Journal of The Korea Society of Computer and Information
Vol. 20, No. 2, February 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.2.137

마이크로어레이 데이터와 PPI 데이터를 이용한 에스트로겐 수용체 음성 유방암 환자의 예후 특이 네트워크 식별 및 예후 예측

황 유 현*, 오 민*, 윤 영 미**

Identification of prognosis-specific network and prediction for estrogen receptor-negative breast cancer using microarray data and PPI data

Youhyeon Hwang*, Min Oh*, Youngmi Yoon**

요 약

본 논문에서는 유전자 네트워크를 기반으로 유방암 환자의 예후를 예측하는 알고리듬을 제안한다. 유방암 환자의 마이크로어레이 데이터와 PPI(Protein-protein interaction)데이터를 이용하여 알고리듬의 분류자로 사용될 예후 특이 네트워크(Prognosis specific gene network)를 추출한다. PPI에 속한 모든 유전자 네트워크에 대하여 각각의 네트워크가 예후 좋음과 나쁨을 잘 구분하는지에 대한 점수를 피어슨 상관계수(Pearson's correlation coefficient)와 마이크로어레이 데이터를 이용하여 계산한다. 이들 중 가장 예후에 유의한 네트워크를 식별하고, 이 네트워크를 분류자로 사용하여 에스트로겐 수용체 음성 유방암 환자의 예후를 분류 분석 한다. 본 연구와 기존 연구의 알고리듬 정확도를 비교 분석 하기 위하여 독립 실험을 진행하고, 본 연구에서 제안된 알고리듬의 성능이 더 우수함을 보인다. 또한, Gene Ontology 데이터베이스를 활용하여 식별된 예후 특이 네트워크를 기능적으로 검증 한다.

▸ Keywords : 데이터마이닝, 분류분석, 마이크로어레이 데이터, 유방암, 예후 예측, PPI

Abstract

This study proposes an algorithm for predicting breast cancer prognosis based on genetic network. We identify prognosis-specific network using gene expression data and PPI(protein-protein interaction) data. To acquire the network, we calculate Pearson's correlation coefficient(PCC) between genes in all PPI pairs

•제1저자 : 황유현 •제2저자 : 오민 •교신저자 : 윤영미
•투고일 : 2014. 11. 12, 심사일 : 2014. 12. 31, 게재확정일 : 2015. 1. 29.
* 가천대학교 컴퓨터공학과(Dept. of Computer Engineering, Gachon University)
**가천대학교 컴퓨터공학과 교수(교신저자)(Dept. of Computer Engineering, Gachon University)
※이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(NRF-2010-0008639).

using gene expression data. We develop a prediction model for breast cancer patients with estrogen-receptor-negative using the network as a classifier. We compare classification performance of our algorithm with existing algorithms on independent data and shows our algorithm is improved. In addition, we make an functionality analysis on the genes in the prognosis-specific network using GO(Gene Ontology) enrichment validation.

▸ Keywords : data mining, classification, microarray data, breast cancer, prognosis prediction, PPI

I. 서 론

암 환자의 치료에 있어서 예후는 아주 중요한 임상 인자 이다. 예후란 암 발병 이후 환자의 상태가 어떠할 것인가에 대한 전망을 말한다. 환자의 예후를 예측하기 위해서 여러 예후 인자들이 사용된다. 예후 인자라 함은 예후 예측 판단에 영향을 주는 인자로서 일반적으로 나이, 임파절 전이 여부, 암의 기수가 있다. 최근에는 분자생물학의 발달로 이와 같은 임상적인 예후인자와 함께 유전자 발현, 혹은 유전자 발현으로 생성된 단백질과 관련된 예후인자들이 함께 사용되고 있다.

환자의 향후 치료법(예 : 화학항암요법, 방사능요법)에 큰 영향을 끼치는 암환자의 예후예측은 매우 중요하다. 특정 환자의 예후가 좋을 것이라고 성공적으로 예측했을 경우 불필요한 화학항암요법을 피할 수 있으며 그에 수반하는 환자의 고통도 줄여줄 수 있다. 실제로 잘못된 예후 예측 판단으로 인해 유방암에서 불필요한 화학항암요법을 받는 환자의 수는 전체의 70-80% 이다[1]. 올바른 예후 예측을 통하여, 가장 적절한 치료법 선택이 가능하다.

본 논문에서는 유방암의 예후 특이 유전자 네트워크 식별 및 샘플의 예후 예측을 진행하였다. 유방암은 전 세계적으로 여성암에서 높은 발병률을 차지하고 있다. 과거 한국의 여성 유방암은 서양에 비해 발병률이 낮은 편이였으나 최근 발병률이 급격히 증가하여 여성암 중 1위를 차지하게 되었다[2]. 또한 유방암은 다른 여러 암에 비해서도 예후 예측이 어렵고 예후가 나쁜 암으로 알려져 있다. 유방암 환자들의 예후 예측을 위한 노력은 최근까지 꾸준히 이루어져 왔고 다수의 연구들이 진행되어 왔다[1],[3]~[6].

유방암 예후인자 중 대표적인 것으로 호르몬 수용체의 존재 유무가 있다. 이는 예후와 치료에 있어서 중요한 역할을 한다고 알려져 있다[7],[8]. 특히 호르몬 수용체 중 에스트로겐 수용체(estrogen receptor, ER)의 양성(ER+), 음성(ER-)에 따라 수술 후 치료법과 예후 예측이 달라진다[9]. 일반적으로 ER+ 유방암 환자가 ER- 유방암 환자 보다 수술 후 보조치료를 하지 않더라도 예후가 좋으며[10],[11], 호르몬 치료 효과도 ER+ 경우가 더 우수한 것으로 보고되고 있다[12],[13]. 그러므로 상대적으로 예후 판단이 어려운 ER- 유방암 환자 치료를 위한 예후 예측이 보다 중요하다. 본 논문에서는 ER- 환자의 예후 특이 유전자 네트워크(Prognosis specific gene network)를 식별하고 환자의 예후를 예측한다.

본 논문에서는 데이터 샘플의 예후를 레이블링 할 때 DMFS(Distant Metastasis Free Survival) 지표를 사용하였다. DMFS란 암 치료 이후 원격 전이 없이 생존한 기간을 나타내는 지표로 해당 값이 5년 미만이면 예후가 나쁜 것으로 판단하고 5년 이상이면 예후가 좋은 것으로 판단한다. 본 논문 뿐만 아니라 기존의 암 연구에서도 DMFS 지표를 사용하여 샘플의 예후를 레이블링 한다.

본 논문은 DNA 마이크로어레이 데이터와 PPI 데이터를 사용하여 ER- 환자의 예후 특이 유전자 네트워크를 식별한다. 기존의 네트워크 기반 혹은 유전자 기반 연구들은 ER의 상태를 고려하지 않은 데이터, 또는 항암화학요법(chemotherapy)을 받은 환자와 받지 않은 환자를 구별하지 않은 데이터를 활용하여 분류분석을 진행하였다[3],[5],[6]. 또한 비교적 적은 수의 샘플로 구성된 단일 데이터 세트만 사용하여 분류자를 추출하고, 정확도를 산출하였기 때문에 모델이 특정 데이터에 과적합화 되어 있어, 다른 독립 데이터를 적용했을 때 정확도가 떨어지는 경향을 보였다. 최근 이러한 점들을 보완하여 복수개의 데이터 세트를 사용하고 ER의 상

태 등을 고려한 연구가 진행되었다[1]. Maxime Garcia et al. 연구에서는 복수개의 데이터 세트를 사용하여 샘플수를 증가 시키고 ER의 상태도 고려하였다[1]. 그러나 이 연구에서는 예후 특이 네트워크를 추출 시 데이터 세트를 통합하여 사용하지 않고 각 데이터 세트에서 개별적으로 네트워크를 추출하여 통합하는 방식을 사용하였다. 이 방법은 각 데이터 세트에 오류가 있을 시 오류를 증폭시킬 가능성이 있으며 실질적으로 복수개의 데이터 세트를 통합하여 분류자를 식별하였다고 보기 어렵다.

본 논문에서는, 복수개의 데이터 세트를 통합하여 기존 논문들의 적은 샘플 수 문제를 해결하였으며, 이 데이터로부터 하나의 네트워크 분류자를 식별하였다. 또한 데이터 샘플의 ER 상태, 항암화학요법 치료 유무를 고려하여 실험을 진행하였다. 실험을 통하여 식별된 예후 특이 네트워크를 사용하여 본 논문의 분류 방법으로 정확도를 계산하였고 기존 연구의 정확도와 비교하여 본 논문의 결과가 향상됨을 보였다. 또한 식별된 예후 특이 네트워크를 구성하는 유전자의 기능적 검증을 위하여 Gene Ontology 데이터베이스를 활용하였다.

II. 관련 연구

1. PPI(Protein protein interaction) 데이터

단백질 상호작용은 생명 현상의 근간이 되며 생체 내에서 특정 기능을 수행하고 세포 기능적인 면에서 중요한 역할을 수행한다[14],[15]. 단백질은 홀로 체내에서 활동하기도 하고 다른 단백질 복합체와 결합하여 특정 기능을 수행하기도 하며 단백질 복합체들 간 상호작용을 통하여 기능을 수행하기도 한다[16]. PPI 데이터는 이러한 여러 단백질 상호작용을 2진 연결로 표현한 데이터이다. 체내에서 다양한 기능을 하는 단백질 들은 유전자의 전사(transcription) 및 mRNA의 번역(translation) 과정을 통하여 만들어진다. 따라서 PPI 데이터를 구성하는 단백질과 해당하는 유전자를 사상(Mapping) 시킬 수 있으며 이를 통하여 유전자 간 상호 작용 네트워크를 만들 수 있다.

본 논문에서는 여러 PPI 데이터베이스 (BIND[17], BioGrid[18], HPRD[19], IntAct[20], MINT[21], DIP[22], MIPS[23])를 포함하고 있는 PPI Meta-database인 OPHID(Online Predicted Human Interaction Database)[24]에서 인간(Human) 단백질에 관련된 최신 PPI를 수집하여 사용하였다[16].

사상을 위한 단백질과 유전자의 정보는 The Universal Protein Resource (UniProt) 데이터베이스 에서 수집했으며 수집된 정보를 통하여 유전자 네트워크 데이터를 도출하였다.

2. 피어슨 상관계수(Pearson's Correlation Coefficient, PCC) [25]

피어슨 상관계수는 상관 분석에서 전통적으로 사용되는 계수로서 두 벡터의 선형적 관계를 찾는 계수이다. 이 값을 사용하여 두 벡터간의 연관 관계의 강도를 나타낼 수 있다. 상관계수 값이 1과 -1에 가까울수록 두 벡터간의 강한 연관성을 나타내며 값이 0에 가까울수록 연관성이 없음을 나타낸다. 값이 1에 가까워질수록 강한 양의 상관관계를 가지며 -1에 가까워질수록 강한 음의 상관관계를 가진다. 피어슨 상관계수 계산식은 아래와 같다.

$$r = \frac{\sum_{i=0}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=0}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=0}^{n}(Y_i - \overline{Y})^2}}$$

(r = 피어슨 상관계수 값, X = X변수의 값, $\overline{X}$ = X변수 값들의 평균, Y = Y변수의 값, $\overline{Y}$ = Y변수 값들의 평균, n = 샘플 수)

본 논문에서는 이를 유전자 벡터간의 상관관계를 계산하는데 사용하였다.

3. Gene ontology(GO) Database

Gene ontology란 유전자의 기능적 정보를 말하며, 크게 세 가지 부분의 ontology로 구분한다. 유전자 생산물(Gene product)이 생체 내에서 하는 기능(바인딩(binding) 및 촉매작용)에 대한 "molecular function ontology", 세포와 세포 밖의 환경에 대한 "cellular component ontology", 체내에서 여러 가지 기능을 하는 세포, 조직, 장기 등의 작용, 여러 분자들의 작용, 단백질의 작용 등이 유전적으로 어떠한 과정을 통하여 만들어지는지에 대한 정보인 "biological process ontology"가 있다[26],[27]. Gene ontology는 여러 GO term들로 이루어져있다. GO term은 체내에서 특정 역할을 하는 것으로 예를 들어 "regulation of cell death"라는 이름의 GO term은 세포 죽음을 조절하는 과정을 말한다. 각 GO term은 해당 과정에 연관된 유전자들을 포함하고 있다.

본 논문에서는 Database for Annotation,

Visualization, and Integrated Discovery (DAVID) 메타 데이터베이스에 있는 Gene ontology 정보를 사용 하였다[28]. DAVID는 Gene Ontology[29]의 정보를 모두 포함하고 있으며 사용자가 유전자 목록을 입력하면, 해당 유전자 정보가 포함된 GO term을 출력해 준다. 각 term이 주어진 유전자들의 기능을 얼마나 적합하게 설명하고 있는지를 나타내는 유의확률을 P-value라 하며, 이 값이 작을수록 해당 term이 유의하다는 의미이다.

본 논문에서는 Gene ontology 데이터베이스를 예후 특이 네트워크를 구성하는 유전자가 실제 유방암의 예후에 관련된 GO term에 부합하는지 판단하기 위하여 사용하였다.

III. 본 론

1. 시스템 개요

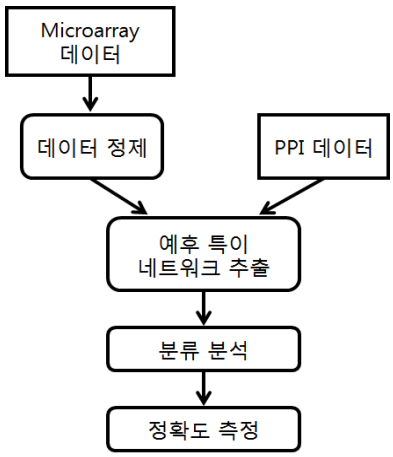

그림 1 . 시스템의 전체적인 개요
Fig. 1. System overview

본 논문의 전체적인 시스템 개요는 그림 1과 같다. 기존 연구로 검증된 다수의 마이크로어레이 데이터 셋을 수집하고 정제 및 통합한다. PPI 데이터와 정제된 마이크로어레이 데이터를 사용하여 예후 특이 네트워크를 식별한다. 식별된 네트워크를 분류자로 사용하여 독립 분류 분석을 시행한다. 분류 분석을 통하여 정확도를 측정하고 기존 연구의 알고리듬 정확도와 비교한다.

2. 데이터 정제 및 통합

본 논문에서 사용한 유방암 예후 관련 마이크로어레이 데이터는 총 4개의 데이터 세트로 이 중 3개는 예후 특이 네트워크를 추출하는데 사용하였고 나머지 1개는 독립 실험에 사용하였다. 실험에 사용된 데이터 세트는 마이크로어레이 데이터뱅크인 NCBI (National Center for Biotechnology Information)의 GEO(Gene expression Omnibus) 데이터베이스에서 다운받아 사용하였으며 공개되어 있다. 정확한 네트워크 추출을 위하여 각 데이터 들은 다음의 기준으로 정제 되었다.

① 에스트로겐 수용체의 상태정보를 포함한 샘플
② DMFS(Distant Metastasis Free Survival) 정보를 포함한 샘플
③ 항암화학요법(Chemotherapy)을 받지 않은 샘플

본 논문은 서론에서 언급한 적은 샘플수의 문제를 해결하기 위하여 이미 검증된 연구의 복수개의 데이터 세트를 통합하였다. 데이터 통합 시 각 데이터 세트의 값은 실험환경에 따라 측정척도(scale)가 다르기 때문에 샘플을 단순히 붙여 사용할 수 없다. 따라서 측정척도를 동일하게 바꿔주는 정규화 작업이 필요하다. 본 논문에서는 Z 점수(Z score, 표준점수) 정규화 방법을 사용하였다. 샘플 각각에 대해 모든 유전자 발현 값(mRNA expression)의 평균 및 표준편차를 계산한 후 아래의 식을 사용하여 Z 점수를 산출하였다.

* $z = \frac{x - \mu}{\sigma}$

(x = 원래의 수치, μ = 한 샘플 내 전체 유전자 발현값들의 평균, σ = 한 샘플 내 전체 유전자 발현값들의 표준편차)

3. 예후 특이 네트워크 추출

본 논문은 다음과 같은 방법을 사용하여 예후 특이 네트워크를 추출하였다.

3.1 예후 특이 점수의 계산

아래 그림2의 (A)와 같이 샘플의 클래스 레이블인 DMFS 지표에 따라 좋은 예후(*Good*) 그룹과 나쁜 예후(*Bad*) 그룹으로 데이터를 분리한 후 PPI를 구성하는 모든 유전자쌍 (gi,gj)에 대해 좋은 예후 그룹에서의 피어슨 상관계수 값 $PCC(V_{gi}^{G}, V_{gj}^{G})$과 나쁜 예후 그룹의 피어슨 상관계수 값 $PCC(V_{gi}^{B}, V_{gj}^{B})$을 구한다. ($PCC(V_{gi}^{G}, V_{gj}^{G})$ = 좋은 예

후 샘플 그룹의 "유전자gi 벡터" 와 "유전자gj 벡터" 간의 피어슨 상관계수 값)

저자의 이전 연구결과인, 다음의 식을 사용하여 그림2의 (B)와 같이 모든 유전자쌍의 예후 특이 점수를 산출한다 [30].

* 예후 특이 점수 = $\left|PCC(V_{gi}^{G}, V_{gj}^{G}) - PCC(V_{gi}^{B}, V_{gj}^{B})\right|$

(V = 벡터, gi,gj = 유전자쌍 $(i \neq j)$,
G = 좋은 예후 샘플 그룹, B = 나쁜 예후 샘플 그룹)

위 식에서 g_i와 g_j는 한 유전자쌍을 구성하는 유전자를 뜻하며 V_{gi}^{G}과 V_{gi}^{B}는 각각 좋은 예후 샘플 그룹과 나쁜 예후 샘플 그룹의 유전자i 발현 값 벡터를 뜻한다. 마찬가지로 V_{gj}^{G}과 V_{gj}^{B}역시 유전자j의 좋은 예후 샘플 그룹, 나쁜 예후 샘플 그룹의 발현 값 벡터를 뜻한다.

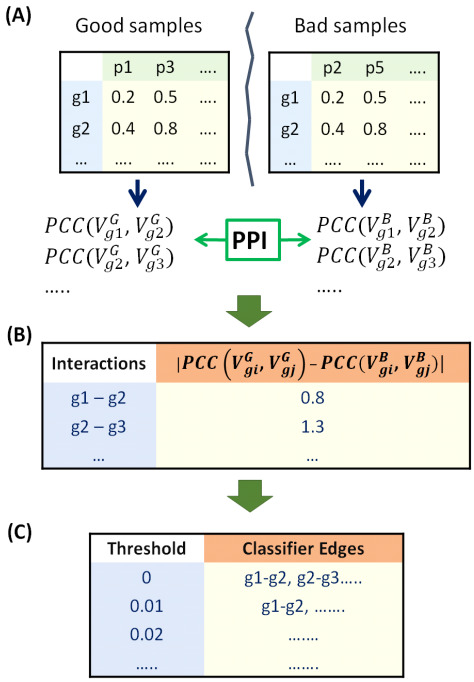

그림 2 . 예후 특이 네트워크 추출
Fig. 2. Identifying a prognosis-specific network

모든 유전자 쌍에 대하여 예후 특이 점수를 산출한 후 그림2의 (C)와 같이 0부터 2까지 0.01씩 증가시키면서 임계값(Threshold)을 설정하여 해당 임계값을 넘는 유전자쌍을 골라 네트워크를 구성한다. 3.2의 샘플 예측 방법과 3.3의 LOOCV(Leave One Out Cross Validataion)를 통하여 가장 좋은 성능을 내는 임계값과 최적의 네트워크를 추출한다.

3.2 샘플 예측(분류)

샘플 예측은 "3.1 예후 특이 점수의 계산"을 통하여 추출된 분류자 네트워크를 사용한다. 그림3 과 같이 분류자 네트워크를 구성하는 유전자의 발현 값만을 사용하여 샘플 분류를 진행한다. 검증용 데이터(Test set)에 속한 샘플 하나를 학습 데이터(Training set)의 좋은 예후 그룹과 나쁜 예후 그룹에 각각 포함시킨다. 포함시킨 후 다음과 같은 점수산출 방법으로 해당 샘플을 좋은 예후 또는 나쁜 예후로 예측한다.

본 논문은 *Good, Bad* 그룹의 검증 샘플 포함유무에 따른 상관관계의 차이를 비교하여 샘플을 분류한다. 즉 검증 샘플을 포함시켰을 때 분류자 네트워크에 속한 유전자들의 상관관계가 더 증가 하는 그룹으로 샘플을 분류한다. 그림3과 같이 *Good, Bad* 두 그룹에 검증 샘플을 포함시킨 후, 분류자 네트워크를 구성하는 모든 유전자쌍(gi,gj)에 대해 상관관계를

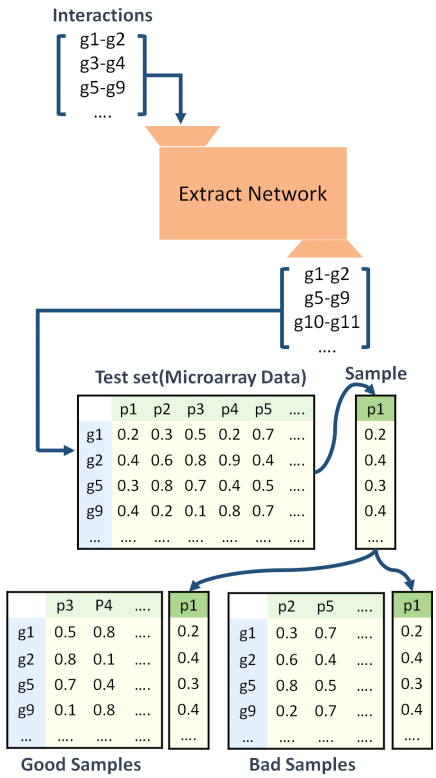

그림 3. 샘플 분류
Fig. 3. Classification of samples

계산한다. 샘플을 포함시키기 전의 상관관계 값과 포함시킨 후의 값을 비교하여 샘플을 포함시킨 후의 유전자쌍 상관관계가 더 큰 그룹으로 해당 샘플을 분류한다. 피어슨 상관 계수는 양의 상관관계와 음의 상관관계가 존재한다. 따라서 샘플을 포함시킨 후의 상관관계가 더 증가되었는지 확인하기 위해서는 샘플 포함 전, 후 상관관계의 부호 비교가 필수적이며, 같은 부호일 때 및 다른 부호일 때 상관관계 차이를 계산하는 방법이 달라진다. 상관관계 차 계산법은 아래 3-1) ~ 3-4)와 같이 네 가지 경우가 도출된다. 본 논문은 이와 같은 방법으로 $scoreG$, $scoreB$ 를 계산하고 그 값에 따라 샘플을 분류한다.

1) 두 개의 점수 $scoreG$과 $scoreB$를 0으로 설정한다.
2) 분류자 네트워크의 모든 유전자쌍(gi,gj)에 대해 샘플을 포함시킨 후의 PCC값과 샘플을 포함시키기 전의 PCC값을 계산한다.
   * 검증용 샘플을 포함시킨 후의 PCC값
   $PCC(V_{gi}^{B},V_{gj}^{B})'$, $PCC(V_{gi}^{G},V_{gj}^{G})'$
   * 검증용 샘플을 포함시키기 전의 PCC값
   $PCC(V_{gi}^{B},V_{gj}^{B})$, $PCC(V_{gi}^{G},V_{gj}^{G})$
   (V = 벡터, g_i,g_j = 유전자쌍 $(i \neq j)$,
   G = 좋은 예후 샘플 그룹, B = 나쁜 예후 샘플 그룹)
3) 한 유전자 쌍에 대하여 아래와 같이 계산한다.
   3-1) $PCC(V_{gi}^{B},V_{gj}^{B})'$와 $PCC(V_{gi}^{B},V_{gj}^{B})$의 부호가 같고 , $PCC(V_{gi}^{G},V_{gj}^{G})'$와 $PCC(V_{gi}^{G},V_{gj}^{G})$의 부호가 같으면 즉, 각 Good, Bad 샘플 그룹에서 샘플을 포함시킨 후의 PCC값과 포함시키기 전의 PCC값이 서로 부호가 같다면 다음 식으로 $\triangle b$과 $\triangle g$을 구한다.
   $\triangle b = |PCC(V_{gi}^{B},V_{gj}^{B})'| - |PCC(V_{gi}^{B},V_{gj}^{B})|$
   $\triangle g = |PCC(V_{gi}^{G},V_{gj}^{G})'| - |PCC(V_{gi}^{G},V_{gj}^{G})|$
   계산된 $\triangle b$의 값과 $\triangle g$의 값을 비교하여 $\triangle b \geq \triangle g$이면 $scoreB$에 1을 더해주고, 그 반대라면 $scoreG$에 1을 더해준다.
   3-2) 위 3-1)의 경우와 달리 $PCC(V_{gi}^{B},V_{gj}^{B})'$와 $PCC(V_{gi}^{B},V_{gj}^{B})$의 부호가 같고, $PCC(V_{gi}^{G},V_{gj}^{G})'$ $PCC(V_{gi}^{G},V_{gj}^{G})$의 부호가 다르다면 같은 부호를 가지는 그룹인 Bad 그룹의 $\triangle b$의 값을 구한다.
   $\triangle b = |PCC(V_{gi}^{B},V_{gj}^{B})'| - |PCC(V_{gi}^{B},V_{gj}^{B})|$
   계산된 $\triangle b$의 값이 0보다 크면 $scoreB$에 1을 더해준다.
   3-3) $PCC(V_{gi}^{B},V_{gj}^{B})'$와 $PCC(V_{gi}^{B},V_{gj}^{B})$의 부호가 다르고, $PCC(V_{gi}^{G},V_{gj}^{G})'$와 $PCC(V_{gi}^{G},V_{gj}^{G})$의 부호가 같다면 같은 부호를 가지는 그룹인 Good 그룹의 $\triangle g$의 값을 구한다.
   $\triangle g = |PCC(V_{gi}^{G},V_{gj}^{G})'| - |PCC(V_{gi}^{G},V_{gj}^{G})|$
   계산된 $\triangle g$의 값이 0보다 크면 $scoreG$에 1을 더해준다.
   3-4) 3-1), 3-2), 3-3) 의 경우와 달리 $PCC(V_{gi}^{B},V_{gj}^{B})'$와 $PCC(V_{gi}^{B},V_{gj}^{B})$의 부호가 다르고, $PCC(V_{gi}^{G},V_{gj}^{G})'$와 $PCC(V_{gi}^{G},V_{gj}^{G})$의 부호가 다르면 다음 식으로 $\triangle b$와 $\triangle g$를 구한다.
   $\triangle b = |PCC(V_{gi}^{B},V_{gj}^{B})' - PCC(V_{gi}^{B},V_{gj}^{B})|$
   $\triangle g = |PCC(V_{gi}^{G},V_{gj}^{G})' - PCC(V_{gi}^{G},V_{gj}^{G})|$
   계산된 $\triangle b$의 값과 $\triangle g$의 값을 비교하여 $\triangle b \geq \triangle g$이면 $scoreG$에 1을 더해주고, 그 반대라면 $scoreB$에 1을 더해준다.
4) 분류자 네트워크를 구성하는 모든 유전자쌍에 대하여 3)을 진행한 후 $scoreG$ 값과 $scoreB$ 값을 비교하여 $scoreB \geq scoreG$ 이면 샘플을 예후 나쁨으로 예측하고, 반대의 경우라면 예후 좋음으로 예측한다.

3.3 최적의(Robust) 예후 특이 네트워크 추출 방법

최적의 예후 특이 네트워크를 추출하기 위해서는 3.1에서 임계값에 따라 추출된 네트워크들을 선별하는 작업이 필요하다. 본 논문에서는 LOOCV(Leave One Out Cross Validataion) 를 사용하여 최적의 네트워크를 선별한다. LOOCV는 아래 그림4와 같이 전체 데이터에서 1개의 샘플을 검증용 데이터라 간주하여 제외시키고 나머지 데이터로 분류기를 학습시켜 제외시킨 1개의 검증 샘플을 분류하는 것을 말하며 이와 같은 과정을 모든 샘플에 대해 진행한다. 모든 샘플을 분류하기 때문에 모델에 대한 정확도 측정이 가능하며 이를 사용하여 본 논문에서는 최적의 정확도를 가지는 임계값을 골라낸다. LOOCV 실험으로 최적의 임계값을 골라내기 때문에 해당 임계값에 따른 분류자 네트워크가 샘플 수만큼 만들어 진다. 전체 샘플의 특성을 고려한 분류자를 만들기 위하여 만들어진 모든 분류자 네트워크를 합집합 하여 하나의 커다란 네트워크를 만들어 최적의 예후 특이 네트워크를 추출한다.

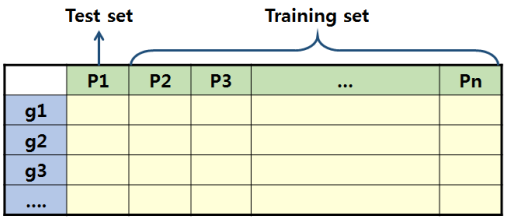

그림 4. LOOCV 모델
(g는 유전자, P_n는 샘플, n은 샘플의 수)
Fig. 4. LOOCV Model
(g is a gene, P_n is samples, n is the number of samples)

IV. 실험 및 결과

1. 실험 데이터 및 실험 환경

본 논문에서는 실험을 위하여 Microsoft visual studio 2010을 사용하였고 실험 환경은 Inter(R) core(TM) i7-4770K CPU @ 3.50GHz, 20GB RAM, 64비트 운영체제이다.

실험에 사용된 데이터 중 PPI 네트워크 데이터는 OPHID(Online Predicted Human Interaction Database)에 공개되어있는 데이터로 인간(Human)단백질 관련 PPI 데이터만 선별해서 사용했으며 총 연결(Interaction)의 수는 81,146개이고 이를 구성하는 단백질의 수는 13,435개 이다. 마이크로어레이 데이터는 [4],[31]~[33] 논문들에서 공개한 데이터들로 모두 다 본론 2의 데이터 정제 기준에 적합한 임상 변수를 가지고 있으며 같은 마이크로어레이 기판(Platform)인 Affymetrix Human Genome U133A Array로 실험되었다. 아래 표1에서 각 데이터의 세부적인 설명과 데이터 정제 기준에 따라 데이터를 정제 한 후의 샘플 수를 확인할 수 있다. 본 논문에서 사용된 마이크로어레이 데이터는 NCBI의 GEO 데이터베이스에서 다운받을 수 있다.

표 1. 데이터 정보
Table 1. Data information

GEO 접근번호	ER-샘플수 (정제후/전)	DMFS 지표*	비고
GSE2034[4]	72 / 77	27 / 45	학습데이터
GSE6532[31]	29 / 195	8 / 21	학습데이터
GSE11121[32]	43 / 44	18 / 25	학습데이터
GSE7390[33]	61 / 64	27 / 34	검증데이터

* DMFS 지표 (전이(Bad) / 비전이(Good))

표1 에서 확인할 수 있듯이 본 논문에서는 적은 샘플 수 문제를 극복하기 위하여 3개의 데이터 셋을 통합하여 학습 데이터로 사용하였고 남은 1개 데이터로 독립 실험을 진행하여 정확도를 산출하였다. 통합 후 학습 데이터의 샘플 수는 205개로 1개의 데이터 셋만 사용한 기존 연구들[3]~[6]의 학습 데이터 샘플 수 보다 월등히 많아졌음을 확인할 수 있다.

2. 정확도 계산

본 논문에서 정확도를 비교하고 임계값에서의 최적 정확도 기준을 삼기 위하여 사용한 척도는 정확도(Accuracy), 특이도(Specificity), 민감도(Sensitivity) 이다. 아래 표2는 예측 클래스와 실제 클래스를 분류기가 얼마나 잘 맞췄는지 보여주는 오분류행렬(confusion matrix)이다. 이 행렬을 통하여 앞의 세 척도를 계산 할 수 있다. 본 논문에서의 TP(True-Positive)는 올바르게 분류된 나쁜 예후 샘플의 수를 의미하며, FP(False-Positive)는 올바르지 않게 분류된 좋은 예후 샘플의 수, FN(False-Negative)은 올바르지 않게 분류된 나쁜 예후 샘플의 수, TN(True-Negative)은 올바르게 분류된 좋은 예후 샘플의 수를 나타낸다.

표 2. 오분류행렬
Table 2. Confusion matrix

		예측 클래스 (Predicted Class)	
		Bad	Good
실제 클래스 (Actual Class)	Bad	TP	FN
	Good	FP	TN

본 논문에서 사용하는 세 가지 척도를 혼돈행렬의 값을 통하여 다음의 식으로 계산하였다.

1) 정확도 $= \dfrac{TN+TP}{TN+TP+FN+FP}$

2) 민감도 $= \dfrac{TP}{TP+FN}$

3) 특이도 $= \dfrac{TN}{TN+FP}$

정확도는 전체 샘플 중 올바르게 분류한 샘플의 비율을 나타내는 척도이고, 민감도는 전체 나쁜 예후 샘플 중 분류기가 실제 나쁜 예후 샘플로 맞춘 비율을 나타내는 척도이다. 특이도는 전체 좋은 예후 샘플 중 분류기가 실제 좋은 예후 샘플로 맞춘 비율을 나타낸다. 본 논문에서는 예후 특이 네트워크를 추출하는데 있어 정확도와 더불어 민감도와 특이도를 함께

고려하였으며 기존 연구의 방법과 비교 시에도 세 척도를 전부 고려하여 비교하였다.

3. 최적 임계값 및 예후 특이 네트워크 추출

본 논문은 LOOCV를 사용하여 0 과 2 사이의 200개 임계값에 대한 정확도, 민감도, 특이도를 계산하였다. 이중 가장 적절한 임계값을 뽑기 위하여 세 척도를 모두 고려하였다. 특정 척도만 사용하여 임계값을 선택하였을 경우 데이터 편중도에 따라 최적의 임계값이 잘못 선택 될 수 있다. 예를 들어 전체 데이터가 10개 이고 이 중 9개 샘플의 클래스가 Good, 1개 샘플의 클래스가 Bad 라고 가정한다. 이 때 분류기가 10개를 다 Good으로 분류한다면 실제 Bad 샘플을 하나도 잘 구분하지 못했음에도 불구하고 정확도는 0.9(90%)의 높은 값을 가지게 된다. 따라서 본 논문에서는 세 가지 척도를 다 고려하여 그 값이 모두 1(100%)에 가까운 값을 가지는 임계값을 뽑았다. 이는 정확도와 민감도와 특이도를 3차원 상의 3개의 축이라고 가정했을 때 (1,1,1)에 가장 가까운 임계값을 말한다. 표 3과 같이 (1,1,1)에 가장 가까운 임계값인 0.29를 골라냈고 이 임계값으로 얻어지는 각 샘플에 대한 유전자 네트워크를 합집합 하여 최적의 예후 특이 네트워크를 추출하였다. 이때 LOOCV 정확도는 0.625(62.5%) 이며 이 임계값을 통하여 얻어진 최적 예후 특이 네트워크의 구성 엣지(Edge)의 수는 18,908개 이고 이를 구성하는 유전자의 수는 총 6,644개 이다.

표 3. LOOCV 결과
Table 3. Results of LOOCV

임계값	TP	TN	FP	FN	정확도	민감도	특이도	*거리
0.29	12	78	13	41	0.625	0.226	0.857	0.871
0.3	12	77	14	41	0.618	0.226	0.846	0.876
0.87	15	67	24	38	0.569	0.283	0.736	0.876
0.28	12	76	15	41	0.611	0.226	0.835	0.881
0.22	14	69	22	39	0.576	0.264	0.758	0.882
…	…	…	…	…	…	…	…	…

* 거리 : 3차원상의 (1,1,1)과의 거리

4. 결과 비교

본 절에서는 앞서 추출된 최적의 예후 네트워크를 분류자로 사용하여, 본 논문의 분류방법으로 독립 데이터를 예후 예측한 결과를 기술한다[33]. 그리고 동일한 데이터로 독립 실험을 진행한 기존 논문의 알고리듬 결과와 본 논문의 결과를 비교 분석한다[1]. 또한 식별된 네트워크를 구성하는 유전자를 Gene Ontology(GO) 데이터베이스인 DAVID에 입력하여 ER- 유방암의 생물학적 기능에 실제로 관여하는지 분석한다.

표 4. 알고리듬의 성능 비교
Table 4. Performance comparison of algorithms

	TP	TN	FP	FN	정확도	민감도	특이도
본 논문 알고리듬	22	13	21	5	0.573	0.814	0.382
ITI[1]	11	22	12	16	0.541	0.407	0.647
70g[1]	27	0	34	0	0.442	1	0
76g[1]	9	14	20	18	0.377	0.333	0.411
GGI[1]	23	6	28	4	0.475	0.852	0.176

표4 는 본 논문 알고리듬과 기존 논문의 알고리듬을 비교 분석한 표이다. 기존 논문의 알고리듬과 비교했을 때 본 논문 알고리듬의 정확도가 0.573으로 가장 좋다. 또한 예후가 나쁜 환자를 예후가 나쁜 환자로 제대로 예측하였는지 판단할 수 있는 척도인 민감도 가 0.814로 기존 알고리듬과 비교해 더 높거나 비슷하다. 암의 예후 판단에서는 예후가 나쁜 환자를 예후가 나쁘다고 잘 판단하는 것이 더 중요하다. 예후가 좋은 환자를 예후가 좋은 환자로 제대로 예측하는지 판단할 수 있는 척도인 특이도는 본 논문 알고리듬이 일부 알고리듬들 보다 낮다. 전체적으로 본 논문의 알고리듬이 기존 알고리듬에 비하여 높은 정확도와 민감도를 보였다.

본 논문에서 식별된 예후 특이 네트워크를 구성하는 유전자들이 ER- 유방암의 예후와 관련된 기능을 하는지 확인하기 위하여 GO 분석을 진행하였다. 네트워크 구성 유전자들의 여러 GO 분석 결과 중 유의미한 GO term 10개를 선별하였고, 본 논문에서 식별된 예후 특이 네트워크가 ER- 유방암의 예후에 관련되었음을 보였다. 유의미한 GO term은 0.01 이하의 P-value 값을 가지며, 기존연구들에서 ER- 유방암의 예후에 관여한다고 입증된 term을 뜻한다.

표 5. ER- 예후 특이 네트워크의 GO term 결과
Table 5. Enriched Gene Ontology annotation of ER- prognosis-specific network

GO	Gene Ontology	P-value
GO:0007049	cell_cycle	4.442E-61
GO:0008283	cell_proliferation	1.985E-32
GO:0007067	mitosis	3.837E-21
GO:0051329	interphase_of_mitotic_cell_cycle	2.523E-19

GO:0051101	regulation_of_DNA_binding	9.296E-13
GO:0043406	positive_regulation_of_MAP_kinase_activity	3.426E-12
GO:0006955	immune_response	2.666E-10
GO:0016064	immunoglobulin_mediated_immune_response	1.816E-5
GO:0002377	immunoglobulin_production	4.376E-4
GO:0060070	Wnt_receptor_signaling_pathway_through_beta-catenin	7.360E-3

표5에서 확인할 수 있는 cell_cycle, mitosis, cell_proliferation은 모두 세포증식(cell proliferation)에 관련된 GO term으로 암세포의 증식 및 전이와 관련이 있다. regulation_of_DNA_binding_term은 ER의 상태에 영향을 끼치는 유전자인 ESR1의 발현량과 관련이 있다[32]. immunoglobulin_mediated_immune_response, immunoglobulin_production, immune_response은 모두 면역(immunity)에 관련된 GO term 으로 암의 전이와 밀접한 관련이 있다[32],[34]. interphase_of_mitotic cell_cycle, positive_regulation_of_MAP_kinase_activity,Wnt_receptor_signaling_pathway_through_beta-catenin은 유사분열시 간기 세포 사이클, kinase에 대한 세포 사이클로 모두 암 세포의 증식과 관련이 있으며 이 term에 속한 유전자 들은 암세포의 유무에 따라 mRNA 발현에 영향을 받는 유전자 들이다[1].

V. 결론

본 논문에서는 예후 특이 네트워크를 식별하고, 이를 분류자로 사용하여 예후 예측을 진행하였다. 본 논문에서는 기존 연구들에서 고려하지 않은 환자의 에스트로겐 수용체 상태를 고려하였고, 상대적으로 예후 예측이 어려운 에스트로겐 수용체 음성 환자 집단에 초점을 맞추어 실험을 진행하였다. 또한 기존 연구에서는 일반적으로 단일 세트의 데이터만 사용하여 모델의 적합화 및 검증을 진행한 반면 본 논문에서는 복수개의 데이터 세트를 통합하여 모델의 적합화 및 검증을 진행하였다. 알고리듬의 성능 비교를 위하여 독립된 검증 데이터를 사용하였고 본 논문의 알고리듬의 성능이 기존 논문의 알고리듬 보다 더 나음을 보였다.

본 논문에서 식별된 예후 특이 네트워크를 구성하는 유전자들의 기능을 검증하기 위하여 Gene Ontology 데이터베이스를 사용하였다. 암, 면역, 에스트로겐 수용체와 관련이 있는 GO term과 유방암 예후에 관련된 다수의 GO term이 선별되었으며 이를 통하여 실제로 본 논문에서 식별된 예후 특이 네트워크가 유방암의 예후와 관련됨을 검증하였다.

REFERENCES

[1] GARCIA, Maxime, et al. "Interactome-transcriptome integration for predicting distant metastasis in breast cancer." Bioinformatics, 28.5: 672-678, Jan. 2012.

[2] Ministry of Health and Welfare. "Annual report of the Central Cancer Registry in Korea, Seoul." Ministry of Health and Welfare, 2001

[3] VAN DE VIJVER, Marc J., et al. "A gene-expression signature as a predictor of survival in breast cancer." New England Journal of Medicine, 347.25: 1999-2009, Dec. 2002.

[4] WANG, Yixin, et al. "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." The Lancet, 365.9460: 671-679, Feb. 2005.

[5] CHUANG, Han-Yu, et al. "Network-based classification of breast cancer metastasis." Molecular systems biology, 3.1, Jan. 2007.

[6] TAYLOR, Ian W., et al. "Dynamic modularity in protein interaction networks predicts breast cancer outcome." Nature biotechnology, 27.2: 199-204, Feb. 2009.

[7] Taylor CR, et al. "Detection of estrogen receptor in breast and endometrial carcinoma by the immunoperoxidase technique." Cancer 47:2634-2640, June 1981.

[8] Scottish Cancer Trials Breast Group and ICRF Breast Unit GHL: "Adjuvant ovarian ablation versus CMF chemotherapy in premenopausal women with pathological stage II breast carcinoma: the Scottish trial." Lancet 341: 1293-1298, May 1993

[9] REIS-FILHO, Jorge S.; PUSZTAI, Lajos. "Gene expression profiling in breast cancer: classification, prognostication, and prediction."

The Lancet, 378.9805: 1812-1823, Nov. 2011.
〔10〕 FISHER, Bernard, et al. "Relative worth of estrogen or progesterone receptor and pathologic characteristics of differentiation as indicators of prognosis in node negative breast cancer patients: findings from National Surgical Adjuvant Breast and Bowel Project Protocol B-06." Journal of Clinical Oncology, 6.7: 1076-1087, July 1988.
〔11〕 MCGUIRE, William L., et al. "How to use prognostic factors in axillary node-negative breast cancer patients." Journal of the National Cancer Institute, 82.12: 1006-1015, June 1990.
〔12〕 BEZWODA, Werner Robert, et al. "The value of estrogen and progesterone receptor determinations in advanced breast cancer. Estrogen receptor level but not progesterone receptor level correlates with response to tamoxifen." Cancer, 68.4: 867-872, Aug. 1991.
〔13〕 ABE, O., et al. "Tamoxifen for early breast cancer: an overview of the randomised trials." Lancet, 351.9114: 1451-1467, May 1998.
〔14〕 MACKAY, Joel P., et al. "Protein interactions: is seeing believing?" Trends in biochemical sciences, 32.12: 530-531, Nov. 2007.
〔15〕 CHATR-ARYAMONTRI, Andrew, et al. "Protein interactions: integration leads to belief." Trends in biochemical sciences, 33.6: 241-242, May 2008.
〔16〕 DE LAS RIVAS, Javier; FONTANILLO, Celia. "Protein-protein interactions essentials: key concepts to building and analyzing interactome networks." PLoS computational biology, 6.6: e1000807, June 2010.
〔17〕 Biomolecular Interaction Network Database, http://bond.unleashedinformatics.com/
〔18〕 Biological General Repository for Interaction Datasets, http://www.thebiogrid.org/
〔19〕 Human Protein Reference Database, http://www.hprd.org/
〔20〕 IntAct Molecular Interaction Database, http://www.ebi.ac.uk/intact/
〔21〕 Molecular INTeraction database, http://mint.bio.uniroma2.it/mint/
〔22〕 Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/dip/
〔23〕 MIPS protein interaction resource on yeast, http://mips.gsf.de/genre/proj/mpact/
〔24〕 Online Predicted Human Interaction Database, http://ophid.utoronto.ca/
〔25〕 Guo Yu, "Statistical issues in microarray data analysis: Array-to-array normalization, Empirical Bayes batch effect adjustment."
〔26〕 ASHBURNER, Michael, et al. "Gene Ontology: tool for the unification of biology." Nature genetics, 25.1: 25-29, Oct. 2000.
〔27〕 GENE ONTOLOGY CONSORTIUM, et al. "The gene ontology project in 2008." Nucleic acids research, 36.suppl 1: D440-D444, Nov. 2007.
〔28〕 DENNIS JR, Glynn, et al. "DAVID: database for annotation, visualization, and integrated discovery." Genome biol, 4.5: P3, Aug. 2003.
〔29〕 Gene Ontology, http://www.geneontology.org
〔30〕 AHN, Jaegyoon, et al. "Integrative gene network construction for predicting a set of complementary prostate cancer genes." Bioinformatics, 27.13: 1846-1853, May 2011.
〔31〕 LOI, Sherene, et al. "Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade." Journal of clinical oncology, 25.10: 1239-1246, April 2007.
〔32〕 SCHMIDT, Marcus, et al. "The humoral immune system has a key prognostic impact in node-negative breast cancer." Cancer research, 68.13: 5405-5413, July 2008.
〔33〕 DESMEDT, Christine, et al. "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series." Clinical cancer research, 13.11: 3207-3214, June 2007.
〔34〕 NAGALLA, Srikanth, et al. "Interactions

between immunity, proliferation and molecular subtype in breast cancer prognosis." Genome Biol, 14.4: R34, April 2013.

저 자 소 개

황 유 현(Youhyeon Hwang)
2015년 : 가천대학교
컴퓨터공학과 졸업 예정
관심분야 : 바이오인포매틱스,
데이터 마이닝,
데이터 베이스
Email : youhyeonhwang@gmail.com

오 민(Min Oh)
2015년 : 가천대학교
컴퓨터공학과 졸업 예정
관심분야 : 바이오인포매틱스,
데이터 마이닝,
그래프 데이터 마이닝
Email :minoh0201@gmail.com

윤 영 미(Youngmi Yoon)
1981년 : 서울대학교
자연과학대학 졸업(학사)
1983년 : 오하이오 주립대학
수학과(학사수료)
1987년 : 스탠포드대학교
컴퓨터과학과 졸업(이학석사)
2008년 : 연세대학교 컴퓨터과학과
졸업(공학박사)
1987년 5월 ~ 1993년 5월 :
IntelliGenetics Inc.,
California, USA,
Software Engineer
1995년 2월 ~ 현재 : 가천대학교
컴퓨터공학과 교수
관심분야 : 데이터베이스 시스템,
데이터 마이닝,
바이오인포매틱스,
소셜미디어 데이터 마이닝
Email : ymyoon@gachon.ac.kr