Long noncoding RNA 예측을 위한 실험적-전산학적 방법

한양대학교 ▌최서원 · 남진우

1. 서 론

인간게놈프로젝트가 끝나고 인체 내에서 모든 생물학적 기작을 이해할 수 있을 거라는 기대가 무산되었고, 이는 곧 게놈 상의 모든 기능적 요소들을 찾으려는 ENCOCE project [1] 으로 이어졌다. ENCODE project는 junk DNA로 불리던 게놈 상의 서열들 대부분이 실제로 전사가 되어 RNA로서 기능을 한다는 놀라운 결과를 발표하였고 [2], 그 이후로 가히 non-coding RNA (ncRNA)의 시대가 열렸다고 해도 과언이 아닐 만큼 ncRNA에 대한 연구가 활발히 이루어졌다. MicroRNA, PIWI-interacting RNA를 비롯해 다양한 ncRNA의 종류가 밝혀져 연구가 되고 있지만, 그 중에서도 단백질 코딩 유전자와 유사한 성질을 가지는 long non-coding RNA (lncRNA)에 관심이 집중되고 있다. lncRNA는 앞서 나열한 ncRNA와 다르게 200nt 이상의 긴 RNA로 분류되며, 단백질을 코딩하는 유전자와 마찬가지로 RNA polymerase Ⅱ에 의해 전사되어 5' cap과 3' polyA tail을 갖는다 [3,4]. 기능적으로는 비록 정확한 분자적 기작이 확인되지는 않았지만, 많은 수의 lncRNA 들이 다른 유전자들의 전사를 조절하거나, 다른 RNA들의 양을 조절하는 등의 조절인자 (regulator) 역할을 한다고 알려져 있다 [5]. 이렇게 lncRNA가 직접 관여하는 생물학적 프로세스는 매우 다양하며, 세포 주기, 세포 분화, 전사, 번역, splicing, 그리고 후성유전체적 변화에 관여하고 있다 [6-10] (그림 1). 더불어 기능이 알려진 lncRNA들은 각종 질병과의 연관성을 보이며, 특히 암에 있어서는 단순한 연관성에 그치지 않고 바이오마커로 활용될 수 있음이 알려졌다 [11].

lncRNA의 생물학적 중요성과 시공간 및 세포주 특이적으로 발현되는 성질 때문에 lncRNA를 동정하려는 노력이 계속되고 있다 (표 1). FANTOM project [12]에서 full-length cDNA sequencing을 이용하여 미보고된 많은 ncRNA를 보고하였고, 2000년 중반 들어 대규모 exon-array 또는 tiling-array 방식을 이용하여 많은 ncRNA를 발굴하기 시작하였다. 2009년에는 히스톤 단백질의 메틸화 패턴을 그린 chromatin state maps로 새로운 ncRNA 유전자를 대규모로 동정한 사례가 있었다 [6]. 또한 lncRNA 역시 RNA polymerase Ⅱ에 의해

전사되기 때문에, RNA polymerase의 occupancy도 lncRNA의 동정에 사용되었다 [13-15]. 그러나 RNA-seq을 이용하면 고해상도의 전사 신호, exon-junction 정보, 전사 방향 정보를 추가적으로 이용할 수 있으며, 다양한 이소체를 갖는 새로운 전사물을 찾아낼 수 있다는 장점이 있어 최근에는 이를 이용한 동정 연구가 대부분을 차지하고 있다. RNA-seq 기반 lncRNA 동정의

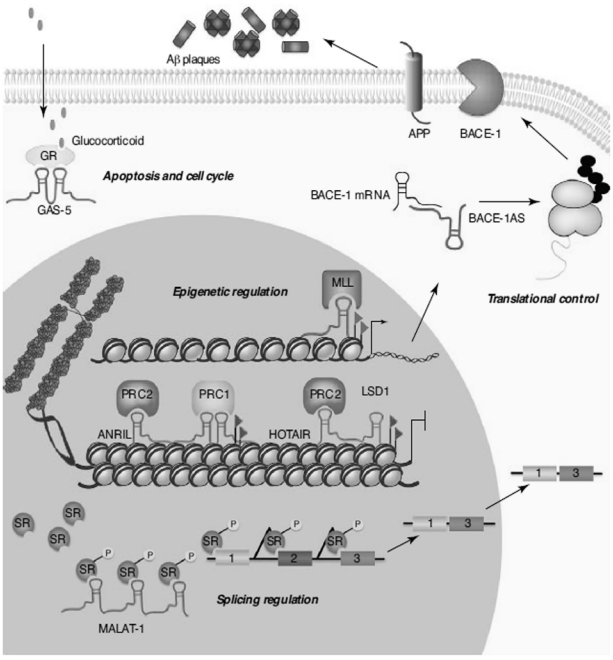

그림 1 lncRNA의 다양한 기능 [31]

† 본 연구는 미래창조과학부에서 제공하고 한국연구재단이 지원하는 연구사업 (NRF-2013R1A1A1010185)와 농촌진흥청 연구사업(세부과제번호: PJ01045303)의 지원에 의해 이루어진 것임.

표 1. lncRNA 데이터베이스 현황

Database	Version (year)	lncRNA	Web Link	Reference
VEGA	55 (2014)	13,349	http://vega.sanger.ac.uk/index.html	Wilming LG et al., Nucleic Acids Res. 2008
NONCODE	4 (2013)	95,135	http://www.noncode.org	Xie et al., Nucleic Acids Res. 2013
LNCipedia	2.0 (2013)	32,183	http://www.lncipedia.org	Volders et al., Nucleic Acids Res. 2012

일반적인 과정은 (1) 먼저 RNA-seq으로 전사체 재 구성을 하여 얻은 전사물 모델들 중 알려지지 않은 새로운 전사물을 분리해내고; (2) Ensembl이나 RefSeq과 같은 데이터베이스에서 제공하는 단백질 코딩 유전자와 위치와 구조 등을 비교하여 알려진 유전자의 새로운 isoform 들을 제거한 후; (3) 발견된 전사물의 단백질 코딩 가능성 (coding potential)을 측정하여 lncRNA와 단백질 코딩 유전자를 분리해낸다. 결국 알려지지 않은 전사물의 coding potential을 정확하게 예측하는 것이 lncRNA 동정에 있어서 가장 핵심적인 부분이다. 따라서 단순히 여러 세포주에서 lncRNA를 동정하는 것뿐만 아니라, coding potential 예측을 위한 알고리즘 개발에 주력한 연구들도 다수 진행되었다. 본 리뷰 논문에서는 기계학습 기법을 적용한 다양한 lncRNA 분류 알고리즘들을 소개하며, 특히 ribosome 신호를 이용한 다양한 lncRNA 분류 기법에 대해서 소개한다.

2. lncRNA 동정을 위한 다양한 접근 방법

2.1 lncRNA 예측을 위한 다양한 연구 분류

전사물의 coding potential을 예측하는 데에는 그 시점에서 알려진 전사물의 거의 모든 정보를 활용할 수 있다. 전사물의 염기서열 정보와 유전체 상에서의 위치, 그리고 다른 연구에서 제공되는 각종 전사체 데이터도 적용이 가능하다. 최근에는 일반적으로 하나의 정보 보다는 여러 가지의 정보를 조합하여 사용하며, 선택하는 정보의 조합은 lncRNA 동정에 대한 접근 방식에 따라 달라질 수 있다. 전산학적인 접근 방식에서는 전사물의 길이나 염기서열 자체, 진화적인 염기 변이, 코돈 사용률 등의 정보를 사용하여 확률적인 분류를 하며, 실험적인 접근 방식에서는 실험에서 얻은 데이터를 분석하여 번역율을 직접 계산한다.

2.2 전산학적 접근 방법

전산학적인 lncRNA 동정은 단백질 코딩 유전자들의 일차원적인 성질들을 조사하여 이들의 분포가 lncRNA로 의심되는 전사물에서 구한 수치와 확률적으로 얼마나 다른지를 계산한다. 예를 들어, (1) 단백질 코딩 유전자들의 ORF 길이; (2) ORF가 전체 전사물에서 차지하는 비율; (3) 염기서열에서 각 염기가 차지하는 비율; (4) 사용되는 코돈의 빈도수; (5) 코돈의 변이율 (codon substitution frequency); (6) 인접하는 코돈 쌍의 빈도수; (7) 전사물의 2차원적 구조 등의 분포를 미리 계산하여, 새로운 전사물에서의 각 수치가 그 분포와 얼마나 다른지를 측정한다. 이 중, ORF 길이와 같은 몇몇 분류자들은 상당히 정확하게 lncRNA를 분리하는 강력한 분류자로 나타나고 있다 (그림 2). 그러나 기본적으로 분포를 사용하기 때문에 하나의 성질만으로는 적절한 수준의 정확성에까지 도달하지는 못하며, 따라서 여러 성질의 조합으로 최적의 해를 찾는 기계학습 알고리즘의 구성을 필요로 한다.

lncRNA 분류를 위해 사용되는 기계학습 알고리즘은 support vector machine (SVM), random forest, logistic regression 등의 여러 기계학습 방법이 존재하지만, 다양한 regression 방법을 제공하는 kernel 개념의 도입 후 생명정보학 분야에서는 SVM이 가장 널리 사용되고 있다. 대표적으로 CPC (Coding Potential Calculator) [16]가

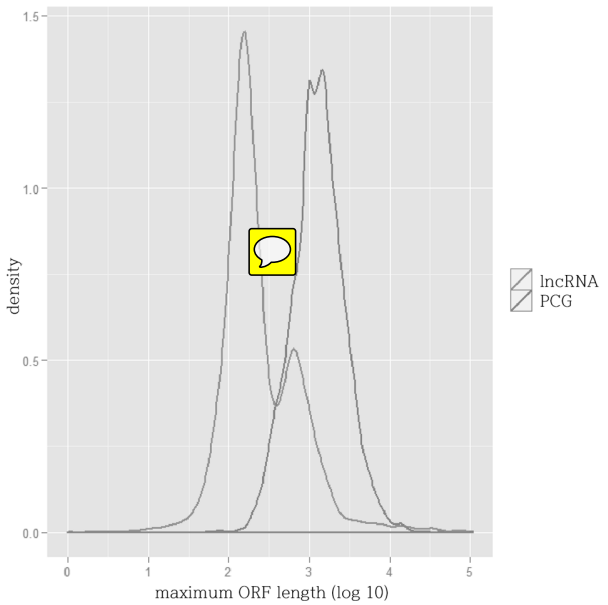

그림 2 lncRNA (GENCODE version 12) 와 단백질 코딩 유전자 (Protein-coding gene, RefSeq 2013.09.09)의 최장 ORF 길이의 분포 비교

표 2 기계학습을 이용하여 lncRNA를 동정한 연구 논문들

연구 논문	사용된 특질	기계학습 알고리즘
[18]	Peptide length, amino acid composition, secondary structure content, and protein alignment information	SVM
[19]	PhastCons score, ORF length, ORF proportion, frequencies of seven di- or tri-nucleotide sequences (GC, CT, TAG, TGT, ACG and TCG)	SVM
[20]	Length, S-score, length-percentage, score-distance, codon-bias	SVM
[21]	Sequence length, CPC, minimum free energy (MFE), frequency of three neighboring bases, content G and C, frequency of a single base, MFE/length, MFE/GC-content	Genetic Algorithm and SVM combined (GA-SVM)
[22]	TE, inside versus outside, fraction length, disengagement score	Random Forest
[23]	Maximum length of ORF, ORF coverage (proportion), Fickett TESTCODE score, hexamer score	Logistic Regression

널리 쓰이는데, 이 역시 전사물의 정보를 받아서 'nr' 데이터베이스에 대해 서열 정렬 방식 (tBlastX)으로 얻은 결과를 SVM을 통해 자체적으로 정한 단백질 코딩 점수로 반환한다. 그러나 사용하는 'nr' 데이터베이스에 등록되지 않은 새로운 단백질 코딩 유전자는 lncRNA로 잘못 분류되는 단점이 있다. 최근에는 코돈의 변이율을 기반으로 coding potential을 예측하는 phlyoCSF [17]가 개발되어 lncRNA 예측에 적용되었지만, ~~기본적으로~~ 서열의 진화적 보존성에 기반 하기 때문에, 진화상에서 새롭게 도입된 단백질 유전자를 lncRNA로 분류하거나 잘 보존된 lncRNA를 단백질 코딩 유전자로 분류하는 오류를 범할 수 있다. 전산학적 ~~접근법이~~ 위와 같이 그 한계가 분명히 ~~있음에도 불구하고~~, 그 사용 비용이 적고, 많은 오류를 선 제거할 수 있으며, 전사체 재구성에서 얻은 전사물의 기본적인 주석 정보 이외에 별도의 데이터가 필요하지 않고 속도가 빠르다는 점 때문에 지속적으로 사용과 개발이 되고 있다 (표 2).

2.3 실험적 접근 방법

실험적 접근 방법의 가장 큰 장점은 전사물의 분류에 생물학적 중요성을 반영할 수 있다는 점이다. 실험으로 생산된 데이터를 이용하면, 전사물이 실제로 세포 안에서 어떤 단백질과 어떤 방식으로 상호작용을 하는지를 알 수 있다. 특히, 단백질 코딩 유전자라면 세포 안에서 반드시 ribosome이 결합하여 번역이 일어나기 때문에 ribosome과의 상호작용은 매우 중요한 분류 기준이 된다. Ribosome과 전사물의 연관성은 최근 ribosome profiling 데이터, 또는 줄여서 Ribo-seq이라고 불리는 프로토콜이 개발되면서 활용이 가능해졌다 [24] (표 3). Ribo-seq은 cyclohexamide와 같은 약물로 번역을 억제시킨 다음, cold condition에서 세포에 RNase를 처리 한 후 ribosome과 결합에 의해 분해되지 않은 RNA 조각들을 클로닝하여 시퀀싱하는 것으로,

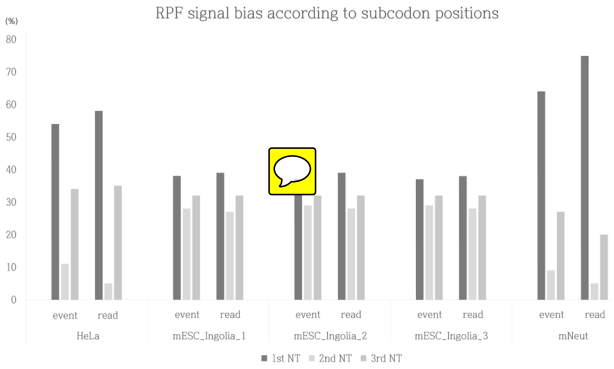

그림 3 [25], [28]에서 제공하는 데이터로 Ribo-seq reads의 5' 말단이 코돈의 첫 번째 염기에 집중적으로 나타나는 현상을 관찰한 그래프

표 3 Ribo-seq 데이터를 사용하여 lncRNA 분류를 수행한 연구 논문들

연구 논문	사용된 특질
[24]	Ribosome association
[25]	Translation efficiency
[32]	Translation efficiency
[26]	Ribosome release score
[27]	ORF score

이 데이터를 유전체에 대응시키면 ribosome과 결합되어 있던 부분을 알아낼 수 있다. 이를 이용하여 전사물의 ribosome과의 결합 정도를 측정할 수 있는데, 초기에는 단순히 Ribo-seq에서 얻은 RPKM만을 측정에 사용했다 [24]. 그러나 곧 세포 내에서의 전사물의 발현 정도에 따라 ribosome과의 결합 빈도가 달라지는 효과를 없애기 위해 RNA-seq에서 얻은 RPKM으로 보정하였고, 이를 번역 효율 (translation efficiency, TE)로 사용하였다 [25]. 하지만 알려진 일부 ncRNA들이 높은 TE 값을 가지는 것이 보고되었고, 특히 몇몇 lncRNA들이 매우 높은 TE 값을 갖는 것으로 보여 [26-27], ribosome이 번역 외의 다른 목적으로 전사물

에 결합하였을 가능성이 제기되고 있다. 따라서 추후의 연구에서는 좀 더 직접적인 번역의 신호를 알아내기 위해 ribosome과의 연관성을 하나의 염기마다 개별적으로 관찰하였다 [27]. 그 결과 단백질 코딩 유전자에서는 ribosome과의 연관성이 3개 염기마다 높게 나타나는 현상을 목격하고 이를 이용하여 coding potential의 계산은 물론이고 짧은 ORF 까지 예측한 결과를 보고하였다. 이러한 3개 염기의 주기성은 이미 2010년에 관찰된 바 있으며 [28] (그림 3), ribosome이 번역을 진행할 때 보이는 3개의 염기로 이루어진 하나의 코돈 단위의 움직임으로 보고되었다.

2.4 전망

현재까지 전산학적 방법으로 분류를 하였던 lncRNA는 대부분 유전자의 사이에 존재하는 long intergenic ncRNA로서, 주변 유전자와의 구분이 용이했다. 그러나 상당수의 lncRNA가 단백질을 코딩하는 유전자와 sense (long intervening ncRNAs), antisense (antisense ncRNAs) 방향으로 겹쳐져서 존재하며, 심지어 두 유전자의 exon이 겹치는 경우도 발견되고 있다. 이러한 전사물은 전산학적 접근 방법에서 활용되는 여러 특성들이 단백질 코딩 유전자와 유사하거나, 이를 공유하기 때문에 분리하기에 많은 어려움이 있다. 이러한 한계점은 현존하는 실험적 접근 방법으로도 완전하게 해소되지는 않는다. 현재로서는 적용 가능한 실험 데이터가 ribosome과의 연관성 외에는 거의 전무하지만, 단백질 코딩 유전자에 특이적인 상호작용이 다양하게 존재하기 때문에 현재의 한계점을 극복하는 새로운 분류자가 개발될 수 있는 가능성이 높다.

3. 결 론

lncRNA를 주석하기 위한 많은 노력이 있었지만 여전히 잘못 주석된 부분도 존재하며 표준화된 workflow도 정립되지 않았다. 실제로 발견 당시에는 lncRNA로 분류되었지만 연구가 진행되면서 단백질을 코딩 유전자로 밝혀진 전사물도 존재한다 [29-30]. 이러한 lncRNA의 정확한 동정은 질병의 발생 기작을 더욱 명확하게 이해하고, 해당 질병의 바이오마커 및 질병 치료제 타깃 발굴을 위해 선제적으로 해결되어야 할 중요한 연구이다.

참고문헌

[1] The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. Science, 2004. 306(5696): p. 636-640.

[2] Gerstein, M.B., et al., Architecture of the human regulatory network derived from ENCODE data. Nature, 2012. 489(7417): p. 91-100.

[3] Khalil, A.M., et al., Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA, 2009. 106(28): p. 11667-11672.

[4] Garber, M., et al., Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods, 2011. 8(6): p. 469-477.

[5] Ponting, C.P., et al., Evolution and functions of long noncoding RNAs. Cell, 2009. 136(4): p. 629-641.

[6] Guttman, M., et al., Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009. 458: p. 223-227.

[7] Mattick, J.S., et al., The genetic signatures of noncoding RNAs. PLoS Genet, 2009. 5: p. e1000459.

[8] Mattick, J.S., et al., Non-coding RNA. Hum. Mol. Genet, 2006. 15: p. R17-29.

[9] Qureshi, I.A., et al., Long non-coding RNAs in nervous system function and disease. Brain Res, 2010. 1338: p. 20-35.

[10] Wilusz, J.E., et al., Long noncoding RNAs: functional surprises from the RNA world. Genes Dev, 2009. 23: p. 1494-1504.

[11] Kunej, T., et al., The decalog of long non-coding RNA involvement in cancer diagnosis and monitoring. Crit Rev Clin Lav Sci, 2014. 15: p. 1-14.

[12] Kawai, J., et al., Functional annotation of a full-length mouse cDNA collection. Nature, 2001. 409(6821): p. 685-690.

[13] De Santa, F., et al., A Large Fraction of Extragenic RNA Pol Ⅱ Transcription Sites Overlap Enhancers. PLoS Biol, 2010. 8(5): p. e1000384.

[14] Garmire, L.X., et al., A global clustering algorithm to identify long intergenic non-coding RNA-with applications in mouse macrophages. PLoS ONE, 2011. 6(9): p. e24051.

[15] Sun, H., et al., Genome-wide mapping of RNA Pol-Ⅱ promoter usage in mouse tissues by ChIP-seq. Nucleic Acids Res, 2011. 39(1): p. 190-201.

[16] Kong, L., et al., CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res, 2007. 35: p. W345-W349.

[17] Lin, M.F., et al., PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 2011. 27: p. i275-i282.
[18] Liu, J., et al., Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines. PLoS Genet, 2006. 2(4): p. e29.
[19] Sun, K., et al., iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics, 2013. 14_Suppl2: p. S7.
[20] Sun, L., et al., Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res, 2013. 41(17): p. e166.
[21] Wang, Y., et al., Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm. Gene, 2014. 533(1): p. 94-99.
[22] Chew, G.L., et al., Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. Development, 2013. 140(13): p. 2828-2834.
[23] Wang, L., et al., CPAT: Coding-Potenial Assessment Tool using an alignement-free logistic regression model. Nucleic Acids Res, 2013. 41(6): p. e74.
[24] Ingolia, N.T., et al., Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 2009. 324(5924): p. 218-223.
[25] Ingolia, N.T., et al., Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell, 2011. 147(4): p. 789-802.
[26] Guttman, M., et al., Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell, 2013. 154: p. 240-251.
[27] Bazzini, A.A., et al., Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. The EMBO journal, 2014. 33(9): p. 981-993.
[28] Guo, H., et al., Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature, 2010. 466(7308): p. 835-840.
[29] Tupy, J.L., et al., Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. Proc Natl. Acad. Sci. USA, 2005. 102: p. 5495-5500.
[30] Kondo, T., et al., Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science, 2010. 329: p. 336-339.
[31] Wapinski, O., et al., Long noncoding RNAs and human disease. Trends Cell Biol, 2011. 21(6): p. 354-361.
[32] Nam, J.W., et al., Long noncoding RNAs in C. elegans. Genome Res, 2012. 22: p. 2529-2540.

약 력

최 서 원

2013 한양대학교 응용수학과 졸업 (학사)
2014~현재 한양대학교 생명과학과 석박통합과정
관심분야 : ncRNA, 전사체/유전체, 생물정보학, 기계학습
E-mail : s2009061@hanyang.ac.kr

남 진 우

2001 연세대학교 생물학과 졸업 (학사)
2004 서울대학교 협동과정 생물정보학 전공 졸업 (석사)
2007 서울대학교 협동과정 생물정보학 전공 졸업 (박사)
2007-2008 서울대학교 microRNA 창의연구소 Postdoctoral Associate
2008~2012 Whitehead Institute for Biomedical Research / MIT
2012~2014 한양대학교 의생명공학전문대학원
2014~현대 한양대학교 자연과학대학 생명과학과
관심분야 : ncRNA, 전사체/유전체, 질병/암 유전체, 생물정보학, 기계학습
E-mail : jwnam@hanyang.ac.kr