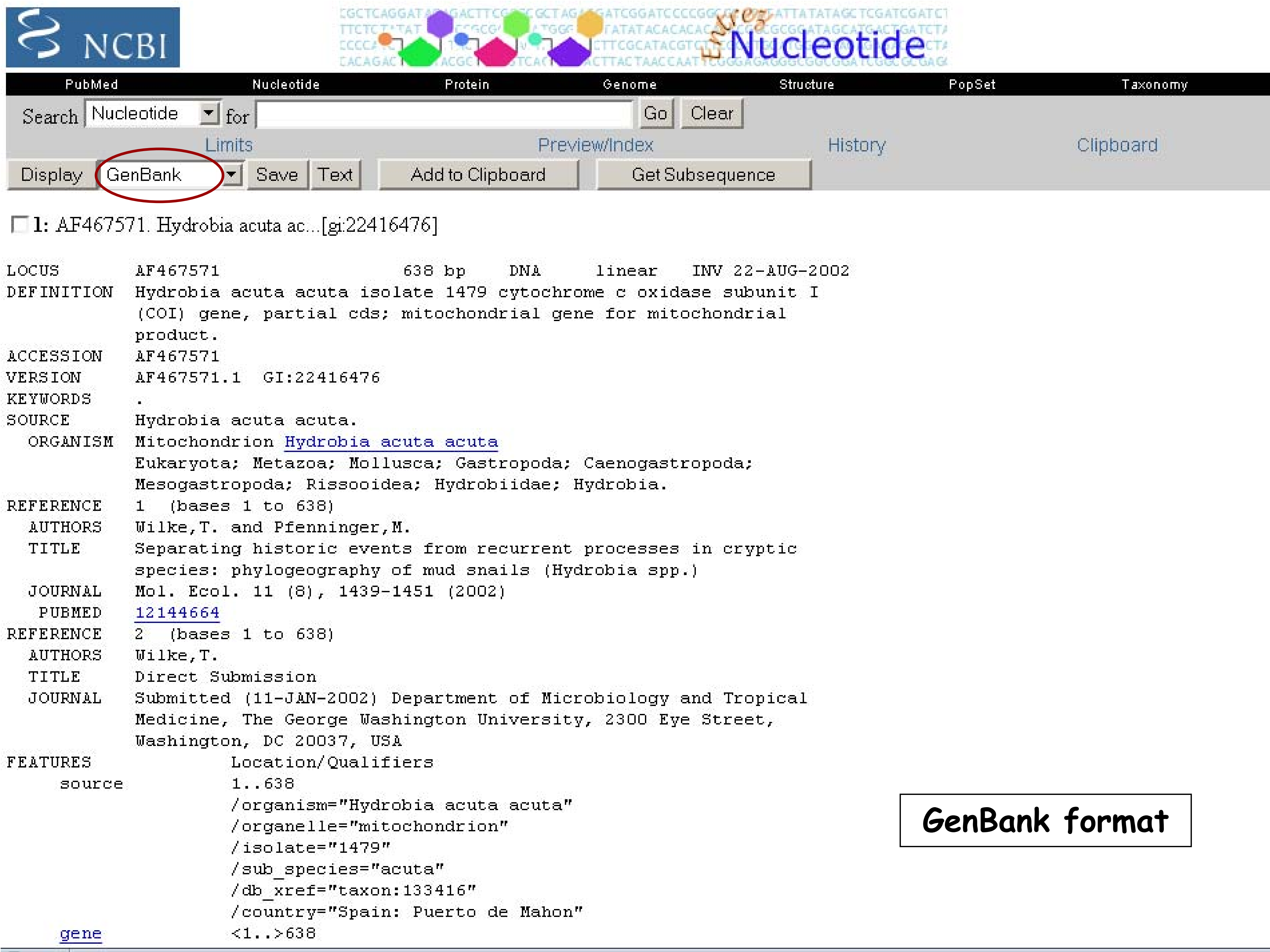
Homework #1

Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

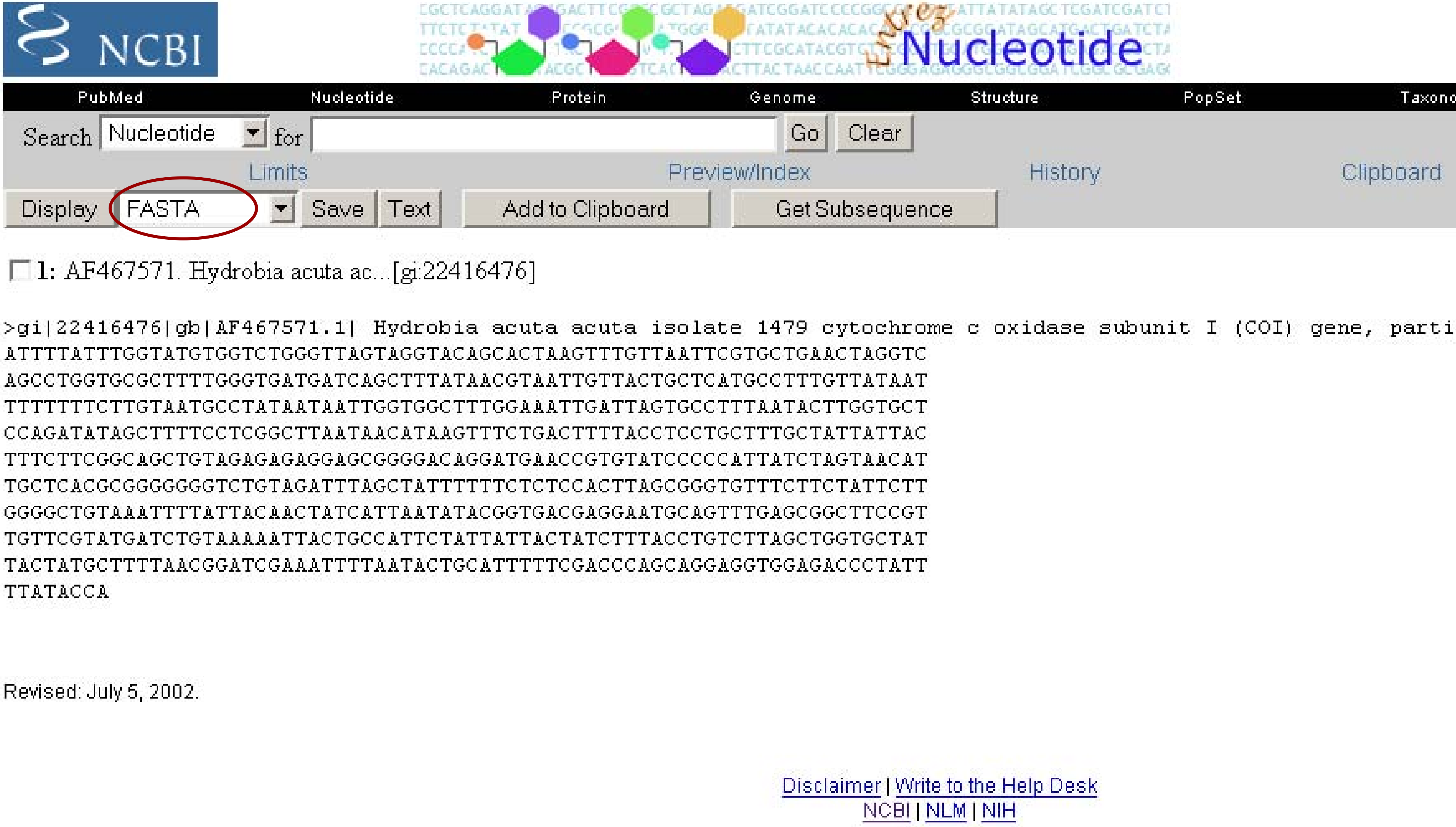
Lab exercises



□ 1: AF467571. Hydrobia acuta ac...[gi:22416476]

```
LOCUS       AF467571                 638 bp    DNA     linear   INV 22-AUG-2002
DEFINITION  Hydrobia acuta acuta isolate 1479 cytochrome c oxidase subunit I
            (COI) gene, partial cds; mitochondrial gene for mitochondrial
            product.
ACCESSION   AF467571
VERSION     AF467571.1  GI:22416476
KEYWORDS    .
SOURCE      Hydrobia acuta acuta.
  ORGANISM  Mitochondrion Hydrobia acuta acuta
            Eukaryota; Metazoa; Mollusca; Gastropoda; Caenogastropoda;
            Mesogastropoda; Rissooidea; Hydrobiidae; Hydrobia.
REFERENCE   1  (bases 1 to 638)
  AUTHORS   Wilke,T. and Pfenninger,M.
  TITLE     Separating historic events from recurrent processes in cryptic
            species: phylogeography of mud snails (Hydrobia spp.)
  JOURNAL   Mol. Ecol. 11 (8), 1439-1451 (2002)
   PUBMED   12144664
REFERENCE   2  (bases 1 to 638)
  AUTHORS   Wilke,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-JAN-2002) Department of Microbiology and Tropical
            Medicine, The George Washington University, 2300 Eye Street,
            Washington, DC 20037, USA
FEATURES             Location/Qualifiers
     source          1..638
                     /organism="Hydrobia acuta acuta"
                     /organelle="mitochondrion"
                     /isolate="1479"
                     /sub_species="acuta"
                     /db_xref="taxon:133416"
                     /country="Spain: Puerto de Mahon"
     gene            <1..>638
```

GenBank format



□ **1:** AF467571. Hydrobia acuta ac...[gi:22416476]

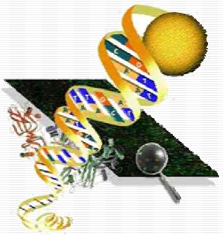
```
>gi|22416476|gb|AF467571.1| Hydrobia acuta acuta isolate 1479 cytochrome c oxidase subunit I (COI) gene, parti
ATTTTATTTGGTATGTGGTCTGGGTTAGTAGGTACAGCACTAAGTTTGTTAATTCGTGCTGAACTAGGTC
AGCCTGGTGCGCTTTTGGGTGATGATCAGCTTTATAACGTAATTGTTACTGCTCATGCCTTTGTTATAAT
TTTTTTTCTTGTAATGCCTATAATAATTGGTGGCTTTGGAAATTGATTAGTGCCTTTAATACTTGGTGCT
CCAGATATAGCTTTTCCTCGGCTTAATAACATAAGTTTCTGACTTTTACCTCCTGCTTTGCTATTATTAC
TTTCTTCGGCAGCTGTAGAGAGAGGAGCGGGGACAGGATGAACCGTGTATCCCCCATTATCTAGTAACAT
TGCTCACGCGGGGGGGTCTGTAGATTTAGCTATTTTTTCTCTCCACTTAGCGGGTGTTTCTTCTATTCTT
GGGGCTGTAAATTTTATTACAACTATCATTAATATACGGTGACGAGGAATGCAGTTTGAGCGGCTTCCGT
TGTTCGTATGATCTGTAAAAATTACTGCCATTCTATTATTACTATCTTTACCTGTCTTAGCTGGTGCTAT
TACTATGCTTTTAACGGATCGAAATTTTAATACTGCATTTTTCGACCCAGCAGGAGGTGGAGACCCTATT
TTATACCA
```

Revised: July 5, 2002.

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Fasta format

Sequence formats



- ASN.1
- DNAStrider
- EMBL
- Fitch
- GCG
- GenBank/GB
- IG/Stanford
- MSF
- NBRF
- Olsen
- PAUP/NEXUS
- Pearson/Fasta
- Phylip
- PIR/CODATA
- Plain/Raw
- Pretty
- Zuker

Convertible in ReadSeq (Web based)

http://bimas.dcrt.nih.gov/molbio/readseq/

or ForCon (stand-alone application)

http://www.hgmp.mrc.ac.uk/embnet.news/vol6_1/ForCon/forcon.html

NOTE:

- FASTA is a popular sequence format

- it also is a sequence similarity and homology search tool (similar to BLAST) used by EMBL-EBI

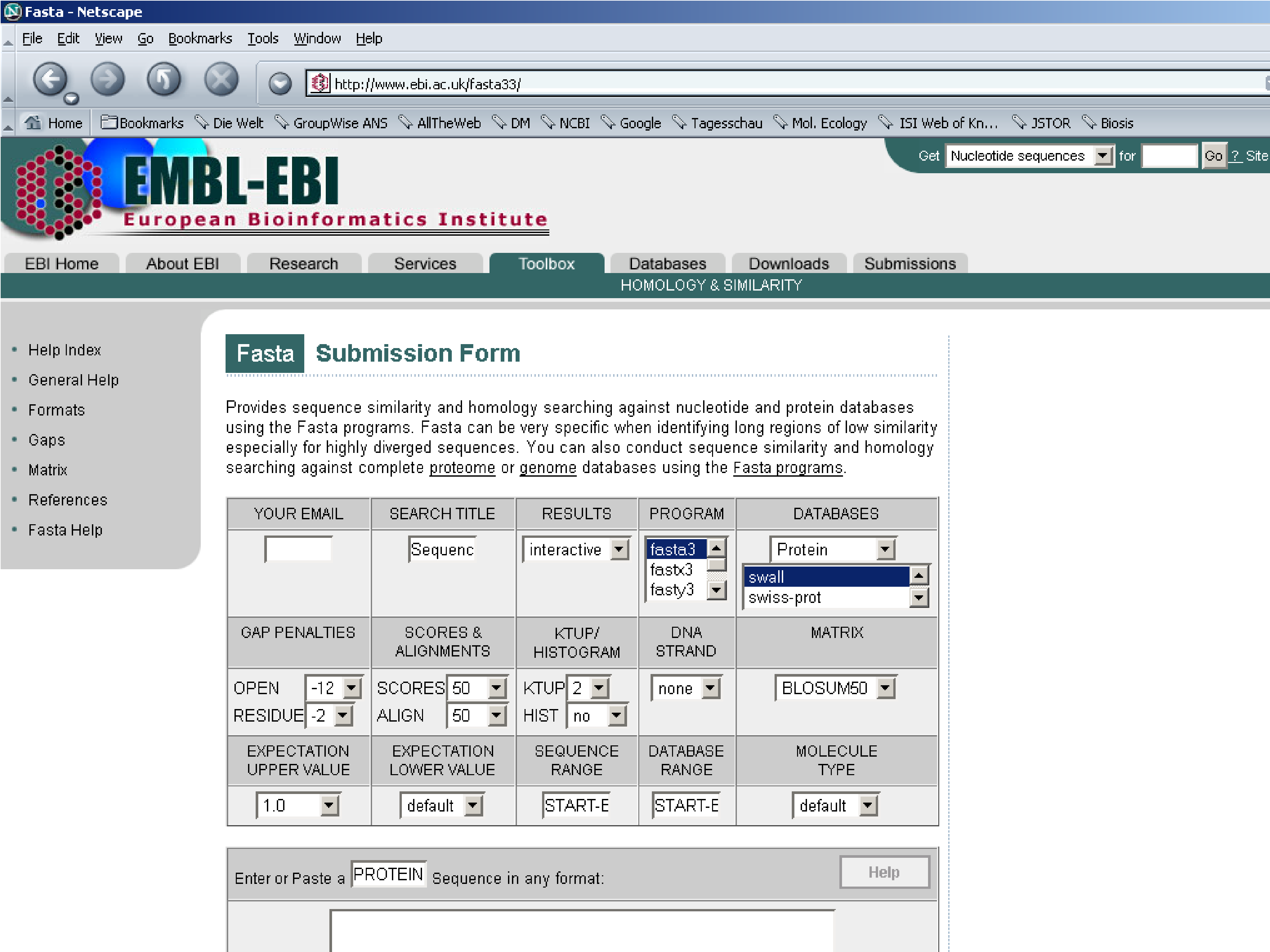
Homework #1

Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises



Fasta Submission Form

Provides sequence similarity and homology searching against nucleotide and protein databases using the Fasta programs. Fasta can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity and homology searching against complete proteome or genome databases using the Fasta programs.

YOUR EMAIL	SEARCH TITLE	RESULTS	PROGRAM	DATABASES
	Sequenc	interactive	fasta3 fastx3 fasty3	Protein swall swiss-prot
GAP PENALTIES	SCORES & ALIGNMENTS	KTUP/ HISTOGRAM	DNA STRAND	MATRIX
OPEN -12 RESIDUE -2	SCORES 50 ALIGN 50	KTUP 2 HIST no	none	BLOSUM50
EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE	SEQUENCE RANGE	DATABASE RANGE	MOLECULE TYPE
1.0	default	START-E	START-E	default

Enter or Paste a PROTEIN Sequence in any format:

Lab exercises

1) How many sequences are available in GenBank for Neanderthals?

Depends on your search strategy …

2) Go to **Entrez nucleotide**. Find all sequences for the following terms:

neander	**0**
Neanderthals	**0**
Neanderthal	**1**
neanderthal	**1**
neanderthal*	**5**
Homo sapiens neanderthalensis	**5**

2) Go to **Entrez taxonomy**. Try to find all sequences for Neanderthals!

5

Homework #1

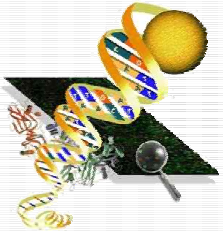
Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises

Lab exercises



4) How many nucleotide sequences are available for the house mouse *Mus musculus*? Try both **Entrez nucleotides** and **Entrez taxonomy**. How do you explain the difference?

Entrez taxonomy	**5.152.704**
Entrez nucleotides	**5.193.354 (*Mus musculus*)**
	5.152.741 (house mouse)
	5.193.375 (*Mus musculsus* OR house mouse)

5) A man is found murdered in Yellowstone National Park. Few hairs of unidentified origin are recovered on the victim's clothes. The samples arrive in the lab and DNA is isolated and sequenced:

CCATGCATATAAGCATGTACATAATATTATATTCTTACATAGGACATATTAACTCAATCTCATAATTCAT

Formulate a hypothesis regarding the origin of the recovered hairs and potential links with the killing!

***Canis lupus* (Gray Wolf)**

Homework #1

Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises

The Poliovirus Problem

Science magazine

VOL 297, 9 August 2002

Cello, J; Paul, A.V. & Wimmer, E.:

Chemical Synthesis of Poliovirus cDNA: Generation of Infectious Virus in the Absence of Natural Template

- they generated about 7.7 kilobases of single-stranded RNA genome based on the know genetic map
- DNA fragments were synthesized from purified oligonucleotides (average length 69: bases)
- the cDNA was then transcribed into highly infectious RNA

Homework #1

Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises

The Poliovirus Problem

washingtonpost.com

17 July 2002

Weiss, R.:

Mail-Order Molecules Brew a Terrorism Debate

- mail-order oligonucleotides can be used to manufacture a deadly virus
- because they are so small, most oligos lack a “fingerprint”
- call for more control and/or institutional oversight

Homework #1

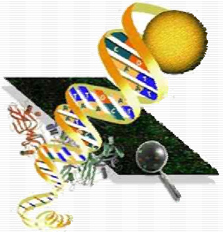
Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises

The Poliovirus Problem



You are a bioinformatician and the U.S. government has asked you to:

- assess the possibility of using oligonucleotides for manufacturing deadly viruses

- test whether small oligos have characteristic “fingerprints”

- design strategies for tracking oligo-orders

- suggest guidelines for the storage and retrieval of sequence information of relevance for bioterrorism

Homework #1

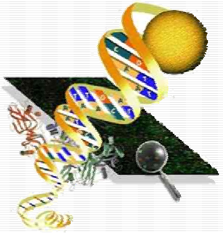
Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises

The Poliovirus Problem



Approach:

- which biological agents are of potential bioterrorism relevance? (See Centers for Disease Control and Prevention at **http://www.bt.cdc.gov/Agent/Agentlist.asp**)

- which genomes of Category A agents are available from public databases (GenBank)?

- how large are those genomes compared to the Polio virus?

- what is the average minimum sequence size identifiable as Polio virus?

- how can this information be used to track oligo orders?

- which information should be withhold from public databases?

Homework #1

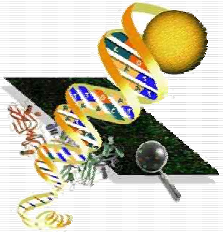
Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises

Homework assignment lecture #5



Explain in your own words and in simple terms the basics of the BLAST tool!

- assignment is due on 3 Mar 2003, 3:30 PM
- send your assignment as e-mail attachment to mtmtxw@gwumc.edu (type your name and the term "homework" in the subject line)
- maximum size: 500 words

Homework #1

Quiz #1

Summary: Nucleotide and protein databases

Sequence formats

Lab exercises