바이오칩을 이용한 간암진단 예측 시스템

이형근* · 김충원* · 이준* · 김성천**

*조선대학교 · **제노프라(주)

Liver cancer Prediction System using Biochip

Hyoungkeun Lee* · Choongwon Kim* · Joon Lee* · Sungchun Kim**

*Chosun University · **GenoProt Inc.

E-mail : afocus@naver.com · cwkim@chosun.ac.kr · jlee@chosun.ac.kr · kimgp@naver.com

요 약

우리나라 암 발생빈도 중 간암은 위암에 이어 두 번째로 흔한 암으로, 초기에는 특이 증상이나 증후 없이 서서히 진행되는 경우가 많아 증상이 생긴 후 간암으로 진단될 경우, 대부분 마땅한 치료방법이 별로 없어 어떠한 치료를 해도 환자의 예후는 불량하나, 조기에 발견될 경우는 치료성적이 우수하여 조기 발견이 대단히 중요시된다. 본 시스템은 간암의 조기발견을 위한 시스템으로, 간암으로 확진된 환자 와 간암이외의 대조군의 혈액을 바이오칩에 반응시켜 바이오칩 프로파일을 기계학습을 통해 분류하는 시스템이다. 본 논문에서는 총 50샘플로 구성된 간암환자 와 100샘플로 구성된 간암이외의 대조군의 혈액시료를 1149의 서로 다른 올리고로 구성된 바이오칩에 반응시켜 획득한 데이터를 인공신경망을 통해 분석한 결과 92~96%의 분류 성능을 보였다.

ABSTRACT

The liver cancer in our country cancerous occurrence frequency to be the gastric cancer in the common cancer, to initially at second unique condition or symptom after the case which is slowly advanced without gets condition many the case which will be diagnosed in the liver cancer, most there was not a reasonable treatment method especially and if what kind of its treated and convalescence of the patient non quantity one, the case which will be discovered in early rising the treatment record was considered seriously about under the early detection. The system which it sees with the system for the early detection of the liver cancer reacts the blood of the control group other than the patient who is confirmed as the liver cancer and the liver cancer to the bio chip and bio chip Profiles mechanical studying leads and it is a system which it classifies. 1149 each other it reacted blood samples of the control group other than the liver cancer patient who is composed of the total 50 samples and the liver cancer which is composed of 100 samples to the bio chip which is composed with different oligo from the present paper and it was a data which it makes acquire worker the neural network it led and it analyzes the classification efficiency of the result 92~96% which it was visible.

키워드

바이오칩, 간암, 기계학습, 마이크로어레이, 압타머

Ⅰ. 서 론

1.1 연구의 배경

우리나라 암 발생빈도 중 간암은 전체의 11.6%로 위암에 이어 두 번째로 흔한 암으로 간암은 특히 55세를 전후하여 많이 나타나 활동적인 사회생활과 가정생활에 많은 피해를 주고 있다. 간암 초기에는 특히 증상이나 증후 없이 서서히 진행되는 경우가 많아 증상이 생긴 후 간암으로 진단될 경우, 대부분 마땅한 치료 방법이 별로 없어 어떠한 치료를 해도 환자의 예후는 불량하거나, 조기에 발견된 경우는 치료성적이 우수하다. 따라

서 간암을 조기에 발견할 수 있는 기술 개발의 필요성이 아주 높다. 간암진단을 위한 본 시스템의 두 가지 핵심 요소는 인공지능/기계학습(machine learning)을 이용한 질병진단 및 압타머 기반 진단이다. 의료분야에서 기계학습을 이용한 진단은 지속적인 응용 목표 분야로 자리 잡고 있다.[5-6] 의료분야에서 큰 흐름 중 하나인 증거기반의료(EBM : evidence-based medicine)를 구체화하는 자연스러운 접근 방법으로서, 기계학습은 축적된 환자 데이터의 복잡한 패턴을 컴퓨터를 통해 학습하여 의료적 결정(질병진단, 처방)의 효율성을 높이고 오류를 줄이게 된다.[2] 인공지능/기계학습을 이용한 질병진단의 과정은 그림 1과 같이 요약할 수 있다. 질병관련 축적된 데이터로부터 기계학습 기법을 이용하여 정상인과 환자, 또는 질병의 단계를 구분할 수 있는 최적의 패턴을 학습하여 진단 엔진을 구축한 후, 이 엔진을 장착한 컴퓨터 시스템을 이용하여 환자여부를 판별하거나 환자의 질병정도를 1차적으로 빠르게 검사할 수 있다. 본 시스템에서는 질병진단을 위한 핵심 엔진구현에 인공신경망을 적용하였다. 진단의 기반이 되는 데이터는 성공적인 질병진단을 위한 필수요소이다. 본 시스템에 사용하는 데이터는 압타머를 이용한 바이오칩으로 최근에 압타머를 이용한 기술이 생체시료의 발현측정에 활용되기 시작했다. 압타머바이오칩은 혈액내에 혈청에 포함된 특정 단백질의 상대적 양을 직접 측정 할 수 있는 바이오칩으로 질병진단 등의 의학적 응용에 활용될 수 도 있으며 기존의 마이크로어레이 분석기법을 그대로 적용할 수 있다는 장점을 가진다.[2] 압타머를 이용한 혈액내의 특정 단백질 양을 측정하는 압타머바이오칩은 유전자를 이용한 기존의 마이크로어레이분석과는 달리 생리적 작용에 보다 가까운 위치에서 영향을 미치는 단백질을 측정함으로써 보다 정확한 인과관계 추정이 가능할 것으로 여겨지고 있다. 더욱이, 최근 질병 진단 방법에 있어서 기존의 임상진단 방법에만 의존하던 것과는 달리 CDSS(Clinical Decision Support System)와 같은 전문 진단환경 시스템을 이용하는 등의 다양한 방향으로 변화가 이루어지고 있다. 이에 본 논문에서는 압타머바이오칩 데이터 및 이러한 진단환경시스템에서 인공신경망을 이용하여 간암을 예측하고 진단보조를 위한 시스템을 구축 하였다.

1.2 관련연구

컴퓨터 기술의 발달로 기계학습 기법을 통한 의료에의 접목은 지속적으로 이루어지고 있으며, 특히 마이크로어레이를 이용한 고속병렬처리 기술과 기계학습의 접목은 지속적인 관심대상이다.[1] CDSS가 의료의 중요한 보조수단으로 대두되고 있으며, CAD(Computer Aided Diagnosis) 등 과거 정보통신 기술이 진료보조의 수단에 머물던 것이 진료에 직접 응용되고 있다. 일반적으로 암을 비롯한 질병은 DNA가 RNA를 통해 단백질로 발현한다고 알려져 있으며, 발현된 단백질을 정확이 확인할 수 있다면 질병을 쉽게 진단할 수 있으며 이를 질병 특이적 바이오마커라 한다. 그러나 분자레벨에서의 단백질의 종류를 알아내는 것은 대단히 어려운 일이며 간암에 관련된 마커로는 AFP가 임상에서 사용되고 있다. 간암에 관련된 마커로 최근 Cystatin B가 보고되고 있으며 AFP 와 보완하여 사용하면 유용할 것으로 예상된다..[3] 그래서 대두된 접근 방법이 단백질의 종류를 정확이 알아내는 목표에서 정확이 종류는 알 수 없지만 단백질의 분포패턴의 해석을 통해 질병을 알아내기 위한 연구가 진행 되고 있다. 최근 바이오칩의 프로파일 특성을 이용하여 유방암의 재발가능성 여부를 예측하는 시스템의 미국 FDA로부터 허가를 얻어 향후 특정 마커중심의 진단에서 여러 요소의 패턴의 분석을 통한 기계적 학습에 의한 진단이 임상에 많이 응용 되고 있다.[4][7] 본연구의 선행연구로 압타머바이오칩을 이용한 심혈관질환의 단계별 예측 시스템이 연구 발표되었다.[1][5][6]

II. 바이오칩 진단 시스템

2.1 구성

본 시스템의 구성은 1. 간암환자의 데이터와 간암 이외의 대조군 데이터를 입력 받는 부분, 2. 입력된 데이터를 이용하여 전처리를 수행하는 부분, 3. 전 처리된 결과를 가지고 학습을 수행하여 간암환자의 모델을 생성하는 과정, 4. 생성도리 모델을 로딩 하여 unknown 환자에 적용하여 진단을 수행하는 부분(Blind Test), 5. 교차검정을 통해 시스템의 성능을 평가하는 부분으로 구성된다.

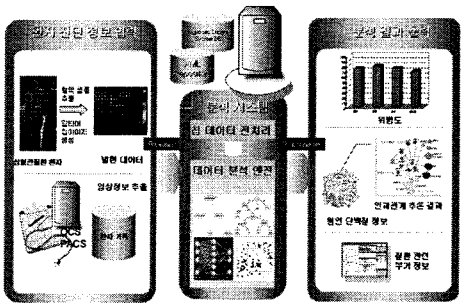

그림 1. 시스템 구성도

본 시스템의 구성 및 실무에서의 응용은 2가지로 나누어서 구성할 수 있는데 진단만을 위한 stand alone 형태의 진단 시스템과 기존의 병원정보시스템 즉. OCS, PACS, LIS 등과 연계된 통합된 시스템 구성을 위해 설계를 반영 하였다. 이를 위해, HL7, 및 DICOM 규약에 의거 시스템을 구성하였다.

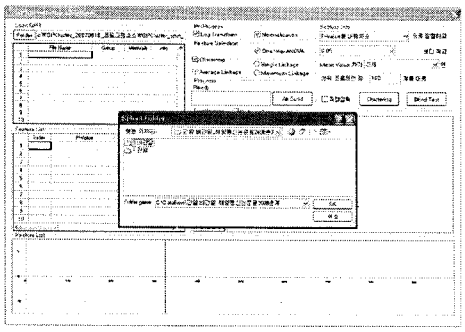

그림 2. 시스템 초기 화면

III. 인공신경망을 이용한 간암 진단

3.1 인공신경망

일반적인 신경망 모델은 그림 3과 같으며 각각의 원은 하나의 artificial neuron을 의미하고, 화살표는 앞 단계에서의 출력이 다음단계의 입력으로 연결되어 있음을 나타낸다. 각각의 연결 화살표는 weight를 가지고 있으며 이 weight를 조절하는 것이 학습을 하는 과정이 된다. Classification문제에서 분류할 class의 개수가 많은 경우에는 그 개수만큼 output layer에 neuron을 둘 수 있다.

3.2 인공신경망을 이용한 바이오칩 분석[5][6]

본 시스템에서는 정상(normal), 간암환자 이렇게 두 가지 중에서 데이터를 분류해야 하므로 output neuron이 2개가 존재한다. input neuron은 60-100개의 단백질을 사용하였고, hidden은 10개, 초기 가중치 범위는 -0.05 ~ 0.05로 1,000회 학습을 수행하였다. 학습 후 Blind 데이터가 입력으로 들어오면 이미 학습된 weight를 이용하여 출력 값을 계산하게 되고 이 경우 2개의 출력 값이 계산될 것이다. 이중에서 가장 큰 출력 값을 내는 output neuron이 어느 것인지 보고 환자의 질병유무를 판단하게 된다.

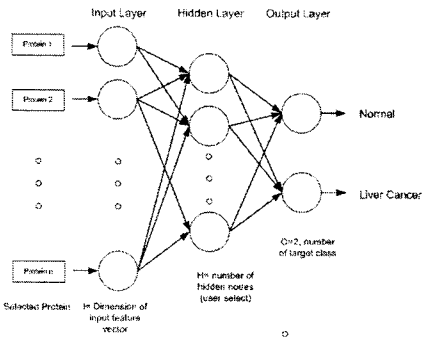

그림 3. 바이오칩 분석을 위한 인공신경망

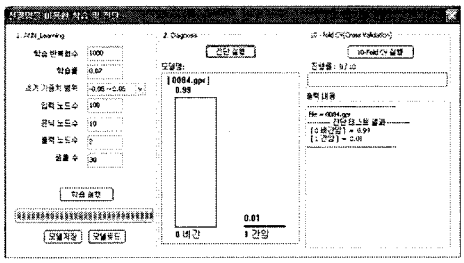

그림 4. 인공신경망을 통한 진단 결과 화면

IV. 시스템 성능 분석

4.1 데이터 및 전처리

본 논문에서 사용된 데이터는 간암으로 확진된 환자50명의 혈액샘플, 간암이외의 대조군 100명의 혈액샘플을 사용하였고, 제노프라(주)가 개발한 1149개의 올리고가 스팟팅된 압타머바이오칩 (이하 1K칩)을 사용하였다. 각 스팟의 median값을 칩 데이터로 사용하였고, ANOVA 테스트후 각 단백질별로 p-value 측정 후 샘플별로 전체 1149개의 단백질 들 중에서 p-value 가 낮은 순으로 상위 60개에서 100개의 단백질을 선택하였다.

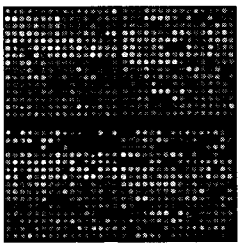

그림 5. 1K 바이오칩 이미지

그림은 1K바이오칩으로 cy3 와 cy5로 염색한 후 획득된 바이오칩의 이미지이다. 염색된 정도의 비율분석을 통해 혈액내 특정 단백질의 양을 추측 할 수 있다.

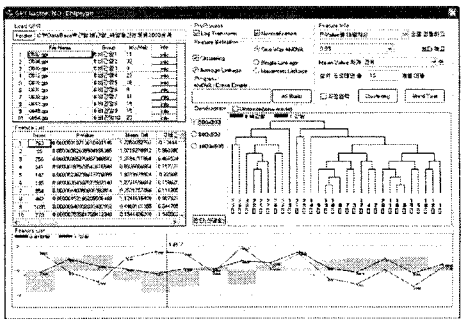

그림 6. 15개의 단백질을 을 골라내는 전처리 과정

4.2 성능평가 결과

전처리결과 선택된 단백질 중 60개에서부터 100개까지 데이터를 학습에 사용하여 성능을 평가 하였다. 성능평가는 10-fold cross validation으로 측정하였으며 결과는 표 1 과 같다.

표 1. 성능평가 결과

학습에 사용된 단백질수	CV 결과
60	92.67
70	92.67
80	92
90	94
100	96

위와 같이 인공신경망을 통한 기계학습을 통해 학습된 단백질의 숫자에 따라 92~96%의 분류 성능을 보였다. 단백질의 숫자를 100개로 하여 학습을 수행했을 경우 가장 좋은 결과를 얻을 수 있었다.

V. 결 론

5.1 결론

혈액 내 단백질의 양을 추측 할 수 있는 기술인 압타머바이오칩 기술과 다량의 생물학적 데이터를 효과적으로 분석할 수 있는 기계학습 기술을 응용하여 혈액내 간암과 연관될 것이라 추측되는 바이오칩 패턴 프로파일 분석을 통해 간암환자를 92~96%를 분류하는 좋은 성능을 보였다. 또한 다양한 기계학습 알고리즘 중 인공신경망이 마이크로어레이 바이오칩 데이터를 분류하는데 좋은 성능을 보임을 알 수 있다. 현재 많이 사용되고 있는 바이오칩의 응용분야중 진단에 직접 응용하기 위해 기계학습 기술을 접목함으로서 기존의 바이오칩 활용에 대해서 효율적인 진단 시스템으로 발전할 수 있음을 알 수 있다. 본 시스템을 기존의 간암표지자인 alpha - fetoprotein (AFP)과 병행, 보완하여 임상에 적용하여 시행할 경우 간암진단의 정확성을 더욱 높이는데 기여할 것이다.

5.2 향후과제

본 기술은 단백질의 분리, 확인, 정량 및 기능해석을 바이오칩 상에서 수행할 수 있어 거의 모든 질병의 진단용 바이오칩 분야에 널리 활용가능하며, 간암뿐만 아니라 수많은 질환진단, 환경모니터링 환경 독성물질 유해성 검색 등으로 시스템의 응용분야를 넓혀나가야 한다. 이를 위해서 보다 많고 다양한 환경에서의 데이터의 확보가 중요하며 인공신경망 이외의 다양한 분류 알고리즘을 적용하여 시스템의 성능을 비교 분석 향상시킬 필요가 있다. 그리고 임상에서의 적용을 위해서 KFDA의 인허가 과정을 거쳐 보다 안정적이고 인증된 시스템으로 구성되어야 할 것이다.

감사의글

본 연구는 재노프라(주)에서 제공된 압타머바이오칩을 사용하였으며 데이터의 소유권은 제노프라(주)에 있음을 밝힙니다.

참고문헌

[1] Eom, J.-H., Kim, S.-C., and Zhang, B.-T., "AptaCDSS : A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction," Expert System with Applications, 34, pp. 2465-2479, 2008.

[2] Cho, S.-B., and Won, H.-H., "Machine Learn in DNA Microarray Analysis for Cancer Classification," Conferences in Research and Practice in Information Technology, Vol. 19, pp. 2003.

[3] Lee, M.-J., Yu, G.-R., Pa가, S.-H., Cho, B.-H., Ahn, J.-S., Park, H.-J., Song, E.-Y., and Kim, D.-G., "Identification of Cystatin B as a Potential serum Marker in Hepatocellular Carcinoma," Clinical Cancer Research, 14, pp.1080-1089, 2008.

[4] Annuska, M.-G., Arno, F., Leonie J.- D., Anke T- W., et al, "Converting a breast cancer microarray signature into a high-throughput diagnostic test," BMC Genomics 2006, 7.278, 2006.

[5] 황규백, 정제균, 남진우, 김병희, 이제근, 장병탁, "바이오데이터 분석을 위한 기계학습 기술", JCCI 2007, pp.35-45, 2007.

[6] 김병희, 김성천, 장병탁, "기계학습에 의한 압타머칩 데이터 기반 심혈관 질환 단계의 예측", 2006한국컴퓨터종합학술대회 논문집, Vol 33 No.1(A), pp.85-87, 2006.

[7] 권형태, 홍진역, 조성배, "마이크로어레이 데이터를 이용한 점증적 유전자 선택기반 암분류", 2007년도 한국정보과학회 가을 학술발표논문집",Vol.34, No.2(B), p.7, 2007.