CMSC858D: Algorithms, data structures & inference for high-throughput genomics

Fall 2020

UNIVERSITY OF MARYLAND

NOTE: This lecture is being recorded

Course Info

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

Office: 3220 IRB (but … COVID so, probably don't look there)

Office Hours: By appointment

Website: https://rob-p.github.io/CMSC858D_F20/

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

Academic Integrity: https://academiccatalog.umd.edu/graduate/policies/academic-record/#text

Piazza Page: https://piazza.com/umd/fall2020/cmsc858d

If you have a class-related e-mail: Please prefix the subject with [CMSC858D], so that My filter will pick it up and it won't be accidentally routed to SPAM.

Note: The lectures (zoom) sessions this semester are being recorded so that they can be provided asynchronously via panopto. This will act as notification of recording.

Coursework & Grading

Coursework and grading: The coursework will consist of 2-3 homework projects, a final project, and a final exam. Students will have an opportunity to select their final project in mid Oct.; there will be a few projects to choose from, and students will also be allowed to propose their own projects. The projects are to be done either alone, or in teams of 2. For the final project, the deliverables will consist of runnable code (including a link to a version-controled repository containing the source), and a short (4-5 page) research-style paper describing the work you've done. The breakdown of weights for these different assignments will be as follows:

- Homeworks — 25%
- Final Project — 50%
- Final Exam — 25%

Academic Integrity maintain it!

Academic integrity: From the University's Graduate Catalog Statement on Academic Integrity :

> *On every examination, paper or other academic exercise not specifically exempted by the instructor, the student will write by hand and sign the following pledge: I pledge on my honor that I have not given or received any unauthorized assistance on this examination.*

> *Failure to sign the pledge is not an honors offense, but neither is it a defense in case of violation of this Code. Students who do not sign the pledge will be given the opportunity to do so. Refusal to sign must be explained to the instructor. Signing or non-signing of the pledge will not be considered in grading or judicial procedures. Material submitted electronically should contain the pledge; submission implies signing the pledge.*

> *On examinations, no assistance is authorized unless given by or expressly allowed by the instructor. On other assignments, the pledge means that the assignment has been done without academic dishonesty, as defined in the Code of Academic Integrity, available online.*

> *The pledge is a reminder that at the University of Maryland students carry primary responsibility for academic integrity because the meaningfulness of their degrees depends on it. Faculty are urged to emphasize the importance of academic honesty and of the pledge as its symbol.*

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

None-required … but :

Genomics algorithms, data structures, and statisitcal models:

- Genome Scale Algorithm Design (Mäkinen, Belazzougui, Cunial, Tomescu 2015)
- Bioinformatics Algorithms: An Active Learning Approach (Pevzner and Compeau, 2018)

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/

Conceptually, what do these machines produce

Original DNA molecule

— ACTACCGACGCCGAATCACTTCTCAGGATCCATAGGCAAC →
← TGATGGCTGCGGCTTAGTGAAGAGTCCTAGGTATCCGTTG —

"Amplification"

— ACTACCGACGCCGAATCACTTCTCAGGATCCATAGGCAAC →
← TGATGGCTGCGGCTTAGTGAAGAGTCCTAGGTATCCGTTG —

— ACTACCGACGCCGAATCACTTCTCAGGATCCATAGGCAAC →
← TGATGGCTGCGGCTTAGTGAAGAGTCCTAGGTATCCGTTG —

— ACTACCGACGCCGAATCACTTCTCAGGATCCATAGGCAAC →
← TGATGGCTGCGGCTTAGTGAAGAGTCCTAGGTATCCGTTG —

— ACTACCGACGCCGAATCACTTCTCAGGATCCATAGGCAAC →
← TGATGGCTGCGGCTTAGTGAAGAGTCCTAGGTATCCGTTG —

— ACTACCGACGCCGAATCACTTCTCAGGATCCATAGGCAAC →
← TGATGGCTGCGGCTTAGTGAAGAGTCCTAGGTATCCGTTG —

— ACTACCGACGCCGAATCACTTCTCAGGATCCATAGGCAAC →
← TGATGGCTGCGGCTTAGTGAAGAGTCCTAGGTATCCGTTG —

"Fragmentation"

– ACTACCGACGCCGAATCACT →
← TGATGGCTGCGGCTTAGTGA –

– TCACTTCTCAGGATCCATAGGCAAC →
← AGTGAAGAGTCCTAGGTATCCGTTG –

– GCCGAATCACTTCTCAGGATCC →
← CGGCTTAGTGAAGAGTCCTAGG –

– CCGACGCCGAATCACTTCTCAGGA →
← GGCTGCGGCTTAGTGAAGAGTCC –

– ATCACTTCTCAGGATCCATAGGCAA →
← TAGTGAAGAGTCCTAGGTATCCGTT –

— TCTCAGGATCCATAGGCAAC →
← AGAGTCCTAGGTATCCGTTG —

Conceptually, what do these machines produce

"Fragmentation"

– ACTACCGACGCCGAATCACT ▸

◂ TGATGGCTGCGGCTTAGTGA –

– TCACTTCTCAGGATCCATAGGCAAC ▸

◂ AGTGAAGAGTCCTAGGTATCCGTTG –

– GCCGAATCACTTCTCAGGATCC ▸

◂ CGGCTTAGTGAAGAGTCCTAGG –

– CCGACGCCGAATCACTTCTCAGGA ▸

◂ GGCTGCGGCTTAGTGAAGAGTCC –

– ATCACTTCTCAGGATCCATAGGCAA ▸

◂ TAGTGAAGAGTCCTAGGTATCCGTT –

— TCTCAGGATCCATAGGCAAC →

◂ AGAGTCCTAGGTATCCGTTG —

"Sequencing"

10bp Paired-end

– ACTACCGACGCCGAATCACT ▸

◂ TGATGGCTGCGGCTTAGTGA –

— TCTCAGGATCCATAGGCAAC →

◂ AGAGTCCTAGGTATCCGTTG —

– ATCACTTCTCAGGATCCATAGGCAA ▸

◂ TAGTGAAGAGTCCTAGGTATCCGTT –

– GCCGAATCACTTCTCAGGATCC ▸

◂ CGGCTTAGTGAAGAGTCCTAGG –

– CCGACGCCGAATCACTTCTCAGGA ▸

◂ GGCTGCGGCTTAGTGAAGAGTCC –

– TCACTTCTCAGGATCCATAGGCAAC ▸

◂ AGTGAAGAGTCCTAGGTATCCGTTG –

```
>foobar1/1
ACTACCGACG
>foobar2/1
TCTCAGGATC
>foobar3/1
ATCACTTCTC
>foobar4/1
GCCGAATCAC
>foobar5/1
CCGACGCCGA
>foobar6/1
TCACTTCTCA
```

```
>foobar1/2
AGTGATTCGG
>foobar2/2
GTATCCGTTG
>foobar3/2
GGTATCCGTT
>foobar4/2
GAGTCCTAGG
>foobar5/2
TGAAGAGTCC
>foobar6/2
GTATCCGTTG
```

Conceptually, what do these machines produce

"Sequencing"

ACTACCGACGCCGAATCACT →

← TGATGGCTGCGGCTTAGTGA

→ TCTCAGGATCCATAGGCAAC →

← AGAGTCCTAGGTATCCGTTG

ATCACTTCTCAGGATCCATAGGCAA →

← TAGTGAAGAGTCCTAGGTATCCGTT

GCCGAATCACTTCTCAGGATCC →

← CGGCTTAGTGAAGAGTCCTAGG

CCGACGCCGAATCACTTCTCAGGA →

← GGCTGCGGCTTAGTGAAGAGTCC

TCACTTCTCAGGATCCATAGGCAAC →

← AGTGAAGAGTCCTAGGTATCCGTTG

10bp Paired-end

```
>foobar1/1
ACTACCGACG
>foobar2/1
TCTCAGGATC
>foobar3/1
ATCACTTCTC
>foobar4/1
GCCGAATCAC
>foobar5/1
CCGACGCCGA
>foobar6/1
TCACTTCTCA
```

```
>foobar1/2
AGTGATTCGG
>foobar2/2
GTATCCGTTG
>foobar3/2
GGTATCCGTT
>foobar4/2
GAGTCCTAGG
>foobar5/2
TGAAGAGTCC
>foobar6/2
GTATCCGTTG
```

Why does it work this way?

Why are there 2 "reads" for every fragment?

Why are the "/2" reads "backward"?

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.

Long-read sequencing

ONT

MinION

- Pocket-sized, portable device for biological analysis
- Up to 512 nanopore channels
- Simple 10-min sample prep available
- Real-time analysis for rapid, efficient workflows
- Adaptable to direct DNA or RNA sequencing
- MinIT available to support IT/software needs

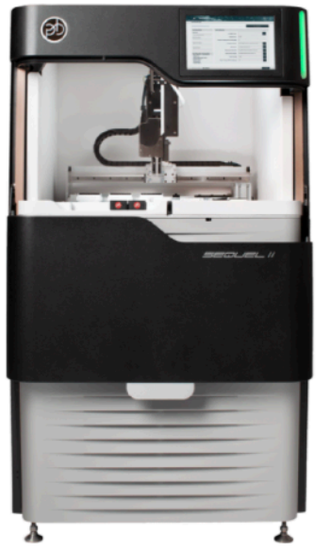
PromethION

- High-throughput, high-sample number benchtop system
- Modular: Up to 48 flow cells, each with up to 3,000 nanopore channels (total up to 144,000)
- Flow cells may be run individually or concurrently
- Same workflow as MinION at larger scale

PacBio

THE AWARD-WINNING SEQUEL II SYSTEM

Delivering Highly Accurate Long Reads



The Sequel II System provides the advantages of SMRT Sequencing and now makes it more affordable for all scientists to drive discovery with comprehensive views of genomes and transcriptomes.

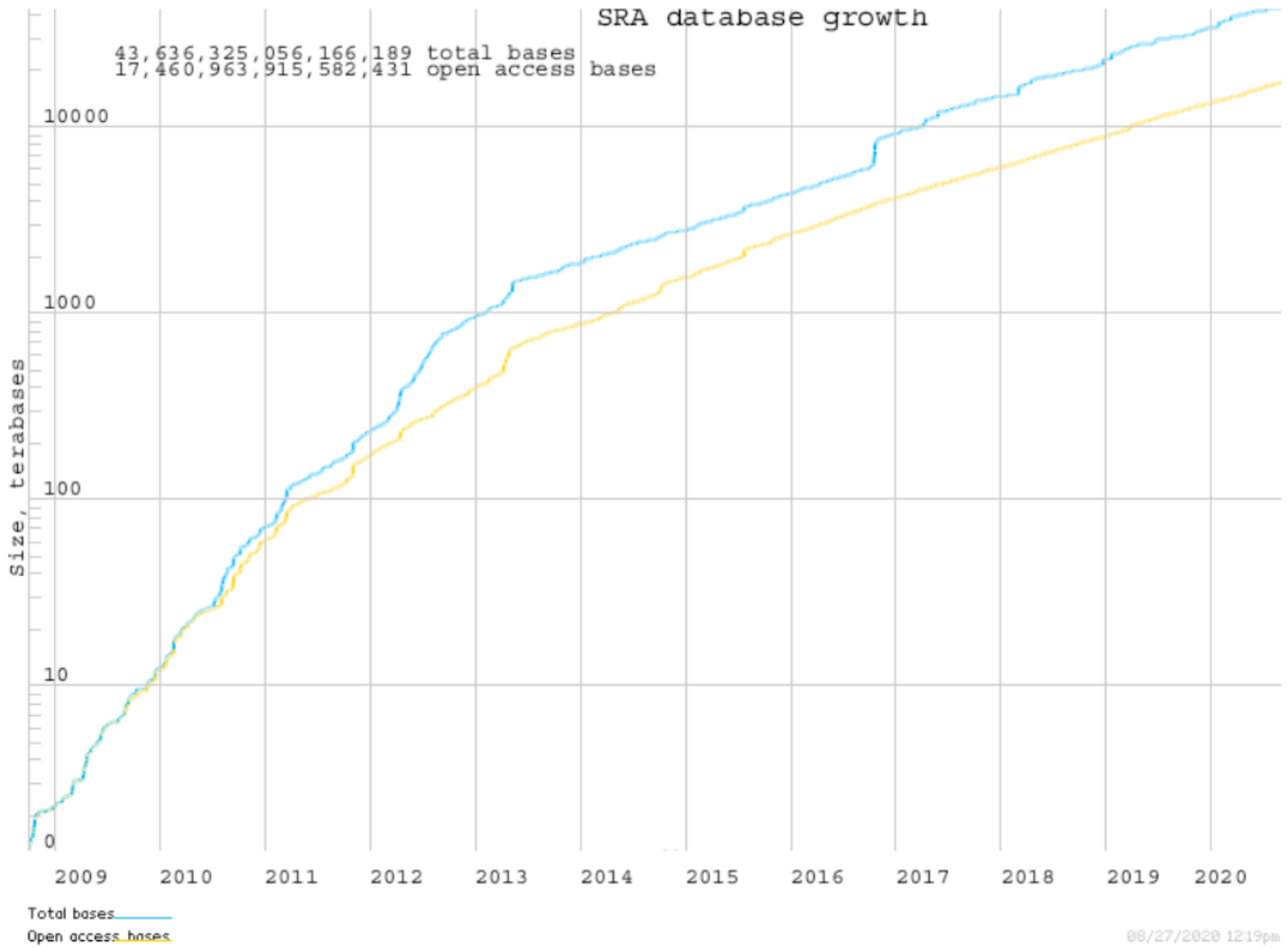
- Produces the exceptional results customers have come to expect from PacBio Systems
- Generates ~8-times more data than the original Sequel System
- Provides access to even more highly accurate long reads (HiFi reads) – Learn more about HiFi reads
- Reduces project time for faster results
- Makes sequencing more affordable
- Supports the range of SMRT Sequencing applications

Much longer reads, but many fewer of them

Generally, much higher error rate (errors are both substitution & indel)

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

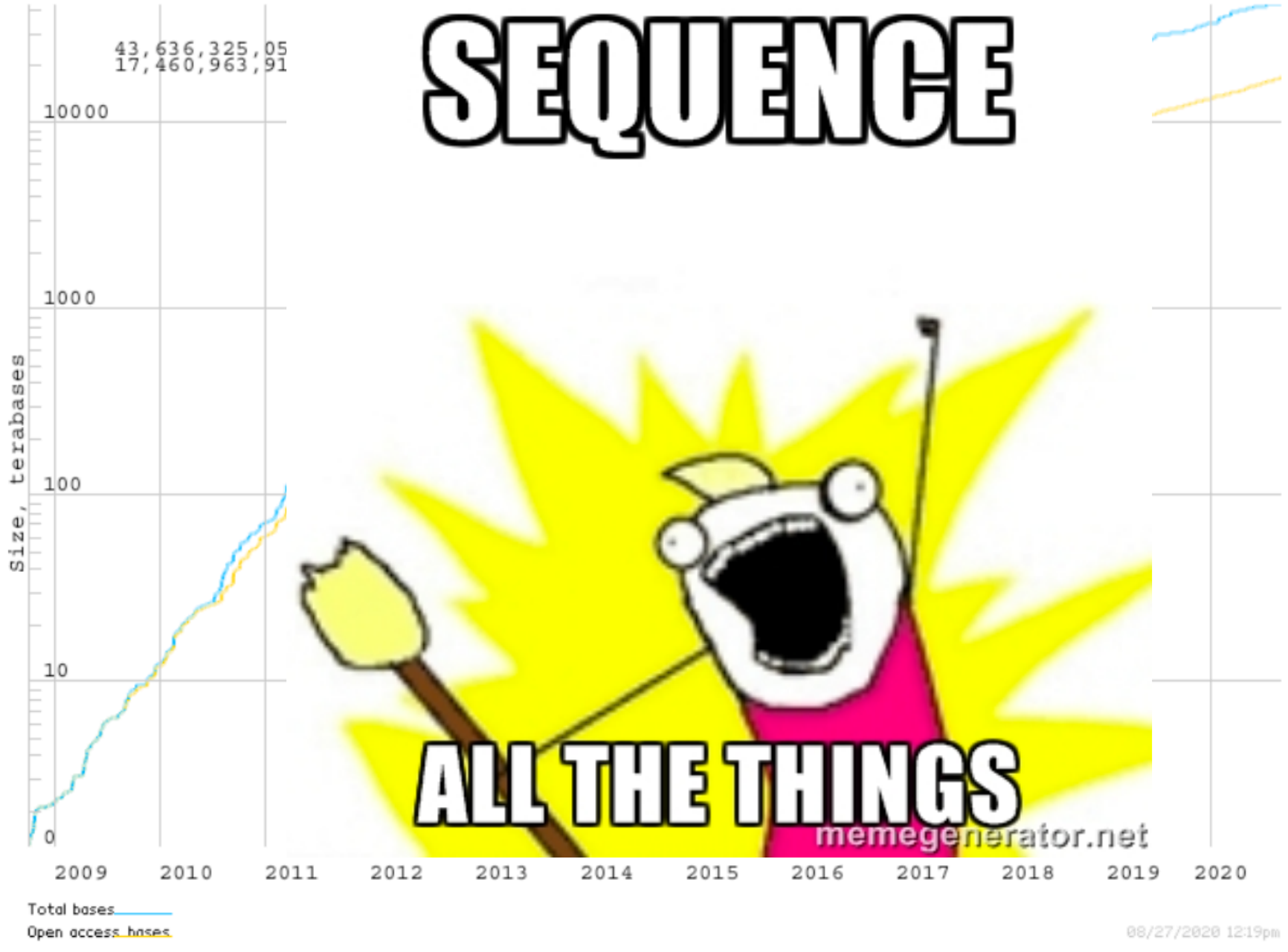
Growth of the Sequence Read Archive (SRA)



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 environmental 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)

CS & Biology, some differences

Some differences between CS & Biology as academic fields, and in terms of culture.

“Scientific” differences

Biology deals with *very* complex natural systems that arise through evolution

Biological systems can be indirect, redundant and counterintuitive

Nothing is “always” true/false — Biological laws are not like Physical or Mathematical laws; more stochastic truths or rules of thumb.

Biological laws *are* a result of Physical laws, but treating them that way is computationally infeasible

Try to understand mechanisms by probing and measuring complex systems and obtaining (often noisy) measurements

Experiments often *very* expensive

“Scientific” differences

Computer Science deals with *less* complex (won’t say simple) systems that arise through design

CS is more about invention than discovery (philosophy aside)

Things are always formally true or false in CS & detailed theoretical analysis allows precise description

Computational outcomes *are* a result of mathematical laws & effective algorithms often have an intuitive explanation

Some subfields of CS (e.g. network measurement) do bear a resemblance to the natural sciences — many are much closer to math.

Experiments often dirt cheap and easy to re-run

“Cultural” differences

Biology	CS
Only journals matter	Selective conferences often preferred to journals
Larger labs: PI → postdoc→grad students	Smaller labs: PI → a few grad students
Student may study a specific gene for their entire PhD	Students typically work on a wide variety of projects in PhD
Focus on being “right” and discovering something interesting about the natural world. (focus on knowledge)	More weight given to being “different”. Need not be 1st often just be “best”, fastest or simplest. (focus on methods)

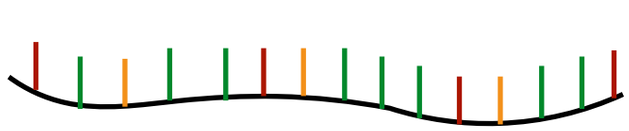
Many of these differences start to vanish at the interface of data-intensive Biology, where computational savvy is a necessity.

“Flow” of information in the cell

DNA

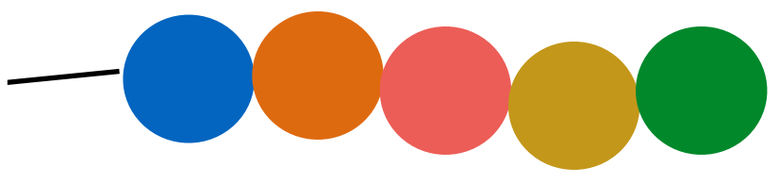
RNA Polymerase
(transcription)

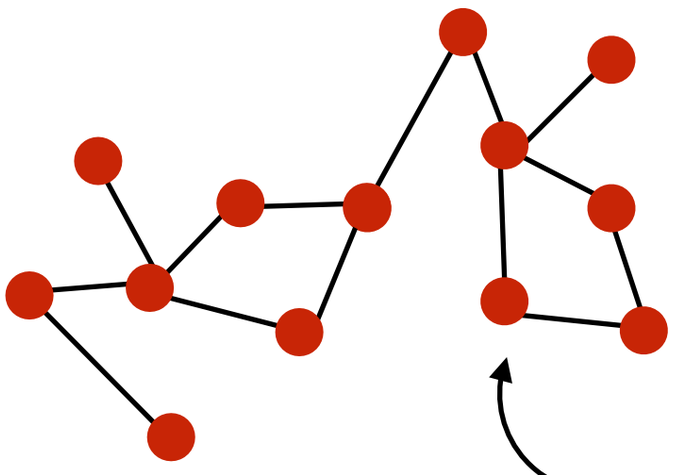
RNA



Ribosomes
(translation)

Protein

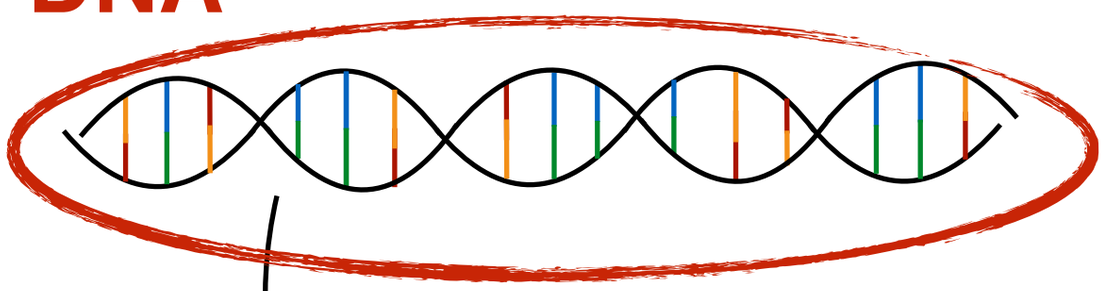




Form networks &
pathways; perform a
vast set of cellular
functions

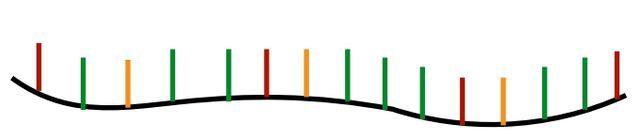
“Flow” of information in the cell

DNA



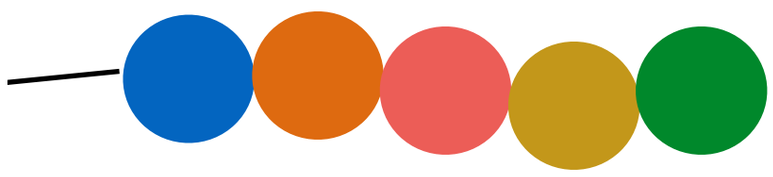
RNA Polymerase
(transcription)

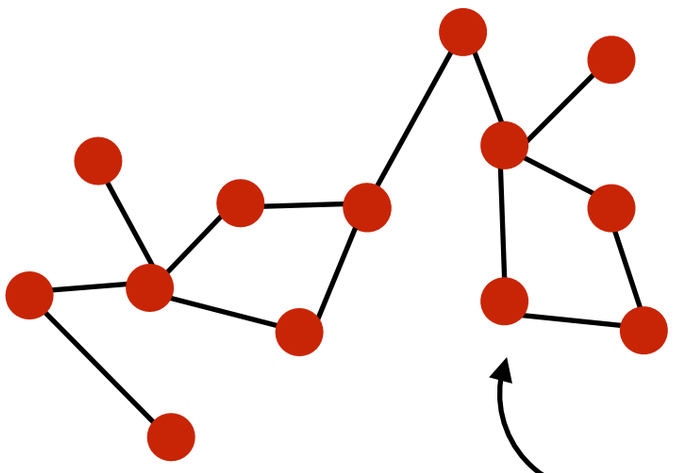
RNA



Ribosomes
(translation)

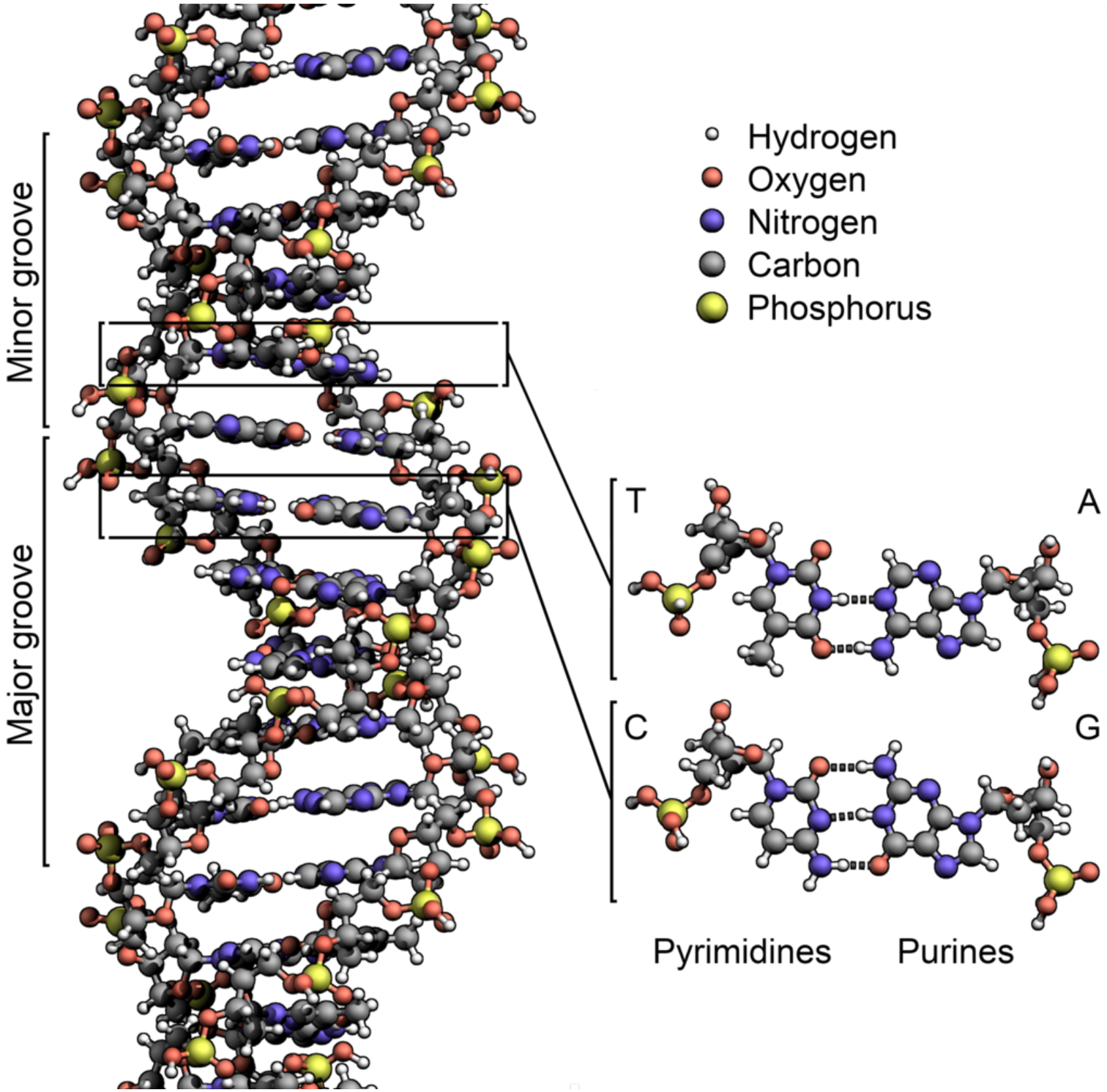
Protein

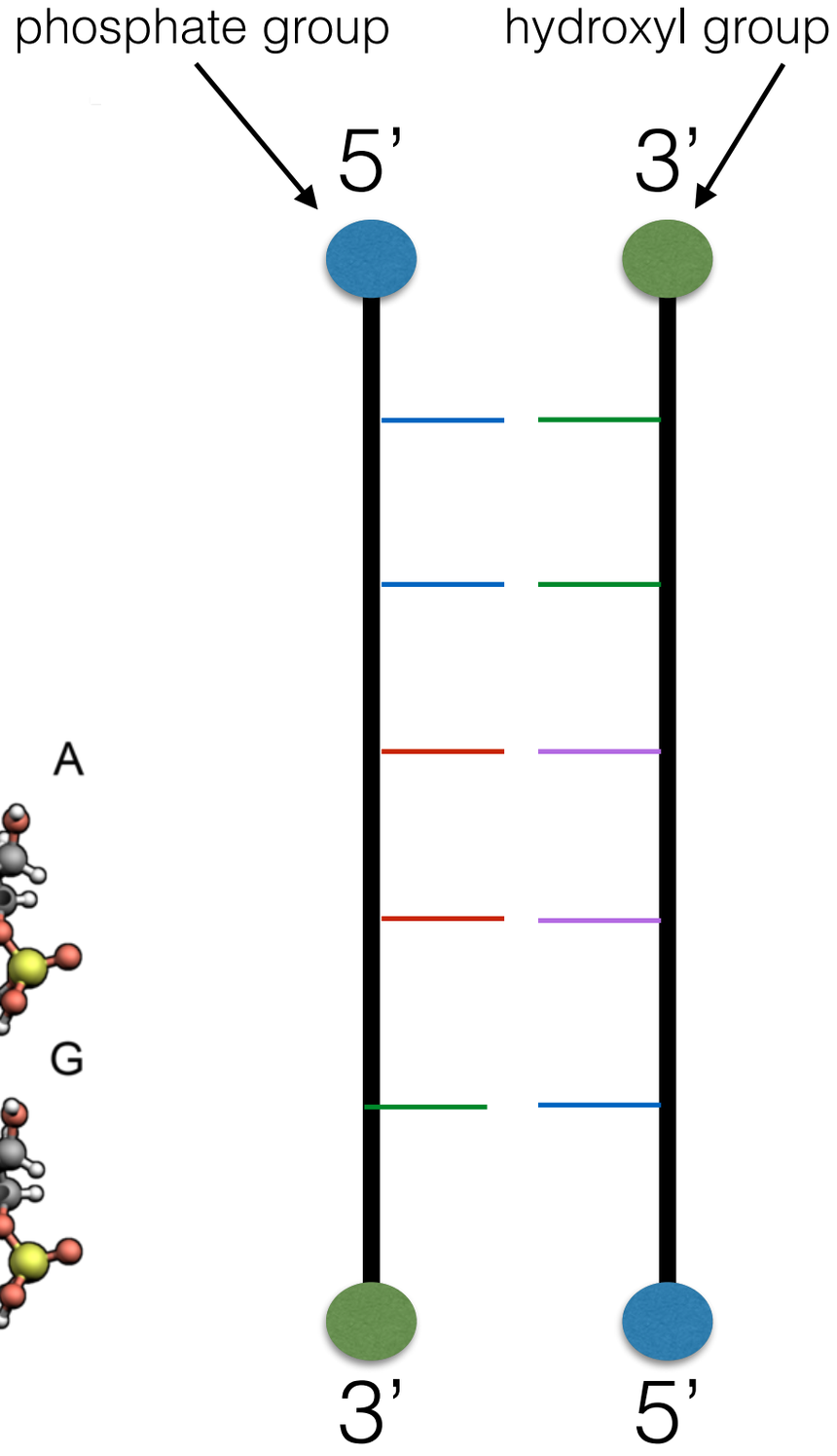




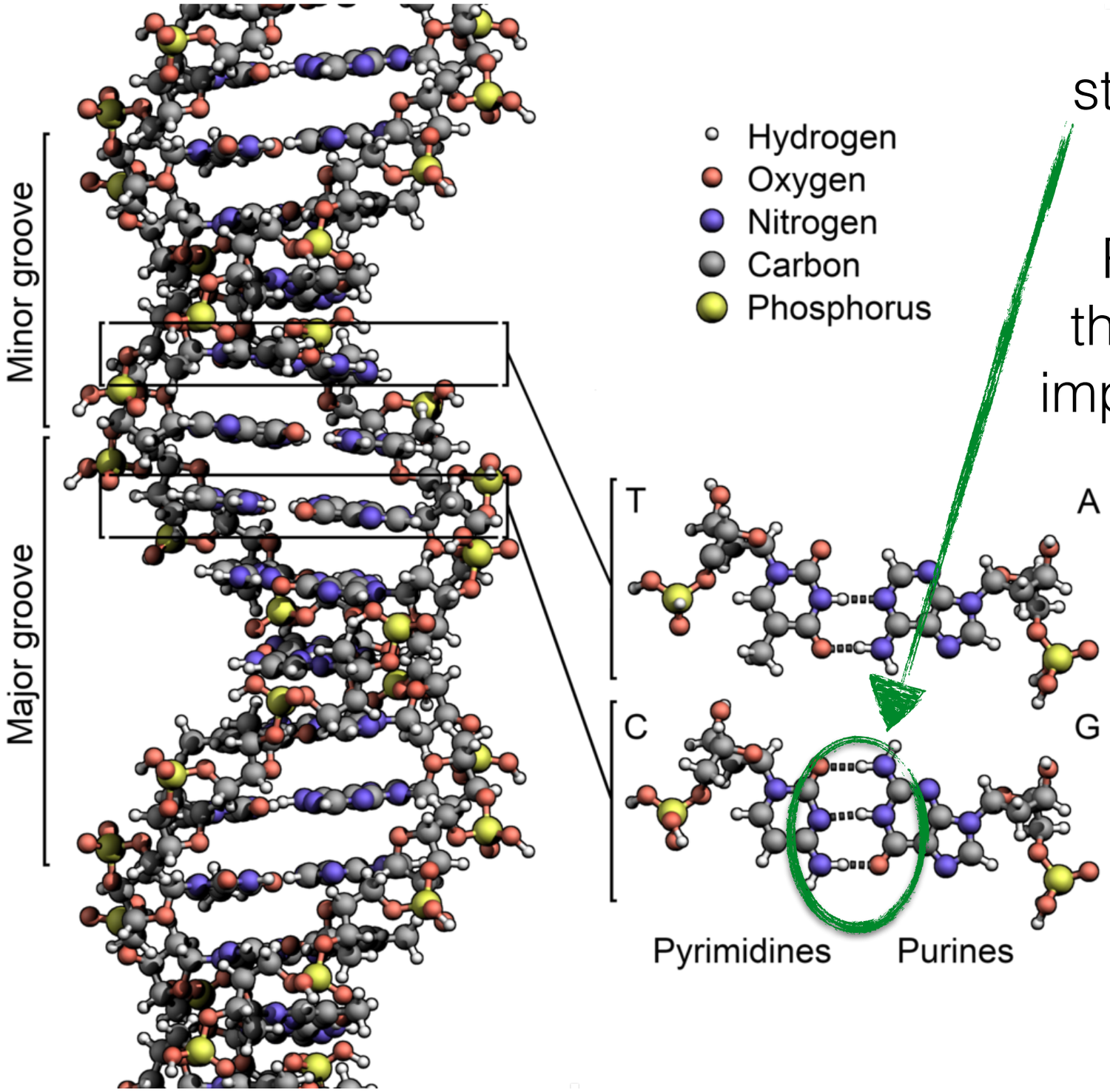
Form networks &
pathways; perform a
vast set of cellular
functions

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”

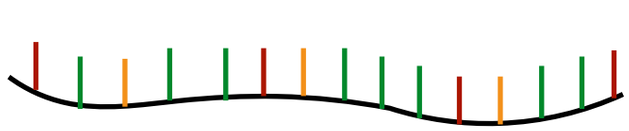
“Flow” of information in the cell

DNA

RNA Polymerase (transcription)

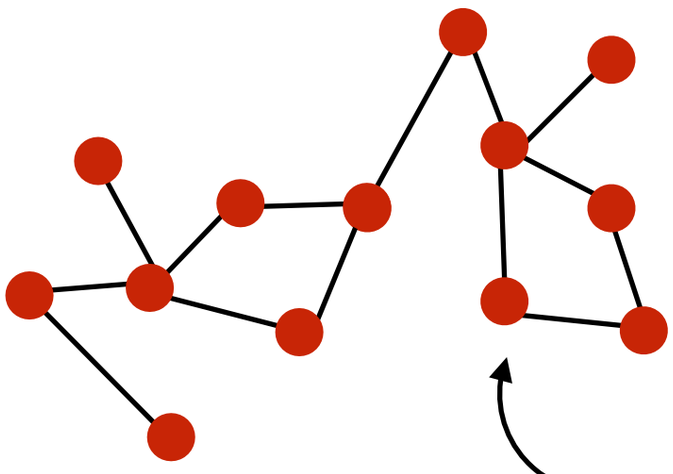
See video on course website

RNA



Ribosomes (translation)

Protein



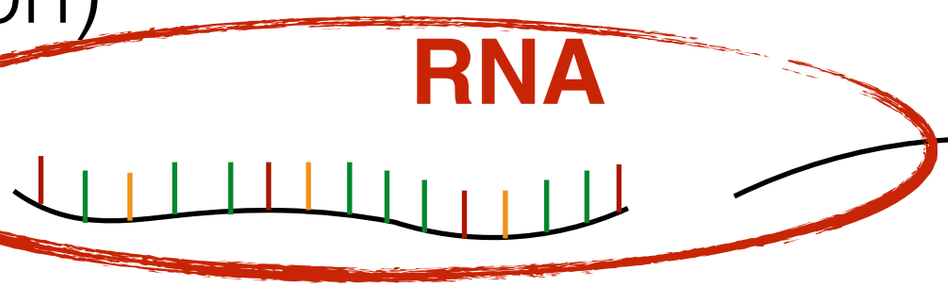
Form networks & pathways; perform a vast set of cellular functions

“Flow” of information in the cell

DNA

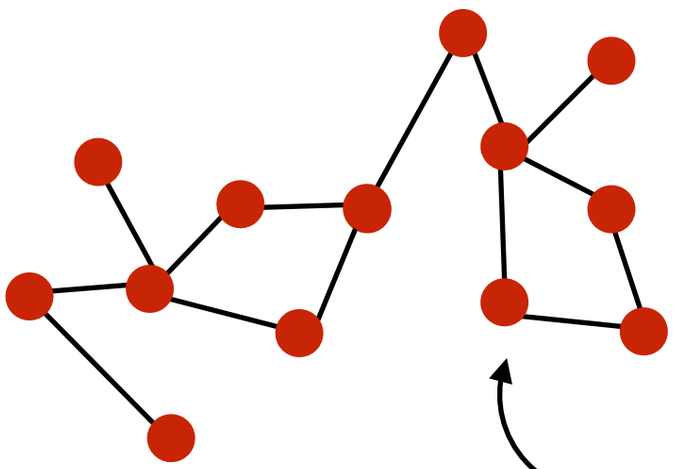
RNA Polymerase
(transcription)

RNA



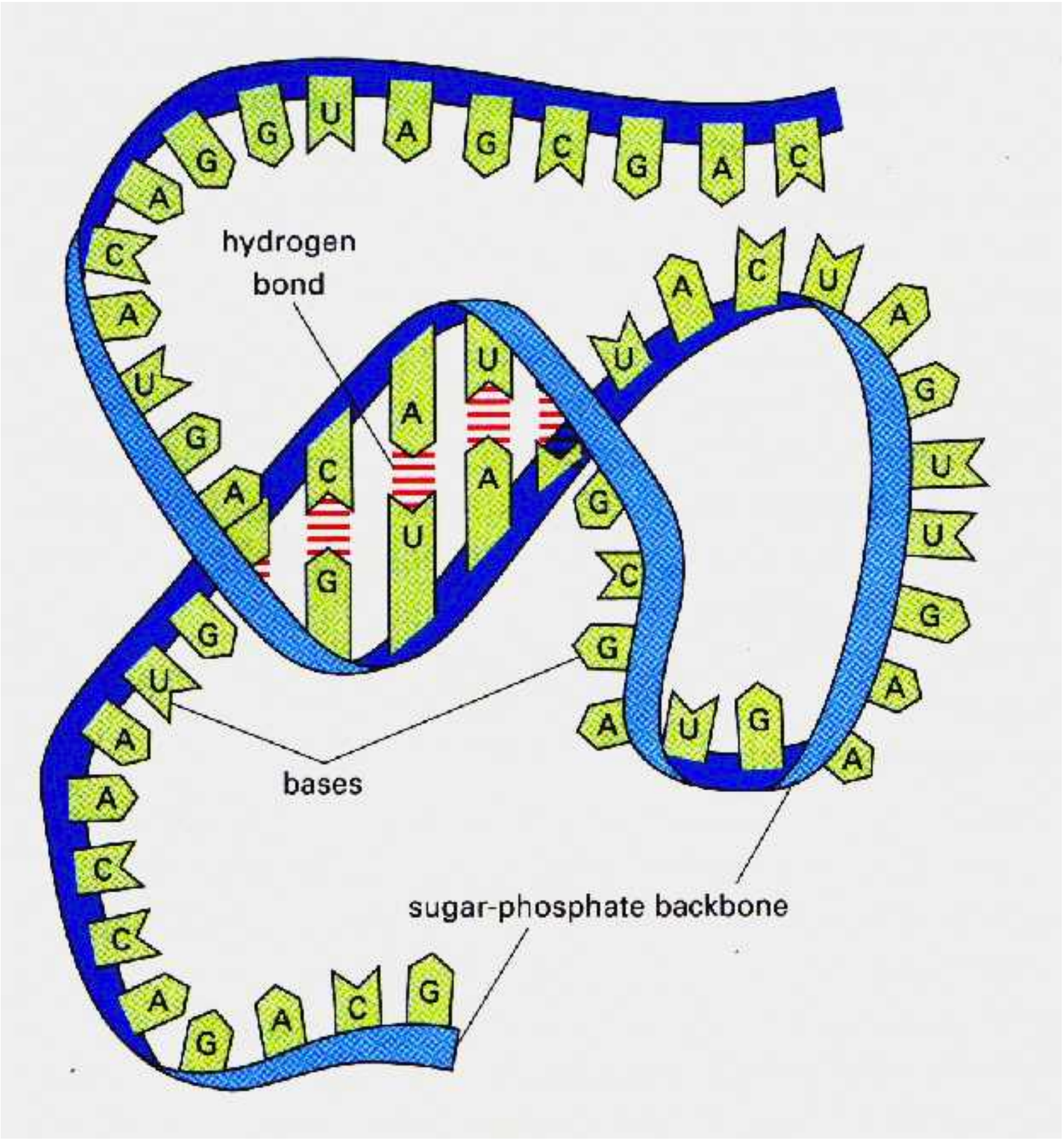
Ribosomes
(translation)

Protein



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

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

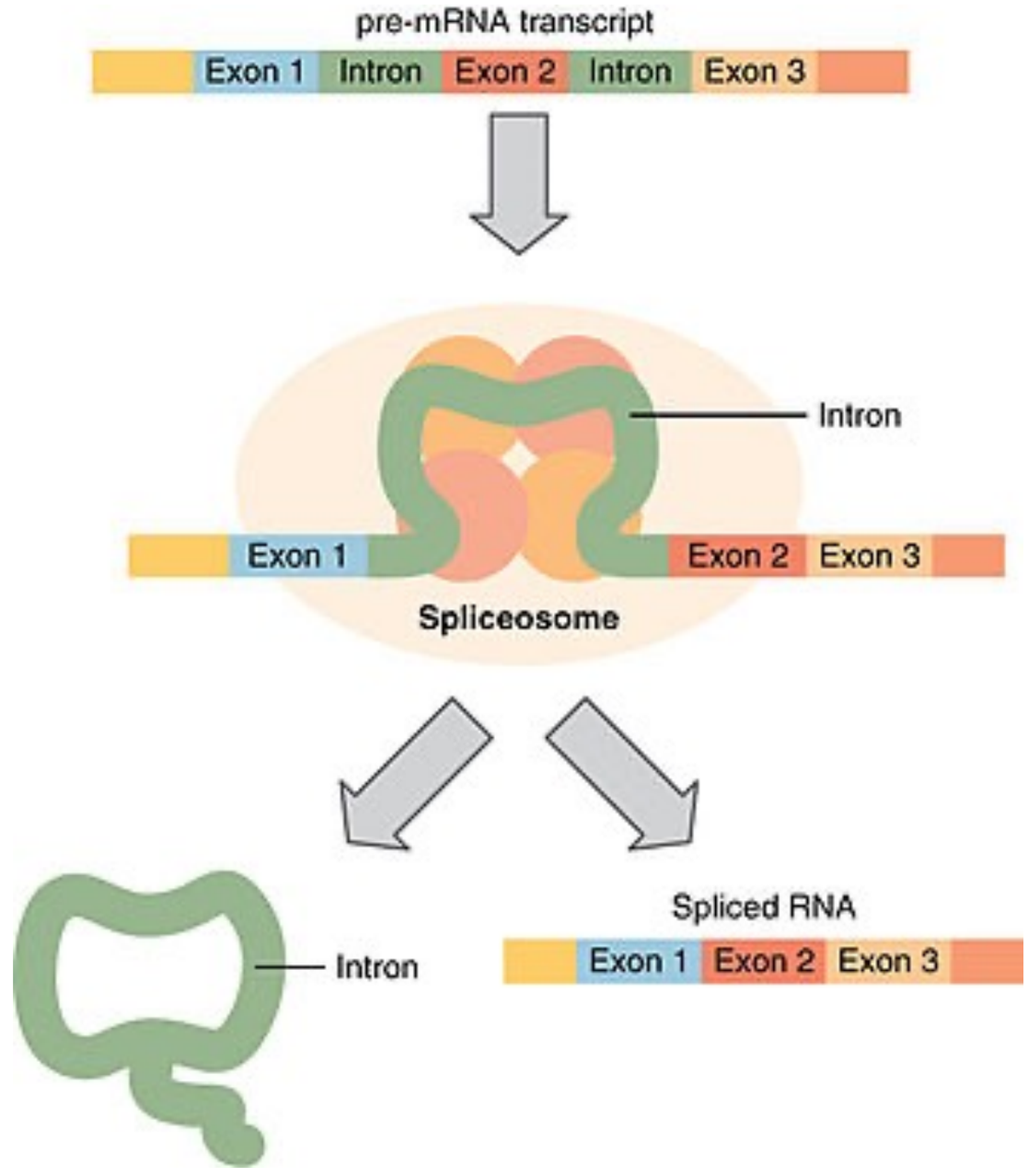
RNA Splicing

DNA transcribed into pre-mRNA

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

Introns removed from pre-mRNA

Introns removed resulting in
mature mRNA

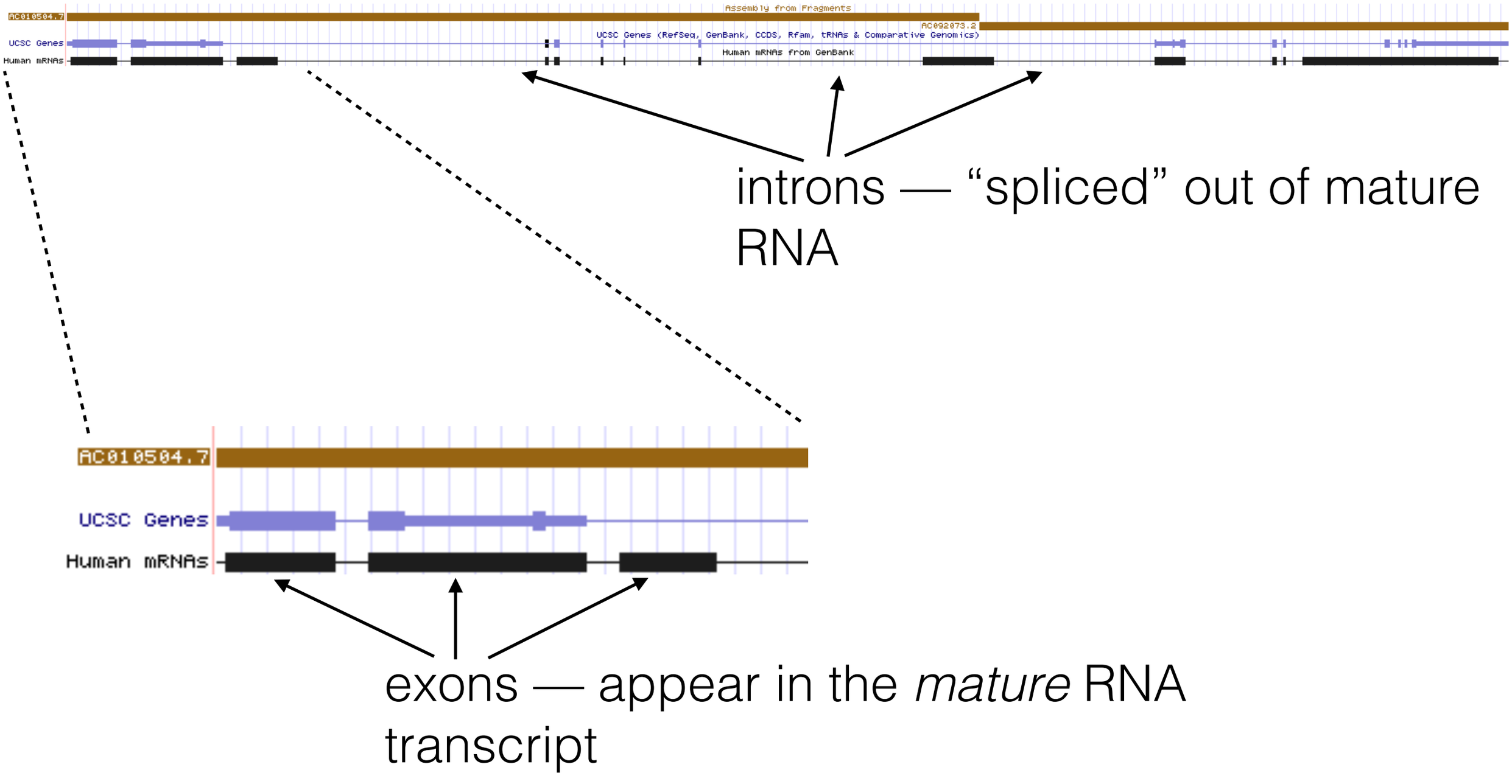


en.wikipedia.org

DNA (the genome)

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

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

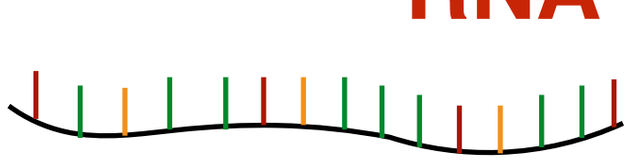


“Flow” of information in the cell

DNA

RNA Polymerase
(transcription)

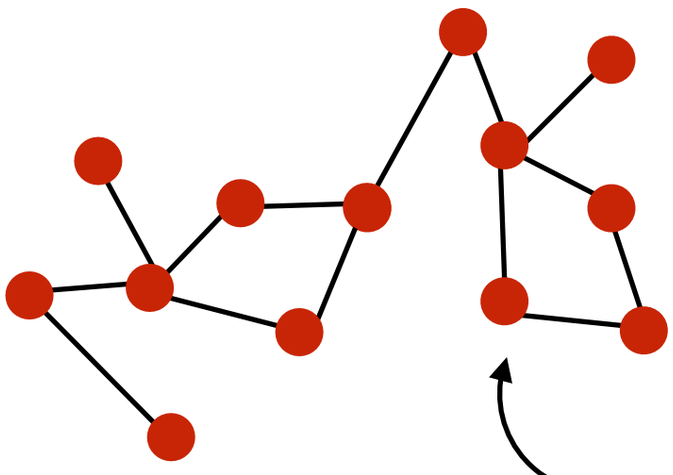
RNA



See video on course website

Ribosomes
(translation)

Protein



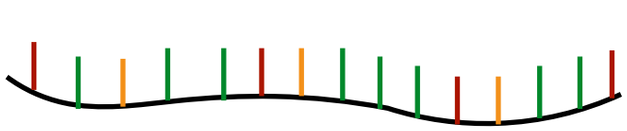
Form networks & pathways; perform a vast set of cellular functions

“Flow” of information in the cell

DNA

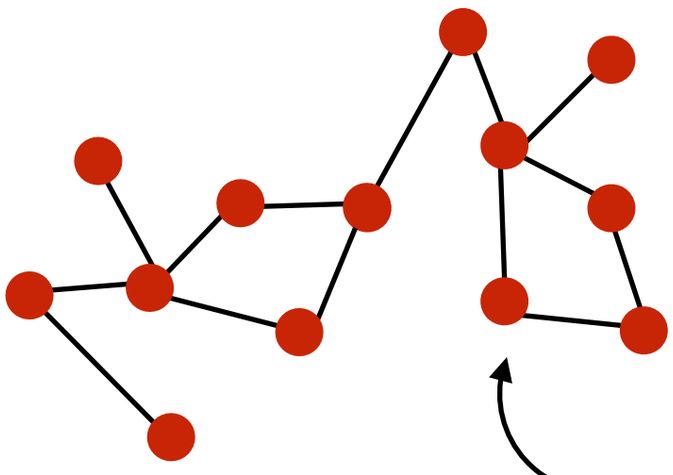
RNA Polymerase
(transcription)

RNA



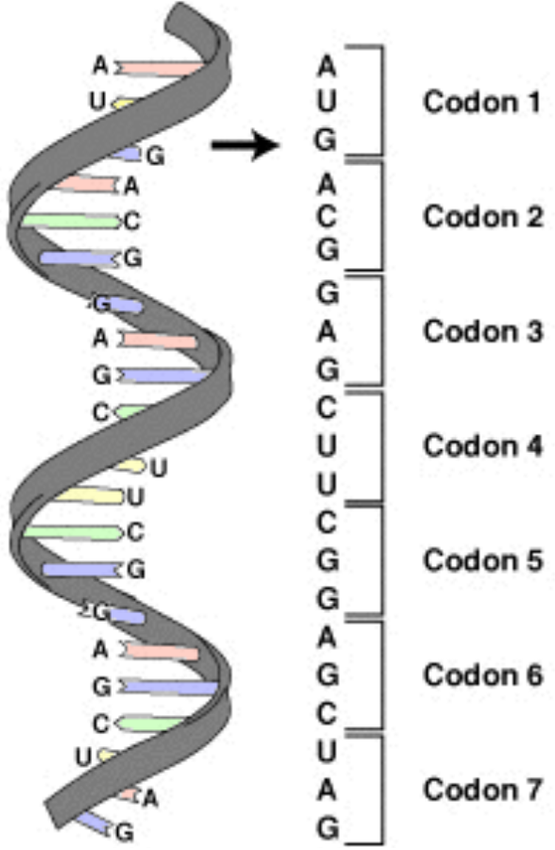
Ribosomes
(translation)

Protein



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

Protein



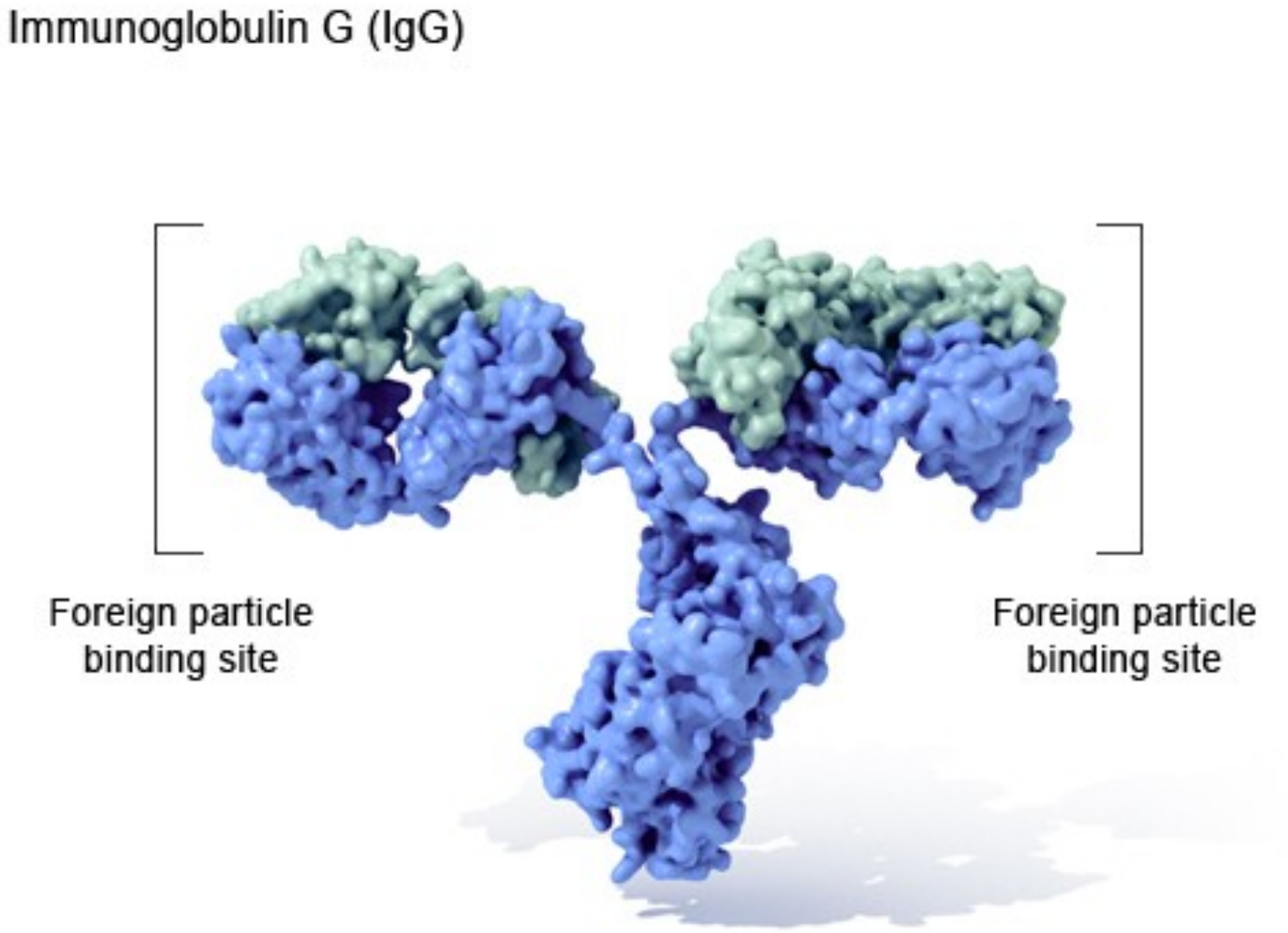
Triplets of mRNA bases (codons) correspond to specific amino acids

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

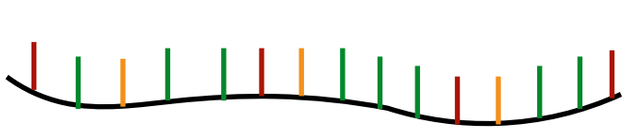
The shape or *conformation* of a protein is intimately tied to its function. Protein shape, therefore, is strongly conserved through evolution — even more so than sequence. A protein can undergo sequence mutations, but fold into the same or a similar shape and still perform the same function.

“Flow” of information in the cell

DNA

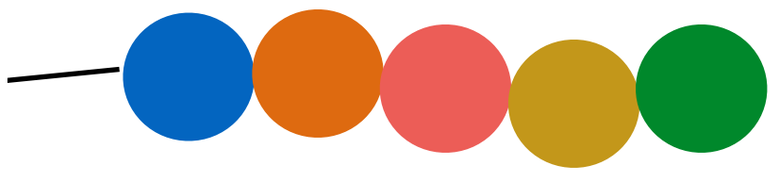
RNA Polymerase
(transcription)

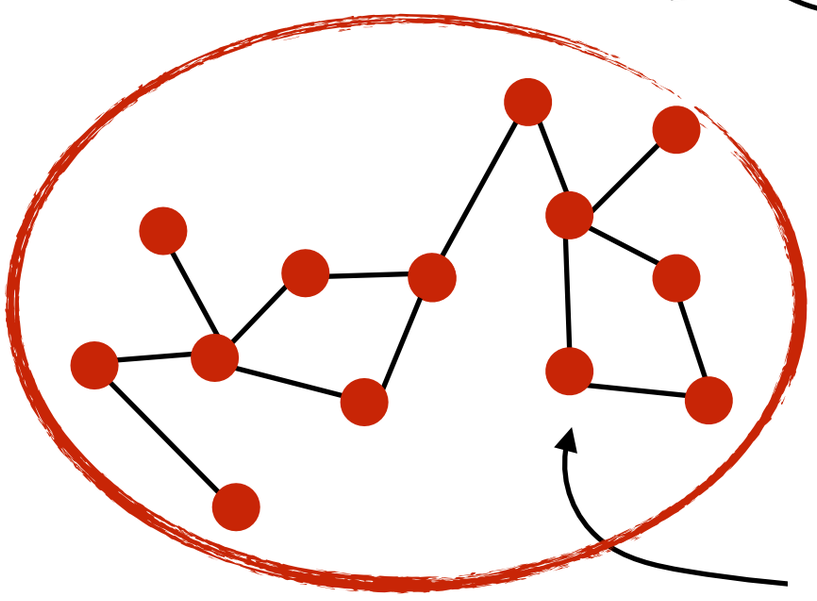
RNA



Ribosomes
(translation)

Protein





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

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

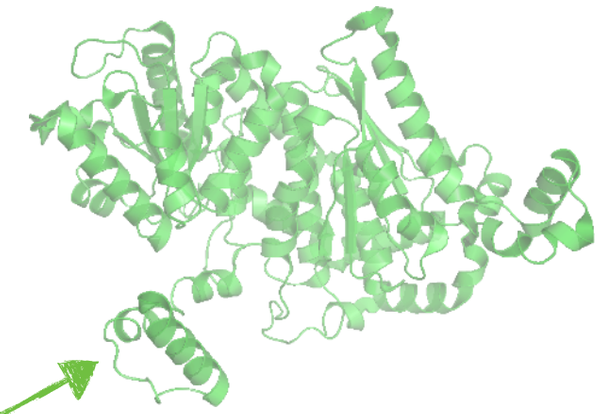
Glycolysis Pathway

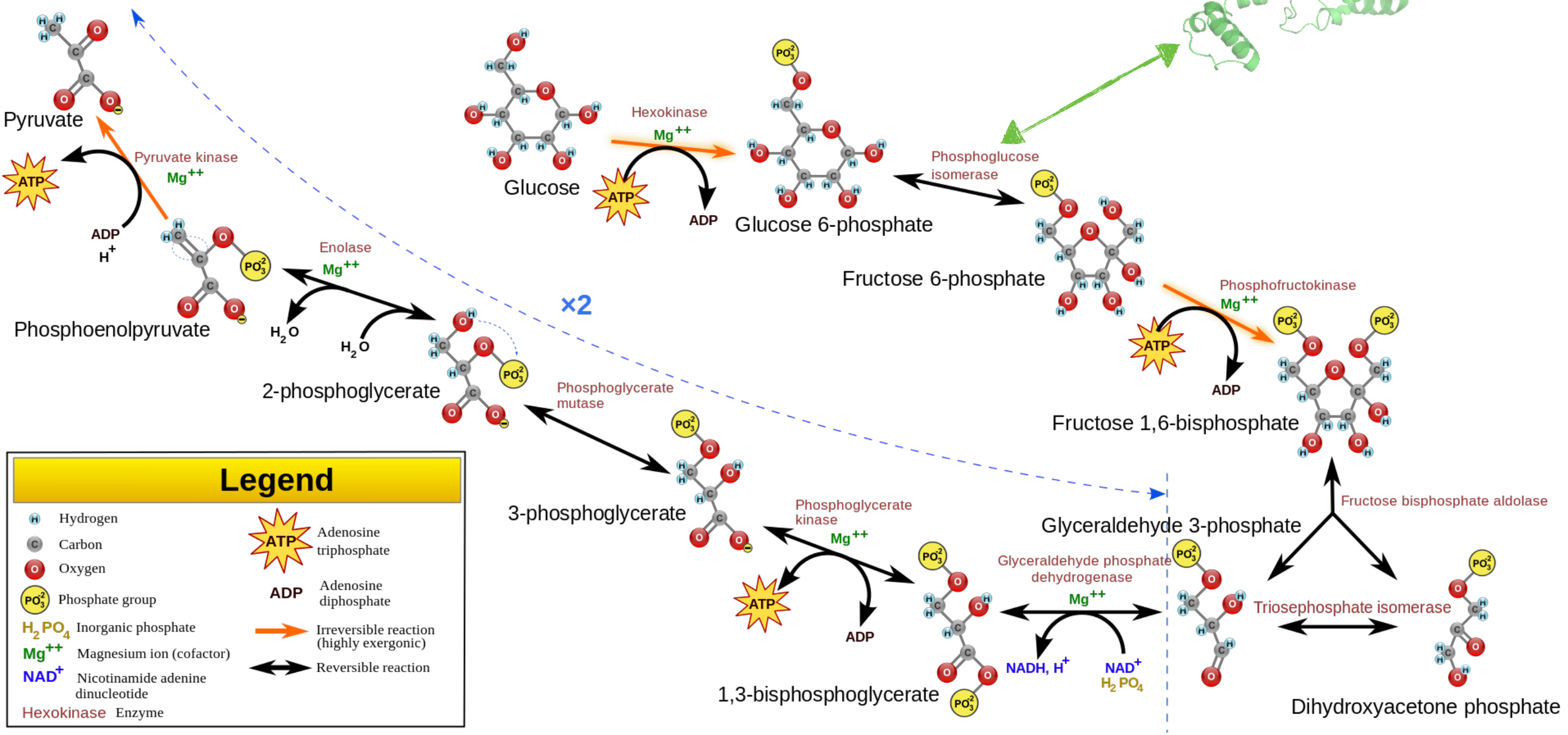
Converts glucose → pyruvate

Generates ATP (“energy currency” of the cell)

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

phosphoglucose isomerase





Some Interesting Facts

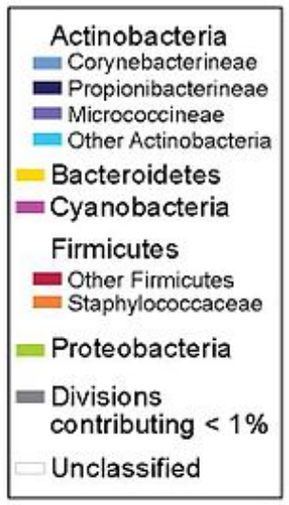
Organism	Genome size	# of genes
ϕX174 (*E. coli* virus)	~5kb	11
E. coli K-12	~4.6Mb	~4,300
Fruit Fly	~122Mb	~17,000
Human	~3.3Gb	~21,000
Mouse	~2.8Gb	~23,000
P. abies (a spruce tree)	~19.6Gb	~28,000

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

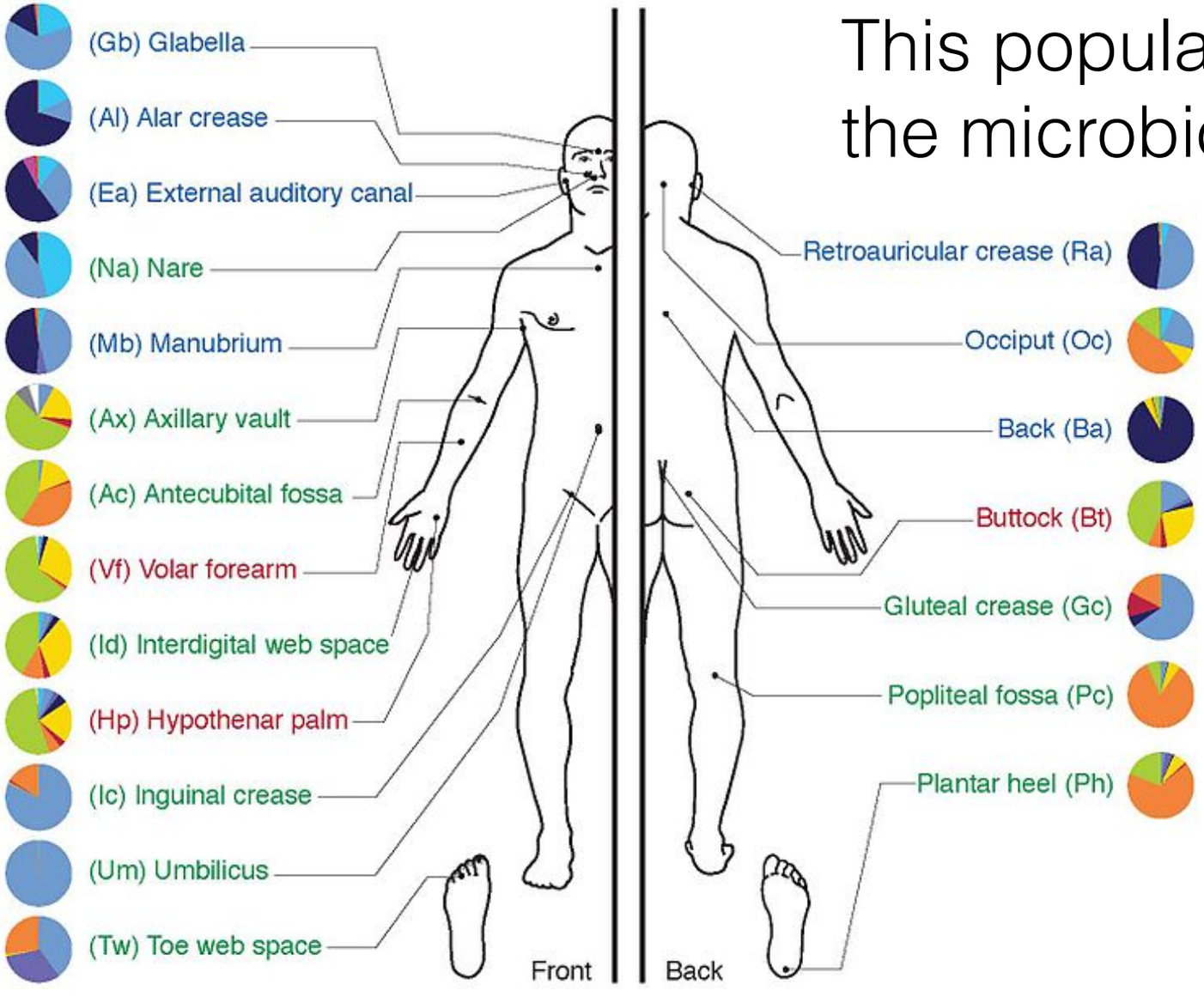
Some Interesting Facts



You are mostly bacteria, fungi & arches

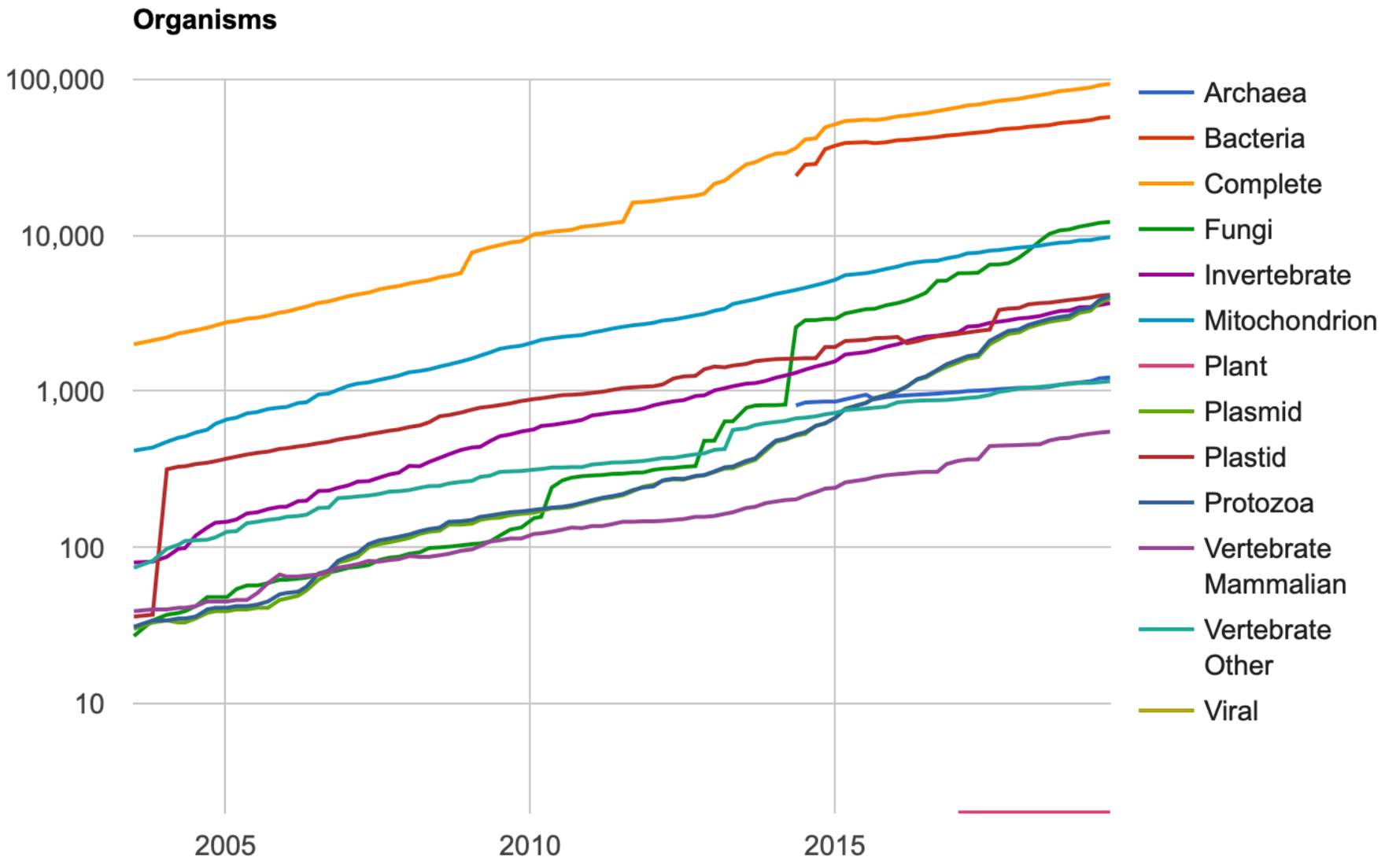
Non-human cells outnumber human cells in the human body

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.