유전체 단위 반복 변이(CNV) 발견을 위한 개선된 SW-ARRAY

문명진[O1] 안재균[1] 윤영미[1,2] 박치현[1] 박상현[1]
1. 연세대학교 컴퓨터과학과 2. 가천의과학대학교 IT학과
{psiwind, ajk, amyyoon, sanghyun}@cs.yonsei.ac.kr

An Enhanced SW-ARRAY Method for Detecting Copy Number Variations(CNVs)

Myungjin Moon[O1] Jaegyoon Ahn[1] Youngmi Yoon[1,2] Chihyun Park[1] Sanghyun Park[1]
1. Dept. of Computer Science, Yonsei University 2. Gachon University of Medicine and Science

요 약

최근 유전체 단위 반복 변이(CNV)의 중요성이 부각되고 있다. CNV란 DNA가 복제될 때 일부가 만들어지지 않거나 혹은 많이 만들어져 그 양이 차이가 나게 되는 것으로, 인간의 질병이나 형질과 밀접한 관련을 가진다고 알려져 있다. 이에 따라 CNV와 관련된 연구가 활발히 진행되었으며, CNV를 찾기 위한 다양한 방법들이 나오게 되었다. 본 논문에서는 CNV를 찾아내는 대표적인 기법 중 하나인 SW-ARRAY에 대해서 알아보고, 여기에 페널티 값과 점수에 따른 가변 임계값을 적용하여 보정함으로써 기존 SW-ARRAY의 문제점을 해결하는 방법을 제안한다. 이를 실제 Array-CGH 데이터에 적용한 결과 긍정 오류 값이 줄어들어 기존의 방식에 비해 정확한 값을 얻게 되었다.

1. 서론

인간의 유전자 변이는 여러 가지 형태로 나타난다. 그 중 유전체 단위 반복 변이(Copy Number Variation, 이하 CNV)는 최근 유전체 연구 분야에서 많은 관심을 받고 있다.

CNV란 DNA가 복제될 때 일부가 만들어지지 않거나 혹은 많이 만들어져 그 양이 차이가 나는 경우가 생기는 것을 의미한다. 기존에는 단일 염기 다형성(SNP)이 개인의 독특한 유전 형질을 나타내는 지표로 알려졌으나 최근에는 CNV가 더 중요한 지표로 인식되고 있다[1]

널리 통용되는 기준에 따르면 마이크로어레이(Microarray) 분석법으로 찾을 수 있는 1 kbp 이상의 변이와, 현미경으로 관측될 수 있는 3Mbp이하의 영역에서 서열이 반복되거나 결실되는 변이를 좁은 의미의 CNV로 정의한다. 처음에는 이러한 종류의 변이가 병리적 상태를 나타내는 것으로 알려졌다. 그러나 2004년에 건강한 사람의 유전체에도 많이 존재한다는 것이 보고되었으며[2][3], 후속 연구들을 통해 CNV가 유전체에 광범위하게 분포된다는 것이 본격적으로 알려지면서[1] CNV가 인간 유전체의 다양성에 어느 정도 기여하는지에 대해 많은 연구가 진행되고 있다.

† 이 논문은 2006년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2006-000-11106-0).

현재까지 CNV를 찾아내기 위한 많은 방법이 개발되어 적용되고 있으나, 대부분 데이터에 오차가 크고 사용자가 임의로 설정한 매개 변수와 임계값에 지나치게 민감하게 반응한다는 단점이 있다. 따라서 CNV를 정확하고 효율적으로 찾아내기 위한 새로운 기법들이 필요하다.

본 논문에서는 CNV를 찾아내는 대표적인 기법인 SW-ARRAY[4]에 대해 살펴보고, 보다 나은 성능을 내기 위해 임계값을 보정하여 개선한 기법에 대해서 알아보도록 한다.

2. 관련 연구

SW-ARRAY는 비교 유전체 보합법(Array Comparative Genomic Hybridizaion, 이하 Array-CGH)을 통해 얻어진 데이터를 바탕으로 CNV를 찾아내는 기법이다. Array-CGH는 사람 유전체의 93.7%를 포함하고 있는 WGTP(Whole Genome Tiling Path) Array에 특정한 색으로 염색한 컨트롤 DNA와 테스트 DNA를 뿌리고, 그 발현 정도의 차이를 분석하는 방법이다[1] 이 실험의 결과값을 이용하여 CNV의 여부뿐만 아니라, 복제 수가 늘어난 것인지 줄어든 것인지도 알아낼 수 있다. Array-CGH를 통해 얻어진 데이터를 바탕으로 CNV를 구하는 기법은 Price et al.[4], Fiegler et al[5], Shah1 et al.[6], 등이 있는데, 그중 대표적인 기법이 SW-ARRAY이다.

SW-ARRAY는 Smith-Waterman의 동적 프로그래밍 알고리즘[7]을 1차원적으로 변형시켜 적용한 기법이다. SW-ARRAY 기법의 목표는 높은 값 혹은 낮은 값이 연