I Need A Title

Contents

1 DNA

A DNA molecule consists of a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, five carbon sugar and a phosphate group. There are four different bases. The pyrimidines are thymine (T) and cytosine (C). The purines are adenine (A) and guanine (G). The DNA occurs as a double helix composed of two polynucleotide strands with bases facing inwards. The two single strands run antiparallel and are connected by hydrogen bonding between complementary bases. Because the two strands are complementary, we can represent a DNA molecule by a sequence of bases in a single strand. The top of the strand is known as the $5'$ and the bottom as $3'$

2 Copy Number Variation CNV

Copy number variants (CNVs) are alternations of DNA of a genome that results in the cell having a less or more than two copies of segments of the DNA [4]. CNVs correspond to relatively large regions of the genome, ranging from about one kilobase to several megabses that are deleted or duplicated [4].

CNVs can be discovered by cytogenetic techniques, array comparative genomic hybridization and by single nucleotide polymorphism (SNP) arrays see [4] and references therein. Furtheremore, CNVs can be identified by next-generation sequencing (NGS) in high resolution [4]. NGS can generate millions of short sequence reads along the whole human genome. When these short reads are mapped to the reference genome, both distances of paired-end data and read-depth (RD) data can reveal the possible structure variations of the target genome [4] Another important source of information useful for inferring CNVs from reads alignment is the read depth (RD). The RD data are generated to count the number of reads that cover a genomic location or a small bin along the genome which provide important information about the CNVs that a given individual carries [4]. When the genomic location or bin is within a deletion, one expects to observe a smaller number

of read counts or lower mapping density than the background read depth. In contrast, when the genomic location or bin is within an insertion or duplication, one expects to observe a larger number of read counts or higher mapping density. Therefore, these RDs can be used to detect and identify the CNVs. The read-depth data provide more reliable information for large CNVs and CNVs flanked by repeats, where accurate mapping reads is difficult. The read depth data also provide information on CNVs based on the targeted sequences where only targeted regions of the genome are sequenced [4].

3 TUF

TUF cores in WGA sequencing samples can be identified as gaps in RD coverage. Thus, they mimic deletions. This means that TUF cores can be represented as deletions in RD data. Thus, we can adapt existing CNV detection models for TUF detection.

4 Modeling

In [6] a window based algorithm is developed for RD data in order to detect CNVs. The basic idea of their approach is to identify regions of consecutive 100-bp windows with significantly increased or reduced RD counts. In order to detect such a scenario, the read count of a window is converted into a $Z-$score according to equation 1

$$Z_{window} = \frac{RC_{window} - \mu}{\sigma} \tag{1}$$

where μ is the mean RD of all windows:

$$\mu = \frac{1}{W} \sum_{W} RC_w \tag{2}$$

and σ is the standard deviation.

The $Z-$score is then converted to its upper-tail probability:

$$p_i^{Upper} = P(Z > z_i) \tag{3}$$

$$p_i^{Lower} = P(Z < z_i) \tag{4}$$

For an interval of consecutive windows A with l windows, they classify an unusual event as duplication if

$$max\{p_i^{Upper} | i \in A\} < \left(\frac{FPR}{L/l}\right)^{1/l} \tag{5}$$

or as a deletion if

$$max\{p_i^{Lower} | i \in A\} < \left(\frac{FPR}{L/l}\right)^{1/l} \tag{6}$$

FPR is the nominal false-positive rate (FPR) desired for the entire chromosome (deletion and duplications are treated separately), L is the number of windows of a chromosome, and l is the size of the interval A.

In [1] a Hidden Markov Model in order to detect CNV from SNP genotyping data is discussed. The model assumes six hidden states; full deletion, single copy deletion, normal heterozygote, normal homozygote, single copy duplication and double copy duplication. The exponential function in equation 7 is used in order to define the a priori transition probabilities.

$$\rho = \frac{1}{2}(1 - exp(-d/2L)) \tag{7}$$

In 7 d is the distance between adjacent SNP loci. The distance between neighbouring SNPs determines the probability of having a copy number state change between them. L is a characteristic length which, as the authors state, could either be inferred directly from the data or adjusted calibrate the model to a given false positive rate in an objective fashion.

The emission probabilities are defined by a simple mixture model of the form

$$p = U + (1 - U)N(\mu, \sigma) \tag{8}$$

where U denotes the uniform distribution ad $N(\mu, \sigma)$ the normal distribution. The uniform distribution is used to model random fluctuations.

In [5] another HMM based model is presented for CNV detection in whole-genome SNP genotyping data. The HMM is based on the HMM model in [1].

5 Our Approach

In this section we describe the modeling approach we follow. We begin by discussing the core elements behind Hidden Markov Models. The discussion is taken from [2]

5.1 Hidden Markov Model

In this section we briefly describe the workings of Hidden Markov Models (HMM). We then proceed in discussing the application of the model.

An HMM has the following core elements [3], [2]

- N the number of states in the model. The set of states is denoted by S. Hence, $S = S1, \cdots S_N$ and $|S| = N$. At time t the state is denoted as s_t.
- M the number of distinct observation symbols per state. The observation symbols correspond to the physical output of the system that we model.
- The state transition probability matrix A where

$$A_{i,j} = P(S_{t+1} = s_j | S_t = s_i) \tag{9}$$

- Emission probabilities with $e_{i,\alpha}$ as the probability of emitting symbol α in state i with $\sum_{\alpha \in A} e_i(\alpha) = 1$
- Initial state probability $\pi = \pi_i$ as the probability of being at state s_i at instant $t = 0$; i.e.

$$\pi_i = P(S_i = s_1), 1 \leq i \leq N \tag{10}$$

For the special case where any state can reach any other state in a single step we have that

$$A_{i,j} > 0, \forall i, j \tag{11}$$

Given the appropriate values of N, M, A, B and π the HMM can be used as model for generating an observation sequence

$$O = O_1 O_2 \dots O_T \tag{12}$$

Each observation $O_i \in V$ and T is the number of observations in the sequence.

1. Choose an initial state s_1 according to π
2. Choose O_t according to the symbol probability when in state S_i that is $e_{i,k}$

The procedure above can be used as both a generator of the observations and as a model for how ta given observation sequence was generated by an appropriate HMM.

An HMM can be summarized by the vector

$$\lambda = (\pi, A, E) \tag{13}$$

that is the HMM is characterized by the initial state probabilities, the transition matrix A, and the emission probabilities E.

For an HMM in order to be useful one should be able to answer the following three questions [2]

1. Given the observation sequence O and a model λ, how can we compute $P(O|\lambda)$. This is the evaluation problem; given a model and a sequence of observations, how do we compute the probability that the observed sequence was produced by the model?
2. Given the observation sequence O and a model λ, how can we choose a corresponding state sequence Q which best explains the observations?
3. How do we adjust the model described by λ to maximize $P(O|\lambda)$? This is the training problem.

The three models are linked. Problem 1 can be solved using the forward-backward algorithm [2]. Problem 2 can be solved using the Viterbi algorithm whilst problem 3 can be solved using Baum-Welch algorithm [2].

5.2 HMM development

We are considering two samples for the same individual. One sample is undergoing WGA (sample m605) before sequencing. Sample m585 does not. From the produced sequenced data we then extract the region that corresponds to the chromosome 1. The extracted data is then aligned into non overlapping windows of length L. Reads of quality less than Q30 were rejected.

A read is considered only if it satisfies a certain quality threshold (The `pysam` function `get_query_qualities` which returns query base quality scores at pileup column position). The extracted RD data for each file is then fitted to a normal distribution

In our approach we assume that the system can be modelled using four states. Hence, the state set is

$$S = \text{NORMAL}, \text{TUF}, \text{INSERTION}, \text{DELETION} \tag{14}$$

We can further refine the S set following previous studies e.g. [1] and [5] where they assumed that the NORMAL state is better described by assuming separate states namely Normal heterozygote and Normal homozygote. Furthermore, we assume every state can lead to any other state. Therefore currently we assume a fully connected model. The observations set V contains only the RD counts. Since the TUF state mimics the DELETION state when considering RD counts (i.e. reduced RD counts) we assume the following:

- Regions of low RD observed in both files do not represent TUF but rather DELETION?
- Regions of low RD observed in m605 that are NORMAL in m585 are indicative of TUF.

.

In our case, we have that the observations consists solely of RD counts. Furthermore, the set of states is the following

We assume the following

We want to fit the an HMM model with K states to the vector of the observed RD counts. We can use the

There are several items we need to clarify.

- How do we initialize the transition probabilities?
- How do we define the emission probabilities?
- How do we differentiate between deletion and TUF as the latter mimics deletion in RD?

References

[1] Stefano Colella, Christopher Yau, Jennifer M. Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S. Bassett, Anneke Seller, Christopher C. Holmes, and Jiannis Ragoussis. Quantisnp: an objective bayes hidden-markov model to detect and accurately map copy number variation using snp genotyping data. 2007.

[2] Rabiner L. R. A tutorial on hidden markov models and selected applications in speech recognition. 2009.

[3] Miguel Rocha and Pedro G. Ferreira. Bioinformatics algorithsm. design and implementation in python. 2012.

[4] Cai T. T., Jeng J. X., and Li H. Robust detection and identification of sparse segments in ultra-high dimensional data analysis. 2012.

[5] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F.A. Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. 2007.

[6] Seungtai Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. 2009.