CMSC 423:
Bioinformatics Algorithms, Databases & Tools

Spring 2020

UNIVERSITY OF MARYLAND

Course Info

Instructor: Rob Patro (rob@cs.umd.edu)

Office: 3220 IRB

Office Hours: Tues. 1-2PM or by appointment

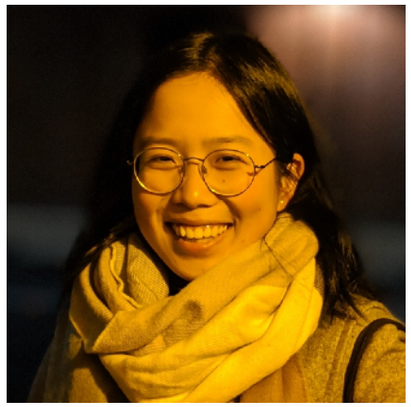
Website: https://rob-p.github.io/CMSC423_S20/

Course Info

TAs:

Yuelin Liu

email: yuelin@cs.umd.edu



Hadi Yami

email: hadihami@cs.umd.edu

TA office hours to be finalized this week. Will be posted on the course website.

Course Info

ADS: https://www.counseling.umd.edu/ads/

Academic Integrity: https://academiccatalog.umd.edu/undergraduate/registration-academic-requirements-regulations/academic-integrity-student-conduct-codes/

Piazza Page: https://piazza.com/umd/spring2020/cmsc423

If you have a class-related e-mail: Please **prefix the subject with [CMSC423]**, so that my filter will pick it up and it won't be accidentally routed to SPAM.

Coursework & Grading

Coursework and grading: The coursework will consist of small programming assignments, a couple of larger projects, a midterm exam and a final exam. The breakdown of weights for these different assignments will be as follows:

- Homeworks & short prgramming assignments — 15%
- Projects - 25%
- Midterm 1 - 15%
- Midterm 2 - 15%
- Final Exam — 30%

Late policy: Assignments that are turned in late will be docked 1% for each hour they are late up to the first 48 hours. After 48 hours, late assignments will not be accepted.

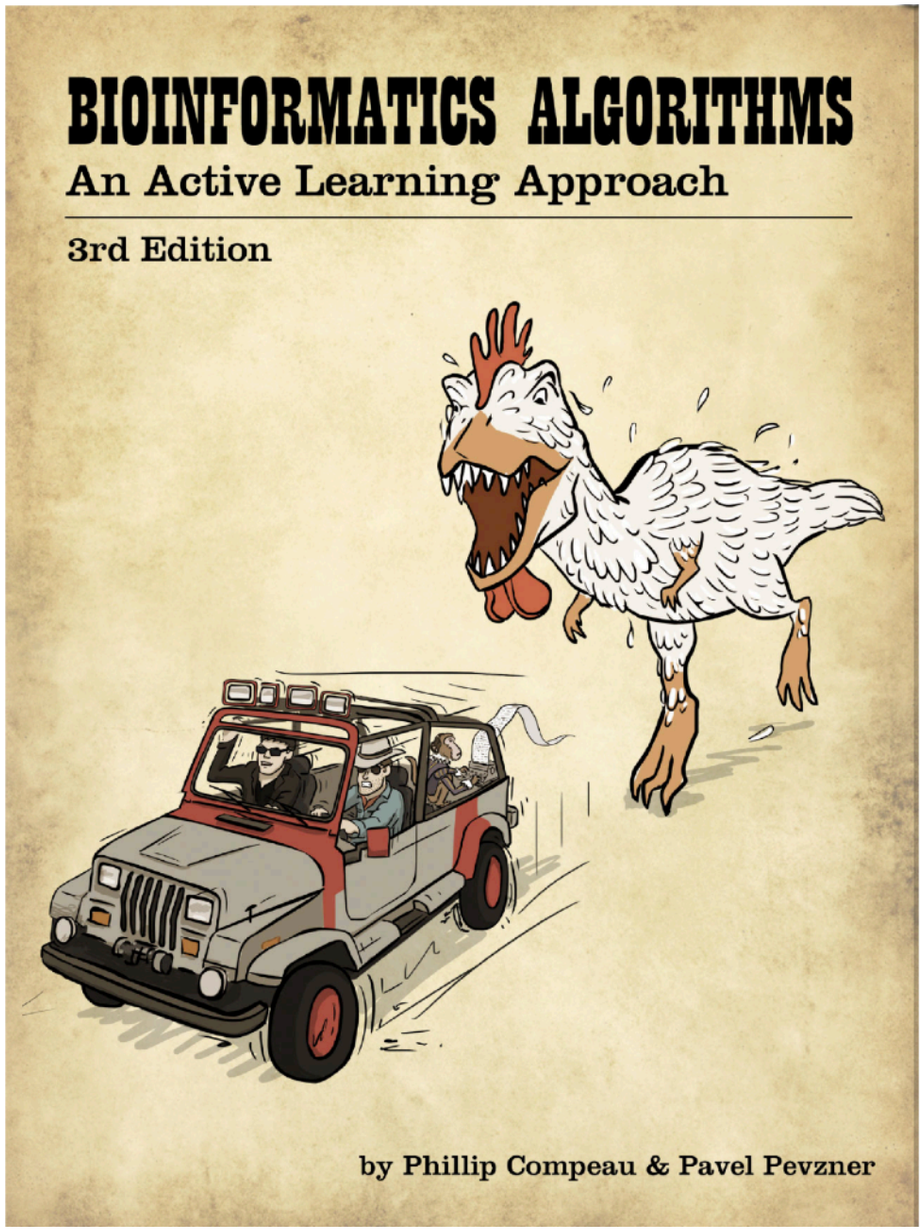
Regrade policy: All requests to re-grade, re-check, or re-mark an assignment or exam question **must be made in writing**. When the assignment is re-graded, it will be re-checked in its entirety. This means that *it is possible to lose points on other problems if they were graded incorrectly or too leniently the first time*. Therefore, I urge you to thoroughly consider each regrade request you make.

Academic Integrity maintain it!

TLDR : Don't cheat. Don't copy code from friends, classmates, or the internet for the short programming assignments or the projects. Don't provide code to classmates for any of the assignments or projects. Don't cheat on the exams. Be cool, and everything will be cool.

Academic integrity is a very serious issue. Any assignment, project or exam you complete in this course is expected to be your own work. If you are allowed to discuss the details of or work together on an assignment, this will be made explicit. Otherwise, you are expected to complete the work yourself. *Plagarism is not just the outright copying of content*. If you paraphrase someone else's thoughts, words, or ideas and you don't cite your source, this constitues plagraism. It is always much better to turn in an incorrect or incomplete assignment representing your own efforts than to attempt to pass off the work of another as your own. **If you are academically dishonest in this course, you will recieve a grade of XF, and you will be reported to the university's Office of Student Conduct**.

Textbooks



This text is required. We won't cover everything in it, and we will cover some things not in it, but we will have assigned readings from it and I will refer to it as a resource.

Other Textbooks

Genomics algorithms, data structures, and statisitcal models:

- Genome Scale Algorithm Design (Mäkinen, Belazzougui, Cunial, Tomescu 2015)
- Biological Sequence Analysis (Durbin, Eddy, Krogh, Mitchinson 1998)

Basics of algorithms and data structures:

This course will assume familiarity with basic algorithms and data structures, though I will attempt to refresh everyone's memory on relevant concepts when we cover them. If you need a refresher on algorithmic basics, I recommend the following resources:

- Algorithms (Dasgupta, Papadimitriou, and Vazirani 2006)
- Algorithm Design (Kleinberg and Tardos 2006)
- Introduction to Algorithms, 3rd edition(Cormen, Leiserson, Rivest and Stein, 2009)

Molecular biology:

We will cover the basic required molecular Biology in the course. However if you're not familiar with basic molecular Biology, there are some useful resources worth reading:

- Molecular Biology of the Cell (Alberts, Johnson, Lewis, Raff, Roberts and Walter, 2002)
- Molecular Biology: Principles of Genome Function 2nd Edition (Craig, Green, Greider, Storz, Wolberger, Cohen-Fix, 2014)
- Molecular Biology (Clark and Pazdernik 2012)

Bioinformatics & Computational Biology

Algorithms & Data Structures
for working with
Biological data

Bioinformatics

Computational Biology

Understanding Biology
via
Algorithmic & Statistical Approaches

Bioinformatics & Computational Biology

We'll treat this as two sides of the same coin

&

try to ignore this distinction

Why Computational Biology?

Our capabilities for *high-throughput* measurement of Biological data has been transformative

1990 - 2000

Sequencing the first human genome took ~10 years and cost ~$2.7 **billion**

Today

Sequencing a genome costs ~$100 - 1,000* (depending on how you count)

~18 Tb per “run” at maximum capacity

Progression of sequencing capacity

$3,000,000,000	**2003** Human Genome Project	
$20,000,000	**2006** 1st individual genome	
$2,000,000	**2007** 1st NGS Genome	
$200,000	**2008** 1st 30x genome	
$10,000	**2010** 1st sub-10K genome	
$1,000	**2014** 1st $1,000 genome	
$100	**2017** 1st $100 genome	

https://blog.dnanexus.com/2017-04-24-dnanexus-celebrates-national-dna-day/

Tons of Data, but we need Knowledge

We'll discuss a bit about how sequencing works soon. But the hallmark *limitations* are:

- Short "reads" (75 — 250) characters when the texts we're interested in are 1,000s to 1,000,000,000s of characters long.

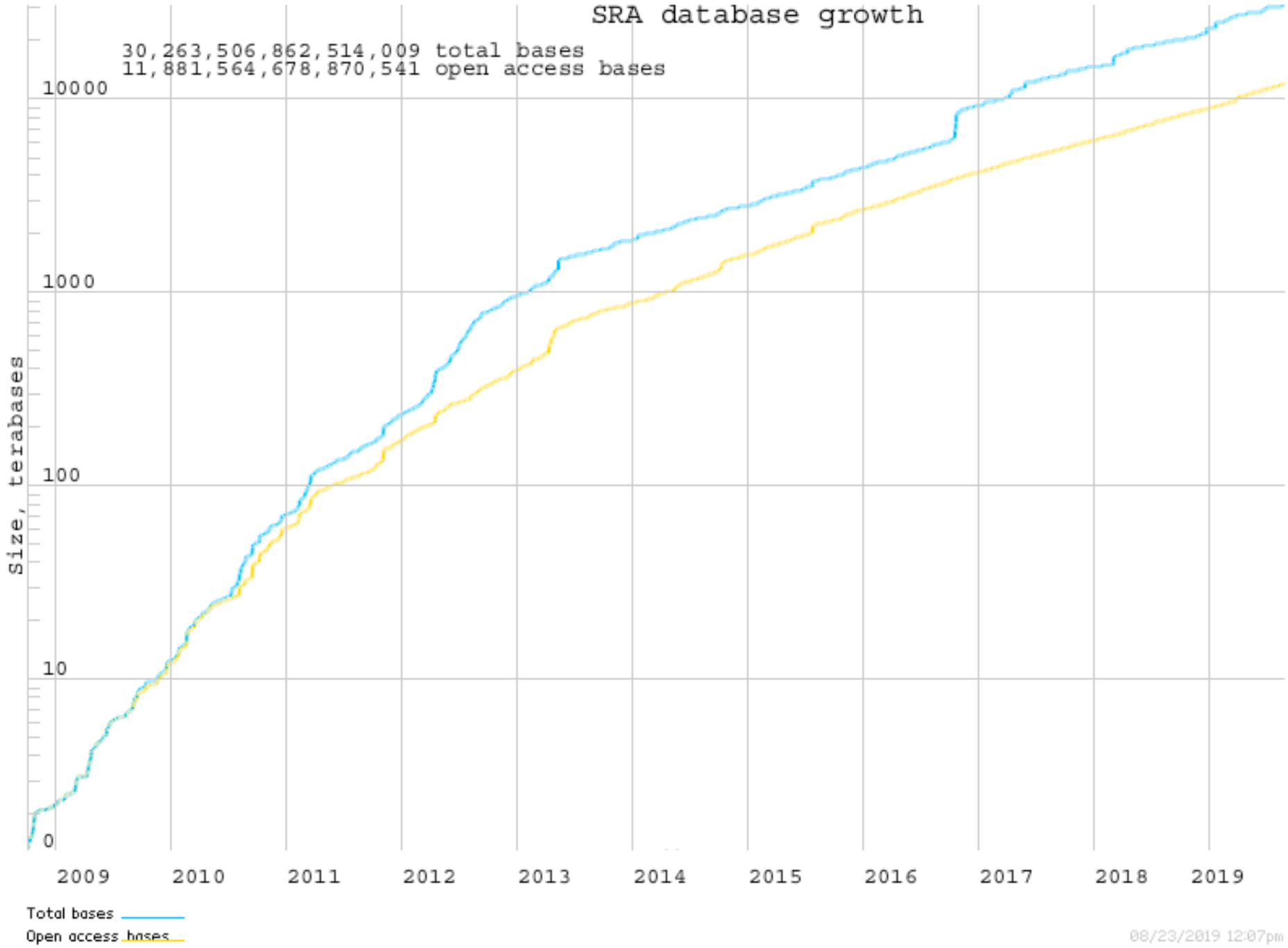
- Imperfect "reads" — results in infrequent but considerable "errors"; modifying, inserting or deleting one or more characters in the "read"

- Biased "reads" — as a result of the underlying chemistry & physics, sampling is not perfectly uniform and random. Biases are not always known.

- Emerging "long read" technologies exist, but have their own set of limitations.

despite these limitations, scientists have taken a very subtle and nuanced approach . . .

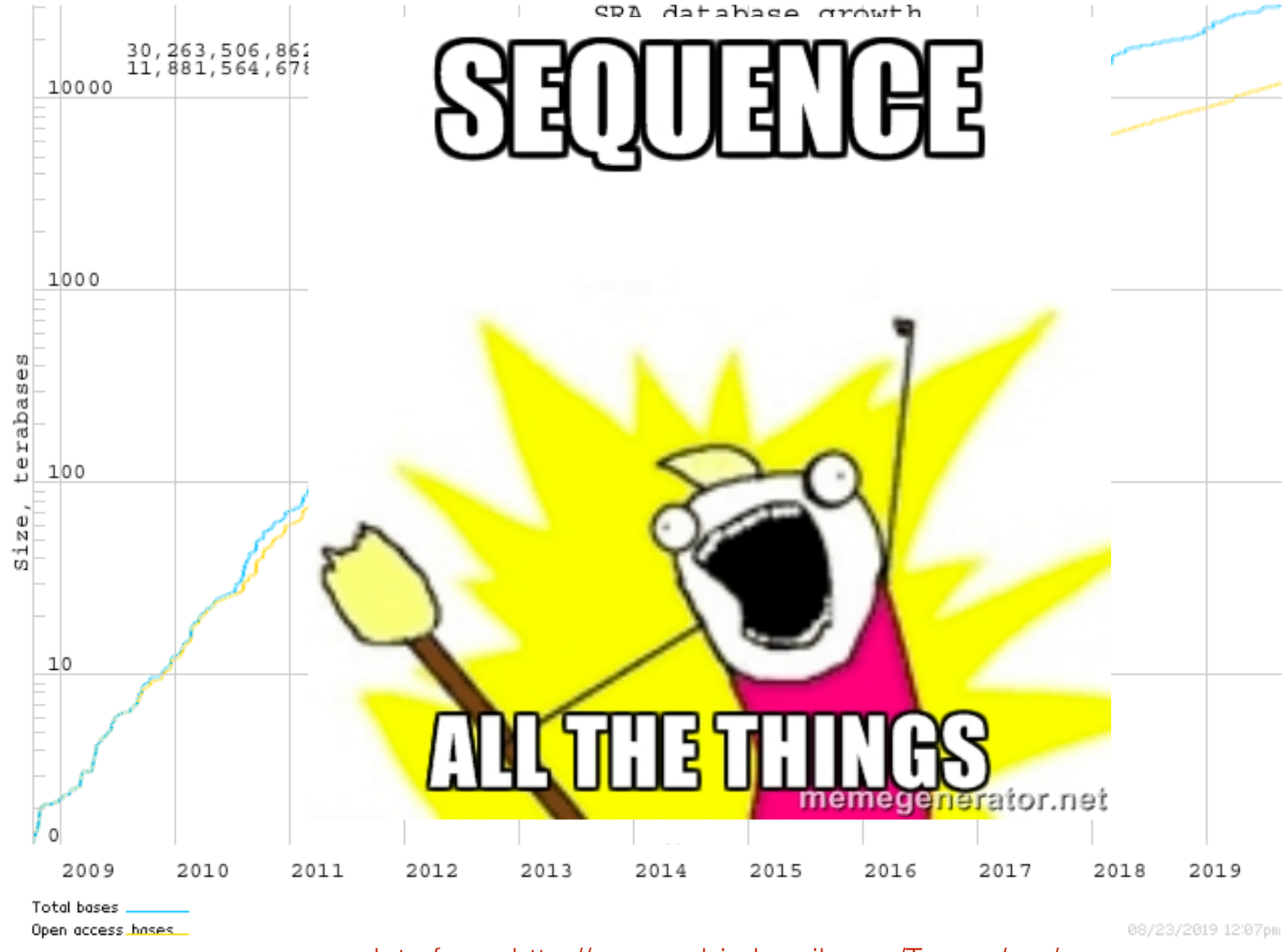
Growth of the Sequence Read Archive (SRA)



data from: http://www.ncbi.nlm.nih.gov/Traces/sra/

As a result, scientists have taken a very subtle and nuanced approach . . .

Growth of the Sequence Read Archive (SRA)



data from: http://www.ncbi.nlm.nih.gov/Traces/sra/

Answer questions “in the large”

What is the genome of the terrapin? (genomics)

Which genes are expressed in healthy vs. diseased tissue? (transcriptomics)

How do environment changes affect the microbial ecosystem of the Chesapeake bay? (metagenomics)

How do genome changes lead to changes & diversity in a population? (population genetics/genomics)

How related are two species if we look at their whole genomes? (phylogenetics / phylogenomics)

Some Computational Challenges

Answering questions on such a scale becomes a *fundamentally* computational endeavor:

Assembly — Find a likely "super string" that parsimoniously explains 200M short sub-strings (string processing, graph theory)

Alignment — Find an *approximate* match for 50M short string in a 5GB corpus of text (string processing, data structure & algorithm design)

Expression / Abundance Estimation — Find the most probable mixture of genes / microbes that explain the results of a sequencing experiment (statistics & ML)

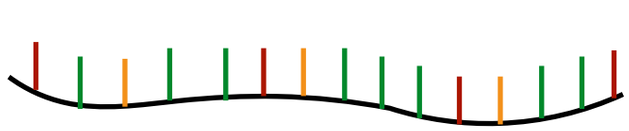
Phylogenomics — Given a set of related gene sequences, and an assumed model of sequence evolution, determine how these sequences are related to each other (statistics & ML)

“Flow” of information in the cell

DNA

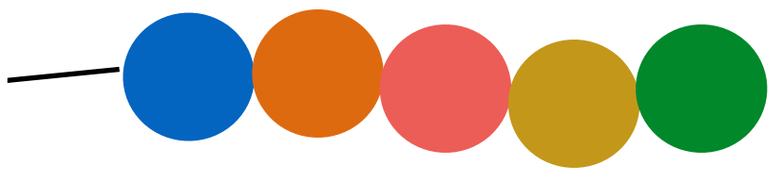
RNA Polymerase
(transcription)

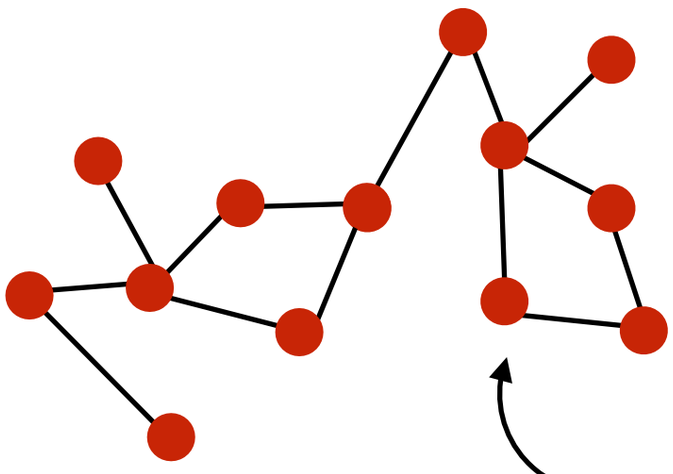
RNA



Ribosomes
(translation)

Protein

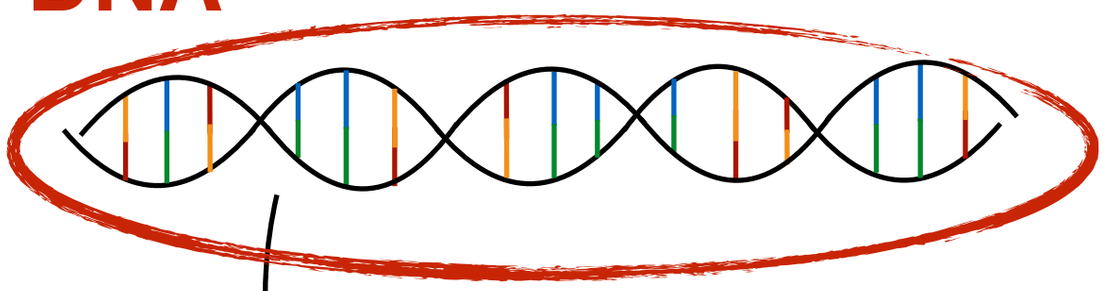




Form networks &
pathways; perform a
vast set of cellular
functions

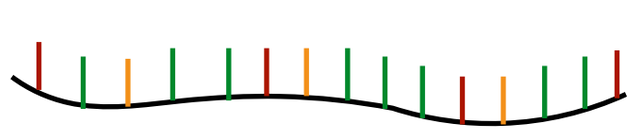
“Flow” of information in the cell

DNA



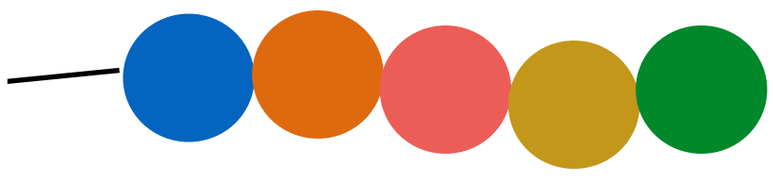
RNA Polymerase
(transcription)

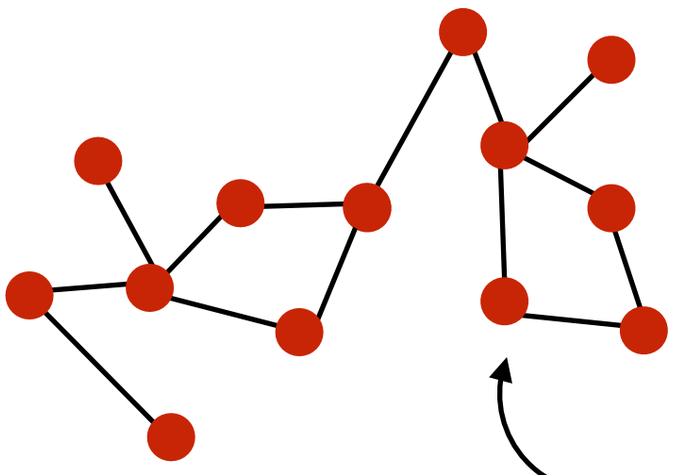
RNA



Ribosomes
(translation)

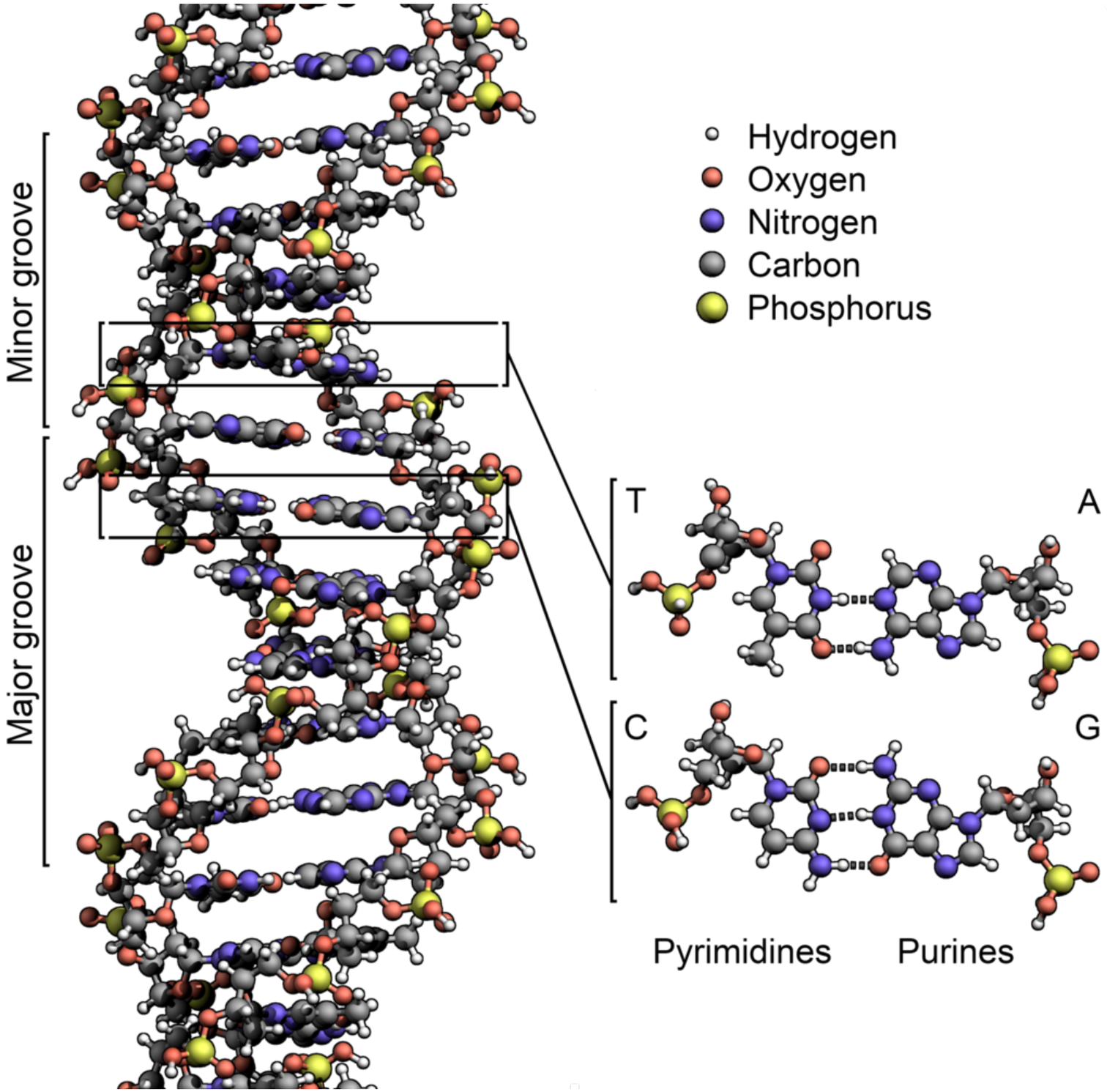
Protein

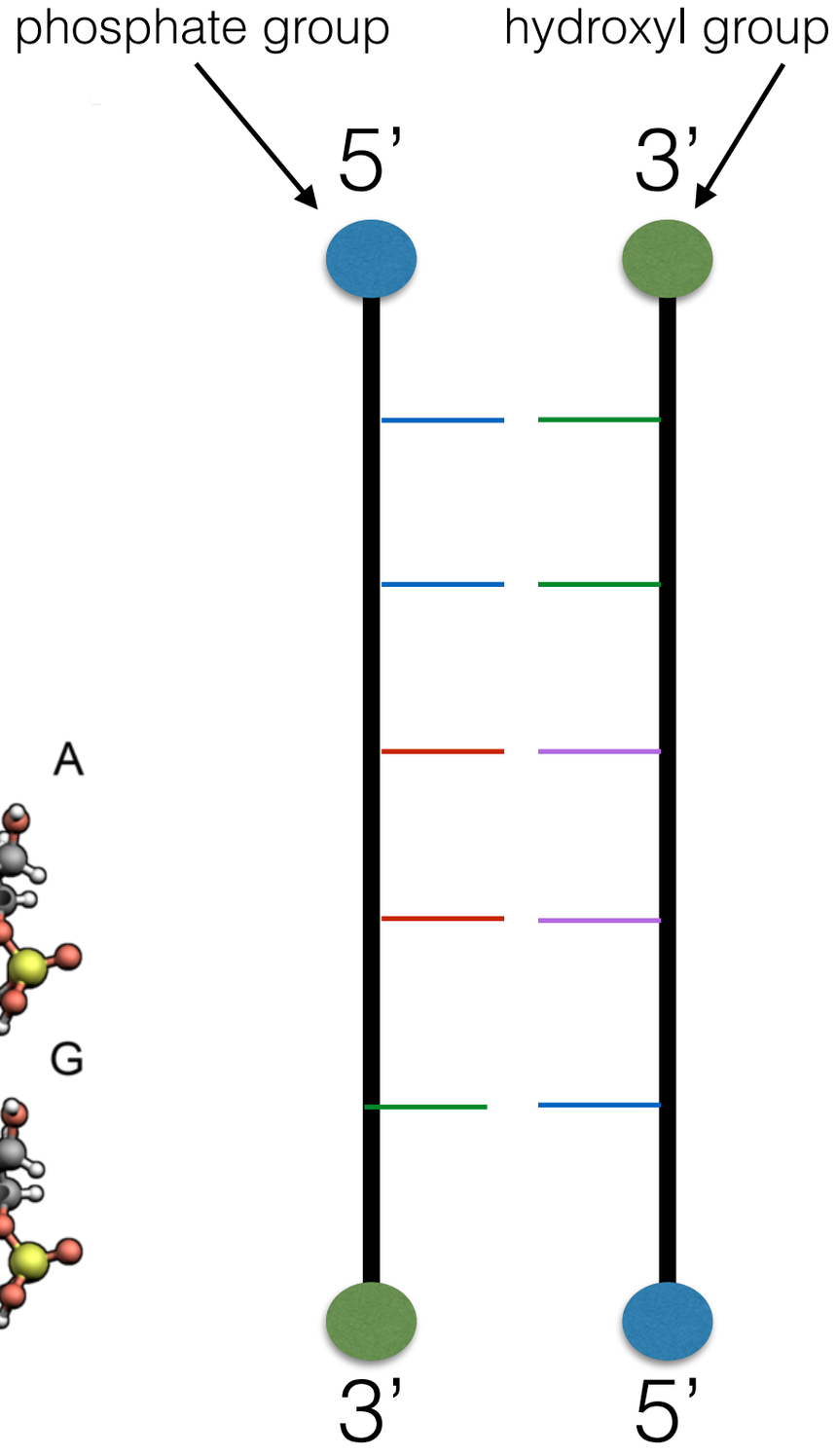




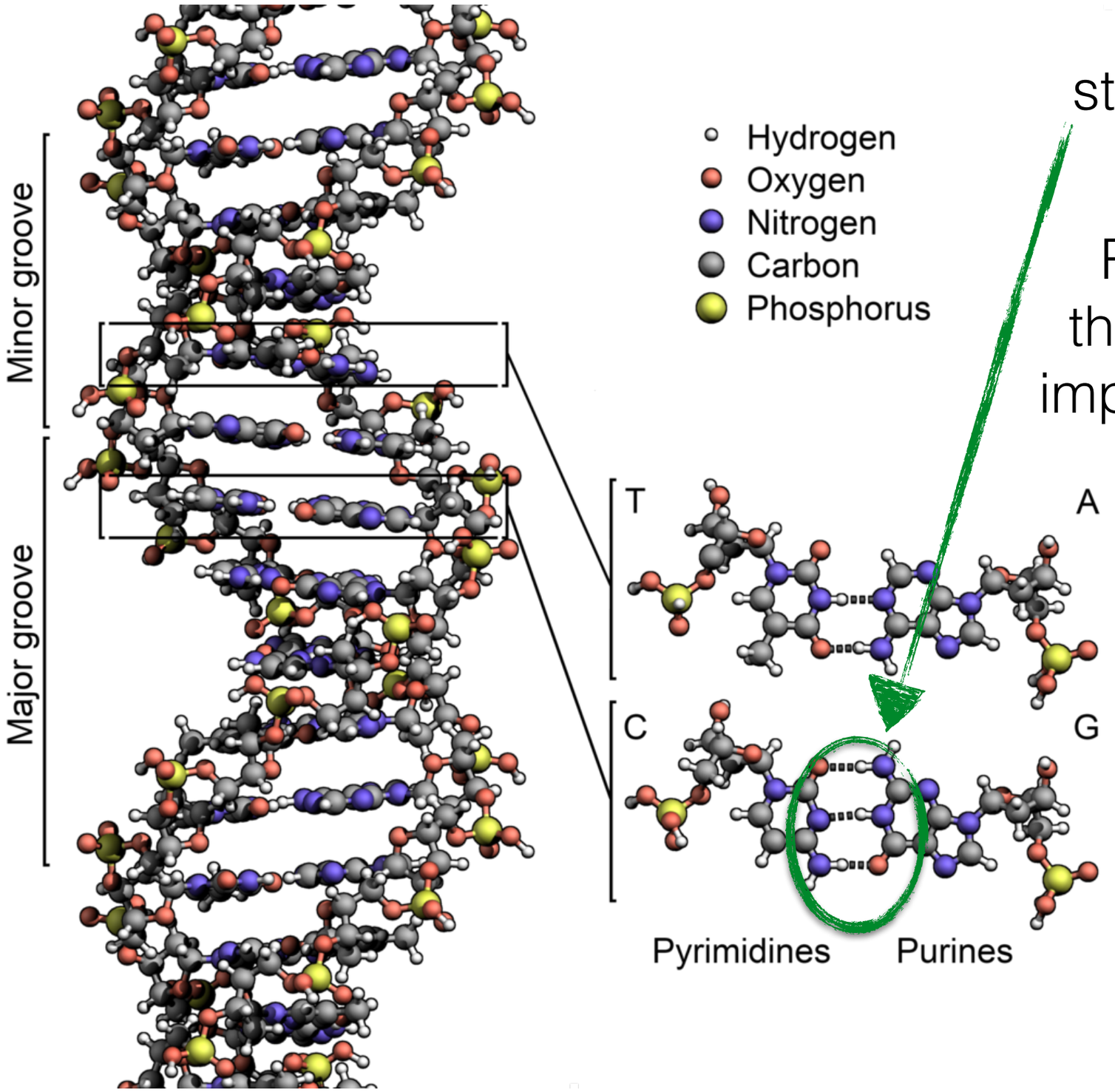
Form networks &
pathways; perform a
vast set of cellular
functions

DNA (the genome)





DNA (the genome)



G-C pairing generally stronger than A-T pairing

Ratio of G+C bases — the “GC content” — is an important sequence feature

DNA (the genome)

gene — will go on to become a protein

“non-coding DNA” — may or may not produce transcripts (e.g. functional non-coding RNA)

In humans, most DNA is “non-coding” ~98%

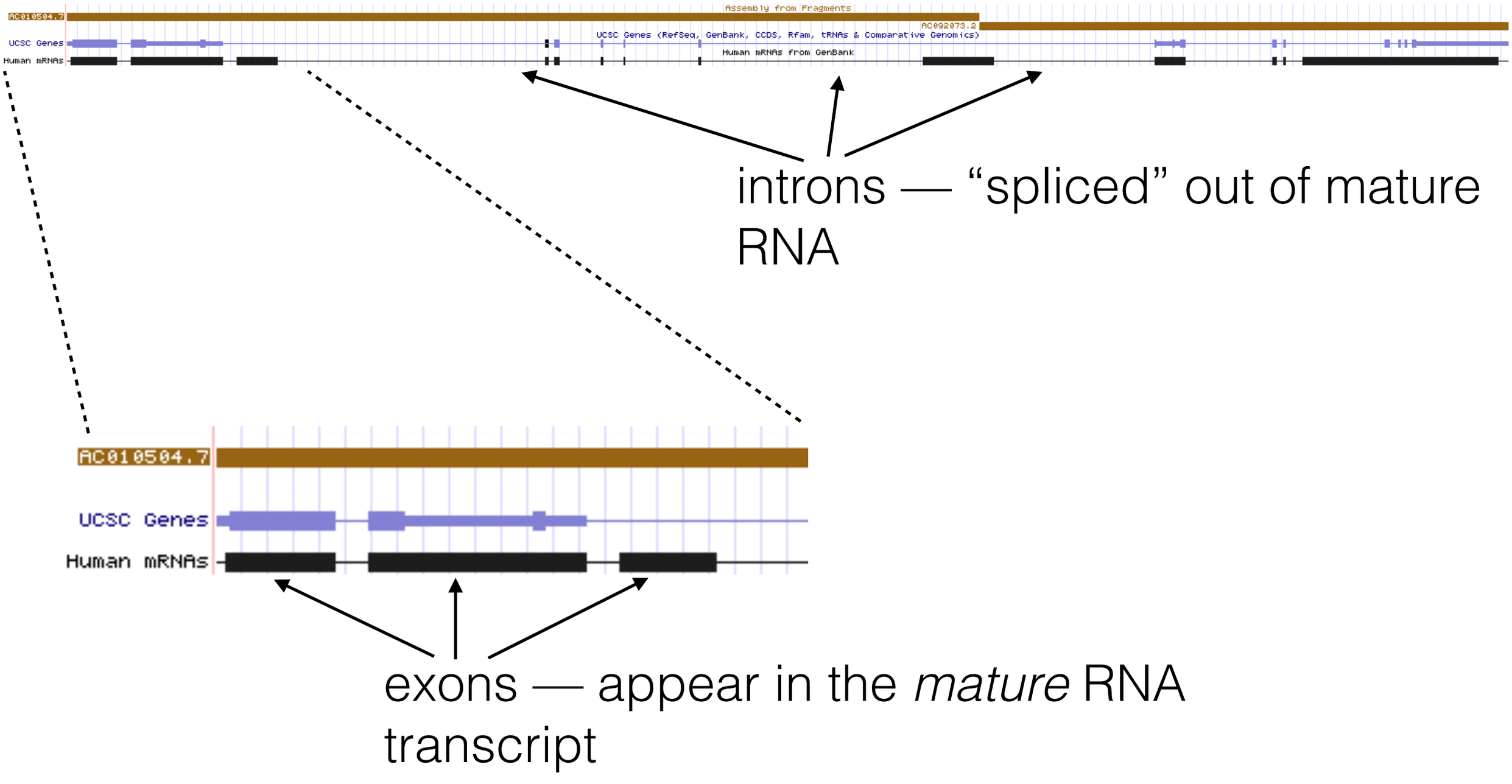
In typical bacterial genome, only small fraction — ~2% — of DNA is “non-coding”

Sometimes referred to as “junk” DNA — much is not, in any way, “junk”

DNA (the genome)

In **prokaryotes**, genes are typically contiguous DNA segment

In **eukaryotes**, genes can have complex structure



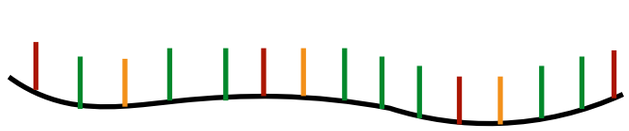
“Flow” of information in the cell

DNA

RNA Polymerase (transcription)

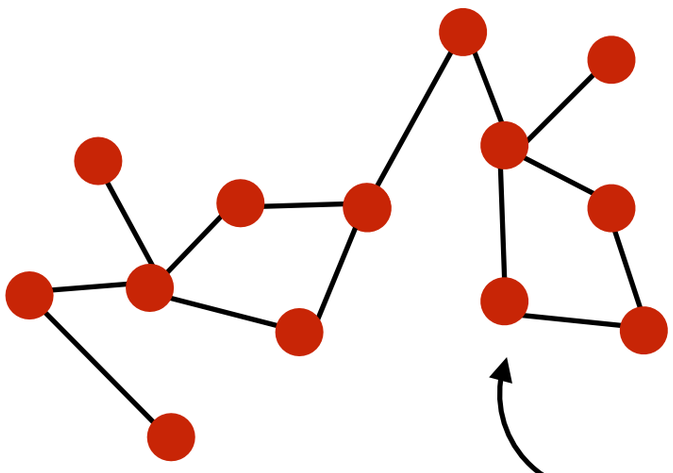
See video on course website

RNA



Ribosomes (translation)

Protein



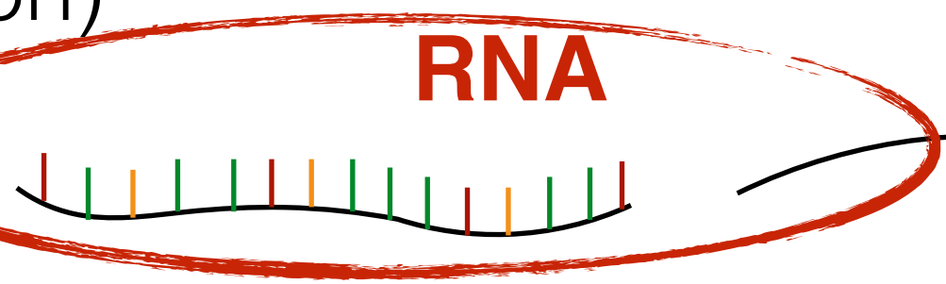
Form networks & pathways; perform a vast set of cellular functions

“Flow” of information in the cell

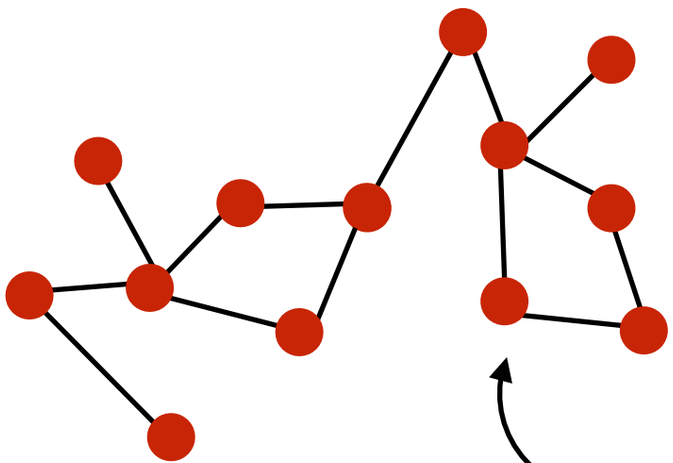
DNA

RNA Polymerase
(transcription)

RNA



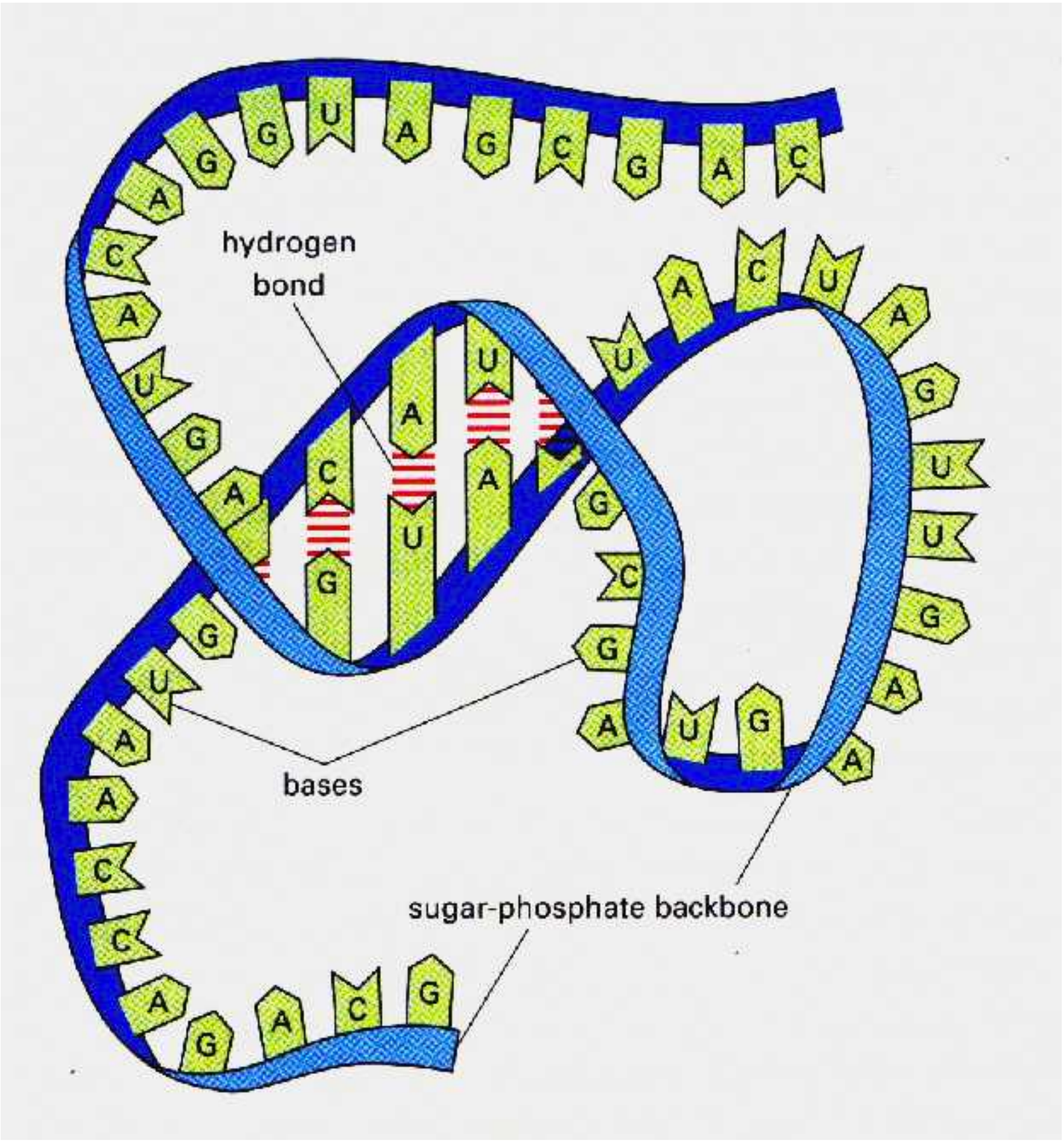
Ribosomes
(translation)



Form networks & pathways; perform a vast set of cellular functions

Protein

RNA



Less regular structure than DNA

Generally a single-stranded molecule

Secondary & tertiary structure can affect function

Act as transcripts for protein, but also perform important functions themselves

Same “alphabet” as DNA, except thymine replaced by uracil

http://tigger.uic.edu/classes/phys/phys461/phys450/ANJUM04/RNA_sstrand.jpg

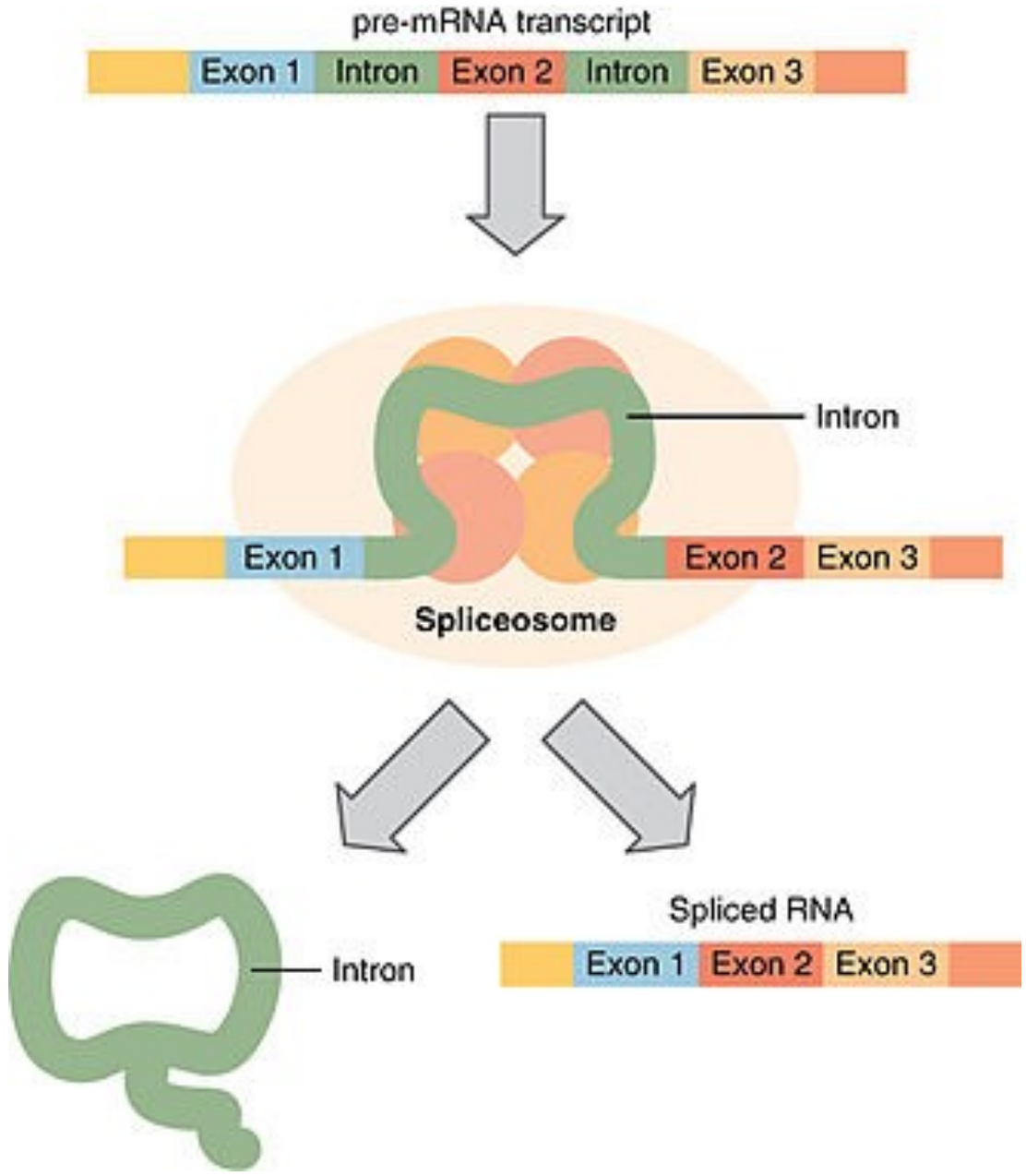
RNA Splicing

DNA transcribed into pre-mRNA

Some “processing occurs”
capping & **polyadenylation**

Introns removed from pre-mRNA

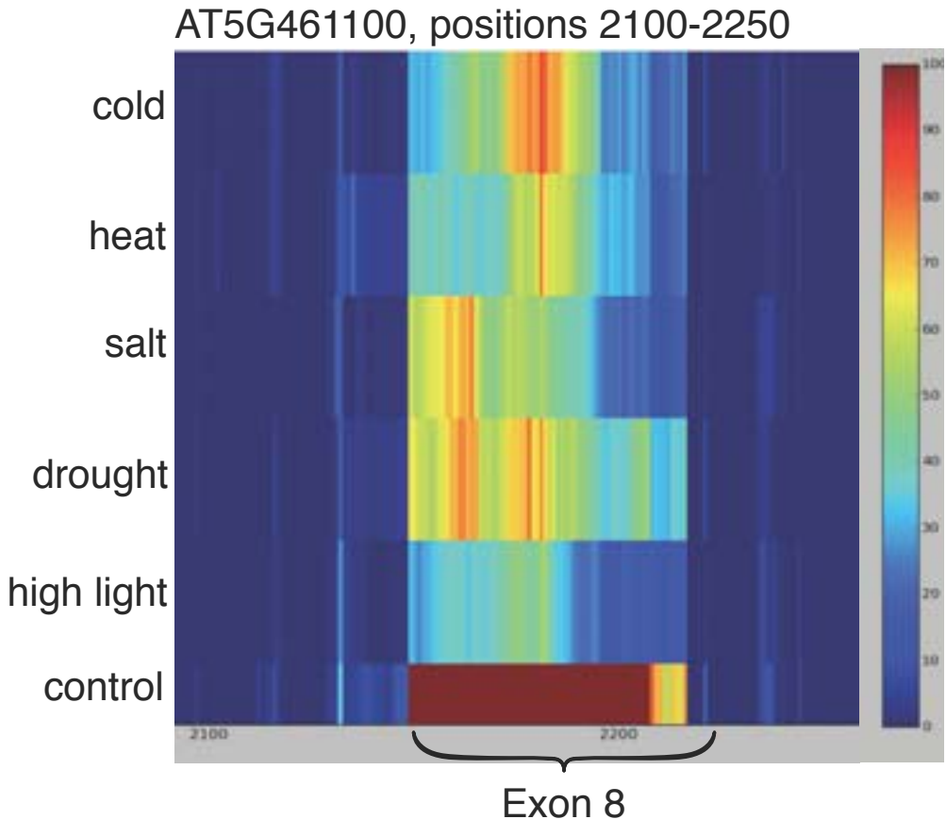
Introns removed resulting in
mature mRNA

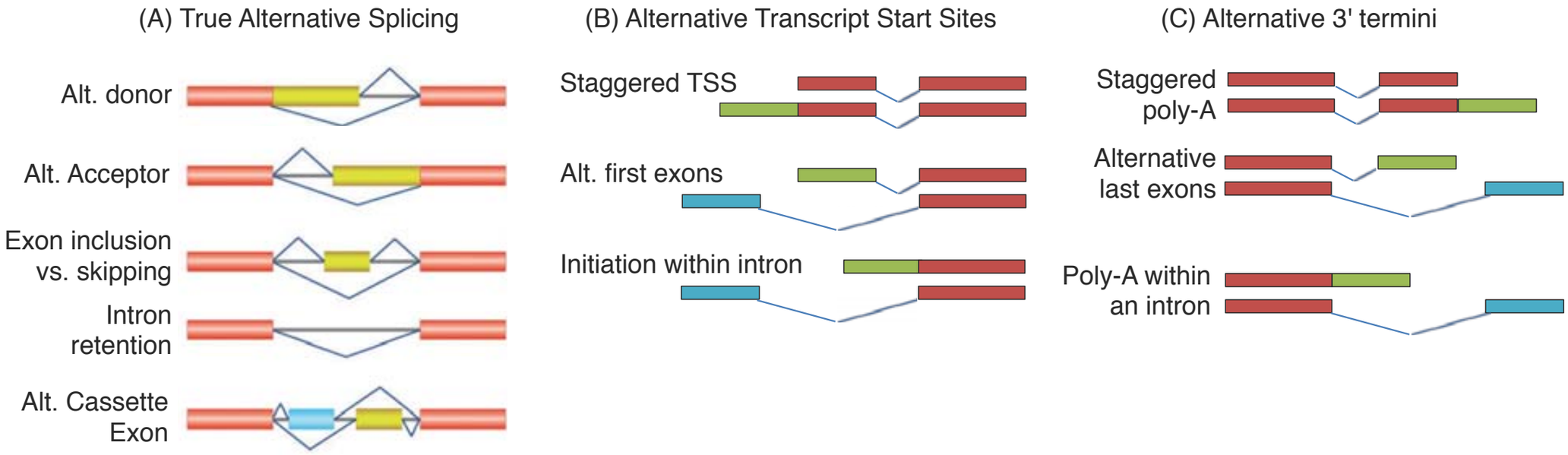


en.wikipedia.org

Alternative Splicing & Isoform Expression

- Expression of genes can be measured via RNA-seq (sequencing transcripts)
- Sequencing gives you short (35-300bp length reads)





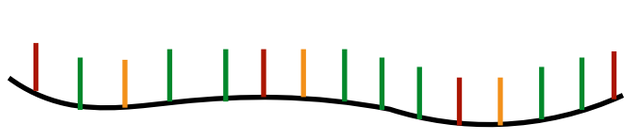
slide courtesy of Carl Kingsford

“Flow” of information in the cell

DNA

RNA Polymerase
(transcription)

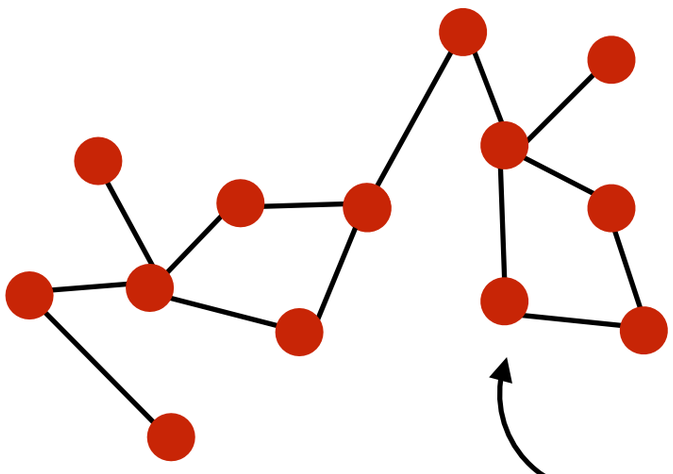
RNA



See video on course website

Ribosomes
(translation)

Protein



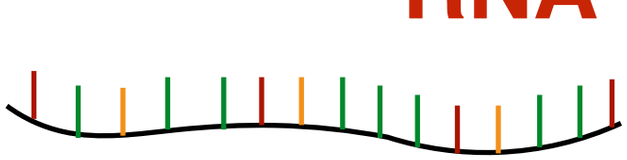
Form networks & pathways; perform a vast set of cellular functions

“Flow” of information in the cell

DNA

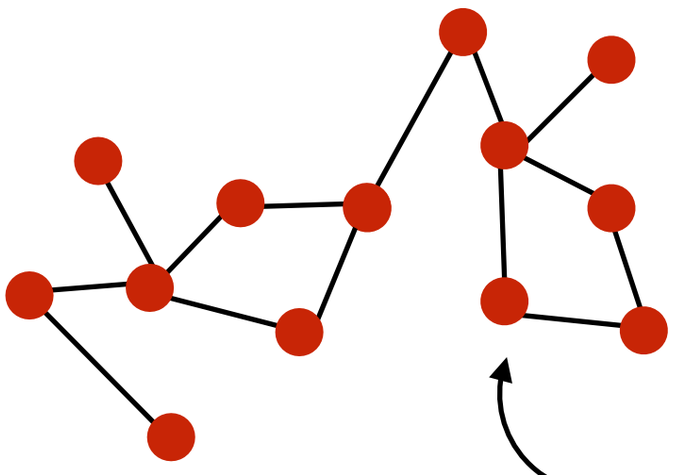
RNA Polymerase
(transcription)

RNA



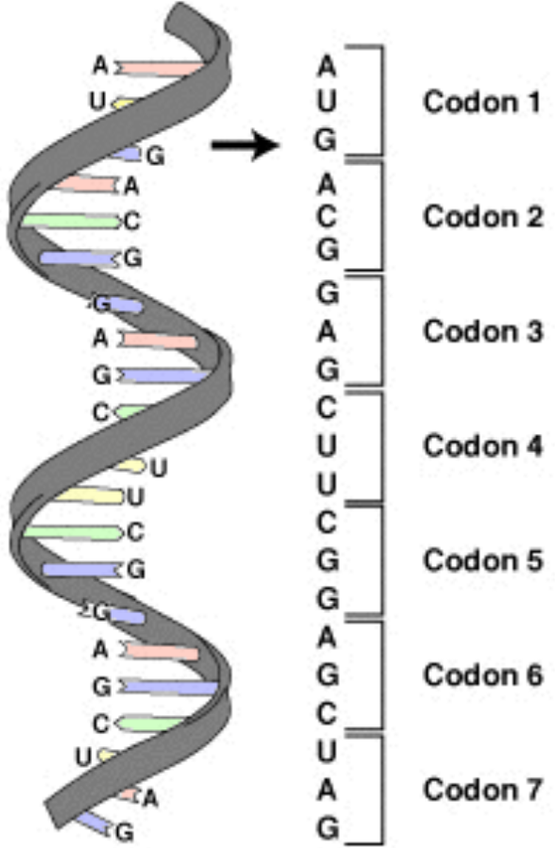
Ribosomes
(translation)

Protein



Form networks &
pathways; perform a
vast set of cellular
functions

Protein



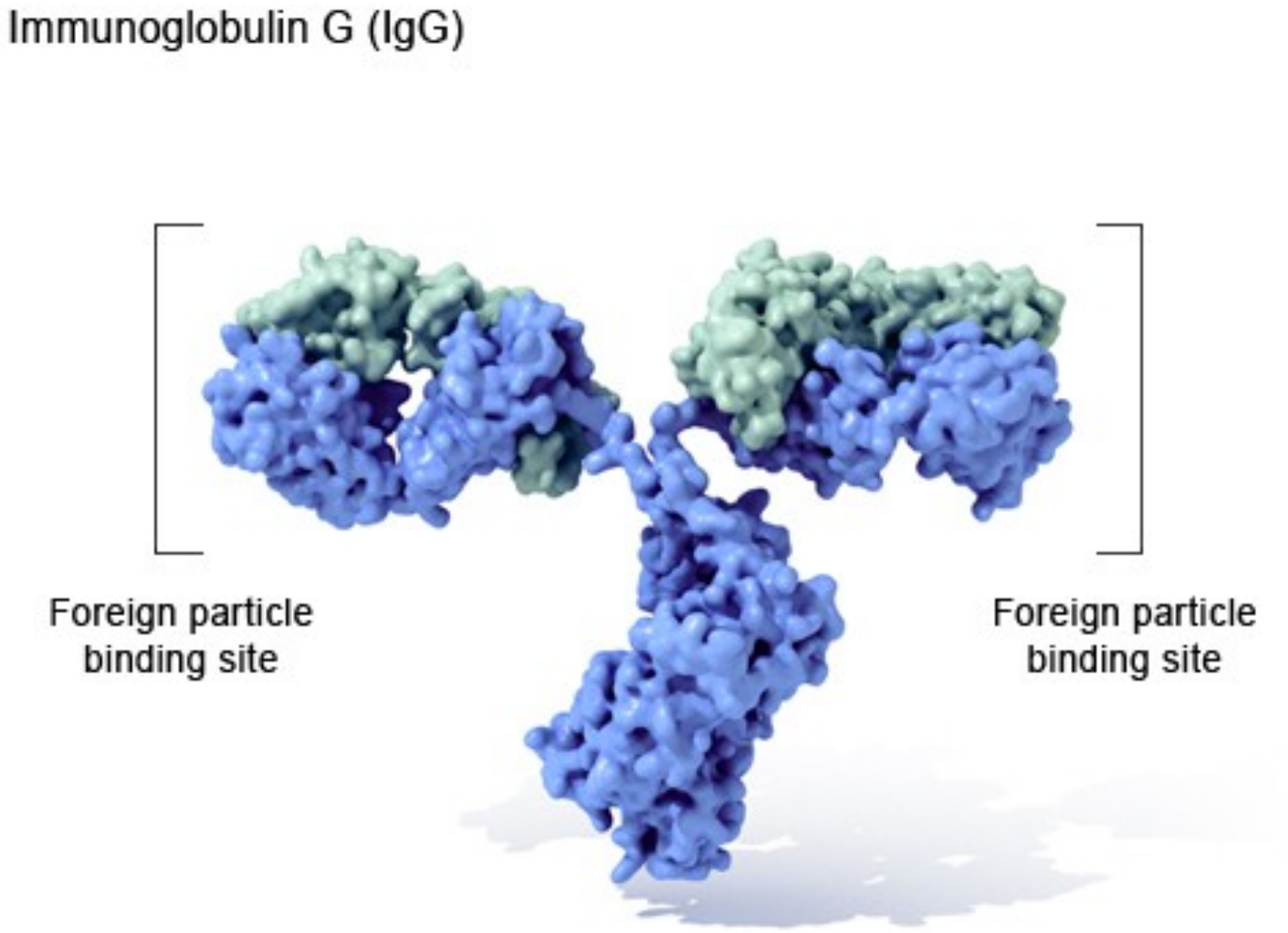
Triplets of mRNA bases (codons) correspond to specific amino acids

This mapping is known as the “genetic code” — an *almost* law of molecular Biology

Inverse table (compressed using IUPAC notation)

Amino acid	Codons	Compressed	Amino acid	Codons	Compressed
Ala/A	GCU, GCC, GCA, GCG	GCN	**Leu/L**	UUA, UUG, CUU, CUC, CUA, CUG	YUR, CUN
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	CGN, MGR	**Lys/K**	AAA, AAG	AAR
Asn/N	AAU, AAC	AAY	**Met/M**	AUG	
Asp/D	GAU, GAC	GAY	**Phe/F**	UUU, UUC	UUY
Cys/C	UGU, UGC	UGY	**Pro/P**	CCU, CCC, CCA, CCG	CCN
Gln/Q	CAA, CAG	CAR	**Ser/S**	UCU, UCC, UCA, UCG, AGU, AGC	UCN, AGY
Glu/E	GAA, GAG	GAR	**Thr/T**	ACU, ACC, ACA, ACG	ACN
Gly/G	GGU, GGC, GGA, GGG	GGN	**Trp/W**	UGG	
His/H	CAU, CAC	CAY	**Tyr/Y**	UAU, UAC	UAY
Ile/I	AUU, AUC, AUA	AUH	**Val/V**	GUU, GUC, GUA, GUG	GUN
START	AUG		**STOP**	UAA, UGA, UAG	UAR, URA

Protein



Perform vast majority of intra & extra cellular functions

Can range from a few amino acids to *very* large and complex molecules

Can bind with other proteins to form protein complexes

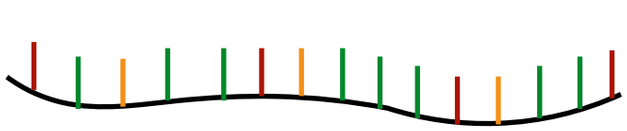
The shape or *conformation* of a protein is intimately tied to its function. Protein shape, therefore, is strongly conserved through evolution — even more so than sequence. A protein can undergo sequence mutations, but fold into the same or a similar shape and still perform the same function.

“Flow” of information in the cell

DNA

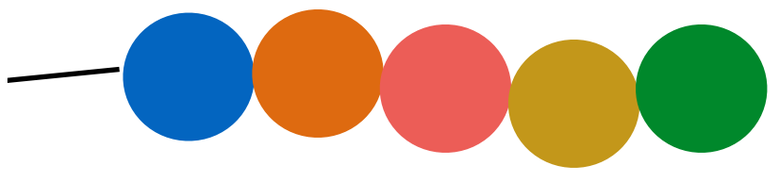
RNA Polymerase
(transcription)

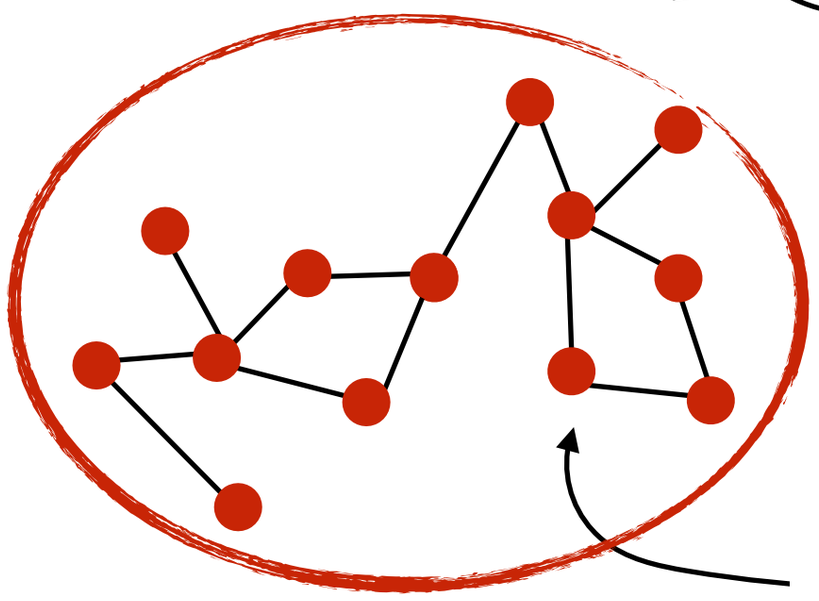
RNA



Ribosomes
(translation)

Protein





Form networks & pathways; perform a vast set of cellular functions

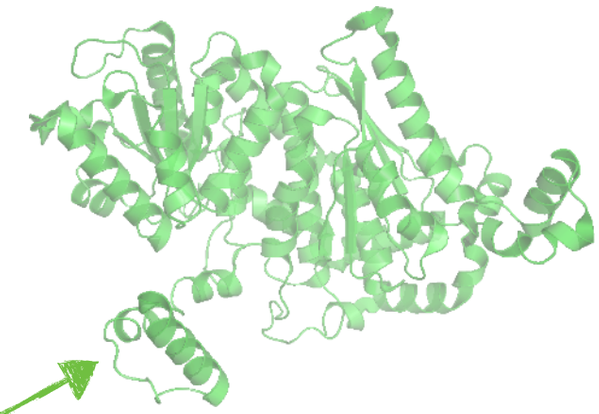
One way in which this “central dogma” is violated … retroviruses

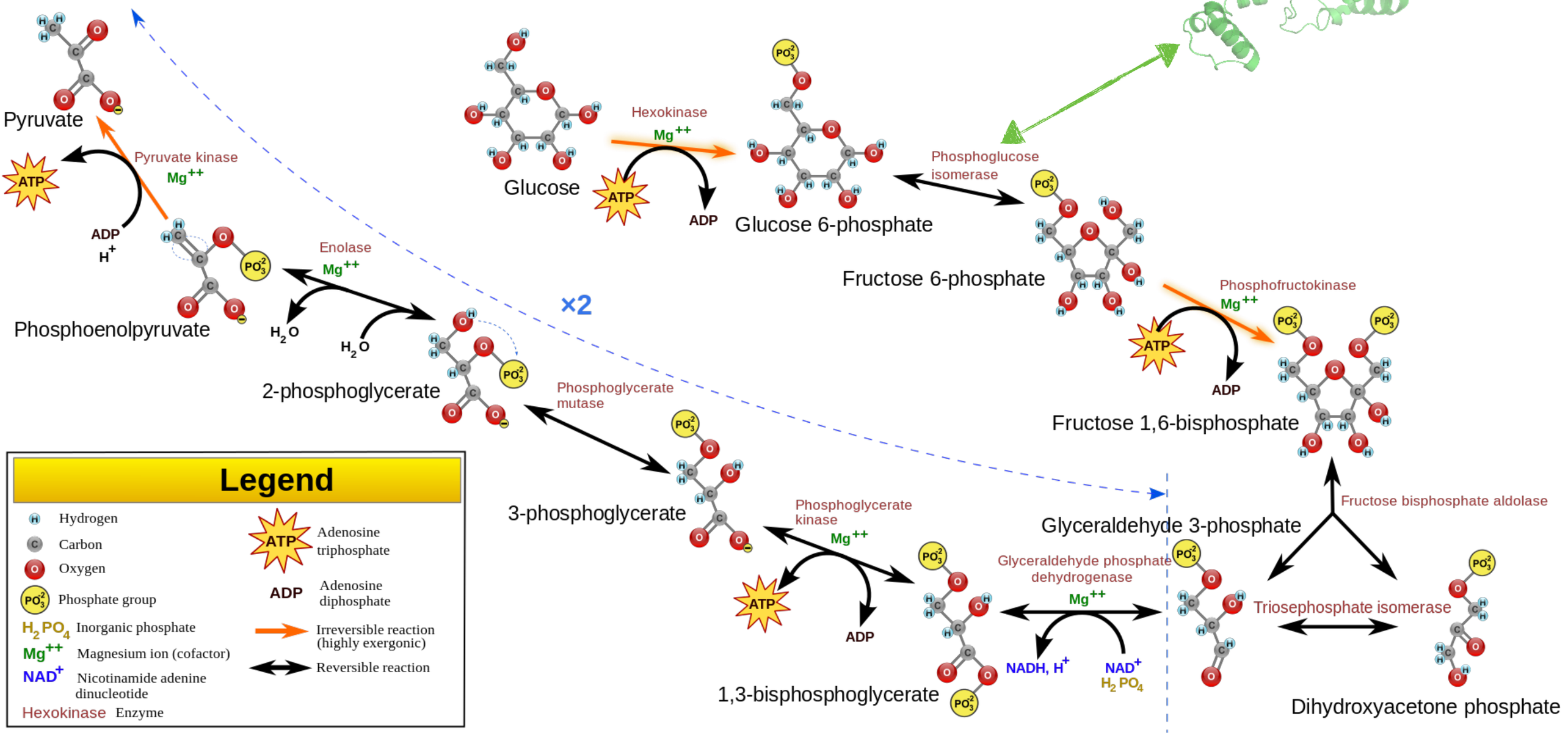
Glycolysis Pathway

Converts glucose → pyruvate

Generates ATP ("energy currency" of the cell)

this is an **example**, no need to memorize this Bio.





Some Interesting Facts

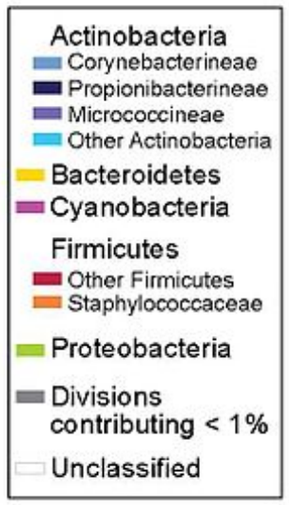
Organism	Genome size	# of genes
ϕX174 (*E. coli* virus)	~5kb	11
E. coli K-12	~4.6Mb	~4,300
Fruit Fly	~122Mb	~17,000
Human	~3.3Gb	~21,000
Mouse	~2.8Gb	~23,000
P. abies (a spruce tree)	~19.6Gb	~28,000

No strong link between genome size & phenotypic complexity

Plants can have **huge** genomes (adapt to environment while stationary!)

http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html

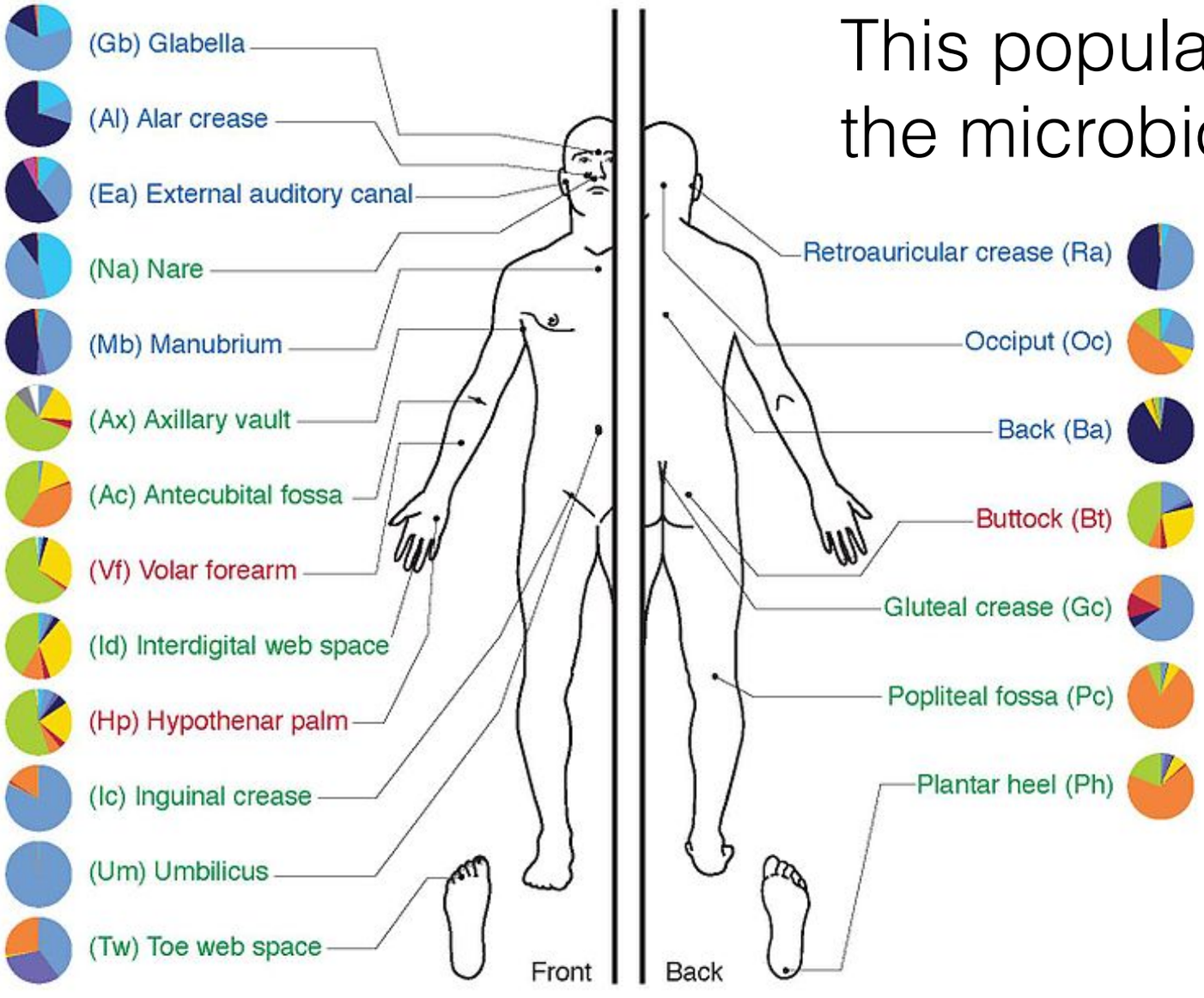
Some Interesting Facts



You are mostly bacteria, fungi & arches

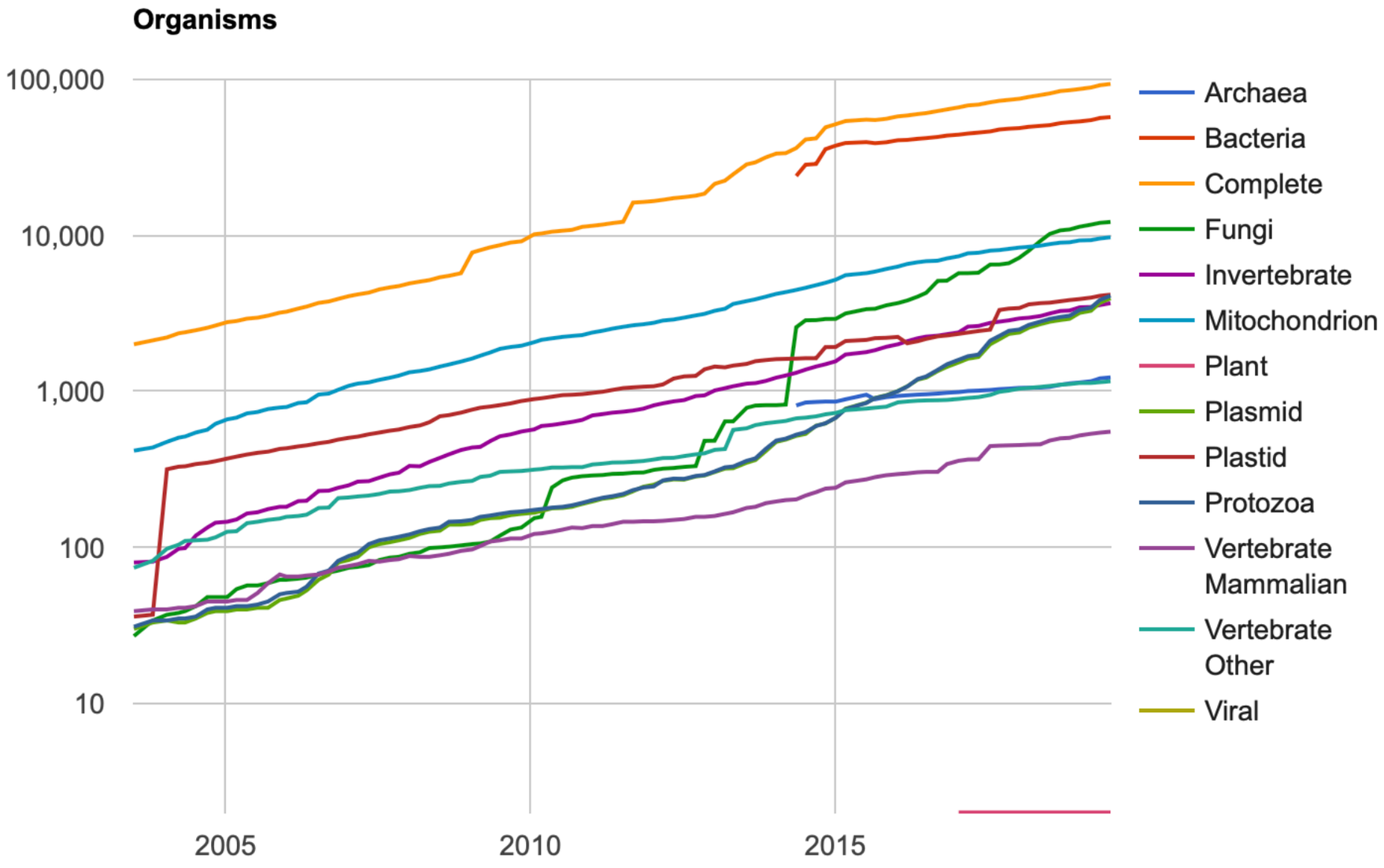
Non-human cells outnumber human cells in the human body

This population of organisms is called the microbiome



en.wikipedia.org : public domain

Some Interesting Facts



https://www.ncbi.nlm.nih.gov/refseq/statistics/

. . . Out of 8.7 ± 1.3 Mil*

Vast majority of species unsequenced & *can not be cultivated in a lab* (one of the many motivations for metagenomics)

*Mora, Camilo, et al. "How many species are there on Earth and in the ocean?." PLoS biology 9.8 (2011): e1001127.