Introduction to Bioinformatics

CPSC 265

Thanks to Jonathan Pevsner, Ph.D.

Textbooks

Johnathan Pevsner, who I stole most of these slides from (thanks!) has written a textbook, *Bioinformatics and Functional Genomics* (Wiley, 2003). The chapters contain content, lab exercises, and quizzes that were developed in his course over the past six years.

All Pevsner's powerpoints are available at:
http://pevsnerlab.kennedykrieger.org

Several other bioinformatics texts are available:
- Baxevanis and Ouellette
- David Mount
- Durbin et al.

What is bioinformatics?

- Interface of biology and computers
- Analysis of proteins, genes and genomes using computer algorithms and computer databases
- Genomics is the analysis of genomes. The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.

Top ten challenges for bioinformatics

[1] Precise models of where and when transcription will occur in a genome (initiation and termination)

[2] Precise, predictive models of alternative RNA splicing

[3] Precise models of signal transduction pathways; ability to predict cellular responses to external stimuli

[4] Determining protein:DNA, protein:RNA, protein:protein recognition codes

[5] Accurate *ab initio* protein structure prediction

Top ten challenges for bioinformatics

[6] Rational design of small molecule inhibitors of proteins

[7] Mechanistic understanding of protein evolution

[8] Mechanistic understanding of speciation

[9] Development of effective gene ontologies: systematic ways to describe gene and protein function

[10] Education: development of bioinformatics curricula

Source: Ewan Birney, Chris Burge, Jim Fickett

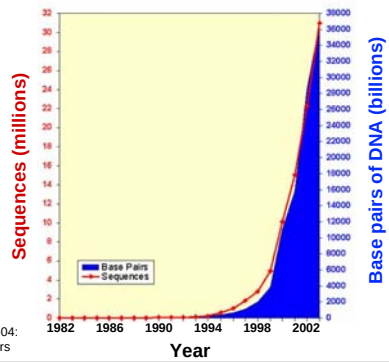


Fig. 2.1
Page 17

There are three major public DNA databases

EMBL ⟷ **GenBank** ⟷ **DDBJ**

Housed at EBI	**Housed at NCBI**	**Housed in Japan**
European Bioinformatics Institute	National Center for Biotechnology Information	

Page 16

>100,000 species are represented in GenBank

all species	**128,941**
viruses	**6,137**
bacteria	**31,262**
archaea	**2,100**
eukaryota	**87,147**

Table 2-1
Page 17

The most sequenced organisms in GenBank

Homo sapiens	11.2 billion bases
Mus musculus	7.5b
Rattus norvegicus	5.7b
Danio rerio	2.1b
Bos taurus	1.9b
Zea mays	1.4b
Oryza sativa (japonica)	1.2b
Xenopus tropicalis	0.9b
Canis familiaris	0.8b
Drosophila melanogaster	0.7b

Updated 8-29-05
GenBank release 149.0

Table 2-2
Page 18

National Center for Biotechnology Information (NCBI)

www.ncbi.nlm.nih.gov

Page 24

www.ncbi.nlm.nih.gov

Fig. 2.5
Page 25

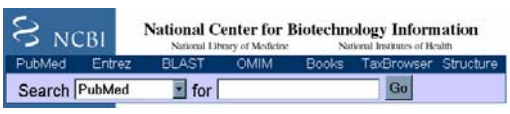
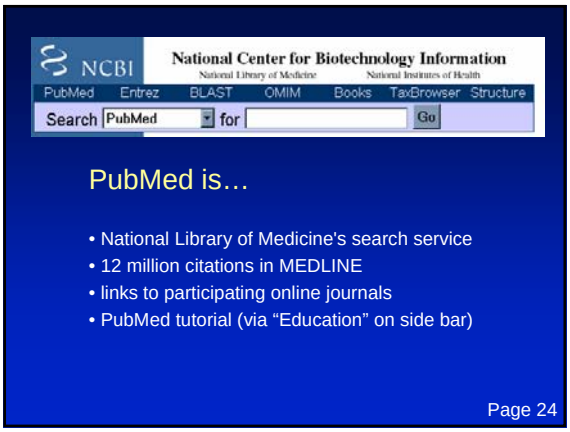
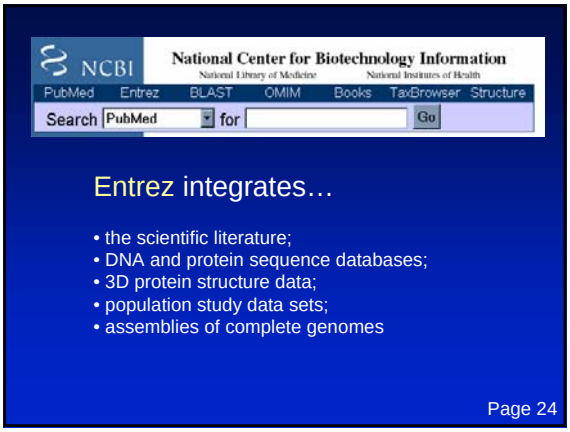
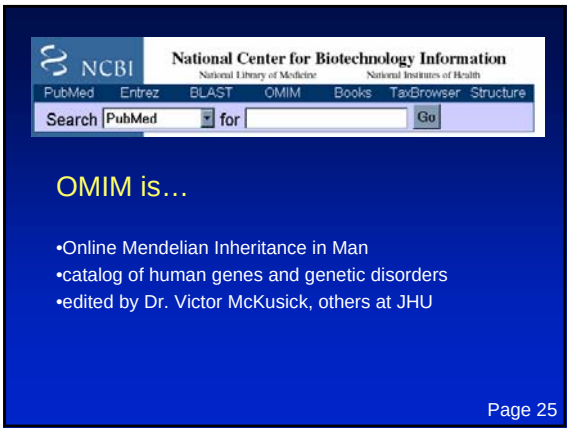


Fig. 2.5
Page 25


NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search PubMed for Go
PubMed is…
• National Library of Medicine's search service
• 12 million citations in MEDLINE
• links to participating online journals
• PubMed tutorial (via "Education" on side bar)
Page 24


NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search PubMed for Go
Entrez integrates…
• the scientific literature;
• DNA and protein sequence databases;
• 3D protein structure data;
• population study data sets;
• assemblies of complete genomes
Page 24

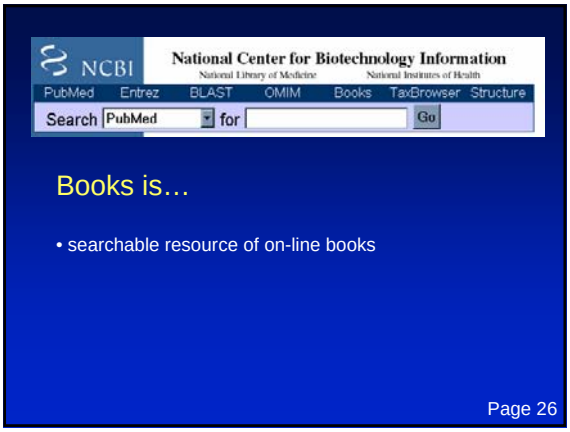

NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search PubMed for Go
BLAST is…
• Basic Local Alignment Search Tool
• NCBI's sequence similarity search tool
• supports analysis of DNA and protein databases
• 80,000 searches per day
Page 25



OMIM is…

- Online Mendelian Inheritance in Man
- catalog of human genes and genetic disorders
- edited by Dr. Victor McKusick, others at JHU

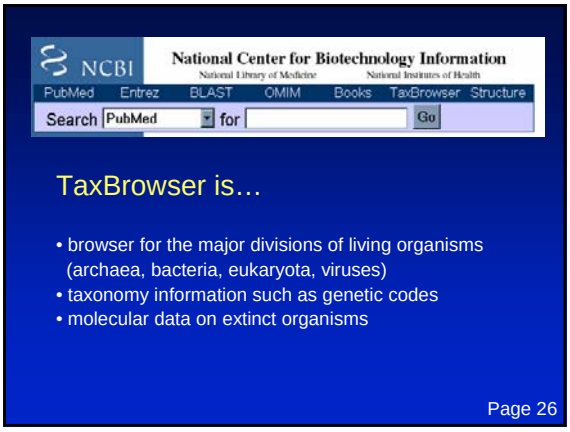
Page 25



Books is…

- searchable resource of on-line books

Page 26



TaxBrowser is…

- browser for the major divisions of living organisms (archaea, bacteria, eukaryota, viruses)
- taxonomy information such as genetic codes
- molecular data on extinct organisms

Page 26

Structure site includes…

- Molecular Modelling Database (MMDB)
- biopolymer structures obtained from the Protein Data Bank (PDB)
- Cn3D (a 3D-structure viewer)
- vector alignment search tool (VAST)

Page 26

Accessing information on molecular sequences

Page 26

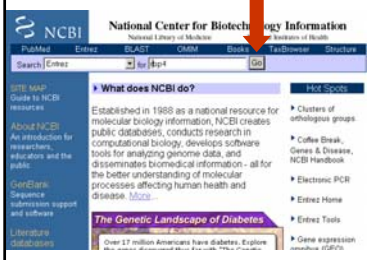
Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences.
You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

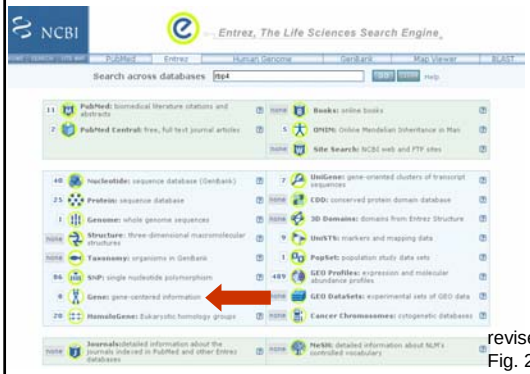
DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.

Page 26

From the NCBI home page, type "lectin" and hit "Go"



revised
Fig. 2.7
Page 29



revised
Fig. 2.7
Page 29

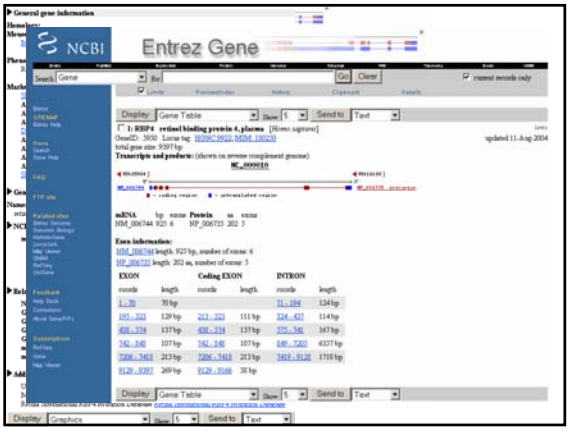


Fig. 2.9
Page 32

FASTA format

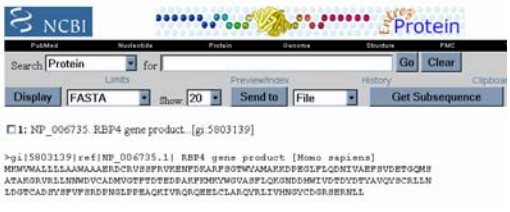


Fig. 2.10
Page 32

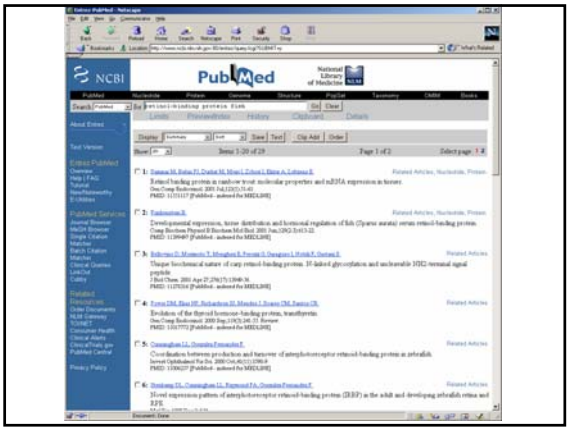
PubMed at NCBI to find literature information

PubMed is the NCBI gateway to MEDLINE.

MEDLINE contains bibliographic citations and author abstracts from over 4,600 journals published in the United States and in 70 foreign countries.

It has 12 million records dating back to 1966.

Page 35



BLAST

BLAST searching is fundamental to understanding the relatedness of any favorite query sequence to other known proteins or DNA sequences.

Applications include
- identifying homologs (orthologs and paralogs)
- discovering new genes or proteins
- discovering variants of genes or proteins
- investigating expressed sequence tags (ESTs)
- exploring protein structure and function

page 88

Four components to a BLAST search

(1) Choose the sequence (query)

(2) Select the BLAST program

(3) Choose the database to search

(4) Choose optional parameters

Then click “BLAST”

page 88

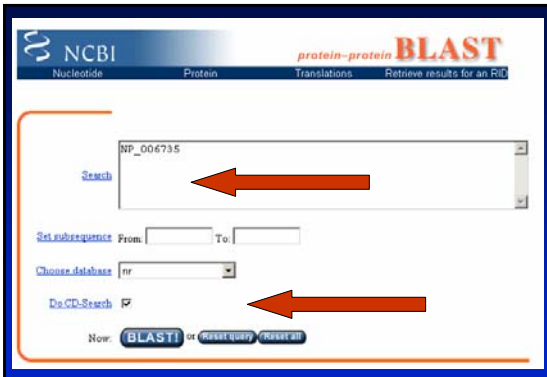


Fig. 4.1
page 89

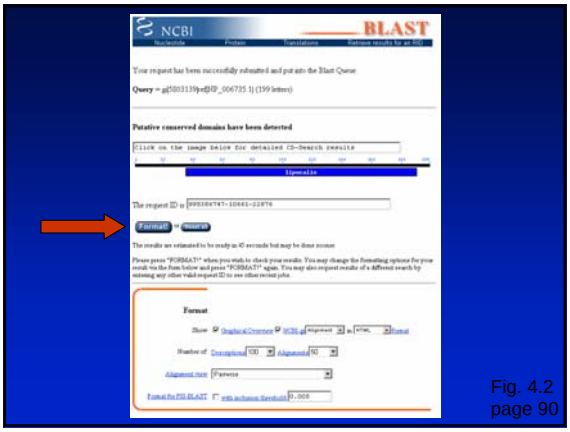


Fig. 4.2
page 90

Step 1: Choose your sequence

Sequence can be input in FASTA format or as accession number

page 89

Example of the FASTA format for a BLAST query

Fig. 2.10
page 32

Step 2: Choose the BLAST program

Step 2: Choose the BLAST program

blastn (nucleotide BLAST)

blastp (protein BLAST)

tblastn (translated BLAST)

blastx (translated BLAST)

tblastx (translated BLAST)

page 90

Choose the BLAST program

Program	Input		Database
blastn	DNA	1	DNA
blastp	protein	1	protein
blastx	DNA	6	protein
tblastn	protein	6	DNA
tblastx	DNA	36	DNA

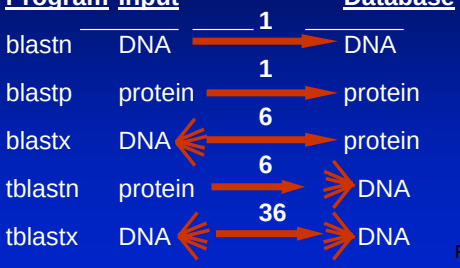
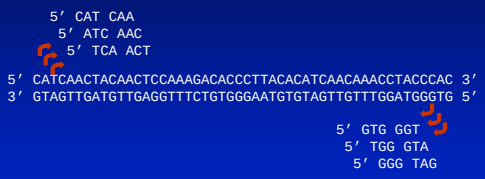


Fig. 4.3
page 91

DNA potentially encodes six proteins



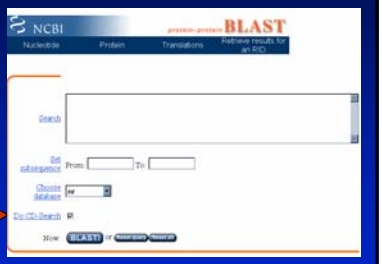
page 92

Step 3: choose the database

nr = non-redundant (most general database)
dbest = database of expressed sequence tags
dbsts = database of sequence tag sites
gss = genomic survey sequences
htgs = high throughput genomic sequence

page 92-93

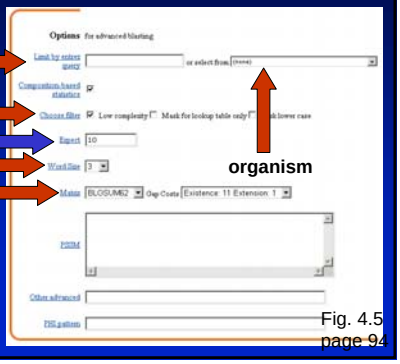
Step 4a: Select optional search parameters



CD search

page 93

Step 4a: Select optional search parameters



Entrez!

Filter

Expect

Word size

Scoring matrix

organism

Fig. 4.5
page 94

BLAST: optional parameters

You can...
- choose the organism to search
- turn filtering on/off
- change the substitution matrix
- change the expect (e) value
- change the word size
- change the output format

page 93

Fig. 4.6
page 95

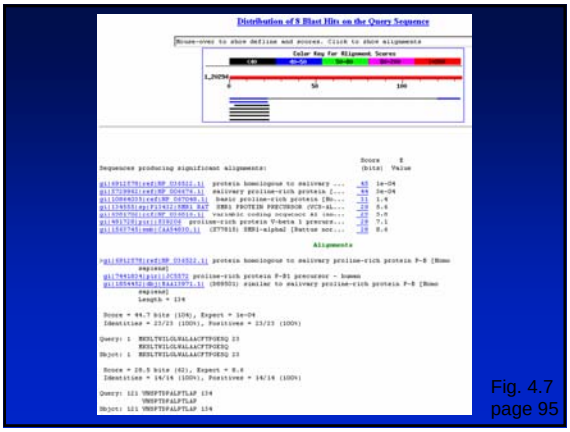


Fig. 4.7
page 95

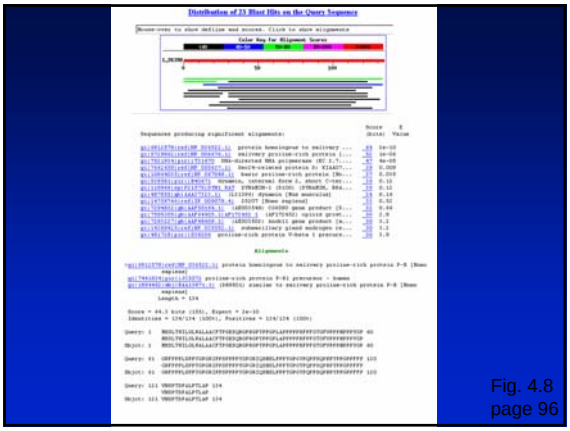


Fig. 4.8
page 96

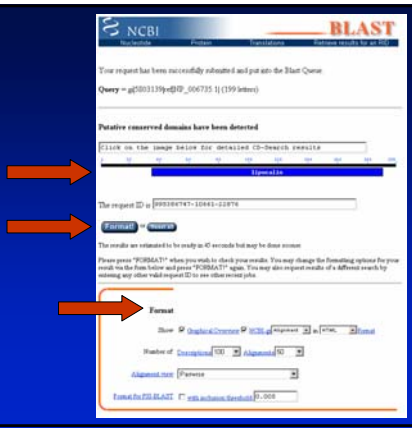
Step 4b: optional formatting parameters

Alignment view

Descriptions

Alignments

page 97



(page 90)

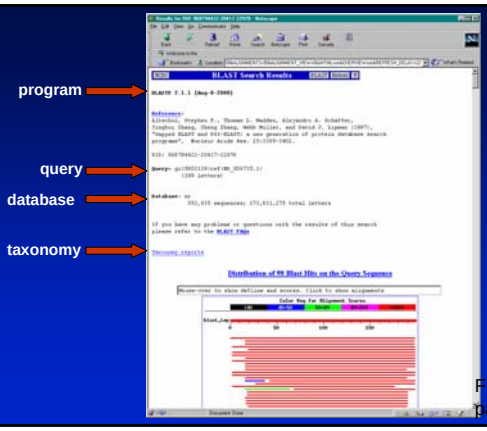
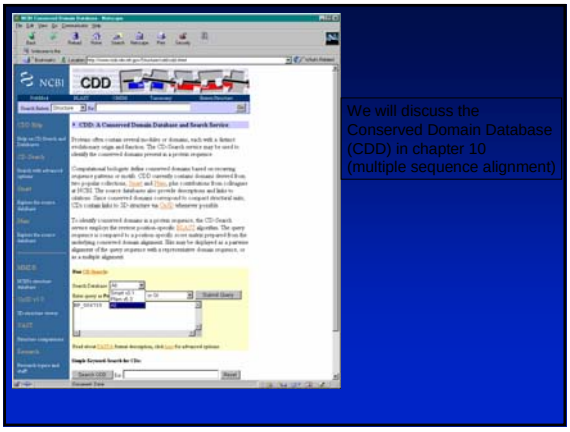
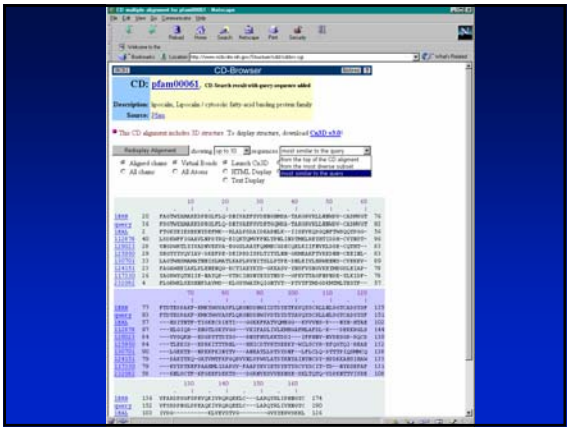


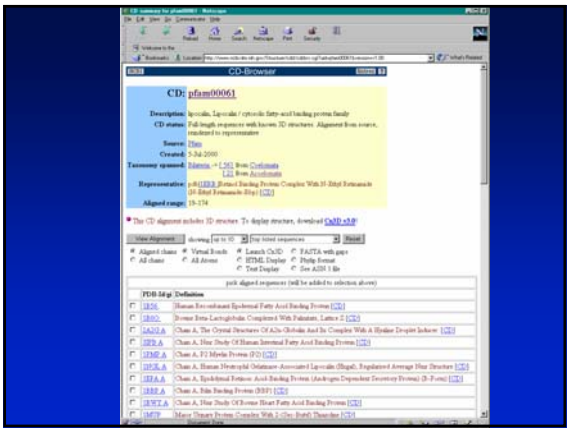
Fig. 4.9
page 98



We will discuss the Conserved Domain Database (CDD) in chapter 10 (multiple sequence alignment)

We will discuss the Conserved Domain Database (CDD) in chapter 10 (multiple sequence alignment)





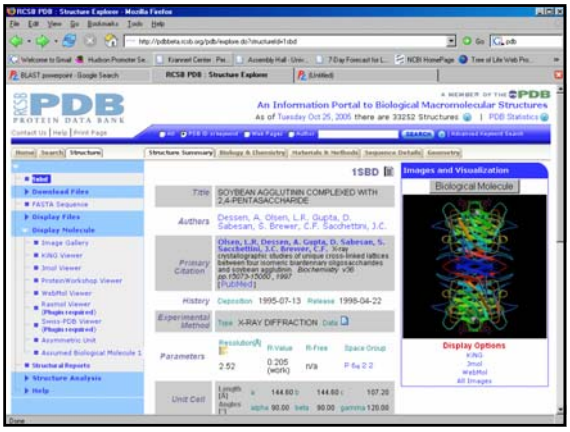
Protein 3D structure

- NCBI – Structure

Unlike mostly everything else, NCBI is not the best

- http://pdbbeta.rcsb.org/pdb/Welcome.do

(latest version of SDSC PDB site)



So now you can

- Find any sequence in the database
- Find relevant publications
- Match DNA to protein sequence
- Find database matches to DNA or protein
- Find conserved domains in protein
- Find the 3D structure of a protein

Without doing any experiments!