균일 격자 구조 탐색을 이용한 마이크로어레이 주소 결정 알고리즘

진희정° 조환규
부산대학교 컴퓨터공학과, ALGORIGENE 연구실
{hjjin°, adagio}@pearl.pusan.ac.kr

An Algorithm for Addressing in Microarray using Regular Grid Structure Searching

Hee-Jeong Jin° Hwan-Gue Cho
Dept. of Computer Engineering, Pusan National University, ALGORIGENE Lab.

요 약

DNA 마이크로어레이(microarray)란 새로운 개념의 기술이 도입되면서, 이를 이용하여 유전체(genome)를 탐색하거나, 동시에 수천 개의 유전자간의 상호작용을 관찰 할 수 있게 되었다. 이러한 이점으로 인하여, 많은 DNA 마이크로어레이 실험이 시행되고 있다. DNA 마이크로어레이 실험으로 생성되는 이미지 데이터는 그 양이 방대하고, 분석하는 연구자에 따라 관점이 달라질 수 있으므로, 이를 효율적으로 분석할 수 있는 방법들이 필요하게 되었다. 하지만, 마이크로어레이 이미지 데이터는 반점(spot) 위치의 변동이나 반점의 모양, 크기가 고르지 않는 것과 같은 다양한 문제로 인하여 자동적으로 분석하기는 어렵다. 본 논문에서는 *마이크로어레이의 균일 격자(regular grid) 구조 탐색을 이용하여 새로운 주소 결정 알고리즘*을 소개한다.

1. 서 론

분자생물학적인 실험 기술의 발달로 인해 등장하게 된 DNA 마이크로어레이(microarray)라는 새로운 개념의 기술 도입으로 하나의 칩(chip)상에서 유전체(genome)의 발현 양상을 탐색하거나, 동시에 수천 개의 유전자들 간의 상호작용을 관찰할 수 있게 되었다.

마이크로어레이 실험의 결과는 일반적으로 Cy3, Cy5 두 형광물질에 따라 두 이미지를 생성하게 된다. 이들 이미지의 가장 큰 특징은 격자 모양으로 구분되어 있다는 것이다. 마이크로어레이 이미지는 격자 모양으로 구분되어 있어, 격자 내부에 존재하는 데이터를 각각 구분하여 분석하기가 용이하다. 일반적으로 격자 구조를 가진 이미지 데이터를 분석하는 경우에는 격자 구조를 결정하는 것을 먼저 수행하여야한다. 이런 과정을 일반적으로 그리딩(griding), 또는 주소 할당(addressing)이라 부르고, 이미지 데이터 분석의 기초가 되므로 아주 중요한 단계라고 할 수 있다. 한번 잘못 계산된 격자의 위치는 결과적으로 잘못된 분석 결과를 도출하기 때문에 정확한 격자 위치 계산이 필요하다. 마이크로어레이 이미지는 블록의 크기나 간격, 반점들의 위치, 모양의 변이등과 같은 문제로 인하여, 마이크로어레이 이미지 분석을 완전 자동화하기가 매우 어렵다. 이로 인하여, 모든 상용 프로그램에서는 매뉴얼 작업을 통해서, 블록과 반점의 주소를 결정하고 있다. 하지만, 최근 자동 마이크로어레이 이미지 분석 시스템들이 소개되기 시작했지만[1,2], 마이크로어레이 이미지의 분석이 자동화되기 위해서는 이미지의 상태가 좋고[3], 일정 비율이상의 유전자들이 발현되어 있어야 한다[4,5]. 마이크로어레이 실험 시에 발생할 수 있는 이러한 문제점을 알아보고 이미지 분석시의 자동화 기능과 그 결과의 정확성을 알아보기 위해서 마이크로어레이의 품질을 측정하는 방법에 대한 연구가 진행되고 있다[8].

2. 점들의 균일화(regularity) 정도

이미지에서 공간상의 균일화(regularity) 정도 값을 찾는 것은 컴퓨터 비전이나, 장면 분석, 적외선 촬영 이미지에서 지뢰를 찾아내는 것과 같은 분야에서 중요하게 다루어져 온 문제이다[6,7]. 서열(sequence)이 정규적(regular 또는 ESCS : Equally-Spaced and Collinear Sequence)이라는 것은 서열에 포함된 점들의 간격이 일정하며, 각 점들이 일직성상에 위치하고 있다는 것이다. 이를 정의하면 다음과 같다[7].

① *Collinear* :
$P \subset E^2$이고, $\overline{P} \subset P$ 일 때($|\overline{P}| \geq 2$),
만약 $\overline{P}$가 collinear하면, $\overline{P}$의 모든 점이 하나의 직선상에 존재한다.

② *Equally spaced* :
$\overline{P} = \{\overline{p_1}, \overline{p_2}, ..., \overline{p_n}\}$가 $|\overline{P}| > 3$ 이고, $\overline{p_i} - \overline{p_{i-1}} = \overline{p_{i+1}} - \overline{p_i}$이면, $\overline{P}$를 equally spaced라 한다.

균일화 서열에 기반 하여, ε을 고려한 균일화 서열(ε-regular sequence)은 서열에 포함된 점들이 균일화 서열이 있는 축에서 최대 ε만큼 떨어져있을 때를 말한다. 이를 식으로 나타내면 다음과 같다[6].

① *ε-regular sequence* : $P \subset E^2$이고, $\overline{P} \subseteq P$일 때, $\overline{P}$가 ε-regular sequence이면, 모든 $1 \leq x_i \leq n$인 $|x_i - \overline{x_i}| \leq \varepsilon, |y_i - \overline{y_i}| \leq \varepsilon$인 $\overline{P} = \{\overline{p_1}, \overline{p_2}, ..., \overline{p_n}\}$인 균일화 서열이 존재한다.

실제, 본 연구에서도 ε을 고려한 균일화 서열을 사용한다. 마이크로어레이의 반점들의 위치나 크기 등이 이상적이지 않으므로, 이를 고려하기 위한 것이다.

3. 격자 주소 결정 알고리즘

본 논문에서는 격자 주소 결정의 자동화를 위하여 마이크로어레이 이미지에서 발현된 반점들의 중점을 입력으로 하여, 가상점을 허용한 ε을 고려한 균일화 서열을 생성하여, 블록과 반점의 격자 주소 알고리즘에 사용한다. 본 논문에서 제안한 격자 주소 알고리즘을 위해서 마이크로어레이 이미지에서 실제 주소 알고리즘을 적용하기 전에 전처리 과정을 거치게 된다. 이는 격자 알고리즘에서 사용할 입력 데이터를 생성하는 단계이다. 입력 데이터는 마이크로어레이 실험에서 발현된 반점들의 중심점의 집합니다. 반점의 중심점을 구하기 위해서는 먼저 고정된 일정한 임계값 T 이상의 강도를 가지는 픽셀의 세그먼트를 찾아서, 중심을 계산한다. 각 세그먼트의 중심점을 s_p라 표기한다. 본 논문에서는 MBR, MASS, Geometry의 세 가지 반점 중심을 고려한다[4,5].

3.1. 기울어진 각과 단위 길이 계산

기울어진 정도와 단위 길이를 구하기 위해서, 본 시스템에서는 최대의 ε을 고려한 균일화 서열을 생성한다. 최대의 ε을 고려한 균일화 서열을 구하기 위한 입력 데이터는 2*ε □ 2*ε로 나눈 셀의 중심을 사용한다. 반점들의 중심점은 반점의 위치 변동이나 모양, 크기의 다양, 샘플의 오염과 같은 다양한 문제로 인하여 동일 간격의 같은 일직선상에 위치하지 않으므로, 이를 그대로 regularity를 구하는데 사용할 수 없다. 따라서 본 시스템에서는 셀의 중심점을 사용한다. 셀 중심점은 ε 만큼의 크기(2*ε □ 2*ε)의 셀(cell)로 전체 이미지를 분할하고, 적어도 하나의 반점을 포함하고 있는 셀을 c_i(valid cell)라고 정의한다. c_i의 중심점들로 이루어진 점집합 C_p(set of points of valid cell)을 구성하고, 이를 최대의 ε을 고려한 균일화 서열을 계산할 때 사용한다.

기울기 정도 θ와 단위 길이 d_u는 이미지가 기울어져 있더라도, 가장 가까운 반점들 사이의 거리가 가장 짧으며, 본 시스템에서는 ε 만큼의 반점의 치우침을 고려하여, C_p를 입력으로 최대의 ε을 고려한 균일화 서열을 구성하였으므로, 획득된 최대의 ε을 고려한 균일화 서열들 중에서 균일화 서열의 원소들 사이의 거리가 가장 작은 균일화 서열들을 이용하여 계산 한다. 가장 작은 거리를 가지는 균일화 서열을 reg_{min}, reg_{min} 집합을 R_{min}이라고 하자. 전체 이미지에서 블록별로 기울어진 정도가 약간씩 다르거나, 블록의 특정 부분이 한쪽으로 기울어져 있을 수 있으므로, reg_{min} 하나만을 사용하여 θ를 구하면 전체 이미지를 왜곡 시킬 수 있다. 따라서 θ를 구하기 위해서 R_{min}의 모든 reg_{min}를 고려한다. 단위 길이 d_u는 R_{min}의 거리가 된다. [알고리즘 1]은 θ를 계산하는 알고리즘이다.

[알고리즘 1]. 기울어진 각 θ를 계산하는 알고리즘

Algorithm: Computation of Rotational Angle
Input: R_{min}; /*set of reg_{min}, which is regularity with minimum distance.*/
Output: θ /*microarray rotational angle*/

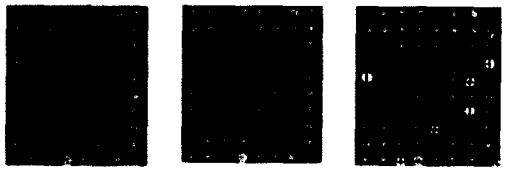
```
· angle1 = 0; /* sum of angles which is larger than 0 */
· angle2 = 0; /* sum of angles which is larger than 0 */
· n1 = 0; /* the set of n1 which is number of angle1 */
· n2 = 0; /* the set of n1 which is number of angle2 */
· n3 = 0; /* the set of n1 which is number of other */

for each element R_min_i in set R_min
    Point p1, p2 = first and second point in R_min_i;
    θ = atan(p1,p2);
    if θ > 0 then
        angle1 += θ; n1++;          endif
    else if θ > 0 then
        angle2 += θ; n2++;          endelsif
    else then
        n3++;                       endels
endfor
if max(n1, n2, n3) = n1 then    θ=angle1/ n1;   endif
else if θ > 0       then        θ=angle2/ n2;   endelsif
else                then        θ = 0;          endels
```

3.2. ε-regularity 생성

기울어진 각이 계산되어지면, 모든 C_p를 $-\theta$만큼 회전시켜 변환한다.

$$c_{p_i}(x_i, y_i) \rightarrow c_{p_i}'(x_i', y_i')$$
$$x_i' = x_i \cdot \cos(-\theta) - y_i \cdot \sin(-\theta)$$
$$y_i' = x_i \cdot \sin(-\theta) + y_i \cdot \cos(-\theta)$$

이렇게 변환된 C_p'는 마이크로어레이 이미지가 기울어지지 않았을 때의 가능한 셀들의 중점이 된다. C_p'를 이용하여, 가로와 세로만을 고려하여 maximal ε-regularity를 생성한다. 각 regularity를 생성할 때에는 하나의 가상 점(pseudo point)을 허용한다. 마이크로어레이 데이터에서는 발현이 되는 유전자와 그렇지 않은 유전자가 존재하는데, 이를 보완하기 위하여, 본 방법론에서는 하나의 가상 점을 허용하여 regularity를 구성한다. [그림 1]은 하나의 블록 (a)에서의 regularity를 생성한 경우 (b)와, 하나의 가상 점을 허용하여 생성한 regularity의 예이다. (c)에서 하나의 균일화 서열 생성 시 하나의 가상점을 허용한 경우, 전체 6개의 가상점이 추가되었다.



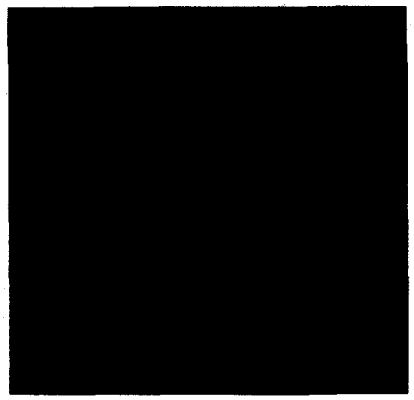
(a) (b) (c)

[그림 1] 하나의 가상 점을 허용한 균일화 서열 : (a) 입력 데이터

3.3. 블록 생성

하나의 가상 점을 허용한 regularity들을 생성하면, 이를 입력으로 하여, 공통 노드를 가지는 regularity들을 하나의 그래프로 연결한다. 모든 공통 노드들을 가진 균일화 서열들을 그래프로 생성하면, 각 그래프들의 MBR을 계산한다. 본 알고리즘에서는 균일화 서열들을 이용하여 블록을 구하기 위해서 연결된 균일

화 서열들을 포함할 수 있는 MBR을 만들고, 그것을 하나의 블록으로 생각한다. 이를 위해서 균일화 서열들에서 서로 공통된 노드들을 지니는 것들을 하나의 그래프로 연결한다. [그림 2]는 본 논문에서 구현한 시스템에서의 블록 주소 결정의 결과이다. [그림 2]의 이미지는 각 블록의 아래쪽 반점들이 한쪽으로 치우쳐져 있지만, 블록 내 부분적인 왜곡이 있는 이미지에서도 블록 주소 결정이 잘 수행됨을 알 수 있다.



[그림 2] standford 이미지 데이터의 블록 인덱싱 : e=1.5화소 이고, 이미지의 해상도는 1024 □1024이다.

3.4. 반점 주소 결정

블록이 결정되면, 각 반점 주소는 일차적으로 블록 주소에서 구해진 단위 길이를 이용하여, 균일하게 블록을 분할하여, 그 중심점을 반점의 주소로 할당한다. 반점의 주소는 블록의 제일 왼쪽 위에 위치하는 반점에서부터 시작하여, 행과 열에서 단위 길이만큼 떨어질 때마다 행과 열에 대해서 1씩 더하면서 주소를 부여한다. 그런 다음, 현재 설정되어 있는 반점 주소에서 반점의 반지름 내에 떨어져 있는 가장 가까운 곳에 위치한 발현된 반점의 중심이 있는 지를 찾아보고, 있다면 현재 주소를 발현된 반점의 중심점으로 변경한다. [그림 3]은 본 논문에서 구현한 시스템에서 반점 주소를 결정한 화면이다. (a)는 생성된 블록에서 일정하게 주소를 부여한 것이고, (b)는 발현된 반점의 중심점을 고려하여, 반점 주소를 변환한 것이다. (c)는 주소가 결정된 반점들의 영역을 표현한 것이다.

(a) (b) (c0

[그림 3] 반점의 주소 할당 : (a) 블록의 위치에서 단위 길이만큼 일정하게 나누어서 각 반점들의 주소를 할당한 화면. (b) 발현된 반점들을 고려하여, 반점 주소를 변환한 화면. (c) 주소가 결정된 반점들의 영역을 표현한 화면.

4. 결론

현재 개발되어진 대부분의 DNA 칩 분석 시스템들은 정확한 이미지 처리를 위해서는 사용자의 많은 개입이 필요하고, 이로 인하여 한 연구자가 처리할 수 있는 이미지의 양이 줄어들며, 여러 사람이 이미지를 처리하게 되므로 이미지 분석에서 일관성이 떨어진다는 것이다. 본 논문에서는 이를 해결하기 위해서 마이크로어레이 이미지 분석에 있어서 완전 자동화된 블록과 반점들의 주소 결정 알고리즘을 제시하였다. 본 논문에서 사용한 알고리즘은 이미지 분석에서 많이 연구되어왔던 점들의 균일화에 대한 연구를 마이크로어레이 이미지에 맞게 하나의 가상 점을 허용한 ε을 고려한 균일화 서열 문제로 변환하여 사용하였다. 본 논문에서 제안한 방법은 반점의 수가 n이고, 적어도 하나의 반점을 포함한 셀의 수가 m일 때, 최악의 조건($n=m$이고, 가장 작은 거리를 가장 나중에 찾을 경우)에서 ε을 고려한 균일화 서열을 생성하기 위한 시간인 $O(m^2)$ 만큼의 시간이 소요된다. 본 논문에서 제시한 알고리즘은 하나의 가상 점을 허용한 ε을 고려한 균일화 서열을 마이크로어레이 이미지 분석에 맞도록 변형하여 사용하였다. 그리고 사용자의 입력을 최소화한 완전 자동 주소 결정론이므로, 사용자가 쉽게 사용할 수 있다. 마지막으로, 본 논문에서 제시한 알고리즘은 마이크로어레이 이미지 뿐 만 아니라, 격자 구조를 가진 다른 이미지 데이터에도 사용할 수 있다. 향후 연구 현재 발현된 반점들의 간격이나, 기울기를 고려하여 ε 값을 자동으로 계산해 줌으로써, 사용자의 개입을 더욱 줄여줄 수 있는 방법에 대한 연구가 필요하다.

5. 참고 문헌

[1] J. Buhler, T. Ideker, and D. Haynor, "Dapple: Improved Techniques for Finding Spots on DNA Microarrays," *University of Washinton CSE Technical Report UWTR 2000-08-05*, 2000.

[2] C.S. Brown, P. C. Goodwin, and P. K. Sorger, "Image metrics in the statistical analysis of DNA microarray data," *Proceedings of National Academy of Sciences of the United States of America*, 98 : 8944-8949, 2001.

[3] M. Steinfath, W. Wruck, H, Seidel, H. Lehrach, U. Radelof, and J. O'Brien, "Automated image analysis for array hybridization experiments," *Bioinformatics*, 17 : 634-641, 2001.

[4] Ho-Youl Jung and Hwan-Gue Cho. "An Automatic Block and Spot Indexing with k-Nearest Neighbors Graph for Microarray Image Analysis". *Bioinformatics(Supple.)*. 18 : 141-151, 2002

[5] 정호열, 황미녕, 유영중, 조환규. "cDNA 마이크로어레이 이미지를 위한 그래프 모델과 분석 알고리즘", *한국정보과학회논문지*. 29 (7/8) : 411-421. 2002

[6] Gabriel Robins, Brian L.Robinson and Bhupinder S.Sethi. "On Detecting Spatial Regularity in Noisy Images". *Information Processing Letters*, 69. 1999

[7] Andrew B.Kahng and Gabriel Robins, "Optimal algorithms for extracting spatial regularity in images", *Pattern Recognition Letters*, 12, 1991

[8] 김판규, 진희정, 조환규, "Quality Measures for Microarray Design and Experiment", 2 : 155p, *KSBI* 2003