Innovations in Computing Technology and Applications
Volume 2 | Year: 2017 | ISBN: 978-967-0194-95-0

Identification of Cancer Copy Number using Array Comparative Genomic Hybridization

Nurul Emira Abdul Aziz[1], Hamimah Mohd Jamil[2], Haslina Hashim[3]

Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia
[1]emiraaziz93@gmail.com, [2]hamimah@utm.my, [3]haslinah@utm.my

Abstract: Somatic alterations in genes such as point mutations, copy number alterations and structural arrangement can lead to various types of cancer. There are many types of technique used to identify copy number alterations such as array Comparative Genomic Hybridization (aCGH) that are widely used in detecting copy number. However, most of the techniques are time consuming and involved high cost experiment. In this research, aCGH technique has been selected to detect copy number alterations that makes use of the log2 intensity ratios in the aCGH data. Moreover, segmentation algorithm is used to identify the same mean log2 intensity ratio of copy number. Bayesian Hidden Markov Model (HMM) method is being used to find genome-wide changes in copy number and posterior inferences about copy number gains and losses are created. To evaluate the effectiveness of the method, aCGH dataset has been used to test the method, demonstrating its reliability. The observation on results obtained in this research indicates the effectiveness of the proposed algorithm based on Bayesian HMM.

Keywords: Somatic alterations, array Comparative Genomic Hybridization, copy number, segmentation method, Bayesian Hidden Markov Model.

1 Introduction

Cancer is one of the most dangerous and common causes of deaths worldwide which is also known as the silent killer that are comprised of unusual cell development that can leads to other diseases and harm the other part of the body. The growth of cancer is being triggered by the changes in somatic genetic alterations such as single base substitutions, translocations, infections and copy number alterations (Beroukhim, 2012). Furthermore, aberrations of this type affect the function of genes and thereby produce a transformed phenotype (Taylor et al. 2008). Many researchers are being focusing in studying and identifying effective methods to identify copy number alterations as it is one of the main factor that leads to cancer.

According to Chiang et al. (2009), the effective ways to detect cancer-causing genes is by identifying genomic regions that show recurrent copy number alterations or known as gains or losses regions in tumor genomes. Copy number alterations or known as CNAs are somatic alterations in chromosome structure that cause the gain and loss in DNA copy sections and leads to various types of cancer (Bierly, 2013). Moreover, changes in copy number of the large genome regions can cause phenotypes that are the cumulative consequences of copy number changes genes, which on their own have little phenotypic significance (Tang et al. 2013). There are different types of CNA exist such as polyploidies,

aneuploidies, partial or segmental aneuploidies, copy number variations and deletions or insertions.

Even though there are a lot of approaches and methods had been studied, a way to identify copy number still be a challenging field to study in achieving the best methods or algorithm in detecting copy number. Some of the techniques based on microarray technologies such as array comparative genomic hybridization (aCGH) can simultaneously measure thousands to millions of loci in the genome for DNA copy number changes (Chen et al. 2005). Array CGH is a method that combines comparative genomic hybridization (CGH) with microarrays techniques for better detection of copy number alterations. This technique uses slides arrayed with small segments of DNA as the targets for analysis of copy number (Zack et al. 2008). There are many advantages in using this technique as it is capable to concurrently identify aneuploidies, mutations, copy number and any other chromosomal aberrations (Theisen et al. 2008).

Then, the data obtained from aCGH process will be further process using segmentations methods such as Circular Binary Segmentation (CBS), Gain and Loss of DNA (GLAD) and Bayesian Hidden Markov Model (HMM). Segmentation methods are developed for partitioning clones into sets with the same copy number segments (Willenbrock et al. 2005). CBS method was developed by Olshen et al. 2004 to segment a chromosome into contiguous regions and bypasses parametric modelling of the data with its use of a permutation reference distribution. Meanwhile, GLAD works by merging segmented levels by iteratively removing excessive breakpoints and subsequently cluster segments across chromosomes to assign levels of copy number gain and loss (Willenbrock et al. 2005). Bayesian HMM proposed by Guha et al. in 2008 is a hybrid of Bayesian learning in identifying genome-wide changes in copy number and it makes uses of hidden Markov model (HMM) that takes account for the dependence between neighboring clones.

In order to solve problems in identifying copy number alterations in genomic regions, we proposed a Bayesian HMM method to identify copy number as it method is proven to be successful by recent published research due to its effectiveness in detecting copy number alterations. This algorithm is being implemented in MATLAB and tested using aCGH dataset to achieve goals and objectives of this research. The performance of the algorithm is being evaluated at the end of this research. Hence, this research is expected to be useful in detecting CNA and hope can be a benefit to cancer studies in the future.

2 Copy Number Alterations Analysis

In this research, Bayesian HMM is used to identify copy number alterations in genomic regions. The Figure 2.1 shows the flowchart for copy number alterations identification process. The crucial steps in the copy number alterations identification process are being explained in the following subtopics.

2.1 Data Acquisition and Preparation

This phase explained about the dataset used to perform the research. The dataset of fibroblast cell line was downloaded from Nature Genetics website in an Excel file format.

The dataset can be reach at http://www.nature.com/ng/journa l/v29 /n3/suppinfo/ng754S1.html. then, the dataset can be execute using MATLAB.

2.2 Pre-Processing Data

The aims of this phase are to retrieve normalized log2 intensity ratio of fibroblast cell line that consists of negative and positive values. Since, aCGH data can be quite noisy, a robust computational method is needed to filter and smoothing the data to get an accurate result. In this phase, a non-parametric filter has been used to perform a high-level smoothing process.

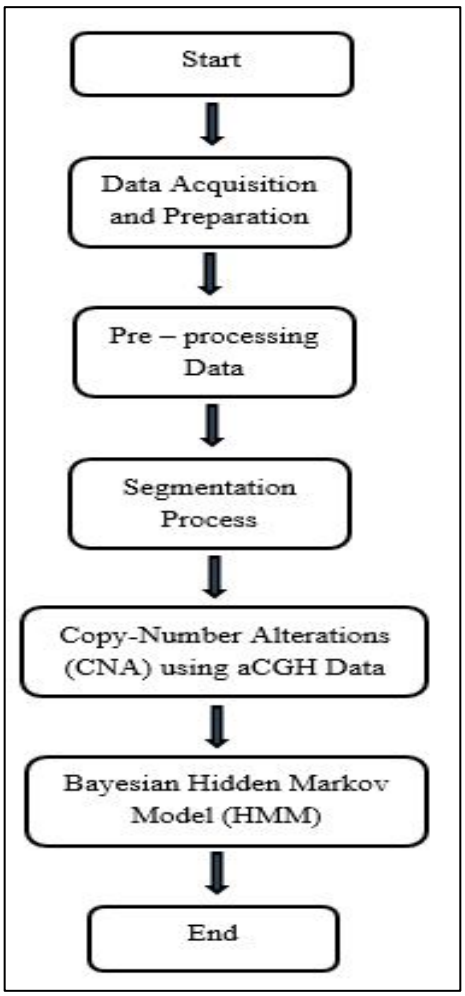

Figure 2.1: Flowchart of CNA identification process

2.3 Segmentation Process

In this phase, genome is divided into contiguous segments to identify segments that has same mean log2 intensity ratio that is assumed to have the same underlying copy number. This phase is also known as smoothing process perform for noise reduction, detection of aberration such as loss, gain or normal copy number and breakpoint analysis purposes.

2.4 Copy – Number Alterations (CNA) using aCGH Data

The main part in analyzing copy number alterations using aCGH data is the detection of segment boundaries of copy number changes and assumption of the copy number state for each segment. In this research, bacterial artificial chromosome (BAC) is used to detect and analyze the gains and losses in copy number alterations. Besides that, log2 intensity ratios used in this research provide useful information about genome-wide CNAs. The log2 intensity ratios of GM01524 cell line from chromosome 1 through 23 has been displayed and studied in this research. Structure array for GM01524 cell line was created for used in further analysis step.

In this phase, change-points of copy number was detected using the derivatives of the smoothed ratio over a certain threshold that usually will specify substantial changes with large peaks and it provides the estimate of the change-point indices. Then, Gaussian Mixture or Expectation-Maximization was used to optimize change-points indices. After that, permutation t-tests was performed to assign the significance of the segments identified from the change-point detection process before. The last process was segmenting the means of the targeted chromosome and plot the data over the original data by using function in MATLAB.

2.5 Bayesian Hidden Markov Model (HMM)

This phase is the most crucial and important phase as it will do the comparison between the segmentation methods. Bayesian learning was used to identify genome-wide changes in copy number from GM01524 cell line dataset. Meanwhile, the posterior inferences are made about the copy number gains and losses. There are four states in Bayesian HMM algorithm defined as single-copy loss state for state 1, copy-neutral state for state 2, single-copy gain for state 3 and amplification or multiple gain state for state 4. In this research, normalized log2 intensity ratios is assumed as

$$Yk \sim N\ (\mu sk, \sigma^2 sk) \tag{4.1}$$

The expected log2 ratio of all clones are defined as μj, where j = 1,…,4. For example, the expected log2 ratio for single-copy loss is μ3. The biological interpretation associated with the state space allows us to assume that the following ordering is:

$$\mu_1 < \mu_2 < \mu_3 < \mu_4 \tag{4.2}$$

Apart from that, the Bayesian approach assumes priors for all unknown parameters. The priors for mean copy number changes are as assumed as below:

$$\mu 1 \sim N\,(-1, \tau^{12})\,.\,\mathrm{I}(\mu_1 < -\epsilon) \tag{4.3}$$

$$\mu_2 \sim N(0, {\tau_2}^2)\,.\,(-\epsilon < \mu_2 < \epsilon) \tag{4.4}$$

$$\mu_3 \sim N\,(0.58, {\tau_3}^2).\,\mathrm{I}(-\epsilon < \mu_3 < 0.58) \tag{4.5}$$

$$[\mu_4|\mu_3, \sigma_3] \sim N\,(1, {\tau_4}^2).\,\mathrm{I}(\mu_4 > \mu_3 + 3\sigma_3) \tag{4.6}$$

where each of the prior represent for single-copy loss state, copy-neutral state, single-copy gains state and multiple-copy gains or amplification state.

3 Experimental Result

Fibroblast cell line dataset has been used in this research to detect copy number alterations in genomic regions. Meanwhile, pancreatic dataset was used to compare Bayesian HMM algorithm analysis with CBS algorithm analysis. Figure 1 show the analysis perform using CBS algorithm while Figure 2 show the analysis perform using Bayesian HMM algorithm. Then, the comparison was done for both algorithm and the result was shown in Figure 3.

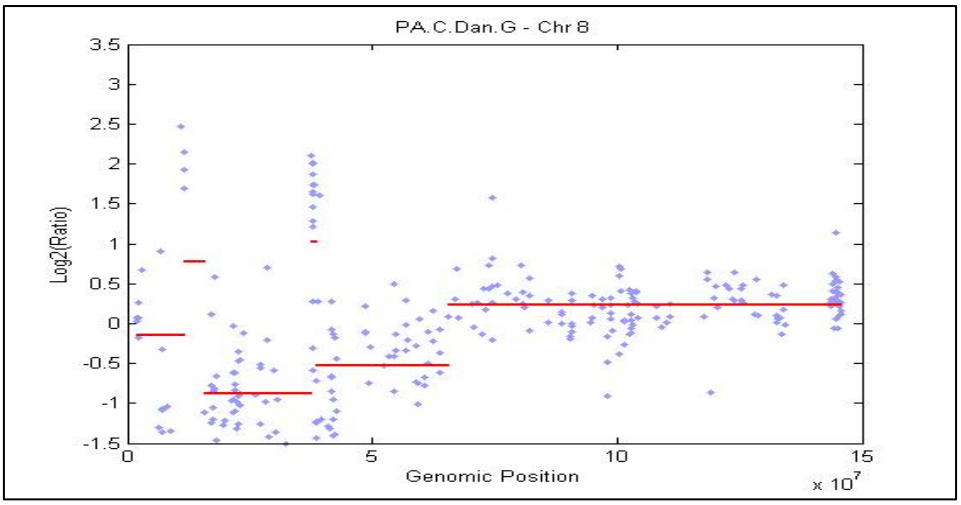

Figure 1: CBS algorithm analysis

The CBS analysis was performed to detect copy number gain and loss in the dataset. The segment was being plotted using ShowPlot parameter in MATLAB and based on analysis done, CBS algorithm found a region that consists of copy number changes in chromosome 8 represent with red lines shown in Figure 3.1. Clones within a segment are assumed to be sharing the same copy number.

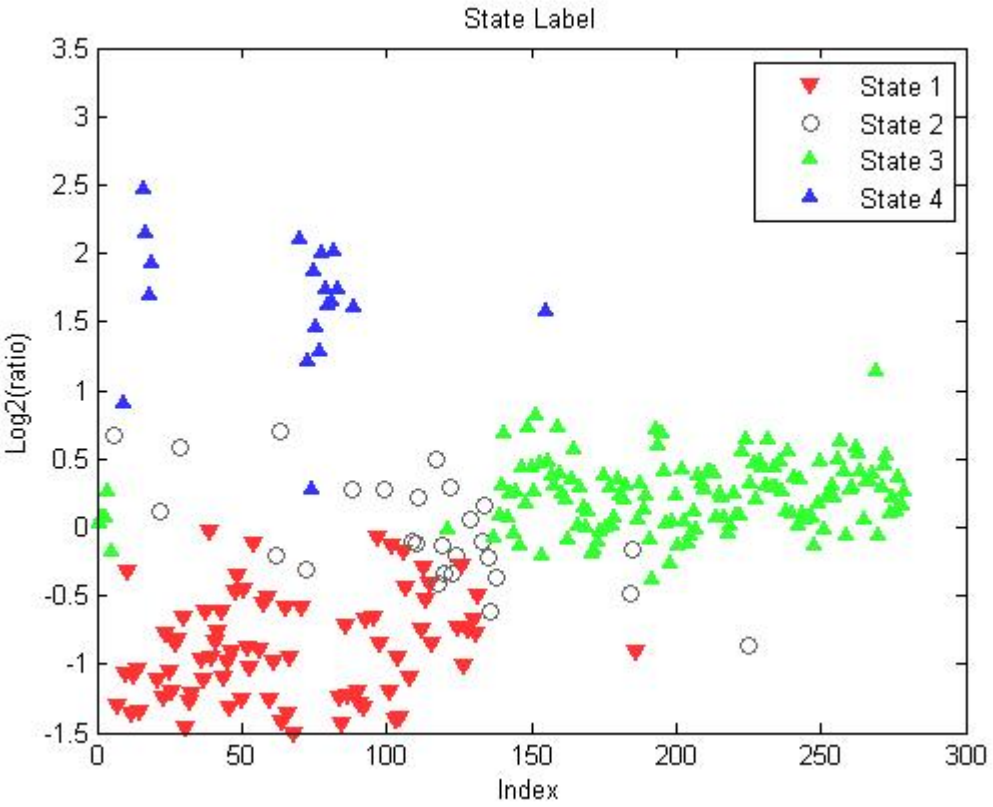

Figure 2: Bayesian HMM analysis

Meanwhile, Bayesian HMM analysis was performed by setting prior parameters first. Then, the number of states used in this analysis was four that represent state in copy number. The first state represent by red colour was copy number loss state, second state represent by grey colour was copy number neutral state, third state represent by green colour was single copy number gain state and the last state, the fourth state represent by blue colour was amplification or multiple gain state. In this analysis, Bayesian HMM are able to detect two-high-level amplification in chromosome 8 represent by the blue colour triangles.

After that, comparison performance was performed between CBS algorithm and Bayesian HMM algorithm. Figure 3.3 illustrated the result of the analysis from the comparison made. From the Figure 3.3, CBS algorithm seem to be failed to detect two-high-level amplification or multiple gain in copy number while Bayesian HMM algorithm successfully detect the two-high-level amplification or multiple gain in copy number. Besides that, CBS algorithm only can segment the aCGH data but it cannot detect gain or

losses in copy number compared to Bayesian HMM algorithm which is more reliable and can immediately recognize the region with copy number changes.

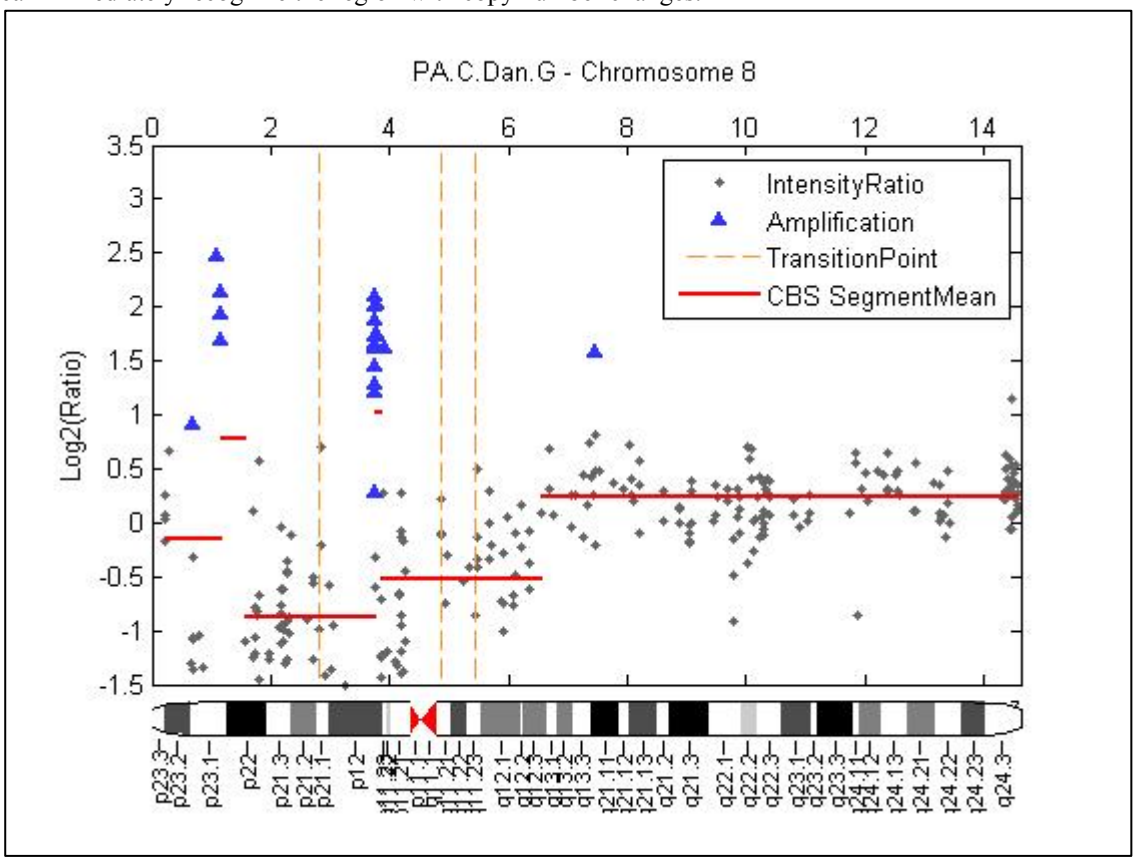

Figure 3: Comparison of CBS and Bayesian HMM algorithm

4 Conclusion and Future Works

Based on the comparison performance done, it is proven that Bayesian HMM algorithm is more reliable and effective in identifying copy number alterations compared to CBS algorithm. Besides that, Bayesian HMM was able to detect amplification or multiple gain in copy number where CBS algorithm was not able to detect it. However, there are still part that can be improved for further research in the future. For example, copy number changes in genomic regions can also be detected by using any other methods by make uses of log2 intensity ratios. Besides that, maybe other programming languages can be implemented to detect copy number alterations for future works.

References

1. Beroukhim, R. (2012). The Landscape of Somatic Copy-Number Alterations Across Human Cancers, *Nature*.
2. Taylor, B. S., Barretina, J., Socci, N. D., DeCarolis, P., Ladanyi, M., Meyerson, M., Singer, S. and Sander, C. (2008). Functional Copy-Number Alterations in Cancer, PLoS ONE.

3. Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meverson, C. and Lander, E. S. (2009). High- Resolution Mapping of Copy-Number Alterations with Massively Parallel Sequencing, *Nature Methods,* Vol. 6, Issue 1, pages 99-103.
4. Bierly, A., (2013), Somatic Mutations and Copy Number Changes in Cancer: Finding the Right Targets.
5. Tang Y. C. and Amon A. (2013). Gene Copy-Number Alterations: A Cost- Benefit Analysis, Vol. 152, Issue 3, p394-405.
6. Chen, W. Y., Houldsworth, J., Olshen, A. B., Nanjangud, G., Chaganti, S., Venkatraman, E. S., Halaas, J., Teruya-Feldstein, J., Zelenetz, A. D. and Chaganti, R. S. K. (2005). Array Comparative Genomic Hybridization Reveals Genomic Copy Number Changes Associated with Outcome in Diffuse Large B-cell Lymphomas, *PMC Publications*.
7. Zack, T., Schumacher, S. E., Carter S. L., Cherniack, A. D., Saksena G., Tabak, B., Lawrence, M. S., Zhang, C. Z., Wala, J., Mermel, C. H., Sougnez, C., Gabriel, S. B., Hernandez, B., Shen, H., Laird, P. W., Getz, G., Meyerson, M. and Beroukhim, R. (2013). Pan-Cancer Patterns of Somatic Copy Number Alteration, *Nature Genetics*, Vol. 45, No. 10, pages 1134-1140.
8. Theisen, A., PhD, (2008). Microarray-Based Comparative Genomic Hybridization (aCGH), *Nature Education* 1(1):45.
9. Willenbrock, H. and Fridlyand, J. (2005). A Comparison Study: Applying Segmentation to Array CGH Data for Downstream Analyses, *Bioinformatics* Vol. 21, No. 22, pages 4084-4091.
10. Guha S., Li Y. and Neuberg D. (2008). Bayesian Hidden Markov Modelling of Array CGH Data, *J Am Stat Assoc.* 103(482): 485-497.