

# 다단계 다층 인공 신경회로망을 이용한 염색체 핵형 분류

<sup>0</sup>장용훈, 이권순, 정형환, \*전계록  
(동아대학교 전기공학과, \*부산대학교 병원 의공학과)

## Chromosome Karyotype Classification using Multi-Step Multi-Layer Artificial Neural Network

<sup>0</sup>Yong Hoon Chang, Kwon Soon Lee, Hyeng Hwan Chong, \*Kye Rok Jun  
Dept. of Electrical Eng., Dong-A Univ.,  
\*Dept. of Medical Eng. Pusan National Univ. Hospital

### Abstract

In this paper, we proposed the multi-step multi-layer artificial neural network(MMANN) to classify the chromosome, Which is used as a chromosome pattern classifier after learning. We extracted three chromosome morphological feature parameters such as centromeric index, relative length ratio, and relative area ratio by means of preprocessing method from ten chromosome images.

The feature parameters of five chromosome images were used to learn neural network and the rest of them were used to classify the chromosome images. The experiment results show that the chromosome classification error is reduced much more, comparing with less feature parameters than that of the other researchers.

### I. 서 론

실세계에 존재하는 모든 생물체는 각각의 형질을 결정하는 유전자가 염색체 내에 존재한다. 즉, 생물체 중 특유의 특성을 유지 계승하게끔 생식세포를 통하여 유전적으로 전달된다. 생물의 종은 각 종에 따라 개체성이 존재하므로 각 염색체도 종에 따라 수와 형태에 있어서 개체성이 존재한다. 따라서 종이 변화하지 않으면 종이 가지고 있는 염색체의 수와 형태는 항상 일정하다.

세포유전학(cytogenetics) 분야에 있어서 유전 형질의 발현을 조절하는 유전자가 염색체 내에 존재하기 때문에 염색체에 대한 연구는 상당히 중요하다. 사람의 염색체 분석은 여러 가지 선천성 질환(다운, 플라인펠터, 에드워드, 파타우 증후군 등) 뿐만 아니라 백혈병, 악성 종양, 방사선 장애, 변위유발소 선량법(mutagen dosimeter)의 연구 그리고 생물학적인 연구 등을 위해 매우 중요하고 필요한 연구이다.

사람의 염색체의 수는 1956년 Tjio와 Levan에 의해 밝혀진 46개이며 그후 1960년 국제 염색체 회의에서 인간의 염색체의 명명 및 분류법에 관해 ISCN이라는 규약이 설정되었다. 이 규약에 의하면 정상인의 염색체 수는 22쌍의 상염색체(autosomes)와 1쌍의 성염색체(X, Y)로 구분되며, 여성의 경우는 XX, 남성의 경우는 XY인 성염색체로 구성된다. 그리고 염색체의 각 쌍은 동원체(centromere)의 위치, 부수체(satellite)의 유무, 상대 길이비 등에 의해 분류되며 이러한 염색체의 수 및 형태학적인 특징을 핵형(karyotype)이라 하며 염색체의 군과 염색체의 분류를 번호순으로 나열하여 테이블화한 도형을 핵형분류도(karyogram)라고 한다[1].

염색체의 핵형분류도의 작성과 해석에는 고도로 숙달된 세포 유전학 전문가가 필요하다. 즉, 세포 배양, 염색, 슬라이드 제작, 관찰에 적합한 염색체 군 선택 및 사진촬영, 가위에 의한 개개의 염색체 분리, 핵형분류도 작성과 염색체 해석이라는 매우 복잡하며 시간과 노력이 많이 필요로 한다. 따라서 컴퓨터를 이용하여 염색체의 핵형을 분류하는 방법이 연구되었다.

염색체 핵형 분석의 자동화에 관한 연구는 1964년 Ledly[2]에 의해 처음 수행된 이후, 지난 30년간 많은 학자들에 의해 수행되었으며 근래에는 상용화된 시스템이 발표되어 사용되고 있다. 그러나 실제 임상에서 사용하기에는 정확도가 낮으며 염색체 자동 분류 과정에 운용자의 개입 등으로 처리 시간의 손실 등 아직 개선하여야 할 부분이 많이 있다[3, 4].

본 연구에서는 영상 패턴인식을 위해 최근 많이 연구되고 있는 인공신경망을 이용하여 다단계 다층 인공신경회로망(multi-step multi-layer artificial neural network : MMANN)을 구성하고 염색체의 핵형 분류기로 사용하였다.

MMANN의 구성은 다량의 입력패턴에 의한 학습 소요시간을 최소화하며, 염색체를 단계적으로 분류하기 위해 두 단계의 신경회로망으로 분리하였다. 첫번째 단계의 신경회로망(MMANN1)은 염색체들을 7개 군(group) 즉, A군(염색체 1-3), B군(염색체 4-5), C군(염색체 6-12, X), D군(염색체 13-15), E군(염색체 16-18), F군(염색체 19-20) 및 G군(염색체 21-22, Y)으로 분류하기 위하여 구성하였고, 두번째 단계의 신경회로망(MMANN2)은 7개 군으로 분류된 염색체를 각각의 염색체로 분류하기 위해 구성된 신경회로망이다.

본 연구에서 제안한 MMANN을 이용하여 염색체의 핵형 분류를 수행한 방법은 다음과 같다. 임상적으로 정상으로 판명된 정상인 10명의 염색체를 수집하고 10명중 5명의 염색체 사진을 영상 전처리(image preprocessing)하여 각 염색체의 동원체 지수, 상대 길이비 및 상대 면적비와 같은 특징 파라메타를 구하고, 이를 입력패턴으로 하여 MMANN을 학습시킨다. 학습이 끝난 MMANN을 패턴 분류기(pattern classifier)로 사용하여, 다른 5명에 대한 염색체를 분류하였다.

### II. 전처리 및 특징 파라메타 추출

#### 1. 영상획득

염색체의 형태학적 구조가 임상적으로 정상으로 판명된 정상인 10명의 염색체 사진을 스캐너(AVR 8800)를 사용하여 염색체 영상을 획득하고 이를 PC(pentium)에 파일로 저장하였다. 이들 중 5명분에 대한 염색체 영상 데이터는 전처리 기법을 수행하여 MMANN을 학습하기 위한 입력패턴으로 사용하였으며, 나머지 5명분의 영상 데이터는 각 염색체를 분류하기 위한 분류용 입력 데이터로 사용하였다.

#### 2. 동원체 검출

염색체 영상의 전체 농도를 나타내는 히스토그램을 구하여 문턱치(thresholding) 처리를 수행하여 2치화된 영상으로 배경과 염색체를 구분하였다. 2치화된 염색체 영상을 지시장치를 사용하여 각 염색체 영상의 서브화일을 구성하여 저장하였다. 염색체의 동원체를 검출하기 위해 세선화 기법(thinning)을 사용하여 염색체의 중앙축(medial axis)을 구하고 중앙축을 형성하는 각 픽셀에 16 방사각( $2\pi/16$  radian) 방향으로 거리를 계산하여 각 염색체의 폭을 구하였다. 이렇게 구하여진 염색체 폭들 중 가장 짧은 폭을 동원체의 위치로 설정하였다.

3. 염색체 특정 파라메타 추출

3.1 동원체 지수(centromeric index : CI)

염색체 길이에 관한 특정 파라메타를 구하기 위해 동원체의 위치를 구한 후, 동원체를 중심으로 염색체의 장완과 단완의 끝 좌표를 구하여 두 좌표간의 유클리디안 거리(euclidean distance)를 계산하면 각 염색체의 길이를 구할 수 있다. 염색체 길이를 이용하면 다음 식에 의해 CI를 구할 수 있다.

$$CI = \frac{\text{염색체단완의길이}}{\text{염색체전체길이의합}} = \frac{l_s}{l_c} \leq 0.5 \quad (1)$$

3.2 상대 길이비(relative length ratio : RL)

염색체의 형태학적 구조에 의하면 각 염색체의 길이는 다르기 때문에 각 염색체의 길이를 비교하여 염색체를 분류할 수 있으므로 RL을 염색체 분류 파라메타로 사용할 수 있다. 염색체의 RL은 하나의 세포(cell)에서 추출된 46개의 염색체 길이를 모두 구한 후 이들을 합한 총 길이로 각 염색체의 길이를 나누면 염색체의 RL을 구할 수 있으며, 염색체의 RL은 다음과 같은 식으로 구할 수 있다.

$$l_t = \sum_{c=0}^{45} l_c \quad (2)$$

$$RL = \frac{l_c}{l_t}$$

여기서  $l_t$ 는 46개 염색체의 총길이,  $l_c$ 는 각 염색체의 길이, RL은 각 염색체의 상대 길이 비를 나타낸다.

3.3 상대 면적비(relative area ratio : RA)

염색체의 형태학적 구조에 의하면 각 염색체의 면적은 다르기 때문에 각 염색체의 면적은 염색체를 분류하는 또 다른 특징 파라메타로 사용할 수 있다. 각 염색체의 면적 계산은 원 영상에서 각 염색체의 샘플을 추출하여 그 화소를 제수하므로써 가능하다. 염색체의 RA는 다음과 같은 식으로 구할 수 있다.

$$RA = \frac{\text{각염색체의면적}}{\text{염색체의전체면적의합}} = \frac{S_c}{S_t} \quad (3)$$

염색체 영상의 전처리 수행 과정과 특정 파라메타(CI, RL, RA)를 구하는 과정과 추출 결과는 그림 1과 같다. 그림 1(a)는 2차화된 염색체 영상, 그림 1(b)는 세선화 기법을 수행한 결과, 그리고 그림 1(c)는 CI, RL 및 RA를 구한 결과를 나타내었다.

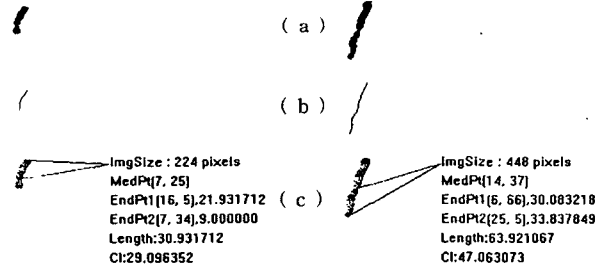


Fig. 1. Chromosome preprocessing and result  
(a) 이치화된 영상 (b) 세선화 기법 (c) 추출된 결과

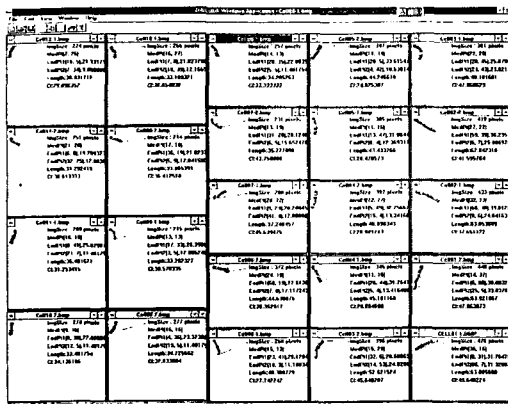


Fig. 2. Extraction of chromosome feature

하나의 염색체 군을 형성하는 46개의 각 염색체를 창(window)으로 구분한 후 그림 1과 같은 방법을 적용하면 영상의 전처리와 염색체 특정 파라메타를 구할 수 있다. 염색체 영상을 23개씩 두 화면으로 나누어 46개 염색체 영상의 전처리와 특정 파라메타 추출을 연속적으로 수행하게 하였다. 그림 2는 한 화면의 수행 결과를 나타내었다. 이렇게 구한 염색체의 특정 파라메타는 패턴 분류기의 입력패턴으로 사용된다.

그림 1과 그림 2의 방법에 의해 영상의 전처리를 수행하고 염색체의 특정 파라메타를 추출한 결과를 정리하면 표 1과 같다. 추출된 염색체 파라메타들의 정확도를 알기 위하여 전체 평균값을 구하였고, 동원체 지수와 상대 길이비는 1986년 ISCN(An International Systems for Human Cytogenetic Nomenclature)에서 발표된 자료와 비교하였고 상대적 면적비는 HC(Human Cytogenetics)의 데이터와 비교하였다[5,6].

Table 1. Centromeric index, relative length and relative area ratio

염색체 번호	CI			RL(%)			RA(%)		
	ISCN	추출	오차	ISCN	추출	오차	HC	추출	오차
1	48.0	48.40	±0.55	9.08	9.23	±0.28	8.44	8.94	±0.16
2	38.1	39.89	±0.38	8.45	8.23	±0.29	8.02	7.79	±0.34
3	45.9	47.51	±0.93	7.06	7.02	±0.22	6.63	7.03	±0.28
4	27.6	27.99	±0.67	6.55	6.44	±0.23	6.30	6.81	±0.32
5	27.4	27.59	±0.74	6.13	5.99	±0.20	6.08	6.08	±0.38
6	37.7	35.57	±1.19	5.84	5.97	±0.19	5.90	5.87	±0.29
7	37.3	37.86	±0.29	5.28	5.24	±0.18	5.36	5.38	±0.34
8	35.9	32.51	±1.34	4.96	4.76	±0.19	4.93	5.01	±0.29
9	33.3	32.31	±0.84	4.83	4.41	±0.15	4.80	4.77	±0.31
10	31.2	33.77	±0.81	4.68	4.67	±0.17	4.59	4.85	±0.25
11	35.6	37.68	±1.37	4.63	4.63	±0.16	4.61	4.83	±0.37
12	30.9	29.43	±0.29	4.46	4.59	±0.15	4.66	4.46	±0.45
13	14.8	14.77	±0.45	3.64	3.76	±0.14	3.74	3.85	±0.37
14	15.5	13.24	±0.67	3.95	3.64	±0.13	3.56	3.57	±0.29
15	14.9	14.82	±0.48	3.36	3.52	±0.12	3.46	3.38	±0.33
16	40.6	38.11	±0.54	3.23	2.85	±0.11	3.36	2.86	±0.19
17	31.4	30.06	±1.11	3.15	2.99	±0.09	3.25	2.86	±0.23
18	26.1	27.30	±0.97	2.76	2.84	±0.08	2.93	2.87	±0.32
19	42.9	43.50	±0.38	2.52	2.37	±0.09	2.67	2.32	±0.18
20	44.6	43.61	±0.75	2.33	2.31	±0.07	2.56	2.52	±0.22
21	25.7	31.99	±0.47	1.83	2.06	±0.08	1.90	1.82	±0.17
22	25.0	33.22	±0.76	1.68	2.14	±0.07	2.04	2.25	±0.24
X	36.9	38.24	±0.59	5.80	5.31	±0.26	5.12	5.63	±0.32
Y	16.3	16.07	±0.67	1.96	2.22	±0.13	2.15	1.90	±0.31

III. 염색체 분류를 위한 신경회로망

1. 염색체 분류를 위한 신경회로망 구현

염색체 분류에 적합한 신경회로망을 구성하기 위해 전처리 과정을 통하여 추출한 3개의 특정 파라메타를 패턴입력으로 하여 MMANN을 학습시켰다.

MMANN의 구성은 3개의 특정 파라메타를 입력패턴으로 사용하여 학습 소요시간을 최소화하며, 염색체를 단계적으로 분류하기 위해 두 단계의 신경회로망으로 분리하였다. 첫 번째 신경회로망(MMANN1)은 염색체를 7개 군으로 분류하기 위해 구성되었으며, 두 번째 신경회로망(MMANN2)은 7개 군으로 나뉘어진 염색체를 각각의 염색체로 분류하기 위해 구성되었다.

염색체의 특정 파라메타는 신경회로망(MMANN1)의 입력 벡터로 사용되며, 이 경우 해당하는 염색체 군의 출력은 1.0으로 두고 나머지 군에 해당하는 출력은 모두 0.0으로 하여 교차신호를 구성하였다. MMANN1에 의해 7개 군으로 분류된 개개의 염색체들은 다시 MMANN2를 이용하여 24개의 염색체 각각을 분류한다. 이 경우 7개 군에 대한 교차신호는 해당 염색체의 출력은 1.0으로 두고 나머지 출력은 모두 0.0으로 한다. 제안된 MMANN에 사용되는 7개의 염색체 군과 염색체 각각의 번호는 표 2와 같다.

Table 2. Chromosome group and number

염색체 군	염색체 번호	염색체 군	염색체 번호
Group A	1번 ~ 3번	Group E	16번 ~ 18번
Group B	4번 ~ 5번	Group F	19번 ~ 20번
Group C	6번 ~ 12번, X	Group G	21번 ~ 22번, Y
Group D	13번 ~ 15번		

2. 다단계 다층 인공 신경회로망 (multi-step multi-layer artificial neural network)

앞 단계에서 추출된 염색체의 특정 파라메타를 입력 패턴으로 하여 염색체를 분류할 수 있는 신경회로망은 그림 3과 같다.

그림 3은 MMANN1의 구성도이며 입력층은 입력 벡터로 사용되는 염색체의 특정 벡터의 수와 같은 3개의 뉴런으로 구성되며, 출력층에는 염색체를 7개의 군으로 분류하기 위하여 7개의 뉴런으로 구성된다.

MMANN1의 출력에 의해 구분된 7개의 군 각각에 염색체의 특정 파라메타를 입력하여 각 염색체 군에 속하는 염색체를 분류하는 MMANN2의 구성은 그림 4와 같다. 그림 4의 입력층은 7개의 뉴런으로 구성되며 입력층의 각 뉴런에 염색체의 특정 파라메타가 입력되며, 출력층은 24개의 염색체를 분류하기 위하여 7개의 신경회로망을 두고 각 신경회로망의 출력은 표 2에 정의된 바와 같이 각 군별로 구분되는 염색체의 수 만큼 구성하였다. 이렇게 구성된 MMANN은 오차 역전파 학습 알고리즘을 사용하여 학습을 하며, 학습을 마친 후 각각의 가중치를 저장한다. 학습이 끝난 MMANN에 임의의 염색체 특정벡터 패턴이 입력으로 들어오면 그에 해당하는 염색체 번호를 분류해 낸다.

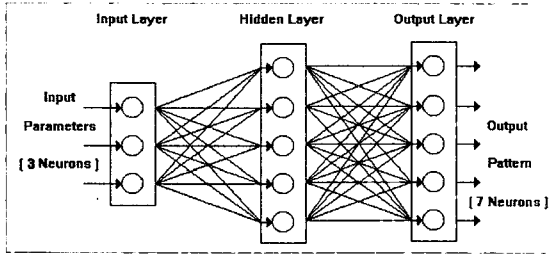


Fig. 3. MMANN1 for chromosome classification

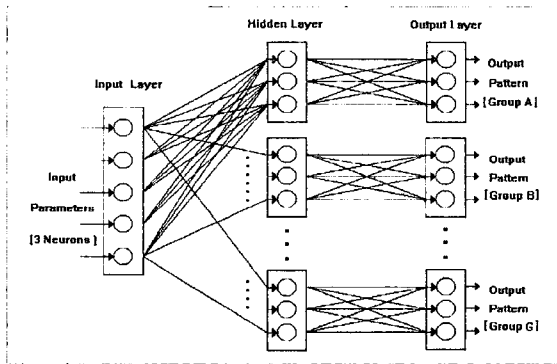


Fig 4. MMANN2 for chromosome classification

이상과 같이 구성된 MMANN은 가중치의 초기값, 활성화함수의 기울기(steeptness of activation function), 학습 상수(learning constant), 중간층 뉴런의 갯수 등이 학습에 큰 영향을 미치게 된다[7]. 즉, 염색체의 특정벡터를 입력으로 하여 학습되는 신경회로망의 최적 상태를 정확히 정의할 수 없으므로 다양한 형태의 실험이 요구된다.

본 연구에서는 역전파 학습 알고리즘의 학습요소들의 적절한 값과 중간층 뉴런의 적합한 수를 찾기 위해 여러가지 실험을 수행한 결과 표 3과 같은 파라메타를 얻을 수 있었다. 또한 학습 시간을 단축하고 신경회로망이 국부적인 극소점(local minimum)에 빠지지 않게 하기 위하여 모멘트법[8]과 적응 학습법[9]을 사용하였다.

Table 3. Parameters of neural network

신경회로망	NN1	Group A	Group B	Group C	Group D	Group E	Group F	Group G
활성함수의 기울기	1.0	1.2	1.1	1.0	0.9	0.8	1.0	1.0
학습 상수	0.05	0.05	0.01	0.05	0.1	0.5	0.05	0.1
중간층 뉴런의 수	30	10	10	30	15	15	10	10

#### IV. 실험결과 및 고찰

염색체의 핵형을 분류하기 위해 구성된 MMANN은 C 언어를 사용하여 구현되었으며, 이를 Pentium 90 컴퓨터에서 시뮬레이션 하였다. 구현된 MMANN은 각 층의 뉴런수, 활성화함수의 기울기, 학습 상수, 가속 상수, 입력 패턴 수, 학습 횟수 및 최대 허용 오차 등의 파라메타들을 키보드를 통하여 입력하므로써 신경회로망의 학습에 유연하게 대처할 수 있도록 설계하였고, 또한 학습되는 과정을 가시화하기 위하여 오차의 변화를 그래픽 처리하였으며 그리고 추출된 결과는 표로 구성하여 화일로 저장하였다.

그림 5는 A군의 염색체 특정 파라메타들을 학습하는 과정에서 자소오차 합이 변화하는 과정을 나타낸 것이다. 학습에 소요된 시간은 MMANN1이 2시간, MMANN2는 2시간 35분이 소요되었다.

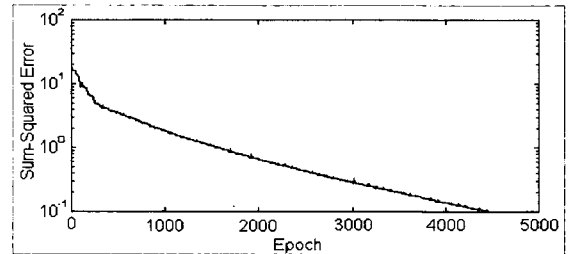


Fig. 5. Variation sum-squared error in NN learning

이상과 같이 학습된 신경회로망을 이용하여 임상적으로 염색체가 정상으로 판명된 5명의 염색체 특정 파라메타를 입력하여 염색체의 핵형 분류를 수행하였다.

1명의 염색체를 MMANN 패턴 분류기를 이용하여 염색체를 분류한 결과는 표 4와 같이 나타내었다. 학습에 사용된 염색체가 아닌 새로운 정상인 5명의 염색체를 분류한 결과 염색체 군의 오분류는 6개, 전체 염색체의 오분류는 21개가 나타났다. 이 결과를 정확도로 나타내면 염색체 군의 분류오차는 약 2.61%이며, 전체 염색체의 분류 오차는 9.13%로 나타났다.

#### V. 결론

패턴인식 분야에 많이 사용되고 있는 인공 신경회로망을 염색체의 영상을 구분하고 염색체의 핵형 분류에 적합한 MMANN 구성에 연구의 중점을 두었다. 영상 전처리 기법을 사용하여 염색체의 형태적인 3개의 특정 파라메타를 연속적으로 추출할 수 있도록 하여 10명에 대한 특정 파라메타를 추출하였다.

Table 4. Result of normal human chromosome classification for one person using MMANN

NO.	Sample 1-1		Sample 1-2		NO.	Sample 1-1		Sample 1-2	
	NO.	Group	NO.	Group		NO.	Group	NO.	Group
1	1	A	1	A	13	13	D	13	D
2	2	A	2	A	14	14	D	14	D
3	3	A	3	A	15	15	D	15	D
4	4	B	4	B	16	16	E	16	E
5	4	B	5	B	17	17	E	16	E
6	6	C	6	C	18	18	E	18	E
7	7	C	7	C	19	19	F	19	G
8	8	C	10	C	20	20	F	20	F
9	9	C	9	C	21	21	G	21	G
10	10	C	11	C	22	22	G	22	G
11	11	C	8	C	X	X	C		
12	12	C	12	C	Y	Y	G		

본 연구에서 구성한 신경회로망을 학습시키기 위하여 5명분의 특정 파라메타를 입력패턴으로 사용하였다. 학습된 신경회로망을 패턴 분류기로 사용하여 다른 5명분의 염색체를 분류하였다. 염색체 군의 분류 오차는 2.61%이며, 전체 염색체의 분류 오차는

9.13%를 나타내었다. 이 결과는 현재까지 발표된 염색체 패턴 분류기의 분류 오차보다 약간 높은 수치이다. 이는 신경회로망의 학습에 사용된 파라메타의 수가 적고, 염색체 영상을 전처리하는 과정에서 특징 파라메타의 추출이 용이하지 않은 몇몇 염색체에 의한 것으로 생각된다. 따라서 분류 오차를 감소시키기 위하여 다양한 종류의 특징 파라메타의 추출이 필요하며, 염색체 원영상의 추출과 전처리 과정에서 특징 파라메타 추출의 정확도를 높여야 할 것이다.

여러가지 특징 파라메타를 사용하여 본 연구에서 구성한 MMANN을 학습시키고, 이를 패턴 분류기로 사용하면 염색체의 분류 오차를 상당히 줄일 수 있을 것이며, 염색체 자동 분류 시스템 개발 시 염색체를 자동으로 분류할 수 있는 분류기로 유용하게 사용할 수 있을리라 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 athens 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 athens semi-automated karyotyping system," Cytometry, vol. 11, pp. 59-72, 1990.
5. An International System for Human Cytogenetic Nomenclature(ISCN), KARGER, 1985
6. D. E. Rooney, B. H. Czepulkowski, "Human Cytogenetics Essential Data," John Wiley & Sons, 1994.
7. 김명원, 방승양, 이수영, 이종호, 전홍관, 정호선, 정홍, "알기 쉬운 신경망 컴퓨터," 전자 신문사, 1992.
8. 박민용, 최항식, "뉴로 컴퓨터," 대영사, 1991.
9. B. Widrow, and R. Winter, "Neural Nets for Adaptive Filtering and Adaptive Pattern Recognition," IEEE Computer, pp. 25-39, March, 1988.
10. 조종만, 신경회로망을 이용한 염색체의 핵형분류에 관한 연구. 인하대학교 박사학위 논문, 1994.