DNA칩 이미지 처리를 위한 완전 그리딩 알고리즘

김판규[0] 정호열 조환규
부산대학교 전자계산학과
e-mail: {pgkim, hyjung, hgcho}@pearl.cs.pusan.ac.kr

A Perfect Gridding Algorithm for DNA Chip Image Processing

Pan-Gyu Kim[0] Ho-Youl Jung Hwan-Gue Cho
Dept. Computer Science, Pusan National University

요 약

본 논문에서는 DNA칩 이미지 처리시스템을 위한 완전 그리딩 알고리즘을 제안한다. DNA칩 이미지를 분석하여 처리할 수 있는 많은 DNA칩 분석 시스템이 있다. 하지만 이전의 시스템들은 정확한 이미지 처리를 통한 올바른 유전자 발현정보를 얻기 위해서 많은 사용자의 개입이 필요한 단점이 있었다. 본 논문에서는 사용자의 개입이 없는 정확한 자동 이미지 처리를 위해서, ϵ-그래프 모델링 기법을 제시하고, MBR, Mass, Geometry 등 세가지 종류의 반점(spot)중심을 이용한 완전 그리딩 알고리즘을 제안한다. 제시된 이미지 처리 기술은 완전한 자동 DNA칩 분석 시스템으로, 사용자의 개입없이도 정확한 DNA 칩 위치 정보를 얻을 수 있다.

1 서론

DNA칩은 유전자의 발현 측정을 목적으로 널리 사용되고 있다 [4, 6]. DNA칩 혹은 DNA microarray는 기계자동화와 전자 제어 기술 등을 이용하여 적게는 수백개부터 많게는 수십만개의 DNA를 유리 박편에 매트릭스 형태로 심어 놓은 것이다. 이러한 DNA칩을 이용한 연구로 동시에 최소한 수백개이상의 유전자를 비교분석하여 검색할 수 있으므로, 유전자의 비교분석에 있어서 획기적인 방법이라 할 수 있다. cDNA microarray 칩은 두가지 다른 환경에서 발현되는 유전자들을 분석하는데 이용 되고 있다. 수천개 이상의 유전자 발현변이를 단 한번의 실험으로 검색 할 수 있는 것이다. 두개의 다른 환경에서 얻은 세포들로부터 mRNA를 추출하여 이들을 역전사 (reverse transcription)시킬 때 각각 다른 색깔의 형광 물질을 띤 염기를 집어 넣어 빨간색(Cy5)이나 녹색(Cy3)을 띤 cDNA를 합성한다. 이와같이 합성된 두 개의 cDNA를 똑같은 양으로 섞어서 하나의 cDNA microarray칩에 결합시킨다 [2, 3, 5]. 각각의 형광 정도는 그 유전자의 발현 정도를 알려주는 것으로 이들 정보를 분석하는 것이 DNA microarray 칩 분석 시스템이다. DNA microarray 칩은 분석 시스템에서는 하나의 이미지로 처리가 되고, 보통 20,000,000이상의 픽셀을 가지는데, 그 용량이 방대하고 정확한 발현 정도 측정을 위해서 효율적인 이미지 처리방법이 필요하다.

Microarray이미지는 그리드내의 반점들의 배열로 이루어져 있으며 이미지 내의 모든 그리드는 같은 수의 행과 열의 반점수를 가진다. 하나의 이미지는 여러개의 그리드로 이루어져 있고, 각각의 그리드를 서브그리드라고 한다. 이상적인 상태의 microarray이미지는 다음과 같은 특성을 가진다.

이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음. (KRF-2000-E00304)

- 모든 서브그리드(sub-grid)의 크기는 같다.
- 서브그리드사이의 간격은 일정하다.
- 반점(spot)들의 중심은 행과 열의 교차점위에 위치한다.
- 반점들의 모양은 기하학적인 원형이고, 모든 반점들의 크기는 일정하다.

하지만 대부분의 경우는 이상적인 상태의 이미지가 아니기 때문에 DNA칩의 분석에 있어서 자동적인 이미지 처리기술은 매우 어렵다. 현재까지 많은 DNA칩 분석 시스템이 연구되어져 왔는데, 반자동 처리시스템인 ScanAlyze[1]와 완전 자동처리 시스템인 AutoGene이 대표적이다[7]. ScanAlyze는 사용자가 직접 손으로 모든 반점들의 위치를 보정해 주어야 하고 각 서브그리드의 행과 열의 반점수를 입력해야하는 단점이 있다. 그리고 AutoGene도 비교적 많은 자동화가 되어 있어 거의 완전 자동처리 시스템에 가깝지만 각 서브그리드의 행과 열의 반점수를 사용자가 입력해야만 하는 단점이 있다.

본 논문에서는 DNA칩 이미지 분석 시스템을 위한 완전 그리딩(perfect gridding) 알고리즘을 제안한다. 우리의 이미지처리 기술은, 본 논문에서 제안한 ϵ-그래프와 세가지의 중심:MBR, Mass, Geometry을 이용하여 다른 시스템들 보다 더 정확하고 사용자의 개입이 적은 방법이다.

그림 1은 효율적인 DNA분석시스템의 개요를 보여준다. DNA칩에서 실제로 계산될 세그먼트를 대략적으로 분리하고, 이것을 이용하여 자동 그리딩을 한다. 그리딩된 결과를 이용하여 배경의 강도를 계산하고, 반점내부의 실제로 유효한 세그먼트를 분리하여, 각 반점의 강도를 계산하여 발현정도를 측정한다. 측정된 발현값을 이용하여 데이터 클러스터링이나 여러가지 데이터 마

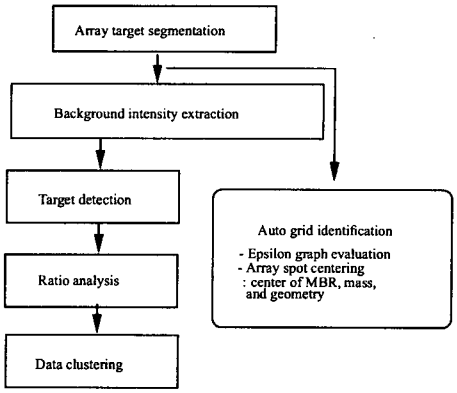

그림 1: DNA칩 분석시스템의 개요 및 제안한 방법

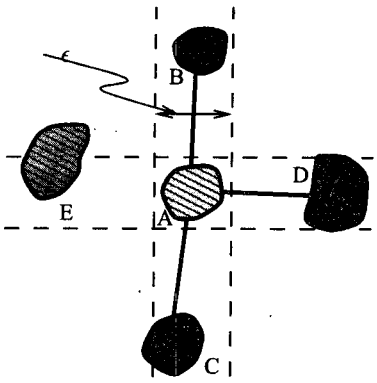

그림 2: 반점 A의 ϵ-그래프

이닝작업을 통한 데이터의 개량화를 하여 유전자의 더 많은 비교분석 정보를 얻을 수 있다. 본 논문에서 제안하는 내용은 ϵ-그래프와 세가지 종류의 중심을 이용하여 자동 그리딩을 하는 과정이다.

2 DNA칩 이미지의 그리딩 문제

대부분의 DNA칩 이미지의 반점들이 앞에서 제시한 이상적인 상태가 아니기 때문에 DNA칩 이미지를 완전 그리딩하는 것은 상당히 어려운 문제다. 완전 그리딩이란 각각의 서브그리딩내의 반점을 정확하게 분리한 그리딩을 말한다. 우리는 기하학적인 방법인 ϵ-그래프와 세가지 종류의 반점중심을 이용해서 완전 그리딩을 하는 방법을 제안한다. 서브그리드내의 모든 반점들은 그래프의 정점으로 나타내어질 수 있고, 각 반점들을 에지로 연결함으로서 이웃한 반점의 정보를 나타낼 수 있다.

2.1 ϵ-그래프

ϵ-그래프, $G_\epsilon(V, E)$는 각 반점의 위치를 정점으로 하는 정점의 집합 V와 반점들 사이의 연결성을 나타내는 에지의 집합 E로 이루어져 있으며, 다음을 만족한다.

- $\forall u, v \in V$, if $(u, v) \in E$ then $\|u.x - v.x\| < \frac{\epsilon}{2}$ or $\|u.y - v.y\| < \frac{\epsilon}{2}$

그림 2에서 반점 A의 상하 및 좌우로 이웃한 반점들에 대한 검색을 할 때 반점 A에서 x좌표의 차나 y좌표의 차가 ϵ범위 이내에 존재하는 반점들 중 가장 가까운 반점들과 에지를 연결하여 ϵ-그래프를 구성하게 된다. 이렇게 구성된 ϵ-그래프에서 내부가 완전히 그리딩된 가장 큰 사각형R을 구하여 단위거리(unit distance)를 계산한다. 실험에서 고려하는 ϵ값은 1, 2, 3, 4의 네가지이다.

2.2 반점(spot)의 중심

ϵ-그래프를 구성하기 위해서는, ϵ-그래프의 정점의 집합이 되는 각 반점 중심의 위치를 찾아야 한다. 반점의 중심을 찾기 위해서 먼저 고정된 일정한 임계값 T 이상의 강도를 가지는 픽셀의 세그먼트를 찾아서 중심을 계산한다. 우리는 MBR, Mass, Geometry의 세가지 반점중심을 고려한다:

- MBR 중심 : 각 세그먼트를 둘러싸는 MBR(Minimum Bounding Rectangle)을 찾아서, MBR의 중심을 반점의 중심으로 한다.
- Mass 중심 : 각 세그먼트의 픽셀들의 평균위치를 반점의 중심으로 한다.
- Geometry 중심 : 각 세그먼트내에서 외부로 부터 차례로 각 픽셀의 레이어를 결정하여 가장 높은 레이어를 가지는 픽셀을 반점의 중심으로 한다.

3 완전 그리딩 알고리즘

반점중심의 종류에 따라 ϵ-그래프의 정점인, 반점의 중심들이 결정되고, 설정된 ϵ값에 따라 이웃한 반점들 사이의 연결성을 에지로 표현하여 ϵ-그래프를 구성할 수 있다. 4종류의 ϵ값과 3종류의 중심이 있으므로, 하나의 DNA칩 이미지에 12개의 ϵ-그래프가 생성된다.

각각의 그래프에서 내부가 완벽하게 그리딩되어 있는 사각형 중 가장 큰 면적의 사각형R을 찾는다. R의 내부는 완벽하게 그리딩이 되어 있으므로 R의 내부에 존재하는 에지의 평균값을 서브그리드의 단위거리(unit distance)로 결정한다.

Algorithm: FindUnitDistance(G_ϵ)
Input: $G_\epsilon = (V, E)$
Output: d, unit distance of given G_ϵ
Objective Function: maximize $f(R)$
1. Find largest rectangle, R, in G_ϵ.
2. Find d by avergaing all edges of R.
3. return d.

여기서 $f(R)$는 평가 함수로써 주어진 R내의 유사 정사각형(pseudo square)의 개수를 계산하는 함수이다. 유

사 정사각형이란 각 변을 a, b, c, d라고 할 때 그 길이의 비가 다음과 같은 사각형을 말한다:

$$\|a\| : \|b\| : \|c\| : \|d\| \simeq 1 : 1 : 1 : 1$$

단위거리가 결정되면 서브그리드를 단위거리로 나누어서 각 반점의 위치를 예측할 수 있다. 이때 서브그리드의 그리딩라인이 반점의 세그먼트와 교차가 일어나게 되면, 그리드 라인의 좌우 혹은 상하로 조금씩 이동시킴으로써 교차가 발생하지 않게 수정한다. 행과 열의 그리드라인 교차점을 계산하면, 각 반점들의 정확한 위치를 계산할 수 있고 반점내의 발현된 세그먼트만을 나타내기 위해서 일정한 임계값이상의 가장 큰 세그먼트를 둘러싸는 원을 그려줌으로해서 반점내의 실제로 유효한 픽셀들을 나타낸다.

4 실험 및 결과

본 논문에서 제안한 완전 그리딩 알고리즘의 성능을 평가하기 위해서 2가지 종류의 실험데이터를 사용하였다. 첫번째는 임의로 만든 DNA칩 이미지를 사용하였고, 두번째는 실제 DNA칩 이미지를 사용하여 실험하였다. 그림3은 실제로 얻어진 DNA칩 이미지를 완전 그리딩한 모습이다. 이때의 ϵ값은 1이고, MBR중심을 이용하였다.

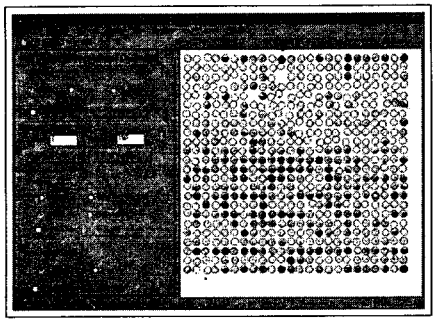

그림 3: 완전 그리딩의 결과 및 시스템 인터페이스

표1은 2가지 종류의 데이터에 대한 실험결과를 보여준다. 각각의 ϵ값과 중심의 종류에 따라서 계산되어진 단위거리와 실제로 완전 그리딩을 한 후의 평균단위거리를 보여준다. Artificial데이터에 대해서는 중심의 종류와 ϵ값에 따른 계산된 단위거리의 변화가 거의 없으나 Real데이터에 대한 결과는 상대적으로 많은 변화가 나타난다. 특히, Real데이터의 경우에는 그리딩후의 평균단위거리값과 가장 가까운 단위거리값을, Mass중심을 이용한 ϵ값1 의 그래프에서 발견할 수 있다. 이처럼 각각의 DNA칩이미지에 가장 알맞은 중심과 ϵ값을 선택한다면 가장 좋은 그리딩을 할 수 있다.

5 결론 및 향후과제

현재 개발되어진 대부분의 DNA칩 분석시스템들의 문제점은 정확한 이미지의 처리를 위해서는 사용자의 개입이 필요하고 이로 인한 처리량의 감소 및 데이터값의 일관성을 유지할 수 없다는 점이다. 본 논문에서는 이러한 문제점을 해결하기 위해서 ϵ-그래프와 세종류의 반점중심을 이용한 완전 그리딩 알고리즘을 이용하여 해결하였다. 우리가 제안한 방법은 사용자의 개입없이 정확한 그리딩을 가능하게 함으로써 이전 DNA칩 분석시스템의 문제점을 해결하였다.

DNA칩 분석시스템은 유전자의 발현정도를 측정하고, 측정된 정보를 이용해서 데이터 클러스터링을 통한 유전자 그룹화 및 유전자의 유사도등을 측정하여 비교분석 데이터를 얻는다. 이를 위해서 정확한 유전자 발현정도의 수치화 및 효율적인 데이터 클러스터링 기술의 개발, 데이터 개량기술의 개발 등이 향후 연구과제로 필요하다.

Data	평균단위거리	ϵ값	중심종류	계산된 단위거리
Artificial	17.91	1	MBR	16.67
Artificial	17.91	2	MBR	16.67
Artificial	17.91	3	MBR	16.67
Artificial	17.91	4	MBR	16.67
Artificial	17.91	1	Mass	16.73
Artificial	17.91	2	Mass	16.73
Artificial	17.91	3	Mass	16.73
Artificial	17.91	4	Mass	16.73
Artificial	17.91	1	Geometry	16.27
Artificial	17.91	2	Geometry	16.38
Artificial	17.91	3	Geometry	16.38
Artificial	17.91	4	Geometry	16.50
Real	17.78	1	MBR	9.00
Real	17.78	2	MBR	12.00
Real	17.78	3	MBR	12.00
Real	17.78	4	MBR	12.00
Real	**17.78**	1	Mass	**17.50**
Real	17.78	2	Mass	9.00
Real	17.78	3	Mass	9.00
Real	17.78	4	Mass	9.00
Real	17.78	1	Geometry	8.25
Real	17.78	2	Geometry	11.50
Real	17.78	3	Geometry	11.50
Real	17.78	4	Geometry	11.50

표 1: 실제의 단위거리와 완전 그리딩 알고리즘을 이용해서 계산된 단위거리의 비교

참고문헌

[1] http://rana.stanford.edu/software/

[2] http://www.gene-chips.com/

[3] http://www.genechip.co.kr/

[4] David J. Duggan et al. Expression profiling using cdna microarray. *Nature genetics supplement*, 21:10–14, 1999.

[5] Erez Hartuv et al. An algorithm for clustering cdnas for gene expression analysis. *RECOMB'99*, pages 188–197, 1999.

[6] J. DeRisi et al. Use of a cdna microarray to analyse gene expression patterns in human caner. *Nature Genetics*, 14:457–460, 1996.

[7] Soheil Shams. Microarray processing technology: Using array image analysis to combat hts bottlenecks. *BioDiscovery News No. 103*, 1999.