Introduction to Bioinformatics: Part 1
NCBI (Entrez)

Goal: The efficient use of online databases for genetic data retrieval.

Bioinformatics: is the study of biological problems through the coordination of techniques from mathematics, statistics, computer science and information technology.

NCBI (National Center for Biotechnology Information): a nationally funded facility that creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biological information.

Entrez: a data retrieval system developed by NCBI that provides integrated access to a wide range of data domains, including literature, nucleotide and protein sequences, complete genomes, three-dimensional structures, and more.

Introduction: The effective and powerful use of Entrez requires an understanding of the available data domains, the variety of data sources and types within each domain, and Entrez's advanced search features. This tutorial uses the human MLH1 gene, implicated in colon cancer, to demonstrate the wide variety of information that we can rapidly gather for a single gene[1].

Sub-Goals: The search goals are to:

- separate the wheat from the chaff – identify a representative, well annotated mRNA sequence record;

- retrieve associated literature and protein records;

- identify conserved domains within the protein;

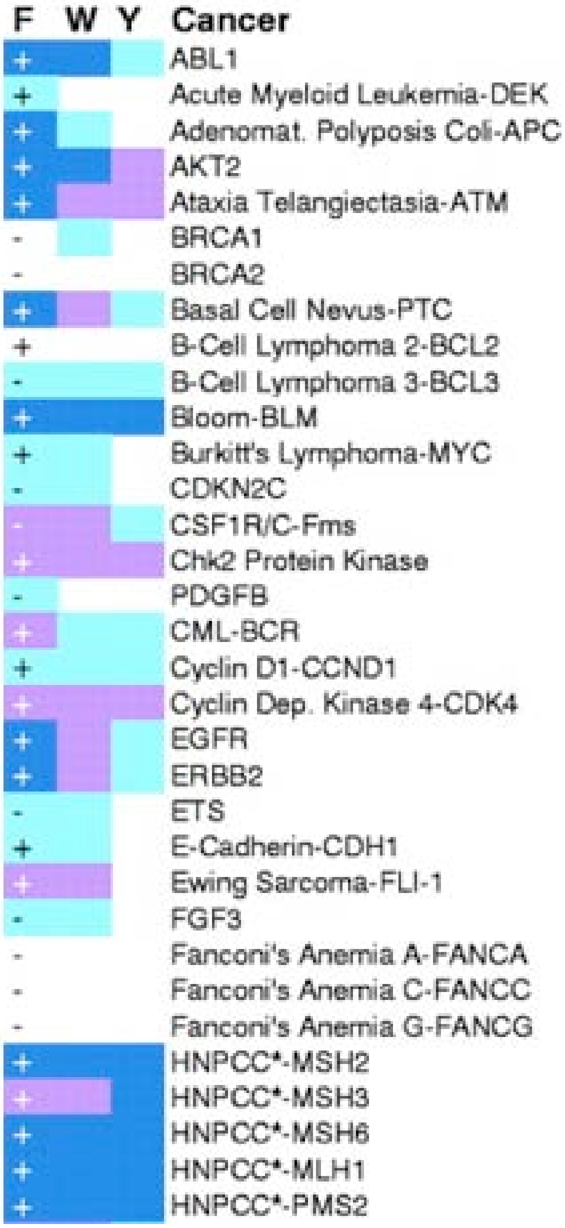
- identify similar proteins;

- identify known mutations within the gene or protein;

- find a resolved three-dimensional structure for the protein or, in its absence, identify structures with homologous sequence;

- view genomic context and download the sequence region.

[1] Geer, R.C. and Sayers, E.W. Entrez: Making use of its power. *Briefings in Bioinformatics*. 2003 June;4(2):1779-184..



F	W	Y	Cancer
+ 🟦	🟦	▨	ABL1
+ ▨			Acute Myeloid Leukemia-DEK
+ 🟦	▨		Adenomat. Polyposis Coli-APC
+ 🟦	🟦	🟪	AKT2
+ 🟦	🟪	🟪	Ataxia Telangiectasia-ATM
-	▨		BRCA1
-			BRCA2
+ 🟦	🟪	▨	Basal Cell Nevus-PTC
+			B-Cell Lymphoma 2-BCL2
- ▨	▨	▨	B-Cell Lymphoma 3-BCL3
+ 🟦	🟦	🟦	Bloom-BLM
+ ▨	▨		Burkitt's Lymphoma-MYC
- ▨	▨		CDKN2C
- 🟪	🟪	▨	CSF1R/C-Fms
+ 🟪	🟪	🟪	Chk2 Protein Kinase
- ▨			PDGFB
+ 🟪	▨	▨	CML-BCR
+ ▨	▨	▨	Cyclin D1-CCND1
+ 🟪	🟪	🟪	Cyclin Dep. Kinase 4-CDK4
+ 🟦	🟪	▨	EGFR
+ 🟦	🟪	▨	ERBB2
- ▨	▨		ETS
+ ▨	▨		E-Cadherin-CDH1
+ 🟪	🟪		Ewing Sarcoma-FLI-1
- ▨	▨		FGF3
-			Fanconi's Anemia A-FANCA
-			Fanconi's Anemia C-FANCC
-			Fanconi's Anemia G-FANCG
+ 🟦	🟦	🟦	HNPCC*-MSH2
+ 🟪	🟪	🟦	HNPCC*-MSH3
+ 🟦	🟦	🟦	HNPCC*-MSH6
+ 🟦	🟦	🟦	HNPCC*-MLH1
+ 🟦	🟦	🟦	HNPCC*-PMS2

322 STUDENTS: ***This exercise will introduce you to the databases that you will need for your FWYH project. We will explore MLH1 (second from bottom of list) as an example. NOTE that in this exericise we will start our ENTREZ search with the general term "colon cancer" and then narrow our hits from there. When you do equivalent searches for information on your gene, you will obviously start with a much narrower search term.***

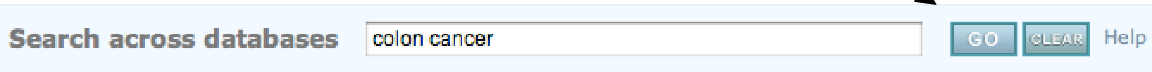
START by looking up gene function on OMIM
http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM

After looking at gene function at OMIM:

1. go to http://www.ncbi.nlm.nih.gov/
(Hint: Google NCBI).

2. Select -Entrez Home-, from the “Hot Spots” (right column).

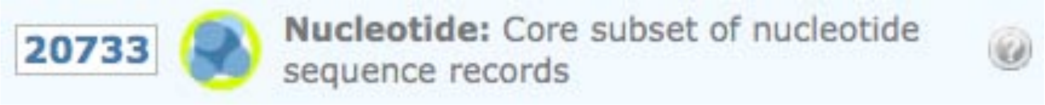
3. Enter “Colon Cancer” in the search field and select GO



Search across databases colon cancer GO CLEAR Help

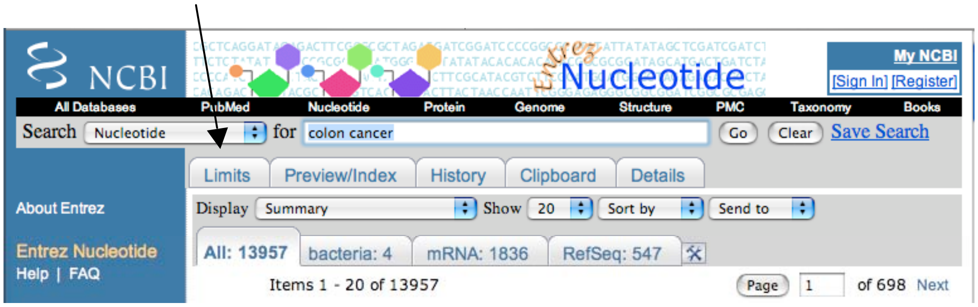
4. You should note that there are over 80,000 PubMed entries, 21,000 Nucleotide entries, and over 1000 Protein entries. Take a moment to familiarize yourself to the databases assayed by Entrez. We will learn how to narrow the search to

5. Click on the “Nuleotide” field



20733 Nucleotide: Core subset of nucleotide sequence records

6. Select the Limits tab to narrow the search.



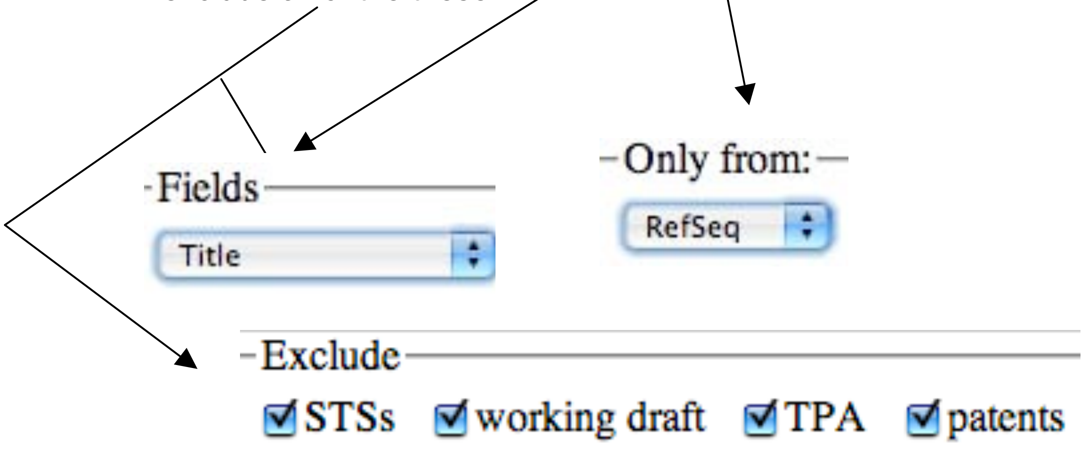
Search Nucleotide for colon cancer

Display Summary Show 20 Sort by Send to

All: 13957 bacteria: 4 mRNA: 1836 RefSeq: 547

Items 1 - 20 of 13957 Page 1 of 698

7. Limit the search to the Title field, to RefSeq (a curated database), and exclude all of the these…



Fields Title

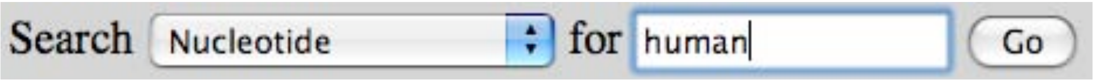
Only from: RefSeq

Exclude

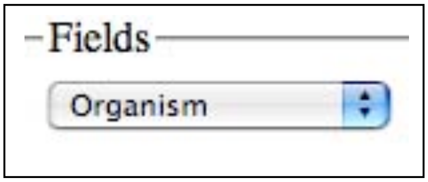
☑ STSs ☑ working draft ☑ TPA ☑ patents

…hit Go.

8. This reduces the number of hits to ~138, still too many unless you are really motivated). To further limit the search, return to the Limits page, search for human after selecting “Organism” for the search field.



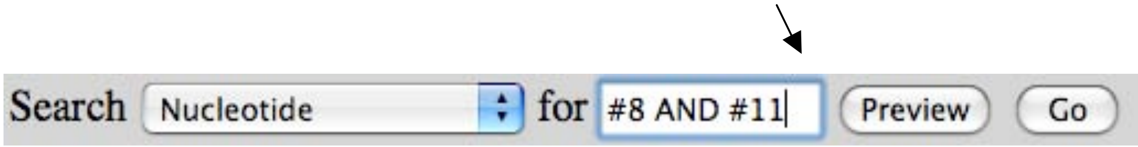
Search Nucleotide for human



Fields

Organism

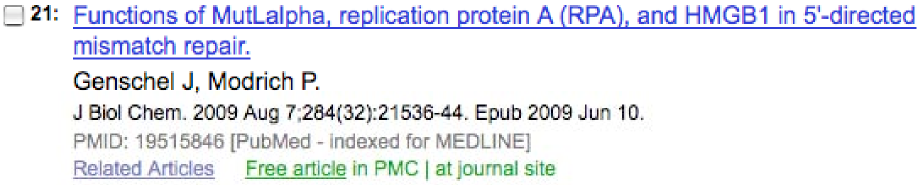
9. Select the “History” tab, and combine the title and organism searches (i.e. #2 and #3) with the Boolean operator AND (must be capitalized).



Search Nucleotide for #8 AND #11

This limits the search to ~ 15 entries. We will pursue **NM_000249**

10. The Links drag down menu (right column) for **NM_000249** provides a list of other Entrez domains. For example, the PubMed entries include curated journal articles entered into RefSeq, i.e. expert selected. Take a moment to look over a PubMed entry, or two. Note, some articles re available free online.



□ 21: Functions of MutLalpha, replication protein A (RPA), and HMGB1 in 5'-directed mismatch repair.

Genschel J, Modrich P.

J Biol Chem. 2009 Aug 7;284(32):21536-44. Epub 2009 Jun 10.

PMID: 19515846 [PubMed - indexed for MEDLINE]

Related Articles Free article in PMC | at journal site

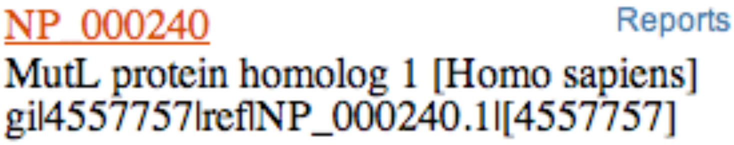
11. Examine and understand the following Links domains:
Homologene: homologous genes in other species.
Reading Assignment: you are responsible for the Wikipedia (www.wikipedia.org) entry for Homology.

OMIM

SNP: **Reading Assignment**: Single Nucleotide Polymorphism (Wikipedia).

Mapviewer: Be sure you can identify the loci that flank MLH1, and know which chromosome codes for MLH1. Hint: zoom in.

12. From the Links menu, choose the Protein domain, then select the NP_000240 link



MutL protein homolog 1 [Homo sapiens]
gi|4557757|ref|NP_000240.1|[4557757]

Note: # of amino acids. Note: at the bottom of the page is the AA sequence.

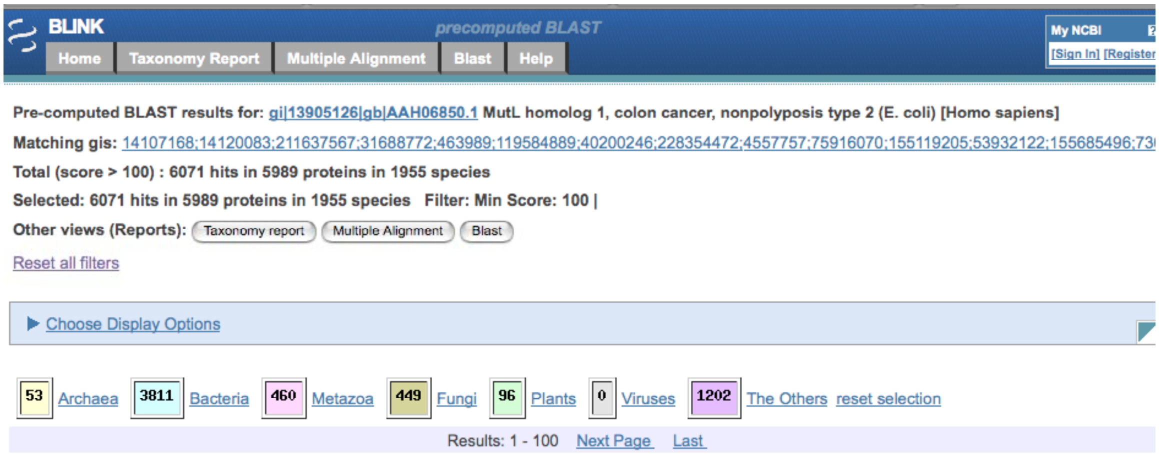
13. IF TIME (ask Instructor) Break Time / Self Paced Tutorial

Important: Keep the MLH1 page open, **and**

open a new browser window and work through BLAST Tutorial
http://www.digitalworldbiology.com/dwb/Tutorials/Entries/2009/1/26_BLAST_for_Beginners.html

Note: additional BLAST information available at Wikipedia.

13. Return to the NP_000240 window, and hit the BLink entry (right side of the page). This is a curated BLAST result. Look at the Multiple Alignment.
14. **Browse thorugh list of sequences and note organisms.**



Pre-computed BLAST results for: gi|13905126|gb|AAH06850.1 MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [Homo sapiens]

Matching gis: 14107168;14120083;211637567;31688772;463989;119584889;40200246;228354472;4557757;75916070;155119205;53932122;155685496;73

Total (score > 100) : 6071 hits in 5989 proteins in 1955 species

Selected: 6071 hits in 5989 proteins in 1955 species Filter: Min Score: 100 |

Other views (Reports): Taxonomy report Multiple Alignment Blast

Reset all filters

▶ Choose Display Options

53 Archaea 3811 Bacteria 460 Metazoa 449 Fungi 96 Plants 0 Viruses 1202 The Others reset selection

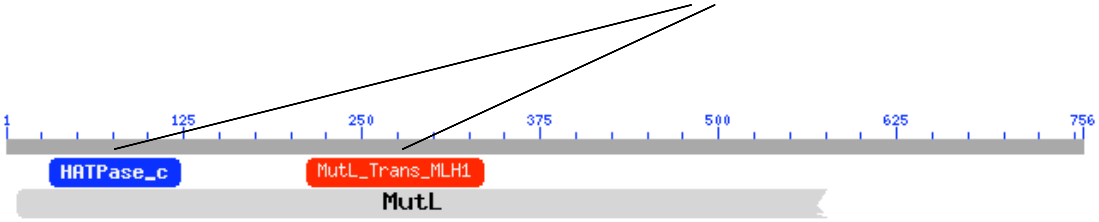
Results: 1 - 100 Next Page Last

USE THE COLORED BOXES TO EXCLUDE OR INCLUDE Categories of organisms

15. ***Under display options choose BEST HITS and see how this affects the numbers of sequences listed.***

16.

16 Return to the NP_000240 window, and hit the Conserved Domains entry (right side of the page). Conserved Domains are comprised of protein sequences that code for a common functional unit, i.e. the active site of a protein, or a transmembrane domain, etc. Click on the domains…



…to see their biological function. Structures can also be observed, when known, from this page. However, a program must be downloaded from NCBI to facilitate viewing.

17. Back up all the way to the NG_008418 Nucleotide page, and select the Reports. Look at the FASTA page, and the Graphic page.

18. This has been a cursory introduction to Bioinformatics. These resources can be drawn on throughout for this course, and may be valuable during your subsequent career as a biologist.