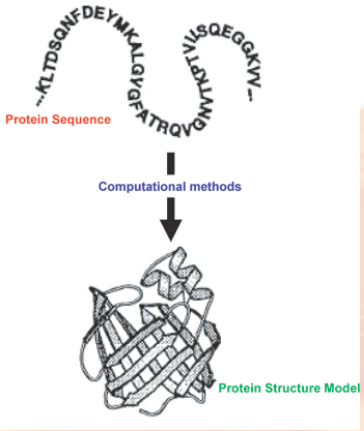
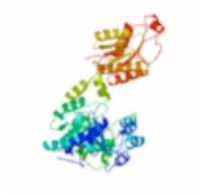
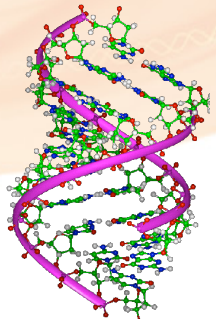




Bioinformatics Computational Methods 1 - BIOL 6308



September 5th 2013

http://155.33.203.128/cleslin/home/teaching6308F2013.php

Lecture 1 - Introduction to Bioinformatics

Goals for the Course

- Introduction to main issues in computational biology
- Opportunity to interact with algorithms, tools, and data in current practice
- Learn Unix/Linux command line interface
- Basic Perl programming –
  - Power scripting
  - Understand how to script to automate bioinformatics applications
- Understand Major Databases
  - NCBI - Gathering sequence information from NCBI, Entrez, and more
  - SWISS-PROT, EMBL, PDB, and more
- Generation of Scoring Matrices
- Position specific scoring matrices
- Log-likelihood Ratio (LLR) and Log-odds
- Sequence Logos and Information Content
- Pairwise Sequence Alignment and Distance measures
- BLAST , PsiBLAST and FASTA and K-mers
- Chance present to your Peers

Grading

- Lab work – 15 points per - show up get the points (40%)
- Pop Quiz – 10 points each (20%)
- Journal Club – 60 points (20%)
- Final Exam – written – 60 points (20%)

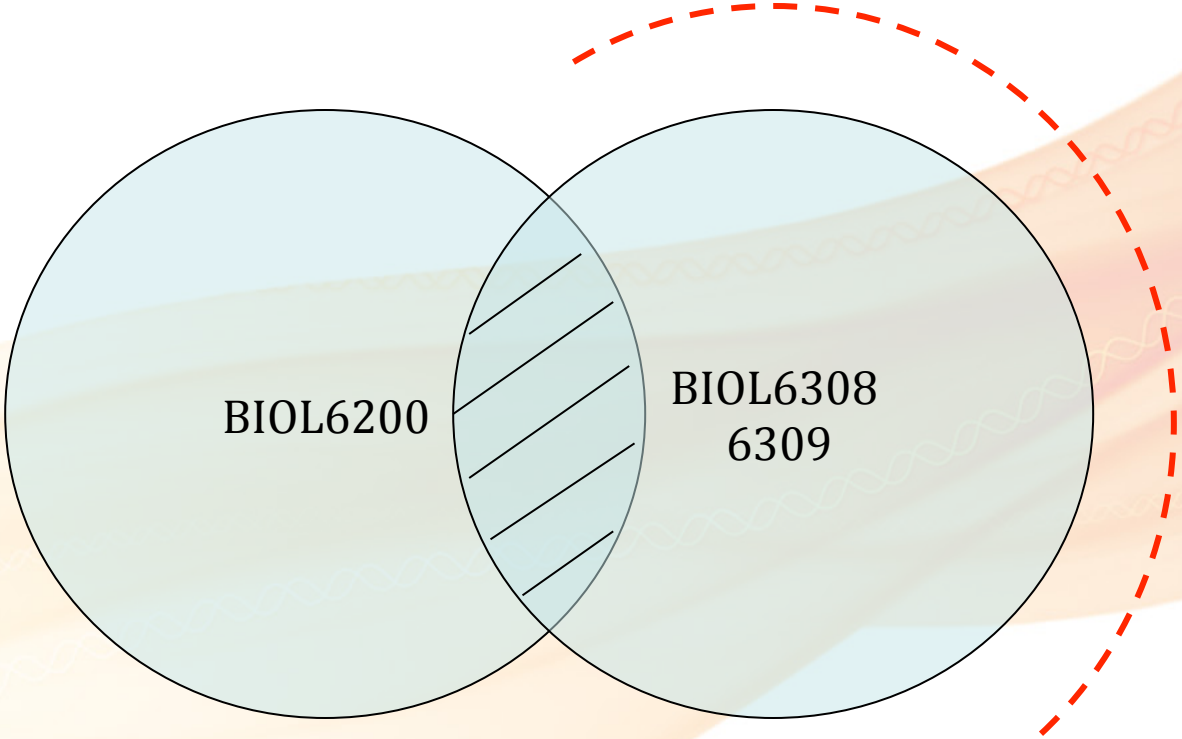
Grading

- **Lab work - 15 points X 8 = 150 points**
- **Pop quizzes - 10 points X 7 = 70 points**
- **Journal Club - 60 points X 1 = 60 points**
- **Final exam - 60 points X 1 = 60 points**
- **~Total 340 points**

Journal Club

- 20 minute presentation
  - Bioinformatics Article
  - From NAR Database issue >=2010 or newer
    - Software, Algorithm, Database, Sequence/Structure Analysis, or Phylogeny
- Must be **approved**
- This should be fun, not a task
- Must get accustomed presenting material, and interacting in discussions:
  - Show your work to others
  - Industry **requires** strong presentation skills
- Graduate education is all about:
  - Reading, Reading, and more Reading
  - Stay current with new algorithms, databases, and theories

Where Does Course Fit?



-Unix/Linux,
-Programming skills used in industry-level positions

-Unix/Linux,
-Power Scripting,
-Methods used in bioinformatics
-Theory behind the Methods
-Running & downstream Analyses of bioinformatics software

Computer Programming

- Programming exercises focus on solving bioinformatics problems with a computer's help
- Examples will be in Perl
  - Perl steep learning curve
    - High level programming language
    - Good for biologists
    - Quickly learn bad habits!!
      - Bad for biologists
    - ***You will learn more advanced Perl programming for bioinformatics in BIOL6200 next year***
  - Syntax is important to practice
  - Perl is known as "glue"
- In both industry/academics
  - New projects are always be initiated
  - Major need for Bioinformaticians
  - Who can & Who can't world

What is Bioinformatics?

- (Molecular) Bio-informatics
- Any ideas for a definition?
- Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale**
- Bioinformatics is a practical discipline with many **applications**

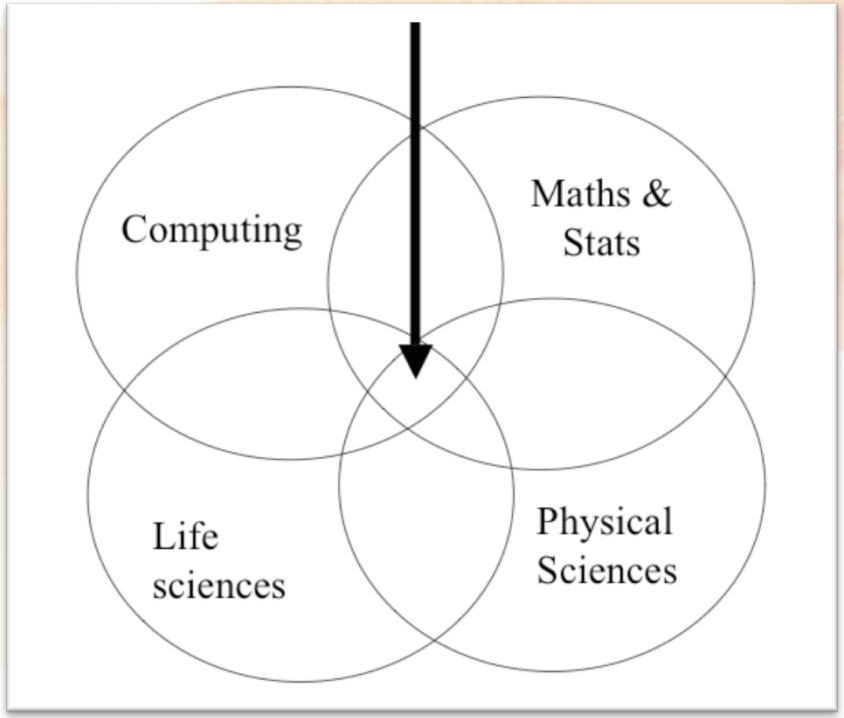
M. Gerstein

Another Definition

- **Molecular Bioinformatics** involves the use of computational tools to discover new information in complex data sets
  - From the one-dimensional information of DNA
    - Genomics
  - Through the two-dimensional information of RNA and the three-dimensional information of proteins
    - Proteomics
  - To the four-dimensional information of evolving living systems
    - Systems Biology

Bioinformatics in Context

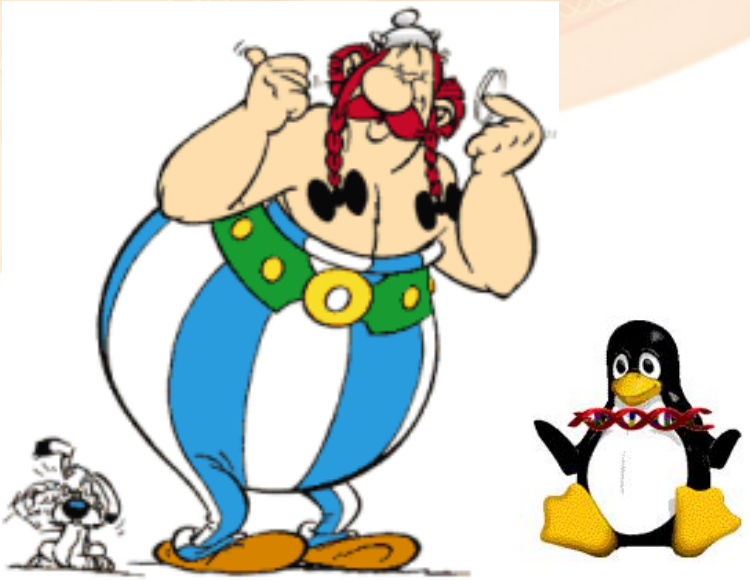
So, I ask you is Bioinformatics a New Discipline?



Profile of a Bioinformatician

- (In-Depth) knowledge of biology and genome sciences
- Translation biology <-> informatics
- Knowledge of Unix-based operating systems
- Programming skills (Java, Perl/Python, Shell, and R)
- (Distributive/Parallel) computing environments
- Data storage and database technology
- Statistics
- Mathematics

Remember without the skills, you'll have the resources and tools, but they can't be used



Adapted from Richter et al (2009) PLoS computational biology

Networking in Bioinformatics

- Internet enables sharing of genomic data among researchers worldwide
- **Encourages easy sharing and access of data by biologists**
- Large public domains of data assist biologists with research
- **In order to become a proficient bioinformatician**
  - **You must learn to transverse the network!**
  - **How to link data**
  - **How to conduct** Hypothesis-driven tests

Adopted from Carlos Moreno

Major Research Areas in Bioinformatics

- Sequence Analysis
- Genome Annotation
- Computational Evolutionary Biology
- Literature analysis
- Analysis of:
  - Gene Expression
  - Regulation
  - Protein Expression
  - Mutations in Cancer
- Prediction of Protein Structure
- Comparative Genomics
- Network and Systems biology
- Computational Genetics
- High-throughput Image Analysis
- Protein-protein docking

Nobody is a master of all these areas……

http://en.wikipedia.org/wiki/Bioinformatics

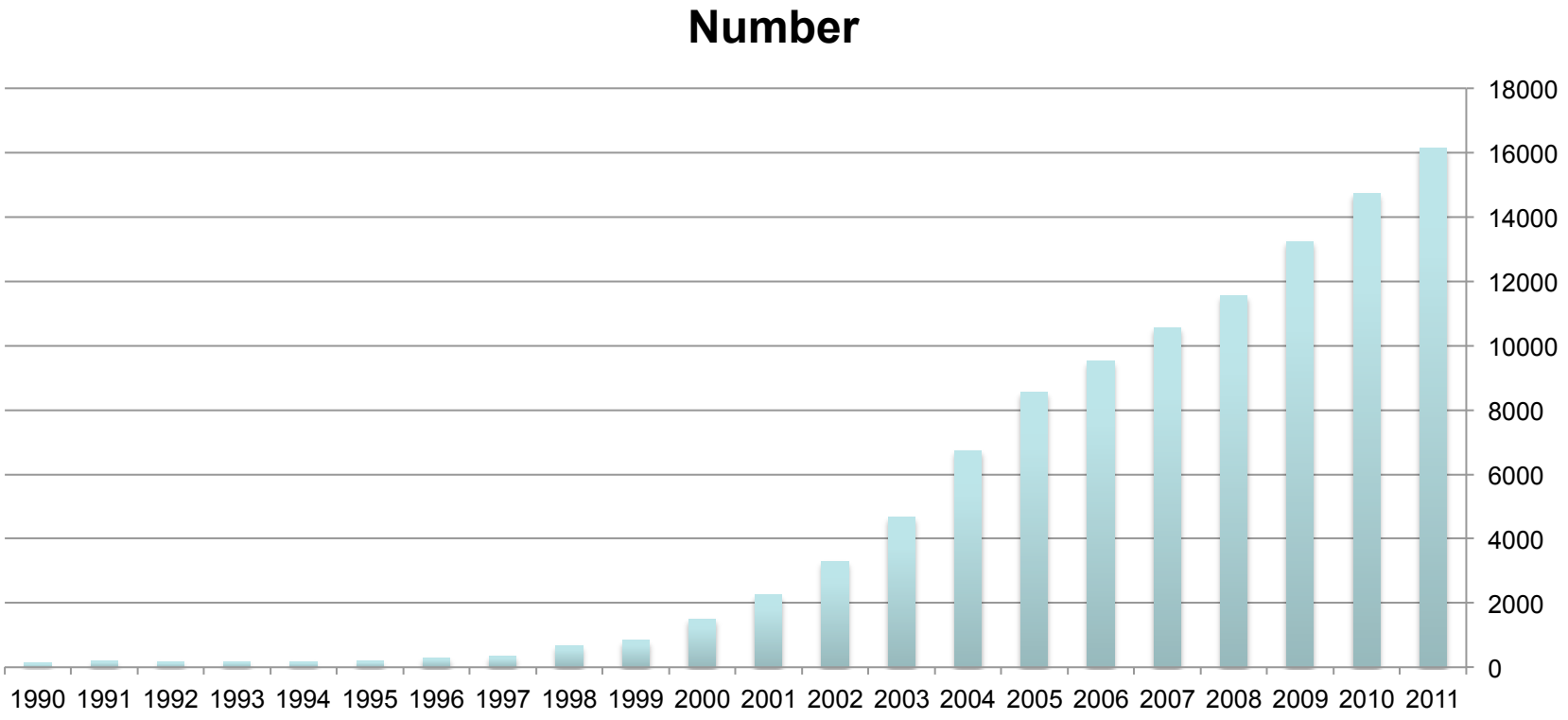
History of Bioinformatics

1965 Margaret Dayhoff's Atlas of Protein Sequences
1970 Needleman-Wunsch algorithm (global alignment)
1977 DNA sequencing and software to analyze it (Staden)
1981 Smith-Waterman algorithm developed (local sequence alignment)
1981 The concept of a sequence motif (Doolittle)
1982 GenBank made public
1983 Sequence database searching algorithm (Wilbur-Lipman)
1987 Perl (Practical Extraction Report Language) is released by Larry Wall.
1988 National Center for Biotechnology Information (NCBI) created at NIH/NLM
1990 BLAST: fast sequence similarity searching
1990 The HTTP 1.0 specification is published. First HTML document.
1994 EMBL European Bioinformatics Institute (EBI), Hinxton, UK
1995 Microsoft version 1.0 of IE. Sun version 1.0 of Java. Version 1.0 of Apache.
1997 PSI-BLAST
1997 International Society for Computational Biology was founded
1998 Worm (multicellular) genome completely sequenced
2000 Gene Ontology (GO)
2001 The human genome (3 Giga base pairs) is published.
2003 myGrid: personalised bioinformatics on the information grid (e.g, Taverna).
2004 Bioconductor: open software development for computational biology and bioinformatics
2005 Reactome: knowledge base of biological pathways

Barbera van Schaik

Medline Trend for Bioinformatics

Get Published!



http://dan.corlan.net/medline-trend.html

Barbera van Schaik

What is Done in Bioinformatics

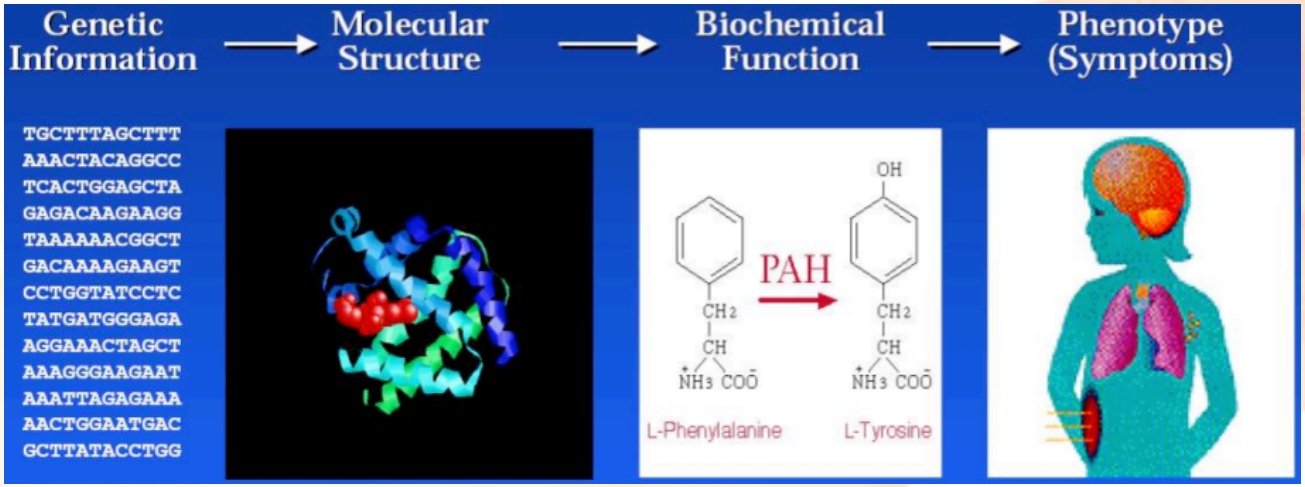
- Understand how living things function, in order to improve the quality of life
- Analysis and interpretation of various types of biological data:
  - Nucleotide sequences
  - Amino acid sequences
  - Protein Domains
  - Protein Structures
- Development of:
  - New algorithms and statistics to assess biological information
  - Tools to enable efficient access and management of the different types of information

Challenges of Working in Bioinformatics

- Need to feel comfortable in interdisciplinary area
- Communication is critical, since you'll be talking with experimentalists
  - Need background in molecular
  - Have be able to understand what's happening in lab
- Depend on others for primary data
- Need to address important biological *and* computer science problems
- Many open problems

Motivation?

- "Biology easily has 500 years of exciting problems to work on." - Donald Knuth



- By developing techniques for analyzing sequence data and related structures, we can attempt to understand the molecular basis of life

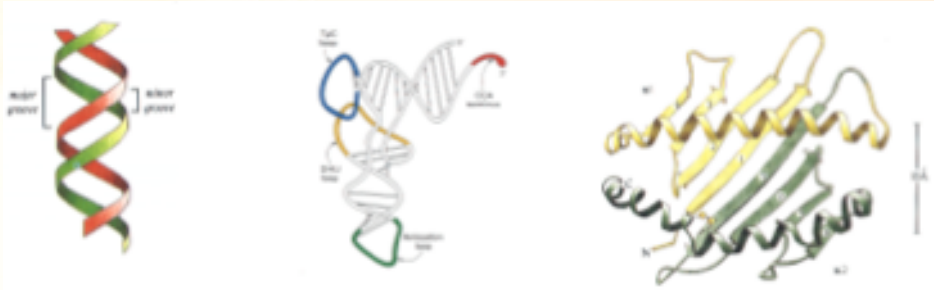
Why use Bioinformatics?

- Find an answer quickly
  - Most *in silico* biology is faster than *in vitro*
  - I'm able simulate millions different….
- Massive amounts of data to analyze
  - Need to make use of all the information
  - Not possible to do analysis by hand
  - Can't organize and store information using only a lab notebook
  - Automation is key
- However - **All results generated by computer analysis should be verified by biologists**

Molecular Biology as an Information Science

What is the Information?

- Central Dogma of Molecular Biology
  - DNA
    - RNA
      - Protein
        - Phenotype
- Molecules
  - Sequence and Structure



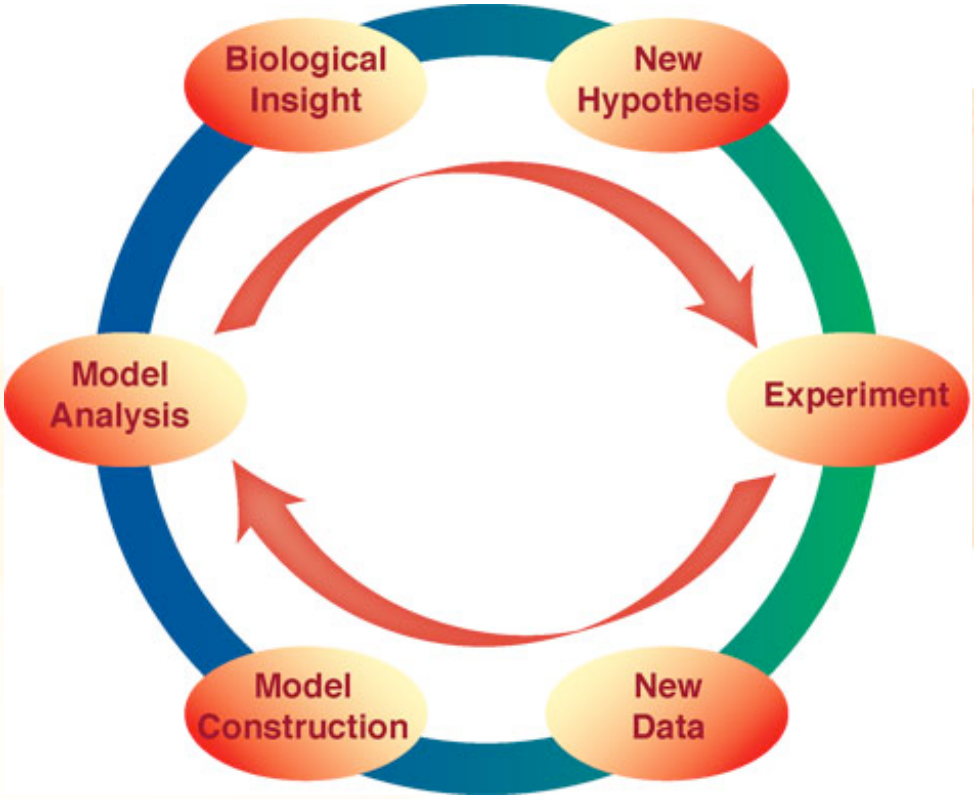
- Central Paradigm for Bioinformatics
  - Genomic Sequence Information
    - mRNA
      - Protein Sequence
        - Protein Structure
          - Protein Function
            - Phenotype
- Large Amounts of Information
  - Standardized
  - Gather Statistics

Must Model

Why Model?

- To understand biological/chemical data
- To share data we need to be able to search, merge, & check via model
- Integrating diverse data types can reduce random & systematic errors
- Design useful modifications and new hypothesis
  - Leads to new:
    - Experiments
    - Data
    - Models
  - Perpetual cycle
    - What do I mean??

Why Model?



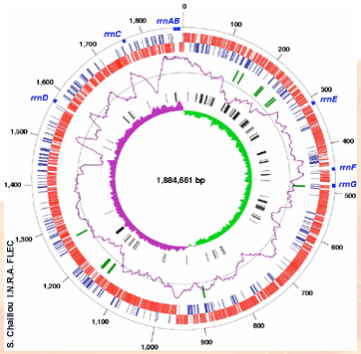
Where does bioinformatics fit

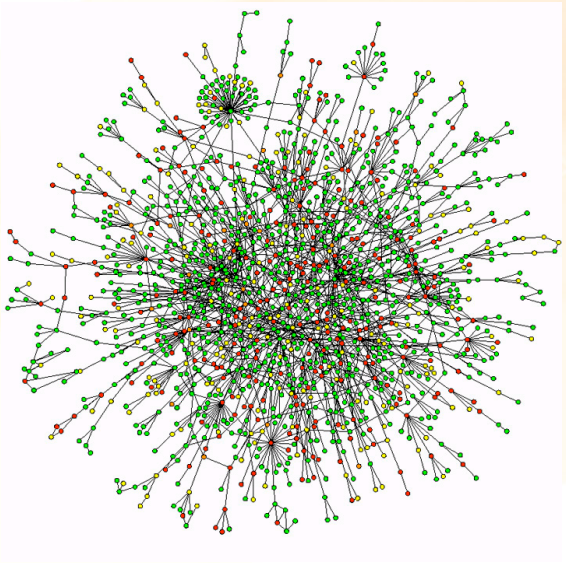
Bioinformatics Methods and Algorithms

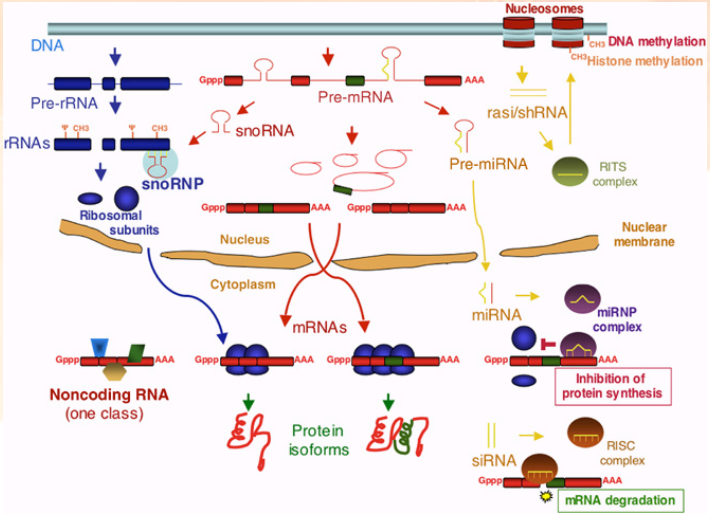
- **Units and Sources of Biological Data for bioinformatics**
- Types of the Data
- Data storage, retrieval and visualization
- General Types of “Informatics Techniques in Bioinformatics”
- Data mining

Basic Units of Biological Data used in Bioinformatics

- DNA (Genome)
- RNA (Transcriptome)
- Protein (Proteome)



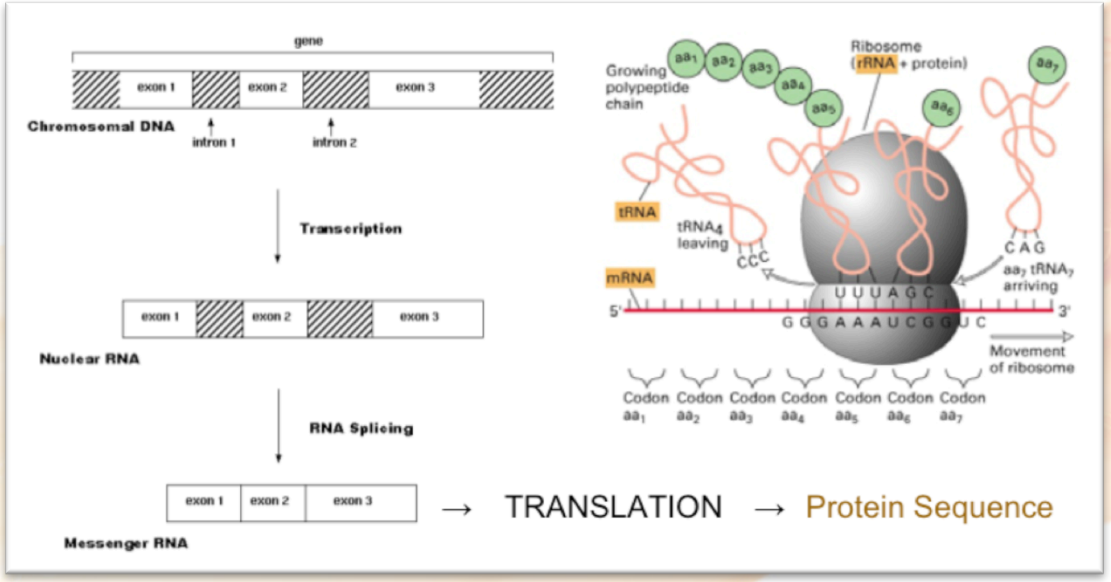




How are these basic units related?

Central Dogma of Molecular Biology?

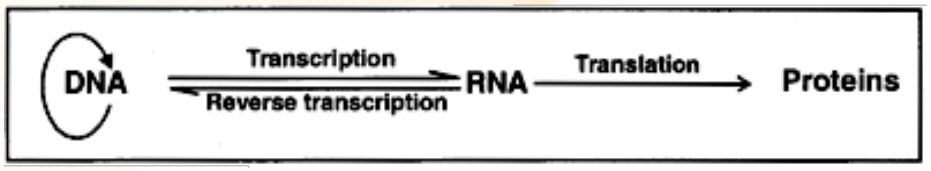
Central Dogma?



- Why have I put a question mark above?

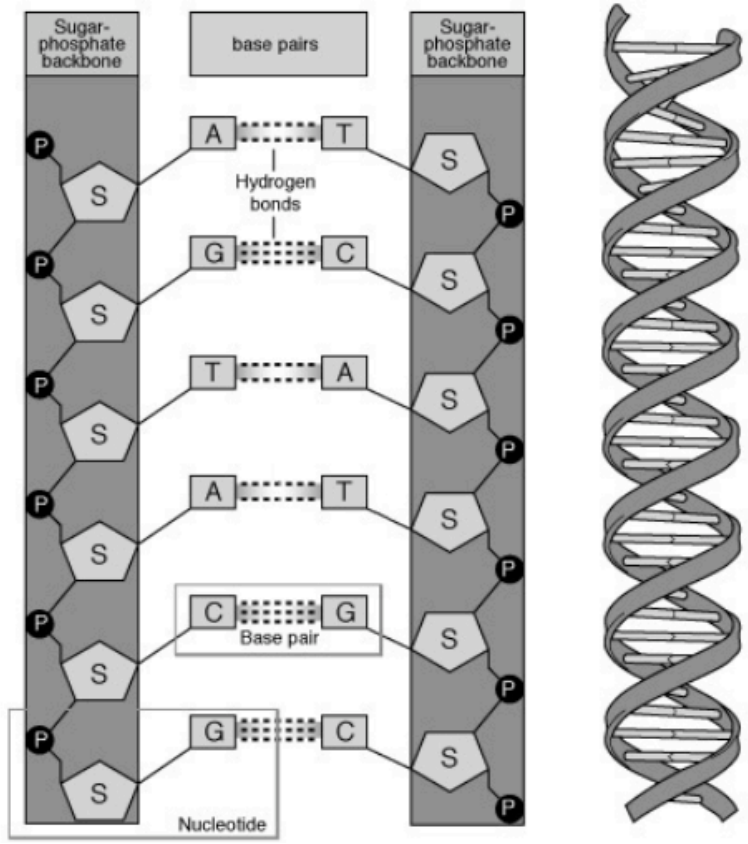
Modified Dogma:

- 1970, D. Baltimore *et al*. brought to light an important modification of this information flow
- Many tumor viruses contain RNA as genetic material
  - Replicate by first synthesizing a complementary DNA
  - Process known: reverse transcription
    - Uses Reverse transcriptase
    - RNA dependent DNA polymerase
- These viruses are known as retroviruses and include HIV
- So the central dogma is now represented as:



Molecular Biology Information: DNA

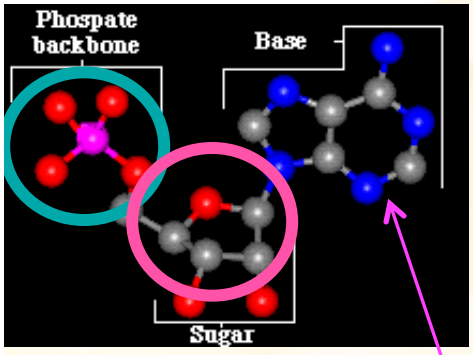
- Deoxyribonucleic acid
  - Helix formed by pairing of bases
  - Four bases
  - Two complement pairs
    - (A) adenine ~ purine
    - (T) Thymine ~ pyrimidine
    - (G) guanine ~ purine
    - (C) cytosine ~ pyrimidine
  - Location
    - Eukaryotes
      - Nucleus
    - Prokaryotes & Archaebacteria
      - chromosome & plasmids within the cytoplasm



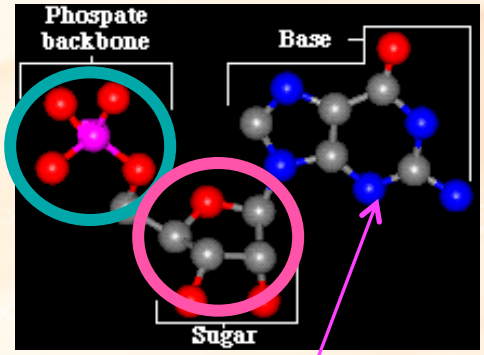
Molecular Biology Information: DNA

2 purines:

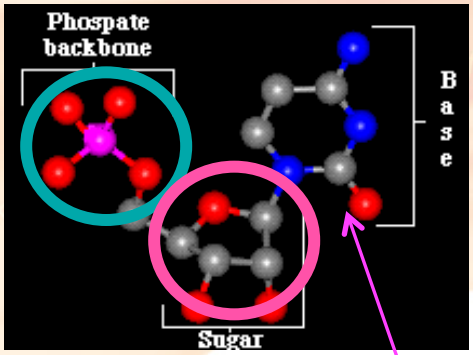
2 pyrimidines:



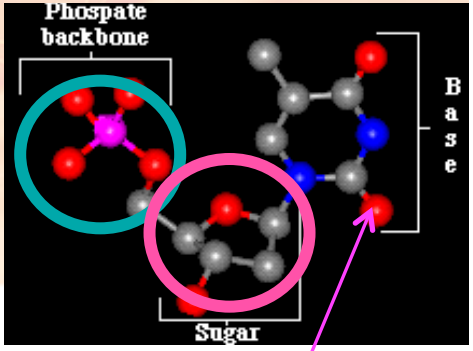
adenine (A)



guanine (G)



cytosine (C)



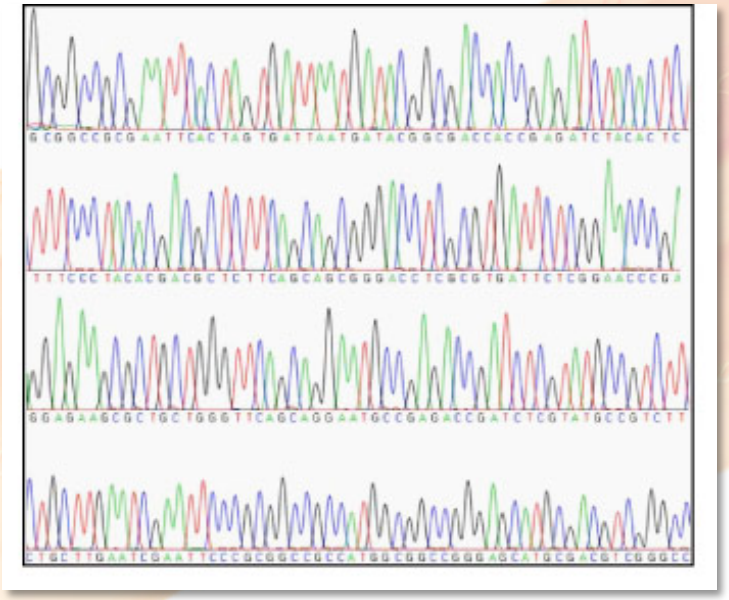
thymine (T)

two rings

one ring

Molecular Biology Information: DNA

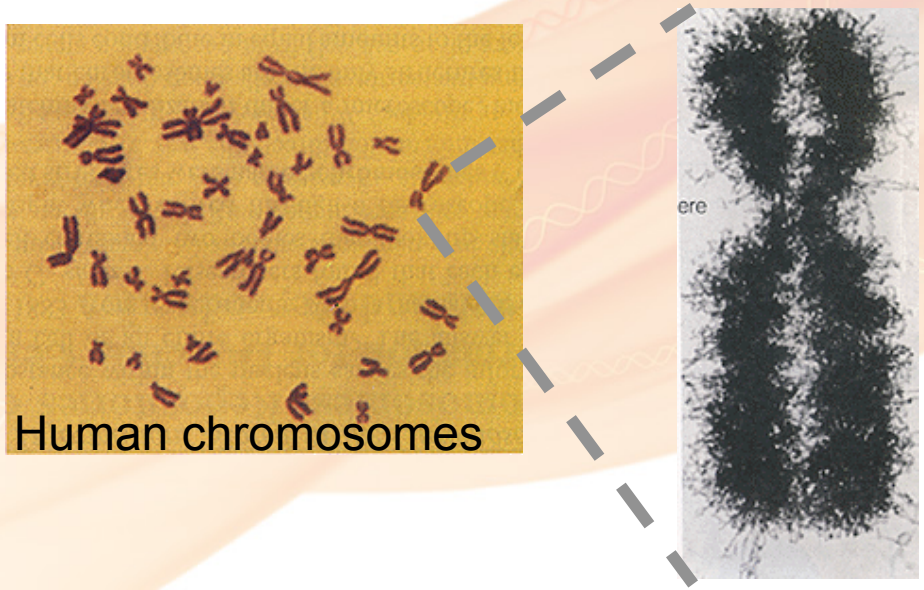
- Raw DNA Sequence
  - Coding/Non-Coding
  - Parse into Genes?
    - Location introns & exons
    - Function
  - Conserved/Non-Conserved
  - GC content
  - 4 bases:
    - 1 K in a gene
    - 10^{4}-10^{12} bases in genomes

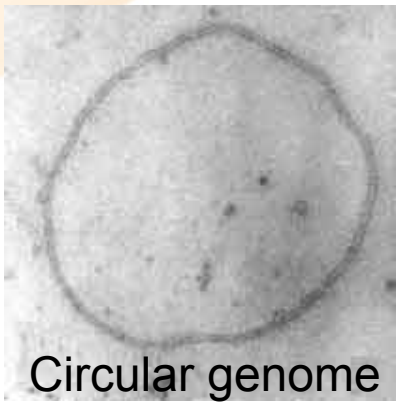


DNA – Genomes

- Entire complement of genetic material carried by an individual: genome
- Eukaryotes may have up to 3 subcellular genomes:
  - 1. Nuclear
  - 2. Mitochondrial
  - 3. Plastid

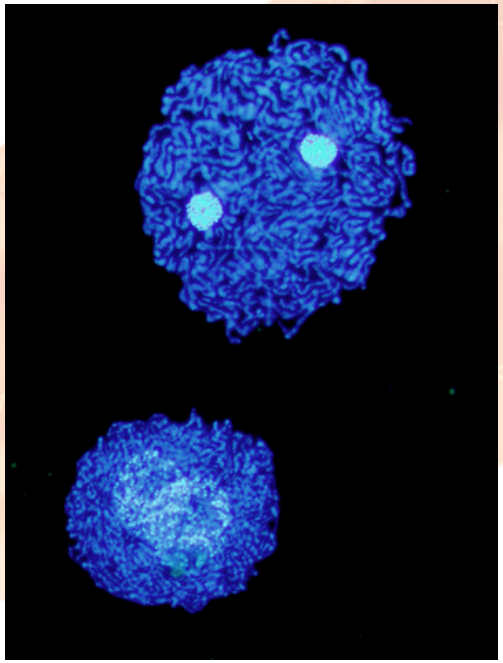
- Bacteria have either circular or linear genomes and may also carry plasmids





DNA – Genomes

- An organism's genome carries all the instructions it needs to make the proteins required for life
- Always making new discoveries
  - *Carsonella ruddii* – smallest bacterium
    - 160,000 base pairs (bp)
    - Codes for 182 proteins
      - < ½ size previously thought to be minimum necessary for life
  - Contrast that with Human Genome
    - 3 billion bp
    - ~25K proteins
- These studies not possible without Bioinformatics



Sequence Vs Structure

RTLAWYAGHLVAGAKDEFGGDFKIWYFGAID...

DFLLVAGAKDEFGKIWYFGGIDAWRTAGDCA...

HLVAGARTLAFGAIDWYAKDEFGGGDFKIWY...

ARTHLVAGFGGGAIDWYFKIWYAKLAFGDED...

GCTAGCTTAAGGCCTTCATGATCTTCTGAG...

AGGGCTCCTTCATGATAGCTTAAGGCTTAA...

AGGCCTTCATGGGGTTAACATATCTTCTGA...

CCTTCATGCTAGCTTAAGGGATCTTAACCG...

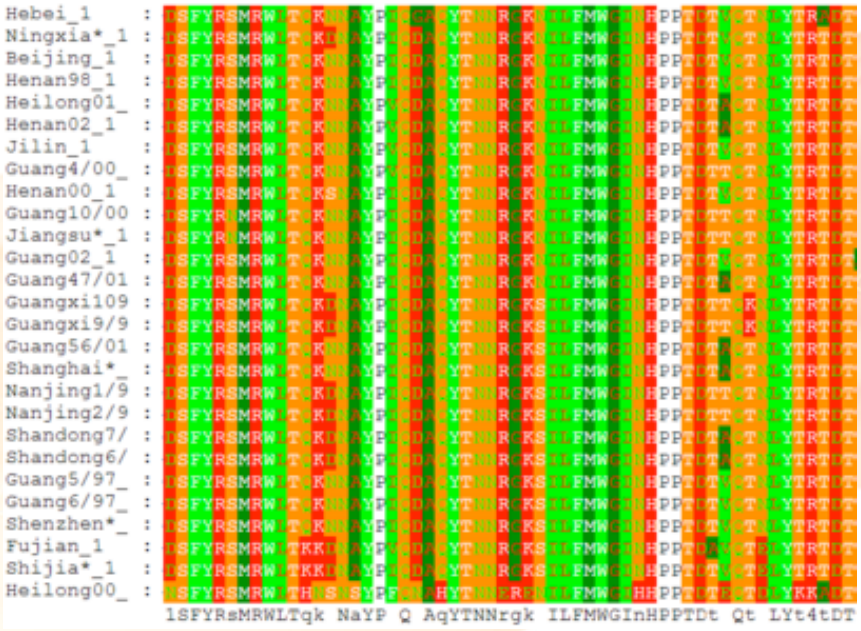
Simple and Linear, with 4 bases

Complex and intricate, with 20 residues

Molecular Biology Information: Protein Sequences

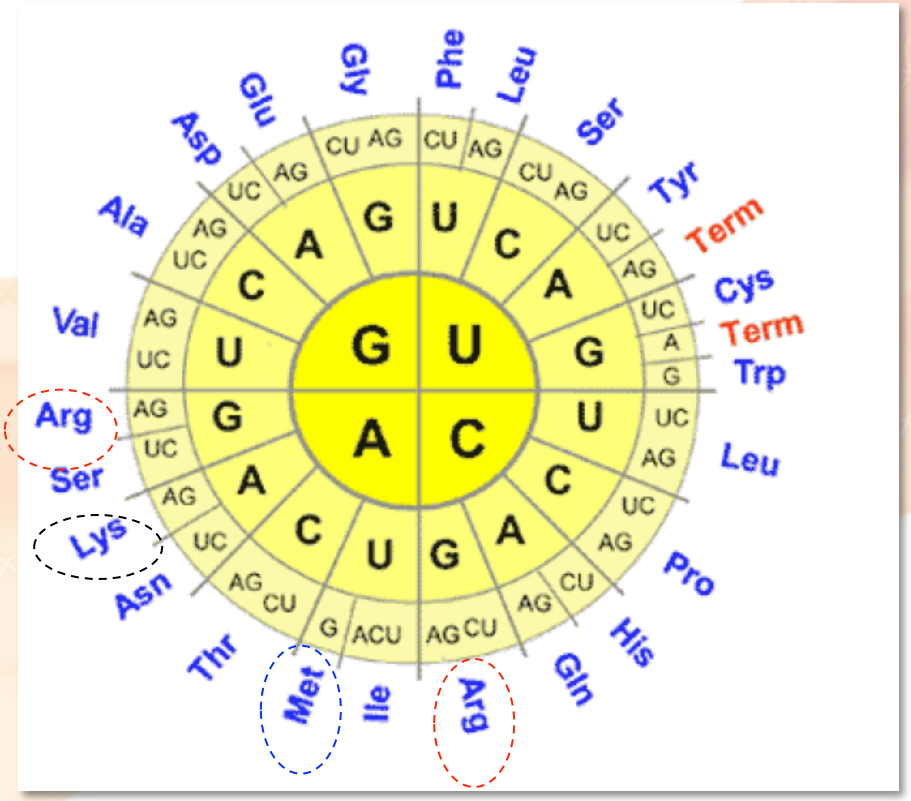
- Organic compounds made up of amino acids (a.a.)
- Also known as polypeptides
- Sequence of a.a. defined by the sequence of a gene (mRNA)
- 20 letter alphabet
  - Some organisms have more (selenocysteine)
  - ACDEFGHIKLMNPQRSTVWY
  - No: **BJOUXZ**
  - ~300 a.a. in average protein (bacteria)
  - ~50-200 a.a. in a protein domain
  - ~12,000,000 known protein sequences
- Shortly after or even during synthesis, the residues in a protein are often chemically modified by post-translational modification

Aligned Protein Sequences



Alphabets of Life

- 00=A
- 01=C
- 10=C
- 11=T (U)



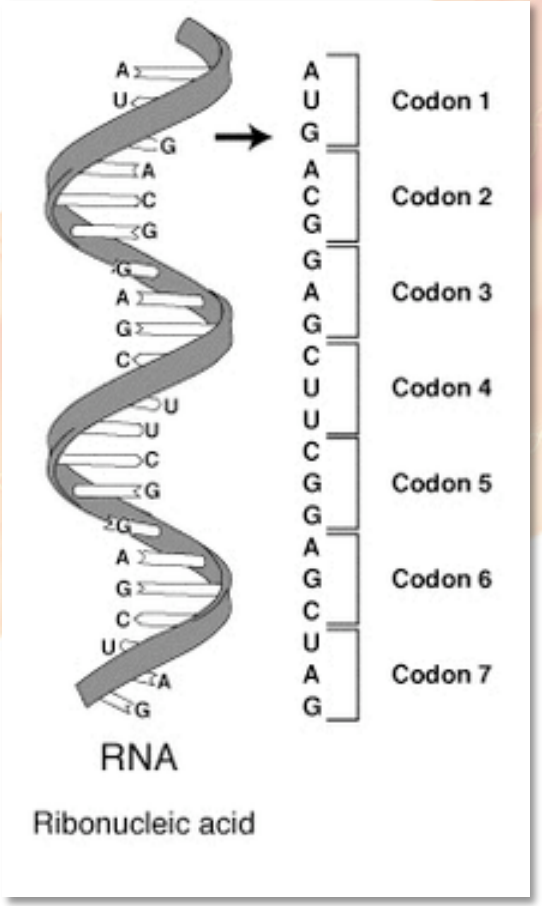
4 DNA letters decode 20 amino acid letters

Degeneracy = 4^4 = 64, but only 20 amino acids

Degeneracy?

Genetic Code - Codons

- Set of rules by which information encoded in genetic material
  - DNA or mRNA sequences are translated into proteins (amino acid sequences) by living cells
- The code defines a mapping between tri-nucleotide sequences
  - Called codons, and amino acids



Degeneracy of the Genetic Code - Codons

- Redundancy, but no ambiguity
  - GAA and GAG -> glutamic acid (redundancy)
  - Neither specifies any other amino acid (no ambiguity)
- Codons may differ in any of their three positions
- Codon position is non-degenerate if any mutation at this position results in amino acid substitution

Types Degeneracy

- Fourfold degenerate site
  - Third position of the glycine codons (GGA, GGG, GGC, GGU)
- Two and threefold
  - Third position of glutamic acid codons (GAA, GAG)
    - The equivalent nucleotides are always either two purines (A/G) or two pyrimidines (C/U)
    - Only **transversional** substitutions (purine to pyrimidine or pyrimidine to purine) in twofold degenerate sites are **nonsynonymous**
  - Only **one** threefold
    - isoleucine codon: AUU, AUC, or AUA all encode isoleucine
      - AUG encodes methionine

Examples of notable Mutations

ΔF508 deletion in cystic fibrosis

1st base \ 2nd base	U	C	A	G
U	UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine
	UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine
	UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (Stop)	UGA Opal (Stop)
	UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (Stop)	UGG (Trp/W) Tryptophan
C	CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine
	CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine
	CUA (Leu/L) Leucine	CCA (Pro/P) Proline	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine
	CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine
A	AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine
	AUC (Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine
	AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
	AUG (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
G	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine
	GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine
	GUA (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine
	GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine

β-Thalassemia

McArdle's disease

β-Thalassemia

Prostate cancer

Colorectal cancer

Sickle-cell disease

Fragile X Syndrome

Huntington's Disease

Selection of notable mutations, ordered in a standard table of the genetic code of amino acids.

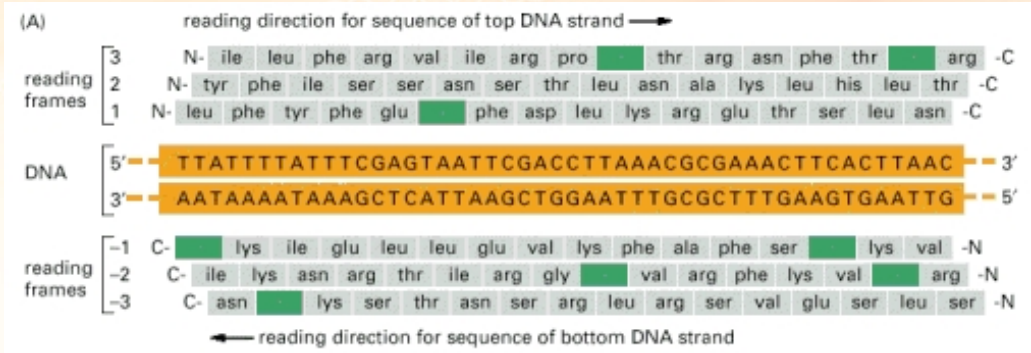
Clinically important missense mutations generally change the properties of the coded amino acid residue between being basic, acidic, polar or nonpolar, while nonsense mutations result in a stop codon.

Amino acids
- Basic
- Acidic
- Polar
- Nonpolar (hydrophobic)

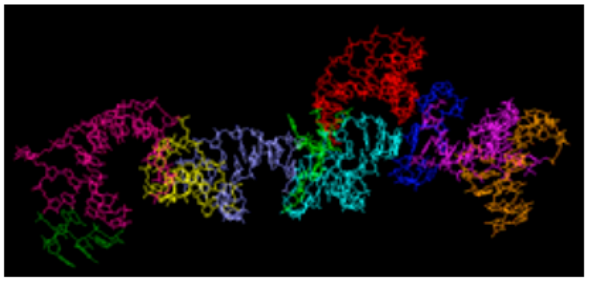
Mutation type
- = Insertion
- = Deletion
- = Missense
- = Nonsense

Its Even More Complicated

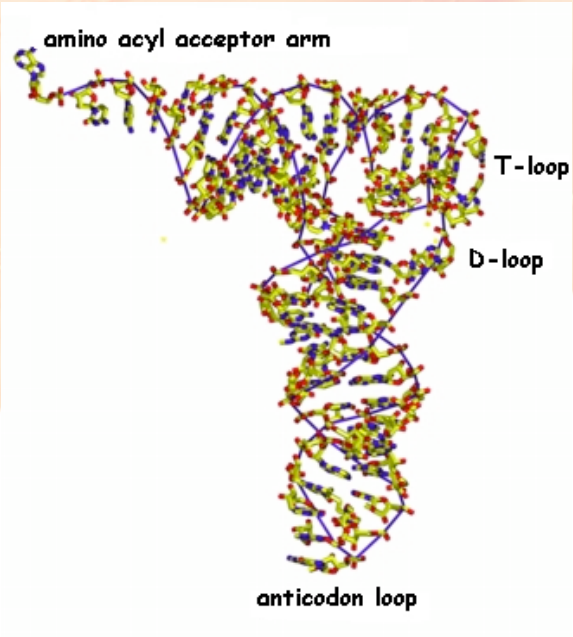
- Any:
  - Region of DNA can, in principle, code for 6 different amino acid sequences
  - One of 3 different reading frames from two strands
- DNA always:
  - Read 5'-to-3' direction
  - Encodes a polypeptide from the amino (N) to the carboxyl (C) terminus



Molecular Biology Information: RNA

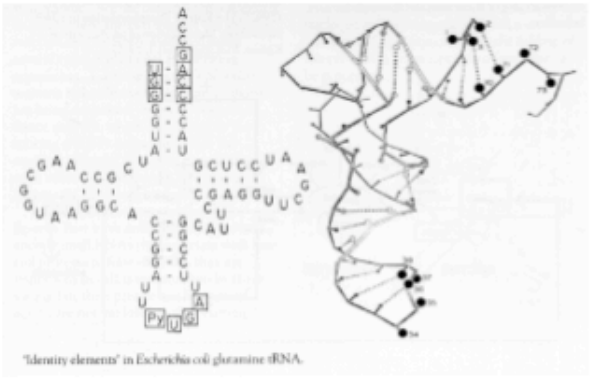


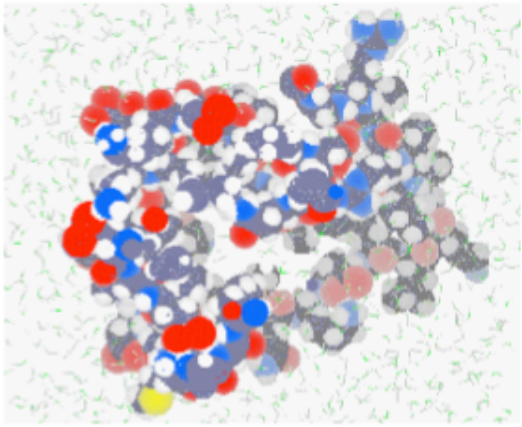
- Ribonucleic Acid
  - Formed with ribose not 2'-deoxyribose as sugar
  - Does not form double helix
  - Can have complicated 3D structure
  - Takes on different forms and functions
    - mRNA - messenger RNA
      - Is the transcribed signal that travels to the ribosome for translation
    - tRNA - transfer RNA
      - Carries amino acid to ribosome
    - rRNA - ribosomal RNA
      - Combines with proteins to form the ribosome
    - sRNA - small RNA
      - Facilitates other functions within the cell

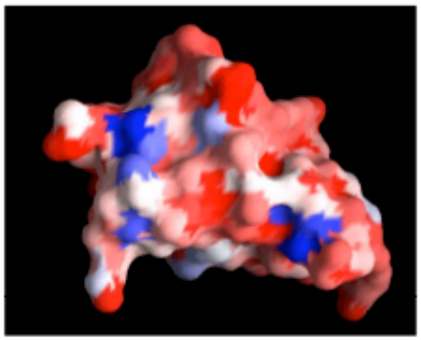


Molecular Biology Information: Macromolecular Structure

DNA/RNA/Proteins

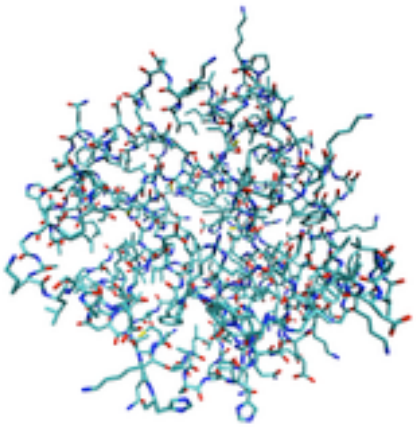




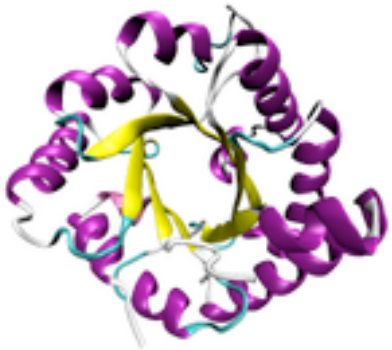


Molecular Biology Information: Macromolecular Structure

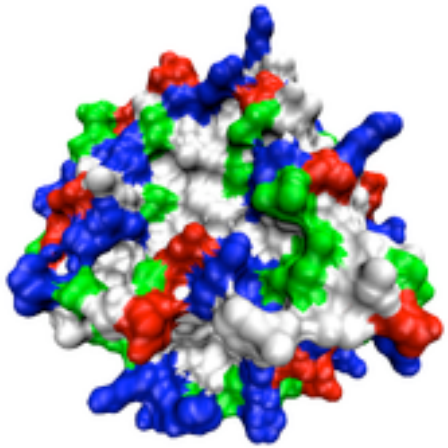
Three possible representations of the three-dimensional structure of the protein triose phosphate isomerase



All-atom representation colored by atom type.

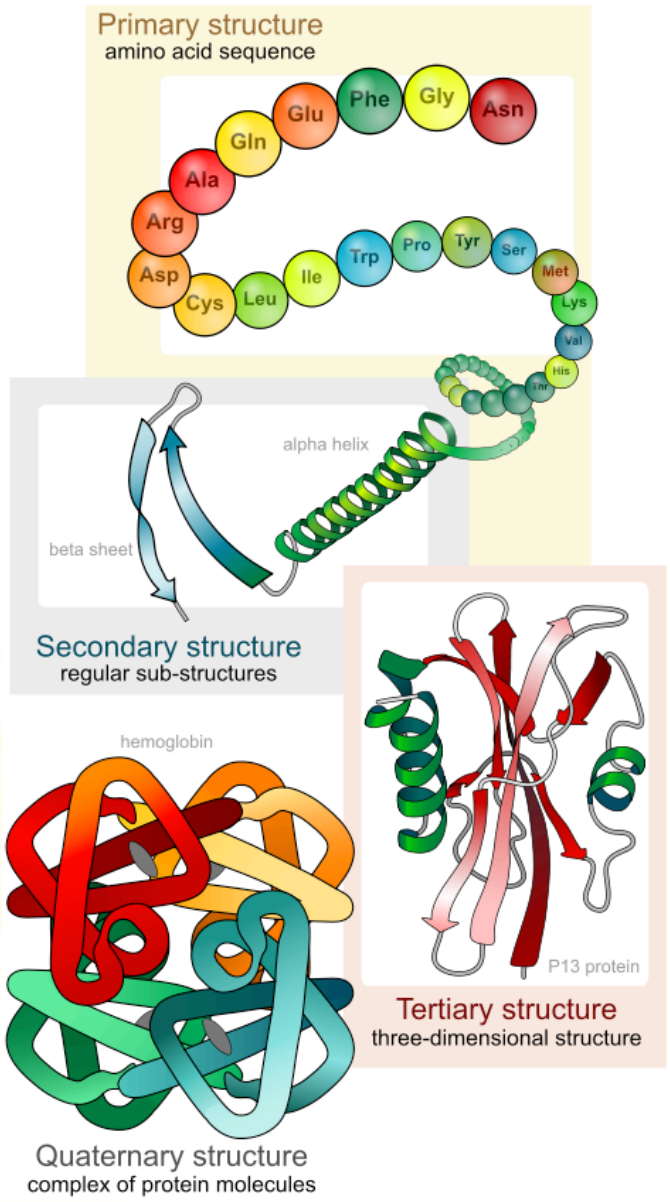


Simplified representation illustrating the backbone conformation, colored by secondary structure



Solvent-accessible surface representation colored by residue type

Molecular Biology Information: Levels of Protein Structure

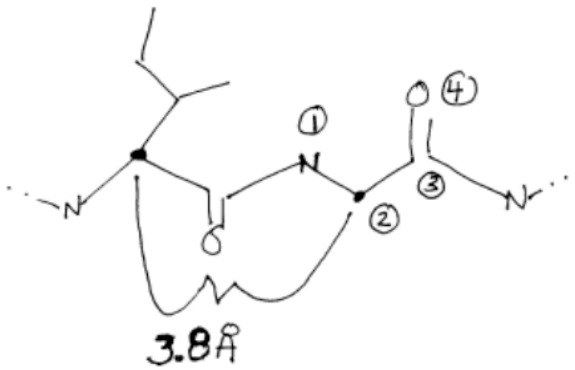


- Primary
  - Refers to the sequence of the different amino acids of the protein
- Secondary
  - Refers to highly regular local sub-structures
- Tertiary
  - Refers to 3-D structure of a single protein molecule
- Quaternary
  - Larger assembly of several protein molecules or polypeptide chains

Molecular Biology Information: Protein Structure Details

- Statistics on Number of XYZ triplets
  - 200 residues/domain
  - 200 CA, separated by 3.8 Angstroms
  - Average residue is Leu:
    - 4 backbone atoms + 4 sidechains atoms
    - 150 cubic Angstroms
      - 1600 xyz triplets per protein domain

```
ATOM      1  C   ACE     0       9.401  30.166  60.595  1.00 49.88      1GKY
 67
ATOM      2  O   ACE     0      10.432  30.832  60.722  1.00 50.35      1GKY
 68
ATOM      3  CH3 ACE     0       8.876  29.767  59.226  1.00 50.04      1GKY
 69
ATOM      4  N   SER     1       8.753  29.755  61.685  1.00 49.13      1GKY
 70
ATOM      5  CA  SER     1       9.242  30.200  62.974  1.00 46.62      1GKY
 71
ATOM      6  C   SER     1      10.453  29.500  63.579  1.00 41.99      1GKY
 72
ATOM      7  O   SER     1      10.593  29.607  64.814  1.00 43.24      1GKY
 73
ATOM      8  CB  SER     1       8.052  30.189  63.974  1.00 53.00      1GKY
 74
ATOM      9  OG  SER     1       7.294  31.409  63.930  1.00 57.79      1GKY
 75
ATOM     10  N   ARG     2      11.360  28.819  62.827  1.00 36.48      1GKY
 76
```



Molecular Biology Information: Protein Structure Classification

- Structure Classification
- Gold Standard - CATH Database

CATH Version 3.4

Based on PDB release: Nov 13, 2010

Changes since v3.3:

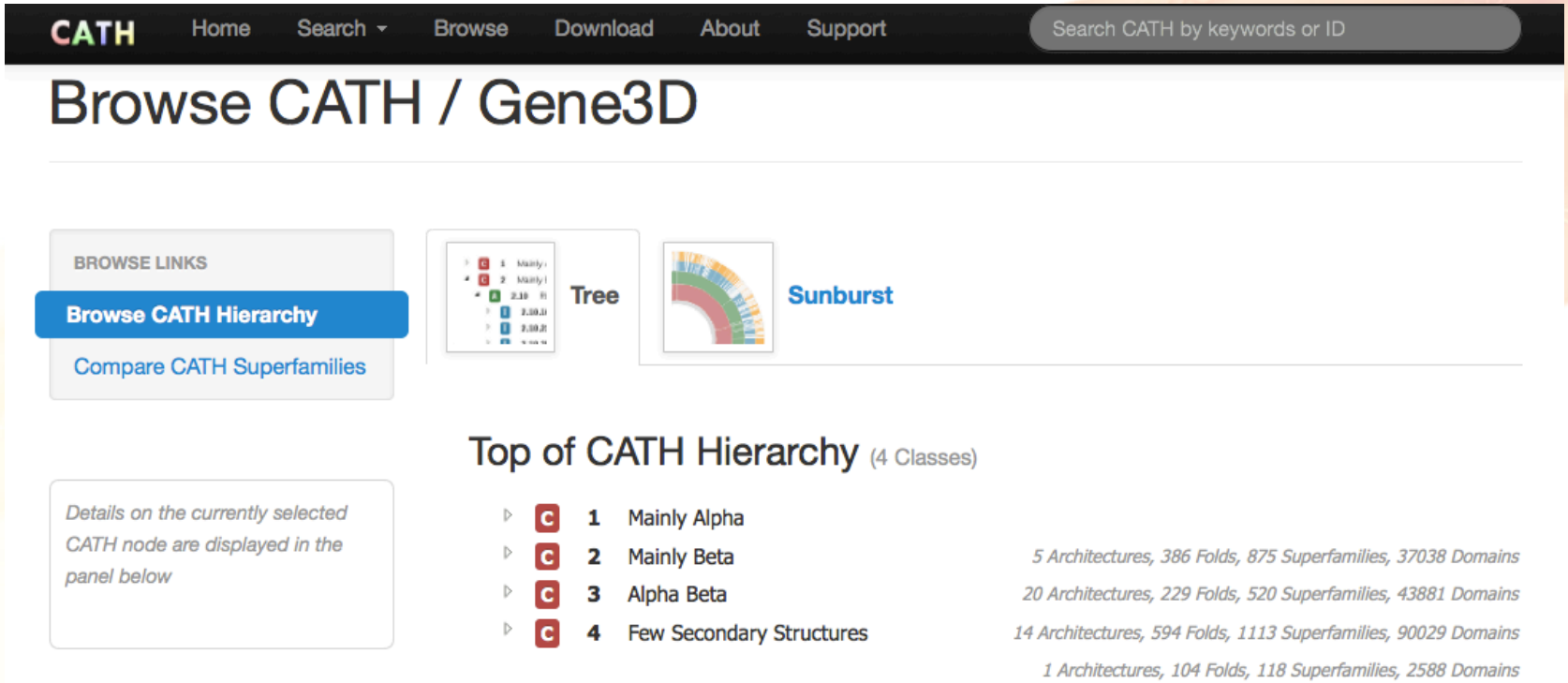
- 24,232 newly assigned domains
- 163 new homologous superfamilies
- 49 new folds (topologies)

The table below summarises the number of clusters within each of the four classes in CATH.

Class	Architecture	Topology	Homologous Superfamily	S35 Family	S60 Family	S95 Family	S100 Family	Domains
1	5	376	839	2763	3571	4679	9217	32396
2	20	228	514	2514	3573	5668	9824	39140
3	14	577	1082	5849	8381	10626	21900	79038
4	1	101	114	204	253	352	547	2346
Total	**40**	**1282**	**2549**	**11330**	**15778**	**21325**	**41488**	**152920**

Molecular Biology Information: Protein Structure Classification

- Structure Classification
- Gold Standard - CATH Database



http://www.cathdb.info/browse/browse_hierarchy_tree

Molecular Biology Information: Whole Genomes

- The Revolution Driving Everything

  Venter et al. (1995) "Whole-genome random sequencing and assembly of Haemophilus influenzae ." *Science* 269: 496-512.

- Integrative Data

  1995:

  HI (bacteria): 1.6 Mb & 1600 genes done

  1997:

  yeast: 13 Mb & ~6000 genes for yeast

  1998:

  worm: ~100Mb with 19 K genes

  1999:

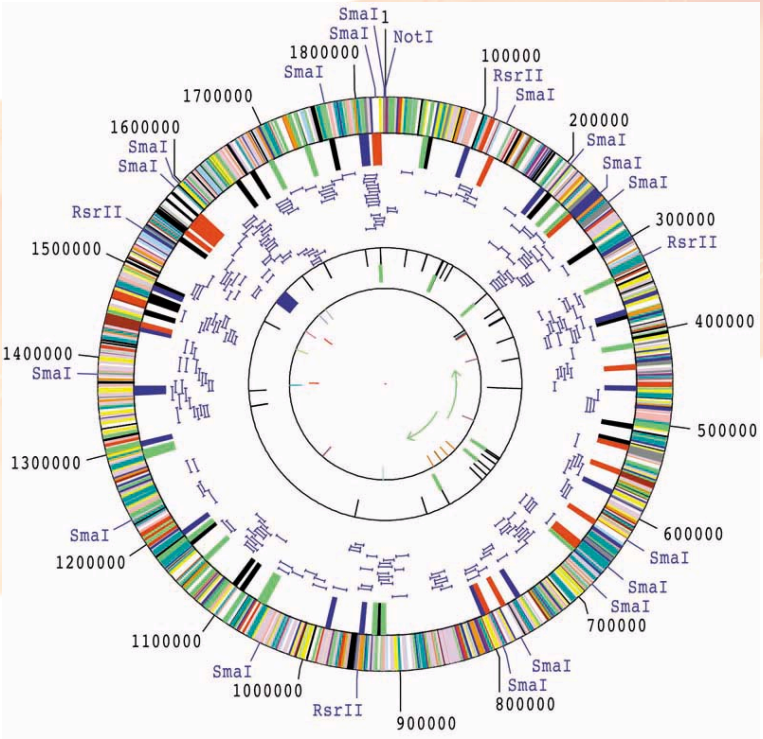
  >30 completed genomes!

  2003:

  human: 3 Gb & 100 K genes

  Today (~2010):

  156 Eukaryotes, 1917 Bacteria, 133 Archaea



Molecular Biology Information: Whole Genomes

- The Revolution Driving Everything

  Venter et al. (1995) "Whole-genome random sequencing and assembly of Haemophilus influenzae ." *Science* 269: 496-512.

- Integrative Data

  1995:

  HI (bacteria): 1.6 Mb & 1600 genes done

  1997:

  yeast: 13 Mb & ~6000 genes for yeast

  1998:

  worm: ~100Mb with 19 K genes

  1999:

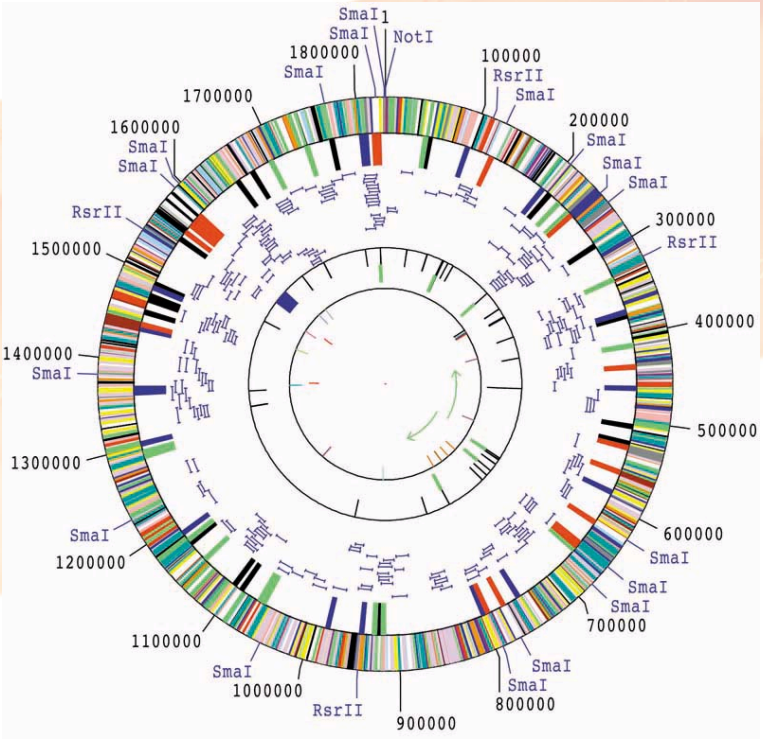
  >30 completed genomes!

  2003:

  human: 3 Gb & 100 K genes

  Today:

  311 Eukaryotes, 6342 Bacteria, 227 Archaea



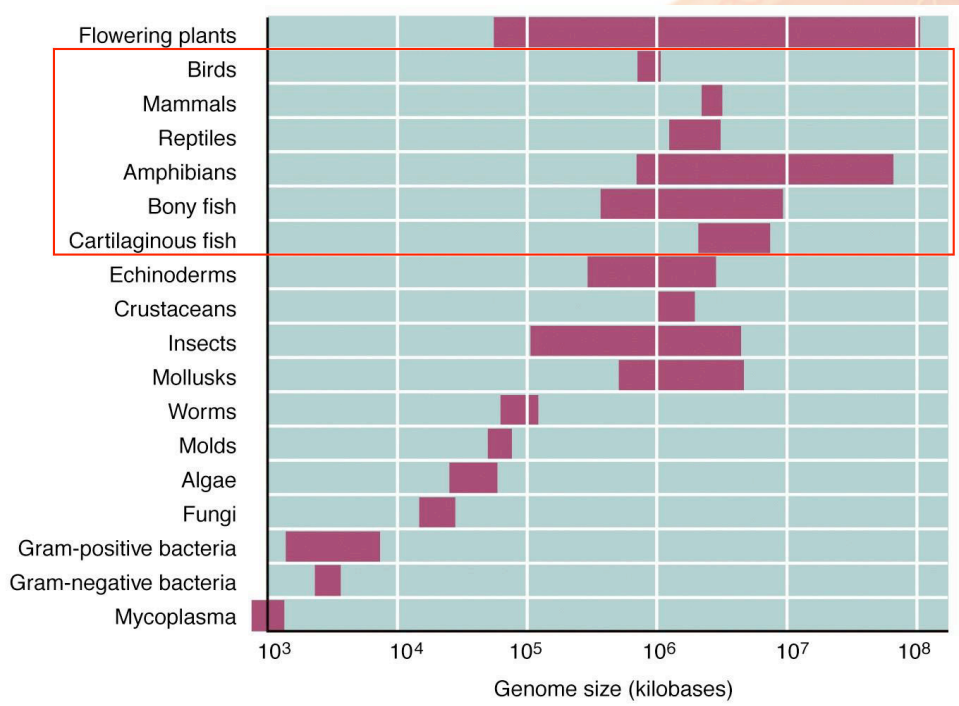
Molecular Biology Information: Whole Genomes

Bacterium	*Buchnera aphidicola*	600,000		
Bacterium	*Wigglesworthia glossinidia*	700,000		
Bacterium	*Escherichia coli*	4,600,000		[14]
Bacterium	*Solibacter usitatus* (strain Ellin 6076)	9,970,000		Largest known Bacterial genome
Amoeboid	*Polychaos dubium* (*"Amoeba" dubia*)	670,000,000,000	737	Largest known genome.[15]
Plant	*Arabidopsis thaliana*	157,000,000		First plant genome sequenced, December 2000.[16]
Plant	*Genlisea margaretae*	63,400,000		Smallest recorded flowering plant genome, 2006.[16]
Plant	*Fritillaria assyrica*	130,000,000,000		
Plant	*Populus trichocarpa*	480,000,000		First tree genome, September 2006
Moss	*Physcomitrella patens*	480,000,000		First genome of a bryophyte, January 2008 [17]
Yeast	*Saccharomyces cerevisiae*	12,100,000		[18]
Fungus	*Aspergillus nidulans*	30,000,000		
Nematode	*Caenorhabditis elegans*	100,300,000		First multicellular animal genome, December 1998[19]
Nematode	*Pratylenchus coffeae*	20,000,000		Smallest animal genome known[20]
Insect	*Drosophila melanogaster* (fruit fly)	130,000,000		[21]
Insect	*Bombyx mori* (silk moth)	530,000,000		
Insect	*Apis mellifera* (honey bee)	236,000,000		
Fish	*Tetraodon nigroviridis* (type of puffer fish)	385,000,000		Smallest vertebrate genome known
Mammal	*Homo sapiens*	3,200,000,000	3	
Fish	*Protopterus aethiopicus* (marbled lungfish)	130,000,000,000	143	Largest vertebrate genome known

base pairs

C-value Paradox

- Observation: genome size does not uniformly increase with respect to perceived complexity of organisms
  - e.g.. vertebrate vs. invertebrate animals, or "lower" versus "higher" vertebrate animals (red box)
  - Some Amphibians have more than 10-fold more DNA than do Mammals, including humans



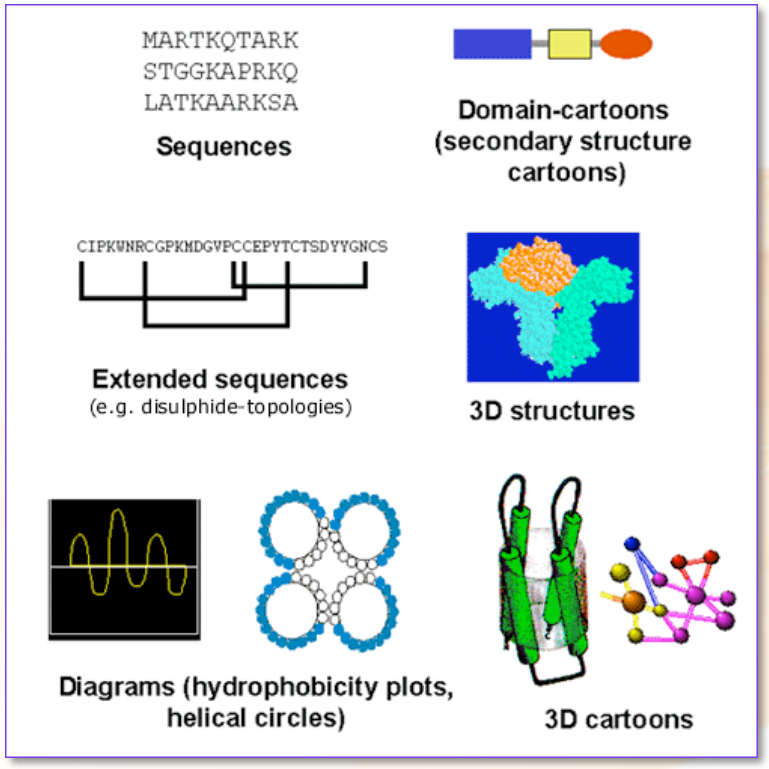
So I ask, "Is it Really a Paradox?"

- There is in fact no "paradox"
- Evolution does not proceed in a linear manner, nor is there a linear succession of organisms from "lower" to "higher"
- Despite differences in DNA content, the number of genes in any vertebrate genome is roughly similar
- Plant and amphibian genomes:
  - polyploid
  - Chromosome number undergoes doubling to two-, four, or eight-fold without a radical change to the form of the organisms

Other Types of Data

- Sequencing
  - Capillary and Next Generation Sequencing
- Microarray
  - Gene Expression
  - 50 M data points to tile the human genome at ~50 bp res.
- Phenotype Experiments
  - UC Davis - Knockout Mouse Project
  - Snyder - Transposons
- Protein Interactions
  - For yeast: 6000 x 6000 / 2 ~ 18M possible interactions
  - maybe 30K real
- Chromatography - Mass spectrometry

Examples Data Stored in Molecular Databases



Will you be the next to design a public database?

If the Molecular Biology is new to you, please read the following:
http://155.33.203.128/teaching/BIOL6308-Fall2013/local/Literature/NucleicAcidWorld.pdf

Bioinformatics Methods and Algorithms

- Units and Sources of Biological Data for bioinformatics
- **Types of the Data**
- Data storage, retrieval and visualization
- General Types of “Informatics Techniques in Bioinformatics”
- Data mining

Types of Biological Data for Bioinformatics

- Continuous data
  - Bimolecular structures from X-ray, NMR experiments, kinetic data on biochemical reactions, images of all kinds
- **Digital or Discrete data**
  - **Sequences of genes, proteins, RNAs**

Bioinformatics Methods and Algorithms

- Units and Sources of Biological Data for bioinformatics
- Types of the Data
- **Data storage**, retrieval and visualization
- General Types of “Informatics Techniques in Bioinformatics”
- Data mining

Bits & Bytes (Discrete Data)

- A binary prefix is a "specifier" that is prepended to the units of digital information, the bit and the byte
- A bit is a binary digit - the smallest increment of data on a computer
  - Can hold one of two values: 0 or 1 | on or off
  - bits usually assembled into a byte
    - A byte contains enough information to store a single ASCII character like ‘A’

bit = binary digit
1 base >= 2 bits **??**
1 byte = 8 bits

00=A
01=C
10=C
11=T (U)

Digital Information - Demystified

- A kilobyte (KB) is 1,024 bytes, not one thousand bytes as might be expected
  - We use a decimal (base ten) system
  - Computers use binary (base two) math
    - Indicates multiplication by a power of 2
    - In practice the powers used are multiples of 10, so the prefixes denote powers of 1024 = 2^{10}
- Computer storage and memory is often measured in megabytes (MB) and gigabytes (GB)
  - A medium-sized novel contains about 1MB of information
  - 1MB is 1,024 kilobytes, or 1,048,576 (1024x1024) bytes, not one million bytes

Binary				
Value	IEC		JEDEC	
1024	Ki	kibi	K	kilo
1024^2	Mi	mebi	M	mega
1024^3	Gi	gibi	G	giga

FYI - Digital Information - (cont'd)

- gigabyte (GB) is 1,024MB, or 1,073,741,824 (1024x1024x1024) bytes
- terabyte (TB) is 1,024GB
  - 1TB is about the same amount of information as all of the books in a large library, or roughly 1,610 CDs worth of data.
- petabyte (PB) is 1,024TB
- exabyte (EB) is 1,024PB
  - Google processes 8 exabytes of data every year
- zettabyte (ZB) is 1,024EB
- yottabyte (YB) is 1,024ZB

A Computer Counts Like This:

We count in base 10 by powers of 10:

$10^1 = 10$
$10^2 = 10*10 = 100$
$10^3 = 10*10*10 = 1,000$
$10^6 = 1,000,000$

Computers count by base 2:

$2^1 = 2$
$2^2 = 2*2 = 4$
$2^3 = 2*2*2 = 8$
$2^{10} = 1,024$
$2^{20} = 1,048,576$

bit = binary digit
1 base >= 2 bits
1 byte = 8 bits

	+ Kilo	Mega	Giga	Tera	Peta	Exa	Zetta	Yotta +
Decimal	3	6	9	12	15	18	21	24
	- milli	micro	nano	pico	femto	atto	zepto	yocto -

		Kibi	Mebi	Gibi	Tebi	Pebi	Exbi
Binary	$1024 = 2^{10}$	2^{20}	2^{30}	2^{40}	2^{50}	2^{60}	

Sizes of the Data

- Number of component types (estimated)

	Mycoplasma	C. elegans	Human
Bases	.58M	>97M	3000M
Genes	.48K	>19K	30K
RNAs	.4k	>30K	.2-3M
Proteins	.6k	>50K	.3-10M
Cells	1	~1000	10^{14}

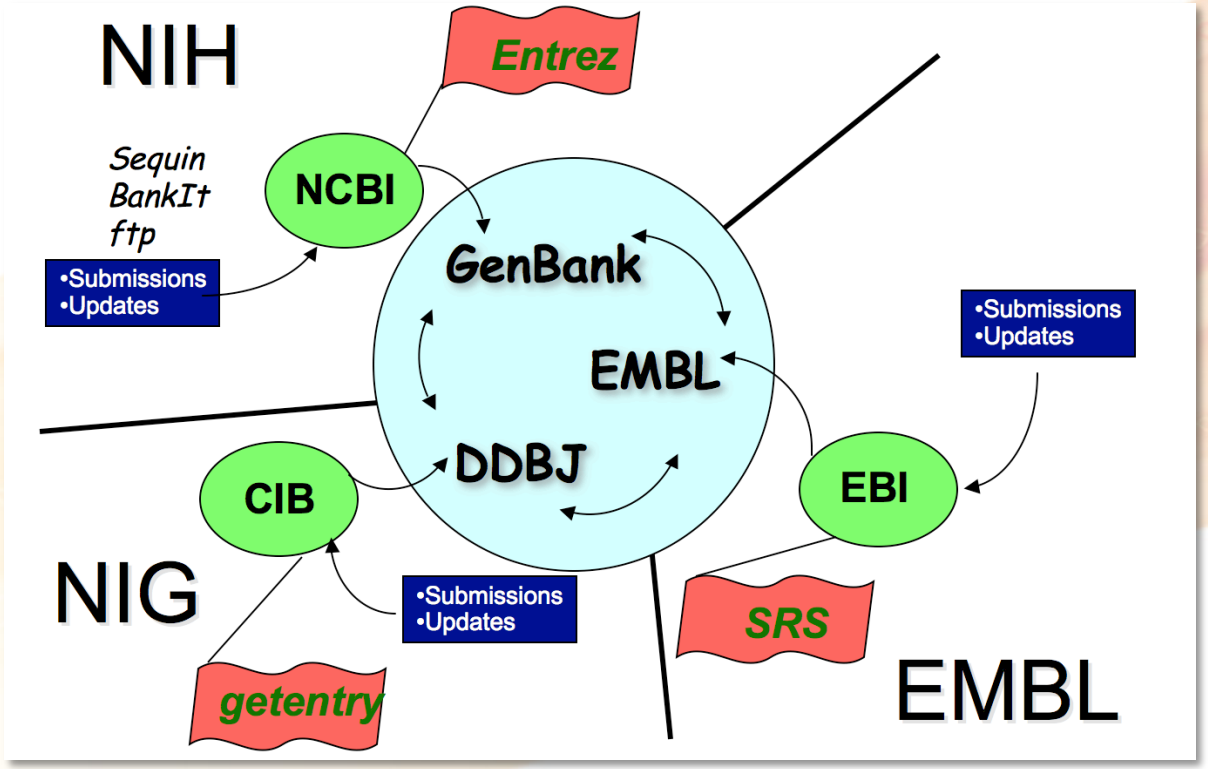
The Take Home - Some Numbers

- So, remember every nucleotide can be coded by 2 bits
  - 1 byte = 8 bits
- A chain of 5,100 nucleotides corresponds to 2 × 5100 = 10,200 bits
  - 1275 bytes
- The E. coli genome is a single DNA molecule consisting of two chains of 4.6 million nucleotides
  - ~9.2 million bits, or 1.15 megabytes, of information
- The 2.9 billion base pairs of the haploid human genome correspond to about 715 megabytes
- So if the average gene size in humans is 3000nts and there are ~30K genes, how much room would be needed to store this?
  - In bits?
  - In bytes
  - In megabytes?

Bioinformatics Methods and Algorithms

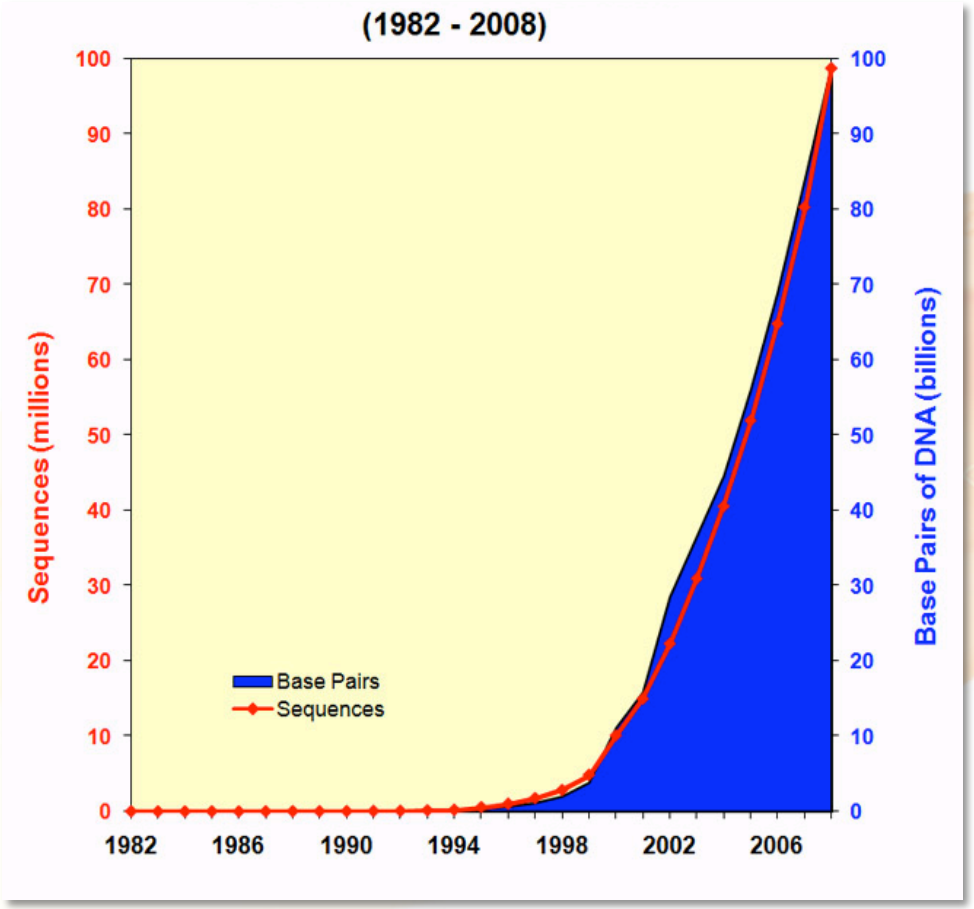
- Units and Sources of Biological Data for bioinformatics
- Types of the Data
- Data storage, **retrieval** and visualization
- General Types of “Informatics Techniques in Bioinformatics”
- Data mining

DNA Data Retrieval



The International Nucleotide Sequence Database Collaboration

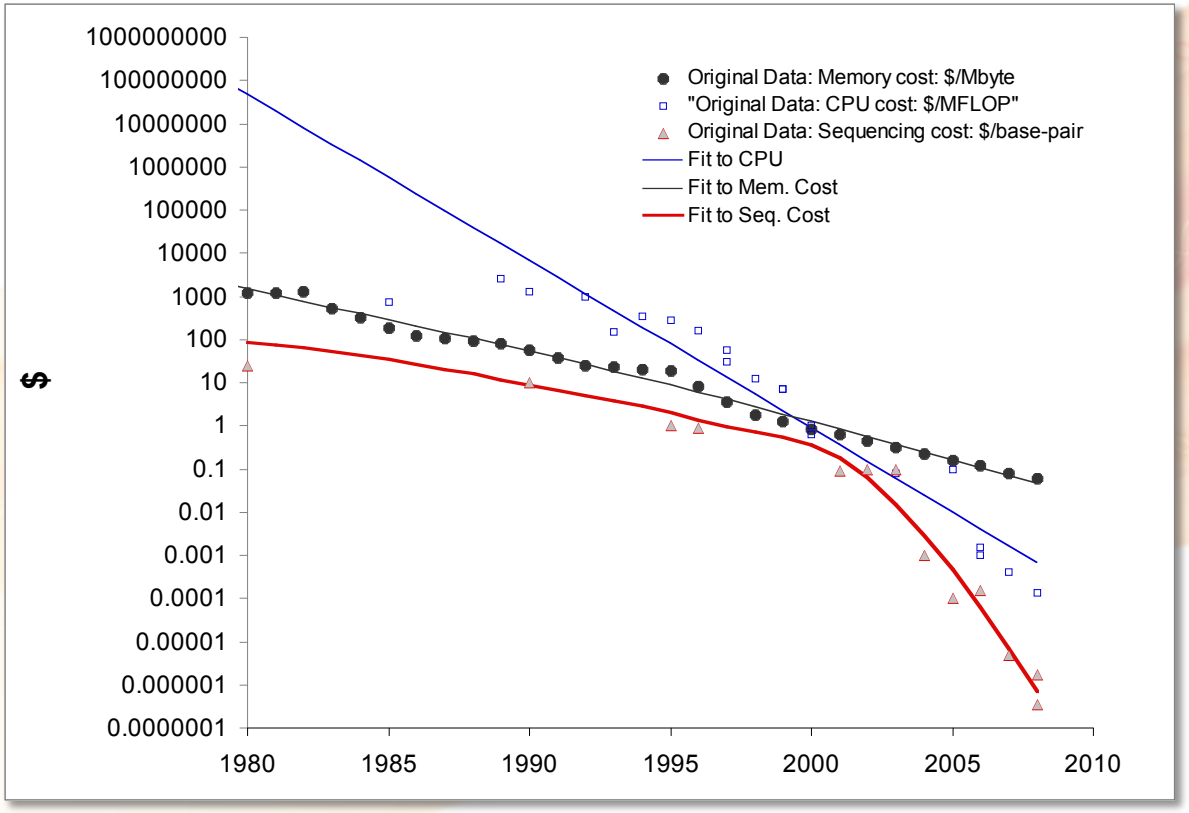
Growth of GenBank



Year	Base Pairs	Sequences
1992	101,008,486	78,608
1993	157,152,442	143,492
1994	217,102,462	215,273
1995	384,939,485	555,694
1996	651,972,984	1,021,211
1997	1,160,300,687	1,765,847
1998	2,008,761,784	2,837,897
1999	3,841,163,011	4,864,570
2000	11,101,066,288	10,106,023
2001	15,849,921,438	14,976,310
2002	28,507,990,166	22,318,883
2003	36,553,368,485	30,968,418
2004	44,575,745,176	40,604,319
2005	56,037,734,462	52,016,762
2006	69,019,290,705	64,893,747
2007	83,874,179,730	80,388,382
2008	99,116,431,942	98,868,465

What's the driving force behind this growth?

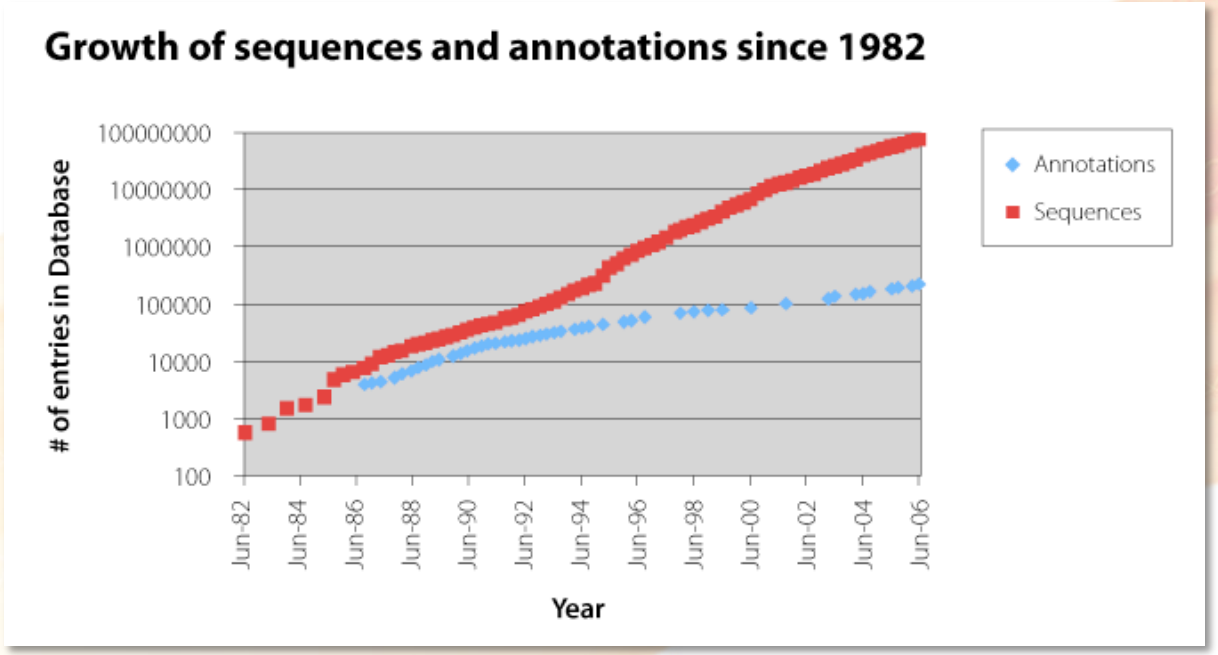
Cost of Sequencing - The Driving Force



So, we're almost of out jobs?

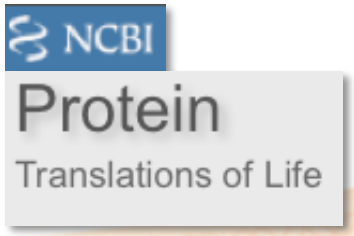
Not exactly!

A Lot of Work to Do!



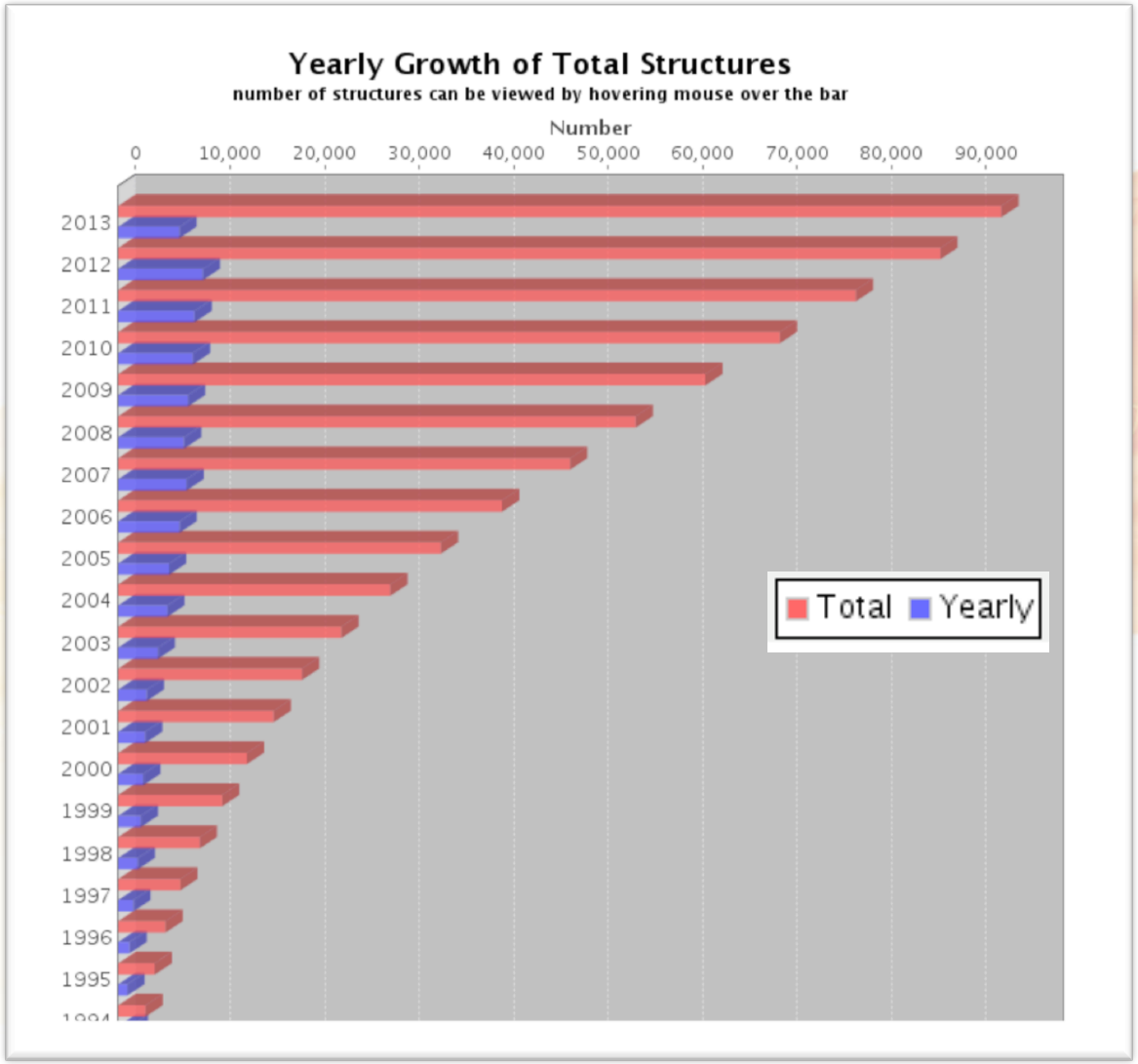
What's the most important feature of this graph?

Protein Data Retrieval



InterProScan

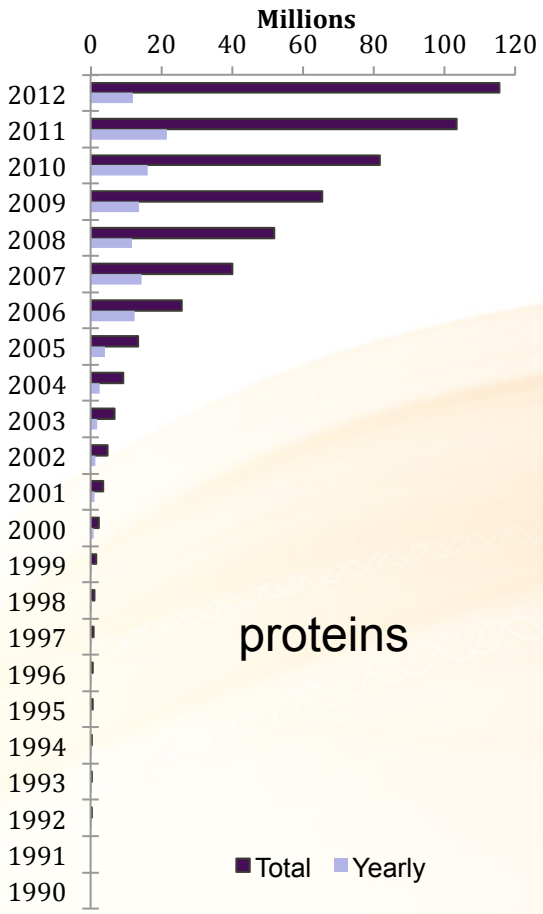
Number PDB Structures

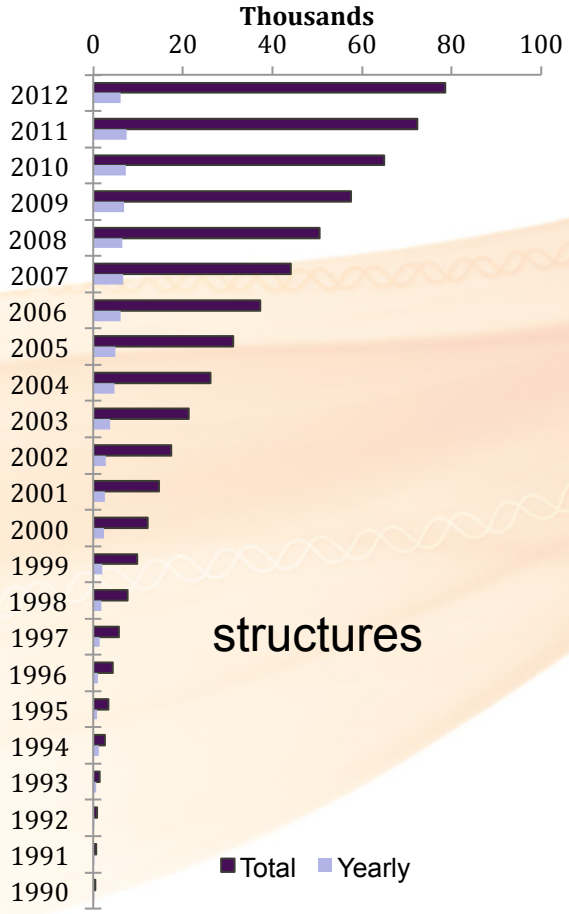


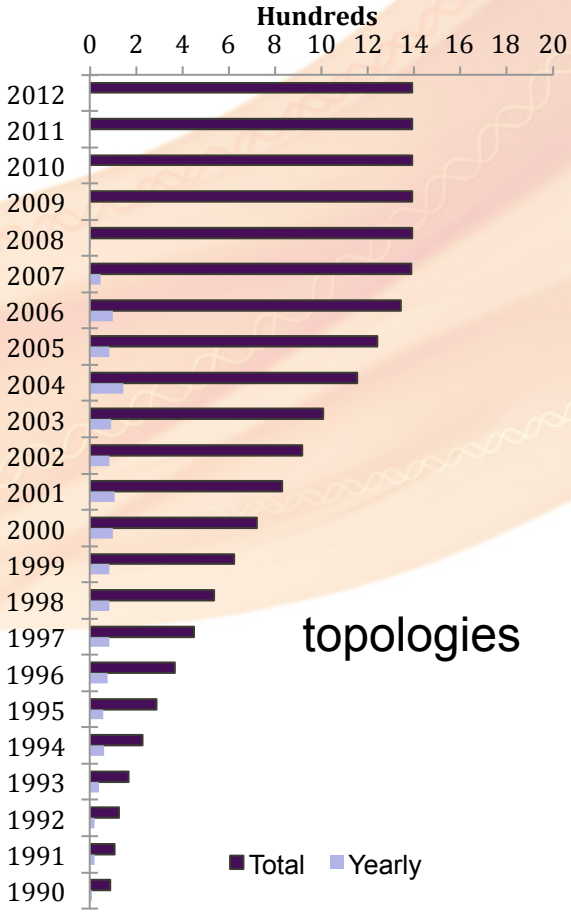
Increase in structure Data, but what about topologies?

http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100

PDB vs. Topology Growth







http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-cath

Bioinformatics Methods and Algorithms

- Units and Sources of Biological Data for bioinformatics
- Types of the Data
- Data storage, retrieval and **visualization**
- General Types of “Informatics Techniques in Bioinformatics”
- Data mining

We Will Use Geneious

http://www.geneious.com

Geneious

- **Bioinformatics Software for Sequence Alignment**
  - For sequence/structure analysis
  - Multiple algorithms for MSA
  - More Complex Trees
  - Primer-Probe Design
  - Sequence Assembly
  - Better GUI
  - Industry level bioinformatics application
  - Connections with NCBI
  - Many file formats supported
  - Not Free
    - Free Basic version with limited functionality
    - Pro version – expensive ($795) or 12mo ($395)
- **But this provides a crutch, so we will not begin to use Geneious until 2nd semester**

But Before Geneious

- Need to become experienced users of the Unix/Linux environment
- Why you ask?
  - Unix/Linux-based software
    - Common languages: Perl/Python
    - Command-Line interface (CLI, **Not very user friendly**)
    - Open source, freeware
    - Majority of our tools are CLI
- In order to get the job done, you will have to be comfortable with the CLI

Bioinformatics Methods and Algorithms

- Units and Sources of Biological Data for bioinformatics
- Types of the Data
- Data storage, retrieval and visualization
- **General Types of "Informatics Techniques in Bioinformatics"**
- Data mining

General Types of "Informatics" Techniques in Bioinformatics

- Databases
  - Building
  - Querying
- Text String Comparison
  - Text Search
  - Pairwise Alignment
- Finding Patterns
  - Machine Learning
  - Clustering
  - Data mining
- Physical Simulations
  - Newtonian Mechanics
  - Electrostatics
  - Molecular Dynamics
- Geometry
  - Comparison and 3D matching

Bioinformatics Topics - Genome Sequences

- Finding Genes in Genomic DNA
  - introns
  - exons
  - promotors
- Characterizing Repeats in Genomic DNA
  - Statistics
  - Patterns
- Duplications in the Genome
  - Large scale genomic alignment
- Whole-Genome Comparisons
- Finding Structural RNAs

Bioinformatics Topics - Protein Sequences

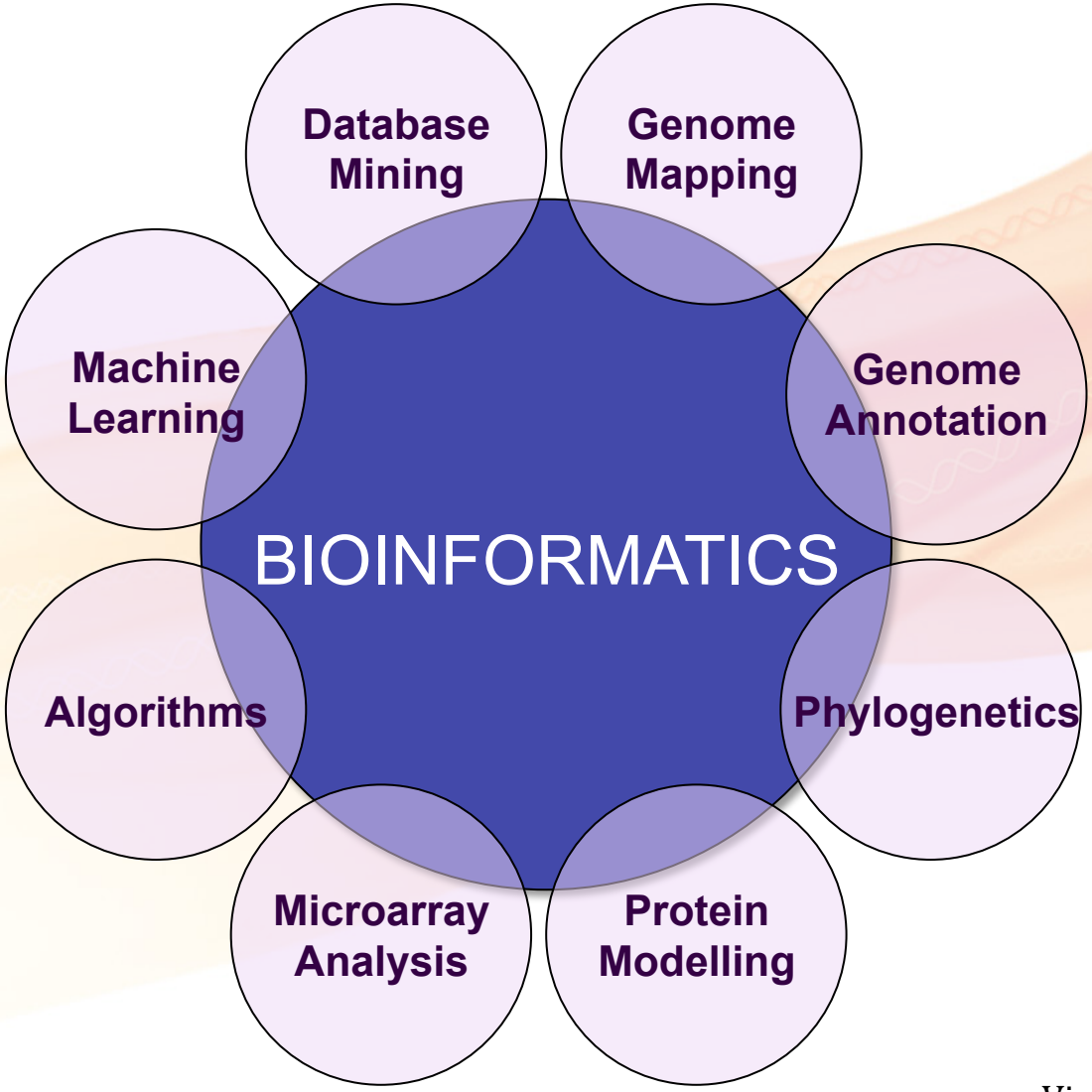
- Sequence Alignment
  - How to align two strings optimally via Dynamic Programming
  - Local vs Global Alignment
  - Suboptimal Alignment
  - Hashing to increase speed (BLAST, FASTA)
  - Amino acid substitution scoring matrices
- Multiple Alignment and Consensus Patterns
  - How to align more than one sequence
    - Fuse the result in a consensus representation
  - HMMs, Profiles
  - Motifs
- Scoring schemes and Matching statistics
  - Is the given alignment or match statistically significant
    - P-value (or an e-value)?
  - Score Distributions (extreme val. dist.)
  - Low Complexity Sequences
- Evolutionary Issues
  - Rates of mutation and change

Bioinformatics Topics - Sequence / Structure

- Secondary Structure “Prediction”
- Structure Prediction: Protein vs RNA
- Tertiary Structure Prediction
  - Fold Recognition
  - Threading
  - Ab initio
  - Quaternary structure prediction
- Direct Functional Prediction
  - Active site Identification
- Relation of Sequence Similarity to Structural Similarity

What's a Bioinformatician to Do?

What Can We Use Bioinformatics For?



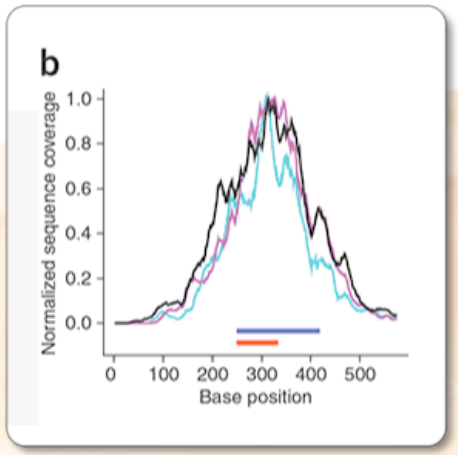
Victoria Offord

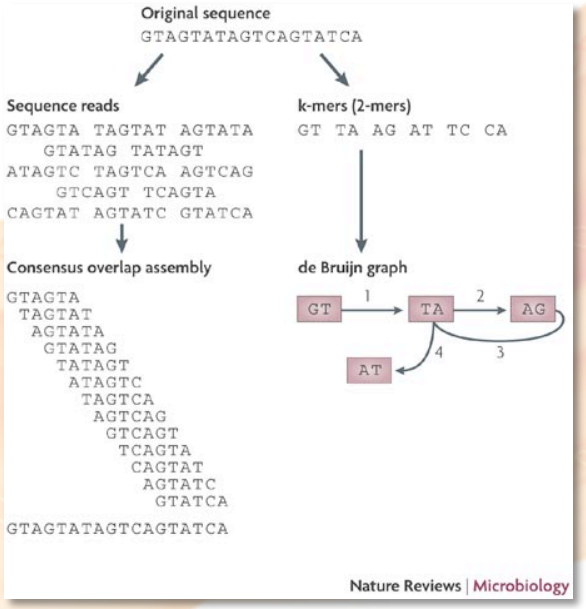
What Format Are You In?

- **Work with many data formats**
  - FASTA
  - EMBL
  - GENBANK
  - FASTQ
  - PIR
  - SWISSPROT
  - PDB
  - XML
  - SAM
  - PILE UP
  - BLAST
  - GFF
  - 100’s more!

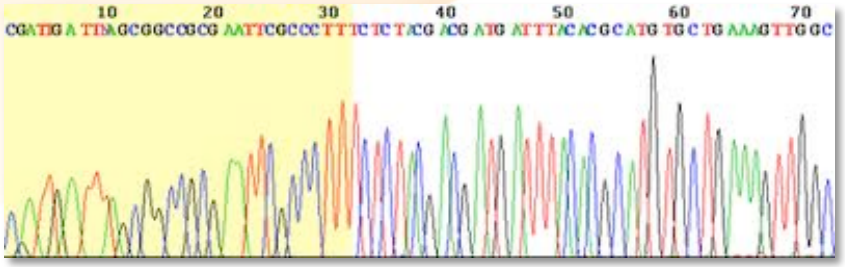
Sequence Assembly

- **Next-Generation Sequencing**

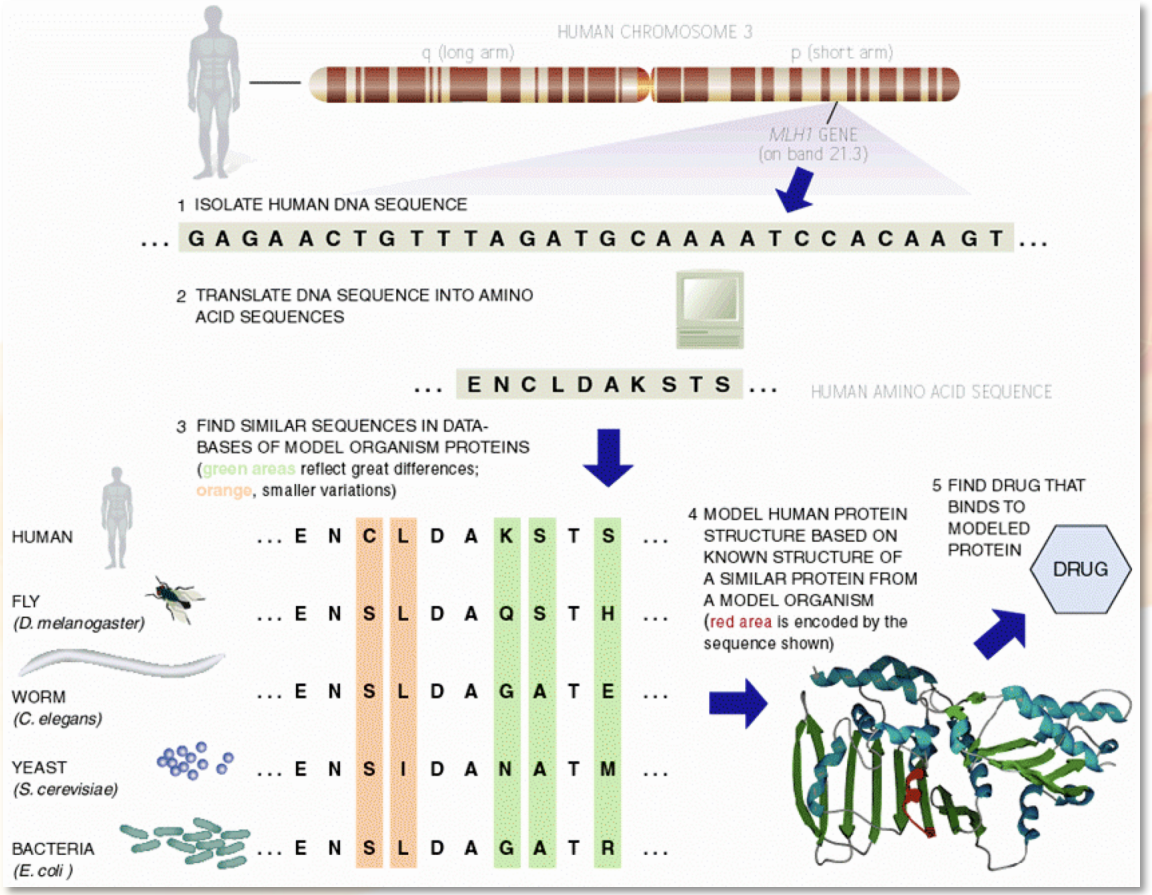




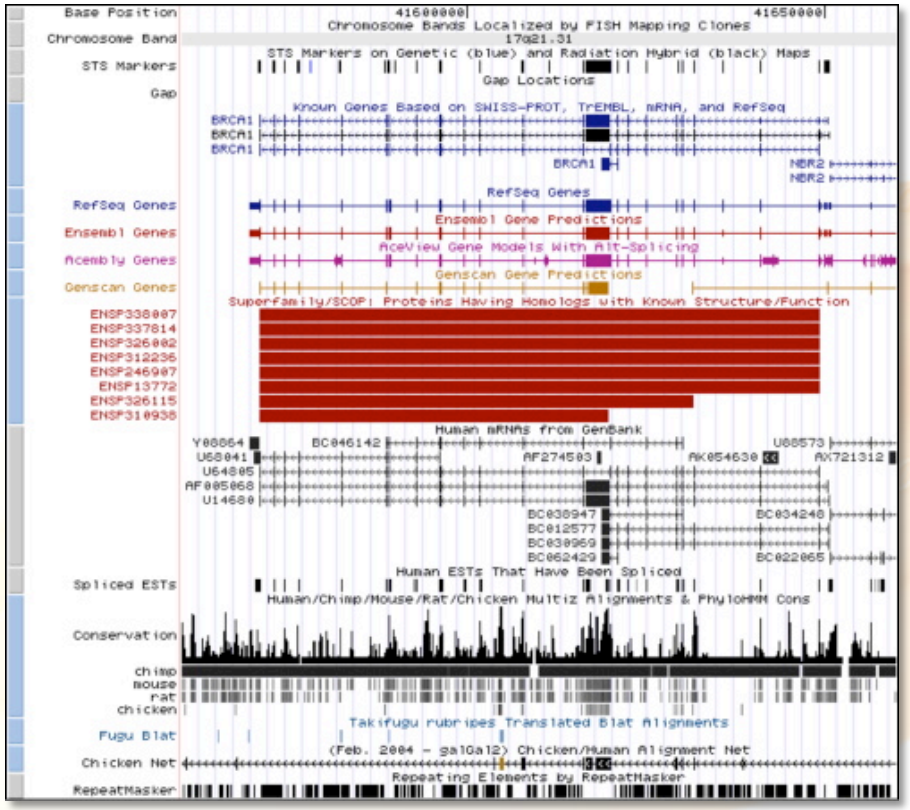
- **Sanger Sequencing**



Finding Homologs

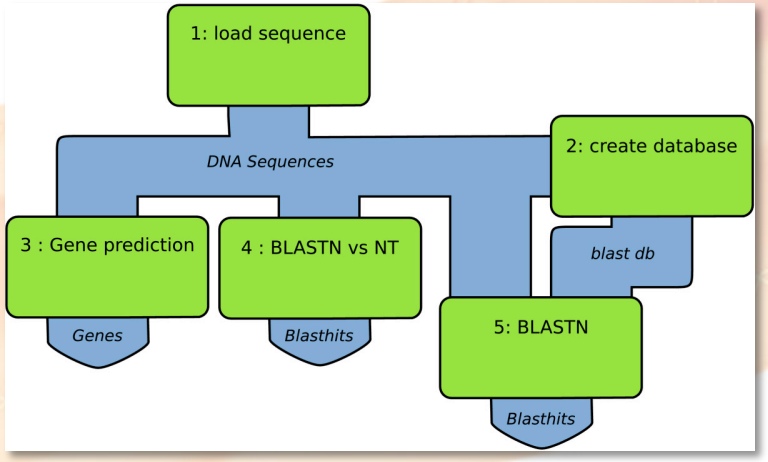


Genome Annotation



BRCA1 in the UC Santa Cruz Genome Browser

http://genome.ucsc.edu/cgi-bin/hgGateway

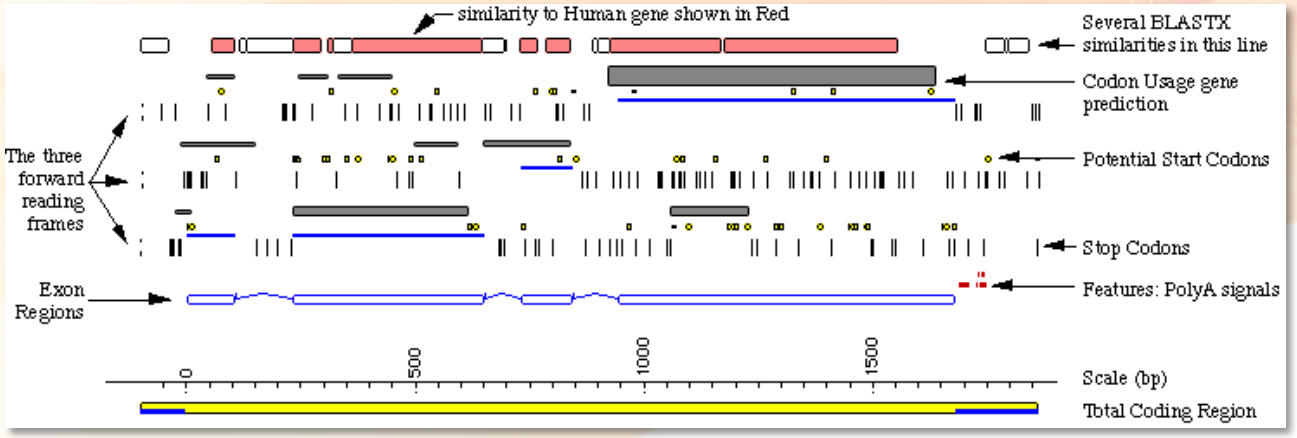


Example of a simple computational pipeline for genome annotation. Green boxes represent nodes. The pipeline describes the execution of a gene predictor (node 3) and two BLAST analyses (nodes 4 and 5) [19] on a set of input DNA sequences (node 1). The BLAST analysis in node 4 compares the incoming sequences against the NCBI NT database. Node 5 uses a BLAST database created by node 2 from the same set of sequences (self-BLAST).

Fiers et al. BMC Bioinformatics 2008 9:96 doi: 10.1186/1471-2105-9-96

Locate all Genes and Regulatory Regions

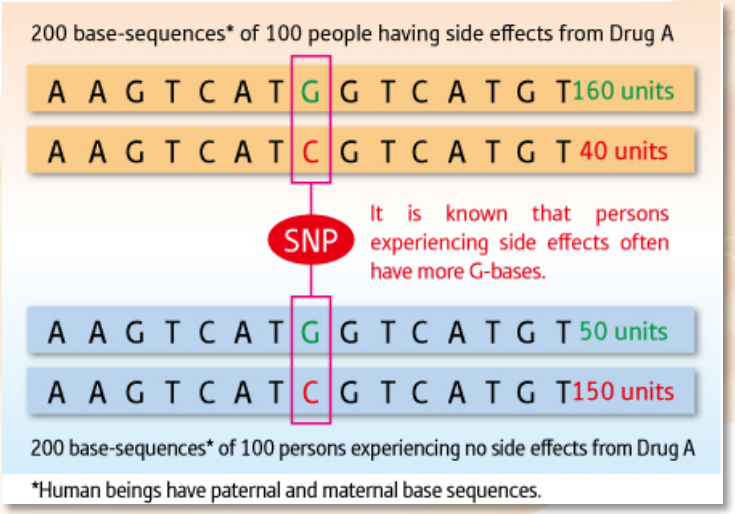
- Location is not enough
- Must bet able to describe their functions
- Bioinformatics plays a critical role



Richard Bruskiewich

Identifying Single Nucleotide Polymorphisms (SNPs)

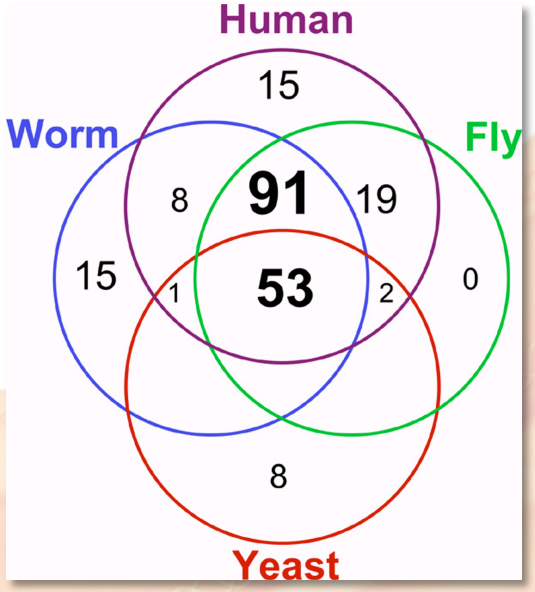
along with other changes between individuals

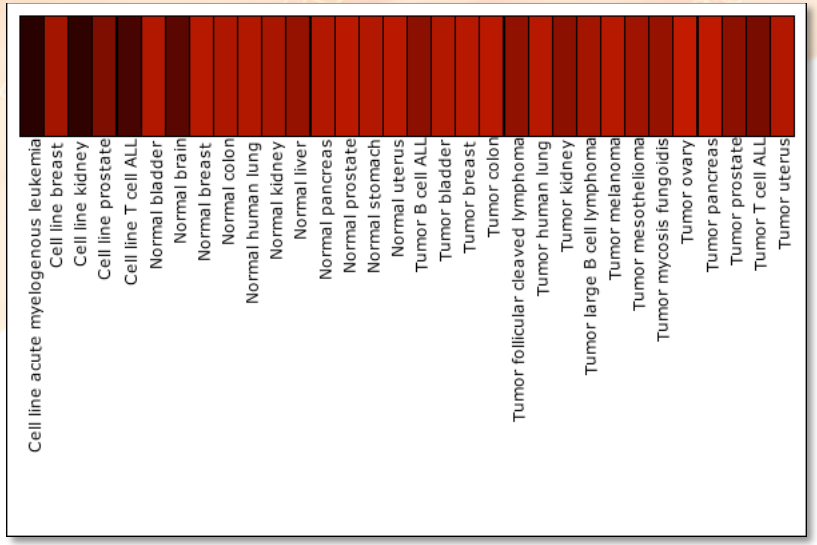


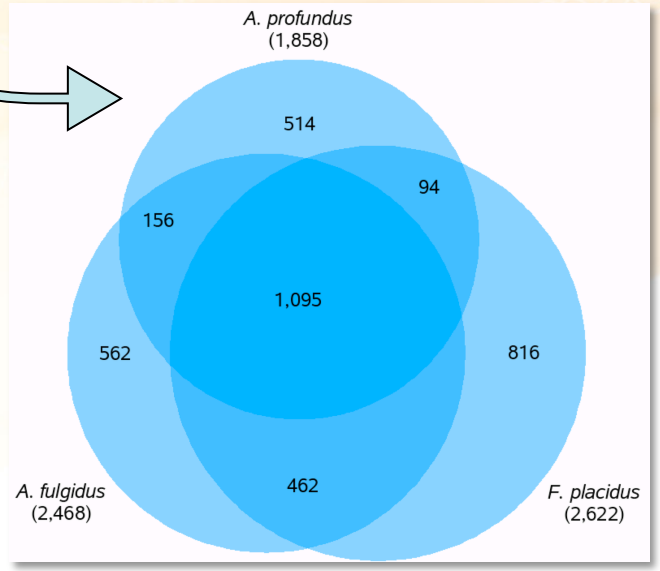
Richard Bruskiewich

Overall Genome Characterization

- Overall occurrence of a certain feature in the genome
  - e.g. how many kinases in Yeast shared with other *spp*
- Compare organisms and tissues
  - Expression levels in cancerous vs normal tissues
- Find shared genes between *spp*

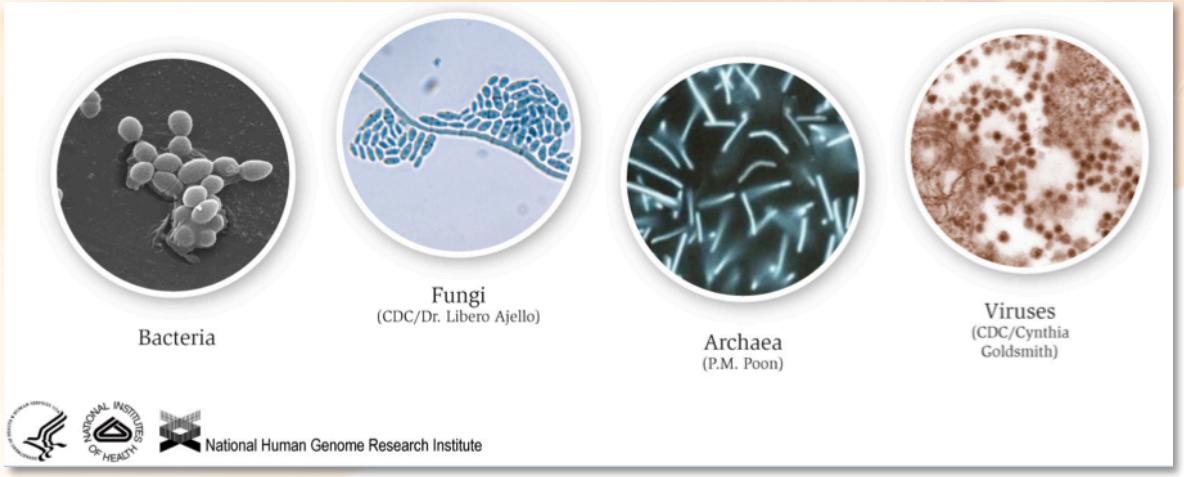






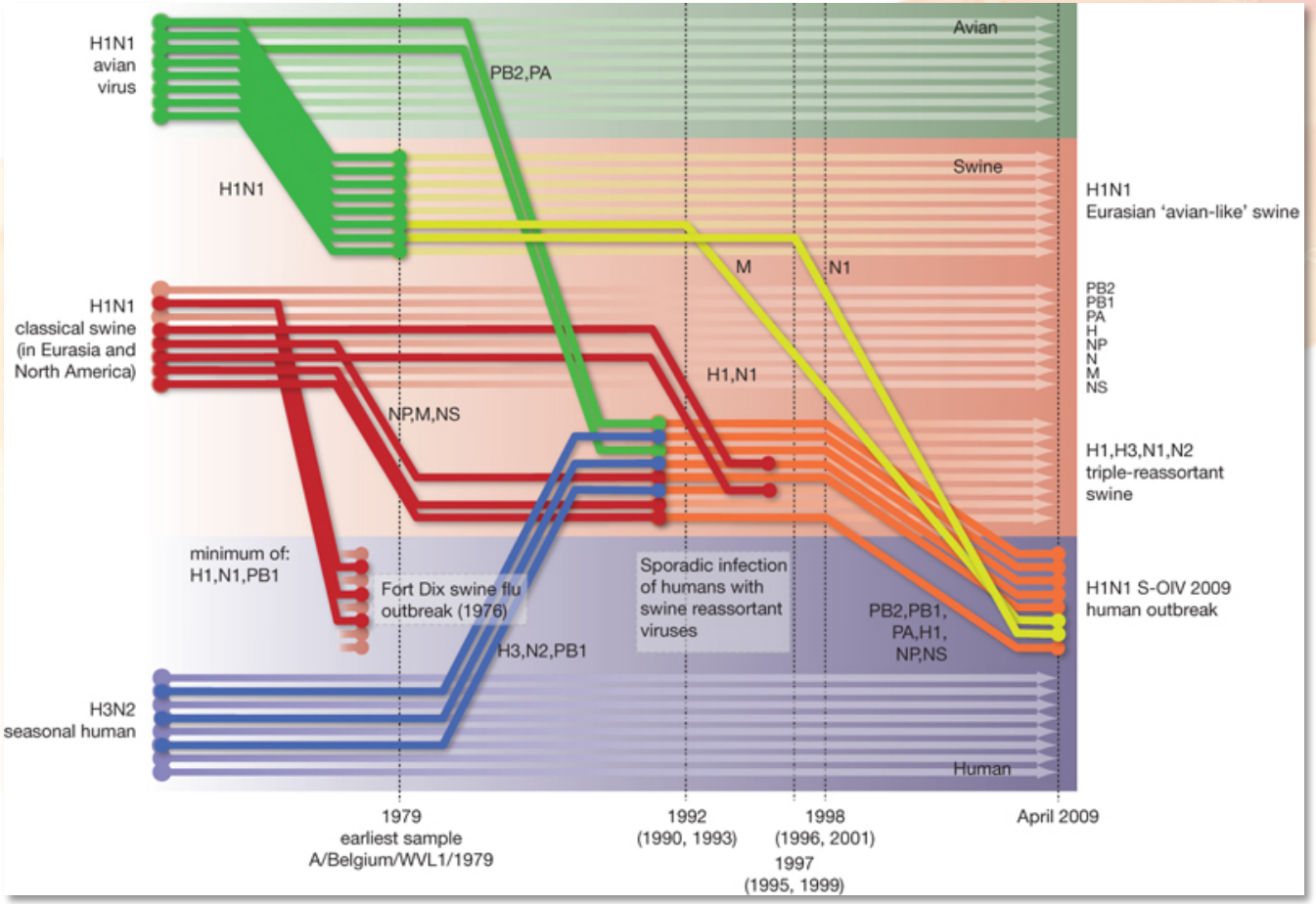
Microbiome Projects

- Characterize microbes and how they interact in an ecosystem
- The cells in our bodies are outnumbered 10:1 by microbes
- About 3 to 4 pounds of microbes live on and in our bodies
- The microbes represent thousands of different species
  - Bacteria
  - Fungi
  - Archaea
  - Viruses



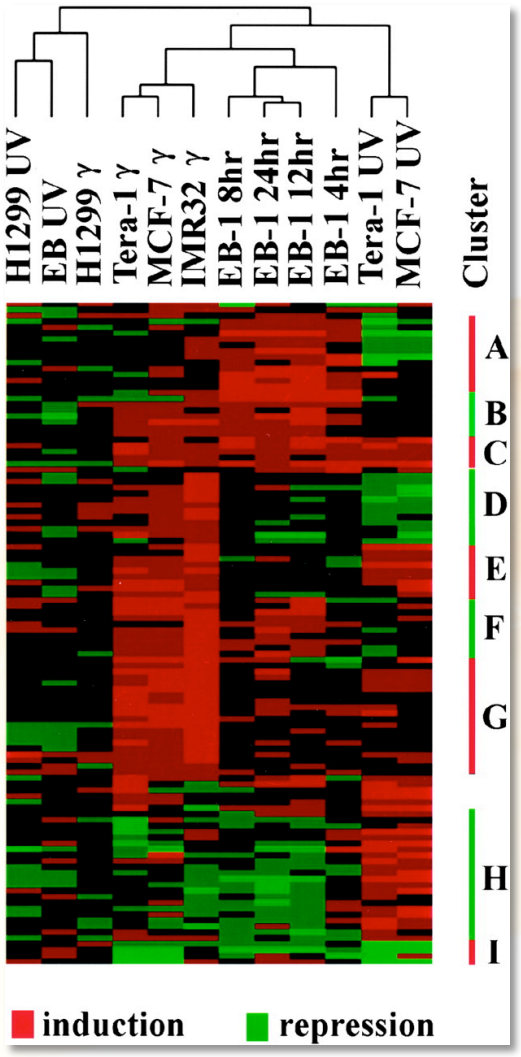
Molecular Evolution

- Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic



Smith et al. (2009) *Nature* ***459, 1122-1125***

Analysis of Gene Expression

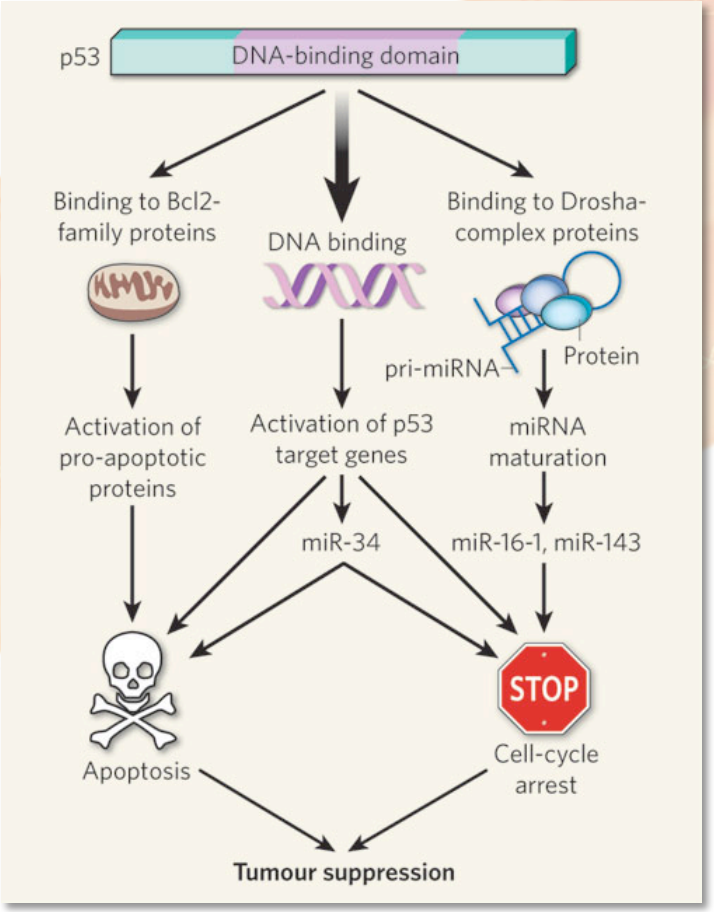


- Cluster analysis of gene expression patterns after UV or γ irradiation
- A total of 112 genes and 5 different cell lines following radiation were included in this analysis
- Cells with wild-type p53 fall into distinct clusters following different irradiation
- The 112 genes were clustered into 9 groups with different expression patterns in cells with wild-type p53 following different types of radiation

http://genesdev.cshlp.org/content/14/8/981.abstract

Analysis of Regulation

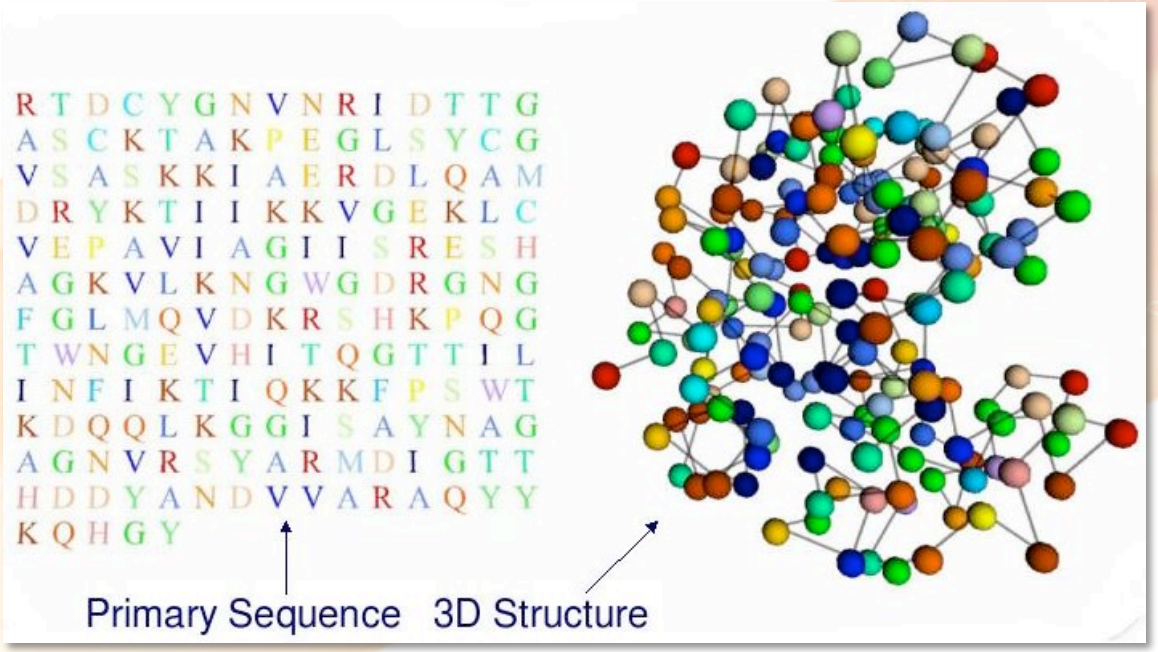
- DNA-binding domain of p53 lies in the core of the protein
- Has three anti-tumour functions.
  - **Binds to DNA** and enables the activation of target genes to induce apoptosis or cell-cycle arrest;
  - **Stimulates apoptosis** through an interaction with proteins of the Bcl2 family at the mitochondrion;
  - **Interacts with proteins** of the Drosha complex to promote processing of a subset of miRNAs, which suppress cell proliferation.
- Most p53 mutations in human cancers lie in the DNA-binding domain and may affect all three functions.



Toledo and Bardot (2009) *Nature* ***460, 466-467***
http://www.nature.com/nature/journal/v460/n7254/full/460466a.html

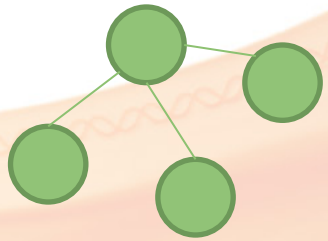
Protein Structure Prediction

- **Structure Prediction aims** to predict the 3D structure of a protein based on its primary sequence, that is, a chain of amino-acids (a string using a 20-letter alphabet)



Protein Motif Finding

- Used to Predict
  - Enzyme Active Sites
  - Protein interaction sites



C-x(2,5)-C-x-[GP]-x-P-x(2,5)-C

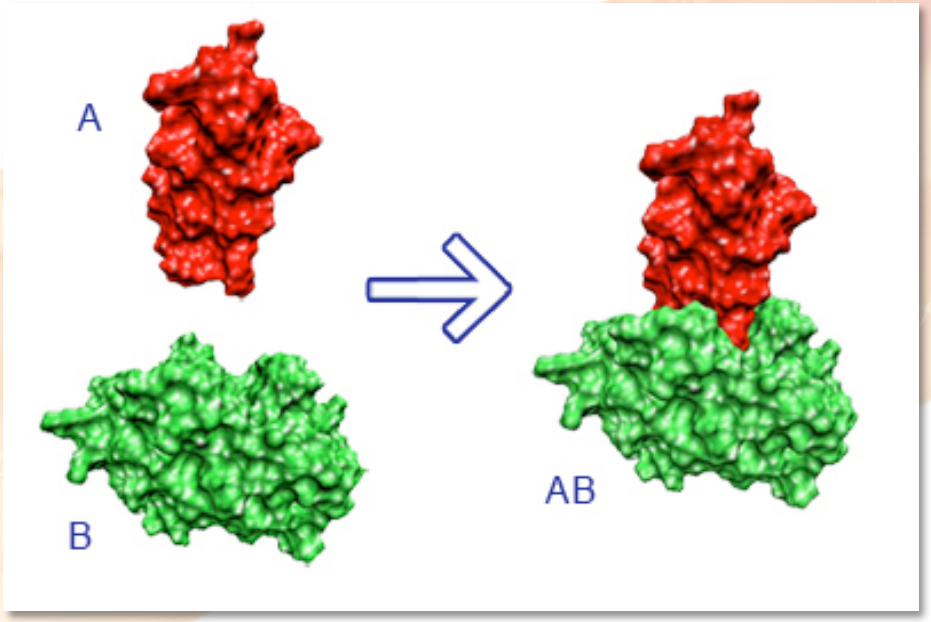
CXXX**C**X**G**X**P**XXXXX**C**

| | | | |

FG**C**AKL**C**A**G**F**P**LRRLP**C**FYG

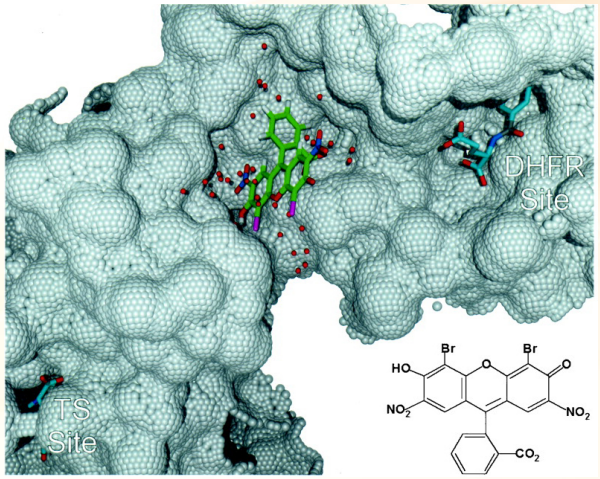
Protein-Protein Docking

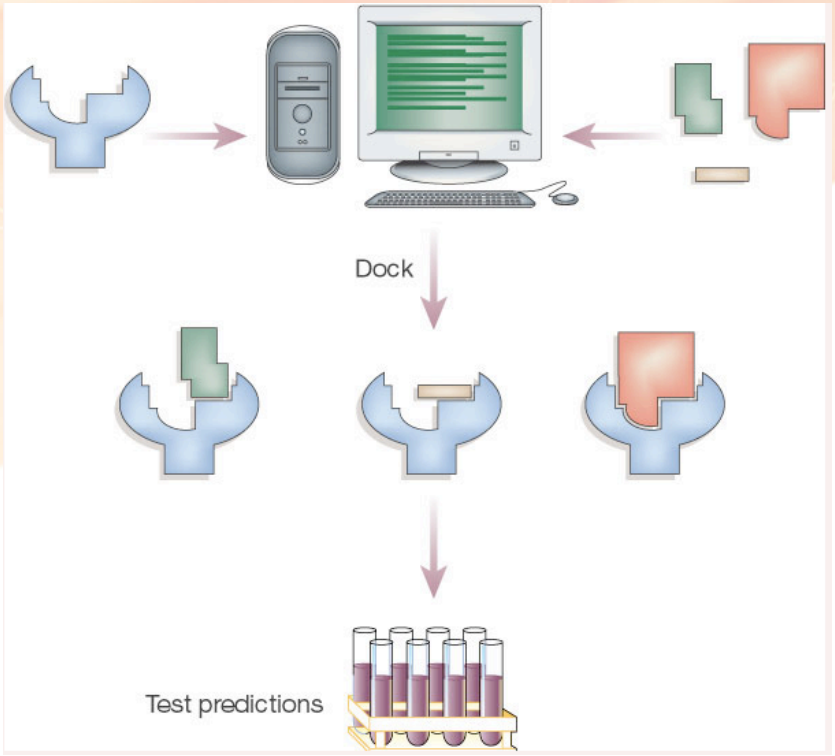
- Computational modeling of the quaternary structure of complexes
  - Two or more interacting biological macromolecules
  - Ultimate goal is the prediction of the 3-D structure of the macromolecular complex of interest as it would occur in a living organism



Drug Docking

- Discover new drug targets - computational docking
- Understanding how structures bind other molecules (Function)
- Designing inhibitors





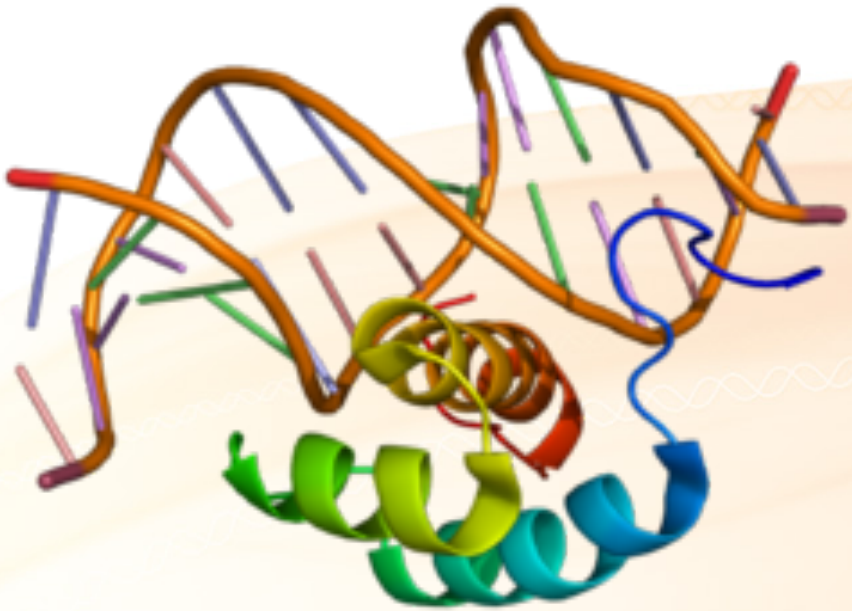
Atreya, C. E. et al. J. Biol. Chem.

FYI - Bioinformatics is Results-Driven

- Cancer and Tumor Progression
- Goals
  - Study affected transcriptional networks in prostate, brain, breast, and ovarian cancers
  - Identify biomarkers for cancer diagnostic purposes
  - Develop effective mRNA and microRNA biomarkers
- Purpose
  - Currently mainly studying prostate cancer, though brain, breast, and ovarian cancers are also of high interest and very similar to prostate cancer

Carlos Moreno

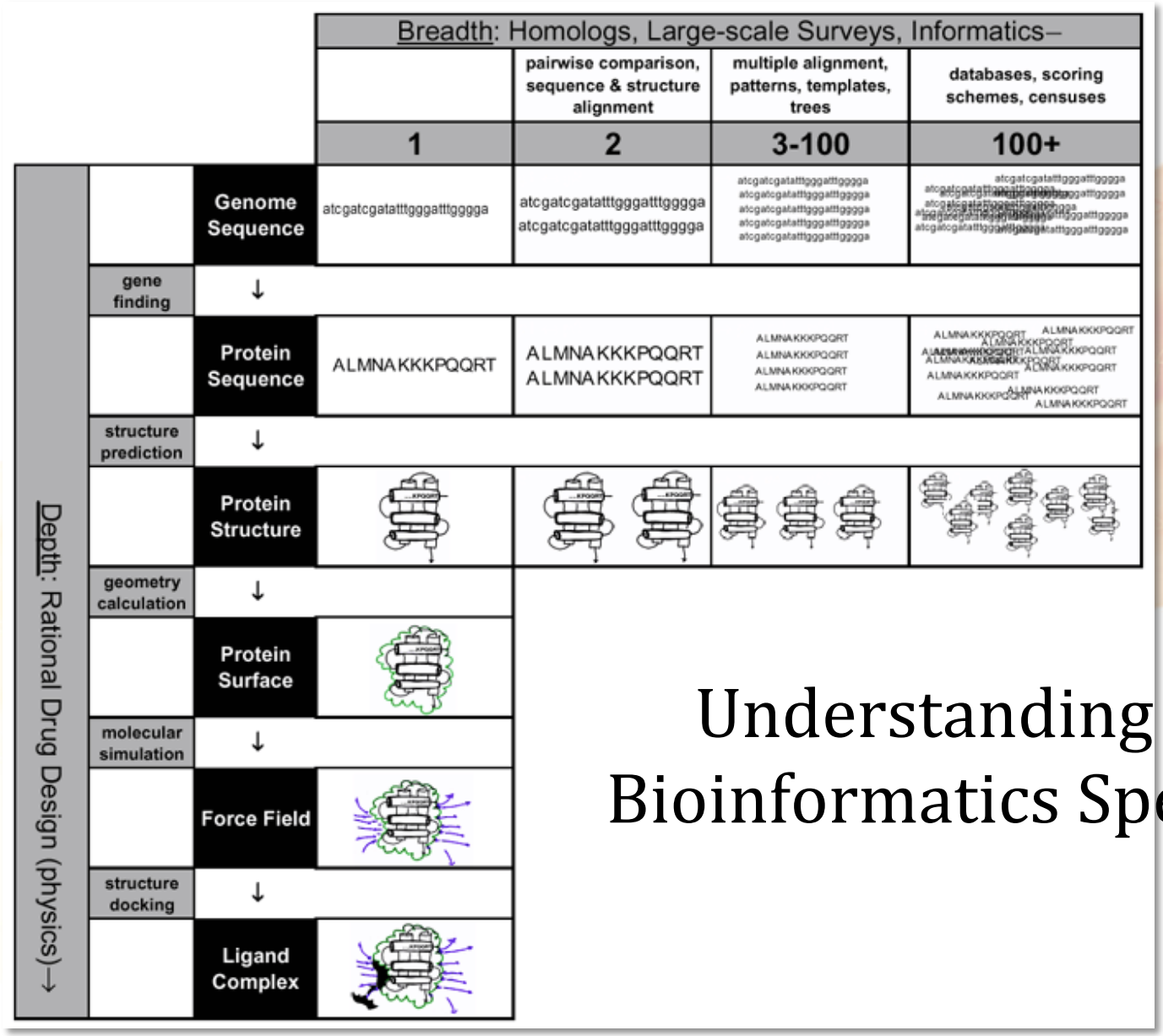
FYI - Results



- His group identified several genes correlating to prostate cancer development
- Presence of two developmental transcription factors
- **HOXC6 and SOX4**
- Responsible for absence of apoptosis in cancer cells
- Provides new potential therapeutic target for treatment of prostate cancer

What will you study?

Carlos Moreno



Understanding the Bioinformatics Spectrum

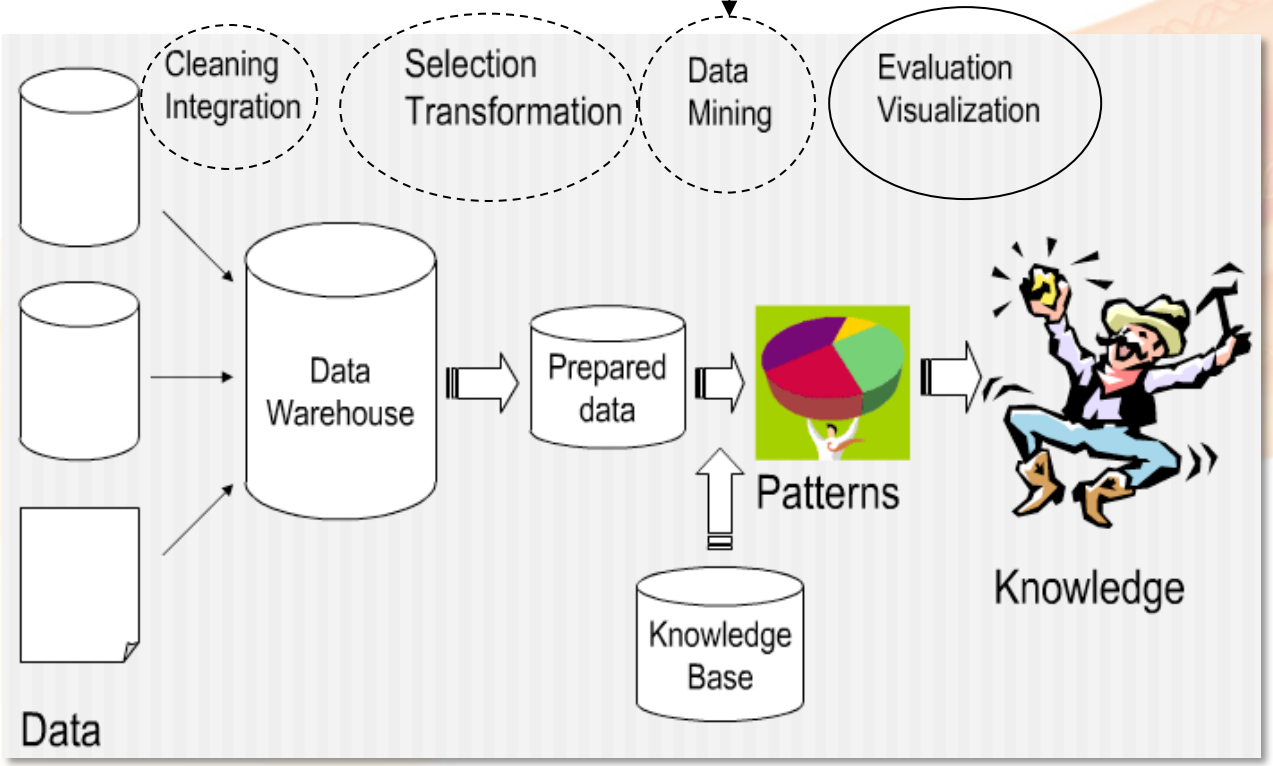
Bioinformatics Methods and Algorithms

- Units and Sources of Biological Data for bioinformatics
- Types of the Data
- Data storage, retrieval and visualization
- General Types of “Informatics Techniques in Bioinformatics”
- **Data mining**

Data Mining - Bioinformatics Tools

- Data mining skills are critical to a bioinformatician
- Thousands of computer programs
  - Many freely available
  - Generally available on Unix or Linux
    - This why you must learn this type of OS
    - **This why we will learn this type of OS!!**
- Often interact with bioinformatics databases
- Many accessible via the WWW
- Some require more powerful servers to run on
  - May need way to automate these programs
  - Perl is great for this
- You should have the ability to code it data mining yourself
  - Sometimes not freely in the public arena

Data Mining



Use Perl

Use Java

Why Perl for Bioinformatics

- Easy string processing for biological data
- No strict rules, helps when biologists begin programming
- File processing made simple
- Combined with SHELL scripts for processing
- CPAN contains many Perl Modules which are specific for Bioinformatics
- Perl used for System administration
- Perl DBI is one of the best modules for database applications
- Processing/Pasing a HTML file made simple with CPA moduels
- File type conversion
- Used throughout industry and academia
- Read the following blog
  - **Why is Perl used so extensively in Biology research?**

Books to own - For a Beginning Bioinformatician

So……..What is Bioinformatics?

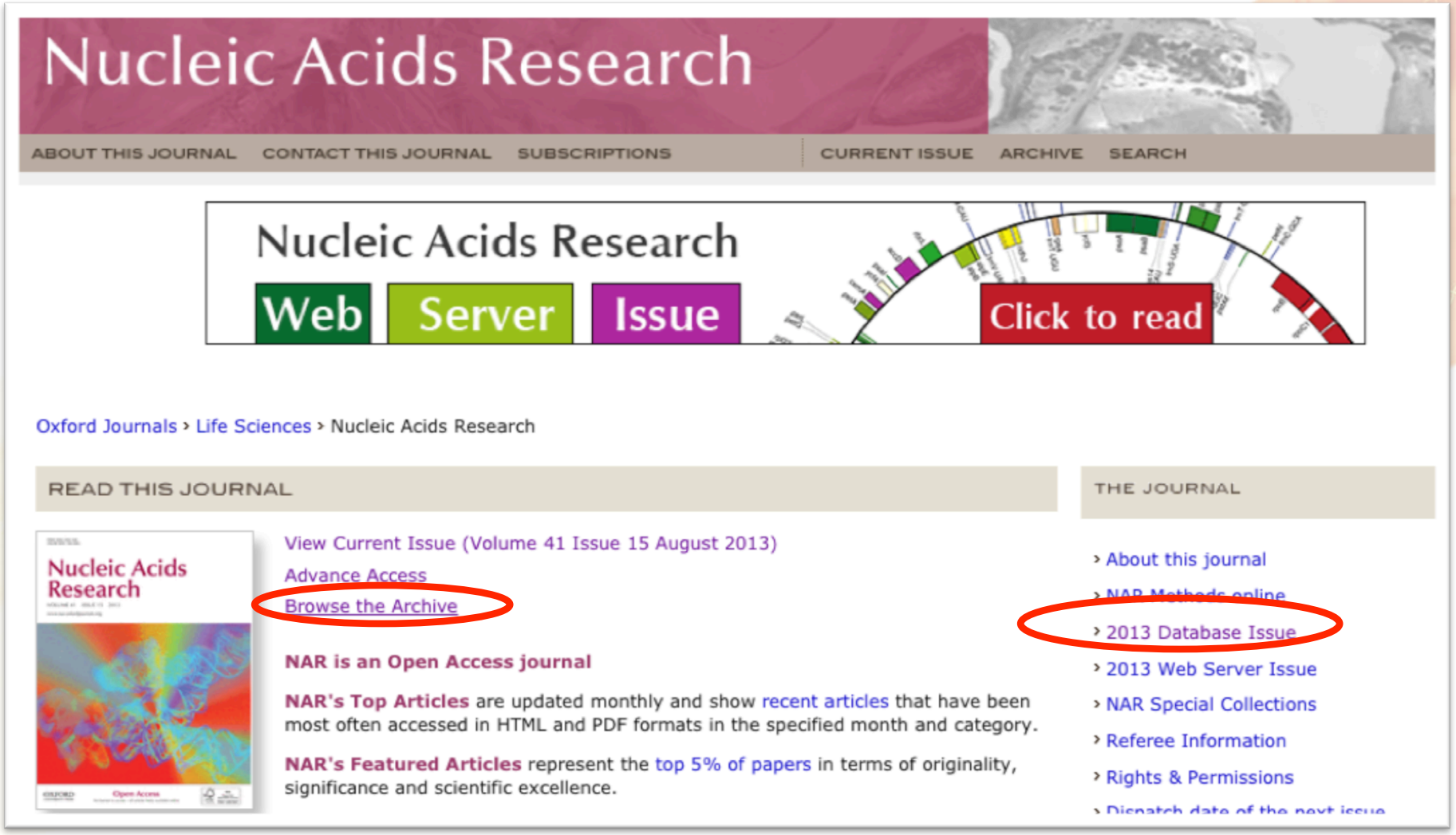
- (Molecular) Bio-informatics
- Any ideas for a definition?

- Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale**

- ***Bioinformatics is a practical discipline with many applications***

Good Reading Material

http://nar.oxfordjournals.org



Check the August edition of Nucleic Acids Research

Computing Versus Biology
Looking Into the Future

- **Like physics, where general rules and laws are taught at the start, biology will surely be presented to future generations of students as a set of basic systems ..... duplicated and adapted to a very wide range of cellular and organismic functions, following basic evolutionary principles constrained by Earth's geological history**

-Temple Smith, Current Topics in Computational Molecular Biology

Course Website

- http://155.33.203.128/teaching/BIOL6308-Fall2013/local/local_BIOL6308_fall2013.html

- user name:
  - First initial + last name
    - i.e. cleslin
  - password – I'll email

For Next Tuesday

- Brush up on your Molecular Biology
  - http://155.33.203.128/teaching/BIOL6308-Fall2013/local/Literature/NucleicAcidWorld.pdf
- Linux Basics
  - http://155.33.203.128/teaching/BIOL6308-Fall2013/local/lab/LinuxBasics.html
- Check out the Unix for Beginners
  - http://www.ee.surrey.ac.uk/Teaching/Unix/