Bioinformatics Databases

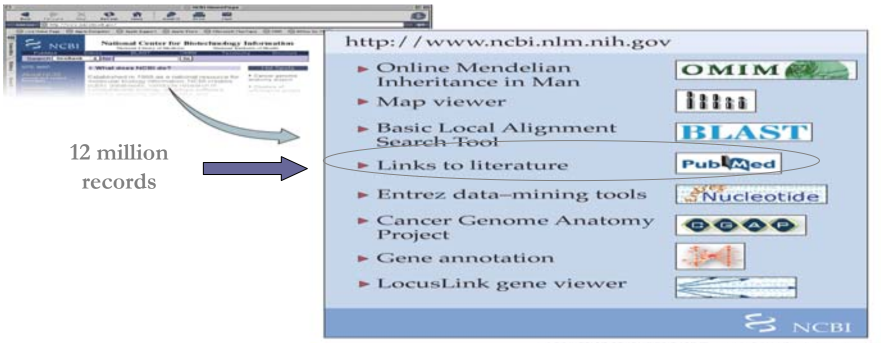
Dr. Taysir Hassan Abdel Hamid
Lecturer, Information Systems Department
Faculty of Computer and Information
Assiut University
taysirhs@aun.edu.eg
taysir_soliman@hotmail.com

Agenda

- Literature databases
- Sequence databases
- Other databases

So, what Computer Scientists do for Bioinformatics?

- Computer scientists are responsible for **INTEGRATING** and **ANALYZING** all literature from both patents and other publications in PubMED (MEDLINE) in NCBI



12 million records

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

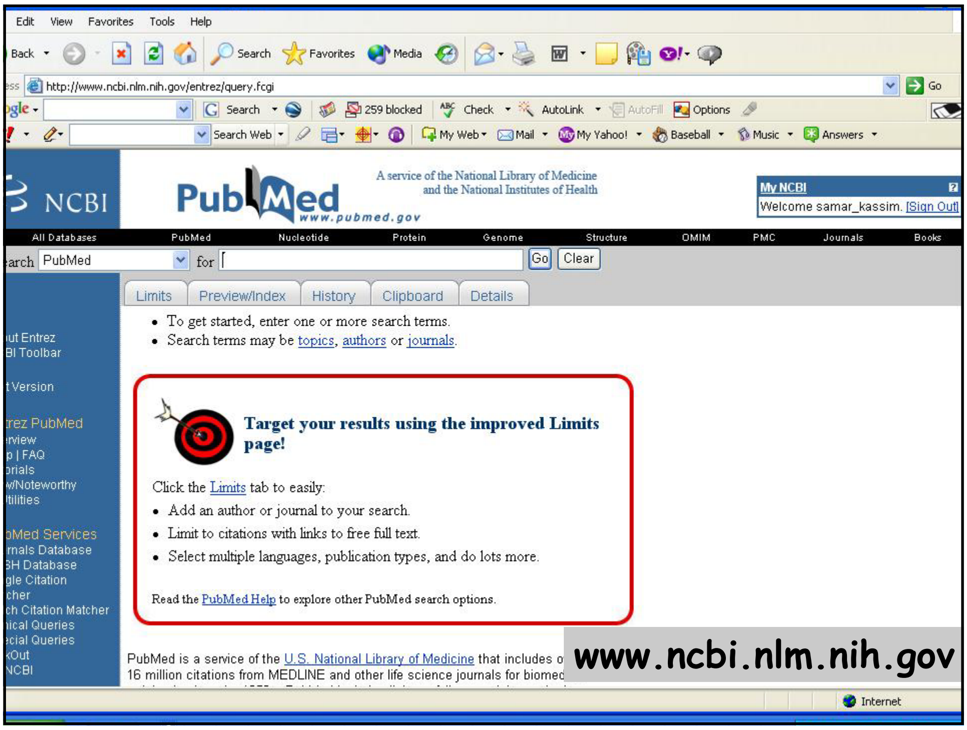
- ► Online Mendelian Inheritance in Man
- ► Map viewer
- ► Basic Local Alignment Search Tool
- ► Links to literature
- ► Entrez data-mining tools
- ► Cancer Genome Anatomy Project
- ► Gene annotation
- ► LocusLink gene viewer

A PRIMER OF GENOME SCIENCE, Figure 1.1 © 2002 Sinauer Associates, Inc.

Where do we get PubMed?

National Center for Biotechnology Information (NCBI)

www.ncbi.nlm.nih.gov


Edit View Favorites Tools Help
Back Search Favorites Media
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
Go
Search 259 blocked Check AutoLink AutoFill Options
Search Web My Web Mail My Yahoo! Baseball Music Answers
NCBI
PubMed www.pubmed.gov
A service of the National Library of Medicine and the National Institutes of Health
My NCBI
Welcome samar_kassim. [Sign Out]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
arch PubMed for
Go Clear
Limits Preview/Index History Clipboard Details
ut Entrez
BI Toolbar
t Version
trez PubMed
erview
p | FAQ
orials
w/Noteworthy
tilities
bMed Services
rnals Database
SH Database
gle Citation
cher
ch Citation Matcher
nical Queries
ecial Queries
kOut
NCBI
Internet

- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

Target your results using the improved Limits page!

Click the Limits tab to easily:

- Add an author or journal to your search.
- Limit to citations with links to free full text.
- Select multiple languages, publication types, and do lots more.

Read the PubMed Help to explore other PubMed search options.

PubMed is a service of the U.S. National Library of Medicine that includes o
16 million citations from MEDLINE and other life science journals for biomed

www.ncbi.nlm.nih.gov

Sequence Databases

Three major database organizations around the world are responsible for maintaining most of this data.

They largely ‘mirror’ one another and share accession codes, but NOT proper identifier names.

Sequence Databases (Cont…)

North America: the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), at the National Institute of Health (NIH).

- The GenBank is one of its databases.

Europe: the European Molecular Biology Laboratory (EMBL), the European Bioinformatics Institute (EBI), and the Swiss Institute of Bioinformatics' (SIB). There are also the expert Protein Analysis System (ExPasy),the SWISS-PROT and TrEMBL amino acid sequence databases.

Asia: The National Institute of Genetics (NIG) supports the Center for Information Biology's (CIG) & DNA Data Bank of Japan (DDBJ).

• All sequence databases contain these elements:

- Name: ID is a unique identifier
- Definition: A brief, one-line, textual sequence description.
- Accession Number: A constant data identifier.
- Source and taxonomy information.
- Complete literature references.
- Comments and keywords.
- The all important FEATURE table!
- A summary or checksum line.
- The sequence itself.

What is an accession number?

An accession number is label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	
Rs7079946	dbSNP (single nucleotide polymorphism)	DNA
N91759.1	An expressed sequence tag (1 of 170)	
NM_006744	RefSeq DNA sequence (from a transcript)	RNA
NP_007635	RefSeq protein	
AAC02945	GenBank protein	
Q28369	SwissProt protein	protein
1KT7	Protein Data Bank structure record	

So how do you access and manipulate all this data?

- Often on the InterNet over the World Wide Web:

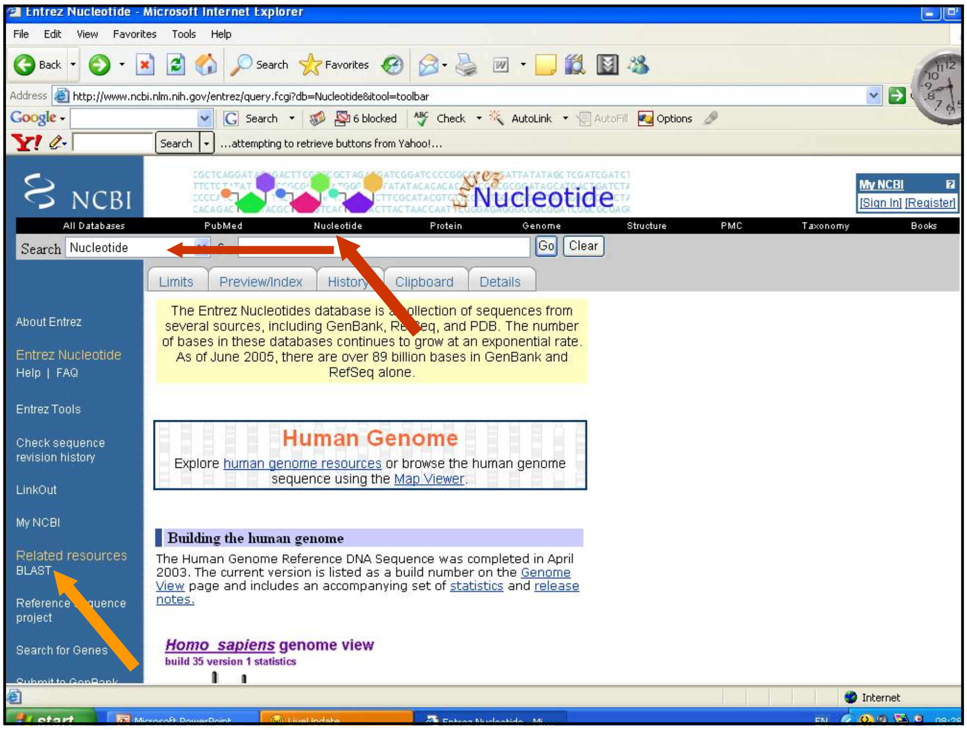
•Site	URL (Uniform Resource Locator)	Content
•Nat'l Center Biotech' Info'		http://www.ncbi.nlm.nih.gov/
	databases/analysis/software	
•PIR/NBRF	http://www-nbrf.georgetown.edu/	protein sequence database
•IUBIO Biology Archive	http://iubio.bio.indiana.edu/	database/software archive
•Univ. of Montreal	http://megasun.bch.umontreal.ca/	database/software archive
•Japan's GenomeNet	http://www.genome.ad.jp/	databases/analysis/software
•European Mol' Bio' Lab'	http://www.embl-heidelberg.de/	databases/analysis/software
•European Bioinformatics	http://www.ebi.ac.uk/	databases/analysis/software
•The Sanger Institute	http://www.sanger.ac.uk/	databases/analysis/software
•Univ. of Geneva BioWeb	http://www.expasy.ch/	databases/analysis/software
•ProteinDataBank	http://www.rcsb.org/pdb/	3D mol' structure database
•Molecules R Us	http://molbio.info.nih.gov/cgi-bin/pdb/	3D protein/nuc' visualization
•The Genome DataBase	http://www.gdb.org/	The Human Genome Project
•Stanford Genomics	http://genome-www.stanford.edu/	various genome projects
•Inst. for Genomic Res'rch	http://www.tigr.org/	esp. microbial genome projects
•HIV Sequence Database	http://hiv-web.lanl.gov/	HIV epidemeology seq' DB
•The Tree of Life	http://tolweb.org/tree/phylogeny.html	overview of all phylogeny

Net access software examples

- **Internet surfing tools**: a World Wide Web browser such as MS Explorer.
  - Advantage: Can access last night's updates

So, what to do with a sequence you had retrieved in your lab?

```
5'GGCCAGTACTGGGCGCACTTGCACTCC
TTTCTCTCCTTCAGGTTGGTAACCATG
ACGATGGTGGCTGAGTTTGTTCCCAGA
TCCATCCGCCAGAAATCATTCACCGTTT
CTTCTTTTGTCCTTGTGCAGCAATGAA
TTTGTTCTTTTCTTGGTAA3'
```


Entrez Nucleotide - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Google Search 6 blocked Check AutoLink AutoFill Options
Search ...attempting to retrieve buttons from Yahoo!...
NCBI
Entrez Nucleotide
My NCBI [Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Entrez Nucleotide
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI
Related resources
BLAST
Reference sequence project
Search for Genes
Submit to GenBank
Internet

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide&itool=toolbar

Search Nucleotide

The Entrez Nucleotides database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB. The number of bases in these databases continues to grow at an exponential rate. As of June 2005, there are over 89 billion bases in GenBank and RefSeq alone.

Human Genome

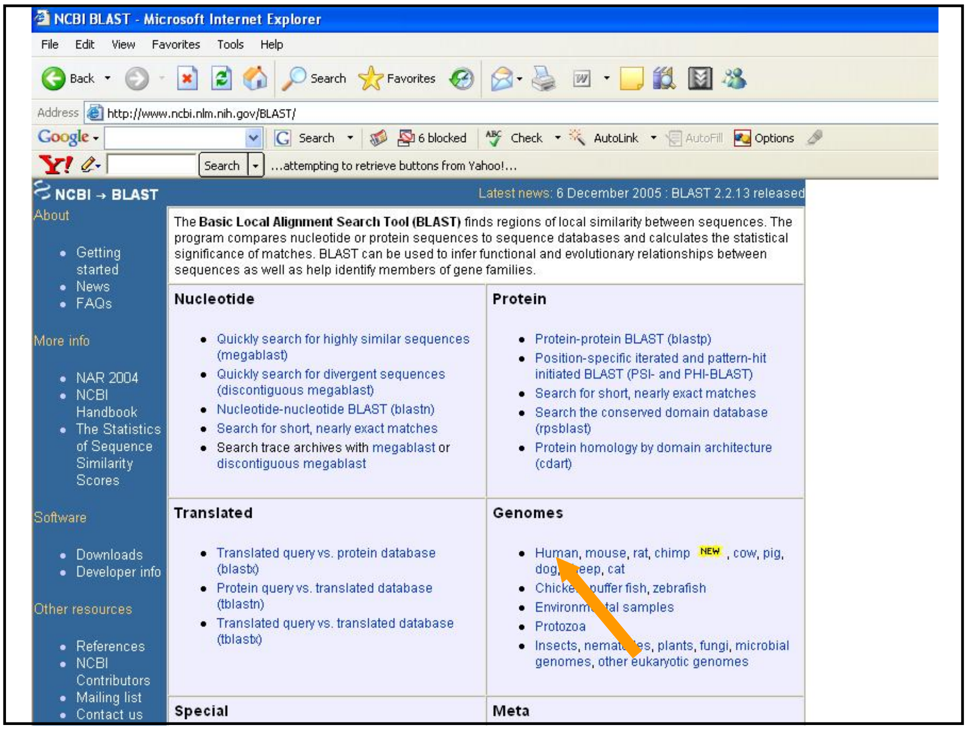
Explore human genome resources or browse the human genome sequence using the Map Viewer.

Building the human genome

The Human Genome Reference DNA Sequence was completed in April 2003. The current version is listed as a build number on the Genome View page and includes an accompanying set of statistics and release notes.

Homo sapiens genome view

build 35 version 1 statistics


NCBI BLAST - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Google Search 6 blocked Check AutoLink AutoFill Options
Search ...attempting to retrieve buttons from Yahoo!...
NCBI → BLAST
About
Getting started
News
FAQs
More info
NAR 2004
NCBI Handbook
The Statistics of Sequence Similarity Scores
Software
Downloads
Developer info
Other resources
References
NCBI Contributors
Mailing list
Contact us

Address http://www.ncbi.nlm.nih.gov/BLAST/

Latest news: 6 December 2005 : BLAST 2.2.13 released

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

Translated

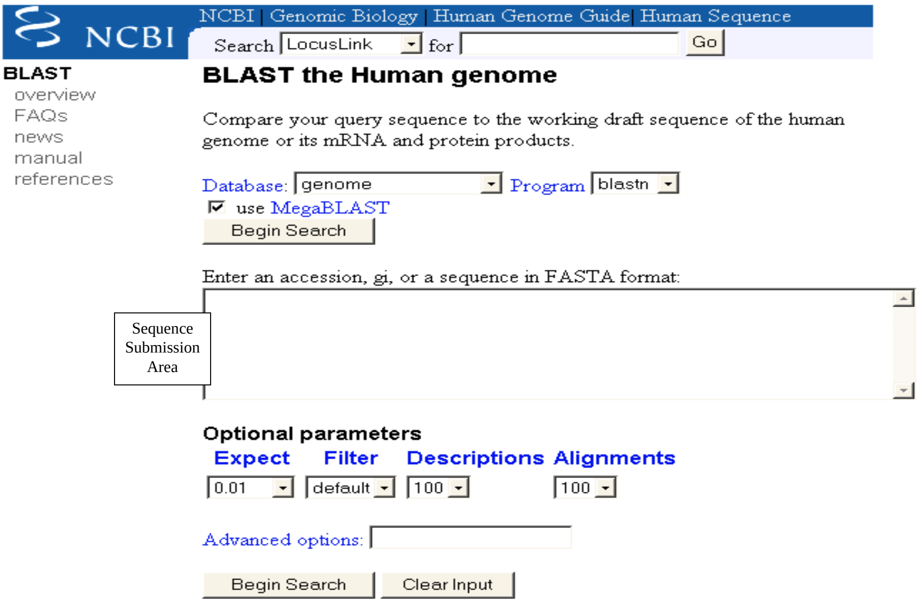
- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat, chimp NEW , cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Environmental samples
- Protozoa
- Insects, nematodes, plants, fungi, microbial genomes, other eukaryotic genomes

Special

Meta



NCBI | Genomic Biology | Human Genome Guide| Human Sequence

Search LocusLink for Go

BLAST
overview
FAQs
news
manual
references

BLAST the Human genome

Compare your query sequence to the working draft sequence of the human genome or its mRNA and protein products.

Database: genome Program blastn

☑ use MegaBLAST

Begin Search

Enter an accession, gi, or a sequence in FASTA format:

Sequence Submission Area

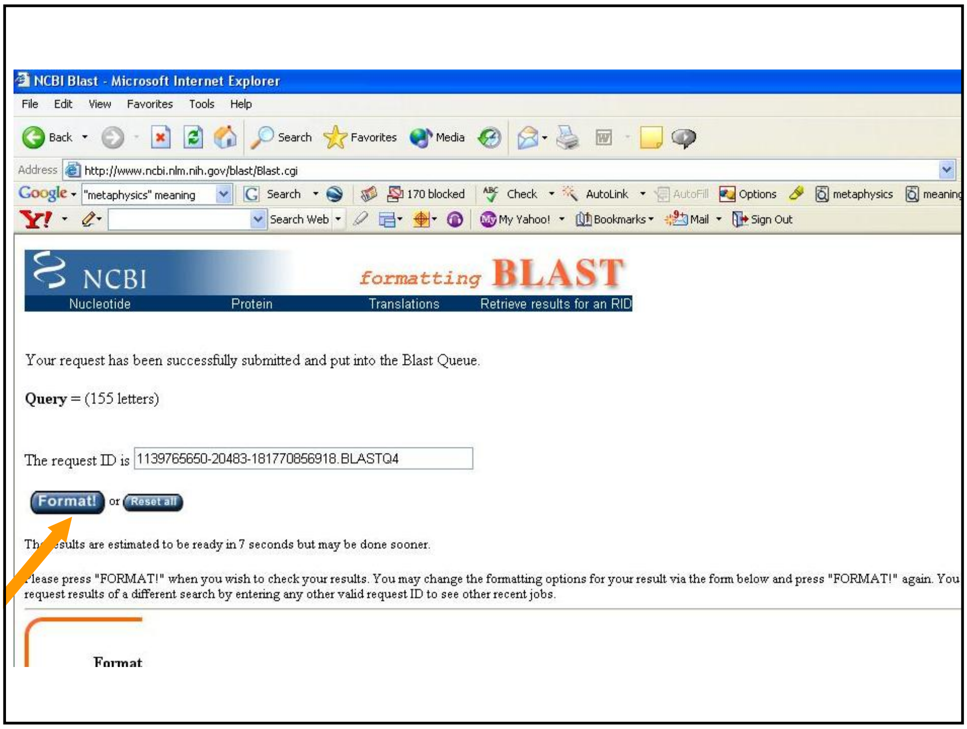
Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options:

Begin Search Clear Input

BLAST Against the Human Genome



NCBI Blast - Microsoft Internet Explorer

Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

formatting BLAST

Nucleotide Protein Translations Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = (155 letters)

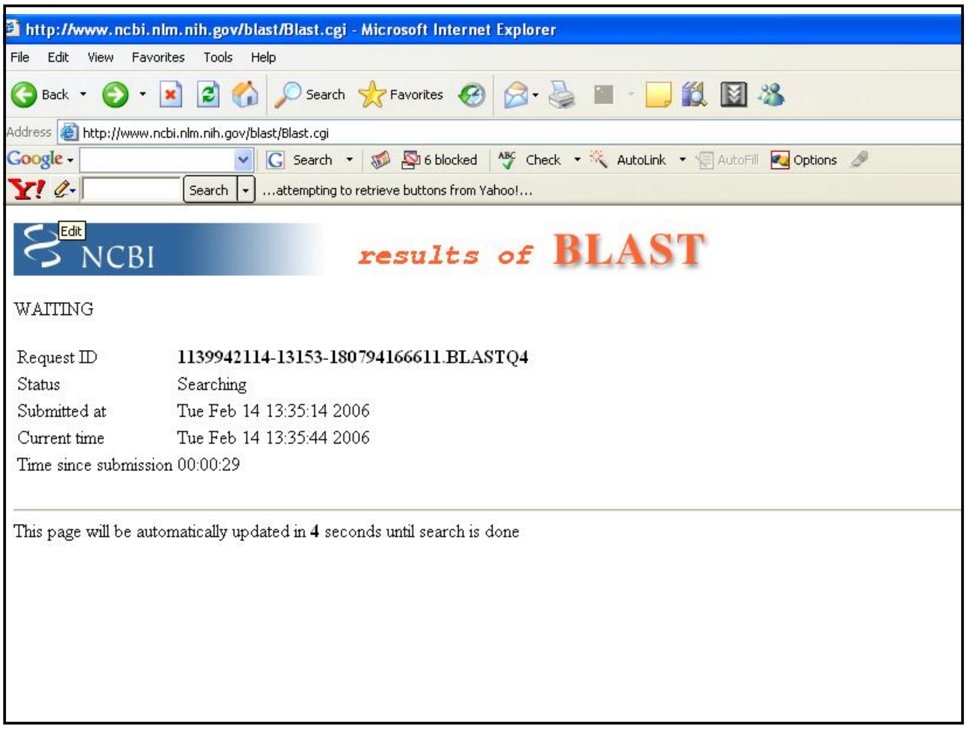
The request ID is 1139765650-20483-181770856918.BLASTQ4

Format! or Reset all

The results are estimated to be ready in 7 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You request results of a different search by entering any other valid request ID to see other recent jobs.

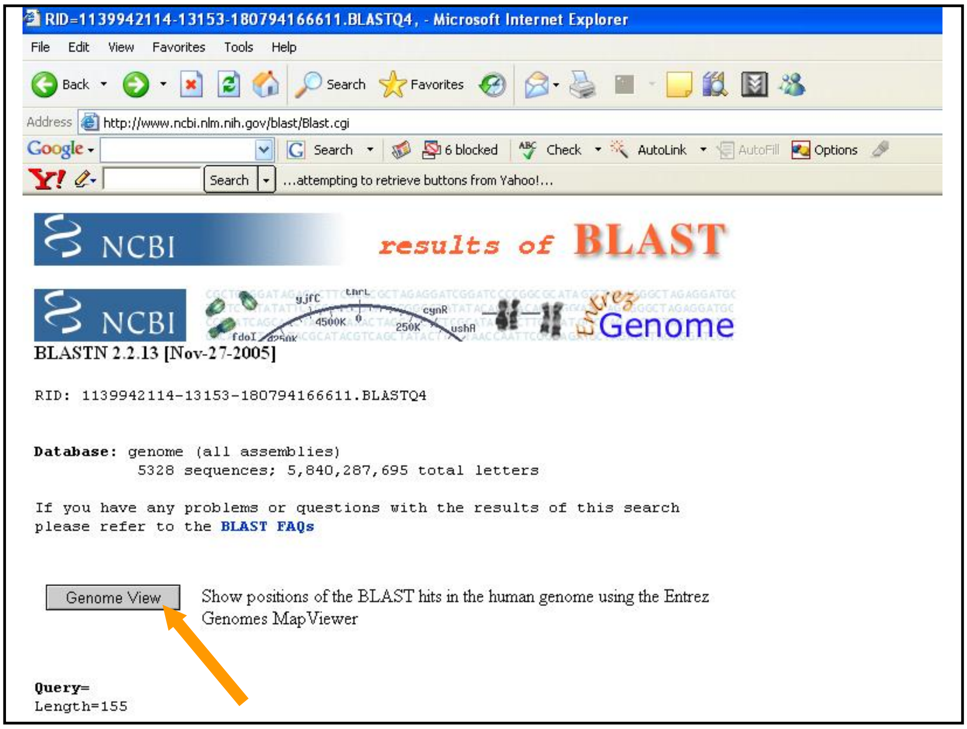
Format


http://www.ncbi.nlm.nih.gov/blast/Blast.cgi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Google Search 6 blocked Check AutoLink AutoFill Options
Y! Search ...attempting to retrieve buttons from Yahoo!...
Edit
NCBI
results of BLAST

WAITING

Request ID	1139942114-13153-180794166611.BLASTQ4
Status	Searching
Submitted at	Tue Feb 14 13:35:14 2006
Current time	Tue Feb 14 13:35:44 2006
Time since submission	00:00:29

This page will be automatically updated in **4** seconds until search is done

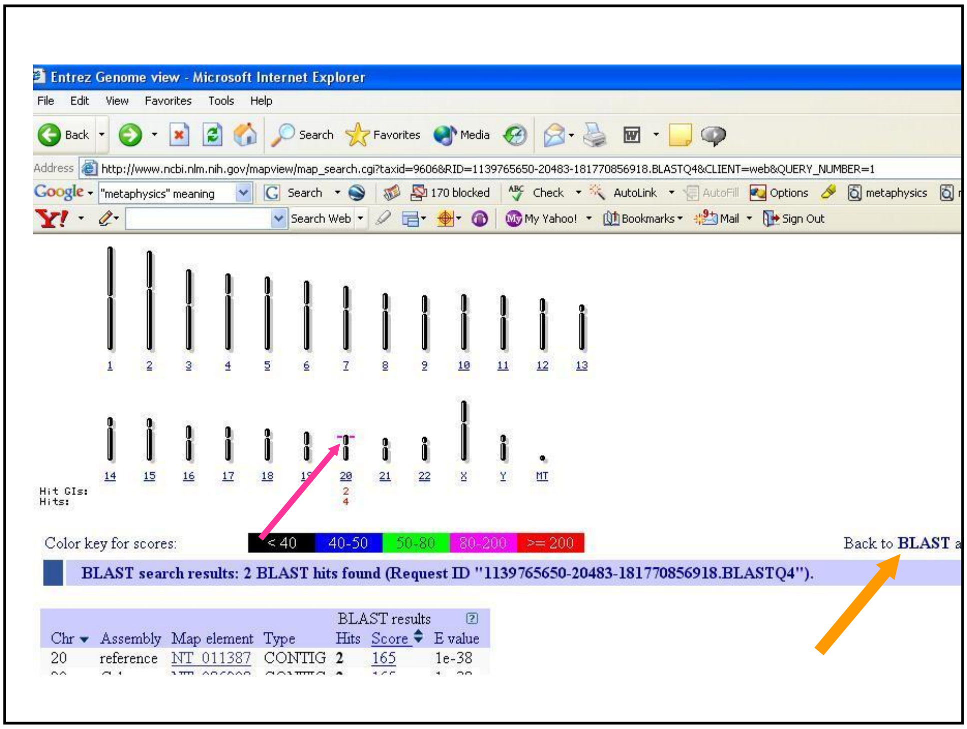

RID=1139942114-13153-180794166611.BLASTQ4, - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Google Search 6 blocked Check AutoLink AutoFill Options
Y! Search ...attempting to retrieve buttons from Yahoo!...
NCBI
results of BLAST
NCBI
Entrez Genome

BLASTN 2.2.13 [Nov-27-2005]

```
RID: 1139942114-13153-180794166611.BLASTQ4


Database: genome (all assemblies)
           5328 sequences; 5,840,287,695 total letters

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

Genome View   Show positions of the BLAST hits in the human genome using the Entrez Genomes MapViewer

```
Query=
Length=155
```



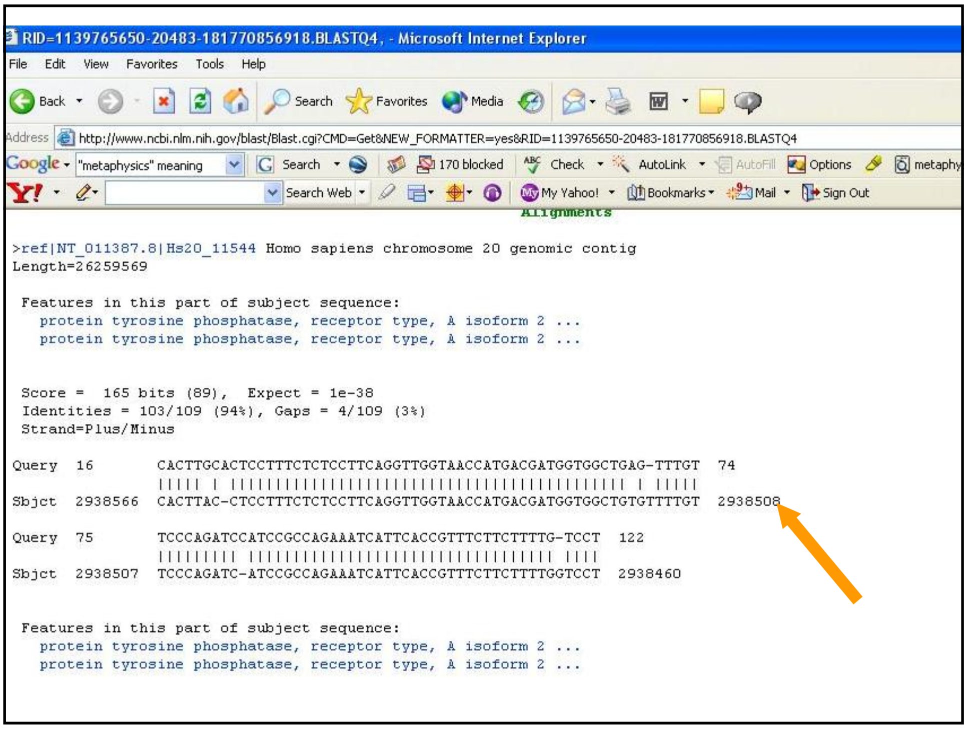
http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606&RID=1139765650-20483-181770856918.BLASTQ4&CLIENT=web&QUERY_NUMBER=1

Color key for scores: < 40 | 40-50 | 50-80 | 80-200 | >= 200

Back to BLAST

BLAST search results: 2 BLAST hits found (Request ID "1139765650-20483-181770856918.BLASTQ4").

				BLAST results		
Chr	Assembly	Map element	Type	Hits	Score	E value
20	reference	NT_011387	CONTIG	**2**	165	1e-38



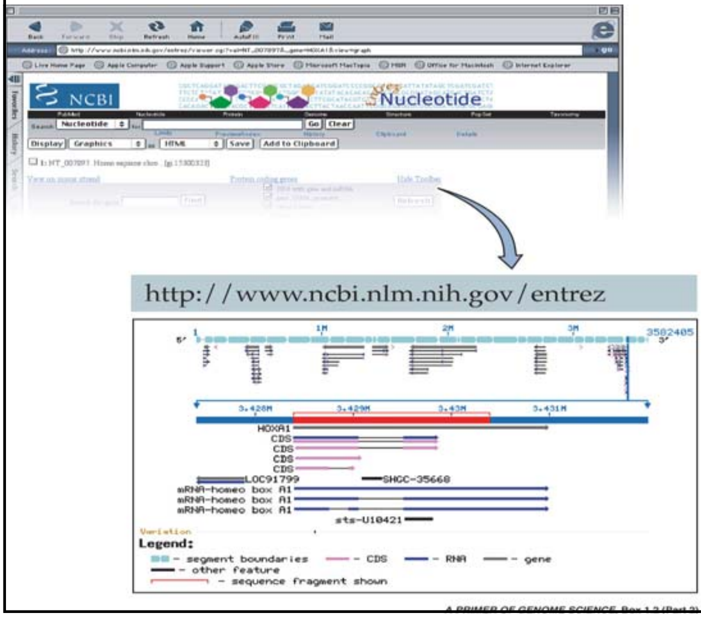
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Get&NEW_FORMATTER=yes&RID=1139765650-20483-181770856918.BLASTQ4

```
Alignments

>ref|NT_011387.8|Hs20_11544 Homo sapiens chromosome 20 genomic contig
Length=26259569

 Features in this part of subject sequence:
   protein tyrosine phosphatase, receptor type, A isoform 2 ...
   protein tyrosine phosphatase, receptor type, A isoform 2 ...


 Score =  165 bits (89),  Expect = 1e-38
 Identities = 103/109 (94%), Gaps = 4/109 (3%)
 Strand=Plus/Minus

Query  16       CACTTGCACTCCTTTCTCTCCTTCAGGTTGGTAACCATGACGATGGTGGCTGAG-TTTGT  74
                ||||| | |||||||||||||||||||||||||||||||||||||||||||| | |||||
Sbjct  2938566  CACTTAC-CTCCTTTCTCTCCTTCAGGTTGGTAACCATGACGATGGTGGCTGTGTTTTGT  2938508

Query  75       TCCCAGATCCATCCGCCAGAAATCATTCACCGTTTCTTCTTTTG-TCCT  122
                ||||||||| |||||||||||||||||||||||||||||||||| ||||
Sbjct  2938507  TCCCAGATC-ATCCGCCAGAAATCATTCACCGTTTCTTCTTTTGGTCCT  2938460


 Features in this part of subject sequence:
   protein tyrosine phosphatase, receptor type, A isoform 2 ...
   protein tyrosine phosphatase, receptor type, A isoform 2 ...
```

Example FASTA sequence

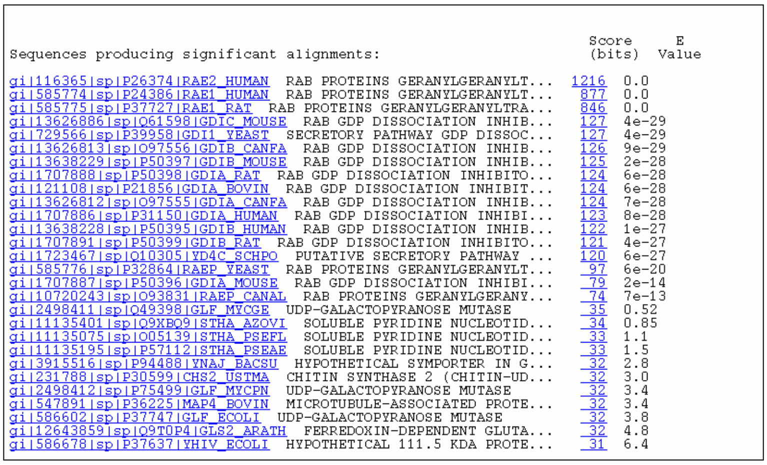
```
>JC2395
NVSDVNLNK---YIWRTAEKMK---ICDAKKFARQHKIPESKIDEIEHNSPQDAAE----
-------------------------QKIQLLQCWYQSHGKT–GACQALIQGLRKANRCDI
AEEIQAM
>KPEL_DROME
MAIRLLPLPVRAQLCAHLDAL-----DVWQQLATAVKLYPDQVEQISSQKQRGRS-----
-------------------------ASNEFLNIWGGQYN----HTVQTLFALFKKLKLHN
AMRLIKDY
>FASA_MOUSE
NASNLSLSK---YIPRIAEDMT---IQEAKKFARENNIKEGKIDEIMHDSIQDTAE----
-------------------------QKVQLLLCWYQSHGKS--DAYQDLIKGLKKAECRR
TLDKFQDM
```

GenBank Files



A PRIMER OF GENOME SCIENCE, Box 1.3 (Part 2) © 2002 Sinauer Associates, Inc.

One-line descriptions



```
                                                                  Score    E
Sequences producing significant alignments:                      (bits)  Value

gi|116365|sp|P26374|RAE2_HUMAN  RAB PROTEINS GERANYLGERANYLT...   1216  0.0
gi|585774|sp|P24386|RAE1_HUMAN  RAB PROTEINS GERANYLGERANYLT...    877  0.0
gi|585775|sp|P37727|RAE1_RAT  RAB PROTEINS GERANYLGERANYLTRA...    846  0.0
gi|13626886|sp|Q61598|GDIC_MOUSE  RAB GDP DISSOCIATION INHIB...    127  4e-29
gi|729566|sp|P39958|GDI1_YEAST  SECRETORY PATHWAY GDP DISSOC...    127  4e-29
gi|13626813|sp|O97556|GDIB_CANFA  RAB GDP DISSOCIATION INHIB...    126  9e-29
gi|13638229|sp|P50397|GDIB_MOUSE  RAB GDP DISSOCIATION INHIB...    125  2e-28
gi|1707888|sp|P50398|GDIA_RAT  RAB GDP DISSOCIATION INHIBITO...    124  6e-28
gi|121108|sp|P21856|GDIA_BOVIN  RAB GDP DISSOCIATION INHIBIT...    124  6e-28
gi|13626812|sp|O97555|GDIA_CANFA  RAB GDP DISSOCIATION INHIB...    124  7e-28
gi|1707886|sp|P31150|GDIA_HUMAN  RAB GDP DISSOCIATION INHIBI...    123  8e-28
gi|13638228|sp|P50395|GDIB_HUMAN  RAB GDP DISSOCIATION INHIB...    122  1e-27
gi|1707891|sp|P50399|GDIB_RAT  RAB GDP DISSOCIATION INHIBITO...    121  4e-27
gi|1723467|sp|Q10305|YD4C_SCHPO  PUTATIVE SECRETORY PATHWAY ...    120  6e-27
gi|585776|sp|P32864|RAEP_YEAST  RAB PROTEINS GERANYLGERANYLT...     97  6e-20
gi|1707887|sp|P50396|GDIA_MOUSE  RAB GDP DISSOCIATION INHIBI...     79  2e-14
gi|10720243|sp|O93831|RAEP_CANAL  RAB PROTEINS GERANYLGERANY...     74  7e-13
gi|2498411|sp|Q49398|GLF_MYCGE  UDP-GALACTOPYRANOSE MUTASE          35  0.52
gi|11135401|sp|Q9XBQ9|STHA_AZOVI  SOLUBLE PYRIDINE NUCLEOTID...     34  0.85
gi|11135075|sp|O05139|STHA_PSEFL  SOLUBLE PYRIDINE NUCLEOTID...     33  1.1
gi|11135195|sp|P57112|STHA_PSEAE  SOLUBLE PYRIDINE NUCLEOTID...     33  1.5
gi|3915516|sp|P94488|YNAJ_BACSU  HYPOTHETICAL SYMPORTER IN G...     32  2.8
gi|231788|sp|P30599|CHS2_USTMA  CHITIN SYNTHASE 2 (CHITIN-UD...     32  3.0
gi|2498412|sp|P75499|GLF_MYCPN  UDP-GALACTOPYRANOSE MUTASE          32  3.4
gi|547891|sp|P36225|MAP4_BOVIN  MICROTUBULE-ASSOCIATED PROTE...     32  3.4
gi|586602|sp|P37747|GLF_ECOLI  UDP-GALACTOPYRANOSE MUTASE           32  3.8
gi|12643859|sp|Q9T0P4|GLS2_ARATH  FERREDOXIN-DEPENDENT GLUTA...     32  4.8
gi|586678|sp|P37637|YHIV_ECOLI  HYPOTHETICAL 111.5 KDA PROTE...     31  6.4
```

Multiple Alignment Formats

- Formats for storing multiple alignments are specified

- FASTA, GCG MSF, ALN, etc

FASTA Format

- Each sequence begins with a description line ‘>’
- Sequence data follows, with gap character ‘-’

Example Fasta sequence

```
>JC2395
NVSDVNLNK---YIWRTAEKMK---ICDAKKFARQHKIPESKIDEIEHNSPQDAAE----
-------------------------QKIQLLQCWYQSHGKT—GACQALIQGLRKANRCDI
AEEIQAM
>KPEL_DROME
MAIRLLPLPVRAQLCAHLDAL-----DVWQQLATAVKLYPDQVEQISSQKQRGRS-----
-------------------------ASNEFLNIWGGQYN----HTVQTLFALFKKLKLHN
AMRLIKDY
>FASA_MOUSE
NASNLSLSK---YIPRIAEDMT---IQEAKKFARENNIKEGKIDEIMHDSIQDTAE----
-------------------------QKVQLLLCWYQSHGKS--DAYQDLIKGLKKAECRR
TLDKFQDM
```

Sequence Conversion Programs

- SEQIO
  - http://bioweb.pasteur.fr/docs/seqio/seqio.html
- READSEQ
  - http://bimas.dcrt.nih.gov/molbio/readseq/

Searching Sequence Databases

- Compare a query sequence against a target database
- Return significant results
  - Possible Homolgous sequences
  - Yields insight into structure and function

FASTA

- First rapid database search utility
- 50 times faster than Dynamic Programming
- Based on a heuristic – not guaranteed to locate optimal solution

FASTA Algorithm

- Hashing approach:
  - Construct a table showing each word of length k (k-tuple) for query and target
    - 1 or 2 for proteins
    - 4 or 6 for DNA
  - Relative positions calculated by subtracting positions
  - Matches in same phase are strung together

FASTA Algorithm

- Identify 10 regions with highest density of hits
  - Trim regions to include only residues contributing to high scores
  - Associate init1 score to each region

Each region is partial alignment without gaps

FASTA Algorithm

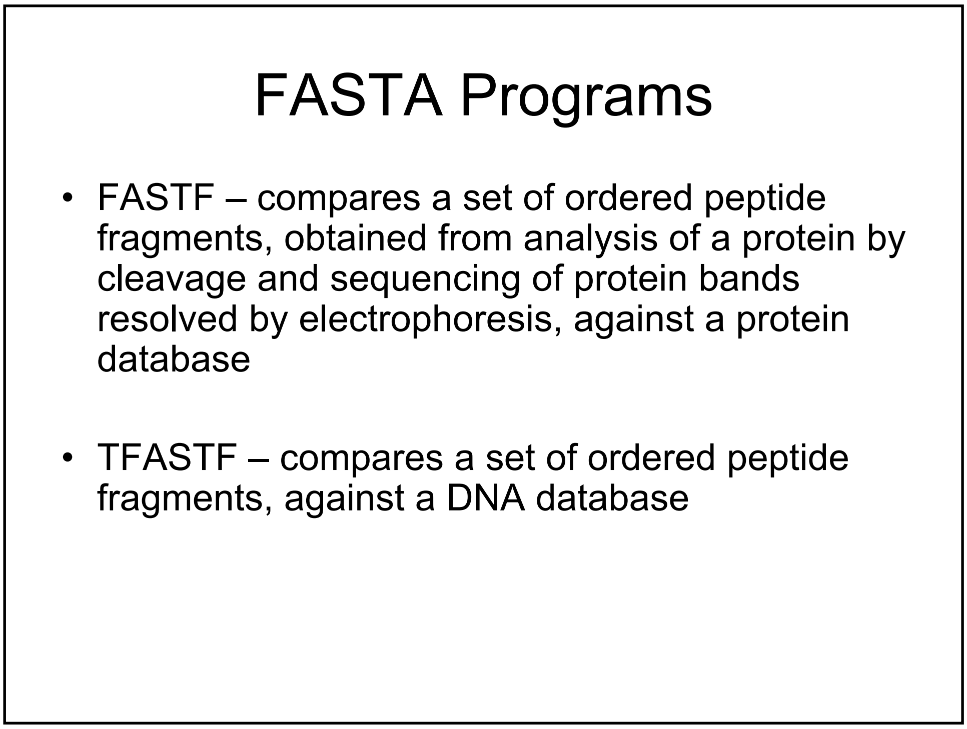
- Join initial regions to form approximate alignments with gaps
- Assign score
  - Sum of init1 scores for initial regions
  - Subtract gap penalty

FASTA Alignment Output

```
>>MERR_STAAU mercuric resistance operon regulatory protei (135 aa)
 initn: 292 init1: 172 opt: 298 Z-score: 373.6 expect() 3.5e-14
Smith-Waterman score: 298;  36.923% identity in 130 aa overlap

               10        20        30        40        50        60
MerR   MENNLENLTIGVFAKAAGVNVETIRFYQRKGLLLEPDKPYGSIRRYGEADVTRVRFVKSA
              . :. .:::  :: ::.:.:.::::.  : .  .. : :.:  . ::::.:
MERR_S      MGMKISELAKACDVNKETVRYYERKGLIAGPPRNESGYRIYSEETADRVRFIKRM
                    10        20        30        40        50

               70          80        90       100       110
MerR   QRLGFSLDEIAELLRL--EDGTHCEEASSLAEHKLKDVREKMADLARMEAVLSELVCACH
       ..: ::: ::  :. .  .:: .:..  ... .: :....:.  : :.. .: ::   :
MERR_S KELDFSLKEIHLLFGVVDQDGERCKDMYAFTVQKTKEIERKVQGLLRIQRLLEELKEKCP
          60        70        80        90       100       110

      120       130       140
MerR   ARRGNVSCPLIASLQGGASLAGSAMP
        ...  .::.: .:.::
MERR_S DEKAMYTCPIIETLMGGPDK
         120       130
```

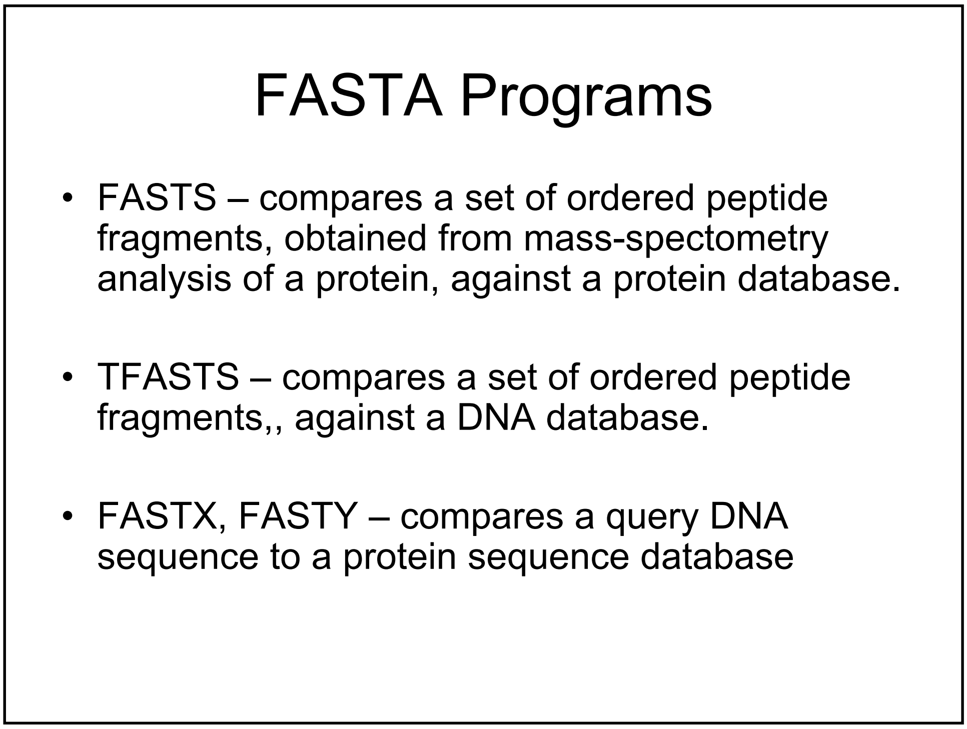
FASTA Programs

- FASTA – protein to protein OR DNA to DNA
- TFASTA –query protein to DNA database
  - the DNA database is first translated in all six reading frames



FASTA Programs

- FASTF – compares a set of ordered peptide fragments, obtained from analysis of a protein by cleavage and sequencing of protein bands resolved by electrophoresis, against a protein database
- TFASTF – compares a set of ordered peptide fragments, against a DNA database



FASTA Programs

- FASTS – compares a set of ordered peptide fragments, obtained from mass-spectometry analysis of a protein, against a protein database.
- TFASTS – compares a set of ordered peptide fragments,, against a DNA database.
- FASTX, FASTY – compares a query DNA sequence to a protein sequence database

FASTA Programs

- TFASTX, TFASTY –protein sequence to a DNA sequence or DNA database
  - DNA sequence translated in all six reading frames
    - Translated from beginning to end
    - Termination codons translated into unknown amino acids

FASTA Programs

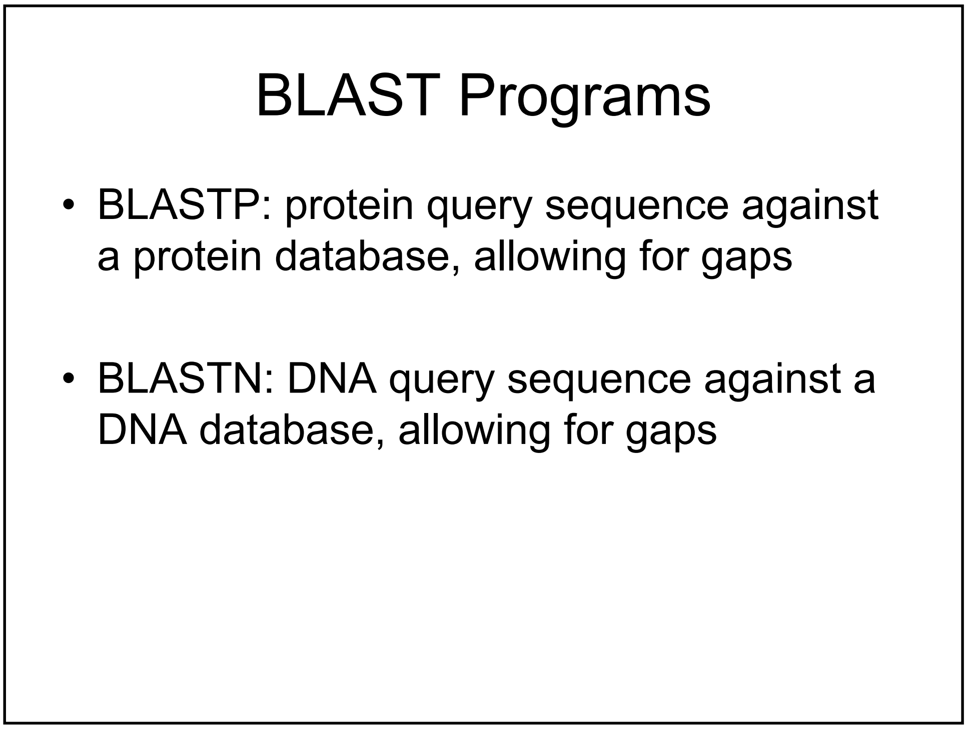
- LALIGN – FASTA, reporting multiple aligning regions
- PLALIGN – dot plot algorithm available through the fasta suite
-
- FAST-pat, FAST-swap: compares a sequence to a pattern database

BLAST Algorithm

- Filter out low complexity regions
- Locate k-tuples (words) in the query sequence
  - Word length 3 for amino acids
  - Word length 11 for nucleotides

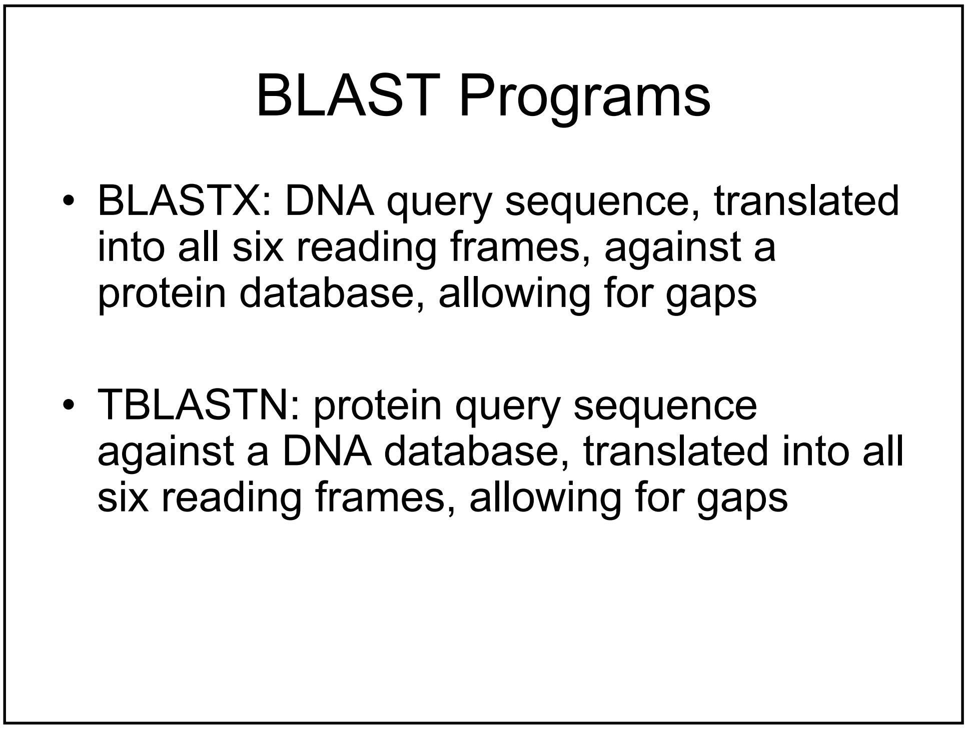
BLAST Options

- http://blast.wustl.edu/blast/README.html



BLAST Programs

- BLASTP: protein query sequence against a protein database, allowing for gaps
- BLASTN: DNA query sequence against a DNA database, allowing for gaps



BLAST Programs

- BLASTX: DNA query sequence, translated into all six reading frames, against a protein database, allowing for gaps
- TBLASTN: protein query sequence against a DNA database, translated into all six reading frames, allowing for gaps

BLAST Programs

- TBLASTX: DNA query sequence, translated into all six reading frames, against a DNA database, translated into all six reading frames (No gaps allowed)

PSI-BLAST

- (position specific iterated blast)
- take in an initial query sequence and find similar sequences to the query
- multiply align to create a scoring matrix
- search the database for more matches

PSI-BLAST

- more sequences are found that can then be added onto the multiple alignment
- caution should be used with PSI-BLAST:
  - a greedy algorithm is used
  - most recently added sequences will influence the next round of sequences

PHI-BLAST

- (pattern hit initiated blast)
- functions in same manner as PSI-BLAST except that the query sequence is first searched for a regular expression
- search for similar sequences is focused on regions containing the pattern

SSAHA

- Sequence Search and Alignment by Hashing Algorithm
- aligns DNA sequences by converting the sequence information into a ‘hash table’ data structure
- word length 10 bases by default

SSAHA

- locating identical or near identical matches
  - SNP detection
  - rapid sequence assembly
  - detecting order and orientation of contigs

SSEARCH

- SSEARCH implements the Smith-Waterman approach to sequence alignment
- SSEARCH is part of the FASTA suite
- compares protein to another protein or protein database (or DNA to DNA sequence or database) using enhanced Smith-Waterman local sequence alignments

BLAT

- (BLAST-Like Alignment Tool)
  - Jim Kent at UCSC
- locate smaller regions of higher identity within genomic assemblies
  - nucleic acids: regions at least 95% similar consisting of 40 bases or more
  - amino acids: sequences at least 80% similar consisting of at least 20 amino acids

BLAT

- Keeps index of entire genome in memory
  - – Non-overlapping k-mers
  - – 1 GB for DNA (11 base k-mers)
  - – 2 GB for amino acids (4-mers)

  - – K-mers in repetitive regions not used

BLAT

- fast tool for localizing highly similar regions

- distant homologies are not detected

- typical use: localize a specific sequence on a genome
  - – BLAT web interface directly ties to the UCSC GoldenPath genomic browser

BLAT

- WEB SERVER:

- http://genome.ucsc.edu/cgi-bin/hgBlat?command=start&org=human

Protein Databases

INTERACT

BIND

Computational sequence analysis

Query secondary databases over the Internet

PRINTS

Protein Fingerprint Database

Pfam

prosite

InterPro

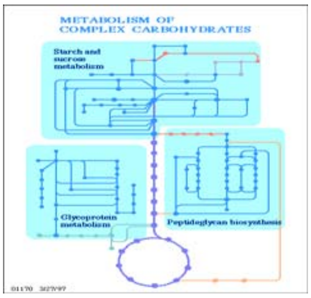
Proteins: Prediction of biochemical function

•Relationships between

```
CGCCAGCTGGACGGGCAC
ACCATGAGGCTGCTGACC
CTCCTGGGCCTTCTG…
```

↓

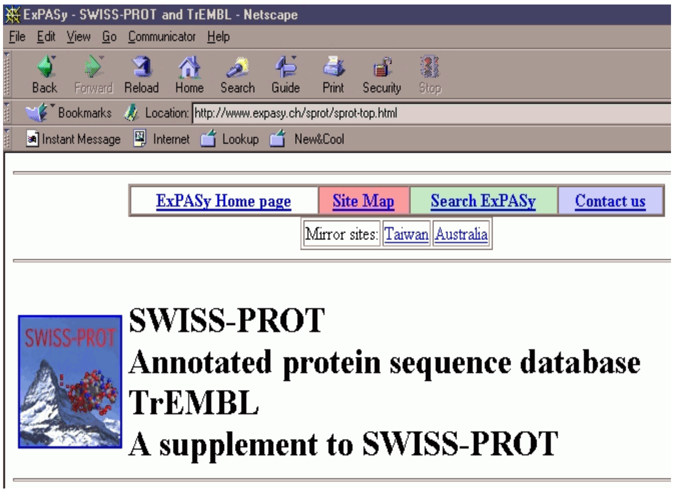
```
TDQAAFDTNIVTLTRFVM
EQGRKARGTGEMTQLLNS
LCTAVKAISTAVRKAGIA
HLYGIAGSTNVTGDQVKK
LDVLSNDLVINVLKSSFA
TCVLVTEEDKNAIIVEPE
KRGKYVVCFDPLDGSSNI
DCLVSIGTIFGIYRKNST
DEPSEKDALQPGRNLVAA
GYALYGSATML
```


METABOLISM OF COMPLEX CARBOHYDRATES
Starch and sucrose metabolism
Glycoprotein metabolism
Peptidoglycan biosynthesis
01170 3/27/97

DNA or amino acid sequence 3D structure protein functions

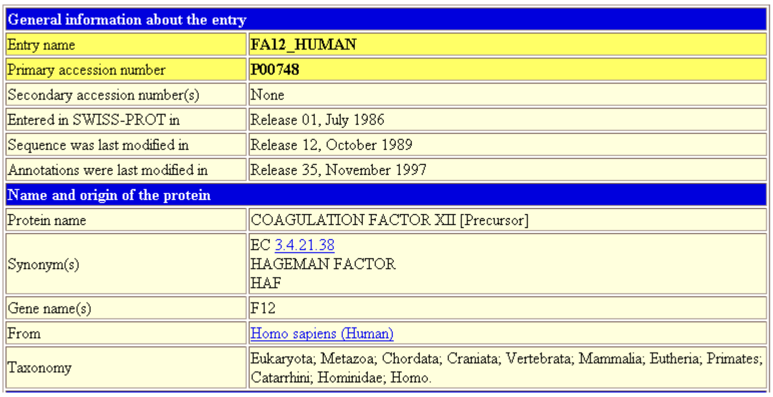
•Use of this knowledge for prediction of function, molecular modelling, and design (e.g., new therapies)

Swiss-Prot


ExPASy - SWISS-PROT and TrEMBL - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location: http://www.expasy.ch/sprot/sprot-top.html
Instant Message Internet Lookup New&Cool
ExPASy Home page Site Map Search ExPASy Contact us
Mirror sites: Taiwan Australia
SWISS-PROT

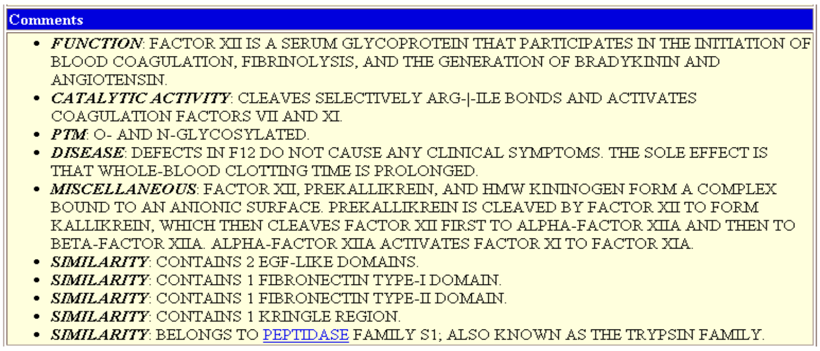
SWISS-PROT
Annotated protein sequence database
TrEMBL
A supplement to SWISS-PROT

SWISS-PROT file format



General information about the entry	
Entry name	**FA12_HUMAN**
Primary accession number	**P00748**
Secondary accession number(s)	None
Entered in SWISS-PROT in	Release 01, July 1986
Sequence was last modified in	Release 12, October 1989
Annotations were last modified in	Release 35, November 1997
Name and origin of the protein	
Protein name	COAGULATION FACTOR XII [Precursor]
Synonym(s)	EC 3.4.21.38 HAGEMAN FACTOR HAF
Gene name(s)	F12
From	Homo sapiens (Human)
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

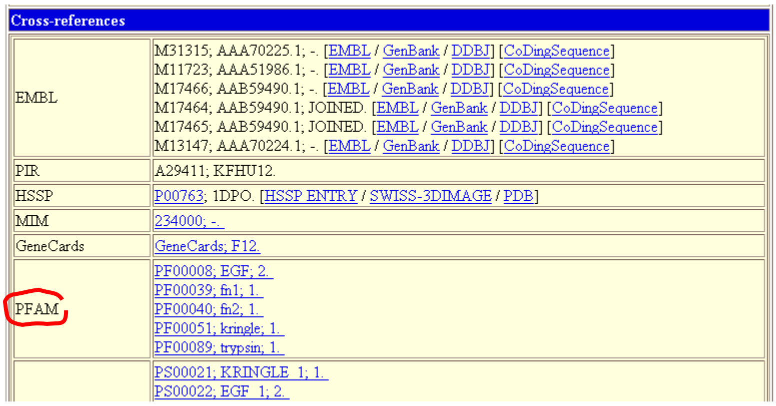
SWISS-PROT file format



Comments

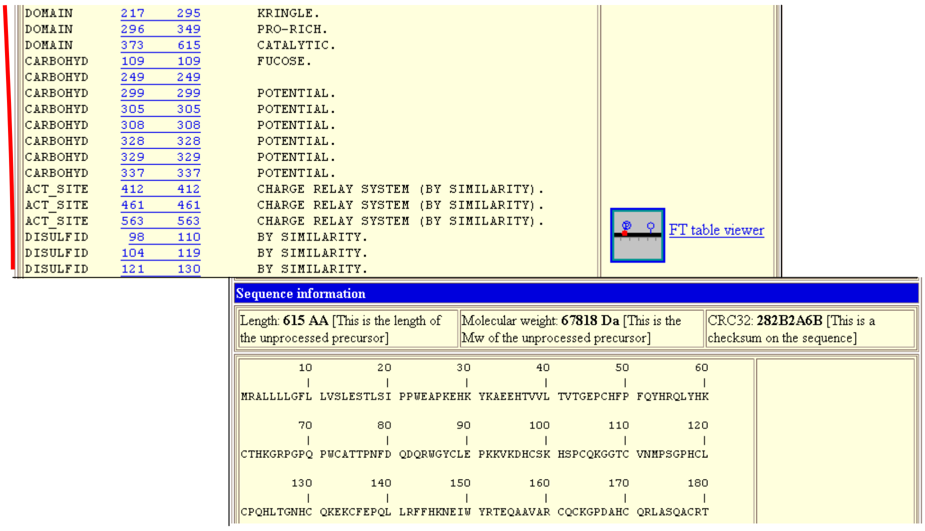
- ***FUNCTION***: FACTOR XII IS A SERUM GLYCOPROTEIN THAT PARTICIPATES IN THE INITIATION OF BLOOD COAGULATION, FIBRINOLYSIS, AND THE GENERATION OF BRADYKININ AND ANGIOTENSIN.
- ***CATALYTIC ACTIVITY***: CLEAVES SELECTIVELY ARG-|-ILE BONDS AND ACTIVATES COAGULATION FACTORS VII AND XI.
- ***PTM***: O- AND N-GLYCOSYLATED.
- ***DISEASE***: DEFECTS IN F12 DO NOT CAUSE ANY CLINICAL SYMPTOMS. THE SOLE EFFECT IS THAT WHOLE-BLOOD CLOTTING TIME IS PROLONGED.
- ***MISCELLANEOUS***: FACTOR XII, PREKALLIKREIN, AND HMW KININOGEN FORM A COMPLEX BOUND TO AN ANIONIC SURFACE. PREKALLIKREIN IS CLEAVED BY FACTOR XII TO FORM KALLIKREIN, WHICH THEN CLEAVES FACTOR XII FIRST TO ALPHA-FACTOR XIIA AND THEN TO BETA-FACTOR XIIA. ALPHA-FACTOR XIIA ACTIVATES FACTOR XI TO FACTOR XIA.
- ***SIMILARITY***: CONTAINS 2 EGF-LIKE DOMAINS.
- ***SIMILARITY***: CONTAINS 1 FIBRONECTIN TYPE-I DOMAIN.
- ***SIMILARITY***: CONTAINS 1 FIBRONECTIN TYPE-II DOMAIN.
- ***SIMILARITY***: CONTAINS 1 KRINGLE REGION.
- ***SIMILARITY***: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE TRYPSIN FAMILY.

SWISS-PROT file format



Cross-references	
EMBL	M31315; AAA70225.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M11723; AAA51986.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M17466; AAB59490.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M17464; AAB59490.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] M17465; AAB59490.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] M13147; AAA70224.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A29411; KFHU12.
HSSP	P00763; 1DPO. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
MIM	234000; -.
GeneCards	GeneCards; F12.
PFAM	PF00008; EGF; 2. PF00039; fn1; 1. PF00040; fn2; 1. PF00051; kringle; 1. PF00089; trypsin; 1.
	PS00021; KRINGLE_1; 1. PS00022; EGF_1; 2.

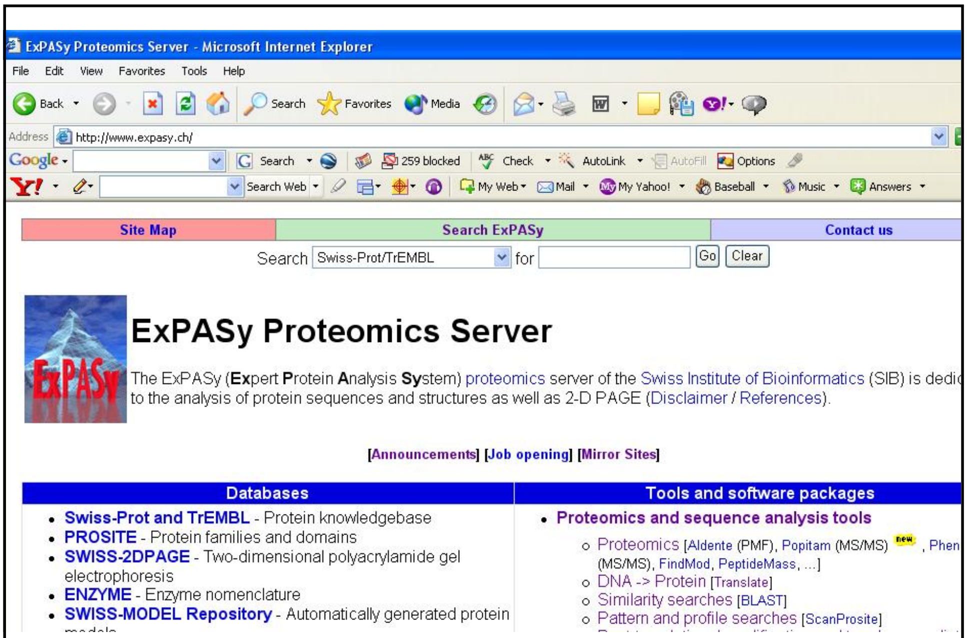
SWISS-PROT file format



```
DOMAIN      217    295       KRINGLE.
DOMAIN      296    349       PRO-RICH.
DOMAIN      373    615       CATALYTIC.
CARBOHYD    109    109       FUCOSE.
CARBOHYD    249    249
CARBOHYD    299    299       POTENTIAL.
CARBOHYD    305    305       POTENTIAL.
CARBOHYD    308    308       POTENTIAL.
CARBOHYD    328    328       POTENTIAL.
CARBOHYD    329    329       POTENTIAL.
CARBOHYD    337    337       POTENTIAL.
ACT_SITE    412    412       CHARGE RELAY SYSTEM (BY SIMILARITY).
ACT_SITE    461    461       CHARGE RELAY SYSTEM (BY SIMILARITY).
ACT_SITE    563    563       CHARGE RELAY SYSTEM (BY SIMILARITY).
DISULFID     98    110       BY SIMILARITY.
DISULFID    104    119       BY SIMILARITY.
DISULFID    121    130       BY SIMILARITY.
```

Sequence information		
Length: **615 AA** [This is the length of the unprocessed precursor]	Molecular weight: **67818 Da** [This is the Mw of the unprocessed precursor]	CRC32: **282B2A6B** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MRALLLLGFL LVSLESTLSI PPWEAPKEHK YKAEEHTVVL TVTGEPCHFP FQYHRQLYHK

        70         80         90        100        110        120
         |          |          |          |          |          |
CTHKGRPGPQ PWCATTPNFD QDQRWGYCLE PKKVKDHCSK HSPCQKGGTC VNMPSGPHCL

       130        140        150        160        170        180
         |          |          |          |          |          |
CPQHLTGNHC QKEKCFEPQL LRFFHKNEIW YRTEQAAVAR CQCKGPDAHC QRLASQACRT
```


ExPASy Proteomics Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.expasy.ch/
Google Search 259 blocked Check AutoLink AutoFill Options
Y! Search Web My Web Mail My Yahoo! Baseball Music Answers
Site Map
Search ExPASy
Contact us
Search Swiss-Prot/TrEMBL for Go Clear
ExPASy

ExPASy Proteomics Server

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedic
to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References).

[Announcements] [Job opening] [Mirror Sites]

Databases

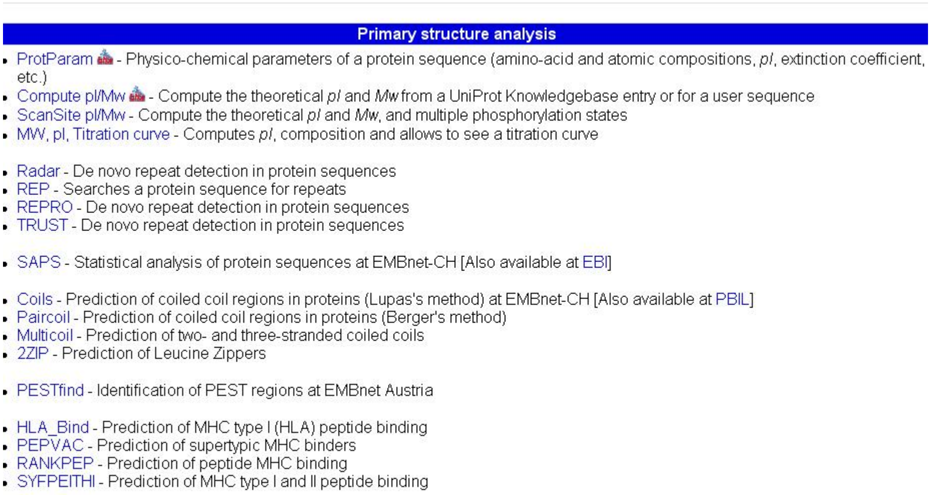
- **Swiss-Prot and TrEMBL** - Protein knowledgebase
- **PROSITE** - Protein families and domains
- **SWISS-2DPAGE** - Two-dimensional polyacrylamide gel electrophoresis
- **ENZYME** - Enzyme nomenclature
- **SWISS-MODEL Repository** - Automatically generated protein

Tools and software packages

- **Proteomics and sequence analysis tools**
  - Proteomics [Aldente (PMF), Popitam (MS/MS) new , Phen (MS/MS), FindMod, PeptideMass, ...]
  - DNA -> Protein [Translate]
  - Similarity searches [BLAST]
  - Pattern and profile searches [ScanProsite]

http://www.expasy.ch/

http://www.expasy.ch/



Primary structure analysis

- ProtParam - Physico-chemical parameters of a protein sequence (amino-acid and atomic compositions, *pI*, extinction coefficient, etc.)
- Compute pI/Mw - Compute the theoretical *pI* and *Mw* from a UniProt Knowledgebase entry or for a user sequence
- ScanSite pI/Mw - Compute the theoretical *pI* and *Mw*, and multiple phosphorylation states
- MW, pI, Titration curve - Computes *pI*, composition and allows to see a titration curve

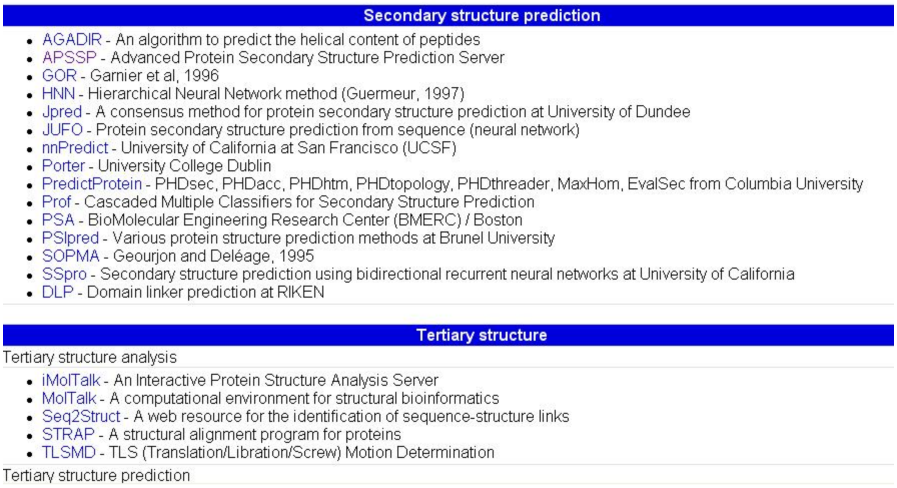
- Radar - De novo repeat detection in protein sequences
- REP - Searches a protein sequence for repeats
- REPRO - De novo repeat detection in protein sequences
- TRUST - De novo repeat detection in protein sequences

- SAPS - Statistical analysis of protein sequences at EMBnet-CH [Also available at EBI]

- Coils - Prediction of coiled coil regions in proteins (Lupas's method) at EMBnet-CH [Also available at PBIL]
- Paircoil - Prediction of coiled coil regions in proteins (Berger's method)
- Multicoil - Prediction of two- and three-stranded coiled coils
- 2ZIP - Prediction of Leucine Zippers

- PESTfind - Identification of PEST regions at EMBnet Austria

- HLA_Bind - Prediction of MHC type I (HLA) peptide binding
- PEPVAC - Prediction of supertypic MHC binders
- RANKPEP - Prediction of peptide MHC binding
- SYFPEITHI - Prediction of MHC type I and II peptide binding



Secondary structure prediction

- AGADIR - An algorithm to predict the helical content of peptides
- APSSP - Advanced Protein Secondary Structure Prediction Server
- GOR - Garnier et al, 1996
- HNN - Hierarchical Neural Network method (Guermeur, 1997)
- Jpred - A consensus method for protein secondary structure prediction at University of Dundee
- JUFO - Protein secondary structure prediction from sequence (neural network)
- nnPredict - University of California at San Francisco (UCSF)
- Porter - University College Dublin
- PredictProtein - PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec from Columbia University
- Prof - Cascaded Multiple Classifiers for Secondary Structure Prediction
- PSA - BioMolecular Engineering Research Center (BMERC) / Boston
- PSIpred - Various protein structure prediction methods at Brunel University
- SOPMA - Geourjon and Deléage, 1995
- SSpro - Secondary structure prediction using bidirectional recurrent neural networks at University of California
- DLP - Domain linker prediction at RIKEN

Tertiary structure

Tertiary structure analysis

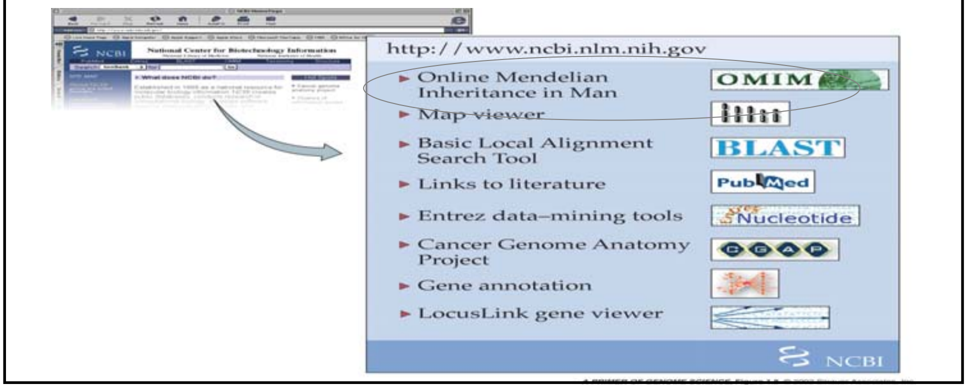
- iMolTalk - An Interactive Protein Structure Analysis Server
- MolTalk - A computational environment for structural bioinformatics
- Seq2Struct - A web resource for the identification of sequence-structure links
- STRAP - A structural alignment program for proteins
- TLSMD - TLS (Translation/Libration/Screw) Motion Determination

Tertiary structure prediction

http://www.expasy.ch/

What else?

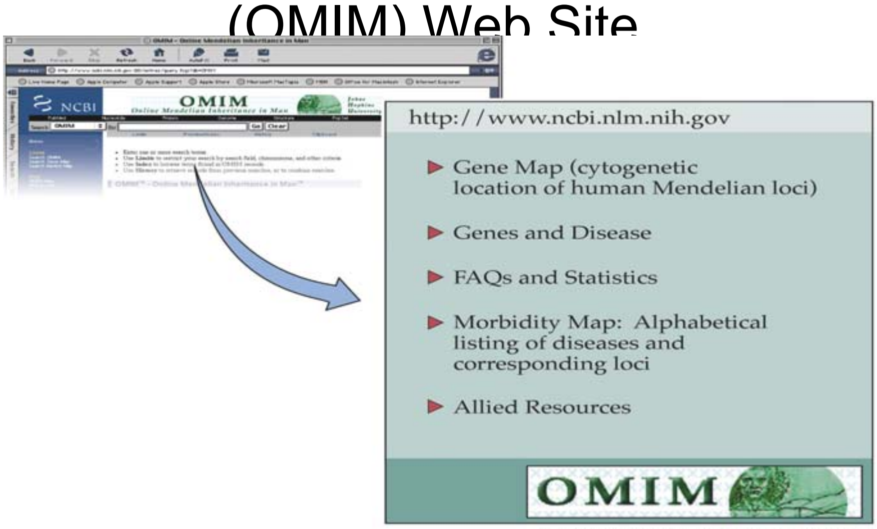
Computer scientists are responsible for developing tools for performing various operations, such as BLAST at NCBI



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

- ► Online Mendelian Inheritance in Man — OMIM
- ► Map viewer
- ► Basic Local Alignment Search Tool — BLAST
- ► Links to literature — PubMed
- ► Entrez data–mining tools — Nucleotide
- ► Cancer Genome Anatomy Project — CGAP
- ► Gene annotation
- ► LocusLink gene viewer

Resources Available through the Online Mendelian Inheritance in Man (OMIM) Web Site

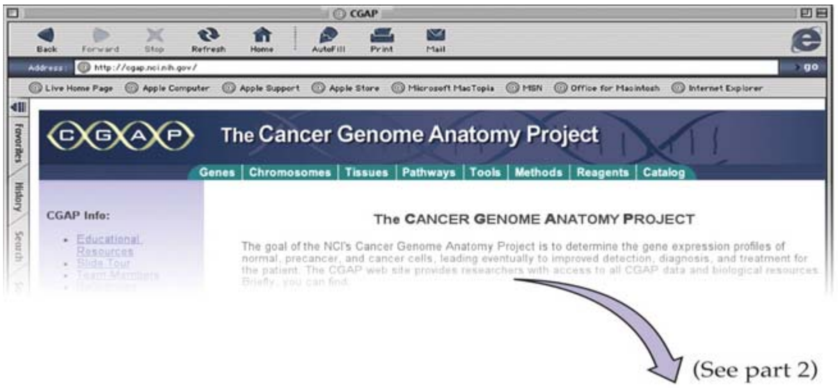


http://www.ncbi.nlm.nih.gov

- Gene Map (cytogenetic location of human Mendelian loci)
- Genes and Disease
- FAQs and Statistics
- Morbidity Map: Alphabetical listing of diseases and corresponding loci
- Allied Resources

A PRIMER OF GENOME SCIENCE, Figure 1.9 © 2002 Sinauer Associates, Inc.

Resources Available through the Cancer Genome Anatomy Project (CGAP) Web Site



http://cgap.nci.nih.gov/

The CANCER GENOME ANATOMY PROJECT

The goal of the NCI's Cancer Genome Anatomy Project is to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient. The CGAP web site provides researchers with access to all CGAP data and biological resources. Briefly, you can find:

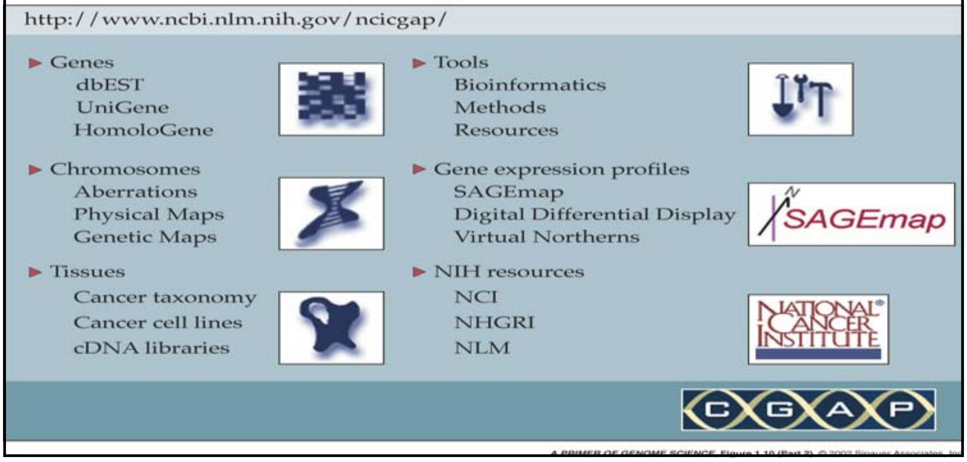
CGAP Info:

- Educational Resources
- Slide Tour
- Team Members

(See part 2)

A PRIMER OF GENOME SCIENCE, Figure 1.10 (Part 1) © 2002 Sinauer Associates, Inc.

Resources Available through the Cancer Genome Anatomy Project (CGAP) Web Site

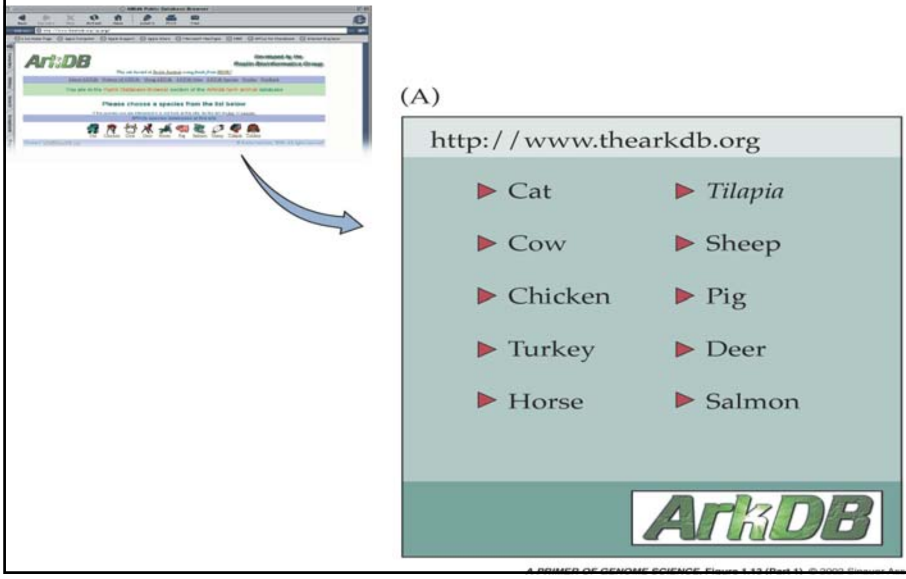


http://www.ncbi.nlm.nih.gov/ncicgap/

- Genes
  - dbEST
  - UniGene
  - HomoloGene
- Chromosomes
  - Aberrations
  - Physical Maps
  - Genetic Maps
- Tissues
  - Cancer taxonomy
  - Cancer cell lines
  - cDNA libraries
- Tools
  - Bioinformatics
  - Methods
  - Resources
- Gene expression profiles
  - SAGEmap
  - Digital Differential Display
  - Virtual Northerns
- NIH resources
  - NCI
  - NHGRI
  - NLM

A PRIMER OF GENOME SCIENCE, Figure 1.10 (Part 2) © 2002 Sinauer Associates, Inc.

Some Online Animal Genome Sources (Part 1)



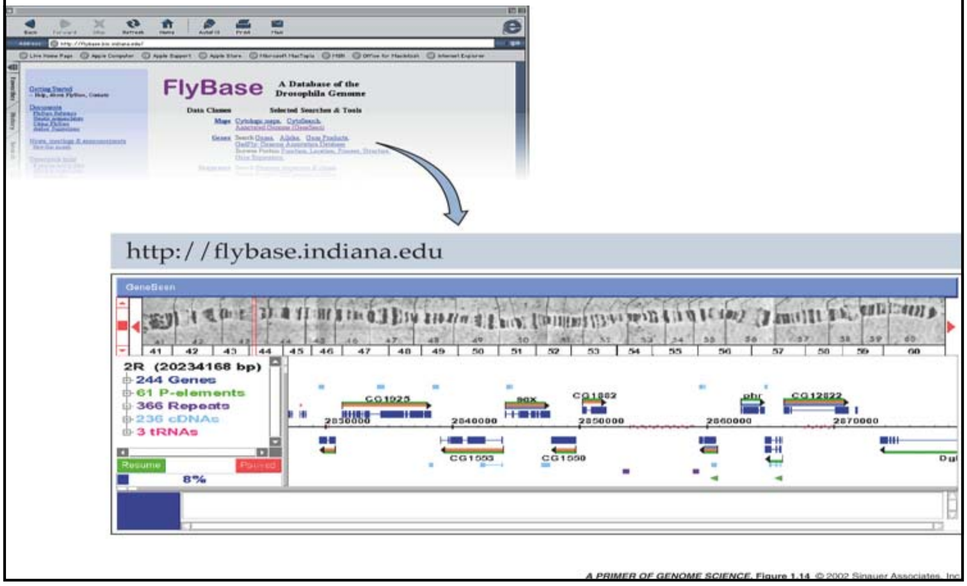
(A)

http://www.thearkdb.org

- Cat
- Cow
- Chicken
- Turkey
- Horse
- *Tilapia*
- Sheep
- Pig
- Deer
- Salmon

A PRIMER OF GENOME SCIENCE, Figure 1.12 (Part 1) © 2002 Sinauer Associates, Inc.

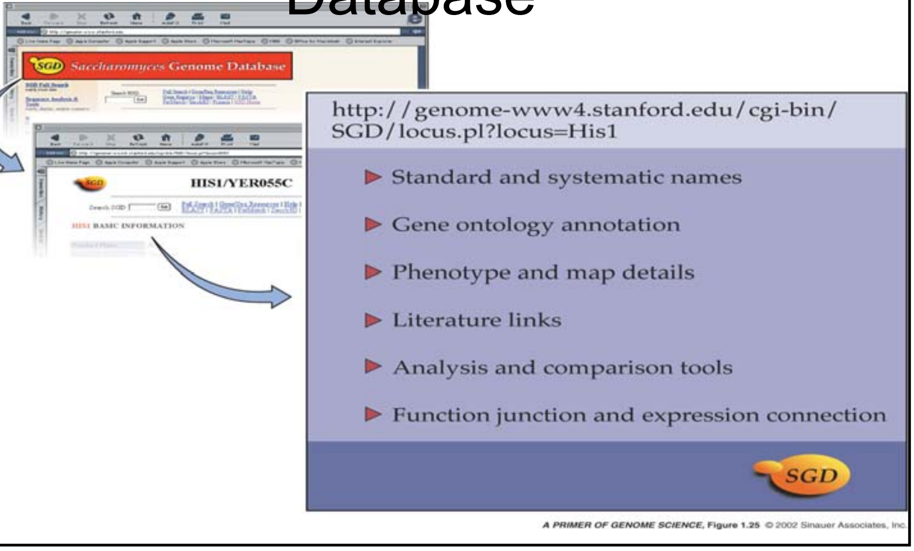
Drosophila Gene Annotation



http://flybase.indiana.edu

A PRIMER OF GENOME SCIENCE, Figure 1.14 © 2002 Sinauer Associates, Inc.

Annotation of Genes on the *Saccharomyces* Genome Database



http://genome-www4.stanford.edu/cgi-bin/SGD/locus.pl?locus=His1

- ▶ Standard and systematic names
- ▶ Gene ontology annotation
- ▶ Phenotype and map details
- ▶ Literature links
- ▶ Analysis and comparison tools
- ▶ Function junction and expression connection

A PRIMER OF GENOME SCIENCE, Figure 1.25 © 2002 Sinauer Associates, Inc.

Thank you for your listening