

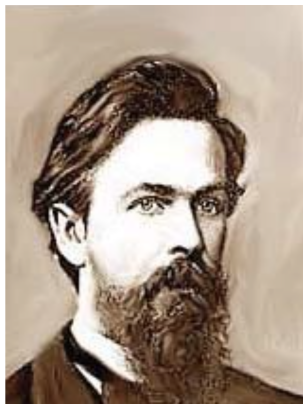
Hidden Markov Models intro, Chromatin segmentation

Mikhail Dozmorov

2021-04-07

Markov Model (aka Markov Chain)

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



Andrey Markov, a Russian mathematician (1856 - 1922)

Markov Model (aka Markov Chain)

- A discrete stochastic process X_1, X_2, X_3, \dots has the Markov property

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

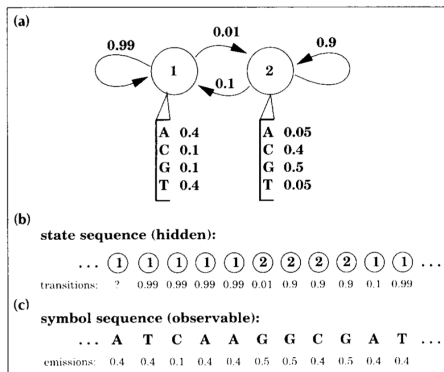
for all x_i , all j , all n

- 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

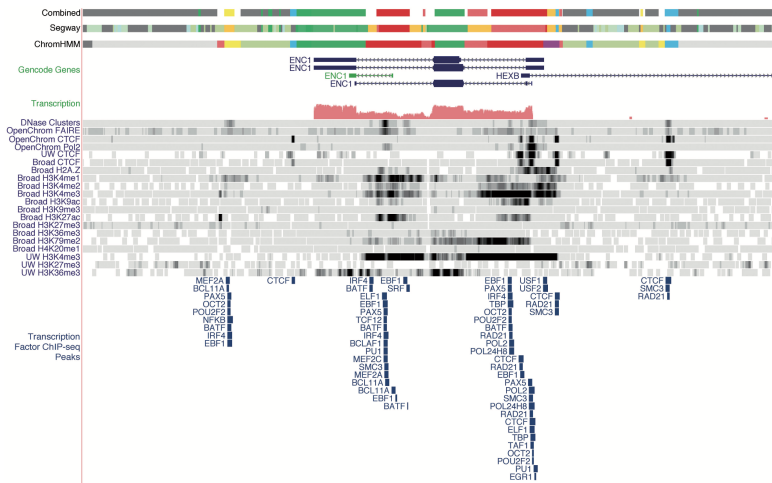
- An alphabet of n emitted symbols (e.g., "A", "T", "C", "G")
- A set of k hidden states (e.g., CG-island, regular sequence)
- Transition = ($transition_{l,k}$) - a $|States| \times |States|$ matrix of **transition probabilities** for changing from state l to state k
- Emission = ($emission_k(symbol_n)$)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

A simple hidden Markov model



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

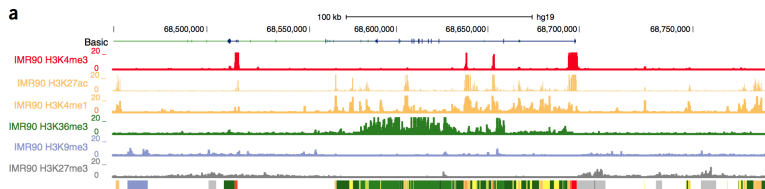


Hoffman, Michael M., Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, et al. "Integrative Annotation of Chromatin Elements from ENCODE Data." *Nucleic Acids Research* 41, no. 2 (January 2013): 827–41. <https://doi.org/10.1093/nar/gks1284>.

Ideas for chromatin track analysis

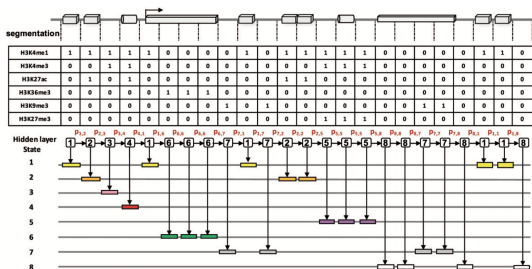
- Hidden Markov Model (ChromHMM)
- Dynamic Bayesian Network (Segway)
 - Bayesian Network that models data sampled at intervals. Still a directed acyclic graph (DAG).
 - Can learn model with Graphical Model Toolkit (GMTK)
 - Can incorporate relationships between variables and handle missing data
 - 1bp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



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 training 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>.

Notes about chromatin segmentation

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

