BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 12 2014, pages 1755–1756
doi:10.1093/bioinformatics/btu070

Genome analysis

Advance Access publication February 19, 2014

MRHMMs: Multivariate Regression Hidden Markov Models and the variantS

Yeonok Lee[1], Debashis Ghosh[1,*], Ross C. Hardison[2] and Yu Zhang[1,*]

[1]Department of Statistics and [2]Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA 16803, USA

Associate Editor: John Hancock

ABSTRACT

Summary: Hidden Markov models (HMMs) are flexible and widely used in scientific studies. Particularly in genomics and genetics, there are multiple distinct regimes in the genome within each of which the relationships among multivariate features are distinct. Examples include differential gene regulation depending on gene functions and experimental conditions, and varying combinatorial patterns of multiple transcription factors. We developed a software package called MRHMMs (Multivariate Regression Hidden Markov Models and the variantS) that accommodates a variety of HMMs that can be flexibly applied to many biological studies and beyond. MRHMMs supplements existing HMM software packages in two aspects. First, MRHMMs provides a diverse set of emission probability structures, including mixture of multivariate normal distributions and (logistic) regression models. Second, MRHMMs is computationally efficient for analyzing large data-sets generated in current genome-wide studies. Especially, the software is written in C for the speed advantage and further amenable to implement alternative models to meet users' own purposes.

Availability and implementation: http://sourceforge.net/projects/mrhmms/

Contact: ghoshd@psu.edu or yuzhang@stat.psu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 30, 2013; revised on January 9, 2014; accepted on January 27, 2014

1 INTRODUCTION

Multivariate features are abundant in genomic and genetic studies. A central theme is to understand the differential relationships among features and their contribution to gene functions and phenotypes. For example, the gene expression variation can be predicted using distinct sets of histone modification levels in gene function groups (Jung and Kim, 2012). The joint occupancy of multiple transcription factors may change depending on the genome context, such that differential regulatory modules are observed and linked to distinct regulation mechanisms. MRHMMs (Multivariate regression hidden Markov models and the variantS) is based on the premise that biological factors do not interact with each other in a uniform manner across the genome. A regression hidden Markov model (rHMM), for example, can be used to segment the genome or genes into groups in each of which there is a unique relationship among biological factors. In addition to rHMM, MRHMMs includes five other hidden Markov model (HMM) variant structures that can be alternatively applied to suit specific studies and data characteristics.

2 MATERIALS AND METHODS

An HMM can be viewed as a mixture model in which the latent (hidden) state has the Markov property. Given the number of states M, an HMM is constructed by specifying the initial state probability, the transition probability and the emission probability distribution. Let $O = (O_1, \ldots, O_T)$ denote the observation of data at time points $t_1, \ldots, t_T$. The time points will correspond to genomic locations in our setting. Let $d \equiv (d_2, \ldots, d_T)$ denote the distance of the two adjacent observations, where $d_k = t_k - t_{k-1}$ for $k = 2, \ldots, T$, and let $q \equiv (q_1, \ldots, q_T)$ denote the hidden states of the HMM. The initial state probability is denoted by $\pi = (\pi_1, \ldots, \pi_M)$, where $\pi_i = P(q_1 = s_i)$ and s_i represents the ith state for $i = 1, \ldots M$. The transition probability from state s_i at time t to state s_j is denoted by $a_{ij}(t) = P(q_{t+1} = s_j | q_t = s_i, d_{t+1})$, where $t \in \{t_1, \ldots, t_T\}$. See the Supplementary material for the details. Let A_t, $t \in \{t_1, \ldots, t_T\}$ denote the transition probability matrix at time t, and let $A = \{A_1, \ldots, A_T\}$.

Let $b_i(O_t) = p(O_t | q_t = s_i, \lambda)$ denote the emission probability distribution given state s_i, where λ denotes a collection of parameters for the emission probability distribution. The joint distribution $p(O, q | \lambda, A, \pi)$ of the HMM is then written as

$$p(O, q | \lambda, A, \pi) = \pi_{q_1} b_{q_1}(O_1) \prod_{t=2}^{T} a_{q_{t-1} q_t}(t) b_{q_t}(O_t). \quad (1)$$

When the emission probability distribution is a mixture of K distributions, the last term in Equation (1) can be replaced by the following:

$$b_{q_t}(O_t) = \sum_{j=1}^{K} u_{q_t j} p(O_t | q_t, h_t = j, \lambda), \quad (2)$$

where h_t is the mixture component at time t, $u_{q_t j} = P(h_t = j | q_t)$ for $q_t = 1, \ldots, M$, $j = 1, \ldots, K$ and $t \in \{t_1, \ldots, t_T\}$.

When the observation O_t consists of explanatory and response variables, denoted by x_t and y_t, respectively, the last term in Equation (2) can be further expressed as

$$p(x_t, y_t | q_t, h_t, \lambda) = p(y_t | x_t, q_t, h_t, \lambda) \cdot p(x_t | q_t, h_t, \lambda), \quad (3)$$

This model is more general than both the regular HMM and the rHMM. The regular HMM has $p(y_t | x_t, q_t, \lambda) = 1$ and the rHMM has $p(x_t | q_t, \lambda) = 1$.

*To whom correspondence should be addressed.

MRHMMs incorporates six different emission probability structures by modifying Equation (3):

(1) $p(x_t|q_t = s_i, \lambda)$

(2) $p(y_t|x_t, q_t = s_i, \lambda) \cdot p(x_t|q_t = s_i, \lambda)$

(3) $p(y_t|x_t, q_t = s_i, \lambda)$

(4) $\sum_j u_{ij} p(x_t|q_t = s_i, h_t = j, \lambda)$

(5) $\sum_j u_{ij} p(y_t|x_t, q_t = s_i, h_t = j, \lambda) \cdot p(x_t|q_t = s_i, h_t = j, \lambda)$

(6) $p(y_t|x_t, q_t = s_i, \lambda) \cdot \sum_j u_{ij} p(x_t|q_t = s_i, h_t = j, \lambda)$

Within a state, the emission probability densities in Models 1–3 consist of a single parametric distribution, while Models 4–6 consist of a mixture of multiple distributions. Model 1 represents the regular HMM, Model 3 corresponds to rHMM and Model 2 incorporates information in both the explanatory variables and their relationships with the response variables. Models 4 and 5 can be considered as an extension of Models 1 and 2 to non-parametric emission densities, i.e. the actual emission densities in unknown distribution families are approximated by a mixture of parametric densities. Model 6 includes a single relationship between explanatory and response variables, whereas the explanatory variables are further modeled non-parametrically using a mixture model.

The unknown parameters in MRHMMs, (π, A, λ), are estimated using the Baum–Welch algorithm (Baum *et al.*, 1970). The hidden states are estimated using the Viterbi algorithm (Viterbi, 1967) that finds the most likely state sequence. The number of states is determined by Bayesian Information Criterion.

The information of the model structure and options that a user would like to build are carried through a text file, which we refer to as an input file. The explanatory, response and location files are space-delimited and need to be prepared separately. MRHMMs automatically generates five output files whose names are indicated by the number of states and the number of mixture components (if Models 4–6 are used). The output files are the initial state and transition probabilities (hmm), the parameter estimates of emission probability distributions (parm), the log likelihood and the (posterior) state probabilities (loglike), the loglikelihood of each independent run (RepLoglike) and the HMM states inferred by the Viterbi algorithm (viterbi). The details of input and output files and available options are provided in the Supplementary material.

3 EXAMPLE

We present an application of MRHMMs to a genome-wide study of the dynamics of GATA1 binding in erythroid cells in mouse. To illustrate, we applied MRHMMs using Model 3 (rHMM) to segment the mouse (mm8) chromosome 7 using the relationships between four explanatory variables: H3K4me3, H3K27me3 (histone modification levels), DNase I hyper sensitivity and Tal1 occupancy, and one response variable: GATA1 occupancy. The data are generated by Wu *et al.* (2011) using the ChIP-seq technology. We used the log-transformed maximums ($\log(x + 1)$) over 1 000-bp non-overlapping windows as the variables. We quantile normalized the explanatory variables and standardized the response variables. MRHMMs found 16 states as determined by Bayesian Information Criterion using 10 independent runs for each number of states $M = 2 \sim 20$. We relabeled the states by an increasing order of the GATA1 averages in each state. The regression coefficients and the boxplots of GATA1 occupancy in each state are shown in Figure 1. The statistically insignificant regression coefficients are marked by 'X'. It is observed that the last two states (states 15 and 16) have notably high GATA1 signals, and the genome segments in these two states are the potential candidates of GATA1 binding sites. The regression coefficients of the two states suggest distinct relationships between GATA1 binding and the four explanatory factors: state 15 demonstrates mild correlation but state 16 indicates strong relationships. Based on the results from McLean *et al.* (2010), we found that state 15 is enriched with *cis*-regulatory regions of genes related with hemoglobin complex, whereas state 16 is enriched with cytosolic part (Binom Raw *P*-values are 8.6×10^{-8} and 4.8×10^{-8}, respectively).

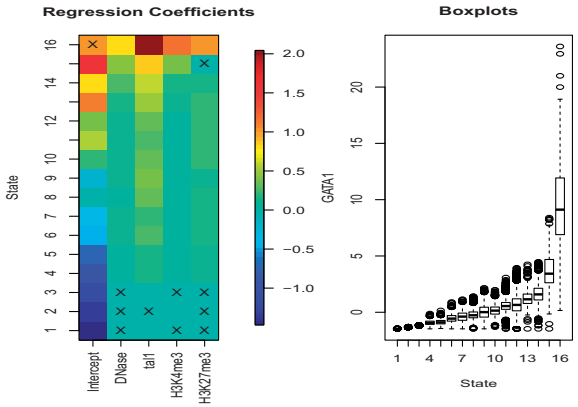

Fig. 1. Regression coefficients and the boxplots for GATA1 occupancy in 16 states identified by MRHMMs. Statistically insignificant regression coefficients are marked by X at $\alpha = .01$ after the Bonferroni correction

Funding: The project described was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through grants UL1TR000127 and NSF ABI-1262538. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflict of Interest: none declared.

REFERENCES

Baum,L.E. *et al.* (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, **41**, 164–171.

Jung,I. and Kim,D. (2012) Histone modification profiles characterize function-specific gene regulation. *J. Theor. Biol.*, **310**, 132–142.

McLean,C.Y. *et al.* (2010) Great improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Viterbi,A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, **13**, 260–269.

Wu,W. *et al.* (2011) Dynamics of the epigenetic landscape during erythroid differentiation after gata1 restoration. *Genome Res.*, **21**, 1659–1671.