Journal of KIIT. Vol. 18, No. 2, pp. 33-43, Feb. 28, 2020. pISSN 1598-8619. eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.2.33

유전자 유사성을 통하여 naive T 면역세포에 특이적인 새로운 발현 유전자 탐색법

고동우*, 양정진**

New Exploration to Identify the Naive T Cell-Specific Gene Using Gene Similarity

Dong-Woo Ko*, Jung-Jin Yang**

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (No. 2016K1A1A2912755).
이 연구는 2019년도 가톨릭대학교 교비연구비의 지원으로 이루어졌음.

요 약

최근에는 생명공학분야는 인공지능분야와의 접목을 통하여 새로운 유전자의 발굴에 큰 성과를 이루고 있으나, 빅 데이터의 처리 및 세포특이성을 부여할 수 있는 새로운 알고리즘의 부족으로 여전히 많은 한계성을 가지고 있다. 본 연구에서는 naive T세포에 특이적인 유전자인 CD27, SELL, IL7R, IL2RG, PTPRC을 기준으로 하여, 유전사들 사이의 유사성을 계산하는 새로운 방법인 Pearson Correlation과 T-SNE를 기존에 사용되고 있는 Fold Change 정렬 방법과 비교분석하였다. 시도한 3가지 분석방법들 중 T-SNE 분석방법을 사용하였을 때, 위의 5개의 유전자들이 모두 가장 높은 순위에 위치하였다. 따라서 특정한 세포에 존재하는 유전자들 중 단백질을 인코딩하는 유전자들만 선별하여 유전자간의 유사성을 계산하는 T-SNE 분석 방법을 사용한다면 타겟 세포의 선택적으로 발현하는 유전자 발굴을 위한 가장 효과적인 방법이 될 수 있음을 제시했다.

Abstract

Recent advancement of DNA/RNA sequencing technology/bioinformatic tools and establishment of various big database combining AI algorithm has made the identification of the essential genes successful. However, the lack of big data processing method and the new algorithm to provide the target cell-specificity need to be overcome. In this study two approaches, Pearson Correlation and T-SNE adopting the calculation of gene similarity, were compared with the current fold change method on the basis of naive T cell-specific surface marker genes (CD27, SELL, IL7R, IL2RG, PTPRC). Among these three methods T-SNE method positioned all of naive T cell-specific genes in the highest rank. Therefore, our results suggest that T-SNE using gene similarity calculation together with the selection of the candidate genes encoding the proteins is the most effective strategy to identify the novel genes specific to the target cells.

Keywords
naive T cells, genes, differential expression, fold change, similarity, pearson correlation, T-SNE

* 가톨릭대학교 컴퓨터공학과 석사과정
- ORCID: https://orcid.org/0000-0003-1444-5038
** 가톨릭대학교 컴퓨터공학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1191-6255

· Received: Jan. 31, 2020, Revised: Feb. 12, 2020, Accepted: Feb. 15, 2020
· Corresponding Author: Jung-Jin Yang
Dept. of Computer Science and Information Engineering
The Catholic University of Korea, Bucheon, 14662, Korea
Tel.: +82-2-2164-4377, Email: jungjin@catholic.ac.kr

Ⅰ. 서 론

인간의 면역세포들은 외부로부터의 다양한 병원균 및 암세포에 대하여 특이적이고 다양한 면역반응들을 유도하여 우리의 몸을 보호하는 가장 중요한 생체현상중의 하나이다. 따라서 면역세포의 다양한 subset들에서 특이적으로 발현하는 유전자를 찾는 것은 다양한 면역반응들에 대한 의생명 기초 연구뿐만 아니라 새로운 면역 질환 치료제 개발 분야에서도 핵심적인 문제이다. 면역세포들 중 naive T 세포는 면역조절기능을 하는 다양한 면역세포 subset들로 분화되는 핵심 면역세포로서, naive T 세포에 특이적으로 발현하는 유전자들의 발굴은 전체적인 면역반응을 이해하고 새로운 면역 질환 치료제의 개발에 핵심적인 과제이다.

특정한 세포에서 선택적으로 높게 발현하는 유전자들의 발굴연구는 다양한 유전자 분석기법, 빅 데이터의 생성과 분석 및 생명정보학의 발전과 더불어 최근에는 인공지능 분야와의 접목을 통하여 새로운 많은 연구개발 전략들이 소개되고 있다. 그러나 면역세포들에서 특이적으로 발현하는 유전자들에 대한 빅 데이터의 처리방법과 면역세포에 특이적으로 발현하는 유전자의 선발 알고리즘이 부족하여, 기존의 연구방법에 의하여 선발된 유전자들을 실험에 의하여 검증하였을 때 대부분 false positive로 판명되고 있다. 또한 면역세포에 특이적으로 발현하는 유전자들 가운데 세포표면에 존재하는 단백질들을 인코딩하는 유전자들만이 면역항체신약으로 개발이 가능하기 때문에 이를 위한 선발알고리즘이 추가되어야 하는 실정이다.

현재 생물학 분야에서 특이적인 유전자를 찾기 위해서는 mRNA의 서열들을 분석한 Raw data를 사용하여 전처리한 후 RPKM, FPKM, TPM 등의 정규화된 값을 사용한다. 이 값을 사용하여 절대적인 발현 값이 큰 순서부터 정렬하거나 Fold Change 값의 크기를 비교하는 방법으로 특이적으로 발현하는 유전자를 찾는다. 이러한 방법은 단순한 크기 비교이기 때문에 특이적인 발현을 하는 유전자를 찾는데는 많은 한계점이 있다.

특이적 발현 유전자를 찾는 또 다른 방법으로는 Deseq2와 edgeR라는 두 가지 툴이 있다[1][2]. 두 가지 툴 모두 특이적 발현 유전자를 찾을 때 실험군과 대조군 데이터를 사용한다. 대조군의 평균 발현량 대비 실험군의 평균 발현량의 증감을 계산하여 특이적 발현 유전자를 찾는다. 정상 조직 세포와 암 조직 세포를 비교한다고 했을 때, 정상 조직 세포 대비 암 조직 세포에서 높은 유전자를 찾아 암 조직 세포 특이적 발현을 갖는 유전자로 분류하는 것이다. 해당 방법은 특이적 발현 유전자를 찾는데 많이 사용되지만 실험군(암 조직 세포)과 대조군 (정상 조직 세포) 데이터가 모두 필요하다. 또한 실험군과 대조군에 사용되는 데이터가 한 개의 세포가 아닌 다양한 세포들이 혼합된 조직을 이용한 데이터이기 때문에 암세포에 특이적으로 발현하는 유전자를 찾는데 많은 한계점들을 가지고 있다.

본 연구에서는 면역세포 중 naive T 세포에 특이적이면서 세포표면 단백질을 인코딩하는 유전자를 발굴하는 새로운 방법을 제시하기 위하여 빅 데이터를 처리하는 새로운 과정과 인공지능에서 사용되고 있는 알고리즘이 접목된 두 가지 방법을 비교분석하였다. 빅 데이터에서 제공된 mRNA 서열분석 데이터를 통해 얻어진 read counts를 이용하여 각 유전자의 발현 정도를 확인할 수 있다. 그러나 각 유전자의 길이에 따라 맵핑된 read counts 수가 다르기 때문에, 본 연구에서는 kallisto를 이용하여 각 유전지의 길이로 정규화된 값인 TPM을 사용하였다. 또한 naive T 세포의 표면에 발현하는 단백질들만을 인코딩하는 유전자들을 분류한 다음, 기존의 단순히 Fold Change 값을 이용하는 방법과 추천 시스템의 기반이 되는 '유사성 계산'방법 두 가지를 사용하여 naive T 세포에서 특이적으로 발현하는 유전자를 찾는 전략을 사용하였다[3]-[6].

유사성을 계산하기 위한 방법으로는 Pearson Correlation을 사용하여 유전자 사이의 선형 상관 계수를 계산하였으며, 다른 방법으로는 T-SNE를 사용하여 유전자 발현 값을 기준으로 유전자 사이의 거리를 측정하였다[7][8]. 기존 연구들로부터 naive T 세포의 표면에 특이적으로 발현한다고 알려진 5개의 유전자들 (CD27, SELL, IL7R, IL2RG, PTPRC)를 기준으로 하여 비교분석한 결과, T-SNE를 이용한 방법이 가장 정확한 것으로 결론을 내리게 되었다[9]-[13].

II. 연구 데이터의 출처 및 방법론

2.1 연구 데이터 출처

본 연구를 위하여 인간의 면역세포 유전자 데이터를 사용하였으며, T 면역세포의 subset 중 naive T 세포, Th0 (활성화된 T 세포), Th1 또는 Th2 세포에서 발현하는 mRNA 염기서열 데이터는 genomic database인 GEO (https://www.ncbi.nlm.nih.gov/geo/)에서 다운로드하여 사용하였다.

2.2 Raw data를 TPM value로의 전환

면역세포에 특이적으로 발현하는 유전자의 발현 정도를 구하기 위해서는 제공된 mRNA 염기서열 데이터를 통해 read counts를 구할 수 있으나, 각 유전자의 길이에 따라 맵핑된 read counts 수가 다르기 때문에 read counts로 발현 값을 정의하기는 어렵다. 생물학 분야에서 사용하는 정규화 값으로는 read counts를 정규화한 Reads per kilobase of transcript per million mapped reads (RPKM), Fragment per kilobase of transcript per million mapped reads (FPKM) 및 Transcripts per million (TPM)이 있다. 본 연구에서는 유전자의 길이를 고려하여 정규화한 TPM값을 사용하였다.

세포에서 mRNA 염기서열 데이터를 각 유전자에 맵핑하여 정규화 발현 값으로 전환하는 툴에는 여러 가지가 있으나, 본 연구에서는 Transcriptome database에 sequence reads를 정열하기 위하여 kallisto를 사용하였다[14]. Kallisto는 그림 1에서 나타난 것과 같이 de Bruijin Graphs를 이용하여 index를 구축하며, 정열단계에서 염기서열을 transcriptome에 빠르게 정열하기 위하여 pseudoalignment를 한다[15].

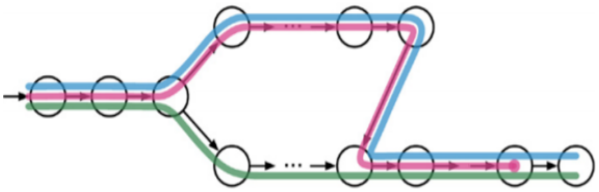


그림 1. de Bruijn 그래프
Fig. 1. de Bruijn graph

그래프에 있는 각 노드는 염기서열인 k-mers를 의미하며, kallisto는 인간의 reference 게놈인 hg38을 기반으로 하는 제작된 index를 제공하고 있다[16].

각 transcript의 양을 정량화하기 위하여 kallisto quant가 이용되었으며, TPM 값은 다음과 같은 단계를 거쳐 계산되었다. 첫째, 각 read count들을 kilobase로 표시된 각 transcript의 길이로 나누어 RPK 값을 환산한다. 그리고 샘플 데이터에 있는 모든 RPK값을 나열한 다음, 이를 scaling factor인 1,000,000으로 나눈다. 마지막으로 각 유전자의 trascript TPM값의 평균값을 구하여, 최종으로 특정 유전자의 TPM 값을 얻게 된다.

2.3 유전자들이 만드는 단백질들의 세포 내 위치분석

면역세포의 기능을 조절하는 새로운 면역신약개발에는 세포표면에 존재하는 단백질을 특이적으로 인식하는 항체신약개발이 가장 중요한 기술로 알려져 있다. 따라서 naive T 세포의 표면에 존재하는 단백질을 만드는 유전자들을 선별하기 위하여, 유전자들이 만드는 단백질들의 세포내 위치는 Gene Cards database에 의해서 분석하였다. Gene Cards (https://www.genecards.org)는 유전자 ID, 기능적 설명, 게놈 또는 세포내의 위치 등 유전자에 대한 포괄적인 정보를 제공하는 database이다[17]. 세포내의 위치는 11개의 위치코드(plasma membrane, extracellular matrix, lysosome, cytoskeleton, endosome, golgi apparatus, mitochondria, peroxisome, endoplasmic reticulum, cytosol, nucleus) 의하여 표시되며 0-5의 가능성 스코어가 각 세포내 위치에 부여된다. 세포표면 단백질을 만드는 유전자들을 분류하기 위하여, plasma membrane 스코어가 4 이상인 유전자들을 필터링하였다.

2.4 유사성 계산 방법

통계학에서 사용하는 유사성을 계산하는 방법으로는 대표적으로 Cosine Similarity, Euclidean Distance, Jaccard Coefficient, Averaged Kullback-Leibler Divergence, Pearson Correlation 등이 있다[18].

각 방법의 특징으로는 Cosine Similarity 발현 값을 벡터 공간에 맵핑하여 두 벡터 간 각도의 cosine 값을 이용하는 방법이다. Euclidean Distance는 발현 값의 절대 값을 기준으로 크기가 유사한 유전자를 찾는 방법이고, Jaccard Coefficient는 두 유전자의 공통된 발현 값을 기준으로 유사성을 구한다. Averaged Kullback-Leibler Divergence는 두 유전자의 확률 분포 차이를 계산하여 유사성을 구하게 된다.

본 실험에서 여러 가지 종류의 T 세포 subset들(Th0, Th1 또는 Th2)과 비교하여 naive T 세포에 특이적으로 발현하는 유전자를 찾기 위하여, 기존에 사용하는 Fold Change 방법과 유전자 사이의 발현 값 상관관계를 계산하는 Pearsion Correlation 및 차원을 축소하는 T-SNE을 사용하여 비교 분석하였다.

2.4.1 Pearson Correlation

본 연구에 사용된 유전자 유사성 계산 방법 중 하나인 Pearson Correlation은 세포에서 특이적으로 발현하는 유전자(해당 세포에서만 발현이 높은 유전자)와 유사한 발현 비율을 갖는 유전자를 찾아 아직 밝혀지지 않은 바이오 마커를 찾는데 유용하게 쓰일 수 있다.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

Pearson Correlation은 식 (1)에서 나타내고 있으며, 두 변수의 평균과 표준편차를 이용해 두 변수 사이의 선형 상관 관계를 계산하는 방법으로 상관관계 값은 -1 ~ 1의 범위를 갖는다. -1은 음의 상관관계, 0은 상관관계 없음, 1은 양의 상관관계를 나타낸다. 따라서 1에 가까운 값을 가질수록 유사성이 높다는 것을 뜻한다. 이를 통해 특이적으로 발현하는 유전자를 찾기 위해, 특정 세포에서 특이적으로 발현하는 유전자를 기준으로 다른 모든 유전자의 유사성을 계산한다. 이후에 특이적으로 발현하는 유전자와 유사성 값이 1에 가까운 유전자를 출력하면 해당 세포에서만 발현 값이 높은 유전자 후보를 얻을 수 있다.

Pearson Correlation은 선형 상관 관계를 계산하는 방법이기 때문에 세포별 유전자 발현 값의 비율이 유사한 유전자를 찾는다. 따라서 기존 방법과 차별적인 방법으로 특이적으로 발현하는 유전자를 찾을 수 있다.

Pearson Correlation을 사용하여 각 세포 별로 정리된 유전자의 TPM 값의 세포별 비율을 기준으로 유전자 간 유사성을 계산하는 과정은 표 1과 같은 코드로 설명한다.

표 1. 유사성을 계산하는 의사 코드
Table 1. Pseudo code to calculate similarity

```
Input = Gene_TPM, standard_gene, similarity

function Pearson_correlation(Input):
    while i is less than number of gene:
        while j is less than number of gene:
            similarity(i,j) = Pearson(i,j)
        end while
    end while
end function

sort(similarity[standard_gene])
```

유사성이 계산된 후에는 특정 세포에서 특이적으로 발현하는 유전자를 기준으로 세포에서 특이적으로 발현될 수 있는 유전자 목록을 출력한다. 특정 세포에서 특이적으로 발현하는 유전자와 유사성이 높은 유전자들은 해당 세포에서 특이적으로 발현될 확률이 높다고 예측할 수 있다.

2.4.2 T-SNE

본 연구에 사용된 유전자 유사성 계산 방법 중 T-SNE는 다음과 같이 계산된다. 각 유전자 데이터는 n개의 세포에서 발현하는 TPM 값을 갖고, 유전자의 TPM 값을 벡터 공간에 나타내면 n차원의 값을 갖는다. 따라서 유전자 별 TPM 값을 사용하여 T-SNE 방법으로 n차원의 발현 값을 2차원으로 축소한 후 벡터 공간에 맵핑한다. 2차원 공간에 맵핑된 유전자 사이의 유사성은 Euclidean Distance를 사용하여 구한다. T-SNE를 사용하면 특이적으로 발현하는 유전자와 발현 값이 유사한 이웃 유전자를 거

리로 표현할 수 있기 때문에 시각화하여 그래프로 나타낼 수 있는 장점이 있다.

T-SNE는 고차원 공간에서 데이터들의 이웃 간 거리를 최대한으로 보존하여 저차원 공간을 만드는 방법론이다. 각각 n차원 값을 갖는 유전자를 T-SNE를 사용하여 2차원 값으로 축소한다. 2차원으로 축소된 벡터공간에서 유전자 사이의 거리를 구하면 특이적으로 발현하는 유전자와 발현 값이 유사한 유전자를 찾는 것이므로 해당 세포에서 발현 값이 높은 유전자를 찾을 수 있다. Pearson Correlation은 선형 상관 계수를 통하여 유사한 유전자를 구하기 때문에 두 방법론은 다른 결과를 출력하게 된다.

$$p_{j|i} = \frac{e^{-\frac{|x_i - x_j|^2}{2\sigma_i^2}}}{\sum_k e^{-\frac{|x_i - x_k|^2}{2\sigma_i^2}}} \quad (2)$$

$$q_{j|i} = \frac{e^{-|y_i - y_j|^2}}{\sum_k e^{-|y_i - y_k|^2}} \quad (3)$$

T-SNE의 식 (2)에서 $p_{j|i}$는 고차원 공간에 존재하는 i번째 데이터 x_i에서 j번째 이웃 데이터인 x_j가 선택될 확률이고, 식 (3)에서 $q_{j|i}$는 저차원에 임베딩된 i번째 데이터 y_i에서 j번째 이웃인 y_j가 선택될 확률이다. 따라서 거리가 가까울수록 확률 값이 커지게 되므로 가까운 이웃일수록 선택될 확률이 높아지게 된다. T-SNE는 $p_{j|i}$와 $q_{j|i}$의 분포 차이를 최대한 작게 만드는 방법을 사용해 차원을 축소한다. T-SNE의 결과 값으로 각 유전자는 이웃 유전자와의 거리를 나타내는 2차원 좌표 값을 갖게 된다. 해당 좌표 값은 2차원 공간에서 이웃 간의 위치를 나타내기 위한 좌표다.

다음으로 T-SNE로 구한 2차원 좌표 값으로 유전자를 벡터 공간에 맵핑한 후 Euclidean Distance를 구하여 기준 유전자와 유사한 유전자를 찾는다.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (4)$$

Euclidean Distance는 식 (4)에서 나타내고 있으며, 두 변수의 좌표를 이용해 두 변수 사이의 거리를 계산할 때 사용하는 방법이다. 유전자를 벡터 공간에 맵핑하여 특정 세포에서 특이적으로 발현하는 유전자와 가까운 거리에 있는 유전자를 탐색한다. 특정 세포에서 특이적으로 발현하는 유전자와 거리가 가까운 유전자들은 해당 세포에서 특이적으로 발현될 확률이 높다고 예측할 수 있다.

T-SNE를 사용하여 각 세포 별로 정리된 유전자의 TPM 값을 기준으로 유전자 간 유사성을 계산하는 과정은 표 2의 코드로 설명한다.

표 2. 차원을 축소하는 의사 코드
Table 2. Pseudo code to reduce dimension

```
Input = Gene_TPM, standard_gene, distance

function dimension_reduction(Input):
    while i is less than number of gene:
        while j is less than number of gene:
            location(i, j) = T_SNE(i, j)
        end while
    end while
end function

function euclidean_distance(location):
    while i is less than number of gene:
        while j is less than number of gene:
            distance(i, j) = euclidean(i, j)
        end while
    end while
end function

sort(distance[standard_gene])
```

Pearson Correlation 방법과 T-SNE와 Euclidean Distance를 사용하여 출력된 유전자를 사용하여 두 방법이 특이적으로 발현되는 유전자를 찾는데 효과적인 방법임을 실험을 통해 검증한다.

III. 실험결과 및 분석

3.1 실험결과

본 연구에서는 기존의 연구에서 단순히 대조군과 실험군 두 가지 데이터를 사용하는 Deseq2와 edgeR과는 다르게, 타겟 세포인 naive T 세포와 대조군

인 Th0, Th1 및 Th2 T 세포 subset들에서 발현하는 mRNA의 발현 값 데이터를 모두 사용하였다. 이들 발현 값에 대한 정규화된 read count를 이용하여, 기존의 Fold Change를 이용하는 방법, 각 유전자 사이의 유사성을 계산하는 두 가지 방법인 Pearson Correlation과 T-SNE를 이용하여 naive T 세포에 특이적으로 발현하는 유전자들을 선별하였다.

다른 T 면역세포 (Th0, Th1 또는 Th2)들에서는 발현이 낮고 naive T 세포에서는 높게 발현하는 유전자들을 찾는 방법들 간의 비교분석과정은 그림 2와 같다. 각 T 면역세포 subset들에서 발현하는 mRNA에 대한 염기서열에 대한 raw data를 GEO database에서 준비를 한 다음, 이들 raw data를 kallisto를 이용하여 각 유전자의 길이를 고려하고 hg38 인간 reference 게놈을 기반으로 하여 정규화된 TPM 값을 구하였다. 이 과정은 그림 2의 (A)에 해당된다.

인간 naive T 세포, Th0, Th1 또는 Th2 T 세포에 발현하는 mRNA에 대한 TPM 값이 유전자 유사성을 구하는 input data로 사용되었다. 이 때 naive T 세포의 TPM 값이 1보다 작은 유전자는 제외하였다. 또한 naive T 세포의 기능을 조절하는 새로운 면역항체신약개발의 핵심적인 타겟인 세포표면단백질들만을 선택하기 위하여, 각 유전자가 만드는 단백질들의 세포 내 위치정보를 제공하는 Gene Card database를 이용하여 유전자들을 필터링하였다. 이 과정은 그림 2의 (B)에 해당된다.

각 유전자들의 TPM 값을 기준으로 기존의 Fold Change 방법 (naive/Th0, naive/Th1 또는 naive/Th2)을 이용하여 naive T 세포에 특이적인 유전자들을 서열화하였다. 또한 유전자간 유사성을 이용하는 방법인 Pearson Correlation 또는 T-SNE를 이용하여 naive T 세포에 특이적으로 발현하는 유전자들을 서열화하였다. 이 과정은 그림 2의 (C), (D)에 해당된다.

Pearson Correlation방법은 유전자 사이의 선형 상관관계를 구한 후 naive T 세포에서 특이적으로 발현하는 유전자를 기준으로 유사성 순위를 정렬하였다. 순위가 높은 유전자는 기준이 되는 유전자와 발현 비율이 비슷한 유전자 (naive T 세포에서 높은 발현 값을 갖는 유전자)이므로 naive T 세포에서 특이적으로 발현되는 유전자 목록을 구할 수 있었다. T-SNE방법은 n차원의 발현 값을 2차원으로 차원을 축소한 후 유전자를 벡터 공간에 맵핑한다. 이후 Euclidean Distance방법으로 유전자 사이의 거리를 구하여 naive T 세포에서 특이적으로 발현하는 유전자를 기준으로 거리가 가까운 유전자를 확인하면 기준이 되는 유전자와 발현 값이 유사한 유전자 목록을 구할 수 있다. 유사성을 구하기 위해 기준이 되는 유전자는 naive T 세포에서만 높은 TPM 값을 갖는 임의의 concept 데이터로 설정하였다.

위의 세 가지 방법을 통하여 작성된 naive T 세포에 특이적으로 발현하는 유전자목록에 대하여, 지금까지의 기존 연구에서 naive T 세포에 특이적으로 발현하는 것으로 알려진 5개의 유전자들 (CD27, SELL, IL7R, IL2RG, PTPRC)의 위치를 비교 평가함으로서 가장 효과적인 방법을 선정하였다.

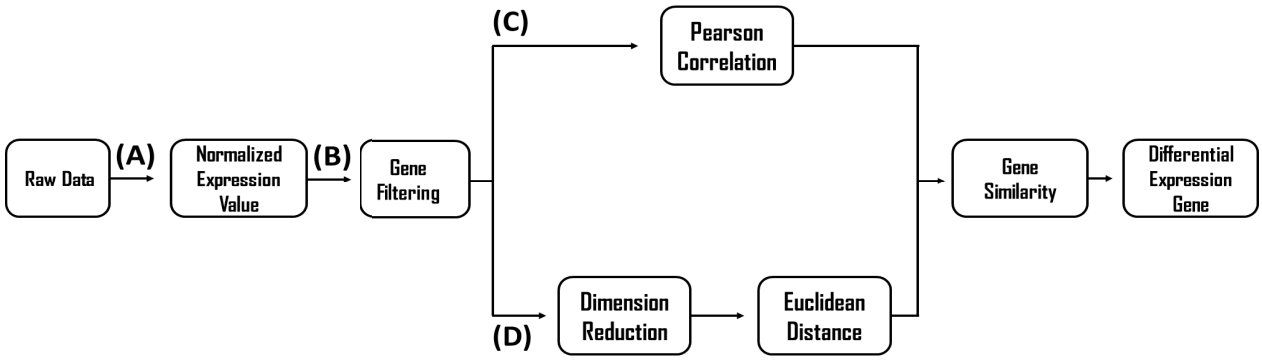


그림 2. 타겟 세포에 특이적으로 발현하는 새로운 유전자 탐색을 위한 데이터 처리와 유전자 유사성 계산 과정
Fig. 2. Flow chart for identification of novel genes specific to target cells using new data processing and gene similarity calculation

3.2 Fold Change에 의한 분석결과

타겟 세포에 특이적 발현 유전자를 찾기 위하여 현재 많이 사용되고 있는 방법인 Fold Change 방법을 이용하여 naive T 세포에 특이적으로 발현하는 표면단백질 유전자 5개의 순위 분석을 진행하였다. $\frac{Naive}{Th0}$, $\frac{Naive}{Th1}$, $\frac{Naive}{Th2}$의 Fold Change 값이 높은 순으로 유전자를 정렬한 후 5개 유전자들의 순위를 비교하며 진행하였다.

그림 3의 (a)는 Fold Change 값이 높은 유전자 100개 중 20개의 TPM 값이 제시되었으며, 제시된 결과 값에는 naive T 세포에 특이적인 유전자들의 TPM 값은 다른 subset T 세포들 대비 높은 것을 확인할 수 있다.

$\frac{Naive}{Th0}$ Fold Change 값을 정렬하였을 때 기존에 알려진 naive T 세포에 특이적인 표면단백질 유전자 5개는 1533위(PTPRC), 2004위(IL7R), 2303위(IL2RG), 3411위(CD27), 3646위(SELL)에 위치하였고, $\frac{Naive}{Th1}$ Fold Change 값을 정렬하였을 때 이 유전자들은 965위(PTPRC), 1203위(IL7R), 2030위(CD27), 2482위(SELL), 6858위(IL2RG)에 위치하였다.

gene_symbol	naive	th0	th1	th2
ECPAS	17.467900	31.911500	37.139800	3.618460e+01
MTF1	2.114035	0.772912	0.000000	4.420000e-10
PDE2A	2.125220	0.437189	0.000000	3.082080e-01
TNIP1	21.782300	17.195400	18.774700	1.031840e+01
STX2	2.389145	2.955610	3.082530	2.014490e+00
MIA2	12.080600	11.121300	16.422300	3.797070e+00
GM14391	3.296735	1.756850	2.404100	1.083070e+00
GM6710	1.860905	2.978220	4.560370	4.822120e+00
GALNT11	3.340305	0.718200	1.087940	7.111410e-01
PEX5	2.712265	1.130020	0.000063	7.510000e-10
SBF1	20.640400	12.718300	11.768300	9.252710e+00
ABR	5.136610	0.822117	1.702150	1.230000e-09
POGZ	11.397750	5.156010	5.887520	1.610090e+00
NFE2L1	2.715540	0.000575	0.000000	4.050000e-10
SMAD7	32.546800	9.316950	8.855530	5.505530e+00
HNRNPD	1.668730	2.660720	2.274690	2.330150e+00
ARHGAP1	5.469010	2.598040	2.676030	4.450000e-10
ABI1	57.363800	40.020100	33.974200	4.711160e+01
PARP6	4.188685	0.755574	1.459790	9.350000e-10
GRK6	6.979915	1.569070	0.000000	4.120000e-10

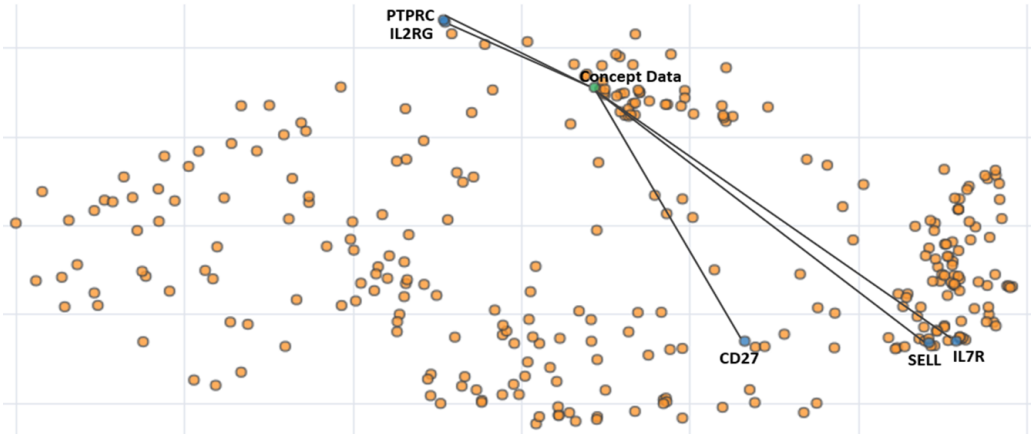
(a) Fold change 결과
(a) Fold change result

gene_symbol	naive	th0	th1	th2
HCST	178.08700	32.66560	29.86480	25.94140
SKAP1	181.85650	22.40970	25.84560	17.90760
IL16	172.06400	58.45020	56.32840	47.78230
RABAC1	145.87900	92.44310	79.98120	83.94440
IL27RA	145.56850	42.64310	49.91990	37.01400
CORO1A	137.25800	54.21590	14.28510	14.97270
LAPTM5	146.25650	0.00000	5.65476	9.14575
PRKCQ	125.78850	55.59590	46.96840	50.72470
EMB	118.54900	57.06820	60.04570	54.53900
GRK2	115.74750	81.23320	80.86860	65.05630
CD96	114.10500	19.02250	17.24770	24.27630
RHOH	104.57950	34.72260	37.58530	29.89430
RASSF2	102.05550	22.83790	21.44910	23.93250
MACF1	98.74420	21.52980	22.68970	13.66000
TGFBR2	94.69935	15.09690	13.40370	8.31015
DGKD	95.42610	39.31010	43.00120	35.32170
TSPAN32	86.73690	19.93330	15.79110	15.87950
ARRB2	91.73820	19.83330	24.14650	18.85180
ARHGEF18	89.59345	13.67230	11.09780	4.36806
SIT1	79.47235	9.42643	7.61647	5.92682

(b) Pearson correlation 결과
(b) Person correlation result

gene_symbol	naive	th0	th1	th2
PTPRC	292.63200	52.9384	50.05020	49.98450
GNAI2	284.76850	153.0530	175.40400	120.74700
ARHGEF1	229.89150	33.0747	29.88080	17.71280
ACTN1	207.60800	140.0610	174.02000	69.15340
GRAMD1A	206.03600	36.2457	40.66950	25.47000
HNRNPF	202.86200	145.3330	278.73100	224.50700
CD27	202.70950	106.6540	100.93100	77.07910
FLNA	189.36850	52.0706	57.05630	40.05490
SKAP1	181.85650	22.4097	25.84560	17.90760
CDC42SE1	150.07450	66.4409	59.92930	66.23320
LAPTM5	146.25650	0.0000	5.65476	9.14575
SRI	143.14350	86.8842	93.51210	72.78370
FMNL1	140.78800	73.0935	82.95710	51.24320
CSK	140.28400	119.6270	129.92200	106.01000
CORO1A	137.25800	54.2159	14.28510	14.97270
LCK	123.93850	24.3439	19.04850	9.37626
RASSF2	102.05550	22.8379	21.44910	23.93250
SELL	101.34720	26.8761	18.31510	7.99223
MACF1	98.74420	21.5298	22.68970	13.66000
TGFBR2	94.69935	15.0969	13.40370	8.31015

(c) T-SNE 결과
(c) T-SNE result



(d) T-SNE를 사용하여 얻은 2차원 공간에서 Concept Data와 naive T 세포에 특이적인 유전자들의 거리
(d) 2-Dimensional distance of concept data and naive T cell-specific genes using T-SNE method

그림 3. 유전자 유사성과 Fold Change 방법을 통해 얻은 naive T 세포 특이적인 발현 유전자 데이터
Fig. 3. naive T cell-specific genes data using gene similarity and Fold Change

마지막으로 $\frac{Naive}{Th2}$ Fold Change 값을 정렬하였을 때 이 유전자들은 631위(SELL), 1286위(IL7R), 1984위(PTPRC), 2529위(IL2RG), 2672위(CD27)에 위치하였다.

3.3 Pearson Correlation에 의한 분석결과

Pearson Correlation을 사용하여 naive T 세포에서 발현 값이 높은 concept 데이터를 기준으로 유전자 사이의 유사성을 구한 후, 유사성이 높은 유전자들을 정렬하였다. Naive T 세포의 표면에 발현하는 단백질을 인코딩하는 유전자를 선별하기 위하여 세포 내 위치의 plasma membrane 값이 4 이상인 유전자들을 사용하였다.

그림 3의 (b)에는 concept 데이터와 유사한 유전자 100개 중 20개 유전자들의 TPM 값이 제시되었으며, 제시된 결과 값에는 naive T 세포에서 특이적인 5개 표면단백질 유전자들의 TPM 값은 다른 subset T 세포들과 비교하여 약 3배~29배 높았다. Pearson Correlation을 이용하여 선별된 유전자 목록에서 기존에 알려진 naive T 세포에 특이적인 표면단백질 유전자 5개는 80위(IL7R), 158위(SELL), 311위(CD27), 424위(PTPRC), 768위(IL2RG)에 위치하였으며, 768순위까지 고려하면 5개의 바이오 마커를 모두 찾을 수 있었다.

3.4 T-SNE에 의한 분석결과

T-SNE를 사용하여 4차원에 해당하는 TPM 값을 2차원으로 축소하여 naive T 세포에서 높게 발현하는 concept 데이터를 기준으로 거리가 가까운 유전자들을 구하였다. 역시 세포 내 위치의 plasma membrane 값이 4 이상인 유전자들을 사용하였다. 그림 3의 (c)는 concept 데이터와 거리가 가까운 유전자 100개 중 20개 유전자의 TPM 값이 제시되었으며, 제시된 결과 값에는 naive T 세포에 특이적인 표면단백질 유전자들의 TPM 값은 다른 subset T 세포 대비 2배~29배 높았다.

그림 3의 (d)는 concept data에서 거리가 가까운 300개의 유전자 중 naive T 세포에 특이적인 표면단백질 유전자인 CD27, SELL, IL7R, IL2RG, PTPRC5와의 거리를 표현한 그래프이다. T-SNE를 이용하여 선별된 유전자 목록에서 기존에 알려진 naive T 세포에 특이적인 표면단백질 유전자 5개는 7위(SELL), 26위(CD27), 38위(PTPRC), 210위(IL2RG), 225위(IL7R)에 위치하였으며, 225순위까지 고려하면 5개의 바이오 마커를 모두 찾을 수 있었다.

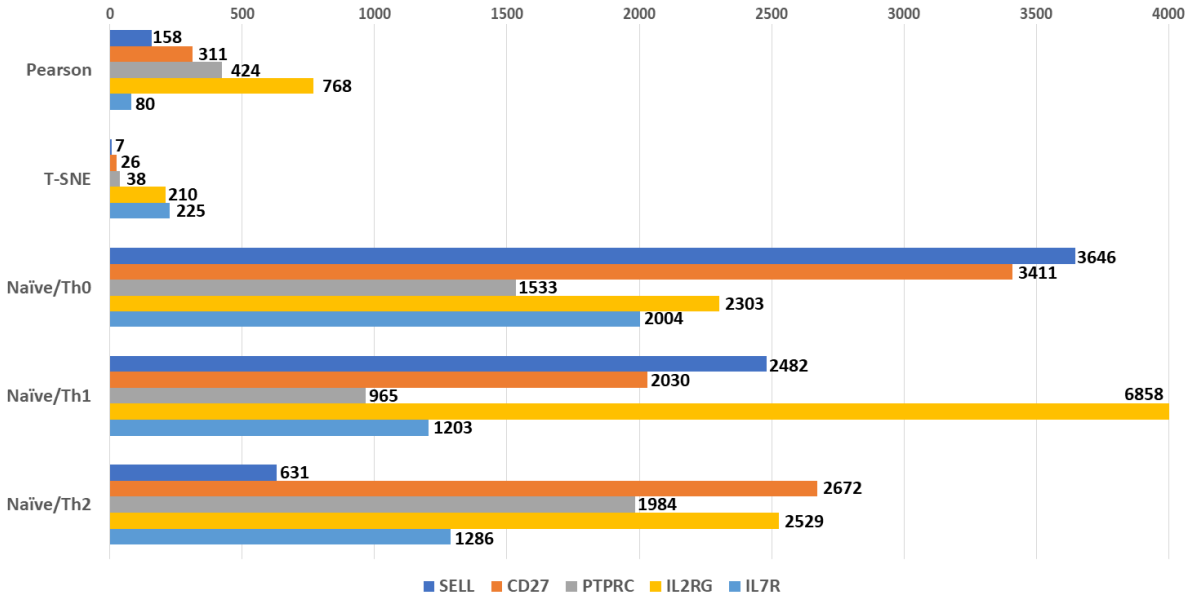


그림 4. Pearson correlation과 T-SNE 방법을 사용한 유전자 유사성 계산 방법과 Fold Change 방법에서 Naive 세포의 특이적 발현 유전자 5개의 순위

Fig. 4. Differential expression hierarchy of five naive T cell-specific genes rank using fold change, pearson correlation or T-SNE methods

3.5 naive T 세포에 특이적으로 발현하는 유전자선별 분석법의 비교 및 검토

특정한 세포에 선택적으로 발현하는 유전자들을 발굴하는 연구에 현재 가장 많이 사용하고 Fold Change 비교분석방법을 Pearson Correlation 방법 또는 T-SNE 방법을 비교분석한 결과를 그림 4와 표 3에 정리하였다. 각 분석법에서 naive T 세포에 특이적으로 발현하는 5개의 유전자 별 순위를 그림 4에 나타냈고, 해당 유전자 순위를 정렬하여 표 3에서 비교하였다.

표 3. 각 분석법 별 naive T 세포에 특이적인 유전자 포함 순위
Table 3. Rank of gene specific for naive T cells by each methods

Methods / Gene	Pearson	T-SNE	naive/ Th0	naive/ Th1	naive/ Th2
1ea	80	7	1,533	965	631
2ea	158	26	2,004	1,203	1,286
3ea	311	38	2,303	2,030	1,984
4ea	424	210	3,411	2,482	2,529
5ea	768	225	3,646	6,858	2,672

naive T 세포에 특이적으로 발현하는 유전자 5개(CD27, SELL, IL7R, IL2RG, PTPRC)를 Fold Change 방법을 이용하였을 때 $\frac{Naive}{Th0}$= 3646위, $\frac{Naive}{Th1}$= 6858위, $\frac{Naive}{Th2}$= 2672위에서 naive T 세포에 특이적으로 발현하는 유전자 5개(CD27, SELL, IL7R, IL2RG, PTPRC)를 모두 찾을 수 있었다.

Pearson Correlation 방법을 이용하면 768위에서 5개의 유전자들을 찾을 수 있었고, T-SNE 방법을 이용하면 225위에서 5개의 유전자들을 찾을 수 있었다. 따라서 단순한 T 면역세포 subset들의 Fold Change 비교분석방법보다는 Pearson Correlation 방법 및 T-SNE 방법이 naive T 세포에 특이적인 유전자들의 선별이 더 효과적이었으며, 특히 T-SNE 방법이 Pearson Correlation 방법보다 기존의 naive T 세포 특이적인 유전자들이 더 높은 순위에 위치하고 있었다. 따라서 특정한 세포에 선택적으로 존재하는 유전자의 발굴방법으로는, 4차원에 해당하는 TPM 값을 2차원으로 축소하여 특정세포에서 높게 발현하는 concept 데이터를 기준으로 분석한 T-SNE방법이 가장 효율적인 것으로 생각된다.

Ⅳ. 결 론

최근 빠르게 발전하고 있는 의생명공학분야의 기초연구 및 신약개발연구에서 특정한 세포에 선택적으로 발현하는 새로운 유전자들을 발굴하는 연구는 가장 핵심적인 요소이다. 특히 세포표면에 발현하는 유전자들의 발굴은 새로운 치료용 항체신약 또는 진단시약의 개발에 관건으로 알려져 있다.

본 연구에서는 면역 미세 환경에 따라 다양한 T 면역세포 subset들로 분화되는 naive T 세포에 특이적으로 발현하는 것으로 알려진 표면 단백질 유전자들인 CD27, SELL, IL7R, IL2RG, PTPRC를 기준으로 하여, 기존의 Fold Change를 단순히 비교하는 방법대신 빅 데이터 처리와 머신러닝을 사용하여 유전자들 사이의 유사성을 계산하는 두 가지 새로운 방법과 비교분석하였다.

빅 데이터의 처리과정으로는 mRNA 염기서열에 대한 raw data를 kallisto를 이용하여 각 유전자에 맵핑하여 정규화 발현 값으로 전환하였다. 또한 세포표면 단백질을 만드는 유전자들만을 이용하기 위하여 Gene Cards database를 이용한 세포 내 단백질 위치 분석 방법을 적용하였다. 빅 데이터 처리와 머신러닝을 사용하여 제시된 방법으로는, concept data를 기준으로 하여 유전자들의 선형 상관계수를 계산하여 특정 세포에서만 높게 발현하는 유전자를 찾는 Pearson correlation방법과 n차원 발현 값을 갖는 유전자를 n차원 벡터 공간에서 2차원으로 차원을 축소하여 거리가 가까운 유전자를 찾는 T-SNE 방법을 사용하였다.

본 연구에서 시도한 3가지 분석방법들 중 T-SNE 분석방법을 사용하였을 때, naive T 세포에 특이적으로 세포표면에 발현하는 5개의 유전자들이 모두 가장 높은 순위에 위치하였다. 따라서 특정한 세포에 선택적으로 발현하는 유전자들의 발굴을 위하여서는 T-SNE 분석방법이 가장 효과적인 방법으로 사용될 수 있음을 제시한 것이다.

본 연구에서의 결과와 같이 빅 데이터의 새로운 처리방법, 인공지능을 이용한 새로운 알고리즘과 생물학을 접목시키는 연구가 활발한 가운데, 유전자 사이의 유사성을 계산하여 타겟 세포에 특이적인 발현을 갖는 유전자들을 찾는 본 연구는 인공지능과 생물학을 접목시키는 발판이 될 수 있다.

본 연구에서의 결과를 기반으로 하여, 특정한 세포에 선택적으로 발현하는 유전자를 찾는 것뿐 만 아니라, 세포 내의 모든 생체기능을 직접 조절하는 단백질들에 대한 발굴 연구에도 중요하게 사용될 것이다. 또한 앞으로도 인공지능에서 사용되는 머신러닝, 딥러닝 방법을 사용해 특이적인 발현을 갖는 유전자 또는 단백질을 찾는 연구는 생명공학분야의 핵심기술이 될 것으로 생각된다.

References

[1] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", Genome Biology, Dec. 2014.

[2] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", Bioinformatics, Vol. 26, No. 1, pp. 139-140, Jan. 2010.

[3] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering", IEEE Computer Society, Vol. 7, No. 1, pp. 76-80, Jan. 2003.

[4] S. Gower, "Netflix Prize and SVD", Apr. 2014.

[5] J. Davidson and B. Liebald, et al., "The YouTube video recommendation system", Proceedings of the fourth ACM Conference on Recommender Systems (RecSys '10), Barcelona Spain, pp. 293-296, Sep. 2010.

[6] P. Covington, J. Adams, and E. Sargin, "Deep Neural Networks for YouTube Recommendations", Proceedings of the 10th ACM Conference on Recommender Systems(RecSys '16), Boston Massachusetts USA, pp. 191-198, Sep. 2016.

[7] G. Hall, "Pearson's correlation coefficient", Feb. 2015.

[8] L. V. D. Maaten and G. Hinton, "Visualizing Data using t-SNE", Journal of Machine Learning Research, Vol. 9, pp. 2579-2605, Nov. 2008.

[9] K. Yakimchuk and K. Institutet, "T cell Markers and B cell Markers", Labome, Mar. 2016.

[10] G. F. Gerberick, L. W. Cruse, C. M. Miller, E. E. Sikorski, and G. M. Ridder, "Selective Modulation of T Cell Memory Markers CD62L and CD44 on Murine Draining Lymph Node Cells Following Allergen and Irritant Treatment", Toxicology and Applied Pharmacology, Vol. 146, No. 1, pp. 1-10, Sep. 1997.

[11] J. T. Tan and E. Dudl, et al., "IL-7 is critical for homeostatic proliferation and survival of naïve T cells", Proceeding of the National Academy of Sciences of the United States of America(PNAS), Vol. 98, No. 15, pp. 8732-8737, Jul. 2001.

[12] C. Ramsey, M. P. Rubinstein, D. M. Kim, J. H. Cho, J. Sprent, and C. D. Surh, "The Lymphopenic Environment of CD132 (Common γ-Chain) - Deficient Hosts Elicits Rapid Homeostatic Proliferation of Naive T Cells via IL-15", The Journal of Immunology, Vol. 180, No. 8, pp. 5320-5326, Apr. 2008.

[13] V. Golubovskaya and L. Wu, "Different Subsets of T Cells, Memory, Effector Functions, and CAR-T Immunotherapy", Caners, Vol. 8, No. 3, 36, Mar. 2016. DOI:10.3390/cancers8030036

[14] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification", Nature Biotechnology, Vol. 34, No. 5, pp. 525-527, Aug. 2016.

[15] P. E. C. Compeau, P. A. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly", Nature Biotechnology, Vol. 29, No. 11, pp. 987-991, Nov. 2011.

[16] V. A. Schneider, T. Graves-Lindsay, K. Howe, and et al, "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly", Genome Research, Vol. 27, No. 5, pp. 849-864, May 2017.

[17] G. Stelzer, R. Rosen, and et al., "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis", Current Protocols in Bioinformatics, Vol. 54, No. 1, pp. 1.30.1-1.30.33, Jun. 2016.

[18] A. Huang, "Similarity Measures for Text Document Clustering", The University of Waikato, Apr. 2008.

저자소개

고 동 우 (Dong-Woo Ko)

2019년 2월 : 가톨릭대학교 컴퓨터정보공학부(공학사)
2019년 3월 ~ 현재 : 가톨릭대학교 컴퓨터공학과 석사과정
관심분야 : 인공지능, 바이오인포매틱스

양 정 진 (Jung-Jin Yang)

1985년 2월 : 이화여자대학교 전자계산학과 졸업(학사)
1992년 6월 : Univ. of Connecticut 대학원 컴퓨터공학과 (공학석사)
1998년 12월 : Univ. of Connecticut 대학원 컴퓨터공학과 (공학박사)
1999년 ~ 2000년 : Univ. of Hartford 컴퓨터학과 조교수
2001년 ~ 2004년 : 가톨릭대학교 컴퓨터공학부 조교수
2005년 ~ 2009년 : 가톨릭대학교 컴퓨터공학부 부교수
2010년 ~ 현재 : 가톨릭대학교 컴퓨터정보공학부 교수
관심분야 : 인공지능, 지능형 에이전트시스템, 바이오인포매틱스, 시맨틱웹