CFHMM: Heterogeneous Tumor CNV Classification by Hidden Markov

Aleksandar Obradovic[1,*], Hongjian Qi[2]
[1]Department of Computer Science, Columbia University azo2104@columbia.edu
[2]Department of Computer Science, Columbia University hq2130@columbia.edu

Abstract

Summary: We here develop and implement a Clonal Fraction Hidden Markov Model (CFHMM), to leverage positional information in classifying Tumor CNVs and their corresponding clonal fraction from log-ratio-normalized Tumor/Normal sequencing data. In simulated data, this approach shows accurate calling of CNVs for high-fraction mutations, and improvement in calling over a naïve clustering benchmark across the board, as well as useful purity estimation for dominant clones.

Availability and Implementation: Source code and documentation is freely available at https://github.com/7lagrange/FCNV implemented in R, with all major operating systems supported.

Contact: azo2104@columbia.edu

Supplementary Information:
Additional tables and figures available at
https://docs.google.com/document/d/1ohbjWaZ20jXX3Tc64BASuZPW-mpnE_ybfwfMjU9jb0mU/edit?usp=sharing

Introduction

Copy Number Variations (CNVs) are duplications or deletions of genome segments, of length greater than one kilobase by convention, which occur normally in the genome but have also been highly implicated in tumor genomes (Zhang). It is therefore important to accurately classify CNVs in tumor genome data. This can be done by modeling next-generation sequencing data, where a natural model to apply is the Hidden Markov Model, with transition matrix of copy number or CNV states and emission of normalized read-counts (Zhao). However, admixture of normal cells in a tumor sample (Gusnanto) and heterogeneity of distinct clones within a tumor (Oesper) complicate analysis, such that apparent copy-number for any given genome region may appear fractional and be misclassified. This, in addition to the loss of prior location-specific variability information that can occur after tumor/normal normalization, is an issue that must be addressed for improved application of a Hidden Markov approach to CNV classification.

In our CFHMM model, we implement a 15-state model, with a commonly used 5-copy-number (0,1,2,3,4) set for tumor and extremes-removed 3-copy-number (1,2,3) set for normal cells, where a normal diploid state of 2 is the most common. We derive an average purity prior from raw log-ratios, from which a set of 15 strong emission distribution priors are constructed that preserve variability information for a given tumor/normal state-pair. CFHMM runs modified unsupervised Viterbi training on the data to give posterior state classifications that, if accurate, may also provide clonal fraction information for each CNV mutation.

Methods

We evaluate the accuracy of our model on simulated whole-genome-sequencing data for which the hidden tumor and normal copy-number states are known. The tumor and normal genomes are segmented into kilobase-length bins for which copy number is generated by Markov Chain with adjustable parameters favoring remaining at the same copy number from one bin to the next and providing equiprobable transition to any other copy number. 60 million reads, corresponding to 10X deep sequencing in (My-

ers) are randomly distributed into these bins with probability weighted by copy number. Clonal fraction k in the tissue for a given bin is randomly selected from a parameter list of options, where we ran several tests with two-tumor-clone data and one with three-tumor-clone data. Log-Normalization proceeds from read-count data as:

$$log.ratio = \log\left(\frac{kT + (1-k)N}{N}\right) \quad (1)$$

Where, conversely:

$$k = \frac{\exp(log.ratio) - 1}{\frac{T}{N} - 1} \quad (2)$$

The purity prior k for input into CFHMM is taken as the mean of values from equation 2, assuming normal state 2 and using only extreme log-ratios over the equation 1 values for pure (k=1) tumor state 3, or under the threshold for pure tumor state 1. With this prior, since the read-counts are Poisson distributed, the log-normalized emission data can be normally approximated in a way that preserves variability information dependent on location along the genome as:

$$\log\left(\frac{kT + (1-k)N}{N}\right) \sim N\left(\log\left(\frac{kp1}{p2} + 1 - k\right), \sqrt{\frac{1-p1}{n*p1} + \frac{1-p2}{n*p2}}\right) \quad (3)$$

Where p1 and p2 are tumor and normal state, respectively, and n is the number of reads. From here, a Hidden Markov Model is created with a 15-state Transition matrix and modified for continuous-emission such that the Emission matrix contains a mean and variance for each of the 15-states. The Viterbi algorithm is then:

$$(4)\quad X_m(n) = max_i\left(X_i(n-1) * T_{im} * dnorm(y(n), E_{mean[m]}, E_{s.d[m]})\right)$$

With probability of state m at position n captured in Xm(n), state transition probability captured in T, and log-ratio emission distributions captured with mean and standard deviation in E. This is applied for classification in unsupervised training with a modified version of the standard iterative Viterbi algorithm (Durbin), where the Maximization step in EM for the continuous-emission-matrix E uses Bayesian update (Lynch) on the normal distributions:

$$posterior\ mean = \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{n}{s^2}}\mu + \frac{\frac{n}{s^2}}{\frac{1}{\sigma^2} + \frac{n}{s^2}}\bar{x} \quad posterior\ var = \frac{2}{\frac{1}{\sigma} + \frac{1}{s}} \quad (5)$$

After 15-state classification, a posterior purity estimate for each bin where tumor state is not equal to normal state can be derived from equation 2, and this distribution can be plotted to visualize clonal fraction, as well as k-means-clustered to group mutations belonging to the same clone sub-population. This is compared in simulated data to the known fractions and the accuracy is evaluated. For CNV-state classification, the 15-state model is compressed by grouping of equivalent states into a 6-state model (see Table 1).

Table 1. State compression table.

state	1	2	3	4	5	6
T/N	0/1, 0/2, 0/3	1/2, 1/3, 2/3	1/1, 2/2, 3/3, 4/4	3/2, 4/3	4/2, 2/1	3/1, 4/1

15-state tumor-normal model is reduced by posterior resolution of ambiguities to 6 distinct CNV states. 1-full deletion, 2-partial deletion, 3-normal, 4-partial amplification, 5-homozygous amplification, 6-large amplification.

As a benchmark for accuracy of CNV state classification, a naïve thresholding method is used, where thresholds are drawn at the log-ratio values from equation 1 with purity prior k and T/N states 0/2, 2/3, 4/3, 4/2, and 3/1, and a genome location is classified into one of the six compressed states

as whatever interval of these thresholds its log-ratio falls inside of. There is no equivalent benchmark for posterior purity from the HMM model, as this is the primary innovation of our method, so it is evaluated on its own.

Results

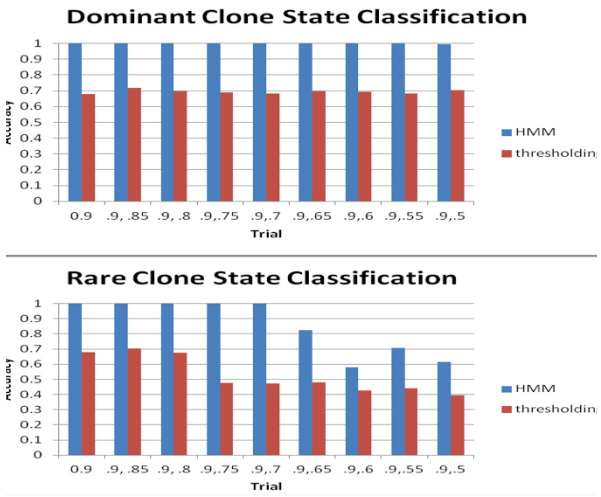

Fig 1. CFHMM v Threshold accuracy for dominant and rare clones.

Simulated data was run for ten matched tumor-normal genomes, with all parameters but purity kept constant at default value. The first trial was homogeneous tumor admixed with normal cells at 0.9 clonal fraction. Each of the next eight trials contained two equiprobable purity values, with a dominant 0.9 fraction clone and a secondary subset of fraction ranging from 0.85 to 0.5 at intervals of 0.05. A final heterogeneous tumor trial was run with three equiprobable mutation fractions of 0.9, 0.7, and 0.5. In terms of state classification accuracy, each trial showed correct classification for clones represented in 0.7 or higher proportion of the tumor sample with accuracy over 99.9%. This is significantly higher than the approximately 70% accuracy seen in the naïve thresholding benchmark. Both methods, however, lose accuracy for rarer CNV mutations, although our CFHMM method remains more accurate than the benchmark across the board (see Fig. 1).

Accuracy of posterior purity estimates (+/ 0.01 of true value) was also better for high-fraction CNVs, showing around 80% accuracy in all trials for the dominant clone and accuracy decreasing to 56% for 0.5 purity. Posterior purity estimates show clear peaks at the true purities for k greater than or equal to 0.7, usefully depicting tumor heterogeneity, but retaining only the dominant-clone peaks for tumors with rarer admixed mutations, such as the 3-state trial tumor, where states 0.9 and 0.7 were well-described and classified with high accuracy, but 0.5 was not (see Table 2).

Table 2. Three-State 0.9, 0.7, 0.5 Purity Trial

Purity states	HMM state accuracy	Thresholding state accuracy	HMM purity accuracy
.9	0.9995894	0.6706574	0.8174977
.7	0.9992959	0.5384039	0.7150379
.5	0.649813	0.3955961	0.4992132

Accuracy is shown per purity state for the HMM classification, threshold-ing classification, and HMM-derived purity.

Areas for future work include the implementation of Baum-Welch rather than Viterbi training for HMM learning (Durbin), in order to ensure globally optimal solution, as well as the application of the algorithm to real matched cancer sequencing datasets for further validation. In addition, the breakdown of the method for low-fraction CNVs is a limitation than should be addressed, possibly by iterative exclusion of classified dominant mutation sites, so that the new prior on subsequent runs of the algorithm is pulled toward lower purity, so long as the limitation for the Markov assumption presented by resulting holes in the data can be addressed.

Acknowledgements

We wish to thank Dr. Itsik Pe'er and the students of CBMFW4761 for providing feed-

back and advisement on project design.

References

Durbin, Richard, ed. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press, 1998.

Gusnanto, Arief, et al. “Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data.” Bioinformatics 28.1 (2012): 40-47.

Lynch, Scott M. Introduction to applied Bayesian statistics and estimation for social scientists. Springer Science & Business Media, 2007.

Myers, Gene. “Whole-genome DNA sequencing.” Computing in Science and Engineering 1.3 (1999): 33-43

Oesper, Layla, Gryte Satas, and Benjamin J. Raphael. “Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data.”Bioinformatics 30.24 (2014): 3532-3540.

Zhang, Nancy R. “DNA Copy Number Profiling in Normal and Tumor Genomes.” Frontiers in Computational and Systems Biology. Springer London, 2010. 259-281.

Zhao, Min, et al. “Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives.”BMC bioinformatics 14.Suppl 11 (2013): S1.