한의학적 유전병변 바이오 마커 추출 알고리즘

김민경[1]· 우성희[2] · 조영복[3*]

[1](주)소노엠 기업부설연구소 · [2]한국교통대학교 · [3]대전대학교

Biomarker Extraction Algorithm for Oriental Genetic Lesion

Min-kang Kim[1] · Sung-hee Woo[2] · Young-bok Cho[3*]

[1]SONOUM Inc · [2]Korea National University of Transportation · [3]Daejeon University

E-mail : minkyoungk79@gmail.com

요 약

한의학의 '과학화'는 세계 속의 K-MEDI를 위해 선행되야 할 과제로 한약의 효능을 과학적으로 입증하는 작업을 통해 유효성 및 안전성이 확보하고자 노력하고 있다. 본 논문은 한의학적 유전병변 판독을 위한 바이오마커 추출 알고리즘을 제안함으로 한의학적 관점에서 응용되있는 다양한 한의학적 치료법의 진단 및 치료 효과의 객관화에 근거를 제시하였다.

ABSTRACT

'Scientificization' of oriental medicine is a task to be preceded for K-MEDI in the world. Also, We are trying to secure efficacy and safety through scientifically proving the efficacy of Oriental medicine. This paper. We propose a biomarker extraction algorithm for genetic lesion reading of Oriental medicine. Also, A variety of applications in terms of Oriental medicine. Oriental medicine was suggested as a basis for the diagnosis and treatment of the treatment.

키워드

바이오마커, 유전병변, 한의학, 딥러닝

Ⅰ. 서 론

바이오 마커란 몸속 세포나 혈관, 단백질, DNA 등을 이용해 몸 안의 변화를 알아낼 수 있는 지표다. 바이오 마커라는 단어를 처음 정의한 곳은 미국 국립보건원(NIH)이다[1]. NIH는 '바이오 마커란 정상적인 생물학적 과정, 질병 진행 상황, 치료방법에 대한 약물의 반응성을 객관적으로 측정하고 평가할 수 있는 지표'라고 정의했다. 여기에 일반적으로 관찰이 어려운 특정 바이오 마커에 색을 입히거나 빛을 내는 물질을 붙이는 등으로 관찰을 용이하게 해 주는 시약들을 바이오 마커라고 부르기도 한다. 과거에는 혈압이나 체온, 혈당 수치 같은 생리학적 지표가 바이오 마커로 주목받았다. 최근에는 생명과학기술이 눈부시게 발전하면서 현대에는 유전물질(DNA, RNA), 단백질, 세균, 바이러스 등이 바이오 마커로 주목받고 있다.

최근에는 Deep learning이라고 불리는 기계학습이 두각을 보이고 있다. 특히 이미지 인식분야에서 다른 방법보다 뛰어나다고 알려져 있다. Deep learning은 deep neural network의 다른 이름인데 이의 기반이 되는 알고리즘들은 의외로 오래 전에 제안되었다[2]. 다만 당시의 알고리즘상의 overfitting 문제점 및 computation cost 등의 문제로 당시의 연구가 정체되어 있었다면 최근에는 overfit 문제를 해결하는 알고리즘의 개선, 하드웨어의 발전, 그에 따라 기하급수적으로 수집된 정보량으로 인해 neural network를 통한 기계학습이 재조명을 받게 되었다. 나아가 인간의 훈련과정을 본뜬 강화

* corresponding author

학습이라는 machine learning 분야가 급성장하면서 구글 딥마인드사의 Deep Q-network라는 알고리즘이 개발되었다. 기계학습 알고리즘들을 통해 고도로 학습된 모델은 인간보다 정확하고 빠른 패턴 인식과 예측을 가능하게 한다는 것을 경험적으로 알게 되었다. 이에 발맞추어 생물 데이터 역시 고급장비의 개발로 다양화, 대량화 되고 있기에 기계학습을 생물데이터의 인식 및 해석에 활용하는 것이 필요하다. 기초과학연구원(IBS)에서는 바이오마커에 대한 기초연구를 진행하고 있다. 최근 IBS RNA연구단에서는 암 등을 진단하는데 바이오 마커로 활용도가 높은 마이크로RNA(miRNA)의 생성 비밀을 밝혀냈다[2,3]. 마이크로RNA는 세포내 물질로 유전자가 과도하거나 부족하게 활동하지 않도록 조절하는 역할을 한다. 마이크로RNA가 제대로 작동하지 않으면 암이나 당뇨 등 질병을 앓을 수 있다. 의료계에서 바이오마커는 기존의 집단적 진단 테스트, 경험이나 통계에 기반을 둔 치료의 범위에서 바이오마커를 통한 개인별 진단 테스트, 즉, 예측&예방 위주의 개인별 맞춤의료 시스템 구축에 큰 기반이 될 패러다임 변화의 주축이다. 이미 글로벌 바이오마커 의료시장은 꾸준히 성장 중이다.

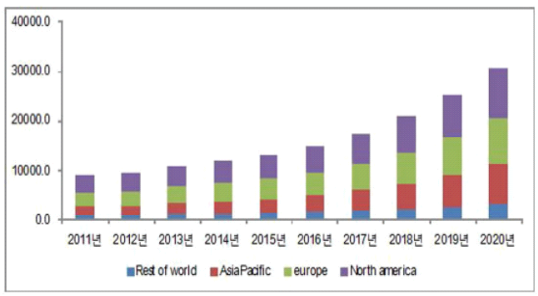

그림 1 바이오마커 시장 전망

[그림1]은 연평균 약 16%의 상장이 예상되고, 특히, 아시아 태평양 지역이 약 20%로 가장 높은 성장률을 전망하고 있다.

II. 관련연구

2.1 바이오 마커의 국내외 연구 동향

바이오마커에 대한 최근의 연구동향은 타겟 물질은 유전자에, 대상 질환은 암에 대한 연구에 집중되어 있다. 특히, 암의 조기진단과 표적치료제의 개발과 연계되어 사용할 수 있는 민감하고 특이적인 암 바이오마커에 대한 관심이 높다. 이 중에서도 바이오마커와 맞춤의약을 동반한 동반진단(Companion Diagnostics)은 성장가능성이 가장 높은 분야로 선진각국에서는 연구지원, 보험적용 등의 정책적인 지원을 통해 괄목할만한 발전을 하였다. 국내에서는 비교적 일찍부터 유전체 연구 육성 및 기술개발지원을 시작하여 맞춤형 치료제 위주의 신약개발사업을 진행하였다[3]. 2012년에는 맞춤의료 연구와 보건의료산업을 위한 핵심 인프라를 구축하고자 국립중앙인체자원은행을 개관하였지만, 현재 맞춤의료를 위한 기초연구와 기술수준은 선진국에 비해 매우 부진한 상황이다. 바이오마커를 이용한 가장 활발한 연구 영역인 암유전체 연구는 인종과 지역 특이성이 있기 때문에 한국인 암환자를 대상으로 한 연구가 필수적인데, 의료현장과 연계된 연구 확대가 정부의 제도적 정비와 함께 동반되어야만 임상 현장에서 상용화되어 사용될 수 있을 것이다[4]. 또한 바이오마커에 대한 연구는 선진국에 비해 초기단계에 있지만, BT, IT 등의 다양한 기술개발이 융합되어 자가 측정용 기기를 중심으로 새로운 체외진단키트에 대한 연구는 지속적으로 이루어지고 있다. 체외진단키트는 해외 기업들이 이미 국내 시장을 상당 부분 선점한 상대로 진입장벽이 높지만 새로운 사업으로 적극 참여하고 있다.

의료산업시장 중 바이오마커 시장의 2020년 예상 규모는 약 300억 달러에 이를 것이고, 연평균 성장률은 약 16%에 이를 것으로 예상되고 있다. 지역별 규모는 미국이 가장 큰 시장을 형성하고 있지만, 아시아 태평양 지역은 약 20.3%로 가장 높은 성장률을 보일 것으로 전망되고 있다. 질환으로는 암질환의 바이오마커가 가장 많은 부분을 차지하고, 분자진단시장의 성장률은 17.6%로 예측되고 있다. 바이오마커에 사용되는 기술은 오믹스 기술(Genomics, Transcriptomics, Proteomics, Metabolomics), 이미징 기술, 생물정보학, 맞춤의료 등을 포함한다. 오믹스 기술과 생물정보학의 연간 성장률은 약 17.1%, 맞춤의료의 연간 성장률은 약 16.39% 일 것으로 예측되고 있다. 맞춤의료와 연관된 바이오마커시장은 다른 바이오마커와 마찬가지로 미국과 유럽이 가장 큰 시장을 형성하고 있으나, 아시아 지역이 21.1%의 가장 높은 연간성장률을 보이고 있다.

최근 IBS RNA연구단에서는 암 등을 진단시 바이오 마커로 활용도가 높은 마이크로RNA(miRNA)의 생성 비밀을 밝혀냈다. 마이크로RNA는 세포내 물질로 유전자가 과도하거나 부족하게 활동하지 않도록 조절하는 역할을 한다. 마이크로RNA가 제대로 작동하지 않으면 암이나 당뇨 등 질병을 앓을 수 있다.여기에 일반적으로 관찰이 어려운 특정 바이오 마커에 색을 입히거나 빛을 내는 물질을 붙이는 등으로 관찰을 용이하게 해 주는 시약들을 바이오 마커라고 부르기도 한다.

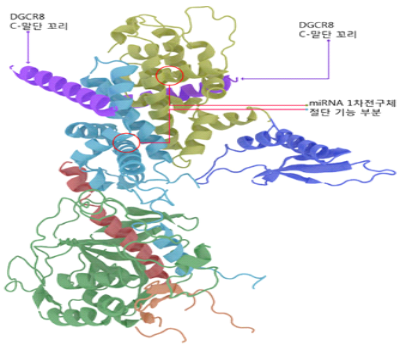

그림 2 드로셔의 3차원 단백질 구조

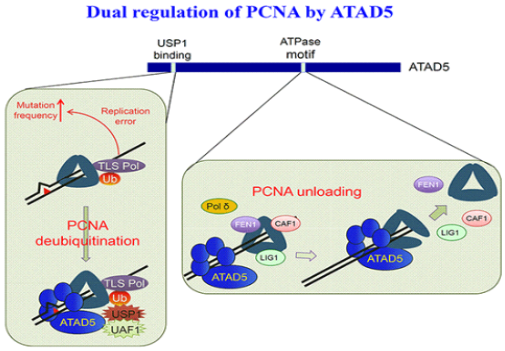

그림 3 ATAD5는 암을 진단하는 바이오 마커

과거에는 혈압이나 체온, 혈당 수치 같은 생리학적 지표가 바이오 마커로 주목받았다. 최근에는 생명과학기술이 눈부시게 발전하면서 현대에는 유전물질(DNA, RNA), 단백질, 세균, 바이러스 등이 바이오 마커로 주목받고 있다. 기초과학연구원(IBS)에서는 바이오 마커에 대한 기초연구를 진행하고 있다. 또한 마이크로RNA는 세포내 물질로 유전자가 과도하거나 부족하게 활동하지 않도록 조절하는 역할을통해 마이크로RNA가 제대로 작동하지 않으면 암이나 당뇨 등 질병을 앓을 수 있다.

2.2 IBS 유전체를 이용한 바이오 마커

BS 유전체 항상성 연구단에서는 DNA의 복구 과정을 연구해 암과 노화의 비밀을 밝혀내고 있다. DNA의 복제 과정에 관여하는 유전자 ATAD5는 종양 억제 유전자로 작용한다. 연구진은 생쥐에서 ATAD5의 발현 양을 의도적으로 줄이면 암이 생기는 것을 확인했다. 자궁내막암, 대장암 등 사람 암세포에서 ATAD5 유전자의 돌변이가 발견되기도 한다. ATAD5 유전자의 이상발현이 암 유발의 시작이 될 수도 있는 것이다. 즉 유전자 ATAD5를 지속적으로 추적한다면 암을 진단하는 바이오 마커로 사용할 수도 있다.IBS 뇌과학 이미징 연구단에서도 뇌 활동에 대한 지도를 그리는 과정에서 바이오마커를 활용하고 있다. 연구단은 최첨단 신경 이미지(뉴로 이미징) 기술을 활용해 뇌의 구조와 기능의 상관관계, 인간의 행동과 신경회로망의 생리학적 메커니즘을 밝혀내고 있다. 연구단이 주로 활용하는 MRI, 광학영상 등에 필수적인 요소가 바로 바이오마커다. 일반적으로 사용하는 나노 조영제는 나노물질과 특정 단백질, 형광 물질 등을 조합해 만든다. 나노 조영제는 뇌 활성화 부분을 측정하거나 암세포를 관찰하는 등 다양하게 사용된다. 연구진은 나노 조영제 외에도 다양한 바이오 마커를 이용해 뇌에서 일어나는 일종의 신경전달 지도(신경회로망)를 구성해가고 있다. 뿐만 아니라 호르몬이나 특정 신경 물질에서 발현되는 바이오 마커를 찾고 개발하는데도 많은 노력을 쏟고 있다.

바이오 마커는 개인맞춤의학·예방의학이 도래하면서 더욱 각광받고 있다. 암이나 심근경색 등 질병의 조기 발견뿐 아니라 신약 개발 분야에서도 주목받고 있다. 최근 미국 NIH는 다운증후군 환자에서 알츠하이머병의 바이오마커를 규명하고 진행하는 연구에 착수하기도 했다. 이 연구에서 발견된 알츠하이머 바이오마커는 치매 환자 치료에 활용될 수 있다. 바이오 마커는 '똑똑한' 센서와 같다. 암이나 심근경색 등 질병과 밀접하게 관련된 마커만 알면 복잡한 생명현상 속에서도 신속하고 정확하게 병을 진단할 수 있다.

III. 한의학적 유전병변 바이오마커 추출을 위한 강화 학습 알고리즘

유전체 서열 분석에서 대표적인 기계학습응용은 바로 유전자 예측(gene prediction)이다. 고품질의 표준유전체를 제작하였다고 가정을 한다면, 염색체의 개수에 해당하는 매우 긴 염기서열의 array를 얻을 수 있다. 여기서 우리가 이미 알고 있는 생물 기능의 단위인 유전자에 해당하는 자리를 찾아야 할 필요가 있다. 일반적으로 쓰이는 알고리즘이 hidden Markov model (HMM)이다[5,6]. HMM을 간단히 설명하기 위해서는 먼저 Markov process라는 개념을 이해할 필요가 있다. 서 HMM은 관찰값이 어떤 상태를 추정해야 하며, 그 상태의 변화는 시간상 바로 직전의 상태에 의존한다고 가정하는 것이 적절하다면 매우 잘 적용된다. 대표적인 적용예는 음성인식이다. 음성데이터를 통해 단어라는 어떤 상태를 추정해야 하며, 그 단어는 잘 정리된 문법을 통해 훈련된 확률로 다음 단어의 가능성과 관찰 값의 가능성을 비교해서 최적의 문장 혹은 구절을 재구성한다.

3.1 GWAS

NGS의 발전으로 분자표지가 대량으로 개발되고 유전체를 관찰할 수 있는 해상력은 상당히 늘어

났으며 육종집단의 개체 혹은 재래종 개체들을 모두 sequencing하여 whole genome QTL mapping 혹은 genome wide association study (GWAS)를 하여 어떤 분자표지가 어떤 표현형에 기여하는지에 관련된 데이터가 쏟아지고 있다。이미 알고 있는 대량의 분자표지를 가중치와 함께 표현형을 추정하는 방법이 필요하게 되었다.

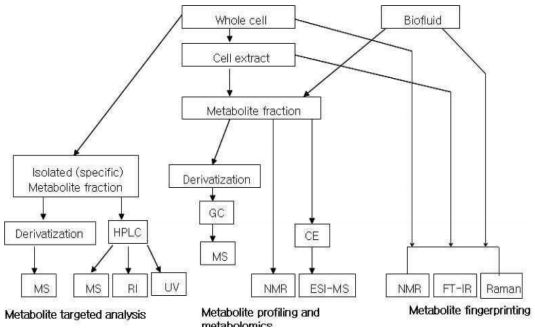

그림 4 생체의 대사물질 분석을위한 플로챠트

3.2 강화학습

GS에서는 기계학습의 지도학습 문제 해결과 마찬가지로 training set의 답안지를 통해 모델을 훈련하고 test set을 통해서 확인한다. Training set에 해당하는 것은 training population이라고 부르며 각 개체의 관찰 가능한 모든 genotype과 원하는 표현형 데이터를 갖추고 있다.

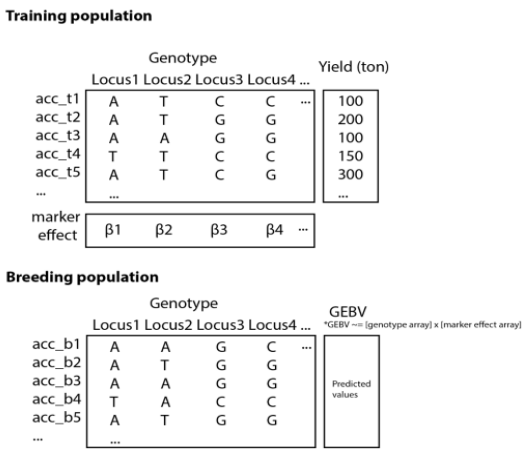

그림 5 GS 모델

[그림4]는 이 데이터를 이용해서 각 genotype의 locus가 표현형에 얼마나 영향을 미치는지를 marker effect를 대표하는 β값을 훈련한다. 그리고 이 marker effect는 test set에 해당하는 breeding population에서 적용되어 genomic estimated breeding values (GEBVs) 계산한다. 이는 기존의 분자표지 육종 시에 육종가가 임의로 분자표지를 선발하거나 임의로 가중치를 결정하는 대신에 모든 분자표지에 대한 가중치를 미리 모델링하기 때문에, 표현형에 약한 영향을 가지는 locus를 놓치지 않고 예측값에 반영할 수 있다.

Ⅳ. 결 론

현재 한의학적으로 바이오마커를 이용한 대사질환등 연관성 판독을 위한 연구가 이루어지고 있다. 한의학적 뇌졸중 진단 방법이나 우울증, 조현병 외에도 양극성장애 조기 진단율을 높이기 위한 뇌영상 연구도 점차 확대되고 있다. 현재 양극성 장애 발병 위험이 높은 청소년부터, 향후 양극성 장애 발병 위험이 상승해 치료가 필요한 청소년까지 분류했다. 정확도는 75%였다. 또 양극성 장애 부모를 둔 청소년과 정상 부모를 둔 청소년도 추가로 구분이 가능하다.

또한 인체의 복잡한 변화를 관찰하기 위한 이론적 근거를 수립하고 동양의학의 과학화를 위해 한의학적 실험방법의 한계를 극복하고자 발전하는 IT 기술을 기반으로 한의학적 예측 및 객관화 즉 표준화를 위해 노력하고 있다. 이렇게 대사체학(Metabolomics)을 한의학적 관점에서 응용하게 된다면 다양한 한 의학 치료법의 진단 및 치료 효과의 객관화에 근거를 제시하였다.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2018R1C1B5083789).

References

[1] Y. D. Joung, “Effective Utilization and Problems of Biometrics”, *Hongik University,* pp. 1-39 2007

[2] J. K. Beck, “The Slope Extraction and Compensation Based on Adaptive Edge Enhancement to Extract Scene Text Region”, *Journal of the Digital Contents Society of Korea*, vol 18, no. 4, pp. 777-789, 2017

[3] Y. B. Cho, and S. H. Woo, "Algorithm for Extraction of ROI Using Fast Binarization Image Processing",*Journal of the Korea Institute of Information and Communication Engineering*,vol 22, no. 04 pp. 0634-0640, 04. 2018

[4] C. Y. Lee, and N. H. Kim, “A Study on the Edge Detection using Region Segmentation of the Mask”, *The Journal of the Korea Information and Communications Society*, vol 17, no.3, pp. 718-723, 2013

[5] A. J. Lorenz, S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi, H. Iwata, K. P. Smith, M. E. Sorrells, J. L. Jannink,“ Genomic Selection in Plant Breeding”. 2011, 110:77-123.

[6] Bogdanov M, Matson WR, Wang L, Matson T, Saunders-Pullman R, Bressman SS, “Metabolomic profiling to develop blood biomarkers for Parkinson's disease”. *Brain.* 2008 ; vol. 131, no. 2, pp. 389-96. 2008