Sequence analysis

StochHMM: a flexible hidden Markov model tool and C++ Library

Paul C Lott and Ian Korf

1Genome Center, One Shields Ave., University of California, Davis, CA 95616, USA

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: 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

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 due to their robust statistical foundation, conceptual simplicity, and malleability that allows researchers to adapt them to fit diverse classification problems (Brejova 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 (Stanke, Keller, et al., 2006; Cawley and Pachter, 2003), and integration of additional data sources into the decoding algorithms to improve gene predictions (Stanke, Schoffmann, et al., 2006).

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), HMMCConverter (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.

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 upon 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 (Brejova 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 to HMMoc in speed and memory usage. In ten side-by-side comparisons with HMMoc using the common Occasionally Dishonest Casino HMM and a 30MB sequence, StochHMM performed Viterbi decoding analysis in an average of 18.15 seconds (1.36GB RAM) compared to average of 10.83 seconds (2.30GB RAM) for HMMoc. Calculating the posterior probability for the same sequence, StochHMM calculated the posterior in an average of 52.8 seconds (1.40GB RAM) compared to 148.1 seconds (0.98GB 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 in order to run or change the model and requires recompilation of the code after each change to model.
4 CASE STUDIES

We’ve 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 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 classify copy number variation using next-generation whole genome sequence from a diploid organism (Porter et al., Manuscript 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 use of StochHMM, we have setup 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.

ACKNOWLEDGEMENTS

Funding: This work was funded by National Institutes of Health [grant number R01-HG004348-01]; and UCDavis Genome Center.

REFERENCES


