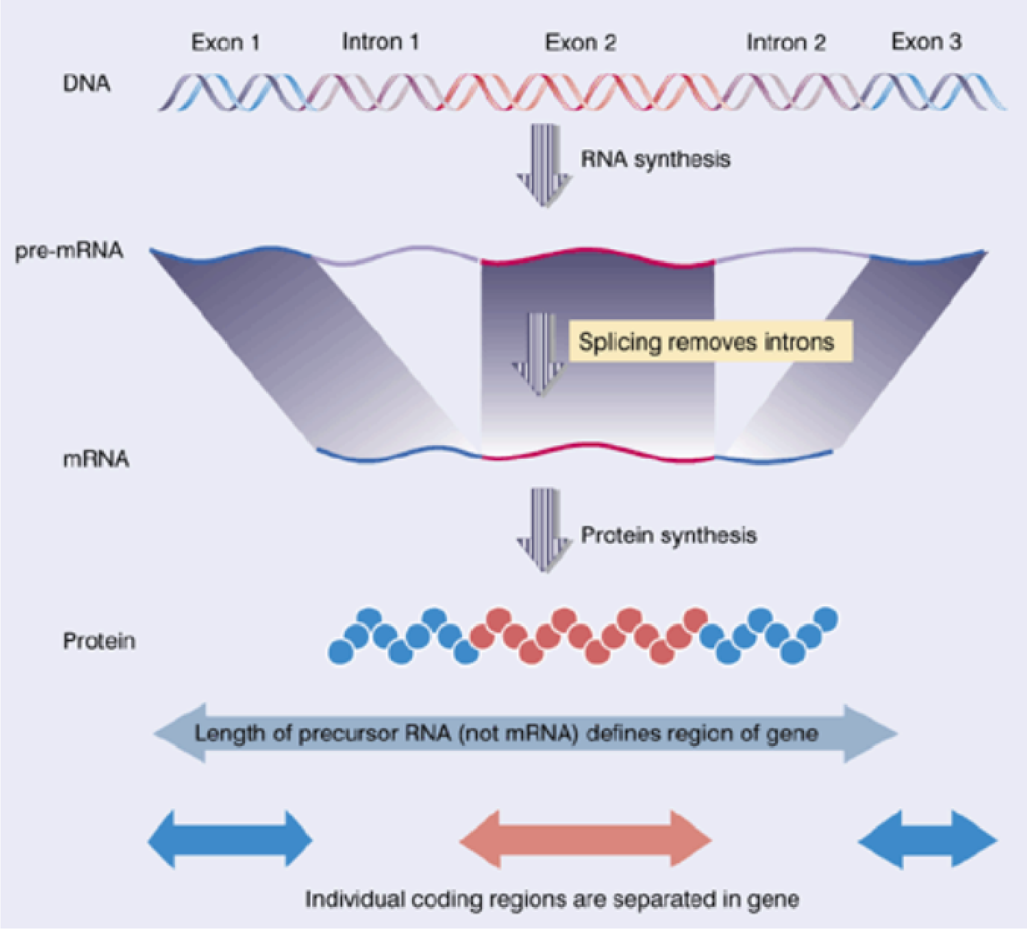
Splicing site recognition

Outline

- Introduction
  - Problem Definition
  - Input
  - Output
- Model Architecture
  - Model Training
  - Model Testing
  - State Path Generation
- Demo

Introduction

- Problem Definition
  - Splicing Site Recognition



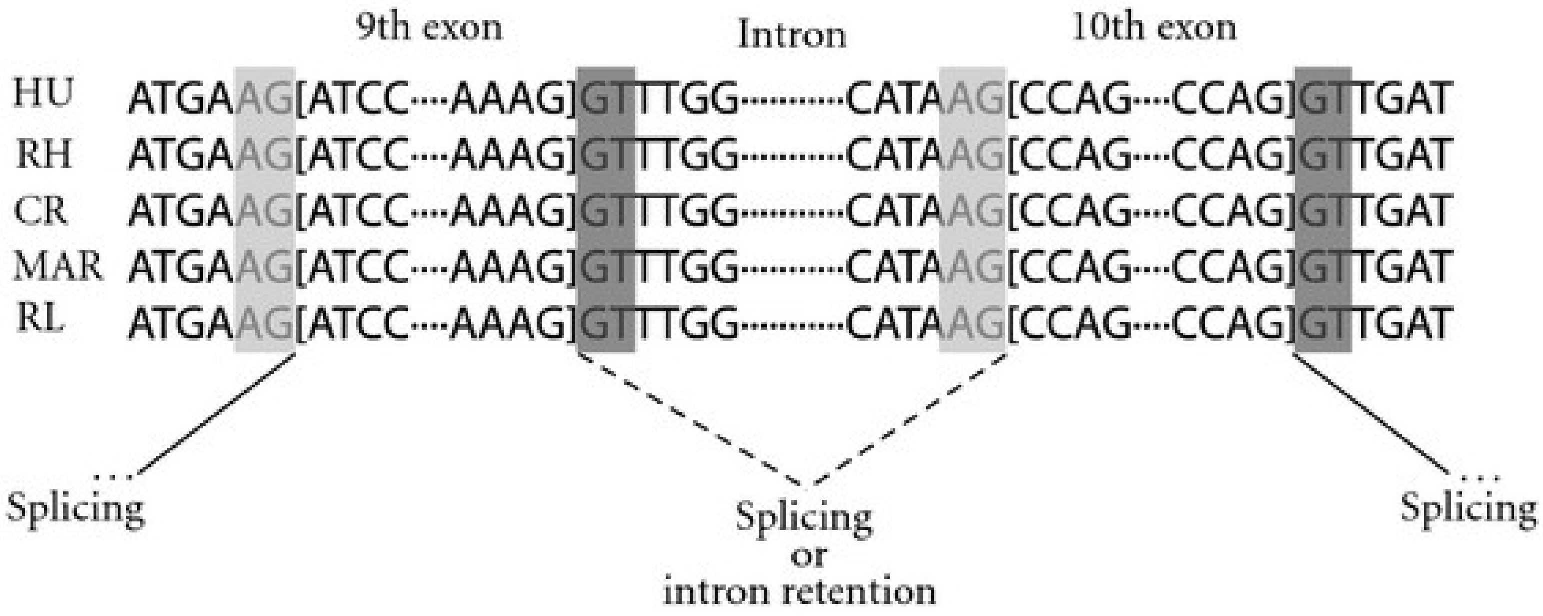
Introduction

- In order to splice out the introns, it is required to identify the splice sites.

- Splice start is known as “Donor sites” which indicates the start of the intron
  - Generally contains “GT”

- Splice end is known as “Acceptor sites” which indicates the end of the intron
  - Generally contains “AG”

Introduction


9th exon
Intron
10th exon
HU ATGAAG[ATCC····AAAG]GTTTGG············CATAAG[CCAG····CCAG]GTTGAT
RH ATGAAG[ATCC····AAAG]GTTTGG············CATAAG[CCAG····CCAG]GTTGAT
CR ATGAAG[ATCC····AAAG]GTTTGG············CATAAG[CCAG····CCAG]GTTGAT
MAR ATGAAG[ATCC····AAAG]GTTTGG············CATAAG[CCAG····CCAG]GTTGAT
RL ATGAAG[ATCC····AAAG]GTTTGG············CATAAG[CCAG····CCAG]GTTGAT
...
Splicing
Splicing
or
intron retention
...
Splicing

Introduction

- Input
  - A Sequence of gene

- Output
  - A sequence where each nucleotide in the gene is labeled as,
    - ‘E’ – for exon
    - ‘I’ - for intron
    - ‘D’ - for donor site
    - ‘A’ - for acceptor site

Introduction

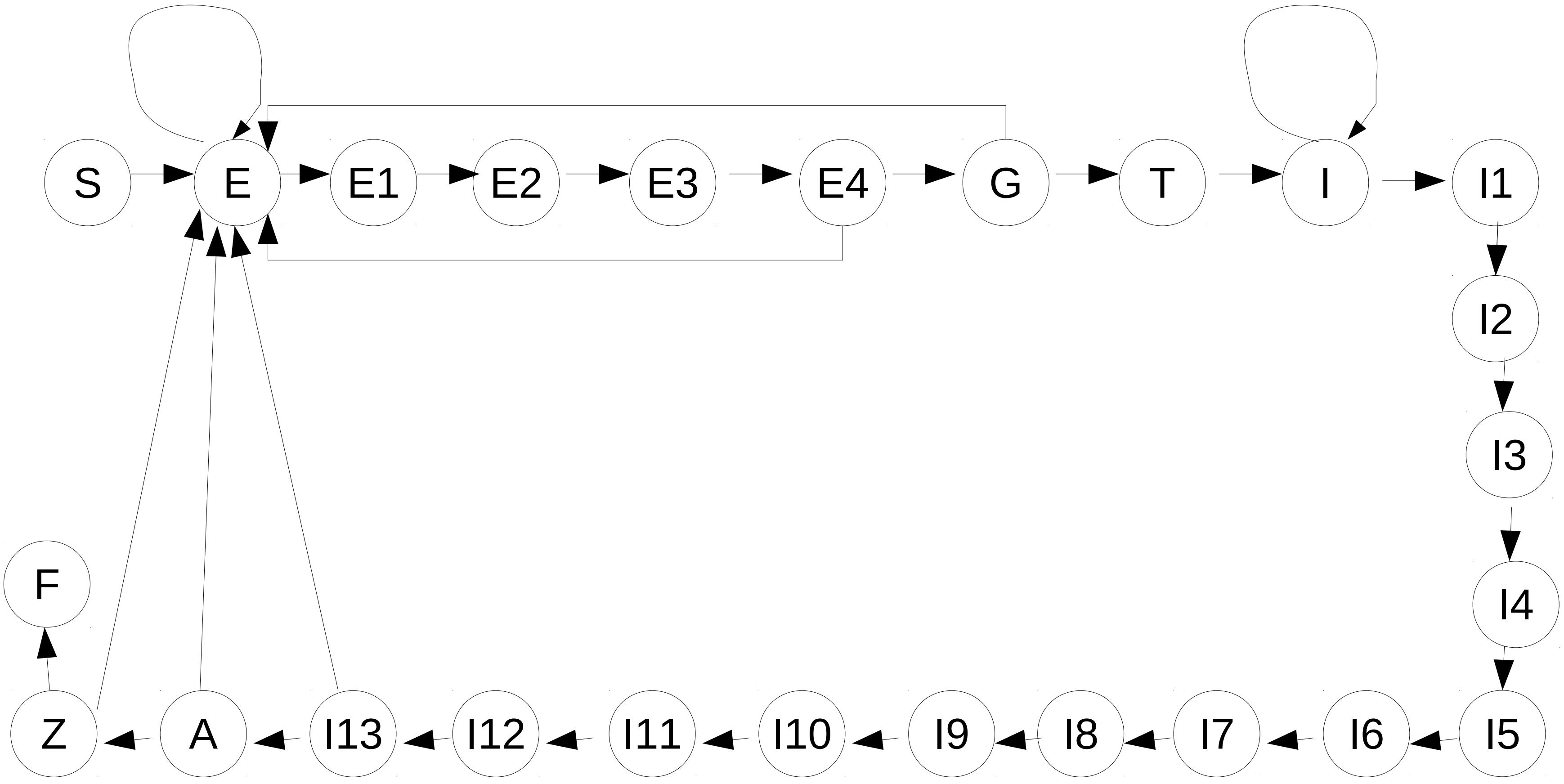
- Dataset
  - DNA sequence for 570 vertebrate gene
  - 570 vertebrate gene information with exons identified

- Implementation
  - Language: MATLAB
  - ToolBox: Kevin Murphy’s HMM Toolbox

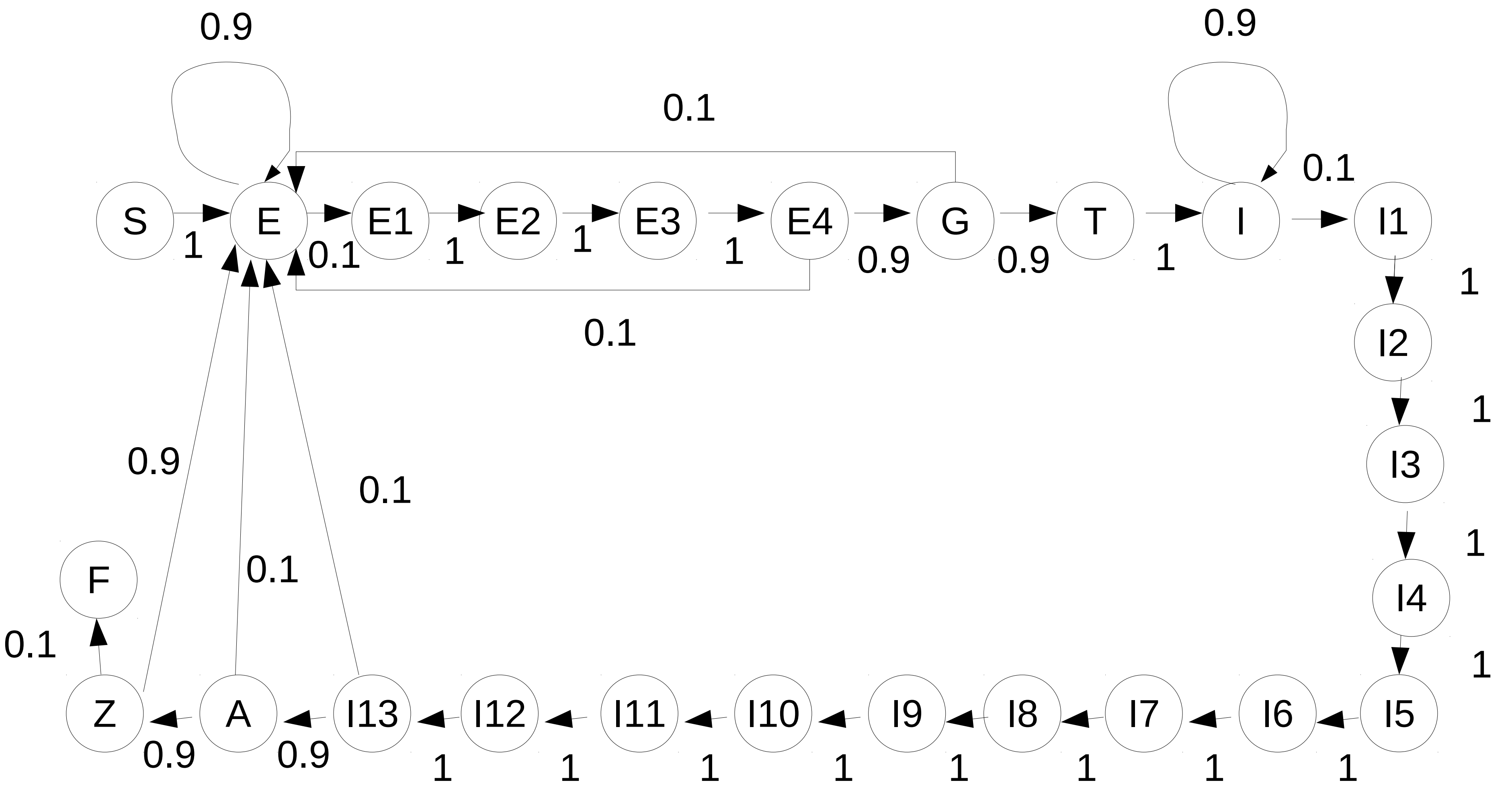
Model Architecture

- Model Definition
  - Order = 1
  - States = 28
    {S, E, E1, E2, E3, E4, G, T, I, I1, I2, I3, I4, I5, I6, I7, I8, I9, I10, I11, I12, I13, A, Z, F}
  - Initial Probability
  - Transition Probability
  - Emission Probability

Model Architecture



Model Architecture



Model Architecture

- Model Training
  - “When all the paths are known, estimation is simple: count the number of times a particular transition or emission is used in the training set”
  - The maximum likelihood estimation

$$a_{s,t} = \frac{f(st)}{\sum_{t' \in \Sigma} f(st')} \qquad e_s(b) = \frac{f_s(b)}{\sum_{c \in \Sigma} f_s(c)}$$

  - Dataset: Last 470 genes of the dataset

Model Architecture

- Model Testing
  - Test Dataset: First 100 gene sequence of the dataset
  - Approach:
    - Generated state path using viterbi algorithm
    - Match the result with the actual exon sequence data
    - Accuracy = TP + TN / All
  - Accuracy = 68.8%

Model Architecture

- State Path Generation
  - Viterbi Algorithm
  - Built-in functions of Kevin Murphy’s HMM Toolbox
    - B = ***multinomial_prob***(test_sequence, emission_probability);
    - [path] = ***viterbi_path***(initial_probability, transition_probability, B);
  - Most probable path sequence

Demo

Thank you