제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003. 5)

TDGS 영상 분석을 통한 자동적 디지털 패턴의 추출

장환, 박유나, 이복주
단국대학교 컴퓨터공학과
{hwan, ypark, blee}@dankook.ac.kr

Automated System on Extracting Digital Pattern for TDGS Image Analysis

Hwan Chang, You-na Park, Bog-Ju Lee
Dept of Computer Engineering, Dan-Kook University

요 약

본 논문은 2차원 전기영동에 의해 나타나는 TDGS 영상을 분석하기 위한 시스템으로 실험적인 특성상 젤 위에 나타나는 반점들의 불규칙한 요소들이 많고 영상의 상태가 좋지 않은 경우 명암도가 떨어지는 반점들의 구분이 힘들게 된다. 기존의 전문가의 육안에 의한 TDGS 영상 분석은 그러한 불안적 요소들에 대해 유연하게 대처할 수 있는 능력이 있었다. 하지만, 그러한 예외적인 경우를 컴퓨터가 처리하기 위해서는 영상의 지역적 상태에 맞는 융통성 있는 영상처리 과정이 필요하고, 실제 분석에 사용되지 않는 반점들을 제외한 유효한 디지털 패턴의 판별이 요구된다. 이에 본 논문에서는 영상의 지역적 특성을 효과적으로 반영한 동적 이진화 방법을 통해 후보 패턴들을 추출하고, 모든 샘플들의 기준이 되는 Reference 패턴과 후보 패턴의 point matching 과정을 통해 디지털 패턴을 추출한다.

1. 서론

현재 생명 과학 분야에서 당면한 중요한 과제는 막대한 DNA 염기 서열 정보를 분석하여 어떻게 인간의 건강과 복지에 이용할 수 있느냐는 것이다 [1]. 사람 개개인의 특정 질병에 대한 차이와 다양한 약물에 관한 반응성 및 효과는 개개인 별로 게놈 상에 나타나는 미묘한 차이 때문인데, 특히 그들 중 가장 흔한 변이인 SNP(Single Nucleotide Polymorphism)를 분석하여 질병의 진단과 예후, 치료와 예방에 이용함으로써 유전학 연구의 강력한 수단으로 이용할 수 있다 [2]. 이러한 SNP의 중요성과 유용성 때문에, 최근 1997년부터 산업계와 학계에서는 SNP의 발굴을 위한 대규모의 노력을 진행하여 왔다. 특히 특정 SNP의 유용성은 인종적으로 상당한 차이를 나타내므로 한국인 특유의 유전적 배경을 바탕으로 하는 SNP의 발굴이 절실히 요구되고 있다 [1].

TDGS는 SNP를 발굴하기 위해 사용되는 전기영동 기술로서 다른 방법에 비해서 정확성과 재현성을 증가시키면서 저비용과 고효율로 특정 유전자에서 지금까지 알려지지 않은 새로운 변이를 발굴하는 것을 가능하게 하는 새로운 방법이다. TDGS를 거쳐 나타나는 영상은 반점(spot)들이 2차원에 걸쳐 분리되며, 서로 다른 유전자마다 모두 개수와 패턴이 다양하게 나타난다. 본 논문에서 실험에 사용한 유전자는 BRCA1이라는 유방암 발현 유전자이다.

유용한 SNP 발굴을 위해서는 수백명의 샘플들을 분석하여야 하는데, 기존에는 이러한 TDGS 영상의 분석을 전문가에 의한 육안으로 식별해 왔기 때문에 소비되는 시간비용과 사람마다 분석의 차이를 나타내는 주관적 판단이 문제가 되지 않을 수 없었다. 본 논문에서는 TDGS 영상 분석의 자동화를 통해 human error를 줄이고 시간 비용을 절약함으로써, 추후 SNP 분석에 관한 연구를 효율적으로 수행할 수 있는 TDGS 영상 자동 분석 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 시스템의 개요에 대해서, 3장에서는 영상처리를 통한 후보 패턴 생성과정을, 4장에서는 Reference 패턴과의 비교를 통한 디지털 패턴 생성과정을 기술하고, 5장에

서는 본 논문에서 제안하는 시스템의 실험결과를 보여준다.

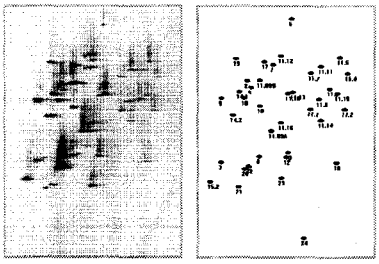


그림 1. TDGS 영상과 디지털 패턴

2. 시스템 개요

해당 유전자의 원활한 SNP 분석을 위해서는 적어도 수 백개 이상의 샘플들이 필요하게 되며, 이를 효율적으로 관리하기 위하여 동일한 유전자들을 하나의 프로젝트 그룹으로 구성한다. 그리고 TDGS 영상의 유효한 반점들로 구성된 디지털 패턴 생성을 위하여 모든 샘플들의 기준이 될 수 있는 Reference 패턴을 생성한다. 이 Reference 패턴은 모든 샘플들과의 point matching을 위한 틀이 되므로 전문가에 의한 수작업이 요구된다. 본 시스템의 대략적 흐름은 그림 2와 같다.

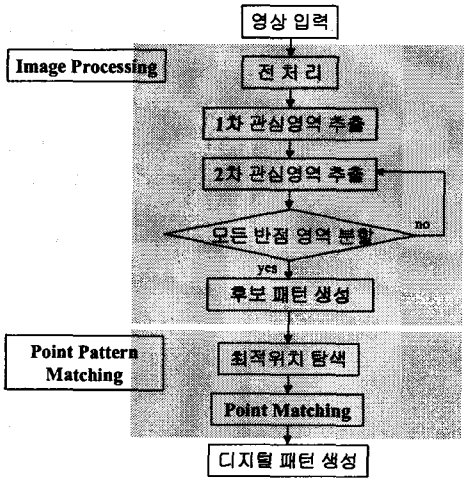


그림 2. 디지턴 패턴 추출 과정

3. 관심영역 추출

입력이 되는 TDGS 영상은 어떠한 전처리 과정도 거치지 않은 원초적인 영상이므로 우선 가우시안 필터 등의 전처리 과정을 통해 잡영을 제거한다. 그리고 영상 내의 반점들을 식별하기 위한 관심영역을 추출하는데 그림 2의 디지털 패턴 추출 과정에서 보는 바와 같이 배경을 제거하기 위한 1차 관심영역의 추출과 1차 관심영역으로 나타나는 각 영역 내의 반점들이 다수 존재하는 지를 검사하여 2차 관심영역을 여러 번에 걸쳐 추출하게 된다.

1차 관심영역은 영상 내의 지역적 특성을 충분히 반영하기 위한 동적 이진화 방법을 사용하게 되며 그 방법은 다음과 같다.

우선 영상을 K개의 일정한 크기의 구역으로 나누고, 각 구역마다 otsu 알고리즘에 의한 지역적 임계치를 구한다[3]. 그리고, N×M 크기의 마스크를 이용한 동적 이진화를 하게 되는데 블록 마스크 내의 지역적 특성을 검사하기 위한 척도(measure)로는 블록 마스크 내의 모든 픽셀들의 분산의 차이, 그리고 해당 픽셀이 위치한 구역 K_i 의 지역 임계치를 사용한다.

■ 분산 척도에 의한 정의는 다음과 같다.

n = 블록 내 픽셀의 개수,
x = 블록 내 픽셀의 명암도
m = 블록 내 명암도 평균

$$p(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

$$g(x, y) = \begin{cases} 1 \text{ if } p(x, y) > V \\ 0 \text{ if } p(x, y) \leq V \end{cases} \quad (1)$$

■ 지역 임계치에 의한 척도는 다음과 같다.

$$g(x, y) = \begin{cases} 1 \text{ if } f(x, y) > T_i \\ 0 \text{ if } f(x, y) \leq T_i \end{cases} \quad (2)$$

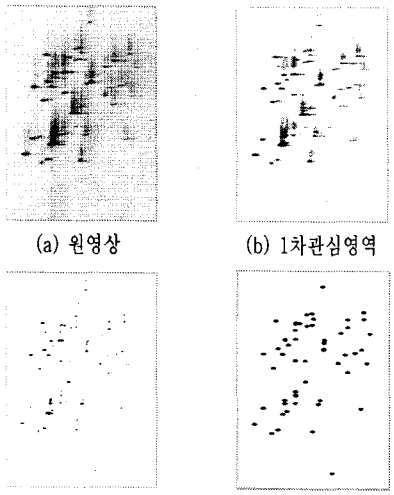
반점들이 특정 위치에 밀집해 있는 경우 1차 관심영역의 추출을 통해서도 다수의 반점들이 하나의 영역 안에 속해 있을 수 있다. 디지털 패턴의 추출을 위해서는 관심영역 내의 반점들이 모두 별개의 영역으로 구분되어야 하므로 이를 검사하여 모든 반점들이 단일 영역을 갖도록 영역별 최적 이진화를 통하여 2차 관심영역을 추출한다.

4. 후보 패턴 생성

최종적으로 나타나는 2차 관심영역은 반점들이 모두 별도의 영역으로 분할된 결과이다. 각 반점들은 디지털 패턴을 구성하는 객체들이 되므로 각 관심영역의 무게중심점을 구하여 이 점을 중심으로 하는

객체들로 구성된 디지털 패턴을 생성한다.

그림 3은 입력된 영상에서 후보 패턴을 추출하는 과정을 그림으로 나타내었다.



(a) 원영상 (b) 1차관심영역

(c) 2차관심영역 (d) 후보 패턴

그림 3. 단계별 영상처리 결과

5. Point Pattern Matching

영상처리 과정으로 생성되는 후보 포인트 패턴들은 유전자의 염기서열을 나타내는 단백질 반점과 무관한 잡영에 해당하는 반점들을 포함한다. 이를 구분하기 위하여 사전에 모든 샘플들의 기준(model)이 되기 위한 Reference 패턴을 정의하였으므로 두 패턴을 비교하여 유효한 디지털 패턴을 추출한다. 하지만 두 패턴은 실험에 의한 물리적 특성 때문에 실제 두 패턴이 정확히 겹쳐지기 힘들고 잡영에 의해 후보 패턴에 나타나는 point의 개수가 Reference 패턴의 point 개수보다 많아지게 된다.

5.1 최적 위치 탐색 (Searching Optimum Area)

정렬 과정에서는 후보 패턴과 Reference 패턴이 가장 잘 겹쳐지는 최적 위치를 탐색하여 이미지의 정렬 기준점을 선정한다.

전역적 위치공간에 대한 탐색은 이미지의 크기에 비례하여 상당히 큰 시간 복잡도가 소요되므로 최소한의 탐색 공간을 이용한 최적 위치 탐색이 요구된다. 본 논문에서는 다음과 같은 제한된 조건에 의한 탐색 공간에 기반하여 최적 위치를 탐색하여 Reference 패턴의 정렬을 수행하게 된다.

■ 최적위치 탐색의 정의

Reference point set R,

$R = \{R_1, R_2, R_3, \ldots, R_n\}$

Candidate point set C,

$C = \{C_1, C_2, C_3, \ldots, C_m\}$

Searching Space Condition:

$R_i(x,y) + S(x,y) = C_j(x,y)$

$$Error\ rate\ e = \sum_{i=1}^{n} \sqrt{(|x_i - x_j|^2 + |y_i + y_j|^2}$$

(C_j = MIN(Distance(R_i,C_j)))

∴ Image alignment of R

= S(x,y) having MIN(e)

Reference 패턴 내의 임의의 점 Ri와 후보 패턴 내의 임의의 점 C_i가 겹쳐지는 탐색 공간의 크기는 n×m이 되고, 탐색 공간에 대한 평가함수(Evaluate Function)는 Reference 패턴 내의 모든 point들과 최단 거리에 있는 후보 패턴들의 거리(Euclidean Distance)의 합이 된다. 여기서 Error rate이 최소가 되는 S가 두 패턴의 정렬을 위한 Reference의 최적 위치가 된다. 그림 4는 두 패턴이 matching하기 위한 최적 위치를 탐색한 후 Reference가 해당 위치로 이동한 그림이다.

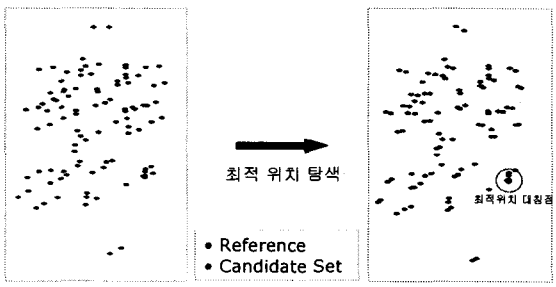


그림 4. 최적 위치 탐색 결과

5.2 Point Matching

Point matching을 위해서는 Reference 패턴의 변형(Transformation)이 필요하며 변형의 종류로는 이동(translation), 회전(rotation), 비율(scale) 등이 있다. Reference 패턴의 변형을 위한 이동, 회전, 비율

의 변형은 다음과 같은 식으로 표현되어 진다[4].

$$P_r=\begin{bmatrix} s\cdot\cos\theta & s\cdot\sin\theta & tx \\ -s\cdot\sin\theta & s\cdot\cos\theta & ty \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}=\begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

TDGS 영상은 특정 크기의 젤 위에서 회전에 의한 변형 없이 일정한 범위(Bound) 내에서 반점들이 나타나므로 변형식에서 θ=0°, s=1로 고정된다. 하지만 반점들이 나타나는 절대적인 위치는 매 실험마다 다소 차이가 있으므로 tx와 ty는 최적 위치 탐색에 의해 변형이 가능하다. point matching을 수행하는 알고리즘은 다음과 같다.

1. Reference point R_i에 대한 변형을 수행
   $T_i = R_i(x,y) + S(x,y)$
2. 모든 T_i에 대해서 최단 거리인 candidate point C_j를 찾아 Ri와 동일한 ID를 부여, T_i와 R_i의 거리를 저장($Di = C_j - T_i$)
3. 동일한 Cj에 ID가 중복될 경우, 중복된 Referenc pont를 R_1, R_2라고 한다면 D_1과 D_2를 비교하여 작은 값을 갖는 R의 ID를 부여
4. Reference point의 모든 ID가 부여될 때까지 1,2,3을 반복

6. 실험 결과

실험에 사용한 시스템은 펜티엄 4 1.8Ghz. 256M RAM의 IBM 호환 PC에 Microsoft Windows 2000 운영체제, Visual C++ 6.0 환경이었다. 모든 TDGS 영상들을 비슷한 유형의 6개의 타입으로 분류하고 그 중 1개씩을 뽑아 시스템의 실험을 위한 입력 영상으로 사용하였다.

표1. 디지털 패턴 추출 결과

샘플	정확도	후보객체 수
9737	82 % (31/38)	73
9748	100 % (38/38)	58
9754	100 % (38/38)	52
9755	87 % (33/38)	59
9759	84 % (32/38)	69
9761	76 % (29/38)	92

실험 결과를 보면 후보 객체의 수가 많아질수록 정확한 추출이 어려워짐을 알 수 있었고, 샘플 9755는 다른 영상에 비해 부분적인 영상의 일그러짐이 많아 다소 정확도가 떨어졌다. 영상이 입력되어 디지털 패턴이 추출되기까지의 시간은 모두 5초 내외로 기존의 전문가에 의한 수작업에 비하면 무시할 수 있는 시간이므로 측정에서 제외하였다.

7. 결론 및 향후과제

본 논문에서는 전기영동에 의한 TDGS 영상을 분석하여 디지털 패턴을 추출하는 자동화 시스템을 구현하였다. SNP 분석을 위해서는 수 백개의 샘플이 사용되어지므로 기존의 작업에 비해 많은 시간 비용을 줄일 수 있고, TDGS 영상의 상태가 좀더 개선된다면 더 좋은 결과를 얻을 수 있다.

TDGS 영상은 전기영동에 의한 실험적 특성에 의해 영상이 일그러지거나 잡영이 심해질 수 있다. 이러한 문제들을 해결을 위해서는 전역적인 패턴의 비교만으로는 완전한 해결이 힘들다. 그러므로 추후 지역적인 패턴의 비교를 통한 알고리즘의 개선이 필요하다.

참고문헌

[1] 서유신, Two-Dimensional Gene Scanning (TDGS)에 의한 SNP (Single Nucleotide Polymorphism) 발굴, 서울대학교 의과대학 암연구소, 2000

[2] N.J. van Orsouw, R.K. Dhanda, R.D. Rines, W.M. Smith, I. Sigalas, C. Eng, and J. Vijg, "Rapid design of denaturing gradient-based two-dimensional electrophoretic gene mutational scanning tests", Nucleic Acids Research, Vol. 26, No. 10, pp. 2398-2406, 1998

[3] *Rafael C. Gonzalez, Richard E. Woods*, "Digital Image Processing", Addison Wesley. pp 443-455, 1992.

[4] J. Denton, J.R. Beveridge, "Two Dimensional Projective Point Matching", Fifth IEEE Southwest Symposium on Image Analysis and Interpretation, 2002, pp.77-81