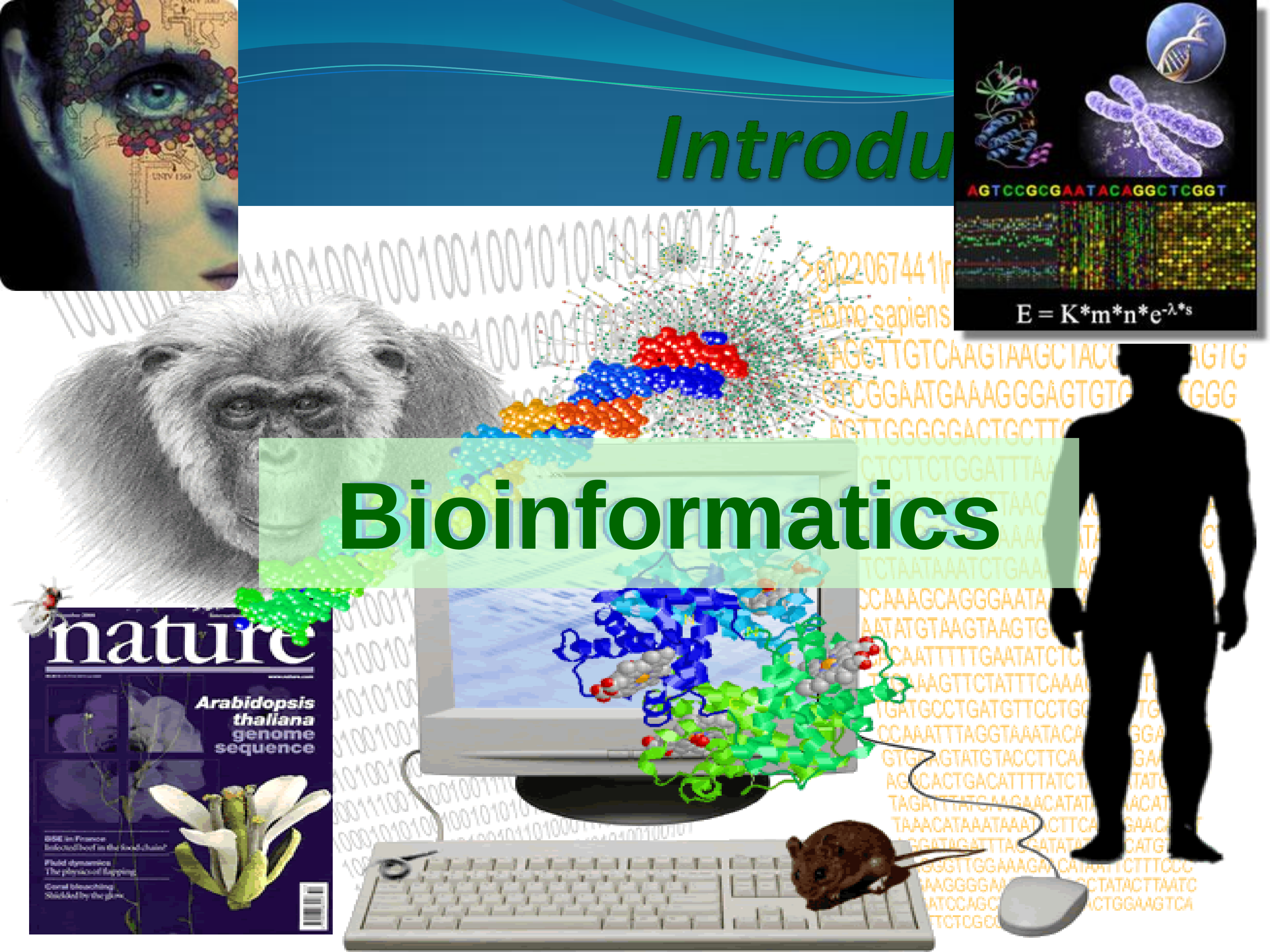
Introdu



Bioinformatics

Introduction to Bioinformatics

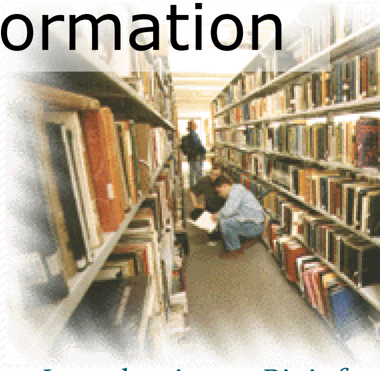
Dr. Taysir Hassan A. Soliman
Associate Professor,
Head of Information System Department,
Faculty of Computers and Information
Assiut University, Egypt
taysirhs2@gmail.com

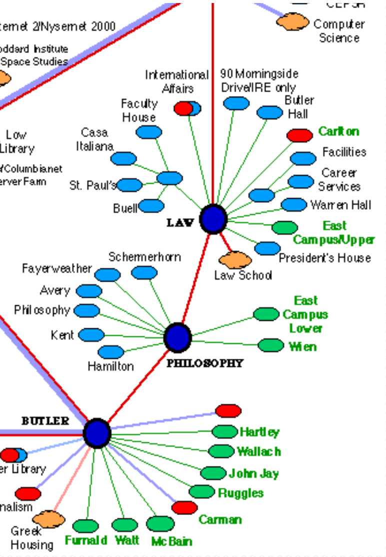
Outline

- Introduction to Bioinformatics
- Bioinformatics Applications
- Bioinformatics databases
- Sequence Alignment

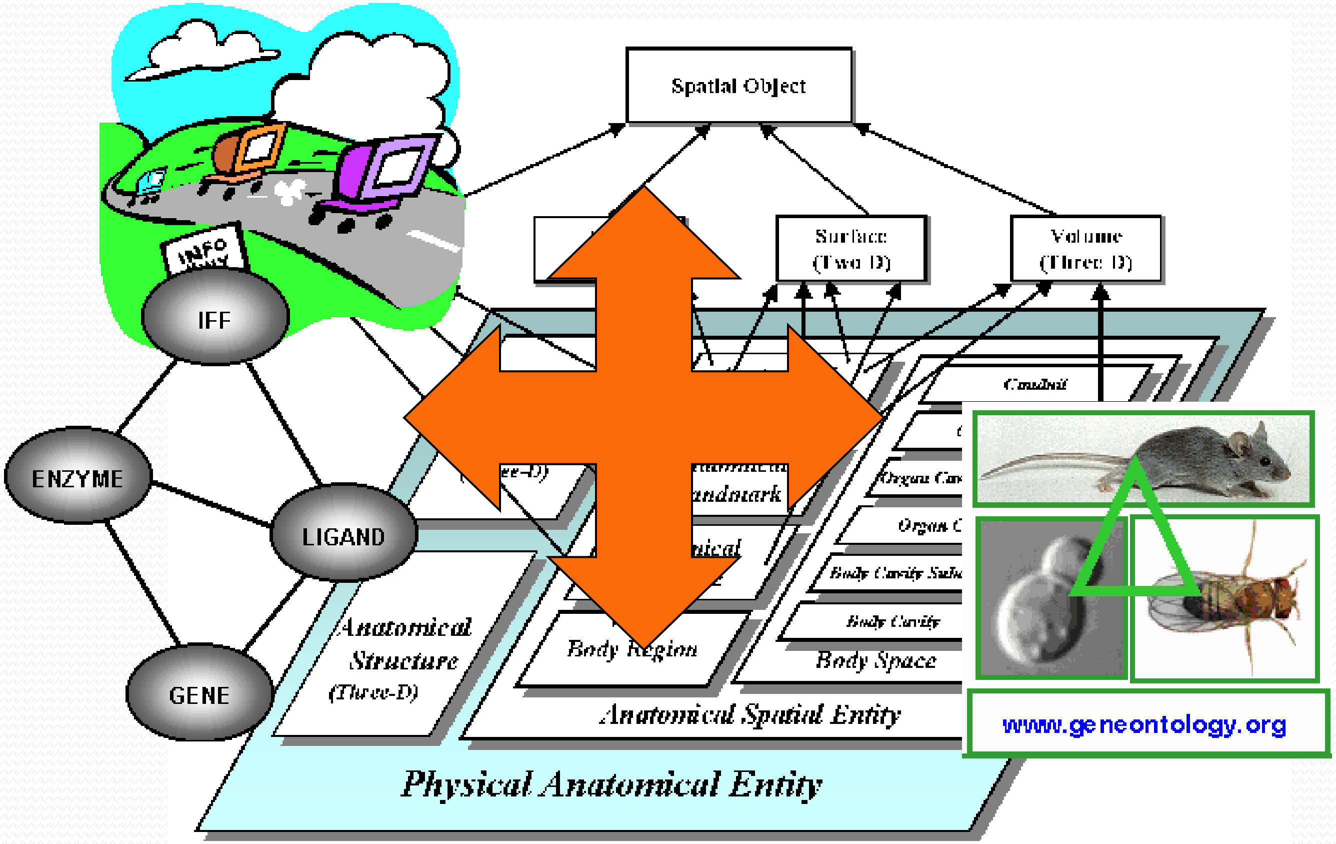
Science then, then and now

A vast amount, rapidly generated related but highly distributed and semantically unconnected information





Science then, then and now



How much Computing skills?

- Bioinformatics can be seen as a tool that the biologist needs to use - like PCR
- Or should biologists be able to write their own programs and build databases?
  - it is a big advantage to be able to design exactly the tool that you want
  - this may be the wave of the future

  "Two months in the lab can easily save an afternoon on the computer."

  —Alan Bleasby, 1997
- Q: Is this school going to train "bioinformatics professionals" or biologists with informatics skills?
- A: Both!

What is Bioinformatics?

- The use of computers to collect, analyze, and interpret biological information at the molecular level.
- A set of software tools for molecular sequence analysis

YES

- DNA & protein sequence databases
- Sequence similarity, alignment, & assembly
- Sequence patterns/motifs
- Phylogenetics
- Microarray gene expression data
- Protein structure prediction
- Mapping metabolic and regulatory pathways (graph theory)

NO

- patient medical charts, billing, hospital payroll, etc.
- X-ray image analysis

MAYBE

- Ontologies

  (biological function, research methods, clinical terminology, etc.)

Bioinformatics - origins

- Driven by experimental molecular biology
  - lab folks generate the data, then need a way to organize and analyze it
- Grabs methods from many different fields
  - biostatistics, machine learning, data mining, linguistics, etc
- Use computer (algorithms) to gain novel biological knowledge.
- Use biological knowledge to construct algorithms.

The Biologist in the Age of Information

Training "computer savvy" scientists

- Know the right tool for the job
- Get the job done with tools available
- Network connection is the lifeline of the scientist
- Jobs change, computers change, projects change, scientists need to be adaptable

The job of the biologist is changing

- As more biological information becomes available …
  - The biologist will spend more time using computers, building and mining databases
  - The biologist will spend more time on data analysis (and less doing lab biochemistry)
  - Biology will become a more quantitative science (think how the periodic table and atomic theory affected chemistry)

Biological Data Characteristics

1. Huge data
2. Heterogeneous distributed data
3. Frequently updated data
4. Defining and representing complex queries are extremely important to the biologist
5. Most biologist will not care or know about the data structure or the schema design
6. Users of biological information often require access to previous versions of existing data.

I. “Traditional” bioinformatics methods

- Conduct online literature and similarity searches (NCBI Entrez and Blast)
- Use desktop sequence analysis tools
  - restriction digest, PCR primer design, ORF finding
- Assembly of automated sequencing reads

II. More advanced stuff

- Multiple alignment
- Phylogenetic trees
- Motif/domain analysis of proteins

  (Pfam, Blocks, ProDom)
- Motif/domain analysis of DNA

  (promoters, transcription factors, intron splice sites)
- Genefinding in genome data

  combining data from ORFs, promoters, and cDNA homology

III. Genome scale data analysis

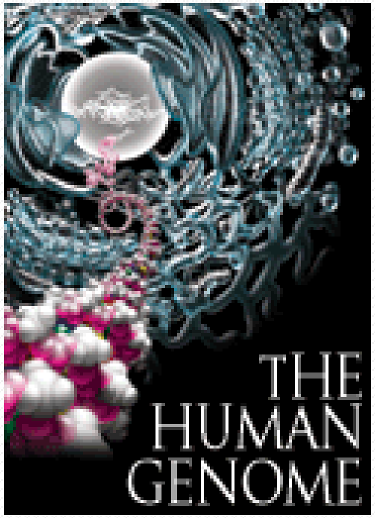
- Handling large amounts of data
  - Create an experiment or lab database
  - use traditional bioinformatics tools on different data
  - scripting languages (simple programming tools, Perl)
- Microarray gene expression analysis
  - differential expression and classification/prediction
  - clustering, principle components
  - functional genomics - pathways, ontology classification
- Genome-wide SNP or genome tiling analysis

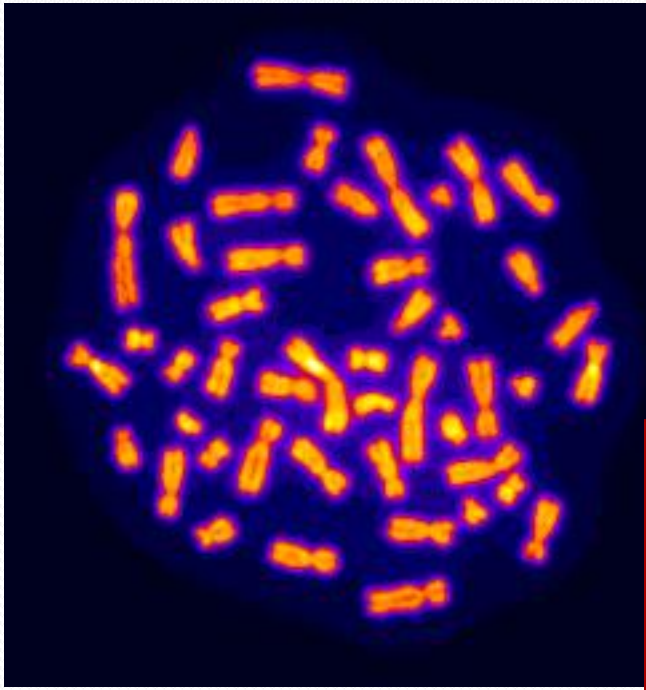
A Genome Revolution in Biology and Medicine

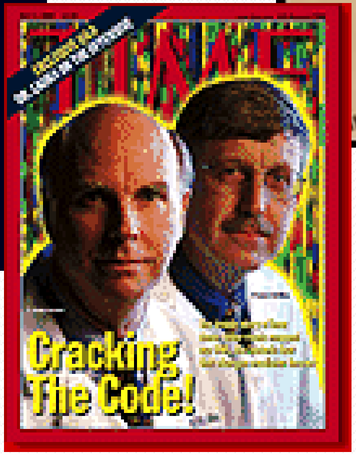
- We are in the midst of a "Golden Era" of biology
- The Human Genome Project has produced a huge storehouse of data that will be used to change every aspect of biological research and medicine
- The revolution is about treating biology as an information science, not about specific biochemical technologies.

Genome Projects

The Human Genome sequence is complete approximately 3.2 billion base pairs



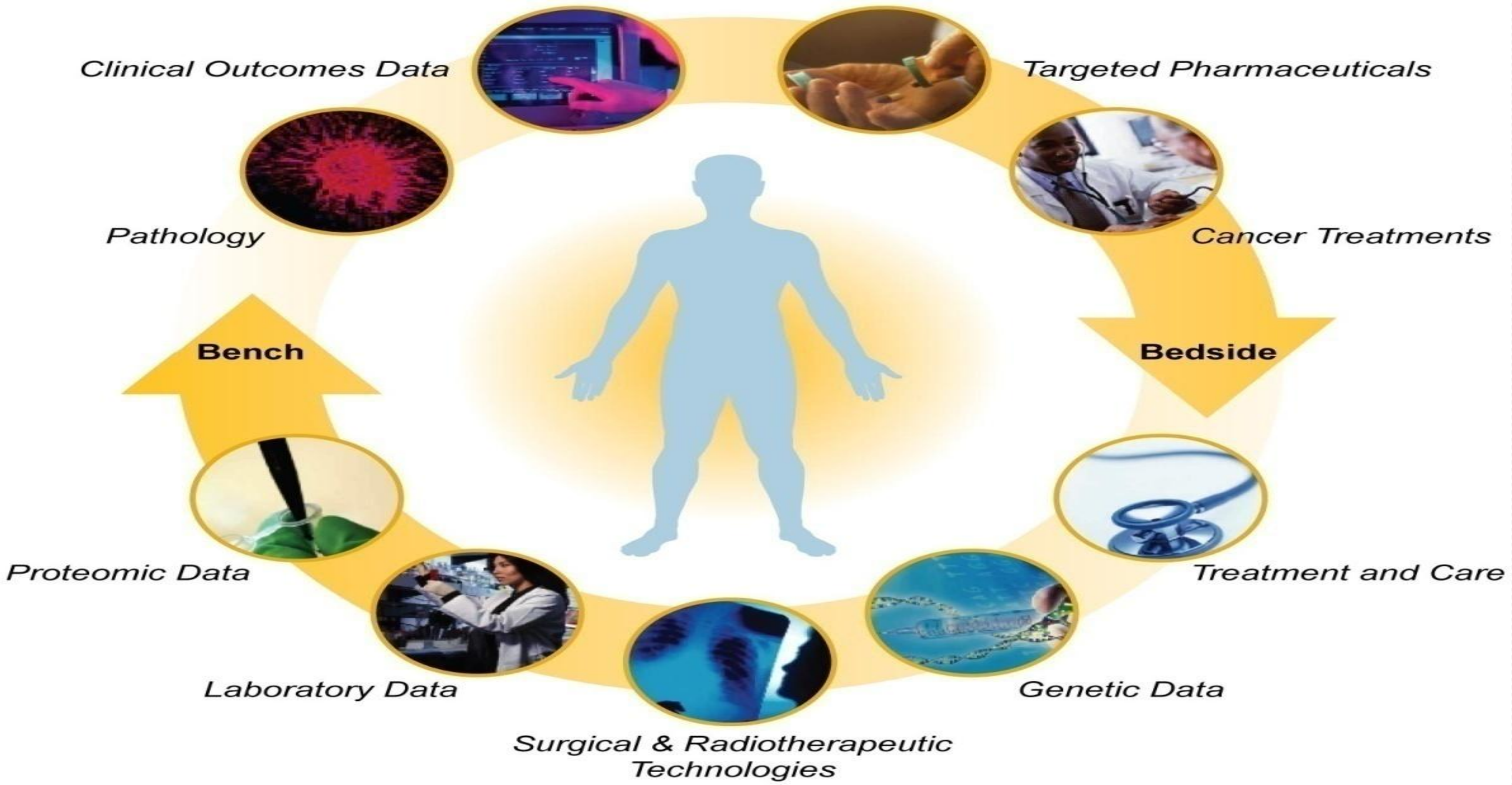


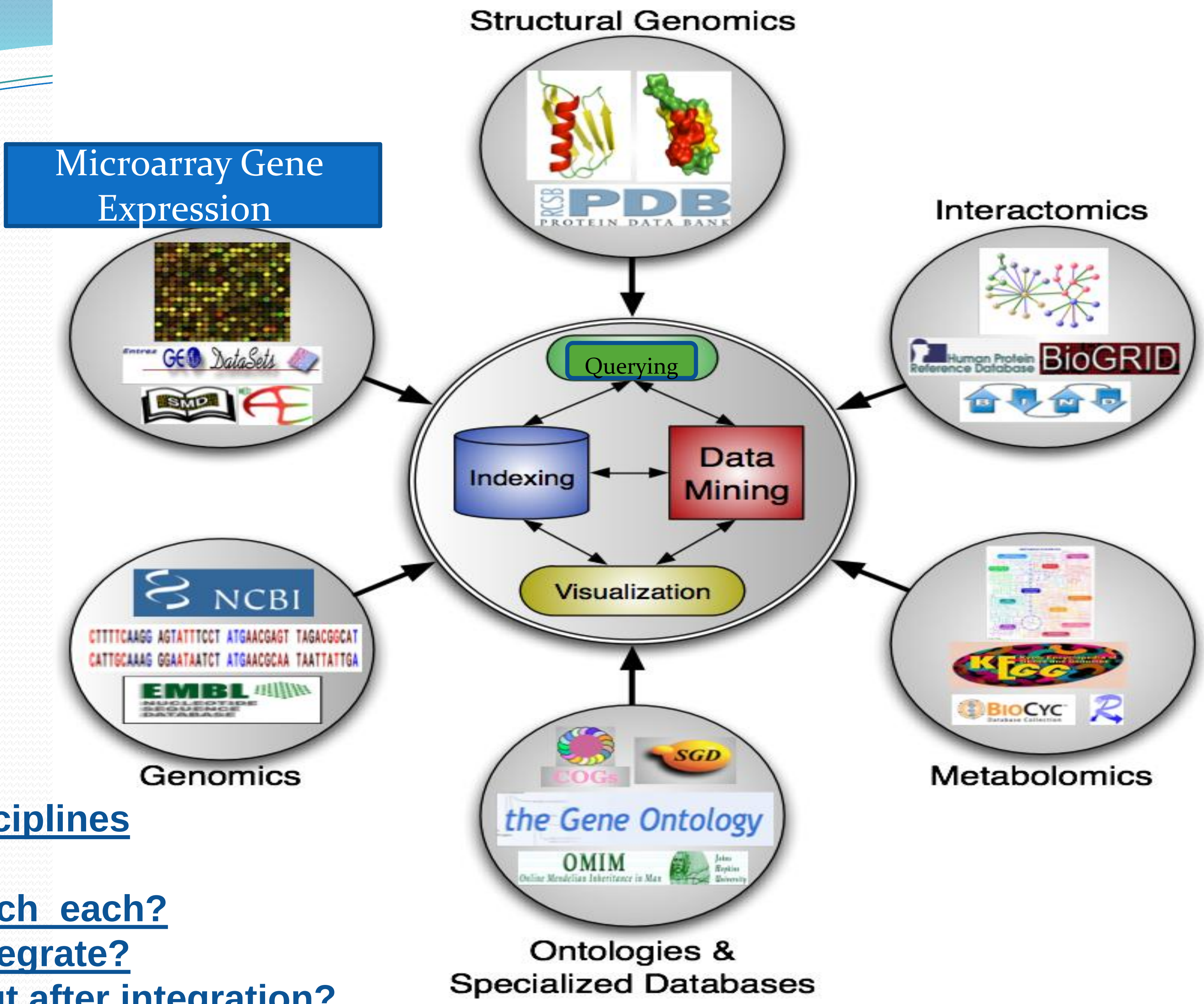


More Genomes

5 December 2002
International weekly journal of science
nature
www.nature.com/nature
Atmospheric CO2
A drop in the ocean
Drag reduction
Flexing in fluids
Regulatory T cells
Basis for persistent infection
The mouse genome
Experimental model for human biology
nature jobs
celebration of the mouse genome
Science
The Drosophila Genome
Science
4 October 2002
Vol. 298 No. 5591
Pages 1–310 $10
The Mosquito Genome
Anopheles gambiae
American Association for the Advancement of Science
14 December 2000
International weekly journal of science
nature
www.nature.com
Arabidopsis thaliana genome sequence
9 December 2004
International weekly journal
nature
£10.00
www.nature.c
The chicken genome
Cracking the code
Stem-cell research
The religious dimension
Quantum physics
Exciton times
Hair-cell trigger
The channel for sound
1 April 2004
natu
www.nature.com/nature
The rat genome
Insights into mammalian evolution
Superconductivity
Diamond springs a surprise
Office life
Make those e-mails count
SARS vaccine
Immunity induced in mice
naturejobs bioinformatics
Genome Research
International weekly journal of science
nature
www.nature.com/nature
Plasmodium genomics
Genomics and proteomics pave the way for controlling malaria
Cold antihydrogen
CERN delivers
Antarctic ice
Flow reversals
Antigen presentation
A 'customizing' protease

Bioinformatics …
A breakthrough towards … Various Applications





Structural Genomics

RCSB PDB
PROTEIN DATA BANK

Interactomics

Human Protein Reference Database
BioGRID
B I N D

Microarray Gene Expression

Entrez GEO DataSets
SMD
AE

Querying

Indexing

Data Mining

Visualization

NCBI
CTTTTCAAGG AGTATTTCCT ATGAACGAGT TAGACGGCAT
CATTGCAAAG GGAATAATCT ATGAACGCAA TAATTATTGA
EMBL NUCLEOTIDE SEQUENCE DATABASE

Genomics

Metabolomics

COGs
SGD
the Gene Ontology
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

Ontologies & Specialized Databases

Various Disciplines

But

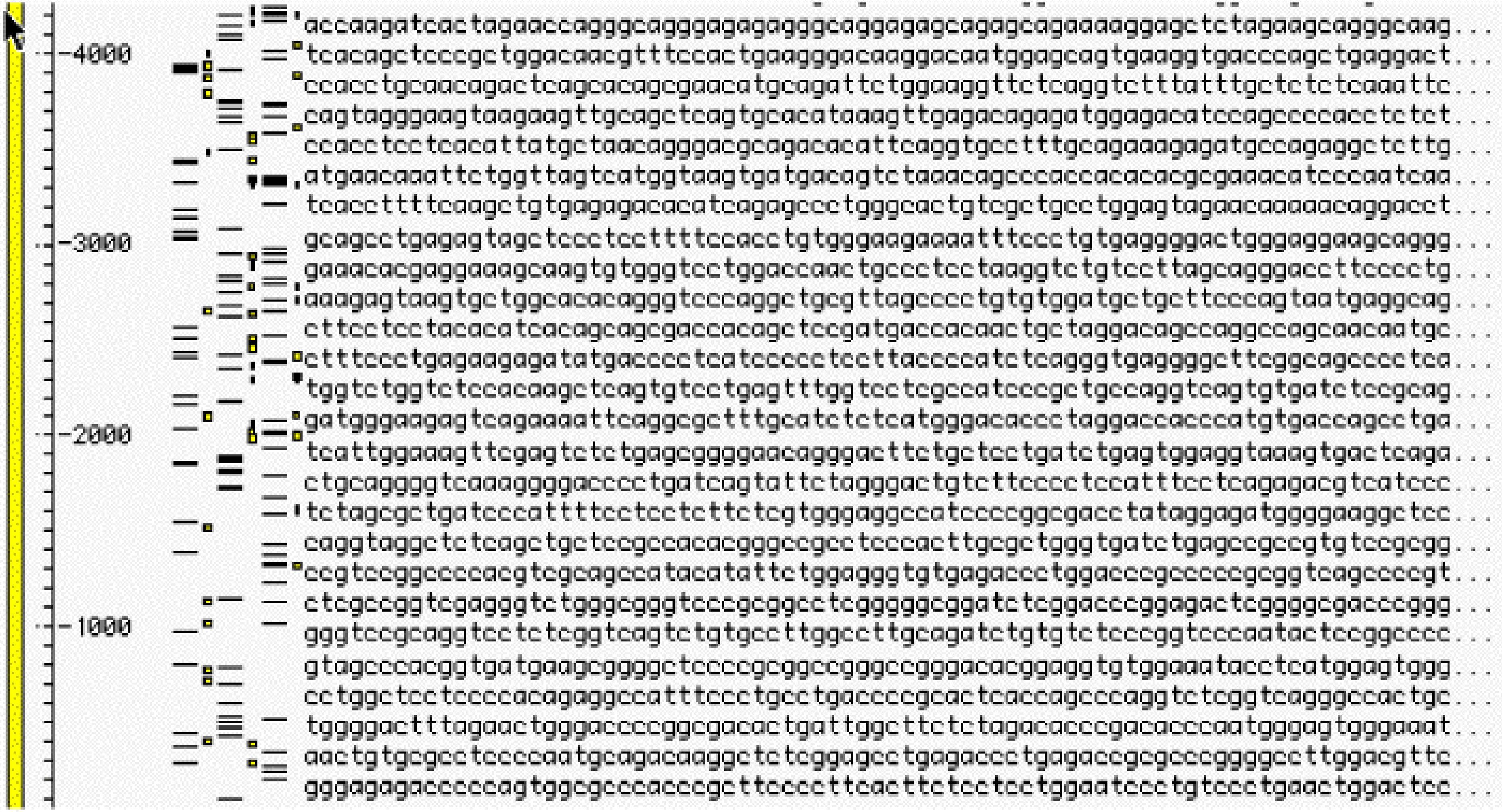
- **How to reach each?**
- **How to integrate?**
- **What about after integration?**

So,…

All the Genes

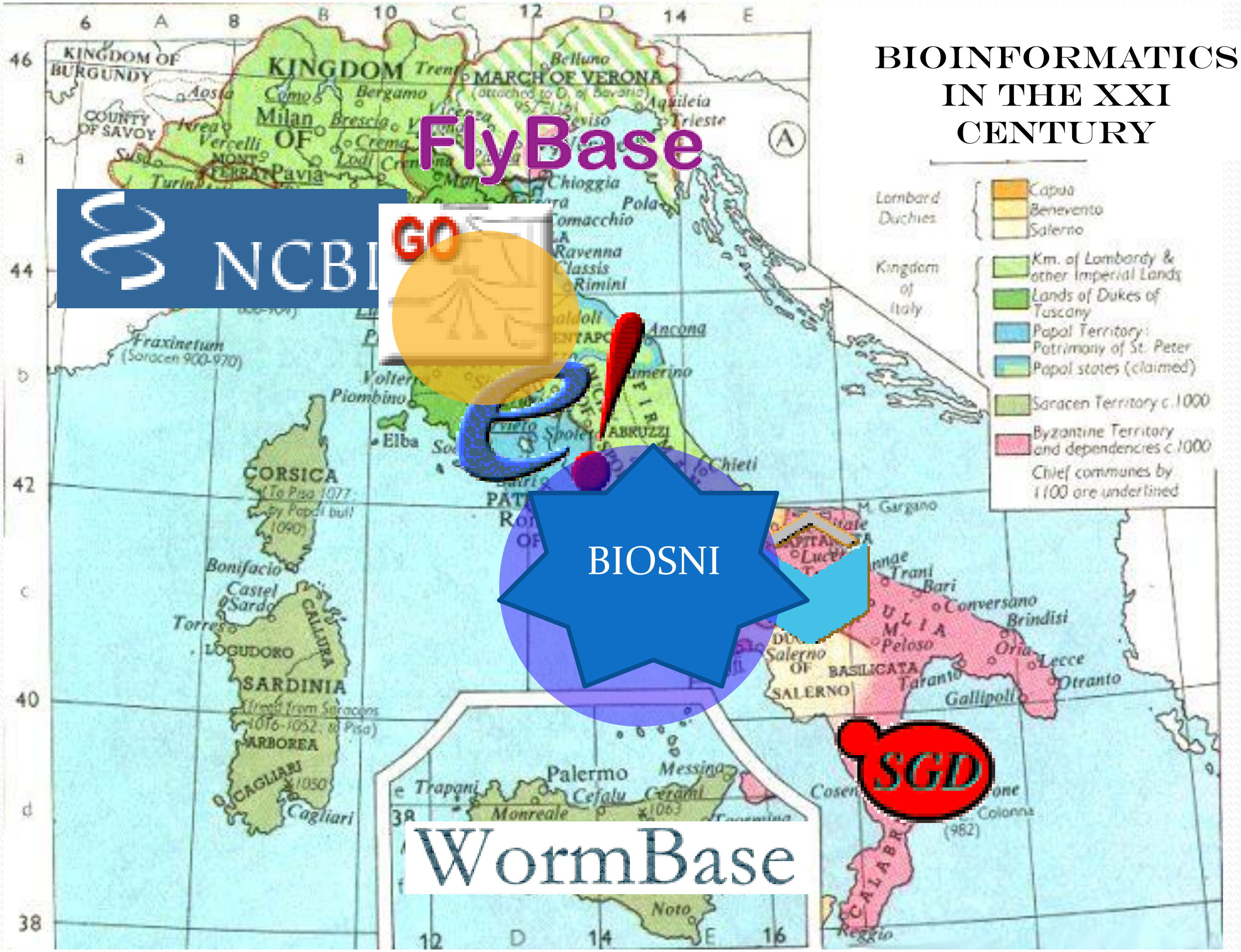
- Any human gene can now be found in the genome by similarity searching with over 99.9% certainty.
- However, the sequence still has gaps
- Still can’t identify pseudogenes, false genes with certainty
  - This will improve as more sequence data accumulates
- We are getting close to a complete list of human genes and proteins
  - Needed as a starting point for gene expression, pattern finding, and systems biology

Raw Genome Data:



```
accaagatcactagaaccagggcagggagagagggcaggagagcagagcagaaaggagctctagaagcagggcaag...
tcacagctcccgctggacaacgtttccactgaagggacaaggacaatggagcagtgaaggtgacccagctgaggact...
ccacctgcaacagactcagcacagcgaacatgcagattctggaaggttctcaggtctttatttgctctctcaaattc...
cagtagggaagtaagaagttgcagctcagtgcacataaagttgagacagagatggagacatccagccccacctctct...
ccacctcctcacattatgctaacagggacgcagacacattcaggtgcctttgcagaaagagatgccagaggctcttg...
atgaacaaattctggttagtcatggtaagtgatgacagtctaaacagcccaccacacacgcgaaacatcccaatcaa...
tcaccttttcaagctgtgagagacacatcagagccctgggcactgtcgctgcctggagtagaacaaaaacaggacct...
gcagcctgagagtagctccctccttttccacctgtgggaagaaaatttccctgtgaggggactgggaggaagcaggg...
gaaacacgaggaaagcaagtgtgggtcctggaccaactgcccctcctaaggtctgtccttagcagggaccttcccctg...
aaagagtaagtgctggcacacagggtcccaggctgcgttagcccctgtgtggatgctgcttcccagtaatgaggcag...
cttcctcctacacatcacagcagcgaccacagctccgatgaccacaactgctaggacagccaggccagcaacaatgc...
ctttccctgagaagagatatgacccctcatcccccctccttaccccatctcagggtgaggggcttcggcagcccctca...
tggtctggtctccacaagctcagtgtcctgagtttggtcctcgccatcccgctgccaggtcagtgtgatctccgcag...
gatgggaagagtcagaaaattcaggcgctttgcatctctcatgggacaccctaggaccaccccatgtgaccagcctga...
tcattggaaagttcgagtctctgagcggggaacagggacttctgctcctgatctgagtggaggtaaagtgactcaga...
ctgcaggggtcaaaggggacccctgatcagtattctagggactgtcttcccctccatttcctcagagacgtcatccc...
tctagcgctgatcccattttcctcctcttctcgtgggaggccatccccggcgacctataggagatggggaaggctcc...
caggtaggctctcagctgctccgccacacgggccgcctccccacttgcgctgggtgatctgagccgccgtgtccgcgg...
ccgtccggcccccacgtcgcagccatacatattctggagggtgtgagaccctggacccgccccccgcggtcagccccgt...
ctcgccggtcgagggtctgggcgggtcccgcggcctcgggggcggatctcggacccggagactcggggcgacccggg...
gggtccgcaggtcctctcggtcagtctgtgccttggccttgcagatctgtgtctcccggtcccaatactccggcccc...
gtagccccacggtgatgaagcggggctccccgcggccgggccgggacacggaggtgtggaaatacctcatggagtggg...
cctggctcctccccacagaggccatttccctgcctgaccccgcactcaccagcccaggtctcggtcagggccactgc...
tggggactttagaactgggaccccggcgacactgattggcttctctagacacccgacacccaatgggagtgggaaat...
aactgtgcgcctccccaatgcagacaaggctctcggagcctgagaccctgagaccgcgcccggggccttggacgttc...
gggagagaccccccagtggcgcccacccgcttccccttcacttctcctcctggaatccctgtccctgaactggactcc...
```

bioinformatics Databases



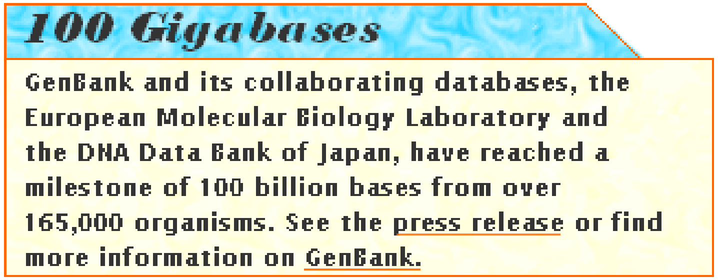
BIOINFORMATICS IN THE XXI CENTURY

FlyBase
NCBI
GO
e!
BIOSNI
SGD
WormBase
KINGDOM OF BURGUNDY
KINGDOM OF
MARCH OF VERONA
COUNTY OF SAVOY
CORSICA
SARDINIA
Fraxinetum (Saracen 900-970)
Lombard Duchies
Capua
Benevento
Salerno
Kingdom of Italy
Km. of Lombardy & other Imperial Lands
Lands of Dukes of Tuscany
Papal Territory: Patrimony of St. Peter
Papal states (claimed)
Saracen Territory c.1000
Byzantine Territory and dependencies c.1000
Chief communes by 1100 are underlined

Bioinformatics Challenges

The huge dataset

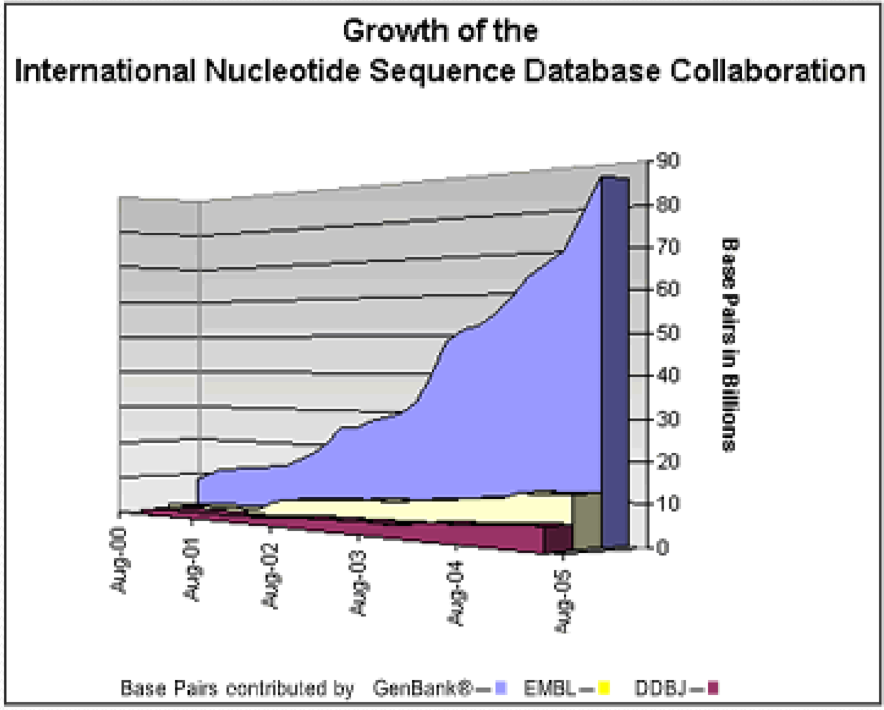
- Lots of new sequences being added
  - automated sequencers
  - genome sequencing
  - EST sequencing
  - environmental/metagenomic sequencing



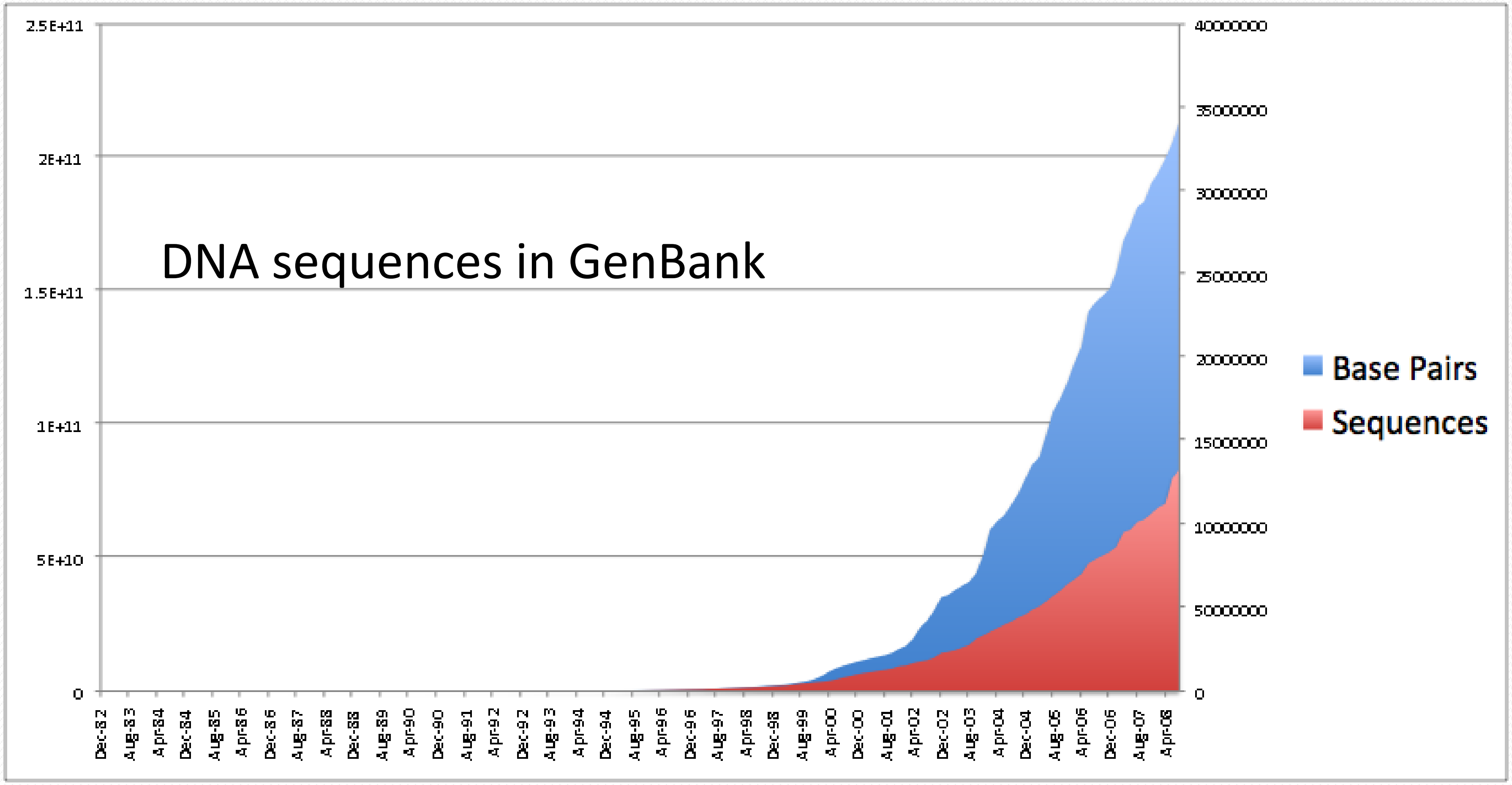
100 Gigabases

GenBank and its collaborating databases, the European Molecular Biology Laboratory and the DNA Data Bank of Japan, have reached a milestone of 100 billion bases from over 165,000 organisms. See the press release or find more information on GenBank.

- GenBank has over 100 **Billion** bases and is doubling every year!!
  - problem of exponential growth
  - how can computers keep up?
  - hard drives are cheaper, but processor speeds are not keeping up



DNA Sequencing capability has grown exponentially

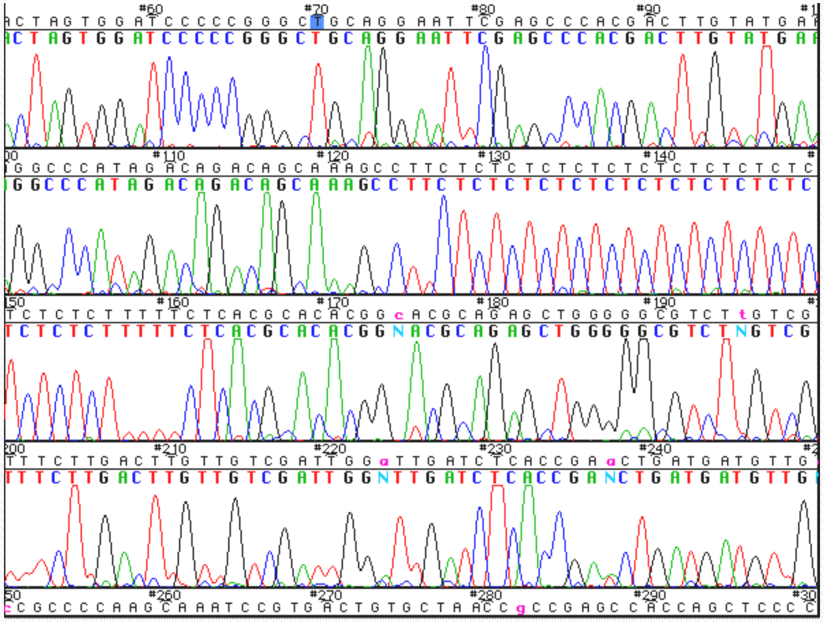


DNA sequences in GenBank

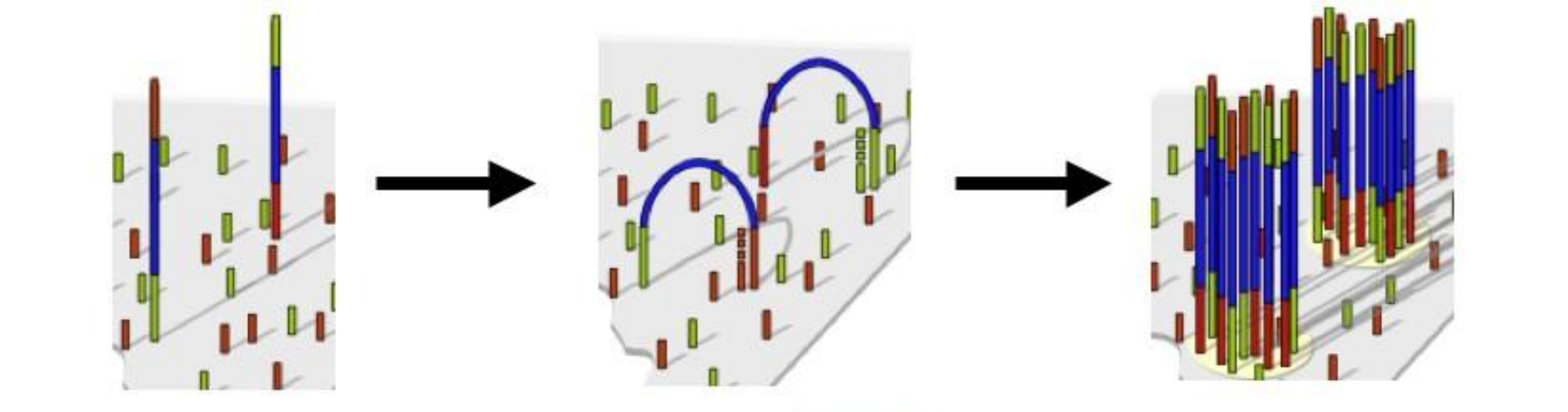
Doubling time = 18 months

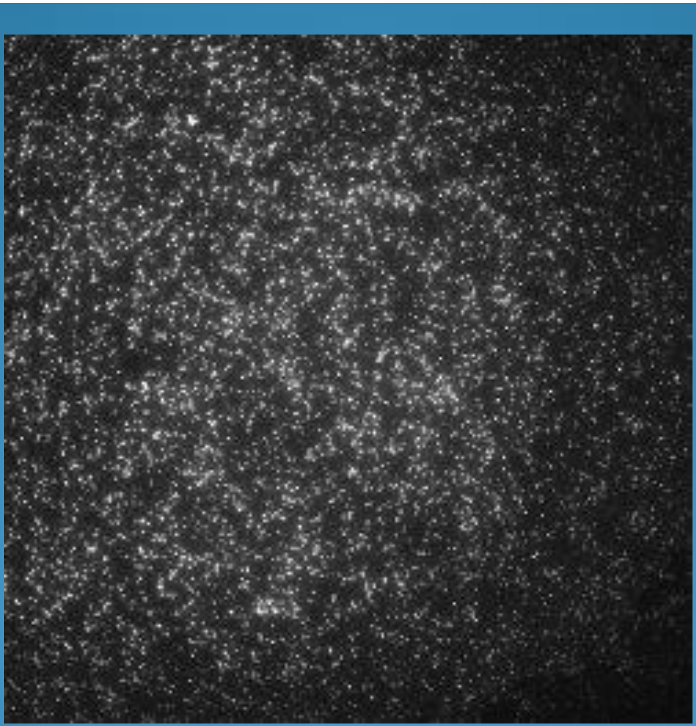
DNA Sequencing & Assembly

- Automated Sequencers
- ~500 bp reads must be assembled into complete genes & genomes
- faster sequencing relies on better software


#60 #70 #80 #90 #1
ACTAGTGGATCCCCCGGGCTGCAGGAATTCGAGCCCACGACTTGTATGAA
ACTAGTGGATCCCCCGGGCTGCAGGAATTCGAGCCCACGACTTGTATGAA
00 #110 #120 #130 #140 #
GGCCCATAGACAGACAGCAAAGCCTTCTCTCTCTCTCTCTCTCTCTCTC
GGCCCATAGACAGACAGCAAAGCCTTCTCTCTCTCTCTCTCTCTCTCTC
150 #160 #170 #180 #190 #
TCTCTCTTTTTCTCACGCACACGGcACGCAGAGCTGGGGGCGTCTtGTCG
TCTCTCTTTTTCTCACGCACACGGNACGCAGAGCTGGGGGCGTCTNGTCG
200 #210 #220 #230 #240 #2
TTTCTTGACTTGTTGTCGATTGGaTTGATCTCACCGAaCTGATGATGTTG
TTTCTTGACTTGTTGTCGATTGGNTTGATCTCACCGANCTGATGATGTTG
50 #260 #270 #280 #290 #30
cCGCCCCAAGCAAATCCGTGACTGTGCTAACCgCCGAGCCACCAGCTCCCC

Next Generation Sequencing

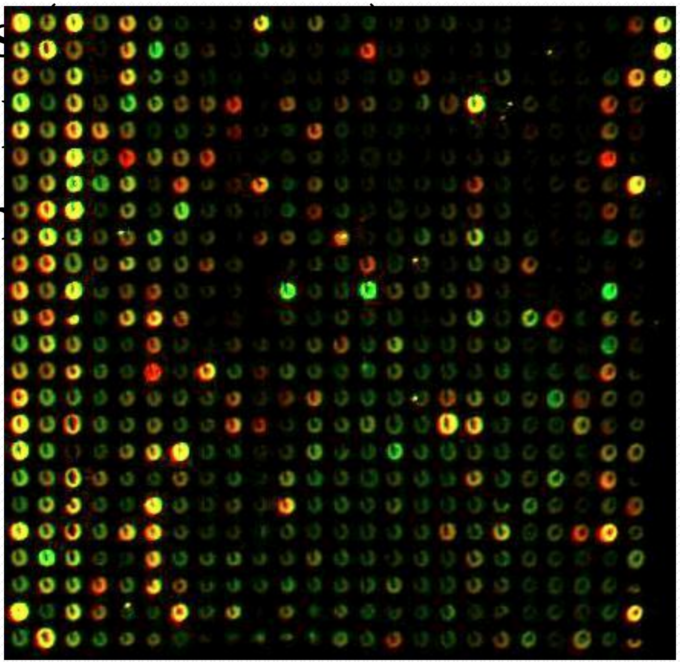




Genomics Technologies

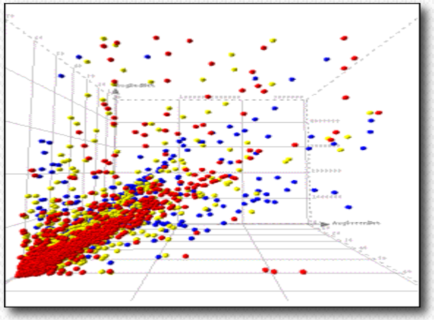
- Next-Generation DNA sequencing
- Automated annotation of sequences
- DNA microarrays
  - gene expression (measure RNA levels)
  - single nucleotide polymorphisms (SNPs)
  - ChIP-chip, genomic tiling, etc
- Proteomics (

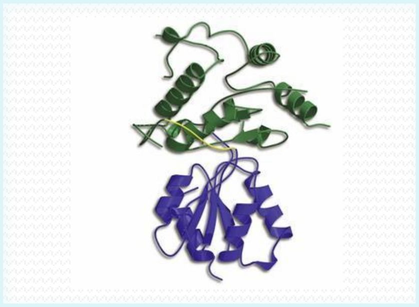
Protein-protein



Biological Information

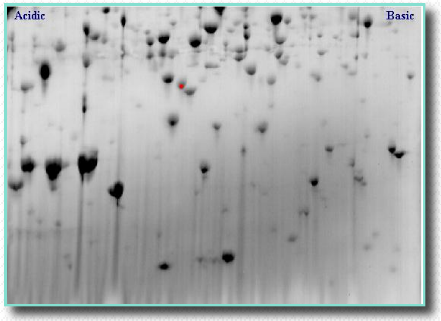
mRNA Expression



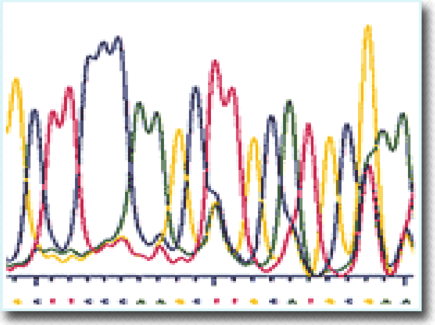


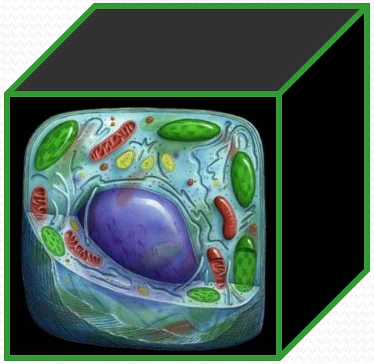
Protein 3-D Structure

Protein 2-D gel



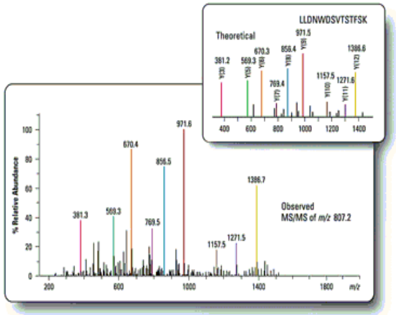
Genome sequence





The Cell

Mass Spec.

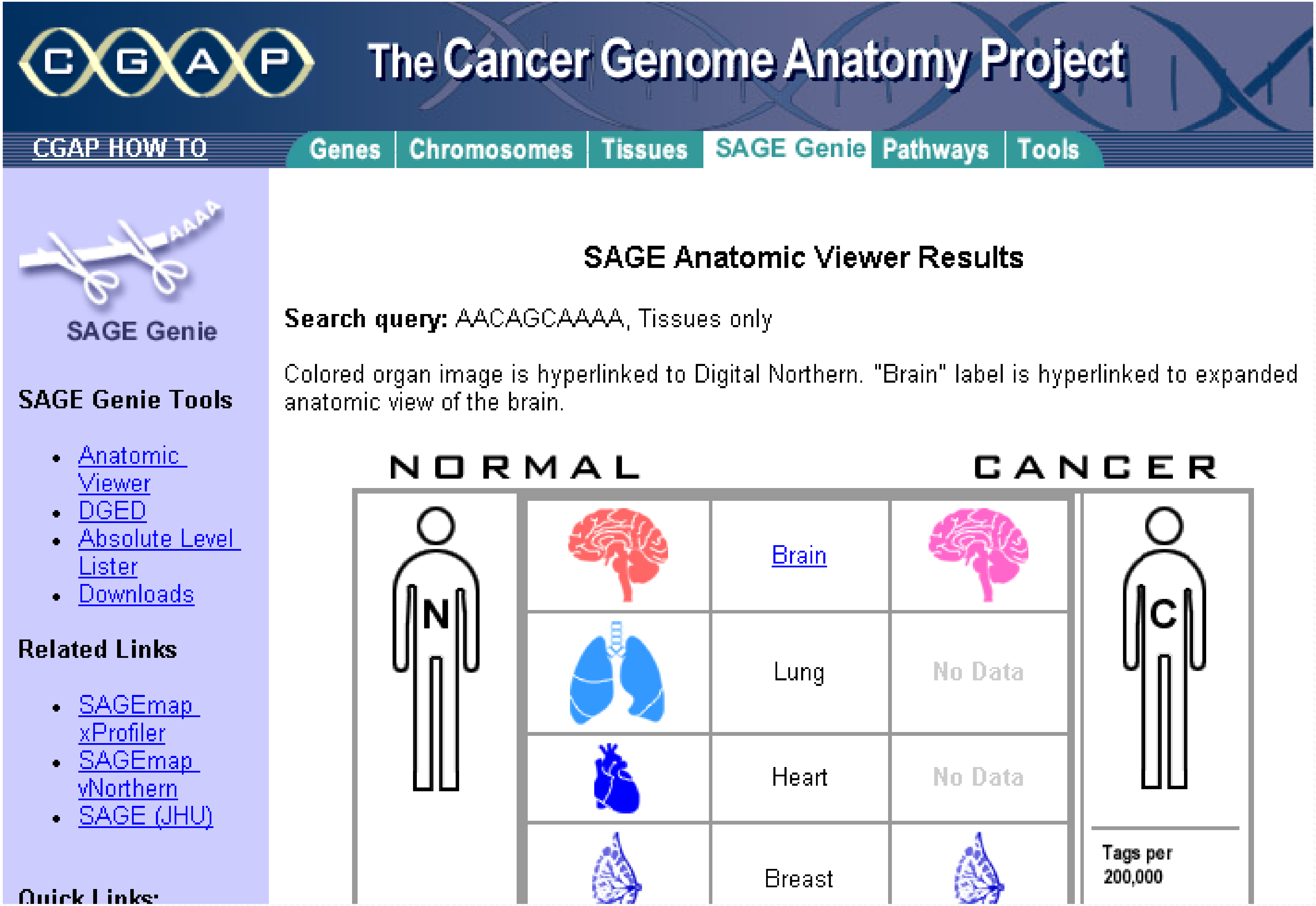


New Types of Big Biological Data

- Microarrays - gene expression
- Networks of protein-protein interactions

Microarray Data Analysis

- Linkage between gene expression data and gene sequence/function/metabolic pathways databases
- Discovery of common sequences in co-regulated genes
- Meta-studies using data from multiple experiments



SAGE Genie Tools

- Anatomic Viewer
- DGED
- Absolute Level Lister
- Downloads

Related Links

- SAGEmap xProfiler
- SAGEmap vNorthern
- SAGE (JHU)

Quick Links:

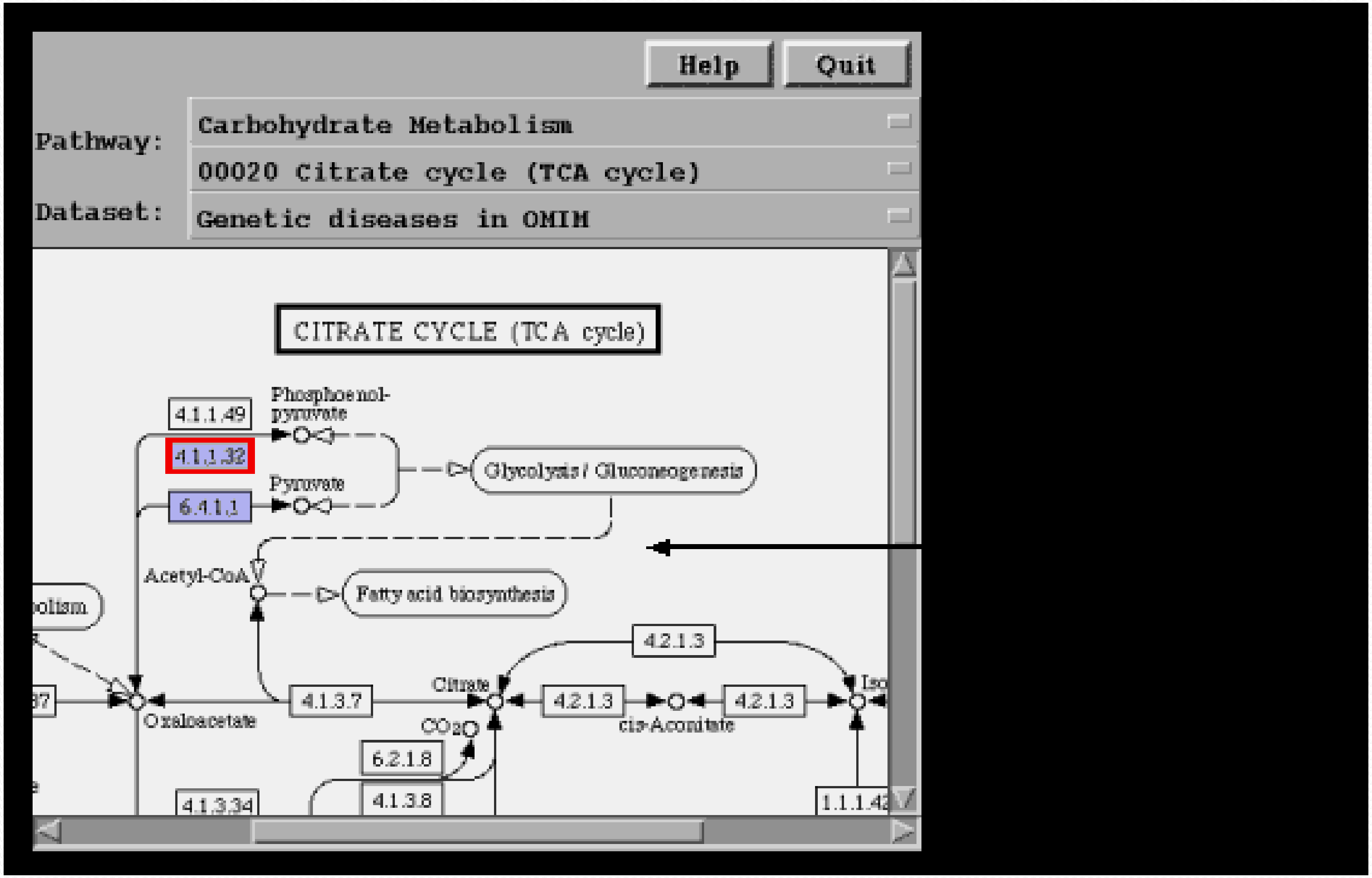
SAGE Anatomic Viewer Results

Search query: AACAGCAAAA, Tissues only

Colored organ image is hyperlinked to Digital Northern. "Brain" label is hyperlinked to expanded anatomic view of the brain.

NORMAL		CANCER
	Brain	
	Lung	No Data
	Heart	No Data
	Breast	

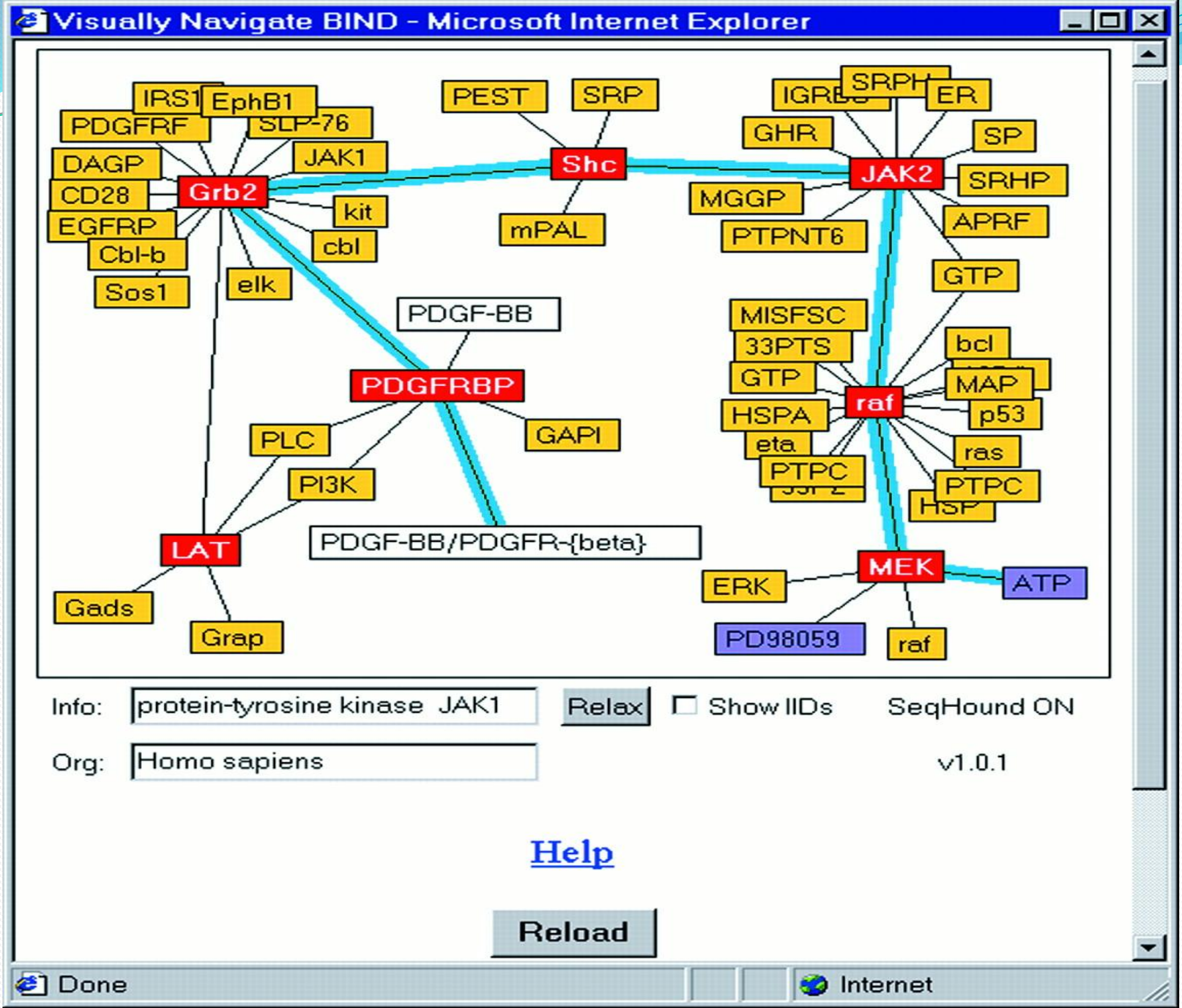
Tags per 200,000



Pathway: Carbohydrate Metabolism

00020 Citrate cycle (TCA cycle)

Dataset: Genetic diseases in OMIM



Visually Navigate BIND - Microsoft Internet Explorer

Info: protein-tyrosine kinase JAK1

□ Show IIDs

SeqHound ON

Org: Homo sapiens

v1.0.1

Done

Internet

Impact on Bioinformatics

- Genomics produces high-throughput, high-quality data, and bioinformatics provides the analysis and interpretation of these massive data sets.
- It is impossible to separate genomics laboratory technologies from the computational tools required for data analysis.

Example of Biological Database Formats

Database	Data Format	Website	Access
IntAct	Flat File	http://www.ebi.ac.uk/intact	ftp://ftp.ebi.ac.uk/pub/databases/intact/current
IntEnz	XML	http://www.ebi.ac.uk/intenz/	ftp://ftp.ebi.ac.uk/pub/databases/intenz/
Pfam	Flat File	http://www.sanger.ac.uk/Software/Pfam/	ftp://ftp.sanger.ac.uk/pub/databases/Pfam/
UniProt	Fasta, Flat File	http://www.expasy.ch/	ftp://ftp.expasy.org/
KEGG	XML	http://www.genome.ad.jp/kegg/	Web Services, ftp://ftp.genome.jp/pub/kegg/
PDB	pdb Flat file, mmCIF, XML	http://www.rcsb.org/pdb	Web Services, ftp://ftp.wwpdb.org/

Sequence Similarity

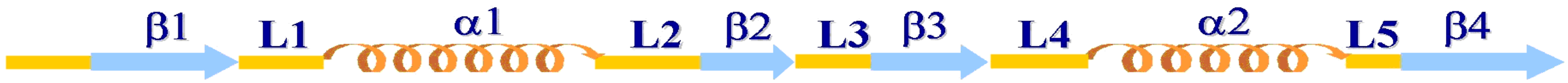
Sequence Alignment

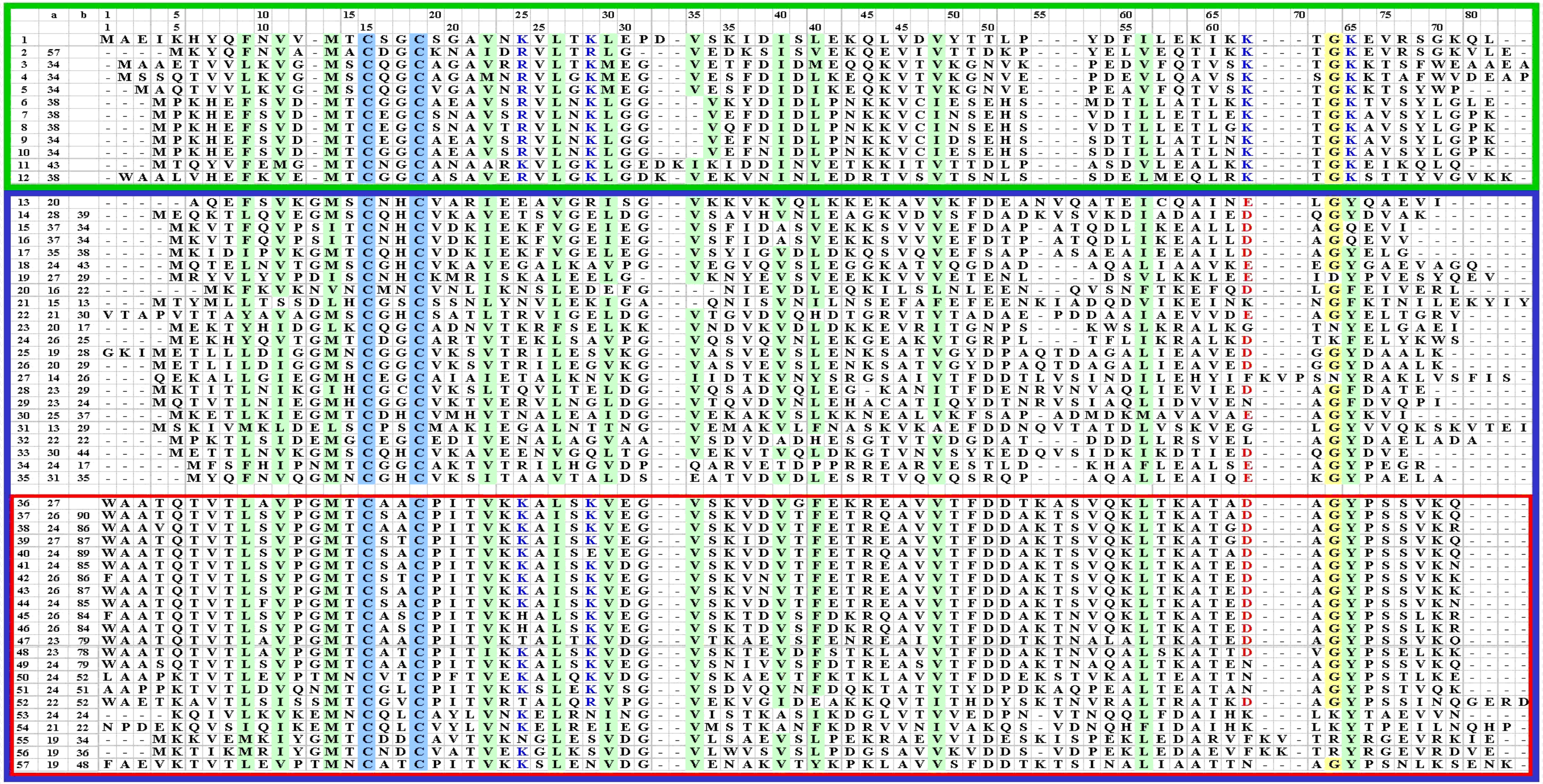
- Definition: Procedure for comparing two or more sequences by searching for a series of individual characters or character patterns that are *in the same order* in the sequences
  - **Pair-wise alignment**: compare two sequences
  - **Multiple sequence alignment**: compare more than two sequences

Bioinformatics

Stuart M. Brown, Ph.D.

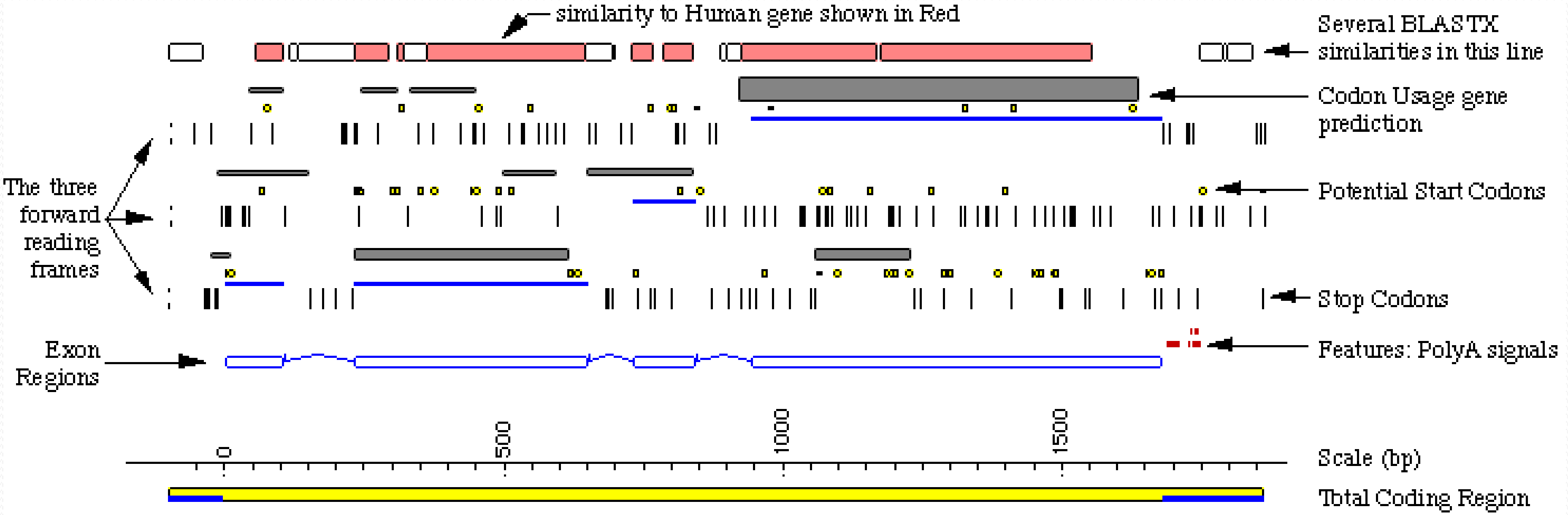
NYU School of Medicine





```
      a   b 1       5         10        15        20        25        30        35        40        45        50        55        60        65        70        75        80
            1       5         10          15        20        25        30          35        40        45        50              55        60              65        70
  1         M A E I K H Y Q F N V V - M T C S G C S G A V N K V L T K L E P D - V S K I D I S L E K Q L V D V Y T T L P - - - Y D F I L E K I K K - - - T G K E V R S G K Q L - -
  2  57     - - - - M K Y Q F N V A - M A C D G C K N A I D R V L T R L G - - - V E D K S I S V E K Q E V I V T T D K P - - - Y E L V E Q T I K K - - - T G K E V R S G K V L E -
  3  34     - M A A E T V V L K V G - M S C Q G C A G A V R R V L T K M E G - - V E T F D I D M E Q Q K V T V K G N V K - - - P E D V F Q T V S K - - - T G K K T S F W E A A E A
  4  34     - M S S Q T V V L K V G - M S C Q G C A G A M N R V L G K M E G - - V E S F D I D L K E Q K V T V K G N V E - - - P D E V L Q A V S K - - - S G K K T A F W V D E A P
  5  34     - - M A Q T V V L K V G - M S C Q G C V G A V N R V L G K M E G - - V E S F D I D I K E Q K V T V K G N V E - - - P E A V F Q T V S K - - - T G K K T S Y W P - - - -
  6  38     - - - M P K H E F S V D - M T C G G C A E A V S R V L N K L G G - - - V K Y D I D L P N K K V C I E S E H S - - - M D T L L A T L K K - - - T G K T V S Y L G L E - -
  7  38     - - - M P K H E F S V D - M T C E G C S N A V S R V L N K L G G - - - V E F D I D L P N K K V C I N S E H S - - - V D I L L E T L E K - - - T G K A V S Y L G P K - -
  8  38     - - - M P K H E F S V D - M T C E G C S N A V T R V L N K L G G - - - V Q F D I D L P N K K V C I N S E H S - - - V D T L L E T L G K - - - T G K A V S Y L G P K - -
  9  34     - - - M P K H E F S V D - M T C E G C A E A V S R V L N K L G G - - - V E F N I D L P N K K V C I D S E H S - - - S D T L L A T L N K - - - T G K A V S Y L G P K - -
 10  34     - - - M P K H E F S V D - M T C G G C A E A V S R V L N K L G G - - - V E F N I D L P N K K V C I E S E H S - - - S D I L L A T L N K - - - T G K A V S Y L G P K - -
 11  43     - - - M T Q Y V F E M G - M T C N G C A N A A R K V L G K L G E D K I K I D D I N V E T K K I T V T T D L P - - - A S D V L E A L K K - - - T G K E I K Q L Q - - - -
 12  38     - W A A L V H E F K V E - M T C G G C A S A V E R V L G K L G D K - V E K V N I N L E D R T V S V T S N L S - - - S D E L M E Q L R K - - - T G K S T T Y V G V K K -

 13  20     - - - - - A Q E F S V K G M S C N H C V A R I E E A V G R I S G - - V K K V K V Q L K K E K A V V K F D E A N V Q A T E I C Q A I N E - - - L G Y Q A E V I - - - - -
 14  28  39 - - - M E Q K T L Q V E G M S C Q H C V K A V E T S V G E L D G - - V S A V H V N L E A G K V D V S F D A D K V S V K D I A D A I E D - - - Q G Y D V A K - - - - - -
 15  37  34 - - - - M K V T F Q V P S I T C N H C V D K I E K F V G E I E G - - V S F I D A S V E K K S V V V E F D A P - A T Q D L I K E A L L D - - - A G Q E V I - - - - - - -
 16  37  34 - - - - M K V T F Q V P S I T C N H C V D K I E K F V G E I E G - - V S F I D A S V E K K S V V V E F D T P - A T Q D L I K E A L L D - - - A G Q E V V - - - - - - -
 17  35  38 - - - - M K I D I P V K G M T C Q H C V D K I E K F V G E L E G - - V S Y I G V D L D K Q S V Q V E F S A P - A S A E A I E E A I L D - - - A G Y E L G - - - - - - -
 18  24  43 - - - - M Q T E L N V T G M S C G H C V K A V E G A L K A V P G - - V E G V Q V S L E G G K A T V Q G D A D - - - A Q A L I A A V K E - - - E G Y G A E V A G Q - - -
 19  27  29 - - - - M R Y V L Y V P D I S C N H C K M R I S K A L E E L G - - - V K N Y E V S V E E K K V V V E T E N L - - - - D S V L K K L E E - - - I D Y P V E S Y Q E V - -
 20  16  22 - - - - - - M K F K V K N V N C M N C V N L I K N S L E D E F G - - - - N I E V D L E Q K I L S L N L E E N - - Q V S N F T K E F Q D - - - L G F E I V E R L - - - -
 21  15  13 - - - M T Y M L L T S S D L H C G S C S S N L Y N V L E K I G A - - - Q N I S V N I L N S E F A F E F E E N K I A D Q D V I K E I N K - - - N G F K T N I L E K Y I Y
 22  21  30 V T A P V T T A Y A V A G M S C G H C S A T L T R V I G E L D G - - V T G V D V Q H D T G R V T V T A D A E - P D D A A I A E V V D E - - - A G Y E L T G R V - - - -
 23  20  17 - - - - M E K T Y H I D G L K C Q G C A D N V T K R F S E L K K - - V N D V K V D L D K K E V R I T G N P S - - - K W S L K R A L K G - - - T N Y E L G A E I - - - -
 24  26  25 - - - - M E K H Y Q V T G M T C D G C A R T V T E K L S A V P G - - V Q S V Q V N L E K G E A K V T G R P L - - - T F L I K R A L K D - - - T K F E L Y K W S - - - -
 25  19  28 G K I M E T L L L D I G G M N C G G C V K S V T R I L E S V K G - - V A S V E V S L E N K S A T V G Y D P A Q T D A G A L I E A V E D - - - G G Y D A A L K - - - - -
 26  20  29 - - - M E T L I L D I G G M S C G G C V K S V T R I L E G V K G - - V A S V E V S L E N K S A T V G Y D P A Q T D A G A L I E A V E D - - - G G Y D A A L K - - - - -
 27  14  26 - - - Q E K A L L G I E G M H C E G C A I A I E T A L K N V K G - - I I D T K V N Y S R G S A I V T F D D T L V S I N D I L E H Y I F K V P S N Y R A K L V S F I S -
 28  23  29 - - - M K T I T L N I K G I H C G C C V K S L T Q V L T E L D G - - V Q S A D V Q L E G - K A N I T F D E N R V N V A Q L I E V I E D - - - A G F D A T E - - - - - -
 29  23  24 - - - M Q T V T L N I E G M H C G G C V K T V E R V L N G L D G - - V T Q V D V N L E H A C A T I Q Y D T N R V S I A Q L I D V V E N - - - A G F D V Q P I - - - - -
 30  25  37 - - - - M K E T L K I E G M T C D H C V M H V T N A L E A I D G - - V E K A K V S L K K N E A L V K F S A P - A D M D K M A V A V A E - - - A G Y K V I - - - - - - -
 31  13  29 - - - M S K I V M K L D E L S C P S C M A K I E G A L N T T N G - - V E M A K V L F N A S K V K A E F D D N Q V T A T D L V S K V E G - - - L G Y V V Q K S K V T E I
 32  22  22 - - - - M P K T L S I D E M G C E G C E D I V E N A L A G V A A - - V S D V D A D H E S G T V T V D G D A T - - - D D D L L R S V E L - - - A G Y D A E L A D A - - -
 33  30  44 - - - - M E T T L N V K G M S C Q H C V K A V E E N V G Q L T G - - V E K V T V Q L D K G T V N V S Y K E D Q V S I D K I K D T I E D - - - Q G Y D V E - - - - - - -
 34  24  17 - - - - - M F S F H I P N M T C G G C A K T V T R I L H G V D P - - Q A R V E T D P P R R E A R V E S T L D - - - K H A F L E A L S E - - - A G Y P E G R - - - - - -
 35  31  35 - - - - - M Y Q F N V Q G M N C G H C V K S I T A A V T A L D S - - E A T V D V D L E S R T V Q V Q S R Q P - - - A Q A L L E A I Q E - - - K G Y P A E L A - - - - -

 36  27     W A A T Q T V T L A V P G M T C A A C P I T V K K A L S K V E G - - V S K V D V G F E K R E A V V T F D D T K A S V Q K L T K A T A D - - - A G Y P S S V K Q - - - -
 37  26  90 W A A T Q T V T L S V P G M T C S A C P I T V K K A I S K V E G - - V S K V D V T F E T R Q A V V T F D D A K T S V Q K L T K A T A D - - - A G Y P S S V K Q - - - -
 38  24  86 W A A V Q T V T L S V P G M T C A A C P I T V K K A I S K V E G - - V S K V D V T F E T R E A V V T F D D A K T S V Q K L T K A T G D - - - A G Y P S S V K R - - - -
 39  27  87 W A A T Q T V T L S V P G M T C S T C P I T V K K A I S K V E G - - V S K I D V T F E T R E A V V T F D D A K T S V Q K L T K A T G D - - - A G Y P S S V K Q - - - -
 40  24  89 W A A T Q T V T L S V P G M T C S A C P I T V K K A I S E V E G - - V S K V D V T F E T R Q A V V T F D D A K T S V Q K L T K A T A D - - - A G Y P S S V K Q - - - -
 41  24  85 W A A T Q T V T L S V P G M T C S A C P I T V K K A I S K V D G - - V S K V D V T F E T R E A V V T F D D A K T S V Q K L T K A T E D - - - A G Y P S S V K N - - - -
 42  26  86 F A A T Q T V T L S V P G M T C S T C P I T V K K A I S K V E G - - V S K V N V T F E T R E A V V T F D D A K T S V Q K L T K A T E D - - - A G Y P S S V K K - - - -
 43  26  87 W A A T Q T V T L S V P G M T C S A C P I T V K K A I S K V E G - - V S K V N V T F E T R E A V V T F D D A K T S V Q K L T K A T E D - - - A G Y P S S V K K - - - -
 44  24  85 W A A T Q T V T L F V P G M T C S A C P I T V K K A I S K V D G - - V S K V D V T F E T R E A V V T F D D A K T S V Q K L T K A T E D - - - A G Y P S S V K N - - - -
 45  26  84 F A A T Q T V T L S V P G M T C A S C P I T V K H A L S K V E G - - V S K T D V S F D K R Q A V V T F D D A K T N V Q K L T K A T E D - - - A G Y P S S L K R - - - -
 46  26  84 W A A T Q T V T L S V P G M T C A S C P I T V K H A L S K V E G - - V S K T D V S F D K R Q A V V T F D D A K T N V Q K L T K A T E D - - - A G Y P S S L K R - - - -
 47  23  79 W A A T Q T V T L A V P G M T C A A C P I T V K T A L T K V D G - - V T K A E V S F E N R E A I V T F D D T K T N A L A L T K A T E D - - - A G Y P S S V K Q - - - -
 48  23  78 W A A T Q T V T L A V P G M T C A T C P I T I K K A L S K V D G - - V S K T E V D F S T K L A V V T F D D A K T N V Q A L S K A T T D - - - V G Y P S E L K K - - - -
 49  24  79 W A A S Q T V T L S V P G M T C A A C P I T V K K A L S K V E G - - V S N I V V S F D T R E A S V T F D D A K T N A Q A L T K A T E N - - - A G Y P S S V K Q - - - -
 50  24  52 L A A P K T V T L E V P T M N C V T C P F T V E K A L Q K V D G - - V S K A E V T F K T K L A V V T F D D E K S T V K A L T E A T T N - - - A G Y P S T L K E - - - -
 51  24  51 A A P P K T V T L D V Q N M T C G L C P I T V K K S L E K V S G - - V S D V Q V N F D Q K T A T V T Y D P D K A Q P E A L T E A T A N - - - A G Y P S T V Q K - - - -
 52  22  52 W A E T K A V T L S I S S M T C G V C P I T V R T A L Q R V P G - - V E K V G I D E A K K Q V T I T H D Y S K T N V R A L T R A T K D - - - A G Y P S S I N Q G E R D
 53  24  24 - - - - K Q I V L K V K E M N C Q L C A Y L V N K E L R N I N G - - V I S T K A S I K D G L V T V V E D P N - V T N Q Q L F D A I H K - - - L K Y T A E V V N - - - -
 54  21  22 N P D E K Q V S I Q I K E M T C Q L C V Y L V N K E L R E I E G - - V M S T K A N F K D R V V N I V A K Q S - V D N Q H F I D A I H K - - - L K Y T P E I L N Q H P -
 55  19  34 - - - M K K V E M K I Y G M T C D D C A V T V K N G L E S V D G - - V L S A E V S L P E K R A E V V I D E S K I S P E K L E D A R V F K V - T R Y R G E V R K I E - -
 56  19  36 - - - M K T I K M R I Y G M T C N D C V A T V E K G L K S V D G - - V L W V S V S L P D G S A V V K V D D S - V D P E K L E D A E V F K K - T R Y R G E V R D V E - -
 57  19  48 F A E V K T V T L E V P T M N C A T C P I T V K K S L E N V D G - - V E N A K V T Y K P K L A V V S F D D T K T S I N A L I A A T T N - - - A G Y P S N L K S E N K -
```

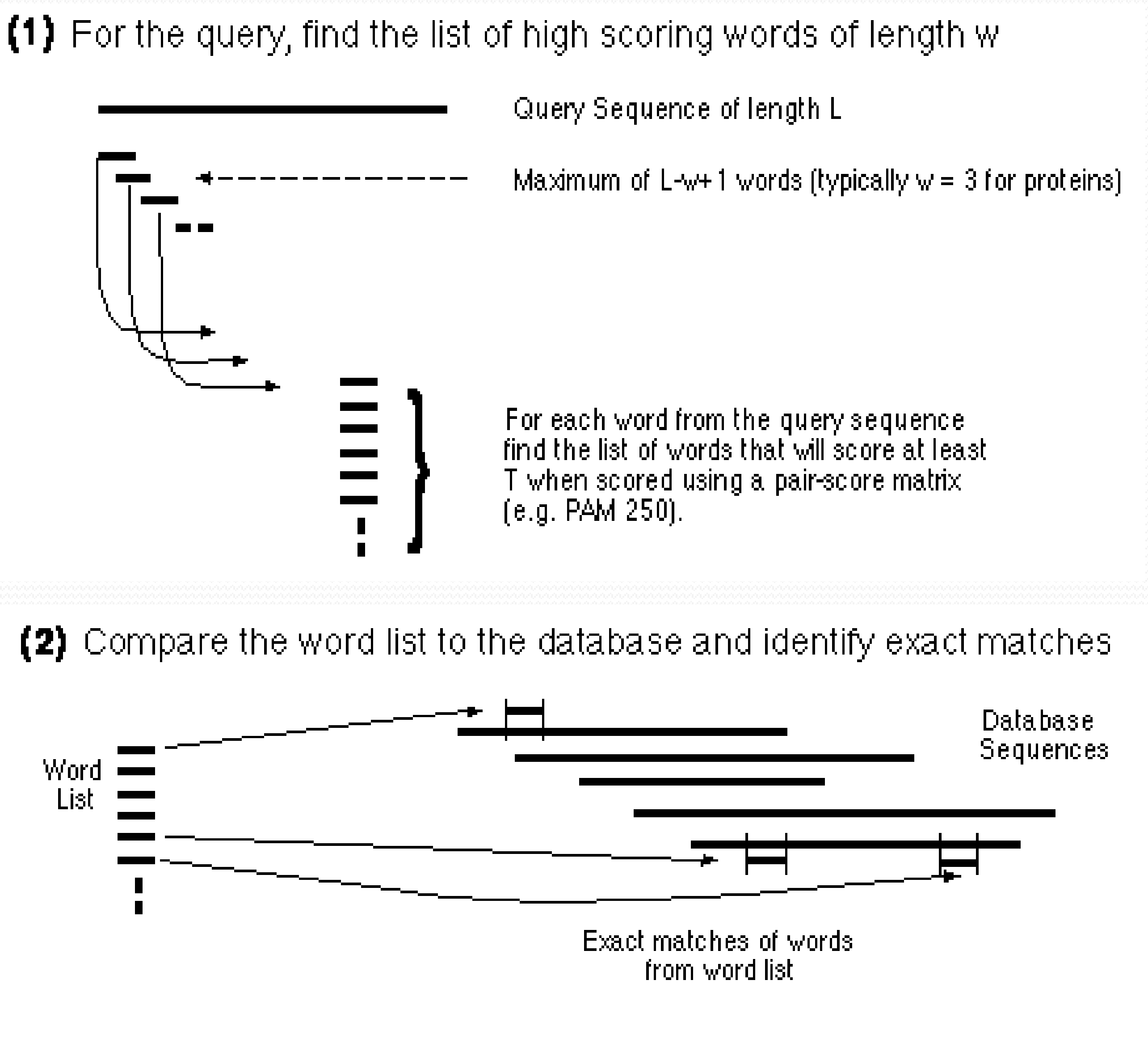
The next step is obviously to locate all of the genes and describe their functions. This will probably take another 15-20 years!


similarity to Human gene shown in Red
Several BLASTX similarities in this line
Codon Usage gene prediction
Potential Start Codons
Stop Codons
Features: PolyA signals
The three forward reading frames
Exon Regions
0
500
1000
1500
Scale (bp)
Total Coding Region

Similarity Searching the Databanks

- What is similar to my sequence?
- Searching gets harder as the databases get bigger - and quality degrades
- Tools: BLAST and FASTA = time saving heuristics (approximate)
- Statistics + informed judgement of the biologist

BLAST Algorithm



(1) For the query, find the list of high scoring words of length w

Query Sequence of length L

Maximum of L-w+1 words (typically w = 3 for proteins)

For each word from the query sequence find the list of words that will score at least T when scored using a pair-score matrix (e.g. PAM 250).

(2) Compare the word list to the database and identify exact matches

Exact matches of words from word list

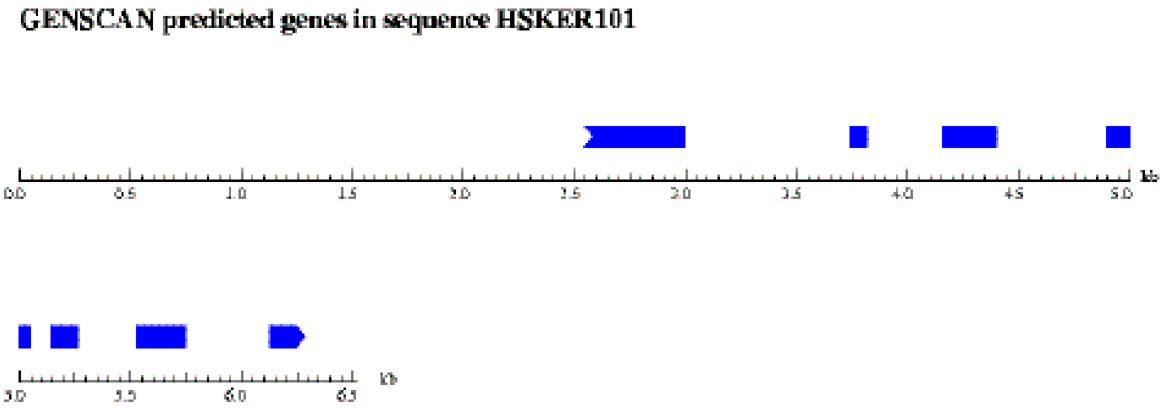
```
>gb|BE588357.1|BE588357 194087 BARC 5BOV Bos taurus cDNA 5'.
 Length = 369
 Score =  272 bits (137), Expect = 4e-71
 Identities = 258/297 (86%), Gaps = 1/297 (0%)
 Strand = Plus / Plus

Query: 17  aggatccaacgtcgctccagctgctcttgacgactccacagataccccgaagccatggca 76
            |||||||||||||||| | ||| | ||| || ||| | ||||  ||||| ||||||||| 
Sbjct: 1   aggatccaacgtcgctgcggctacccttaaccact-cgcagacccccgcagccatggcc 59
.
Query: 77  agcaagggcttgcaggacctgaagcaacaggtggaggggaccgcccaggaagccgtgtca 136
            |||||||||||||||||||||||| | || ||||||||| | ||||||||||| ||| ||
Sbjct: 60  agcaagggcttgcaggacctgaagaagcaagtggagggggcggcccaggaagcggtgaca 119
.
Query: 137 gcggccggagcggcagctcagcaagtggtggaccaggccacagaggcggggcagaaagcc 196
             |||||||| | || | ||||||||||||||| ||||||||||| || ||||||||||||
Sbjct: 120 tcggccggaacagcggttcagcaagtggtggatcaggccacagaagcagggcagaaagcc 179
.
Query: 197 atggaccagctggccaagaccacccaggaaaccatcgacaagactgctaaccaggcctct 256
            ||||||||| | |||||||| |||||||||||||||||| ||||||||||||||||||||
Sbjct: 180 atggaccaggttgccaagactacccaggaaaccatcgaccagactgctaaccaggcctct 239
.
Query: 257 gacaccttctctgggattgggaaaaaattcggcctcctgaaatgacagcagggagac 313
            || || ||||| ||  ||||||||||| | |||||||||||||||||| ||||||||
Sbjct: 240 gagactttctcgggtttgggaaaaaacttggcctcctgaaatgacagaagggagac 296
```

Finding Genes in genome Sequence is Not Easy

- About 1% of human DNA encodes functional genes.
- Genes are interspersed among long stretches of non-coding DNA.
- Repeats, pseudo-genes, and introns confound matters

Pattern Finding Tools

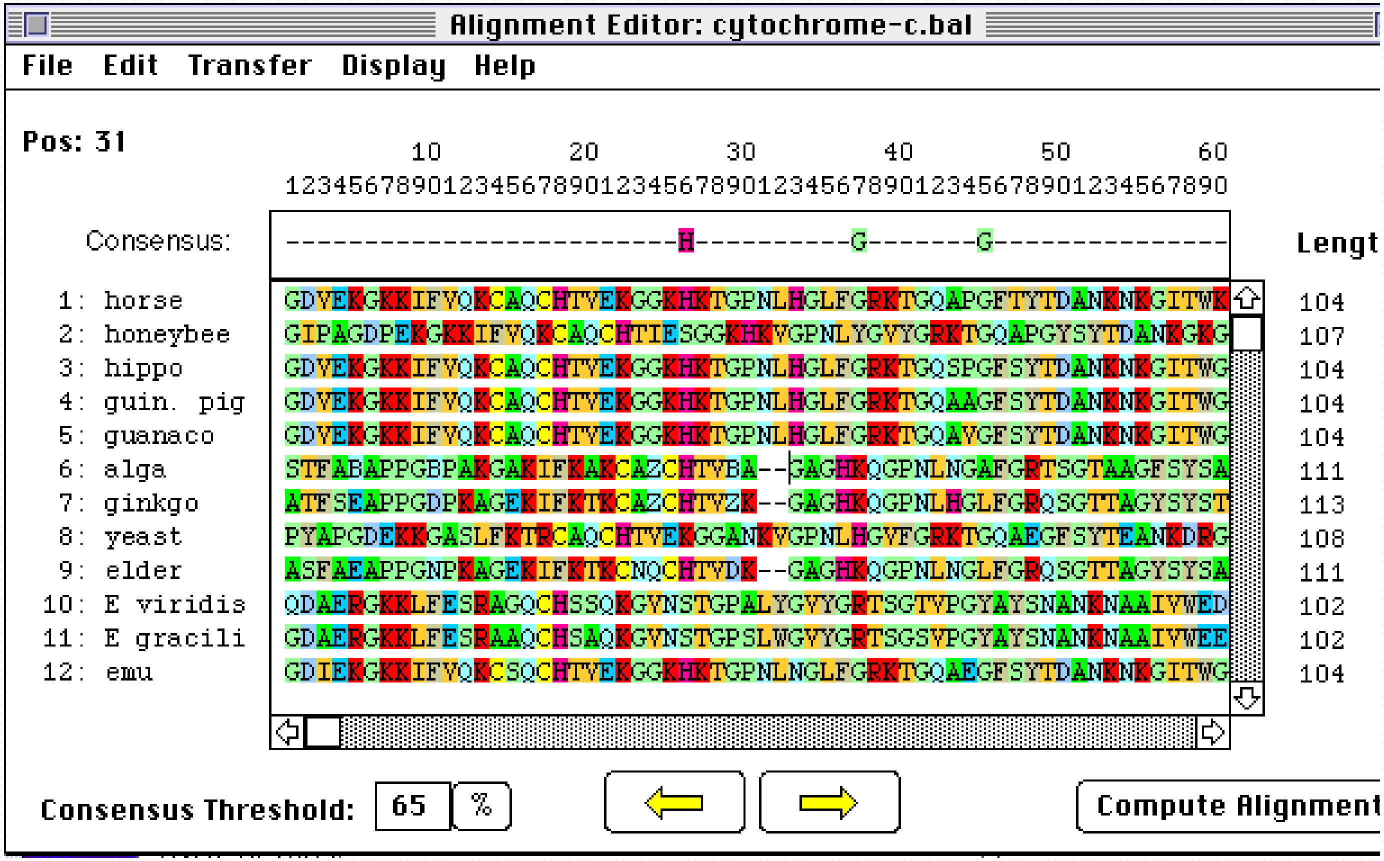
- It is possible to use DNA sequence patterns to predict genes:
  - promoters
  - translational start and stop codes (ORFs)
  - intron splice sites
  - codon bias
- Can also use similarity to known genes/ESTs



Alignment

- Alignment is the basis for finding similarity
- Pairwise alignment = dynamic programming
- Multiple alignment: protein families and functional domains
- Multiple alignment is "impossible" for lots of sequences
- Another heuristic - progressive pairwise alignment

Sample Multiple Alignment



```
Pos: 31
                          10        20        30        40        50        60
                 123456789012345678901234567890123456789012345678901234567890

    Consensus:   -------------------------H----------G-------G---------------   Lengt

 1: horse        GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWK   104
 2: honeybee     GIPAGDPEKGKKIFVQKCAQCHTIESGGKHKVGPNLYGVYGRKTGQAPGYSYTDANKGKG   107
 3: hippo        GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQSPGFSYTDANKNKGITWG   104
 4: guin. pig    GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAAGFSYTDANKNKGITWG   104
 5: guanaco      GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAVGFSYTDANKNKGITWG   104
 6: alga         STFABAPPGBPAKGAKIFKAKCAZCHTVBA--GAGHKQGPNLNGAFGRTSGTAAGFSYSA   111
 7: ginkgo       ATFSEAPPGDPKAGEKIFKTKCAZCHTVZK--GAGHKQGPNLHGLFGRQSGTTAGYSYST   113
 8: yeast        PYAPGDEKKGASLFKTRCAQCHTVEKGGANKVGPNLHGVFGRKTGQAEGFSYTEANKDRG   108
 9: elder        ASFAEAPPGNPKAGEKIFKTKCNQCHTVDK--GAGHKQGPNLNGLFGRQSGTTAGYSYSA   111
10: E viridis    QDAERGKKLFESRAGQCHSSQKGVNSTGPALYGVYGRTSGTVPGYAYSNANKNAAIVWED   102
11: E gracili    GDAERGKKLFESRAAQCHSAQKGVNSTGPSLWGVYGRTSGSVPGYAYSNANKNAAIVWEE   102
12: emu          GDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLNGLFGRKTGQAEGFSYTDANKNKGITWG   104

Consensus Threshold: 65 %
```

Structure- Function Relationships

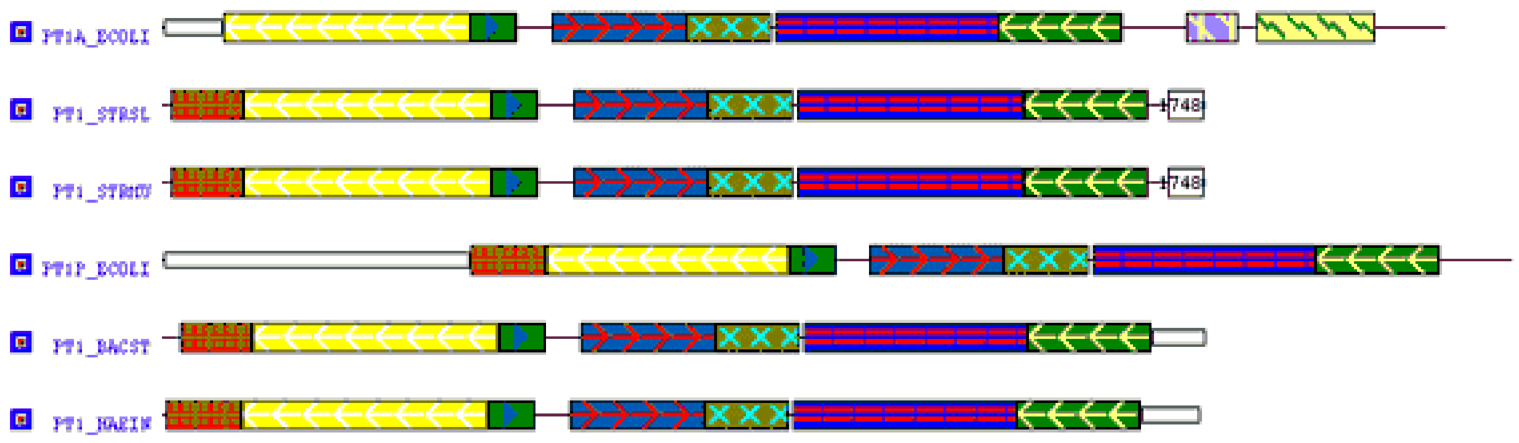
- Can we predict the function of protein molecules from their sequence?

sequence > structure > function

- Conserved functional domains = motifs
- Prediction of some simple 3-D structures (α-helix, β-sheet, membrane spanning, etc.)

Protein domains

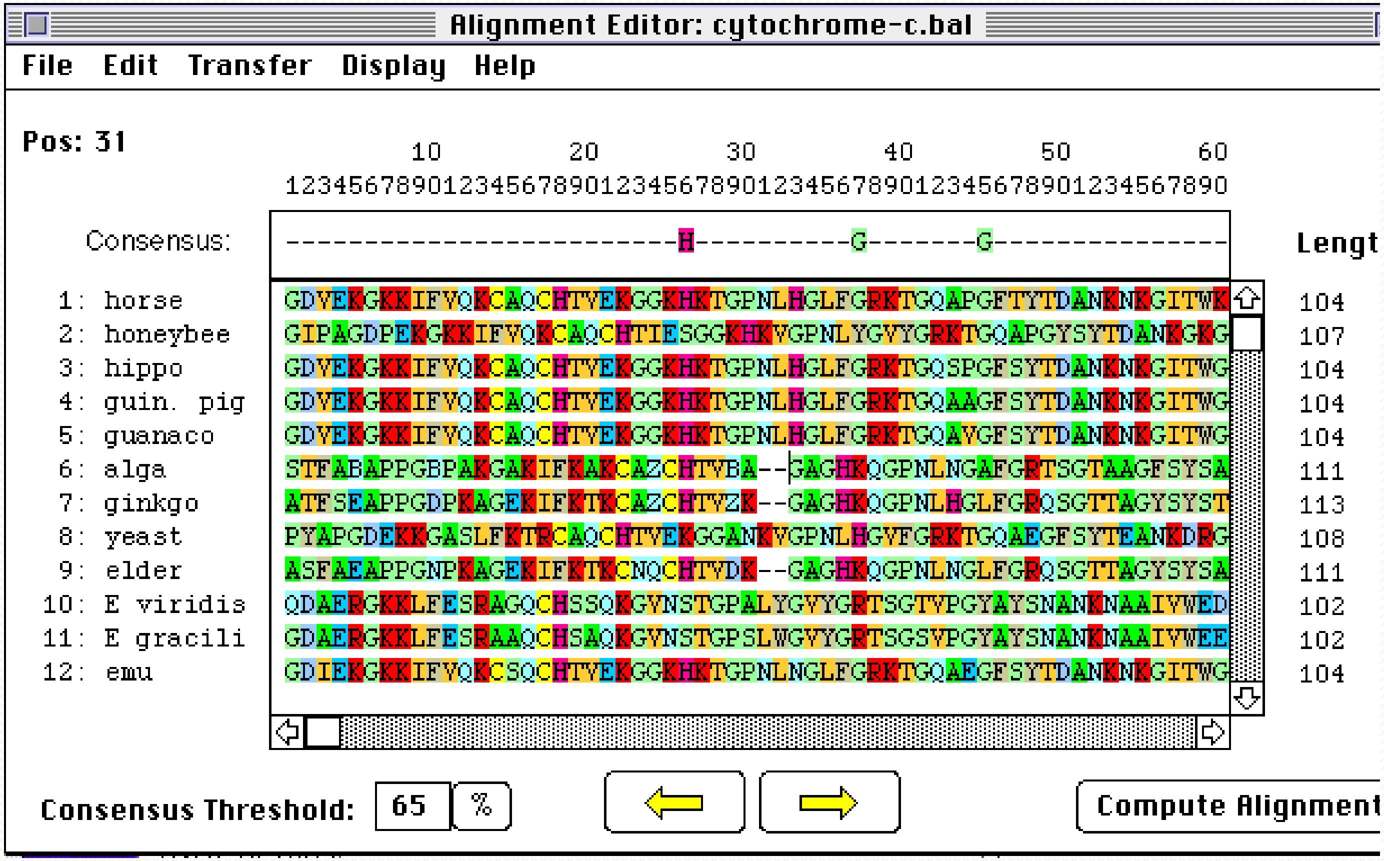
(from ProDom database)



Main Method for Pairwise Alignment

- Word or *k*-tuple methods (FASTA and BLAST)

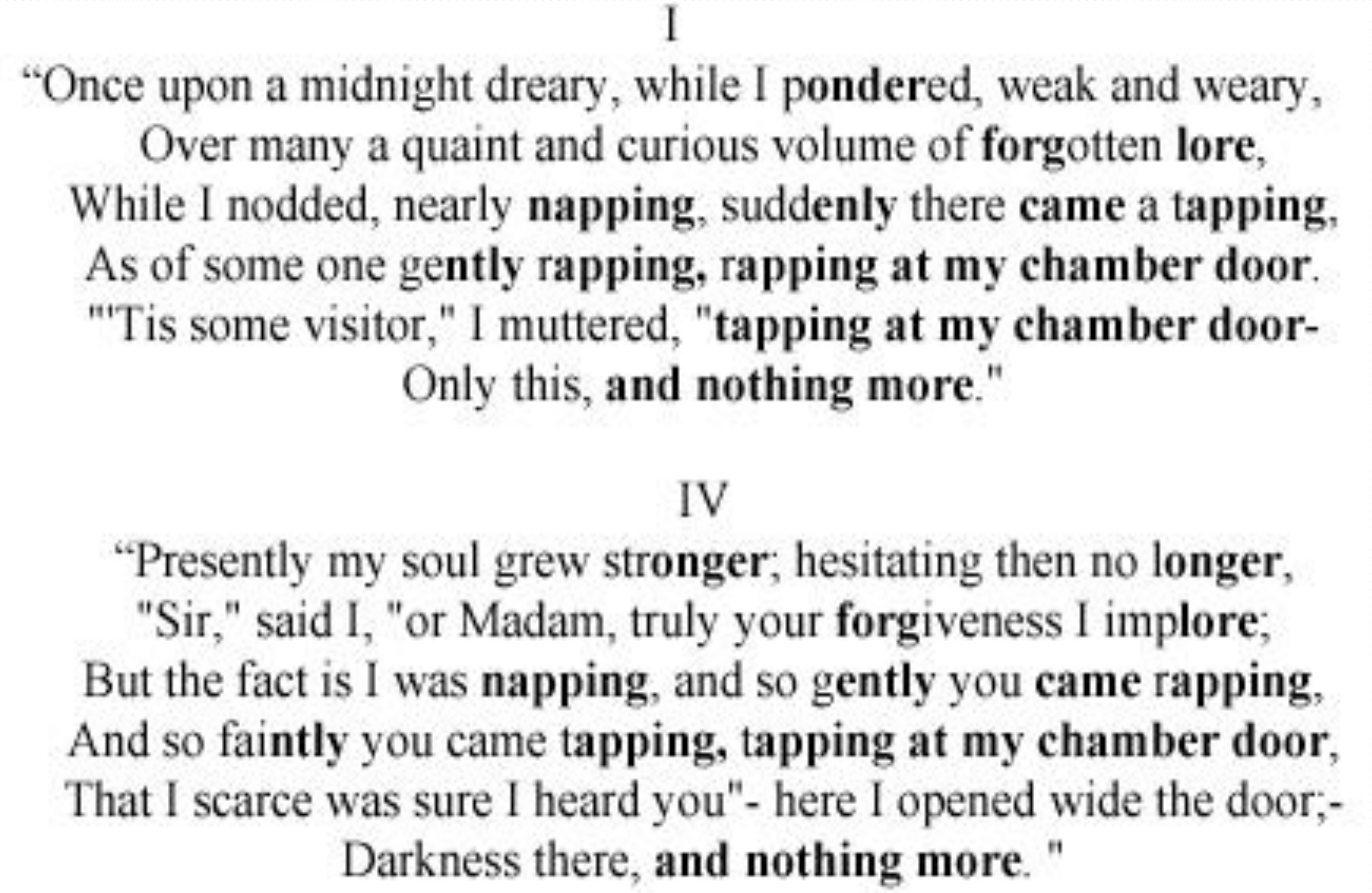
Sample Multiple Alignment



Alignment Editor: cytochrome-c.bal
File Edit Transfer Display Help
Consensus Threshold:
65
%
Compute Alignment

```
Pos: 31
                        10        20        30        40        50        60
               123456789012345678901234567890123456789012345678901234567890

   Consensus:  -------------------------H----------G-------G---------------     Lengt

 1: horse      GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWK     104
 2: honeybee   GIPAGDPEKGKKIFVQKCAQCHTIESGGKHKVGPNLYGVYGRKTGQAPGYSYTDANKGKG     107
 3: hippo      GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQSPGFSYTDANKNKGITWG     104
 4: guin. pig  GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAAGFSYTDANKNKGITWG     104
 5: guanaco    GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAVGFSYTDANKNKGITWG     104
 6: alga       STFABAPPGBPAKGAKIFKAKCAZCHTVBA--GAGHKQGPNLNGAFGRTSGTAAGFSYSA     111
 7: ginkgo     ATFSEAPPGDPKAGEKIFKTKCAZCHTVZK--GAGHKQGPNLHGLFGRQSGTTAGYSYST     113
 8: yeast      PYAPGDEKKGASLFKTRCAQCHTVEKGGANKVGPNLHGVFGRKTGQAEGFSYTEANKDRG     108
 9: elder      ASFAEAPPGNPKAGEKIFKTKCNQCHTVDK--GAGHKQGPNLNGLFGRQSGTTAGYSYSA     111
10: E viridis  QDAERGKKLFESRAGQCHSSQKGVNSTGPALYGVYGRTSGTVPGYAYSNANKNAAIVWED     102
11: E gracili  GDAERGKKLFESRAAQCHSAQKGVNSTGPSLWGVYGRTSGSVPGYAYSNANKNAAIVWEE     102
12: emu        GDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLNGLFGRKTGQAEGFSYTDANKNKGITWG     104
```

Examples



I

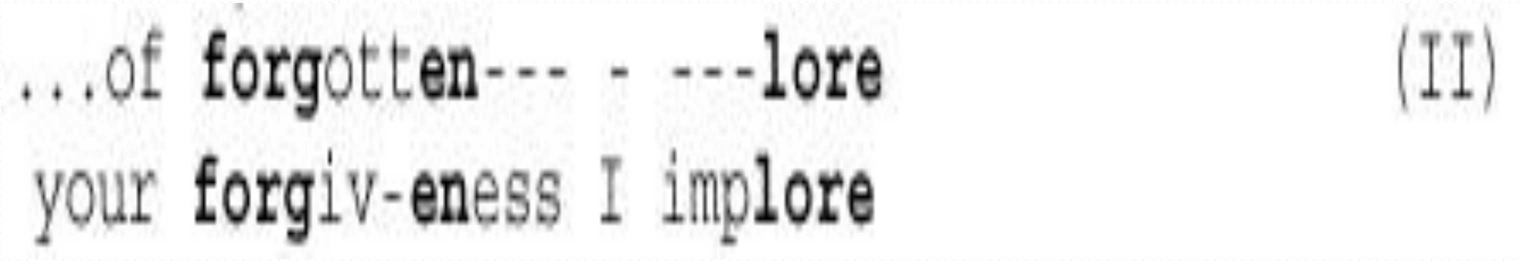
“Once upon a midnight dreary, while I p**onder**ed, weak and weary,
Over many a quaint and curious volume of **forg**otten **lore**,
While I nodded, nearly **napping**, sudd**enly** there **came** a t**apping**,
As of some one ge**ntly rapping, rapping at my chamber door**.
"'Tis some visitor," I muttered, "**tapping at my chamber door-**
Only this, **and nothing more**."

IV

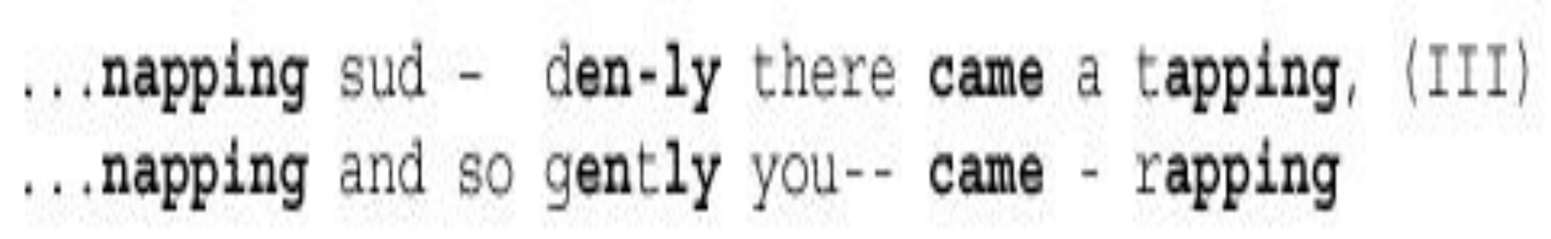
“Presently my soul grew str**onger**; hesitating then no l**onger**,
"Sir," said I, "or Madam, truly your **forg**iveness I imp**lore**;
But the fact is I was **napping**, and so g**ently** you **came rapping**,
And so fai**ntly** you came t**apping, tapping at my chamber door**,
That I scarce was sure I heard you"- here I opened wide the door;-
Darkness there, **and nothing more**. "

Examples (Cont…)

...I pondered ... (I)
...stronger...

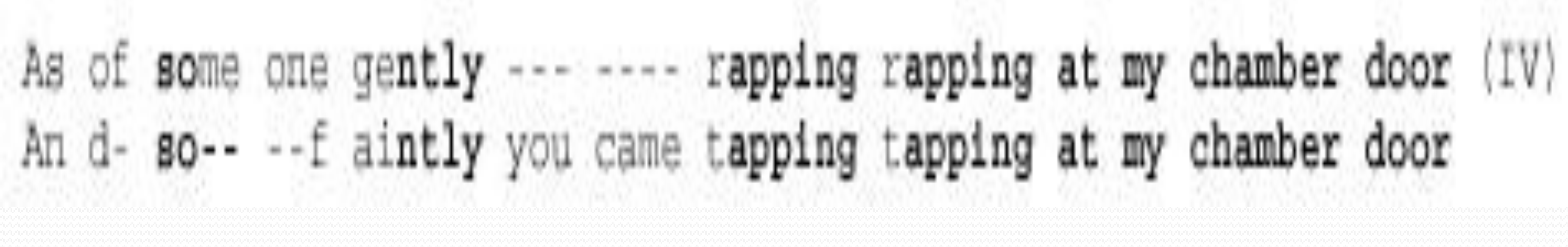


...of forgotten--- - ---lore (II)
your forgiv-eness I implore

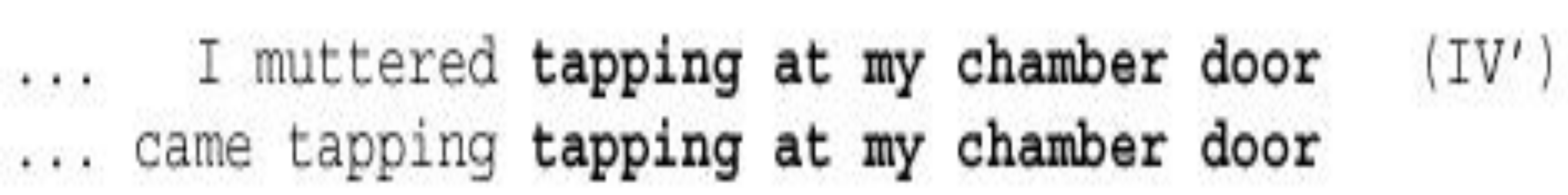


...napping sud - den-ly there came a tapping, (III)
...napping and so gently you-- came - rapping

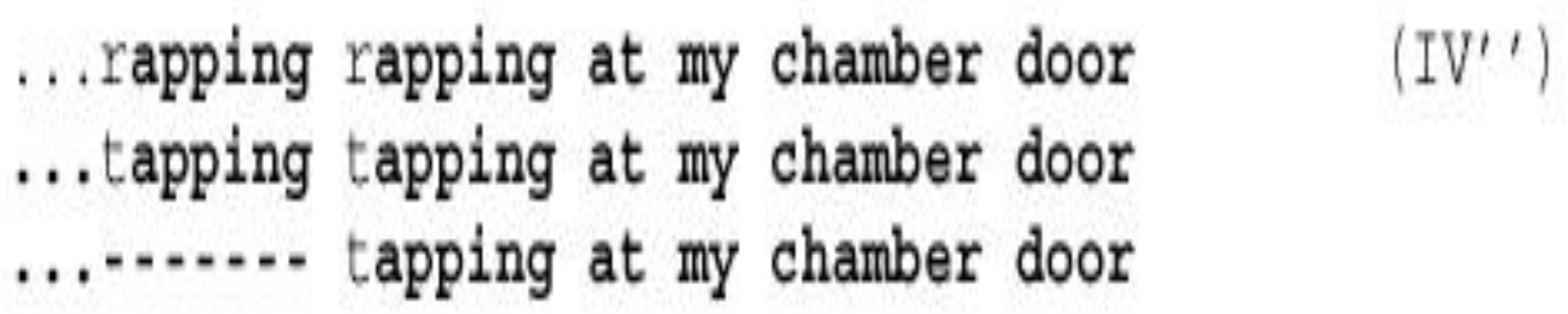
Examples (Cont…)



```
As of some one gently --- ---- rapping rapping at my chamber door (IV)
An d- so-- --f aintly you came tapping tapping at my chamber door
```



```
...    I muttered tapping at my chamber door    (IV')
... came tapping tapping at my chamber door
```



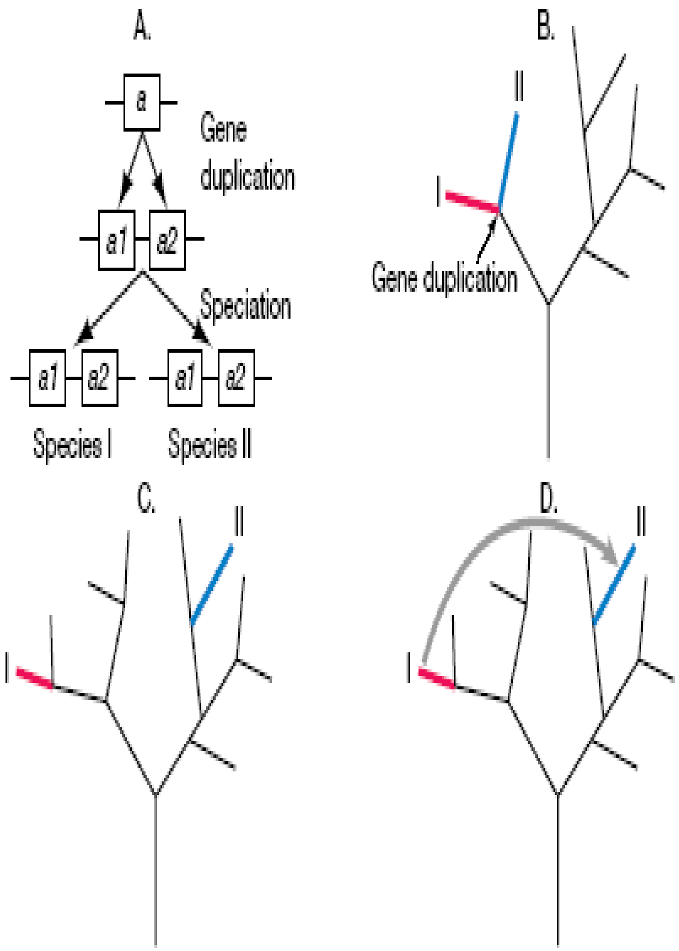
```
...rapping rapping at my chamber door          (IV'')
...tapping tapping at my chamber door
...------- tapping at my chamber door
```

Why do sequence alignments?

- To find out whether homologs of this gene (protein) are already available, and if they are, what is known about them
- To find whether two (or more) genes or proteins are evolutionarily related to each other
- To find structurally or functionally similar regions within proteins

Origin of similar genes

- Similar genes arise by **gene duplication**
- Copy of a gene inserted next to the original
- Two copies mutate independently
- Each can take on separate functions
- All or part can be transferred from one part of genome to another



Example sequence alignment

- Task: align **“abcdef”** with **“abdgf”**
- Write second sequence below the first

```
abcdef
abdgf
```

- Move sequences to give maximum match between them
- Show characters that match using vertical bar

Example sequence alignment

abcdef

||

abdgf

- Insert gap between **b** and **d** on lower sequence to allow **d** and **f** to align

Example sequence alignment

abcdef

|| | |

ab-dgf

Example sequence alignment

```
abcdef
|| | |
ab-dgf
```

- Note **e** and **g** don’t match

An alignment of two sequences t and s must satisfy:

- All symbols (residues) in the two sequences have to be in the alignment, and in the same order they appear in the sequences
- We can align one symbol from one sequence with one from the another
- A symbol can be aligned with a blank ('-')

- Two blanks cannot be aligned
- t: c g g g t a t c c a a
- s: c c c t a g g t c c c a

- t: c g g g t a - - t - c c a a
- s: c c c - t a g g t c c c - a

Matching Similarity vs. Identity

- Alignments can be based on finding only identical characters, or (more commonly) can be based on finding *similar* characters
- More on how to define *similarity* later

Global vs. Local Alignment

- We distinguish
  - **Global** alignment algorithms which optimize *overall* alignment between two sequences
  - **Local** alignment algorithms which seek only relatively *conserved* pieces of sequence
    - Alignment stops at the ends of regions of strong similarity
    - Favors finding conserved patterns in otherwise different pairs of sequences

Global vs. Local Alignment

- Global

```
LGPSSKQTGKGS-SRIWDN
|    | |||  | |
LN-ITKSAGKGAIMRLGDA
```

- Local

```
--------GKG--------
        |||
--------GKG--------
```

Global vs. Local Alignment

- Global

```
LGPSSKQTGKGS-SRIWDN
|   | |||  | |
LN-ITKSAGKGAIMRLGDA
```

- Local

```
--------TGKG---------
        |||
--------AGKG---------
```

Sequence FASTA Format

- In the process of writing a similarity searching program (in 1985), William Pearson designed a simple text format for DNA and protein sequences
- The FASTA format is now universal for all databases and software that handles DNA and protein sequences

One header line, starts with > with a [return] at end

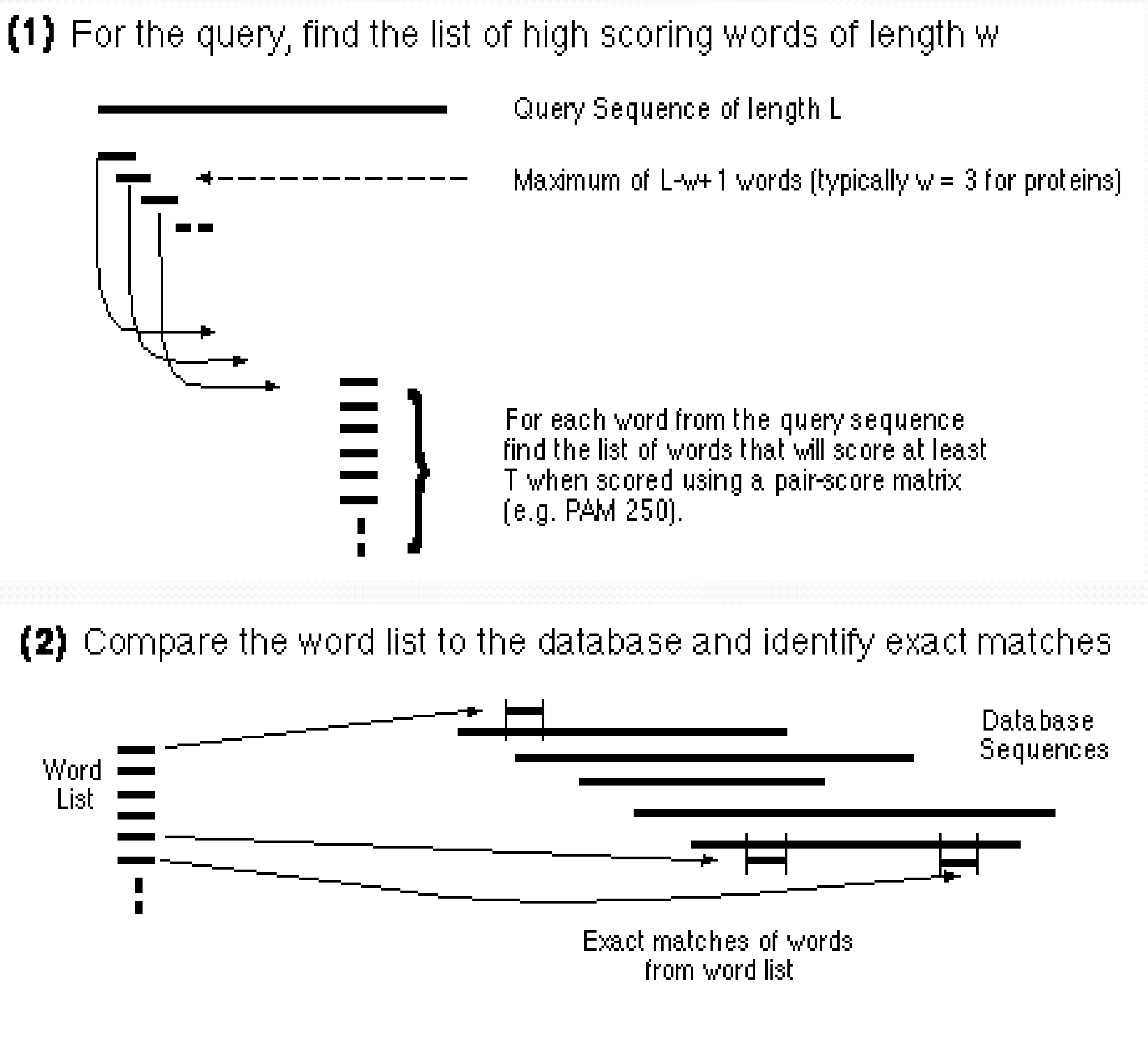
All other characters are part of sequence.
Most software ignores spaces, carriage returns.
Some ignores numbers

```
>URO1 uro1.seq  Length: 2018  November 9, 2000 11:50  Type: N  Check: 3854  ..
CGCAGAAAGAGGAGGCGCTTGCCTTCAGCTTGTGGGAAATCCCGAAGATGGCCAAAGAC
A
ACTCAACTGTTCGTTGCTTCCAGGGCCTGCTGATTTTTGGAAATGTGATTATTGGTTGTT
GCGGCATTGCCCTGACTGCGGAGTGCATCTTCTTTGTATCTGACCAACACAGCCTCTACC
CACTGCTTGAAGCCACCGACAACGATGACATCTATGGGGCTGCCTGGATCGGCATATTTG
TGGGCATCTGCCTCTTCTGCCTGTCTGTTCTAGGCATTGTAGGCATCATGAAGTCCAGCA
GGAAAATTCTTCTGGCGTATTTCATTCTGATGTTTATAGTATATGCCTTTGAAGTGGCAT
CTTGTATCACAGCAGCAACACAACAAGACTTTTTCACACCCAACCTCTTCCTGAAGCAGA
TGCTAGAGAGGTACCAAAACAACAGCCCTCCAAACAATGATGACCAGTGGAAAAACAATG
```

Multi-Sequence FASTA file

```
>FBpp0074027 type=protein; loc=X:complement(16159413..16159860,16160061..16160497); ID=FBpp0074027; name=CG12507-PA;
        parent=FBgn0030729,FBtr0074248; dbxref=FlyBase:FBpp0074027,FlyBase_Annotation_IDs:CG12507
        PA,GB_protein:AAF48569.1,GB_protein:AAF48569; MD5=123b97d79d04a06c66e12fa665e6d801; release=r5.1; species=Dmel; length=294;
MRCLMPLLLANCIAANPSFEDPDRSLDMEAKDSSVVDTMGMGMGVLDPTQ
PKQMNYQKPPLGYKDYDYYLGSRRMADPYGADNDLSASSAIKIHGEGNLA
SLNRPVSGVAHKPLPWYGDYSGKLLASAPPMYPSRSYDPYIRRYDRYDEQ
YHRNYPQYFEDMYMHRQRFDPYDSYSPRIPQYPEPYVMYPDRYPDAPPLR
DYPKLRRGYIGEPMAPIDSYSSSKYVSSKQSDLSFPVRNERIVYYAHLPE
IVRTPYDSGSPEDRNSAPYKLNKKKIKNIQRPLANNSTTYKMTL
>FBpp0082232 type=protein; loc=3R:complement(9207109..9207225,9207285..9207431); ID=FBpp0082232; name=mRpS21-PA;
        parent=FBgn0044511,FBtr0082764; dbxref=FlyBase:FBpp0082232,FlyBase_Annotation_IDs:CG32854-
        PA,GB_protein:AAN13563.1,GB_protein:AAN13563; MD5=dcf91821f75ffab320491d124a0d816c; release=r5.1; species=Dmel; length=87;
MRHVQFLARTVLVQNNNVEEACRLLNRVLGKEELLDQFRRTRFYEKPYQV
RRRINFEKCKAIYNEDMNRKIQFVLRKNRAEPFPGCS
>FBpp0091159 type=protein; loc=2R:complement(2511337..2511531,2511594..2511767,2511824..2511979,2512032..2512082); ID=FBpp0091159; name=CG33919-
        PA; parent=FBgn0053919,FBtr0091923; dbxref=FlyBase:FBpp0091159,FlyBase_Annotation_IDs:CG33919-
        PA,GB_protein:AAZ52801.1,GB_protein:AAZ52801; MD5=c91d880b654cd612d7292676f95038c5; release=r5.1; species=Dmel; length=191;
MKLVLVVLLGCCFIGQLTNTQLVYKLKKIECLVNRTRVSNVSCHVKAINW
NLAVVNMDCFMIVPLHNPIIRMQVFTKDYSNQYKPFLVDVKIRICEVIER
RNFIPYGVIMWKLFKRYTNVNHSCPFSGHLIARDGFLDTSLLPPFPQGFY
QVSLVVTDTNSTSTDYVGTMKFFLQAMEHIKSKKTHNLVHN
>FBpp0070770 type=protein; loc=X:join(5584802..5585021,5585925..5586137,5586198..5586342,5586410..5586605); ID=FBpp0070770; name=cv-PA;
        parent=FBgn0000394,FBtr0070804; dbxref=FlyBase:FBpp0070770,FlyBase_Annotation_IDs:CG12410-
        PA,GB_protein:AAF46063.1,GB_protein:AAF46063; MD5=0626ee34a518f248bbdda11a211f9b14; release=r5.1; species=Dmel; length=257;
MEIWRSLTVGTIVLLAIVCFYGTVESCNEVVCASIVSKCMLTQSCKCELK
NCSCCKECLKCLGKNYEECCSCVELCPKPNDTRNSLSKKSHVEDFDGVPE
LFNAVATPDEGDSFGYNWNVFTFQVDFDKYLKGPKLEKDGHYFLRTNDKN
LDEAIQERDNIVTVNCTVIYLDQCVSWNKCRTSCQTTGASSTRWFHDGCC
ECVGSTCINYGVNESRCRKCPESKGELGDELDDPMEEEMQDFGESMGPFD
GPVNNNY
…
```

BLAST Algorithm



(1) For the query, find the list of high scoring words of length w

Query Sequence of length L

Maximum of L-w+1 words (typically w = 3 for proteins)

For each word from the query sequence find the list of words that will score at least T when scored using a pair-score matrix (e.g. PAM 250).

(2) Compare the word list to the database and identify exact matches

Exact matches of words from word list

```
>gb|BE588357.1|BE588357 194087 BARC 5BOV Bos taurus cDNA 5'.
 Length = 369
 Score =  272 bits (137), Expect = 4e-71
 Identities = 258/297 (86%), Gaps = 1/297 (0%)
 Strand = Plus / Plus

Query: 17  aggatccaacgtcgctccagctgctcttgacgactccacagataccccgaagccatggca 76
           |||||||||||||||| | ||| | ||| || ||| | ||||  ||||| |||||||||
Sbjct: 1   aggatccaacgtcgctgcggctacccttaaccact-cgcagacccccgcagccatggcc 59
.
Query: 77  agcaagggcttgcaggacctgaagcaacaggtggaggggaccgcccaggaagccgtgtca 136
           |||||||||||||||||||||||| | || ||||||||| | ||||||||||| ||| ||
Sbjct: 60  agcaagggcttgcaggacctgaagaagcaagtggagggggcggcccaggaagcggtgaca 119
.
Query: 137 gcggccggagcggcagctcagcaagtggtggaccaggccacagaggcggggcagaaagcc 196
            |||||||| | || | ||||||||||||||| ||||||||||| || ||||||||||||
Sbjct: 120 tcggccggaacagcggttcagcaagtggtggatcaggccacagaagcagggcagaaagcc 179
.
Query: 197 atggaccagctggccaagaccacccaggaaaccatcgacaagactgctaaccaggcctct 256
           ||||||||| | |||||||| |||||||||||||||||| ||||||||||||||||||||
Sbjct: 180 atggaccaggttgccaagactacccaggaaaccatcgaccagactgctaaccaggcctct 239
.
Query: 257 gacaccttctctgggattgggaaaaaattcggcctcctgaaatgacagcagggagac 313
           || || ||||| ||  ||||||||||| | |||||||||||||||||| ||||||||
Sbjct: 240 gagactttctcgggttttgggaaaaaacttggcctcctgaaatgacagaagggagac 296
```

Two classes of widely used protein scoring matrices

PAM = % Accepted Mutations:

1500 changes in 71 groups w/ > 85% similarity

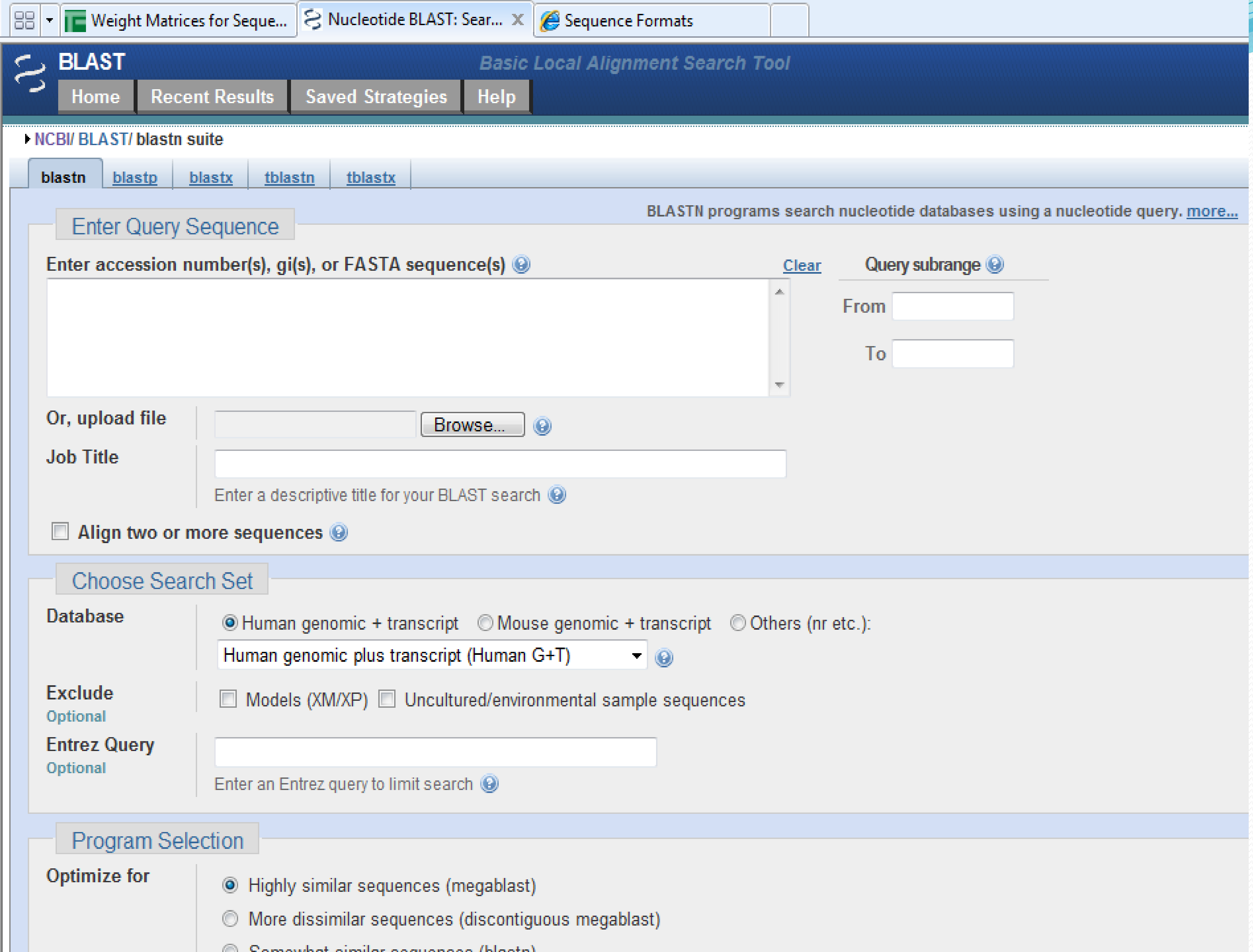
BLOSUM = Blocks Substitution Matrix:

2000 “blocks” from 500 families

```
>mysequence1
atggaggatgatttcatgtgcgatgatgaggaggactacgacctggaatactctga
agatagtaactccgagccaaatgtggatttggaaaatcagtactataattccaaag
cattaaaagaagatgacccaaaagcggcattaagcagtttccaaaaggttttggaa
cttgaaggtgaaaaaggagaatggggatttaaagcactgaaacaaatgattaagat
taacttcaagttgacaaactttccagaaatgatgaatagatataagcagctattga
cctatattcggagtgcagtcacaagaaattattctgaaaaatccattaattctatt
cttgattatatctctacttctaaacagatggatttactgcaggaattctatgaaac
aacactggaagctttgaaagatgctaag
```

Use **Blast** - there are different varieties, depending on what kind of sequence you have and what kind of sequence you are looking for

blastn	Search nucleotide database using a nucleotide query
blastp	Search protein database using a protein query
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query


Weight Matrices for Seque...
Nucleotide BLAST: Sear...
Sequence Formats
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastn suite
blastn
blastp
blastx
tblastn
tblastx
Clear
Browse...

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ● Human genomic + transcript ○ Mouse genomic + transcript ○ Others (nr etc.):

Human genomic plus transcript (Human G+T)

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional)

Enter an Entrez query to limit search

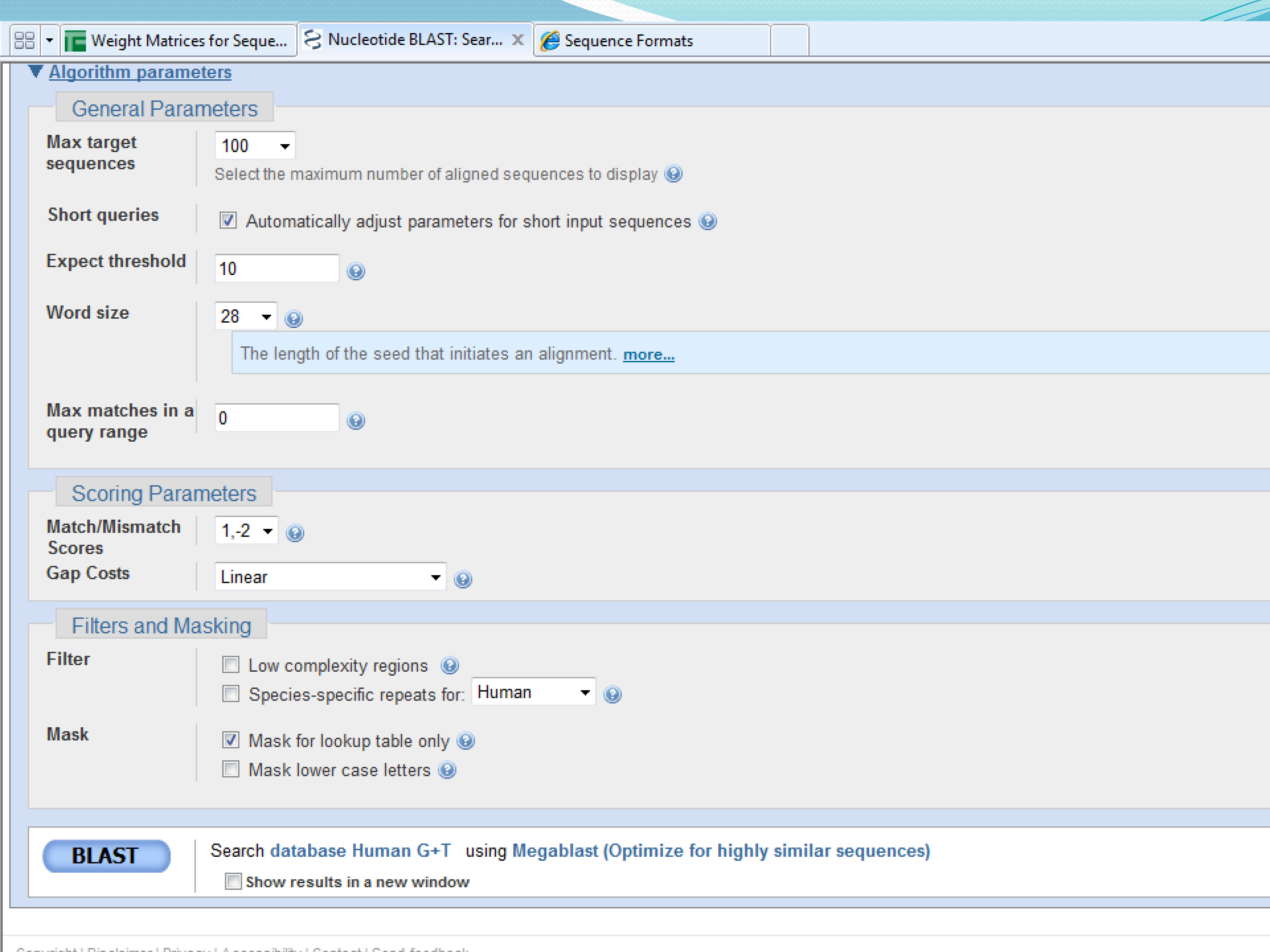
Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)



▼ Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 28
The length of the seed that initiates an alignment. more...

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear

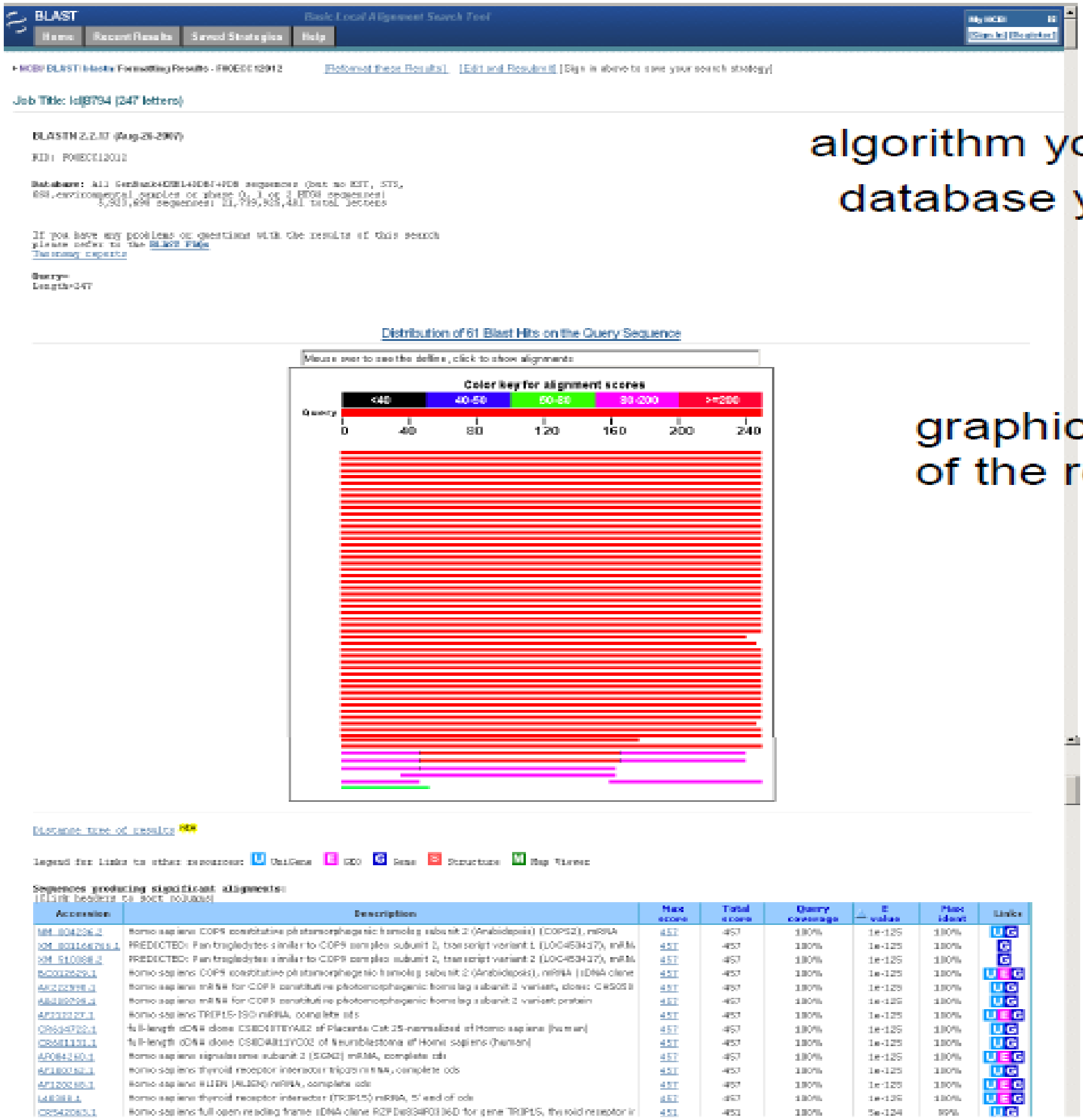
Filters and Masking

Filter: □ Low complexity regions
□ Species-specific repeats for: Human

Mask: ☑ Mask for lookup table only
□ Mask lower case letters

BLAST | Search database Human G+T using Megablast (Optimize for highly similar sequences)
□ Show results in a new window

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback



▸ NCBI/ BLAST/ blastn/ Formatting Results - F0OEC012012 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search strategy]

Job Title: lcl|8794 (247 letters)

BLASTN 2.2.17 (Aug-26-2007)

RID: F0OEC012012

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,environmental samples or phase 0, 1 or 2 HTGS sequences) 5,303,698 sequences; 21,719,925,481 total letters

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

Query=
Length=247

Distribution of 61 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Distance tree of results NEW

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_004236.2	Homo sapiens COP9 constitutive photomorphogenic homolog subunit 2 (Arabidopsis) (COPS2), mRNA	457	457	100%	1e-125	100%	U G
XM_001164761.1	PREDICTED: Pan troglodytes similar to COP9 complex subunit 2, transcript variant 1 (LOC453437), mRNA	457	457	100%	1e-125	100%	G
XM_510088.2	PREDICTED: Pan troglodytes similar to COP9 complex subunit 2, transcript variant 2 (LOC453437), mRNA	457	457	100%	1e-125	100%	G
BC012629.1	Homo sapiens COP9 constitutive photomorphogenic homolog subunit 2 (Arabidopsis), mRNA (cDNA clone	457	457	100%	1e-125	100%	U E G
AK222998.1	Homo sapiens mRNA for COP9 constitutive photomorphogenic homolog subunit 2 variant, clone: CAS05	457	457	100%	1e-125	100%	U G
AB189298.1	Homo sapiens mRNA for COP9 constitutive photomorphogenic homolog subunit 2 variant protein	457	457	100%	1e-125	100%	U G
AF212227.1	Homo sapiens TRIP15-ISO mRNA, complete cds	457	457	100%	1e-125	100%	U E G
CR614722.1	full-length cDNA clone CS0DI077YA02 of Placenta Cot 25-normalized of Homo sapiens (human)	457	457	100%	1e-125	100%	U G
CR601131.1	full-length cDNA clone CS0DA011YC02 of Neuroblastoma of Homo sapiens (human)	457	457	100%	1e-125	100%	U G
AF084260.1	Homo sapiens signalosome subunit 2 (SGN2) mRNA, complete cds	457	457	100%	1e-125	100%	U E G
AF180762.1	Homo sapiens thyroid receptor interactor trip15 mRNA, complete cds	457	457	100%	1e-125	100%	U G
AF120268.1	Homo sapiens ALIEN (ALIEN) mRNA, complete cds	457	457	100%	1e-125	100%	U E G
L40388.1	Homo sapiens thyroid receptor interactor (TRIP15) mRNA, 5' end of cds	457	457	100%	1e-125	100%	U E G
CR542063.1	Homo sapiens full open reading frame cDNA clone RZPDo834F0316D for gene TRIP15, thyroid receptor ir	451	451	100%	5e-124	99%	U G

algorithm you used

database you searched

graphical representation of the results

individual sequences found in the database

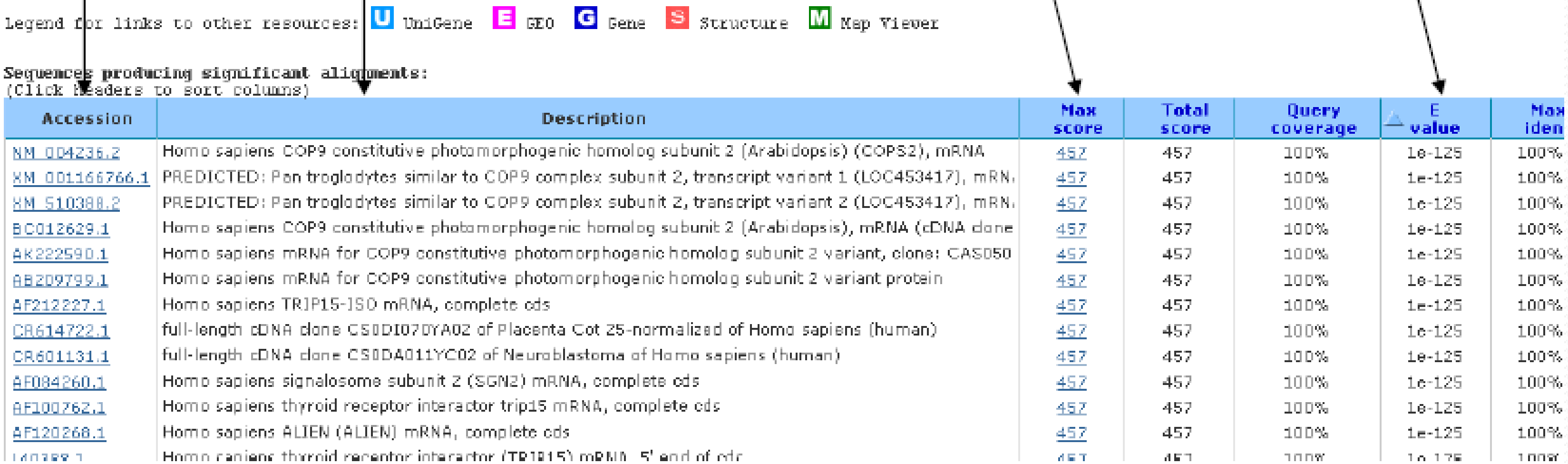
How to read a BLAST result

score – indicates how similar the query sequence is to the results, **larger number is better** BUT: longer sequences lead to higher scores

e-value – expectation val how often would you exp to find this sequence in th database randomly (this i particularly relevant if you query sequence is short c contains many repeats, e **smaller number is bette** Note: 2e-3 = $2*10^{-3}$ = 0.0

gene id of hit (accession number)

description



Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max iden
NM_004236.2	Homo sapiens COP9 constitutive photomorphogenic homolog subunit 2 (Arabidopsis) (COPS2), mRNA	457	457	100%	1e-125	100%
XM_001166766.1	PREDICTED: Pan troglodytes similar to COP9 complex subunit 2, transcript variant 1 (LOC453417), mRN	457	457	100%	1e-125	100%
XM_510388.2	PREDICTED: Pan troglodytes similar to COP9 complex subunit 2, transcript variant 2 (LOC453417), mRN	457	457	100%	1e-125	100%
BC012629.1	Homo sapiens COP9 constitutive photomorphogenic homolog subunit 2 (Arabidopsis), mRNA (cDNA clone	457	457	100%	1e-125	100%
AK222590.1	Homo sapiens mRNA for COP9 constitutive photomorphogenic homolog subunit 2 variant, clone: CAS050	457	457	100%	1e-125	100%
AB209799.1	Homo sapiens mRNA for COP9 constitutive photomorphogenic homolog subunit 2 variant protein	457	457	100%	1e-125	100%
AF212227.1	Homo sapiens TRIP15-ISO mRNA, complete cds	457	457	100%	1e-125	100%
CR614722.1	full-length cDNA clone CS0DI070YA02 of Placenta Cot 25-normalized of Homo sapiens (human)	457	457	100%	1e-125	100%
CR601131.1	full-length cDNA clone CS0DA011YC02 of Neuroblastoma of Homo sapiens (human)	457	457	100%	1e-125	100%
AF084260.1	Homo sapiens signalosome subunit 2 (SGN2) mRNA, complete cds	457	457	100%	1e-125	100%
AF100762.1	Homo sapiens thyroid receptor interactor trip15 mRNA, complete cds	457	457	100%	1e-125	100%
AF120268.1	Homo sapiens ALIEN (ALIEN) mRNA, complete cds	457	457	100%	1e-125	100%
L40388.1	Homo sapiens thyroid receptor interactor (TRIP15) mRNA, 5' end of cds	457	457	100%	1e-125	100%

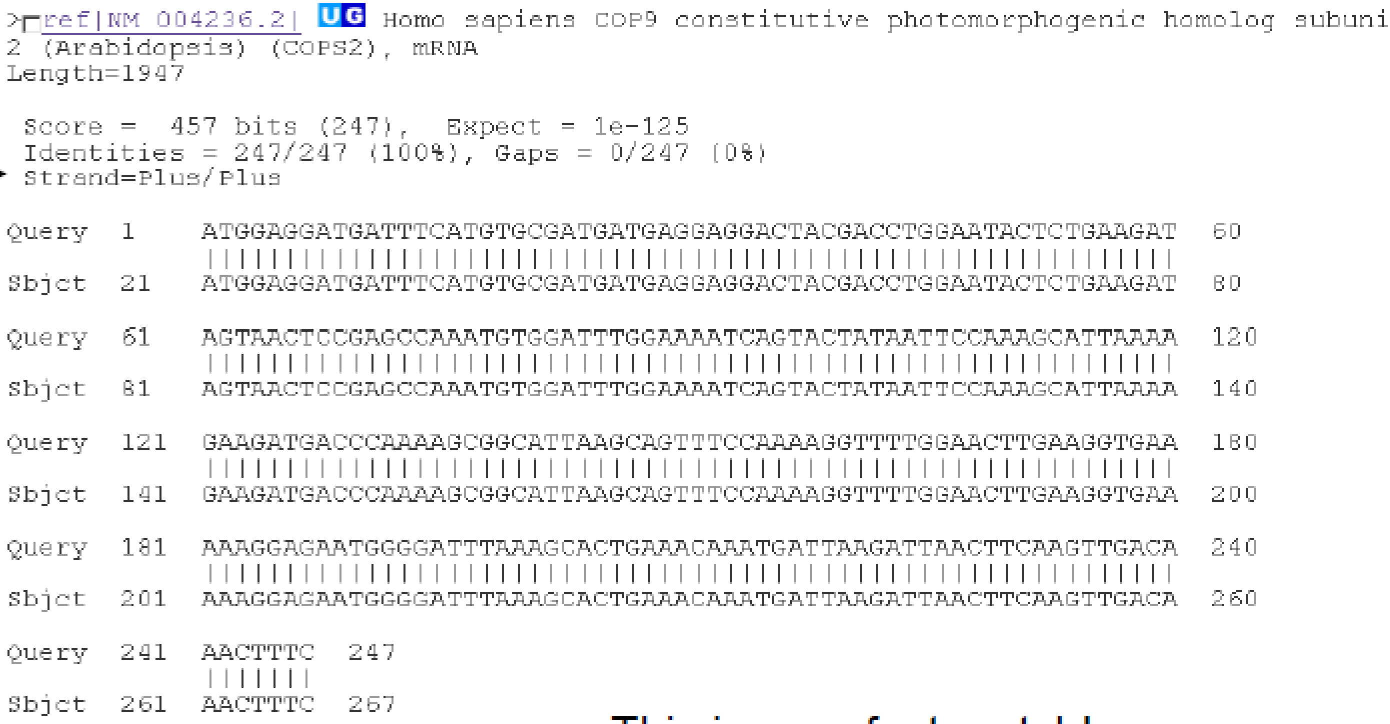
An individual “BLAST hit” in more detail

accession number + description

score, e-value, identical nucleotides, gaps, orientation

your sequence

blast hit



```
>ref|NM_004236.2| UG Homo sapiens COP9 constitutive photomorphogenic homolog subuni
2 (Arabidopsis) (COPS2), mRNA
Length=1947

 Score =  457 bits (247),  Expect = 1e-125
 Identities = 247/247 (100%), Gaps = 0/247 (0%)
 Strand=Plus/Plus

Query  1    ATGGAGGATGATTTCATGTGCGATGATGAGGAGGACTACGACCTGGAATACTCTGAAGAT  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  21   ATGGAGGATGATTTCATGTGCGATGATGAGGAGGACTACGACCTGGAATACTCTGAAGAT  80

Query  61   AGTAACTCCGAGCCAAATGTGGATTTGGAAAATCAGTACTATAATTCCAAAGCATTAAAA  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  81   AGTAACTCCGAGCCAAATGTGGATTTGGAAAATCAGTACTATAATTCCAAAGCATTAAAA  140

Query  121  GAAGATGACCCAAAAGCGGCATTAAGCAGTTTCCAAAAGGTTTTGGAACTTGAAGGTGAA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  141  GAAGATGACCCAAAAGCGGCATTAAGCAGTTTCCAAAAGGTTTTGGAACTTGAAGGTGAA  200

Query  181  AAAGGAGAATGGGGATTTAAAGCACTGAAACAAATGATTAAGATTAACTTCAAGTTGACA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  201  AAAGGAGAATGGGGATTTAAAGCACTGAAACAAATGATTAAGATTAACTTCAAGTTGACA  260

Query  241  AACTTTC  247
            |||||||
Sbjct  261  AACTTTC  267
```

This is a perfect match!

References

[1] T. Etzold, A. Ulyanow, and P. Argos, "SRS: Information Retrieval System for Molecular Biology Data Banks," *Methods Enzymol.*, vol. 226, pp. 114–128, 1996.

[2] W. Fujibuchi, S. Goto, H. Migimatsu, I. Uchiyama, A. Ogiwara, M. Kanehisa, "DBGET/LinkDB: An Integrated Database Retrieval System," *Proc. Pacific Symp. Biocomputing*, pp. 681–692, 1998.

[3] C. Ramu, "SIR: A Simple Indexing and Retrieval System for Biological Flat File Databases," *Bioinformatics*, vol. 17, no. 8, pp. 756–758, 2001.

[4] L. M. Haas, P. Schwarz, M. Kodali, E. Kotlar, J. E. Rice, and W. C. Swope, "DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources," *IBM Syst. J.*, vol. 40, pp. 489–510, 2001.

[5] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, *et al*, "The TSIMMIS Project: Integration of Heterogeneous Information Sources," *Proceedings of IPSJ conference*, Tokyo, Japan; pp. 7–18, 1994.

[6] A. Freier, R. Hofestadt, M. Lange, U. Scholz, A. Stephanik, "BioDataServer: a SQL-based Service for the Online Integration of Life Science Data," *Silico Biol*, 2(2):37–57, 2002.

[7] R. D. Stevens, P. Baker, S. Bechhofer, G. Jacoby, A. N., Ng, W. Paton, C. A. Goble, and A. Brass, "TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources," *Bioinformatics*, vol. 16, no. 2, pp. 184–186, 2000.

[8] J. Köhler, S. Philippi, and M. Lange, "SEMEDA – Ontology Based Integration of Biological Databases," *Bioinformatics*, vol. 19, no. 18, pp.2420–2427, 2003.

Thank YOU