RNA 슈도유리딘 특징 추출 분석

임대영*,**, 박세희**, 정길도*,**
*전북대학교 전자공학부
**전북대학교 전자정보 신기술 연구센터
e-mail: tamudylim@gmail.com, cebast@naver.com, kitchong@jbnu.ac.kr

Analysis of feature extraction of RNA Pseudouridine.

Dae Yeong Lim*,**, Sehi Park**, Kil To Chong*,**
*Dept. of Electronics & Information Engineering, Jeonbuk National University
**Advanced Electronic Research Information Center, Jeonbuk National University

요 약

본 논문은 RNA의 슈도유리딘 여부를 판단하기 위해 시퀀스 데이터로부터 특징을 추출하는 기법에 대해 다루고 있다. EIIP, ANF, k-mer, ENAC, CKSNAP, DAC 기법에 대해 소개하고 머신러닝 기법인 RBF-SVM을 적용하여 각 추출 기법의 정확도를 계산하였다. 실험 결과 DNA/RNA 임베딩에 자주 사용되는 k-mer (59.4%)를 비롯하여 ENAC(60%)과 CKSNAP(59.2%)기법에서 좋은 성능을 보였다. 이러한 특징 추출 알고리즘은 수학적 통계와, 물리화학적 이론에 기반을 두고 있다. 따라서 슈도유리딘 여부를 정확하게 결정지을 수 있는 방법을 찾는 것은 역으로 DNA/RNA의 구성 및 특징 해석에 귀중한 정보를 제공해 줄 수 있는 기회가 될 수 있다.

1. 서론

본 연구는 암, 유전질환 등의 세포과정 이해에 필요한 슈도유리딘의 위치를 정확히 식별하기 위해 기존 방법이 갖는 비용, 노동집약적 문제점을 해결할 수 있는 슈도유리딘 식별을 목표로 한다.

DNA의 일부가 전사되어 만들어지는 리보핵산(RNA)[1]은 오탄당의 일종인 리보스를 기반으로 뉴클레오타이드를 이루는 핵산의 한 종류로 아데닌(A), 유라실(U), 사이토신(C), 구아닌(G)의 4가지 타입의 뉴클레오타이드를 갖는다. 유라실의 뉴클레오타이드의 기본 구조는 인산(P), 오탄당(ribose), 염기(base)로 구성되어 있으며, 오탄당의 1번 탄소와 염기의 1번 질소가 결합된 구조를 갖지만 슈도유리딘의 경우 RNA의 변형에 의해 오탄당의 1번 탄소와 염기의 5번 탄소가 결합된 구조를 갖는다. 슈도유리딘은[2] mRNA를 생성에 필요한 스플라이소좀(spliceosome)의 반응을 활성화 하고 결합력을 증대하며, tRNA 구조 안정화에 기여한다고 알려져 있다. 따라서, 슈도유리딜화는 RNA 기능에 영향을 주며 유전질환, 암 등에 연관성이 있는 슈도유리딘 여부를 정확하게 식별 하고 특성을 분석하는 연구가 필요하다. 논문 구성은 먼저 실험에 사용된 데이터 세트를 설명하고 각 특징 추출 알고리즘에 대해 소개 하였다. 2.3장에서는 6개의 특징 추출 알고리즘에 대해 평균 정확도를 분석하였다.

2. 본론

2.1 슈도유리딘 데이터 세트

슈도유리딘 데이터 세트는 공개데이터[3]를 사용하며 다음 표와 같은 구성으로 되어 있다. 데이터는 반절 씩 양성과 음성(슈도유리딘 여부)으로 균등하게 구성되어 있으며, 각 데이터의 중앙에 유라실(U)이 위치하여 있다.

[표 1] 데이터 세트

이름	개수	길이(nt)	용도
H.Sapiens	990	21	학습
H.Sapiens	200	21	테스트
S.Cerevisiae	628	31	학습
S.Cerevisiae	200	31	테스트
M.musculus	944	21	학습

2.2 데이터 추출 알고리즘

RNA 시퀀스 데이터는 {A, C, G, U} 4개의 뉴클레오타이드(nt)로 표현된다. one-hot 인코딩 등을 이용하여 0과 1로 표현한 데이터를 머신러닝에 사용되긴 하지만 복잡한 RNA의 특징을 찾아내는데 정보량이 부족하다. 따라서 주어진 시퀀스에 대해 통계적 방법으로 nt의 빈도나 유사성을 평가하거나 물리화학적 특징을 반영한 시퀀스 데이터 임베딩 기법들이 연구 되었다. 이번 장에서는 이러한 데이터 추출방법에 대해

소개 하도록 한다.

▪ **EIIP**(Electron-ion interaction pseudopotentials of trinucleotide)

EIIP는 뉴클레오타이드(nt)의 평균 에너지 상태로 원자의 개수와 에너지 크기를 바탕으로 계산되며, 각 nt의 값은 다음 표와 같다. 2개 이상의 nt는 합 연산으로 계산되며, 예를 들어 AG(=GA) = 0.1260 + 0.0806 = 0.2066이 된다.

[표 2] 뉴클레오타이드의 EIIP 값

Nucleotide	EIIP value(Ry)
A	0.1260
T/U	0.1335/0.0562
C	0.1340
G	0.0806

▪ **ANF**(Accumulated Nucleotide Frequency)

ANF기법은 RNA 시퀀스에서 뉴클레오타이드의 빈도 및 분포 특징을 추출하는 기법으로 식 (1)과 같다.

$$d_i = \frac{1}{|s_i|}\sum_{j=1}^{l} f(s_i), f(q) = \begin{cases} 1 \text{ if } s_i = q \\ 0 \; other case \end{cases} \quad (1)$$

예를 들어, Sq=[G, G, U, G, C, U, A, A, C, A]가 있을 때, 계산된 값은 표 3의 val 값과 같다.

[표 3] ANF 연산

i_{th}	1	2	3	4	5	6	7	8	9	10
nt	G	G	U	G	C	U	A	A	C	A
val	1	1	0.33	0.75	0.20	0.33	0.17	0.25	0.22	0.30

▪ **k-mer**

k-mer 기법은 k의 값에 따라 nt 분류세트가 결정되고, 각 시퀀스에 해당 분류 세트의 개수를 카운팅 하는 방법이다. 일반적으로 k=3을 사용하며, k에 따른 nt세트는 다음과 같다.

$k=1: t \in \{A, C, G, U\}$
$k=2: t \in \{AA, AC, AG, AU, CA, \cdots, UU\}$
$k=3: t \in \{AAA, AAC, AAG, AAU, \cdots, UUU\}$

▪ **ENAC**(Enhanced Nucleic Acid Composition)

ENAC 기법은 window(=5)를 이용하여 각 스텝에서 nt의 빈도수를 측정하는 방식이다. 길이가 31인 시퀀스가 다음과 같이 있을 때, 스텝별 계산방법은 표 4와 같다.

$sq = \{\underline{G, G, U, G, C}, U, A, A, \cdots, G, G, C, U, A\}$

[표 4] ANF 연산

Step	Calculation of frequency	Output
1	A:0/5, C:1/5, G:3/5, U:1/5	[0.0, 0.2, 0.6, 0.2]
2	A:0.5, C:1/5, G:2/5, U:2/5	[0.0, 0.2, 0.4, 0.4]
3	A:1/5, C:1/5, G:1/5, U:2/5	[0.2, 0.2, 0.2, 0.4]
Last	A:1/5, C:1/5, G:2/5, U:1/5	[0.2, 0.2, 0.4, 0.2]

▪ **CKSNAP**(Composition of k-spaced Nucleic Acid Pairs)

CKSNAP 기법은 ENAC와 마찬가지로 window(=5)를 사용하며, 2개의 nt쌍의 개수를 식 (2)를 이용하여 계산한다.

$$\left(\frac{N_{AA}}{N_t}, \frac{N_{AC}}{N_t}, \frac{N_{AG}}{N_t}, \cdots, \frac{N_{UU}}{N_t}\right)_{16} \quad (2)$$

이때, N_t는 k=0, 1일 때, 5, 4이며, k-paced는 nt쌍 사이에 빈 공간을 만드는 역할을 하며, k값에 따른 계산의 차이는 다음과 같다.

길이가 6인 시퀀스가 있을 때, $sq = \{G, G, U, G, C, U\}$

$k=0:$
$[0.0_{AA}, 0.0_{AC}, 0.0_{AG}, 0.0_{AU}, 0.0_{CA}, 0.0_{CC}, 0.0_{CG}, 0.2_{CU},$
$0.0_{GA}, 0.2_{GC}, 0.2_{GG}, 0.2_{GU}, 0.0_{UA}, 0.0_{UC}, 0.2_{UG}, 0.0_{UU}]$

$k=1:$
$[0.0_{A\_A}, 0.0_{A\_C}, 0.0_{A\_G}, 0.0_{A\_U}, 0.0_{C\_A}, 0.0_{C\_C}, 0.0_{C\_G}, 0.0_{C\_U},$
$0.0_{G\_A}, 0.0_{G\_C}, 0.25_{G\_G}, 0.5_{G\_U}, 0.0_{U\_A}, 0.25_{U\_C}, 0_{U\_G}, 0_{U\_U}]$

예를 들어, 시퀀스에서 G_U에 해당하는 nt조합은 {G,G,U}, {G,C,U} 두 개가 존재하며, k=1이기 때문에 N_t=4가 된다. 따라서 2/4 = 0.5가 계산된다.

▪ **DAC**(Dinucleotide-based Auto Covariance)

DAC 기법은 2개의 nt쌍과 기하학 정보를 가지고 있는 인덱스의 평균사이의 확률변수의 상관정도를 나타내는 방법으로, 식 (3)과 같이 정의 된다.

$$DAC(u, lag) = \sum_{i=1}^{L-lag-1} \frac{(P_u(R_i R_{i+1}) - \overline{P_u})(P_u(R_{i+lag} R_{u+lag+1}) - \overline{P_u})}{L - lag - 1} \quad (3)$$

$$\overline{P_u} = \sum_{j=1}^{L-1} \frac{P_u(R_j R_{j+1})}{L-1}$$

여기서, u는 물리화학적 인덱스 값을 lag는 autocorrelation 연산 시 우측 항에 지연된 값을 사용하는데 사용된다. u의 인덱스 값은 그림 3과 같이 nt 한쌍의 기하학적 물리량을 나타낸다.

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
Rise	3.18	3.24	3.3	3.24	3.09	3.32	3.3	3.3	3.38	3.22	3.32	3.24	3.26	3.38	3.09	3.18
Roll	7	4.8	8.5	7.1	9.9	8.7	12.1	8.5	9.4	6.1	12.1	4.8	10.7	9.4	9.9	7
Shift	-0.08	0.23	-0.04	-0.06	0.11	-0.01	0.3	-0.04	0.07	0.07	-0.01	0.23	-0.02	0.07	0.11	-0.08
Slide	-1.27	-1.43	-1.5	-1.36	-1.46	-1.78	-1.89	-1.5	-1.7	-1.39	-1.78	-1.43	-1.45	-1.7	-1.46	-1.27
Tilt	-0.8	0.8	0.5	1.1	1	0.3	-0.1	0.5	1.3	0	0.3	0.8	-0.2	1.3	1	-0.8
Twist	31	32	30	33	31	32	27	30	32	35	32	32	32	32	31	31

[그림 1] A physicochemical index

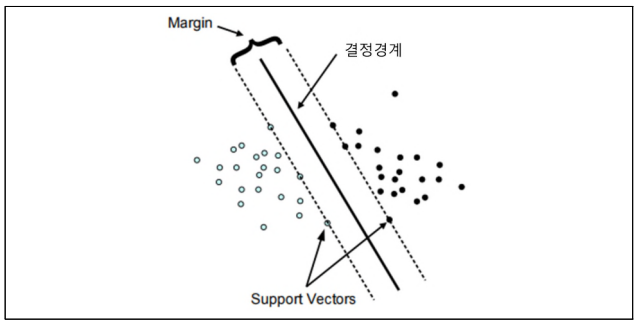
이 장에서 설명한 특징 추출알고리즘은 iLearn에서 제공하

는 함수를 사용하였다. (https://github.com/Superzchen/iLearn)

2.3 추출 특징 정확도 분석

▪ SVM(Support vector machine)

SVM 알고리즘은 대표적인 머신러닝 기법중 하나로 딥러닝에 비해 간단하고 효과적인 퍼포먼스를 보여주기 때문에 2.2장에서 소개한 기법들의 단순 성능 비교에 적합하다.

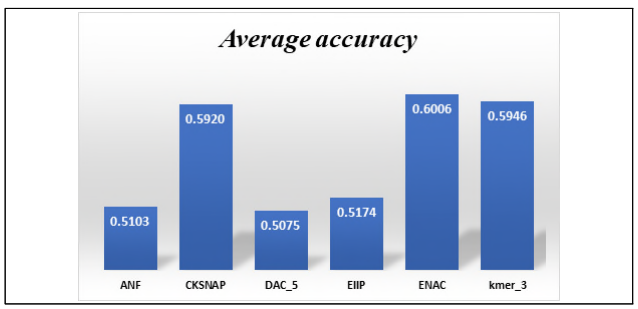


[그림 2] 마진, 결정 경계, 서포트 벡터[4]

SVM 알고리즘은 데이터를 특징에 따라 분류하기 위해 결정 경계를 사용한다. 이때, 오버피팅을 방지하기 위해 마진이 가장 큰 선을 찾도록 한다. 마진은 어느 한쪽 특징에 치우치지 않고 중간의 값을 선택한다. SVM은 선형분류와 비선형 분류로 구분되는데, 여기에서는 RBF-커널을 사용한 비선형 분류기법을 사용하였다. RBF SVM은 최적화를 위해 C와 gamma의 파라미터 값을 갖는다. C는 결정경계로 나누어진 특징 집단에 다른 클래스의 값이 들어오는 것을 얼마나 허용할 것인지를 결정하고, gamma는 결정 경계의 곡률을 결정하는데 사용된다. 적절한 C와 gamma 파마리터 조정을 통해 오버피팅을 방지 하고 성능을 향상 시킬 수 있다. 파라미터 값 결정은 grid_search 기법을 이용하여 최적의 C와 gamma 파라미터 값을 설정하였다.

▪ 데이터셋 성능 분석

데이터 추출 알고리즘을 통해 임베딩된 데이터를 SVM 학습시킨 결과는 그림 3과 같다.



[그림 3] 특징 추출 알고리즘 별 평균 정확도

슈도유리딘 데이터 세트의 경우 ENAC, k-mer, CKSNAP 알고리즘이 평균적으로 높은 정확도를 내었다. 현재 RNA 시퀀스 데이터 세트로부터 슈도유리딘 유무를 구별하는 논문들은 다양한 특징 추출 방법을 병합하고 추가 알고리즘을 적용하여 약 70%의 정확도를[5] 내고 있음을 감안했을 때 ENAC 단독 기법의 60% 정확도는 비교적 높은 편에 속한다.

3. 결론

본 논문에서는 RNA의 슈도유리딘 여부를 식별하기 위해 데이터를 확장하고 새로운 특징을 보여줄 수 있는 6가지의 특징 추출 알고리즘에 대해 살펴보고 정확도를 측정해 보았다. 실험 결과 DNA/RNA 임베딩에 흔히 사용되는 k-mer (59.4%)를 비롯하여 ENAC(60%)과 CKSNAP(59.2%)기법에서 좋은 성능을 보였다. 이러한 특징 추출 알고리즘은 수학적 통계와, 물리화학적 이론에 기반을 두고 있다. 따라서 슈도유리딘 여부를 정확하게 결정지을 수 있는 방법을 찾는 것은 역으로 DNA/RNA의 구성 및 특징 해석에 귀중한 정보를 제공해 줄 수 있는 기회가 될 수 있다.

현재 모든 특징 추출 기법에서 계산된 결과물을 병합하여 데이터 세트를 구축하였고, 정보량에 따른 우선순위를 계산할 수 있는 CHI2, xgboost와 같은 알고리즘을 적용한 연구를 진행하고 있다.

Acknowledgements

이 논문은 2017년도 정부(과기정통부)의 한국연구재단 뇌과학원천기술개발사업(NRF-2017M3C7A1044816) 및 2019년도 정부(교육부)의 한국연구재단의 지원(2019R1A6A3A01094685)을 받아 수행된 연구임.

참고문헌

[1] 위키백과, “RNA”, https://ko.wikipedia.org/wiki/RNA, 2020년

[2] Michael Charette Michael W. Gray, “Pseudouridine in RNA: What, Where, How, and Why”, IUBMB Journals, 제 49권 5호, pp. 341-351, 1월, 2008년.

[3] 슈도유리딘 데이터셋: http://www.bioml.cn/data.html

[4] 스카이비전, “서포트 벡터 머신”, 10월, 2017년.

[5] Zhibin Lv, Jun Zhang, Hui Ding and Quan Zou, “RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites”, 8권, 134호, 2월, 2020년.