Applications of HMMs in Epigenomics

Yuzhen Ye
School of Informatics and Computing
Indiana University, Bloomington
Spring 2013
Contents

- Background: chromatin structure & DNA methylation
- Epigenomic projects
  - ENCODE
  - modENCODE
  - Human epigenome atlas
- Techniques
  - ChIP-Seq
  - MeDIP (MeDIP-chip & MeDIP-seq)
- Two applications
  - MeDIP-HMM (second-order HMM)
  - ChromHMM (multivariate HMM)
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)
Figure 1. The Organization of the ENCODE Consortium.

http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046
Table 1. Experimental assays used by the ENCODE Consortium.

<table>
<thead>
<tr>
<th>Gene/Transcript Analysis</th>
<th>Method/Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene annotation</td>
<td>GENCODE; Wellcome Trust</td>
</tr>
<tr>
<td>PolyA+ coding regions</td>
<td>RNA-seq; tiling DNA microarrays; PET; CSHL; Stanford/Yale/Harvard; Caltech</td>
</tr>
<tr>
<td>Total RNA coding regions</td>
<td>RNA-seq; tiling DNA microarrays; PET; CSHL</td>
</tr>
<tr>
<td>Coding regions in subcellular RNA fractions (e.g. nuclear, cytoplasmic)</td>
<td>PET; CSHL</td>
</tr>
<tr>
<td>Small RNAs</td>
<td>short RNA-seq; CSHL</td>
</tr>
<tr>
<td>Transcription initiation (5'-end) and termination (3'-end') sites</td>
<td>CAGE; ditAGs; RIKEN, GIS</td>
</tr>
<tr>
<td>Full-length RNAs</td>
<td>RACE; University of Geneva; University of Lausanne</td>
</tr>
<tr>
<td>Protein-bound RNA coding regions</td>
<td>RIP; CLIP; SUNY-Albany; CSHL</td>
</tr>
</tbody>
</table>

Transcription Factors/Chromatin

<table>
<thead>
<tr>
<th>Elements/Regions</th>
<th>Method(s)</th>
<th>Group(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription Factor Binding Sites (TFBS)</td>
<td>ChIP-seq</td>
<td>Stanford/Yale/UC-Davis/Harvard; HudsonAlpha/Caltech; Duke/UT-Austin; UW; U. Chicago/Stanford</td>
</tr>
<tr>
<td>Chromatin structure (accessibility, etc.)</td>
<td>DNasel hypersensitivity; FAIRE</td>
<td>UW; Duke; UNC</td>
</tr>
<tr>
<td>Chromatin modifications (H3K27ac, H3K27me3, H3K36me3, etc.)</td>
<td>ChIP-seq</td>
<td>Broad; UW</td>
</tr>
<tr>
<td>DNasel footprints</td>
<td>Digital genomic footprinting</td>
<td>UW</td>
</tr>
</tbody>
</table>

Other Elements/Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Method(s)</th>
<th>Group(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA methylation</td>
<td>RRBS; Illumina Methyl127; Methyl-seq</td>
<td>HudsonAlpha</td>
</tr>
<tr>
<td>Chromatin interactions</td>
<td>5C; ChIA-PET</td>
<td>UMass; UW; GIS</td>
</tr>
<tr>
<td>Genotyping</td>
<td>Illumina 1M Duo</td>
<td>HudsonAlpha</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pbio.1001046.t001

http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046
ENCODE data

chr21:33,031,597-33,041,570 9,974 bp.

Click on a feature for details. Click or drag in the base.

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
Figure 4. ENCODE chromatin annotations in the HLA locus.

http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046
Figure 5. Occupancy of transcription factors and RNA polymerase 2 on human chromosome 6p as determined by ChIP-seq.

http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046
The modENCODE Project will try to identify all of the sequence-based functional elements in the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes.
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
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 NotI), or methylated (e.g. McrBC).
  - 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.

Image from wikipedia
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:

\[ b_i(\bar{d}) := \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_i)}} \exp \left( -\frac{1}{2} (\bar{d} - \bar{\mu}_i) \cdot \Sigma_i^{-1} \cdot (\bar{d} - \bar{\mu}_i)^T \right) \]
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>
<thead>
<tr>
<th></th>
<th>ChromHMM</th>
<th>Segway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeling framework</td>
<td>Hidden Markov model</td>
<td>Dynamic Bayesian network</td>
</tr>
<tr>
<td>Genomic resolution</td>
<td>200 bp</td>
<td>1 bp</td>
</tr>
<tr>
<td>Data resolution</td>
<td>Boolean</td>
<td>Real value</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>Length modeling</td>
<td>Geometric distribution</td>
<td>Geometric plus hard and soft constraints</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 six cell types</td>
<td>Single model for all cell types</td>
<td>One model per cell type</td>
</tr>
</tbody>
</table>

Fib 1b. Segmentation results from ChromHMM

Hoffman M M et al. Nucl. Acids Res. 2012;nar.gks1284

© The Author(s) 2012. Published by Oxford University Press.
ChromHMM is a *multivariate* HMM

- ChromHMM uses a multivariate HMM that explicitly models the combination of marks

<table>
<thead>
<tr>
<th>Cell</th>
<th>chr1</th>
<th>Mark1</th>
<th>Mark2</th>
<th>Mark3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

absent

present
Multivariate HMM
Multivariate HMM (formal definition)

- A multivariate HMM $\theta$ has
  - $N$ sets of observation symbols, each for one given observation sequence $n$ ($n=1, 2, \ldots, N$)
  - A set of hidden states
  - Transition probabilities $a_{ij}$, for any pair of hidden states $i$ and $j$
  - Initial probabilities $B_j = a_{0j}$ for any hidden states $j$
  - $N$ sets of emission probabilities $e^n_{s}(x_n)$ 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 $L$, $X=\{(x_{1,1}...x_{1,L}), ..., (x_{N,1}...x_{N,L})\}$ and the hidden state sequence $S=(s_1...s_L)$, the full probability from the multivariate HMM is,

$$P(S,X \mid \theta) = \prod_{j=1}^{L} \left[ a_{s_{j-1}s_j} \prod_{n=1}^{N} e_{s_j}(x_{n,j}) \right]$$

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