

염색체 영상의 재구성과 특징 파라메타 추출

장용훈, 이권순, 이영진, *전계록, **엄상희
동아대학교 전기공학과, *부산대학교 병원 의공학과, **부산대학교 전자공학과

Chromosome images Reconstitution and Feature Parameter Extraction

Y. H. Chang, K. S. Lee, Y. J. Lee, *K. R. Jun, **S. H. Eom

Dept. of Electrical Eng., Dong-A Univ.

*Dept. of Medical Eng., Pusan National Univ. Hospital

**Dept. of Electronic Eng., Pusan National Univ.

Abstract

In this paper, We propose an algorithm for reconstitution of chromosome images to extract its morphological feature parameters. It is reconstituted from 460 chromosome images using the 32 direction line algorithm. We extract three morphological feature parameters such as centromeric index, relative length ratio, and relative area ratio. The experiment results show that our method is better than that of other researchers comparing with the error of feature parameters.

I. 서 론

인간의 염색체 분석은 임상의학과 세포 유전학(cytogenetics)분야에서 여러 가지 선천성 질환(다운, 클라인펠터, 에드워드, 파타우 증후군 등)과 백혈병, 악성종양, 방사선 장애, 변위유발소 선량법(mutagen dosimeter)의 연구 그리고 생물학적 연구 등을 위하여 매우 중요하고 필요한 연구이다.

염색체 수는 1956년 Tijo와 Levan에 의해 46개로 밝혀졌고, 그후 1960년 국제 염색체 회의에서 염색체의 명명법과 분류법에 관하여 ISCN(An International System for Human Nomenclature)이라는 규약이 설정되었다. 이 규약에 의하면 정상인의 염색체 수는 22쌍의 상염색체(autosome)와 1쌍의 성염색체(sex chromosomes)로 구분되며 여성의 경우는 XX, 남성은 XY인 성염색체로 구성된다. 염색체의 분류는 형태학적 특징인 동원체(centromere)의 위치, 부수체(satellite)의 유무 및 상대 길이비 등에 의하여 분류되며, 염색체의 군(group)과 염색체를 번호순으로 나열한 도형을 핵형 분류도(karyogram)라고 한다[1].

염색체 핵형 분석의 자동화에 관한 연구는 1964년 Ledly에 의해 수행된 이후[2], 지난 30여년 간 많은 학자들에 의하여 수행되고 있으며, 근래에는

상용화된 시스템이 발표되어 사용되고 있지만 분류에의 정확도가 낮아 염색체 분류의 정확도를 개선하기 위한 연구가 계속 진행되고 있다[3, 4].

컴퓨터를 이용한 염색체의 자동분류는 형태학적인 특징 파라메타(morphological feature parameters)를 추출하여 패턴 분류기의 입력으로 사용하고 있다[5, 6]. 형태학적인 특징 파라메타는 동원체의 위치 검출, 동원체 지수(centromeric index : C.I), 상대 길이비(relative length : R.L), 상대 면적비(relative area : R.A) 등이며, 이는 세포 배양 과정에서의 불 균일, 3차원적인 영상의 2차원적인 검출 등으로 인하여 동일한 번호의 염색체에서도 같은 값의 특징 파라메타를 검출하기가 용이하지 않다[3]. 염색체의 형태학적인 특징 파라메타를 추출하기 위해 윤곽선[3], MAT (medial axis transform)[5, 6, 7] 및 밴드 패턴 정보[8] 등을 이용하여 특징 파라메타를 추출하는 연구들이 수행되었다. 이러한 연구들은 염색체의 중앙 축을 검출하여 중앙 축의 각 화소에서 수직선을 그어 염색체 폭을 결정하였다. 그러나 염색체는 형태적으로 굽어져 나타나므로 그 정도에 따라 추출된 특징 파라메타 값에 많은 오차가 발생한다. 따라서 본 연구에서는 중앙축의 각 화소에 등 각도로 32방향의 선을 연장하여, 이들 중 가장 짧은 선분을 결정하여 염색체를 재구성하는 알고리즘을 제안한다.

본 연구에서는 임상적으로 정상인으로 판명된 460개의 염색체를 사용하여, 중앙축에 수직선을 연장한 방법과 제안한 염색체를 재구성하는 방법으로 형태학적 특징 파라메타인 C.I, R.L 및 R.A를 구하였고, 이를 비교하였다. 이때 제안된 방식에 의하여 구한 특징 파라메타들의 오차가 감소되었음을 알 수 있었다.

II. 염색체의 형태학적 특징 파라메타

1. 동원체 지수(centromeric index : C.I.)

염색체 길이에 관한 특징 파라메타를 구하기 위해 먼저 동원체의 위치를 구한 후, 동원체를 중심

으로 염색체의 장완(long arm)과 단완(short arm)의 끝좌표를 구하여 두 좌표간의 유클리드 거리(euclidean distance)를 계산하여 그 길이를 개개의 염색체 길이로 정하였다. 이와같이 구한 염색체의 길이를 이용하여 식 (1)과 같이 C.I.를 계산할 수 있다.

$$C.I = \frac{\text{염색체 단완의 길이}}{\text{염색체의 길이}} = \frac{l_s}{l_c} \leq 0.5 \quad (1)$$

여기서, l_s 는 염색체 단완의 길이, l_c 는 염색체의 길이를 나타낸다.

2. 상대 길이비(relative length ratio : R.L.)

염색체는 형태학적으로 서로 다른 길이를 가지기 때문에 각 염색체의 길이를 비교하여 염색체를 분류할 수 있으므로 R.L.을 염색체의 분류 파라미터로 사용할 수 있다. 하나의 세포에서 추출된 46개의 염색체 길이를 구한 후, 44개의 상염색체들의 길이를 합한 총 길이로 각 염색체의 길이를 나누면 식 (2)와 같이 각 염색체의 R.L.을 구할 수 있다.

$$l_t = \text{상염색체의 총길이} = \sum_{c=1}^{44} l_c \quad (2)$$

$$R.L. = \frac{\text{염색체의 길이}}{\text{상염색체의 총길이}} = \frac{l_c}{l_t}$$

여기서, l_t 는 44개의 상염색체의 총 길이를 나타내며, R.L.은 각 염색체의 상대 길이비를 나타낸다.

3. 상대 면적비(relative area ratio : R.A.)

각 염색체의 면적도 서로 다른 값을 가지므로, 염색체의 면적을 분류 파라미터로 사용할 수 있다. 염색체 하나의 면적은 개개의 염색체 영상 정보에서 화소 수를 계수(Counting)하여 구할 수 있다. 따라서 식 (3)과 같이 염색체의 상대 면적비를 구할 수 있다.

$$S_t = \text{상염색체의 총면적} = \sum_{c=1}^{44} S_c \quad (3)$$

$$R.A. = \frac{\text{각염색체의 면적}}{\text{상염색체의 총면적}} = \frac{S_c}{S_t}$$

여기서, S_c 는 각 염색체의 면적, S_t 는 상염색체의 총면적, RA는 염색체의 상대 면적비를 나타낸다.

III. 염색체 영상의 전처리

1. 영상 획득 및 필터링

임상적으로 정상인으로 판명된 염색체(46개)를 현미경상(microscopic image)에서 촬영한 염색체 필름을 스캐너(AVR8800)를 사용하여 염색체 영상을 획득하여 파일로 저장하였다. 그림 1에서 나타난 것과 같이 저장된 염색체 영상은 반전된 농도치(negative value)의 정보를 가지므로, 전체 염색

체 영상 정보의 농도치를 반전 필터링(inverse filtering)하였으며, 결과는 그림 2와 같다.

그림 3은 배경과 염색체의 정보를 분리하기 위해 반전된 염색체 영상 정보의 히스토그램(histogram)을 구하여 문턱치(thresholding)처리를 수행한 결과를 나타낸다.

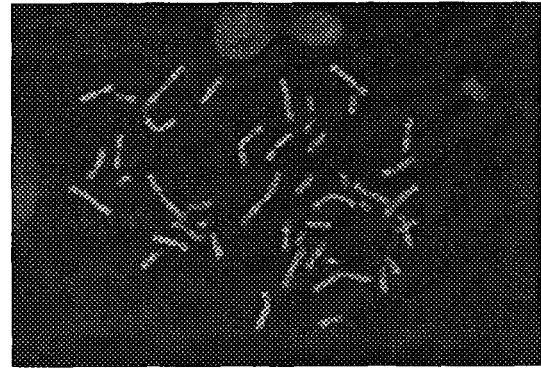


그림 1. 염색체 필름을 스캐너로 획득한 염색체 영상

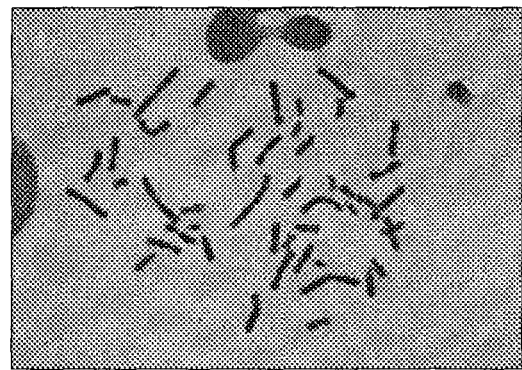


그림 2. 반전 필터링을 사용한 염색체 영상

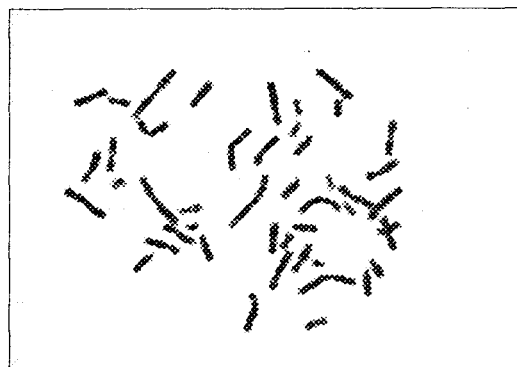


그림 3. 문턱치 처리를 수행한 염색체 영상

2. 중앙축에 수직선을 연장한 특징 파라미터 추출

전처리를 수행한 염색체 영상을 이치화하여 Zhang[9]등이 제안한 세선화 알고리즘(thinning algorithm)을 사용하여 염색체의 중앙축을 구하였고, 중앙축의 각 화소를 따라 수직선을 연장하면

염색체의 폭(width)을 구할 수 있다. 이와같이 구한 염색체의 폭들 중 가장 짧은 폭을 염색체의 동원체(centromere)로 설정하여 염색체의 형태학적 특징 파라메타(C.I., R.L., R.A.)를 구하였다. 그림 4는 이의 수행과정과 그 결과를 나타내었으며, 그림 4(a)는 이치화한 염색체 영상, 그림 4(b)는 세선화 기법을 수행한 결과, 그림 4(c)는 C.I., R.L. 및 R.A.를 구한 결과를 나타내었다.

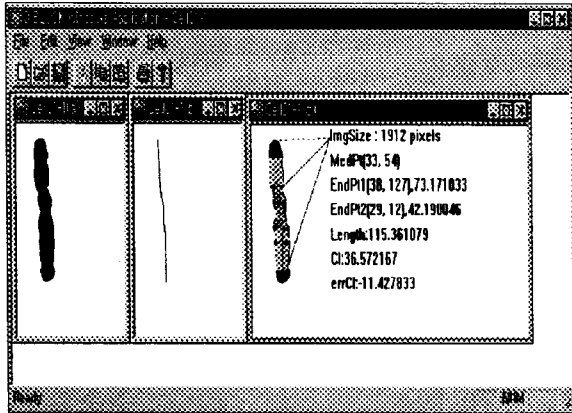


그림 4. 특징 파라메타 추출 과정

- (a) 이치화된 영상
- (b) 세선화 기법에 의한 결과
- (c) 특징 파라메타 추출 결과

이와 같이 개개의 염색체 영상을 추출하는 방법은 상당히 많은 시간이 소요되며 지속적인 운영자의 개입이 필요하다. 따라서 본 연구에서는 염색체 영상을 23개씩 두 화면으로 나누어 46개의 염색체 영상의 이치화 처리와 특징 파라메타 추출을 연속적으로 수행하도록 하였다.

3. 염색체의 재구성 기법

염색체는 형태적으로 굵어져 있으므로 특징 파라메타의 추출이 까다로우며, 동일한 번호의 염색체라도 같은 값이 구해지지 않는다. 따라서 염색체의 재구성은 안정된 특징 파라메타의 추출을 가능하게 하며, 염색체의 농도 정보도 사용할 수 있는 장점이 있다.

3.1. 재구성을 위한 선분의 결정

염색체의 재구성을 위한 선분의 결정을 위하여 먼저 2절에서 사용한 방법과 동일한 세선화 알고리즘(thinning algorithm)을 이용하여 염색체의 중앙축을 추출하였으며, 다음으로 중앙축의 각 화소에 32방향으로 선을 연장하여 염색체 영역의 가장자리까지 32방향의 누적된 길이를 구하였다. 이렇게 구한 32개의 길이들을 비교하여 이들 중에서 가장 짧은 선분의 좌표와 중앙축의 원점과의 유클리드 거리를 계산하고, 이와 π 라디안만큼 떨어진 선분의 합을 구하여 염색체를 재구성하기 위한 선분으로 결정하였다. 또한 이 선분에 있는 각 화소의 농도 정보를 합하여 선분에 있는 총화소 수로 나누어 평균 농도를 구하여 염색체를 재구성하였다. 이와 같은 알고리즘은 다음과 같이 수행된다.

- (a) 세선화 알고리즘을 수행하여 염색체의 중앙축을 구한다. 이때 염색체 양끝에서 잘린 부분(양끝의 약 10%씩)은 맨 끝점의 좌표와 다음 점의 좌표와의 각도를 이용하여 연장한다.
- (b) 설정된 염색체 중앙축의 화소(x_c, y_c)를 원점으로 선택한다.
- (c) 원점(x_c^k, y_c^k)로 하여 $\Delta\theta = 2\pi / M$ 의 등 각도로 M개의 길이를 계산한다. 여기서 k는 중앙축의 화소에 대한 알고리즘의 반복 수행 횟수이며, M은 32이다.
- (d) 길이 성분 i에 대하여 식(4)를 사용한 문턱치 에지 검출을 행하여 H_{edg} 값이 문턱치 H_{th} 보다 크면 길이 성분 i에 대한 수행을 종료하고 이의 좌표 l_i^r 을 저장한다.

$$H_{edg}(r^i) = \frac{g(r^i+1) + g(r^i) - g(r^i-1) - g(r^i-2)}{2} \quad (4)$$

여기서 r^i 는 길이 성분 i에 대한 방사방향의 중분값이며, $g(\)$ 는 염색체 각 화소의 농도 정보 값을 나타낸다.

- (e) M개의 방향에 대해 (d)를 반복 수행한다.
- (f) (e)가 종료되면 식(5)와 같이 원점의 좌표(x_c^k, y_c^k)와 저장된 좌표 l_i^r 의 유클리드 거리를 계산한다.

$$l_{ik} = \sqrt{(x_{k1} - x_{k2})^2 + (y_{k1} - y_{k2})^2} \quad (5)$$

- (g) (f)의 수행이 완료되면 식(6)과 같이 l_{ik} 와 π 간격만큼 떨어진 $l_{ik-\pi}$ 의 길이를 합하여 재구성에 대한 길이 l_i 를 구한다.

$$l_i = l_{ik} + l_{ik-\pi} \quad (6)$$

- (h) 길이 l_i 중에서 가장 짧은 길이 L_i 를 식(7)과 같이 구한다.

$$L_i = \text{Min}[\text{Sort}(l_i)] \quad (7)$$

- (i) 중앙축의 총화소 수 k개를 만족할 때까지 (d)에서 (h)까지 반복 수행한다.

이상과 같이 수행하면 염색체를 재구성하기 위한 전체 선분이 결정된다.

3.2. 재구성된 염색체 영상의 특징 파라메타 추출

염색체 영상의 재구성을 위하여 전절에서 설명한 알고리즘을 먼저 수행하여 염색체 재구성을 위한 선분을 결정한다. 결정된 선분 상에 있는 각 화소의 농도정보를 합하고, 그 선분상의 총화소 수로 나누면 평균 농도정보를 구할 수 있으며, 이 농도 정보를 사용하여 염색체를 재구성하였다.

재구성된 염색체 영상에서 양끝의 약 10%씩을 제외하고, 나머지 부분의 폭을 정렬(sort)하여 가장 짧은 폭을 동원체의 위치로 설정하며, 이를 중심으로 염색체의 장완, 단완의 길이 및 염색체의 길이를 구한다.

동원체의 위치와 염색체의 길이를 이용하면 염색체의 형태학적 특징 파라메타인 C.I. 와 R.L.을 구할 수 있다. 또한 재구성된 염색체 영상의 전화소 수를 계수하여 염색체의 면적 비인 R.A.를 구한다. 그림 5는 원염색체 영상, 재구성된 염색체 영상과 특징 파라메타를 나타낸다.

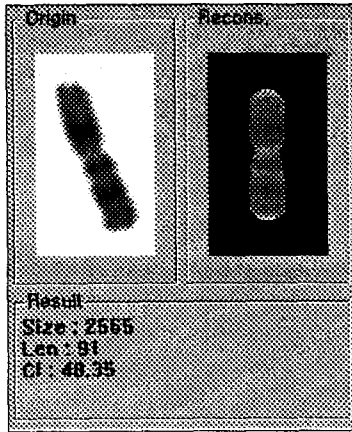
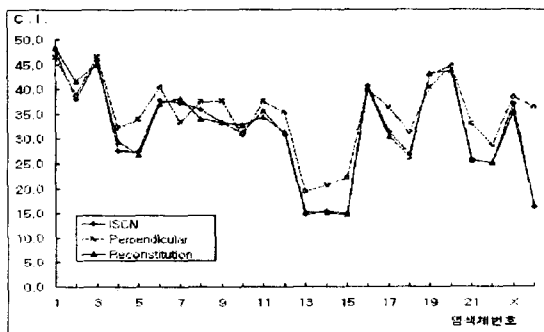


그림 5. 원염색체 영상과 재구성된 염색체영상

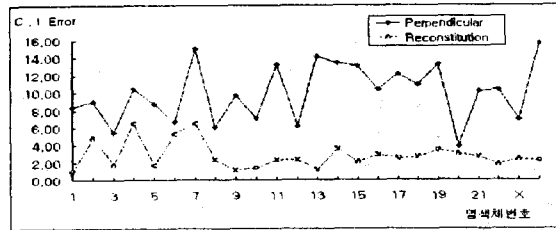
이와같은 방법으로 동일한 460개의 염색체를 사용하여 염색체의 중앙축에 수직선을 사용한 방법과 재구성된 염색체의 특징 파라메타를 추출한 결과를 정리하여 고찰하였다. 추출한 염색체의 특징 파라메타의 정확도를 알기 위하여 전체 평균값을 구하였고, 동원체 지수와 상대 길이비는 1986년 ISCN(An International Systems for Human Cytogenetic Nomenclature)에서 발표한 자료와 비교하였고, 상대적 면적비는 HC(Human Gytogenetics)의 데이터와 비교하였다[10, 11].

IV. 실험결과 및 고찰

염색체의 특징 파라메타는 염색체를 자동분류하기 위해 사용되는 여러 가지 파라메타중 가장 중요한 요소이며, 염색체 자동분류에 사용되는 여러 가지 인공지능 기법의 학습 데이터 및 입력 데이터로 이용되고 있다. 따라서 염색체의 특징 파라메타들의 안정된 값과 오차율을 줄이는 것이 무엇보다 중요하다. 본 연구에서는 임상적으로 정상인으로 판명된 460개의 염색체 영상으로 제안한 기법을 사용하여 염색체를 재구성하였으며, 염색체의 특징 파라메타 값을 더욱 안정하게 정량화하고 오차율을 감소시키기 위하여 재구성된 염색체 영상에서 특징 파라메타 C.I., R.L 및 R.A의 값을 추출하였다. ISCN과 HC에서 발표한 데이터와 동일한 염색체 영상으로 중앙축에 수직선을 연장한 방법과 비교하였으며, 그 결과는 그림 6, 그림 7 및 그림 8에 나타내었다.

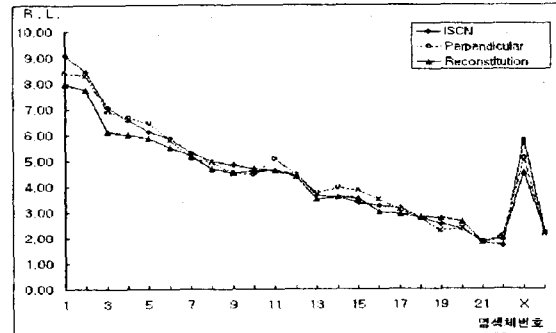


(a) 동원체 지수(C.I.)

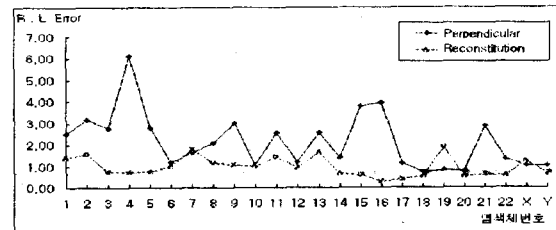


(b) 동원체 지수의 오차

그림 6. ISCN, Perpendicular data 및 염색체 재구성에 의해 추출된 C.I.의 비교

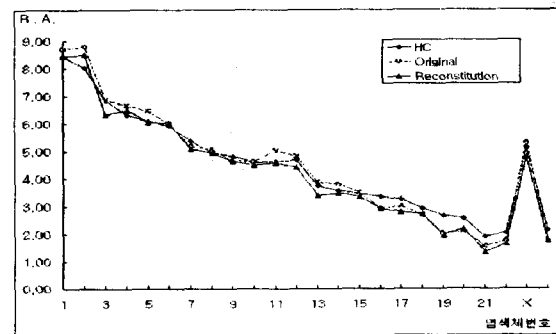


(a) 상대 길이비(R.L.)

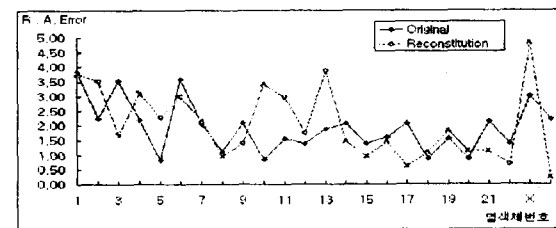


(b) 상대 길이비의 오차

그림 7. ISCN, Perpendicular data 및 염색체 재구성에 의해 추출된 R.L.의 비교



(a) 상대 면적비(R.A.)



(b) 상대 면적비의 오차

그림 8. HC, 염색체의 원 영상 및 32 염색체 재구성에 의해 추출된 R.A.의 비교

그림 6은 동원체 지수(C.I.)를 나타내며, 재구성된 염색체 영상에서의 동원체 지수값이 기존의 방법에 의한 결과보다 오차율이 더욱 감소되었음을 알 수 있고, 또한 특징 파라메타로 사용하기에 큰 차이가 나지 않을 만큼 안정화되었다. 그림 7은 상대 길이비(R.L.)이며 염색체의 길이가 큰 부분에서는 다소 차이가 나지만 길이가 작은 염색체에서는 안정된 값을 나타낸다. 염색체 자동 분류시 길이가 긴 염색체들은 쉽게 분류되는 반면, 길이가 짧은 작은 염색체들은 염색체 분류시 정확도 다소 낮다. 따라서 길이가 작은 염색체의 분류 정확도를 높일 수 있다. 또한 상대 길이비의 오차율이 상당히 감소되었음을 볼 수 있다. 그림 8은 상대 면적비(R.A.)로서 이의 결과는 두가지 방법이 거의 유사하다. 이는 염색체의 화소를 계수하여 구한 값으로 염색체를 재구성한 결과값이 원염색체 영상의 면적비와 유사함을 알 수 있다.

V. 결론

염색체 분석은 임상의학 분야, 세포 유전학분야 등에서 많은 연구가 진행되고 있다. 염색체의 핵형 분류도 작성시 현미경상에서 적절한 염색체 군의 탐색, 사진촬영 및 현상, 사진 상에서 각 염색체의 선택 등, 많은 수 작업의 수행으로 인하여 하나의 핵형 분류도 작성에 시간이 많이 소요되고 있는 실정이다. 이로 인하여 염색체를 자동 분류하는 시스템이 요구되어 왔으며 현재까지 많은 연구가 진행되고 있다. 염색체 자동 분류시 특징 파라메타는 여러 가지 인공 지능기법의 분류 파라메타로 사용되고 있어 오차율이 작고 안정한 파라메타의 추출이 요구된다. 본 연구에서 수행하였던 특징 파라메타의 결과 값은 C.I. 및 R.L.에서 오차율이 작고 안정한 값을 가짐을 알 수 있었다. R.A.는 염색체의 원영상에서 추출된 값과 거의 유사하여 안정된 결과 값을 나타낸다. 재구성된 염색체 영상의 농도정보를 이용하면 염색체 자동분류의 정확도를 더욱 높일 수 있으리라 생각되며, 또한 이 영상 정보를 사용하여 ISCN에서 염색체를 도식화하여 발표한 형태로 구성하면 염색체의 이상유무에 관한 진단도 가능하리라 생각된다.

References

1. Robert S. Ledly, "High-speed automatic analysis of biomedical picture," Science, vol. 146, pp. 216-223, 1964.
2. J. M. Cho, and D. H. Hong, "Computer-assisted karyotyping system of Giemsa-stained chromosomes (II)," Proc. of 1989 Korea-Japan Joint Conference on MBE, pp. 19-23, Sep. 21-22, 1989.
3. Lucas J. van Vliet, Ian T Young, and Brian H. Mayall, "The athena semi-automated karyotyping system," Cytometry, vol. 11, pp. 51-58, 1990.
4. Brian H. Mayall, James D. Tucker, Mari L. Christensen, Lucas J. van Vliet, and Ian T Young,

"Experience with the athena semi-automated karyotyping system," Cytometry vol. 11, pp. 59-72, 1990.

5. Lerner, B., Levinstein, M., Rosenberg, B., Guterman, H., Dinstein, I. and Romem, Y., "Feature selection and chromosome classification using a multilayer perceptron neural network," IEEE International Conference on Neural Networks, vol. II.2/7, pp. 3540-3545, Jun. 28-Jul. 2, 1994.
6. 조종만, 신경회로망을 이용한 염색체의 핵형분류에 관한 연구. 인하대학교 박사학위 논문, 1994.
7. J. Piper, E. Granum, D. Rutovitz, and H. Rutledge, "Automation of chromosome analysis," Signal Processing, vol. 2, pp. 203-221, 1980.
8. Jens Gregor and Erik Granum, "Finding chromosome centromeres using band pattern information," Comput. Biol. Med., vol. 21, No. 1/2, pp. 55-67, 1991.
9. Rafael C. Gonzalez and Paul Wintz, "Digital image processing," Reading. Mass., Addison-Wesley, pp. 398-403, 1987.
10. An International System for Human Cytogenetic Nomenclature(ISCN), KARGER, 1985.
11. D. E. Rooney, B. H. Czepulkowski, "Human Cytogenetics Essential Data," John Wiley & Sons, 1994.