Hidden Markov Models intro, Chromatin segmentation

Mikhail Dozmorov

Spring 2018

Markov Model (aka Markov Chain)

- Stochastic process - a random process, or a sequence or random variables

Andrey Markov, a Russian mathematician (1856 - 1922)

Markov Model (aka Markov Chain)

- A discrete stochastic process $X_1, X_2, X_3, ...$ has the Markov property

$$P(X_{n+1} = j | X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = j | X_n = x_n)$$

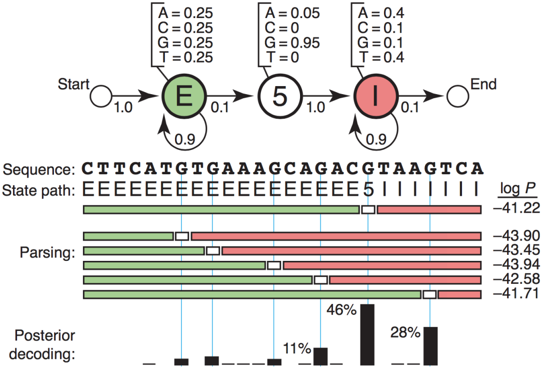
$$\textit{for all } x_i, \textit{all } j, \textit{all } n$$

- A random process which has the property that the future (next state) is conditionally independent of the past given the present (current state)

Elements of Hidden Markov Models

- An alphabet of n emitted symbols (e.g., “A”, “T”, “C”, “G”)
- A set of k hidden states (e.g., CG-island, regular sequence)
- Transition $= (transition_{l,k})$ - a $|States| \times |States|$ matrix of **transition probabilities** for changing from state l to state k
- Emission $= (emission_k(symbol_n))$ - a $|States| \times n$ matrix of **emission probabilities** (of emitting $symbol_n$ then the HMM is in state k)

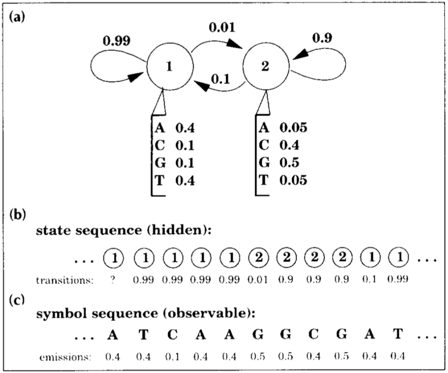
A simple hidden Markov model



The model generates two strings of information. One is the underlying *state path* (the labels), as we transition from state to state. The other is the *observed sequence* (the DNA), each residue being emitted from one state in the state path. The efficient Viterbi algorithm is guaranteed to find the most probable state path given a sequence and an HMM. The Viterbi algorithm is a dynamic programming algorithm quite similar to those used for standard sequence alignment.

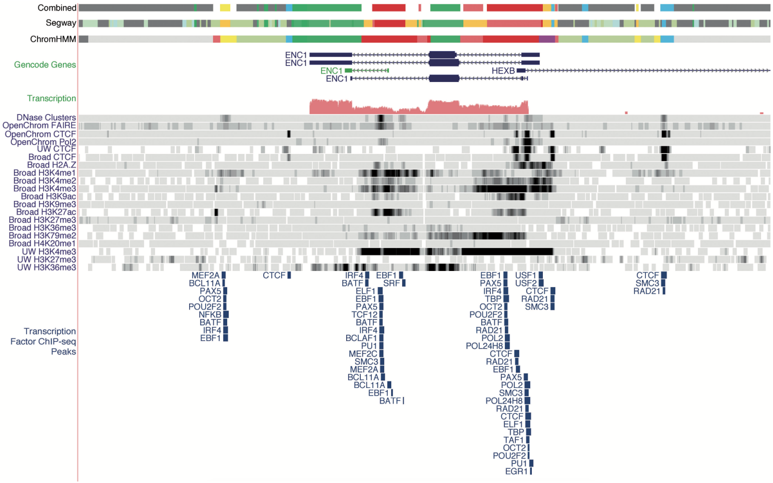
Eddy, Sean R. "What Is a Hidden Markov Model?" Nature Biotechnology 22, no. 10 (October 2004): 1315–16. https://doi.org/10.1038/nbt1004-1315

A simple hidden Markov model


(a)
0.99
0.01
0.9
1
2
0.1
A 0.4
C 0.1
G 0.1
T 0.4
A 0.05
C 0.4
G 0.5
T 0.05
(b)
state sequence (hidden):
... 1 1 1 1 1 2 2 2 2 1 1 ...
transitions: ? 0.99 0.99 0.99 0.99 0.01 0.9 0.9 0.9 0.1 0.99
(c)
symbol sequence (observable):
... A T C A A G G C G A T ...
emissions: 0.4 0.4 0.1 0.4 0.4 0.5 0.5 0.4 0.5 0.4 0.4

A two-state HMM describing DNA sequence with a heterogeneous base composition. (a) State 1 generates AT-rich sequence, and state 2 generates CG-rich sequence. State transitions and their associated probabilities are indicated by arrows, and symbol emission probabilities for A,C,G and T for each state are indicated below the states. (b) This model generates a state sequence as a Markov chain and each state generates a symbol according to its own emission probability distribution (c). The probability of the sequence is the product of the state transitions and the symbol emissions. For a given observed DNA sequence, we are interested in inferring the hidden state sequence that 'generated' it, that is, whether this position is in a CG-rich segment or an AT-rich segment.

Chromatin segmentation

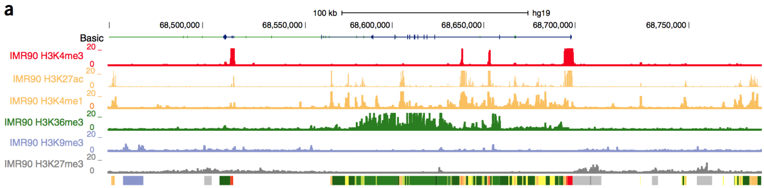


Hoffman, Michael M., Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, et al. "Integrative Annotation of Chromatin Elements from ENCODE Data." Nucleic Acids Research 41, no. 2 (January 2013): 827–41. https://doi.org/10.1093/nar/gks1284.

Ideas for chromatin track analysis

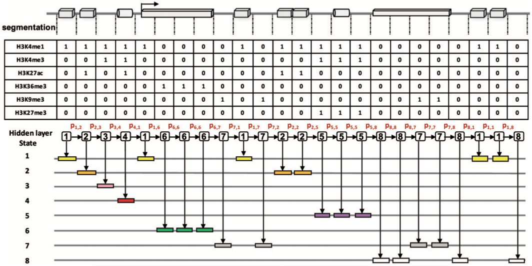
- Hidden Markov Model (ChromHMM)
- Dynamic Bayesian Network (Segway)
  - Bayesian Network that models data sampled at intervals. Still a directed acyclic graph (DAG).
  - Can learn model with Graphical Model Toolkit (GMTK)
  - Can incorporate relationships between variables and handle missing data
  - 1bp analysis resolution

ChromHMM



- chromHMM learns chromatin-state signatures using a multivariate hidden Markov model (HMM) that explicitly models the combinatorial presence or absence of each mark
- chromHMM uses these signatures to generate a genome-wide annotation for each cell type by calculating the most probable state for each genomic segment
- chromHMM provides an automated enrichment analysis of the resulting annotations to facilitate the functional interpretations of each chromatin state

ChromHMM

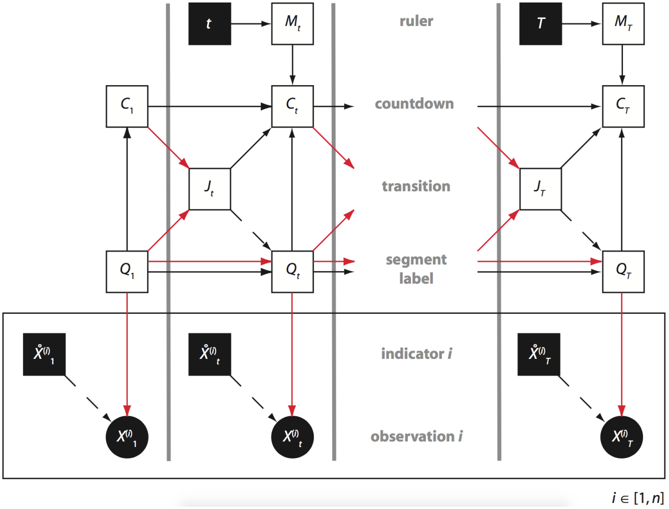

segmentation
Hidden layer State
$p_{1,2}$ $p_{2,3}$ $p_{3,4}$ $p_{4,1}$ $p_{1,6}$ $p_{6,6}$ $p_{6,6}$ $p_{6,7}$ $p_{7,1}$ $p_{1,7}$ $p_{7,2}$ $p_{2,2}$ $p_{2,5}$ $p_{5,5}$ $p_{5,5}$ $p_{5,8}$ $p_{8,8}$ $p_{8,7}$ $p_{7,7}$ $p_{7,8}$ $p_{8,1}$ $p_{1,1}$ $p_{1,8}$
1 2 3 4 1 6 6 6 7 1 7 2 2 5 5 5 8 8 7 7 8 1 1 8
1
2
3
4
5
6
7
8

H3K4me1	1	1	1	1	1	0	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	1	1	0
H3K4me3	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
H3K27ac	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
H3K36me3	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3K9me3	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
H3K27me3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0

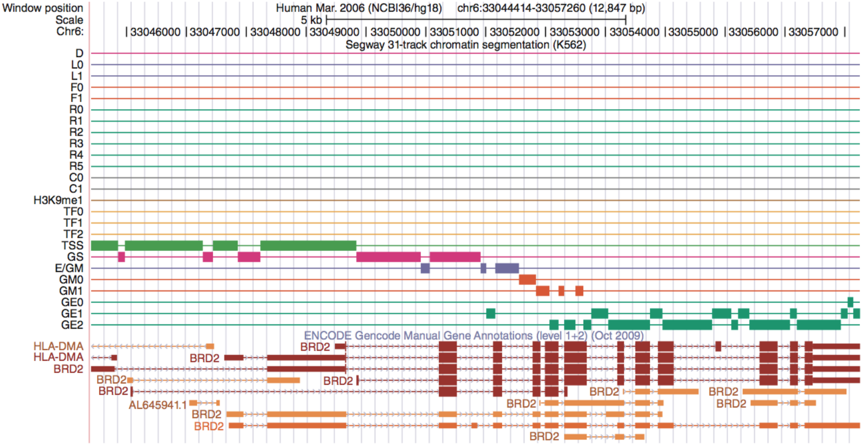
The genome is split into nonoverlapping segments, and ChIP- seq signal for histone modifications is binarized (0 or 1) and collected for each segment, which are further built into input matrix for HMM training. The hidden state of the current segment is dependent on the state of the previous one, and the transition probabilities (in red) of changing from one state to another are learnt from train- ing on the input matrix. ChromHMM outputs trained hidden states for each segmentation, which are then interpreted as chromatin states based on the chromatin profile and gene annotations, such as active promoter/enhancer, transcriptional elongation or repressive states.

Jiang, Shan, and Ali Mortazavi. “Integrating ChIP-Seq with Other Functional Genomics Data.” Briefings in Functional Genomics, March 20, 2018. https://doi.org/10.1093/bfgp/ely002.

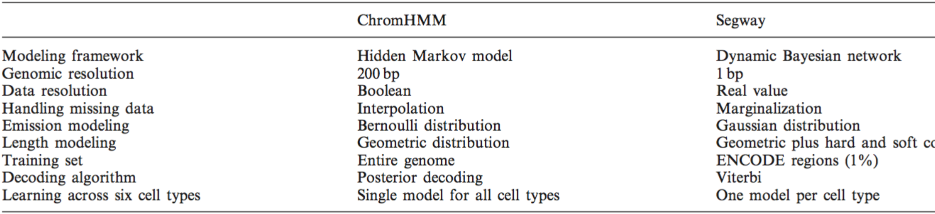
Graphical model representation of the default Segway Dynamic Bayesian Network



Segway segmentation



ChromHMM vs. Segway

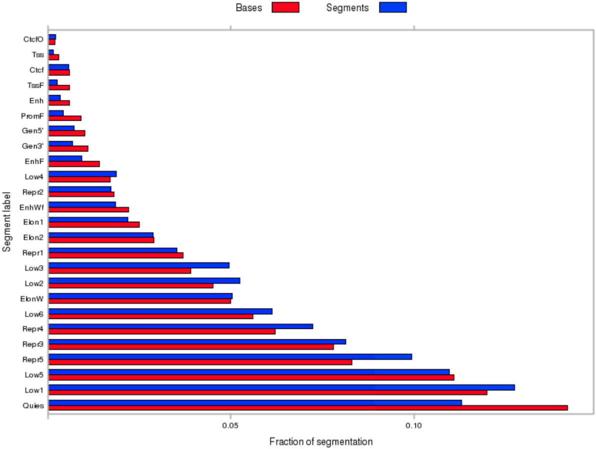


	ChromHMM	Segway
Modeling framework	Hidden Markov model	Dynamic Bayesian network
Genomic resolution	200 bp	1 bp
Data resolution	Boolean	Real value
Handling missing data	Interpolation	Marginalization
Emission modeling	Bernoulli distribution	Gaussian distribution
Length modeling	Geometric distribution	Geometric plus hard and soft co
Training set	Entire genome	ENCODE regions (1%)
Decoding algorithm	Posterior decoding	Viterbi
Learning across six cell types	Single model for all cell types	One model per cell type

Hoffman, Michael M., Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, et al. “Integrative Annotation of Chromatin Elements from ENCODE Data.” Nucleic Acids Research 41, no. 2 (January 2013): 827–41. https://doi.org/10.1093/nar/gks1284.

Notes about chromatin segmentation

A large portion of the human genome exists in a quiescent state, which holds across multiple cell types.



Bases 🟥
Segments 🟦
Segment label
CtcfO
Tss
Ctcf
TssF
Enh
PromF
Gen5'
Gen3'
EnhF
Low4
Repr2
EnhWf
Elon1
Elon2
Repr1
Low3
Low2
ElonW
Low6
Repr4
Repr3
Repr5
Low5
Low1
Quies
0.05
0.10
Fraction of segmentation