BIOINFORMATICS APPLICATIONS NOTE
Vol. 23 no. 11 2007, pages 1424–1426
doi:10.1093/bioinformatics/btm096

Genome analysis

Unsupervised segmentation of continuous genomic data

Nathan Day[1,†], Andrew Hemmaplardh[1,†], Robert E. Thurman[2,3,*], John A. Stamatoyannopoulos[3] and William S. Noble

[1]Department of Computer Science and Engineering, [2]Division of Medical Genetics, and [3]Department of Genome Sciences, University of Washington, Seattle, WA, USA

Received on November 22, 2006; revised on February 9, 2007; accepted on March 7, 2007

Advance Access publication March 23, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: The advent of high-density, high-volume genomic data has created the need for tools to summarize large datasets at multiple scales. HMMSeg is a command-line utility for the scale-specific segmentation of continuous genomic data using hidden Markov models (HMMs). Scale specificity is achieved by an optional wavelet-based smoothing operation. HMMSeg is capable of handling multiple datasets simultaneously, rendering it ideal for integrative analysis of expression, phylogenetic and functional genomic data.

Availability: http://noble.gs.washington.edu/proj/hmmseg

Contact: rthurman@u.washington.edu

1 INTRODUCTION

The convergence of the genomic era and the advent of high-throughput biological and chemical assays has created a wealth of genomic data, much of which is presented in continuous, time-series-like fashion across the genome. Often, it is desirable to extract simplifying summary information from such data. One summarization approach involves segmenting the data into a small number of discrete states based on the continuous output values. This segmentation may be accomplished in an unsupervised fashion using hidden Markov models (HMMs). For example, a chromosome-wide continuous profile of bulk RNA output generated using tiling DNA microarrays may be partitioned by segmenting the chromosomal coordinates into three states, corresponding to regions of low, medium and high transcription levels. This type of categorization is often desirable in the context of elucidating broad, large-scale trends in the data. In this case, it may be preferable to smooth the data to a specified scale before segmentation, in order to eliminate spurious state transitions resulting from isolated fine-scale features (see Fig. 1).

HMMSeg is a tool for segmenting continuous genomic datasets on a scale-specific basis using HMMs. Scale specificity is achieved by an optional smoothing step using wavelets (see below). HMMSeg provides multivariate capability, computing a single segmentation based on multiple datasets simultaneously defined on a common set of genomic coordinates.

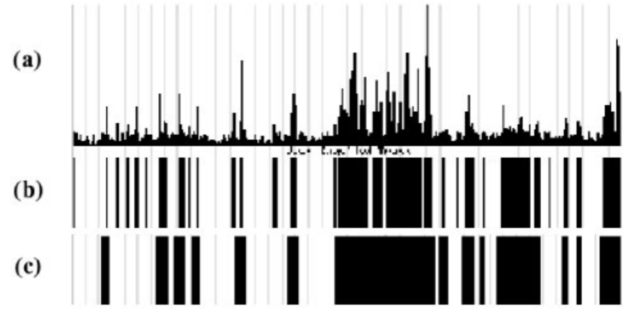

Fig. 1. Effect of wavelet smoothing on HMM segmentations. (**a**) Histone modification H3K4me1 (average raw data resolution ~900 bp). (**b**) Two-state segmentation of the data in (a) with black and white bars representing the two different states. (**c**) Two-state segmentation of data in (a) following 20 kb wavelet smoothing. (Data excerpted from Thurman *et al*., 2006.)

As a platform for segmenting a wide variety of genomic data, HMMSeg is distinguished from existing programs using HMMs, which typically fall under two categories: toolboxes for applications in any field, such as htk (Young *et al*., 1995) and GHMM (http://ghmm.org); or biological application-specific tools that use HMMs, such as glimmerHMM (Majoros *et al*., 2004), for gene finding and HMMer (Eddy, 1995) for sequence analysis.

1.1 Hidden Markov models

An HMM is a statistical model in which data are assumed to be generated by a stochastic process defined by a predetermined number of hidden states (Rabiner, 1989). Each state is defined by an *emission distribution*, from which data values are generated. The model also specifies probabilities for transitioning between states. The parameters defining the emission and transition probabilities are typically learned from the data by expectation maximization (EM). Given the learned parameters, there are two common methods for determining the state labels for each observation: the Viterbi algorithm, which finds the single most probable path (sequence of states); and posterior

[*]To whom correspondence should be addressed.
[†]The authors wish it to be know that, in their opinion, the first two authors should be regarded as First Authors.

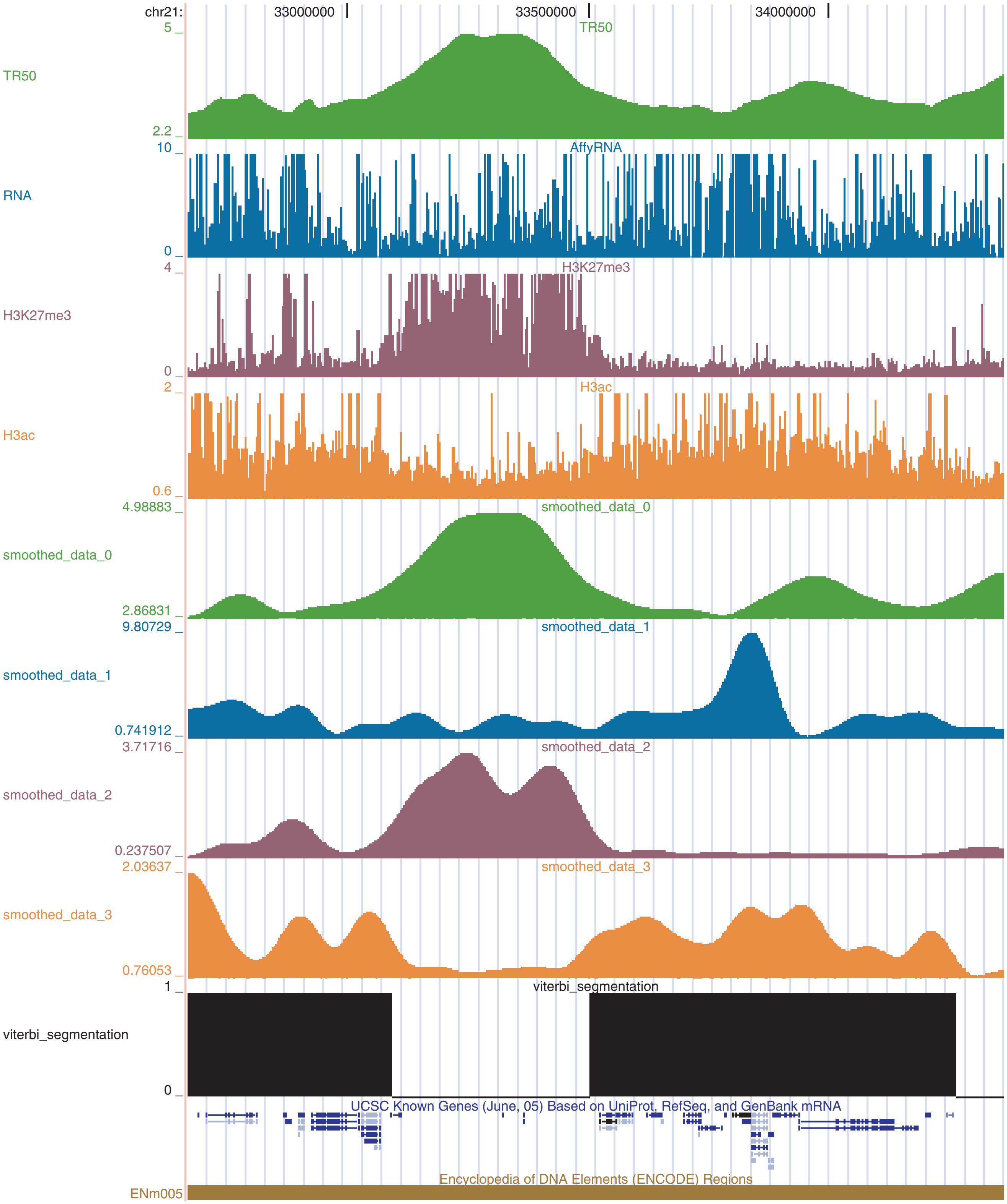

Fig. 2. Multi-datatype functional domains defined using HMMSeg. Shown in this 1.7 Mb region on chr 21 (ENCODE region ENm005) are raw data (top) and 64 kb smoothed data (bottom) for DNA replication timing (green), RNA transcription (blue) and histone modifications H3K27me3 (purple) and H3ac (orange). Row nine shows a two-state Viterbi segmentation based on all four datasets with active domains in black and inactive in white. Note the concentration of genes (bottom row) within active domains. (Data displayed in UCSC Genome Browser with colors added later.)

decoding, which computes the most likely state at each point of the sequence. HMMSeg uses Gaussian emission distributions, with diagonal covariance for multiple datasets (assuming independence between variables), and supports both the Viterbi and posterior decoding methods for state assignments.

1.2 Wavelet smoothing

Wavelets are a mathematical tool for multi-scale analysis (Percival and Walden, 2000). Though first used in practical applications in the fields of engineering and signal processing, in recent years wavelets have found many applications in computational biology (Liò 2003). We apply scale-specific smoothing using a variant of the discrete wavelet transform (DWT) called the maximal overlap discrete wavelet transform (MODWT) (Percival and Walden, 2000). Both the DWT and MODWT can be used to decompose a given signal via *multiresolution analysis* into a sum of scale-specific signals. In contrast with other smoothing techniques, wavelet smoothing is essentially the process of subtracting out the small-scale behavior rather than averaging it. HMMSeg uses the LA(8) family of wavelets for all wavelet transforms. The choice of wavelet scale is application dependent, and can be informed by prior biological information about the scale of features of interest, or by trial-and-error to achieve, say, a desired segment length distribution. See the website for further details on wavelet smoothing.

2 DESCRIPTION OF FUNCTIONALITY

HMMSeg provides a command-line interface. The input to HMMSeg is one or more collections of files in either single column or tab-delimited BED format. Each collection represents a different dataset; multiple collections trigger a multivariate segmentation. After reading the data from the input files, HMMSeg optionally smooths the data at a user-specified scale using the MODWT. In this case, wavelets require that the input data be evenly spaced. There is also an option to smooth the data without HMM training.

HMMSeg proceeds to train a completely connected HMM on the data by using EM. By default, the HMM has two states; models with more states may also be specified. The Gaussian parameters and transition probabilities are initialized randomly, although the user may provide model parameters to replace or initialize EM training. Training may be repeated multiple times from different random starts, in which case the model with the highest total likelihood is selected. Based on the final model, observations are assigned to states using the Viterbi algorithm or posterior decoding.

The program outputs the trained model plus the state assignment for each observation. If the user provides input data in BED format, then the segmentation is output in *wiggle* format, suitable for display in the UCSC Genome Browser (Kent *et al*., 2000). The wiggle file contains separate tracks for the original data, smoothed data, the state assignments and (for the posterior decoding method) the probabilities of each data point belonging to each state.

HMMSeg is implemented in Java for platform independence. It has been successfully tested on Windows and UNIX-style systems. Validation and accuracy test results are available on the website.

3 EXAMPLE

In a recent study (Thurman *et al*., 2006), we analyzed a number of independently generated experimental datasets produced under the NHGRI ENCODE project (ENCODE Consortium, 2004), whose ultimate goal is to identify all of the functional elements in the human genome. Currently the ENCODE project is in its pilot phase, analyzing 44 regions spanning 30MB ($\sim$1%) of the genome (ENCODE Consortium, 2006). Our aim was to integrate multiple functional datatypes to create a functional domain map of the ENCODE regions. We used HMMSeg to segment the data at the 64 kb scale into two states, interpreted *a posteriori* as functionally ‘active’ or ‘inactive’. (Note, however, that in the study wavelet smoothing was performed on all data except for replication timing, as a pre-processing step.) We successfully applied this technique to individual datatypes and up to five datasets simultaneously. Examples of large-scale domains delineated by HMMSeg using MODWT smoothing on all datasets are pictured in Figure 2 (see website for details). Here we see the advantages of using HMMs over simple thresholding techniques, the logic of which breaks down in scenarios with multiple datasets and few states. This approach highlights the potential for integrating multiple functional genomic datatypes with widely varying experimental resolution.

ACKNOWLEDGEMENTS

We thank Charles Grant and Tobias Mann for helpful advice. Funding for this work and Open Access publication charges was provided by NIH grants U01 HG003161, R01 GM071923 and R01 GM71852.

Conflict of Interest: none declared.

REFERENCES

Eddy,S.R. (1995) Multiple alignment using hidden Markov models. In Rawlings, C. (ed.) *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, pp. 114–120.

ENCODE Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.

ENCODE Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, in press.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.

Liò,P. (2003) Wavelets in bioinformatics and compuataional biology: state of art and perspectives. *Bioinformatics*, **19**, 2–9.

Majoros,W.H. *et al*. (2004) Tigrscan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

Percival,D.B. and Walden,A.T. (2000) *Wavelet Methods for Time Series Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.

Thurman,R.E. *et al*. (2007) Stamatoyannopoulos. Identification of higher-order functional domains in the human genome. *Genome Research*, in press.

Young,S. *et al*. (1995) *The HTK Book*. Cambridge University.