Applications of HMMs in Epigenomics
Yuzhen Ye
School of Informatics and Computing
Indiana University, Bloomington
Spring 2013

Background
- Genomic DNA is packaged into a complex molecular structure known as chromatin. This structure mediates the interaction between the genome and all types of regulatory and transcriptional molecules.
- In vertebrate genomes, methylation at position 5 of the cytosine in CpG dinucleotides is a heritable "epigenetic" mark that has been connected with both transcriptional silencing and imprinting
  - Ref: DNA methylation patterns and epigenetic memory (Genes & Dev. 2002. 16: 6-21)

ENCODE
- Encyclopedia of DNA Elements
  - "The ENCODE Consortium is integrating multiple technologies and approaches in a collective effort to discover and define the functional elements encoded in the human genome, including genes, transcripts, and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns."
- Initial phase launched in 2003—1% of the human genome
  - Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project (Nature, June 13, 2007)
Human epigenome atlas
• Successive releases of the Atlas will provide progressively more detailed insights into locus-specific epigenomic states, including histone marks and DNA methylation marks across specific tissues and cell types, developmental stages, physiological conditions, genotypes, and disease states.

ChIP-Seq
• Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) has become a valuable and widely used approach for mapping the genomic location of transcription-factor binding and histone modifications in living cells.
  • Genome-Wide Mapping of in Vivo Protein-DNA Interactions (Science, 2007); 1946 binding sites of the Neuron-restrictive silencer factor (NRSF) were mapped at ~50bp resolution

• There are considerable differences in how these experiments are conducted, how the results are scored and evaluated for quality, and how the data and metadata are archived for public use.
  • Genome Res., 2012 Sep;22(9):1813-31
Barcoded ChIP-Seq

Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing (BMC Genomics 2009, 10:37)

CHIP-Seq: peak detection

Typical ChIP-seq peak

Low-complexity ChIP-seq peak

Genome-wide DNA methylation profiling

- **Restriction enzyme-based methods**
  - Use one or more enzymes that will restrict DNA only if it is unmethylated (e.g. HpaII or NsiI), or methylated (e.g. MspI).
  - Limited to the analysis of CpG sites located within the enzyme recognition site(s).

- **Bisulfite-conversion based approaches**
  - Unmethylated cytosines are converted to uracil; offer single CpG resolution; the gold standard for DNA methylation analysis.
  - Con: reduction of sequence complexity following bisulfite conversion (Bi-chip) & Bi-seq approach is expensive.
  - Align BS-treated reads to a reference genome

- **Immunoprecipitation-based methods**
  - Use either 5-methylcytosine-specific antibodies (MeDIP) or methyl-binding domain proteins, to enrich for the methylated (or unmethylated) fraction of the genome.

Methylation analysis by DNA immunoprecipitation (MeDIP)

MeDIP can be coupled with microarray or high-throughput sequencing

Long-range chromatin interaction

Long-range Chromatin Interactions: Chromosome Conformation Capture Carbon Copy (5C)

A HMM application for the inference of DNA methylation

- MeDIP-HMM: genome-wide identification of distinct DNA methylation states from high-density tiling arrays
- MeDIP-HMM utilizes a higher-order state-transition process improving modeling of spatial dependencies between chromosomal regions
- Enables a differentiation between unmethylated, methylated and highly methylated genomic regions.
- Training algorithm: a Bayesian Baum-Welch algorithm integrating prior knowledge on methylation levels.
- Application of MeDIP-HMM to the analysis of the Arabidopsis root methylome and systematically investigate the benefit of using higher-order HMMs.
MeDIP-HMM: three-state architecture

Second-order HMM
Multivariate Gaussian Emission Distribution:
\[ k(\theta) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left(-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1} (\theta - \mu)\right) \]

Chromatin-state decoding
- Automated mapping of large-scale chromatin structure in ENCODE
- ChromHMM: automating chromatin-state discovery and characterization

Integrative annotation of chromatin elements from ENCODE data

Table 1. Major differences between ChromHMM and Segway as applied to the ENCODE data

<table>
<thead>
<tr>
<th>Modeling framework</th>
<th>ChromHMM</th>
<th>Segway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic resolution</td>
<td>100 bp</td>
<td>1 bp</td>
</tr>
<tr>
<td>Handling missing data</td>
<td>Interpolation</td>
<td>Marginalization</td>
</tr>
<tr>
<td>Emission modeling</td>
<td>Bernoulli distribution</td>
<td>Gaussian distribution</td>
</tr>
<tr>
<td>Training set</td>
<td>Entire genome</td>
<td>ENCODE regions (1%)</td>
</tr>
<tr>
<td>Decoding algorithm</td>
<td>Posterior decoding</td>
<td>Viterbi</td>
</tr>
<tr>
<td>Learning across cell types</td>
<td>Single model for all cell types</td>
<td>One model per cell type</td>
</tr>
</tbody>
</table>


ChromHMM is a multivariate HMM
- ChromHMM uses a multivariate HMM that explicitly models the combination of marks
Multivariate HMM (formal definition)

- A multivariate HMM $\theta$ has
  - $N$ sets of observation symbols, each for one given observation sequence $n$ ($n=1, 2, ..., N$)
  - A set of hidden states
  - Transition probabilities $a_{ij}$ for any pair of hidden states $i$ and $j$
  - Initial probabilities $b_i$ for any hidden states $j$
  - $N$ sets of emission probabilities $e_{is}$ for the observation symbol being emitted in the $n$th observation sequence from the hidden state $s$.

Multivariate HMM

- Given $N$ observation sequences of the same length $L$, $X=\{(x_{1,1}, x_{1,2}, ..., x_{1,L}), ..., (x_{N,1}, x_{N,2}, ..., x_{N,L})\}$ and the hidden state sequence $S=\{s_1, s_2, ..., s_L\}$, the full probability from the multivariate HMM is,

$$P(S, X | \theta) = \prod_{n=1}^{N} \prod_{j=1}^{L} e_{is}(x_{n,j})$$

Let $e_{s_1, s_2, ..., s_L} = \prod_{n=1}^{N} e_{is}(x_{n,j})$, the multivariate HMM can be reduced to conventional HMM, except the observation symbol becomes a vector $(x_{n,1}, x_{n,2}, ..., x_{n,L})$ at position $j$. The same algorithms for model inference (Viterbi and forward/backward) and learning can be used.