

---

# Segway: simultaneous segmentation of multiple functional genomics data sets with heterogeneous patterns of missing data

---

**Michael M. Hoffman, Orion J. Buske, Jeff A. Bilmes, William Stafford Noble**

Departments of Genome Sciences, Electrical Engineering, and Computer Science and Engineering  
University of Washington  
Seattle, WA

mmh1@uw.edu, stasis@u.washington.edu

bilmes@ee.washington.edu, william-noble@u.washington.edu

## Abstract

New functional genomics methods enabled by high-throughput DNA sequencing have begun to produce an unprecedented amount of data anchored to the genome of humans and other species. We have developed a method to identify joint patterns in the results of multiple classes of functional genomics experiments. The method partitions the genome into variable-length segments using a dynamic Bayesian network where the dynamic (or “time”) axis represents genomic position. Segments are assigned one of a finite number of labels such that the vectors of observations are similar in segments with the same label. A multinet switching structure allows inference on sequences with combinations of missing data in different tracks that vary at each position, without downsampling or interpolation. This permits us to take full advantage of the high-resolution data generated by sequencing assays, working at up to 1-base-pair resolution. Our system can also incorporate other kinds of data into its classification, including lower-resolution continuous data such as microarray data, or discrete data such as the dinucleotide sequence beginning at each position. We demonstrate the use of the method in both unsupervised and semisupervised training of segment parameters.

## 1 Introduction

Recently, the genomics community has seen an explosion in the availability of large-scale functional genomics data. Researchers have produced genome-wide data sets on the locations of transcription factor binding and histone modifications via chromatin immunoprecipitation (ChIP), open chromatin, and RNA transcription, among several others, using high-throughput microarray (\*-chip) and sequence census assays (\*-seq). For the first time, our representation of the whole human genome expands beyond a sequence of nucleotides occasionally annotated with discrete features to a collection of numerical data tracks defined at almost every part of every chromosome. At the same time, we have moved beyond treating the cellular state determined by functional genomics experiments as constitutive and universal. The ENCODE Project is completing most of its assays in at least six different cell types, and for some assays, we will soon have access to data from dozens of cell types. How can we make sense out of this multitude of data, unprecedented in experimental biology?

A conceptually simple approach to find patterns in these data is through a segmentation procedure [1, 2]. One partitions the genome so that it is completely covered by non-overlapping contiguous segments. Each segment is assigned a label taken from a finite set such that the observed data in regions sharing the same label also share certain properties. Applying Occam’s Razor, it is desirable to make the set of segment labels as small as possible while still accurately modeling the observations.

There are regions of the genome that cannot be interrogated by any particular genome-scale experiment. Unfortunately, these regions vary by assay, and individual analysis methods might result in the reporting of only a subset of those regions. This means that the respective data values for different experiments at a particular position cannot be represented simply as a vector of real numbers, as the missing data varies by position. An example of this situation is shown in Figure 1, which displays the pattern of missing data in 31 experimental data sets produced by five ENCODE collaborations. To circumvent this frustration, researchers have previously used some sort of smoothing or downsampling, but this sacrifices the fidelity and resolution of new experimental methods capable of reporting data at 1-base-pair (bp) resolution in some parts of the genome. Here, we introduce a novel method for segmentation that handles heterogeneous missing data, where each position might have an entirely different pattern of defined data, using a dynamic Bayesian network (DBN) with a switching multinet structure.

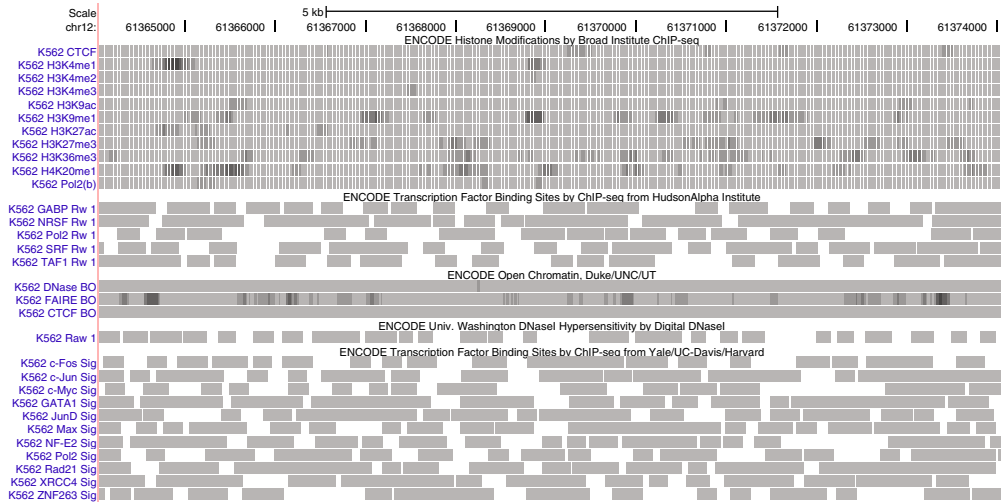


Figure 1: A heterogeneous pattern of missing data in 31 ENCODE data sets collected in the chronic myeloid leukemia cell line K562 along a 10,000-bp region of human chromosome 12, as shown in the UCSC Genome Browser. The first two rows show a scale bar and position along the chromosome. The remaining rows are divided into five supertracks containing data from different collaborations. The supertracks contain one track for each experiment. A white position indicates missing data, and the darkness of a grey position indicates the magnitude of the data value at that position.

DBNs provide a powerful framework for modeling the hidden relationships between observed data defined along an axis of arbitrary length. They are often used in automated speech recognition, where the axis corresponds to time in a sequence of recorded speech, but now are being used to solve biological problems, where the axis can represent either time, such as to model temporal changes in gene expression, or instead the residue position within a macromolecule. We can use multinet techniques to create a probabilistic structure that essentially edits itself at individual positions depending on the variable values at those positions, and we can use this flexibility to model missing data. Wherever a data track is missing at a particular point, we edit the DBN at that frame to remove the dependence of the undefined observation on the segment label, which means that the missing observations are not considered when inferring the hidden segment label.

## 2 Methods

The core of the default Segway DBN is similar to an HMM, with multiple observation tracks and a number of discrete hidden variables. An observation track is a sequence of numerical observations, such as the number of \*-seq tags overlapping successive genomic positions or the intensity of a microarray probe associated with a position. In the default model, shown in Figure 2, the  $i$ th observation track is represented by the sequence of random variables  $X_{0:T}^{(i)} = (X_0^{(i)}, X_1^{(i)}, \dots, X_T^{(i)})$ . Some positions  $t$  may not correspond to a defined value of  $X_t^{(i)}$ . To explicitly model these missing

data, we use an indicator variable  $\hat{X}_t^{(i)}$  to mark whether  $X_t^{(i)}$  is defined ( $\hat{X}_t^{(i)} = 1$ ) or undefined ( $\hat{X}_t^{(i)} = 0$ ). The observation variables at every position depend both on the indicator variable and the hidden segment label, except that when  $\hat{X}_t^{(i)} = 0$ , the the observation and segment label are rendered independent. There are a number of additional structures in the model which allow explicit duration modeling and optional semisupervised training.

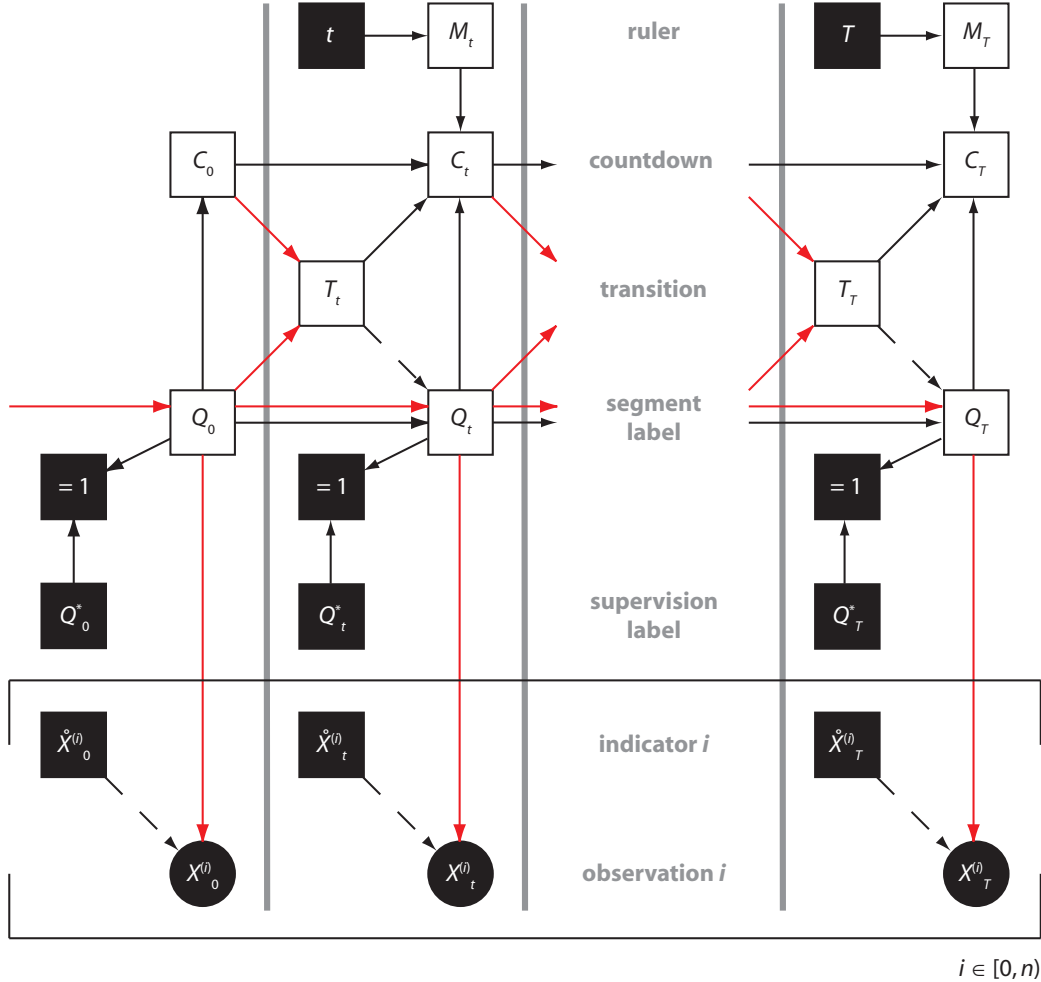


Figure 2: Graphical model representation of the default Segway DBN for  $n$  observation tracks over a sequence of length  $T + 1$ . Nodes represent random variables. The shape represents the domain of the variable (squares: discrete; circles: continuous), and color represents the visibility of the variable (white: hidden; black: observed). Black arcs represent a deterministic conditional dependence, and red arcs represent a stochastic conditional dependence. A dashed arc indicates that, depending on the value of the parent variable along that arc, the conditional parents of the child variable switch deterministically.

### 3 Results

We performed unsupervised expectation maximization (EM) training on 0.15% of the human genome (4,500,065 bp, 32,462,925 observations) using the Segway model and 31 ENCODE signal tracks containing data on histone modifications, open chromatin, and transcription factor binding. We arbitrarily fixed the number of labels to be 25 so that the set of labels would be sufficiently small to be interpretable by biologists. The discovered parameters for the Gaussian probability dis-

tributions  $P(X_t^{(i)} | Q_t = q_t)$  for each value of  $i \in [0, 31)$  and  $q_t \in [0, 25)$  are shown in Figure 3. Many segment labels recapitulate known patterns in the chromatin literature. We used these discovered parameters to perform Viterbi decoding on 93% of the genome (2,847,093,241 bp), and then to demonstrate the correlation of particular segment labels with various known features such as particular parts of protein-coding genes (data not shown).

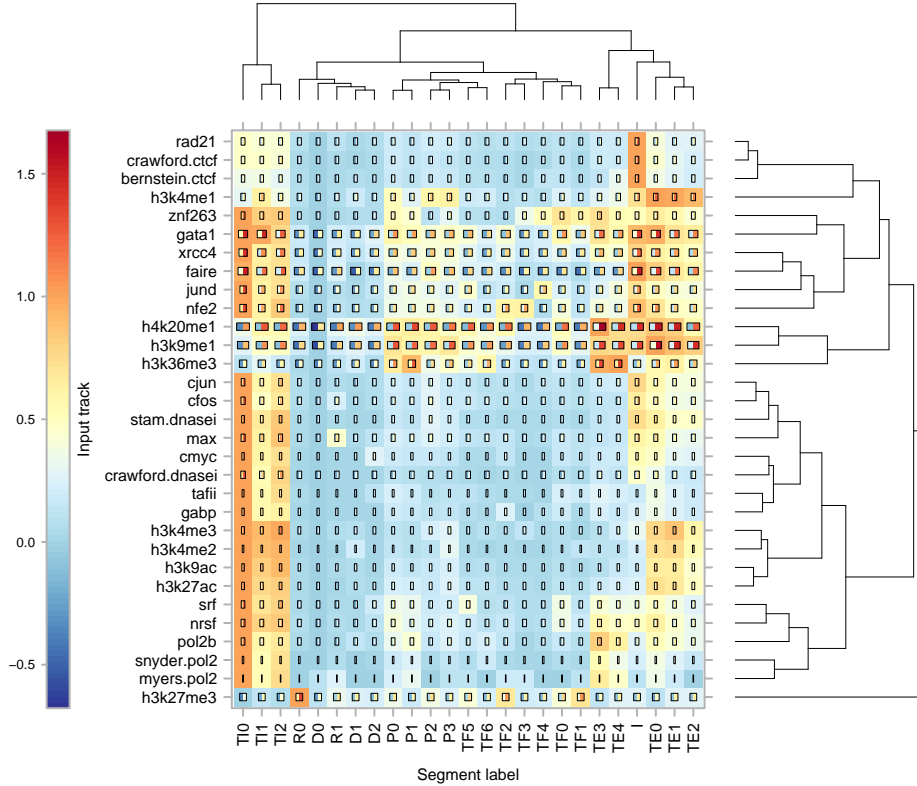


Figure 3: Heat map of discovered Gaussian parameters in an unsupervised 25-label segmentation of 31 ENCODE signal tracks. Each row contains parameters for one signal track, and each column parameters for one segment label. Within each row, we performed an affine transformation such that the largest mean was 1 and the smallest 0. The colors in each cell indicate the mean parameter  $\mu$  and the variance parameter  $\sigma^2$  according to the color bar on the left: the outer color represents  $\mu$ , and the inner colors  $\mu \pm \sigma$ . The width of the inner boxes is proportional to  $\sigma$ , after multiplying by the linear factor used in the transformation of  $\mu$ . We hierarchically clustered both rows and columns, and assigned mnemonics to groups of segment labels based on similarity to patterns in published literature (TI: transcription initiation, R: repression, D: dead zone, P: unknown, TF: unknown, I: insulator, TE: transcription elongation).

We then used the Segway model and the same signal tracks to perform semisupervised EM training with supervision labels drawn from a set of enhancers around the *HBB* locus curated from published literature. With the discovered parameters from the semisupervised learning, a Viterbi segmentation successfully identified known enhancers not in the set of supervision labels (data not shown).

## References

- [1] Day, N., Hemmaphard, A., Thurman, R.E., Stamatoyannopoulos, J.A. & Noble, W.S. (2007) Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**(11):1424–1426.
- [2] Jaschek, R. & Tanay, A. (2009) Spatial clustering of multivariate genomic and epigenomic information. In S. Istrail, P. Pevzner and M. Waterman (eds.), *Research in Computational Molecular Biology*, pp. 170–183. Berlin: Springer.