BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 11 2014, pages 1625–1626
doi:10.1093/bioinformatics/btu057

*Sequence analysis*

Advance Access publication January 30, 2014

# StochHMM: a flexible hidden Markov model tool and C++ library

Paul C. Lott and Ian Korf*

Genome Center, One Shields Ave., University of California, Davis, CA 95616, USA

Associate Editor: John Hancock

**ABSTRACT**

**Summary:** Hidden Markov models (HMMs) are probabilistic models that are well-suited to solve many different classification problems in computation biology. StochHMM provides a command-line program and C++ library that can implement a traditional HMM from a simple text file. StochHMM provides researchers the flexibility to create higher-order emissions, integrate additional data sources and/or user-defined functions into multiple points within the HMM framework. Additional features include user-defined alphabets, ability to handle ambiguous characters in an emission-dependent manner, user-defined weighting of state paths and ability to tie transition probabilities to sequence.

**Availability and implementation:** StochHMM is implemented in C++ and is available under the MIT License. Software, source code, documentation and examples can be found at http://github.com/KorfLab/StochHMM.

**Contact:** ifkorf@ucdavis.edu

Received on July 10, 2013; revised on December 9, 2013; accepted on January 24, 2014

## 1 INTRODUCTION

Since the first application of hidden Markov models (HMMs) to biological sequences in the 1980s, they have become a fundamental tool in bioinformatics. This is because of their robust statistical foundation, conceptual simplicity and malleability that allow researchers to adapt them to fit diverse classification problems (Brejová and Brown, 2008). Many HMM-based programs have adapted the basic HMM framework to solve unique biological problems, such as development of generalized HMM (GHMM) to simplify state durations (Kulp *et al*., 1996), stochastic sampling to predict alternative splicing (Cawley and Pachter, 2003; Stanke *et al*., 2006a) and integration of additional data sources into the decoding algorithms to improve gene predictions (Stanke *et al*., 2006b).

Currently, there are a small number of different applications, libraries and compilers available to quickly develop HMMs, such as MAMOT (Schütz and Delorenzi, 2008), HMMoc (Lunter, 2007), HMMlib (Sand *et al*., 2010), GHMM (Kulp *et al*., 1996), HMMConverter (Lam and Meyer, 2009), R HMM and Matlab.

Each tool differs in the level of expertise required to use them and the features available to the users (see Table 1).

StochHMM is a program and C++ library that provides the ability to implement HMMs quickly from a simple text file. It provides accessibility to a broader audience by providing many of the features available in the higher-level libraries, without requiring the same level of programming skill. In addition to providing traditional HMM algorithms, StochHMM implements a few stochastic decoding algorithms and provides multiple ways to integrate additional data sources into the model either as multiple independent or joint emissions. The C++ library provides the flexibility to integrate additional user-defined functions into the HMM framework, thereby allowing us to integrate existing bioinformatics tools into the HMM.

*To whom correspondence should be addressed.

## 2 HMM ARCHITECTURE

StochHMM implements the traditional HMM architecture (discrete or continuous emissions, and constant transitions). It also supports multiple joint or independent discrete emissions and continuous emissions per state, lexical transitions (transitions that are based on the observation sequence) and user-defined emission/transition functions (see Table 1). Like other HMM application/libraries, StochHMM supports user-defined alphabets. However, StochHMM also allows users to define ambiguous alphabets and provides multiple ways to score ambiguity in an emission-dependent manner. Finally, StochHMM allows researchers to violate the strict probabilistic framework of an HMM, thereby giving them the power to produce conditional random field models (Brejová and Brown, 2008). These additional features provide the researchers the freedom to adapt the HMM architecture to generate more powerful and accurate models.

## 3 PERFORMANCE

StochHMM is comparable with HMMoc in speed and memory usage. In 10 side-by-side comparisons with HMMoc using the common occasionally dishonest casino HMM and a 30 Mb sequence, StochHMM performed Viterbi decoding analysis in an average of 18.15 s (1.36 GB RAM) compared with average of 10.83 s (2.30GB RAM) for HMMoc. Calculating the posterior probability for the same sequence, StochHMM calculated the posterior in an average of 52.8 s (1.40 GB RAM) compared with 148.1 s (0.98 GB RAM) for HMMoc. An additional factor for researchers to consider is the amount of development time that is required to get a model to run. Starting with model in hand, StochHMM provided the ability to run, test and tweak the model without any additional code development, whereas HMMoc requires additional code development to run or change the model and requires recompilation of the code after each change to model.

## 4 CASE STUDIES

We have used StochHMM to develop models to solve a diverse set of problems. StochHMM allowed us to quickly develop a 4-state model called SkewR, which predicts R-loop formation

**Table 1.** Comparison of features available in different HMM packages, including StochHMM

| Tools | Supported HMM types | Features | | | | | | | | | | | | | | | | Portability | Dependencies |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Definable alphabet | Ambiguous alphabet | Silent states | Discrete emission | High - order emissions | Continuous emission | Multiple emissions | Joint emissions | Lexical transitions | Link to user - defined functions | Viterbi decoding | Posterior decoding | N - best decoding | Stochastic decoding | Weighted path | Explicit path | | |
| Compilers | | | | | | | | | | | | | | | | | | | |
| HMMoc | Traditional, pair, generalized | ✓ | | ✓ | ✓ | ✓ | | ✓ | | ✓ | | ✓ | ✓ | | [a] | ✓ | | Linux, Windows, OS X | None |
| HMMConverter | Traditional, pair, generalized | ✓ | | ✓ | ✓ | | | ✓ | | | | ✓ | | | | ✓ | | Linux | None |
| Libraries | | | | | | | | | | | | | | | | | | | |
| GHMM | Traditional, pair, generalized | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ | | | | | | Linux, OS X | Libxml2, python |
| HMMlib | Traditional | ✓ | | | ✓ | | | | | | | ✓ | ✓ | | | | | Linux, Windows, OS X | Cmake, Boost |
| StochHMM | Traditional | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Linux, Windows, OS X | None |
| Toolboxes | | | | | | | | | | | | | | | | | | | |
| R-HMM | Traditional | ✓ | | | ✓ | | ✓ | | | | | ✓ | | | | | | Linux, Windows, OS X | n/a |
| Matlab | Traditional | ✓ | | | ✓ | | | | | | | ✓ | ✓ | | | | | Linux, Windows, OS X | n/a |
| Implementaion applications | | | | | | | | | | | | | | | | | | | |
| MAMOT | Traditional | ✓ | | | ✓ | | | | | | | ✓ | ✓ | | | | | Linux, Windows, OS X | None |
| StochHMM | Traditional | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Linux, Windows, OS X | None |

[a]Stochastic decoding available with further code implementation. "✓" indicates support for feature within the application. (Kulp *et al.*, 1996; Lam and Meyer, 2009; Lunter, 2007; Sand *et al.*, 2010; Schütz and Delorenzi, 2008).

in the human genome (Ginno *et al.*, 2012, 2013). In collaboration with Diane Schroeder, we created a simple 2-state model, which uses a unique alphabet based on percent methylation of CpG sites, to classify high and partial methylation domains in multiple Methyl-Seq datasets (Schroeder *et al.*, 2011, 2013). The StochHMM library was used to implement Ploidamatic, a 12-state HMM to identify copy-number variation using next-generation whole genome sequence from a diploid organism (Porter *et al.*, submitted for publication). We are currently using StochHMM to create gene prediction models that integrate existing bioinformatics tools and additional data sources into the HMM framework to improve gene prediction.

## 5 CONCLUSIONS

StochHMM is a flexible hidden Markov model program and C++ library that gives researchers the ability to implement traditional HMMs from a simple text file. The application provides similar performance and features previously available only in libraries, such as HMMoc. It provides multiple ways to integrate additional dataset or user-defined functions into the HMM framework. To encourage the use of StochHMM, we have set up forums on Google Groups to assist researchers in development of HMMs using StochHMM. We look forward to comments, suggestions and future collaborative development of StochHMM.

*Funding:* National Institutes of Health (grant number R01-HG004348-01); and UC Davis Genome Center.

*Conflict of Interest*: none declared.

## REFERENCES

Brejová,B. and Brown,D. (2008) Advances in hidden Markov models for sequence annotation. In: Mandouiu,I. and Zelikovski,A. (eds) *Bioinformatics Algorithms: Techniques and Applications*. pp. 55–92.

Cawley,S.L. and Pachter,L. (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19** (**Suppl. 2**), ii36–ii41.

Ginno,P.A. *et al.* (2013) GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res.*, **23**, 1590–1600.

Ginno,P.A. *et al.* (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell*, **45**, 814–825.

Kulp,D. *et al.* (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **4**, 134–142.

Lam,T.Y. and Meyer,I.M. (2009) HMMCONVERTER 1.0: a toolbox for hidden Markov models. *Nucleic Acids Res.*, **37**, e139.

Lunter,G. (2007) HMMoC–a compiler for hidden Markov models. *Bioinformatics*, **23**, 2485–2487.

Sand,A. *et al.* (2010) HMMlib: a C++ library for general hidden Markov models exploiting modern CPUs. In: *CORD Conference Proceedings*. pp. 126–134.

Schroeder,D.I. *et al.* (2011) Large-scale methylation domains mark a functional subset of neuronally expressed genes. *Genome Res.*, **21**, 1583–1591.

Schroeder,D.I. *et al.* (2013) The human placenta methylome. *Proc. Natl Acad. Sci. USA*, **110**, 6037–6042.

Schütz,F. and Delorenzi,M. (2008) MAMOT: hidden Markov modeling tool. *Bioinformatics*, **24**, 1399–1400.

Stanke,M. *et al.* (2006a) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.*, **34**, W435–W439.

Stanke,M. *et al.* (2006b) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.