대용량 공간 자료들의 세그먼테이션에서의 모수들의 최적화

오미라, 이현주
광주과학기술원 정보통신공학과
e-mail : *omr@gist.ac.kr, hyunjulee@gist.ac.kr*

Optimization of parameters in segmentation of large-scale spatial data sets

Mi-Ra Oh, Hyunju Lee
Dept. of Information and Communications,
Gwangju Institute of Science and Technology

Abstract

Array comparative genomic hybridization (aCGH) has been used to detect chromosomal regions of amplifications or deletions, which allows identification of new cancer related genes. As aCGH, a large-scale spatial data, contains significant amount of noises in its raw data, it has been an important research issue to segment genomic DNA regions to detect its true underlying copy number aberrations (CNAs). In this study, we focus on applying a segmentation method to multiple data sets. We compare two different threshold values for analyzing aCGH data with CBS method [1]. The proposed threshold values are p-value or $Q \pm 1.5IQR$ and $Q \pm 1.5IQR$.

I. 서론

어레이 비교 유전체 보합법 (array comparative genomic hybridization: aCGH)은 염색체(chromosome) 상의 증폭(amplification) 또는 결손(deletion)된 유전자 구간을 측정함으로써, 암과 관련된 새로운 암유전자 (oncogene) 및 암 억제 유전자 (tumor suppressor gene)를 찾을 수 있다. aCGH는 암과 다른 질병들에 대한 조기진단 뿐만 아니라 새로운 유전자의 발굴이나 기능 예측에도 사용된다 [2].

aCGH 데이터는 염색체에서 0.5-20 Mbase 단위로 유전자 개수의 이상 (copy number aberrations, CNAs)을 측정하며, 사용된 마이크로어레이 (microarray)에 따라서 한 샘플 당 수천 개에서 수만 개의 공간 데이터를 생산한다. 이 데이터들의 각 위치에서의 값은 정확한 이상 변이를 측정하지 못하므로, 공간상의 주위 값들을 사용한 세그먼테이션기법을 적용해야 한다.

본 논문에서는 aCGH에서 유전체 변동을 검출하는 방법들 중 하나인 circular binary segmentation (CBS) [1]을 표본의 수가 큰 자료에 적용하여, 증폭 또는 결손을 판단하는데 쓰이는 두 개의 임계 값을 제시한다.

II. 본론

aCGH는 전체 염색체에서 유전자의 증폭이나 결손, 염색체상의 특정 부위의 증감이나 결손을 나타내는 방법으로 circular binary segmentation (CBS) [1],

hidden markov model [3] 등이 있다. 이와 같은 방법들은 임의의 임계 값, 즉, 예를 들어 자료의 logratio 값이 0.25 이상인 값은 증폭, -0.25이하인 값은 결손으로 판단하는데 자료들마다 임계 값이 달라지므로 객관성이 없다. 따라서, 본 논문에서는 객관적인 임계 값을 제시하기 위해서 CBS 방법을 사용하여 각 aCGH의 자료에서 단일평균 t-검정을 실시해 유의확률(p)값과 전체 자료의 제1사분위수(Q_1), 제3사분위수(Q_3), 사분위수범위(IQR)을 구하여 두 가지 방법으로 임계 값을 제시하였다. 첫 번째 임계 값은 CNAs의 값이 양수인 경우 $p<0.0001$ 또는 $Q_3+1.5IQR$보다 크면 증폭, 세포분열의 값이 음수인 경우 $p<0.0001$ 또는 $Q_1-1.5IQR$보다 작으면 결손이라고 판단한다. 두 번째 임계 값은 $Q_3+1.5IQR$보다 크면 증폭, $Q_1-1.5IQR$보다 작을 때는 결손이라고 판단한다.

Ⅲ. 구현

두 가지 임계 값을 비교하기 위해서 다형성 아교모세포종 (glioblastoma multiforme: GBM)을 가진 143명 환자(표본)로부터 얻어진 자료를 사용하였다 [4]. 이 자료를 각 염색체별로 분석하여, 143명 환자 DAN 상의 증폭 또는 결실의 백분율 결과는 그림1과 그림2와 같다.

그림1은 $p<0.0001$ 또는 $Q\pm1.5IQR$값을 기준으로 그림2는 $Q\pm1.5IQR$ 값만을 기준으로 한 결과이다. 그림에서 증폭은 빨간색 (Gain), 결실은 초록색 (Loss)로 표시하였다. 그림1에서 7번과 20번 염색체에서는 증폭의 비율이 높고, 10번, 13번, X 염색체에는 결손의 비율이 높다, 그림2에서는 7번과 12번 염색체는 증폭, X 염색체에는 결손을 나타내고 있다. 그리고, 그림1은 그림2를 포함한다.

Ⅳ. 결론 및 향후 연구 방향

aCGH는 유전자 및 염색체의 증폭이나 결손에 대한 방법은 많이 연구 되어왔으나, 증폭과 결손을 판단하는 객관적인 방법은 중요하지만 아직까지는 연구가 미비하다. 그러므로, 본 논문에서는 CBS 방법을 사용할 경우 객관적인 두 가지 방법을 제시하여 서로 비교하였다. 다른 aCGH 자료를 사용하더라도 쉽게 임계 값을 가지고 증폭 또는 결손을 나타낼 수 있다.

감사의 글

본 연구는 광주과학기술원의 GIST faculty start-up fund의 지원에 의한 것입니다.

참고문헌

[1] Olshen, A. B. *et al*, Circular binary segmentation for the analysis of arrray-based DNA copy number data, Biostatistics, Vol. 5, pp. 557-572, 2004.

[2] Pinkel, D., Albertson, D. G., Array comparative genomic hybridization and its applications in cancer, Nature Genetics, Vol. 37, pp. s11-s17, 2005.

[3] Fridlyand, J. *et al*, Hidden Markov models approach to the analysis of array CGH data, Journal of Multivariate Analysis, Vol. 90, pp. 132-153, 2004.

[4] Kotliarov, Y. *et al,* High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. Cancer Res., Vol. 66, pp. 9428-9436, 2006.

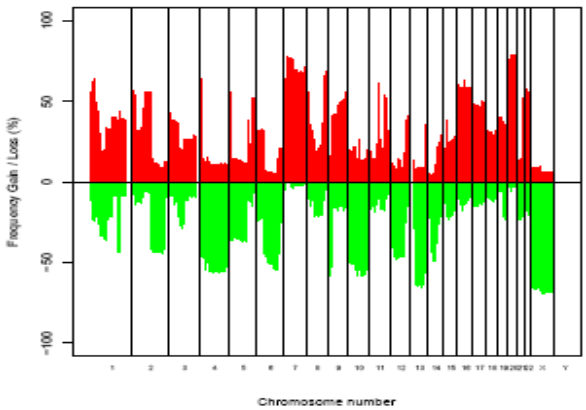

그림 1. CBS : $\alpha=0.0001$ 또는 $Q\pm1.5IQR$

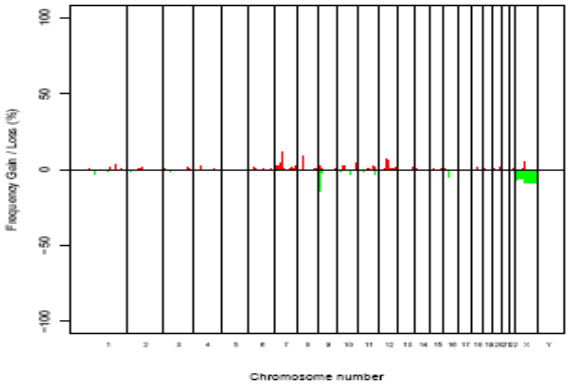

그림 2. CBS : $Q\pm1.5IQR$