A Novel Approach to detect Copy Number Variation using Segmentation and Genetic Algorithm

Chihyun Park[1], Youngmi Yoon[1,2], Jaegyoon Ahn[1], Myungjin Moon[1] and Sanghyun Park[1]

1. Department of Computer Science, Yonsei University, South Korea
2. Department of Information Technology, Gachon University of Medicine and Science, South Korea

{tianell, amyyoon, ajk, psiwind, sanghyun}@cs.yonsei.ac.kr

ABSTRACT

Among many forms of genomic variations, copy-number variations (CNVs) can be defined as gains or losses of several kilobases to hundreds of kilobases of genomic DNA. Since many CNVs include genes that result in differential levels of gene expression, CNVs may account for a significant proportion of normal phenotypic variation. Some scientists demonstrated that a large portion of overlapping, currently known common human CNVs, were smaller in his dataset. However, previous experimental studies, performed primarily by a-CGH techniques, are limited to detection of CNVs of large-sized CNVs. Efficient algorithms for finding small-sized CNVs are essential. In our paper, we propose a novel approach to find small-sized CNVs on a-CGH data which is a sequential 2-dimensional clustering method. The algorithm we propose is robust to some level of noise. And regardless of the size of probes, our algorithm can find CNVs consisting of small number of probes.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – *Data Mining*; J.3 [Life and Medical Sciences] : *Biology and genetics*;

Keywords

Copy number variation, a-CGH, WGTP, Segmentation, Genetic Algorithm, Parameter estimation

1. INTRODUCTION

Variation in the human genome is present in many forms, including Single-Nucleotide Polymorphisms (SNP), small insertion-deletion polymorphisms, variable numbers of repetitive sequences, and genomic structural alterations [16]. Among these genomic variations, Copy-Number Variations (CNVs) can be defined as gains or losses of several kilobases to hundreds of kilobases of genomic DNA among phenotypically normal individuals [6]. Since many CNVs include genes that result in differential levels of gene expression, CNVs may account for a significant proportion of normal phenotypic variation [4].

Recently, the researches relating to the genomic variation of human genome are being actively carried out. Two representative platforms to assess CNVs are as follows: (1) comparative analysis of hybridization intensities on SNP genotyping array (2) comparative genomic hybridization with a Whole Genome TilePath (WGTP) array. WGTP platform comprises more than 90% of the euchromatic portion of the human genome and microarray Comparative Genomic Hybridization (a-CGH) data of human genome are being generated. In this paper, we analyzed the data from WGTP platform because this platform is prevalent in current CNV assay.

Perry [9] repeated the comparison with CNVs called by the Redon [12], revealed that 213 of 264 overlapping CNVs (80%) were smaller in his dataset, with 154 of the 264 CNVs (58%) smaller by more than 50%, and concluded that the total genomic content of currently known common human CNVs is likely to be smaller than previously thought. However, previous experimental studies, performed primarily by a-CGH techniques, are limited to detection of CNVs of large-sized CNVs, tens or hundreds of kilobases [14]. Efficient algorithms for finding small-sized CNVs are essential, and we focused on this problem.

In our paper, we propose a sequential 2-dimensional clustering method to find small-sized CNVs. Our algorithm uses log ratio value and position information from WGTP sample and finds segments which are used for scoring phase with six parameters. We assign scores to the probes based on the average log ratio value of the segments, and find CNVs by selecting top scoring probes. Genetic Algorithm (GA) helps to estimate six optimal parameters because their search spaces are wide. GA has excellent exploration power that provides the capability of escaping from local optima and working well when solutions to a problem contain complex interacting part [2].

There are two types of CNVs. One is called gain, which means relatively duplicated part compared with a reference sample. The other is called loss, which means relatively deleted part compared with a reference sample. There could be a possibility that the segments having high intensity ratio which is supposed to be a gain could not be a genomic duplication since a-CGH data have some level of noise in acquired hybridization intensity ratios. The proposed algorithm is robust to some level of noise. Regardless of the size of probe, our algorithm can find CNVs consisting of small number of probes. In other words, if the a-CGH dataset is from higher-resolution tiling arrays, our algorithm can detect small-sized CNVs.

2. RELATED WORKS

2.1 SW-ARRAY

SW-ARRAY [11] is a popular method to find CNVs on a-CGH

SAC'09, March 8-12, 2009, Honolulu, Hawaii, U.S.A.