CSE 549:
Computational Biology

Fall 2016

Course Info

Instructor: Rob Patro (rob.patro@cs.stonybrook.edu)

Office: 259 New Computer Science

Office Hours: Tues 1:00 — 3:00 and by appointment

TA: Laraib Malik (laraib.malik@stonybrook.edu)

Office Hours: Wed 3:00 — 5:00 (NCS 159)

Website: www3.cs.stonybrook.edu/~cse549/

DSS: http://studentaffairs.stonybrook.edu/dss/

Academic Integrity: http://www.stonybrook.edu/commcms/academic_integrity/

Coursework & Grading

Coursework and grading: The coursework will consist of two exams, a midterm (whose tentative date is Tues, October 11) and a final (on the University-scheduled date). In addition there will be a few homework assignments (some small programming assignments, and one or two written homeworks that will require you to devise / explain an algorithm, or prove some properly about an algorithm or data structure that we cover in class) and a final course project. The final project can be selected from a list of projects that will be distributed in a few weeks. The project will be done in teams of 3 - 4 students (a team of 2 is OK, but no solo projects and **no teams > 4 students**). For the final project, there will be a brief (7 min) presentation by each group, a deliverable as runnable code, 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:

- Midterm – 30%
- Final – 30%
- Final Project – 20%
- Homeworks – 20%

Academic Integrity maintain it!

Academic integrity: From the University's Academic Integrity Syllabus Statement:

> *Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Any suspected instance of academic dishonesty will be reported to the Academic Judiciary. For more comprehensive information on academic integrity, including categories of academic dishonesty, please refer to the academic judiciary website at www.stonybrook.edu/academicintegrity.*

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 (i.e. this is just as bad as copying someone's answer on an exam or code on a homework). 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. I have a lot of tolerance for those who are making a significant effort but may be having trouble understanding a particular concept or completing a certain assignment. However, there will be no tolerance of academic dishonesty. **If you are academically dishonest in this course, you will recieve a grade of F, and you will be reported to the department's academic integrity committee**.

Textbooks

Required

- **Bioinformatics Algorithms: An Active Learning Approach Volume I** (Compeau and Pevzner 2015)
- **Bioinformatics Algorithms: An Active Learning Approach Volume II** (Compeau and Pevzner 2015)

Other great resources

- **Biological Sequence Analysis** (Durbin, Eddy, Krogh, Mitchinson 1998)
- **Genome Scale Algorithm Design** (Mäkinen, Belazzougui, Cunial, Tomescu 2015)
- **Molecular Biology of the Cell, by Bruce Alberts*** (Alberts et al. 2002)
- **Molecular Biology**(Clark and Pazdernik 2012)

* free

Textbooks

CS

Algorithms* (Dasgupta, Papadimitriou, and Vazirani 2006)

Algorithm Design (Kleinberg and Tardos 2006)

The Algorithm Design Manual (Skiena 2008).

* free

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**

2014

Today, s*equencing* a genome costs ~$1,000✢ and a "run" takes <3 days✢

~18 Tb per "run" at maximum capacity

✢ on an Illumina HiSeq X Ten — the machine costs ~$10M and sample prep takes a little extra time.

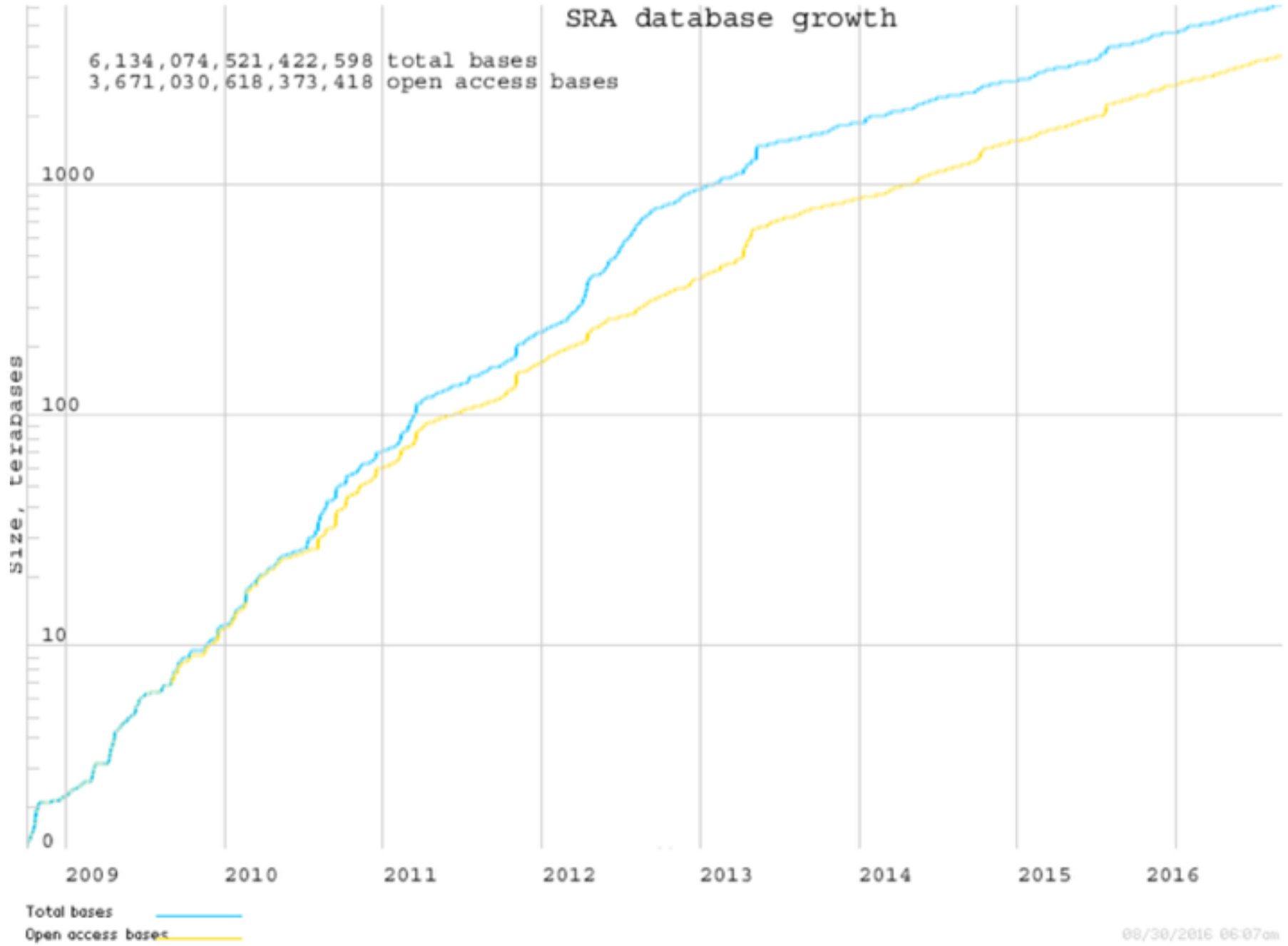
Tons of Data, but we need Knowledge

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 1,000,000s 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.

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

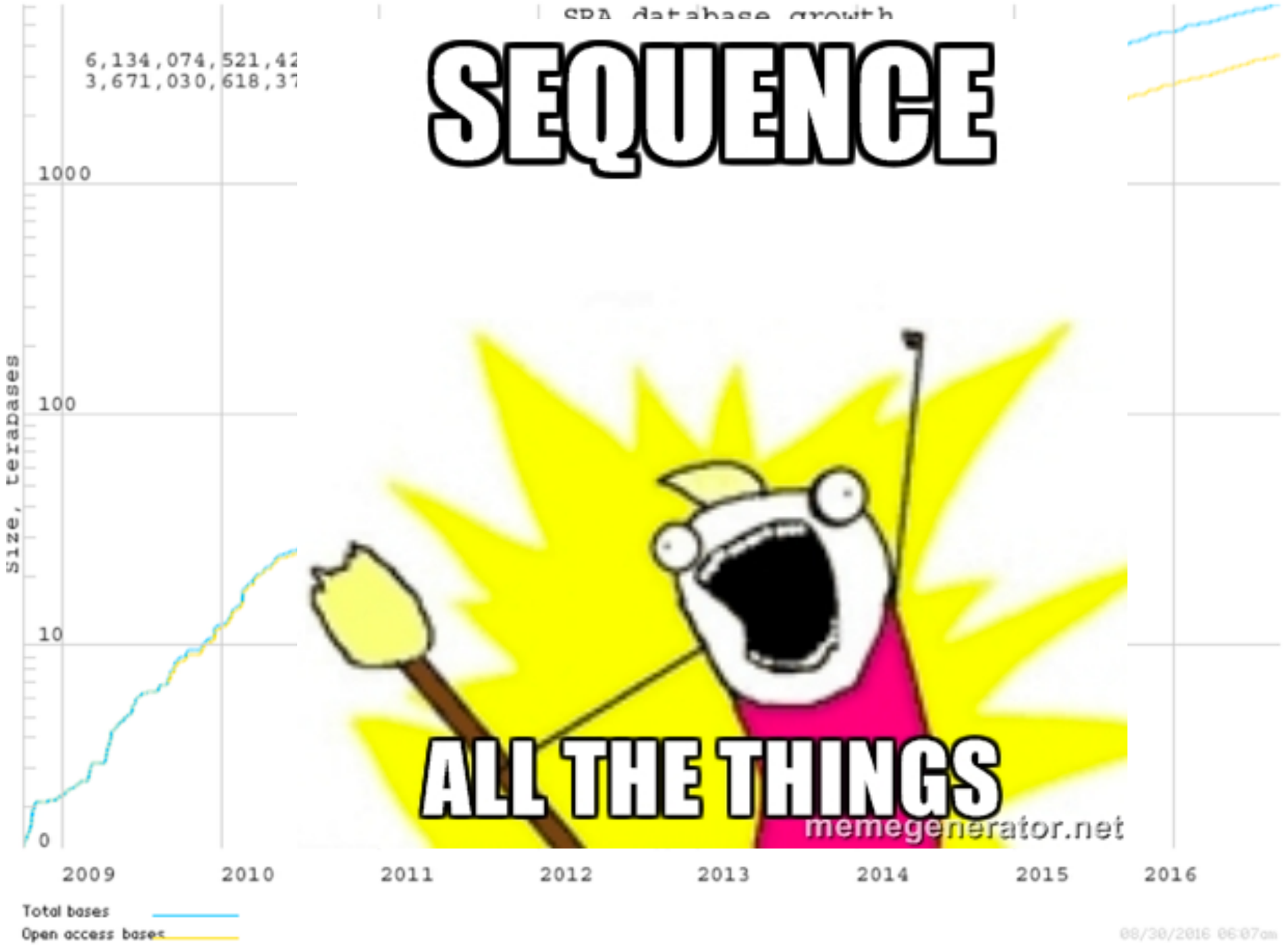
Growth of the Sequence Read Archive (SRA)



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

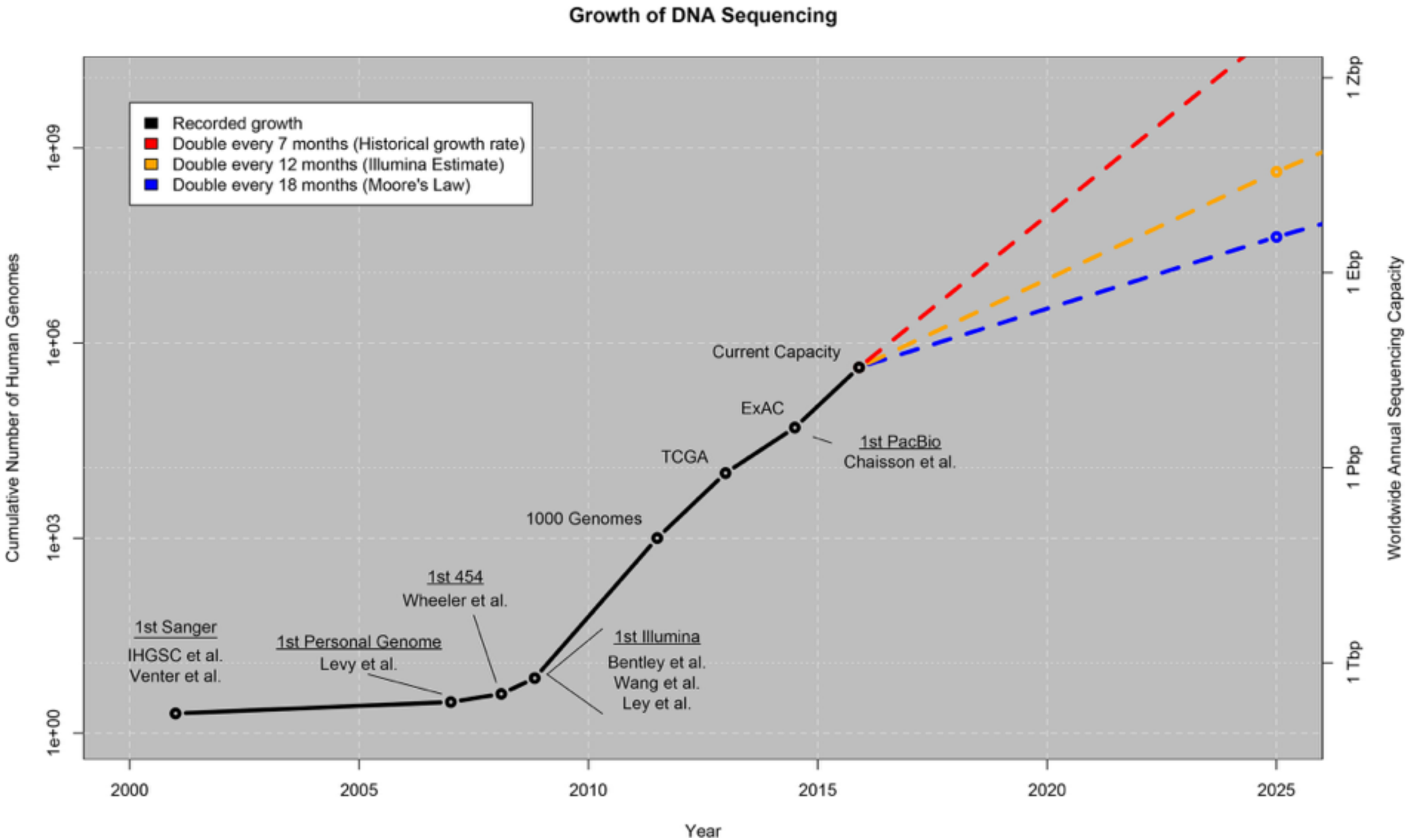
As a result, scientists have taken a very subtle and nuanced approach . . .

Growth of the Sequence Read Archive (SRA)



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

Growth becomes its own problem



Stephens, Zachary D., et al. "Big Data: Astronomical or Genomical?." PLoS Biol 13.7 (2015): e1002195.

Answer questions “in the large”

What is the genome of the beaver (state animal of NY)? (genomics)

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

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

How do environment changes affect the microbial ecosystem of the Long Island sound? (metagenomics)

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

The Computational Part

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)

SB is a great place for Comp Bio

Location, Location, Location:
~20 min from Brookhaven
~40 min from CSHL
~1.5 hours from NY Genome Center

This course

Broad survey of some main areas of computational biology:

Genomics

Genome assembly

Search:
Homology detection
Read mapping
BWT, suffix arrays etc.

Gene finding (HMMs)

Transcriptomics (RNA-seq)

Motif finding (Gibbs sampling, statistical inference)

Phylogenetics

Character inference

Tree building

Current Topics

Network analysis / alignment

Genome folding & structure ({3,4,5Hi}-C)

Metagenomics

CS Topics

Many techniques broadly applicable in CS:

Dynamic Programming

String search & indexing (full-text indices):

- Suffix trees / arrays
- Burrows-Wheeler transform & FM-Index

Discrete Optimization & Network Analysis

Statistical Inference (frequentist & Bayesian)

Hidden Markov Models

Next ~2 Lectures

How Biology and CS differ as fields

Biology for Computer Scientists

- Some fundamentals about molecular Biology
- Basics of sequencing techniques and experiments

Computer Science for Biologists

- How CS differs from Biology
- Some fundamentals notions about Computer Science

“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

Only journals matter

Large labs:
PI → postdoc→grad students

Student may study a specific gene for their entire PhD

Focus on being “right” and discovering something interesting about the natural world. (focus on knowledge)

CS

Selective conferences often preferred to journals

Smaller labs:
PI → a few grad students

Students typically work on a wide variety of projects in PhD

More weight given to being “different”. Need not be 1st often just be “best”, fastest or simplest. (focus on methods)

Immense Spatial & Time Scales

The scale, in both space and time, of the Biological systems we're interested in studying are **truly expansive**.

Time:

Protein folding can happen on the order of microseconds

Evolution works over the span of hundreds, thousands and tens of thousands of years

Space:

A cell nucleus is measured in micrometers

Population migrations happen over tens of thousands of miles

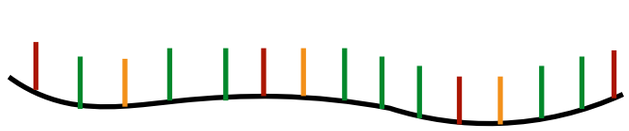
Computational Biology encompasses the study of all of these problems.

“Flow” of information in the cell

DNA

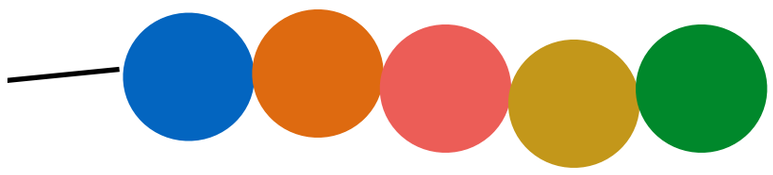
RNA Polymerase
(transcription)

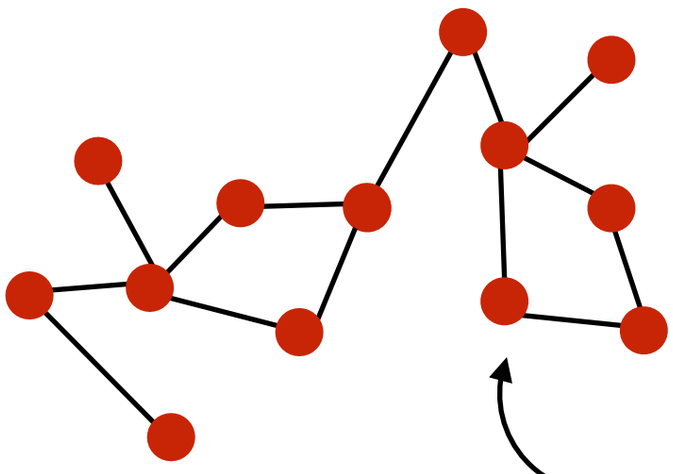
RNA



Ribosomes
(translation)

Protein

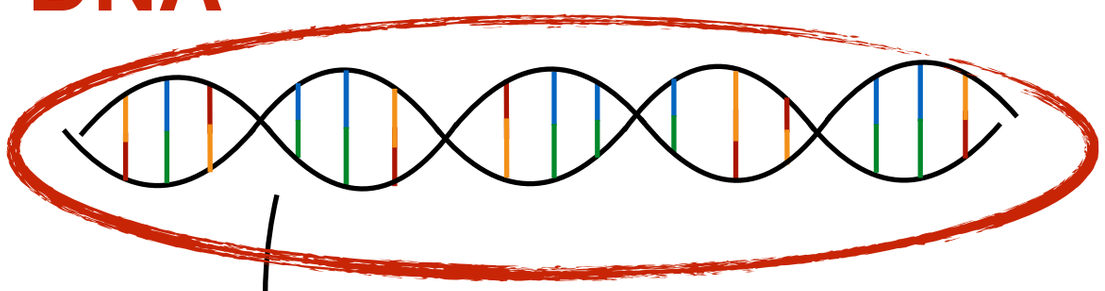




Form networks &
pathways; perform a
vast set of cellular
functions

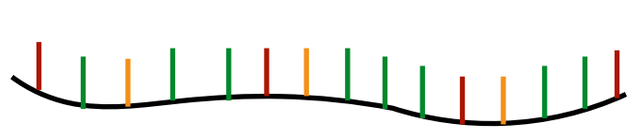
“Flow” of information in the cell

DNA



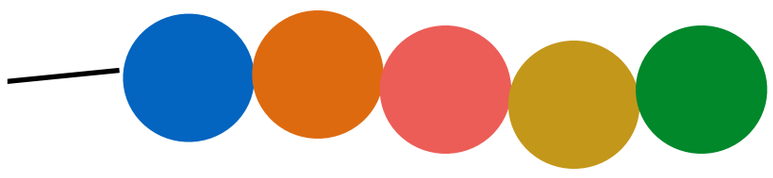
RNA Polymerase
(transcription)

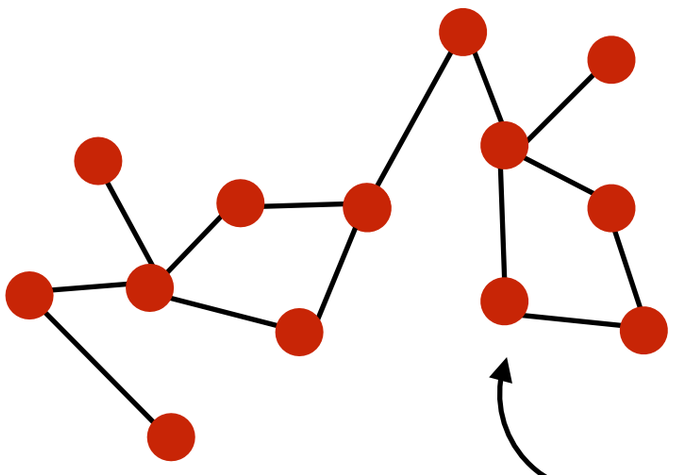
RNA



Ribosomes
(translation)

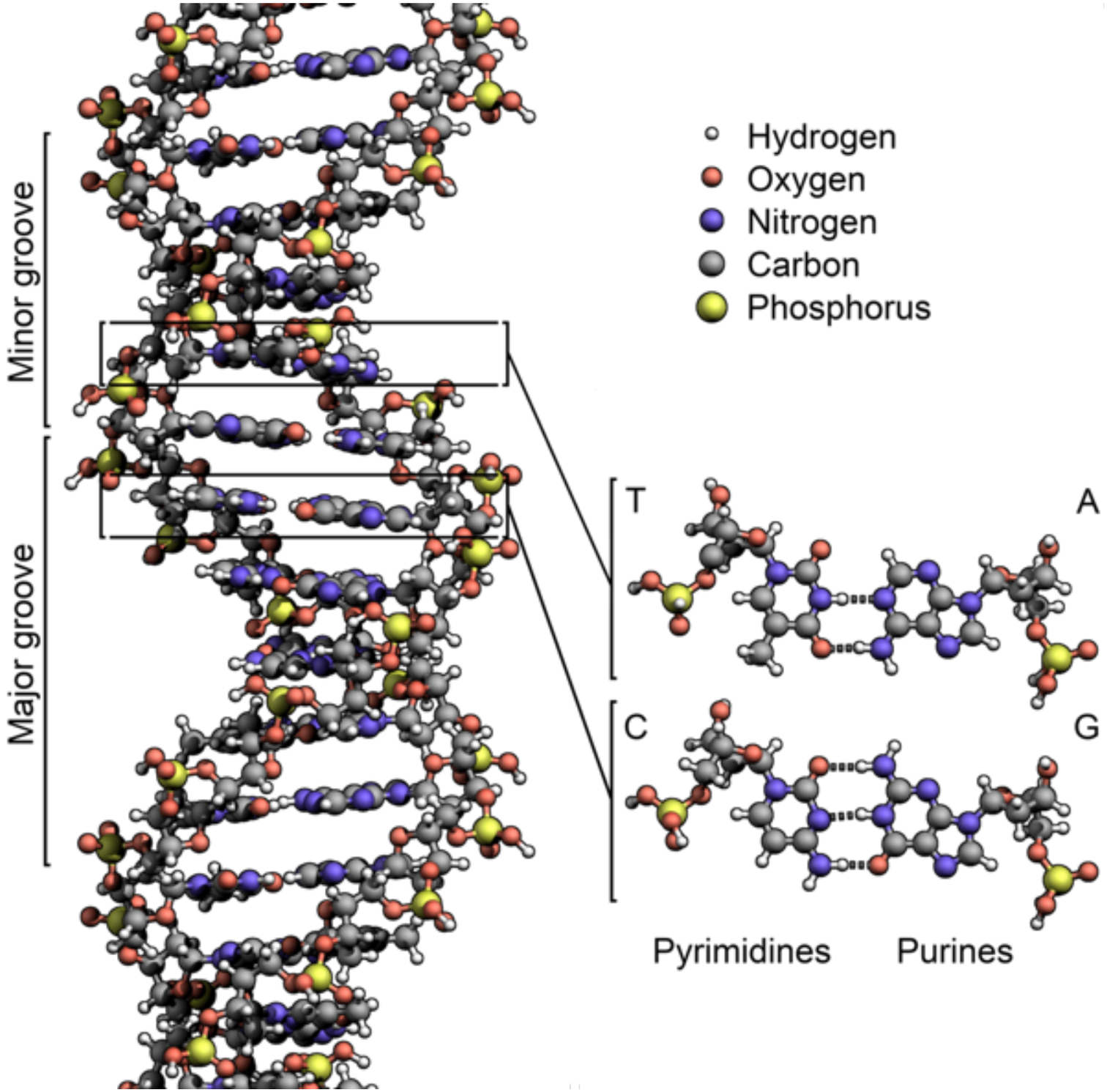
Protein

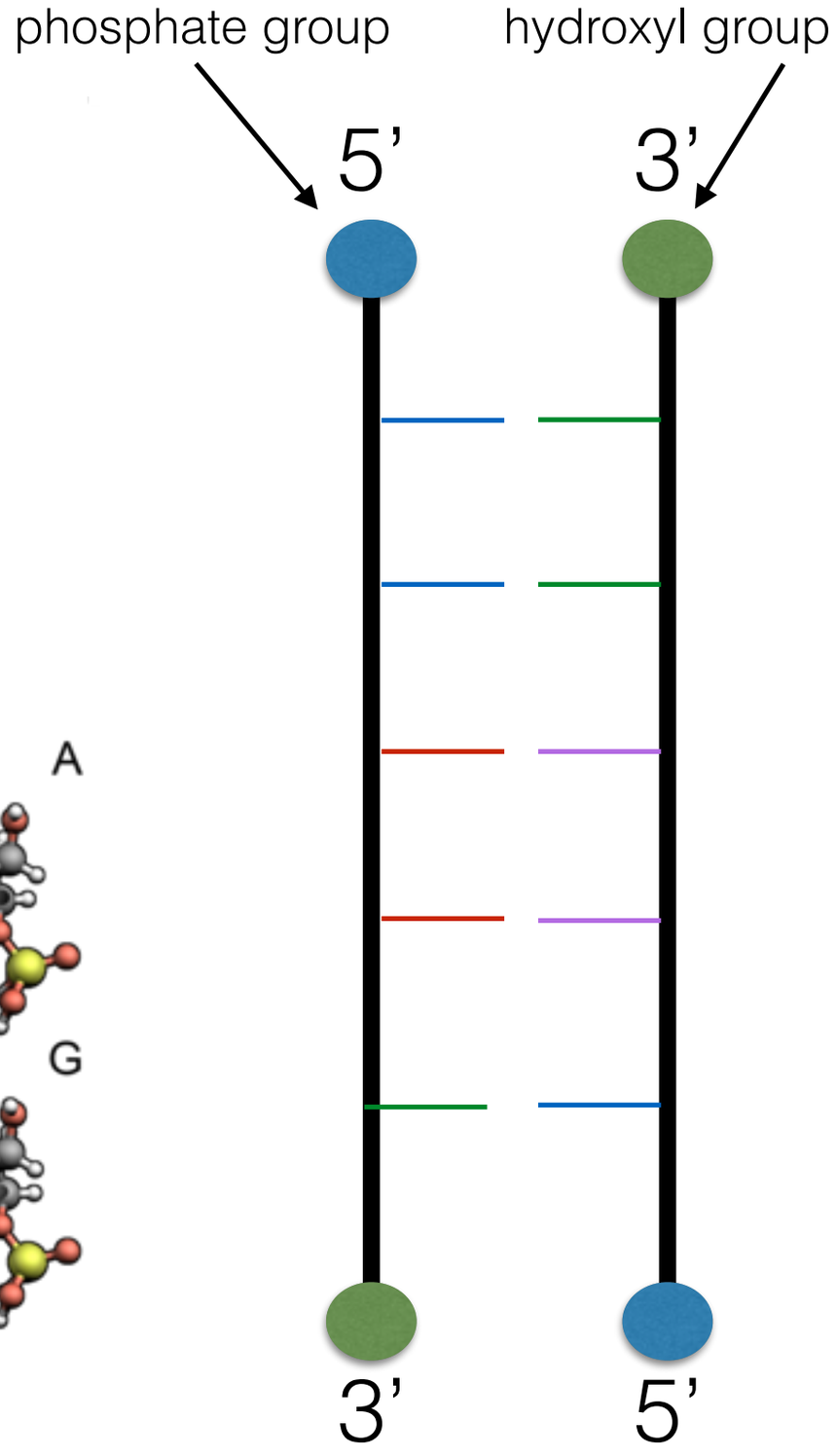




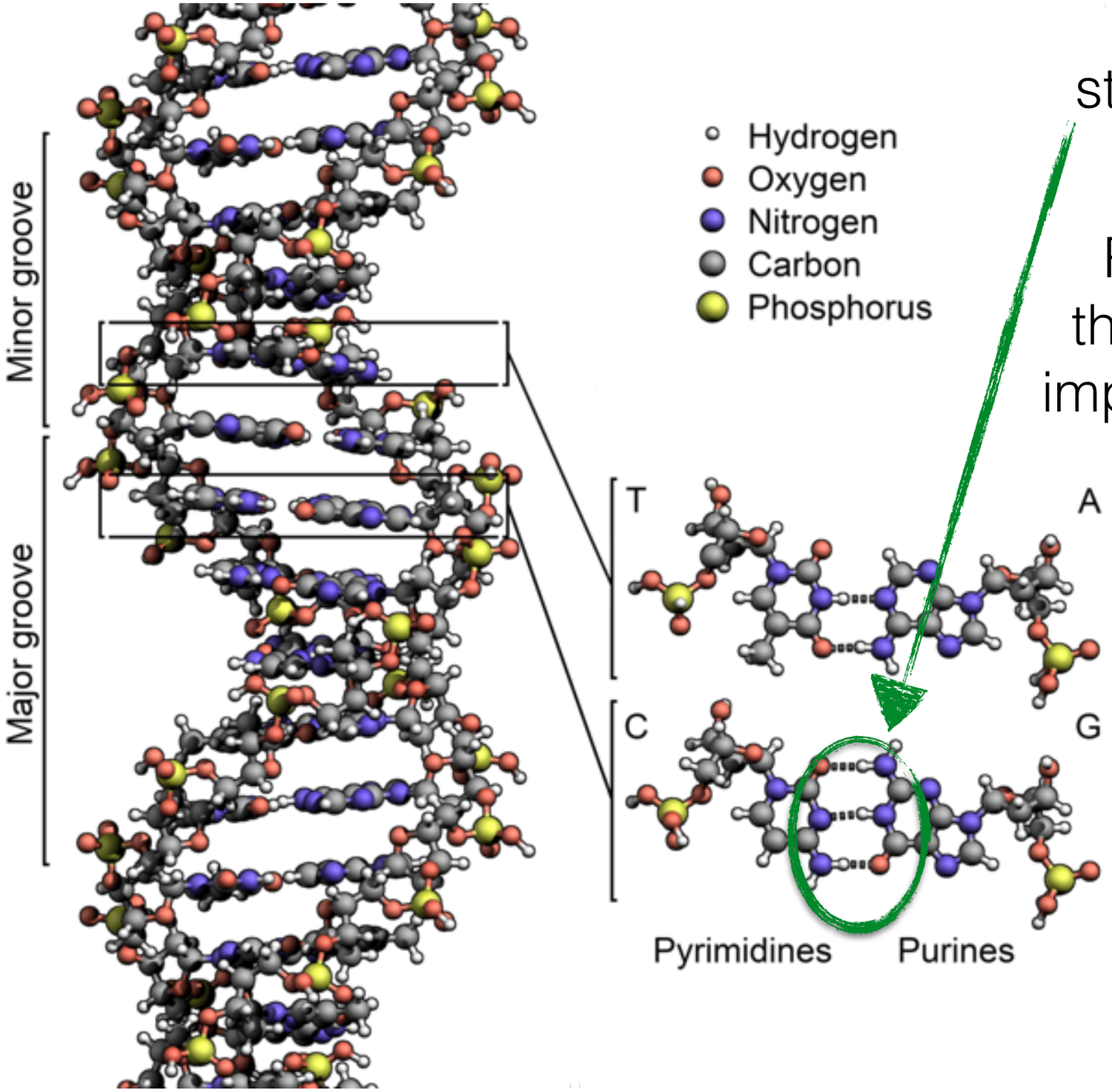
Form networks &
pathways; perform a
vast set of cellular
functions

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%

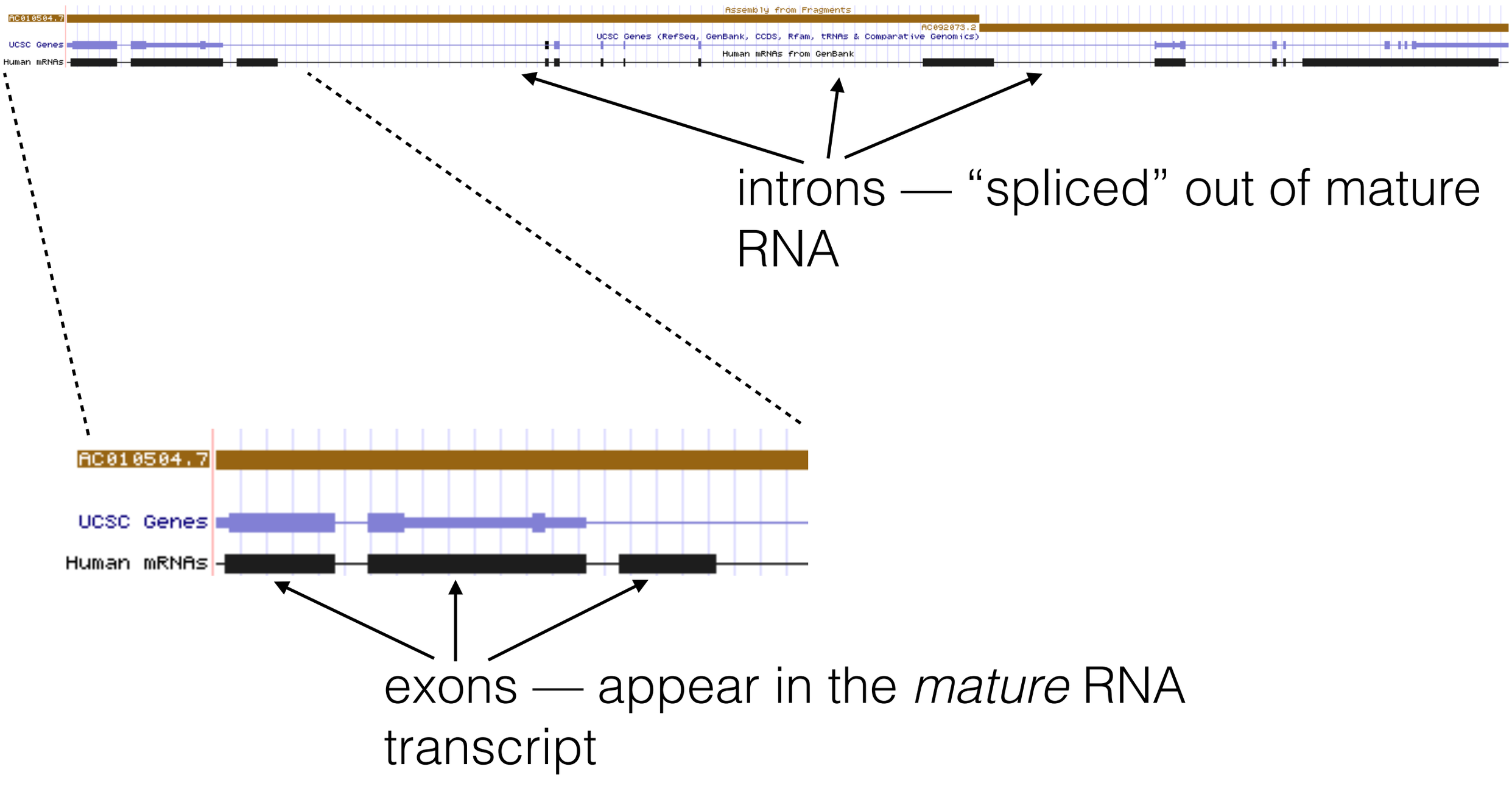
In typical bacterial genome, only small fraction — ~2% — of DNA is “non-coding”

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

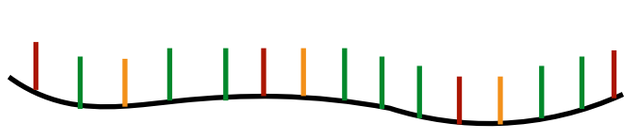


“Flow” of information in the cell

DNA

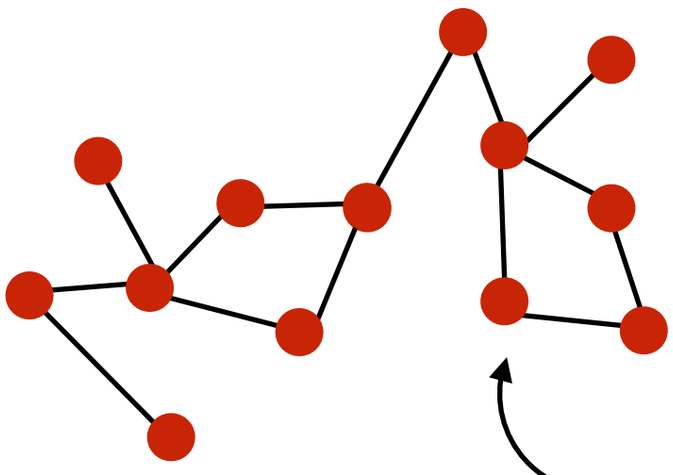
RNA Polymerase
(transcription)

RNA



Ribosomes
(translation)

Protein



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

DNA → RNA : Transcription

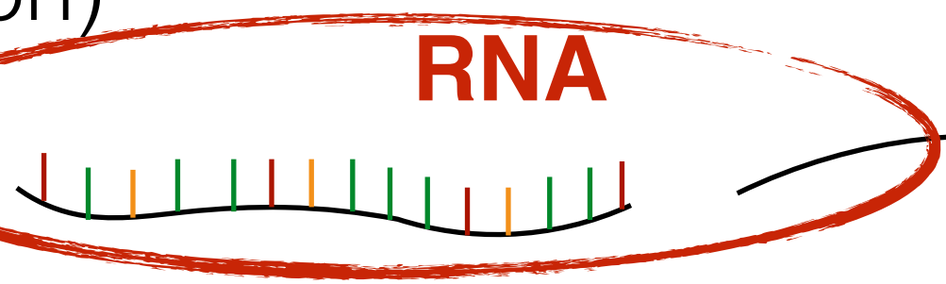
https://www.dnalc.org/resources/3d/12-transcription-basic.html

“Flow” of information in the cell

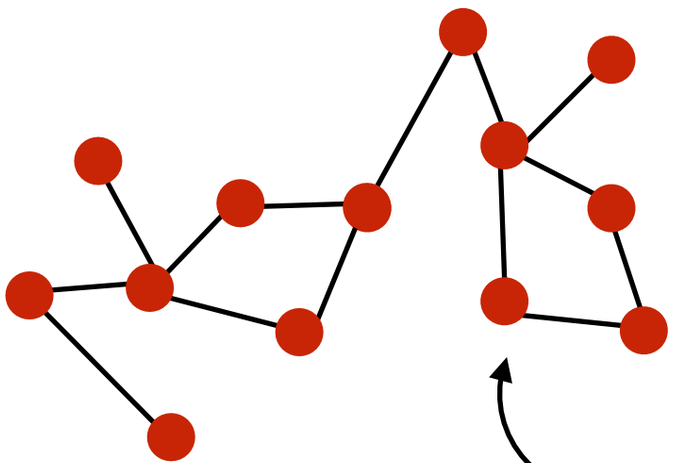
DNA

RNA Polymerase
(transcription)

RNA



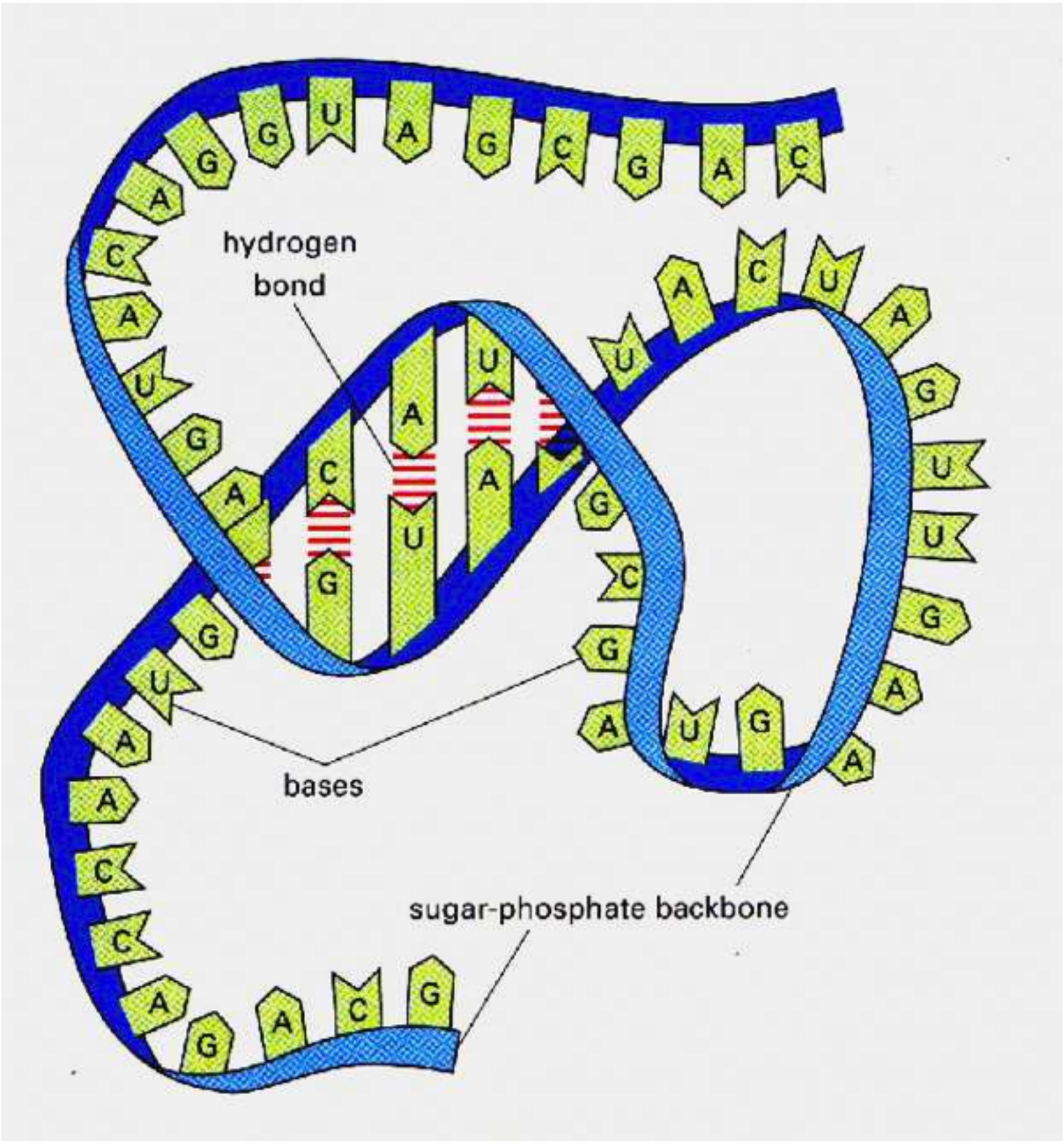
Ribosomes
(translation)



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

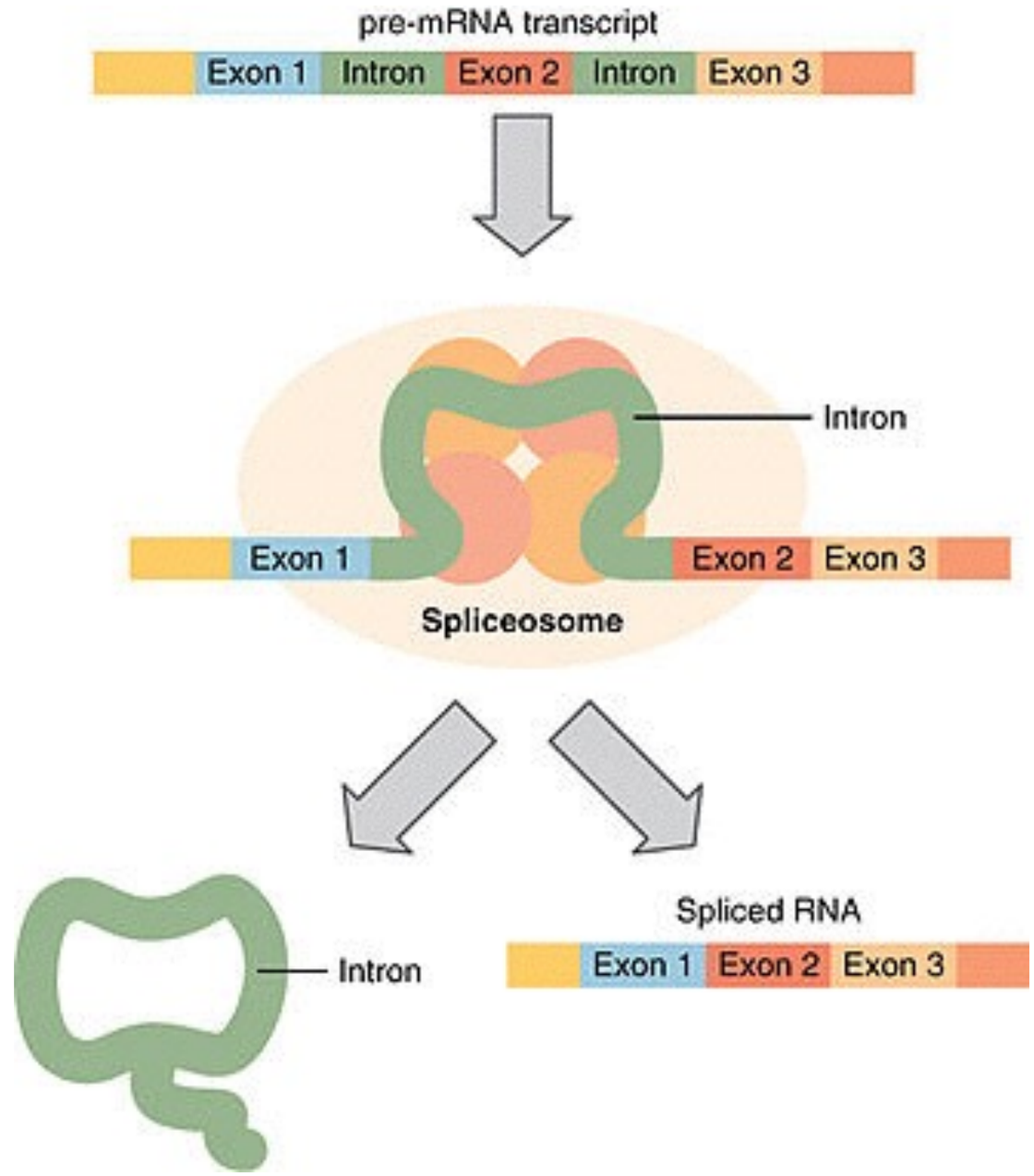
RNA Splicing

DNA transcribed into pre-mRNA

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

Introns removed from pre-mRNA

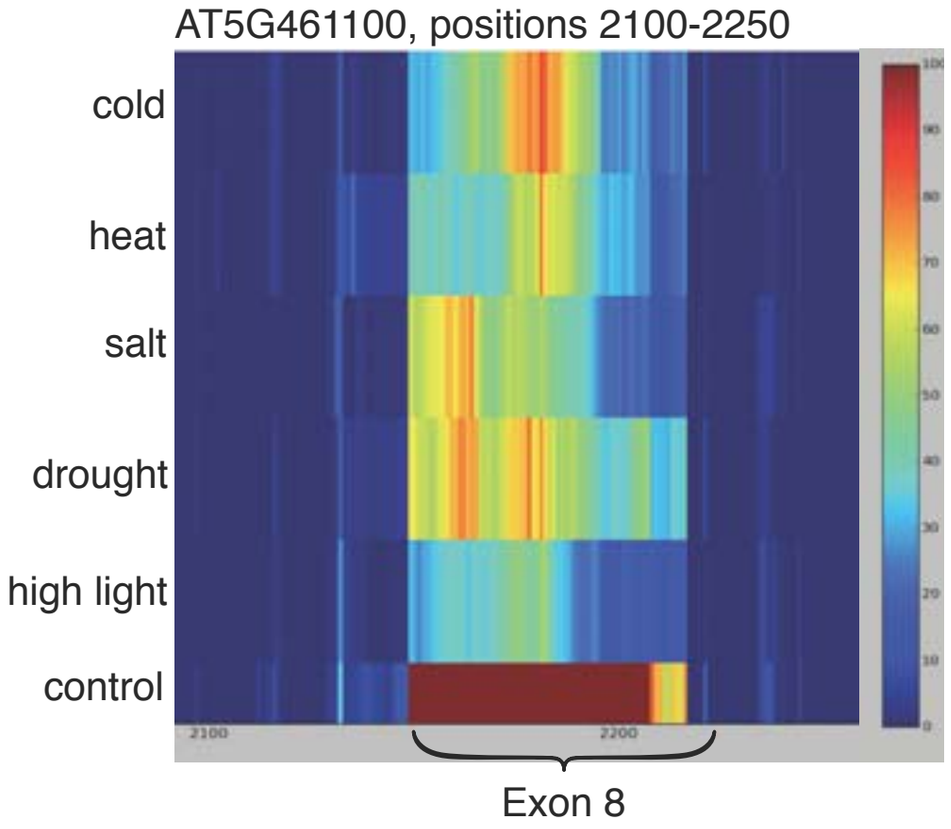
Introns removed resulting in
mature mRNA

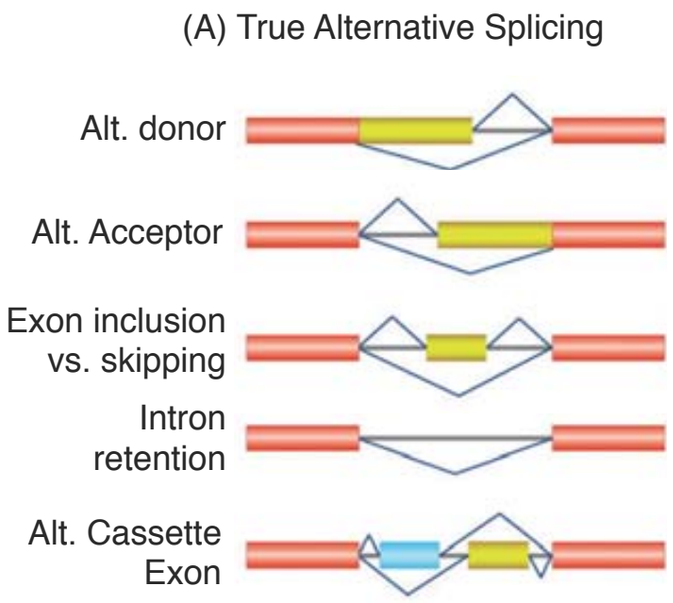


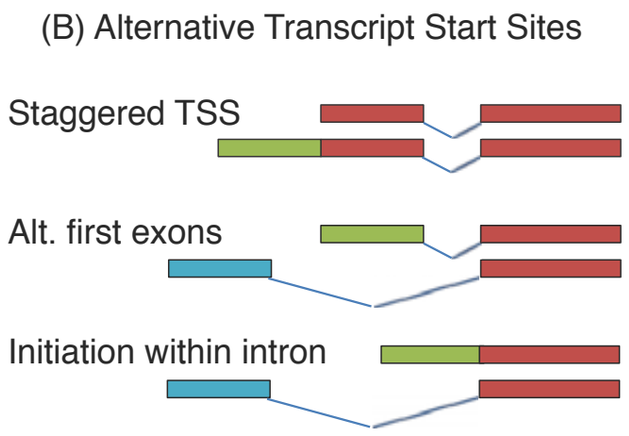
en.wikipedia.org

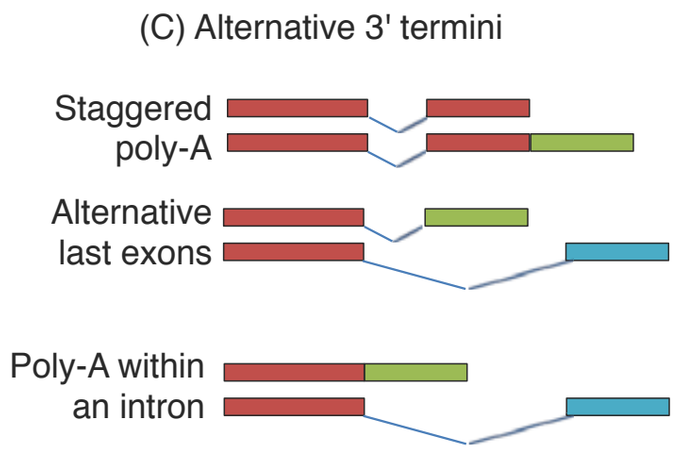
Alternative Splicing & Isoform Expression

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









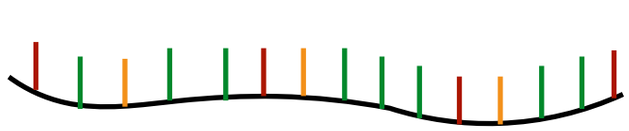
slide courtesy of Carl Kingsford

“Flow” of information in the cell

DNA

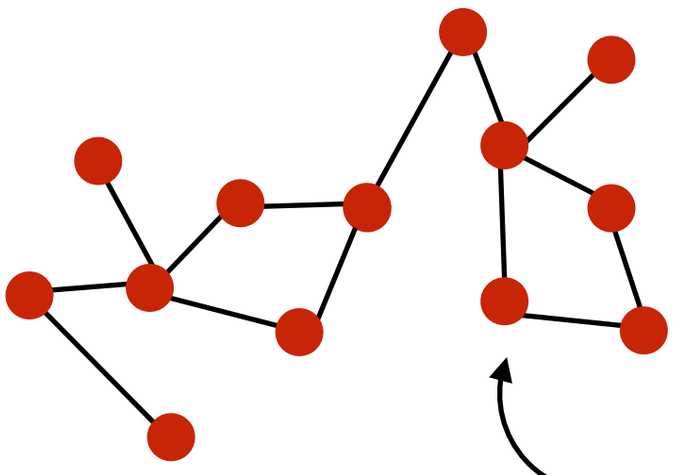
RNA Polymerase
(transcription)

RNA



Ribosomes
(translation)

Protein



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

mRNA→ Protein: Translation

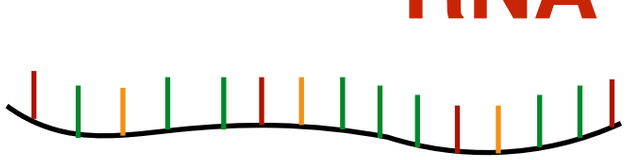
https://www.dnalc.org/resources/3d/15-translation-basic.html

“Flow” of information in the cell

DNA

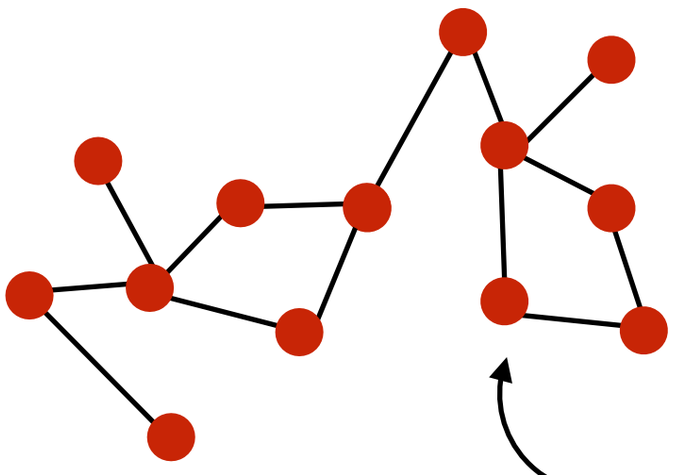
RNA Polymerase
(transcription)

RNA



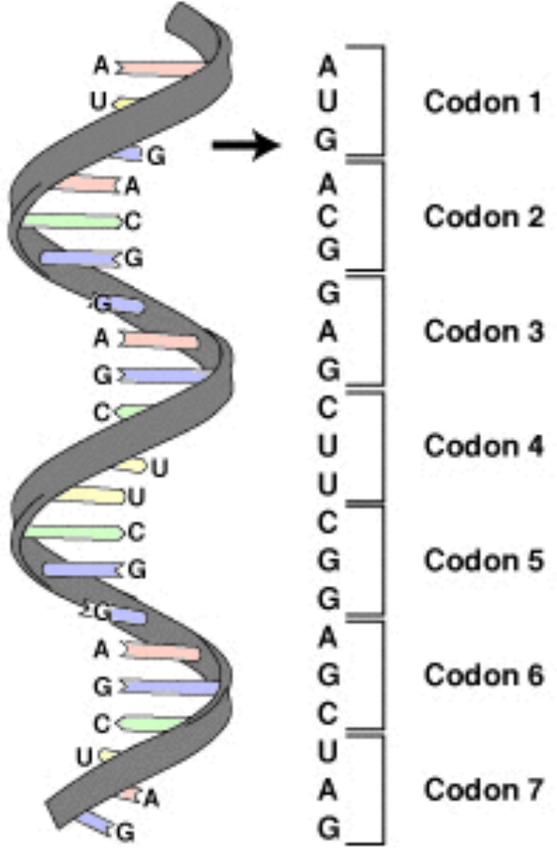
Ribosomes
(translation)

Protein



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

Protein



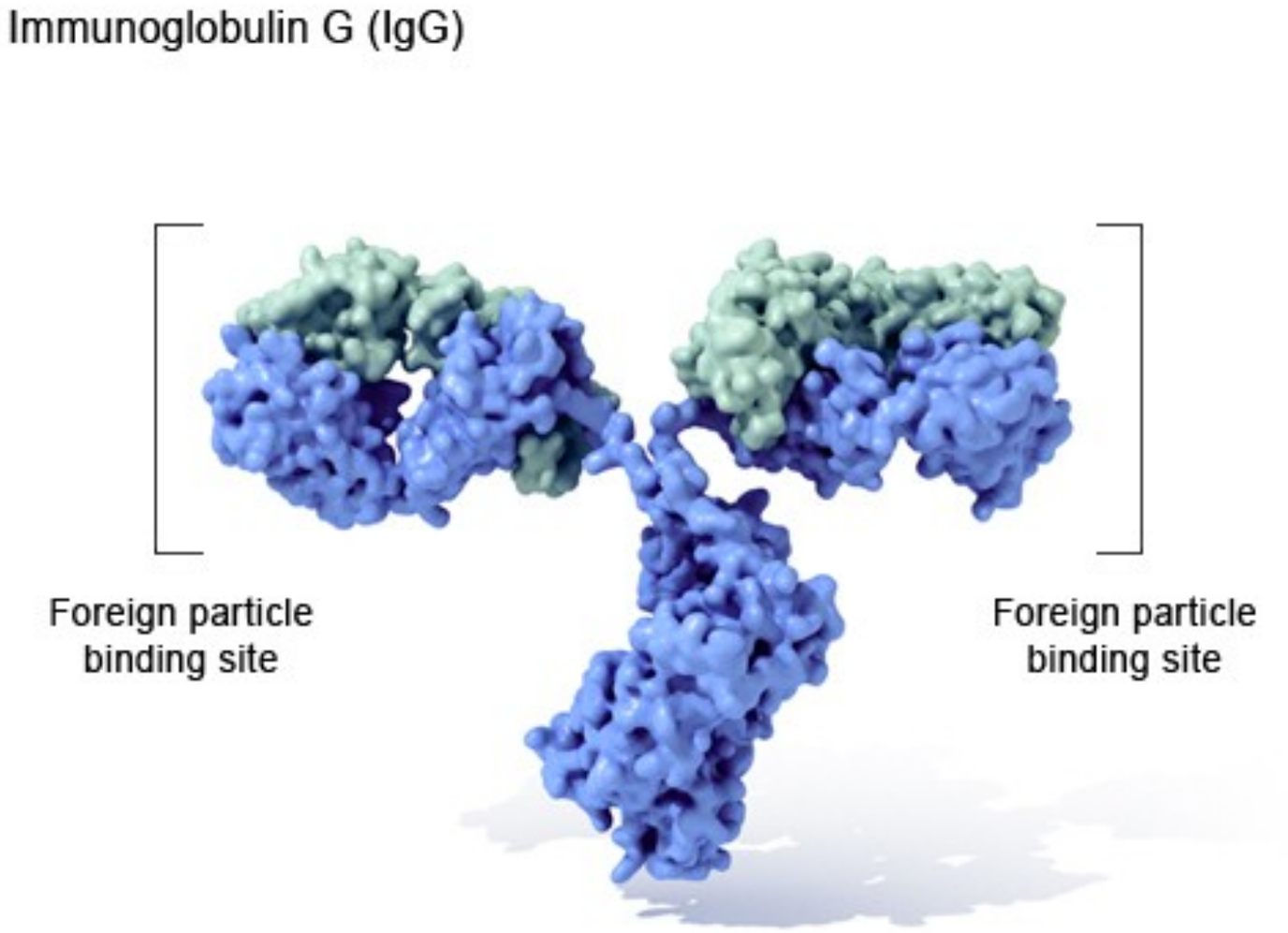
Triplets of mRNA bases (codons) correspond to specific amino acids

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

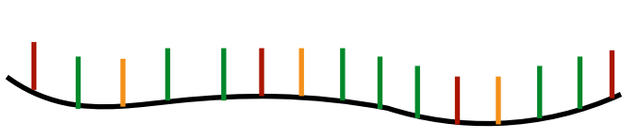
The shape or *conformation* of a protein is intimately tied to its function. Protein shape, therefore, is strongly conserved through evolution — even moreso than sequence. A protein can undergo sequence mutations, but fold into the same or a similar shape and still perform the same function.

“Flow” of information in the cell

DNA

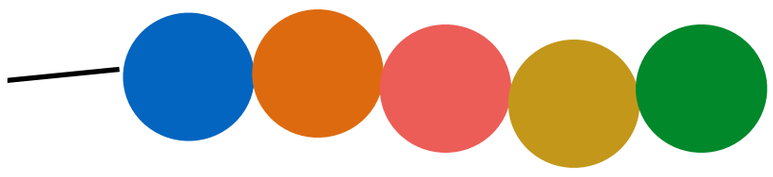
RNA Polymerase
(transcription)

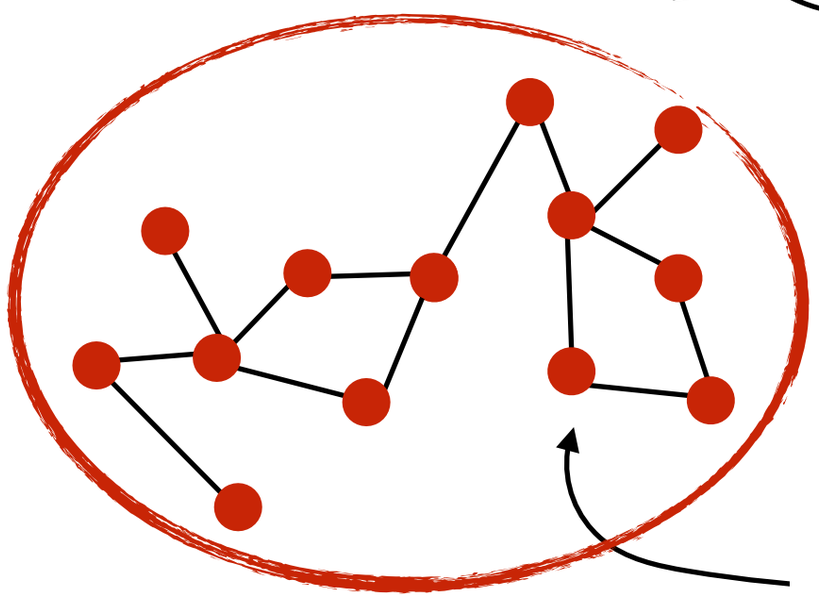
RNA



Ribosomes
(translation)

Protein





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

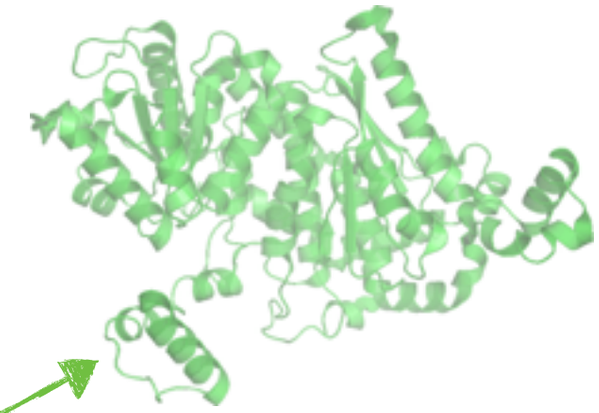
Glycolysis Pathway

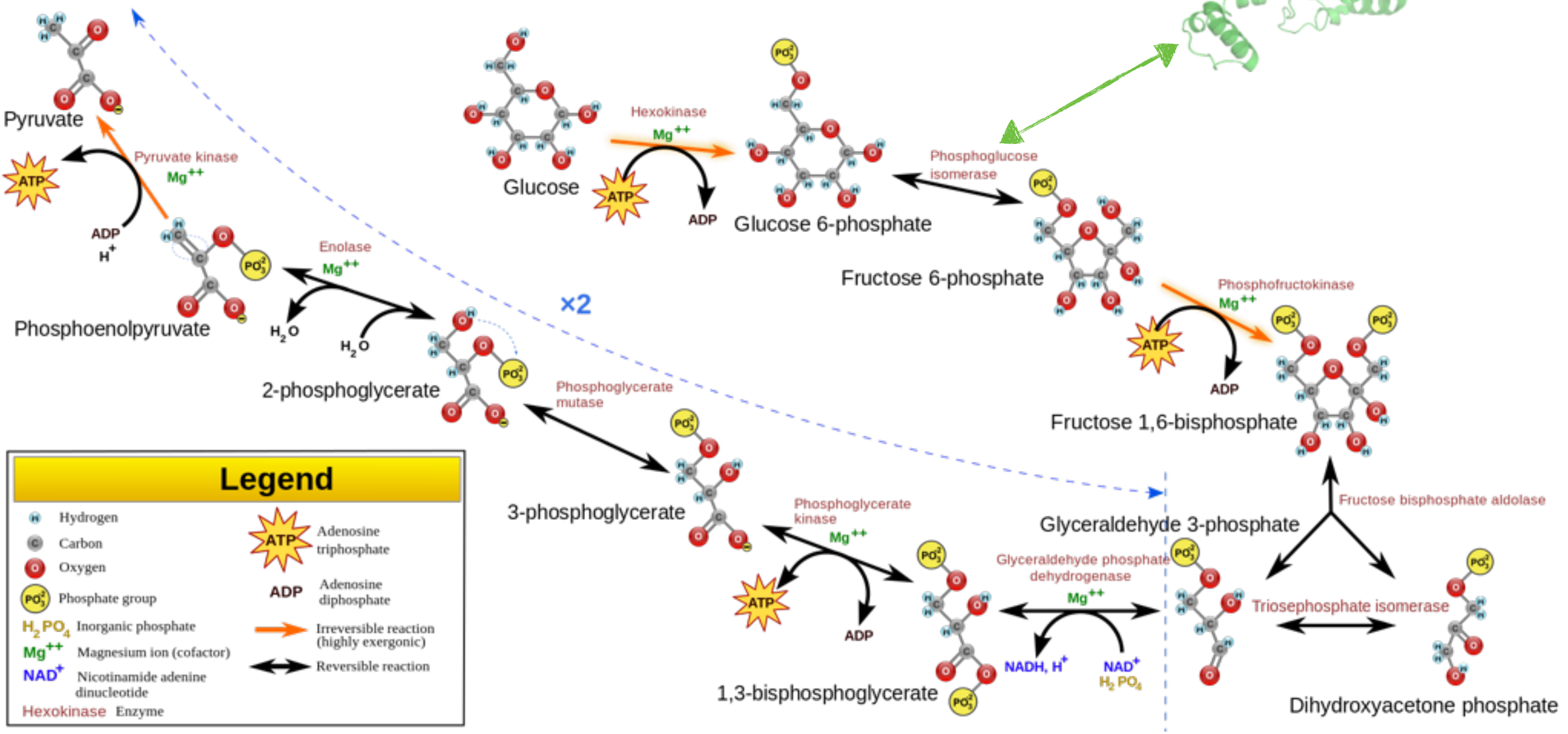
Converts glucose → pyruvate

Generates ATP (“energy currency” of the cell)

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

phosphoglucose isomerase





Some Interesting Facts

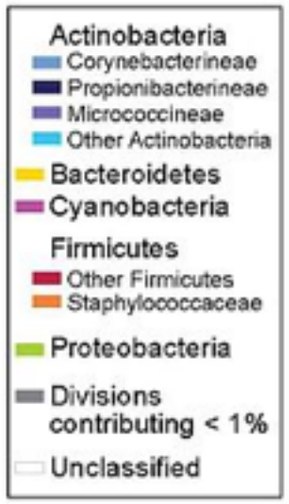
Organism	Genome size	# of genes
ϕX174 (*E. coli* virus)	~5kb	11
E. coli K-12	~4.6Mb	~4,300
Fruit Fly	~122Mb	~17,000
Human	~3.3Gb	~21,000
Mouse	~2.8Gb	~23,000
P. abies (a spruce tree)	~19.6Gb	~28,000

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

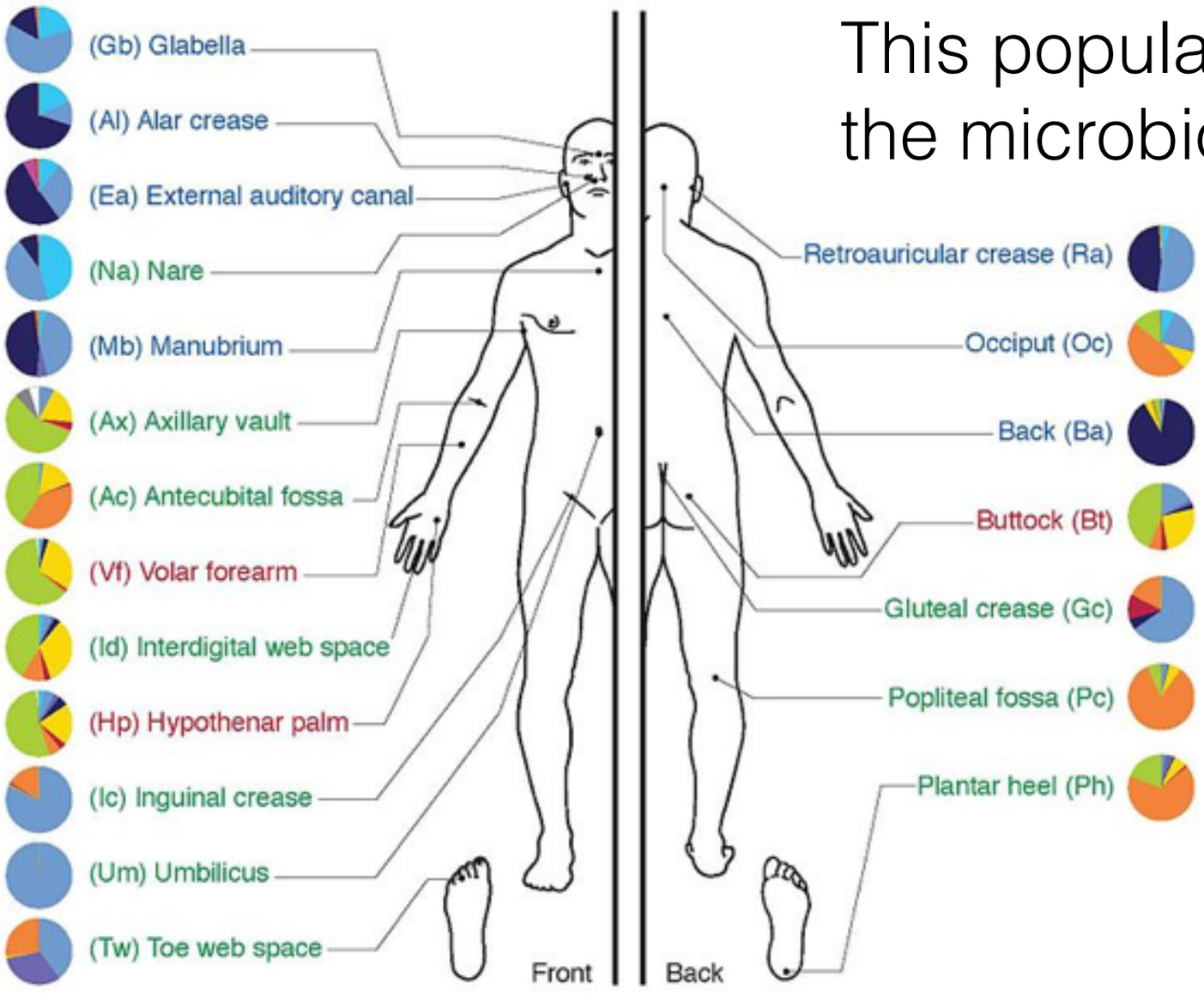
Some Interesting Facts



You are mostly bacteria, fungi & arches

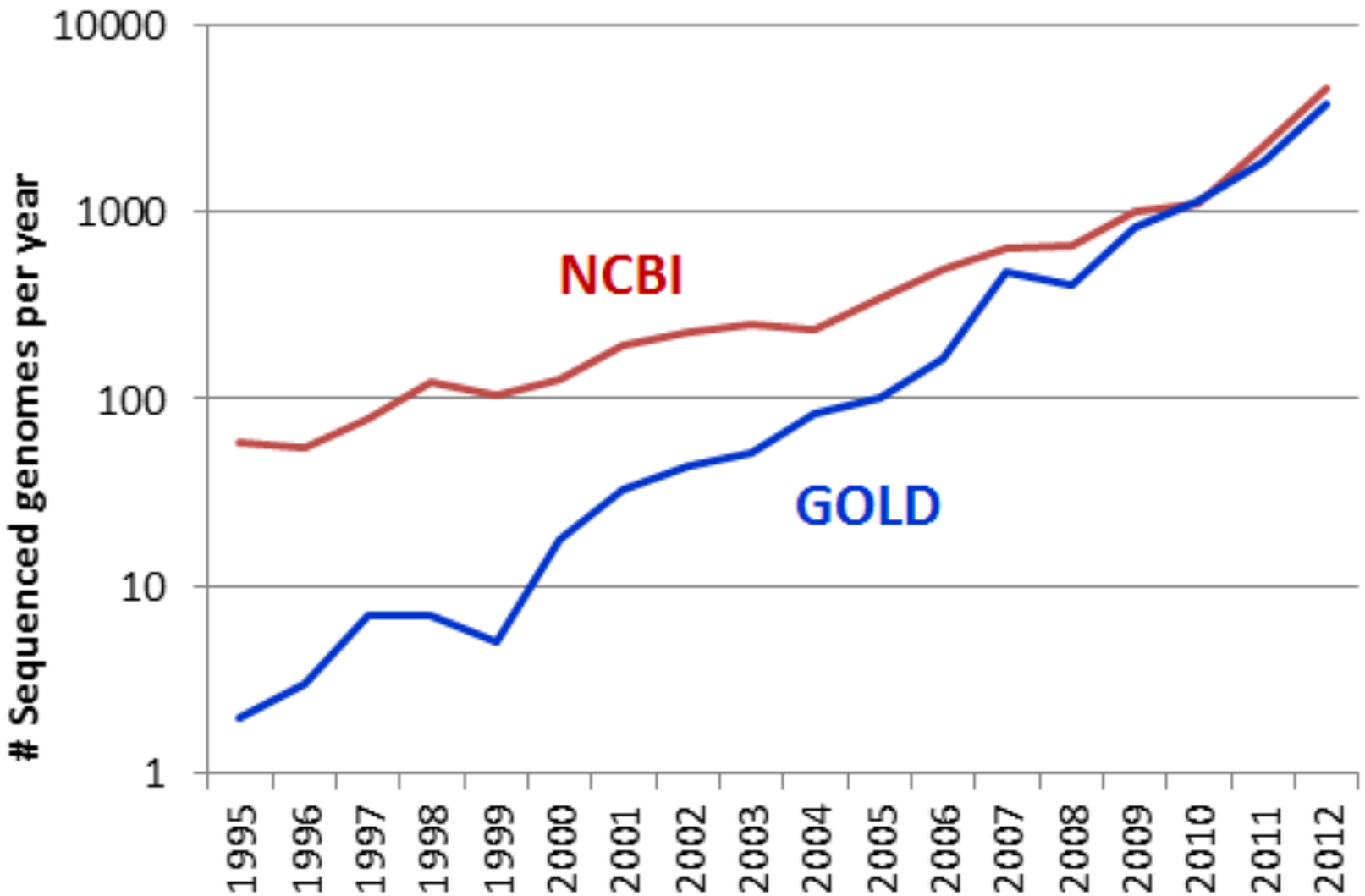
Non-human cells outnumber human cells ~10:1 in the human body

This population of organisms is called the microbiome



en.wikipedia.org : public domain

Some Interesting Facts



http://figshare.com/articles/Sequenced_Genomes_by_Year/715898

. . . Out of 8.7 ± 1.3 Mil*

Vast majority of species unsequenced & *can not be cultivated in a lab* (motivation for metagenomics)

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