MLfinal

Hsiang-Yu Tsai

November 2025

A Toy Model for Future AI-Based Genome Annotation
Mathematical Modeling and Machine Learning Framework

1. Introduction

We assume that AI systems in the next 20 years will possess the ability to fully interpret the human genome, identifying all genes, splice variants, and regulatory features with near-perfect accuracy. As an initial step toward this long-term objective, we construct a simplified and solvable *toy model* that captures essential components of genome annotation while remaining computationally feasible with current machine learning techniques.

2. Mathematical Modeling of Genomic and Expression Data

2.1 Genome Sequence Model

The human genome is represented as a long sequence:

$$X = (x_1, x_2, \ldots, x_L), \qquad x_i \in \{A, C, G, T\}.$$

The objective is to determine whether position i belongs to an exon:

$$y_i \in \{0, 1\}.$$

We seek a parametric mapping

$$f_\theta : X \to \hat{Y},$$

where the prediction at position i is computed from a local window:

$$\hat{y}_i = f_\theta(x_{i-w:i+w}).$$

2.2 RNA-seq Expression Model

RNA-seq provides a coverage value at each genomic position:

$$c_i \in R_{\geq 0}.$$

Exonic regions typically display higher coverage, motivating the model:

$$c_i \sim \{\ \ Poisson(\lambda_E), y_i = 1, Poisson(\lambda_I), y_i = 0, \qquad \lambda_E > \lambda_I.$$

This provides weak supervision complementary to sequence information.

2.3 Splice Site Probability Model

A splice site motif can be represented using a Position Probability Matrix (PPM):

$$P(x_{i-k:i+k} \mid iisasplicesite) = \prod_{j=-k}^{k} P_j(x_{i+j}),$$

capturing donor and acceptor sequence patterns.

3. Machine Learning Framework

3.1 Unsupervised Learning: Masked Language Modeling

Genome-wide self-supervised training enables the model to learn biological syntax, including promoters, enhancers, motifs, and splice elements. The objective is:

$$\theta = \arg\min_{\theta} E\big[-\log P_\theta(x_i \mid X_{\setminus i})\big]\,.$$

3.2 Deep Learning for Sequence-to-Function Mapping

Nucleotide sequence tokens are first embedded:

$$h_0 = Embed(X).$$

Contextual representations are extracted by a deep neural architecture:

$$h = Transformer(h_0).$$

The probability that position i belongs to an exon is computed as:

$$\hat{y}_i = \sigma(w^\top h_i + b).$$

3.3 Supervised Fine-Tuning

With known gene annotations, the model is fine-tuned via:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{L} BCE(y_i, \hat{y}_i) + \alpha \sum_{i=1}^{L} |c_i - \hat{c}_i| \,.$$

The auxiliary coverage loss enforces consistency with RNA-seq evidence.

4. Proposed Toy Model (Solvable Today)

Task 1: Exon Boundary Detection

The model predicts exon membership using sequence and coverage windows:

$$\hat{y}_i = f_\theta(x_{i-w:i+w},\ c_{i-w:i+w}).$$

To encourage contiguous exon segments, a smoothness prior is applied:

$$\mathcal{R} = \beta \sum_{i=1}^{L} |y_i - y_{i-1}|.$$

Task 2: Splice Site Classification

The model outputs:

$$P(iisadonor), \qquad P(iisanacceptor),$$

based on a local sequence window centered at i.

Task 3: Gene Proposal Construction

A predicted gene consists of a sequence of exon intervals:

$$G = \{(s_1, e_1), (s_2, e_2), \ldots\}.$$

Coverage consistency provides a scoring function:

$$Score(G) = \sum_{i \in G} \log P(c_i \mid y_i = 1).$$

5. Evaluation Metrics

- Exon-level precision, recall, and F1-score (compared with GENCODE)
- Cross-tissue reproducibility of predicted gene structures
- Splice site classification accuracy
- Discovery of novel genes supported by RNA-seq evidence

6. Significance

This toy model simulates key components of future AI genome annotation systems: unsupervised extraction of sequence grammar, integration of expression evidence, and reconstruction of gene structures. It represents an early prototype of AI not only as a computational tool but as an autonomous scientific reasoner capable of generating testable biological hypotheses.