

# CMSC 423: Bioinformatics Algorithms, Databases & Tools

Fall 2021



UNIVERSITY OF  
**MARYLAND**

# Course Info

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

Office: 3220 IRB

Office Hours: Wed. 10-11AM (but please e-mail me before to let me know, as I cannot accommodate > 2 students in the office at once)

Website: [https://rob-p.github.io/CMSC423\\_F21/](https://rob-p.github.io/CMSC423_F21/)

# Course Info

TAs:

Noor Pratap Singh

email: [npsingh@umd.edu](mailto:npsingh@umd.edu)



Shramay Palta

email: [spalta@umd.edu](mailto:spalta@umd.edu)



TA office hours are available 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/fall2021/cmsc423>

**If you have a class-related e-mail:** Please **prefix the subject with [CMSC423\_F21]**, so that my filter will pick it up and it won't be accidentally routed to SPAM.

# COVID Related

- **Masking policy** : President Pines [provided clear expectations](#) to the University about the wearing of masks for students, faculty, and staff. Face coverings over the nose and mouth are required while you are indoors at all times. There are no exceptions when it comes to classrooms and laboratories. Students not wearing a mask will be given a warning and asked to wear one, or will be asked to leave the room immediately. Students who have additional issues with the mask expectation after a first warning will be referred to the Office of Student Conduct for failure to comply with a directive of University officials.
- **Other COVID-related information** : More specific health information appears below in this syllabus. However, the following bears repeating; please use both *common sense* and the *precautionary principle* in determining if you should get tested or attend a specific lecture in person. Our goal this semester is to maintain in person instruction as long as is feasible (ideally the whole semester). In addition to mask and vaccine mandates, retaining in person instruction and avoiding serious outbreaks will require the consistent vigilance of all of those in our university community. If you are feeling at all symptomatic, or if you have been in “contact” (< 6 feet for  $\geq$  15 minutes) with someone who has tested positive for SARS-CoV-2 (regardless of whether or not they are symptomatic), please *do not attend* this class or any other class until you have tested negative. Only with our collective effort and continued vigilance will we have an opportunity at a successful *in person* semester.

# Coursework & Grading

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

- Programming assignments - 40%
- 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. Each student is allowed *one* free late assignment turnin (you can turn the assignment in up to 48 hours late with no penalty). However, you must let us know (e-mail the TAs) that you are using your free late assignment when you turn in your assignment, and the decision is non-revocable (if you decide to use the free late assignment for assignment 3, you can't than request to take the late penalty for 3 and use the late assignment for assignment 4).

**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. *Plagiarism 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 constitutes plagiarism. 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 receive a grade of XF, and you will be reported to the university's Office of Student Conduct.**

# Textbooks

Based on student feedback from previous offerings of this course, there *is no required text.*

Additional material will be made available on the course website as needed.

**However**, this is an *upper-level* course, and you should *absolutely* seek out other sources explaining these topics from different angles, using different notations and examples, etc.

If you seek out other sources and still are having difficulty with an idea, *please* reach out to us (myself & TAs). Also, consider reaching out to your peers *via* Piazza.

# Other Textbooks

## Genomics algorithms, data structures, and statistical models:

- [Bioinformatics Algorithms: An Active Learning Approach](#)
- [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)

# More syllabus stuff

## Course Objectives

The main objective of this course will be to provide an understanding of some of the algorithms, data structures, and methods that underlie modern computational genomics. This course is intended as a broad introduction to bioinformatics and computational biology. However, this is a huge field, so we will not cover everything, and what we do cover will not all be at the same depth (e.g. we will spend more time discussing indexing than clustering). Our perspective will be a computational and algorithmic one, though we will take the time to understand the necessary biology and motivation for the problems we discuss. At the end of this course, you should have a good understanding of how new challenges in genomics drive algorithmic innovations and how algorithmic innovations enable new and improved biological analyses.

About 35,400,000 results (0.91 seconds)

<https://www.medicaltechnologyschools.com> › bioinfor... ⋮

## Bioinformatics vs. Computational Biology: A Comparison

As both fields rely on the availability and accuracy of datasets, they usually help one another reach their respective project goals. While **computational** ...

### People also ask ⋮

[What is the difference between bioinformatics and computational biology?](#) ⋮

[Is bioinformatics better than computational biology?](#) ⋮

[Is computational biology a good field?](#) ⋮

[What can I do with a masters in computational biology?](#) ⋮

Feedback

<https://www.northeastern.edu> › graduate › blog › comp... ⋮

## Computational Biology vs. Bioinformatics: What's the Difference?

May 28, 2021 — While Kaluziak notes that there is a great deal of overlap between **computational biology** and **bioinformatics**, the latter requires programming and ...

<https://www.reddit.com> › bioinformatics › comments ⋮

## Computational Biology versus Bioinformatics - Reddit

Jan 22, 2016 — There's considerable overlap, but in general **Computational Bio** is more concerned with developing the theory and writing the software to answer ...

[Bioinformatics vs Computational Biology - Reddit](#) Sep 23, 2017

[Bioinformatics vs Computational Biology. The skills you need ...](#) Mar 5, 2017

[And also the job market for it: bioinformatics - Reddit](#) Jun 30, 2021

[Systems Biology vs Computational Biology vs Bioinformatics?](#) Sep 13, 2015

[More results from www.reddit.com](#)

<https://www.facom.ufms.br> › ~diego › bioinformatics-v... ⋮

## Bioinformatics vs. Computational Biology - Facom - UFMS

**Bioinformatics** applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and ...

<https://www.quora.com> › What-are-the-differences-bet... ⋮

## What are the differences between bioinformatics and ... - Quora

Nov 15, 2015 — **Computational biology:** The study of biology using computational techniques.

# 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**

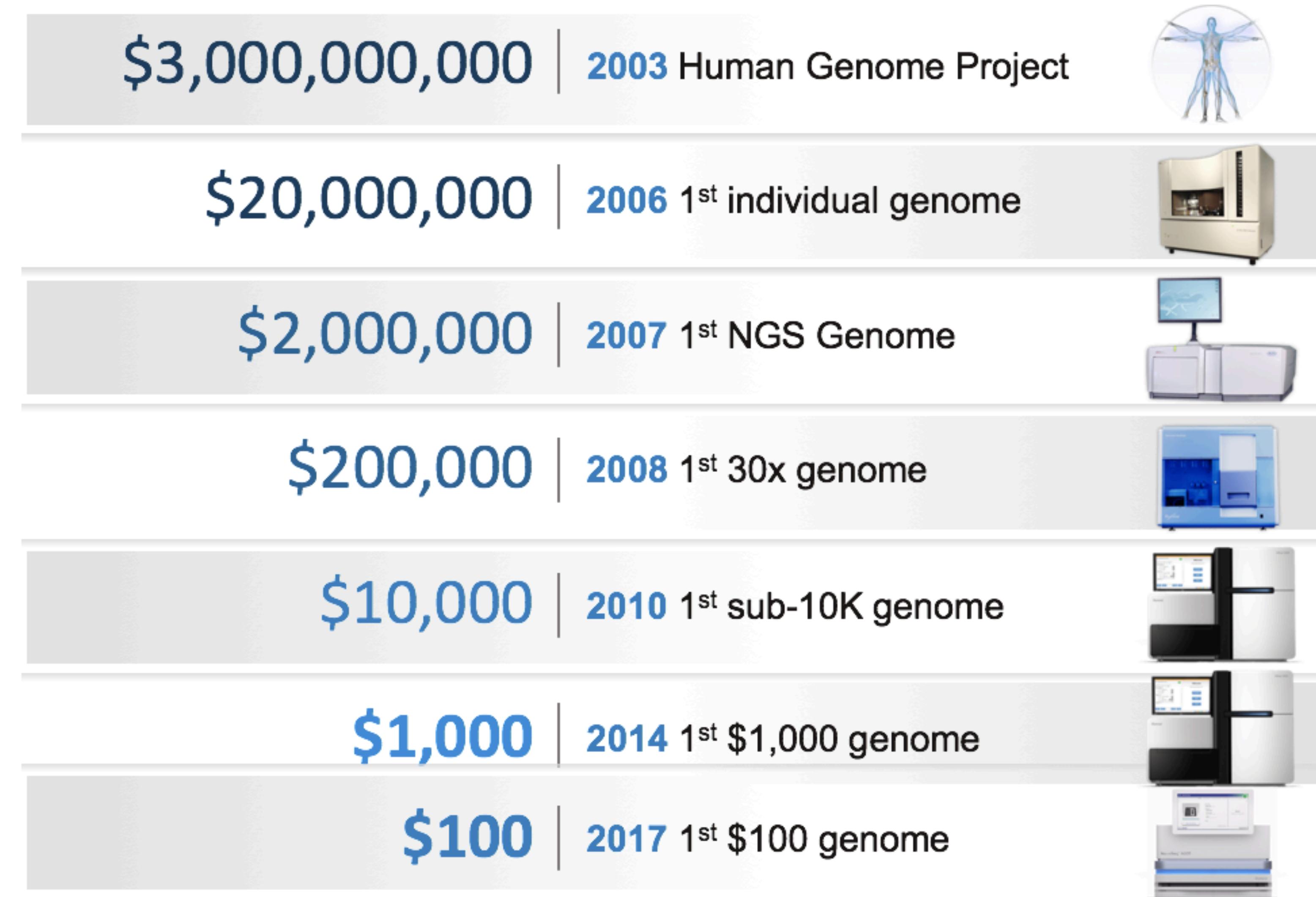
*Sequencing* the first human genome took ~10 years and cost ~\$2.7 **billion**

**Today**

*Sequencing* a genome costs ~\$100 - 1,000<sup>+</sup> (depending on how you count)

~18 Tb per “run” at maximum capacity

# Progression of sequencing capacity



# 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.

# For the first time when teaching this class ...



# bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Follow this preprint

## The complete sequence of a human genome

Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G.S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpeljiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F.A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, Adam M. Phillippy

**doi:** <https://doi.org/10.1101/2021.05.26.445798>

This article is a preprint and has not been certified by peer review [what does this mean?].



Abstract

Full Text

Info/History

Metrics

Preview PDF

[HOME](#) | [ABOUT](#) | [SUBMIT](#) | [NEWS & NOTES](#) | [ALERTS / RSS](#)

| [CHANNELS](#)

Search



[Advanced Search](#)

Previous

Posted May 27, 2021.

Next

[Download PDF](#)

[Email](#)

[Supplementary Material](#)

[Share](#)

[Data/Code](#)

[Citation Tools](#)

[XML](#)

[Tweet](#)

[Like 2.6K](#)

## COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv

### Subject Area

**Genomics**

### Subject Areas

### All Articles

Animal Behavior and Cognition

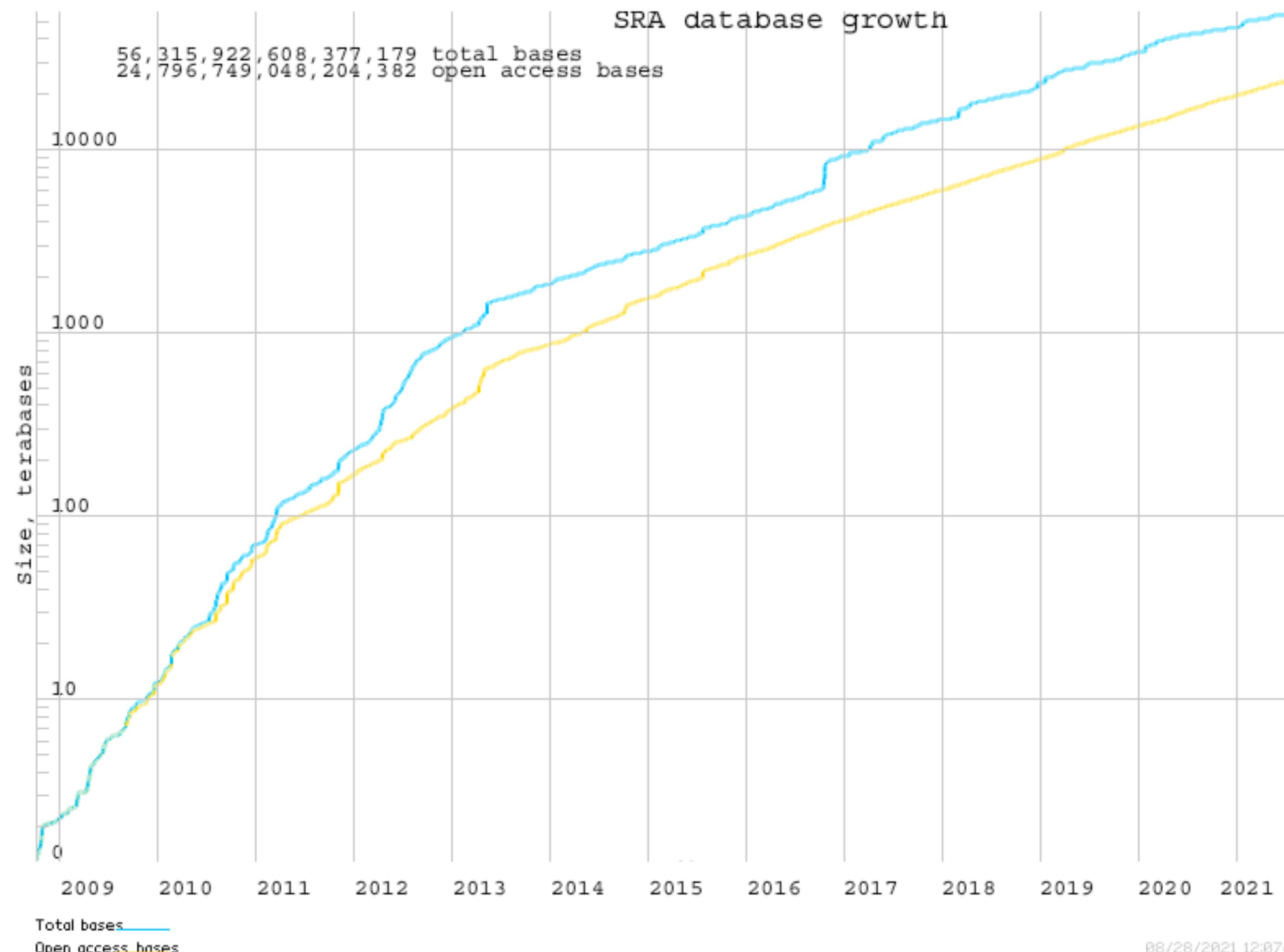
Biochemistry

Bioengineering

Bioinformatics

despite these limitations, scientists have used sequencing at a breakneck pace

## 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)

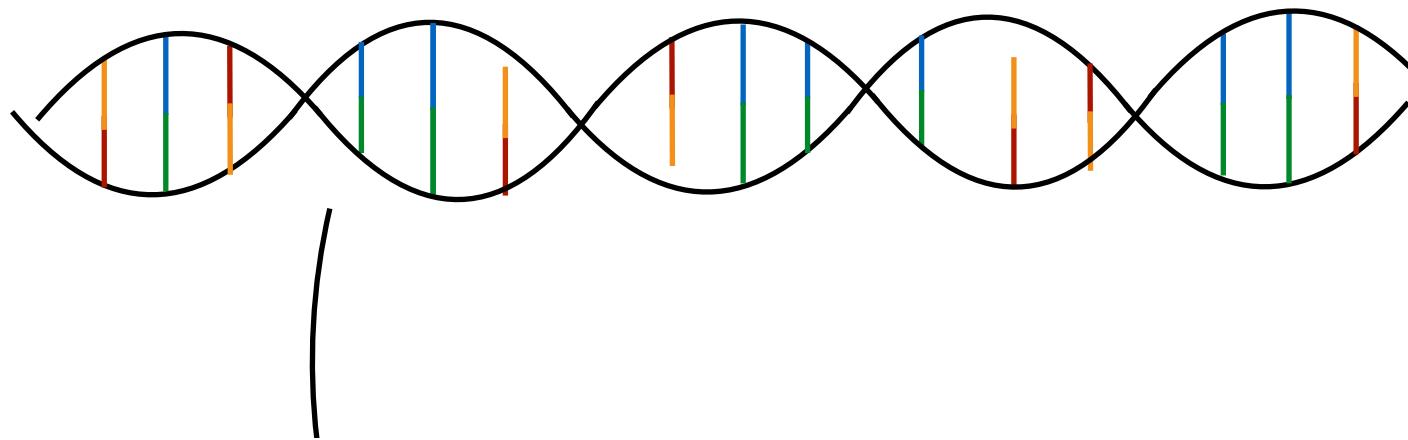
# Establishing a basic lexicon

This course will focus on algorithms and data structures (with a little bit of probability & statistics), but the **problems** we will explore derive directly from biological questions.

In order to motivate our Computer Science, we will need a basic understanding of the Molecular Biology in which the problems are phrased.

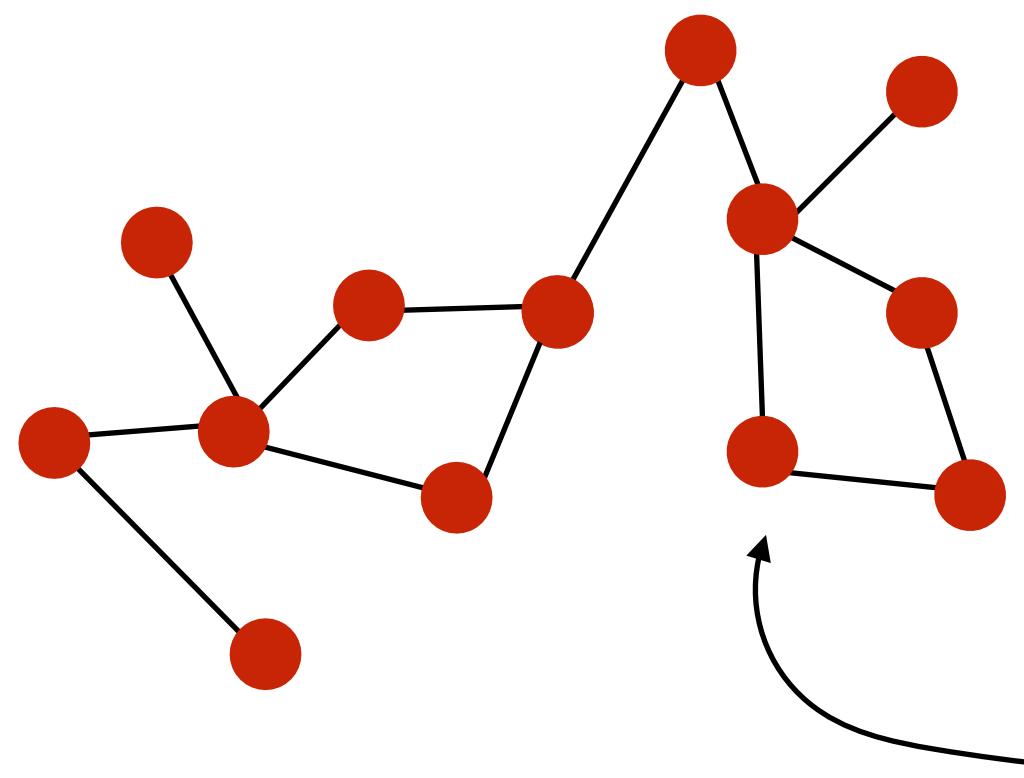
# “Flow” of information in the cell

**DNA**



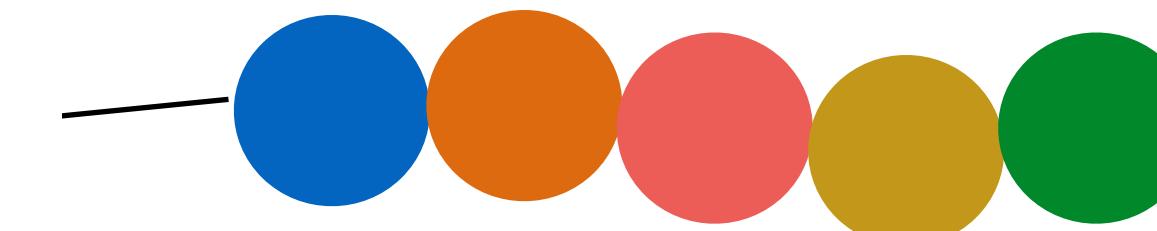
RNA Polymerase  
(transcription)

**RNA**



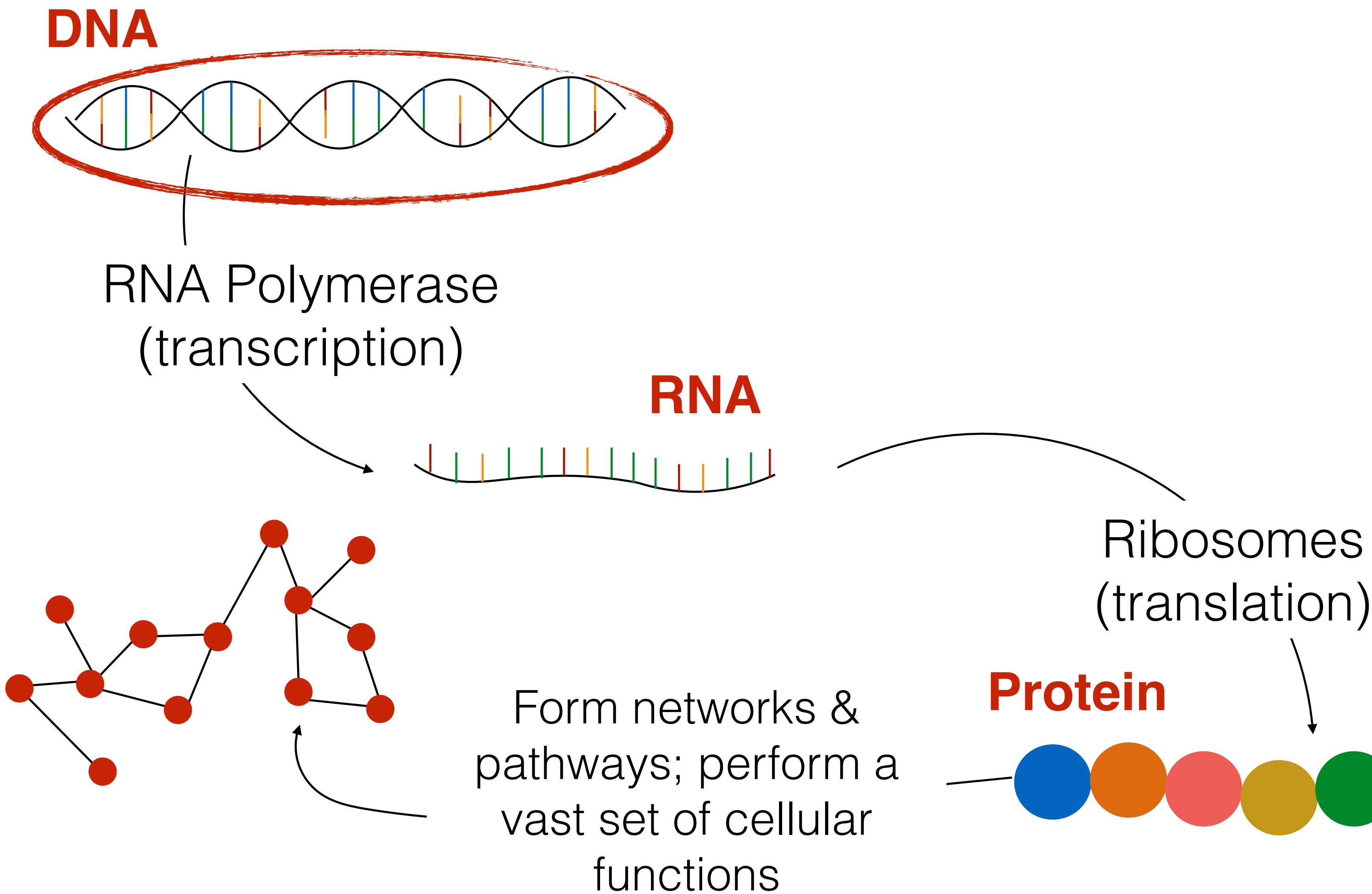
Form networks &  
pathways; perform a  
vast set of cellular  
functions

**Protein**

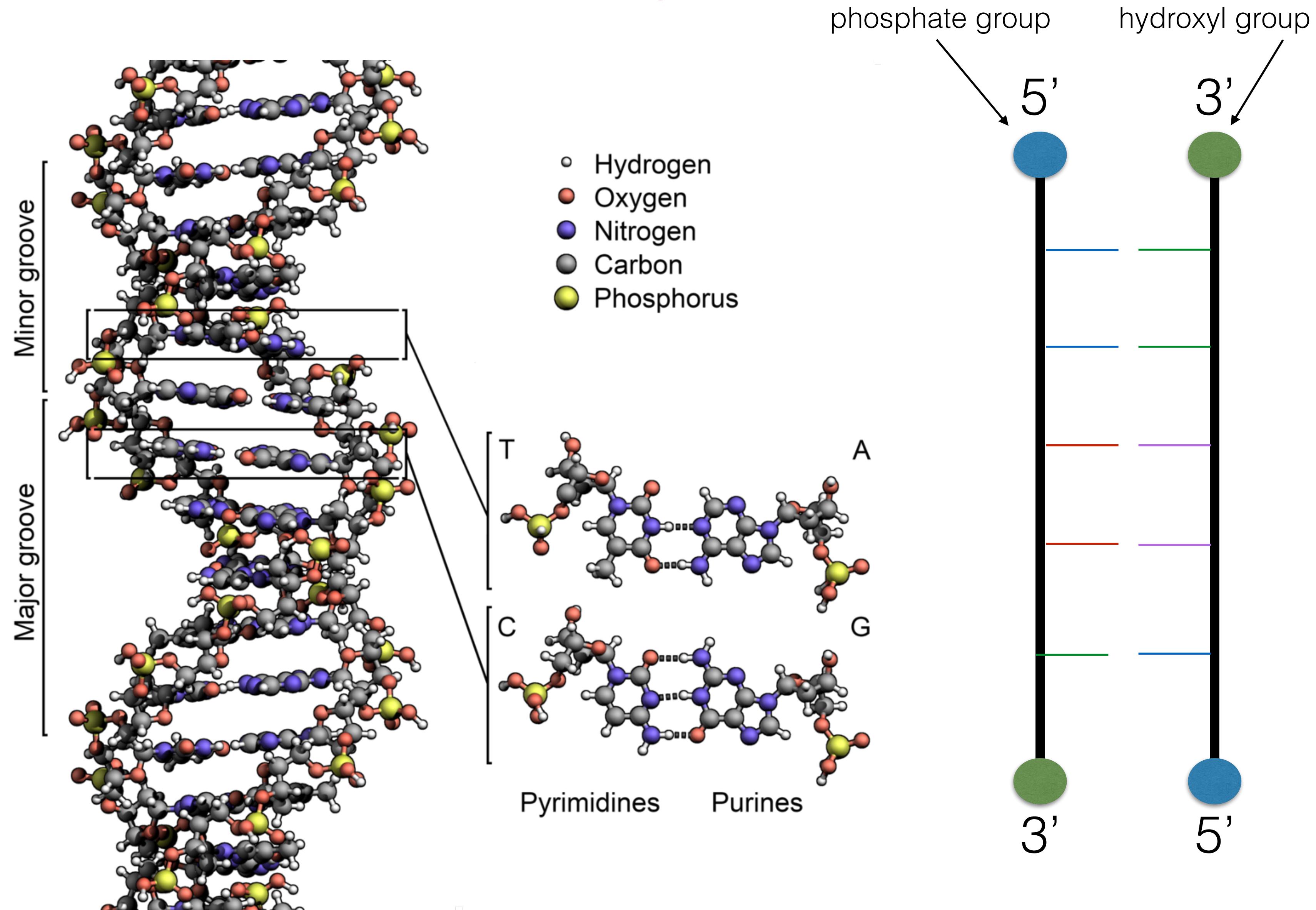


Ribosomes  
(translation)

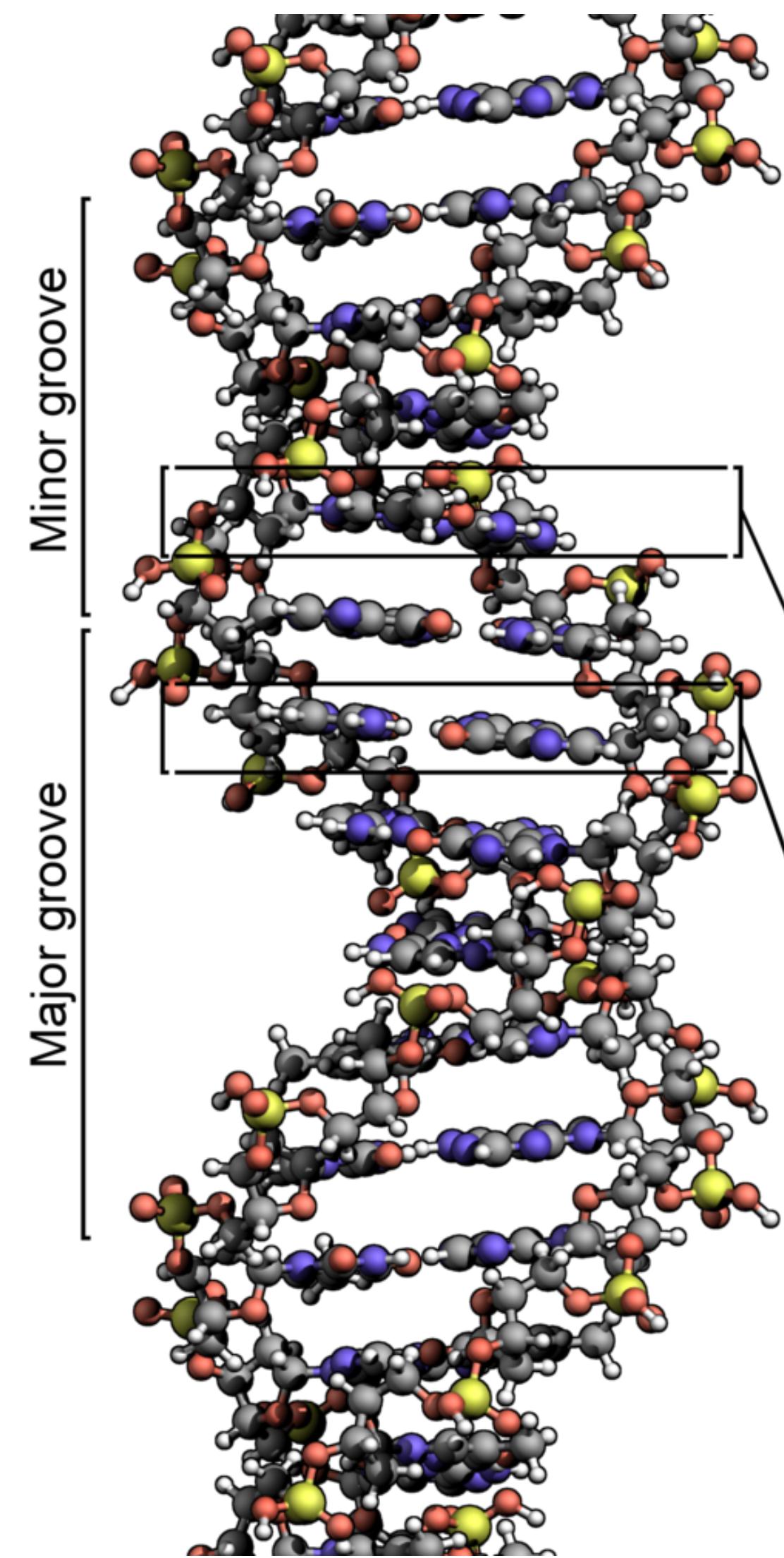
# “Flow” of information in the cell



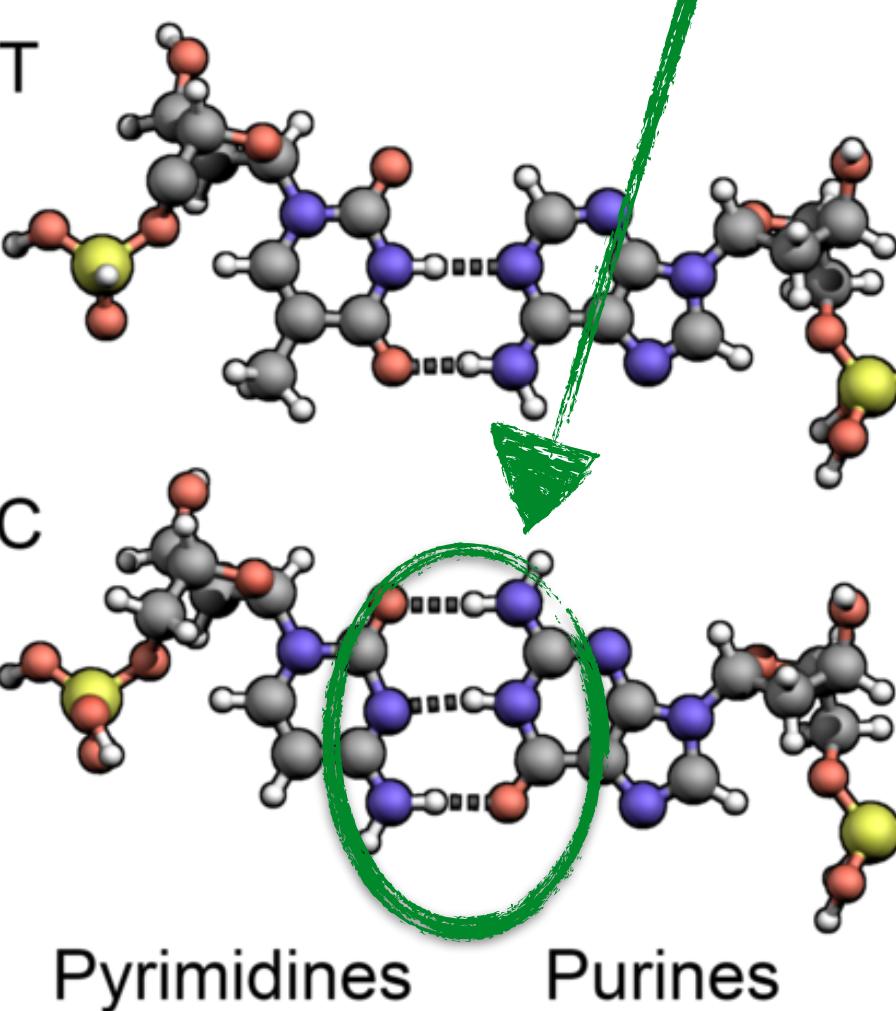
# DNA (the genome)



# DNA (the genome)



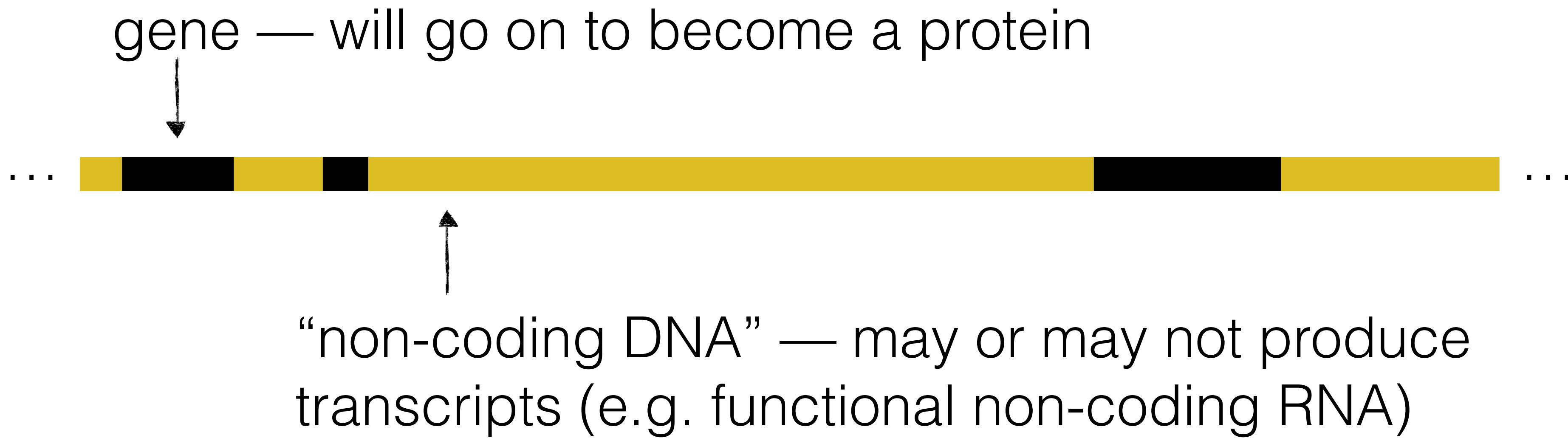
- Hydrogen
- Oxygen
- Nitrogen
- Carbon
- Phosphorus



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)



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

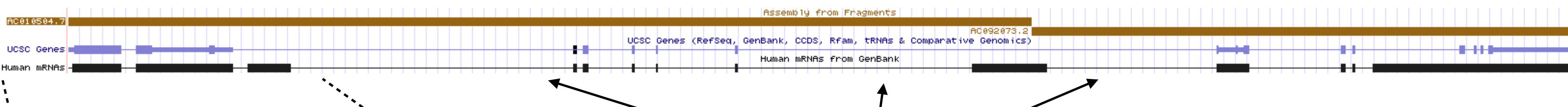
In typical bacterial genome, only small fraction —  
~2% — of DNA is “non-coding”

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

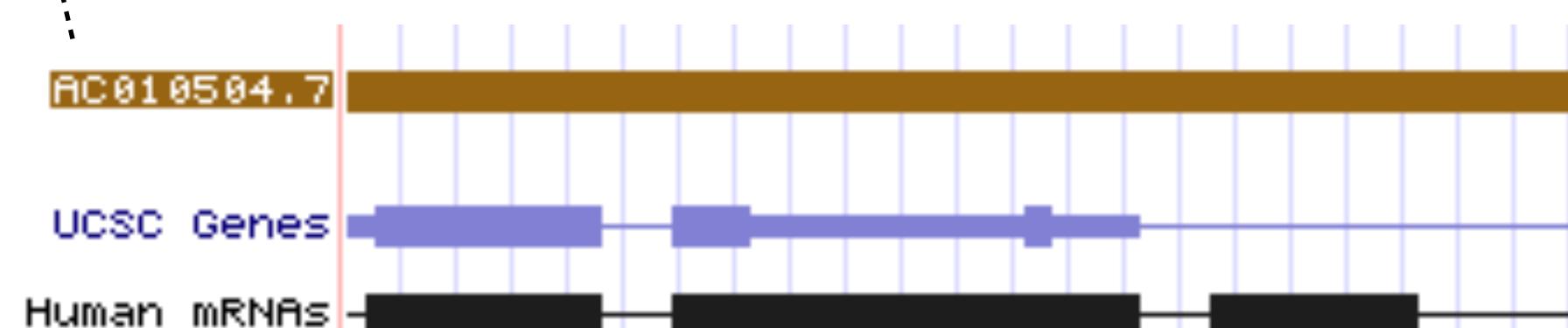
# DNA (the genome)

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

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

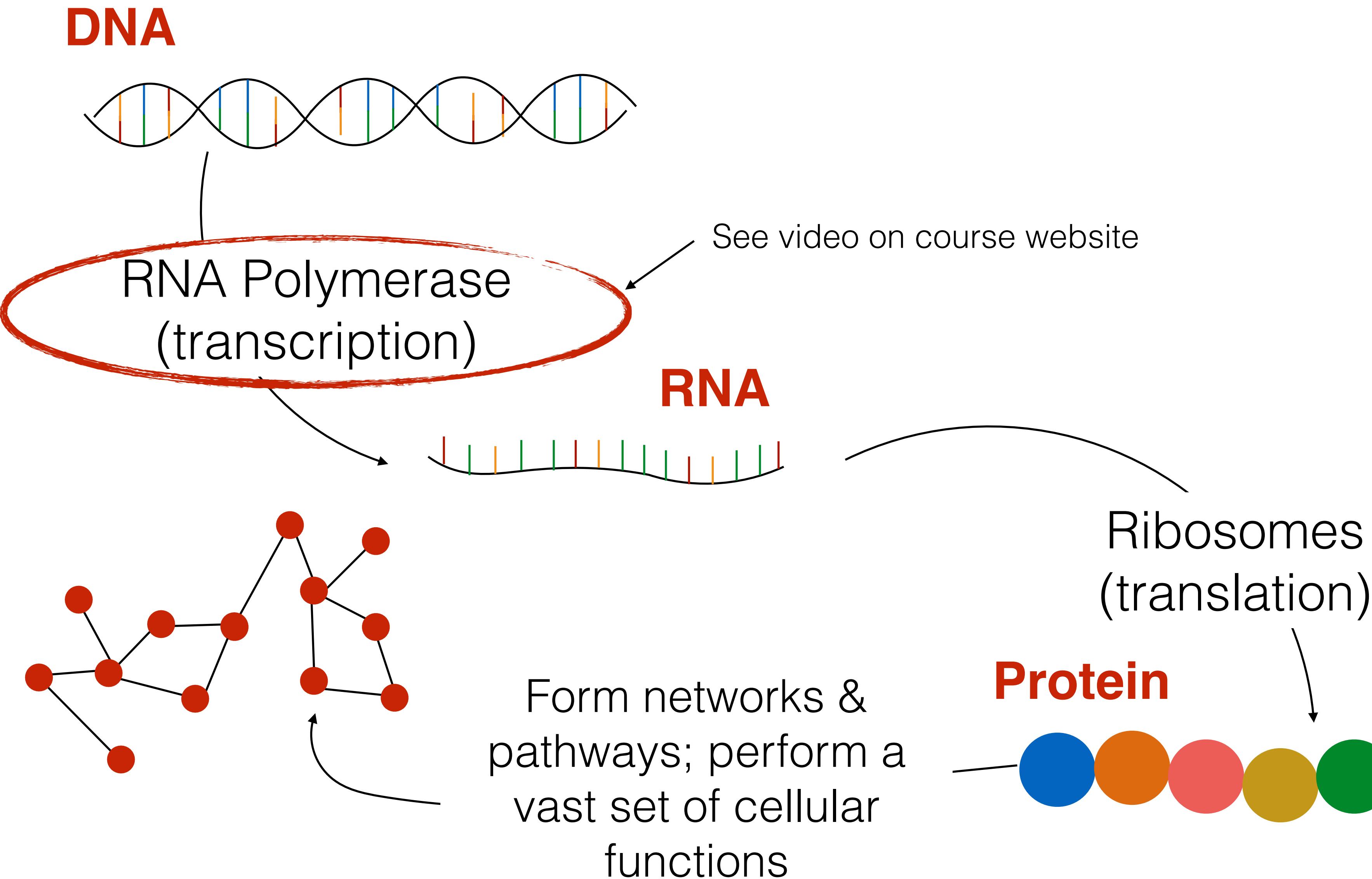


introns — “spliced” out of mature RNA

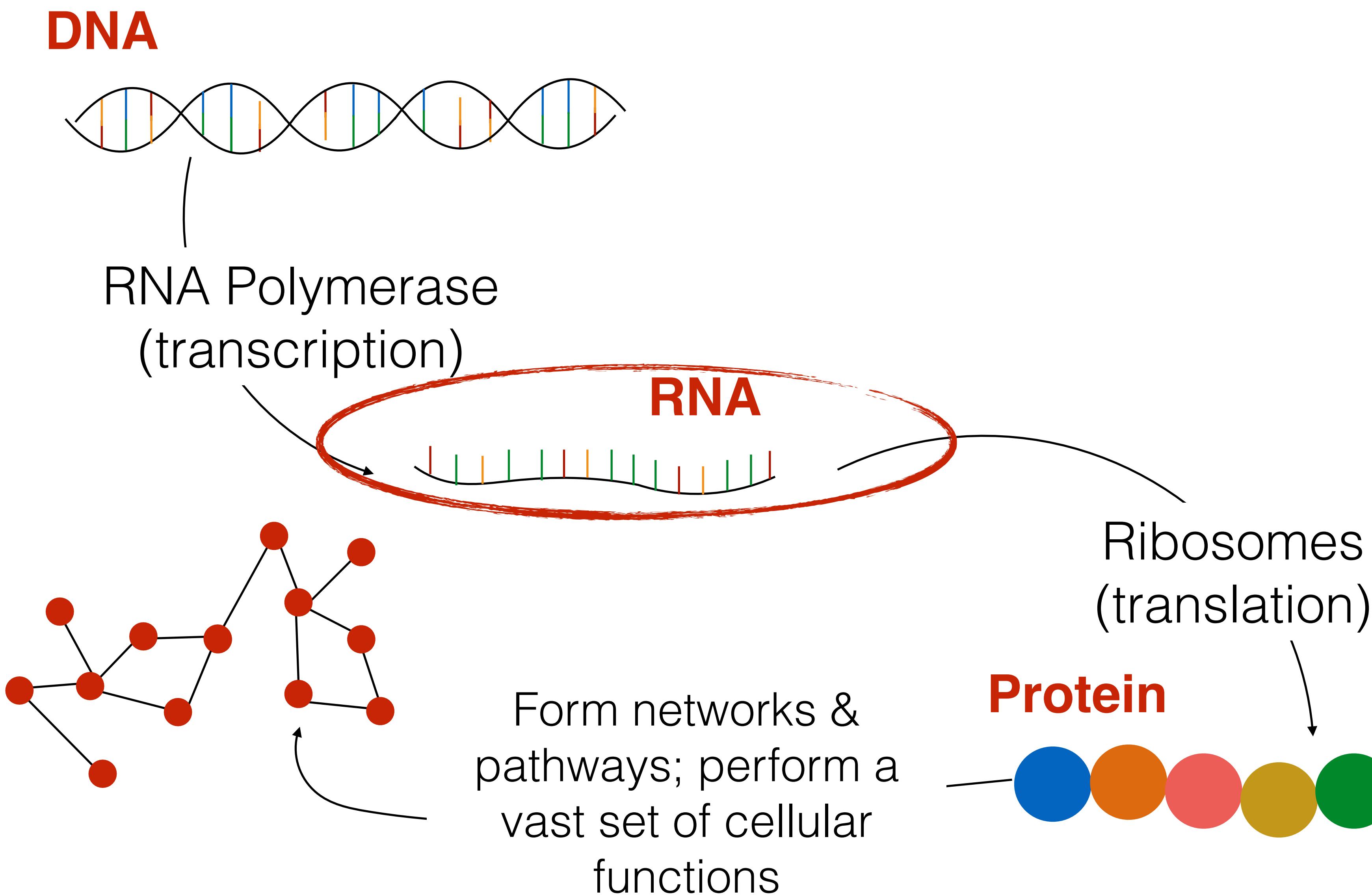


exons — appear in the *mature* RNA transcript

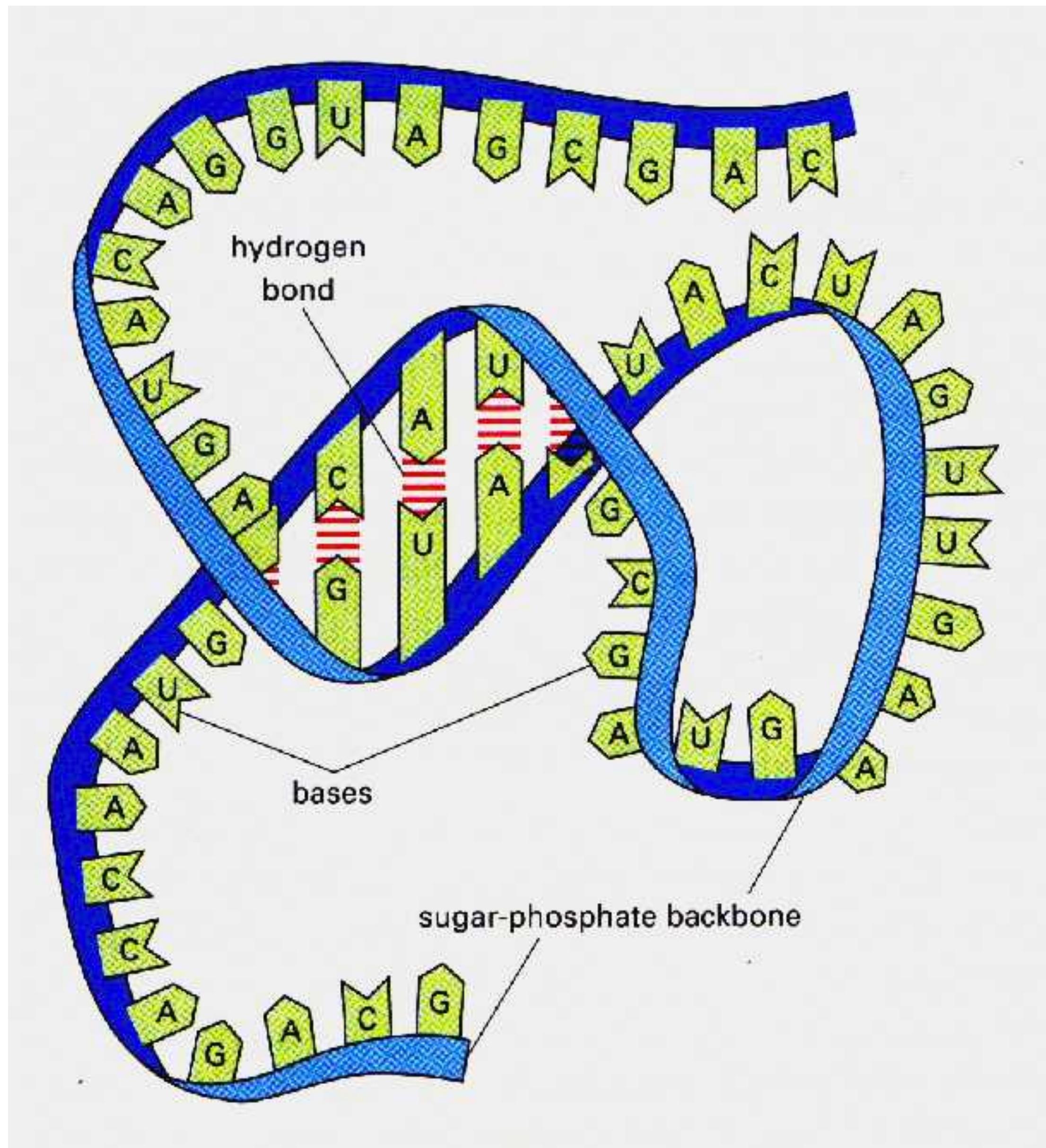
# “Flow” of information in the cell



# “Flow” of information in the cell



# 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

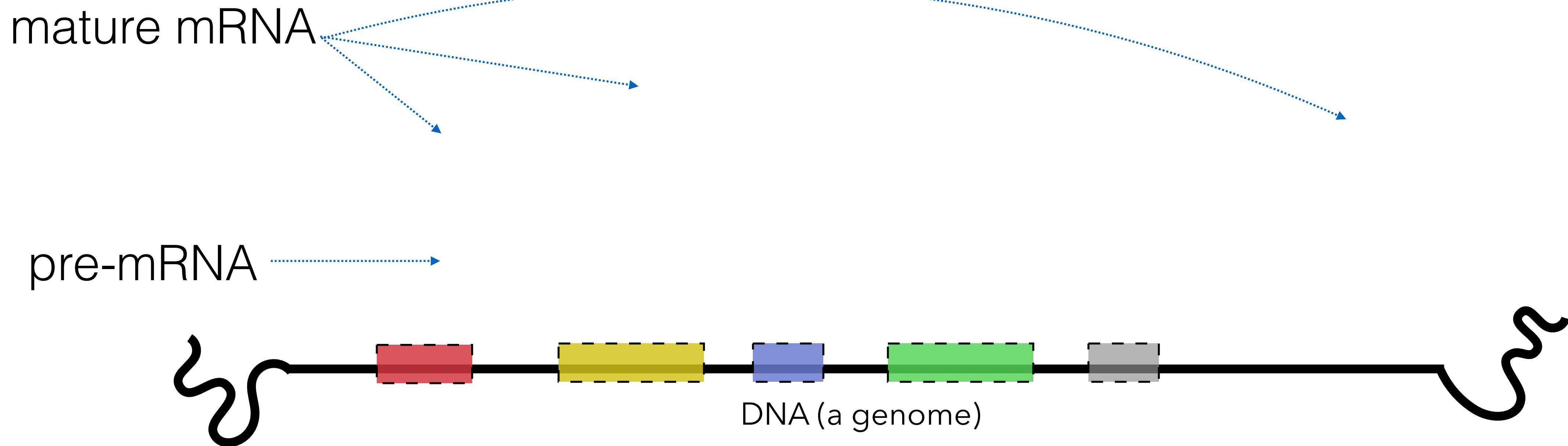
Same “alphabet” as DNA,  
except thymine replaced by  
uracil

# RNA Splicing

DNA transcribed into pre-mRNA

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

Introns removed from pre-mRNA resulting in mature mRNA

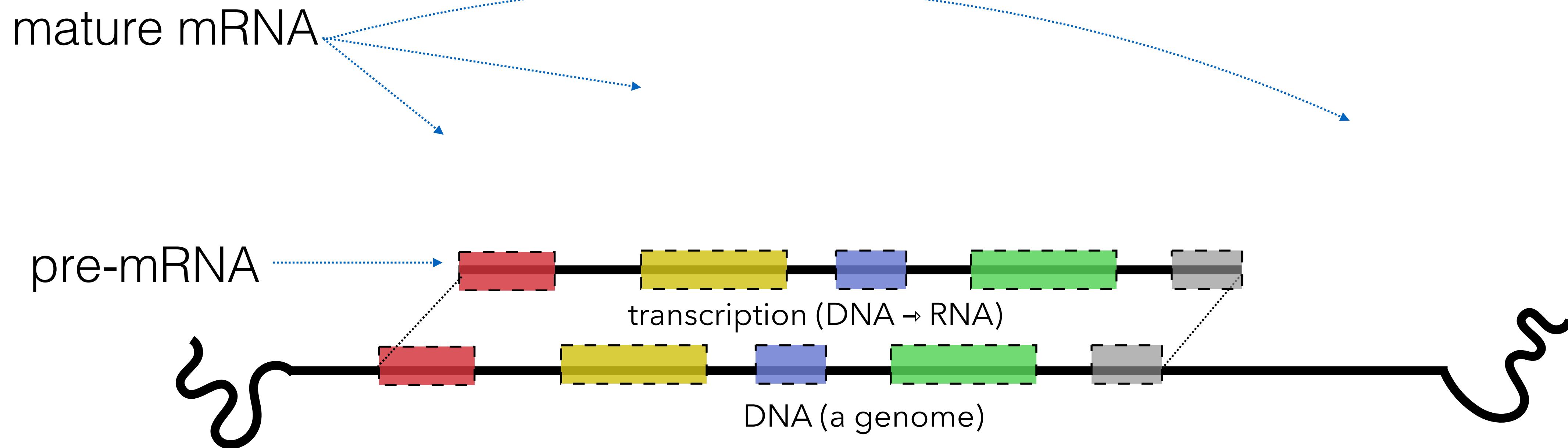


# RNA Splicing

DNA transcribed into pre-mRNA

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

Introns removed from pre-mRNA resulting in mature mRNA

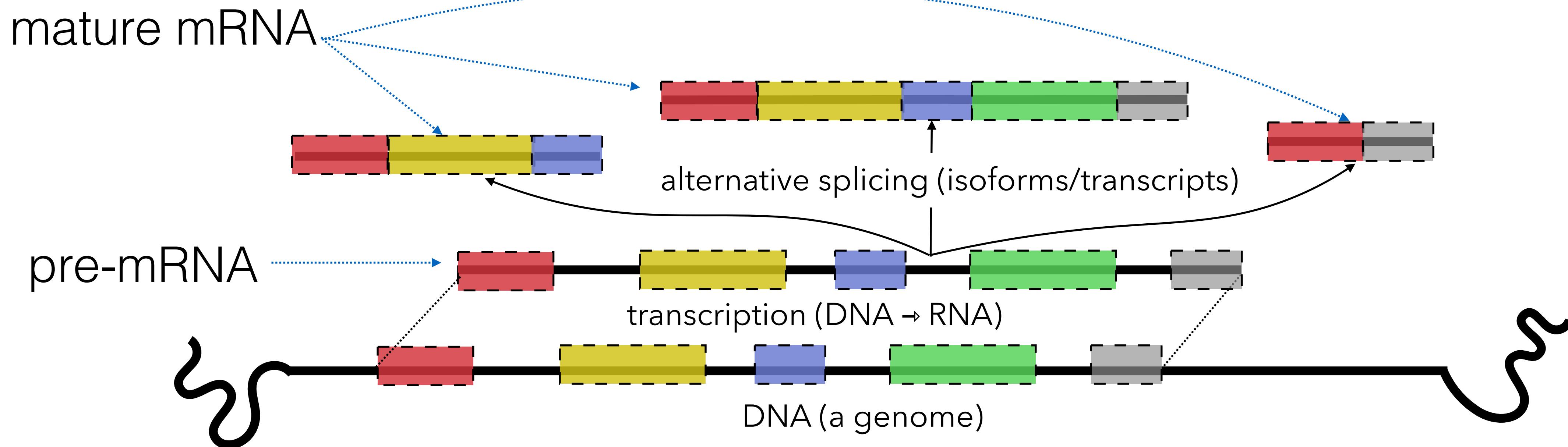


# RNA Splicing

DNA transcribed into pre-mRNA

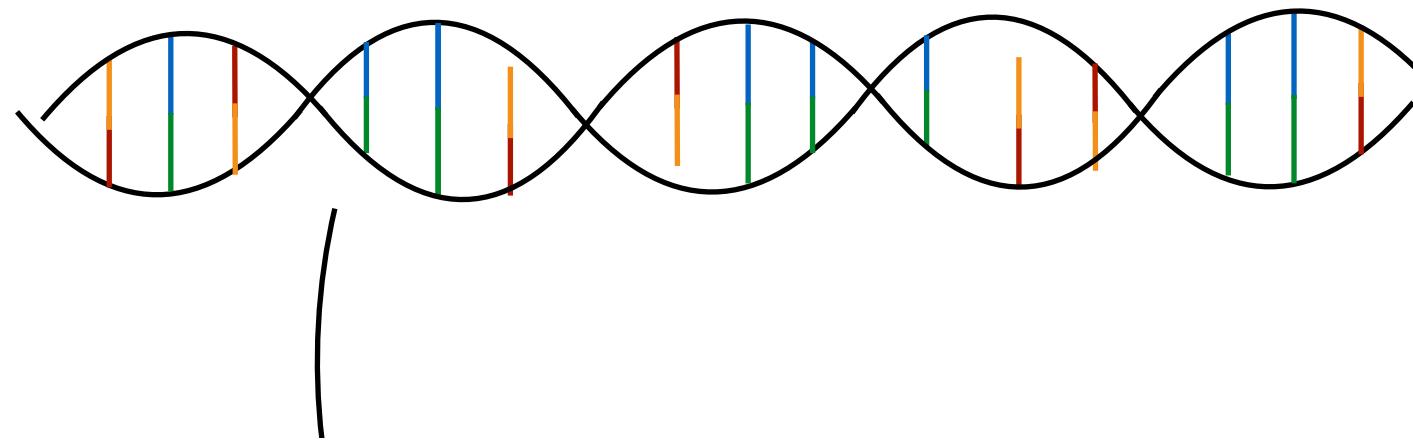
Some “processing” occurs **capping** & **polyadenylation**

Introns removed from pre-mRNA resulting in mature mRNA



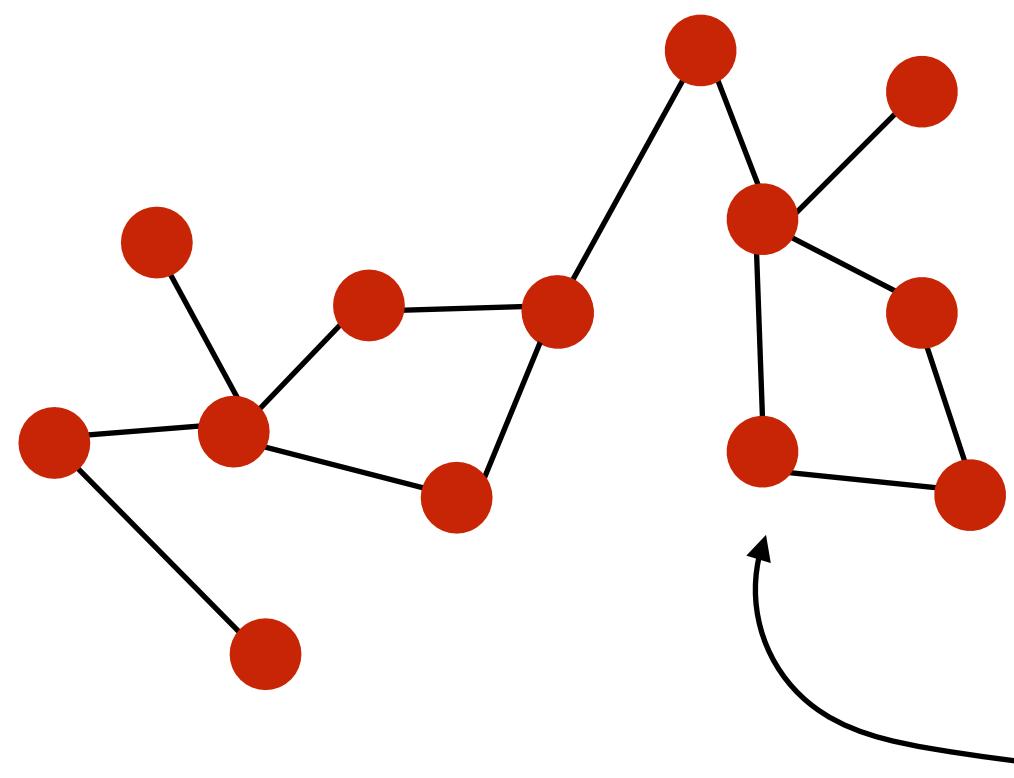
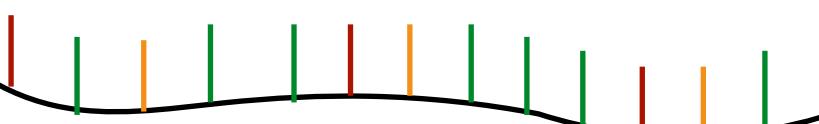
# “Flow” of information in the cell

DNA



RNA Polymerase  
(transcription)

RNA

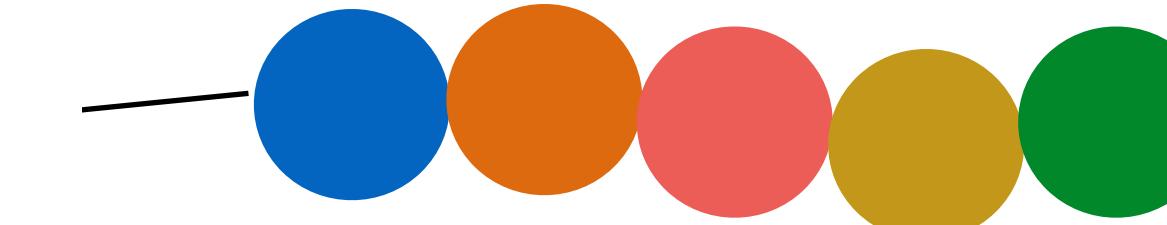


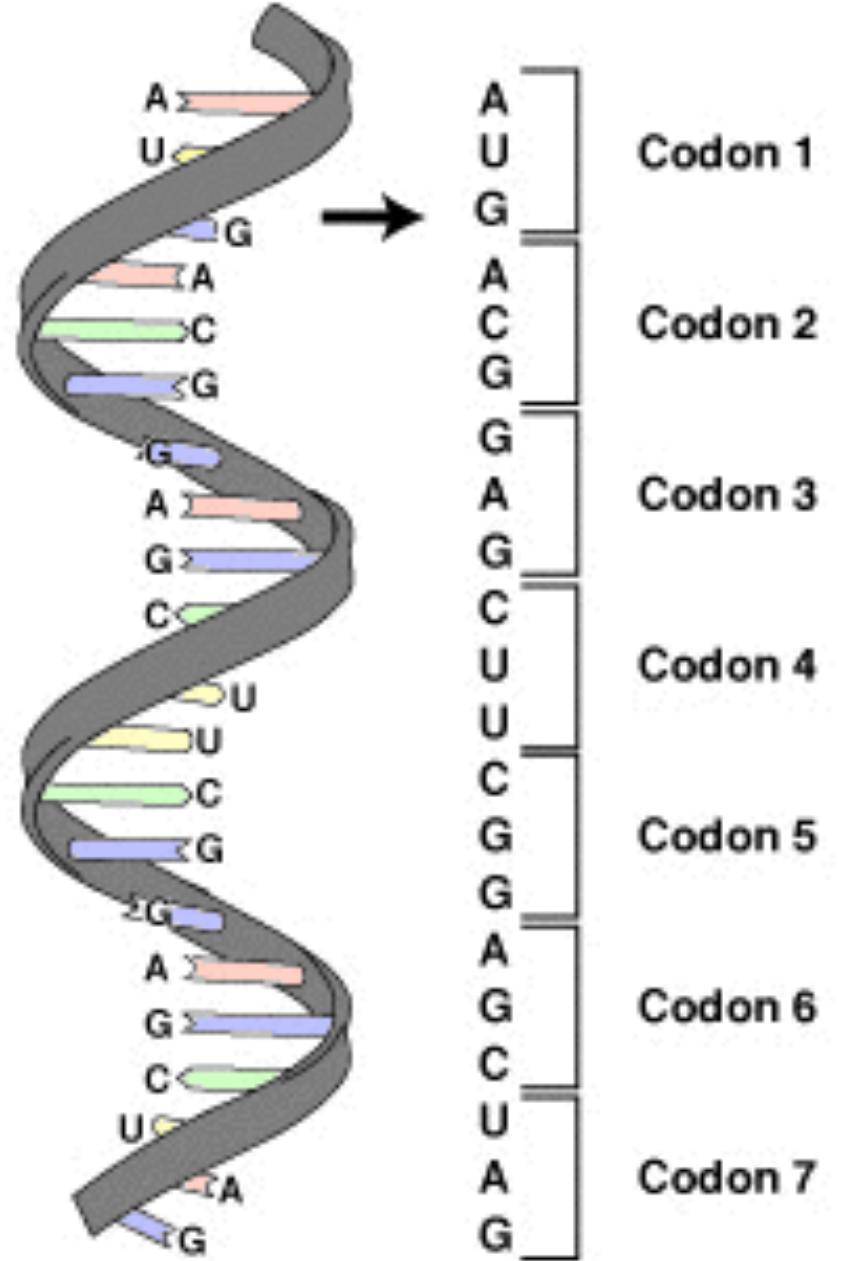
Form networks &  
pathways; perform a  
vast set of cellular  
functions

See video on course website

Ribosomes  
(translation)

Protein





# Protein

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

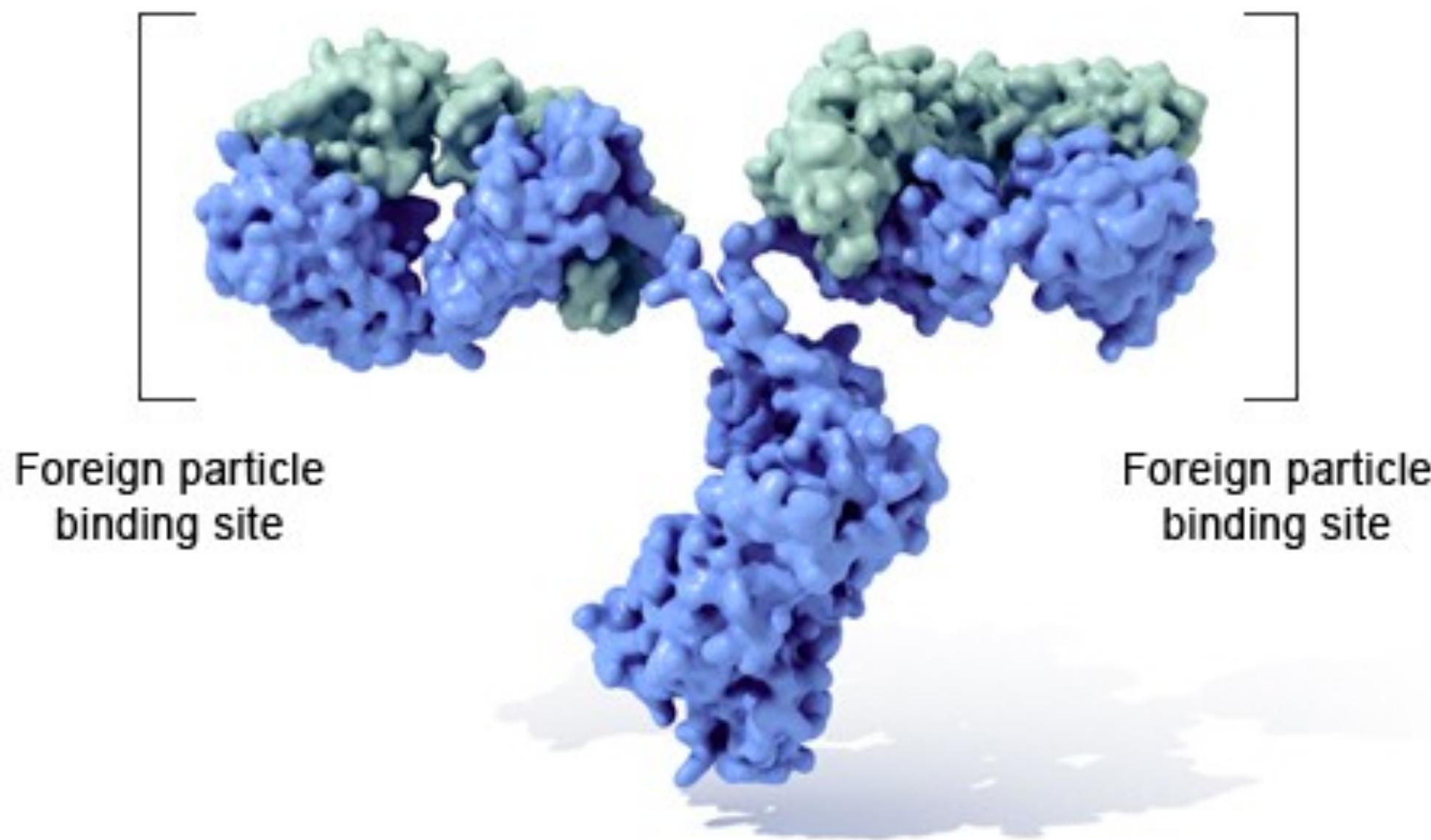
This mapping is known as the “genetic code” — an *almost* law of molecular Biology

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	UY
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
<b>START</b>	AUG		<b>STOP</b>	UAA, UGA, UAG	UAR, URA

# Protein

Immunoglobulin G (IgG)



U.S. National Library of Medicine

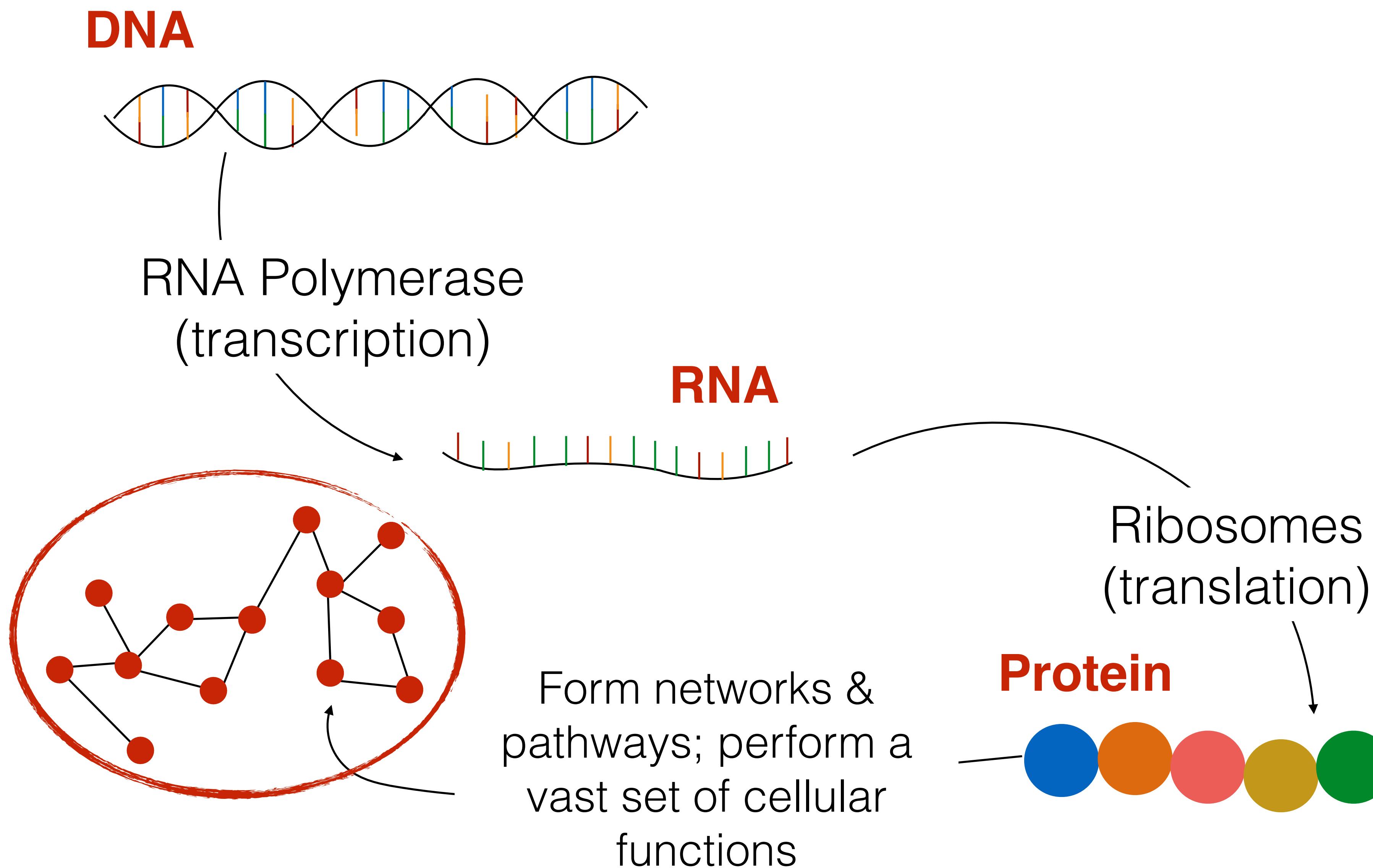
Perform vast majority of intra & extra cellular functions

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



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

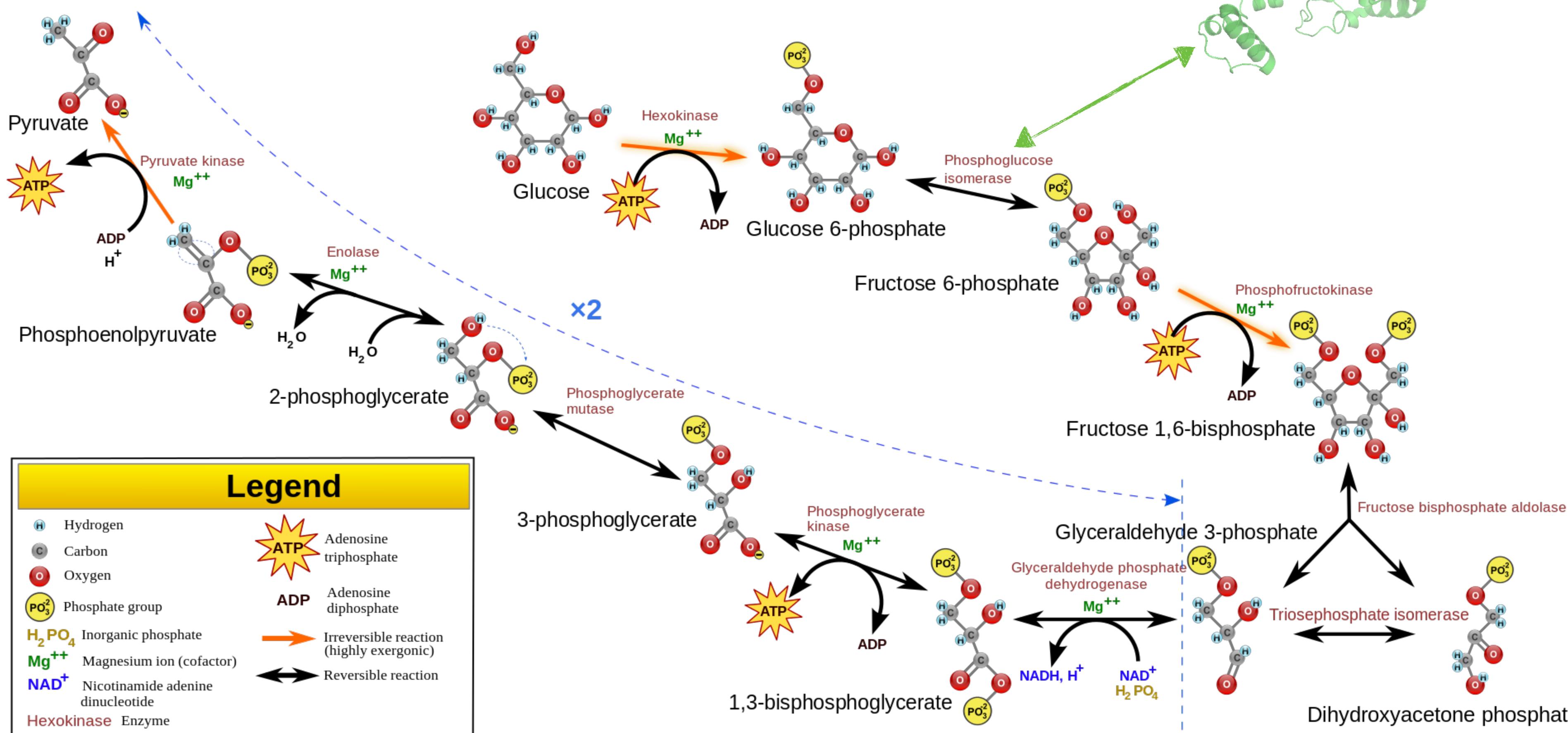
# Glycolysis Pathway

Converts glucose → pyruvate

phosphoglucone isomerase

Generates ATP ("energy currency" of the cell)

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



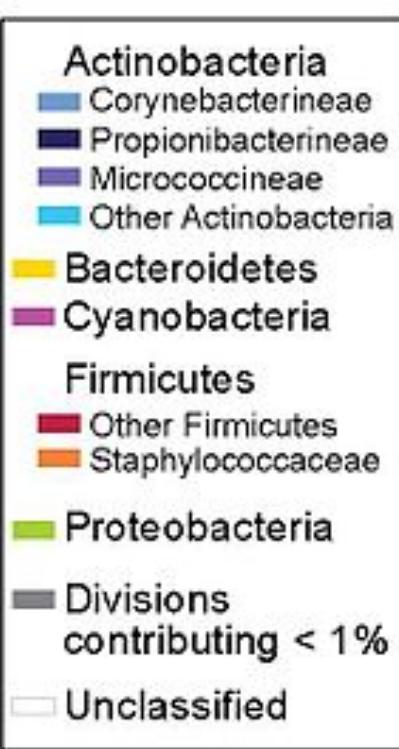
# Some Interesting Facts

Organism	Genome size	# of genes
$\phi$ X174 ( <i>E. coli</i> virus)	~5kb	11
<i>E. coli</i> K-12	~4.6Mb	~4,300
Fruit Fly	~122Mb	~17,000
Human	~3.3Gb	~21,000
Mouse	~2.8Gb	~23,000
<i>P. abies</i> (a spruce tree)	~19.6Gb	~28,000

No strong link between genome size & phenotypic complexity

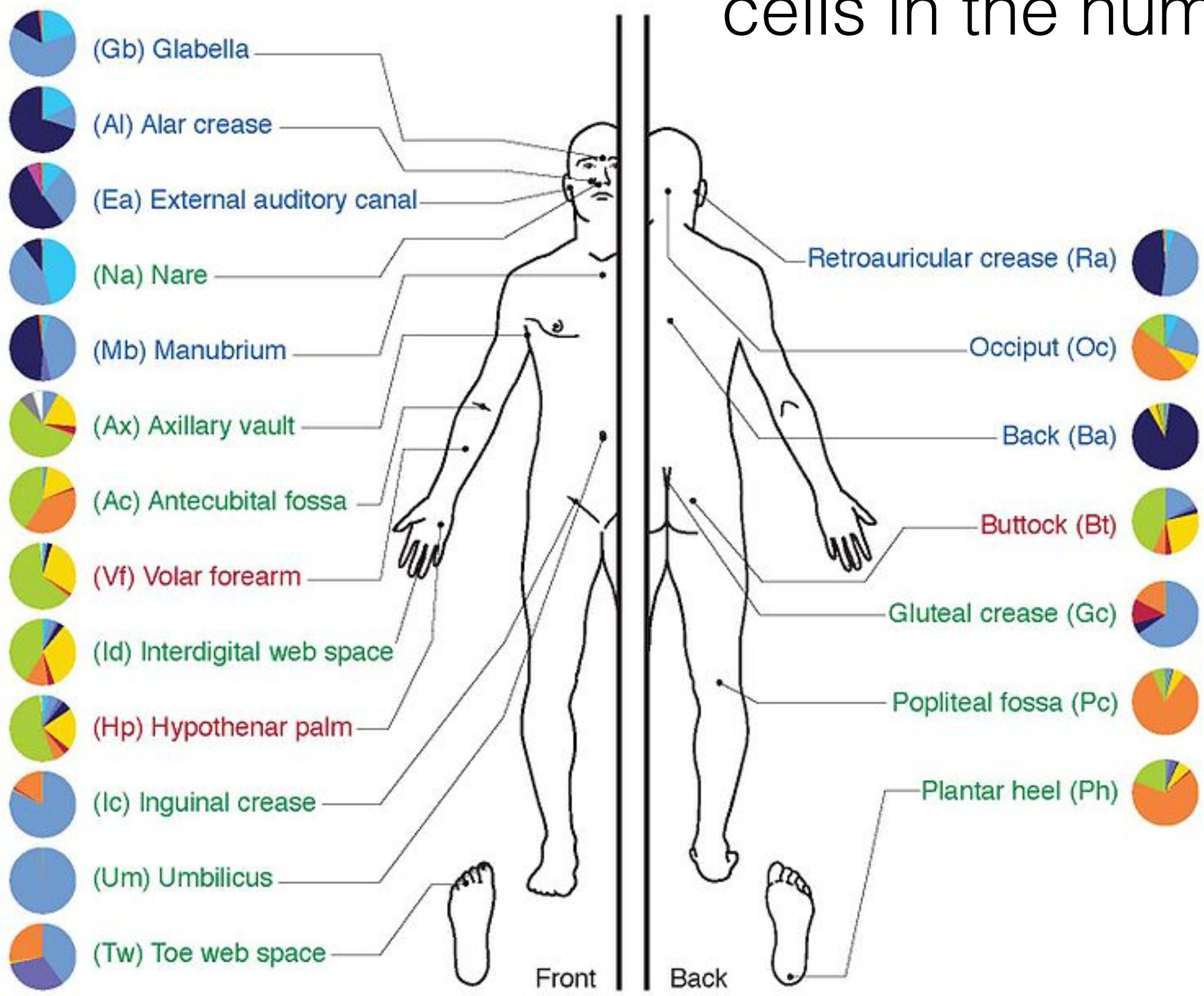
Plants can have **huge** genomes (adapt to environment while stationary!)

# Some Interesting Facts



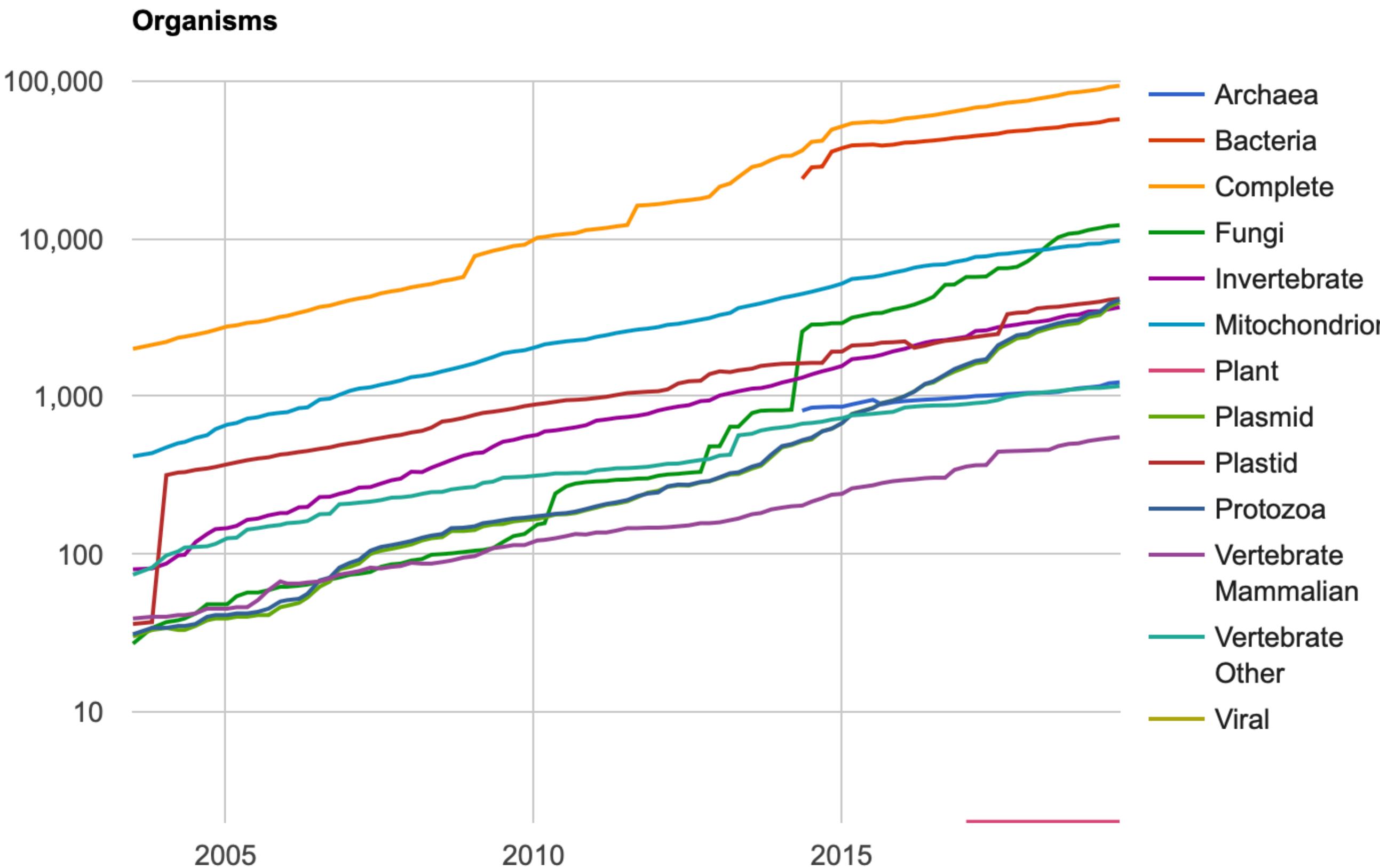
You are a good part non-human cells (e.g. bacteria)

Non-human cells equal or outnumber human cells in the human body



This population of organisms is called the microbiome

# Some Interesting Facts



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

. . . Out of  $8.7 \pm 1.3$  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.