Deciphering DNA using AI

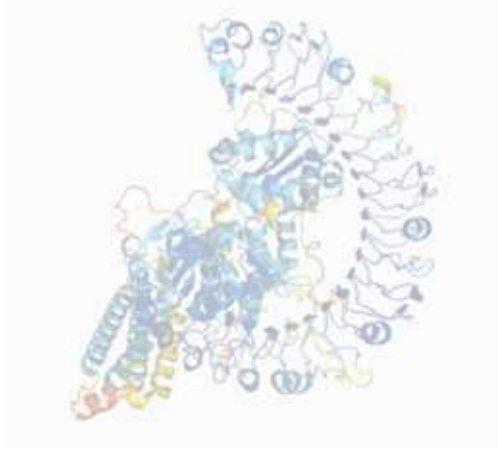
Hibiki Kato, James Yorke, Chirag Adwani, Aleksey Zimin, Brian Hunt, Bowen Chen

University of Maryland, Johns Hopkins University

Deciphering DNA using AI

Hibiki Kato, James Yorke, Chirag Adwani, Aleksey Zimin, Brian Hunt, Bowen Chen

University of Maryland, Johns Hopkins University



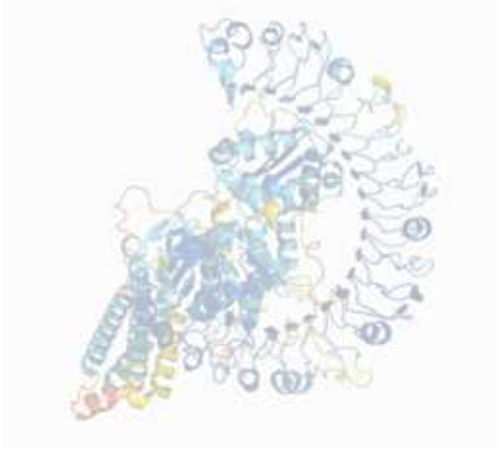
Deciphering DNA using AI

Hibiki Kato, James Yorke, Chirag Adwani, Aleksey Zimin, Brian Hunt, Bowen Chen

University of Maryland, Johns Hopkins University

Motivation:

- Genomic Medicine
- Basic biology: evolution



Deciphering DNA using AI

Hibiki Kato, James Yorke, Chirag Adwani, Aleksey Zimin, Brian Hunt, Bowen Chen

University of Maryland, Johns Hopkins University

Motivation:

- Genomic Medicine
- Basic biology: evolution

Our goal

Identify Protein-Coding Regions in DNA

Deciphering DNA using AI

Hibiki Kato, James Yorke, Chirag Adwani, Aleksey Zimin, Brian Hunt, Bowen Chen

University of Maryland, Johns Hopkins University

Motivation:

- Genomic Medicine
- Basic biology: **evolution**

Our goal

Identify Protein-Coding Regions in DNA

Our strategy

- Generate candidates from evidence
- Use binary classification to filter unreliable ones

DNA

Genes are 50% of DNA

Coding sequence is 2% of genes

gene gene

…**ATGAC**GTCAGA……GATATCAG**ATAGATCA**……**AGCATCATA**GTGACATA……AG**CATATAGCGAT**…

- **sequence of 4 symbols** {**A**, **C**, **G**, **T**}
- carrying the **genome information** (proteins, RNA, regulation, etc.)

For human genomes…

- The whole sequence is about **3 billion** bases (3GB text)
- About 50% are **gene** regions
- Roughly 2% of the gene is protein-coding

DNA

Genes are 50% of DNA

Coding sequence is 2% of genes

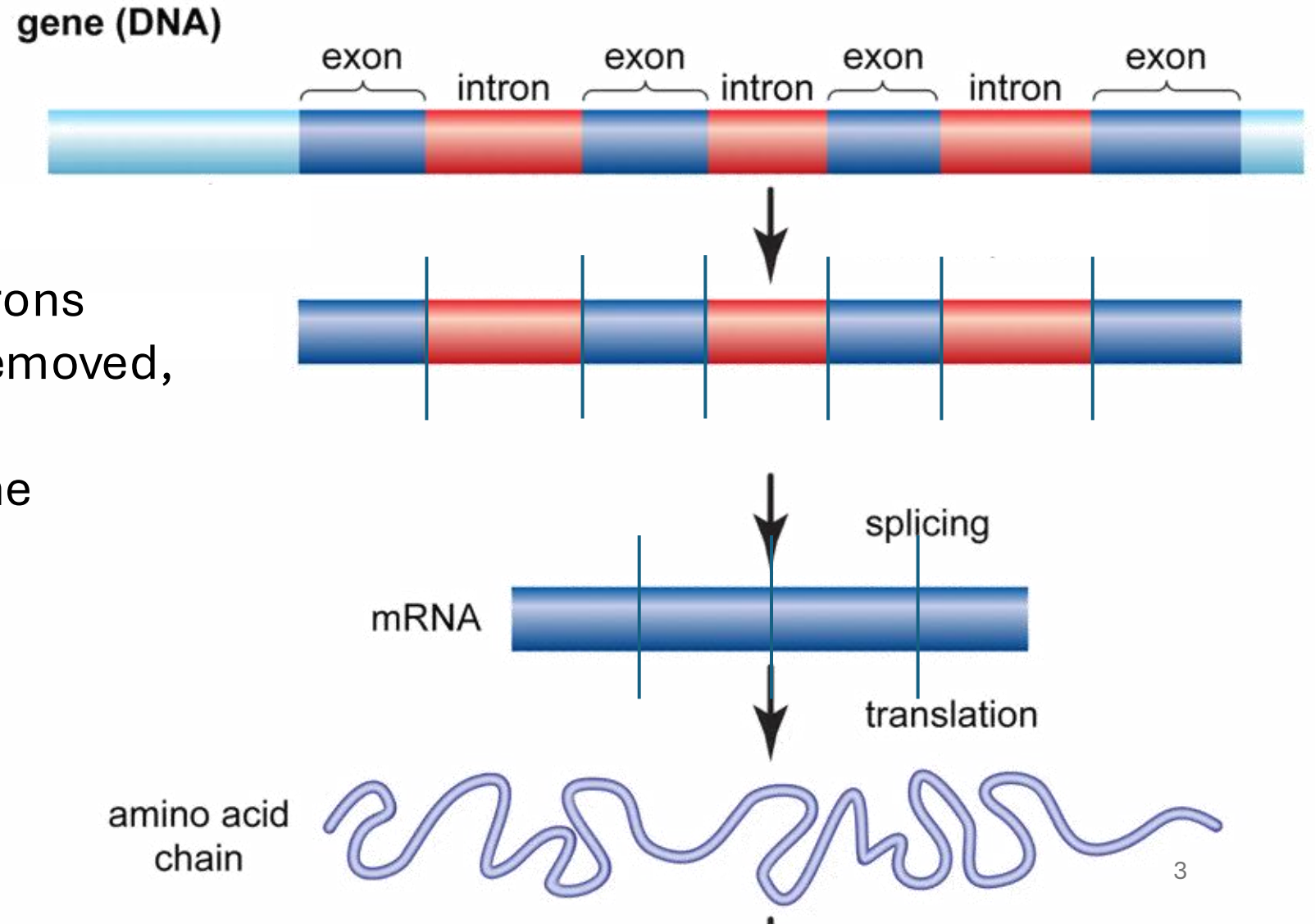
- **sequence of 4 symbols** {**A**, **C**, **G**, **T**}
- carrying the **genome information** (proteins, RNA, regulation, etc.)

For human genomes...

- The whole sequence is about **3 billion** bases (3GB text)
- About 50% are **gene** regions
- Roughly 2% of the gene is protein-coding

Biological Process: Splicing

- Genes contain exons and introns
- During splicing, introns are removed, and exons are joined to mRNA
- What we want to identify is the boundary in the genome

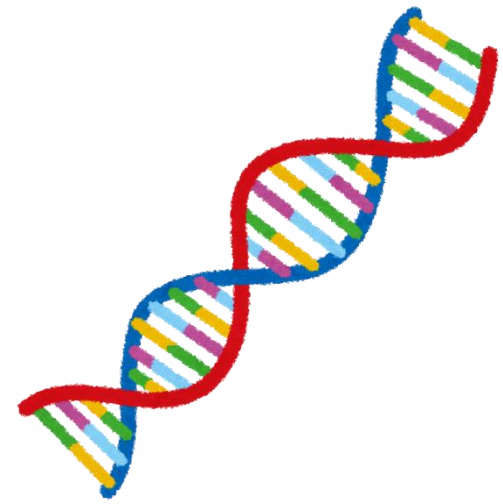


Problem:
Identify Splice Boundaries from DNA

Input

Output

...ATGACGTCAGAGATATCA
GATAGATCAAGCACATAGT
GACATAAGCATAGTGACAT
AAGCATATCATATTCATAG
TGACATAAGCATATAGCGA
T...



Splicing Machinery

GTACGTCAGAGATATCA
GATAGAG | GTGACATAAG
CATAGTGACATAAGCATAT
CATATTCATAG | GTACGTC
AGAGATATCAGATAGAG

Problem:
Identify Splice Boundaries from DNA

Input

Output

...ATGACGTCAGAGATATCA
GATAGATCAAGCACATAGT
GACATAAGCATAGTGACAT
AAGCATATCATATTCATAG
TGACATAAGCATATAGCGA
T...

AI

GTACGTCAGAGATATCA
GATAGAG | GTGACATAAG
CATAGTGACATAAGCATAT
CATATTCATAG | GTACGTC
AGAGATATCAGATAGAG

Problem:
Identify Splice Boundaries from DNA

Input

Output

...ATGACGT
GATAGATC
GACATAAC
AAGCATAT
TGACATAA
T...

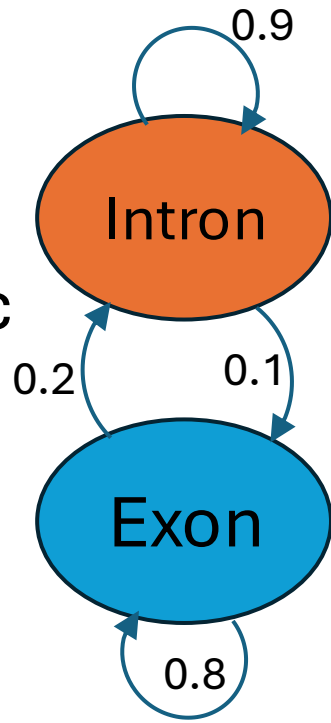
GATATCA
CATAAG
AGCATAT
GTACGTC
TAGAG

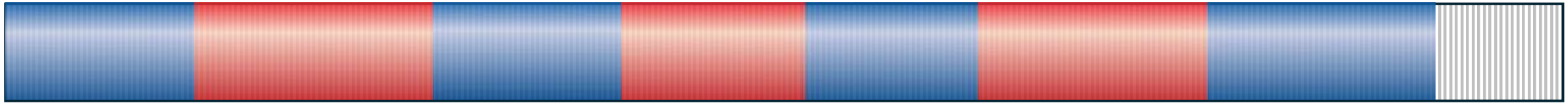
Our task:

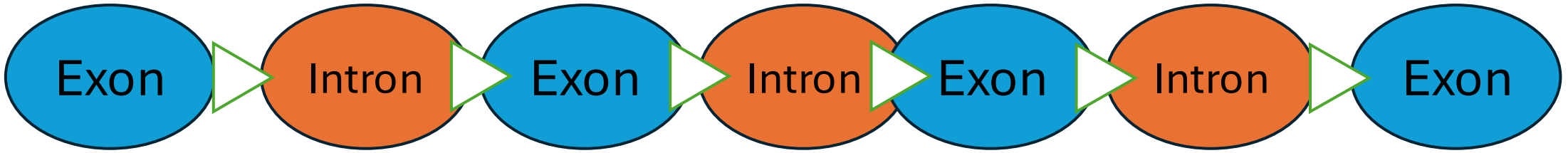
- **Choose the data for training/testing**
- Choose the model architecture
- **Define how to use AI in the pipeline**
- Train AI
- Evaluate AI

One approach:
Predict transition of intron/exon

- Model transitions between Exons and Introns with a stochastic model. (Hidden Markov Model)
- Problem: false positive, stochastic assumption



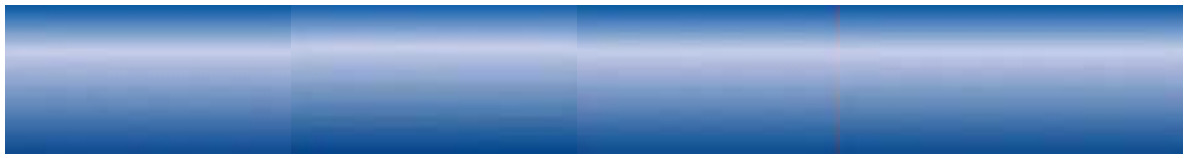




Our approach:
Build candidate transcript → Filter them with AI

- Collect RNA data
- Build transcript by mapping RNA into genome
- Score what the boundary look like (**AI used here!**)
- Score the whole transcript by aggregating boundary scores
- Classify the built transcript **`true/false`**

mRNA

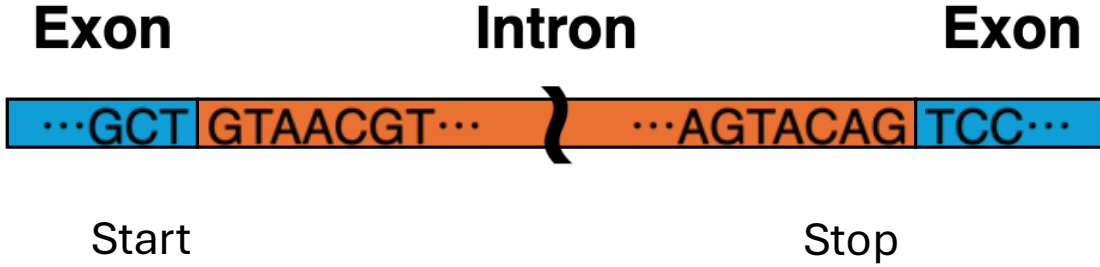


Genome

AI Task:
Binary Classification of Splice Boundaries

Input

- DNA sequence around a boundary around intron
- Fixed-length window of 100 characters



Training data

- Positive examples: “reliable” splice boundaries
- Negative examples: unreliable candidate boundaries

Output

- A score between 0 and 1 indicating confidence

Evaluation is ongoing.
Preliminary results suggest improvement over the previous model.

Summary

- We aim to identify protein-coding regions in DNA
- Instead of scanning the whole genome directly, we use evidence to build candidate transcripts
- We then apply AI to filter unreliable candidates

Take-home message

Reframing the problem can make it more suitable for modern ML models