http://dx.doi.org/10.14400/JDC.2016.14.4.349

유전체 생태계 분석을 위한 알고리즘 구현: 미토콘드리아 사례

최성자*, 조한욱**
충남대학교 융복합시스템공학과*, 충남대학교 전기 · 전자 · 통신공학교육과**

The Algorithm of implementation for genome analysis ecosystems : Mitochondria's case

Sung-Ja Choi*, Han-Wook Cho**
Dept. of Convergence System Engineering Chungnam National University*
Dept. of Electric, Electronic and Comm. Eng. Edu. Chungnam National University**

요 약 융복합 패러다임의 도입은 방대한 유전체 정보의 분석을 위한 컴퓨팅 기술의 연구 및 개발 또한 활발히 진행되고 있다. 최근 유전체 분석 서비스 유형은 개인의 유전체 정보(personal genome analysis)를 읽어서 특정 질환들의 발병 확률 등을 알려주고, 해당 질병을 예방할 수 있도록 식습관, 라이프 스타일등의 변화를 꾀하도록 맞춤형의 서비스를 제공하고 있다. 생물의 특성을 결정하는 정보는 유전자이며, 이 유전자는 DNA 염기서열에 따라 결정되므로, 유전체 정보의 분석기술은 정확하고 빠르게 수행되어야 한다. 정확한 유전체 분석을 빠르게 수행하기위해 K-Mean 클러스터링 기법을 활용하였으며, 코돈 데이타 패턴을 추출하여 유전체 정보 분석에 적용하였다. 또한, 미토콘드리아 데이타군을 실험사례로 제공한다. 본 연구의 결과, 제공된 분석 데이타를 통해 기존의 문자열 형태의 유전체 분석 기법을 이미지 패턴 형태로 추출이 가능하며, 패턴형태의 이미지는 분석시간의 단축과 정확도를 높인다.

주제어 : 바이오인포메틱스, 클러스터링, K-Mean, 유전체학, 헬스케어

Abstract The studies on the human environment and ecosystem analysis is being actively researched. In recent years, The service of genome analysis has been offering the customized service to prevent the disease as reading an individual's genome information. The genome information by analyzing technology is being required accurate and fast analyses of ecosystem-dielectrics due to the spread of the disease, the use of genetically modified organism and the influx of exotic. In this paper the algorithm of K-Mean clustering for a new classification system was utilized. It will provide new dielectrics information as quickly and accurately for many biologists.

Key Words : Bio Informatics, Clustering, K-Mean, Genomics, Health care

* 본 논문은 2015 년 충남대학교의 학술연구비에 의하여 지원되었음(No. 2015115601)
Received 26 February 2016, Revised 28 March 2016
Accepted 20 April 2016, Published 28 April 2016
Corresponding Author: Han-Wook, Cho
(Chungnam National University)
Email: hwcho@cnu.ac.kr

ISSN: 1738-1916

1. 서론

최근 바이오기술과 컴퓨팅 기술의 융합을 통한 새로운 패러다임이 등장하고 있으며[1,2], 대규모 바이오 데이타의 생성과 이를 분석하기위한 컴퓨팅 기술의 중요성이 더욱 증가하고 있다. 개개의 유전자나 단백질을 분석하는 기존 연구방식에서 유전체(genome) 및 단백질(proteme) 전체를 다루는 Omics 연구, 세포내 전체 분자들의 상호작용망의 분석이 가능한 네트워크 생물학 또는 시스템 생물학 연구가 이루어지고 있다.

유전체 정보의 분석을 위해 시퀀싱 과정을 거쳐야 하는데 염기의 순서를 화학적으로 읽어서 염기서열해독과정을 수행한다. 시퀀싱기법은 생명체가 가지고 있는 전체 유전체를 해독하는 방법이며, 이를 통해 서로 다른 개체들간의 특이적인 변이를 찾거나 질환 특이적인 변이를 찾는데 유용하게 적용되고 있다. 시퀀싱 기법의 응용 예는 다음과 같다. 질병진단을 위한 Maker 개발에 활용 가능한 DNA Marker, 유전체 전체 지도 구축과 양적 형질 연구, HapMap construction & GWAS, 돌연변이에 대한 연구(Mutation Research), 개체의 진화와 타개체간의 진화연구(Evolution Stury)에 응용되고 있다. 대표적인 시퀀싱 기법중의 하나인 Sanger 기법은 생물체(대장균, 박테리아 등)을 증폭하여 순차적인 촬영과 이미지를 판독하여 상대적으로 긴 Read를 사용한다. 정확성이 높은 반면 매우 번거롭고 증폭량이 적으며, 시간과 비용이 많이 든다. 최근 시퀀싱 기법으로는 NGS(Next Generation Sequencing) 기법이 활발히 적용되고 있다. NGS는 Nano 기술의 비생물체 증폭방식으로 다중 촬영/판독 처리 방식이며 상대적으로 짧은 Read를 사용한다. 빠른 처리속도와 대량 증폭이 가능하며, 저렴한 처리 비용을 강점인 반면, 중복이 짧은 대량의 Read의 조합/데이타 분석이 어렵다.

또한, 유전정보 분석을 위해 클러스터링 기법을 적용하는데, 클러스터링을 통해 유사한 기능을 갖는 유전자군을 찾아낸다. Tamayo[3]는 SOM technique를 사용하여 대식세포 분화에 관해 알려진 관련 유전자의 수를 포함한 클러스터 평균치를 보였으며, Alizadeh[4]는 유전자 발현 프로파일링에 의해 식별 확산 큰 B 세포 림프종 유형을 구별하기 위해 클러스터링 기법을 사용하였다. 또한, Furlong[5]은 야생형 대 돌연변이의 유전자 발현 변화 패턴 을 찾기 위해 클러스터링을 사용하였다. 대표적인 클러스터링 알고리즘은 계층응집성 클러스터링[14,15,16], K-Mean 클러스터링[6], 자기조직화 지도[7] 등이 있으며, 본 논문에서는 매틀랩을 활용하여 유전체 생태계 정보를 K-Mean 클러스터링기법을 적용하여 구현하였다. 미트콘드리아 유전체정보는 유전자 재조합이 거의 없기 때문에 실험 데이터로 활용하였으며, 제시된 알고리즘을 활용하여 코돈 조합의 데이터 군을 클러스터링화 하였고, 유전체 분석 패턴이 가능한 이미지를 생성하여 결과를 확인할 수 있다.

2. K-Mean 클러스터링과 실험데이타군

클러스터링은 계층적 구조와 분할적 구조로 구분되며, 계층적 구조는 단일 링크와 완전링크기법이 있다. 또한, 분할적 구조를 사용한 클러스터링은 그래프이론, 혼합해법, 모드탐색, 스케어 에러등의 기법을 적용한다.

2.1 K-Mean 클러스터링 알고리즘

K-Mean 클러스터링 알고리즘은 분할법을 적용한다. 분할법은 주어진 데이타를 여러 그룹으로 분할하며, 입력된 데이타 보다 작거나 같은 k개의 군집으로 나눈다. 그룹을 나누는 과정에서 거리기반의 그룹 간 비유사도 방식으로 비용함수를 최소화한다. 즉, 같은 그룹 내의 데이타 오브젝트끼리의 유사도는 증가하고, 다른 그룹의 데이타 오브젝트의 유사도는 감소한다. 수식1은 i번째 클러스터의 중심을 μi, 클러스터에 속하는 점의 집합을 Si 일 때 최소값의 Si를 찾는다.

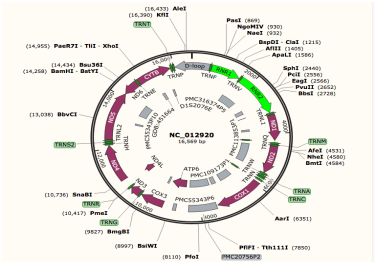
$$V = \sum_{i=1}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2 \quad (1)$$

2.1.1 미토콘드리아 유전체 데이타군

유전체분석[8,9]을 위한 미토콘드리아 데이타 군을 적용하였으며, 미토콘드리아의 기본 데이타 군을 분석하기 위해 MatLab 2015b 버전을 사용하여 분석하였다. [Fig. 1]에서는 미트콘드리아 유전체정보를 [Fig. 2]에서는 유전체 정보간의 연관성을 지도로 보여준다.

```
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTC
TGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATC
TGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACAT
ACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGC
ACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTC
```

[Fig. 1] Mitochondria's genome



[Fig. 2] Mitochondira's map

2.2 코돈(Codon) 조합

코돈은 아미노산을 지정하여 유전정보를 표현하며 [10,11,12], 미트콘드리아의 유전체 정보는 텍스트 유형으로 구성되어 있으며, 코돈의 조합을 <Table 1>에서 보여준다. 코돈은 유전자 발현에서 하나의 아미노산을 지정하는 전령 RNA 유전정보이다. DNA 유전체 정보로부터 구해진다.

〈Table 1〉 Codon data combination

U	UUU PhenylAlanine	UCU Serine	UAU Tyrosine	UGU Cysteine
	UUC PhenylAlanine	UCC Serine	UAC Tyrosine	UGC Cysteine
	UUA Leucine	UCA Serine	UAA -	UGA -
	UUG Leucine	UCG Serine	UAG -	UGG Tryptophane
C	CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine
	CUC Leucine	CCC Proline	CAC Histidine	CGC Arginine
	CUA Leucine	CCA Proline	CAA Glutamine	CGA Arginine
	CUG Leucine	CCG Proline	CAG Glutamine	CGG Arginine
A	AUU IsoLeucine	ACU Threonine	AAU Asparagine	AGU Serine
	AUC IsoLeucine	ACC Threonine	AAC Asparagine	AGC Serine
	AGC Serine	ACA Threonine	AAA Lysine	AGA Arginine
	AUG Methionine	ACG Threonine	AAG Lysine	AGG Arginine
G	GUU Valine	GCU Alanine	GAU AsparticAcid	GGU Glycine
	GUC Valine	GCC Alanine	GAC AsparticAcid	GGC Glycine
	GUA Valine	GCA Alanine	GAA GlutamicAcid	GGA Glycine
	GUG Valine	GCG Alanine	GAG GlutamicAcid	GGG Glycine

3. 실험

클러스터링 예측 및 분석을 위해 제공된 K-Mean 함수에 대한 적용알고리즘은 다음과 같다.

(1) 미트콘드리아 데이타군에서 k개의 데이타오브젝트 임의 추출

(2) 클러스터의 중심 설정(초기값지정)

(3) 각 데이타 오브젝트에 대해 k개의 클러스터 중심 오브젝트와의 거리 구함

(4) 찾아낸 중심점에 각 데이타 오브젝트 할당

(5) 클러스터 중심점 재계산

(6) 소속 클러스터가 바뀌지 않을 때 까지 2, 3과정 반복

클러스터링 데이타 셋을 얻기 위해 genBank 데이타군의 휴먼 미트콘드리아 유전체의 클러스터링 셋을 얻기 위한 코드를 수행하였으며, 적용된 알고리즘을 수행하여 <Table 2>의 코돈 클러스터링 데이타 셋을 추출하였다. 코돈 데이타 셋은 리신(AAA)가 167의 오브젝트 군으로 구성되어 있으며, 아스파라긴(AAC)는 171개의 오브젝트 군으로 구성된 데이타 셋을 확인할 수 있다. 알고리즘을 통해 수행된 코드는 다음과 같다.

〈Table 2〉 Codon clustering data set

AAA- 167	ACA - 137	AGA - 59	ATA - 126
AAC - 171	ACC - 191	AGC - 87	ATC - 131
AAG - 71	ACG - 42	AGG - 51	ATG - 55
AAT -130	ACT - 153	AGT - 54	ATT - 113
CAA - 146	CCA - 141	CGA - 40	CTA - 175
CAC - 145	CCC - 205	CGC - 54	CTC - 142
CAG - 68	CCG - 49	CGG - 29	CTG - 74
CAT - 148	CCT - 173	CGT - 27	CTT - 101
GAA - 67	GCA - 81	GGA - 36	GTA - 43
GAC - 53	GCC - 101	GGC - 47	GTC - 26
GAG - 49	GCG - 16	GGG - 23	GTG - 18
GAT - 35	GCT - 59	GGT - 28	GTT - 41
TAA - 157	TCA - 125	TGA - 64	TTA - 96
TAC - 118	TCC - 116	TGC - 40	TTC - 107
TAG - 94	TCG - 37	TGG - 29	TTG - 47
TAT - 10	TCT - 103	TGT - 26	TTT - 78

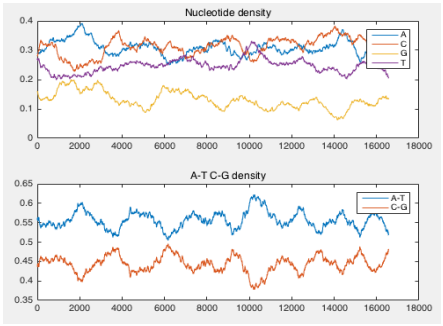
[실행소스 코드]

```
mitochondria = getgenbank('NC_012920','SequenceOnly',true);
load mitochondria
ntdensity(mitochondria)
basecount(seqrcomplement(mitochondria))
codoncount(mitochondria)
seqshoworfs(mitochondria);
rng(1)
num = xlsread('codonMitodata.xlsx',1)
x=num(1:end,1)
rng(1)
[idx, C]=kmeans(x,4);
cluster1 = x(idx ==1, :);
cluster2 = x(idx ==2, :);
cluster3 = x(idx ==3, :);
cluster4 = x(idx ==4, :);
```

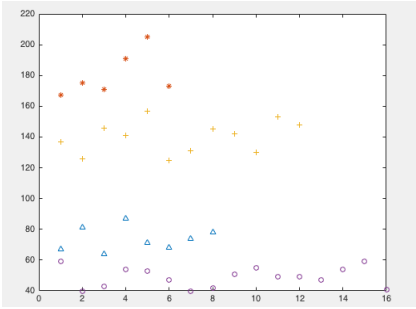
인간의 몸은 약 60조개의 세포들로 구성되어 있으며, 0.02-0.03밀리미터의 세포막으로 구성되어 있으며, 핵과 세포질이 세포막 안에 있다. 핵 안에 인이 중심을 기준으로 둘러져 있고 인 주위에는 염색체(chromosome)이 꼬인 실타래 모양으로 둘러 싸여 있다[13]. 핵을 둘러싸는 세포질에는 유전암호를 해독하여 단백질을 합성하는 리보좀, 세포의 에너지를 공급하는 미토콘드리아, 생산된 단백질을 저장하는 골지체, 독을 제거하는 리소좀, 핵과 세포질 사이를 출입하는 소포체등이 있다. 본 논문에서는 미트콘드리아의 하나의 아미노산을 지정하는 세 개의 뉴클레오드 배열의 코돈(Codon)을 활용하여 클러스터링화 하였다.

4. 분석

실험을 통해 클러스터링 군을 구성하였으며, A,C,G,T 염기군에 대한 기본 밀집도에 대한 분석그래프를 [Fig. 3]에서 보여주고 있다. 본 논문에서는 휴먼 미트콘드리아의 K-Mean 클러스터링을 통해 분석된 클러스터링 군 밀집도를 [Fig. 4]에서 보여주고 있다. 클러스터링 군을 표현하기 위해 {^,*,+,o} 문자셋을 클러스터링 표현을 위해 플럿상에 구분하였으며, cluster1, cluster 2, cluster3, cluster4에 각각 매핑 되었다. 클러스터링 결과 화면을 통해, 코돈 데이타 셋은 뉴클레오드유형의 새로운 데이타 패턴을 제공하며, 코돈 데이타 셋을 통해 유전체 분석 및 비교가 가능하며, 패턴의 유형에 따라 유전체 예측이 가능할 것으로 보여 진다.



[Fig. 3] Density of ucleotide & A-T-C-G



[Fig. 4] Result view of clustering

5. 결론

텍스트 유형의 유전체정보를 코돈 조합을 추출하여 클러스터링 후 이미지 형태의 패턴을 생성하여 유전체 정보를 분석하였다. 이는 기존의 텍스트 유형의 방법보다 빠르고 손쉽게 유전체 정보를 추출 할 수 있다. 향후 연구방향으로 쥐의 미트콘드리아 코돈 데이타 셋과 휴먼 미트콘드리아의 코돈 데이타 셋의 클러스터링 패턴 비교를 통해 유전체 예측의 솔루션을 제공하고자 한다.

REFERENCES

[1] Keun-Ho Lee, “A Method of Defense and Security

Threats in U-Healthcare Service", Journal of the Korea Convergence Society, Vol. 3, No. 4, pp. 1-5, 2012.

[2] Eun-Hee Park, Hye-Suk Kim, Ja-Ok Kim, "The Effect of Convergence Action Learning techniques in Simulation Class", Journal of the Korea Convergence Society, Vol. 6, No. 5, pp. 241-248, 2015.

[3] P. Tamayo et al., Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, PNAS 96: 2907-12, 1999.

[4] A. Alizadeh et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403: 503-11, 2000.

[5] E. Furlong et al., Patterns of Gene Expression During Drosophila Development, Science 293: 1629-33, 2001.

[6] Evanno, Guillaume, Sebastien Regnaut, and Jérôme Goudet. "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." Molecular ecology 14.8 (2005): 2611-2620.

[7] Oja, Merja, Samuel Kaski, and Teuvo Kohonen. "Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum." Neural computing surveys 3.1 (2003): 1-156.

[8] Wang, K. et al. Monitoring gene expression pro®le changes in ovarian carcinomas using cDNA microarray. Gene 229, 101±108 (1999).

[9] Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA 95, 14863±14868 (1998).

[10] Nakamura, Yasukazu, Takashi Gojobori, and Toshimichi Ikemura. "Codon usage tabulated from international DNA sequence databases: status for the year 2000." Nucleic acids research 28.1 (2000): 292-292.

[11] Crick, Francis HC. "Codon—anticodon pairing: the wobble hypothesis." Journal of molecular biology 19.2 (1966): 548-555.

[12] Ikemura, Toshimichi. "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system." Journal of molecular biology 151.3 (1981): 389-409.

[13] Wellmer, Frank, and José Luis Riechmann. "Gene network analysis in plant development by genomic technologies." International Journal of Developmental Biology 49.5/6 (2005): 745.

[14] Jobson, J. (1992) Applied Multivariate Data Analysis: Categorical and Multivariate Methods (Springer, NewYork).

[15] Hartigan, J. (1975) Clustering Algorithms (Wiley, New York).

[16] Gordon, A. E. (1981) Classification: Methods for the Exploratory Analysis of Multivariate Data (Chapman & Hall, New York).

최 성 자(Choi, Sung Ja)

- 1991년 2월 : 한남대학교 컴퓨터공학과(공학사)
- 1997년 2월 : 한남대학교 컴퓨터공학과(공학석사)
- 2005년 8월 : 한남대학교 컴퓨터공학과(공학박사)
- 관심분야 : Bio 센서, 뇌공학
- E-Mail : irecomm@naver.com

조 한 욱(Cho, Han Wook)

- 2002년 2월 : 충남대학교 전기공학교육과 (공학사)
- 2004년 2월 : 충남대학교 전기공학과 (공학석사)
- 2007년 8월 : 충남대학교 전기공학과 (공학박사)
- 2007년 9월 ~ 2010년 8월 : 한국기계연구원 시스템엔지니어링연구본부 선임연구원
- 2010년 8월 ~ 현재 : 충남대학교 전기·전자·통신공학교육과 부교수
- 관심분야 : 전기기기, 전력전자
- E-Mail : hwcho@cnu.ac.kr