miRNA 데이터베이스 통합 및 순위 결정에 의한 특정 질병 관련 microRNA의 추출 방법

하지환, 김현진, 박상현
연세대학교 컴퓨터과학과
e-mail : jh0499@yonsei.ac.kr

Finding Specific Disease Related microRNA Using by Ranking Score with Integrated miRNA Database

Ji-Hwan Ha, Hyun-Jin Kim, Sang-Hyun Park
*Dept of Computer Science, Yonsei University

요 약

최근 MicroRNA(miRNA)가 질병 발생과 밀접한 연관성이 있다고 밝혀진 이래, 이와 관련된 연구가 활발히 진행되고 있다. 하지만 각종 질병 관련 miRNA의 기능과 역할 그리고 질병 발생 메카니즘 등이 명백히 밝혀진 것이 없는 실정이다. 본 논문에서는 여러 종류의 miRNA 데이터베이스(miRecords, miRTarBase, miR2Disease 등)를 통합하고, 본 논문에서 새로이 제안하는 scoring 방법과 특정 질병과 관련된 miRNA의 순위결정과정을 통하여 질병과 연관성이 높은 miRNA을 밝혀내는 방법을 제안한다. 새로이 제안하는 방법을 바탕으로 miRNA와 특정 질병과의 연관성을 효과적으로 밝혀냈다.

1. 서론

생명과학의 연구과정에 있어서 가장 큰 발견의 하나로 DNA가 유전물질이라는 것이 밝혀진 사례를 들 수 있다. 이후 DNA와 단백질에 관한 많은 연구가 활발하게 진행되었고, 이 과정에서 microRNA(miRNA)라는 물질이 발견된 것은 불과 얼마 되지 않은 일이다. miRNA는 19-25개의 뉴클레오티드(nucleotide)로 이루어진 단일 염기가닥으로 유전자를 조절하는데 있어서 중요한 역할을 하는 인자이다. 이전의 연구결과에 의하면, miRNA의 역할은 단순히 DNA에 담겨있는 유전정보를 단백질로 변환시키는 과정에서 정보를 전달하는 물질로만 알려져 있었다. 하지만 많은 실험과 연구 결과 miRNA가 단순히 유전정보 전달 역할만 하는 것이 아니라, 세포증식 및 발달 등을 포함한 거의 모든 생명현상에 관여함으로써 질병 발생과 밀접한 연관이 있다고 밝혀진바 있다.

최근 miRNA의 기능과 질병과의 연관성을 찾기 위한 시도가 계속 되고 있지만, miRNA의 데이터 수의 부족과 miRNA의 질병 관련 정보의 부족으로 인하여 관련 연구내용은 매우 미흡한 실정이다. 본 논문에서는 여러 종류의 miRNA 데이터 셋(miRecords, miRTarBase, miR2Disease 등)을 통합하고, target gene을 이용한 접근 방법을 통하여 miRNA가 특정 질병과 연관이 있는지를 밝혀내는 시도를 수행하였다. 새로이 제안하는 scoring 방법과 특정 질병과 관련된 miRNA의 순위 결정과정을 통하여 질병과 연관성이 높은 miRNA을 밝혀내는 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서는 데이터 통합과 실험방법(새로운 scoring 방법, 순위 결정과정)을 기술하고, 3장에서는 실험결과를 정리 해석한다. 4장에서는 추후 연구 방향을 제시함과 동시에 결론을 맺는다.

※ 이 논문은 2012년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2012R1A2A1A01010775).

2 데이터 출처 및 실험 방법

2.1 데이터 출처

miRNA 데이터의 수가 다른 생물학적 데이터보다 상대적으로 훨씬 적기 때문에, 본 논문에서는 3가지 miRNA 데이터를 통합(miRecords, miRTarBase, miR2Disease)하여 중복되는 것은 제외하고 총 982개의 miRNA 데이터와 이와 관련된 39642개의 target gene을 확보하였다.

<표 1>은 본 논문에서 사용한 데이터의 출처와 종류와 그리고 총 데이터 수를 나타낸다. miRecords에서 제공하는 2120개의 miRNA- target gene pair 데이터를, miR2Disease에서는 753개의 miRNA- target gene pair 데이터를, miRTarBase에서는 39139개의 miRNA- target gene pair 데이터를, 그리고 hmdd에서 5076개의 질병 관련

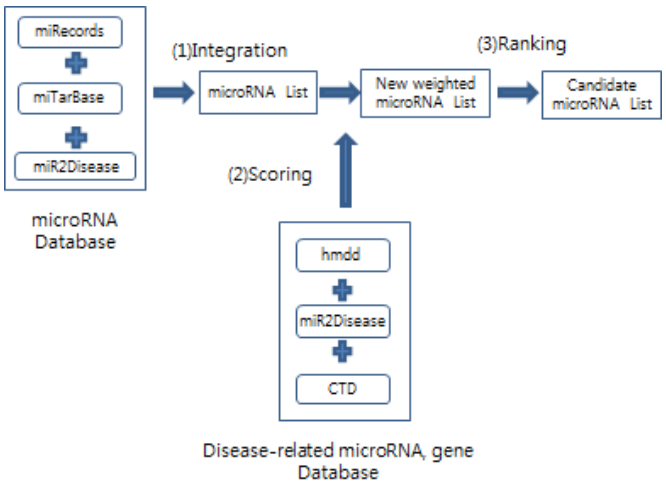
miRNA를 포함하는 데이터를 다운로드 받을 수 있었다. 또한, miR2Disease에서는 질병 관련 miRNA 2877개를 포함하는 데이터를 다운로드 받았고, CTD에서 Alzheimer Disease와 관련이 있다고 밝혀진 유전자(disease related gene) 총 3349개도 다운로드 받았다.

<표 1> 실험에 사용한 데이터 종류 및 출처

데이터 종류	총 수	출처
miRNA target gene pair	2120 개	miRecords
miRNA target gene pair	753 개	miR2Disease
miRNA target gene pair	39139 개	miRTarBase
질병 관련 miRNA	5076 개	hmdd
질병 관련 miRNA	2877 개	miR2Disease
Alzheimer Disease 관련 유전자	3349 개	CTD

2.2 실험 방법

실험은 그림 1에 나타낸 바와 같이 크게 분류하여 살펴보면, (1) miRNA 데이터를 통합(integration)하고, (2) 이미 알려진 질병 관련 miRNA와 유전자를 토대로 점수화(scoring)한 후, (3) 점수화 된 miRNA를 순위과정을 결정(ranking score)하는 순서로 구성되어 진행한다.



(그림 1) 실험 개요도

즉, 본 논문에서는 특정 질병이 주어졌을 때 우선, 본 논문에서 제안하는 2가지 점수화(scoring) 방법을 이용하여 각각의 miRNA가 얼마나 질병과 연관성이 있는지 점수화 하고, 그 점수(score)를 토대로 1위부터 순서대로 순위결정과정(ranking)을 수행하였다. miRNA를 하나의 개별 node로 나타내었고, miRNA 간에 서로 공유하는 target gene가 있으면 둘 사이를 edge로 이어 나타내었다. miRNA끼리의 유사성(similarity)은 공통으로 공유하는 target gene가 많을수록 높다고 가정하여 실험을 수행하였다.

2.2.1 데이터 통합

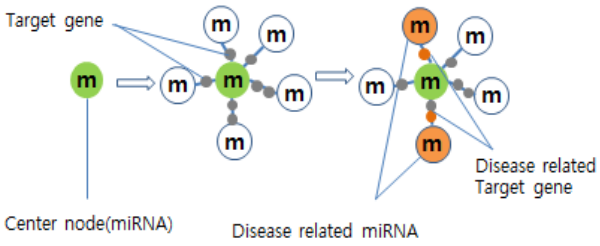
miRNA 데이터를 통합(miRecords, miRTarBase, miR2Disease)하여 중복되는 것은 제외하고 총 982개의 miRNA 데이터와 39642개의 target gene을 확보하였다.

2.2.2 점수화

본 단계에서는 miRNA가 얼마나 특정 질병과 관련이 있는지를 각 miRNA 노드 당 score를 부여하였다. 아래에 제안하는 두 가지 점수화(scoring) 방법을 통해 얻은 점수를 Min-max의 정규화 방법으로 0~1사이의 값으로 정규화 하여 1에 가까울수록 질병과 관련이 높은 miRNA일 것이라고 추정하였다. miRNA는 두 miRNA 사이에 target gene을 많이 공유 할수록 서로 기능적으로 유사성(similarity)이 높다고 본다. 따라서 center node(miRNA)에 edge로 이어진 주변 miRNA들이 질병과 연관이 많을수록 center node도 질병과 관련이 높을 것이라 가정하였다. 본 논문에서는 알츠하이머병과 관련이 있는 miRNA를 점수화하였다.

(Scoring 방법 1):

그림 2에 나타낸 바와 같이, miRNA를 하나의 center node로 두고, miRNA 간에 서로 공유하는 target gene이 존재하면 둘 사이에 edge를 추가한다. 공유하는 target gene을 바탕으로 이웃 miRNA는 중심 mi RNA 주변으로 edge로 연결되어 miRNA 하나당 개별 네트워크를 형성한다. 공유하는 target gene의 수에 따라 node 간 거리(distance)로 유사성(similarity)을 나타내었다.



(그림 2) miRNA 개별 네트워크 형성 과정

$$Score_1 = number\ of\ disease\ related\ target\ genes + number\ of\ disease\ related\ miRNA\ connected\ with\ disease\ related\ target\ genes * 0.5 \quad (1)$$

두 miRNA간 공유하는 target gene이 특정 질병과 연관이 있는 유전자라면 Center node에 점수 1점을 부여한다. Target gene와 특정 질병간의 관계 유무 판별은 CTD에서 다운로드 받은 데이터를 기반으로 수행하였다. 질병 관련 Target gene을 공유하는 이웃 miRNA가 질병 관련 mi RNA라면 추가적으로 중심 miRNA score 1 점수에 0.5를 합산한다.

(Scoring 방법 2):

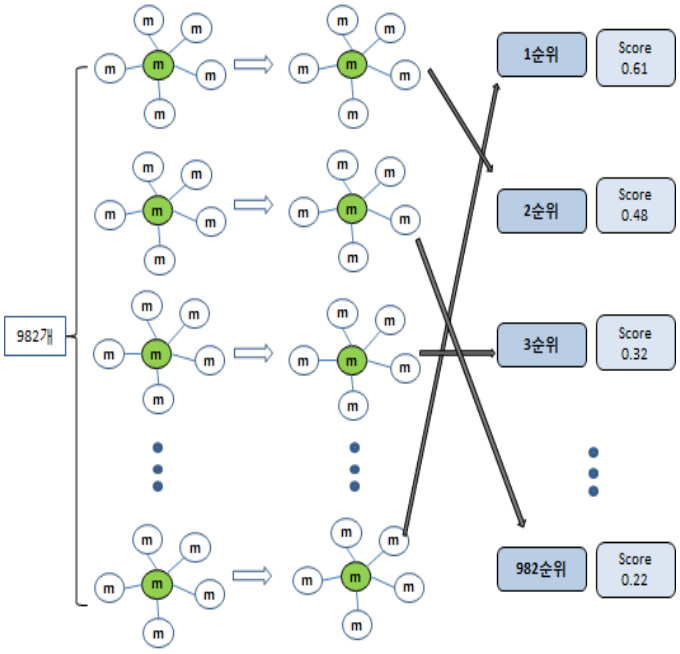
기본적으로 두 mi RNA 간에 공통적으로 공유하는 target gene을 바탕으로 개별 네트워크를 형성하는 방법은 scoring 방법 1과 동일하다. 중심 miRNA의 질병 연관성을 살펴보기 위해서 edge로 연결된 주변 miRNA가 질병과 관련이 있는지를 확인한다. miRNA의 질병 관련 유무는 miR2Disease 데이터를 기반으로 확인하였다. Edge로 연결된 주변 miRNA가 질병과 관련이 있으면 그 수만큼 중심 miRNA에 1점을 부여한다. 그리고 거리를 기반으로 한 두 miRNA 사이의 유사성(similarity)을 바탕으로 질병과 관련된 miRNA와의 거리를 점수에 포함하였다. miRNA 사이의 거리는 target gene의 수에 비례하며 min-max 정규화를 통하여 거리를 0~1로 나타내었다.

$$Score_2 = number\ of\ disease\ related\ miRNA + Distance\ between\ center\ node\ and\ disease\ related\ miRNA(similarity) \quad (2)$$

$$Total\ Score_i = w_1 * Score_1 + w_2 * Score_2 \quad (3)$$
w_1: Score1의 가중치
w_2: Score2의 가중치

2.2.3 순위화

위의 과정에서 구한 각각의 Scoring 방법 1과 Scoring 방법 2를 합산하여 Total Score을 구하였다. 이 Total score을 바탕으로 (그림 3)에 나타낸 바와 같이 miRNA를 1위부터 982위까지 순위화를 수행하였다.



(그림 3) 순위 결정 과정도

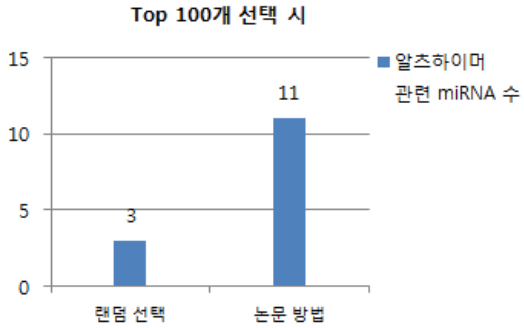
3. 실험과정 및 결과

본 논문에서는 질병과 연관이 있는 miRNA를 찾는 실험을, 알츠하이머 질병 관련 데이터를 표본으로 하여 수행하였다. 본 논문에서 새로이 제안하는 scoring 방법 1과 2를 통하여 질병과 관련된 miRNA의 순위 결정과정을 수행한 후, 알츠하이머 병과 연관성이 높은 miRNA를 추출하였다. 이 과정에서 $Score_1$ 과 $Score_2$의 가중치 w_1, w_2을 각각 0.2, 0.8을 사용하였다. 가중치는 각각 0.1부터 0.9까지 0.1씩 증가하여 차례대로 실험을 수행한 후, 가장 결과가 좋은 값으로 선택하였다. 총 982개의 miRNA 중 알츠하이머와 관련된 miRNA가 30개 있을 때, 실험을 수행하여 다음 <표 2>와 같은 결과를 얻었다.

<표 2> 실험결과

	랜덤 선택 방법	본 논문 방법
Top 30	1	3
Top 50	1.5	6
Top 100	3	11
Top 200	6	20

<표 2>는 차례로 상위 30,50,100,200을 선택 했을 시 실제로 알츠하이머 관련 miRNA를 포함하는 수를 나타낸 것이다. <표 2>에서 기존의 랜덤 선택 방법보다 본 논문에서 제시하는 방법이 3~4배 정도 더 많은 알츠하이머 관련 miRNA를 찾아냄을 확인할 수 있다.



(그림 4)랜덤 선택 시와 논문 제시방법 비교분석

(그림 4) 상위 100개의 miRNA 추출 시 랜덤 선택 방법은 3개의 miRNA가 알츠하이머와 관련이 있다고 검증되었다. 본 논문에서 제시하는 방법으로는 100개중 11개의 miRNA가 알츠하이머병과 관련이 있다고 검증되었다. 이는 본 논문에서 제시하는 방법이 정확성이 높은 miRNA 추출 방법임을 입증하였다.

4. 결론 및 추후 연구 방향

본 논문에서는 질병 관련 miRNA 추출 및 예측하는 방법으로, miRNA 데이터를 통합하고 miRNA의 Target gene을 이용하여 두 miRNA 사이의 거리를 기반으로 한 유사성(similarity)을 구하여 질병 관련 miRNA 추출 및 예측하

는 것이다. 실험 결과, 논문에서 제시하는 질병 관련 miRNA 추출 방법이 랜덤 선택 방식보다 약 3~4배 많은 질병 관련 miRNA를 찾을 수 있었다. 이번 연구에서 miRNA의 target gene과 miRNA 사이의 거리 기반에 따른 유사성(similarity)에 중점을 두고, 개별 miRNA 네트워크에 대해서만 실험을 수행했다면 차후의 연구에는 miRNA의 binding site와 Gene Ontology(GO) 까지 고려하여 통합적인 miRNA 네트워크를 형성, miRNA의 기능과 질병 관련성을 더욱 효과적으로 해석하고자 할 계획이다.

참고문헌

[1] V. Narry Kim, "MicroRNA biogenesis: coordinated cropping and dicing," Nature Reviews Molecular Cell Biology, vol.6, 2005

[2] David P.Barte, "MicroRNAs: genomics, biogenesis, mechanism and function," Cell, vol.6, 2005.

[3] Sheng-Da Hsu1, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel, Chih-Hung Chou1, Chao-Fang Chu1, Hsi-Yuan Huang, Ching-Min Lin, Shu-Yi Ho, "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions", Nucleic Acids Research, vol.42, D78 – D85, 2014

[4] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao and Tongbin Li, "miRecords: an integrated resource for microRNA-target interactions", Nucleic Acids Research, vol.37, D105 – .D110, 2009

[5] Yang Li, Chengxiang Qiu, Jian Tu, Bin Geng, Jichun Yang, Tianzi Jiang,Qinghua Cui, "HMDD v2.0: a database for experimentally supported human microRNA and disease associations",Nucleic Acids Research,Vol. 42,D1070 – .D1074, 2014

[6]]Qinghua Jiang, Yangyang Hao, Guohua Wang, Liran Juan, Tianjiao Zhang, Mingxiang Teng, Yunlong Liu,Yadong Wang ,"Prioritization of disease microRNAs through a human phenome-microRNAome network", BMC Systems Biology, 2010

[7] Xia Li1, Qianghu Wang, Yan Zheng, Sali Lv, Shangwei Ning, Jie Sun, Teng Huang, Qifan Zheng, Huan Ren, Jin Xu, Xishan Wang and Yixue Li," Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer",Nucleic Acids Research, vol.39, 2011

[8] miRecords [Internet], http://mirecords.biolead.org/

[9] miR2Disease [Internet] http://www.mir2disease.org

[10] mirTarbase [Internet], http://mirtarbase.mbc.nctu.edu.tw/php/download.php