Example applications of HMMs

Colin Dewey
BMI/CS 576
www.biostat.wisc.edu/bmi576/
colin.dewey@wisc.edu
Fall 2015
Three applications of HMMs

• Profile HMMs
  – Models of families of sequences
  – Textbook Chapter 5

• Phylogenetic HMMs
  – Better models of sequence in space and time
  – Siepel et al., 2005

• ChromHMM:
  – Aims to segment the genome based on multiple chromatin modifications
  – Ernst et al., 2010
Profile HMMs

- Profile HMMs are commonly used to model families of sequences

*Delete states* are silent; they account for characters missing in some sequences.

*Insert states* account for extra characters in some sequences.

*Match states* represent key conserved positions.

*Insert and match states* have emission distributions over sequence characters.
A Profile HMM Trained for the SH3 Domain

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences
Multiple Alignment of SH3 Domain

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences
Profile HMMs

• To classify sequences according to family, we can train a profile HMM to model the proteins of each family of interest.

• Given a sequence \( x \), use Bayes’ rule to make classification

\[
\Pr(c_i \mid x) = \frac{\Pr(x \mid c_i) \Pr(c_i)}{\sum_j \Pr(x \mid c_j) \Pr(c_j)}
\]
Phylogenetic Hidden Markov Models (phylo-HMM)

• Marriage of phylogenetic trees and Hidden Markov models
• A statistical model that considers both the process of nucleotide substitution at each site as well as how this changes from one site to another
  – The first being the phylogenetic part
  – The second being the HMM part
• Provides a rigorous way to model molecular evolution
• Have been extended for secondary structure prediction, site-specific substitutions, gene prediction, functional element identification
HMMs and PhyloHMMs

**A**

A simple HMM DNA sequence

\[ X = \text{TAAACCGGCAGA} \ldots \]

**B**

A phylo-HMM for DNA Multiple sequence alignment

\[ X = \text{TTAGGCAAGG} \ldots \text{AAGGCGCCGA} \ldots \]
Generative model of phylo-HMM

• A phylo-hmm’s emission is a column of a multiple sequence alignment
• A character is drawn from a background distribution and assigned to the root
• Character substitutions happen on their way from the root to the leaf
  – What is left at the leaves is what constitutes our observed sequence
  – A state of a phylo-HMM controls the substitution/evolutionary model
Formal definition of a phylo-HMM

• A phylo-HMM is a four-tuple

\[ \theta = \{ S, \psi, A, b \} \]

• \( S = \{ S_1, \cdots, S_M \} \): set of \( M \) states

• \( A \): set of state transition probabilities

• \( b \): set of initial probabilities

• \( \psi = \{ \psi_1, \cdots, \psi_M \} \): set of phylogenetic models, one for each state
The phylogenetic model of Phylo-HMM

- A phylogenetic model is made up a rate matrix $Q_j$, initial probabilities, $p_j$, tree $\tau_j$, branch lengths, $\beta_j$.
  \[ \psi_j = \{Q_j, p_j, \tau_j, \beta_j\} \]

- $Q_j$: Rate matrix over alphabet of size $K$
  - $K$-by-$K$ matrix

- This is really the emission model

- Probability that a column $X_j$ is emitted by a state $k$ is
  \[ P(X_j \mid \psi_k) \]

- This in turn is computed how?
  - Felsentein’s algorithm!
Probability of an alignment and path

Path: sequence of states

\[ P(\pi, X|\theta) = b_{\pi_1} P(X_1|\psi_{\pi_1}) \prod_{i=1}^{L} a_{\pi_{i-1}\pi_i} P(X_i|\psi_{\pi_i}) \]

A multiple sequence alignment

- As in a standard HMM, \( P(X|\theta) \) is computed using the forward algorithm
- The most likely sequence of states is computed by the Viterbi algorithm
A Phylo-hmm for gene finding

$A_4$ - State phylo-HMM for gene finding. States $s_1, s_2, s_3$ represent the three codon positions and state $s_4$ represents non-coding sites. The associated phylogenetic models capture characteristic properties of the different types of sites - e.g., the higher average rate of substitution, and the greater transition/transversion ratio, in non-coding and 3rd-codon-position sites than in 1st- and 2nd-codon-position sites.

Example 1. A toy gene finder

This example is meant to demonstrate, in principle, how a phylo-HMM can be used for gene finding. Consider a simple 4-state phylo-HMM, with states for the three codon positions and noncoding sites (Fig. 2A). The problem is to identify the genes in a synthetic data set based on this model, using nothing but the aligned sequence data and the model (this is a multiple-sequence version of the ab initio gene prediction problem). For simplicity, we assume the model parameters $\psi$ are given, along with the data set $X$. In practice, the parameters have been set to reasonable values for a phylogeny of $n = 8$ mammals (Fig. 2B), and the data set has been generated according to these values. The state path was recorded during the generation of the data, so that it could be used to evaluate the accuracy of predictions. The synthetic data set consists of $L = 100,000$ sites and 74 genes.

The Viterbi algorithm can be used for prediction of genes in this data set in a straightforward way. For every site $i (1 \leq i \leq L)$ and state $j (1 \leq j \leq M)$, the emission probability $P(X_i | j)$ is computed using Felsenstein's algorithm. These $L \times M$ values, together with the state-transition probabilities $A$ and initial-state probabilities $b$, are sufficient to define the joint probability $P(\lambda, X | \psi)$ for any path $\lambda$, and can be simply plugged into the standard

Parameter estimates from [44] were used for the phylogenetic models, and the state-transition probabilities were approximately based on estimates from [43] (the probability from $s_4$ to $s_1$ was inflated so that genes would not be too sparse). A uniform distribution was assumed for initial-state probabilities.
Phastcons: A Phylo-hmm for finding conserved sequence elements

• Motivation
  – About 5% of the genome is under purifying selection (based on comparative study of human and mouse)
  – Only 1.5% of this codes for proteins

• How to go from 2-way comparison to n-way comparisons to identify conserved sequence elements
  – That may not look like proteins
  – Based on a phylogenetic hidden Markov model (phyloHMM)
A Phylo-hmm for finding conserved sequence elements

- Two state Phylo-HMM
  - $c$: for conserved elements
  - $n$: for non-conserved elements
- Conserved elements were predicted using the Viterbi algorithm
- Single base pair conservation was obtained using the posterior probability of a state given a sequence
Applying Phastcons to real data

- Four insect species
  - 472,000 conserved elements
- Seven yeast species
  - 68,000 conserved elements
- Two worm species
  - 98,000 conserved elements
- Five vertebrate species
  - 1.31 million conserved elements
Types of conserved elements found in different genomes
Phastcons score of conservation profile

Figure 4. Screen shots of the conservation tracks in the (A) human and (B) S. cerevisiae UCSC Genome Browsers. Each conservation track has two parts, a plot of conservation scores, and beneath it, a display showing where each of the other genomes aligns to the reference genome. (Darker shading indicates higher BLASTZ scores; white indicates no alignment.) A separate track labeled "PhastCons Conserved Elements" shows predicted conserved elements and log-odds scores. In A, exons 7–11 of the RNA-edited human gene GRIA2 are shown. Peaks in the conservation plot generally correspond to exons and valleys to noncoding regions, but a 158-bp conserved noncoding element can be seen near the 3′ end of exon 11. This conserved element includes the editing complementary sequence (ECS) of the RNA editing site in exon 11. The displays seen when zooming in to the base level at a typical exon (left) and in the region of the RNA editing site (right; see arrow) are shown as insets. On the left, several synonymous substitutions are visible (highlighted bases) and the elevated conservation abruptly ends after the splice site, while on the right, there are fewer synonymous substitutions and the elevated conservation extends into the intron. In the base-level display, the vertical orange bars and numbers above them indicate "hidden" indels and their lengths—i.e., deletions in the human genome or insertions in other genomes. In B, the S. cerevisiae GAL1 gene and 5′-flanking region are shown. Strong cross-species conservation can be seen in the regulatory region upstream of the promoter, as well as in the protein-coding portion of the gene. The conserved element shown at bottom overlaps three GAL4-binding sites (highlighted in base-level view). A fourth GAL4-binding site also is reflected by a small bump in the conservation scores (left arrow), as is the promoter itself (right arrow).
Two applications of HMMs

• Phastcons: Based on PhyloHMMs
  – Aims to find conserved sequence elements in the genome

• ChromHMM:
  – Aims to segment the genome based on multiple chromatin modifications
Chromatin organization and gene expression

Transcriptional regulation

Epigenetic regulation

Post-transcriptional regulation

RNA Polymerase

Gene

DNA

RNA

Binding site

Histone modifications

Chromatin & histone binding proteins

[modENCODE consortium et al. Science, 2010]

http://www.youtube.com/watch?v=eYrQ0EhVCYA
ChIP-seq to measure histone data

Measuring the regulome (e.g., protein-binding of the genome)

Chromatin Immunoprecipitation (ChIP) + fragmentation

DNA

Protein-bound DNA fragments

reads

peak caller

genomic intervals bound by proteins

Adapted from Peter Park Nature Genetics Review
ChIP-seq data for multiple chromatin marks

Chromatin state: A specific combinations of mark values. Important because it can be used to segment the genome into biologically meaningful units.
Problem definition

• Given
  – A collection of genome-wide measurements of $m$ chromatin marks

• Do
  – Segment the genome into $M$ chromatin states
Binarizing the chromatin data

- Each mark \( j \) at bin \( t \) is represented by a binary variable \( v_{tj} \):
  - 1: mark is present
  - 0: mark is absent
An HMM for segmenting genomes using chromatin marks

- The ChromHMM model is defined as the tuple

\[ \theta = \{ S, E, A \} \]

- **Collection on** \( M \) **states**

- **Emission probabilities**

- **State transition probabilities**

- \( S = \{ S_1, \cdots, S_M \} \)

- **\( E \)** needs to model a collection of \( m \) binary variables

\[ E = \{ p_1, \cdots, p_M \} \]

\[ p_k = \{ p_{k,1}, \cdots, p_{k,m} \} \]

- \( p_{k,m} \): the probability of a mark \( m \) being present in state \( k \)
Emission probability of Chrom HMM

• Emission probability of $m$ marks per state is a product of $m$ Bernoulli random variables.

$$P(v_{t1}, \cdots, v_{tm} | \pi_t = k) = \prod_{j=1}^{m} p_{k,m}^{v_{tm}} (1 - p_{k,m})^{1-v_{tm}}$$

Collection of binary variables representing mark presence/absence
Learning the ChromHMM

• The number of states, $M$ are unknown
• Learn HMMs for $M=2$ to 80 states with a penalty factor to penalize the number of parameters
• State transitions: start with the fully connected HMM, and set parameters to zero if $<10^{-10}$
• Final model had 51 states
Learned Emission parameters for 51 states
Example output around CAPZA2 gene from ChromHMM

Input chromatin marks

Inferred state sequences

Promoter states

Transcribed states

Active intergenic

Repressed

Repetitive

Chromatin marks

Chromatin states

State 3
State 6
State 7
State 8
State 10
State 11
State 13
State 15
State 16
State 17
State 18
State 19
State 24
State 25
State 26
State 36
State 37
State 38
State 39
State 43
State 44
State 45

CAPZA2

Repetitive

Repressed

Active intergenic

Transcribed states

Promoter states
Posterior probability distributions of all 51 states around CAPZA gene

Max posterior state

Posterior probability values of each state
Summary

• HMMs are powerful models to capture sequential data
  – Sequence data can be different types
    • DNA sequence
    • Chromatin signals
• ChromHMM
  – Emission: \( m \) binary values
  – Number of states were learned
• PhyloHMM (Phastcons)
  – Emission: A multiple sequence alignment
  – Two states: conserved/not conserved