Model-Based Clustering of Digital PCR Droplets using Expectation Maximization

A Thesis Proposal
Presented to
the Faculty of the College of Science
De La Salle University Manila

In Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Statistics

by

GUIAO, Joyce Emlyn B.

Frumencio F. Co
Adviser

March 10, 2020

Abstract

From 150 to 200 words of short, direct and complete sentences, the abstract should be informative enough to serve as a substitute for reading the thesis document itself. It states the rationale and the objectives of the research.

In the final thesis document (i.e., the document you'll submit for your final thesis defense), the abstract should also contain a description of your research results, findings, and contribution(s).

Keywords can be found at `http://www.acm.org/about/class/class/2012?pageIndex=0`. Click the link "HTML" in the paragraph that starts with "The **full CCS classification tree**...".

Keywords: Keyword 1, keyword 2, keyword 3, keyword 4, etc.

Contents

List of Figures

List of Tables

Chapter 1

Research Description

1.1 Introduction

1.2 Background of the Study

Quantification of Nucleic acids (NA) is a developing research field in molecular biology for the detection and quantification expression levels of genes (Huggett, O'Grady, & Bustin, 2015). These NA molecules are found in deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), which carries genetic information and is used as biomarkers for the detection of diseases (Cao et al., 2017). Additionally, along with the rise of bioinformatics tools, NA quantification methods are also utilized in rare mutation detection, copy number variation detection, single-cell gene and microRNA expression analysis, and next-generation sequencing (Quan, Sauzade, & Brouzes, 2018). Outside the scope of molecular biology, its application has also found its way in forensic research (Whale, Cowen, Foy, & Huggett, 2013), medical diagnosis, environmental monitoring, and food safety analysis (Cao et al., 2017).

To be able to determine the concentration of target NAs, NA detection is naturally a pre-requisite. There are, however, NAs of interests that have very low concentrations to the point that it becomes undetectable in existing detection technologies. This problem is solved by amplifying the NA sequences using Polymerase Chain Reaction (PCR), a widely-used method for NA amplification since its invention in the 1980s (Cao et al., 2017). PCR can multiply specific NA sequences in DNA or RNA from low concentrations to millions of copies. This method exposes the NA sequences mixed with chemical components in a series of 20 to 40 temperature cycles. In each cycle, PCR doubles the NA molecule;

theoretically producing 2^n molecules after n cycles (Quan et al., 2018).

After PCR amplification, absolute NA quantification is achieved using digital Polymerase Chain Reaction (dPCR) technique. This equally divides the NA samples into thousands of partitions; each of these partitions is evaluated as either off or on, or in this context, labeled as positive or negative, hence the term "digital" (Cao et al., 2017).

The dPCR workflow, as illustrated in Figure 1.1, is usually a sequential procedure of extracting the sample from an organism, concocting the sample with several chemical components into a reaction mix, distributing the reaction mix to equal partitions, amplifying and detecting the target molecules using PCR, and the concentration is then finally estimated using a Poisson correction factor. In (Jacobs, Goetghebeur, & Clement, 2014), it was emphasized that every step of the dPCR workflow inevitably allows for the introduction of different sources of variation. These variance components within the dPCR workflow is shown in Figure 1.2.

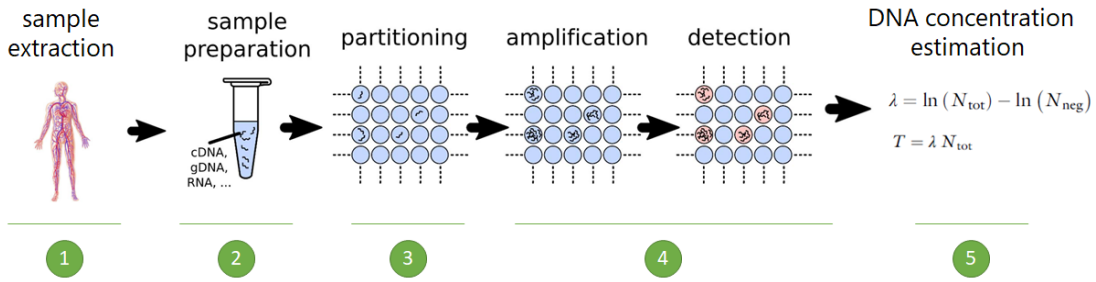

Figure 1.1: The dPCR workflow

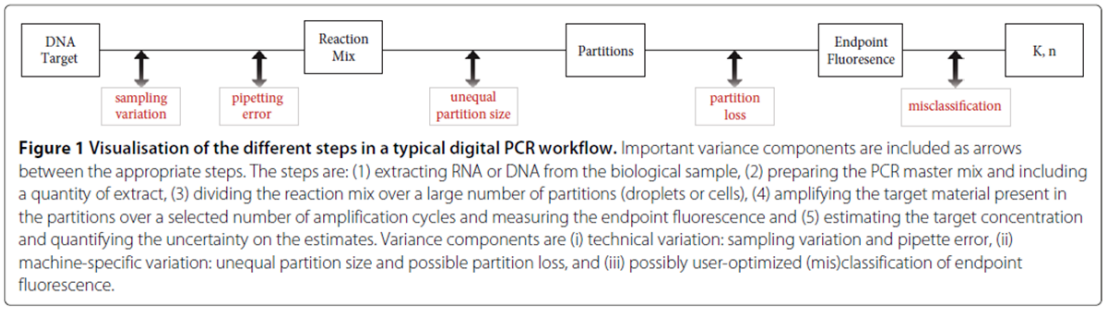

Figure 1 Visualisation of the different steps in a typical digital PCR workflow. Important variance components are included as arrows between the appropriate steps. The steps are: (1) extracting RNA or DNA from the biological sample, (2) preparing the PCR master mix and including a quantity of extract, (3) dividing the reaction mix over a large number of partitions (droplets or cells), (4) amplifying the target material present in the partitions over a selected number of amplification cycles and measuring the endpoint fluorescence and (5) estimating the target concentration and quantifying the uncertainty on the estimates. Variance components are (i) technical variation: sampling variation and pipette error, (ii) machine-specific variation: unequal partition size and possible partition loss, and (iii) possibly user-optimized (mis)classification of endpoint fluorescence.

Figure 1.2: Potential variation components between steps of the dPCR workflow

Sampling variation stems from the fact that only a small sample of the organism is extracted; and although there is an expected number of target molecules per liters of a sample, drawing equally sized samples will result in different target molecules that are more or less near the average. Tzonev (n.d.) demonstrates the

number of target molecules that can be drawn from extraction is distributed as Poisson. Besides the sampling error, samples may also exhibit imperfections, and thus have inhibited amplification.

Preparing the reaction mix is a delicate process that strictly requires the accuracy of the pipetting volume; and yet, technical variation still occurs that results in pipetting errors. The next variation components are the possibility of the distributed partitions to be of unequal volumes and the loss of partitions due to physical interventions. Finally, upon PCR amplification, each droplet partition emits an endpoint fluorescence that would be used to classify the partition as positive or negative. However, some partitions are difficult to classify due to inhibition, delayed reactions, primer depletion, and other biological factors.

Each variance component accumulates to the bias and variance of the final estimated target molecule concentration, and thus, this gives rise to the importance of providing solutions that would increase precision in every step. To increase the sensitivity and specificity of the estimate, the misclassification of droplet partition should be minimized as much as possible. A high presence of false negative droplets reduces sensitivity, while specificity is lowered for high false positive count.

The primary problem in misclassification lies in 'rain' droplets; these are partitions that emits a fluorescence signal that is not clearly classified as positive or negative. Figures 1.3 and 1.4 demonstrates two different DNA targets with the former showing a visually clear distinction of the positive and negative population, and the latter possessing multiple rain droplets. These figures came from the publicly available dataset from the study of Lievens et al. (2016).

Figure 1.3: Fluorescence readings of 4 repititions of DNA target cru

Figure 1.4: Fluorescence readings of 4 repititions of DNA target TC1507

1.3 Statement of the Problem

This thesis aims to classify dPCR droplet partitions into positive or negative by exploring Model-Based clustering with Expectation Maximization. The specific objectives of this study are to:

1. fit two-component finite mixture densities on the datasets with varying amounts of rain and increasing target DNA concentration;
2. classify partitions using the fitted models for each experiment and estimate DNA concentration using the standard formula;
3. and evaluate and compare the precision and bias of the estimates amongst the existing classification methods.

1.4 Significance of the Study

1.5 Scope and Limitations

Chapter 2

Review of Related Literature

2.1 dPCR Droplet Classification Methods

2.1.1 Threshold

2.1.2 Population Detection of Gaussian Kernel Densities

2.1.3 K Nearest Neighbors

2.1.4 Non-parametric Mixture Models

2.2 Model-Based Clustering

Chapter 3

Theoretical Framework

3.1 G-component Finite Mixture Density

3.2 Expectation Maximization Clustering

Chapter 4

Methodology

4.1 Data

4.1.1 Rain Experiment Dataset

Plate 2 - Primer and Probe Concentration Gradient

Plate 4 - PCR Enhancers Experiment

Plate 5 - Cycle Gradient

Plate 6 - Sonication Gradient

Plate 7 - Annealing Temperature Gradient

4.1.2 DNA Quantification Dataset

Plate 3 - Rain Dilution Series

Albumin

4.2 Model Fitting and Classification

4.3 Performance Evaluation

Appendix A

Diagrams and Other Documentation Tools

This appendix may consist of proposed architectural design, algorithms, scientific formula for MSCS and Data Flow Diagrams, Fishbone for MSIT.

Appendix B

Theoretical and/or Conceptual Framework

Discusses the basic framework/foundation the thesis is based on. This section is normally referred to when discussing Scope and Limitations, and Research Methodology

Appendix C

Resource Persons

Dr. Firstname1 Lastname1
Adviser
College of Computer Studies
De La Salle University-Manila
emailaddr@dlsu.edu.ph

Mr. Firstname2 Lastname2
Role2
Affiliation2
emailaddr2@domain.com

Ms. Firstname3 Lastname3
Role3
Affiliation3
emailaddr3@domain.net

References

Cao, L., Cui, X., Hu, J., Li, Z., Choi, J. R., Yang, Q., ... Xu, F. (2017). Advances in digital polymerase chain reaction (dPCR) and its emerging biomedical applications. *Biosensors and Bioelectronics*, *90*(November 2018), 459–474. Retrieved from `http://dx.doi.org/10.1016/j.bios.2016.09.082` doi: 10.1016/j.bios.2016.09.082

Huggett, J. F., O'Grady, J., & Bustin, S. (2015). QPCR, dPCR, NGS - A journey. *Biomolecular Detection and Quantification*, *3*(March 2007), A1–A5. doi: 10.1016/j.bdq.2015.01.001

Jacobs, B. K., Goetghebeur, E., & Clement, L. (2014). Impact of variance components on reliability ofl absolute quantification using digital PCR. *BMC Bioinformatics*, *15*(1), 1–13. doi: 10.1186/1471-2105-15-283

Lievens, A., Jacchia, S., Kagkli, D., Savini, C., & Querci, M. (2016). Measuring Digital PCR Quality: Performance Parameters and Their Optimization. *PloS one*, *11*(5), e0153317. doi: 10.1371/journal.pone.0153317

Quan, P. L., Sauzade, M., & Brouzes, E. (2018). DPCR: A technology review. *Sensors (Switzerland)*, *18*(4). doi: 10.3390/s18041271

Tzonev, S. (n.d.). Chapter 3 Fundamentals of Counting Statistics in Digital PCR : I Just. , *1768*, 25–43.

Whale, A. S., Cowen, S., Foy, C. A., & Huggett, J. F. (2013). Methods for Applying Accurate Digital PCR Analysis on Low Copy DNA Samples. *PLoS ONE*, *8*(3). doi: 10.1371/journal.pone.0058177