Machine Learning

10-701/15-781, Spring 2008

Machine Learning in Computational Biology

Eric Xing

Reading:

The Central Dogma

[Diagram showing the process of DNA replication, transcription, RNA translation, and protein synthesis]

[Image credit: Xue et al., 2005]
**Genome and Proteome**

- Genes on DNA
- The inference problem: predicting locations of the genes on DNA

**Gene Structure in DNA**

- Promoter
- Exon
- Intron
- Transcription start
- 5' UTR
- Start codon
- Donor site
- Amino acid translation
- Stop codon
- 3' UTR
- Acceptor site
- Poly-A site
- UTRs

---

Eric Xing
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

- The inference problem: predicting the structures from sequences

**Primary Structure**

**Secondary Structures**

**Tertiary Structures**

**Quaternary Structures**

---

**APAFSVSPASGCPECA**
Genetic Polymorphisms

The ABO Blood System

<table>
<thead>
<tr>
<th>Blood Type (genotype)</th>
<th>Type A (AA, AB)</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>A agglutinin only</td>
<td>B agglutinin only</td>
<td>A and B agglutinogens</td>
<td>No agglutinogens</td>
</tr>
<tr>
<td>Plasma Antibodies (phenotype)</td>
<td>b agglutinin only</td>
<td>a agglutinin only</td>
<td>No agglutinin</td>
<td>a and b agglutinins</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!
**Haplotype Inference**

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

![Haplotype Inference Diagram]

Solution seems 'better' since it uses fewer haplotypes

**Finite mixture model**

- The probability of a genotype $g$:
  \[
p(g) = \sum_{h_1, h_2 \in H} p(h_1, h_2) p(g | h_1, h_2)
  \]

- 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

Essentially a clustering problem, but ...

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. ICML 2004)

• 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 $\mathcal{H} \times \mathbb{R}^J$.

• Inference: Markov Chain Monte Carlo

Chinese Restaurant Process

$$P(c_i = k \mid c_d) = \begin{array}{ccc} \frac{1}{1+\alpha} & \frac{0}{\alpha} & \frac{0}{\alpha} \\ \frac{1}{2+\alpha} & \frac{1}{2+\alpha} & \frac{\alpha}{2+\alpha} \\ \frac{1}{3+\alpha} & \frac{2}{3+\alpha} & \frac{\alpha}{3+\alpha} \\ \frac{m_i}{i+\alpha-1} & \frac{m_i}{i+\alpha-1} & \frac{\alpha}{i+\alpha-1} \end{array}$$

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

![Diagram of haplotype configurations]

- 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

![Graph showing convergence]

Eric Xing
Results on simulated data

- DP vs. Finite Mixture via EM

![Graph showing comparison between DP and EM for different data sets.]

Results

The Gabriel data

![Two graphs showing haplotyping performance for different regions with DRP and PHASE.]
Population structure

- DATA: 256 European individuals with 103 loci

Computation Biology and ML

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

Typical structure of a gene

Eric Xing
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
  \[ p(y_t^i = 1 \mid y_{t-1}^j = 1) = a_{ij}, \]
  or
  \[ p(y_t^i \mid y_{t-1}^j = 1) \sim \text{Multinomial} \{ a_{1j}, a_{2j}, \ldots, a_{Mj} \}, \forall i \in I. \]
- **Start probabilities**
  \[ p(y_1^i) \sim \text{Multinomial} \{ \pi_1, \pi_2, \ldots, \pi_M \} \]
- **Emission probabilities** associated with each state
  \[ p(x_t^i \mid y_t^j = 1) \sim \text{Multinomial} \{ b_{1j}, b_{2j}, \ldots, b_{Mj} \}, \forall i \in I. \]
  or in general:
  \[ p(x_t^i \mid y_t^j = 1) \sim f(\cdot \mid \theta_i), \forall i \in I. \]

**GENSCAN (Burge & Karlin)**

![Graphical model of HMM](image)
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.
**Phylogeny 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:
  \[ P(D^{(i)} | T) = \sum_{x} \sum_{y} \sum_{z} \sum_{w} P(A, A, C, C, x, y, z, w | T) \]
- Due to the Markov property of the tree, we can factorize the complete likelihood according to the tree topology:
  \[ P(A, A, C, C, x, y, z, w | T) = P(x) \left( \sum_{y} P(y | x, t_4) \right) \left( \sum_{z} P(C | y, t_3) P(C | y, t_2) \right) \left( \sum_{w} P(z | x, t_4) P(C | z, t_1) \right) \left( \sum_{w} P(z | w, t_3) P(A | w, t_4) P(A | w, t_3) \right) \]
- 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:
  \[ P(D^{(i)} | T) = \sum_{x} \sum_{y} \sum_{z} \sum_{w} P(A, A, C, C, x, y, z, w | T) = \]
  \[ = \sum_{x} P(x) \left( \sum_{y} P(y | x, t_4) \right) \left( \sum_{z} P(C | y, t_3) P(C | y, t_2) \right) \left( \sum_{w} P(z | x, t_4) P(C | z, t_1) \right) \left( \sum_{w} P(w | z, t_3) P(A | w, t_4) P(A | w, t_3) \right) \]
Felsenstein’s Pruning Algorithm

- To calculate $P(x_1, x_2, ..., x_N | T, t)$

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

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

  - If $k$ is a leaf node:
    
    Set $P(L_k | a) = 1(a = x_k)$

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

  **Termination:**

  Likelihood at this column = $P(x_1, x_2, ..., x_N | T, t) = \sum_a P(L_{2N-1} | a) P(a)$

---

Modeling rate variation among sites

- 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.
Recall the HMM

The shaded nodes represent the observed nucleotides at particular sites of an organism’s genome.

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
  - Alphabet 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
  \[ p(y_i | y_{i-1}) = \frac{a_{y_i y_{i-1}}}{a_{y_i}} \]
- $p(Y_i | Y_{i-1}) \sim \text{Multinomial}(a_{1}, a_{2}, \ldots, a_{M}) \forall i \in I$.
- Start probabilities
  \[ p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M) \]
- Emission probabilities associated with each state
  \[ p(x_i | y_i) = 1 \sim \text{Multinomial}(b_1, b_2, \ldots, b_X) \forall i \in I. \]
- Or in general:
  \[ p(x_i | y_i) = 1 \sim f(\cdot | \theta_i) \forall i \in I. \]
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%.