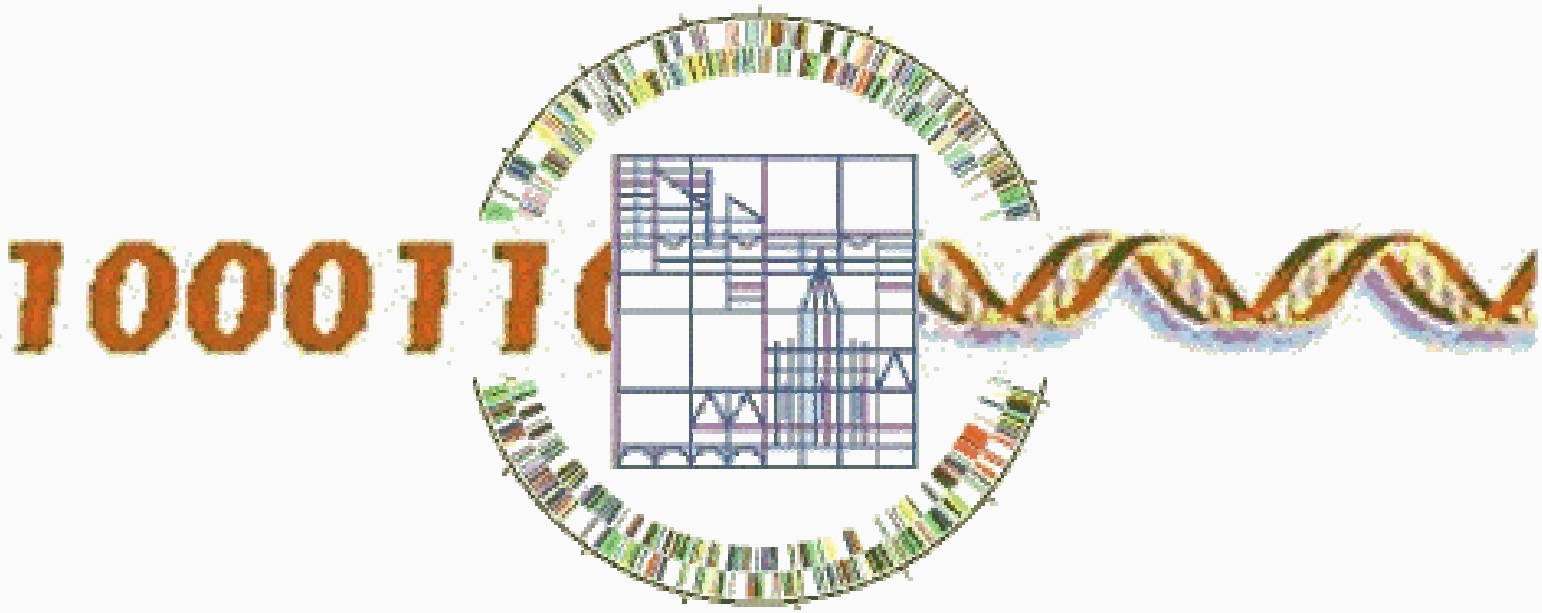
Bioinformatics and Drug Research

Klaus P. Schaefer



Databases in the WorldWide Web and Database Searching

Summer Course - Lecture 02 - Version of May 26, 2003

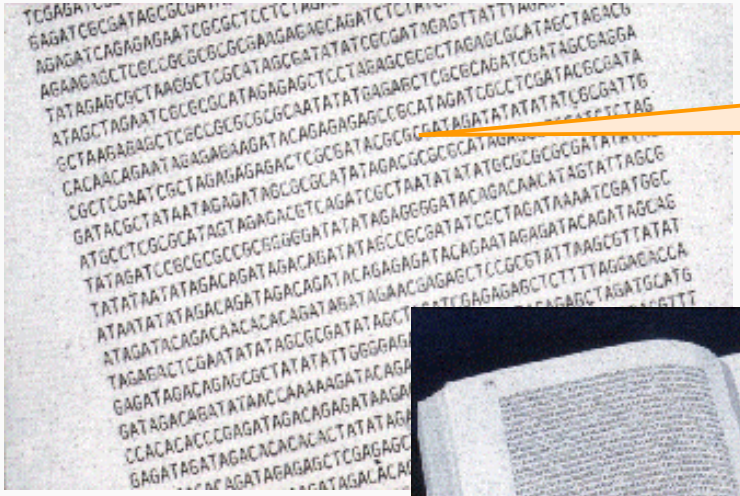
http://www.uni-konstanz.de/FuF/Bio/Bioinformatik/



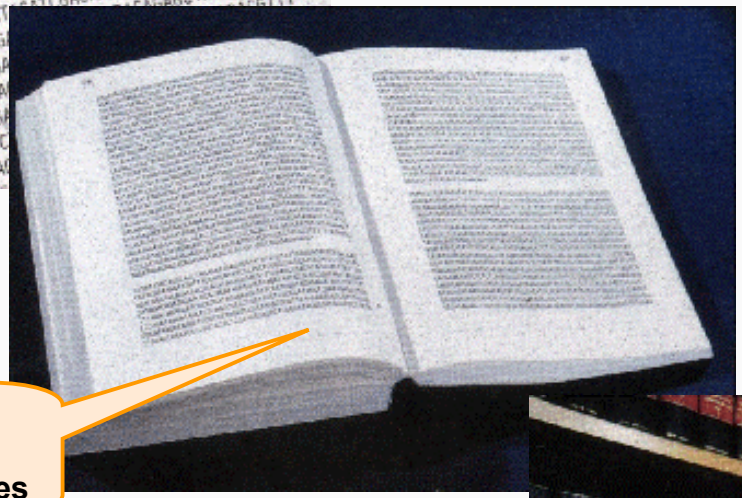
Bioinformatics DataBanks

- DataBanks store biological data in various forms
  - Sequences
    - DNA, RNA (Nucleic Acids)
    - Proteins
    - Structural data
    - X-ray crystallography data
  - Expression Data
    - Transcription of all genes
  - Interactions Data
    - Protein-Protein interactions
    - Receptor Binding Data
    - Substrate Binding Data

The Ultimate Challenge – the Human Genetic Blueprint



Virus Genome
3000 Base Pairs
1 Page with 2000 Characters



Bacterial Genome
3 Million Base Pairs
1 Book with 1000 Pages



Human Genome
3 Billion Base Pairs
Library with 1000 Books

Before You Can Start You Need Access to the Internet ...

http://www.microsoft.com/windows/ie/default.asp



http://www.netscape.com/

http://www.opera.com

The three leading Browsers
Microsoft Internet Explorer
Netscape Navigator
Opera

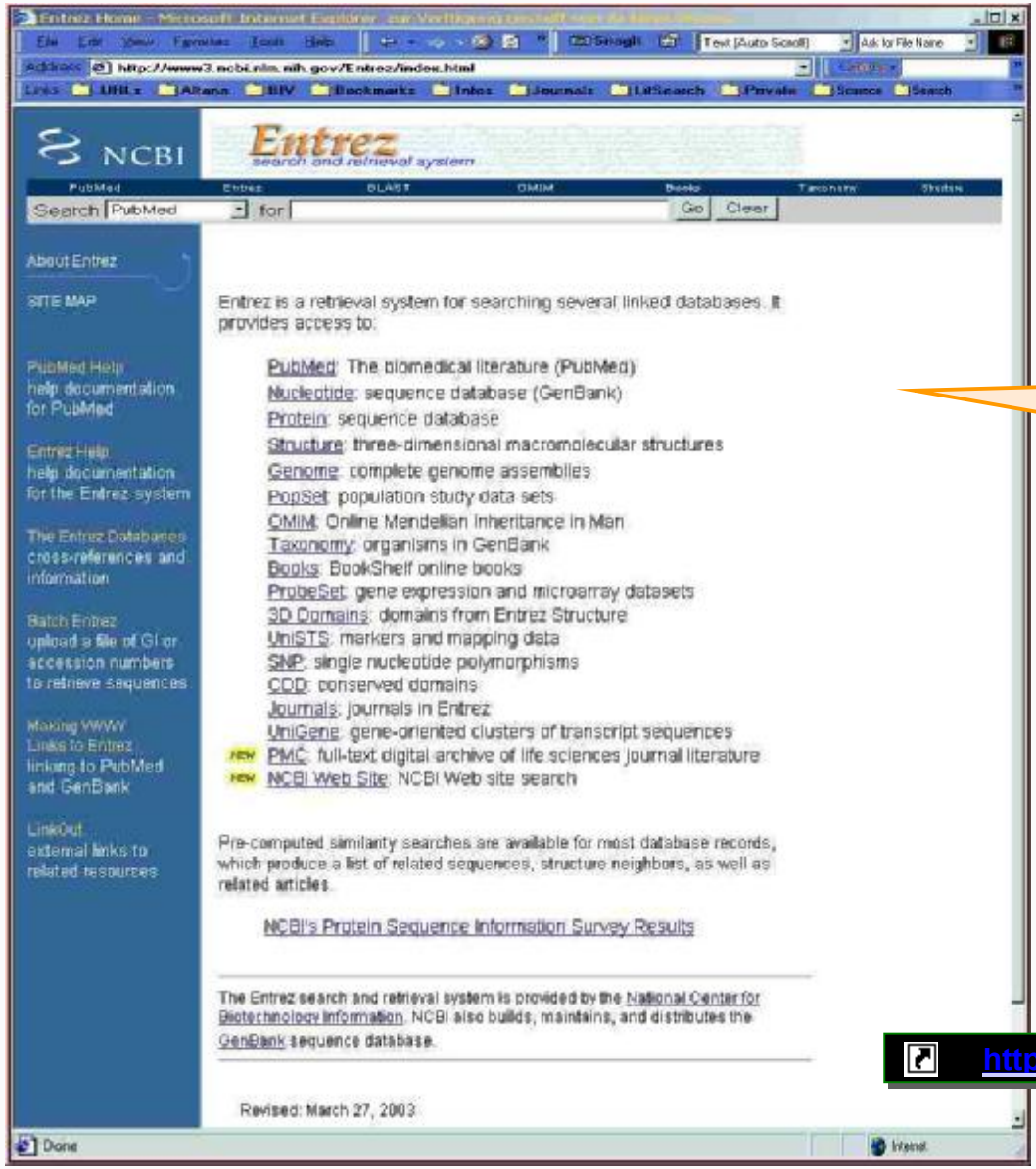
The National Center for Biotechnology Information (NCBI)



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

http://www3.ncbi.nlm.nih.gov/

ENTREZ HomePage



Via the ENTREZ Site you have access to a whole collection of related information, e.g. the MEDLINE / PubMed Literature DataBank

WebSite Infos - Entrez
Entrez is a search and retrieval system that integrates information from databases at NCBI. These databases include nucleotide sequences, protein sequences, macromolecular structures, whole genomes, and MEDLINE, through PubMed.

http://www3.ncbi.nlm.nih.gov/Entrez/index.html

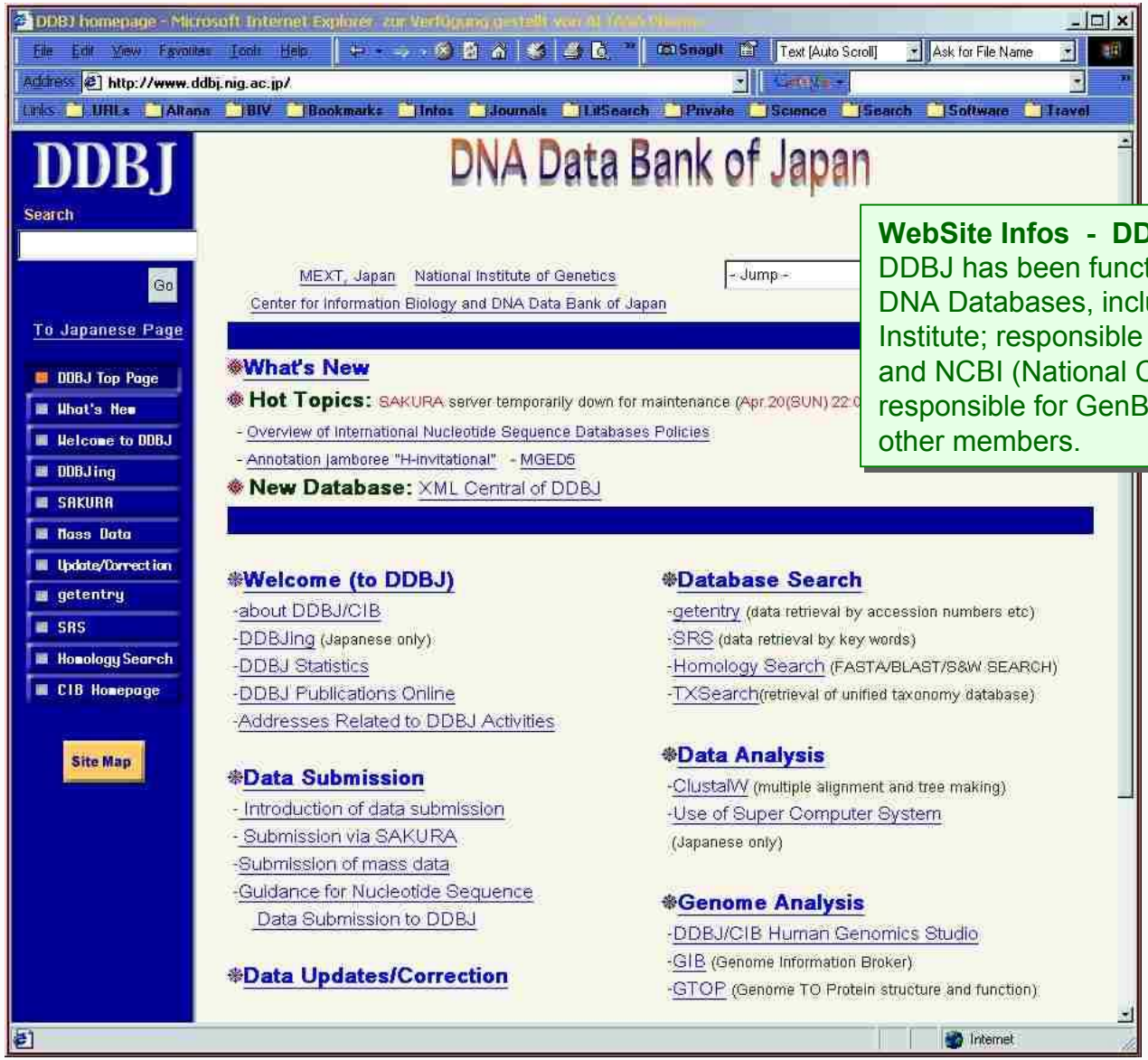
The European Bioinformatics Institute (EBI - EMBL Database)



WebSite Infos – EBI The EMBL Nucleotide Sequence Database
The EMBL Nucleotide Sequence Database constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

http://www.ebi.ac.uk/embl/index.html

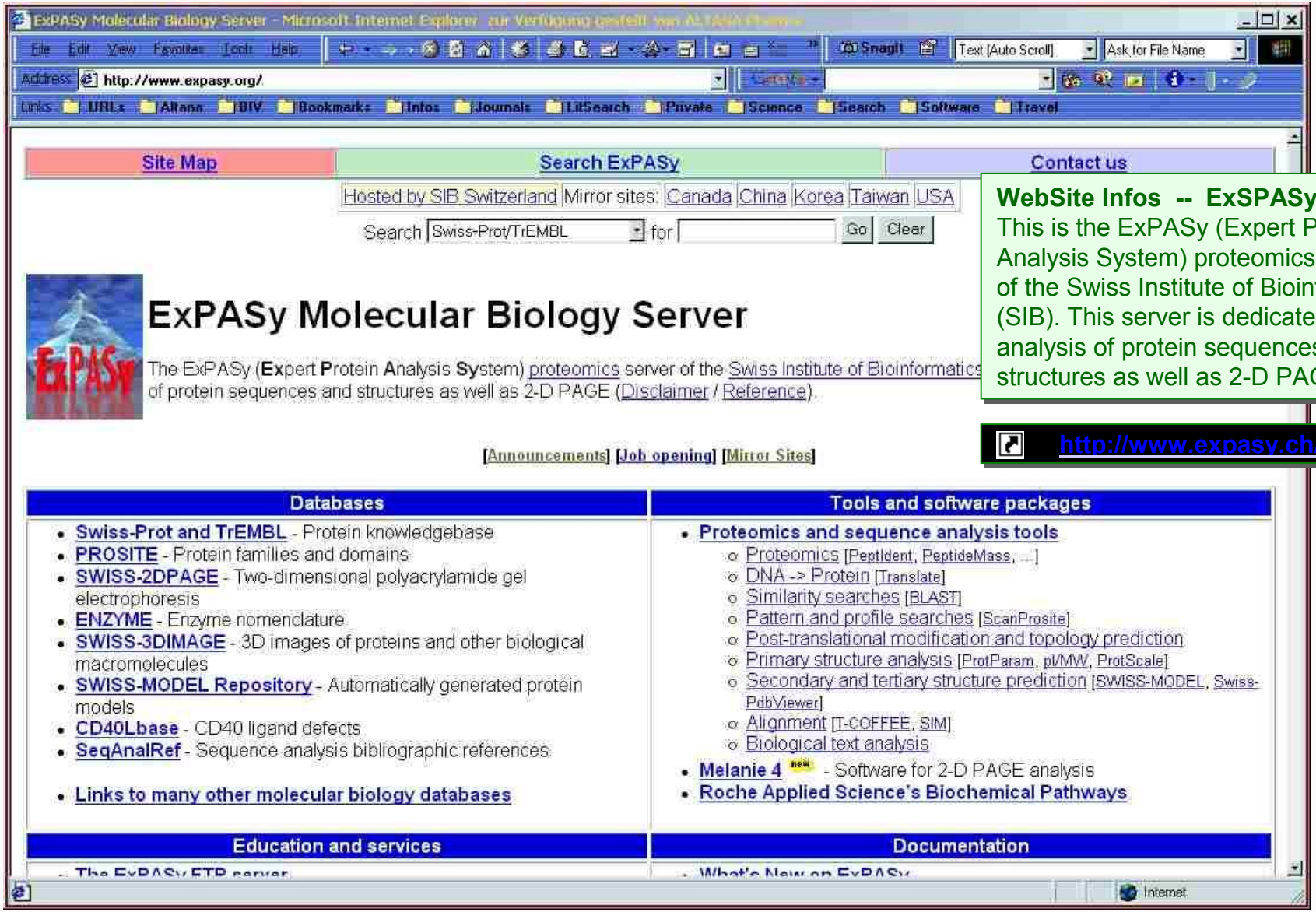
The DDBJ (DNA DataBank of Japan)



WebSite Infos - DDBJ (DNA Data Bank of Japan)
DDBJ has been functioning as one of the International DNA Databases, including EBI (European Bioinformatics Institute; responsible for the EMBL database) in Europe and NCBI (National Center for Biotechnology Information; responsible for GenBank database) in the USA as the two other members.

http://www.ddbj.nig.ac.jp/

The ExPASy HomePage

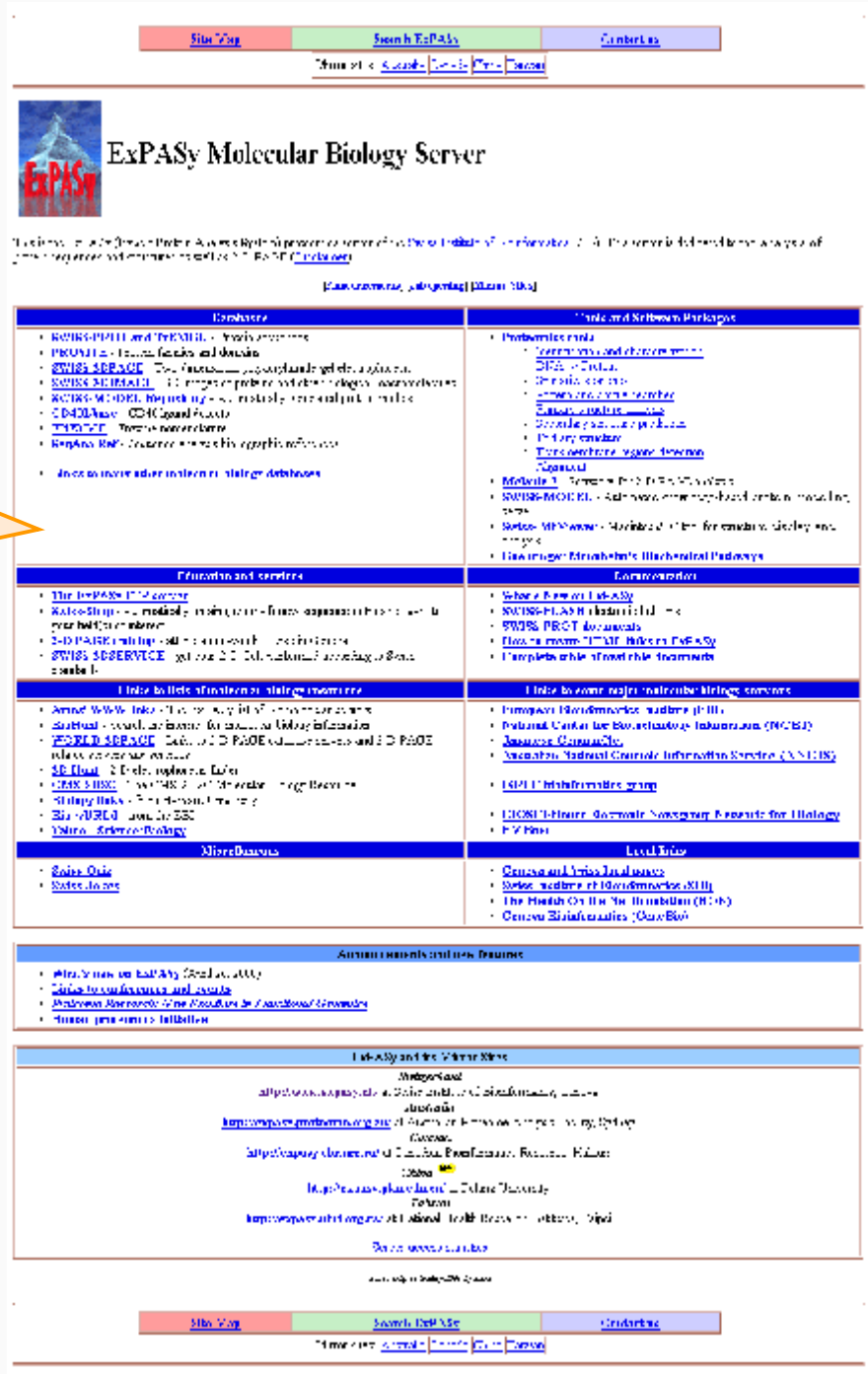


WebSite Infos -- ExSPASy
This is the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB). This server is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE

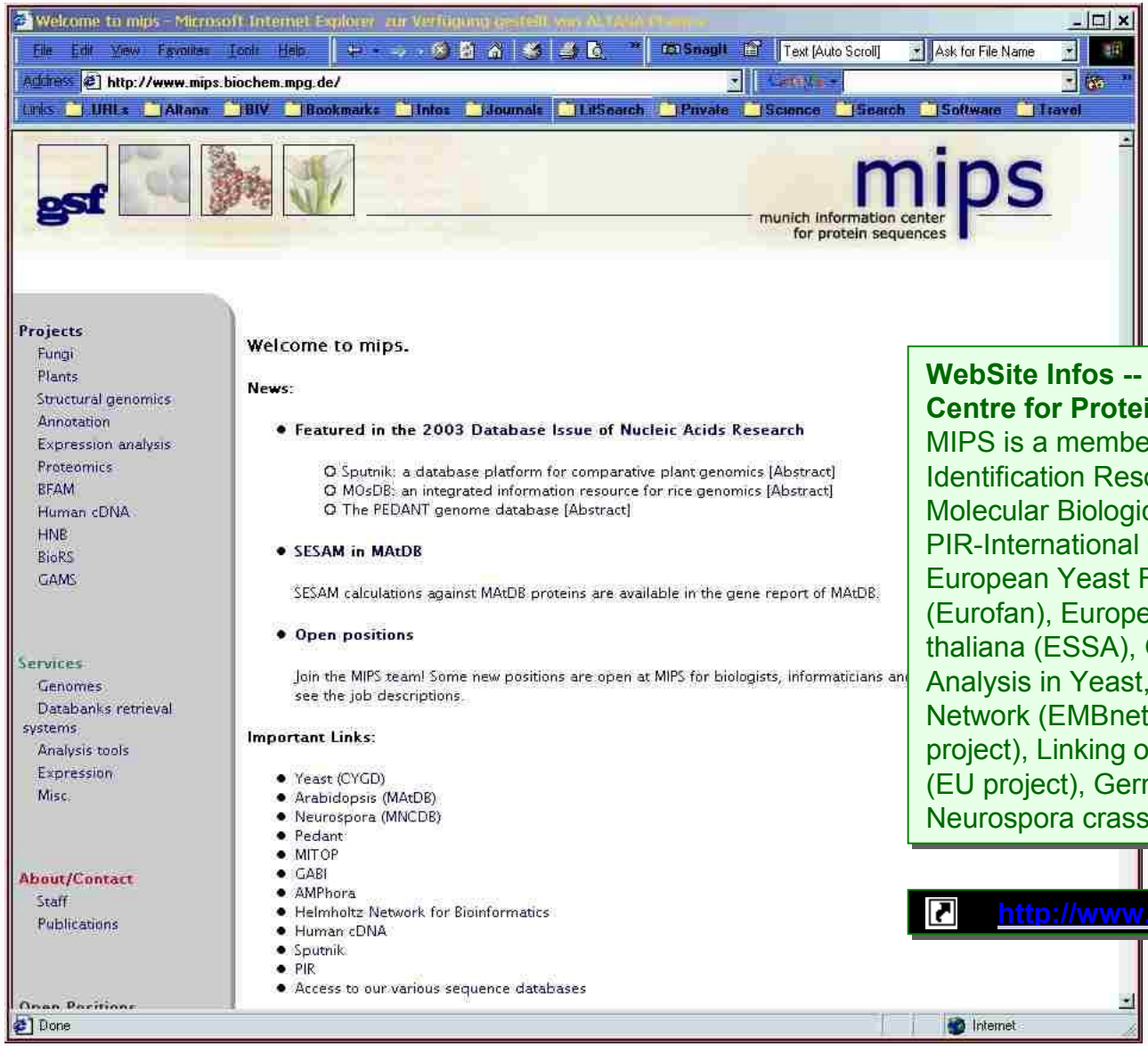
http://www.expasy.ch/

The ExPASy HomePage II

This server originated from a protein sequence database, but has extended its suite of programs into nucleic acids and especially structures.
A new addition is the access to proteomic databases.



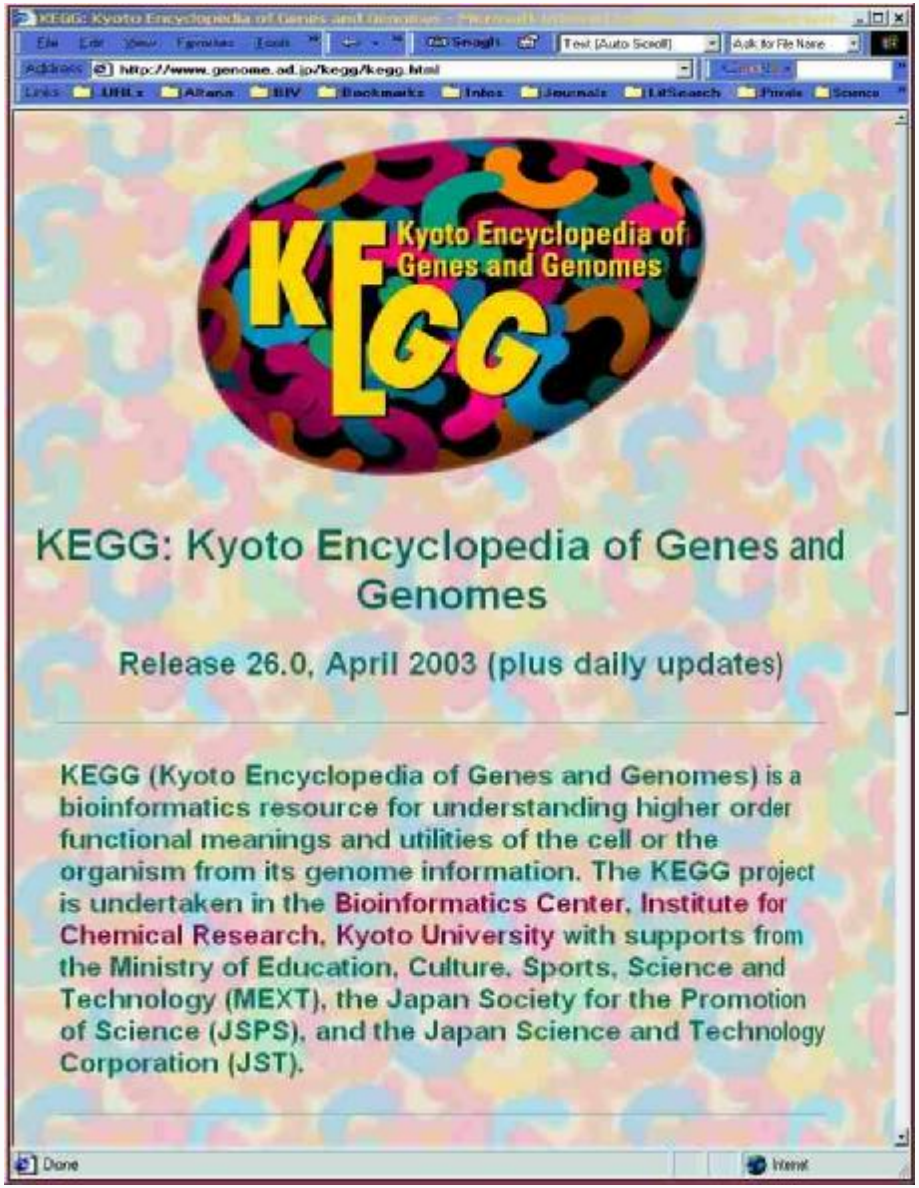
The Munich Information Center for Protein Sequences (MIPS)



WebSite Infos -- MIPS is the Munich Information Centre for Protein Sequences.
MIPS is a member of PIR-International (Protein Identification Resource) and of EMBNET (European Molecular Biological Network). Contributed Projects: PIR-International Protein Sequence Database European Yeast Functional Analysis Program (Eurofan), European Sequencing of Arabidopsis thaliana (ESSA), German Network for Functional Analysis in Yeast, European Molecular Biology Network (EMBnet), Integrated Databases (EU project), Linking of Biological Databases, CORBA (EU project), German Sequencing project of Neurospora crassa (DFG funded) .

http://www.mips.biochem.mpg.de/

The KEGG DataBank Resource



WebSite Infos - Kyoto Encyclopedia of Genes and Genomes (KEGG)
KEGG is an effort to computerize current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting molecules or genes and to provide links from the gene catalogs produced by genome sequencing projects. The KEGG project is undertaken in the Bioinformatics Center, Institute for Chemical Research, Kyoto University with supports from the Ministry of Education, Culture, Sports, Science and Technology and the Japan Society for the Promotion of Science.

http://www.genome.ad.jp/kegg/kegg.html

KEGG Links You to Metabolic DataBases ...

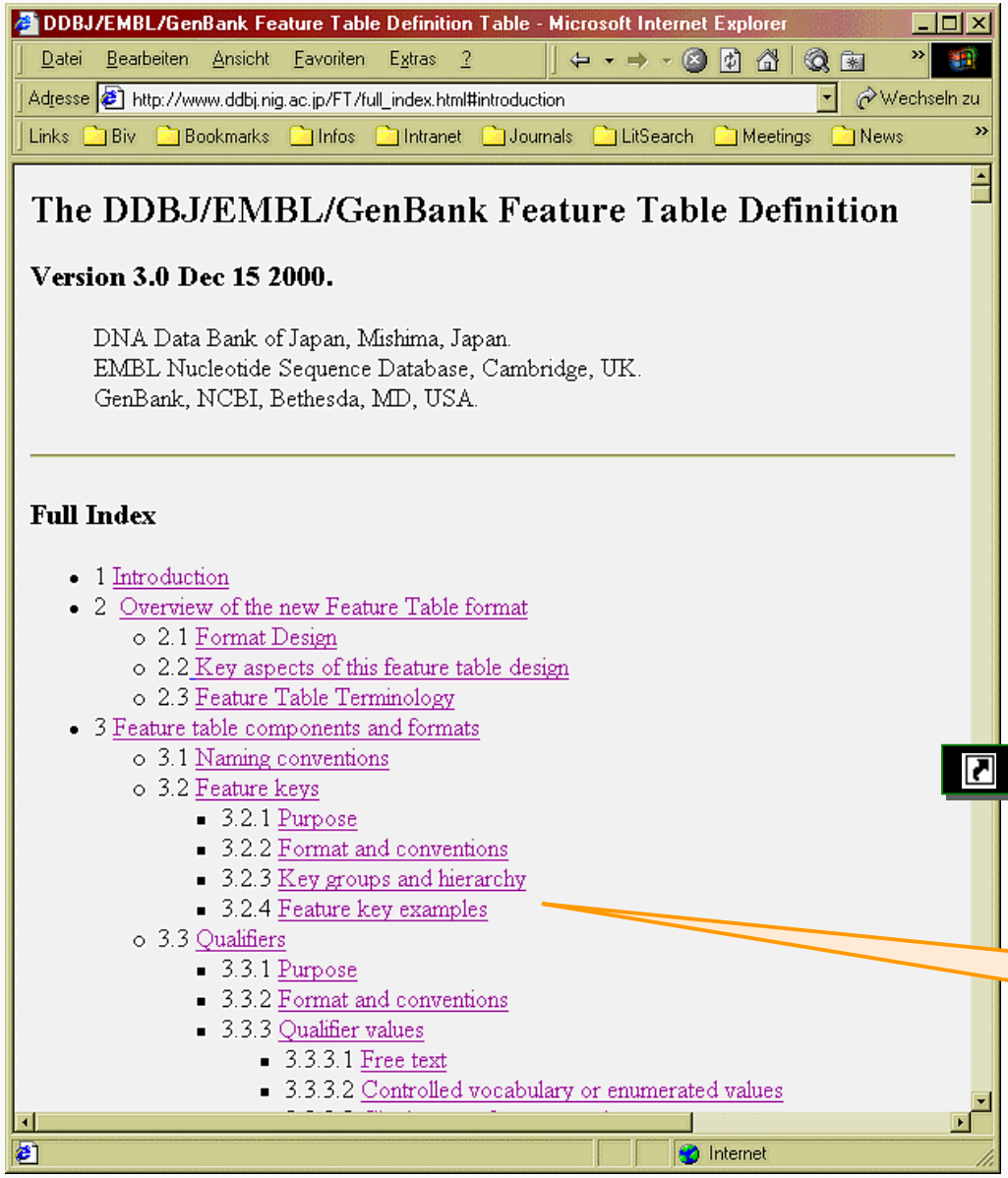


Pathway Information

Maps and Ortholog Tables

Cate	Pathway Map / Ortholog Table	Search & Compute	DBGET Search
Pathway	Metabolic pathways	Search objects in pathway maps Color objects in pathway maps Search or color genes in ortholog tables	PATHWAY GENES
	Regulatory pathways	Search similar sequences in pathway maps Search similar sequences in ortholog tables Generate possible reaction pathways	

Information on DDBJ/EMBL/GenBank Sequence Files

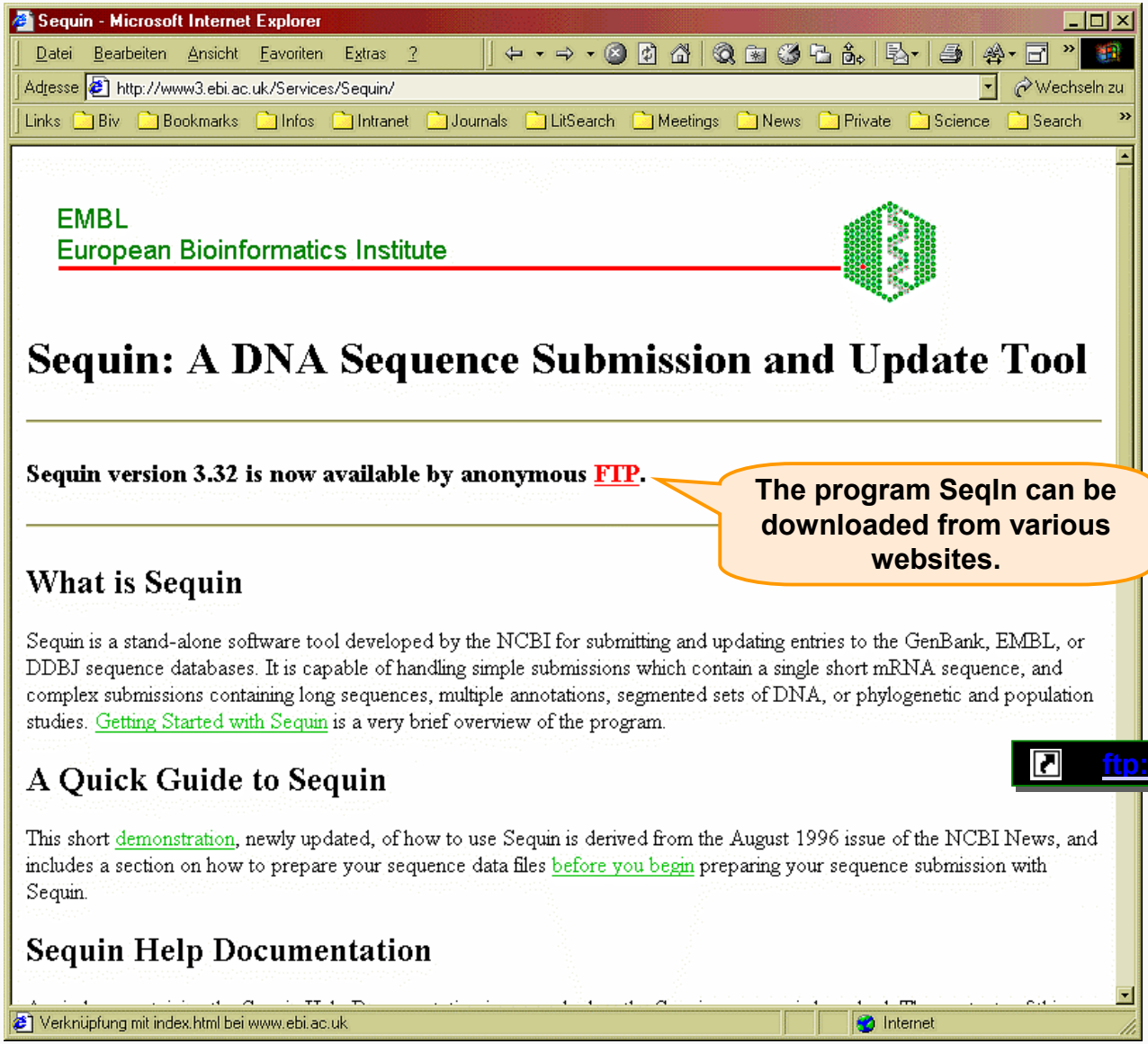


WebSite Infos - Nucleic acid sequences provide the fundamental starting point for describing and understanding the structure, function, and development of genetically diverse organisms. The GenBank, EMBL, and DDBJ nucleic acid sequence data banks have from their inception used tables of sites and features to describe the roles and locations of higher order sequence domains and elements within the genome of an organism.
In February, 1986, GenBank and EMBL began a collaborative effort (joined by DDBJ in 1987) to devise a common feature table format and common standards for annotation practice.

http://www.ddbj.nig.ac.jp/FT/full_index.html

This site offers detailed instruction on DataBanks and their files as well as info on search strategies

How to Submit Data?

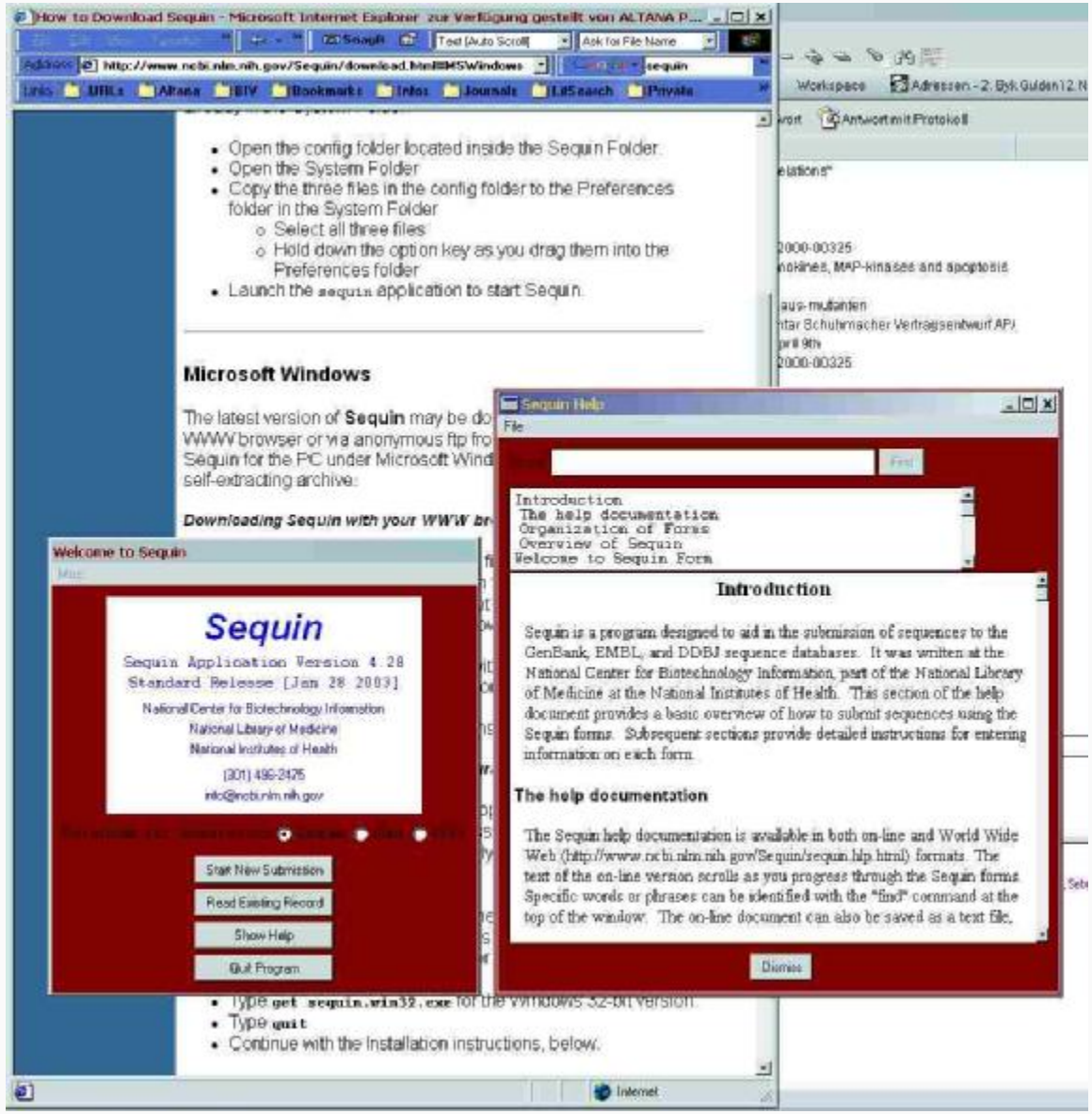


WebSite Infos - Sequin: A DNA Sequence Submission and Update Tool

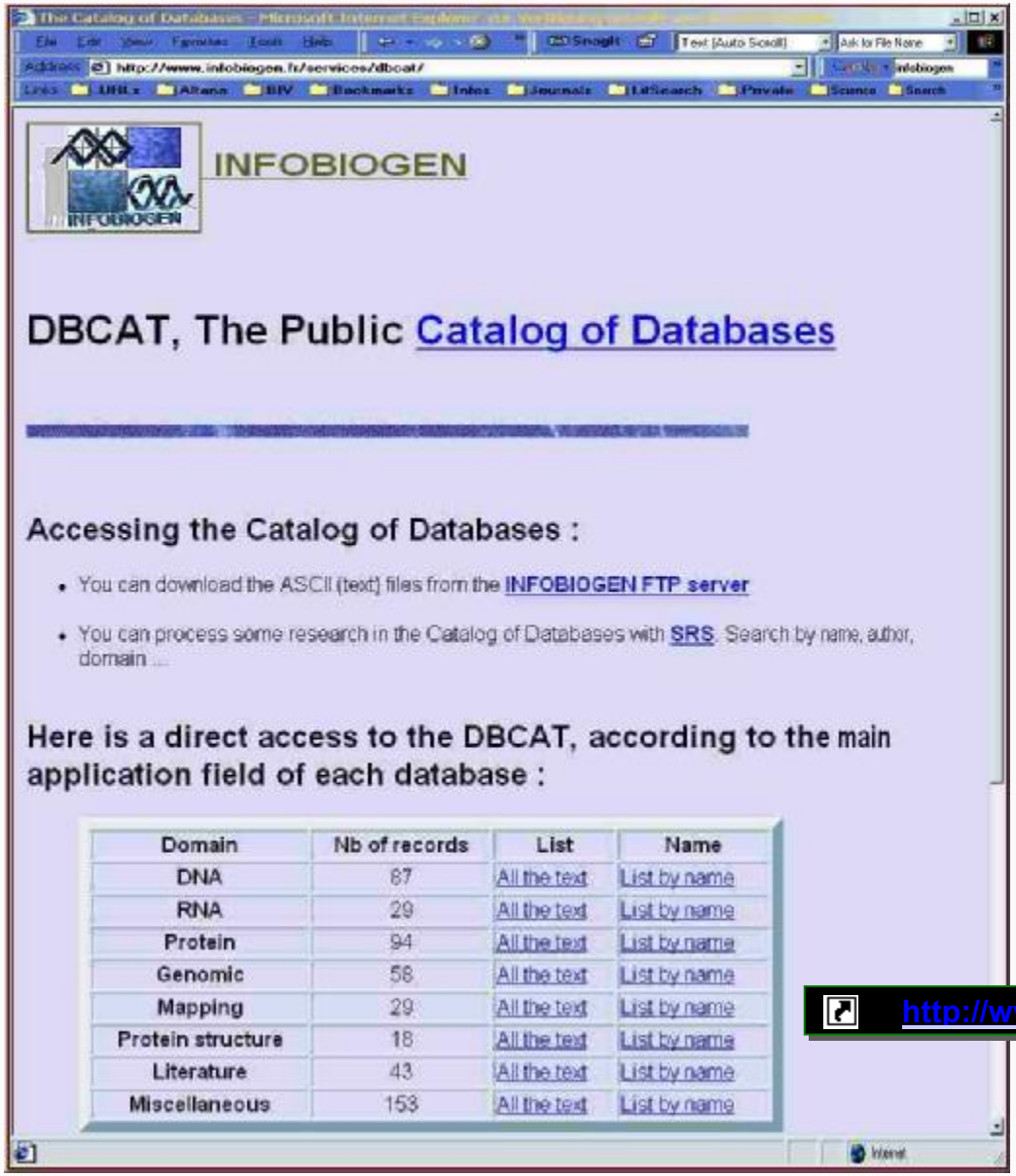
Sequin is a stand-alone software tool developed by the NCBI for submitting and updating entries to the GenBank, EMBL, or DDBJ sequence databases. It is capable of handling simple submissions which contain a single short mRNA sequence, and complex submissions containing long sequences, multiple annotations, segmented sets of DNA, or phylogenetic and population studies. Getting Started with Sequin is a very brief overview of the program.

ftp://ftp.ebi.ac.uk/pub/software/sequin

SeqIn Can Be Used as a Stand-Alone Application on any PC



The Infobiogen Site Offers Access to a Catalog of DataBanks



INFOBIOGEN

DBCAT, The Public Catalog of Databases

Accessing the Catalog of Databases :

- You can download the ASCII (text) files from the INFOBIOGEN FTP server
- You can process some research in the Catalog of Databases with SRS. Search by name, author, domain ...

Here is a direct access to the DBCAT, according to the main application field of each database :

Domain	Nb of records	List	Name
DNA	67	All the text	List by name
RNA	29	All the text	List by name
Protein	94	All the text	List by name
Genomic	58	All the text	List by name
Mapping	29	All the text	List by name
Protein structure	18	All the text	List by name
Literature	43	All the text	List by name
Miscellaneous	153	All the text	List by name

WebSite Infos - DBCAT, The Catalog of Databases
DBCAT is an attempt to gather all existing databases, as the Biocatalog for the software. The project is a collaboration with the EBI (European Bioinformatics Institute).

http://www.infobiogen.fr/presentation/servindex.html

The NCGR Site Offers Access to Whole Genome Resources



http://www.ncgr.org/

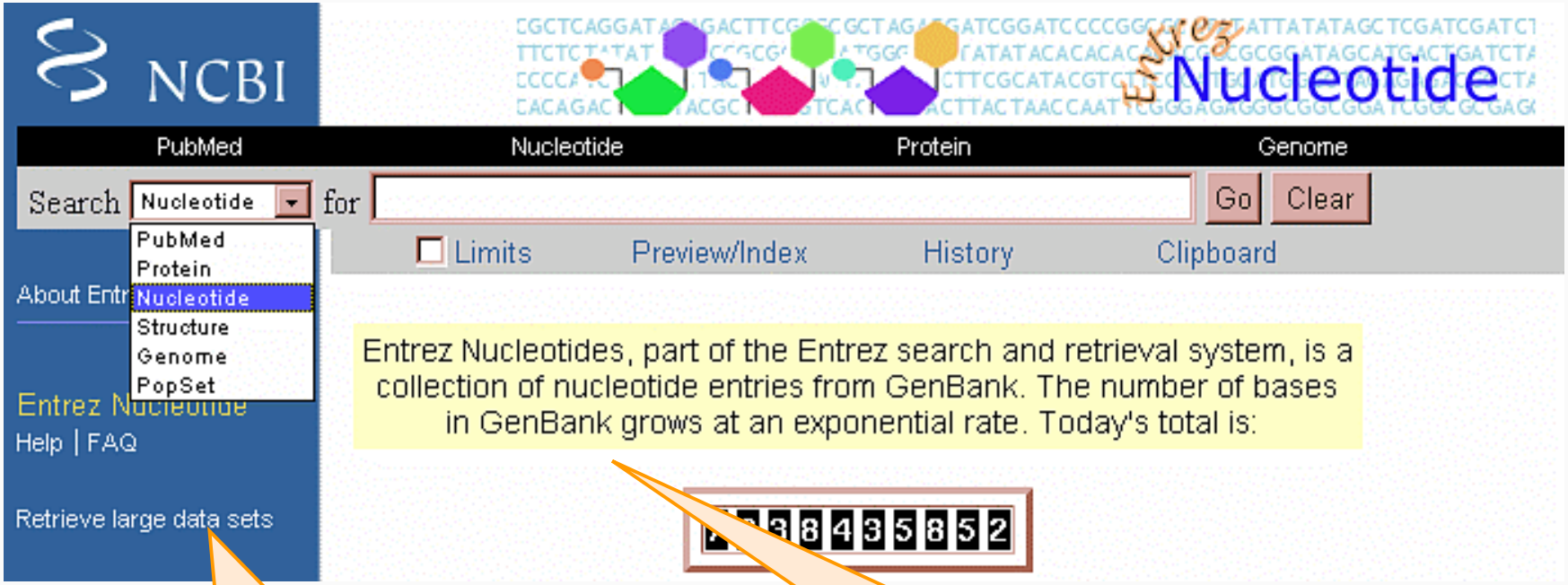
WebSite Infos - The National Center for Genome Resources is deciphering the secret languages of agricultural plenty, a cleaner environment, and improved health -- for everyone. Our staff work in a new field, bioinformatics, that merges biology and computers to apply genetic data to humanity's best possible advantage. NCGR is a nonprofit research organization in Santa Fe, New Mexico, USA.



Basic Rules for Searching DataBases - Search concepts

- Boolean search
  - An advanced query technique used when searching with two or more terms.
    - Terms are combined using the Boolean operators AND, OR, NOT.
    - The default Boolean operator is AND.
- Broadening the search
  - If the results of a search produce no useful entries, change or remove terms that have been entered.
- Narrowing the search
  - If the results of a search produce tao many entries, change or add terms that have been entered.
- Proximity searching
  - To search with multiword terms or phrases, place quotes around the terms.
- Wild card
  - The character * can be prepended or appended to a search term to make a search less specific.
    - For example, to look tor all authors whose last name begins with Zav, search using Zav*.

Searching for or Comparing a Nucleic Acid Sequence

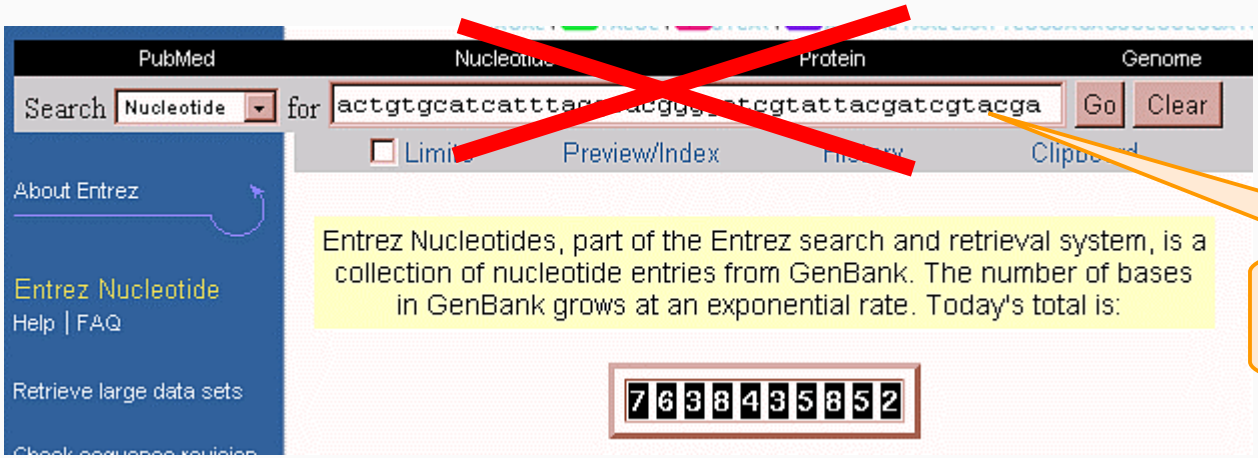


Searching the Nucleic Acid Sequence DataBases via ENTREZ

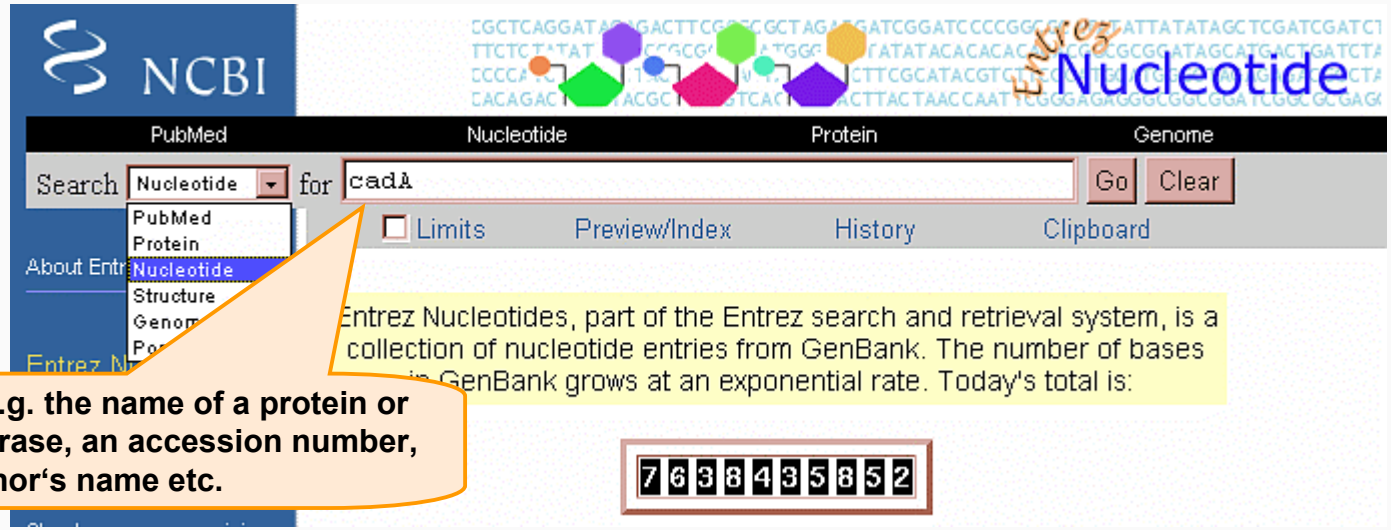
Window for query input, e.g. author's name, accession number, search phrase. See NOTE for more details!

Note

Chose the Correct Input to get the Correct Answer



Notabene:
this window is not a sequence input window!



Input your query, e.g. the name of a protein or a gene, a search phrase, an accession number, an author's name etc.

Text-based searching example

A database search using the ENTREZ Suite and the BLAST set of programs.
The example is of historical importance only since the Drosophila genome has been fully sequenced in 2000!



- Goal
  - ➔ To find the human homolog of the Drosophila *per* gene.
- Procedure
  - ➔ In Entrez, start by searching 'All Fields' in the nucleotide database. Enter the terms human per in the text box. The search finds three hits. In looking at the results, none has anything to do with the human homolog of the Drosophila per gene. They are found because they discuss human topoisomerase and the last name of one of the authors is Per. The search should now be 'broadened' by using the term period rather than per, as period more broadly describes the function of the gene whereas per is the actual gene name in Drosophila. This query results in 24 hits. (Note that 113 hits are obtained if the search is made with the terms period and drosophila.) All 24 entries could be looked at, or the search could be 'narrowed' by adding a third term - drosophila - to human and period, as it is the human homolog of the Drosophila period gene that is being searched for. This results in three hits, all of which are what the search set out to find. The gene in humans is named RIGUI.
  - ➔ These searches can be repeated using SRS and DBGET to see how the search engines differ. SRS found two entries. Using DBGET, the final search produced no hits in the nucleotide database, and the other searches did not find the human homolog, despite the fact that the human RIGUI entry could be retrieved directly using the bget command. The problem is that not all fields of the database are available for searching.

Search for <human per>



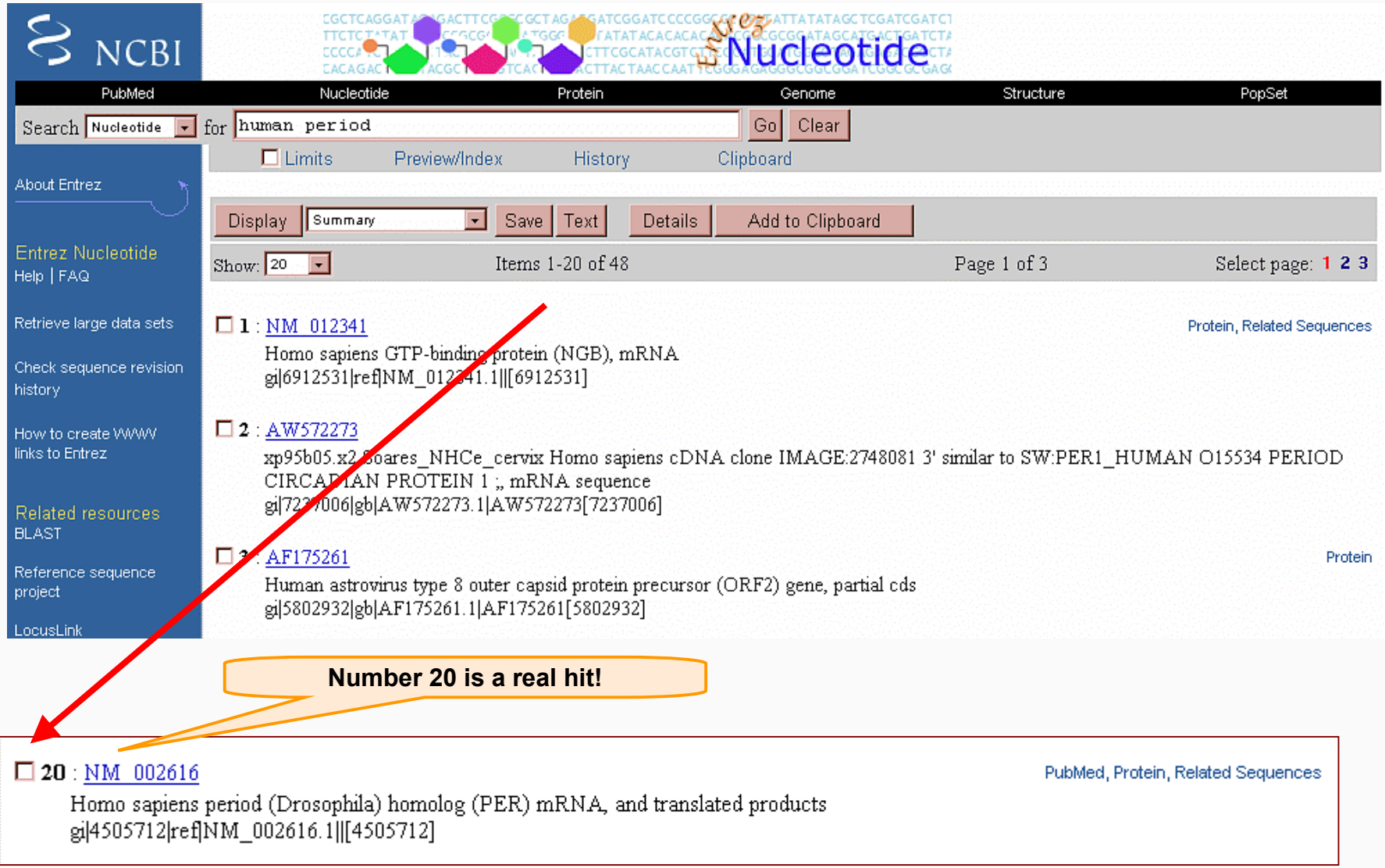
Broadening your Search for <human period>

48 hits for this search
This reflects always the current status of the sequencing efforts.
Progress is measured by the day!

Hence check your databanks often - when working on a <hot> topic - every day at least twice!



Search for <human period> II

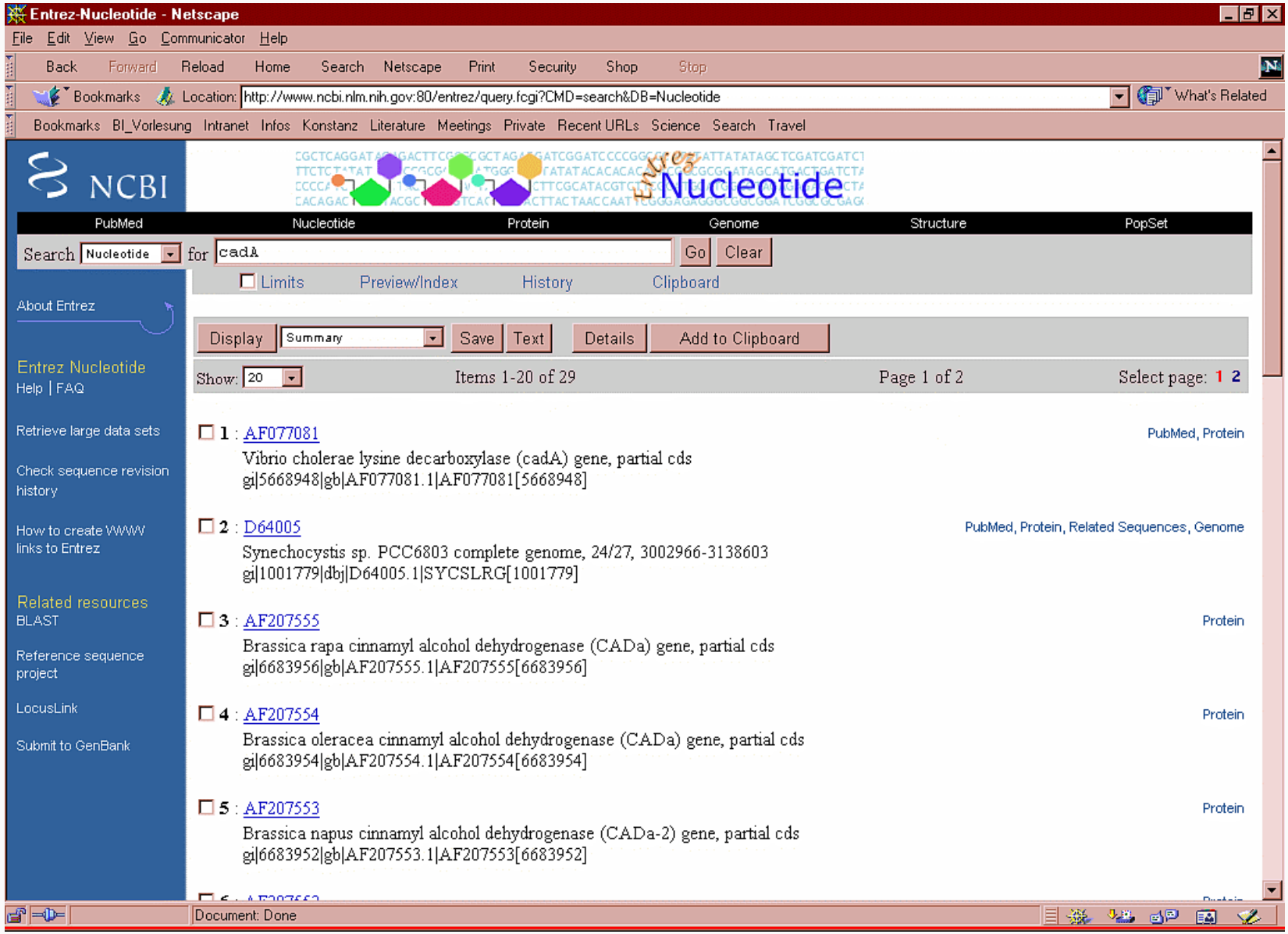




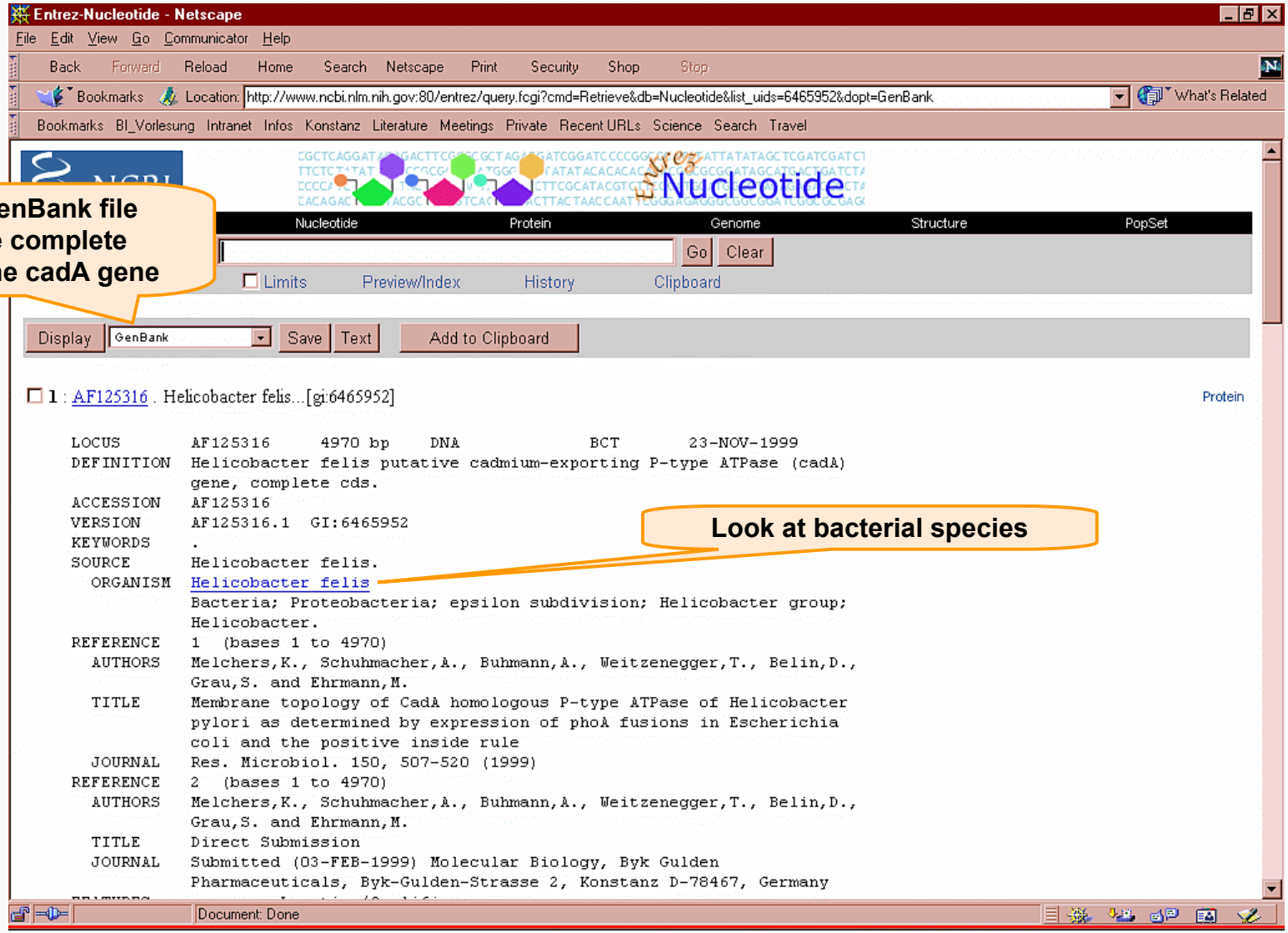
Database Searching

- **Basic Considerations**
  - **Think about every step**
  - **Search a large current database**
  - **Compare as protein rather than DNA**
  - **Filter query for low-complexity regions**
  - **Interpret scores with E-values**
  - **Recognize that most homologs are not found by pairwise sequence comparison**
  - **Consider slower and more powerful methods, but use iterative programs with great caution**

Search for <cadA>



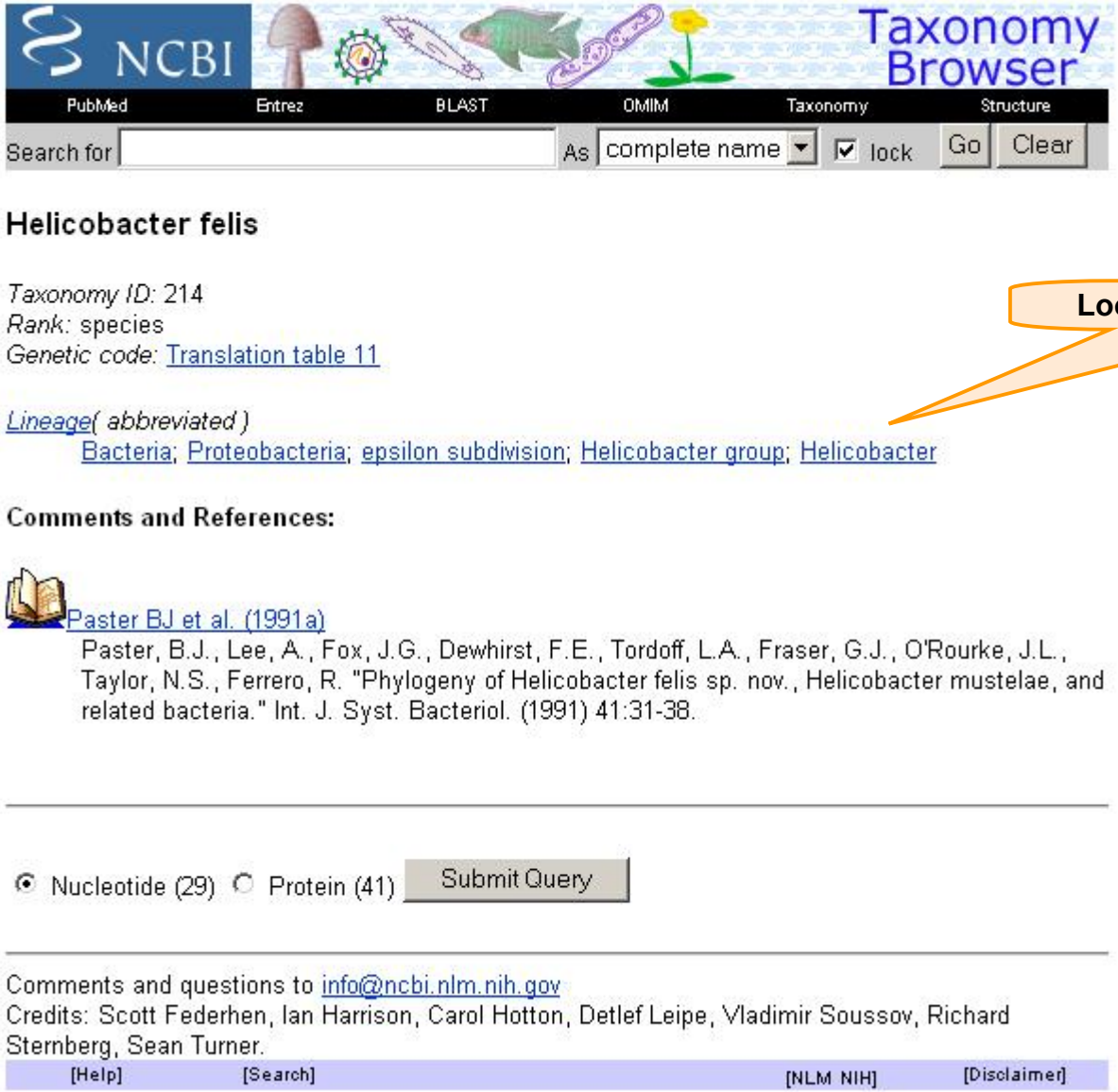
Search for <cadA>: Select a Hit for Closer Inspection



This is the GenBank file showing the complete sequence of the cadA gene

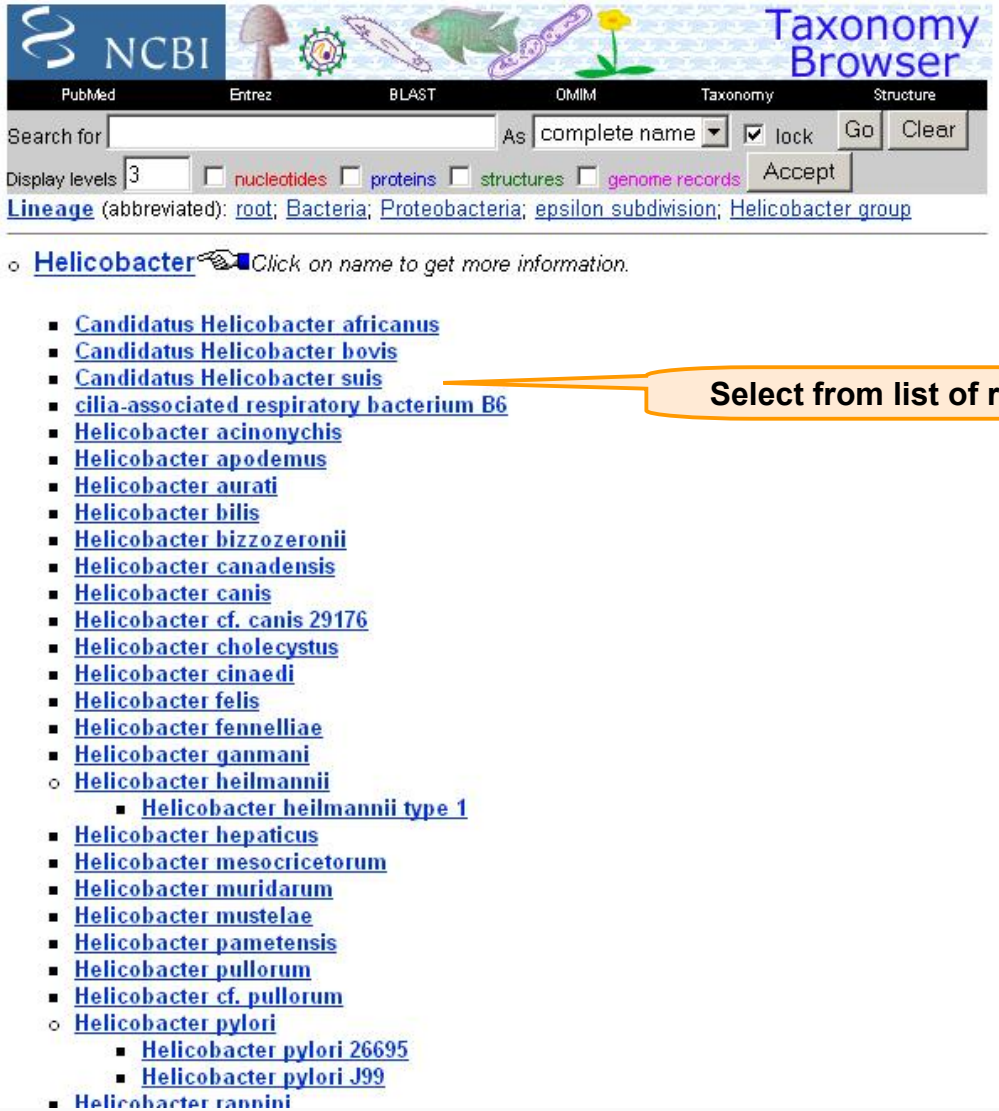
Look at bacterial species

You Want to Kow More About the Bacterial Host Species?



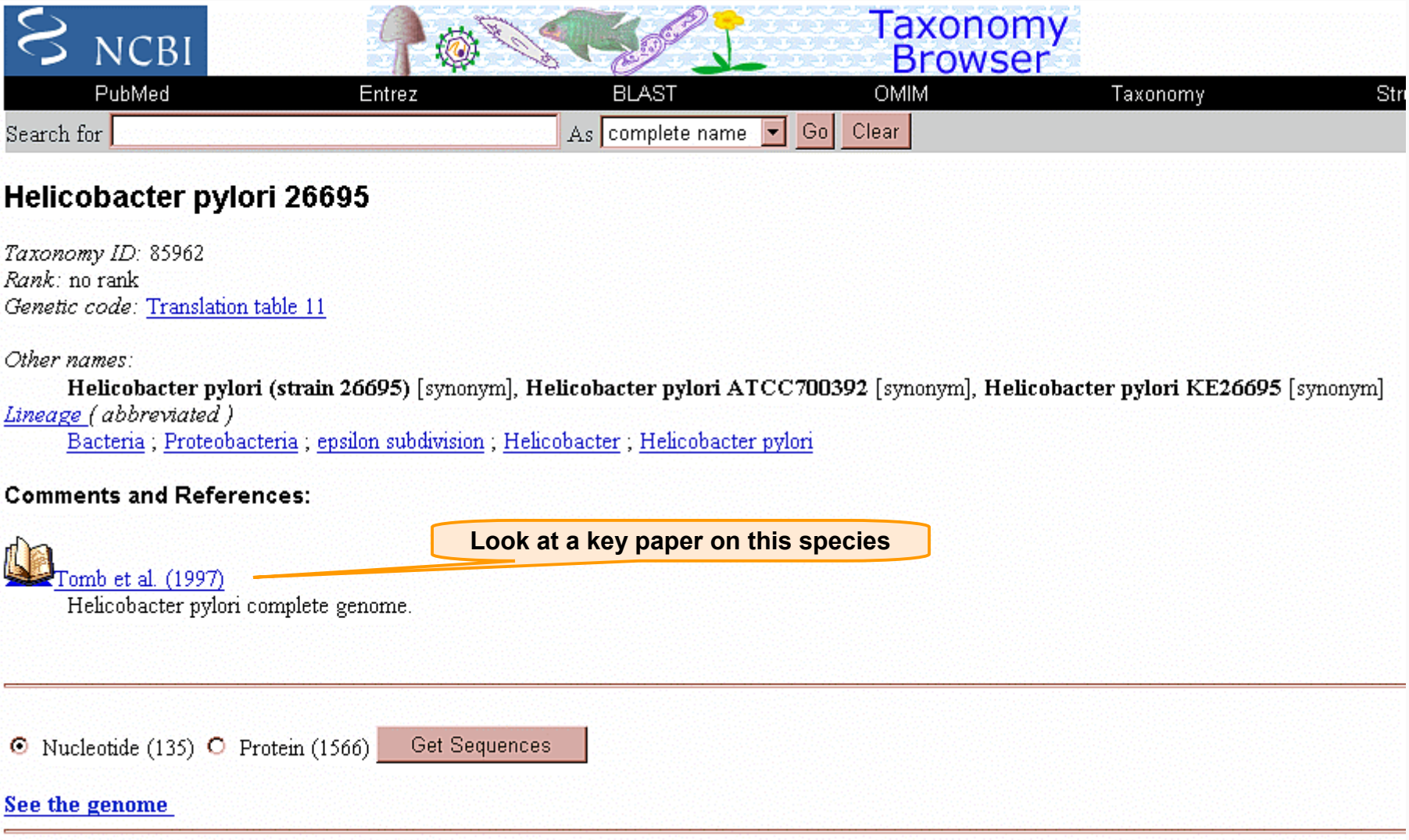
Look at related bacterial species

There May Be An Interesting Relationship Between *Helicobacter pylori* and *Helicobacter felis*

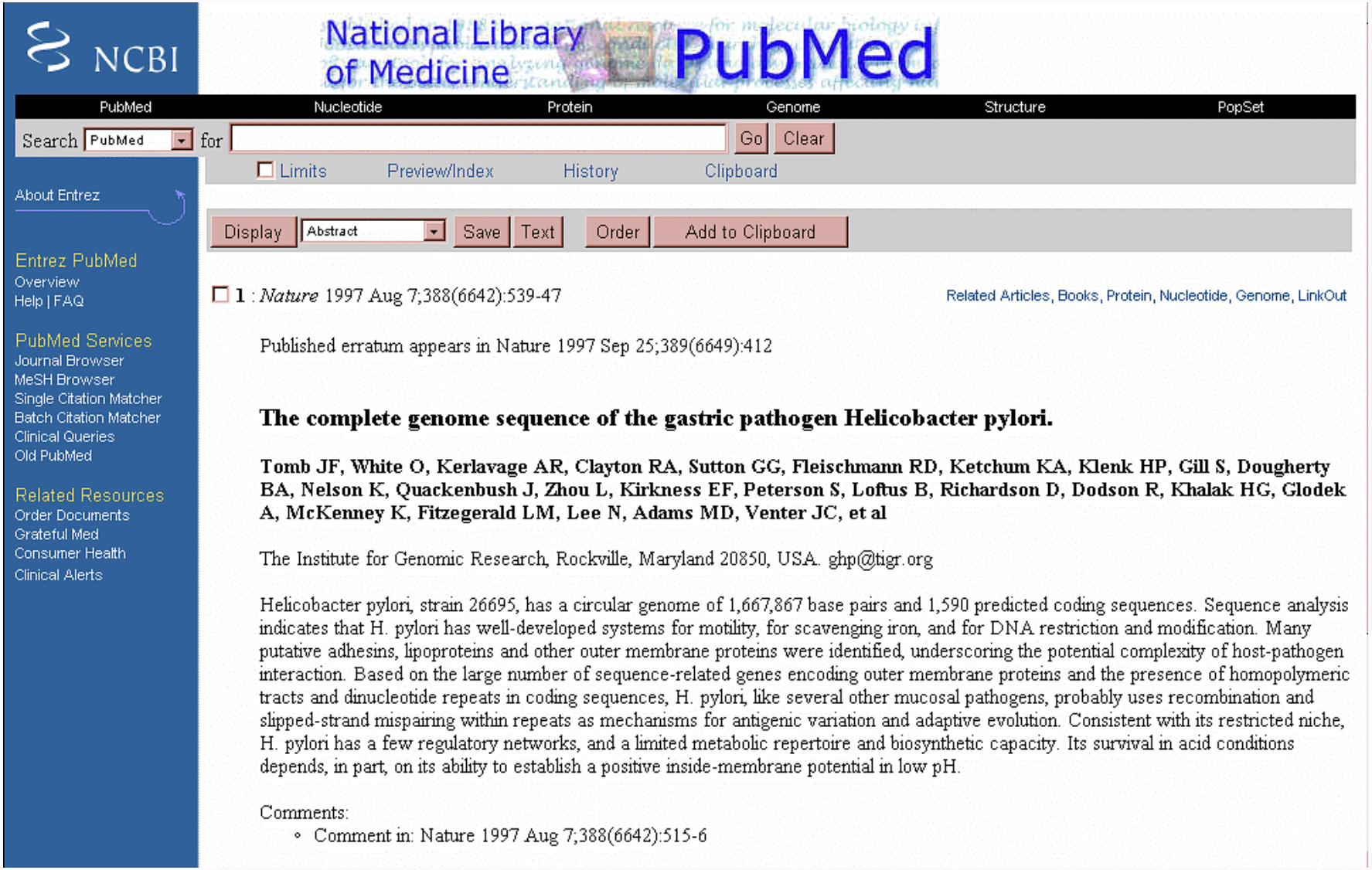


Select from list of related species

You Want to Have a Look at a Key Publication?



Linking Sequences and Published Data



NCBI

National Library of Medicine PubMed

PubMed | Nucleotide | Protein | Genome | Structure | PopSet

Search PubMed for [ ] Go Clear

Limits Preview/Index History Clipboard

About Entrez

Display Abstract Save Text Order Add to Clipboard

Entrez PubMed
Overview
Help | FAQ

PubMed Services
Journal Browser
MeSH Browser
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Old PubMed

Related Resources
Order Documents
Grateful Med
Consumer Health
Clinical Alerts

1 : *Nature* 1997 Aug 7;388(6642):539-47 Related Articles, Books, Protein, Nucleotide, Genome, LinkOut

Published erratum appears in Nature 1997 Sep 25;389(6649):412

The complete genome sequence of the gastric pathogen Helicobacter pylori.

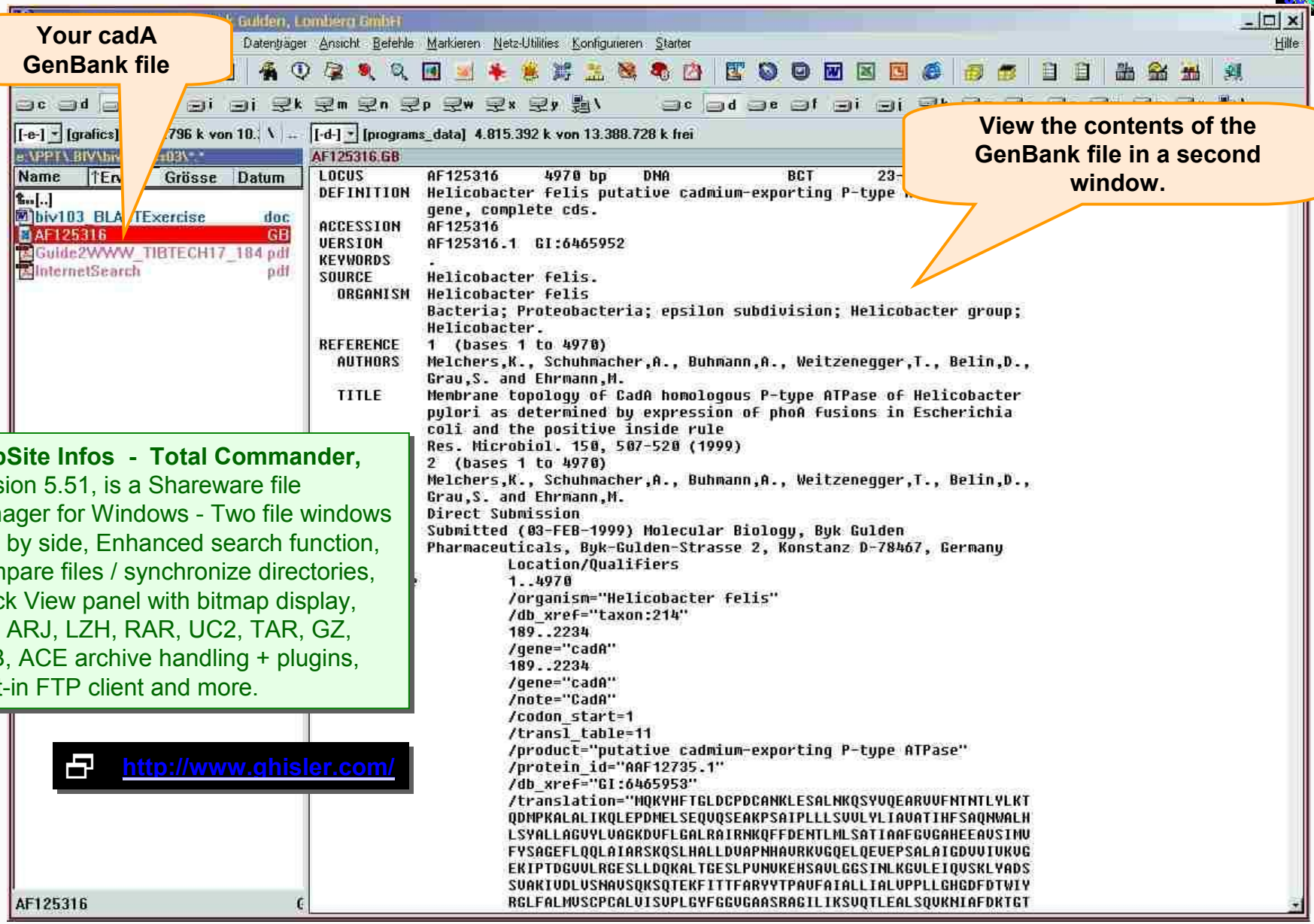
Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Venter JC, et al

The Institute for Genomic Research, Rockville, Maryland 20850, USA. ghp@tigr.org

Helicobacter pylori, strain 26695, has a circular genome of 1,667,867 base pairs and 1,590 predicted coding sequences. Sequence analysis indicates that H. pylori has well-developed systems for motility, for scavenging iron, and for DNA restriction and modification. Many putative adhesins, lipoproteins and other outer membrane proteins were identified, underscoring the potential complexity of host-pathogen interaction. Based on the large number of sequence-related genes encoding outer membrane proteins and the presence of homopolymeric tracts and dinucleotide repeats in coding sequences, H. pylori, like several other mucosal pathogens, probably uses recombination and slipped-strand mispairing within repeats as mechanisms for antigenic variation and adaptive evolution. Consistent with its restricted niche, H. pylori has a few regulatory networks, and a limited metabolic repertoire and biosynthetic capacity. Its survival in acid conditions depends, in part, on its ability to establish a positive inside-membrane potential in low pH.

Comments:
- Comment in: Nature 1997 Aug 7;388(6642):515-6

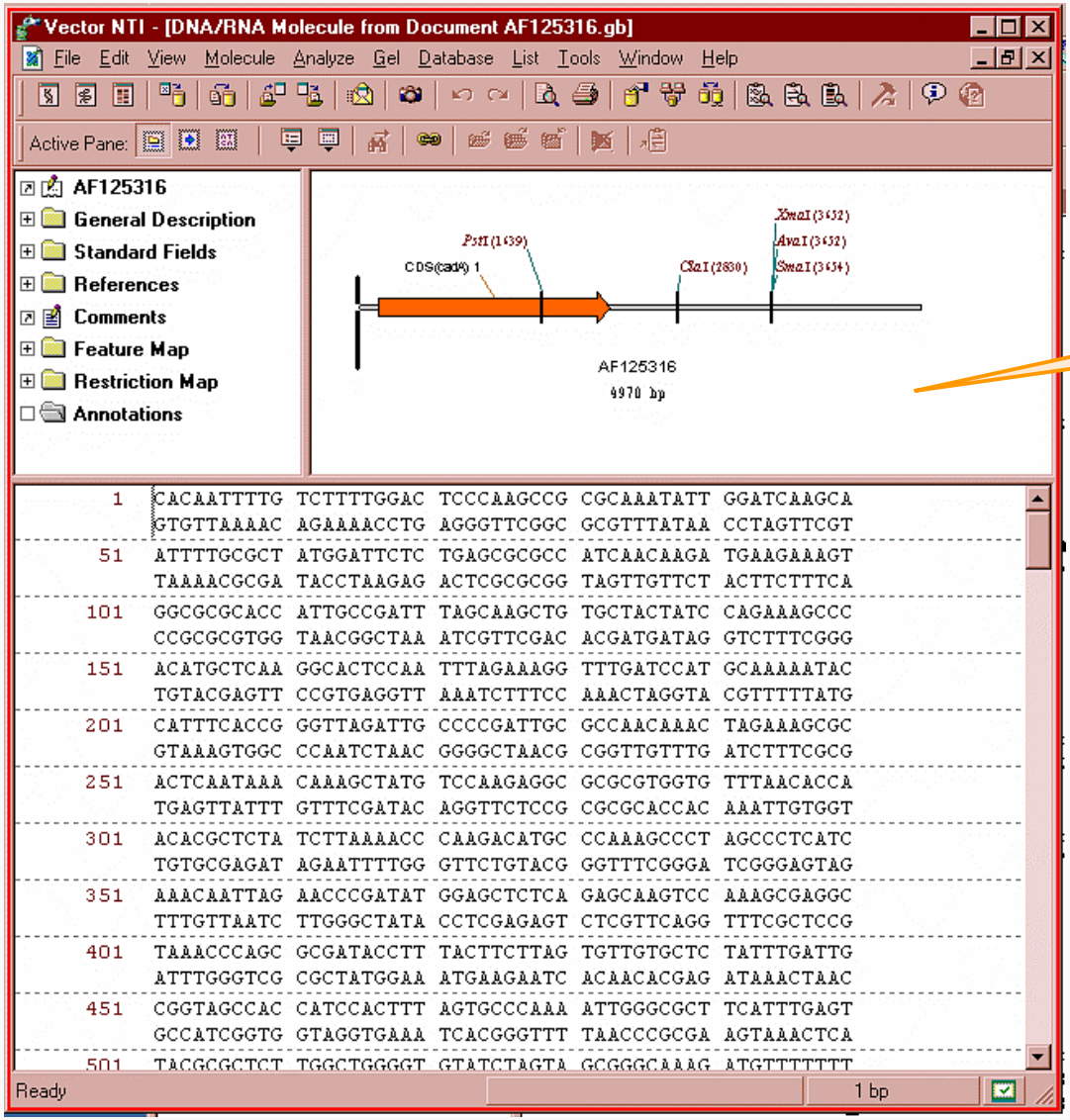
Visualize Your Sequences and Data



WebSite Infos - Total Commander, Version 5.51, is a Shareware file manager for Windows - Two file windows side by side, Enhanced search function, Compare files / synchronize directories, Quick View panel with bitmap display, ZIP, ARJ, LZH, RAR, UC2, TAR, GZ, CAB, ACE archive handling + plugins, Built-in FTP client and more.

http://www.ghisler.com/

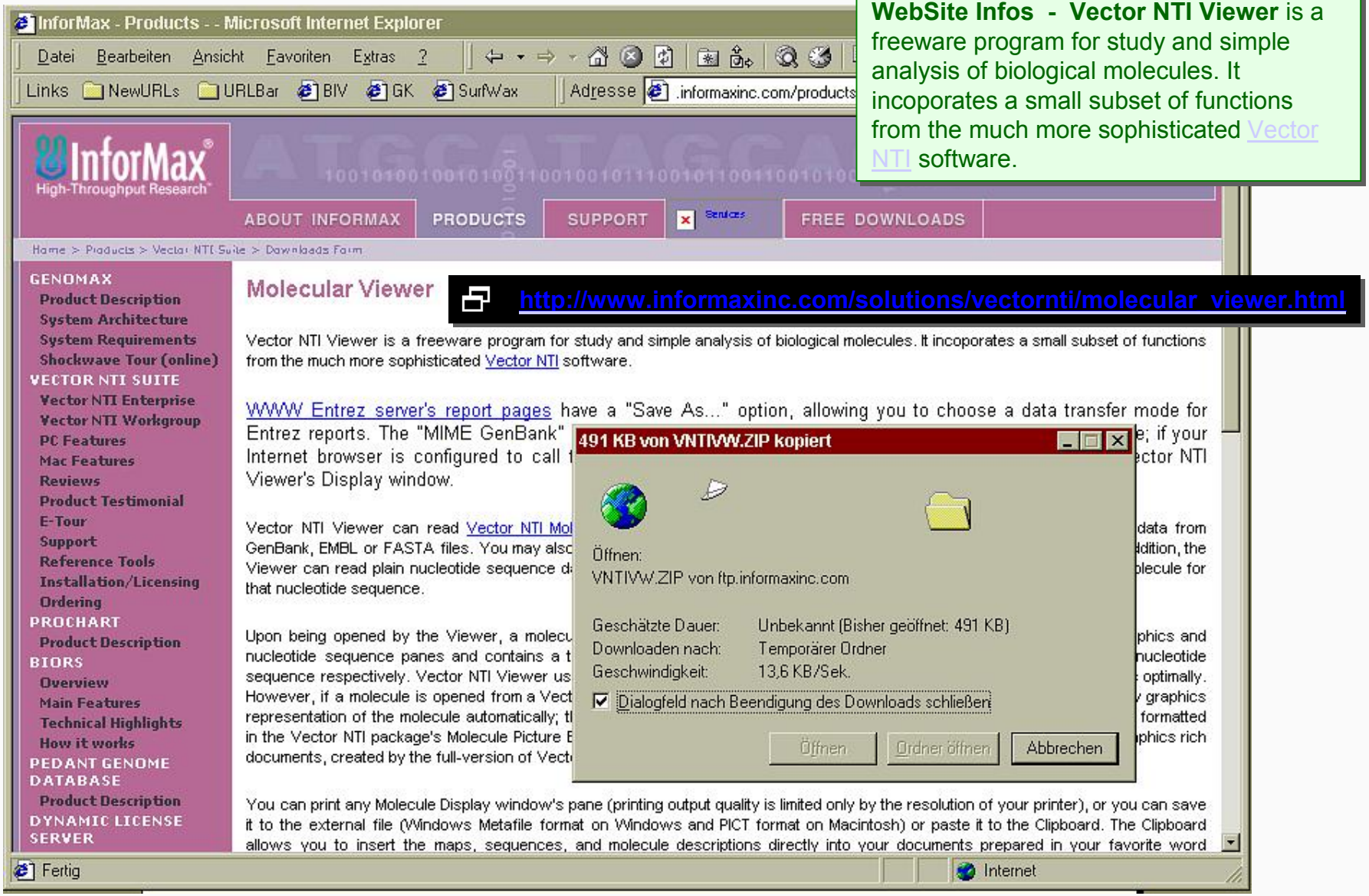
The Vector NTI Viewer Shows the Organization of the Locus



Graphical view of the locus

The Vector NTI Viewer is FreeWare

WebSite Infos - Vector NTI Viewer is a freeware program for study and simple analysis of biological molecules. It incoporates a small subset of functions from the much more sophisticated Vector NTI software.

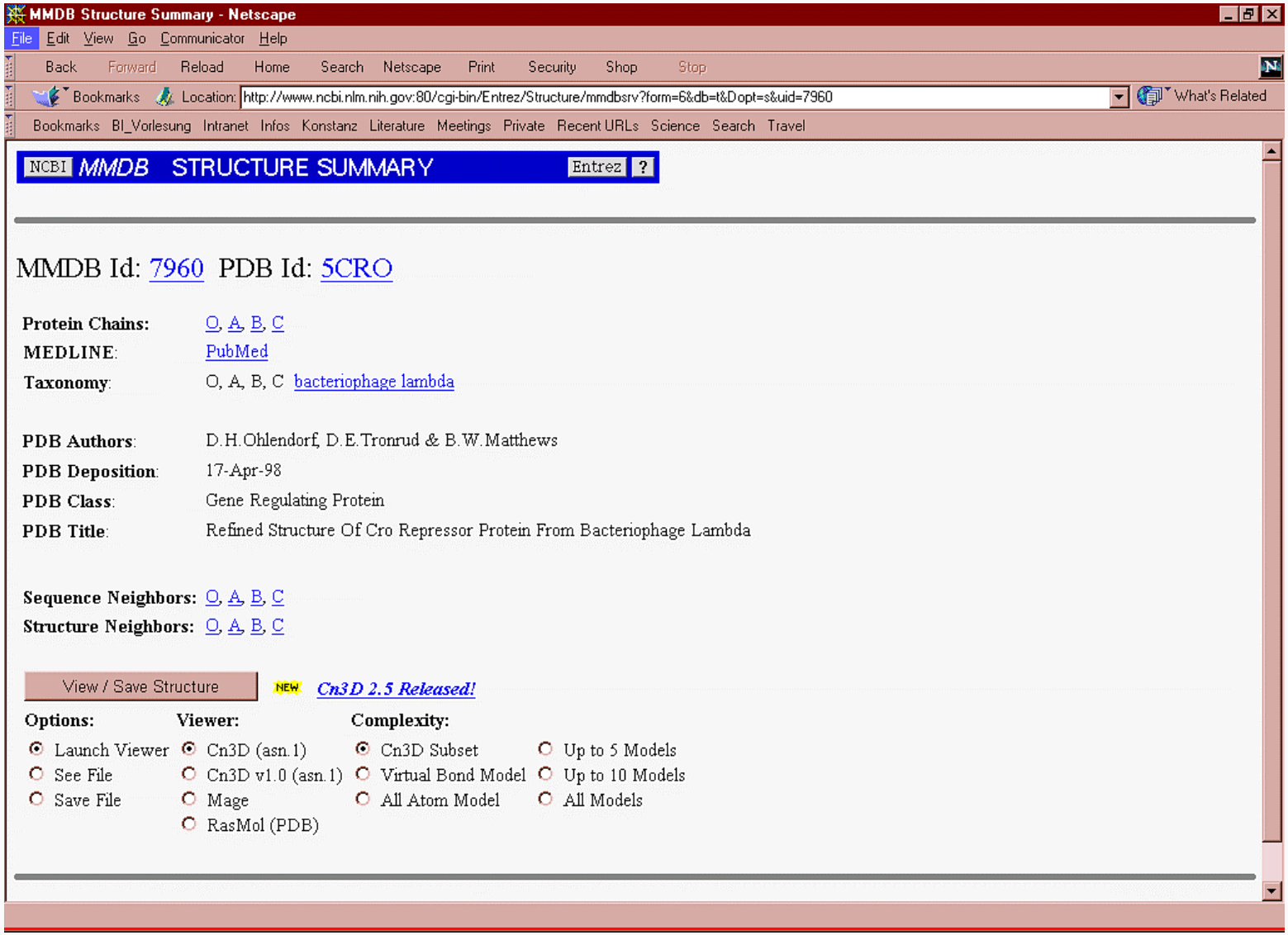


Looking at Protein Sequences

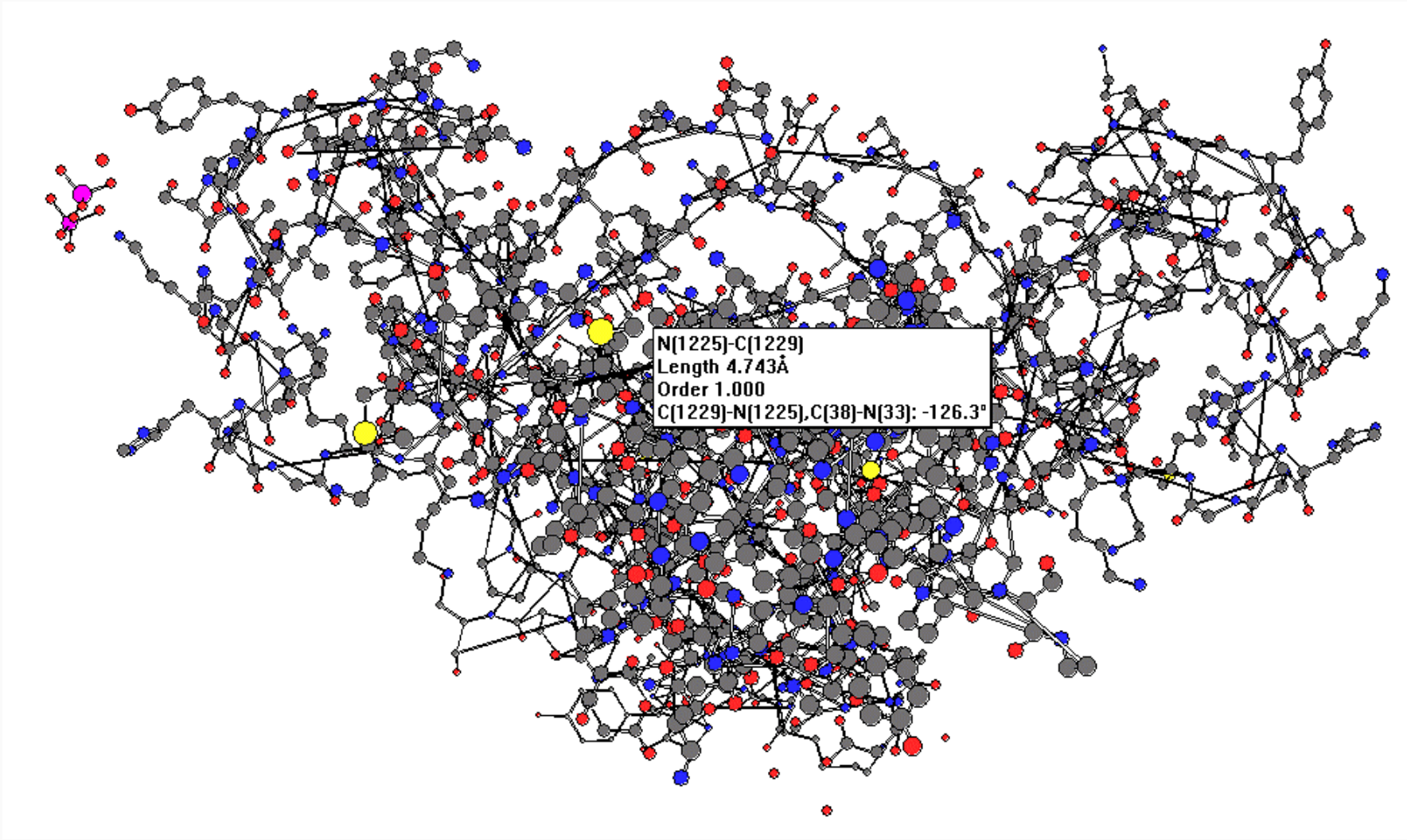
```
FEATURES             Location/Qualifiers
     source          1..4970
                     /organism="Helicobacter felis"
                     /db_xref="taxon:214"
     gene            189..2234
                     /gene="cadA"
     CDS             189..2234
                     /gene="cadA"
                     /note="CadA"
                     /codon_start=1
                     /transl_table=11
                     /product="putative cadmium-exporting P-type ATPase"
                     /protein_id="AAF12735.1"
                     /db_xref="GI:6465953"
                     /translation="MQKYHFTGLDCPDCANKLESALNKQSYVQEARVVFNTNTLYLKT
                     QDMPKALALIKQLEPDMELSEQVQSEAKPSAIPLLLSVVLYLIAVATIHFSAQNWALH
                     LSYALLAGVYLVAGKDVFLGALRAIRNKQFFDENTLMLSATIAAFGVGAHEEAVSIMV
                     FYSAGEFLQQLAIARSKQSLHALLDVAPNHAVRKVGQELQEVEPSALAIGDVVIVKVG
                     EKIPTDGVVLRGESLLDQKALTGESLPVNVKEHSAVLGGSINLKGVLEIQVSKLYADS
                     SVAKIVDLVSNAVSQKSQTEKFITTFARYYTPAVFAIALLIALVPPLLGHGDFDTWIY
                     RGLFALMVSCPCALVISVPLGYFGGVGAASRAGILIKSVQTLEALSQVKNIAFDKTGT
                     LSKGEFNVIDVVPIAPFSKEDVLQHATCAQILSTHPIAISIQKAYKRQCQHEVKDYQE
                     IGGLGVQASCHSHLIIAGNDKMLHKYNIPHDTCSLEGTIVHVAVDGKHIGYIVVADTL
                     KDNAKECLDGLKHAGIEHMCILSGDHEYSTKRVAKELDCPYYANLLPEDKLNAFKDFQ
                     AQHAHKSMFVGDGINDAPTLARADVSMSMGSASQISKESADIVITNNSLESVLKVFKI
                     AKKTKRIIIENIIFALAIKAMFIVLGLSGDASLWEAVLGDVGVTLIALANSMRTMRI"
BASE COUNT     1351 a   1200 c   1241 g   1178 t
ORIGIN
        1 cacaattttg tcttttggac tcccaagccg cgcaaatatt ggatcaagca attttgcgct
       61 atggattctc tgagcgcgcc atcaacaaga tgaagaaagt ggcgcgcacc attgccgatt
      121 tagcaagctg tgctactatc cagaaagccc acatgctcaa ggcactccaa tttagaaagg
      181 tttgatccat gcaaaaatac catttcaccg ggttagattg ccccgattgc gccaacaaac
      241 tagaaagcgc actcaataaa caaagctatg tccaagaggc gcgcgtggtg tttaacacca
      301 acacgctcta tcttaaaacc caagacatgc ccaaagccct agccctcatc aaacaattag
      361 aacccgatat ggagctctca gagcaagtcc aaagcgaggc taaacccagc gcgatacctt
      421 tacttcttag tgttgtgctc tatttgattg cggtagccac catccacttt agtgcccaaa
```

The GenBank file contains the translated reading frames

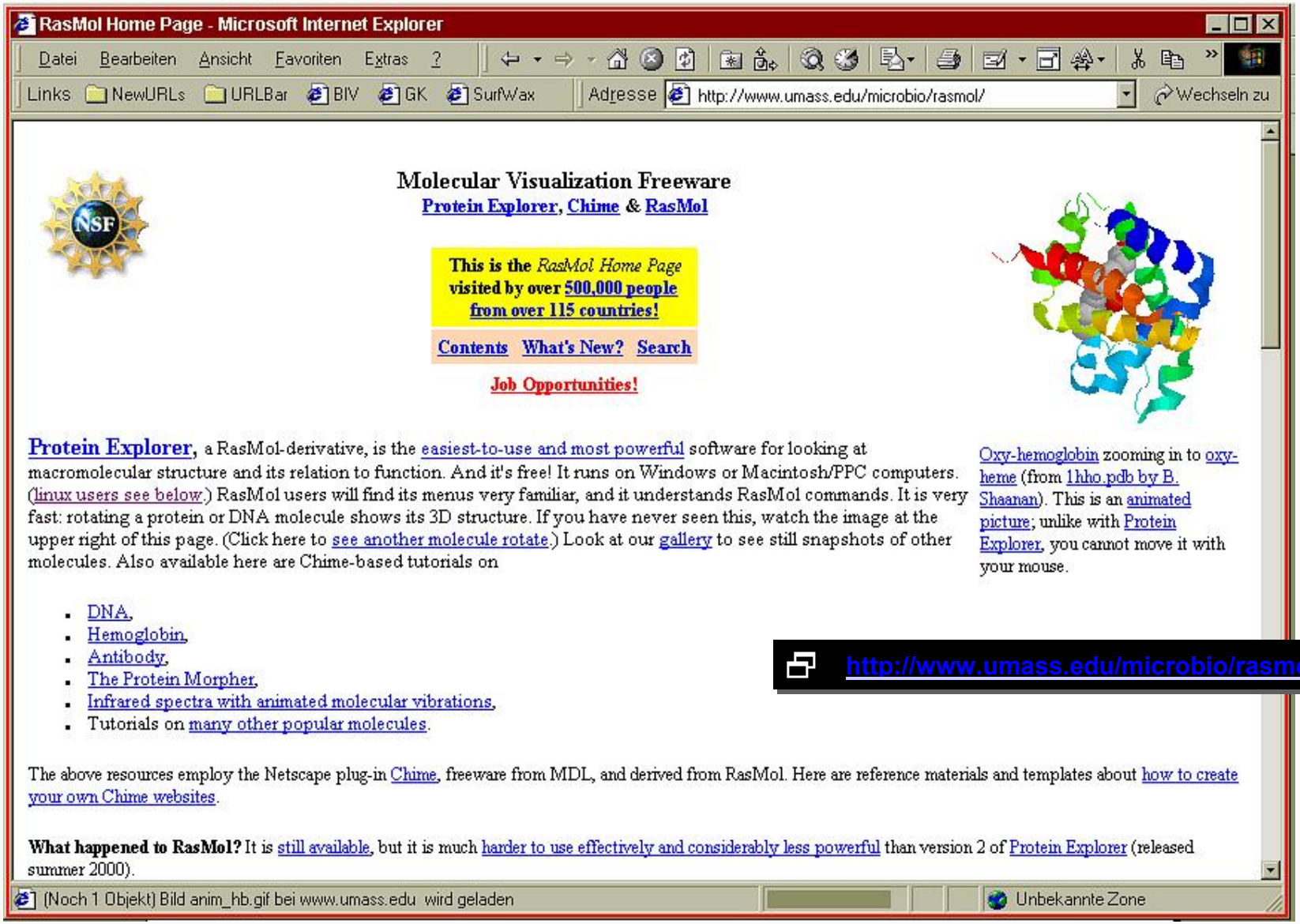
If the X-ray Structure of a Protein is Known ...



Viewing a Structure with RasMOL



One of Many RasMol Download Sites



The TIGR Site

The Institute for Genomic Research (TIGR) is a not-for-profit research institute whose primary research interests are in structural, functional and comparative analysis of genomes and gene products from a wide variety of organisms including viruses, eubacteria (both pathogens and non-pathogens), archaea (the so-called third domain of life), and eukaryotes (plants, animals, fungi and protists such as the malarial parasite).

http://www.tigr.org/

Besides listing sequences TIGR strives to become an academic Center of Excellence by offereing high ranking scientific courses and meetings

