Machine Learning

10-701/15-781, Fall 2006

Machine Learning in Computational Biology

Eric Xing

Lecture 21, November 28, 2006

Reading:
The Central Dogma

(Griffiths et al. 1996)
Genome and Proteome
Gene Structure in DNA

- The inference problem: predicting locations of the genes on DNA
Proteins are coded by DNA

- There are between 30,000 to 40,000 genes in the human genome

- The human gene inventory corresponds to ~1.5% of the genome (coding regions)
Protein Structure Hierarchy

<table>
<thead>
<tr>
<th>Primary Structure</th>
<th>Secondary Structures</th>
<th>Tertiary Structures</th>
<th>Quaternary Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Amino acid structure" /></td>
<td><img src="image2" alt="Alpha-helix" /></td>
<td><img src="image3" alt="Beta-helix" /></td>
<td><img src="image4" alt="Triple beta-spiral" /></td>
</tr>
<tr>
<td><img src="image5" alt="Carboxy group" /></td>
<td><img src="image6" alt="Anti-parallel beta-sheet" /></td>
<td><img src="image7" alt="Parallel beta-sheet" /></td>
<td></td>
</tr>
<tr>
<td>… LACA A EECS …</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The inference problem: predicting the structures from sequences

APAFSVSPASGACGPECA

![Structure prediction](image8)
Genetic Polymorphisms

### The ABO Blood System

<table>
<thead>
<tr>
<th>Blood Type (genotype)</th>
<th>Type A (AA, AO)</th>
<th>Type B (BB, BO)</th>
<th>Type AB (AB)</th>
<th>Type O (OO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Blood Cell Surface Proteins (phenotype)</td>
<td><img src="#" alt="Type A" /></td>
<td><img src="#" alt="Type B" /></td>
<td><img src="#" alt="Type AB" /></td>
<td><img src="#" alt="Type O" /></td>
</tr>
<tr>
<td>A agglutinogens only</td>
<td>B agglutinogens only</td>
<td>A and B agglutinogens</td>
<td>No agglutinogens</td>
<td></td>
</tr>
<tr>
<td>Plasma Antibodies (phenotype)</td>
<td><img src="#" alt="Type A" /></td>
<td><img src="#" alt="Type B" /></td>
<td><img src="#" alt="Type AB" /></td>
<td><img src="#" alt="Type O" /></td>
</tr>
<tr>
<td>b agglutinin only</td>
<td>a agglutinin only</td>
<td>No agglutinin</td>
<td>a and b agglutinin</td>
<td></td>
</tr>
</tbody>
</table>
Genetic Demography

- Are there genetic prototypes among them?
- What are they?
- How many? (how many ancestors do we have?)
Single Nucleotide Polymorphism (SNP)

The inference problem: "haplotypes" and population diversity
Computation Biology and ML

- Mixture and infinite mixture
  - clustering of genetic polymorphisms

- Hidden Markov Models
  - gene finding

- Trees
  - sequence evolution

- Conditional Random Fields
  - protein structure prediction
Computation Biology and ML

- Mixture and infinite mixture
  - clustering of genetic polymorphisms
- HMMs
  - gene finding
- Trees
  - sequence evolution
- CRMs
  - protein structure prediction
**Biological Terms**

- **Genetic polymorphism**: a difference in DNA sequence among individuals, groups, or populations

- **Single Nucleotide Polymorphism (SNP)**: DNA sequence variation occurring when a single nucleotide - A, T, C, or G - differs between members of the species
  
  - Each variant is called an “allele”
  - Almost always bi-allelic
  - Account for most of the genetic diversity among different (normal) individuals, e.g. drug response, disease susceptibility
From SNPs to Haplotypes

- Alleles of adjacent SNPs on a chromosome form **haplotypes**

- Useful in the study of **disease association or genetic evolution**
Phase ambiguity of SNPs "haplotypes"

- This is a mixture modeling problem!

A heterozygous diploid individual

The Genotype:
- pairs of alleles with association of alleles to chromosomes unknown

Haplotype \( h=(h_1, h_2) \)
- possible associations of alleles to chromosome
Haplotype Inference

Why is it approachable?
- Many of the haplotypes appear many times
- Data for many individuals allows inference

Solution seems ‘better’ since it uses fewer haplotypes
Finite mixture model

- The probability of a genotype $g$:

$$ p(g) = \sum_{h_1, h_2 \in H} p(h_1, h_2) p(g \mid h_1, h_2) $$

- Standard settings:
  - $p(h_1, h_2) = p(h_1)p(h_2)$  
    Hardy-Weinberg equilibrium
  - $|H| = K$  
    fixed-sized population haplotype pool

- Problem:  $K$  \?  $H$  \?
Ancestral Inference

- Better recovery of the ancestors leads to better haplotyping results (because of more accurate grouping of common haplotypes)
- True haplotypes are obtainable with high cost, but they can validate model more subjectively (as opposed to examining saliency of clustering)
- Many other biological/scientific utilities
Being Bayesian about ...

- Population haplotype identities
- Population haplotype frequencies
- Number of population haplotypes
- Associations between population haplotype and individual haplotype/genotype
A Hierarchical Bayesian Infinite Allele model

Bayesian Haplotype Inference via the Dirichlet Process (Xing et al. ICML2004)

• Assume an individual haplotype $h$ is stochastically derived from a population haplotype $a_k$ with nucleotide-substitution frequency $\theta_k$:

$$h \sim p(h \mid \{a, \theta\}_k).$$

• Not knowing the correspondences between individual and population haplotypes, each individual haplotype is a mixture of population haplotypes.

• The number and identity of the population haplotypes are unknown
  
  – use a Dirichlet Process to construct a prior distribution $G$ on $\mathcal{H} \times \mathcal{R}^J$.

• Inference: Markov Chain Monte Carlo
Chinese Restaurant Process

\[ P(c_i = k \mid c_{-i}) = \begin{align*}
\frac{1}{1+\alpha} & \quad 0 & \quad 0 \\
\frac{1}{2+\alpha} & \quad \frac{\alpha}{1+\alpha} & \quad \frac{\alpha}{2+\alpha} \\
\frac{1}{3+\alpha} & \quad \frac{1}{2+\alpha} & \quad \frac{\alpha}{3+\alpha} \\
\frac{m_i}{i+\alpha-1} & \quad \frac{m_2}{i+\alpha-1} & \quad \ldots \quad \frac{\alpha}{i+\alpha-1}
\end{align*} \]

CRP defines an exchangeable distribution on partitions over an (infinite) sequence of integers
The DP Mixture of Ancestral Haplotypes

- The customers around a table form a cluster
  - associate a mixture component (i.e., a population haplotype) with a table
  - sample \( \{a, \theta\} \) at each table from a base measure \( G_0 \) to obtain the population haplotype and nucleotide substitution frequency for that component

With \( p(h|\{A, \theta\}) \) and \( p(g|h_1,h_2) \), the CRP yields a posterior distribution on the number of population haplotypes (and on the haplotype configurations and the nucleotide substitution frequencies)
Convergence of Ancestral Inference
Results on simulated data

- DP vs. Finite Mixture via EM
Results

The Gabriel data

Haplotyping performance

<table>
<thead>
<tr>
<th>Region</th>
<th>CRP</th>
<th>PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>16a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Haplotyping performance

<table>
<thead>
<tr>
<th>Region</th>
<th>CRP</th>
<th>PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>16a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Computation Biology and ML

- Mixture and infinite mixture
  - clustering of genetic polymorphisms
- HMMs
  - gene finding
- Trees
  - sequence evolution
- CRMs
  - protein structure prediction
The challenge

Eric Xing 26
Typical structure of a gene
**Gene Finding**

- Given un-annotated sequences,
- delineate:
  - transcription initiation site,
  - exon-intron boundaries,
  - transcription termination site,
  - a variety of other motifs: promoters, polyA sites, branching sites, etc.

- The hidden Markov model (HMM)
Hidden Markov Models

The underlying source: genomic entities, dice,
The sequence: Ploy NT, sequence of rolls,
Definition (of HMM)

- **Observation space**
  - Alphabetic set: \( C = \{c_1, c_2, \ldots, c_K\} \)
  - Euclidean space: \( \mathbb{R}^d \)
- **Index set of hidden states**
  - \( I = \{1, 2, \ldots, M\} \)
- **Transition probabilities** between any two states
  \[
  p(y_t^i = 1 | y_{t-1}^j = 1) = a_{i,j},
  \]
  or
  \[
  p(y_t | y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \ldots, a_{i,M}), \forall i \in I.
  \]
- **Start probabilities**
  \[
  p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M).
  \]
- **Emission probabilities** associated with each state
  \[
  p(x_t | y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \ldots, b_{i,K}), \forall i \in I.
  \]
  or in general:
  \[
  p(x_t | y_t^i = 1) \sim f(\cdot | \theta_i), \forall i \in I.
  \]
GENSCAN (Burge & Karlin)

\[
p(\cdot | y) = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}
\]
The Idea Behind a GHMM

GeneFinder

- **States** represent standard gene features: intergenic region, exon, intron, perhaps more (promotor, 5’UTR, 3’UTR, Poly-A,..).

- **Observations** embody state-dependent base composition, dependence, and signal features.

- In a GHMM, **duration** must be included as well.

- Finally, **reading frames** and **both strands** must be dealt with.
The HMM Algorithms

Questions:

- **Evaluation**: What is the probability of the observed sequence? **Forward**

- **Decoding**: What is the probability that the state of the 3rd position is Bk, given the observed sequence? **Forward-Backward**

- **Decoding**: What is the most likely parsing? **Viterbi**

- **Learning**: Under what parameterization are the observed sequences most probable? **Baum-Welch (EM)**
Computation Biology and ML

- Mixture and infinite mixture
  - clustering of genetic polymorphisms
- HMMs
  - gene finding
- Trees
  - sequence evolution
- CRMs
  - protein structure prediction
A pair of homologous bases

Typically, the ancestor is unknown.
Homology identification via multiple alignment
The shaded nodes represent the observed nucleotides at a given site for a set of organisms.
The unshaded nodes represent putative ancestral nucleotides.
Transitions between nodes capture the dynamic of evolution.
Phylogeney methods

- **Basic principles:**
  - Degree of sequence difference is proportional to length of independent sequence evolution
  - Only use positions where alignment is pretty certain – avoid areas with (too many) gaps

- **Major methods:**
  - Parsimony phylogeny methods
  - Likelihood methods
Likelihood methods

- A tree, with branch lengths, and the data at a single site.

Since the sites evolve independently on the same tree,

\[ L = P(D \mid T) = \prod_{i=1}^{m} P(D^{(i)} \mid T) \]
Likelihood at one site on a tree

- We can compute this by summing over all assignments of states $x$, $y$, $z$ and $w$ to the interior nodes:

$$P(D^{(i)} \mid T) = \sum_{x} \sum_{y} \sum_{z} \sum_{w} P(A, A, C, C, x, y, z, w \mid T)$$

- Due to the Markov property of the tree, we can factorize the complete likelihood according to the tree topology:

$$P(A, A, C, C, x, y, z, w \mid T) =$$

$$P(x) \cdot P(y \mid x, t_6) \cdot P(C \mid y, t_1) \cdot P(C \mid y, t_2) \cdot P(z \mid x, t_8) \cdot P(C \mid y, t_3) \cdot P(w \mid z, t_7) \cdot P(A \mid y, t_4) \cdot P(A \mid y, t_5)$$

- Summing this up, there are 256 terms in this case!
Getting a recursive algorithm

- When we move the summation signs as far right as possible:

\[
P(D^{(i)} | T) = \sum_{x} \sum_{y} \sum_{z} \sum_{w} P(A, A, C, C, x, y, z, w | T) =
\]

\[
\left( \sum_{x} P(x) \right) \\
\left( \sum_{y} P(y | x, t_6) P(C | y, t_1) P(C | y, t_2) \right) \\
\left( \sum_{z} P(z | x, t_8) P(C | z, t_3) \right)
\]

\[
\left( \sum_{w} P(w | z, t_7) P(A | w, t_4) P(A | w, t_5) \right)
\]
Felsenstein’s Pruning Algorithm

- To calculate $P(x_1, x_2, \ldots, x_N \mid T, t)$

**Initialization:**
Set $k = 2N - 1$

**Recursion:** Compute $P(L_k \mid a)$ for all $a \in \Sigma$

  - If $k$ is a leaf node:
    - Set $P(L_k \mid a) = 1(a = x_k)$
  - If $k$ is not a leaf node:
    1. Compute $P(L_i \mid b), P(L_j \mid b)$ for all $b$, for daughter nodes $i, j$
    2. Set $P(L_k \mid a) = \sum_b, c P(b \mid a, t_i) P(L_i \mid b) P(c \mid a, t_j) P(L_j \mid c)$

**Termination:**

Likelihood at this column = $P(x_1, x_2, \ldots, x_N \mid T, t) = \sum_a P(L_{2N-1} \mid a)P(a)$
There are a finite number of rates (denote rate i as $r_i$).

There are probabilities $p_i$ of a site having rate i.

A process not visible to us ("hidden") assigns rates to sites.

The probability of our seeing some data are to be obtained by summing over all possible combinations of rates, weighting appropriately by their probabilities of occurrence.
The shaded nodes represent the observed nucleotides at particular sites of an organism's genome.

For discrete $Y_i$, widely used in computational biology to represent segments of sequences:

- gene finders and motif finders
- profile models of protein domains
- models of secondary structure
Definition (of HMM)

- **Observation space**
  
  - Alphabetic set: \( C = \{c_1, c_2, \ldots, c_K\} \)
  
  - Euclidean space: \( \mathbb{R}^d \)

- **Index set of hidden states**
  
  \( I = \{1, 2, \ldots, M\} \)

- **Transition probabilities** between any two states
  
  \[ p(y_t^i = 1 | y_{t-1}^j = 1) = a_{i,j}, \]
  
  or
  
  \[ p(y_t | y_{t-1}^j = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \ldots, a_{i,M}), \forall i \in I. \]

- **Start probabilities**
  
  \[ p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M). \]

- **Emission probabilities** associated with each state
  
  \[ p(x_t | y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \ldots, b_{i,K}), \forall i \in I. \]
  
  or in general:

  \[ p(x_t | y_t^i = 1) \sim f(\cdot | \theta_i), \forall i \in I. \]
Hidden Markov Phylogeny

- Replacing the standard emission model with a tree
  - A process not visible to us ("hidden") assigns rates to sites. It is a Markov process working along the sequence.
  - For example it might have transition probability \( \text{Prob}(j|i) \) of changing to rate \( j \) in the next site, given that it is at rate \( i \) in this site.

- These are the most widely used models allowing rate variation to be correlated along the sequence.
Hidden Markov Phylogeny

- this yields a gene finder that exploits evolutionary constraints
Based on sequence data from 12-15 primate species, McAuliffe et al (2003) obtained sensitivity of 100%, with a specificity of 89%.

- Genscan (state-of-the-art gene finder) yield a sensitivity of 45%, with a specificity of 34%.