FactorialHMM: Fast and exact inference in factorial hidden Markov models

Regev Schweiger[1,2,*], Yaniv Erlich[2,3,4,5] and Shai Carmi[6]

[1] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

[2] MyHeritage, Or Yehuda, Israel

[3] Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA

[4] Center for Computational Biology and Bioinformatics (C2B2), Department of Systems Biology, Columbia University, New York, NY, USA

[5] New York Genome Center, New York, USA

[6] Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel.

*To whom correspondence should be addressed, at regev.schweiger@myheritage.com

Motivation: Hidden Markov models (HMMs) are powerful tools for modeling processes along the genome. In a standard genomic HMM, observations are drawn, at each genomic position, from a distribution whose parameters depend on a hidden state; the hidden states evolve along the genome as a Markov chain. Often, the hidden state is the Cartesian product of multiple processes, each evolving independently along the genome. Inference in these so-called Factorial HMMs has a naïve running time that scales as the square of the number of possible states, which by itself increases exponentially with the number of sub-chains; such a running time scaling is impractical for many applications. While faster algorithms exist, there is no available implementation suitable for developing bioinformatics applications.

Results: We developed FactorialHMM, a Python package for fast exact inference in Factorial HMMs. Our package allows simulating either directly from the model or from the posterior distribution of states given the observations. Additionally, we allow the inference of all key quantities related to HMMs: (1) the (Viterbi) sequence of states with the highest posterior probability; (2) the likelihood of the data; and (3) the posterior probability (given all observations) of the marginal and pairwise state probabilities. The running time and space requirement of all procedures is linearithmic in the number of possible states. Our package is highly modular, providing the user with maximal flexibility for developing downstream applications.

Availability: https://github.com/regevs/factorial_hmm

1 Introduction

Hidden Markov models (HMMs) are instrumental for modeling sequential data across numerous disciplines, such as signal processing, speech recognition, and climate modeling. HMMs are also widely popular in bioinformatics (Durbin *et al.*, 1998; Ernst and Kellis, 2012; Li *et al.*, 2014; Shihab *et al.*), due to the sequential nature of the genome. In a standard HMM, the system is characterized by an internal discrete state variable, which evolves as a Markov chain between time points. In typical applications in bioinformatics, the state describes a slowly varying characteristic of the genome that is not directly observed, such as the identity or population of origin of the ancestor, or the chromatin state. Under the model, observed data points (which may be continuous or discrete) are drawn, at each genomic position, from a distribution whose parameters depend on the hidden state at that position. Algorithms exist for the efficient inference, given a sequence of observations, of the most likely sequence of hidden states or the marginal probability distribution of the state at each position, among other quantities.

An important generalization of the HMM is the factorial HMM (Ghahramani and Jordan, 1997), in which there are multiple independent Markov chains of latent variables, and the distribution of the observed variable at a given time step is conditional on the states of all of the latent variables at that same time step. Factorial HMMs have been particularly useful in genetics, where they have been used in classic linkage algorithms (Lander and Green, 1987), admixture mapping (McKeigue *et al.*, 2013), local and global ancestry inference (Bercovici *et al.*, 2012; Baran *et al.*, 2012; Pei *et al.*, 2018), estimating identity-by-descent (IBD) (Bercovici *et al.*, 2010), identifying relationships (Kyriazopoulou-Panagiotopoulou *et al.*, 2011), and detecting recombination events (Husmeier, 2005).

In a general implementation of an HMM, the time and space complexity is quadratic in the number of states. For a factorial HMM, the number of states is exponential in the number of latent Markov chains. However, the independence of the hidden chains in the factorial HMM can lead to reduced complexity of several standard operations. In (Ghahramani and Jordan, 1997), an exact calculation is presented to perform the Forward-Backward algorithm in $O(TMK^{M+1})$, instead of $O(TK^{2M})$, where T is the number of time steps (or genomic positions), K is the alphabet size of each latent chain, and M is the number of latent chains.

We extended the Forward-Backward efficient calculations presented in (Ghahramani and Jordan, 1997) to perform fast exact calculation of all standard HMM operations, effectively avoiding both time and space quadratic complexity. We present our implementation of these methods in FactorialHMM, a Python program supporting the full set of standard HMM operations. FactorialHMM is expected to be particularly useful in the context of genetics, providing a flexible and generic framework for researchers.

2 Features

We highlight the main distinct features of FactorialHMM. First, the latent Markov chains may be inhomogeneous, i.e., a distinct transition matrix per chain and per step. This is an important feature in genetics applications, where the (physical or genetic) distance between consecutive genomic sites may vary. Second, the alphabet size of each latent chain may be distinct, which allows for flexibility of modeling. The observed states may be continuous or discrete, and univariate or multivariate. Finally, the library supports exact and efficient calculations of all standard HMM operations, as follows. (i) Simulation from the model. (ii) Calculation of the marginal posterior probabilities of the latent hidden states, using the Forward-Backward algorithm. (iii) Calculation of the maximum posterior sequence of hidden states given the observed sequence, using the Viterbi algorithm. (iv) Calculation of the posterior pairwise probability for the hidden states, per chain, conditional on the observed data, used for constructing an efficient M-step in the Expectation-Maximization (EM) algorithm. (v) Sampling from the posterior distribution of hidden states, again given the observed sequence. The full description of these methods and the details on their efficient implementation are given in the Supplementary information.

3 Performance

We are not aware of an available stable implementation of the factorial HMM to which to compare, although several repositories contain some support for it (Johnson and Willsky, 2013; Ghahramani and Jordan, 1997). To demonstrate the computational efficiency of FactorialHMM, we compared it to a naïve implementation of the HMM using *hmms*, whose core algorithms are implemented in Cython. We tested a factorial HMM with M chains with binary states, with a symmetric transition matrix, uniform initial state distributions, and 100 observations simulated from the HMM with random initial states. We measured the computation time for the Viterbi and the Forward algorithms for increasing values of M. As expected, the computation time grows much faster for the naïve HMM than for FactorialHMM, allowing the user to scale up the number of hidden states.

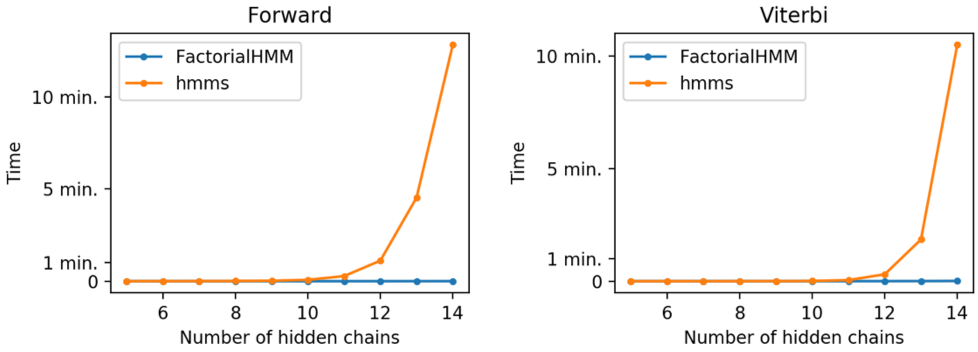

***Figure 1**. A benchmark comparing FactorialHMM and a naïve implementation by hmms. We tested a factorial HMM with M chains with binary states and 100 observations simulated from the HMM with random initial states. We compared the performance of both implementations on the Forward and the Viterbi algorithms. The computation time grows much faster for the naïve HMM than for FactorialHMM. All computation times of FactorialHMM were <1sec.*

4 Discussion

Future directions for development include: (i) implementing the EM for specific forms of the transition and emission matrices; (ii) low-level language implementation; and (iii) an implementation of the linear-time approximate inference procedures detailed by (Ghahramani and Jordan, 1997). We look forward to comments, suggestions and future collaborative development of FactorialHMM.

Factorial HMM is available at https://github.com/regevs/factorial_hmm.

Conflict of Interest: S.C. is a paid consultant to MyHeritage.

References

Baran,Y. *et al.* (2012) Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, **28**, 1359–1367.

Bercovici,S. *et al.* (2012) Ancestry Inference in Complex Admixtures via Variable-Length Markov Chain Linkage Models. In, *Research in Computational Molecular Biology*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 12–28.

Bercovici,S. *et al.* (2010) Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics*, **26**, i175–i182.

Durbin,R. *et al.* (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids Cambridge university press.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, **9**, 215–216.

Ghahramani,Z. and Jordan,M.I. (1997) Factorial Hidden Markov Models. *Machine Learning*, **29**, 245–273.

Husmeier,D. (2005) Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics*, **21**, ii166–ii172.

Johnson,M.J. and Willsky,A.S. (2013) Bayesian Nonparametric Hidden Semi-Markov Models. *Journal of Machine Learning Research*, **14**, 673–701.

Kyriazopoulou-Panagiotopoulou,S. *et al.* (2011) Reconstruction of genealogical relationships with applications to Phase III of HapMap. *Bioinformatics*, **27**, i333–i341.

Lander,E.S. and Green,P. (1987) Construction of multilocus genetic linkage maps in humans. *PNAS*, **84**, 2363–2367.

Li,Y. *et al.* (2014) Expansion of Biological Pathways Based on Evolutionary Inference. *Cell*, **158**, 213–225.

McKeigue,P.M. *et al.* (2013) Extending Admixture Mapping to Nuclear Pedigrees: Application to Sarcoidosis. *Genetic Epidemiology*, **37**, 256–266.

Pei,J. *et al.* (2018) Inferring the ancestry of parents and grandparents from genetic data. *bioRxiv*, 308494.

Shihab,H.A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, **34**, 57–65.