Sequence analysis

MAMOT: hidden Markov modeling tool

Frédéric Schütz and Mauro Delorenzi*

Swiss Institute of Bioinformatics and Bioinformatics Core Facility of the NCCR Molecular Oncology, CH–1015 Lausanne, Switzerland

Received on February 22, 2008; revised and accepted on April 21, 2008

ABSTRACT

Summary: Hidden Markov models (HMMs) are probabilistic models that are well adapted to many tasks in bioinformatics, for example, for predicting the occurrence of specific motifs in biological sequences. MAMOT is a command-line program for Unix-like operating systems, including MacOS X, that we developed to allow scientists to apply HMMs more easily in their research.

One can define the architecture and initial parameters of the model in a text file and then use MAMOT for parameter optimization on example data, decoding (like predicting motif occurrence in sequences) and the production of stochastic sequences generated according to the probabilistic model. Two examples for which models are provided are coiled-coil domains in protein sequences and protein binding sites in DNA. A wealth of useful features include the use of pseudocounts, state tying and fixing of selected parameters in learning, and the inclusion of prior probabilities in decoding.

Availability: MAMOT is implemented in C++, and is distributed under the GNU General Public Licence (GPL). The software, documentation, and example model files can be found at http://bcf.isb-sib.ch/mamot

Contact: Mauro.Delorenzi@isb-sib.ch

1 INTRODUCTION

Hidden Markov models (HMMs) were first introduced in bioinformatics at the end of the eighties, for the analysis of biological sequence data (Durbin et al., 1998, Koski, 2001), and are widely used for sequence alignment and motif finding, as well as many other applications not related to sequence analysis, including linkage analysis and the interpretation of aCGH gene copy number data. HMMs are particularly effective when patterns consist of submotifs that vary in order and length, like introns, exons and intergenic regions in DNA. While gene finding requires rather sophisticated models, there are many problems for which fairly simple first-order HMMs are adequate and might outperform other methods. For example, in a recent benchmarking (Gruber et al., 2006), the HMM MARCOIL (Delorenzi and Speed, 2002) compared favorably to various methods for the prediction of coiled-coil domains in protein sequences. MARCOIL has been created for a comparison between HMMs and position-specific scoring matrices, a widely used procedure based on fixed-length windows. The HMM approach has shown evident advantages for short domains, for which an optimal choice of starting and ending points are more critical. While very popular among bioinformaticians, HMMs are under-used at large. Some existing programs are well suited for specific applications (e.g. HMMER or SAM for profile HMMs), with limitations as to which models can be implemented. Other programs might be complicated to use or are not free software. With the release of MAMOT, a free, general-purpose tool for working with HMMs, users can define models or any topology and our intention is to facilitate application of HMMs in teaching and research by a growing number of scientists. In our experience, to invent, train and test an HMM tailored to a specific research question is frequently possible in a relatively short time.

In the following sections, we first describe the main functionalities of the program and then present an application that we used to introduce students to the practical use of HMMs.

2 FUNCTIONALITIES

The models used by MAMOT are described in files written in simple text format using straightforward conventions. It is easy to write, understand and modify them, or to create them with a separate program. The description starts with the indication of the different symbols that can be emitted by the states (alphabet), followed by a list of the states. For each state, transition and emission probabilities are specified. An optional feature of a model is a list of ‘tied’ states that during training have the same commonly estimated transition and/or emission probabilities. Another feature is ‘reverse-tying’ for emission probabilities in DNA models, where the frequency of emission of the symbols at one of the states equals the frequency of emission of the complementary symbol (A for T, C for G) at another state. This is useful for models that concomitantly scan both strands of DNA for motif occurrence. It is also possible to fix emission or transition probabilities that should not change during training. Typically, this is useful for states set to a constant symbol, even when pseudocounts are used, or for states that represent an assumed background model, or for ‘forbidden’ transitions.

All the classical algorithms associated with HMMs are implemented in MAMOT. Given an (initial) model and some example data, optimized parameters can be obtained by the Baum–Welch’s maximum likelihood EM algorithm or by the Viterbi training algorithm. The data likelihood, that is, the probability of an observed sequence given the model, can be obtained by the Forward algorithm. The user can specify the desired degree of regularization by the addition of pseudocounts, which is

*To whom correspondence should be addressed.
generally recommended and prevents null probabilities. The two common decoding methods to assign hidden states to each position of a sequence of observed symbols are available: the Viterbi algorithm finds the most probable state path, while the Forward–Backward algorithm computes the state probabilities from which the most likely state corresponding to each symbol (‘posterior decoding’) can be deduced. Finally, random sequences of symbols can be generated according to a chosen model.

There are also some more specific features that will appeal to some users. A possibility inspired by the coiled-coil recognition problem (Gruber et al., 2006) is to define state weights used in the decoding process. Baum–Welch estimates maximum likelihood parameters, which are not necessarily optimal for discrimination, so it can be interesting to explore the effect of different weights for some states of an HMM. For example, high weights to hydrophobic states reduce the wrong prediction of certain hydrophilic fragments as potential coiled-coil domains. The weights are implemented as multiplicative factors of the log probabilities and we advise to optimize them by cross-validated performance testing. A model file for running the MARCOIL model with weights is available on our website.

One can also provide external probabilities for a list of selected state-position combinations to be used in the decoding of an observed sequence. These probabilities are used as multiplicative factors to the emission probabilities. By setting selected values to zero, one can obtain the posterior state probabilities, or the Viterbi path, under the constraint that a particular set of position-state combinations is excluded. Conversely, one can increment the probability of a path that has a given position-state combination. These additional factors can be derived from experimental or in silico evidence, for example primer elongation sequences or sequence homology data proving or supporting the existence of an exon in a particular DNA position in gene finding.

MAMOT commands typically produce a result on the screen, often accompanied by an additional text file with more information. The posterior state distributions are delivered in a matrix format that can easily be imported into a statistical package such as R (http://www.r-project.org/) for additional analyses or graphical representation (see example below). An example R script is included in the distribution.

3 EXAMPLE: PROTEIN BINDING SITES

Here, we shortly expose how one can proceed to generate and use a model by reporting a training and prediction example for the binding site of the p53 protein. The site has a structure similar to those described by Sandelin and Wasserman (2005) for the DNA response element of nuclear hormone receptor proteins. The binding site model was set up with one emitting state for each nucleotide of two half-sites of 10 consecutive binding positions, one state for a possible spacer and one state for non-binding DNA. The half-sites were connected directly and through the spacer. The spacer state was connected with itself to allow for spacers of different lengths (but parameters will favor short ones). Another series of states represented binding sites with reversed polarity and corresponding positions were ‘reverse-tied’ in training. A model file for transcription factor binding sites that occur in multiple configurations is available from the webpage.

Initial parameter values were extracted from tables summarizing 37 well documented sites (Hoh et al., 2002). A refined HMM was obtained by training on chromatin immunoprecipitation fragments (Wei et al., 2006). We obtained a model with sensitivity and specificity almost identical to those reported by the authors, which were obtained with a similar approach. The trained model was applied to large genomic sequences around some transcription start sites to search for presumptive p53 target genes. Figure 1 shows as example the profile of posterior state probabilities obtained by MAMOT and graphed in R. Shown is a region of 50 kb upstream of the transcription start site of the CDKN1A (p21) gene, a well known target gene. Two immunoprecipitation fragments reported in this region (removed from the examples used to refine the model) correspond well to the location of p53 sites predicted by the model.

4 CONCLUSIONS

MAMOT is a free, easy to use, flexible tool for developing HMMs, useful for research and for teaching. While MAMOT provides some features that are specifically targeted at predicting coiled-coil domains or transcription factor binding sites, it is a generic tool for HMM modeling that is ideally suited for retraining existing models and for creating new ones.

MAMOT is still being developed and the authors are keen on hearing about users’ experiences and suggestions.

ACKNOWLEDGEMENTS

Funding: This work was funded by National Centre for Competence in Research (Molecular Oncology), Swiss National Science Foundation.

Conflict of Interest: none declared.

REFERENCES