**Course: Advance Bio Informatics**

**Module Title: HMM Gene Finding**

**Module No: 104**

**Gene Finding**

**Objective:** To find coding & non-coding regions of an unlabeled string of DNA nucleotides

**Motivation**

**(1)** Assist in annotation of genomic data produced by genome sequencing methods

**(2)** Gain insight into the mechanisms involved in transcription, splicing and other processes

**Example**

Assume we are given a DNA sequence that begins in an exon, contains one 5' splice site and ends in an intron

Identify where the switch from exon to intron occurs

Where is the splice site?

**Statistical Properties**

**Exons** have a uniform base composition on average A/C/T/G: 25% for each base

**Introns** are A/T rich

A/T: 40% for each

C/G: 10% for each

**5' Splice site** consensus nucleotide is almost always a G...

G: 95% A: 5%

**HMM**

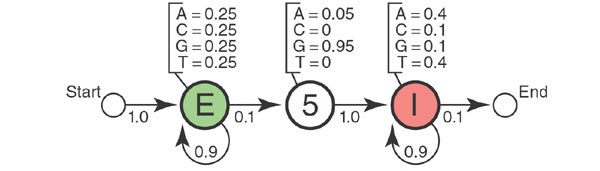
We have three states

"E" for Exon

"5" for 5' SS

"I" for Intron

Each State has its own emission probabilities which model the base composition of exons, introns and consensus G at the 5'SS. Each state also has transition probabilities



* We generate two strings of information
  + Observed Sequence
  + Underlying State Path

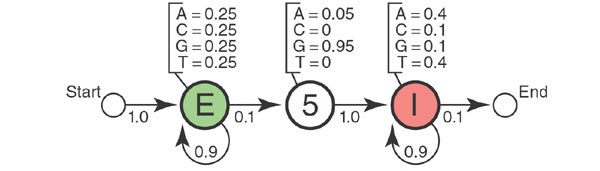
**HMM Structure**

*S* – Observed sequence

π – State Path

Θ – Parameters

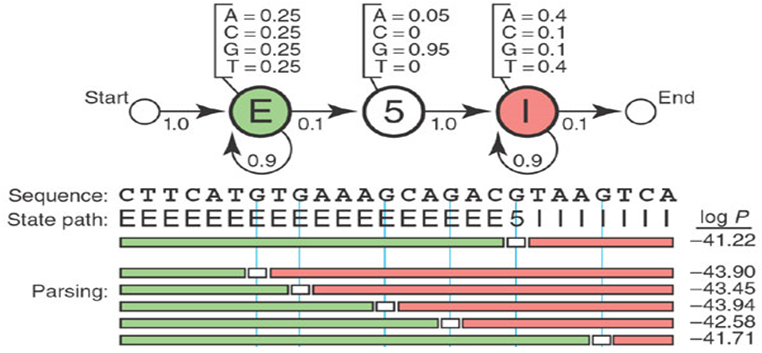
The probability P(*S*,π|HMM, Θ) is the product of all emission probabilities and transition probabilities.





* 27 transitions & 26 emissions.
* log P(*S*, π|HMM, Θ) = -41.22

**Discussion about All Observed Paths**

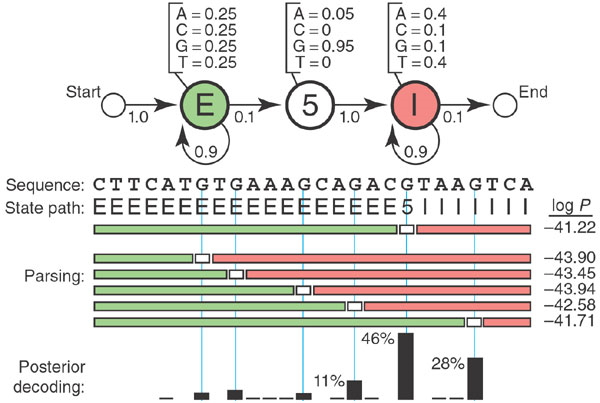


**Confidence Score**

Fifth G is the right choice?

The probability that nucleotide *i* was emitted by state *k* is the sum of the probabilities of all the states paths use state *k* to generate *i*, normalized by the sum over all possible state paths.

**Posterior Decoding**



**Further Possibilities**

But we can go further; we could add a more realistic consensus GTRAGT at the 5' splice site. We could put a row of six HMM states in place of '5' state to model a six-base un-gapped consensus motif.

Possibilities are not limited.