Introduction to Bioinformatics

S. Ravichandran, PhD, PMP
FNLCR

Short Biography

- I go by the name **Ravi**
  - S. Ravichandran
- I work for FNLCR (Leidos Biomedical Inc)
  - Joined FNLCR (back then NCI) around early 2002

- I am a scientist. I support and carry out my own research. I am also involved in teaching/training/Research for the past 15 years.
  - Taught *Biocomputing* in Hood in 2010
  - BIFX-550, since 2015

Short Biography

- Also have a Project Management Professional certification (a.k.a PMP)

WELCOME
Please Introduce Yourselves

- Name?
- What do you do?
- Experience with Bioinformatics?
- Experiences beyond Windows?
  - Linux, Mac etc.
- Any programming experience?
  - C, Fortran, C++, Java, Python etc.
- If you haven't done so,
  - Please also send me a short introduction/goals via email

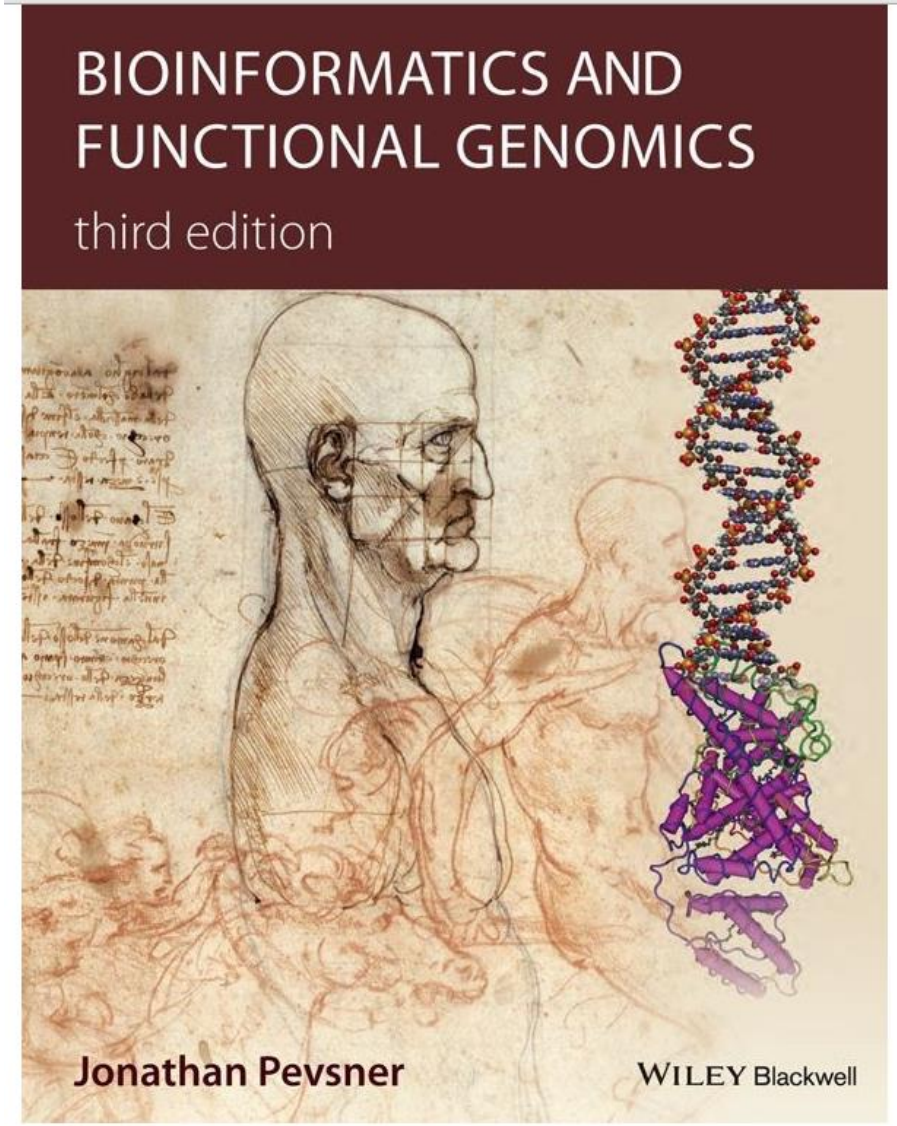
Book

- Mostly (Parts I and II)
- Few classes from Part III and outside the book

book web link (http://www.bioinfbook.org)

Ebook: Vitalsource.com (not affiliated and no monetary/other benefits)

Note that we need the THIRD Edition for this class



Bioinformatics and Functional Genomics

3rd ed.
Jonathan Pevsner

[M]: Contain some Math

Sometimes the class content for a week can spill over to the next week

Tentative Course Overview

1. **Introduction to Bioinformatics**
   - Syllabus-Book-Final Project-Grades-Introduction to Bioinformatics-Basics
2. **Command line resources introduction**
   - Linux OS-Eutils-R/Rstudio-Hands-on exercises
3. **Access to Data**
   - Where is the data coming from?-How is it stored?-NCBI-UniProt;-Ensembl-How to retrieve data?-Common prompt/Web-interface-Handson/Quiz
4. **Pairwise Sequence Alignment-I [M]**
   - Homology-Orthologs/Paralogs-Scoring Matrices-Derivation-Popular scoring matrices PAM/BLOSUM-Algorithm used for Alignment-Hands-on exercises-Quiz
5. **Pairwise Sequence Alignment-II [M]**

[M]: Contain some Math

Course Overview

6. **BLAST [M]**
   - Applications of sequence alignment-NCBI Blast interface-Details of BLAST-E-scores-Databases used for search-How to carry out meaningful searches

7. **Multiple Sequence Alignment (MSA) [M] (book MSA chapter comes after Adv. DB searching)**
   - What is MSA?-MSA using ClustalW-Alternative applications, Tcoffee, ProbCons, MUSCLE etc.-MSA in the genomic context-Hands-on and Quiz

8. **Advanced Database Searching**
   - PSSM, PSI-BLAST; HMM; Sensitivity in Searches; NextGen sequencing

9. **Molecular Phylogeny and Evolution**
   - Phylogenetic tress and explain their parts; How trees are created and what are the different methods-Positive and Negative selection and evolution

Course Overview

Mid-term

- Multiple choice/web site search-Open Book/Web in-class exam

10. NextGen Sequencing

- NGS data generation-FASTQ, SAM/BAM and VCF file formats-How reads are aligned to reference genome-Genome variants-Variant calling-Consequence of variants in individual genomes

11. Bioinformatic approaches to RNA and Gene Expression

- RNA-types, measuring RNA (Microarrays technique; RNAseq)-Exploratory data Analysis-Visualization-Statistics of quantifying RNA-basics of t-test

Course Overview

12. Structural Bioinformatics/Functional Genomics

- Why the need for 3D structures-How to connect to 1D sequence to 3D structures-How 3D structure is related to biological function-Protein Data Bank (PDB)-Why need 3D structure-Analysis?

13. Genomic Variations and Phenotypic Effect Predictions [M]

- Theory behind variant impact prediction tools (MutationTaster, SIFT and Polyphen2)-Pros and Cons of using prediction tools for impact analysis

14. Introduction to Molecular Modeling

- Intro to PubChem-Connection of PubChem to 3D database and UniProt-Simulate Protein-Protein, Protein-Ligand interactions-Visualization and Analysis

Find a Gene; Final Project Student Presentations

- **Final Project-Student Presentations-More on first class**

Grades

- In-class work, Computer Lab/Problems (end of each relevant chapter in the book) and ongoing assignments (help you complete final project) 50%

- Mid-term 25%

- Final Exam (Presentation/Write-up): 25%
  - 15 minutes presentation and 5 minutes for Q&A

- Contact Email:

  ravichandran@hood.edu

More on Mid-term (open-web exam; in-class exam)

- Two parts
- First part (Basic Biology/Bioinformatics)
  - Multiple-choice
  - Some brief answer type questions
- Second part (carries more points)
  - Have to use online servers to answer the questions
  - 4 questions
  - Expect to provide elaborate answers

General Class Structure

- Lecture (~ half of the class time)
  - Lectures (*helpful if you can read the chapter before the class*)
- Remaining time
  - Discussion Questions
  - Computer Labs (from the book assigned in class)
  - Assignments (mostly every alternate week)
    - Upload your solutions to the instructor before leaving the class
  - Self-Quiz (end of each chapter)
    - Optional but helpful

General Class Structure

- Absence
  - Planned absence: Please email me ahead of time
    Emergency: I will work with you

- I **will not** cover **all the** materials for each chapter in the book.
  - Will cover important topics
    - what is important, is based on my experience
- Depending on the time and progress, I may decide to expand/drop a chapter

General Class Structure

- Math
  - Basic knowledge with motivation
- R
  - I am sure you all have exposure to R/Other programming languages(?)
  - R-Programming is not needed but basic knowledge will be helpful
- **Any Questions, so far????**

Technology

- Windows (default)
  - Modern Windows
- Linux OS
  - Via bio-linux (account is created and ready)
- Software for the class will be available in the class computers
- Please use Class Computers for the class work

Communication

- BlackBoard
- Each class materials will be stored in a separate folder
- I mostly upload materials for the class a day before the class

Online servers that require user accounts

- NCBI
  - myNBCI (user account)
  - https://www.ncbi.nlm.nih.gov/
- Ensembl
- UCSC etc.
- Galaxy
  - https://usegalaxy.org/ (user account)
- Create the accounts ahead of time

Teaching Bioinformatics

- Software/Browser/OS
  - Version issues
  - Scripts sometimes fail
  - Web connectivity issues
  - NCBI/Ensembl might change their genomic browsers without notice; Genomic browsers behave sometimes differently with different browsers/OS
  - Book has few typos; some exercises might be different
- Please be patient, thanks!

Applications

Tagalog
Українська
Tiếng Việt
粵語
中文
Edit links

The 1918 flu caused an unusual number of deaths, possibly due to it causing a cytokine storm in the body.[8] (The current H5N1 bird flu, also an influenza A virus, has a similar effect.)[8] The Spanish flu virus infected lung cells, leading to overstimulation of the immune system via release of cytokines into the lung tissue. This leads to extensive leukocyte migration towards the lungs, causing destruction of lung tissue and secretion of liquid into the organ. This makes it difficult for the patient to breathe. In contrast to other pandemics, which mostly kill the old and the very young, the 1918 pandemic killed unusual numbers of young adults, which may have been due to their healthy immune systems mounting a too-strong and damaging response to the infection.[9]

The term "Spanish" flu was coined because Spain was at the time the only European country where the press were printing reports of the outbreak, which had killed thousands in the armies fighting World War I. Other countries suppressed the news in order to protect morale.[10]

Fort Dix outbreak [edit]

Main article: 1976 swine flu outbreak

In 1976, a novel swine influenza A (H1N1) caused severe respiratory illness in 13 soldiers with 1 death at Fort Dix, New Jersey. The virus was detected only from January 19 to February 9 and did not spread beyond Fort Dix.[11] Retrospective serologic testing subsequently demonstrated that up to 230 soldiers had been infected with the novel virus, which was an H1N1 strain. The cause of the outbreak is still unknown and no exposure to pigs was identified.[12]

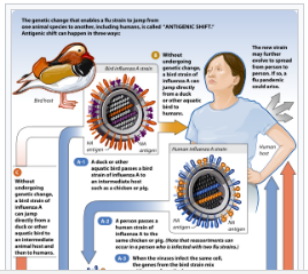
Russian flu [edit]

The 1977–1978 Russian flu epidemic was caused by strain *Influenza A/USSR/90/77 (H1N1)*. It infected mostly children and young adults under 23 because a similar strain was prevalent in 1947–57, causing most adults to have substantial immunity. Because of a striking similarity in the viral RNA of both strains – one which is unlikely to appear in nature due to antigenic drift – it was speculated that the later outbreak was due to a laboratory incident in Russia or Northern China, though this was denied by scientists in those countries.[13][14][15] The virus was included in the 1978–1979 influenza vaccine.[16][17][18][19]

See also 1889–1890 flu pandemic for the earlier Russian flu pandemic caused either by H3N8 or H2N2

2009 A(H1N1) pandemic [edit]

Main article: 2009 flu pandemic



In the 2009 flu pandemic, the virus isolated from patients in the United States was found to be made up of genetic elements from four different flu viruses – North American swine influenza, North American avian influenza, human influenza, and swine influenza virus typically found in Asia and Europe – "an unusually mongrelised mix of genetic sequences."[20] This new strain appears to be a result of reassortment of human influenza and swine influenza viruses, in all four different strains of subtype H1N1.

Preliminary genetic characterization found that the hemagglutinin (HA) gene was similar to that of swine flu viruses present in U.S. pigs since 1999, but the neuraminidase (NA) and matrix protein (M) genes resembled versions present in European swine flu isolates. The six genes from American swine flu are themselves mixtures of swine flu, bird flu, and human flu viruses.[21] While viruses with this genetic makeup had

Year 1977
Host Human
Protein HA
Subtype H1N1

5 protein sequences after collapsing (7 total)

Accession	Length	Host	Protein	Subtype	Country	Region	Date	Virus name
ABD60944	566	Human	HA	H1N1	Hong Kong	N	1977	Influenza A virus (A/Hong Kong/117/1977(H1N1))
ABO44134	566	Human	HA	H1N1	China	N	1977	Influenza A virus (A/Tientsin/78/1977(H1N1))
ABD95350	566	Human	HA	H1N1	Russia	N	1977	Influenza A virus (A/USSR/90/1977(H1N1))
APC57869	566	Human	HA	H1N1	USSR		1977	Influenza A virus (A/USSR/90/1977(H1N1))
ABD60933	566	Human	HA	H1N1	Russia	N	1977	Influenza A virus (A/USSR/92/1977(H1N1))

NCBI

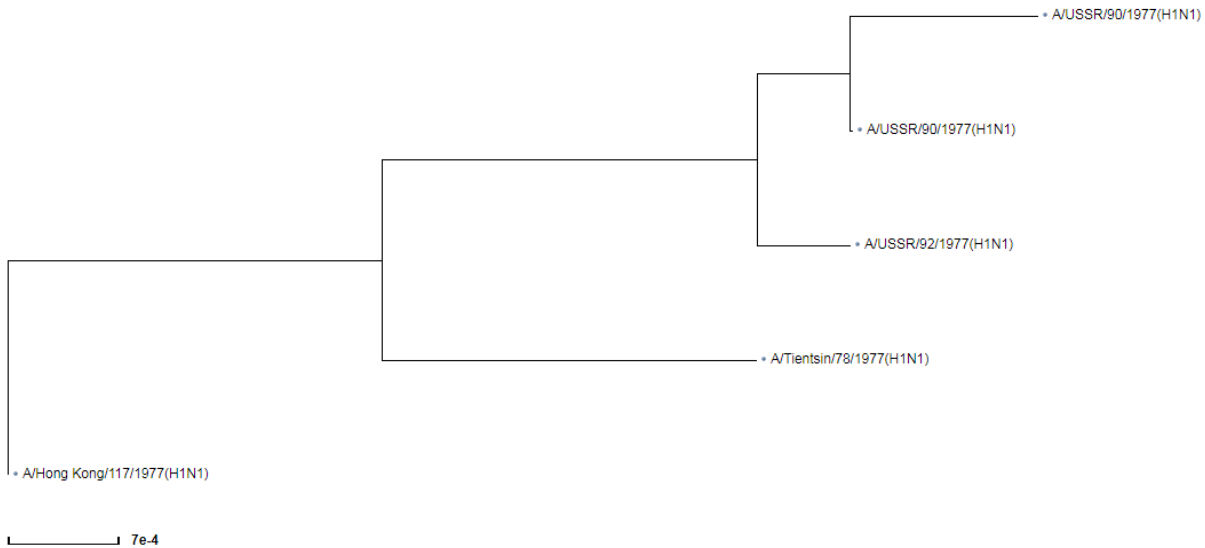
Influenza Virus Resource
Information, Search and Analysis

HOME | SEARCH | SITE MAP | Flu home | Database | Genome Set | Alignment | Tree | BLAST | Annotation | FTP | Help | Contact us

Multiple alignment for 5 protein sequences. Alignment length is 566.

Show tree | Download alignment | Print-friendly version | Go to position 1 Go | BLAST 2 seq. | Clear

```
Position. 1........10........20........30........40........50........60........70........80........90........100.......110.......120.......130.......140.......150.......160.......170.......180.....
Consensus MKAKLLVLLCALSATDADTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCRLKGIAPLQLGKCSIAGWILGNPECESLFSKKSWSYIAETPNSENGTCYPGYFADYEELREQLSSVSSFERFEIFPKERSWPKHNVTRGVTASCSHKGKSSFYRNLLWLTEKNGSYPNLSKSYVNNKEKEVL
ABD60944 ..........................................................................................................................................
ABO44134 ..........................................................................................................................................
ABD95350 ..........................N.............................................................................................................
APC57869 ..........................N.....................................A.....................................D.....................
ABD60933 ..........................................V.....................................................................................
```



"it is better 100 guilty Persons should escape than that one innocent Person should suffer" **Benjamin Franklin**

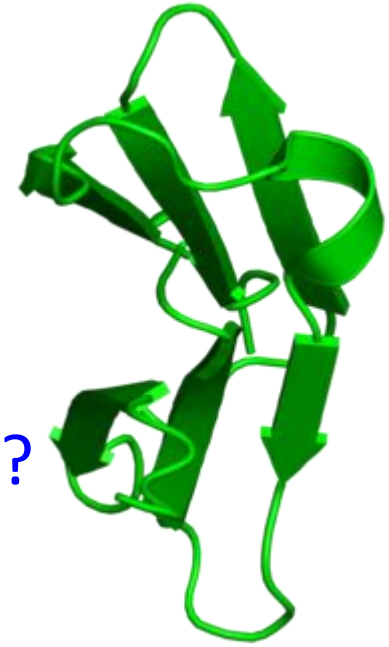
Applications

- Criminal Justice system https://www.innocenceproject.org/dna-exonerations-in-the-united-states/
  - Genetic evidence and exoneration
  - First event happened in 1989
  - Since then ~367 cases had been resolved using DNA exoneration
  - DNA evidence is admitted in criminal trials in almost all states in USA
  - Disease, Therapy(?)
    - Breast cancer: Mutations in BRCA1/2 genes
    - Achondroplasia (Dwarfism): Mutations in FGFR3 gene
- Evolving area
  - Systems' view is lacking

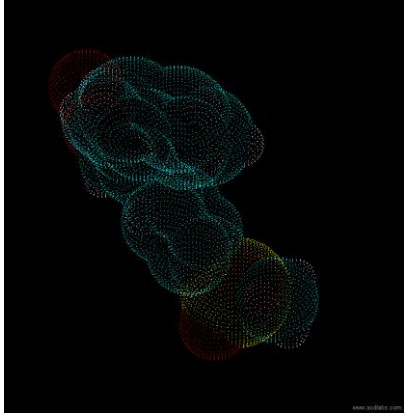
Outline for today

- Basics
  - What is bioinformatics?-Formal definitions
- Information
  - Sequence
    - Databases
    - Approximation and relevance to chemistry
  - Structure (3D)
    - Experiments
    - Models
- What can we learn from the information?
  - Function

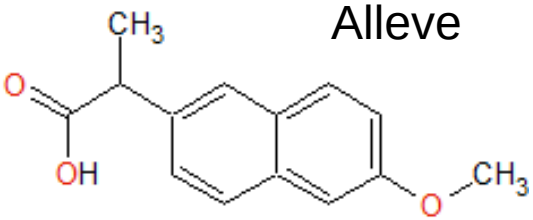
```
>sp|P04792|HSPB1_HUMAN Heat shock protein beta
MTERRVPFSLLRGPSWDPFRDWYPHSRLFDQAFGLPRLPEEWSQWLGGSSWPGYVRPLPP
AAIESPAVAAPAYSRALSRQLSSGVSEIRHTADRWRVSLDVNHFAPDELTVKTKDGVVEI
TGKHEERQDEHGYISRCFTRKYTLPPGVDPTQVSSSLSPEGTLTVEAPMPKLATQSNEIT
IPVTFESRAQLGGPEAAKSDETAAK
```



Outline for today



- How can we learn from the information?
  - Comparison (Sequence)
  - And Lots of help from Math/Statistics Information (structure)
  - 1D → 3D → Function → Drug Discovery



- Two selected genetic disorders/disease (Sickle Cell Anemia, Lactose Intolerance)

Common Definitions of Bioinformatics

- *"Use of computer databases and algorithms to analyze the proteins, genes and the complete collection of DNA that comprises the organism (genome)"* *Pevsner, Bioinformatics and Functional Genomics, 2015 III Ed.*

- *Research, Development **or** Application of data to analyze and/or build models to understand the biological mechanisms*
  - *Collection, maintenance and analyzing*

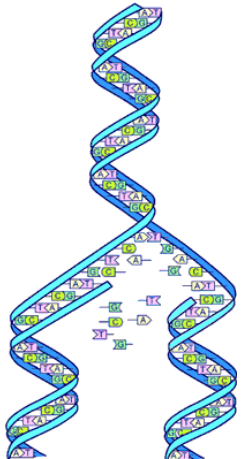
Three perspectives

- First
  - Cell
    - Study of individual genes/proteins and their collections
- Second
  - Individual organisms
  - Different regions/different development stages/different times
- Third
  - Genomics: The tree of life
  - How to group many organisms

Motivation

- Existing Diversity
  - People (different races, different traits etc.)
  - Animals
  - Fungi etc.

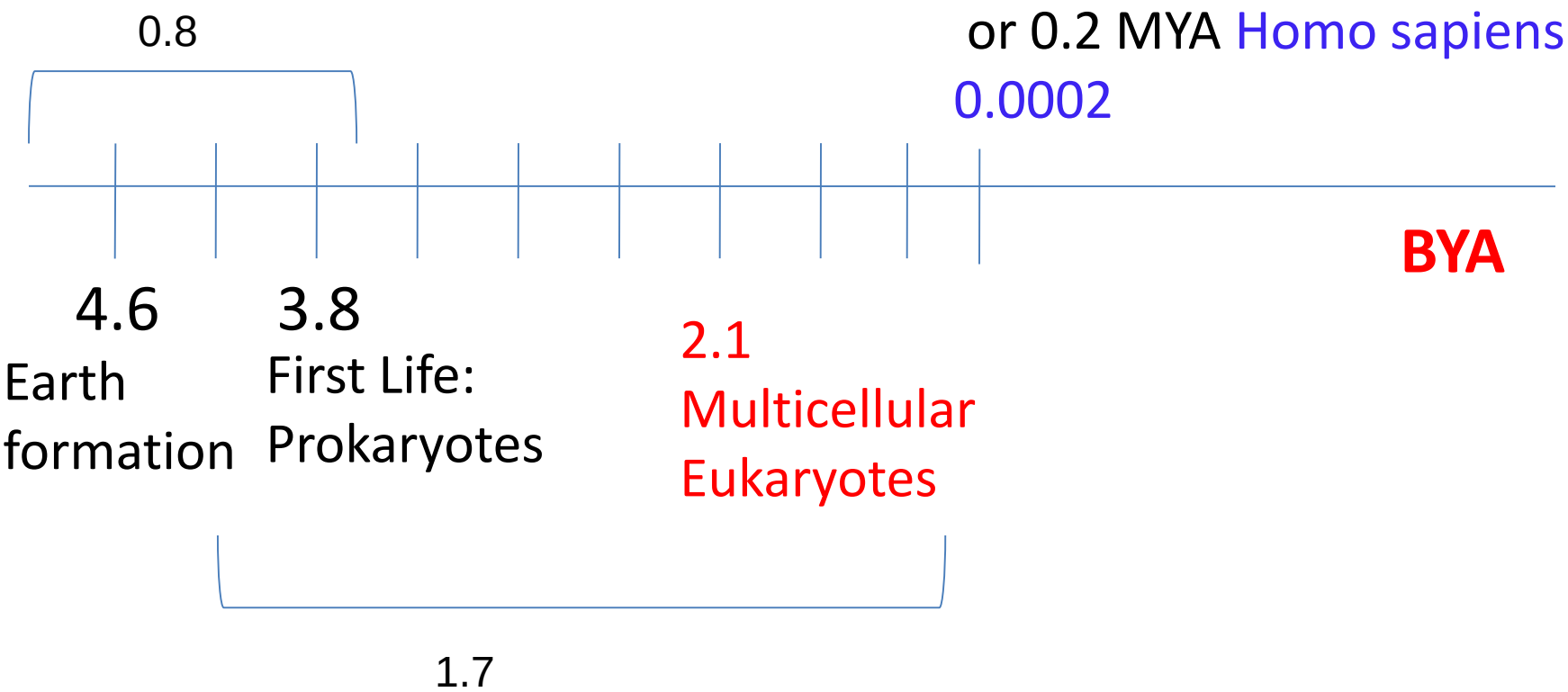
- At the micro cellular level
  - There is so much (??) similarity
  - We will spend almost all our time in this world
    - Key players: Proteins, DNA & RNA
      - Their interaction(s) in a limited way



Motivation

- Use the genomic differences to explain the phenotypical differences
  - Cell structure is same; Small differences make us who we are

History of Life on Earth



Life

- Ability to reproduce itself
  - Many other definitions!
- All life evolved from a common ancestor
  - Evolution
    - inheritance, passing of characteristics from parents, variation
  - Variation
    - Mutation-Genetic modification-sexual recombination-viruses etc

Prokaryotes
Dr. Woese proposed a division (1977)

"The branch of science concerned with classification, especially of organisms"

Taxonomy

Molecular Biology for Computer Scientists, Chapter 1, Lawrence Hunter

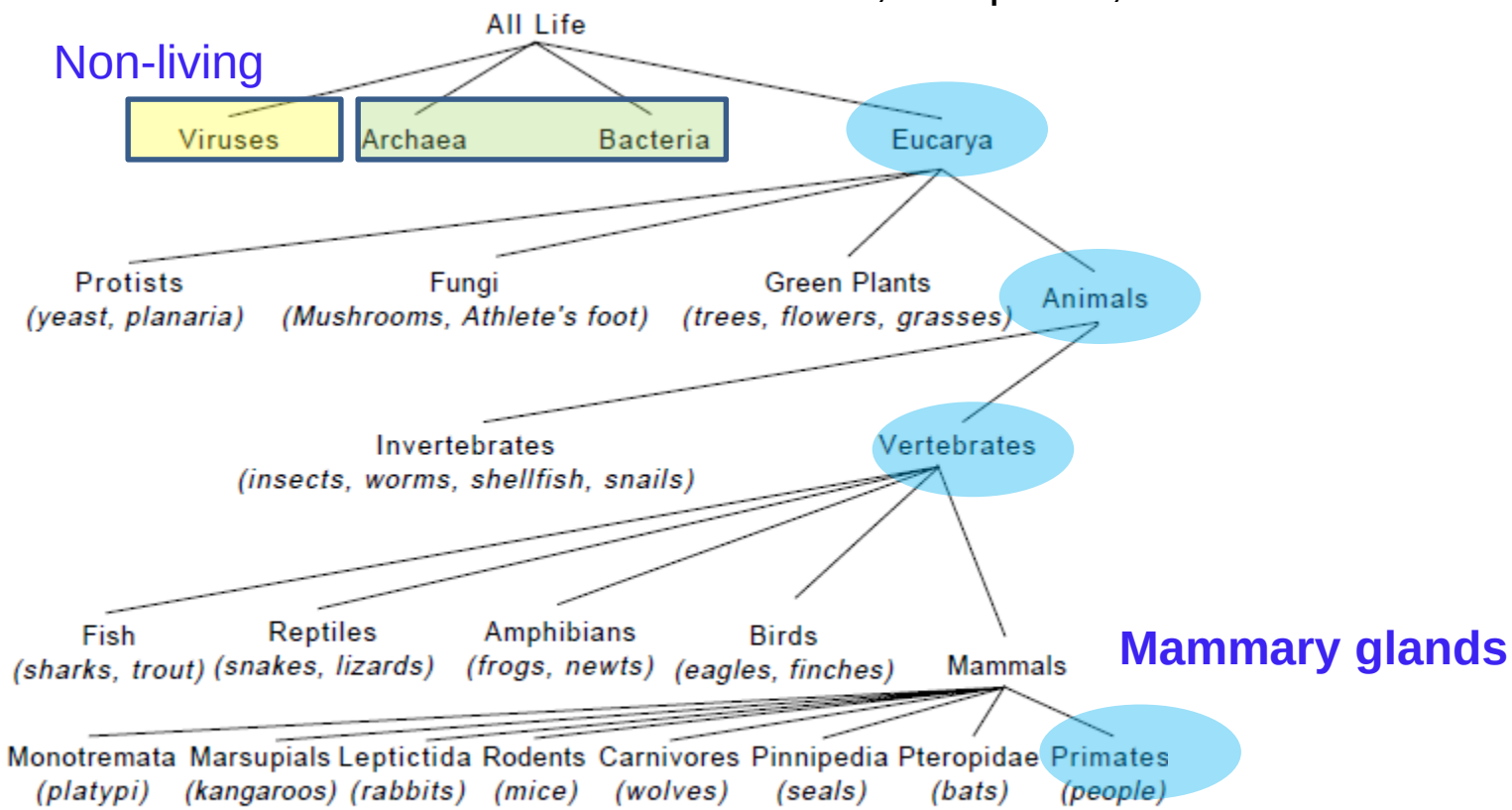
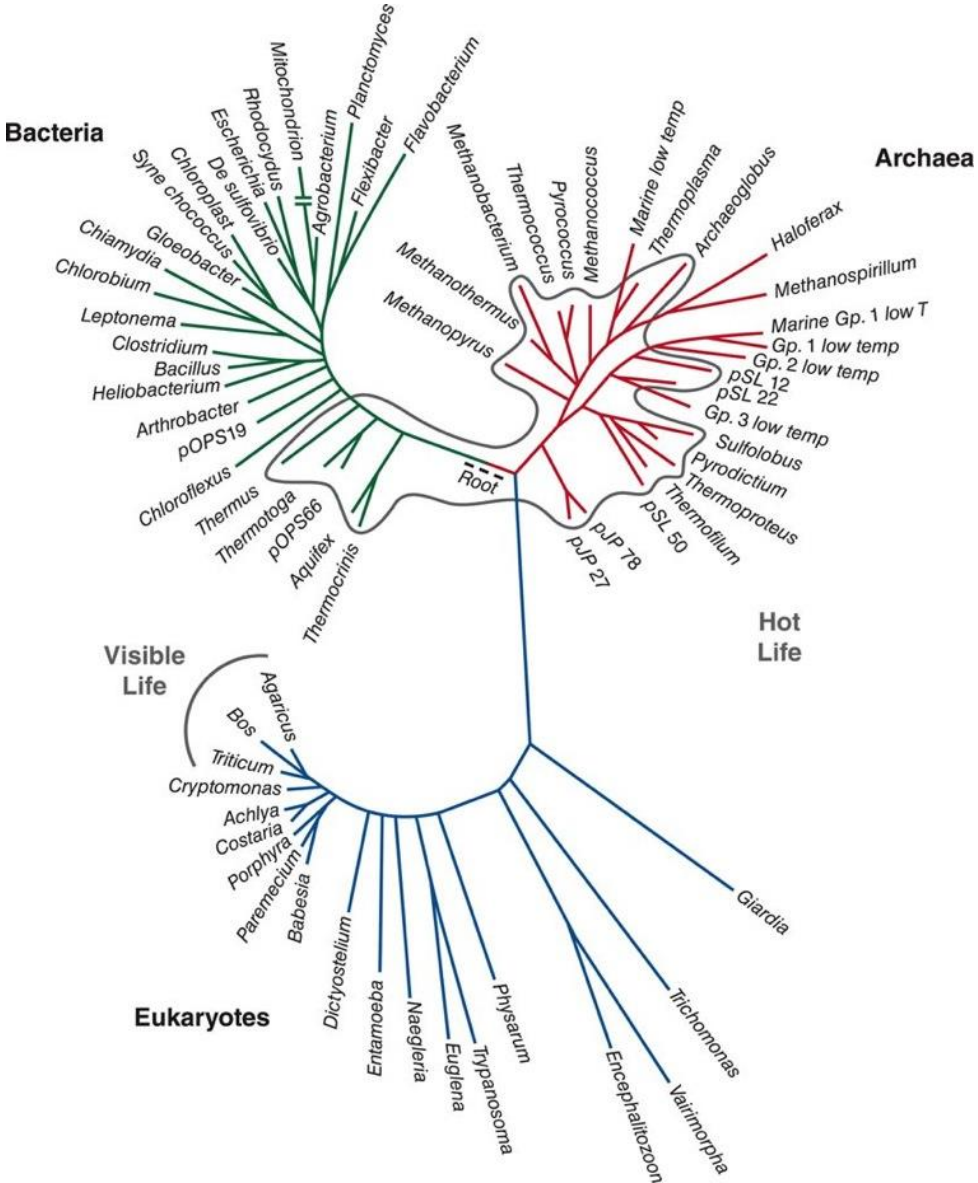


Figure 1. A very incomplete and informal taxonomic tree. Items in italics are common names of representative organisms or classes. Most of the elided taxa are Bacteria; Vertebrates make up only about 3% of known species.

Tree of life

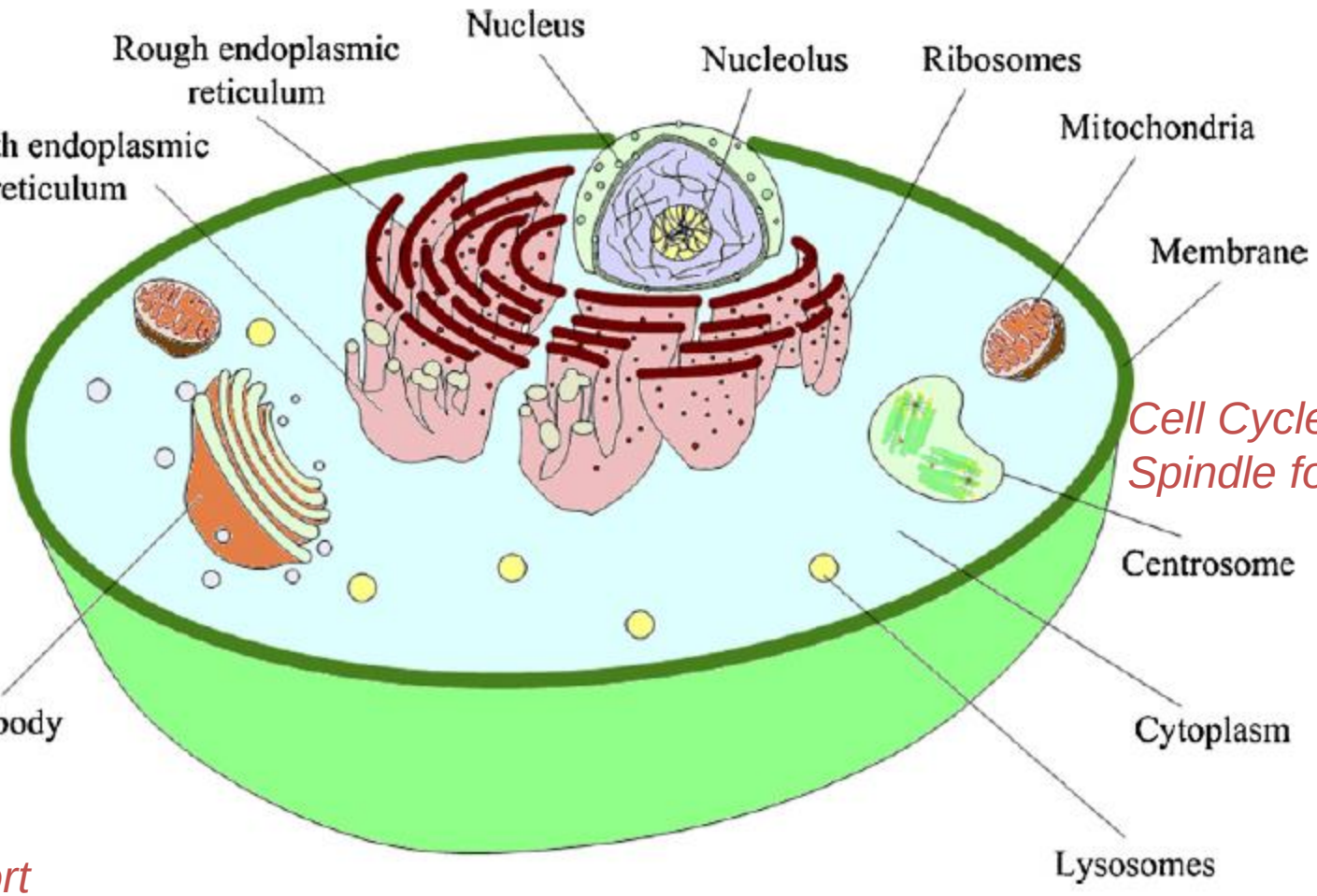
Viruses not part of the tree



Cell as a factory

Human

K. Khoshmanesh et al. / Computational Biology and Chemistry 32 (2008) 315–331



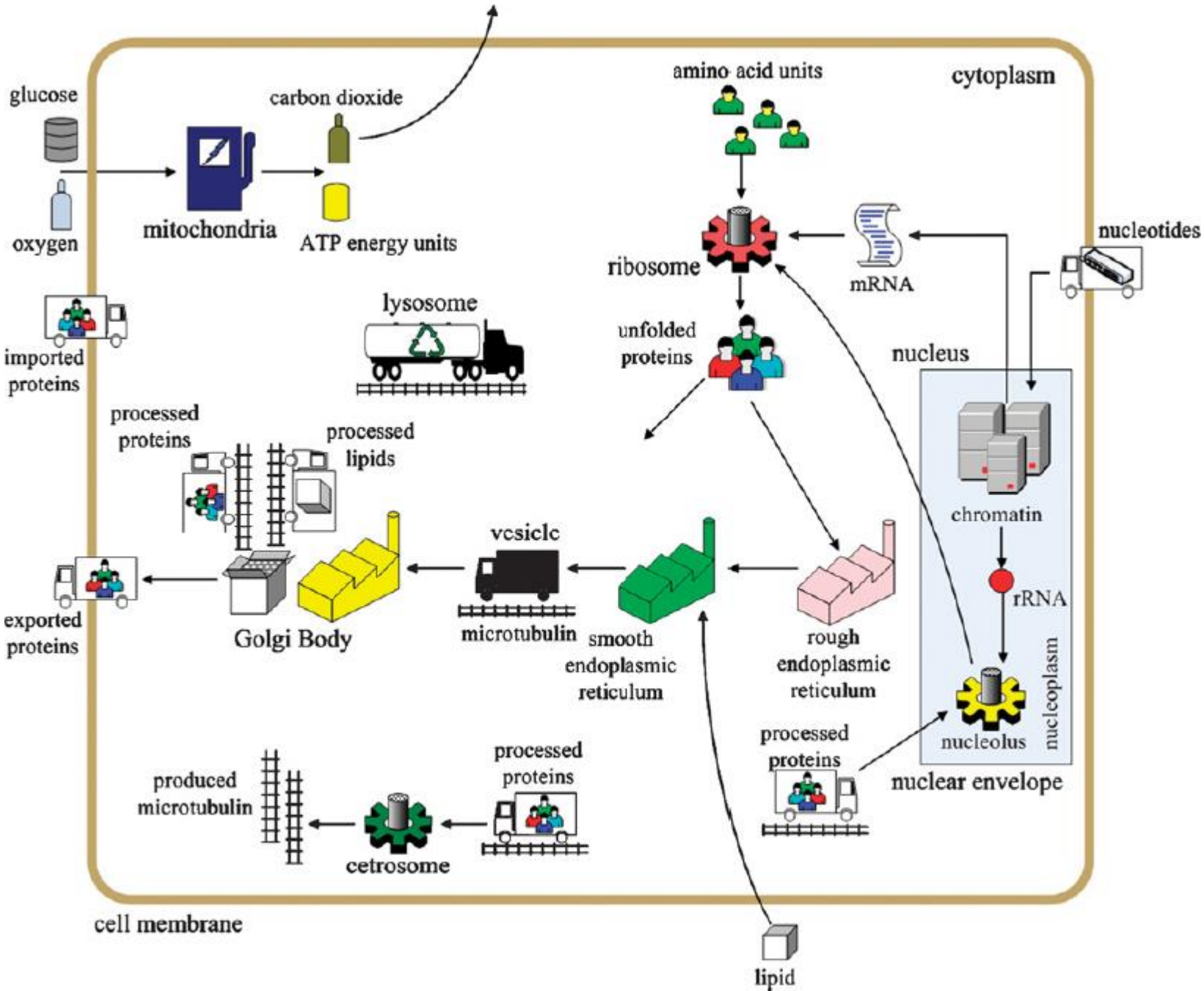
Prodn. of lipids and steroids

Cell Cycle- Spindle formation

Processing of proteins after synthesis/export

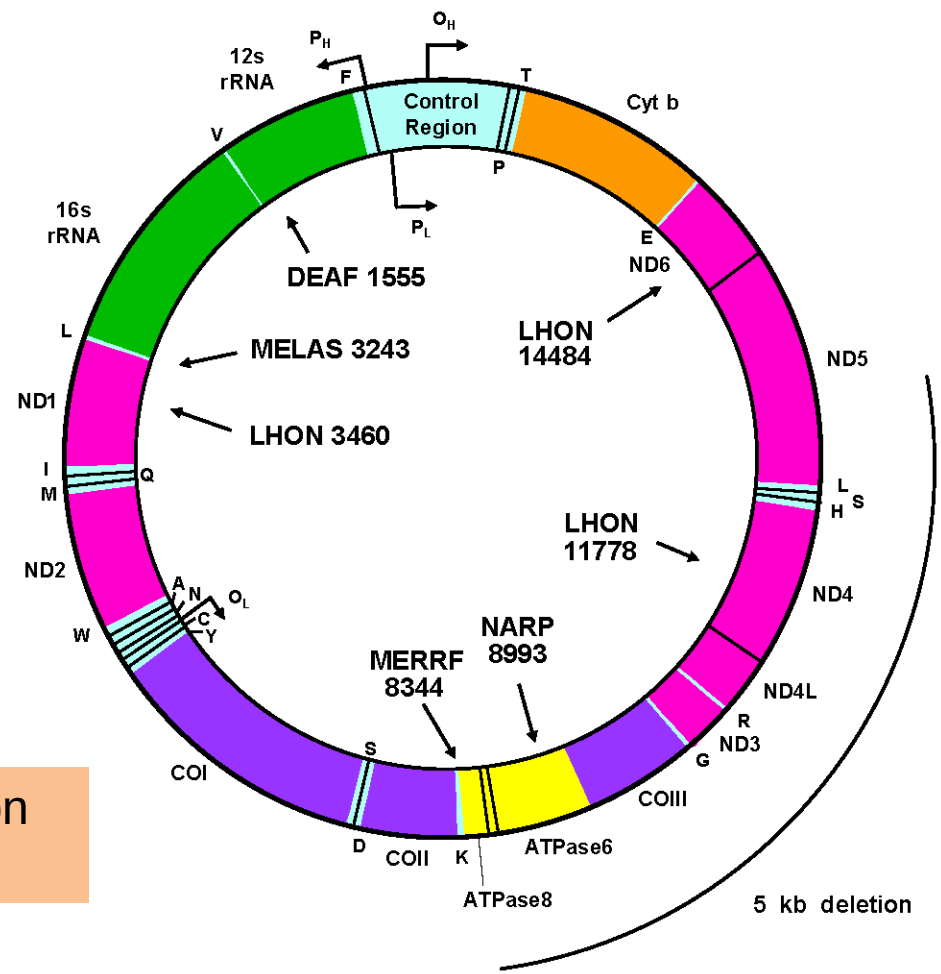
Roughly 37.2 trillion cells in our body

Typical cell (across length) 10×10^{-6}m


glucose
oxygen
carbon dioxide
mitochondria
ATP energy units
amino acid units
cytoplasm
ribosome
mRNA
nucleotides
lysosome
imported proteins
unfolded proteins
nucleus
processed proteins
processed lipids
chromatin
vesicle
rRNA
exported proteins
Golgi Body
microtubulin
smooth endoplasmic reticulum
rough endoplasmic reticulum
nucleolus
nucleoplasm
processed proteins
nuclear envelope
produced microtubulin
processed proteins
cetrosome
cell membrane
lipid

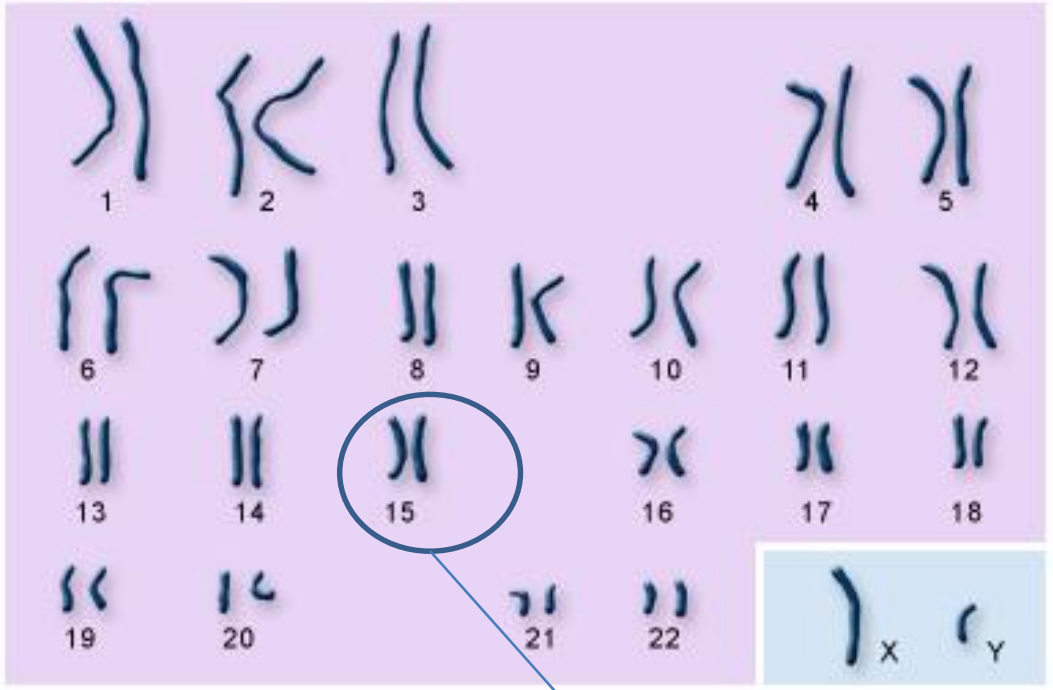
Morbid Map of the Human mtDNA Genome

Inherited from
Mom to children



Fingerprints of Evolution
Circular Genome

Chromosome



U.S. National Library of Medicine

Book of Life: 2 Multi-Volume Set

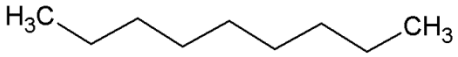
DAD

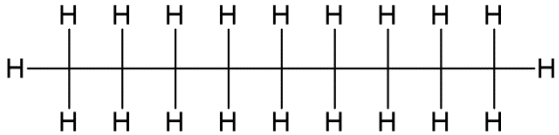
MOM

To understand the language of DNA, we need to understand some Chemistry/Biochemistry

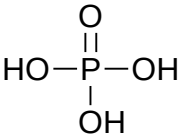
Brief Introduction

Periodic Table

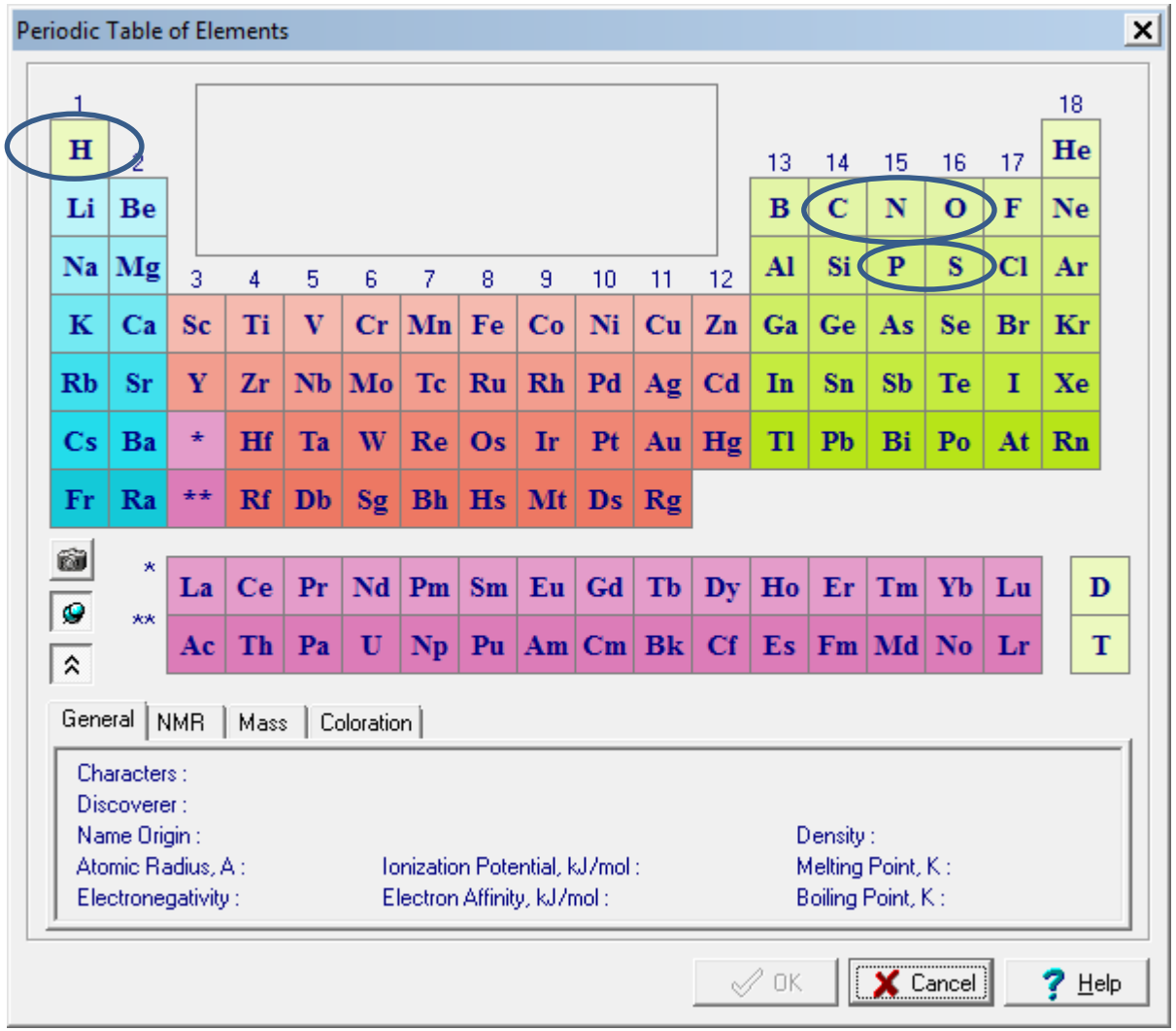


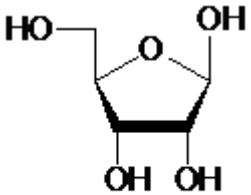


H-O-H



Phosphate

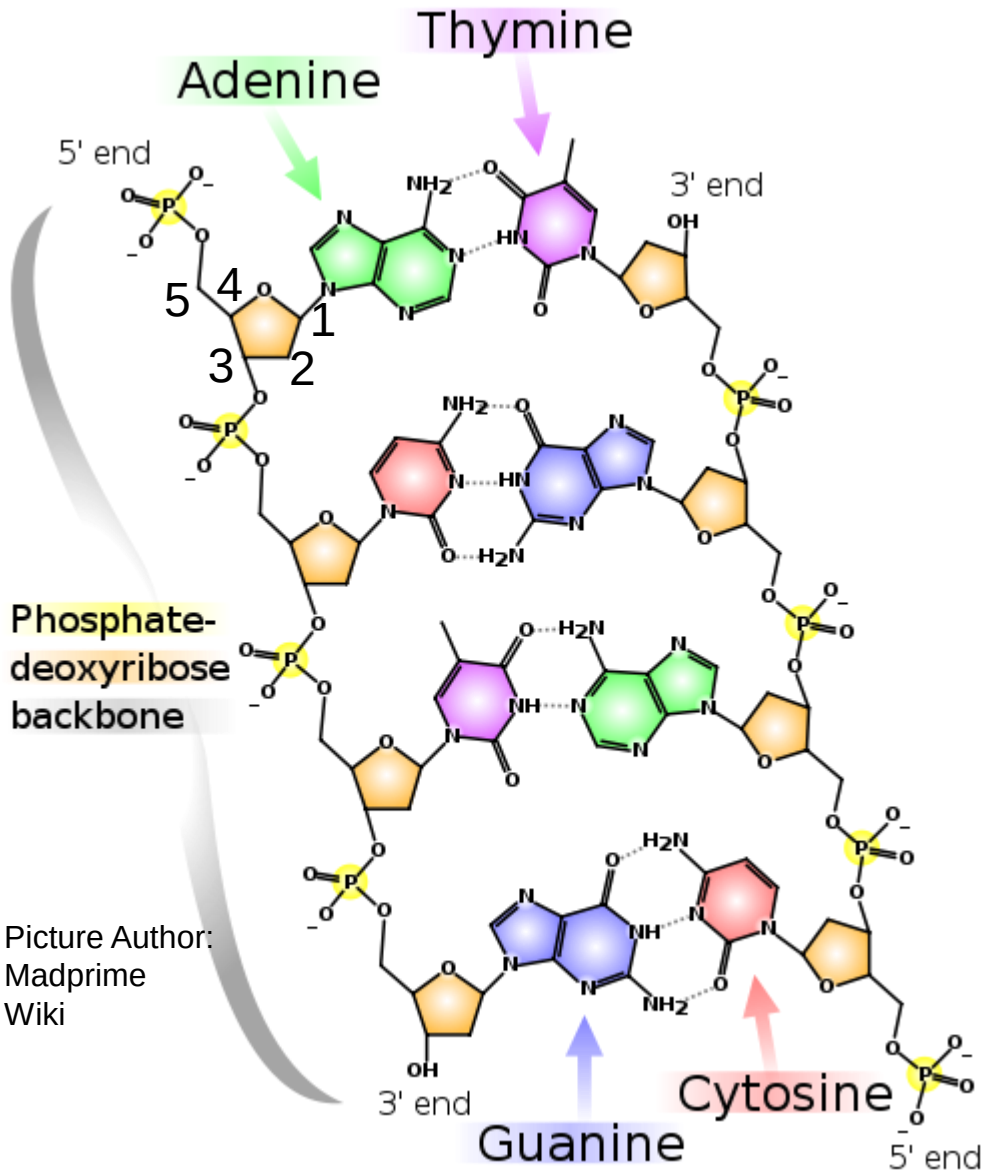




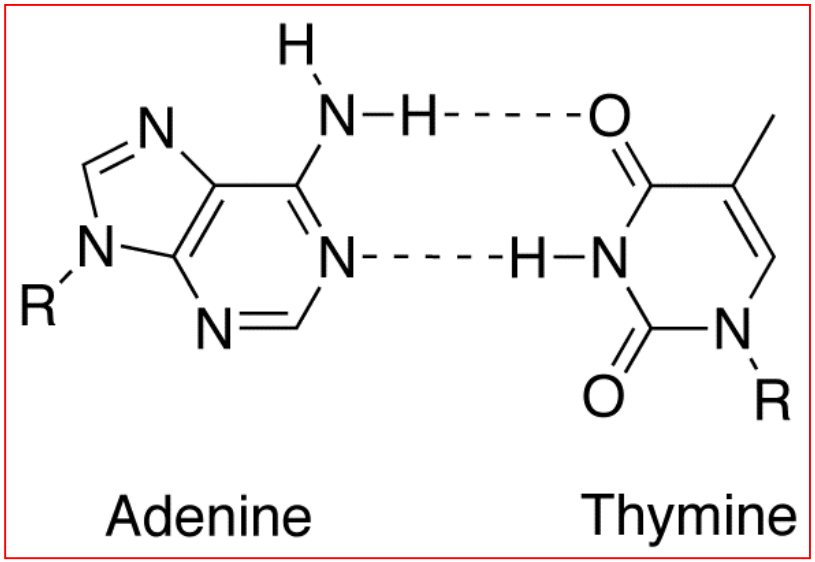
Ribose

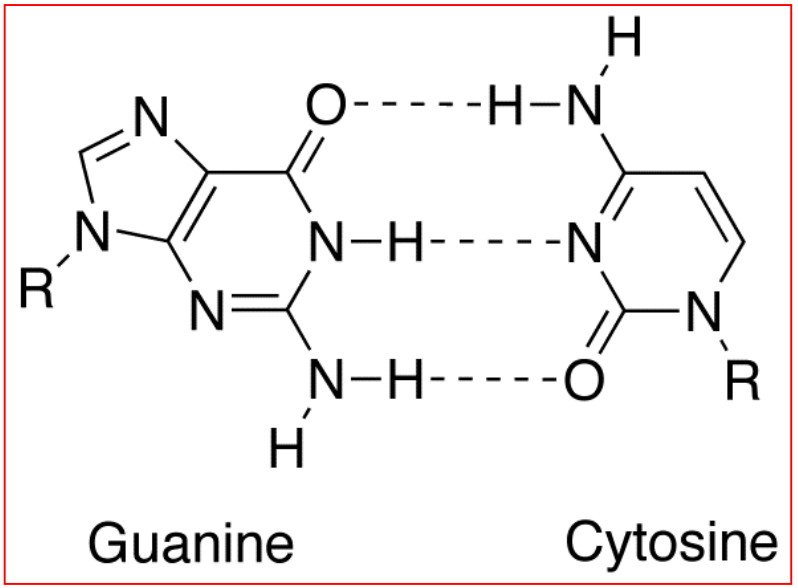
Periodic Table from ChemSketch 2015 (Freeware)

DNA has direction



Picture Author:
Madprime
Wiki





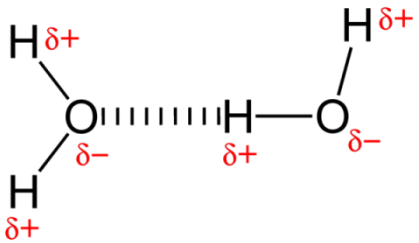
http://en.wikipedia.org/wiki/File:GC_base_pair_jypx3.png

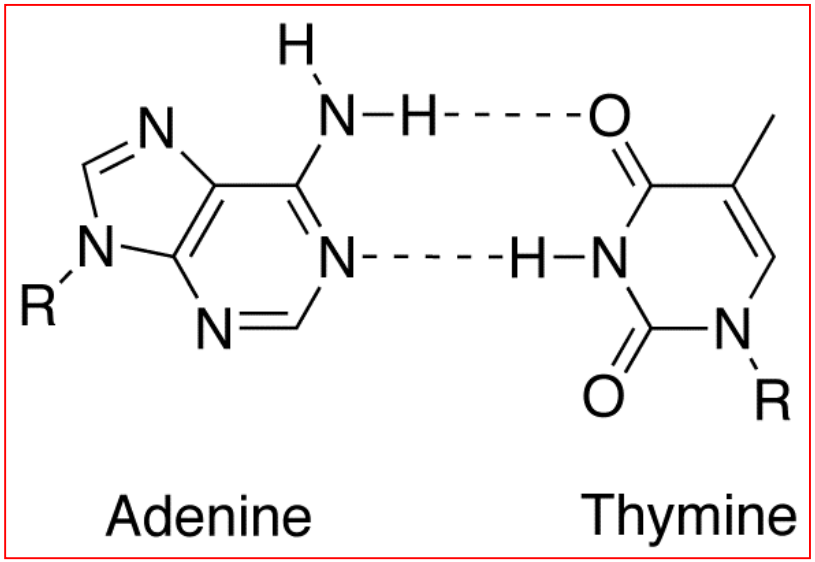
Biology is the chemistry that crawls

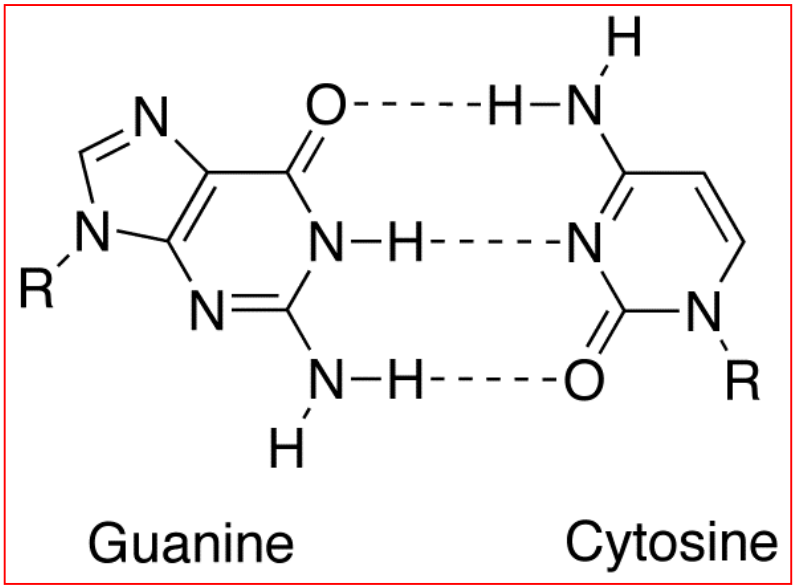
What holds the molecules together?

Atomic level: Chemistry Rules

- Bonded interactions
  - Covalent bonds
- Non-bonded interactions
  - H-bonds
    - Holds DNA
    - Makes drug binding work
  - Ionic, VDW etc.







H-bonds

Which pair is easy to break?

A-T

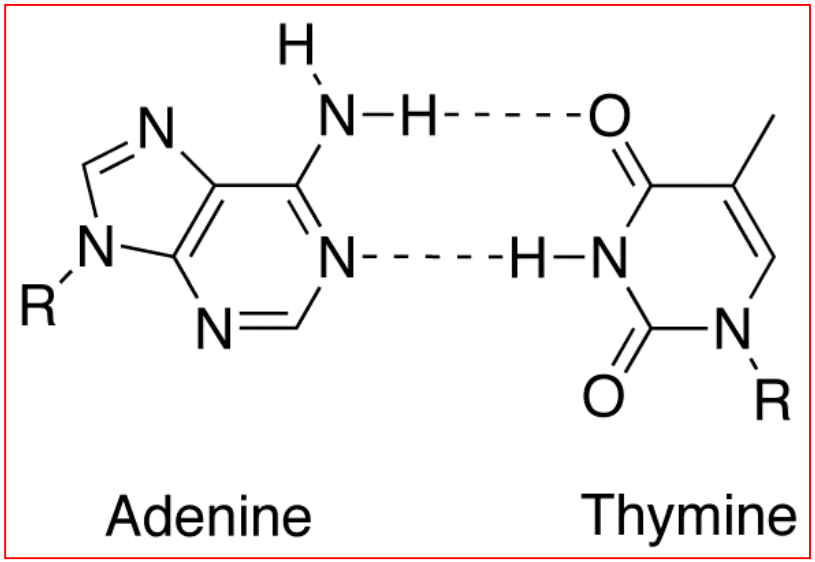
Or

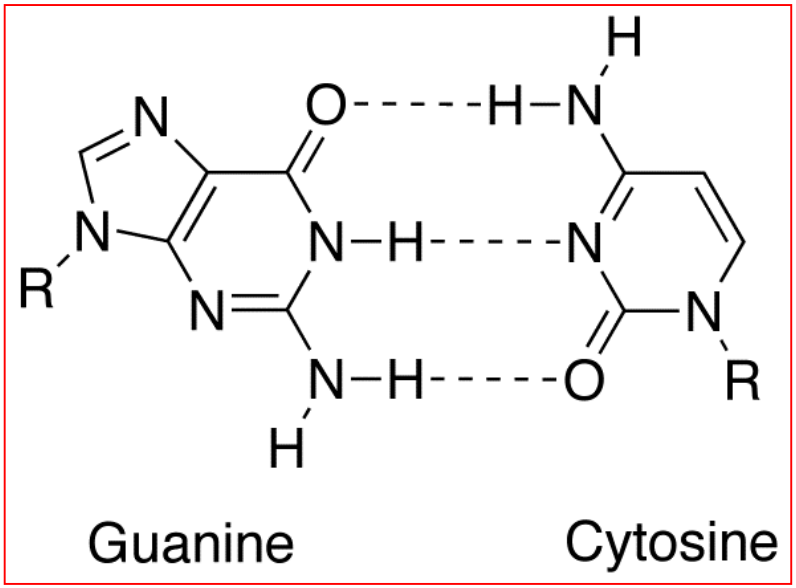
G-C

# of H-bonds

Can we think of A-T being the site for DNA actions such as
double-stranded → single-stranded

Replication and Transcription etc.

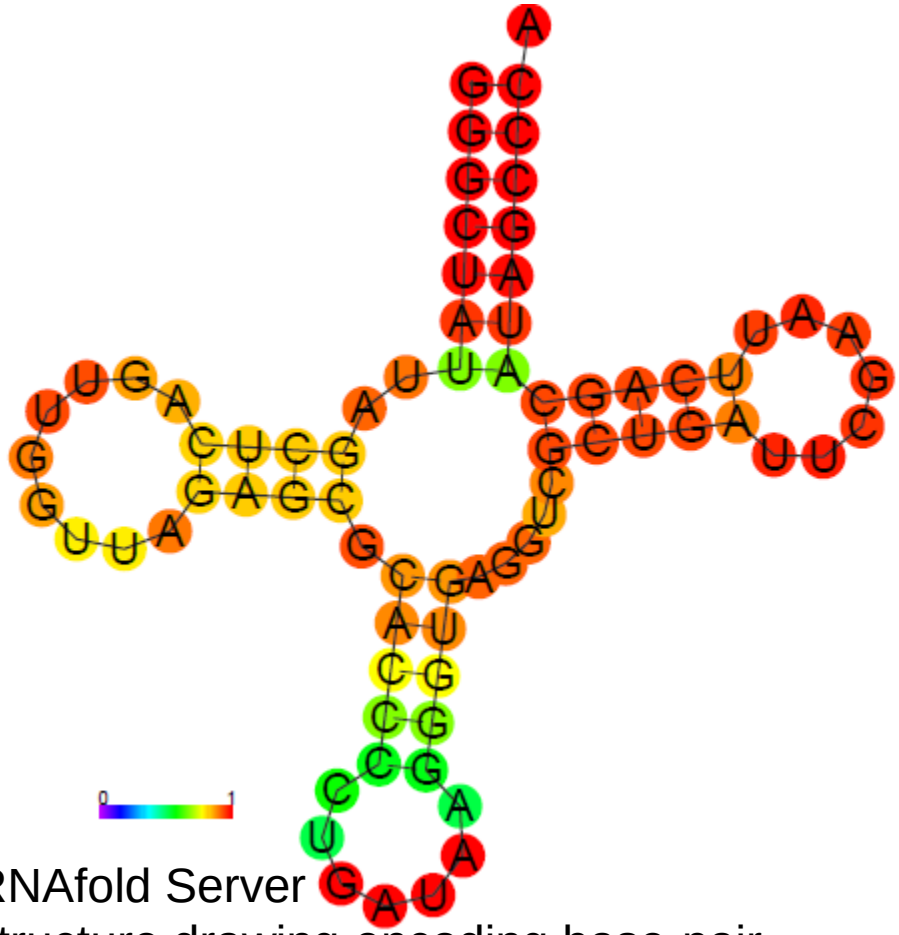




DNA/RNA 3D Structures

CGCGAATTCGGG
GCGCTTAAGCCC

GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCC
CUGAUAAGGGUGAGGUCGCUGAUUCGAAUUC
AGCAUAGCCCA



RNAfold Server structure drawing encoding base-pair probabilities

DNA/RNA

- Different types of DNA
  - B, -Z etc
- Different RNAs
  - mRNA
    - Nucleus → Ribosomes
  - NR (non-coding; 95% of all RNAs)
    - tRNA
    - rRNA

e-Genome
Genome → Computers

- Ecoli

a

0	1	1	0	0	0	1	1

  - 4 million bases
    - 4,000,000 * 1 Byte = 4,000,000
      - ~4 MB Hard drive

- Human
  - **~ 3,000,000,000** (3 Billion) bases (one copy)
    - Each cell
    - 3×10^9 * 1 Byte = 3×10^9 = 3 GB

How much DNA in a Genome?

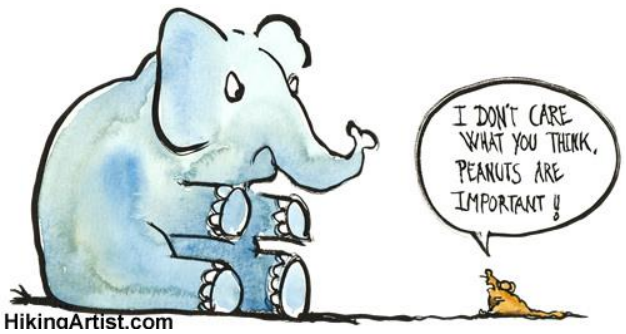
- Eukaryotic genome (haploid) size **vary widely**
- The common unit for quantifying DNA is C-value
- What is C-value?
  - # of base pairs in DNA or picograms (pg) of DNA
  - 1 pg $\cong$ 1Gb (gigabases)
    - more precisely 1 pg * 0.9869×10^9

Giga = 10^9
Mega = 10^6

10^{-12} g approximate equal to 1 picogram

Compare Eukaryotic species using C-Value

Complexity of organisms doesn't correlate with C-values



"*I Don't care what you think, PEANUTS are important*"

Species	Common name	C value (Gb)
S. Cerevisiae	Yeast	0.012
Dysidea Crawshagi	Sponge	0.054
Drosophila melanogaster	Fruit Fly	0.12
Oryza Sativa	Rice	0.47
Gallus domesticus	Chicken	1.23
Canis familiaris	Dog	2.9
Rattus norvegicus	Rat	2.9
Xenopaus laevis	African clawed frog	3.1
Homo sapiens	**Human**	**3.3**
Allium cepa	Onion	15
Lilium formosanum	Lily	36
Protopterus aethiopicus	Marbled lungfish	140
Amoeba proteus	Amoeba	290

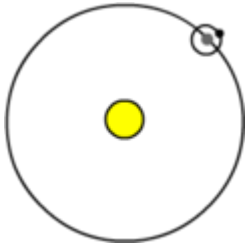
Based on Table 8.3 *from* Bioinformatics and Functional genomics III Ed.

Total length of DNA present in one adult human

(length of 1 bp) * (# of bp/cell) * (# of cells in body)

$= (0.34 \times 10^{-9}\ m)\ (6 \times 10^{9})\ (10^{13})$

$= 2.0 \times 10^{13} m = 12.42$ Billion Miles



~ 70 trips from Earth to the Sun & back

How can this lengthy molecule fit inside a tiny cell?

How can this lengthy molecule fit inside a tiny cell

http://commonfund.nih.gov/epigenomics/figure.aspx

Genome Organization

1) Chromatin (Histone wrapping)
2) Chromosome

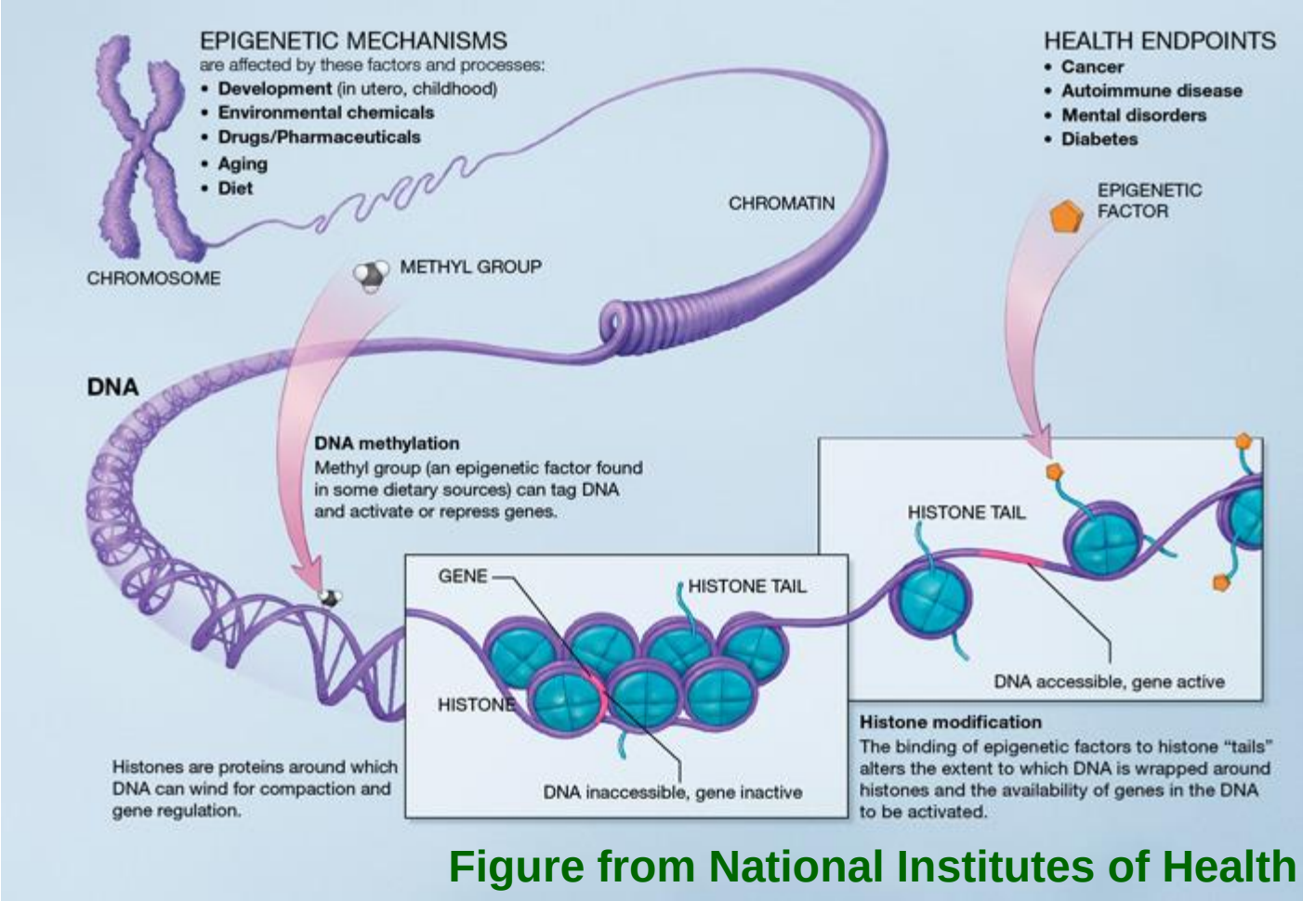
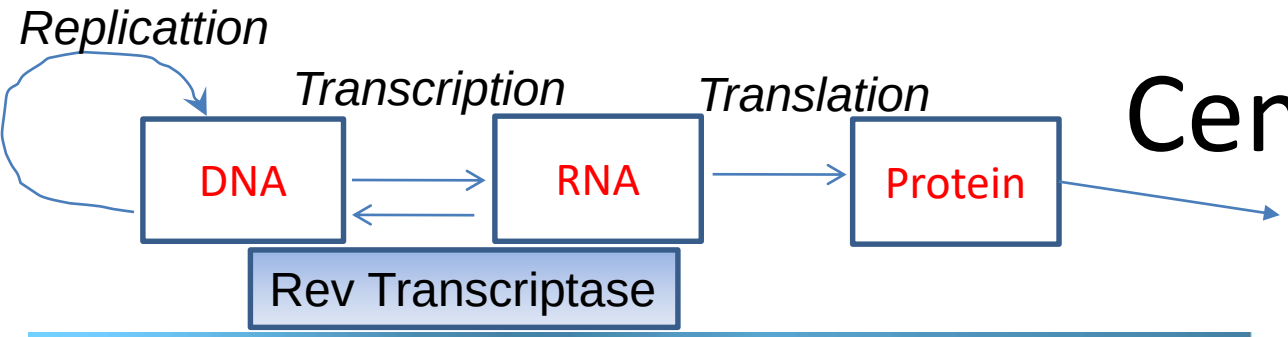
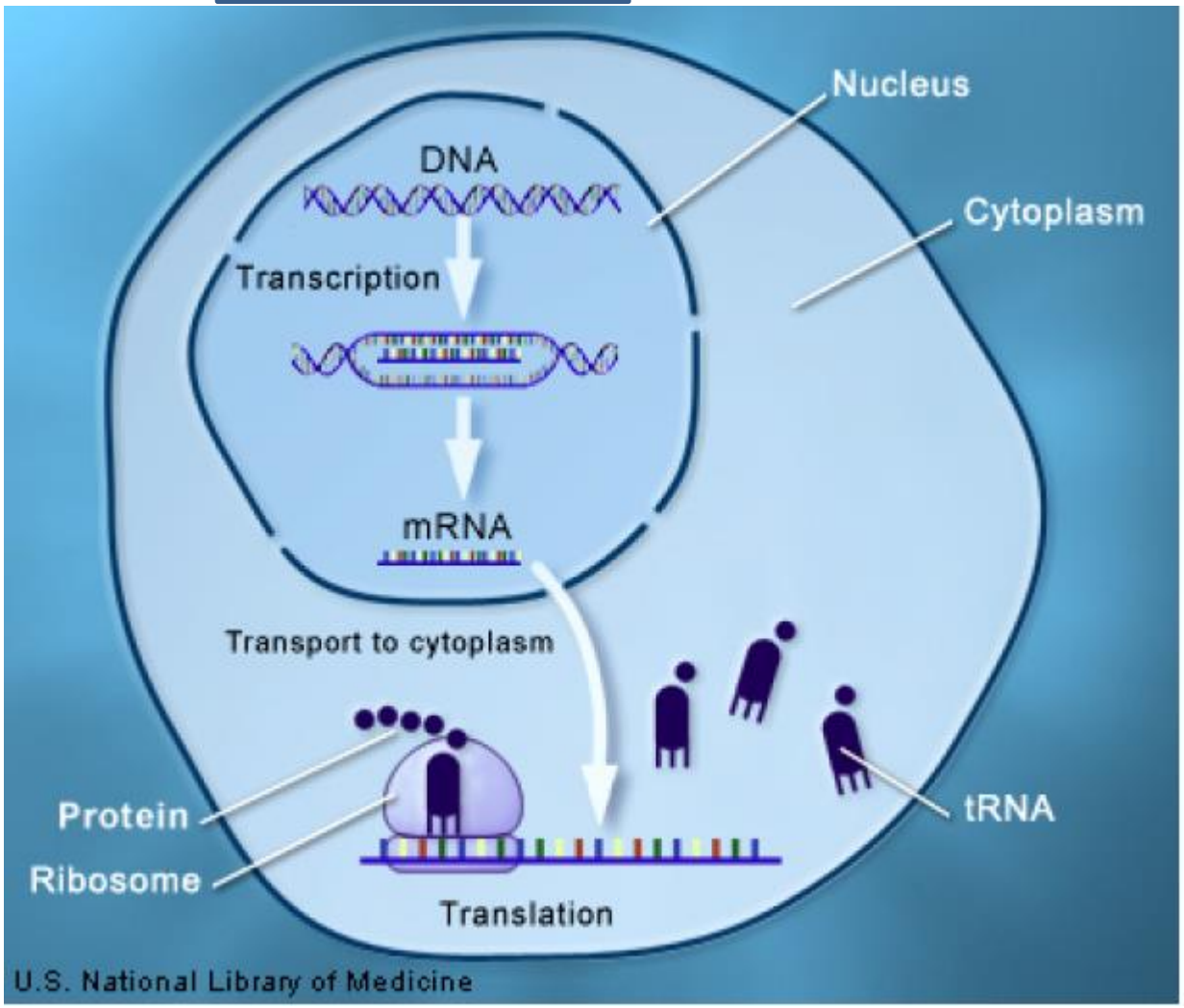


Figure from National Institutes of Health

Central Dogma



Cellular Phenotype



- pre-mRNA
- 7-methylguanosime placed at 5' end (prevent RNA degradation)
- Poly A tail is added at the 3'end (200 bps)
- splicing
- Final product, mRNA

Theory published by **Crick** 1958 (Yes, the same Crick that worked with Watson)

Splicing

- Archaea and Bacteria
  - Usually have one chromosomes
  - Chromosomes are circular
    - some can be linear
- Eukaryotes
  - Multiple chromosomes
  - Linear
  - Packed into cell nucleus

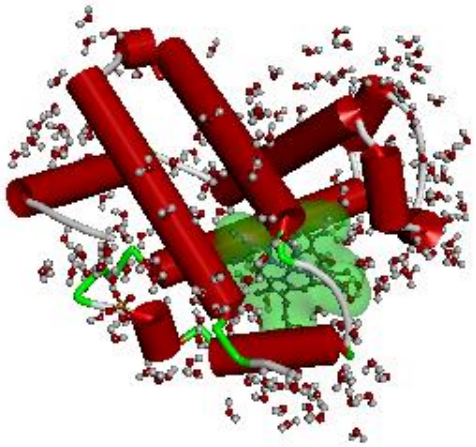
Chromosome (highlight: Gene)

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence						.....gcaggagccagggctgggcataaaagtcagggcagagccatctattgctt
1	ENSE00001829867	5,227,071	5,226,930	-	2	142	ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATC TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG TTGGTGGTGAGGCCCTGGGCAG
	Intron 1-2	5,226,929	5,226,800			130	gttggtatcaaggttacaagacagg.....tattggtctattttcccacccttag
2	ENSE00001057381	5,226,799	5,226,577	2	0	223	GCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCAC TCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGC CTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGA GCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGG
	Intron 2-3	5,226,576	5,225,727			850	gtgagtctatgggacgcttgatgtt.....catacctcttatcttcctcccacag
3	ENSE00001600613	5,225,726	5,225,464	0	-	263	CTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCA CCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAG TATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAG TCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATA AAAAACATTTATTTTCATTGCAA
	3' downstream sequence						tgatgtatttaaattatttctgaatattttactaaaaagggaatgtggga.....

Gene sequence

mRNA (cDNA) and protein sequences

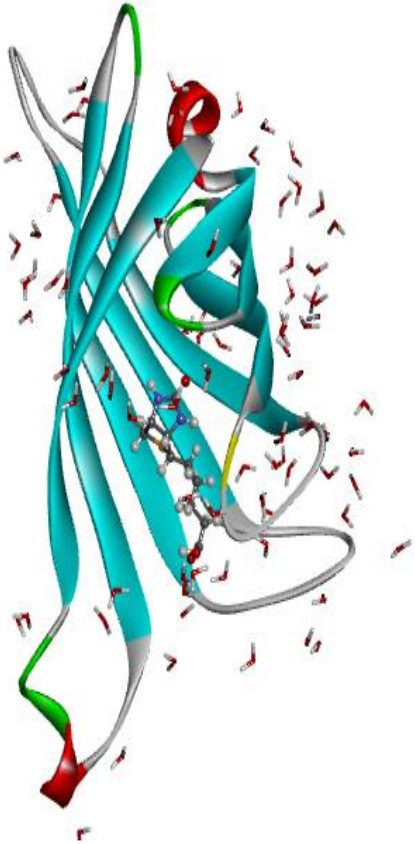
```
  1 ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATC   60
    ..................................................ATGGTGCATC   10
    ..................................................-M--V--H--    3

 61 TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG  120
 11 TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG   70
  4 L--T--P--E--E--K--S--A--V--T--A--L--W--G--K--V--N--V--D--E--   23
```

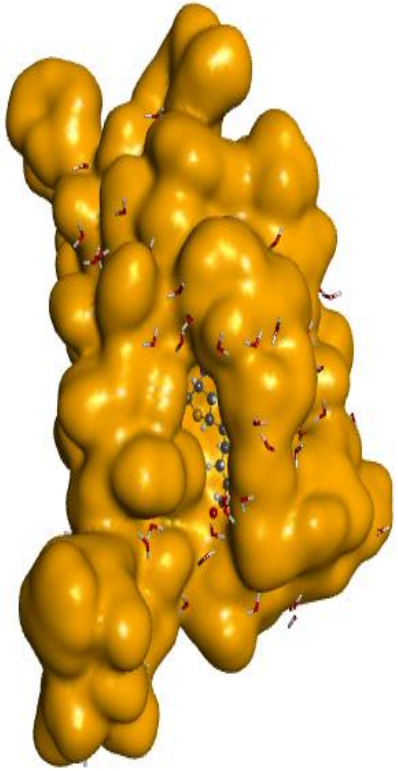


Proteins in 3D

Water molecules

Drug binding inhibition





Vitamin-H bound protein (183 aa protein)

Proteins vs DNA

Proteins

- Unstable
  - Seconds to months
    - Depends on protein & organism
    - <> life span of human proteins: 1 day
  - Destroyed after some time and recycled

DNA

- Stable
- DNA can be stable even for 100,000 years!!!

Related Dogmas

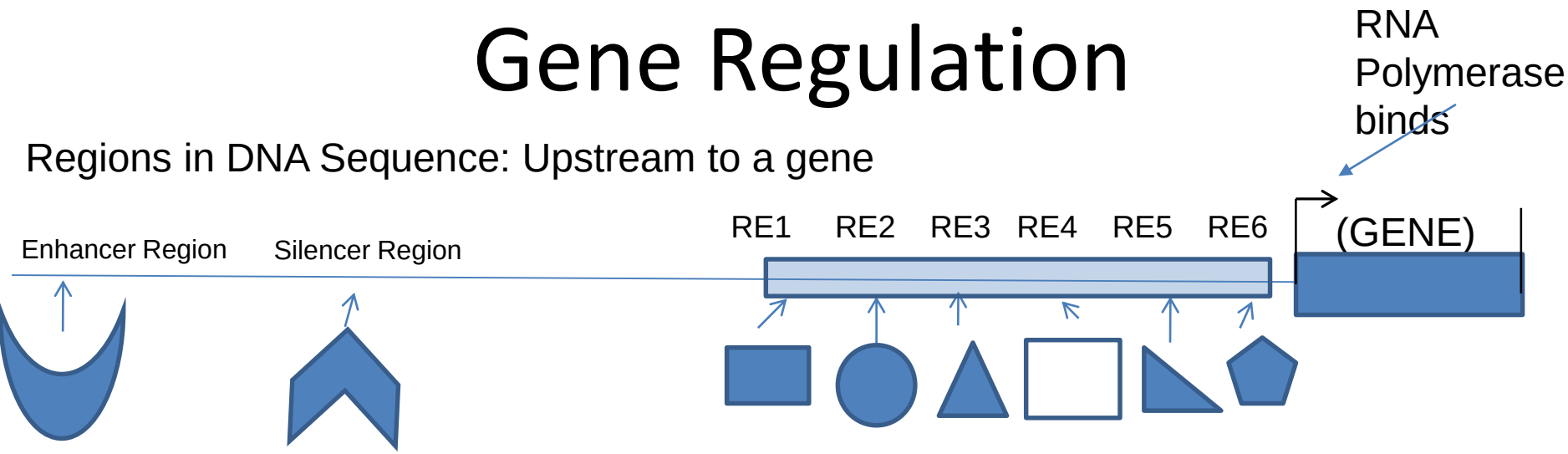
- Central Dogma of Genomics
  - Genome → Transcriptome → Proteome → Cellular Phenotype

- Study of organisms that inhabit the human body
  - microbiome

What differentiates different cells?

Gene Regulation

Regions in DNA Sequence: Upstream to a gene



Activator Activates gene (Oct-1)

Repressor- bind and block RNA polymerase binding/tran scription

Transcription Factors (TF) bind to Regulatory Element (RE)

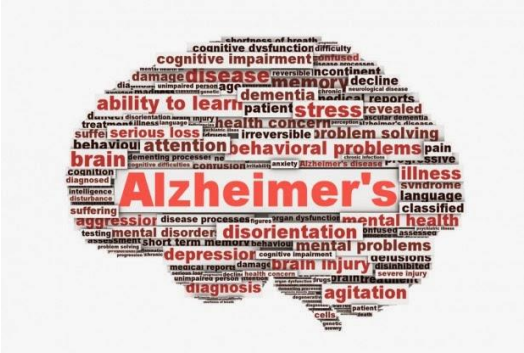
PROMOTER REGION

- Promoters, Transcriptor factor binding regions are identified by experiments.
  - Collecting samples and sequence alignments etc.

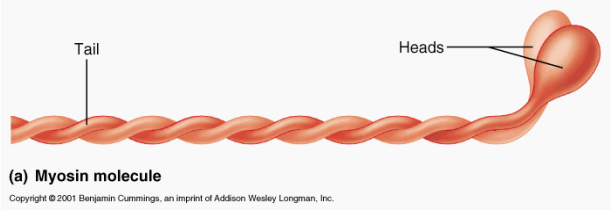
Same content (~25K Genes), different function

Active
NonActive

- **Brain Cells**
  - Amyloid, myosin, α-amylase

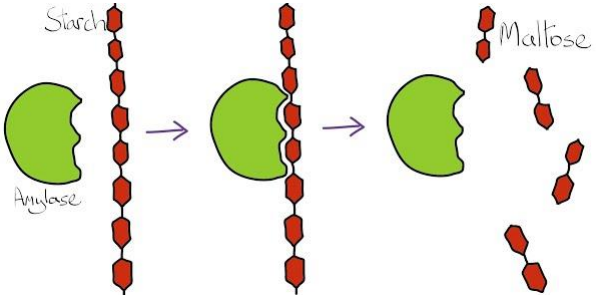


- **Muscle Cells**
  - Amyloid , myosin, α-amylase



- **Salivary Gland Cells**
  - Amyloid, myosin, α-amylase

Starch Sugars



Data based on NCBI Mol Biology Review Class
Gene Expression

Approximations in Bioinformatics

- 3 → 1 Letter (don't forget the chemistry)
  - Protein (20 alphabet)
    - Glu-Leu-Val-Ile-Ser-Thr-His-Glu-Lys-Ile-Gln-Gly
    - ELVISTHEKING
  - DNA/RNA (4 letter alphabet)

- Redundant information of DNA
  - storage

ATGGAGCTGTCTTG
TACCTCGACAGAAG

What is the key issue in Bioinformatics?

Release	Date	Bases (GB)	Seq (GB)	Bases(WGS)	Seq (WGS)
235	12/2019	388,417,258,009	215,333,020	6,277,551,200,690	1,127,023,870

Data Growth in numbers

Enormous data

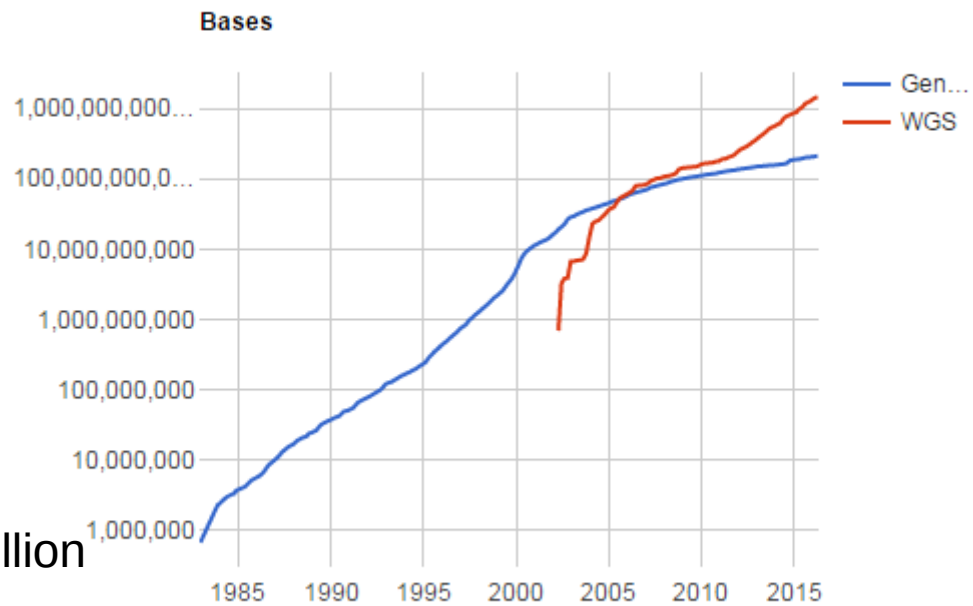
Need for algorithms to model and carry out analysis

1 Billion

Need for Computers

Storage, Retrieval and Analysis

1 Million



Whole Genome Shotgun (WGS)
High-throughput sequencing (Illumina etc.)

Image from NCBI GenBank

Storage/Retrieval

What can we do with the information?
Can we compare/align them?
What do we learn from the alignments?

Comparing 3D is one way
Not all 3D information is available

So, sequence comparison is the common approach

How to find out whether two proteins are related?

Sequence Alignment

DNA can be compared instead of protein. But, for **most cases** proteins have more information than DNA

- Relatedness (homology) among proteins/DNAs
  - Common function?
  - Homology (common ancestor)
    - When two sequences (proteins/genes) are highly similar, they might be homologous
    - Converse is not true (lack of similarity != No Homology)
  - What is homology?

How can I compare two sequences?

- Not possible without the help of Math and Statistics

- Luckily for us the problem is addressed and some framework is available for us to use

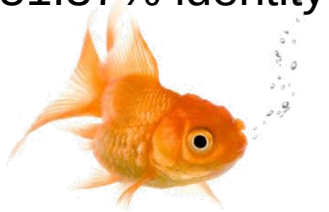
Temple F. Smith

Michael S. Waterman

82.88% identity

```
Human          MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
RT-Lemur       -TFLTPEENGHVTSLWGKVNVEKVGGEALGRLLVVYPWTQRFFESFGDLSSPDAIMGNPK
                ..*****:. **:*******::****************************:***:*****

Human          VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
RT-Lemur       VKAHGKKVLSAFSEGLHHLDNLKGTFAQLSELHCVALHVDPENFKLLGNVLVIVLAHHFG
               *********.***:** ********** ******  ********:******* *******

Human          KEFTPPVQAAYQKVVAGVANALAHKYH
RT-Lemur       NDFSPQTQAAFQKVVTGVANALAHKYH
               ::*:* .***:****:***********
```

51.37% identity



```
Human          MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Goldfish       -VEWTDAERSAIIGLWGKLNPDELGPQALARCLIVYPWTQRYFATFGNLSSPAAIMGNPK
                *. *  *:**: .****:* **:* :**.* *:*******:* :**:**:* *:*****

Human          VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Goldfish       VAAHGRTVMGGLERAIKNMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFG
               * ***:.*:*.:. .: ::**:*.*:* ** :*.:******:*****.: :.   * :**

Human          -KEFTPPVQAAYQKVVAGVANALAHKYH
Goldfish       PSGFNADVQEAWQKFLSVVVSALCRQYH
                . *.  ** *:**.:: *..**.::**
```

21.74% identity

```
Human        MVHLTPEEKSAVTALWGKVNVDEV----GGEALGRLLVVYPWTQRFFESFGDLSTPDAVM
Bloodworm    -MGLSAAQRQVVASTWKDIAGSDNGAGVGKECFTKFLSAHHD---IAAVFGFSG-----A
              : *:  ::..*:: * .:  .:     * *.: ::* .:     :   **  .

Human        GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPENFRLLGNVL
Bloodworm    SDPGVADLGAKVLAQIGVAVSHLGDEGKMVAEMKAVGVRHKGYGYKHIKAEYFEPLGASL
             .:* *   * ***. :. .::**.:    .* :.     .   .  *:. * *. **  *

Human        VCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-
Bloodworm    LSAMEHRIGGKMTAAAKDAWAAAYADISGALISGLQS
             :..: *::* ::*  .: *:  . *.::.**    :
```

19.85% Identity

```
Human       MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
Soybean     MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA----NGVDPTN
            ** :* ::.: *:: :  .:  : : .      :*   * :: :*. :.     ..   *

Human       PKVKAHGKKVLGAFSDGLAHLDNLKG--TFATLSELHCDKLHVDPENFRLLGNVLVCVLA
Soybean     PKLTGHAEKLFALVRDSAGQLKASGTVVADAALGSVHAQKAVTDPQ-FVVVKEALLKTIK
            **:..*.:*::. . *. .:*.      : *:*..:*.:*  .**: * :: :.*: .:

Human       HHFGKEFT----PPVQAAYQKVVAGVANALAHKYH
Soybean     AAVGDKWSDELSRAWEVAYDELAAAIKKA------
           .*.:::       :.**:::.*.: :*
```

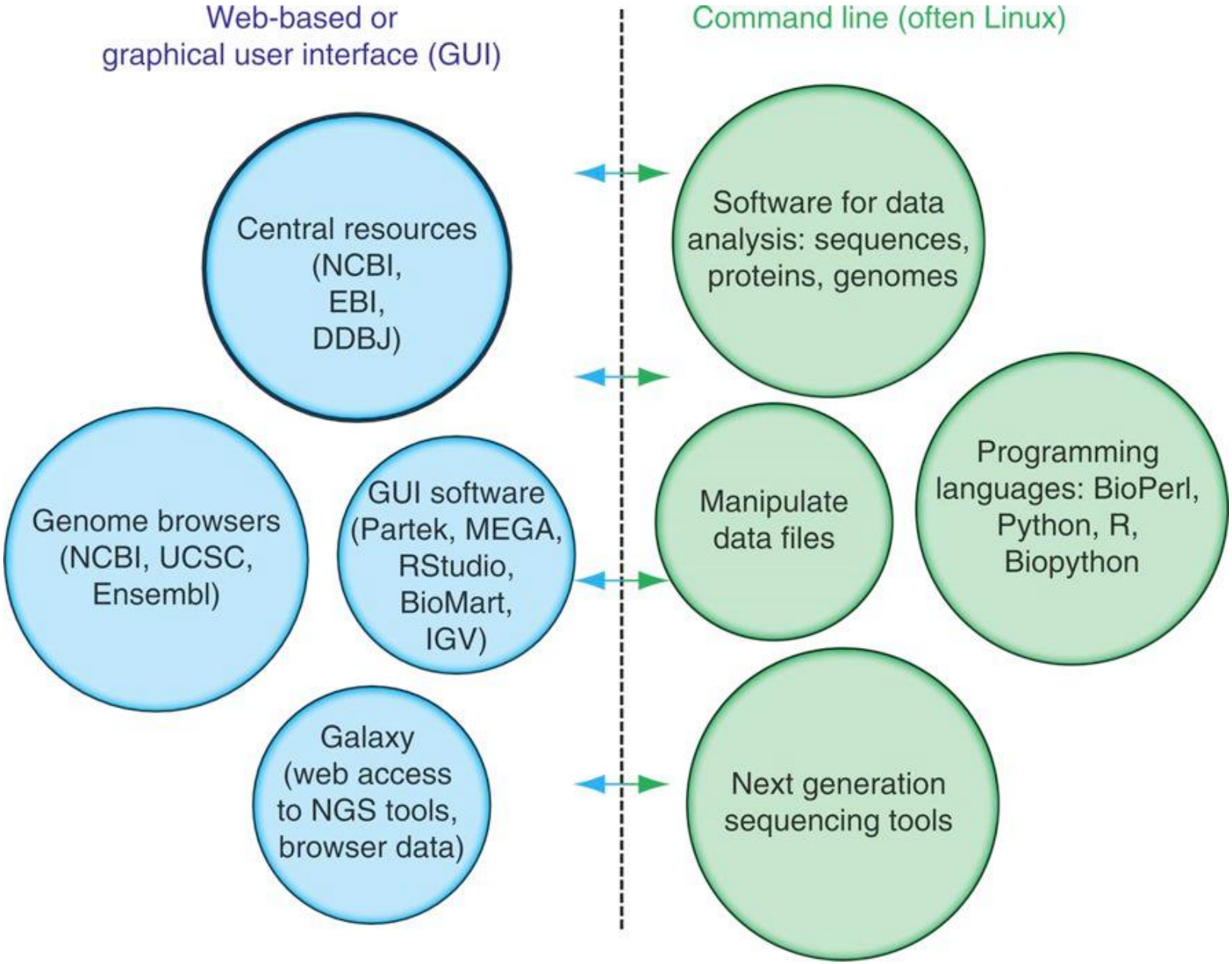
Questions

- What sequences to use and why?
- What types of alignments
  - Global and Local
- Statistics of alignments
  - Scores and matrices

Two Cultures in Bioinformatics

- Two cultures
  - Web-based
    - Point-and-Click (no programming effort)
  - Command line
    - Sometimes steep learning curve (some programming)
- Which one is better?

Fig 1.5 from Bioinformatics and functional genomics / Jonathan Pevsner.—Third edition. Copyright Figure. Please do not distribute.



Validity of Predictions

- Can we use a software as a black-box?
- How do we know whether a software method is working properly?
- Each software team will in most cases do self evaluation
- Sensitivity(TPR) and Specificity(TNR)
  - Sensitivity: Detecting true cases
  - Specificity: Excluding those without disease

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Evaluations

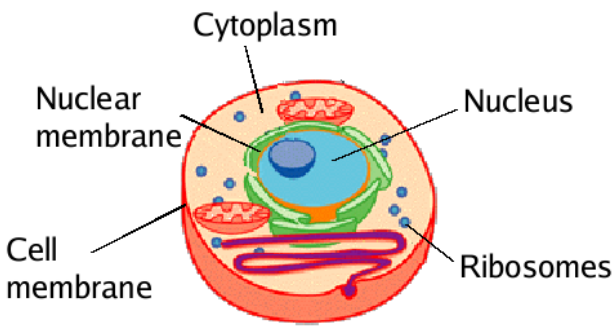
Name	Competition
Alignathon	Compare whole-genome sequence alignment methods
EGASP	ENCODE Genome Annotation Assessment Project
Assemblathon	Compare the performance of genome assemblers
GAGE	Genome Assembly Gold-standard Evaluations
ANRF	Assn. of Biomolecular Resource Facilities (ABRF) assessment of phosphorylation
CASP	Critical Assessment of Structure Prediction
CAFA	Critical Assessment of protein Function Annotation algorithms
CAGI	Critical Assessment of Genome Interpretation. Assess computational methods for predicting phenotypic impacts of genomic variation

New Paradigms for Learning Bioinformatics

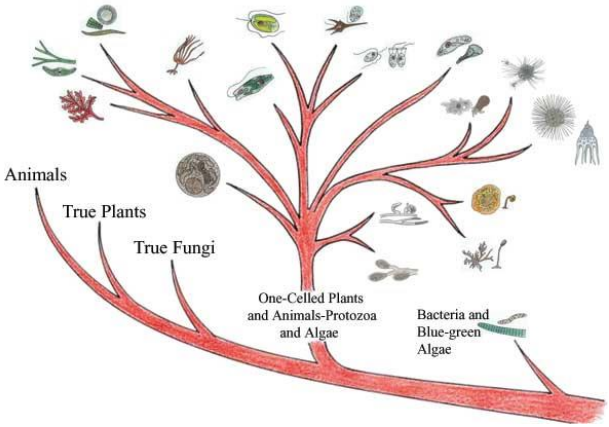
- Online resources
  - Google
- Online Classes
  - MOOC
  - EdX, Stanford, Coursersa, Udacity etc.
- Programming
  - R, Python, Perl
  - Linux Shell scripting

Two perspectives in Bioinformatics

- Cell Perspective
  - Contents of the cell
    - DNA/RNA/Protein
  - Analysis of the sequences



- Organism Perspective
  - How genes are expressed?
    - Age
    - Different tissues/cells
    - Race
    - Disease vs non-disease states



Command-line interface

- Command line or point-click or application focused
  - What OS?
    - Windows or Linux
  - Why?
    - Cost, ability to carry out tasks

Reproducibility

"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments"

Is there a reproducibility crisis? M.Baker, Nature, 533, 452, 2016

Reproduce another scientist's experiments (failed to reproduce their own experiment)
Chemistry: 90% (60%)
Biology: 80% (60%)
Physics & Engineering: 70% (50%)
Medicine: 70% (60%)
Earth and Env. Science: 60% (40%)

Reproducibility in Published Papers

- Script availability
  - Supplemental pages is a good place
  - Useful for checking the results
  - Useful for learning/teaching
  - Useful for reviewers
  - Etc.

Based on ideas from Hadley Wickham

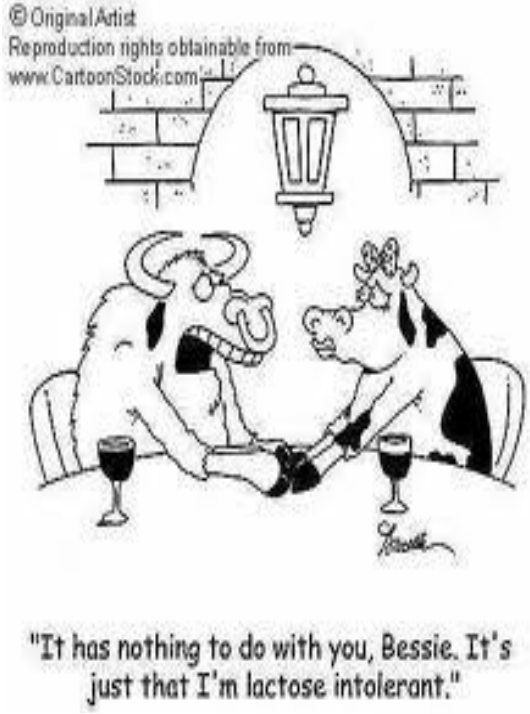
Reproducibility using R

- R session
  - sessionInfo()
- What packages was used
  - library(??)
- Show the code
  - Use R command “dput” to make the user copy and use your code
- Show comments

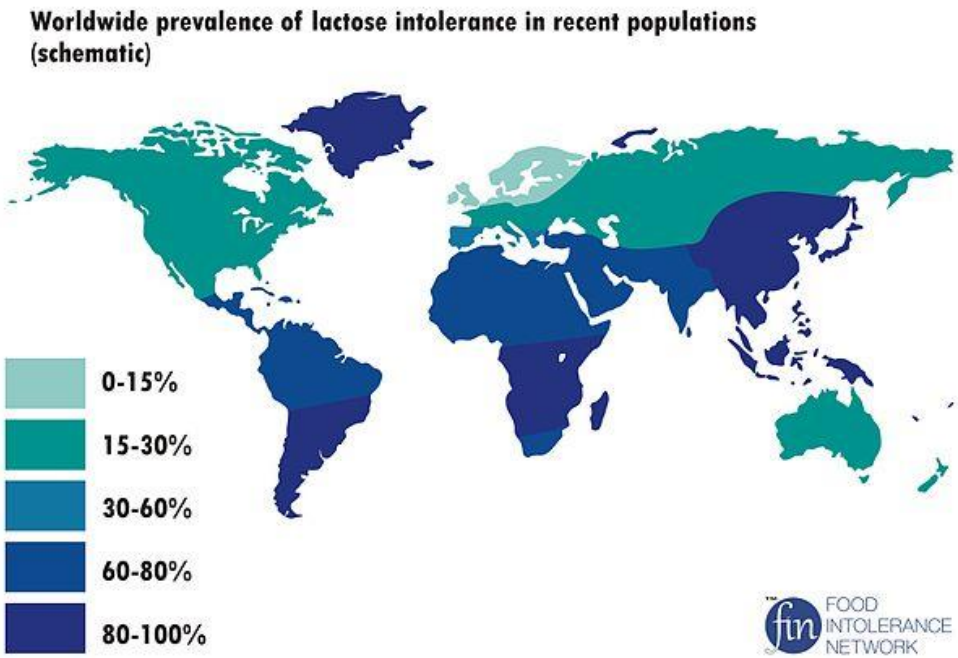
```
# Reproducible Code ; S. Ravichandran, Ph.D. 01/23/2017

# load libraries at the top of the script
library(rafalib)

# if not installed
# install.packages("rafalib")
# set seed for reproducibility
set.seed(100)

# create 100 uniform normally distributed random numbers
x <- rnorm(100)
y <- rnorm(100)

# use mypar from rafalib to plot 2 figs in 1 row
mypar(1,2)
hist(x,col="red"); hist(y,col="blue") # use two lines; for lack of space used 1 line
ttest <- t.test(x,y, alternative = "two.sided", conf.level = 0.95) # p-value = 0.94

# add an outlier in y and call it my
m_y <- c(y, 150)
# use dput(data) if you want to send data via email; ex. dput(m_y)

#use two lines; for lack of space used 1 line
hist(x,col="red"); hist(m_y,col="blue")
ottest <- t.test(x, m_y, alternative = "two.sided", conf.level = 0.95) # p-value = 0.94
ottest
sessionInfo() # provide sessionInfo()
```

Application-1
Sequence-based approach

Lactose Intolerance



"It has nothing to do with you, Bessie. It's just that I am Lactose intolerant"



Finland: 1/60K inborns have LCT intolerance

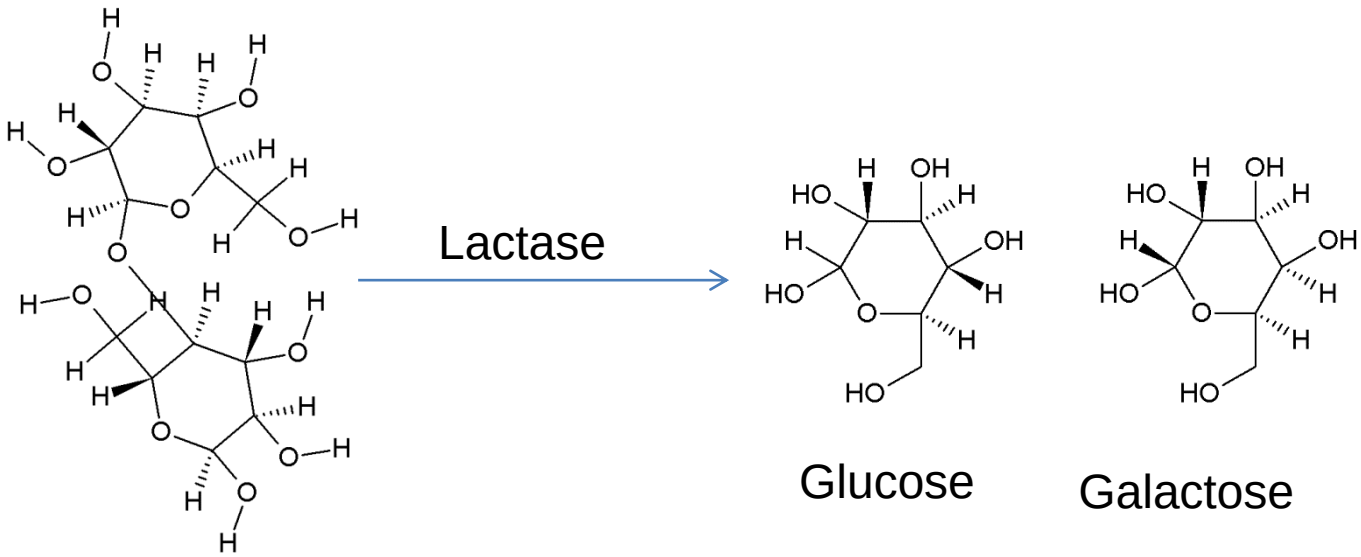
Very common in people of
-West African,
-Arabs,
-Jewish,
-Greek and
-Italian descent.

(ghr.nlm.nih.gov)

Lactose intolerance

- Lactase is the gene that produces Lactase (protein enzyme)

- Lactase digests Lactose → simple sugars



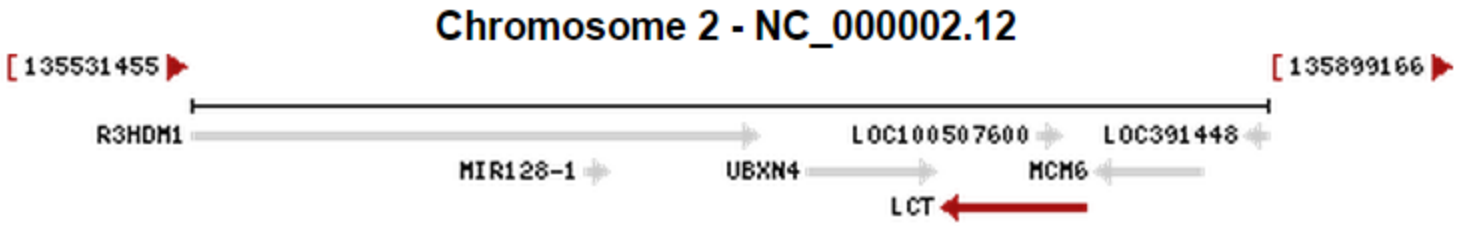
LCT Gene

- Intolerant
  - Gene turned off
- Tolerant (adult)
  - Persistence
- Remedy
  - Lactase pill
  - Which does not interfere with transcription but just provide a supply of Lactase enzyme
  - Need to be taken before the Lactose food

LCT Gene

- Lactase is active during childhood but slows or stops when child grows up for some people
- LCT (short name for the Lactase gene)

  – Chr 2; 17 exons



Note the gene direction?

Hum Genet (2011) 130:483–493
DOI 10.1007/s00439-011-0966-0

The −14010*C variant associated with lactase persistence is located between an Oct-1 and HNF1α binding site and increases lactase promoter activity

Tine G. K. Jensen · Anke Liebert · Rikke Lewinsky · Dallas M. Swallow · Jørgen Olsen · Jesper T. Troelsen

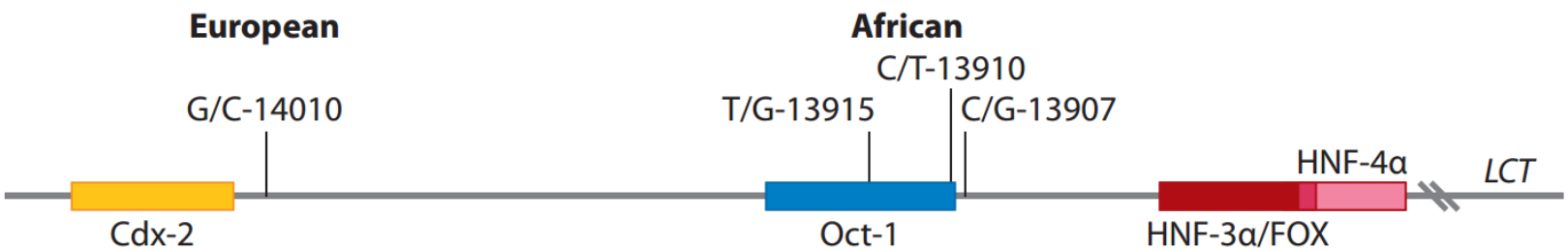


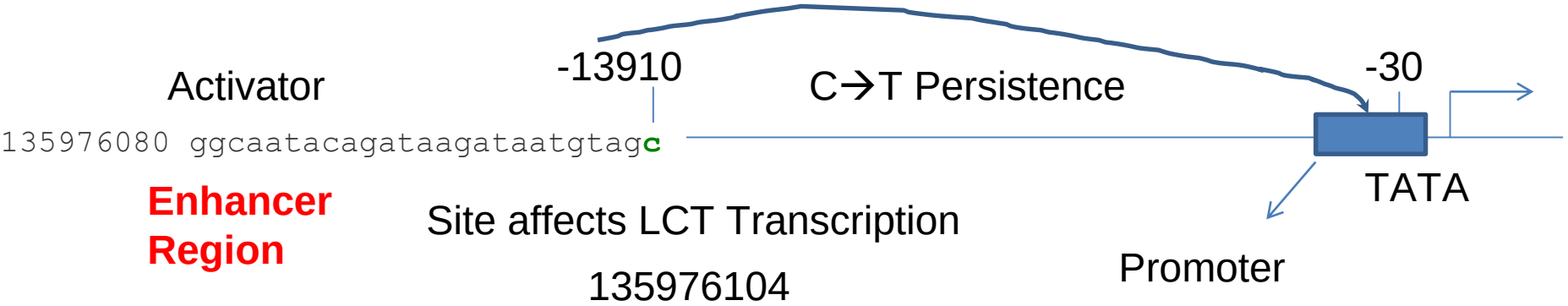
Figure 3

Locations of transcription factor-binding sites and predicted adaptive alleles upstream of *LCT*, the lactase gene. Three alleles were identified as potentially causal alleles in the African pastoral populations, whereas C/T-13910 was predicted to be the causal allele in Northern Europeans. Additionally, the T/G-13915 allele is correlated with lactase persistence in the Saudi Arabian population. The transcription factors and the sequence they bind in a supershift assay (48) are: HNF-4α (−13854 to −13830), HNF-3α and FOX (−13872 to −13848), Oct-1 and GAGA (−13933 to −13909), and Cdx-2 (−14040 to −14016).

Kelley and Swanson, Human Genet, 2008:9:143-160

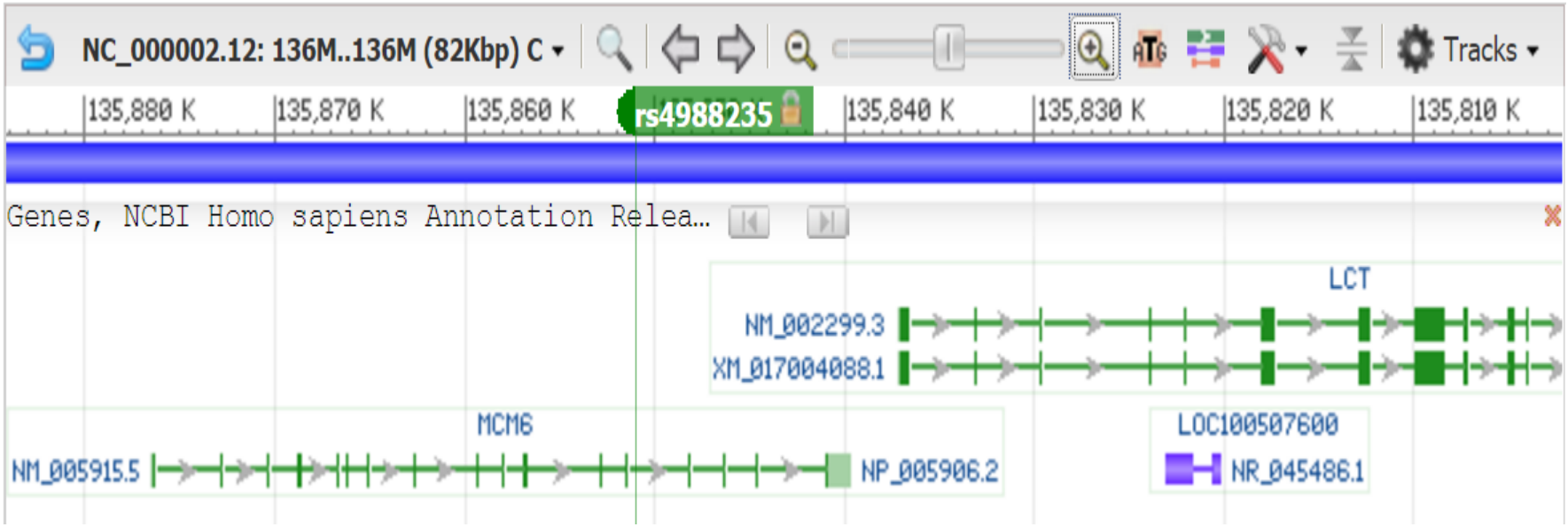
Possible Mechanism

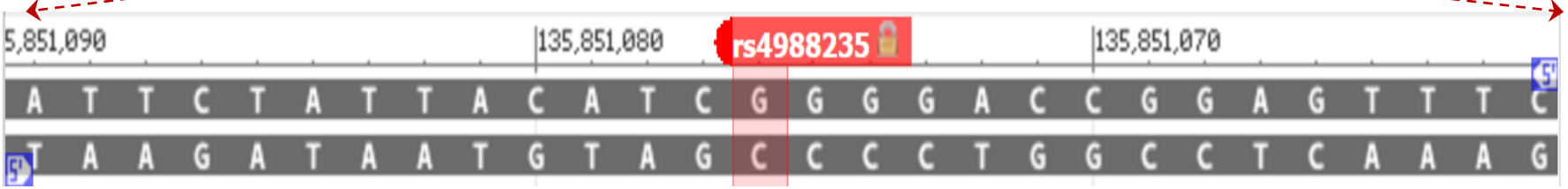
<u>Protein (Activator) that</u> bind in Enhancer regions far away from Promoter (non-binding) and bends and interacts with Promoter (TATA region) to positively affect transcription



http://www.fda.gov/forconsumers/consumerupdates/ucm094550.htm

Where is the variation (or lack of) that causes Lactose Persistence (Lactose intolerance)?





Application-2
Sickle Cell Disease (SCD): Structure-based approach

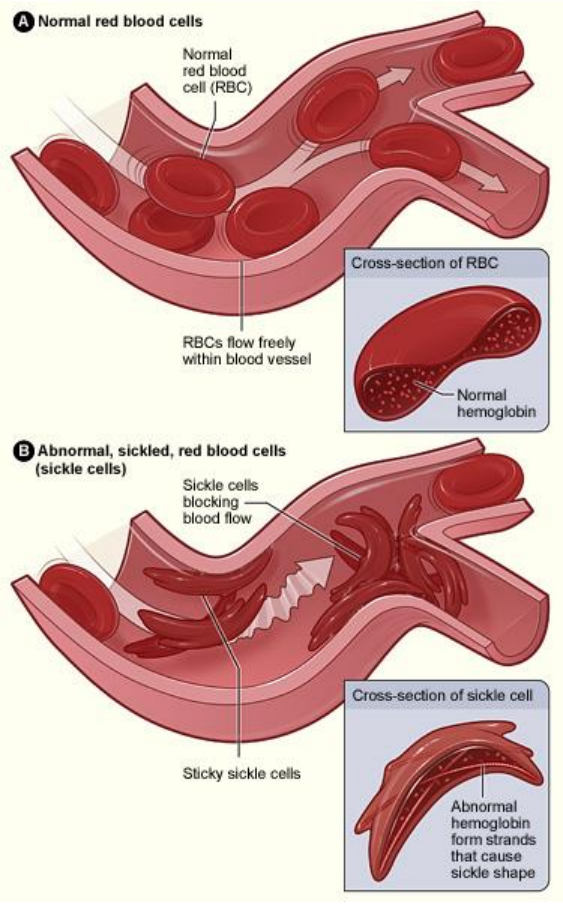


Figure taken from NCBI/NIH

Data Source CDC

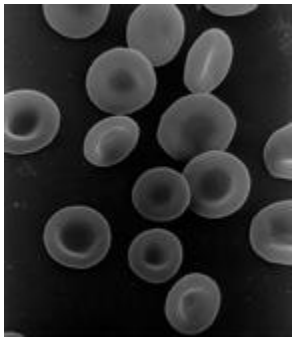
Malaria (2015)

214M (World-wide)
~1500 cases in US every year

Sickle-Cell Disease (SCD)

affects ~ 100,000 in US
occurs 1/365 black or African-American births
Occurs 1/16,300 Hispanic-American births

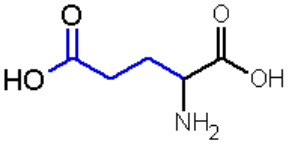
Biology of RBC



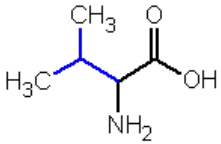
S.E.M of RBC Wikipedia

- RBC (a.k.a. erythrocytes, haematids etc.)
- Nucleus lacking in human
- 2.4 M raw RBC are produced/second
- Produced in bone marrow and travel all over body carrying O_2 (& CO_2)
- Carrier protein complex: Hemoglobin
  - Not a single gene product ; 2 α and 2 β chains
  - HBA: α ; HBB:β

https://en.wikipedia.org/wiki/File:Redbloodcells.jpg

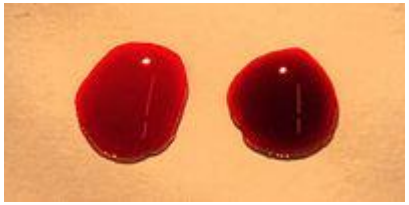


Glutamic Acid **(Glu)**



Valine **(Val)**

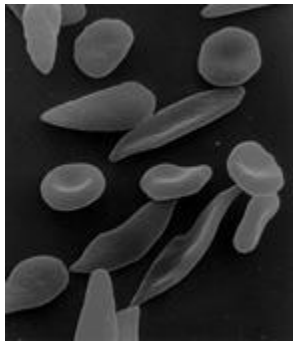
SCD

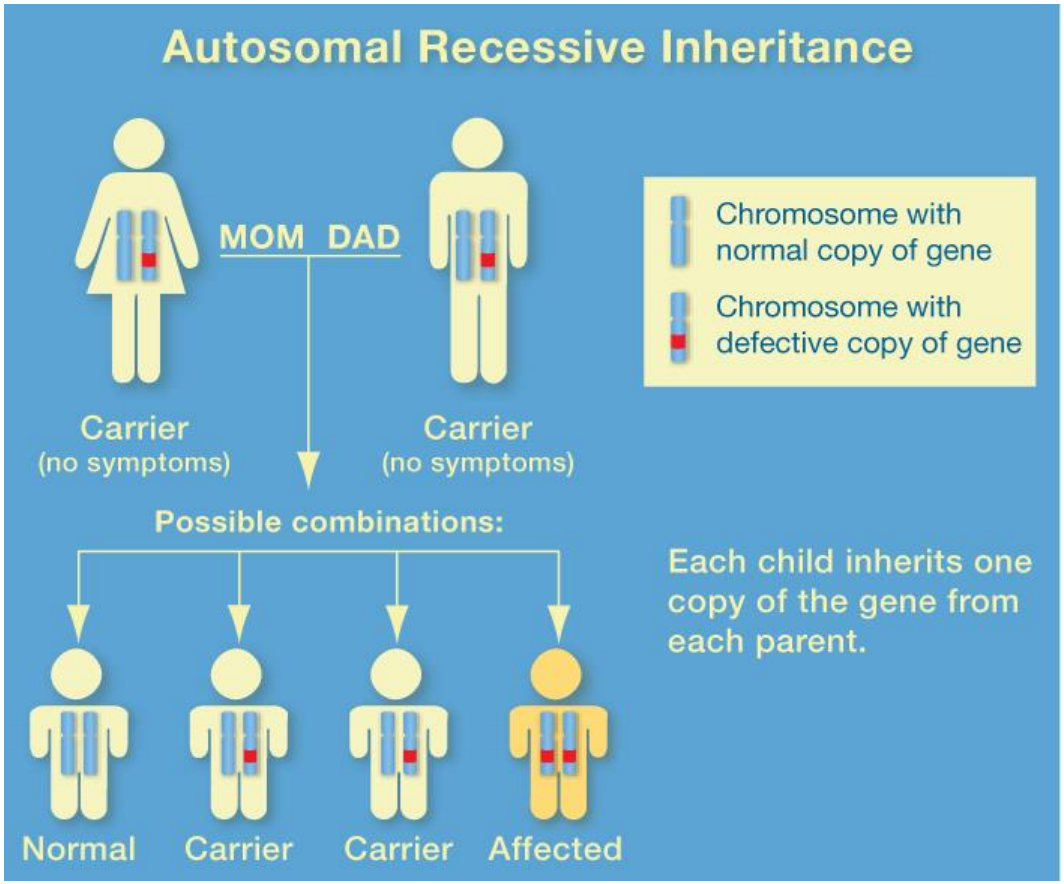


2 drops of oxygenated/deoxygenated blood

https://en.wikipedia.org/wiki/Red_blood_cell

- Genetic disorder
- HBB:Glu7→Val
- Cell sickle & dies sooner

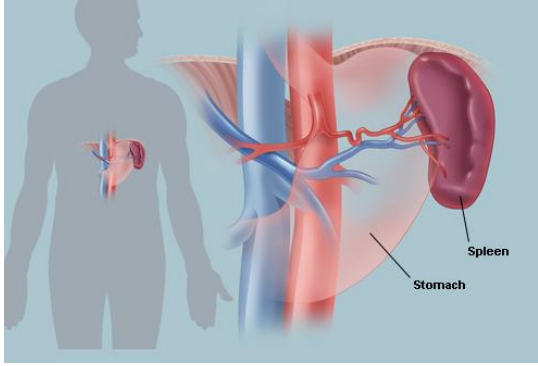




Picture taken from

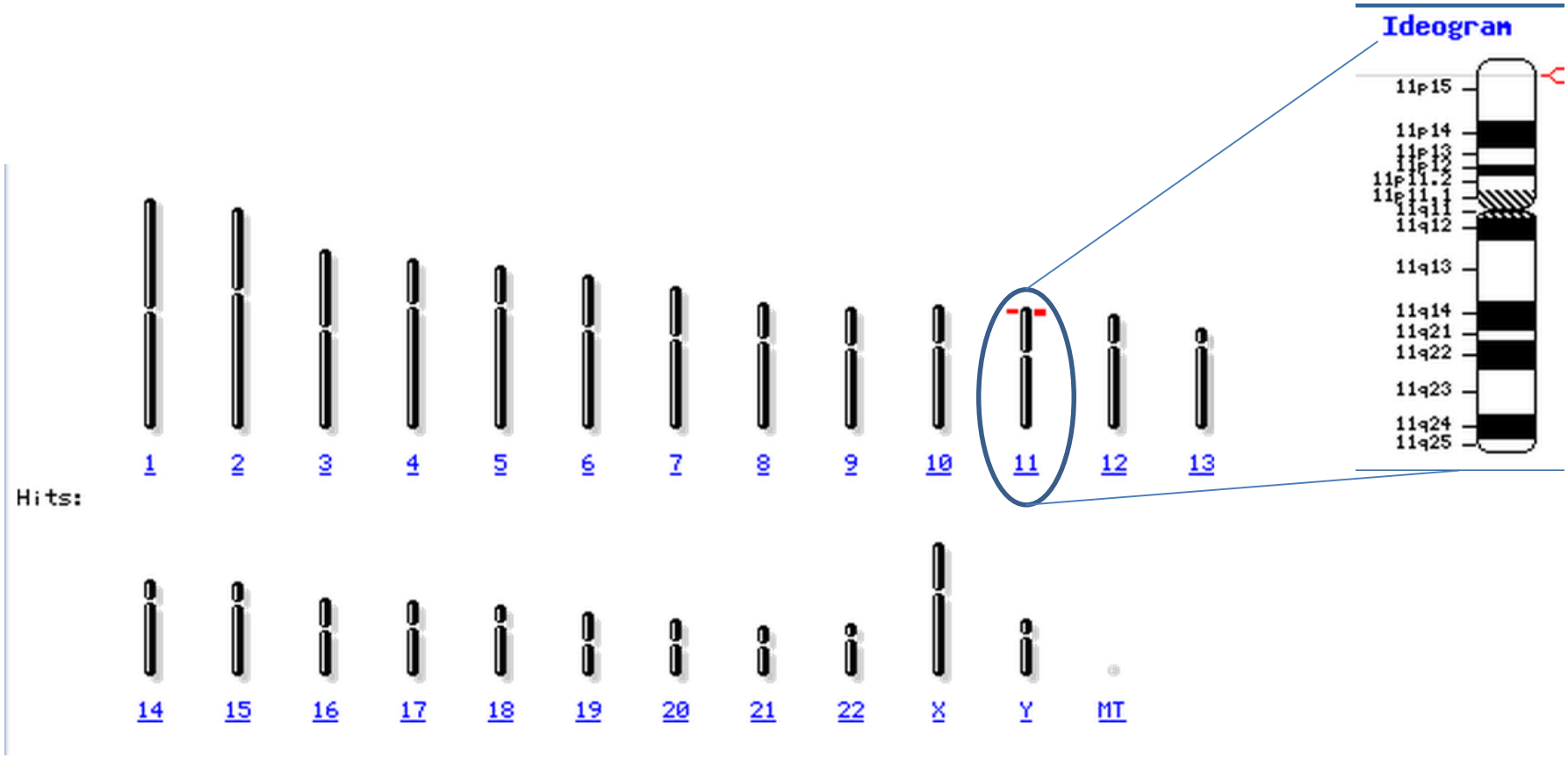
http://learn.genetics.utah.edu/content/disorders/singlegene/sicklecell/

Related Disease



- SCD → Anemia
- Spleen the blood filter will be clogged by the sickle cells and damages → infections
- **Malaria** is caused by mosquito bites. Parasite invades the RBC and **destroys** it.
- In US disease most commonly affects African American

HBB: Where is it?



NCBI Database

Hemoglobin alpha/beta sequences

```
        10         20         30         40         50
MVLSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHFDLS
        60         70         80         90        100
HGSAQVKGHG KKVADALTNA VAHVDDMPNA LSALSDLHAH KLRVDPVNFK
       110        120        130        140
LLSHCLLVTL AAHLPAEFTP AVHASLDKFL ASVSTVLTSK YR
```

Alpha

E→V → Sickle Cell Anemia

```
        10         20         30         40         50         60
MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK
        70         80         90        100        110        120
VKAHGKKVLG AFSDGLAHLD NLKGTFATLS ELHCDKLHVD PENFRLLGNV LVCVLAHHFG
       130        140
KEFTPPVQAA YQKVVAGVAN ALAHKYH
```

Beta

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


sp|P68871|HBB_HUMAN      VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
sp|P69905|HBA_HUMAN      -----KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLP
                              *** .*::::.:**:*:: .::::**:**. **.*** **.**.: *: .** *:

sp|P68871|HBB_HUMAN      KEFTPPVQAAYQKVVAGVANALAHKYH
sp|P69905|HBA_HUMAN      AEFTPAVHASLDKFLASVSTVLTSKYR
                          ****.*:*: :*.:*.*:..*: **.
```

43.9% identity

Summary

- What is Bioinformatics?
- Basics of Bioinformatics
- Cell biology
  - DNA/Proteins/RNA
- Central Dogma
- Expression → Function
- Homology, sequence alignment (1D → 3D)
- Applications

FINAL PROJECT

- FIND A GENE PROJECT

Thanks

S. Ravichandran, Ph.D
ravichandran@hood.edu