Goals for Lecture

Key concepts:

• Related genomes as an additional source of evidence for gene finding
• Pair hidden Markov models
• Extending GENSCAN to emits pairs of observed variables
Why Use Comparative Methods?

- Genes are among the most conserved elements in the genome
  - use conservation to help infer locations of genes

- Some signals associated with genes are short and occur frequently in the genome
  - use conservation to eliminate false candidate sites from consideration
TWINSCAN
Korf et al., Bioinformatics 2001

• Extend GENSCAN using pre-computed conservation

• Prediction with TWINSCAN
given: a sequence to be parsed, $x$
using BLAST, construct a conservation sequence, $c$
have HMM simultaneously parse (using Viterbi) $x$ and $c$
Conservation Sequences in TWINSCAN

• Before processing a given sequence, TWINSCAN first computes a corresponding conservation sequence

```
ATTTAGCCTACTGAAATGGACCGCTTCAGCATGGTATCC
```

matched  unaligned  mismatched

• Based on BLAST matches sorted by alignment score
Conservation Sequence Example

input sequence

significant BLAST matches ordered from best to worst

resulting conservation sequence
Modeling Sequences in TWINSCAN

• Each state “emits” two sequences
  – the given DNA sequence, $x$
  – the conservation sequence, $c$

• Treats them as conditionally independent given the state

$$\Pr(q_i, d_i, x_i, c_i) \approx \Pr(d_i \mid q_i) \Pr(x_i \mid q_i, d_i) \Pr(c_i \mid q_i, d_i)$$
SLAM
Pachter et al., *RECOMB* 2001

- Doesn’t require a pre-computed alignment
- Combine generalized HMM (GENSCAN) and pair HMM
  - GPHMM

- Prediction with SLAM
  given: a pair of sequences to be parsed, $x$ and $y$
  find approximate alignment of $x$ and $y$
  run constrained Viterbi to have HMM simultaneously
  parse and align $x$ and $y$
Pair Hidden Markov Models

- Each non-silent state emits one or a pair of characters

Transition probabilities

H: homology (match) state
I: insert state
D: delete state
PHMM Paths = Alignments

sequence 1: AAGCGC
sequence 2: ATGTC

hidden: B H H I I H D H E
observed: A A G C G C

AT G T C
Generalized Pair HMMs

- Represent a parse $\pi$, as a sequence of states and a sequence of associated lengths for each input sequence.

Sequence of hidden states:

$\tilde{q} = \{q_1, q_2, \ldots, q_n\}$

Sequence of pairs of duration times generated by hidden state:

$\tilde{d} = \{d_1, d_2, \ldots, d_n\}$

Sequence of symbols:

$\tilde{e} = \{e_1, e_2, \ldots, e_n\}$

May be gaps in the sequences.
TWINSCAN vs SLAM

- Both use multiple genomes to predict genes
- Both use generalized HMMs
- TWINSCAN
  - takes as an input a genomic sequence, and a conservation sequence computed from an informant genome
  - models probability of both sequences; assumes they’re conditionally independent given the state
  - predicts genes only in the genomic sequence
- SLAM
  - takes as input two genomic sequences
  - models joint probability of pairs of aligned sequences
  - can simultaneously predict genes and compute alignments

- More detailed slides in Spring 2015 syllabus
  - https://www.biostat.wisc.edu/bmi776/spring-15/syllabus.html