Density Profile 추출 방법에 따른 염색체 분류정확도 비교분석

최광원*, 송혜정**, 김종대**, 김유섭**, 이완연**, 박찬영**

한림대학교 컴퓨터공학과*
한림대학교 유비쿼터스 컴퓨팅학과**

sodagu@hallym.ac.kr, hjsong@hallym.ac.kr, kimjd@hallym.ac.kr
yskim01@hallym.ac.kr, cypark@hallym.ac.kr, wanlee@hallym.ac.kr

Comparison of Accuracy for Chromosome Classification using Different Feature Extraction Methods based on Density Profile

Kwang-Won Choi, Hae-Jung Song, Jong-Dae Kim,
Yu-Seop Kim, Wan-Yeon Lee, Chan-Young Park
Dept of Computer Engineering, Hallym University*
Dept. of Ubiquitous Computing, Hallym University**

요 약

본 연구에서는 다양한 density profile 특징추출에 기반한 염색체 자동분류방법들의 성능을 비교분석하였다. density profile은 염색체의 밴드패턴을 가장 잘 표현한 특징으로 염색체의 중심축을 구성하는 화소들의 밝기 값을 추출하는 방법이다. 염색체의 밴드패턴은 염색체의 끝단까지를 잘 표현해주어야만 정확한 염색체 번호를 확인할 수 있다. 따라서 염색체의 중심축을 추출하여 염색체 끝단까지 확장 처리한 방법에 대한 성능을 확인하였다. 염색체 중심축에 위치한 화소만을 이용한 프로파일은 잡음에 민감할 수 있으므로 이를 해결하기 위하여 염색체의 중심축에 대한 화소 값 대신 주변 밝기 값들에 대한 평균을 이용한 국소평균방법과 중심축의 수직라인 상에 존재하는 화소 값들에 대한 평균을 구한 수직평균방법을 비교하였다. 분류알고리즘은 k-NN을 사용하였고, 실험데이터는 (주)Gendix 로부터 제공받은 임상적으로 정상인 100명(남자 50명, 여자 50명)으로부터 추출한 4600개의 염색체 영상을 훈련데이터와 테스트데이터로 각각 50%씩 랜덤하게 분리하여 실험하였다. 실험결과 중심축을 확장하고 수직평균에 대한 프로파일을 특징으로 추출하여 분류한 경우가 가장 좋은 성능을 보였다.

1. 서 론

유전정보를 가지는 DNA는 세포분열 중기(metaphrase)에서 염색체라는 형태를 가지면서 광학 현미경을 통해 관찰이 가능해 진다. 인간 염색체는 ISCN (International system for human cytogenetic nomenclature)에 염색체 번호와 분류방법에 대한 규약이 설정되어 있으며 쌍으로 된 22개의 상염색체와 X, Y에 해당하는 성염색체로 구성되어 총 46개의 염색체를 갖는다[1]. 염색체를 통한 세포 유전학적 해석을 위해서는 46개의 개별 염색체를 찾아서 배열하는 핵형 분석(karyotype analysis)이 필요하다. 핵형분석은 초기단계의 암이나 유전적 이상을 진단하는 임상을 위해 중요한 수단이므로 빠르고 정확하게 핵형을 분류하는 자동화된 시스템이 요구된다. 핵형분석 영상은 그림 1과 같다.

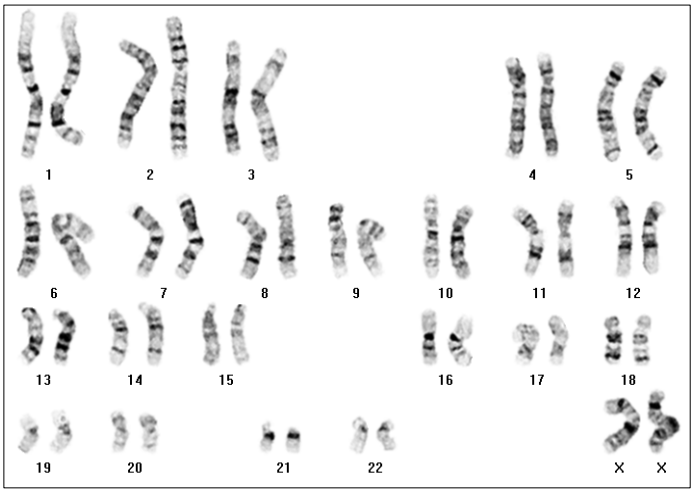

그림 1. 정상 남성의 Karyogram

Giemsa 염색된 염색체는 각 염색체 고유의 밴드패턴을 지니고 있으며 이 밴드패턴은 핵형분류에서 가장 중요한 특징으로 증명되었다[2].

염색체의 밴드패턴은 중심축을 따라 나타나는 연속적인 패턴이므로 중심축에 대한 밝기 값들을 density profile로 표현하여 염색체 분류를 위한 특징으로 사용한다[2-7].

density profile 특징 추출 방법으로는 영상잡음에 대한 민감도를 줄이기 위한 국소평균(Local Averaging - 5 화소 이동 평균법) 방법[5]과 중심축 수직선상에 일부의 점만을 평균하는 방법[6] 또는 수직선상의 전체의 점을 평균화하는 방법[7] 등이 있다. density profile 특징에 염색체의 밴드패턴이 잘 표현되려면 염색체의 끝단까지를 잘 표현해주어야만 하기 때문에 추출된 염색체의 중심축을 염색체 끝단까지 확장 처리하여 성능을 향상시킬 수 있다[6].

본 연구에서는 다양한 density profile 특징추출 방법들에 대한 성능을 비교하였다. 성능 비교를 위하여 세선화 알고리즘으로 얻은 염색체 중심축의 밝기 값들을 기본 density profile 특징으로 추출하고 , 여기에 국소평균방법, 중심축 수직 평균 방법에 대한 성능을 비교하였다. 또한 염색체 중심축의 양끝을 확장하여 최적의 성능을 보이는 조합을 찾고자 하였다. 실험 결과 중심축의 양 끝을 확장하고 중심축 수직 평균 방법으로 추출된 density profile이 다른 방법에 비하여 더 향상된 성능을 보였다. 본 실험은 (주)Gendix에서 제공한 염색체 영상 샘플로 진행하였고 분류기(classifier)로는 k-NN (k-nearest neighbor)을 사용하였다.

2장에서는 염색체의 중심축을 추출하여 확장하는 방법을 설명하고, 3장에서는 density profile 특징 추출 방법을 설명한다.

2. 염색체 중심축의 추출

염색체의 중심축 검출은 Zhang Suen의 세선화 알고리즘을 사용하여 기본 중심축(axis)을 생성한다. 생성된 중심축은 그림 2에서와 같이 염색체의 양끝의 패턴을 포함하지 못한다.

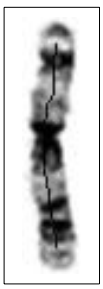

그림 2. 양끝 밴드패턴을 포함시키지 못한 중심축

따라서 염색체의 양 끝단까지 중심축을 확장하는 알고리즘이 필요하다. 확장된 세선화 알고리즘은 추출된 중심축 끝에서 1화소씩의 위치를 확인하여 확장시킨 것과 3화소를 보고 확장시킨 2가지를 구현하였다. 그러나 1개의 화소를 이용하여 확장한 방법은 그림 2에서 처럼 염색체 끝단에서 중심축을 벗어난 결과를 보이므로 3개의 화소를 복사하여 연장하는 방식을 사용하였다. 그림 3에서 3-화소를 사용한 확장 방법이 염색체의 끝단까지 중심축을 잘 표현하고 있음을 확인할 수 있다.

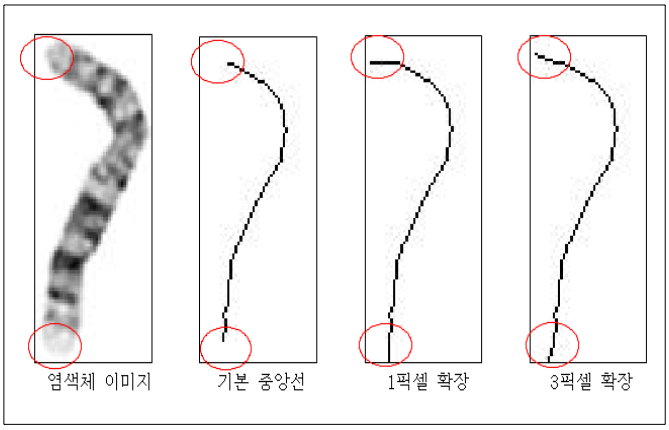

그림 3. 세선화 알고리즘의 중심축과 확장된 알고리즘이 적용된 중심축

3. Density profile

Density profile은 다음 세 가지 방법을 사용하여 추출하였다. 염색체의 중심축을 염색체 영상과 1:1 매핑하여 대응되는 화소의 밝기 값을 이용한 1:1추출방법과 염색체 중심축 상의 잡음에 대한 민감도를 줄이기 위해 중심축상의 전후화소들의 평균을 구한 국소 평균 추출방법(5-화소 이동 평균법)을 사용하였다. 또한 그림 4와 같이 중심축에 패턴이 존재하지 않을 수 있으므로 이를 해결하기 위해 중심축에 대한 수직라인을 구하여 수직라인 상에 위치한 화소들의 평균 밝기 값을 추출하였다. 이것을 중심축 수직평균 프로파일이라 부르며 염색체의 패턴을 잘 표현해주는 특징으로 보여진다.

그림 4. 중심축에 패턴정보가 없는 경우

그림 5는 다양한 방법으로 추출된 density profile이다. 염색체 중심축에 대한 확장 전, 후의 영상에 대한 추출결과이다.

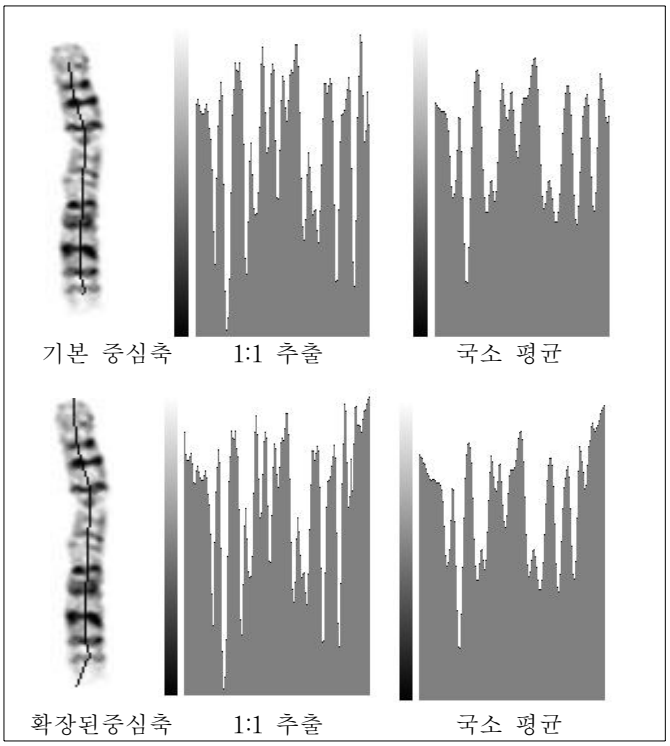

그림 5. 염색체의 중심축 1:1 density profile과 국소 평균을 적용시킨 density profile

그림 6은 중심축 수직 평균 프로파일을 구하기 위해 추출된 중심축의 수직라인을 보인 것이다 영상의 밝은 부분이 추출된 수직라인에 해당된다

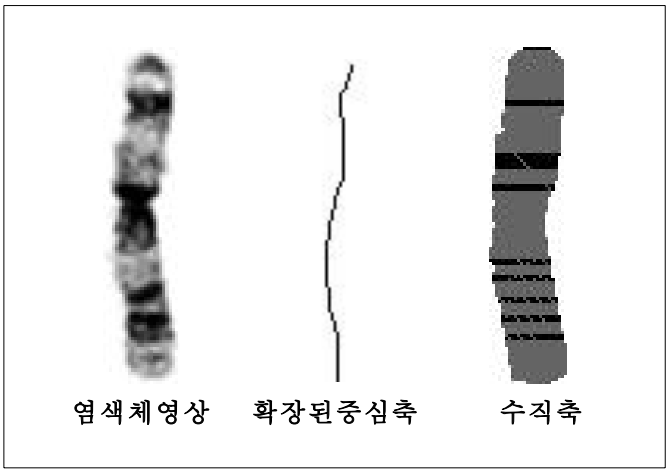

그림 6. 염색체 중심축에 대한 수직축

4. 실험 및 결과

실험방법은 다양한 density profile 추출방법에 대해 중심축 확장 전후의 결과를 비교하였다.

기본 1:1추출방법에서의 확장전후, 국소평균 추출방법에서의 확장 전/후, 수직평균 추출방법에서의 확장 전/후로 나누어 실험하였다.

실험데이터는 (주)Gendix 로부터 제공받은 임상적으로 정상인 100명(남자 50명, 여자 50명)으로부터 얻은 4600개의 염색체 영상을 이용하여 훈련데이터와 테스트데이터는 각각 50%식 랜덤하게 분리하여 실험하였다.

개별 염색체에 대하여 density profile을 추출하였고, 서로 다른 염색체 길이를 규격화하기 위하여 density profile을 32로 샘플링 하였고 샘플링 방법은 선형보간법을 사용 하였다. 또한 염색체의 밝기 값의 차이를 특징값을 0~1로 표준화 하여 데이터로 사용하였다.

염색체 자동분류의 정확도를 확인하기 위하여 3-NN 알고리즘을 사용하였다. 3-NN은 테스트 데이터와 가까운 3개의 훈련데이터를 구하여 빈도가 가장 높은 염색체 번호로 테스트 데이터를 분류하는 방식이다.

<표 1>은 분류정확도를 보인 것이며 10회 반복 실험으로 얻은 평균정확도이다.

표 1 염색체 자동분류 정확도 실험 결과표

	중심축 확장 전	중심축 확장 후
기본 1:1추출	67.9%	75.8%
Local Averaging	65.3%	81.1%
수직 프로파일	68.9%	81.9%

모든 방법에서 중심축 확장방법이 확장전보다 평균 12% 향상되었음을 확인할 수 있고, 중심축 수직평균 프로파일 추출방법이 81.9%로 가장 좋은 결과를 보였다.

5. 결론

염색체 핵형분류에 중요한 특징인 밴드패턴은 density profile로 추출되어 자동분류에 이용된다.
다양한 density profile 추출방법을 실험한 결과 염색체 중심축에 수직방향으로 나타나는 밴드를 표현한 수직평균 방법이 가장 우수한 성능을 보였으며 염색체의 끝단까지 확장하여 분류한 경우가 확장전보다 더 정확한 분류성능을 보였다.
본 연구에서는 염색체의 중심축 수직평균을 이용한 density profile 특징이 다른 방법으로 추출된 density profile보다 더 정확하게 밴드패턴을 표현하고 있음을 확인할 수 있었고, 염색체의 끝단에 포함된 밴드패턴정보가 중요하다는 사실을 확인할 수 있었다.

6. 감사의 글

1. 본 연구는 ㈜gendix의 실험데이터 지원으로 수행되었으며, 이에 감사합니다.(http://www.gendix.com),
2. 이 논문은 2009년도 정부(지식경제부) 지역전략기획 기술개발사업의 지원으로 수행되었음 (No.70007355).

7. 참고논문

[1] ISCN. An international system for human cytogenetic nomenclature. Karger, 1992.

[2] J. Piper, E. Granum, D. Rutovitz, H. Ruttledge,“ Automation of chromosome analysis, “ Signal Processing, vol. 2, pp. 203-221, 1980.

[3] Lerner, B., Levinstein, M., Rosenberg, B., Guterman, H., Dinstein, I., and Romem, Y., “Feature selection and chromsome classification using a multilayer perceptron neural network“, IEEE Internatinal Conference on Neural Networks, vol. II .2/7, pp.3540-3545, Jun. 28-Jul. 2, 1994.

[4] Jens Gregor and Erik Granum, “Finding chromosome centromeres using band pattern information,“ Comput. Biol. Med., vol. 21, No. 1/2, pp.55-67, 1991.

[5] Phil A. Errington and Jim Graham, “Application of Artificial Neural Networks to Chromosome Classification,“ Cytometry 14, pp627-639, 1993.

[6] K. Jau-hong, C. Jen-gui, W. Tsaipei, “Chromosome classification based on the band profile similarity along approximate medial axis“, Pattern Recognition 41, pp. 77-89, 2008.

[7] J.Piper. E.Granum, “On Fully Automatic Feature Measurement for Banded Chromosome Classification“, Alan R. Liss, Inc. Cytometry 10, pp. 242-255, 1989.