2 Hidden Messages

23 October 2025

Prof. Dr. Sebastian Wild

CS594 (Winter 2025-26)
Philipps-Universität Marburg
version `2025-10-24 01:50` H

2.1 Biology Big Picture

Biology

- *biology* = the scientific study of *living* things
  - originally *naturalists*: individual people manually **observing** plants and animals
    e. g., *Darwin's finches*
  - gradually more scientific: controlled experiments, isolated mechanisms
    e. g., *Mendel's inheritance experiments on peas*
  - gradually more focus on molecular/chemical mechanisms: microscopes, biochemisty

- now clear: fundamental mechanisms (and origins!) of life are microscopic

⇝ fundamental mechanisms to be found in *molecular biology*

Bioinformatics

- 20th Century: discovery of DNA and genes
  - DNA stores information about biomolecules in **discrete form**
    human genome: 3.055 billion letter string over alphabet `{A, C, G, T}` (!)
  - ⇝ genetic information can **copied** precisely
    mutations are errors in the copying
  - double strands (backup!) and "coiling up" into chromosomes protects data
  - production of chemicals in living cells (*proteins*) is determined by *genes* (parts of DNA)

▶ Zoom in on DNA
https://youtu.be/wZoz0rFluiw

⇝ *Life itself has inherently **computational** components!* 🤯

⇝ Computer science can contribute to the understanding these! ⇝ *bioinformatics*

- But also: biology increasingly a data-centric field
  - much of knowledge discovery intrinsically reliant on computational analysis of collected data
  - e. g., reading the 3 billion letters of DNA is not possible with current lab techniques
    ⇝ use computers to puzzle it together (see *Sequencing Unit*)
  - "*in silico*" experiments

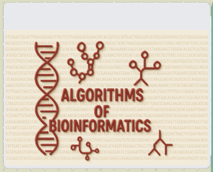
Collection of (more or less) Fun Sources

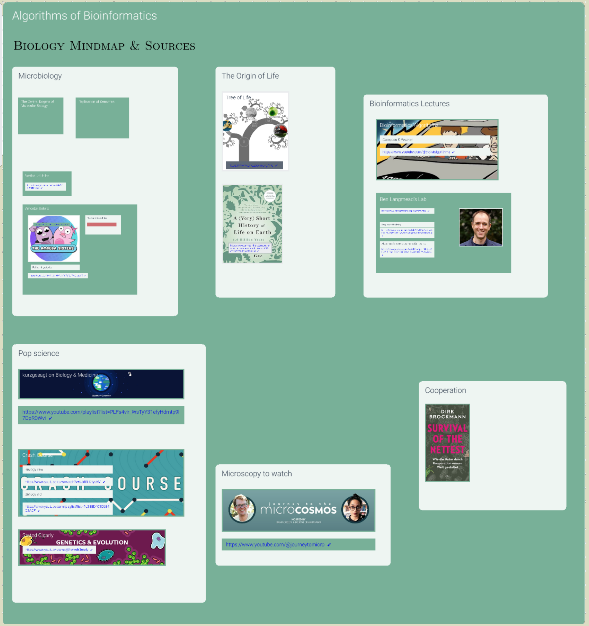
Collaborative Mindmap

on infinity maps

- ► Share useful resources
- ► Structure knowledge hierarchically
- ► Link on Campuswire / ILIAS

There's tons to learn,
new things discovered every day,
and it's about life itself!




Algorithms of Bioinformatics

Biology Mindmap & Sources

Microbiology

The Central Dogma of Molecular Biology

Replication of Genomes

Amoeba Sisters

The Origin of Life

Tree of Life

A (Very) Short History of Life on Earth
4.6 Billion Years
Gee

Bioinformatics Lectures

Bioinformatics Algorithms

Compeau & Pevzner

Ben Langmead's Lab

Pop science

kurzgesagt on Biology & Medicine

https://www.youtube.com/playlist?list=PLFs4vir_WsTyY31efyHdmtp9l7DpR0Wvi

Crash Course

Stated Clearly

GENETICS & EVOLUTION

Microscopy to watch

journey to the microcosmos

https://www.youtube.com/@journeytomicro

Cooperation

DIRK BROCKMANN
SURVIVAL OF THE NETTEST

Molecular Biology 101

Molecular Biology (Britannica concise)

- concerned with chemical structures and processes of biological phenomena at the molecular level
- developed out of biochemistry, genetics, and biophysics
- particularly concerned with the study of **proteins**, nucleic acids, and enzymes

Biology = lots of terminology and names . . .

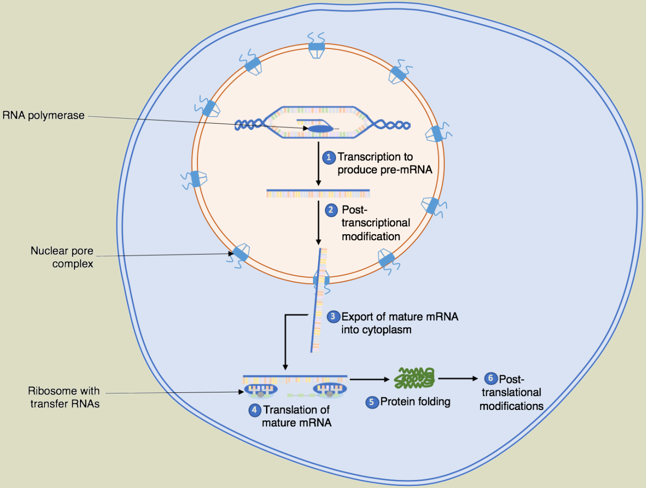
We will focus on mechanisms over terms, but a bit of context helps
let’s make it at least whimsical (and maybe memorable)

▶ Biomolecules (Updated 2023)
https://youtu.be/1Dx7LDwINLU

2.2 What are Genes?

The Central Dogma of Molecular Biology

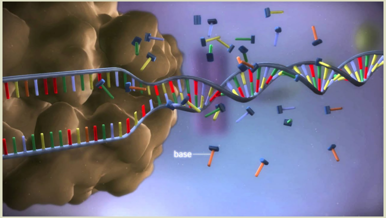
DNA makes **RNA** makes **Protein**


RNA polymerase
1 Transcription to produce pre-mRNA
2 Post-transcriptional modification
Nuclear pore complex
3 Export of mature mRNA into cytoplasm
Ribosome with transfer RNAs
4 Translation of mature mRNA
5 Protein folding
6 Post-translational modifications

https://commons.wikimedia.org/wiki/File:Summary_of_the_protein_biosynthesis_process.png

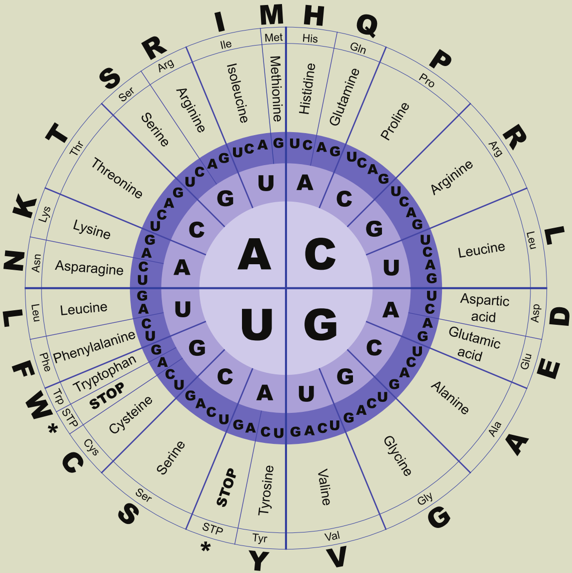
Protein Biosynthesis

- mechanism to produce *protein* a according to recipe stored in a *gene*



From DNA to protein - 3D
https://youtu.be/gG7uCskUOrA

Genetic Code



Compeau & Pevzner, *Bioinformatics Algorithms*, Fig. 4.1
`https://cogniterra.org/lesson/29910/step/2?unit=22007`

Within *ribosomes* (protein factories)

- translation
  - from RNA bases `{A,C,G,U}`
  - to amino acids (peptide) $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$
- uses *transfer RNA* *"chemical finite state transducer"*
- *Genetic Code*: 3-base *codons* → amino acid

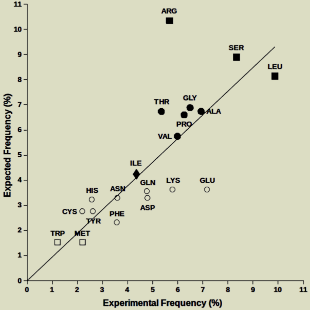
Inverse Codon Table

#Codons	Amino Acid (abbr.)		Codons
1	*Start*	>	AUG
4	Ala	*A*	GCU GCC GCA GCG
2	Cys	*C*	UGU UGC
2	Asp	*D*	GAU GAC
2	Glu	*E*	GAA GAG
2	Phe	*F*	UUU UUC
4	Gly	*G*	GGU GGC GGA GGG
2	His	*H*	CAU CAC
3	Ile	*I*	AUU AUC AUA
2	Lys	*K*	AAA AAG
6	Leu	*L*	CUU CUC CUA CUG UUA UUG
1	Met	*M*	AUG
2	Asn	*N*	AAU AAC
4	Pro	*P*	CCU CCC CCA CCG
2	Gln	*Q*	CAA CAG
6	Arg	*R*	CGU CGC CGA CGG AGA AGG
6	Ser	*S*	UCU UCC UCA UCG AGU AGC
4	Thr	*T*	ACU ACC ACA ACG
4	Val	*V*	GUU GUC GUA GUG
1	Trp	*W*	UGG
2	Tyr	*Y*	UAU UAC
3	*Stop*	<	UAA UAG UGA
1	Sec	*U*	(UGA)
1	Pyl	*O*	(UAG)

Some amino acids have several codons
(most frequent amino acids receive strongest error protection!)

Sometimes, stop codon UGA instead codes 21st amino acid *Selenocystein*. . .

Amino Acid Frequencies in Human Proteins



https://doi.org/10.1371/journal.pone.0148174.g001

But:

- non-ribosomal peptides (proteins not made according to central dogma)
- epigenetics (which genes are expressed)
- horizontal gene transfer (change genome during lifetime)
- retro viruses (inserts its one genes into host's genome!)
- proteins are also not the only active molecules (e. g., functional RNA)

Life finds a way . . . or a few dozen, just to be sure

2.3 Gene Detection

How can we find genes?

Recall: Gene = protein-coding region of DNA

Central options:

1. *ab initio:* Just using the DNA
   - search for start and stop codons (base triples) ⇝ *open reading frame*
   - search for promoter binding sites (docking station for transcription molecules)
   - bias of base frequencies in coding vs non-coding regions
2. extrinsic methods: using additional (lab) data
   - e. g.sequencing messenger RNA from live cells (many more options)
   - comparison of genome to other species with known genes

Focus for today: Ab initio options

Why should there be any hope of finding hidden messages?

- Evolution!
- Random mutations always at play
- If functional part becomes dysfunctional, individual does not produce offspring
- other parts might be subject to random modifications

⇝ *signal*: property in a text that us unlikely to be present in random strings (noise)

⇝ noise / null model: unused DNA is random

2.4 Frequent words

Random strings

Expected number of occurrences of a *k*-mer

Expected *distance* of occurrences