microRNA 발현 데이터의 상관관계 분석을 통한 microRNA Functional Family 탐색

남진우[O,1,2] 장병탁[1,2,3]
서울대학교 대학원 생물정보학 협동과정[1]
서울대학교 바이오정보기술 연구센터(CBIT)[2]
서울대학교 컴퓨터공학부 바이오지능연구실[3]
{jwnam, btzhang}@bi.snu.ac.kr

Defining microRNA functional families through correlation analysis of microRNA microarray data

Jin-Wu Nam[O,1,2] Byoung-Tak Zhang[1,2,3]

Graduate Program in Bioinformatics[1]
Center for Bioinformation Technology (CBIT)[2]
Biointelligence Laboratory, School of Computer Science and Engineering[3]
Seoul National University, Seoul 151-742, Korea

요 약

microRNA는 유전자의 전사 후 과정에서 negative regulation을 담당하는 small noncoding RNA의 한 종류이다. 최근 까지 330여개의 인간 microRNA가 발견되었지만 그들의 기능이 밝혀진 것은 소수에 불과하다. microRNA의 기능은 3'UTR에 불완전 상보결합을 통해 negative regulation을 받게 되는 유전자의 기능으로부터 유추되는 것이 일반적이다. 특별히 유전체상에 군집화 된 microRNA들은 하나의 전사체로부터 발현되는 것으로 판단되며, 같은 또는 관련된 기능을 하거나 같은 목표 유전자를 조절하기 위한 functional family일 가능성이 높다. 또한 이러한 functional family는 하나의 전사체로부터 발현되기 때문에, 조직별로 조건별로 같은 발현 패턴을 보여야 한다. 본 연구에서는 발현데이터로부터 microRNA functional family를 탐색하기 위해, 5개의 연구 그룹에서 공개한 조직별 microRNA 발현데이터를 표준화 작업을 거친 후 통합하고, k-nearest neighbor 알고리즘을 이용해 결측치를 보정한 후 microRNA 발현사이의 correlation을 계산한다. 이때 데이터 통합에서 생기는 문제에 robust한 결과를 얻기 위해 실제 발현 데이터가 아닌 rank 데이터부터 correlation을 측정한다. 계산된 spearman ranked correlation 결과와 microRNA의 genomic coordination 정보로부터 34개의 functional family를 정의할 수 있었다.

1. 서 론

microRNA (miRNA)는 약 21 nt 길이의 small RNA의 한 종류로, 일반적으로 mRNA의 3'untranslated region에 binding하여 mRNA의 번역을 저해 시키며[1], 특히 최근 보고에 의하면, miRNA에 의한 mRNA의 deadenylation은 mRNA의 stability를 저하시키는 것으로 보고되었다 [2,3]. miRNA는 인간 유전체에서 현재까지 약 330여개가 발굴 되었으며, 유전자 개수의 약 1~3%의 수가 존재 할 것으로 예상되고 있다 [4]. miRNA는 RNA polymerase II에 의해 전사되며 (pri-miRNA라 부른다), 하나의 전사체 내에서 복수의 pre-miRNA가 발현되는 polycistronic 한 구조로 되어 있다 [5]. 이렇게, miRNA는 약 50% 이상이 유전체상에서 tandem array 형태로 나타나며, 발현 패턴이 조직별 시기별로 연관되어 있으므로 관련된 기능을 할 것으로 유추되고 있다 [5].

최근 microarray, bead assay, quantitative PCR 그리고 serial analysis gene expression (SAGE)와 같은, 발현 분석 방법을 이용하여 miRNA의 기능을 연구하기 위한 대량의 발현 데이터가 보고되었다 [6,10]. 특히 암과 miRNA의 상관성을 분석하기 위한 노력이 최근 이루어지고 있다 [7]. miRNA의 이상 발현 패턴은 regulation을 하는 oncogene의 발현 패턴을 변화시켜 암을 발생시키는 것으로 보고되었으며, 이러한 이상 발현 패턴은 암을 진단하는 signature로서 주목 받고 있다 [7].

Hayahsita et al.은 최근 폐암 세포에서 polycistronic miRNA, mir-17-92이 과 발현되고 있음을 보고 하였다 [8]. 이로 인해 폐암 세포의 proliferation이 증가됨을 보였다. 또한 Chen et al.은 clustered miRNA mir-1-133이 muscle의 발생 과정을 담당하며, 발생 과정 중 세포 특이적이 발현 패턴이 유사하게 나타남을 보였다 [9]. 이렇듯 polycistronic miRNA들은 같은 기능을 하기위해 같은 시기에 같은 조직에서 특이적인 발현 패턴을 보이는 functional family일 가능성이 높으며, 대부분 중요한 세포내 기능을 담당할 것으로 생각되고 있다. 이러한 polycistronic한 miRNA의 발굴은 miRNA 기능을 연구하는데 필수적이다.

본 논문에서는 polycistronic miRNA set을 발굴하기

위해 공개된 miRNA 발현 데이터를 이용하여 조직별로 특이적인 발현 상관성을 보이는 miRNA set중 genome 상에 clustered 된 miRNA set을 찾는다. 이를 위해 서로 다른 그룹에서 생산된 5 종류의 miRNA 발현 데이터들을 조직별로 통합하여 사용한다. 상관분석의 결과는 수치의 유사성을 보이기 위해 euclidean distance matrix와 비교 한다.

2. Correlation analysis

2.1 Expression profile dataset of microRNA

miRNA의 조직 특이적 발현의 상관관계를 분석하기 위해 다양한 발현 분석 방법을 통해 생산된 5종류의 발현 데이터를 이용한다. 우선 세 곳의 연구그룹에 의해서 만들어진 microarray 데이터는 Kim et al. [5] 논문에서 통합된 220 miRNA에 대한 37개 조직별 데이터를 사용하며, 또 다른 microarray 데이터는 He et al. 에 의해 보고된 190개 miRNA에 대한 40 sample의 lymphoma 데이터와 [10], Bentwich et al.에 의해 보고된 254개 miRNA에 대한 5개 조직의 데이터를 사용 한다 [6].

2.2 Integration of multi-source data

위에서 언급한 5개의 multi-source 데이터를 통합하는 것은 쉬운 일이 아니다. 다양한 실험 방법으로 인한 수치의 다양성과 오차, control이 달라 데이터의 비교 및 표준화의 문제점 등은 multi-source 데이터의 통합에 가장 큰 걸림돌이다. 또한 데이터의 통합에서 생기는 결측치들은 meta 분석의 성능을 크게 좌우하고 있다. 우선 데이터의 표준화를 위해 수치들을 sample 별로 표준화 한다 (식 1).

$$e_{ij} = \frac{x_{ij} - \mu_{\cdot j}}{\sigma_{\cdot j}} \qquad (1)$$

e_{ij}는 j번째 sample에서 i번째 miRNA의 발현데이터의 표준화된 수치를 말하며, 이것은 표준편차(σ)와 평균값(μ)으로 계산된다. 이때 결측치를 보정하기위해 k-nearest neighbor (k-nn) 알고리즘을 적용한다. 여기서는 k를 5로 사용한다. 결측치 e에 대해서 관측된 데이터 사이에서 높은 상관성을 보이는 k 개의 nearest miRNA를 $x_1 \ldots x_k$ 라고 할 때 e는 다음 식 2와 같이 평균값으로 구해진다. 이때 계산되는 상관성은 수치가 아닌 rank를 사용하며, 아래에서 자세히 설명된다.

$$F(e) = \frac{1}{k}\sum_{i=1}^{k} x_i \qquad (2)$$

이렇게 5개의 발현 데이터는 254 miRNA에 대한 82 sample 데이터인 254x82 matrix로 통합된다.

2.3 Spearman rank correlation

데이터의 상관분석은 comparable한 데이터가 주어졌을 때 가능하다. 하지만 본 연구를 위해 수집된 발현데이터의 수치들은 다양한 방법으로 만들어져, 각 방법에 대한 bias가 존재할 뿐 아니라, control의 발현데이터가 없기 때문에 비교가 가능하지 않다. 이를 극복하기 위해 본 연구에서는 발현데이터의 rank를 사용한다. rank를 사용하면 실제 수치를 사용할 수 없게 되는 단점이 있지만, 등분산성을 보장해주며 이종 데이터들 간의 비교가 가능해지는 장점이 있다.

여기서, 상관분석을 위해 Spearman rank correlation 방법을 사용한다. 상관계수 ρ 는 아래 식 3처럼 관찰된 rank의 차이(D)와 pair의 개수(N)에 의해서 계산될 수 있다.

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} \qquad (3)$$

또한, N의 수가 충분히 클 때는, 아래 식 4의 t는 상관계수가 0인(null case)일 때 student t' 분포를 가진다.

$$t = \frac{\rho}{\sqrt{(1-\rho^2)/(n-2)}} \qquad (4)$$

이 통계치를 이용해 우리는 그 상관성이 통계적으로 유의한지 무의미 한지를 판단할 수 있게 된다.

2.4 Euclidean distance matrix

correlation은 발현의 패턴이 유사한지를 보여주기는 하지만 그 값들이 유사한지는 알 수 없다. 이것은 euclidean distance로 알 수 있는데, rank로 변환한 데이터에서 i번째 miRNA와 j번째 miRNA 간의 euclidean distance는 차이 제곱의 합의 제곱근으로 계산되며, 여기서는 결측치를 제외한 거리를 구하기 위해 식 5와 같이 평균값을 이용한다.

$$mean(d_{i,j}) = \frac{\sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2}}{n} \qquad (5)$$

여기서 n은 결측치를 제외한 차이를 구한 샘플 수이다. 이렇게 m개의 miRNA 데이터가 있을 때 m x m matrix를 만들 수 있다.

3. 결 과

3.1 Defining functional family

Polycistronic miRNA들은 시기별 조직별 특이적 발현 패턴에 높은 상관계수 값을 갖는다. 하지만, Polycistronic 하지 않은 유전체 내에 다른 loci에 위치하더라도 functional family일 가능성은 여전히 남아 있다. 또한 clustered miRNA가 하나의 같은 전사체로 발현된다는 것 또한 장담할 수 없다. 이러한 이유로 polycistronic miRNA set을 결정하기 위해서는 유전체 상의 위치 정보와 발현데이터 간의 상관계수가 모두 필요하게 된다.

그림 1은 254 x 254 symmetric correlation matrix 중 일부의 결과를 보여준다. 각 miRNA들은 인간 염색체 3번에서 7번상의 위치에 따라 순서적으로 배치되었다. 그림 1은 correlation의 절대 값이 0.4보다 클 때만 녹색과 빨간색의 ratio heat map으로 표시하였다. 빨간색일수록 높은 양의 상관성을 녹색일수록 높은 음의 상관성이 있음을 의미한다. 또한 왼쪽의 색깔 별 바는 각 miRNA가 해당하는 염색체 번호를 보여주며, 파란색 박스는 염색체상에서 clustered miRNA를 표시하고 있다.

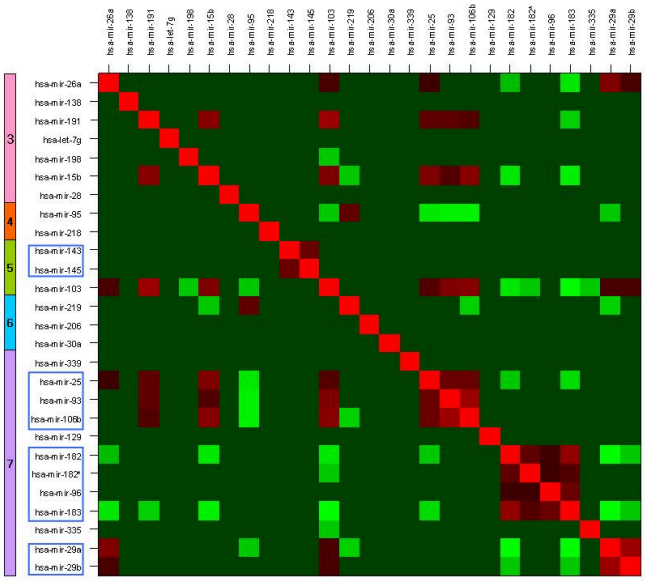

그림 1. Correlation matrix와 functional family

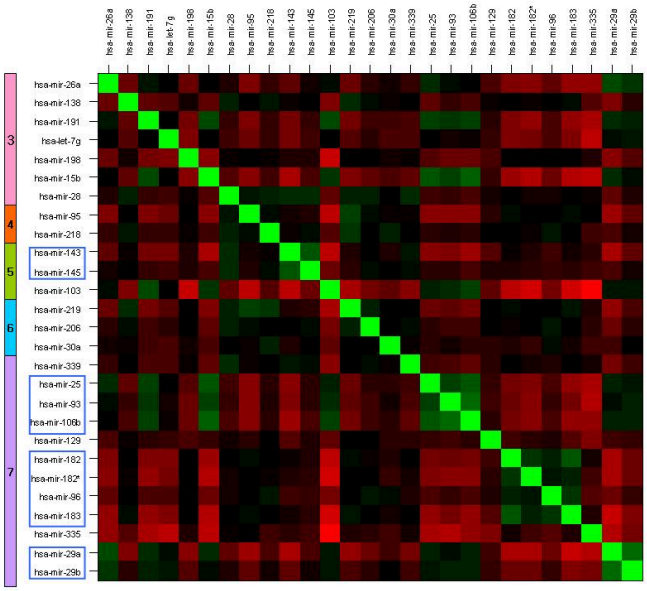

그림 2. Euclidean distance matrix와 functional family

그림 1에서 clustered miRNA들은 correlation matrix 상에서 높은 상관관계를 보이고 있으며, 이것은 clustered miRNA들이 조직별 시기별 특이적으로 발현상의 상관성을 높게 보임을 분명히 말해주며, functional family로서 정의가 가능한 set이라 할 수 있다.

또한 clustered되어 있지 않거나 다른 염색체 상에 있는 miRNA 중 이러한 functional family들과 발현 상관성을 보이는 것들이 존재한다. 예를 들어 mir-182-96-183 functional family과 높은 발현 상관성을 보이는 mir-103, mir-15b, mir-191들 또한 functional family라고 예측 될 수 있다. 전자를 clustered functional family (CFF)라고 하다면 후자를 distant functional family (DFF)로 정의할 수 있다. 또한 그림 1에서 CFF인 mir-29a-29b의 DFF는 mir-103 과 mir-26a로 보여 진다. 전체 genome wide 분석에서 총 34개의 CFF를 찾을 수 있었으며 백여 개의 DFF를 찾을 수 있었다.

한편 위와 같은 CFF의 발현 상관성이 발현 수치의 유사성에 기인하는 것인지 보기 위해 euclidean distance matrix를 구하여, 그림 2와 같이 heat map으로 표시하였다. 여기서 녹색일수록 발현 수치가 유사함을 의미한다. 이 결과는 그림 1의 상관성 결과와 동일하게 CFF 내에서 낮은 distance 값을 가짐을 설명하고 있다.

4. 결론

5개의 그룹에서 생성된 데이터의 통합으로부터 만들어진 샘플별 miRNA 발현데이터는 miRNA 사이의 발현 상관성의 meta 분석을 위해 사용되었다. 거의 모든 clustered miRNA의 set이 발현상의 높은 상관성을 보였으며, 이 상관성 결과로부터 functional family를 정의 할 수 있었고, distant한 functional family들도 함께 찾을 수 있었다. 이러한 genome wide 분석은 miRNA들의 polycistronic transcript를 찾는 데 크게 기여하게 될 것이다.

감사의 글

이 논문은 산업자원부 차세대 신기술 과제 및 과학기술부 국가지정연구실사업(NRL)에 의하여 지원되었음.

References

[1] Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116,** 281-297.

[2] Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J. and Schier, A.F. (2006) Zebrafish mir-430 promotes deadenylation and clearance of maternal mRNAs. *Science*.

[3] Wu, L., Fan, J. and Belasco, J.G. (2006) From the Cover: MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A*, **103,** 4034-4039.

[4] Griffiths-Jones, S. (2004) The microRNA Registry. *Nucleic Acids Res*, **32,** D109-111.

[5] Kim, V.N. and Nam, J.W. (2006) Genomics of microRNA. *Trends Genet*, **22,** 165-173.

[6] Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E. *et al.*(2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, **37,** 766-770.

[7] Meltzer, P.S. (2005) Cancer genomics: small RNAs with big impacts. *Nature*, **435,** 745-746.

[8] Hayashita, Y., Osada, H., Tatematsu, Y., Yamada, H., Yanagisawa, K., Tomida, S., Yatabe, Y., Kawahara, K., Sekido, Y. and Takahashi, T. (2005) A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Res*, **65,** 9628-9632.

[9] Chen, J.F., Mandel, E.M., Thomson, J.M., Wu, Q., Callis, T.E., Hammond, S.M., Conlon, F.L. and Wang, D.Z. (2006) The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat Genet*, **38,** 228-233.

[10] He, L., Thomson, J.M., Hemann, M.T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S.W., Hannon, G.J. *et al.*(2005) A microRNA polycistron as a potential human oncogene. *Nature*, **435,** 828-833.