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

속되는 지역을 구하는 것이다.

입력 값으로는 Array-CGH의 데이터를 보정 및 정규화하여 사용한다. 우선 임계값인 t_0를 설정한다. 일반적으로 t_0의 값은 median+0.2MAD(Median Absolute Deviation)가 사용된다. 복제 수의 증가를 찾을 경우 이 값을 Array-CGH를 통해 얻어진 데이터에서 빼 보정하며 감소를 찾을 경우 반대로 데이터에 값을 더해서 보정해준다. 전자의 경우 보정을 통해 데이터 값 평균이 음수가 되며, 이는 비교적 더 높은 값이 연속되는 지역만을 구하겠다는 뜻이 된다.

그 이후 보정된 값에 Smith-Waterman 알고리즘을 적용하여 연속되는 지역 즉 아일랜드(island)를 구한다. 이때 사용되는 공식은 다음과 같다.

$$S(p)=\begin{cases} S(p-1)+X(p) \text{ if } S(p-1)+X(p)>0 \\ 0 \qquad otherwise \end{cases}$$

$$B(p)=\begin{cases} B(p-1) \text{ if } S(p)>0 \\ p \qquad otherwise \end{cases}$$

여기서 p는 p번째 입력이라는 것을 나타내며, X(p)는 p번째 입력이 가진 값, 즉 점수를 의미한다. B(p)는 아일랜드가 시작되는 지점이며 S(p)는 p 위치까지의 아일랜드 점수이다. S(0) = 0이고, p > 0이다. 이 공식을 통해 전체 데이터 영역에서 가장 높은 점수를 가진 아일랜드를 구할 수 있다. 가장 높은 점수 값을 가진 아일랜드란 영역이 늘어나도 줄어들어도 점수가 더 이상 늘어날 수 없는 곳을 의미한다. 가장 높은 값을 가진 아일랜드를 구한 후에는 이 영역의 데이터를 전부0으로 바꾼 후 알고리즘을 다시 적용하여 그 다음으로 높은 값을 가진 아일랜드를 구한다. 이를 0 이상의 아일랜드의 점수가 반복될 때까지 적용한다. 복제 수의 감소를 찾을 경우 부호를 바꿔서 적용하면 된다. 그림을 통해 살펴보면 다음과 같다.

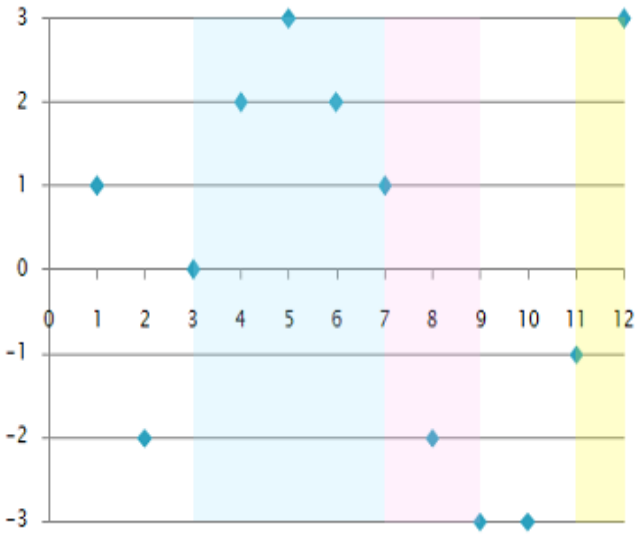


그림 1 : SW-ARRAY의 예(x축 : p, y축 : X(p))

표 1 : SW-ARRAY의 예

p	1	2	3	4	5	6	7	8	9	10	11	12
X(p)	1	-2	0	2	3	2	1	-2	-3	-3	-1	3
S(p)	1	0	0	2	5	7	8	6	3	10	0	3
B(p)	0	2	3	3	3	3	3	3	3	10	11	11

그림과 표에서 p=3~9, p=11~12 구간이 각각 아일랜드에 해당하며, p=3~9 구간의 일부인 p=3~7 구간이 가장 높은 점수를 가진 아일랜드에 해당한다. 그림을 살펴보면 가장 높은 점수를 가진 아일랜드 내에서도 음수 값이 나올 수 있는데, 이는 어느 정도 긍정 오류(false positive)와 부정 오류(false negative) 허용하게 된다.

이 알고리즘의 p-value는 1000개의 데이터의 임의 순열을 통해 구하게 된다. 그리고 임계값에 대한 신뢰도(robustness)는 median과 median+0.4MAD 사이의 100개의 값을 이용해서 테스트한다. 신뢰도의 값이 특정 위치에서 1에 가까우면 이는 CNV를 의미하며, 0에 가까우면 정상에 가깝다는 것을 의미한다. 따라서 신뢰도 0.5 이상인 경우 CNV가 있다고 판단할 수 있다.

3. 개선된 SW-ARRAY 기법과 실험 결과

기존의 SW-ARRAY 기법은 단일 임계값을 사용한다. 그러나 전체적인 점수가 높은 구간일수록 CNV일 확률이 높고, 낮은 구간일수록 CNV가 아닐 확률이 높다. 따라서 여기에서는 아일랜드의 점수의 순위 따라 차등적으로 임계값을 설정함으로써 보정해주는 방법을 사용하기로 하였다.

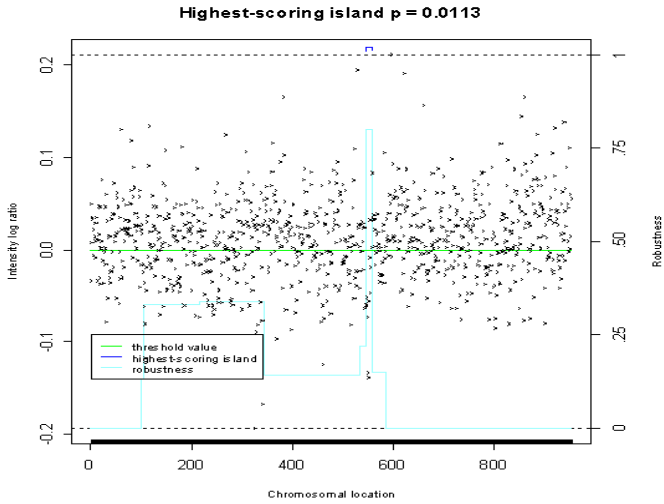
또한 위에서 언급했듯이 가장 높은 점수를 가진 아일랜드 내에서도 음수 값이 나오는 경우를 허용한다는 문제점이 있다. 이를 해결하기 위한 방식은 다음과 같다.

우선 기본 임계값인 median+0.2MAD 값을 이용하여 아일랜드를 찾는다. 이때 아일랜드 내에 음수 값이 나올 경우 X(p)의 값을 2배로 설정하여 페널티를 부여한다. 이런 식으로 찾은 첫 번째 아일랜드, 즉 전체 중에서 가장 큰 점수를 가진 아일랜드의 순위를 1로 설정하고, 알고리즘을 반복 적용한다. 그러면 두 번째로 찾게 되는 아일랜드의 순위는 2가 되며, 마지막에 찾은 아일랜드의 순위는 아일랜드의 전체 수와 같은 n이 된다.

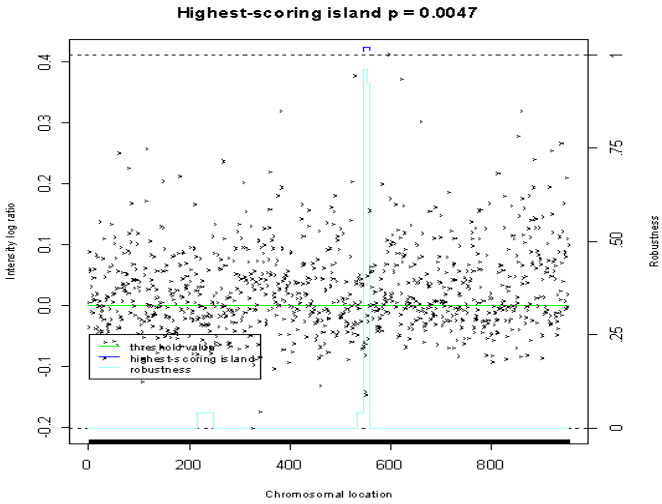
아일랜드를 전부 찾고 나면 각각의 순위에 따라 차등적으로 임계값을 설정한다. 임계값의 범위는 median+0.2MAD와 median+0.4MAD이다. 아일랜드의 수를 n, 순위 값을 r이라 하면 임계값 t_r은

$$t_r=\frac{(median+0.4MAD)-(median+0.2MAD)}{n+1}\times(n-r)$$

이 된다. 이는 낮은 점수를 가진 아일랜드일수록 엄격한 임계값을 적용한다는 것을 의미한다. 이 값을 적용하여 실제 데이터를 통해 출력해본 결과는 다음과 같다. 프로그램은 Thomas S. Price의 cgh R 패키지를 이용하였으며, 복제 수의 감소를 측정하였다.

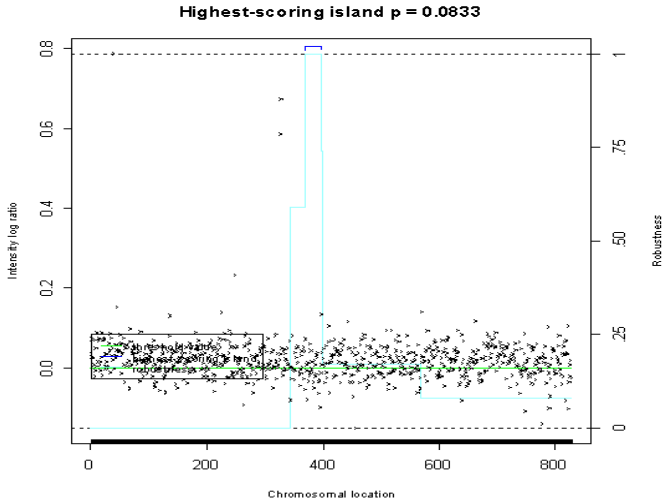


(a) 9번 염색체에 SW-ARRAY 적용

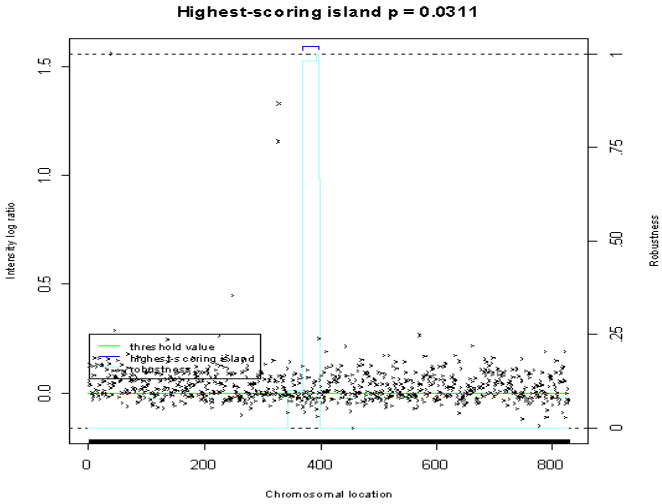


(b) 9번 염색체에 보정된 SW-ARRAY 적용

그림 2 : 9번 염색체에 두 가지 기법을 적용한 결과

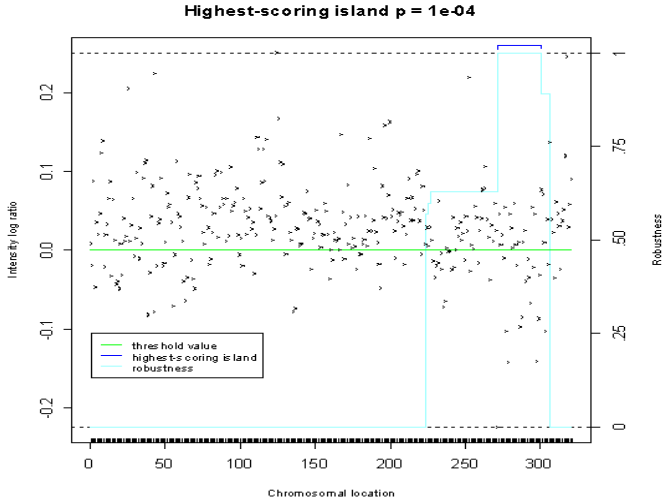


(a) 13번 염색체에 SW-ARRAY 적용

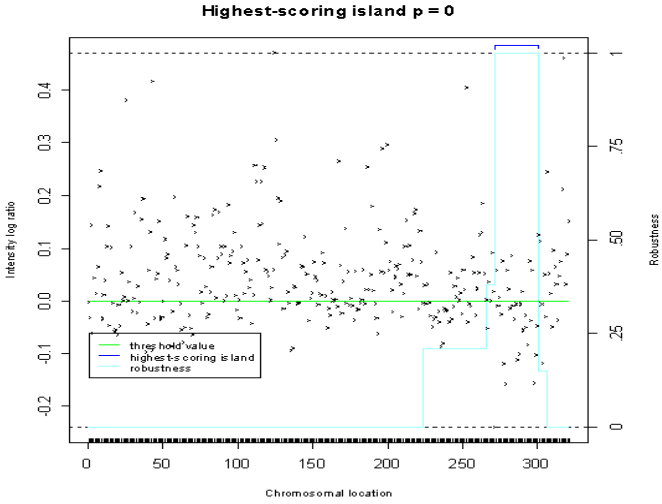


(b) 13번 염색체에 보정된 SW-ARRAY 적용

그림 3 : 13번 염색체에 두 가지 기법을 적용한 결과



(a) 19번 염색체에 SW-ARRAY 적용



(b) 19번 염색체에 보정된 SW-ARRAY 적용

그림 4 : 19번 염색체에 두 가지 기법을 적용한 결과

그림에서 x 축은 몇 번째 데이터인지를 나타내며 y 축은 그 데이터의 발현 값 즉 점수를 나타낸다. Highest scoring island는 가장 높은 점수를 가진 아일랜드를 뜻하며, threshold value는 y 값이 0인 축이다. Robustness 선은 데이터의 신뢰도를 나타내며 이 값이 0.5가 넘어가는 구간에 CNV가 있다고 판단하게 된다. 그림의 p 값은 유의 확률을 의미한다. 위의 예에서는 복제수의 감소를 측정하였기에 y 값이 0 이하인 값이 연속해서 나오는 경우 CNV로 판단하게 된다.

데이터를 살펴보면 기존 방식에 비해 복제 수의 감소가 확연히 나타나는 구간 즉 y 축의 값이 0 아래쪽에 연속해서 나오는 구간을CNV로 판단한다는 것을 확인할 수 있으며, 보정된 SW-ARRAY의 p 값이 기존의 p 값보다 작다는 것을 확인할 수 있다. 이는 보정된 SW-ARRAY가 기존의 방식에 비해서 긍정 오류를 줄일 수 있다는 것을 나타낸다.

4. 결론 및 향후 연구 과제

지금까지 SW-ARRAY에 페널티 값과 가변 임계값을 적용하는 방법에 대해 알아보았다. 기존의 SW-ARRAY 방식은 임계값으로 단 하나의 값을 사용하였기에 유동성이 떨어지고, 아일랜드 내에 점수를 떨어뜨리는 요소가 있어도 그대로 포함시키기에 긍정 오류가 많이 발생한다는 문제점이 있다. 본 논문에서 제안한 기법에서는 다양한 아일랜드의 순위에 따라 차등적으로 임계값을 적용하고, 아일랜드의 점수를 떨어뜨리는 값에 대해 페널티를 부여함으로써 긍정 오류를 줄여 보다 정확한 값을 높은 확률로 도출할 수 있게 되었다.

향후에는 SW-ARRAY 외에 다양한 기법을 연구하여 보다 정확한 CNV 값을 찾을 수 있도록 할 계획이다.

5. 참고 문헌

[1] Redon et al., "Global variation in copy number in the human genome, " Nature, Vol. 444, pp. 444-454, 2006.

[2] Sebat et al., "Large-Scale Copy Number Polymorphism in the Human Genome, " Science, Vol. 305, pp. 525-528, 2004.

[3] Iafrate et al., "Detection of large-scale variation in the human genome, " Nature Genet., Vol. 36, pp. 949-951, 2004.

[4] Price et al., "SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data", Nucleic Acids Research, Vol. 33, No. 11, pp. 3455-3464, 2005

[5] Fiegler et al., "Accurate and reliable high-throughput detection of copy number variation in the human genome, " Genome Res., Vol. 16, pp. 1566-1574, 2006.

[6] Shah1 et al., "Integrating copy number polymorphisms into array CGH analysis using a robust HMM, " Bioinformatics, Vol. 22, No. 14, pp. e431-e439, 2006.

[7] Smith et al., "Identification of Common Molecular Subsequences, " J. Mol. Biol., Vol. 147, pp. 195-197, 1981