Integrative Analysis of Epigenomics Data using the bidirectional hidden Markov Model in the R package STAN

Benedikt Zacher[1,2], Rafael Campos-Martin[1,3], Julien Gagneur[2], Achim Tresch[1,2,3,*]
[1]*Department of Biology, University of Cologne;* [2]*Gene Center, Ludwig-Maximilians-University Munich;* [3]*Max-Planck Institute for Plant Breeding Research, Cologne*
[*]send correspondence to: tresch@mpipz.mpg.de

Introduction

Current sequencing-based experimental techniques like RNA-seq and ChIP-Seq generate a wealth of data which can be aligned to the genome of the targeted organism. Yet the integrated analysis of multiple such datasets requires methods for their automated and efficient statistical analysis. One major goal is to annotate the genome, i.e., to cluster genomic positions into functional groups based on the observations made at these positions. One might, e.g., want to dissect the process of RNA trancription into distinct phases characterized by the presence of different protein complexes that change their composition as the RNA Polymerase moves along the DNA. Ideally, such a clustering accounts for the dependency of observations induced by the linear structure of the DNA and the processes associated to it. Hidden Markov models (HMMs) have been used extensively to partition the genome into discrete functional states that can be interpreted as DNA-associated protein complexes. They have been used to infer chromatin states, and annotate enhancers, promoters and transcribed and quiescent regions in human [TDNS07, EK12] and fly [FvBB$^+$10].

Current HMM-based approaches ignore the fact that DNA-related processes may occur in forward or reverse direction. [KGP14] use time-reversible Markov chains to alleviate this drawback. Still, this model is not able to infer the diretionality of DNA-related processes, nor do they properly integrate strand specific (e.g., RNA expression) with non-strand-specific (e.g., ChIP) data. In order to address these points, our present contribution highlights the bidirectional hidden Markov model (bdHMM) introduced in [ZLC$^+$14].

Results

The main idea of the bdHMM is to have so-called twin states, one for each strand and genomic state. Transitions between twin states are coupled by a generalized time-reversibility condition, which replaces the ordinary time-reversibility constraint for reversible HMMs (see Methods for a precise definition). bdHMMs can identify forward and reverse directed states by taking into account directional information contained in each single observation. We derived an efficient analog of the Baum-Welch expectation-maximization (EM) algorithm for bdHMM parameter learning. The bdHMM model along with the EM algorithm is implemented in the open source R/Bioconductor package STAN [ZGT14]. STAN allows the modeling of multivariate Gaussian, Poisson, negative binomial, and multinomial emission distributions and arbitrary independent combinations thereof. It thereby provides a general and flexible framework for obtaining a directed functional state annotation from genomics data.

We applied the bdHMM to a combined RNA transcription and ChIP data set of RNA Polymerase II-associated general transcription factors in yeast. The bdHMM annotated the genome with transcription states (Figure 1), which were characterized by different compositions of the Polymerase II complex. Searching this sequence of states with regular expressions recovers the majority of transcribed loci. We reveal gene-specific variations in the yeast transcription cycle and we find an alternative transcription termination pathway for antisense transcripts. Application of the bdHMM to chromatin modification

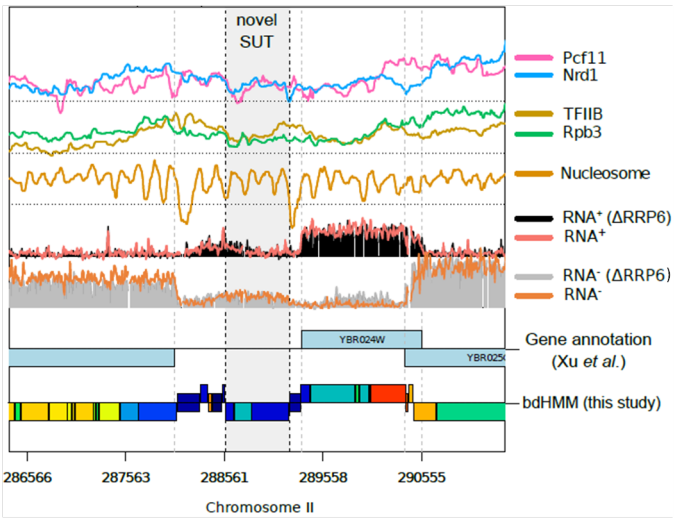

Figure 1: De novo transcript annotation by the bdHMM. Top panels: The data which was used to train the bdHMM include termination factors Pcf11 (pink) and Nrd1 (blue), initiation factor TFIIB (ocre), the RNA Polymerase II subunit Rpb3 (green), Nucleosomes (orange), and strand-specific RNA-Seq expression data in wild type cells (dark and light red) and cells deficient for the nuclear exosome (black and grey). Middle panel: The transcript annotation by [XWG+09], which was not known to the bdHMM, was used as a gold standard. Bottom panel: Viterbi path derived from the bdHMM. Different colors indicate different states. States above (below) the baseline indicate reverse (forward) states, the other states are undirected. The grey area highlights a novel SUT (Stable unannotated transcript, a stable non-coding RNA) region predicted to be expressed on the + strand by the bdHMM yet not captured by former annotations based on the wild-type RNA levels alone. (Modified after [ZLC+14])

data in human T cells provides evidence for existence of directed chromatin state patterns around transcribed regions in the human genome.

Methods

A hidden Markov model is a tuple $\theta = (\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ such that $\mathcal{K}$ is a finite state set, $\pi = (\pi_i)_{i\in\mathcal{K}}$ is the initial state distribution, $A = (a_{ij})_{i,j\in\mathcal{K}}$ is a $\mathcal{K} \times \mathcal{K}$ transition matrix, and $\Psi = \{\psi_i;\ i \in \mathcal{K}\}$ is a set of probability distributions on the observation space $\mathcal{D}$. An HMM defines a probability distribution on a sequence of observations $\mathcal{O} = (o_1, ..., o_T)$. Each observation o_t is *emitted* by a corresponding hidden (unobserved) state variable s_t which can assume values in $\mathcal{K}$. The value of s_t determines the probability of observing o_t by $\Pr(o_t \mid s_t) = \psi_{s_t}(o_t)$. The hidden variables are assumed to form a homogenous Markov chain $\mathcal{S} = (s_1, ..., s_T)$ with time-independent transition probabilities $\Pr(s_t = j \mid s_{t-1} = i) = a_{ij}$, $i, j \in \mathcal{K}$, $t = 2, ..., T$, and with initial state distribution $\Pr(s_1 = i) = \pi_i$, $i \in \mathcal{K}$. The full likelihood of an HMM is

$$\begin{aligned}\Pr(\mathcal{O}, \mathcal{S};\, \theta) &= \Pr(\mathcal{S};\, \theta) \cdot \Pr(\mathcal{O} \mid \mathcal{S};\, \theta) = Pr(s_1;\, \pi) \cdot \prod_{t=2}^{T} Pr(s_t \mid s_{t-1};\, A) \cdot \prod_{t=1}^{T} Pr(o_t \mid s_t;\, \Psi) \\ &= \pi_{s_1} \cdot \prod_{t=2}^{T} a_{s_{t-1}s_t} \cdot \prod_{t=1}^{T} \psi_{s_t}(o_t)\end{aligned}$$

Given a sequence of observations $\mathcal{O}$, the Viterbi algorithm can be used to find the maximum likelihood hidden state sequence $\mathcal{S}$, thus assigning to each position a state in $\mathcal{K}$. This Viterbi path is commonly

used as annotation of the genome (Figure 2a,b). The main idea of the bdHMM is to split the state space $\mathcal{K}$ into undirected states, and pairs of directed (forward and reverse) twin states. Symmetry conditions couple the emission and transition probabilities of twin states in a meaningful way (Figure 2a,c).

Definition. A **bidirectional hidden Markov model** (bdHMM) is a tuple $\theta = ((\mathcal{K}, \kappa), \pi, A, (\mathcal{D}, \delta), \Psi)$ such that $(\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ is an HMM. Additionally, $\kappa : \mathcal{K} \to \mathcal{K}$,$k \mapsto \bar{k}$ and $\delta : \mathcal{D} \to \mathcal{D}$, $o \mapsto \bar{o}$ are involutions ($\kappa^2 = \mathrm{id}$, $\delta^2 = \mathrm{id}$). The involustion κ defines the directed twin states by mapping a state j to its direction-reversed twin state $\bar{j}$, while leaving undirected states fixed. The involution δ maps an observation o to $\bar{o}$ by swapping strand-specific observations. Finally, the following symmetry conditions hold:

1. Generalized detailed balance relation: The transition matrix A and the initial state distribution π satisfy
$$\pi_i a_{ij} = \pi_{\bar{j}} a_{\bar{j}\bar{i}} \quad , \; i, j \in \mathcal{K} \tag{1}$$
2. Initiation symmetry: The initial state distributionπ satisfies
$$\pi_i = \pi_{\bar{i}} \quad , \; i \in \mathcal{K} \tag{2}$$
3. Observation symmetry: Ψ satisfies
$$\psi_i(o) = \psi_{\bar{i}}(\bar{o}) \quad , \; i \in \mathcal{K}, \, o \in \mathcal{D} \tag{3}$$

Why did we specifically choose conditions (1)-(3) as the defining properties of a bdHMM? To motivate our definition, we give an alternative characterization of the bdHMM in terms of a biologically motivated condition. It is natural to require that a directionality-aware HMM marginally cannot distinguish between a forward transition $i, \; j$ from position $t-1$ to t when observing $x, \; y$ at the corresponding positions, and the reverse transition $\bar{j}, \; \bar{i}$ at position $t-1$ to t when observing $\bar{y}, \; \bar{x}$ at the corresponding positions (Figure 2d). In other words, we require that

$$\Pr(s_{t-1} = i, s_t = j, o_{t-1} = x, o_t = y; \, \theta) = \Pr(s_{t-1} = \bar{j}, s_t = \bar{i}, o_{t-1} = \bar{y}, o_t = \bar{x}; \, \theta) \tag{4}$$

holds for all $i, j \in \mathcal{K}$, $x, y \in \mathcal{D}$, $t = 1, 2, \ldots$. Under very mild additional assumptions that are always met in practice, this condition characterizes a bdHMM:

Theorem. Let $\theta = ((\mathcal{K}, \kappa), \pi, A, (\mathcal{D}, \delta), \Psi)$ be a tuple with involutions $\kappa : \mathcal{K} \to \mathcal{K}$, $\delta : \mathcal{D} \to \mathcal{D}$, such that $(\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ is an HMM. Let each $\psi_i \in \Psi$ be non-constant, and and let A be irreducible, i.e., there exists some positive integer r such that $(A^r)_{ij} > 0$ for all $i, j \in \mathcal{K}$. Then, θ is a bdHMM if and only if Condition (4) holds.

Discussion

Bidirectional Hidden Markov Models (bdHMMs), are a novel method for de novo and unbiased inference of directed genomic states from genome-wide profiling data. It allows for the integration of strand-specific data such as RNA expression together with non-strand-specific data such as ChIP occupancy. It can jointly model nominal, continuous and count data by a variety of emission distributions. The open-source package STAN provides a fast, multiprocessing implementation that can process the human chromatin data set in less than one day using a 20 CPU compute cluster. The most significant advance of bdHMM analysis over previous methods is its potential to de novo identify characteristic sequences (patterns) of directed states on the genome. The explicit modeling of forward and reverse states detected an alternative transcription termination pathway which is primarily associated with antisense transcripts. We find that directed patterns of histone modifications are ordered according to the direction of RNA transcription. A bdHMM has essentially the same number of parameters as a comparable standard HMM, and its learning is done at the same speed. Thus, the inference of directionality is without additional costs. We therefore expect the bdHMM to have a broad range of applications in genomics and epigenomics.

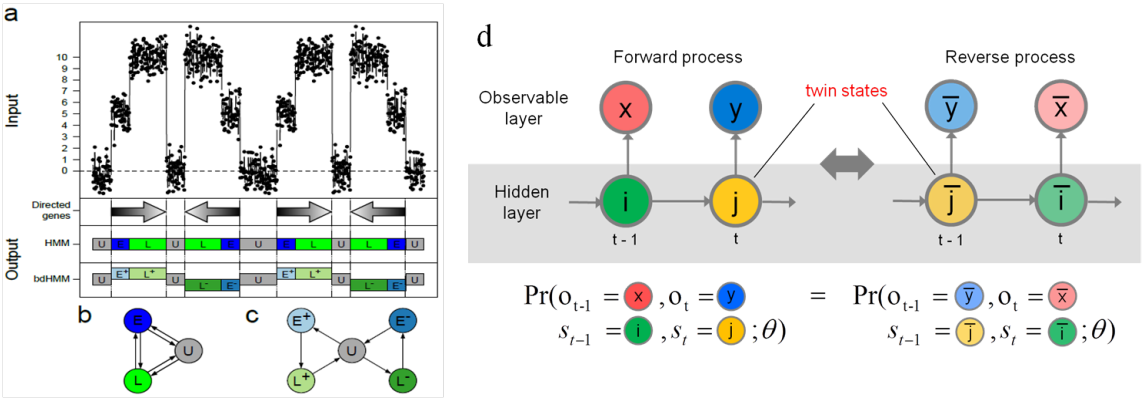

Figure 2: Toy example of a bdHMM model. (a) Simulated occupancy signal (1st track from the top) for a putative factor with a low level (centered at 0) in untranscribed regions (state U), an intermediate level in 5' part of genes (state E), and a high level in 3' part of genes (state L). Arrows (2nd track) depict boundaries and orientation of transcription. Unlike standard HMMs (3rd track) bdHMM (4th track) infer strands (+ or -) to expressed states (E, L). (b) HMM transition graph. Because orientation of transcription is not modeled by standard HMMs, the spurious reverse transitions (E⇒ U, L⇒ E, and U⇒ L) are as likely as the correctly oriented transitions (U⇒ E, E⇒ L, and L⇒ U). (c) bdHMM transition graph. In contrast to HMMs, bdHMMs explicitly model strand-specific expression states (E^+/E^- and L^+/L^-), which results in the correct inference of oriented transitions. (d) Illustration of condition (4), the defining property of a bdHMM. (Modified after [Zacher et al. 2014])

References

[EK12] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9(3):215–216, Mar 2012.

[FvBB+10] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, and B. van Steensel. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, 143(2):212–224, Oct 2010.

[KGP14] David Knowles, Zoubin Ghahramani, and Konstantina Palla. A reversible infinite HMM using normalised random measures. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1998–2006, 2014.

[TDNS07] R. E. Thurman, N. Day, W. S. Noble, and J. A. Stamatoyannopoulos. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.*, 17(6):917–927, Jun 2007.

[XWG+09] Z. Xu, W. Wei, J. Gagneur, F. Perocchi, S. Clauder-Munster, J. Camblong, E. Guffanti, F. Stutz, W. Huber, and L. M. Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037, Feb 2009.

[ZGT14] Benedikt Zacher, Julien Gagneur, and Achim Tresch. STAN: STrand-specific ANnotation of genomic data. R package version 1.2.0. *Bioconductor 3.0*, 2014.

[ZLC+14] Benedikt Zacher, Michael Lidschreiber, Patrick Cramer, Julien Gagneur, and Achim Tresch. Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. *Molecular systems biology*, 10(12):768, 2014.