What is a Hidden Markov Model?

- A Hidden Markov Model (HMM) is a type of machine learning algorithm.

- With respect to genome annotation, HMMs label individual nucleotides with a nucleotide type. Possible nucleotide types include:
  - Introns
  - Exons
  - Splice Sites (3' and 5')

- HMMs are used in speech recognition, facial recognition and many other applications.

HMM Probabilities

- The probability of switching from one nucleotide type to another (ex. Exon → Intron) is called a transition probability.

- The probability of observing a nucleotide (A, T, C, G) that is of a certain nucleotide type (exon, intron, splice site) is called an emission probability.

- Think of an emission probability as the probability of:
  - Observing an adenine in an exon
  - Observing an adenine in a splice site

HMM Features

- Transition Probabilities
- Emission Probabilities

- A state path is the list of nucleotide type labels assigned to each nucleotide in the sequence.

- An HMM can produce many state paths for a single sequence.
Determining the Correct Splice Site

- A HMM will identify many splice sites for one sequence, but how do we measure which splice site is most likely to be correct?
- One way is to calculate the probability of each splice site.
- Splice site probabilities are calculated by multiplying all transition and emission probabilities in the state path.
- The splice site with the highest probability is most likely the correct splice site.

HMMs and Gene Prediction

- Hidden Markov Models are the core of a number of gene prediction algorithms.
  - GENSCAN
  - Augustus
  - Gened
  - Genemark
  - GRAIL
  - Twinscan

Conclusions

- Hidden Markov Models have proven to be useful for finding genes in unlabeled genomic sequence.
- Hidden Markov Models are machine learning algorithms that have nucleotide types, transition probabilities and emission probabilities.
- Hidden Markov Models label a series of observations with a state path, and they can create multiple state paths.
- It is mathematically possible to determine state paths that are likely to be correct.

Challenges

- How do transition probabilities affect the length of predicted ORFs?
- How do emission probabilities for specific states affect the accuracy of splice site predictions?
- Do gene predictions give the final word on correct splice sites? What other pieces of information would be useful for annotating genes?