BIOINFORMATICS APPLICATIONS NOTE
Vol. 20 no. 5 2004, pages 788–789
DOI: 10.1093/bioinformatics/btg489

Algorithms for variable length Markov chain modeling

Gill Bejerano

Center for Biomolecular Science and Engineering, School of Engineering, University of California, Santa Cruz, CA 95064, USA

Received on August 4, 2003; revised on September 30, 2003; accepted on October 15, 2003
Advance Access publication January 29, 2004

ABSTRACT

Summary: We present a general purpose implementation of variable length Markov models. Contrary to fixed order Markov models, these models are not restricted to a predefined uniform depth. Rather, by examining the training data, a model is constructed that fits higher order Markov dependencies where such contexts exist, while using lower order Markov dependencies elsewhere. As both theoretical and experimental results show, these models are capable of capturing rich signals from a modest amount of training data, without the use of hidden states.

Availability: The source code is freely available at http://www.soe.ucsc.edu/~jill/src/

Contact: jill@soe.ucsc.edu

INTRODUCTION

Clustering sequences of discrete, or quantized values into groups of related sequences is of great importance in bioinformatics. Relevant data ranges from macromolecular composition, through gene expression time series measurements, to observed transcription factor binding site or protein domain combinations. One way of quantifying the notion of relatedness is to obtain a training set of examples from a group of interest, and fit these with a generative probabilistic model that captures statistical correlations shared by the sequences in the set. Then, whenever the trained model M is presented with a novel query sequence $s = s_1, \ldots, s_l$, it assigns to it a score, the probability that M would emit s out of all possible sequences of the same length. Subsequently, we can set a threshold above which a novel sequence is considered to be related to the training set, we can compare the predictions of two or more such models for multi-classification purposes, or we can search for high-scoring sequence segments that parse a given query sequence into one or more known elements.

Using the chain rule we cast this computation into a series of estimates of the next symbol in the query sequence, given its past

$$P(s) = P(s_1) \prod_{i=2}^{l} P(s_i \mid s_1, \ldots, s_{i-1}).$$

An order-d Markov chain models the future of a partial sequence from its immediate past, approximating

$$P(s_i \mid s_1, \ldots, s_{i-1}) \approx P_M(s_i \mid s_{i-d}, \ldots, s_{i-1}).$$

Such models are often used in bioinformatics to capture relatively simple sequence patterns, such as genomic CpG islands, or serve as background distributions for more complex signals (Durbin *et al.*, 1998, Ch. 3). The complex signals are often modeled using hidden Markov models (HMMs), which introduce additional hidden (unobservable) states that replace the context altogether. A major reason for not using Markov chains to model these signals lies with the fact that the memory and training set size requirements of an order-d Markov chain grow exponentially with d. As a result, while low order Markov chains are poor classifiers, higher order chains are often impractical to implement or train.

However, this approach overlooks an intermediate class of variable length Markov models (VMM), which offer the ability to capture statistical correlations of different length scales in a single probabilistic model. Rather than estimating all contexts of length $d, C = \sum^d$, the VMM models a selected set of contexts of different lengths, $C \subset \sum^*$. The chosen context set C is determined by the training data, and includes longer contexts where these appear in the data and shorter contexts elsewhere. Prediction using such a model matches the longest memorized context at every point in the query sequence,

$$P(s_i \mid s_1, \ldots, s_{i-1}) \approx P_M(s_i \mid \max_{d_i \geq 0} s_{i-d_i}, \ldots, s_{i-1} \in C). \quad (1)$$

This context selection scheme avoids the exponential explosion of higher order Markov chains altogether. Furthermore, theoretical evidence shows that it is much less demanding than HMMs in terms of data abundance and quality (Ron *et al.*, 1996), deeming the VMM attractive even in cases where hidden states may appeal as encoding some underlying biological process, such as the ‘ancestral’ sequence in profile HMMs.

Due to space limitations we cannot elaborate on context selection and refer the reader to a detailed introduction in a bioinformatic context given by Bejerano and Yona (2001).

Alphabet = {0,1}   Training sequence = 0100100100111101011000010111

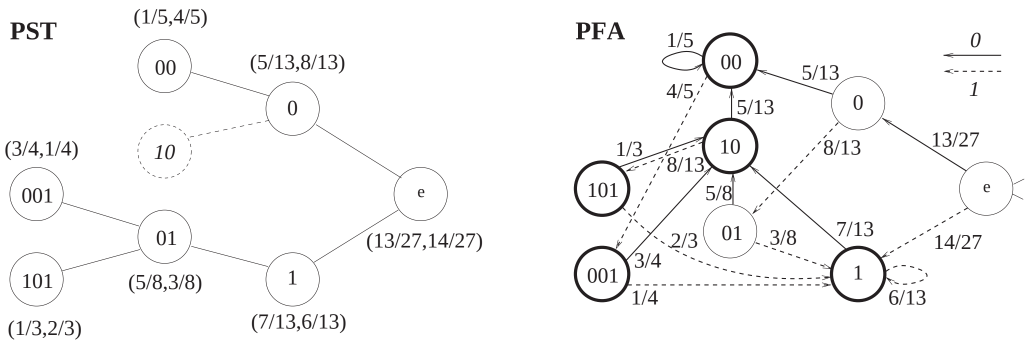

Fig. 1. Models found in directory `example/`. (left) A PST over the binary alphabet generated from the training sequence above. Landscape mode makes Equation 1 predictions easier to follow. Node labels represent VMM contexts. The empty context is marked e. Each node is associated with a vector that holds its predictions for $\{0, 1\}$ respectively. For example, the probability to observe 1 after a subsequence whose longest learned context is 00, is 4/5. The dashed node is added to the PST only for generating the equivalent PFA (right). As such it inherits its parent prediction vector. Both models assign the same probability to any given sequence, e.g. $P_M(1010) = 14/27 \cdot 7/13 \cdot 8/13 \cdot 1/3$. Bold PFA nodes form its ergodic part whose stationary distribution we also compute.

DESCRIPTION

We report here on a general purpose software package that implements VMMs over a user defined alphabet, utilizing a data structure termed probabilistic suffix tree (PST) to hold a set of chosen contexts (or suffixes) together compactly with their prediction vectors. Four main modules are implemented:

train: Train a VMM from a given training set. The output is a PST model. Both algorithms from (Bejerano and Yona, 2001) are implemented, as well as incremental tree growing.

predict: Generates a likelihood score, symbol by symbol, for a given query sequence, and a PST model, using Equation 1. Prediction depth at every step, d_i, is also reported, as it is also a good indicator of query sequence similarity to model.

emit: Stochastic generation of a sequence of symbols from a given PST model, useful for synthetic data generation.

2pfa: Converts a given PST into an equivalent probabilistic finite automaton (PFA). The conversion corrects (Ron *et al.*, 1996), whereas the PFA nodes are the union of all PST nodes (not just leaves, as is stated there) together with all PST leaf prefixes and their parent nodes. Figure 1 shows an example of a PST model and its equivalent PFA. The PFA stationary distribution is also derived, allowing one to compute the relative abundance of different short subsequences in a typical sequence, e.g. for novel binding site detection.

A detailed discussion of these models, as well as applications, extensions and related approaches, such as context tree weighting and prediction by partial matching are found in (Bejerano, 2003).

ACKNOWLEDGEMENTS

Nir Friedman, Ron Begleiter and the referees have helped clarify the manuscript. G.B. has performed this work while at the Hebrew University, where he was supported by a grant from the ministry of science, Israel.

REFERENCES

Bejerano,G. (2003) Automata learning and stochastic modeling for biosequence analysis. Ph.D. thesis, Hebrew University.

Bejerano,G. and Yona,G. (2001) Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*, **17**, 23–43.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Ron,D., Singer,Y. and Tishby,N. (1996) The power of amnesia: learning probabilistic automata with variable memory length. *Mach. Learn.*, **25**, 117–149.