COS 126

Introduction to Computational Biology

December 1, 2025

Outline

- What is computational biology and bioinformatics?
- Computational genomics:
  - Central dogma of molecular biology – hardware and software of life
  - Computational study of biological sequences
  - Finding hidden messages in DNA and protein sequences
  - Sequence comparison algorithms to study evolution and beyond
  - CRISPR for “programmable” genome editing
- Computational biology beyond sequence analysis

Goal of today’s lecture: give general (non-comprehensive) overview, provide context, pose computational challenges

Life Sciences and Biomedicine

Life Sciences: Scientific studies of living organisms:
- Biology
- Biochemistry
- Anatomy
- Physiology
- Neuroscience
- Zoology
- Ecology
- Anthropology, etc.

https://medium.com/@Innoplexus/the-life-science-revolution-f24a00d2ca0

Across many dimensions:
- Molecules, cells, organs, organisms, populations, environments
- Milliseconds to millions of years
- Across domains of life

Biomedicine:
- Biological science to improve human health and clinical practice
- Prevent, diagnose, control and treat diseases, modify human health

Data and computation in life sciences and biomedicine

Data: New technologies enable generation and collection of massive amounts of data

Computation: to collect, store, analyze, visualize, interpret data in biomedicine

Computer science in life sciences and biomedicine:

- Software development and compute architectures for data collection, storage, handling
- Algorithms for efficient computation
- Statistics and machine learning for biomedical data analysis and modeling

https://datamanagement.hms.harvard.edu/about/what-research-data-management/biomedical-data-lifecycle

Computational biology and bioinformatics

Computational biology and *bioinformatics* – related, often interchangeable terms

Computational biology: A field of *biology* concerned with the development of techniques for the collection and manipulation of biological data, and the use of such data to make biological discoveries or predictions.

Focus: *biology*

NIH/NLM glossary: https://www.ncbi.nlm.nih.gov/mesh/68019295

https://www.scripps.edu/science-and-medicine/cores-and-services/center-for-computational-biology-bioinformatics/index.html

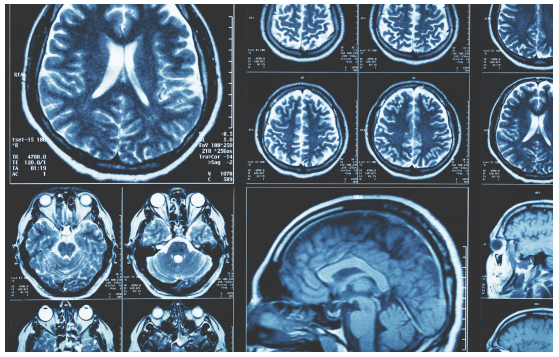
Bioinformatics: A scientific discipline that involves using computer technology to collect, store, analyze and disseminate biological data and information.

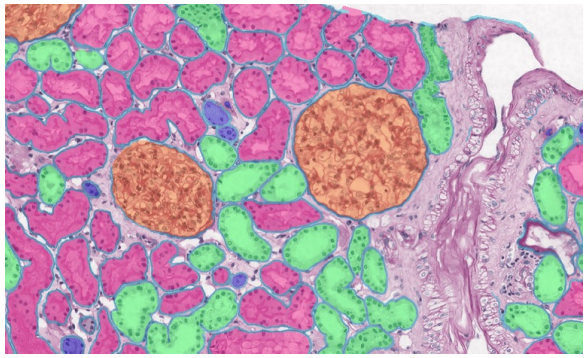
Focus: *informatics*

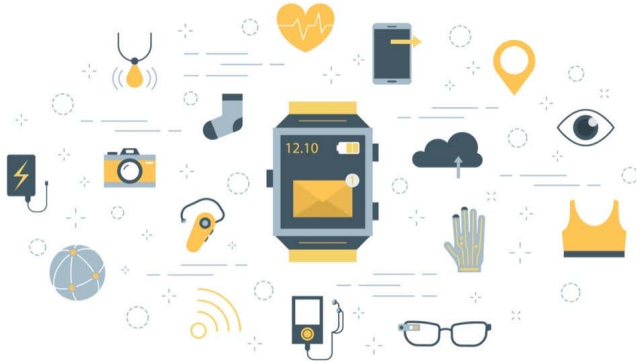
NIH/NHGRI glossary: https://www.genome.gov/genetics-glossary/Bioinformatics

Computer science is essential for progress in biomedicine

- Genomic data analysis and genome editing – *today's lecture*
- Medical imaging
- Personalized medicine
- Drug and therapy discovery
- Genetic forensics
- Robotics in surgery
- Telemedicine
- Wearable devices and biohacking



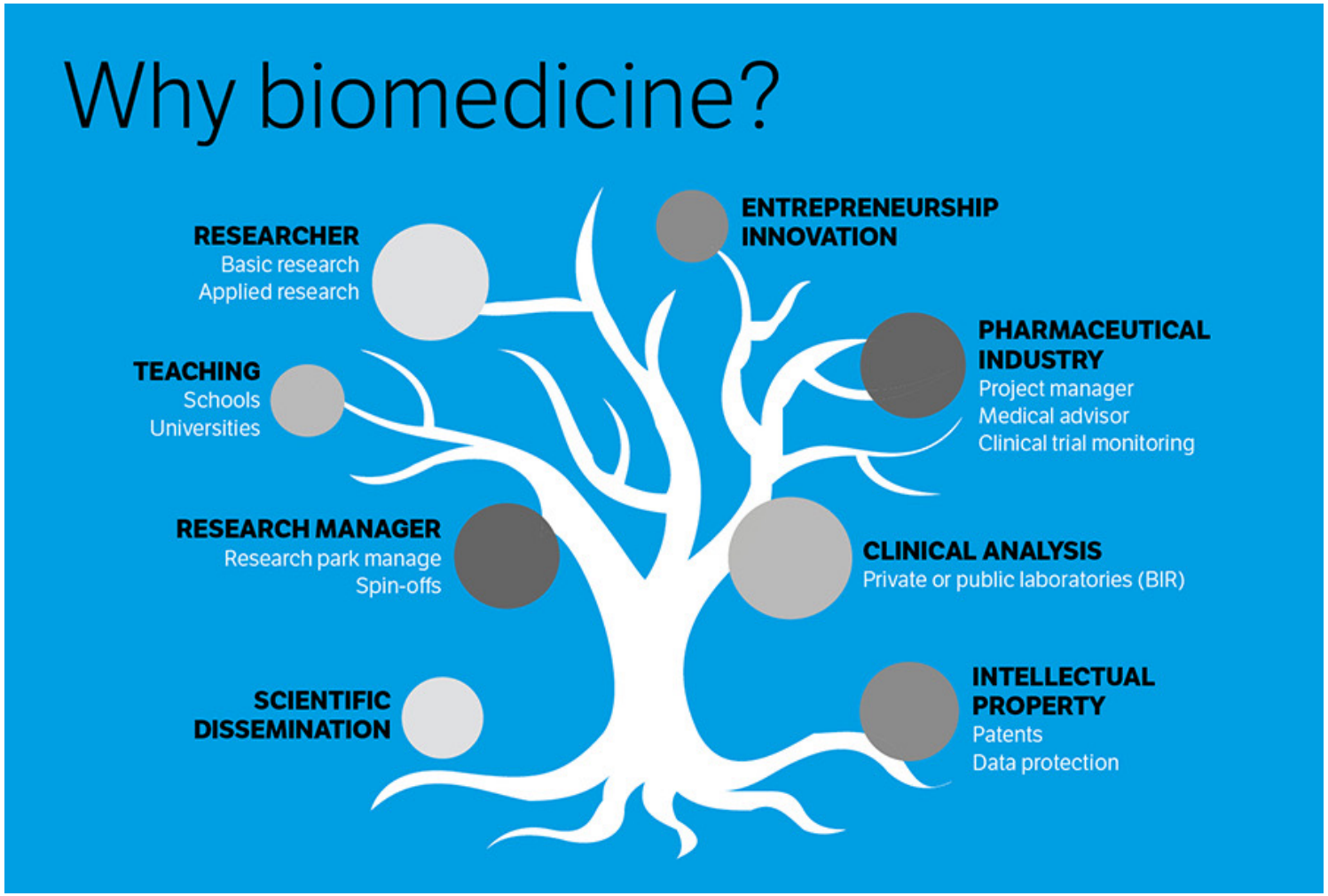




https://newsinhealth.nih.gov/2013/12/personalized-medicine
https://www.computationalpathologygroup.eu/news/nat_med_comp_path_review/
https://www.bme.jhu.edu/research/research-areas/imaging-and-medical-devices/
https://journal.getabstract.com/en/2021/03/04/biohacking-for-beginners
https://www.online-tech-tips.com/gadgets/7-coolest-wearable-electronics-you-need-to-have/
https://www.nytimes.com/2021/08/16/well/live/robotic-surgery-benefits.html

Many career paths in computational biology and bioinformatics

- Scientific research
- Teaching
- Software development
- Pharmaceutics
- Clinical work
- Biotechnology
- Consulting
- Editing
- Management
- Startups and investing



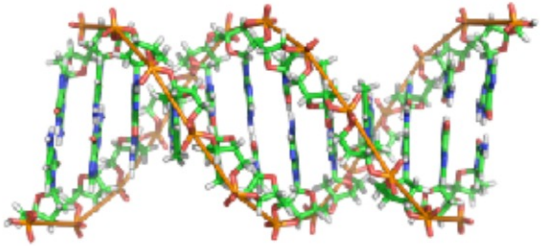
https://www.uic.es/en/estudis-uic/medicine-and-health-sciences/bachelors-degree-biomedical-science-biomedicine

Computational genomics

Genome

"The complete set of genes or genetic material present in a cell or organism."

Oxford dictionaries



"Blueprint" or "recipe" of life

TAGCCCGACTTG

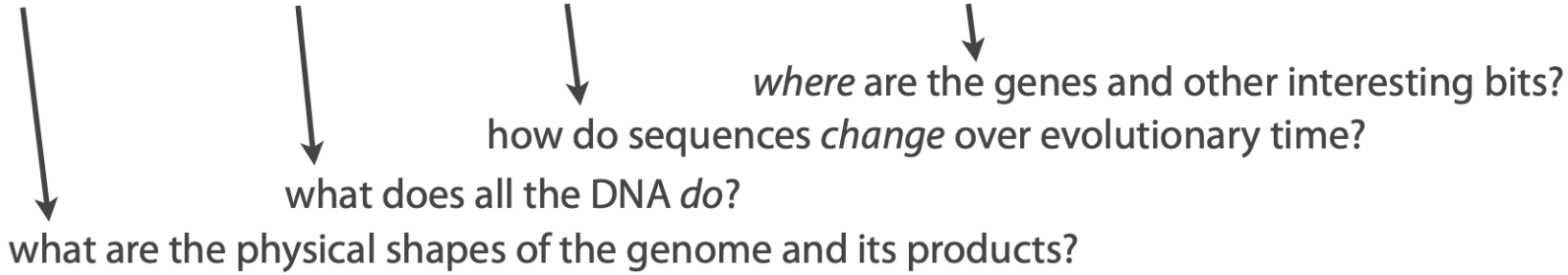
Self-copying store of read-only information about how to develop and maintain an organism

From teaching materials of Ben Langmead: https://langmead-lab.org/teaching-materials/

Genomics

Oxford dictionaries

"The branch of molecular biology concerned with the **structure**, **function**, **evolution**, and **mapping** of **genomes**."



Collins English Dictionary

"The branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes and the **application** of this knowledge in **medicine**, **pharmacy**, **agriculture**, etc."

From teaching materials of Ben Langmead: https://langmead-lab.org/teaching-materials/

Computational genomics

Addresses crucial problems at the intersection of genomics and computer science

The intersection:

Key biological models are straight out of computer science: **circuits** and **networks** for molecular interactions, **trees** for evolution and pedigrees, **strings** for DNA, RNA and proteins

Thanks to sequencers and microarrays, research bottlenecks increasingly hinge on computational issues: **speed**, **scalability**, **energy**, **cost**

With large, noisy, biased high-throughput datasets comes a critical need for **machine learning** and **statistical reasoning**

From teaching materials of Ben Langmead: https://langmead-lab.org/teaching-materials/

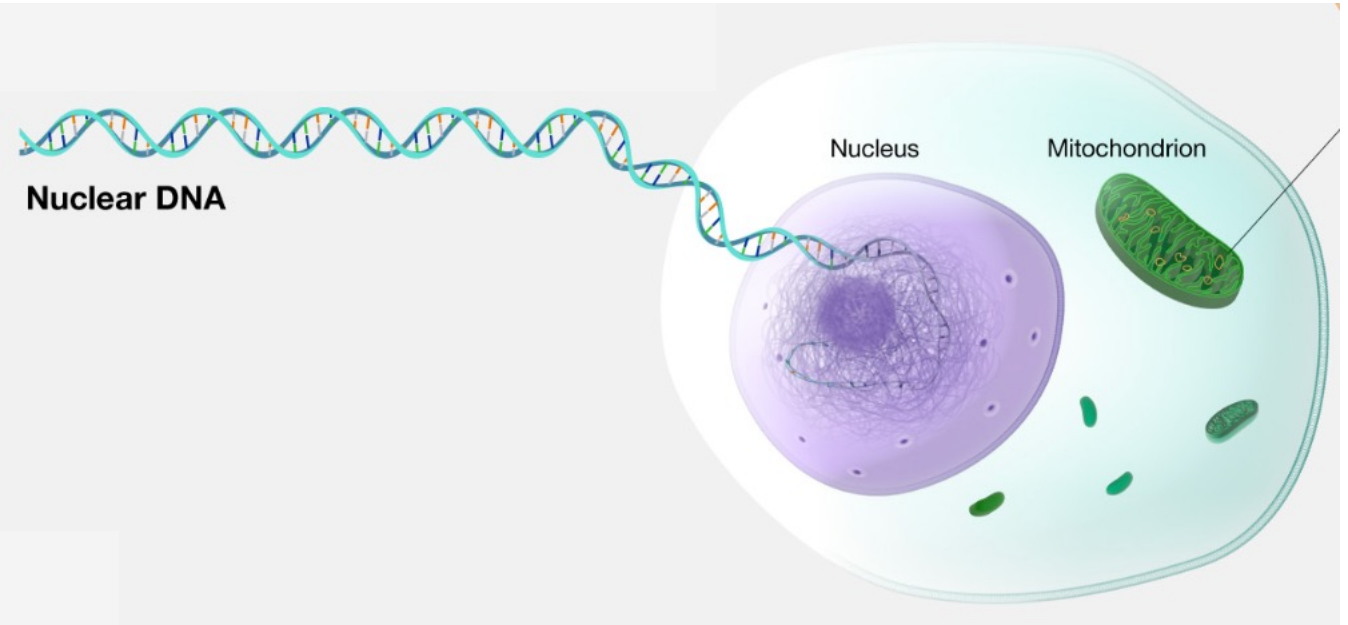
Central dogma of molecular biology – hardware and software of life

Molecular biology of the cell

- Cell is the basic functional unit of life

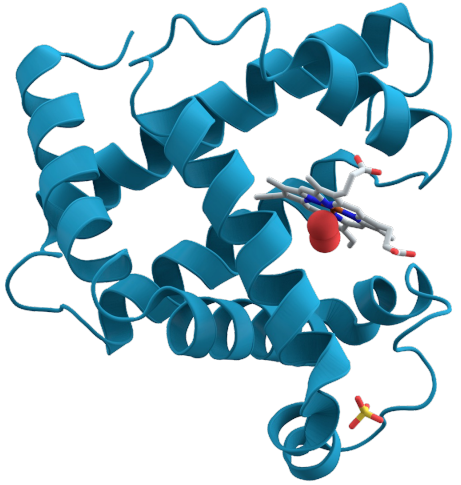
- Macromolecules DNA, RNA, proteins are the main molecules of life, define structure and function of cells

- Genome is all DNA in the cell



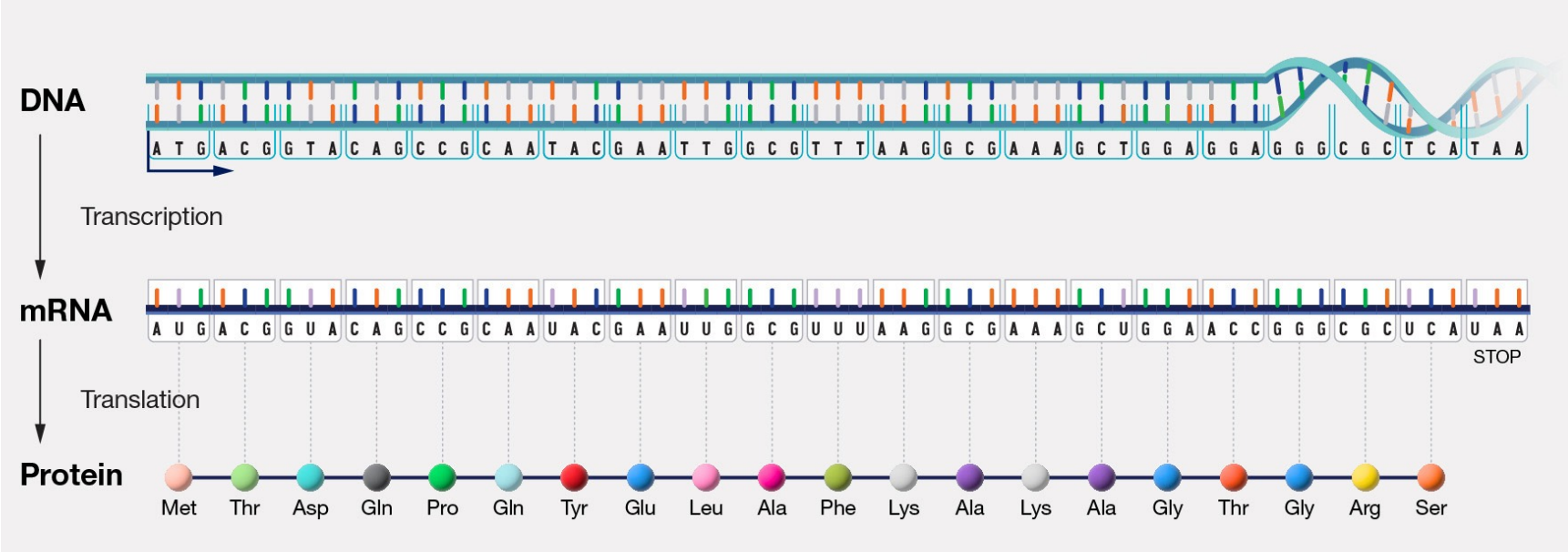
DNA

protein



https://www.genome.gov/genetics-glossary/
https://en.wikipedia.org/wiki/Protein#/media/File:Myoglobin.png

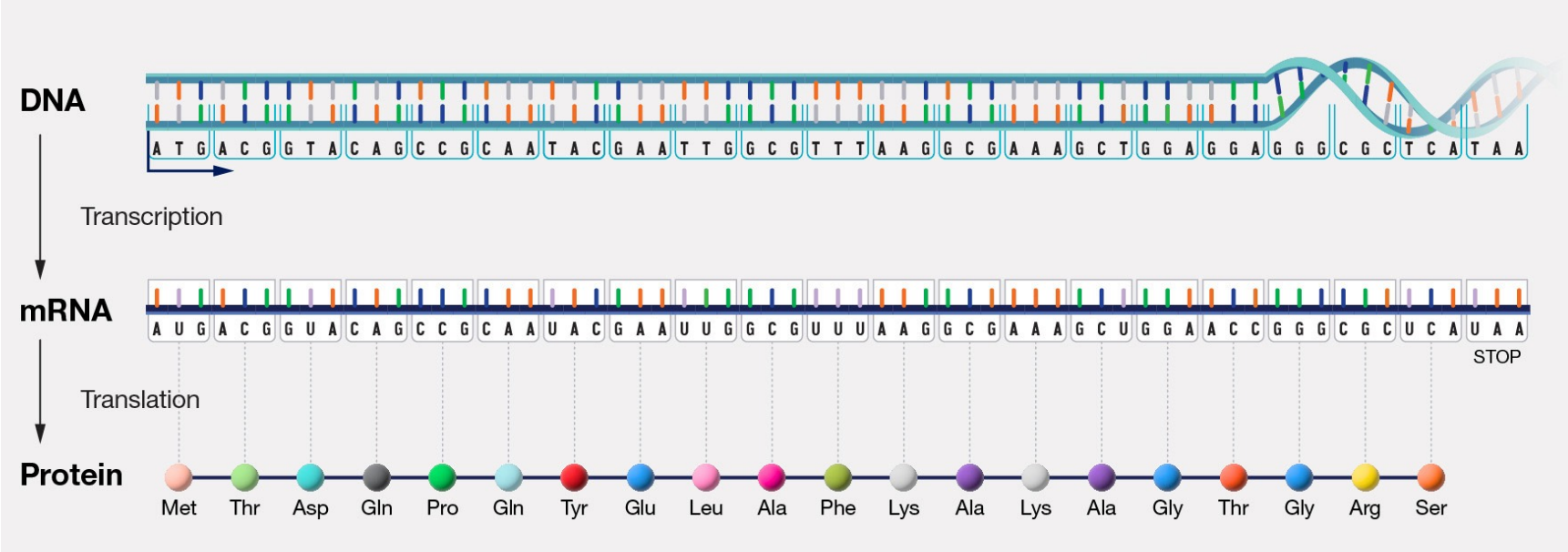
Biological sequences



https://www.genome.gov/genetics-glossary/

- Stereotypical organization of macromolecules:
- DNA is a sequence in a 4-letter alphabet of nucleotides: `A C G T`
- RNA is a sequence in a 4-letter alphabet of nucleotides: `A C G U`
- Protein is a sequence in a 20-letter alphabet of amino acids: `Met Thr Asp` … or `M T D` …
- For example, Human genome: 3.2 billion nucleotides, 20,000 protein-coding genes

Central dogma of molecular biology



https://www.genome.gov/genetics-glossary/

- Surprising analogy with engineered computer systems: ”Hardware” and “software” of life
- DNA: “hard drive” – long-term storage of information about life
- RNA: intermediate, temporary storage, “memory” or “buffer”
- Proteins: main functional and structural molecules for nearly everything in the cell, “executable”
- DNA transcribes to RNA, RNA translates to protein – “software”

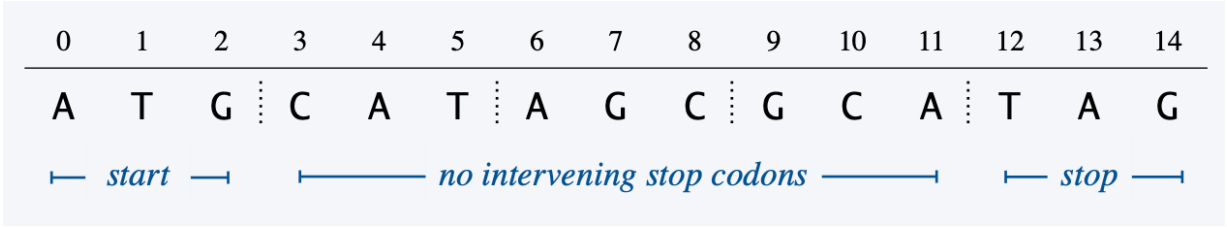
Finding hidden messages in DNA and protein sequences

Find patterns in biological sequences

- Biological processes depend on patterns in biological sequences

- **CS challenge:** given a pattern, find (likely) instances of this pattern in DNA sequence

- Example (review):



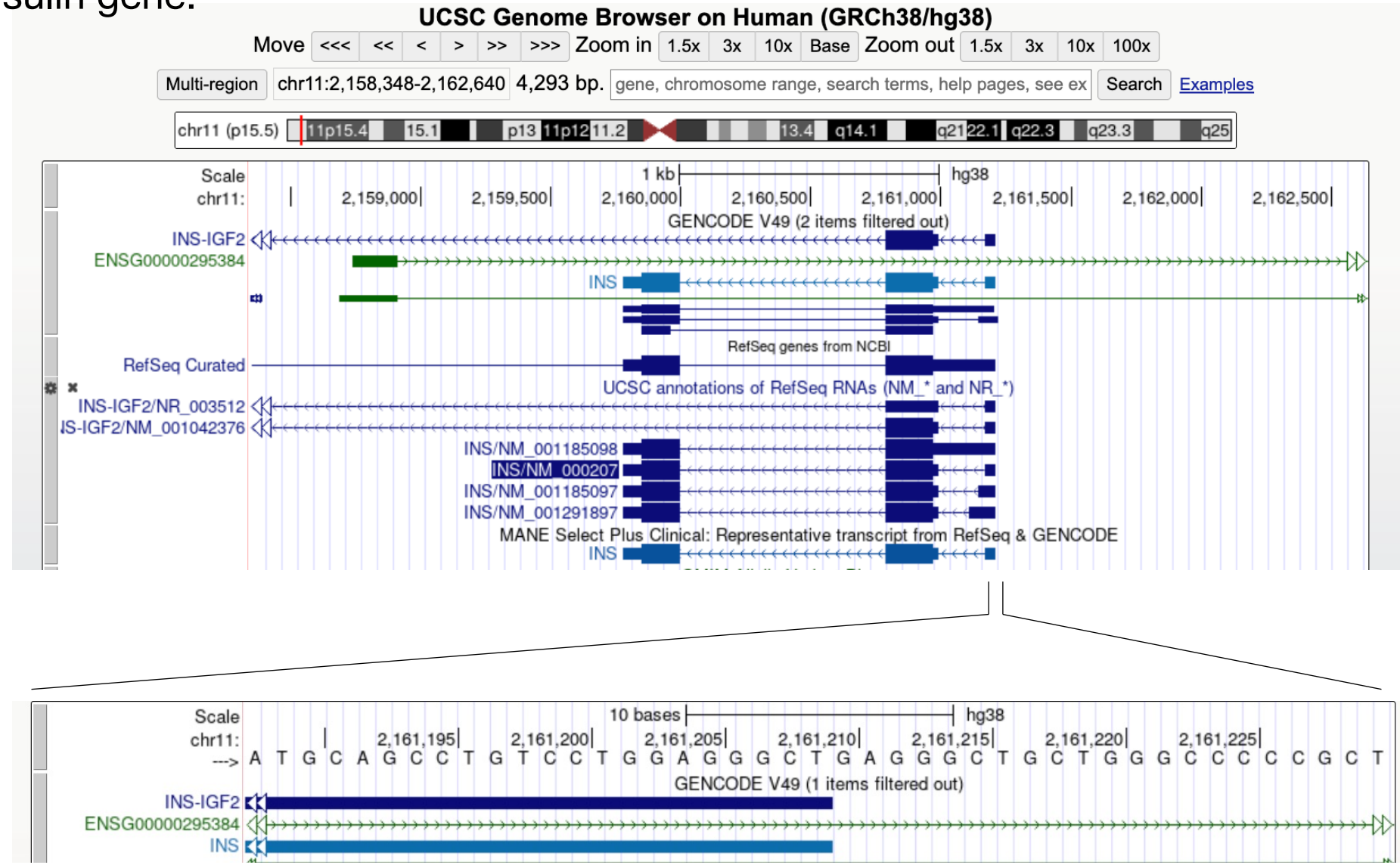
A gene: a substring of genomic DNA that represents a functional unit, i.e. encodes a protein
  - Made up of codons (three `A C T G` nucleotides)
  - Begins with start codon `ATG` (or alternative start codon `GTG`)
  - Ends with a stop codon (`TAG`, `TAA` or `TGA`)
  - No intervening stop codons

- *Note:* This is a simplified formulation of a real gene search problem

- *Note:* Bioinformatics methodology: start with a simplified model, then iteratively refine it to maximize biological realism while maintaining computational feasibility

Genome browser

Human insulin gene:

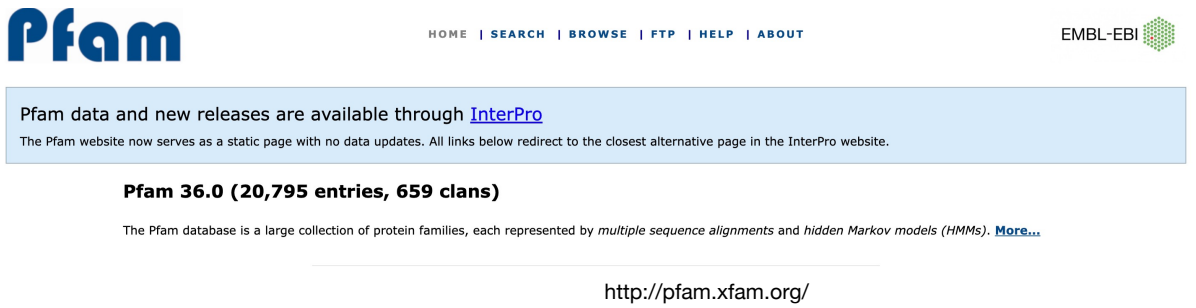


https://genome.ucsc.edu/

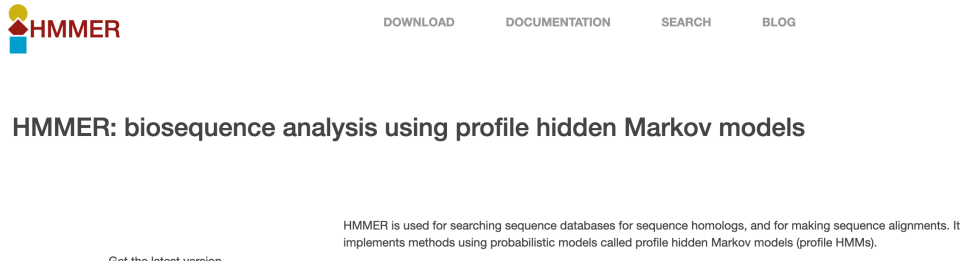
Biological processes associated with patterns in DNA sequences

- Other patterns:
  - Splicing sites
  - Promoters of gene expression
  - Binding sites of regulatory proteins (transcription factors)
  - Origin of replication
  - Restriction enzyme sites
- Realistic biological patterns are more complex, require probabilistic solutions
  - Algorithms: Hidden Markov Models, Gibbs sampling, Convolutional Neural Networks, etc. (*see courses COS 343, COS 455*)

CTCCAGCACAACCGTATATGTTTCATAAACATGCAGGTAAACAGATAAATGTTGAGCTTATTA
CGTCATCTCCTGTGCTGTGAATTTTGCTCCTCCCCATCACTTCCGTCTGAAGATTCAAGTCTA
CGTGGCGAAGAGCCAAAGCAACACTGGAACTGTCGACATATAAACACCGTTACTCAAGACTGC
GCCTCTTGTGAGGAGTGGGACAGAAACAGCCAGCAAAACGGTGGCACATGTCCGGTCTCTGGC
TCGCGTAAAGTCCAGCTCAGAACCTGGCCGTTCTACTCGATACAAAATGTAGGGTCCACGGTT
CTCCAAATCGCACCAAATAATTGTAAACAGAGAGGGAGTTGAGCACGTCCCAACCAGGGGCAC
TACGTCCTAGCTCAACATAAACCTGCACGCACTTGCGCGTCTCGTGCCGTCCCTACGAGCCCA
GCGTGATAGGCTTAGAGAGAACTTGCGACTTTGCTGGATGCCAACATCTTCACTATTCACAAT
CTCGCGCAAACAGTTTCGATATCGGCAACTCTACCGCGGCTTGAGGTATGGAGACCGAACATC
GATTTCATCGTCAATCCCCATTGAAGCCTGTGTTGCTAGACAATGTCCTATAGTGTTAATTCA



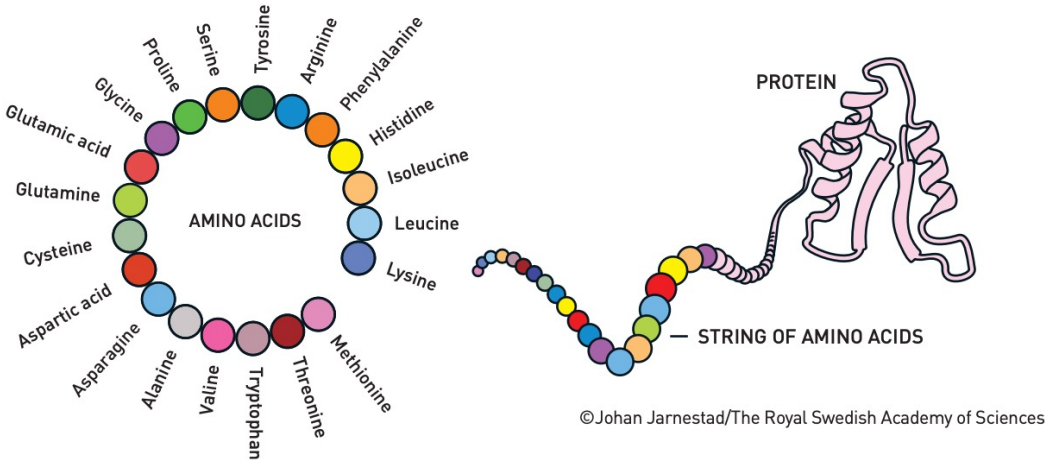
http://pfam.xfam.org/



http://hmmer.org/

Patterns in protein sequence

- Proteins fold into complex 3-dimensional structures
- Protein 3d structure determines its function
- **CS challenge:** given protein sequence, define 3d protein structure
- Difficult problem, long history of research
- AlphaFold – breakthrough algorithm using deep learning and multiple sequence alignments (*more about alignments later*)
- Nobel Prize 2024



Article | Open access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli & Demis Hassabis

— Show fewer authors

Nature **596**, 583–589 (2021) | Cite this article

2.57m Accesses | **37k** Citations | **4089** Altmetric | Metrics

Sequence comparisons / alignments

DNA replication and evolution

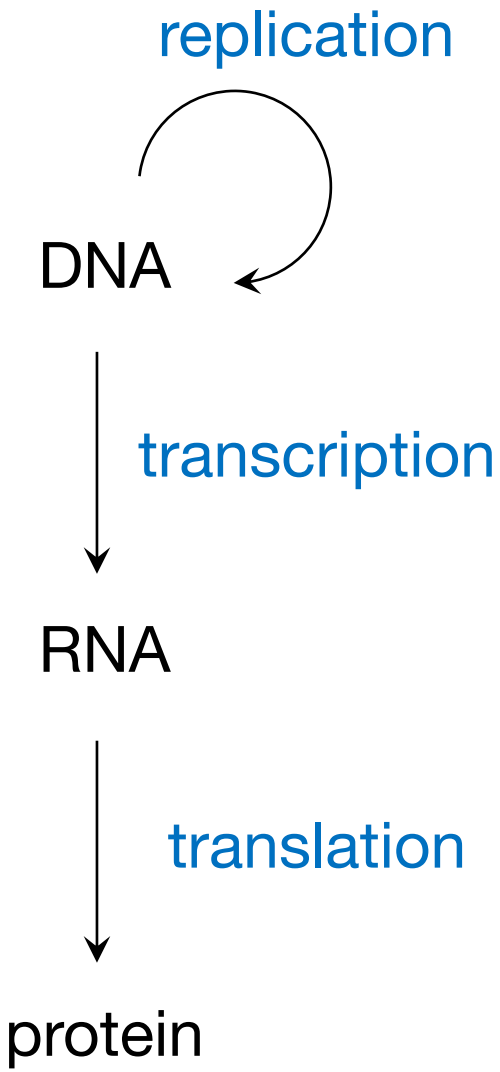
- When a cell divides, DNA is copied, or **replicates**
- *DNA replication* is a high fidelity process but errors occur (~1 error per 10^9 nucleotides) – **mutations**

GATCCGTAGAGCTAGCTAGCTTTAAAGGCGAAAGCTGGAGGAGGGCGCTC

replication

GATCCGTAGAGCTAGCTAGCTTTAAAGGCGAAAGCTGGAGGAGGGCGCTC
GATCCGTAGAGCTAGCTAGCTTTAAAGGCGAAAGCTAGAGGAGGGCGCTC

mutation



- **Mutation** – change in DNA of an organism
- **Evolution** – accumulation of DNA sequence mutations over time

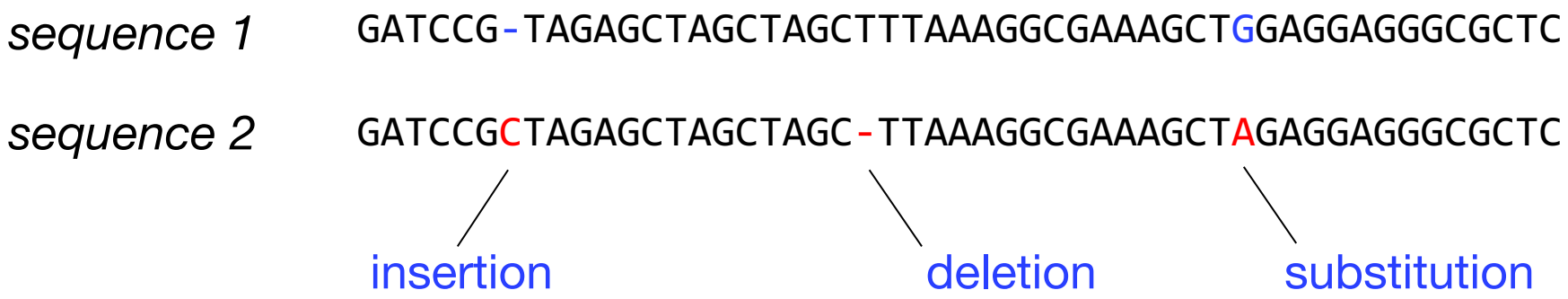
DNA mutations and evolution

- **Evolution –** the process by which living organisms change over time through changes in the genome

  NIH/NHGRI glossary: https://www.genome.gov/genetics-glossary/Bioinformatics

- Mutations – **substitutions, insertions, deletions**

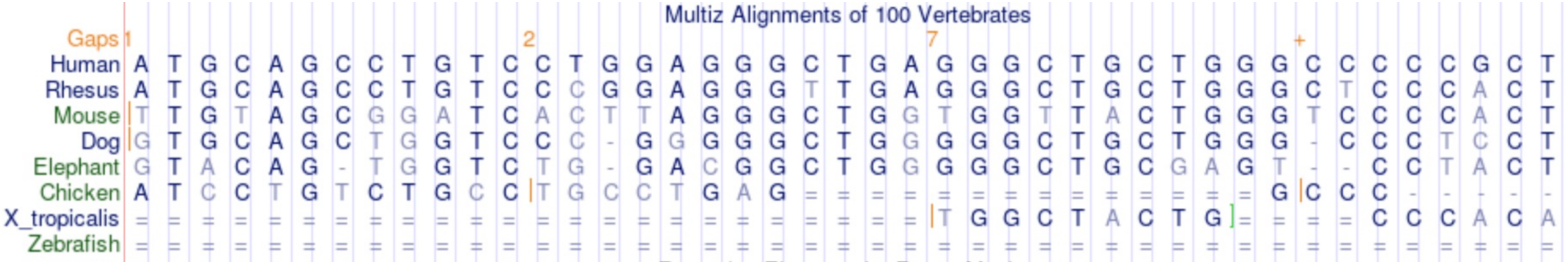
- **Indel –** insertion/deletion



Sequence comparisons to study evolution

- Compare genes, genomes to understand evolution
- *Idea:* more similar genomic sequences have more recent common ancestor

Comparison of (the starting part of) insulin genes:



- **Problem:** how similar are two DNA sequences?
- Convert this biological problem to a well-defined computational problem

https://genome.ucsc.edu/

How to define a DNA sequence comparison / alignment problem?

- **Problem:** Align two DNA sequences to minimize *distance* between them
- How to define the distance so that it's biologically meaningful but feasible to calculate?
- 1st approximation: **Hamming distance** between sequences of the same length – number of positions with *a mismatch*, i.e. different nucleotide
- The calculation is straightforward – linear run time for sequences of size *n*

```
AGATGCATGCT
AGATTCATGAT
```

Hamming distance = 2

mismatch

adapted from: https://compeau.cbd.cmu.edu/teaching-experience/great-ideas-in-computational-biology/

How to define a DNA sequence comparison / alignment problem?

- **Problem:** Align two DNA sequences to minimize *distance* between them
- How to define the distance so that it's biologically meaningful but feasible to calculate?
- 1st approximation: **Hamming distance** between sequences of the same length – number of positions with *a mismatch*, i.e. different nucleotide
- The calculation is straightforward – linear run time for sequences of size *n*

```
AGATGCATGCT
AGATTCATGAT
```

Hamming distance = 2

```
AGATGCATGCT
GATGCATGCTC
```

Hamming distance = 11

```
AGATGCATGCT
GATGCATGCTC
```

Long shared substring, so is Hamming distance appropriate?

adapted from: https://compeau.cbd.cmu.edu/teaching-experience/great-ideas-in-computational-biology/

How to define a DNA sequence comparison / alignment problem?

- **Problem:** Align two DNA sequences to minimize *distance* between them

- Better definition – **edit (Levenshtein) distance:** the minimum number of single-nucleotide edits (insertions, deletions or substitutions) to transform one sequence to another

```
AGATGCATGCT
GATGCATGCTC
```

Edit distance = 2

```
AGATG-ATGCT
AG-TGCATCCT
```

Edit distance = 3

- Optimal alignment is non-trivial to find, what's the best algorithm?

Algorithm to calculate edit distance between sequences

- **CS problem:** Calculate edit distance between two sequences of size *n* (find optimal alignment with minimal number of insertions, deletions, substitutions)

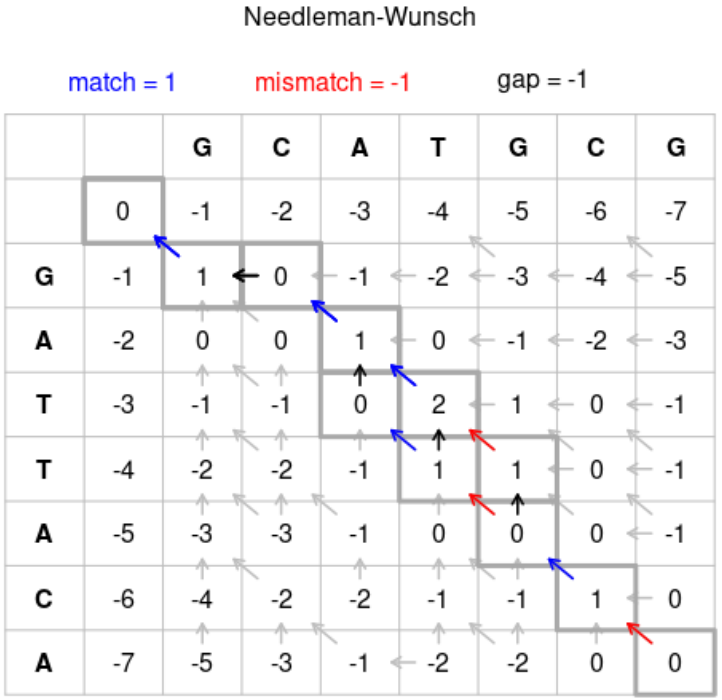
```
AGATG-ATGCT
AG-TGCATCCT
```

Edit distance = 3

- A naïve algorithm: try all possible alignments – exponential running time

- **Needleman-Wunsch algorithm** using dynamic programming – quadratic running time

- **Underlying idea:** find optimal alignment for prefixes of increasing size (*see courses COS 226, COS 343, COS 455*)



https://commons.wikimedia.org/wiki/File:Needleman-Wunsch_pairwise_sequence_alignment.png

jmb Journal of Molecular Biology

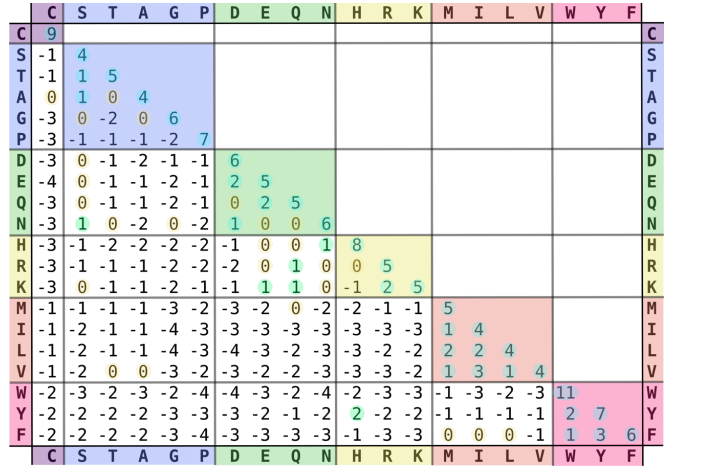
Volume 48, Issue 3, 28 March 1970, Pages 443-453

A general method applicable to the search for similarities in the amino acid sequence of two proteins ☆

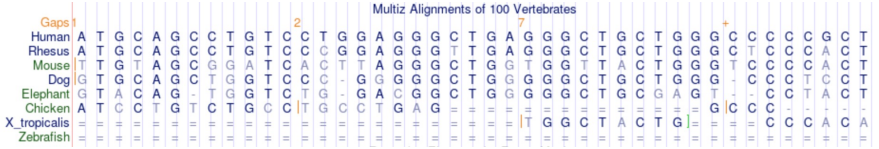
Saul B. Needleman, Christian D. Wunsch

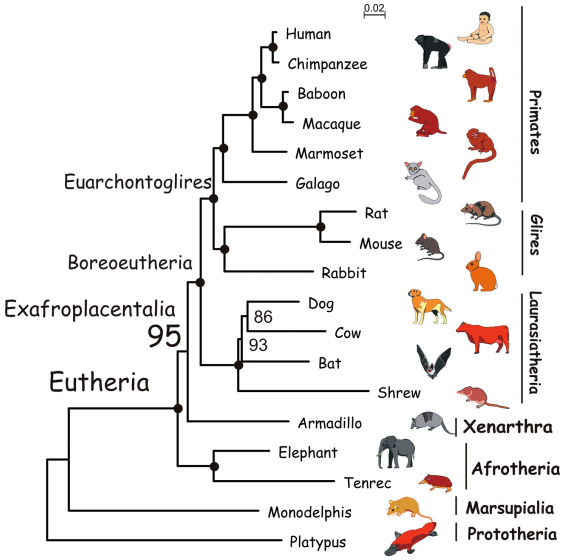
Other sequence alignment and related problems

- **Edit distance:** the minimum number of single-nucleotide edits (insertions, deletions or substitutions – penalty 1 for each) to transform one sequence to another
- Generalizations – variable cost / penalty for substitutions and indels, variable award for a match
- Global (entire sequence should be aligned) vs. local (best aligned substring) alignment
- Multiple sequence alignments
- After alignment – phylogeny reconstruction
- For some applications quadratic run time is too slow – then additional assumptions, heuristics
- BLAST – search for sequence in a database
- *See courses COS 226, COS 343, COS 455*



https://commons.wikimedia.org/wiki/File:Blosum62-dayhoff-ordering.svg





https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030002

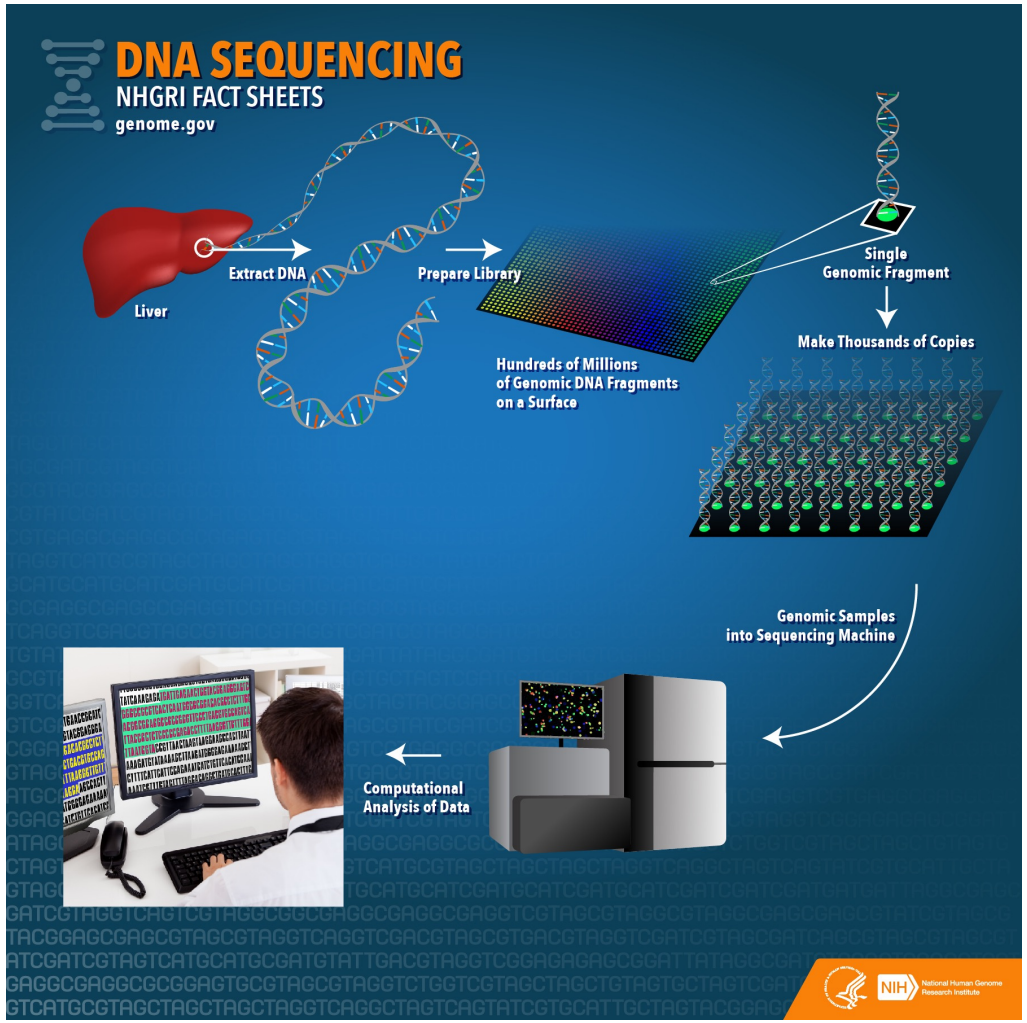
National Institutes of Health (.gov)
https://pubmed.ncbi.nlm.nih.gov › ...

Basic local alignment search tool

by SF Altschul · 1990 · Cited by 108452 — A new approach to rapid sequence comparison, basic local alignment search tool (**BLAST**), directly approximates alignments that optimize a ...

Other type of DNA alignment? High-throughput DNA sequencing

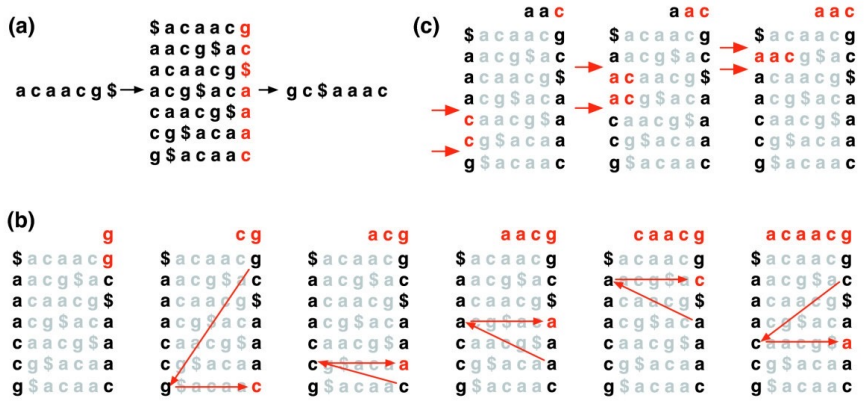
- To study genome, need DNA sequence
- Human Genome Project (completed 2003) took 13 years and ~$3 Billion to read one human genome
- Modern high-throughput sequencing is massively parallel
- A single machine (e.g. NovaSeq) generates ~6 Terabytes of data in under 2 days (20 billion *reads*)
- **Read** – short DNA sequence, up to ~250 nt
- Sequencing of another human genome ~$600
- **CS challenge:** align reads back to the original genome (if known)
- **CS challenge:** assemble original genome (if unknown)



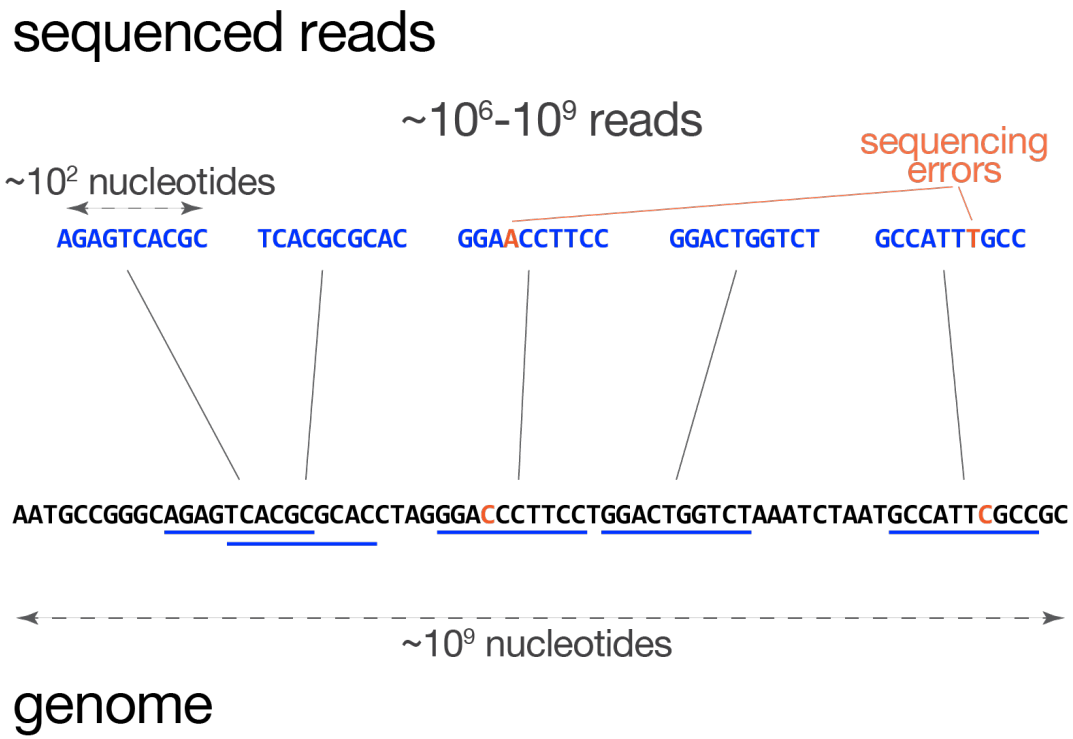
By the National Human Genome Research Institute

Algorithms for mapping reads to genome are critical in genomics

- A DNA alignment problem: align millions or billions of short reads back to the original genome
- Account for mutations and sequencing errors
- Naïve solution prohibitively slow
- *Underlying idea*: preprocess the genome and create an index using the **Burrows-Wheeler Transform** (BWT) data structure
- BWT invented in 1983, adopted in genomics in 2009



https://doi.org/10.1186/gb-2009-10-3-r25



Brief Communication | Published: 04 March 2012

Fast gapped-read alignment with Bowtie 2

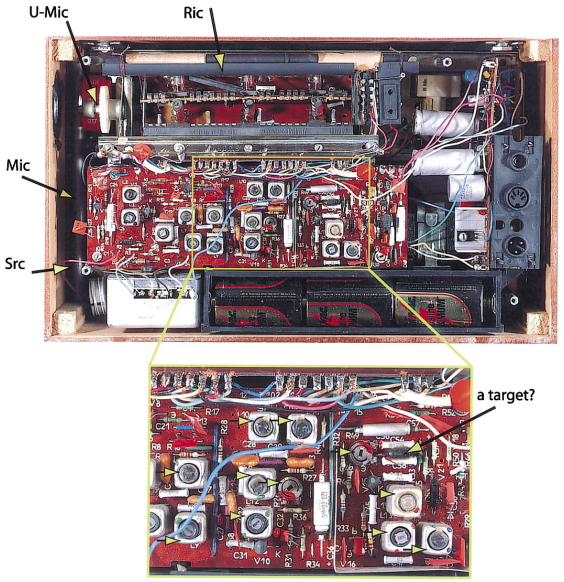
Ben Langmead & Steven L Salzberg

Nature Methods **9**, 357–359 (2012) | Cite this article

CRISPR for “programmable” genome editing

Genome editing in research and therapeutics

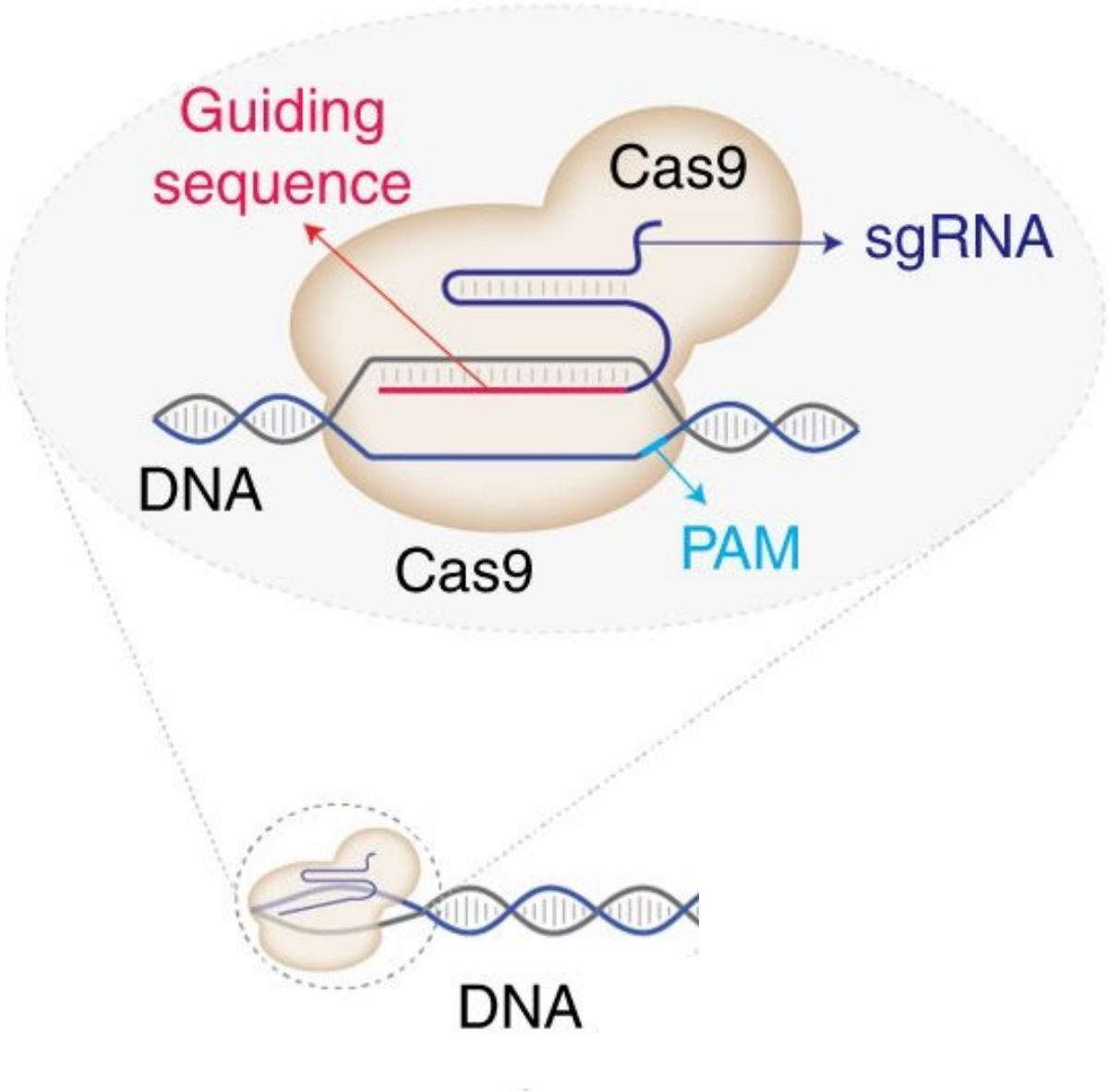
- Genome editing is a traditional tool in genetics: to study gene function, perturb or break it and observe the effect
- Recently genome editing has moved to global headlines, thanks to CRISPR
- CRISPR is the leading genome editing technology, achieving historic milestones
- CRISPR was not invented; it was discovered as bacterial immune system, then adopted widely as a biotechnology



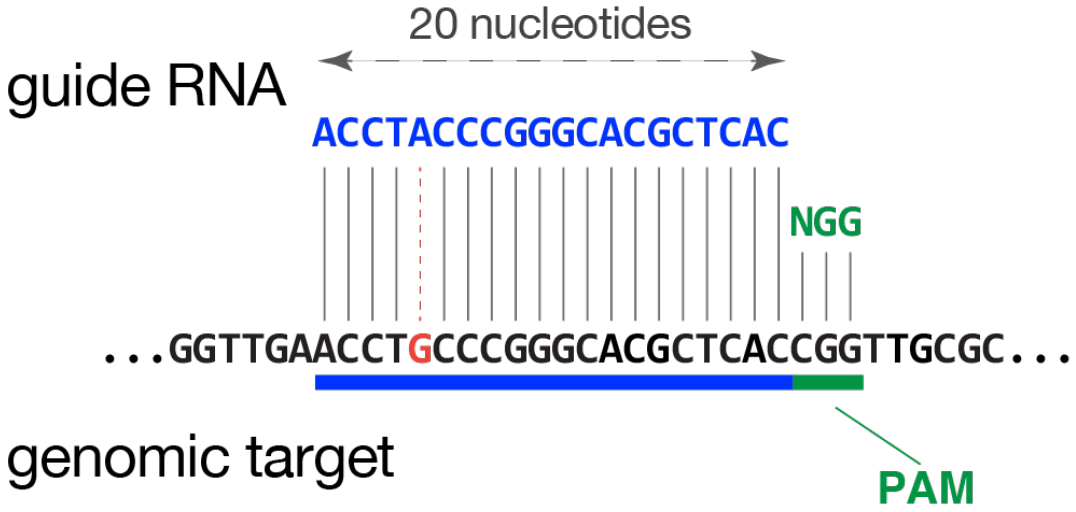
Gemini summary of recent CRISPR headlines:

• **First "On-Demand" Custom Cure:** a personalized CRISPR therapy for an infant with a rare genetic disorder in just six months, successfully treating the child.
• **Hereditary Angioedema Functional Cure:** A single-dose CRISPR therapy shown to eliminate swelling attacks in 97% of patients in a Phase 1/2 trial.
• **Sickle Cell & Thalassemia:** FDA approval of **Casgevy**, the first CRISPR drug.
• **Drought-Resistant Staples:** New CRISPR-edited varieties of wheat and rice capable of withstanding extreme heat and drought, boosting yields by up to 20%.

“Programmable” genome editing with CRISPR-Cas9



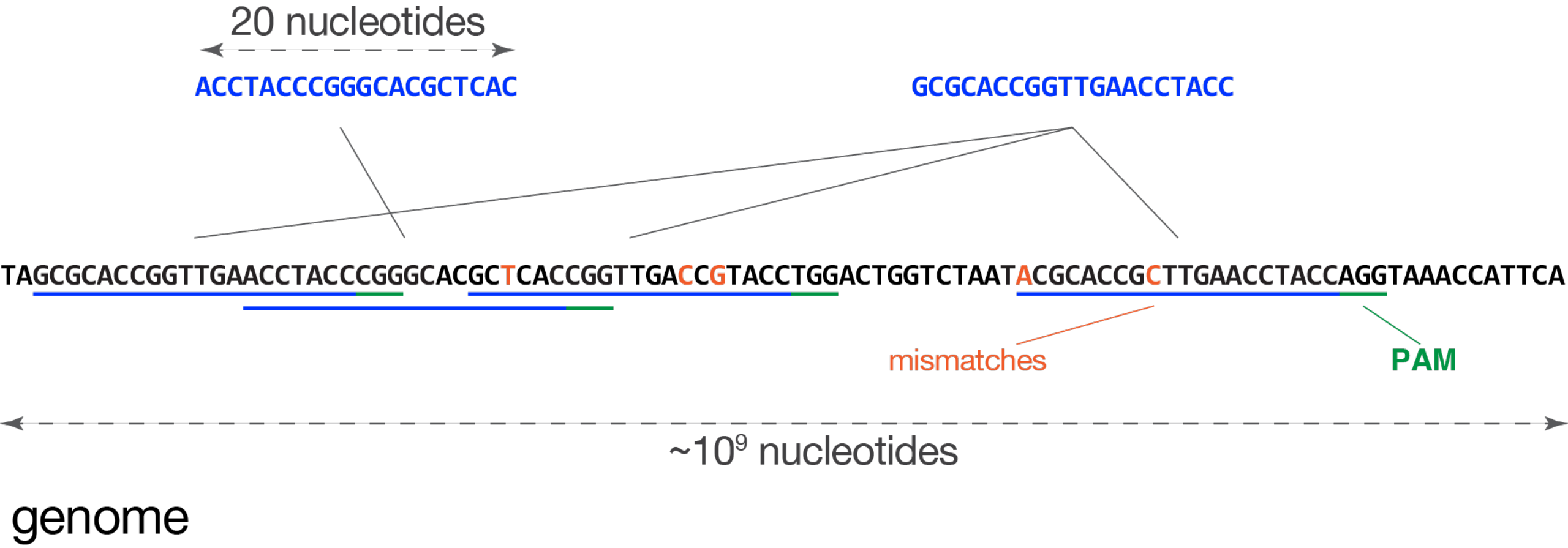
CRISPR molecular machinery is **programmed**, or directed, to specific genomic locations by guide RNA (gRNA)



Adapted from Adli *et al.* Nat Commun 2018

CRISPR targeting

- gRNA may (nearly) match multiple genomic sites
- Need to find all potential gRNA off-targets
- **CS challenge:** genome is large, many possible gRNA targets and off-targets



Computational methods for CRISPR experimental design

Example from my research:

- GuideScan for improved CRISPR guide RNA design
- Genome-wide gRNA targeting, identifies potential off-targets better than other tools, allows batch gRNA design
- Software, databases and web interface freely available
- *Key underlying idea*: **retrieval tree (trie)** data structure to preprocess and store the space of potential targets and off-targets

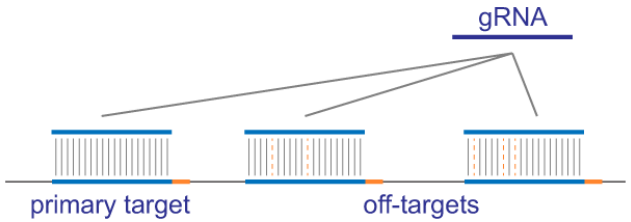
Brief Communication | Published: 06 March 2017

GuideScan software for improved single and paired CRISPR guide RNA design

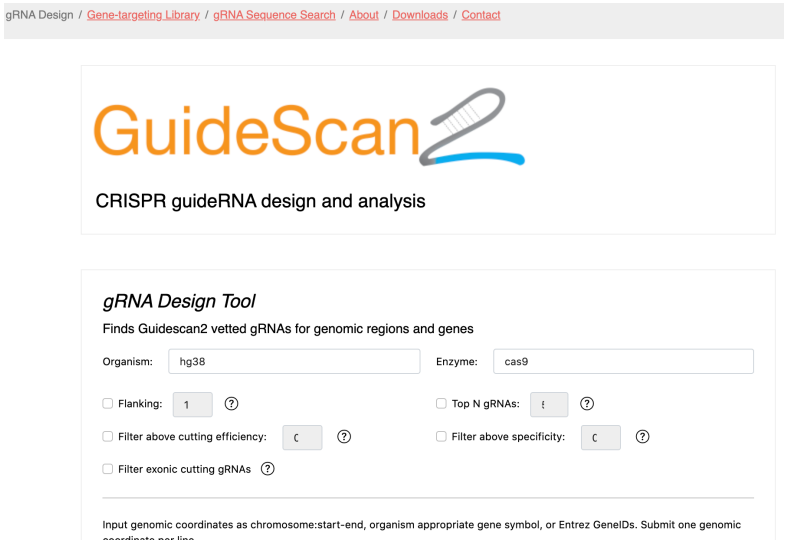
Alexendar R Perez, Yuri Pritykin, Joana A Vidigal, Sagar Chhangawala, Lee Zamparo, Christina S Leslie & Andrea Ventura

Nature Biotechnology **35**, 347–349 (2017) | Cite this article

Computational methods for CRISPR design and analysis



https://www.guidescan.com



Example from my research:

- GuideScan2: Memory-efficient, parallelizable construction of high-specificity CRISPR guide RNA databases
- *Key underlying idea*: **Burrows-Wheeler Transform** index to preprocess and access the genome

Schmidt *et al. Genome Biology* (2025) 26:41
https://doi.org/10.1186/s13059-025-03488-8

Genome Biology

METHODOLOGY **Open Access**

Genome-wide CRISPR guide RNA design and specificity analysis with GuideScan2

Henri Schmidt[1,2†], Minsi Zhang[3†], Dimitar Chakarov[4], Vineet Bansal[5], Haralambos Mourelatos[3,6], Francisco J. Sánchez-Rivera[3,7], Scott W. Lowe[3], Andrea Ventura[3*], Christina S. Leslie[2*] and Yuri Pritykin[1,2,4*]

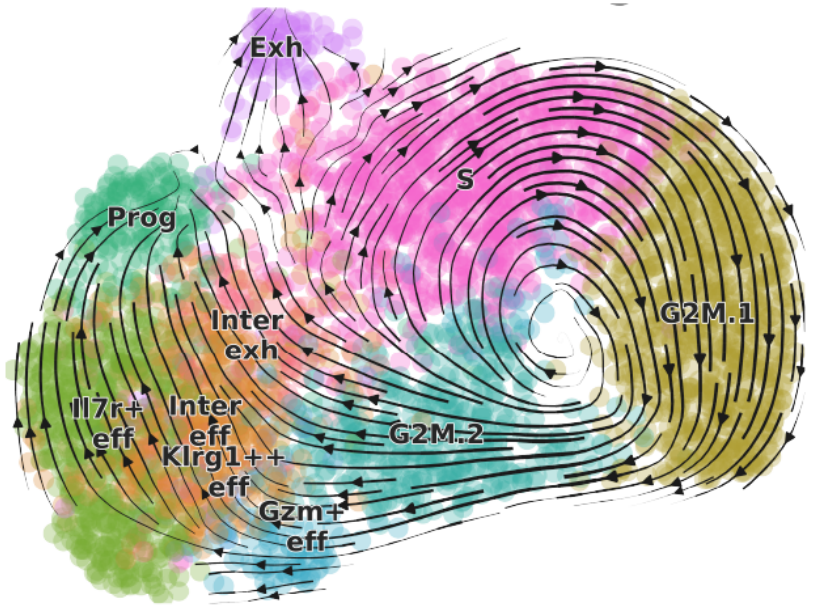
Computational biology beyond sequence analysis

Computational biology beyond sequence analysis

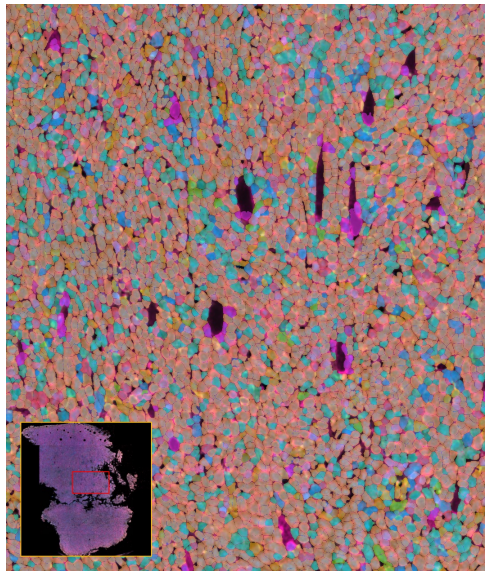
Examples from my lab's research:

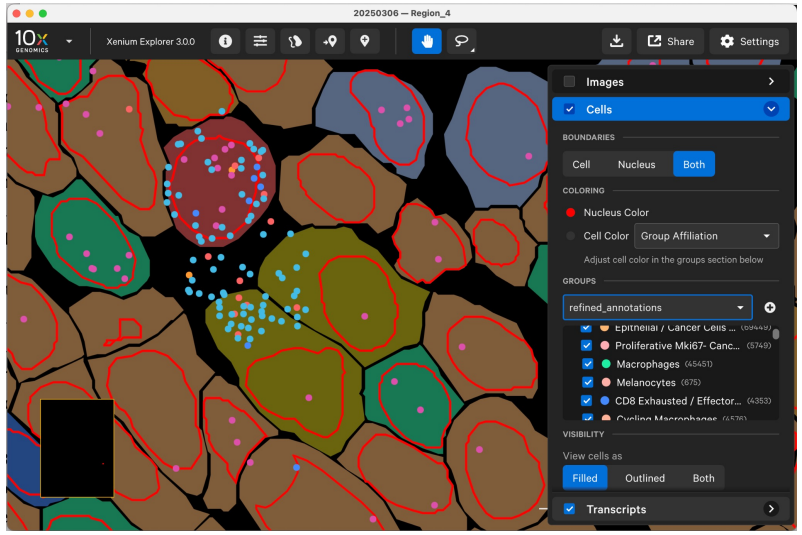
- Why cells in a multicellular organism have the same genome but different shapes and functions?
- Different genes are active in different cells – how is it regulated?
- How do cells communicate?
- How do cells organize into organs?
- How are cells continuously changing when responding to signals?
- What happens with genomes and cells in disease (e.g. cancer)?

To address these questions, need measurements and manipulations using modern technologies (e.g. high-throughput sequencing, imaging, CRISPR etc.) and new algorithms to analyze the resulting massive datasets



Avdeeva et al. bioRxiv 2025.09.14.676182





Unpublished research

Summary

- Data, algorithms and computation essential for progress in life science and biomedicine
- Computational biology and bioinformatics: study of living organisms using computers
- Computational genomics addresses crucial problems at intersection of genomics and CS:
    - Patterns in biological sequences
    - Sequence comparisons / alignments
    - Prediction of structure and function from sequence
- Bioinformatics methodology: start with a simplified model, then iteratively refine it to maximize biological realism while maintaining computational feasibility

Other courses:

QCB / COS 311: Genomics

COS 343: Algorithms for Computational Biology

QCB / COS 455: Introduction to Genomics and Computational Molecular Biology

COS 226: Algorithms and Data Structures

Credits

COS 126 : Computer Science
An Interdisciplinary Approach

Creators of COS 126:
Bob Sedgewick and Kevin Wayne

Undergrad graders and lab TAs:
apply to become one next semester!

Thank **you** for taking COS 126 !

Fall 2025 Staff:

Senior Staff


Yuri Pritykin


Kobi Kaplan


Donna Gabai


Alan Kaplan

Assistant Instructors


Tim Kosfeld
Graduate Student


Sowmya Thanvantri
Graduate Student


Ryan Oet
Graduate Student


Reva Hirave
Graduate Student


Nobline Yoo
Graduate Student


Max Gonzalez Saez-Diez
Graduate Student


Matthew Munoz
Graduate Student


Kylie Zhang
Graduate Student


Kara Schechtman
Graduate Student


David Shustin
Graduate Student


Daniel Williams
Graduate Student


Alan Zhang
Graduate Student


Abhishek Joshi
Graduate Student