ALGORITHMS OF BIOINFORMATICS

2

Hidden Messages

23 October 2025

Prof. Dr. Sebastian Wild

CS594 (Winter 2025-26)
Philipps-Universität Marburg
version 2025-10-30 12:02

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

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/wZozOrFluiw

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/wZozOrFluiw`

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

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/wZozOrFluiw`

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

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

Bioinformatics

Zoom in on DNA
https://youtu.be/wZozOrFluiw

- 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)

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

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

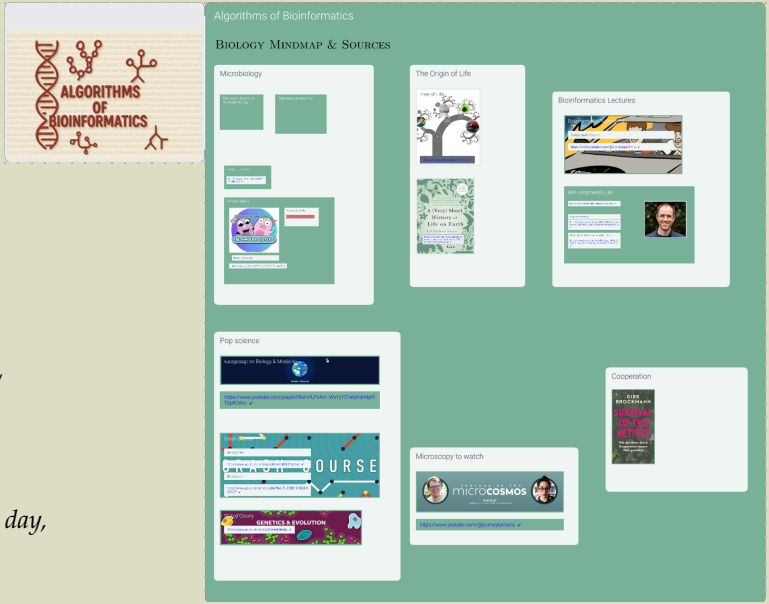
- But also: biology increasingly a data-centric field
  - much of knowledge discovery intrinsically reliant on computational analysis of collected data
  - e. g., reading the 3 billion letters of DNA is not possible with current lab techniques
    - ⇝ use computers to puzzle it together (see *Sequencing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



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

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 . . .

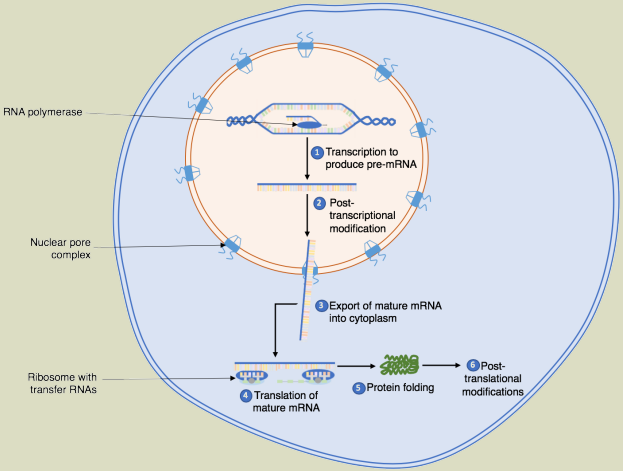
We will focus on mechanisms over terms, but a bit of context helps
let's make it at least whimsical (and maybe memorable)

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

2.2 What are Genes?

The Central Dogma of Molecular Biology

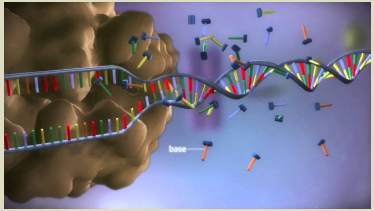
DNA makes **RNA** makes **Protein**



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

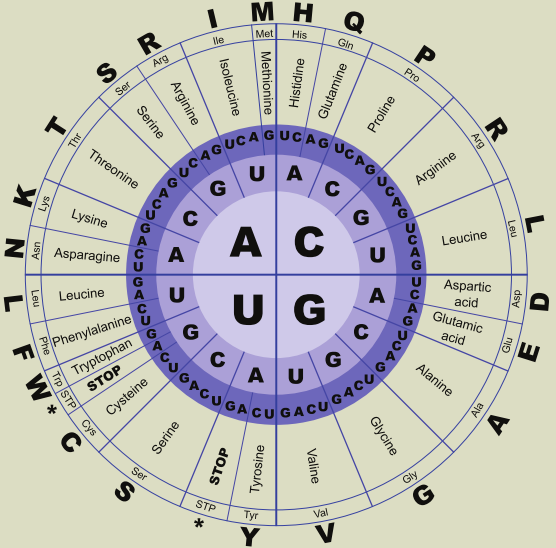
Protein Biosynthesis

- 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

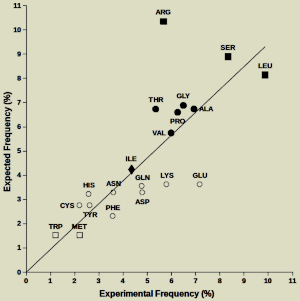
Inverse Codon Table

#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)

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)

Amino Acid Frequencies in Human Proteins



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

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

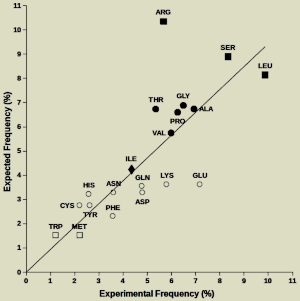
Inverse Codon Table

#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* ("from the beginning"): just using the DNA
   - search for start (AUG) and stop codons (UAA, UAG, UGA) ⇝ *open reading frame*
   - search for promoter binding sites (docking station for transcription molecules)
   - bias of base frequencies in coding vs non-coding regions (hidden Markov models)
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?

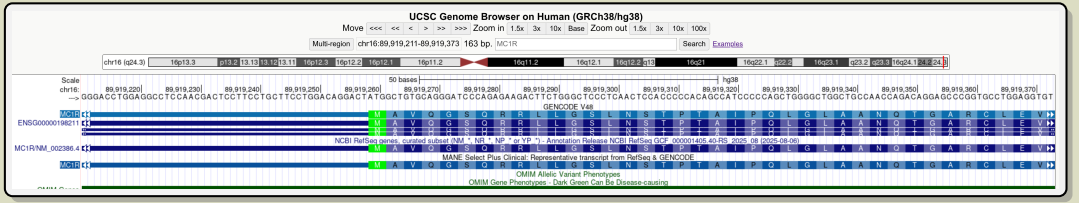
- Evolution!
- Random mutations always at play
- If functional part becomes dysfunctional, individual does not produce offspring
- other parts might be subject to random modifications

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

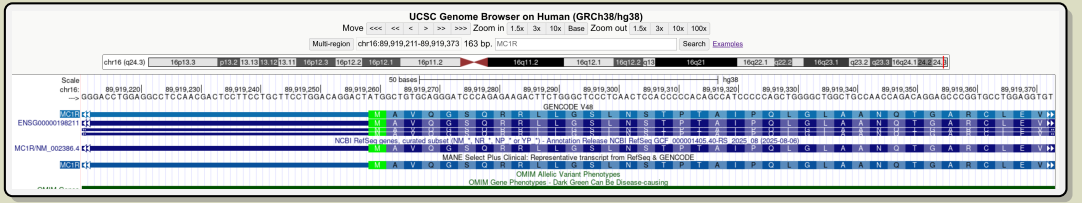
⇝ noise / *null model:* unused DNA is random

2.4 Waiting for Words

How big are genes?



How big are genes?

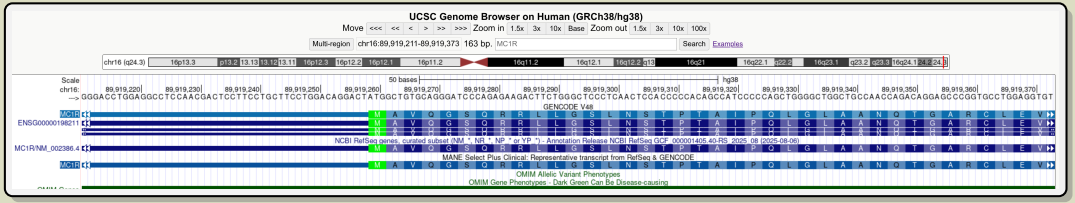


- only ~10% of human genome are genes
- length of (human) genes highly variable

base pairs (DNA strands!)

  - shortest known gene (*U7 snRNA*) has only 63 bp
  - longest gene (dystrophin) over 2M bp
    - but: 99% of that are *introns* (cut out before translation)! "split genes"
    - ⇝ transcription takes several hours (!)
  - moretypical: ~20K bp

How big are genes?



- only ~10% of human genome are genes
- length of (human) genes highly variable
  - shortest known gene (*U7 snRNA*) has only 63 bp
  - longest gene (dystrophin) over 2M bp
    - but: 99% of that are *introns* (cut out before translation)! "split genes"
    - ⇝ transcription takes several hours (!)
  - moretypical: ~20K bp

base pairs (DNA strands!)

⇝ *Can we reliably distinguish genes from randomly occurring open reading frames?*

Open Reading Frame Gene Detection

- ***Random RNA Model:*** String $D[0..N)$ generated i.i.d. uniformly
  i.e., each $D[i] \overset{\mathcal{D}}{=} \text{Uniform}(\{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\})$
- ***Random Open Reading Frame:*** How many bp should we expect in **random RNA** between occurrences of the start codon `ATG` and **first** occurrence of any stop codon (`TAA`, `TAG`, `TGA`)?

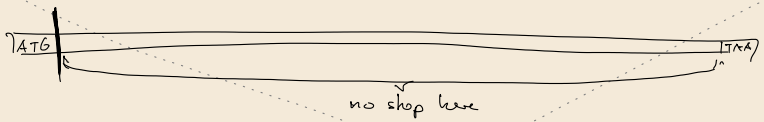
  (Recall: `U` in mRNA is `T` in DNA)

Clicker Question

In a string of random (i.i.d. uniform) DNA, what is the expected length of an open reading frame?

→ *sli.do/cs594*

Back-of-the-envelope

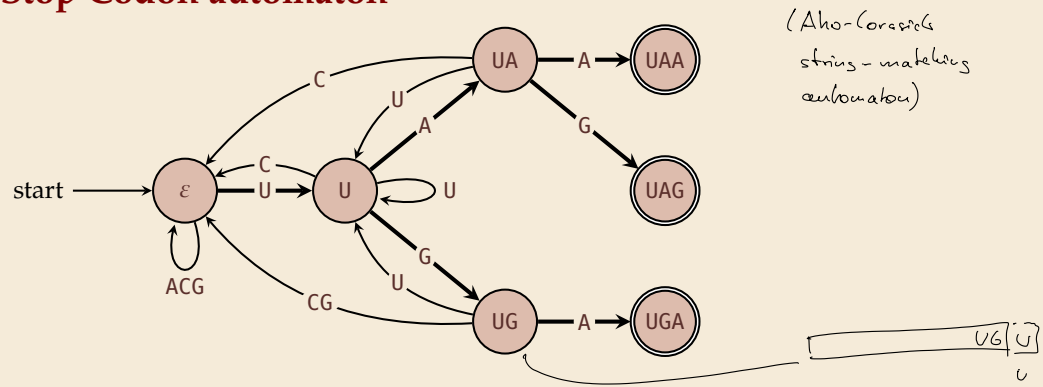


$$P\left[\boxed{?}\boxed{?}\boxed{?} \in \{TAA, TGA, TAG\}\right] = \frac{3}{64} \approx 0.05$$

maybe $\simeq$ 20 codons before stop

$\simeq$ 60 bp

Stop Codon automaton

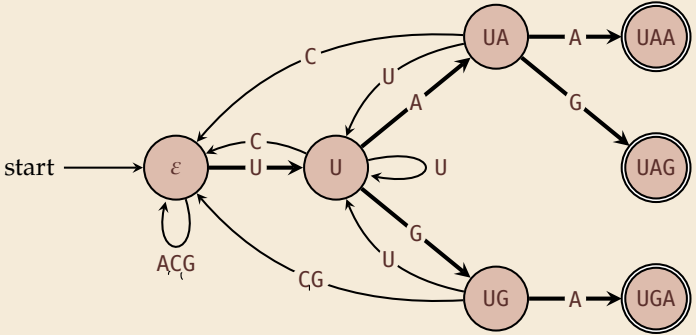


(Aho-Corasick string-matching automaton)

After seeing a start codon AUG, we accept the language of all strings that

- end with a stop codon **and**
- do not contain a stop codon earlier.

Stop Codon automaton

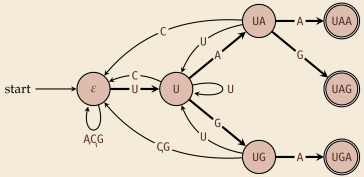


Expected number of characters from state q to some accepting state. (We want q_ε)

T_x := expected # chars from q_x to ◎

After seeing a start codon AUG, we accept the language of all strings that

- end with a stop codon **and**
- do not contain a stop codon earlier.



$$T_{UAA} = T_{UAG} = T_{UGA} = 0$$

$$T_{UA} = \frac{1}{4} \cdot T_{UAA} + \frac{1}{4} T_{UAG} + \frac{1}{4} T_U + \frac{1}{4} T_\varepsilon + 1$$

$$T_{UG} = \frac{1}{4} T_{UGA} + \frac{1}{4} T_U + \frac{1}{2} T_\varepsilon + 1$$

$$\vdots$$

$$T_\varepsilon = \frac{64}{3} = 21.\overline{3} \quad << 60$$

2.5 Probability Generating Functions

Probability Generating Functions

Expected values do not tell the full story . . . can we get at the distribution?

Probability Generating Functions

Expected values do not tell the full story . . . can we get at the distribution?

Definition 2.1 (pgf)

For $X \in \mathbb{N}_{\geq 0}$ a random variable, define its *probability generating function (pgf)* as

chars

$$G_X(z) = \sum_{k \geq 0} \mathbb{P}[X = k] \cdot z^k$$

Probability Generating Functions

Expected values do not tell the full story . . . can we get at the distribution?

Definition 2.1 (pgf)

For $X \in \mathbb{N}_{\geq 0}$ a random variable, define its *probability generating function (pgf)* as

$$G_X(z) \;=\; \sum_{k \geq 0} \mathbb{P}[X = k] \cdot z^k$$

$$\mathbb{E}[X] = \sum_{k \geq 0} \mathbb{P}[X=k] \cdot k$$

$$G_X'(z) = \sum_{k \geq 0} \frac{d}{dz} \mathbb{P}[X=k] \cdot z^k$$

$$= \mathbb{P}[X=k] \cdot k \cdot z^{k-1}$$

Lemma 2.2 (Moments from pgf)

1. The expected value of X is $\boxed{\mathbb{E}[X] \;=\; G_X'(1)}$
2. The variance of X is $\boxed{\mathrm{Var}[X] \;=\; G_X''(1) + G_X'(1) - \left(G_X'(1)\right)^2}$