Regular paper

Journal of the KIECS. pp. 579-586, vol. 13, no. 3, Jun. 30 2018, t. 89, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2018.13.3.579

효소 반응 예측을 위한 유사도 모델 분석 및 구현

오주성* · 나도균** · 박춘구* · 정희택***

Similarity Model Analysis and Implementation for Enzyme Reaction Prediction

Joo-Seong Oh* · Do-Kyun Na** · Chun-Goo Park* · Hyi-Thaek Ceong***

요 약

빅데이터에 대한 관심이 증가하면서 데이터로부터 의미 있는 정보의 추출 및 예측은 중요한 연구분야가 되고 있다. 본 연구에서는 신약개발과정에서 필요한 후보약물의 약리적인 활성을 분석하기 위한 데이터를 획득하고 이를 기반으로 의미 있는 예측 분석을 하고자 한다. 신약개발과정에서 대사반응 된 신약후보물질의 약리적인 활성 연구는 신약개발 성공률을 높이기 위해 필요한 단계이다. 본 연구에서, 약용 후보물질의 체내 효소 반응 유무를 예측하기 위해, 유사도 모델들을 적용 분석하였다. 유사도 모델의 군집별 특성을 반영하여 13개의 모델을 선택하여 효소 반응 예측을 수행하였다. 이들 모델들을 민감도와 AUC를 기반으로 비교 평가하였다. 평가 모델들 중, 효소 사이의 반응성을 예측하는데 있어서 Simpson coefficient 모델이 가장 좋은 성능을 보였다. 분석된 유사도 모델 전체를 웹 서비스로 구축하였다. 제안된 모델은 반응정보의 추가에 동적으로 대응 할 수 있으며 신약개발시간 단축 및 비용 절감에 기여할 것으로 여겨진다.

ABSTRACT

With the beginning of the new era of bigdata, information extraction or prediction are an important research area. Here, we present the acquisition of semi-automatically curated large-scale biological database and the prediction of enzyme reaction annotation for analyzing the pharmacological activities of drugs. Because the xenobiotic metabolism of pharmaceutical drugs by cellular enzymes is an important aspect of pharmacology and medicine. In this study, we apply and analyze similarity models to predict bimolecular reactions between human enzymes and their corresponding substrates. Thirteen models select to reflect the characteristics of each cluster in the similarity model. These models compare based on sensitivity and AUC. Among the evaluation models, the Simpson coefficient model showed the best performance in predicting the reactivity between the enzymes. The whole similarity model implement as a web service. The proposed model can respond dynamically to the addition of reaction information, which will contribute to the shortening of new drug development time and cost reduction.

키워드

Enzyme Reaction Prediction, Similarity Model, Sensitivity, Web Service
효소 반응 예측, 유사도 모델, 민감도, 웹 서비스

* 전남대학교 생명과학기술학부(ojooseong@gmail.com, chungoo@jnu.ac.kr)
** 중앙대학교 융합공학부(blisszen@lile.cau.ac.kr)
*** 교신저자 : 전남대학교 멀티미디어전공
· 접 수 일 : 2018. 03. 14
· 수정완료일 : 2018. 04. 29
· 게재확정일 : 2018. 06. 15
· Received : Mar. 14, 2018, Revised : Apr. 29, 2018, Accepted : Jun. 15, 2018
· Corresponding Author : Hyi-Thaek Ceong
Dept. of Multimedia, Chonnam National University.
Email : htceong@jnu.ac.kr

Ⅰ. 서 론

빅데이터와 머신러닝에 대한 관심이 증가하면서 다양한 연구 분야에서 데이터의 축적은 매우 중요한 화두가 되고 있다. 생물학, 의학, 약동학 등의 분야에서도 이와 같은 데이터의 축적은 꾸준히 이어져 오고 있고 이를 활용하여 유의미한 정보를 찾기 위한 연구들이 활발히 진행중이다. 이로 인해 새로운 물질들이 발굴되고 천연물을 활용하는 등 기존에 사용되지 않았던 약재들에 대한 관심이 증가하고 있다. 특히나 천연물 같은 경우는 동의보감, 본초강목처럼 예전부터 그 효능에 대해 경험적으로 알려진 부분들이 있기 때문에 관련된 연구가 활발히 진행되고 있다[1-6].

천연물은 물질들 간에 서로 상승작용을 일으키는 것들이 체내의 여러 효소들에 작용하는 MCMT(: Multi-Component Multi-Target)을 함으로 인해서 약효를 상승시키거나 독성을 감소시키는 등의 기작을 일으킨다고 알려져 있고, 신약의 개발을 위해 이러한 기작들을 규명하기 위한 연구가 이루어지고 있다. 특히 천연물은 표적에 직접 작용하는 것뿐만 아니라 효소에 의해 대사변환을 일으켜 효능을 발하는 것들도 존재한다. 이와 같은 천연물이 인체 내에서 일으키는 변화에 대해서 실험적으로 검증을 하기 위해서는 많은 시간과 노력이 필요하다. 이로 인해 천연물을 기반으로 하는 외인성 대사물질이 인체 내에서 효소와의 작용으로 인해 일으키는 변화와 해당 물질의 효능에 대한 활용을 위해 효소의 대사반응에 대한 특징을 지식베이스로 구축하는 연구가 진행되었다. 본 연구는 효소의 대사반응에 대한 데이터베이스를 활용하여 천연물이나 새로운 물질이 체내에 들어왔을 때, 어떤 효소에 의해 변환될지를 예측하고자 한다. 예측을 위해서 유사도 모델들을 사용하여 분석하고, 다양한 유사도 모델 중에서 어떤 모델이 효소의 반응성을 예측하는데 적합한지 제안한다. 이를 위해 2장에서는 연구를 위해 사용된 효소 대사반응과 관련된 데이터베이스와 유사도 분석에 대한 개념을 서술하고, 3장에서는 유사도 분석 방법과 유사도 모델 별 분석 결과를 비교한다. 4장에서는 이를 활용한 효소 반응성 예측 시스템을 소개한다. 마지막으로 5장에서는 결론 및 향후 연구에 대해 제시한다.

Ⅱ. 관련연구

2.1 효소의 대사정보 데이터베이스

2.1.1 HMDB

HMDB(: Human Metabolome Database)는 인체에서 발견되는 작은 분자의 대사 정보를 제공하는 데이터베이스로 화학적 데이터, 임상 데이터, 분자 생물학과 생화학 데이터를 포함한다. 데이터베이스는 총 114,110 개의 대사물질에 대한 정보를 제공하고, 이는 5,702 개의 단백질 서열 정보와 연결되어 있다. 대사물질에 대한 정보는 총 130가지 제공되고, 대표적으로 화학식, SMILES(: Simplified Molecular-Input Line-Entry System, 화학 물질 종의 구조를 ASCII문자열로 기술한 것), InChI(: International Chemical Identifier, 데이터베이스와 웹에서 검색을 용이하게 하기위해 분자 정보를 암호화한 화학 물질의 텍스트 식별자), InChiKey(웹 검색을 쉽게 하기 위해 해쉬된 InChI 코드), 물리적 특성, 생물학적 위치, 관련된 질병 정보들을 제공한다. 이 중 효소 반응 예측을 위해 효소정보와 반응하는 대사물질의 정보를 획득하였다. 이 데이터는 유사도 모델 분석을 통해서 효소의 대사반응 예측을 위한 초기 데이터로 사용하였다[7].

2.1.2 BRENDA

BRENDA(: BRaunschweig ENzyme DAtabase)는 가장 많은 효소 정보를 담고 있는 데이터베이스로 83,000개의 효소 정보와 이와 반응하는 207,000 개의 효소 화합물 정보를 포함한다. 각 효소 정보는 EC(: Enzyme Commission) 번호를 기준으로 분류되어 있으며 대사 대상이 되는 대사물 집합인 기질(substrate), 반응 결과에 의한 대사물 집합인 생성물(product)과 억제 화합물(inhibitor), 보조인자(cofactor), 활성 화합물(activating compound), 관련 종(organism) 등의 정보를 제공한다[8].

효소와 대사물의 반응에 대한 정보는 실험을 통해 발견한 대사반응정보와 생체 내에서 실제로 나타나는 대사반응정보를 모두 포함하고 있다. 이러한 정보는 관련 문헌들을 통해 수집하여 텍스트 마이닝, 외부 데이터 통합, 예측 알고리즘을 통해 확장되었다.

2.2. 유사도 모델

일반적으로 유사도는 두 개체간의 특징들이 일치하는 정도를 나타낸다. 두 개체 A와 B에서, 각 개체들의 특징에 대한 집합을 $\{X_{jA}\},\{X_{jB}\}$로 나타낼 때, 두 개체간의 유사도는 수식(1)과 같이 표현 할 수 있다[9].

$$S_{A,B} = X_{iA} \bigcap X_{iB},\ i = 1, 2, ..., n \qquad (1)$$

2.2.1 유사도 모델 유형

유사도 모델은 크게 거리 기반 방법, 연관 계수 방법, 상관 계수 방법 3 가지 유형으로 구분된다. 첫 번째로 거리 기반 방법은 두 개체 사이의 차이 정도에 대해서 정량화하는 방법으로 다양한 분야에서 많이 사용되고 있다. 이는 개체들 간의 유사도를 기하학적인 개념을 통해 간단히 계산할 수 있기 때문이다. 거리 기반 방법은 두 개체 사이의 차이를 거리로 나타내기 때문에 값이 작을수록 유사도가 크다. 두 번째로 연관 계수 방법은 개체 사이의 서술자(descriptor)의 존재 유무에 따라 이진 데이터(0과 1)로 표현한다. 또한 비-이진 데이터를 함께 사용하기 위해서는 적합한 변환식을 사용하여야 한다. 마지막으로 상관 계수 방법은 두 개체 각각을 표현하는 특징들의 집합에 대한 상관관계를 측정하는 방법이다[10].

표 1. 유사도 모델과 수식
Table 1. Similarity models and formula

Model name	Formula	Model name	Formula
Chebyshev	$d_{A,B} = \max_i \lvert X_{iA} - X_{iB} \rvert$ (2) $S_{A,B} = 1 - d_{A,B}$	Taneja	$d_{A,B} = \sum_{i=1}^{n} \left(\frac{X_{iA} + X_{iB}}{2} \right) \ln \left(\frac{X_{iA} + X_{iB}}{2\sqrt{X_{iA} X_{iB}}} \right)$ (9) $S_{A,B} = 1 - d_{A,B}$
Cosine	$S_{A,B} = \frac{\sum X_{iA} X_{iB}}{\sqrt{\sum X_{iA}^2} \sqrt{\sum X_{iB}^2}}$ (3)	Tanimoto	$d_{A,B} = \frac{\sum (\max(X_{iA}, X_{iB}) - \min(X_{iA}, X_{iB}))}{\sum \max(X_{iA}, X_{iB})}$ (10) $S_{A,B} = 1 - d_{A,B}$ (10)
Euclidean	$d_{A,B} = \sqrt{\sum \lvert X_{iA} - X_{iB} \rvert^2}$ (4) $S_{A,B} = 1 - d_{A,B}$		
Jeffreys	$d_{A,B} = \sum (X_{iA} - X_{iB}) \ln \frac{X_{iA}}{X_{iB}}$ (5) $S_{A,B} = 1 - d_{A,B}$	Wave Hedges	$d_{A,B} = \sum_{i=1}^{n} \frac{\lvert X_{iA} - X_{iB} \rvert}{\max(X_{iA}, X_{iB})}$ $S_{A,B} = 1 - d_{A,B}$ (11)
Pearson χ^2	$d_{A,B} = \sum \frac{(X_{iA} - X_{iB})^2}{X_{iB}}$ (6) $S_{A,B} = 1 - d_{A,B}$	Pearson correlation coefficient	$\gamma_{A,B} = \frac{n \sum X_{iA} X_{iB} - \sum X_{iA} \sum X_{iB}}{\sqrt{[n \sum X_{iA}^2 - (\sum X_{iA})^2][n \sum X_{iB}^2 - (\sum X_{iB})^2]}}$ $S_{A,B} = \frac{\gamma_{A,B} + 1}{2}$ (12)
Sorensen	$d_{A,B} = \frac{\sum \lvert X_{iA} - X_{iB} \rvert}{\sum (X_{iA} + X_{iB})}$ (7) $S_{A,B} = 1 - d_{A,B}$	Simpson coefficient	$K_{A,B} = \frac{\sum \min(X_{iA}, X_{iB})}{\min[\sum X_{iA}, \sum X_{iB}]}$ (13) $S_{A,B} = K_{A,B}$
Squared chord	$d_{A,B} = \sum \left(\sqrt{X_{iA}} - \sqrt{X_{iB}} \right)^2$ (8) $S_{A,B} = 1 - d_{A,B}$	Spearman correlation coefficient	$\rho_{A,B} = 1 - \frac{6 \sum D_{iA,iB}}{n(n^2 - 1)}$ (14) $S_{A,B} = \frac{\rho + 1}{2}$

2.2.2 유사도 모델 선정 및 유사도 계산

본 연구에서는 유사도를 분석할 수 있는 다양한 모델 중 [11-12]에서 분석한 유사도 모델들의 군집화 결과를 기본으로 동일한 군집에서 중복 사용되지 않도록 모델을 선정하였다. 이 과정을 통해 Chebyshev L infinite similarity, Cosine similarity, Jeffreys similarity, Pearson χ^2 similarity, Sorensen similarity, Squared chord similarity, Taneja similarity, Tanimoto similarity, Wave Hedges similarity 모델 8개를 선정하였다. 한편 가장 널리 이해되고 많이 사용되는 Euclidean similarity, Tanimoto similarity 모델을 추가하였다. 또한 앞서 언급한 연관계수 및 상관 계수 방법을 제공하기 위해 Pearson correlation coefficient, Simpson coefficient, Spearman correlation coefficient 모델 3개를 선정하였다. 선택된 유사도 모델은 거리기반, 연관계수, 상관계수 방법을 모두 포함하는 13개 모델을 선정하였다. 선정된 13개 모델의 수식은 표 1과 같다.

표 1에서 거리기반인 경우 유사도를 식(15)와 같이 정의한다.

$$S_{A,B} = 1 - d_{A,B} \tag{15}$$

Pearson과 Spearman 상관계수는 $-1 \leqq \gamma_{A,B}, \rho_{A,B} \leqq 1$범위의 값이기 때문에 이를 0과 1사이로 정규화 하기 위해 식(16), 식(17)과 같이 정의한다. Spearman에서 $D_{iA,iB}$는 i에서 등위의 차이다.

$$S_{A,B} = \frac{\gamma_{A,B}+1}{2} \tag{16}$$

$$S_{A,B} = \frac{\rho_{A,B}+1}{2} \tag{17}$$

III. 유사도 모델을 기반으로 한 효소 반응성 예측 실험

효소는 기질과 반응하여 생성물을 만들어 내는 대사반응을 하고, 억제 화합물과 활성 화합물을 통해 대사반응의 활성화 정도가 조절된다.

본 연구에서는 기질이 반응하는 효소를 예측하기 위한 데이터 집합으로, 대사반응에 대한 데이터가 축적되어 있는 BRENDA와 HMDB에서 정보를 수집했다. 이 중 인간과 관련된 효소의 대사반응 데이터를 추출하여 분석에 사용하였다. 인간과 관련된 데이터는 많은 연구로 인해 효소의 대사경로에 대한 신뢰도가 높아 효소 반응성 예측을 위한 모델을 설계하는데 있어서 적합하다. 분석에 사용된 데이터는 효소 2,531개와 기질 6,069개를 포함하고, 각 효소는 최소 1개에서 최대 57개의 기질 정보를 가진다.

이렇게 수집된 기질 정보에 대해, 효소 반응성 예측을 위해 각 기질들에 대한 물리화학적 정보가 필요하다. 기질의 물리화학적 정보를 기반으로 유사도 예측을 통해, 효소의 반응성을 예측할 수 있기 때문이다. 기질의 물리화학적 정보를 추출하기 위해 PaDEL descriptor[13]을 사용하였다. PaDEL descriptor는 가장 많은 물리화학적 정보를 추출할 수 있고 가장 널리 사용되기 때문에 선택하였다. PADEL descriptor는 기질 당 1,444개의 descriptor를 생성한다.

추출된 물리화학적 정보를 기반으로 13개 유사도 모델에 대한 평가를 수행한다. 이들 모델들의 비교는 민감도(Sensitivity)와 AUC(: Area Under an ROC Curve)[14] 분석을 기반으로 한다. 이 값을 활용하여 모델간의 성능을 비교한다.

3.1 효소와 기질 정보의 특성

하나의 효소는 하나 이상의 기질에 반응할 수 있다. 분석 대상이 되는 데이터에서는 효소 하나가 최대 57개의 기질과 반응한다는 정보를 표함하고 있다. 달리 표현하면 57개의 기질 정보가 하나의 효소를 설명할 수 있음을 의미한다. 또한 각 기질은 1444개의 물리화학적 정보로 표현될 수 있다. 정리하면, 하나의 효소가 57개의 기질과 반응한다는 것은 82,308 (= 57 ×1444)개의 물리화학적 정보로 효소를 설명하고 있음을 의미한다.

3.2 질의 기질에 대한 효소 반응성 계산

질의 기질에 대해 2,531개의 효소 중 가장 반응성이 높은 효소를 예측하기 위해, 본 연구에서는 유사도 모델을 적용한다. 이는 질의 기질과 효소를 직접 비교하는 것이 아니고, 효소가 반응한다고 알려져 있는 효소의 기질들과 질의 기질사이에 유사도를 계산하는 것을 의미한다.

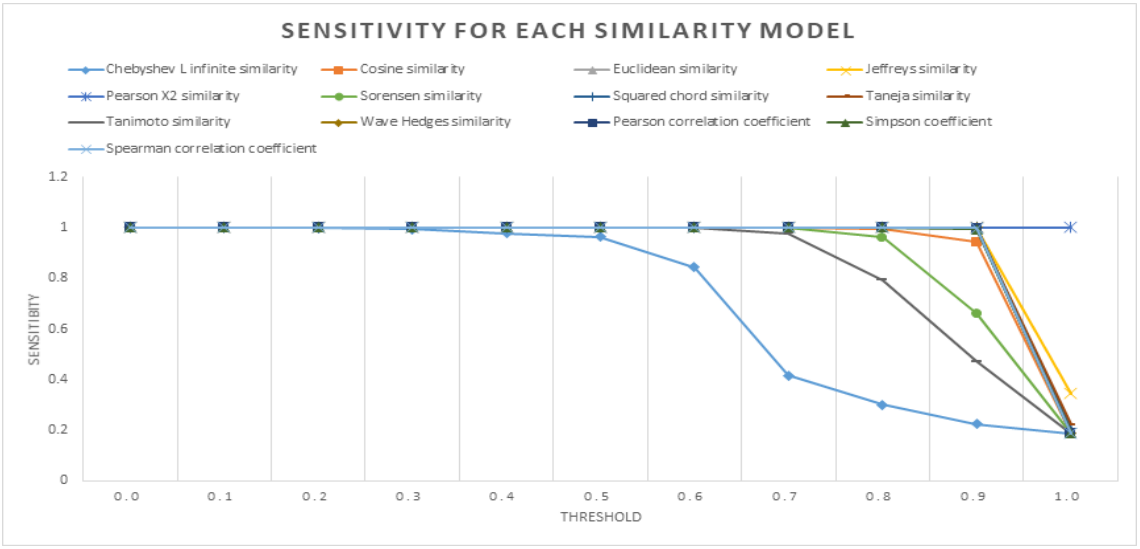


그림 1. 유사도 모델에 대한 민감도

Fig. 1 Sensitivity for each similarity model

효소가 하나 이상의 기질들과 반응하기 때문에, 질의 기질에 대해 효과적인 반응성 계산 방안이 필요하다. 예를 들어 효소와 그 효소가 반응하는 기질들을 $E_1=\{s_{11},s_{12}\}$, $E_2=\{s_{21},s_{22}\}$라 하고 질의 기질을 Qs라 할 때, 임의의 유사도 모델에 의해 계산된 유사도를 $SE_1=\{S_{S_{11},Qs},S_{S_{12},Qs}\}=\{0.5,0.5\}$, $SE_2=\{S_{S_{21},Qs},S_{S_{22},Qs}\}=\{0,1\}$ 이라 가정한다. 이때 효소와 질의 기질 사이의 반응성을 평균에 의해 계산한다면 Qs와 E_1사이의 반응성은 0.5이고 Qs와 E_2사이의 반응성은 0.5이다. 그러나 의미적으로 기질 차원의 물리화학적 정보 관점에서 E_2가 높은 유사도 값(즉, 1)을 가지고 있기 때문에, E_2와 Qs사이에 보다 높은 반응성으로 계산해야 한다. 이를 위해 식(18)을 제안한다.

$$p_{E_i,Qs}=\overline{S_{E_i,Qs}}+\sum_{j=1}^{k}\left(S_{S_{i,j},Qs}-\overline{S_{E_i,Qs}}\right)\times\sqrt{\frac{\sum_{j=1}^{k}\left(S_{S_{i,j},Qs}-\overline{S_{E_i,Qs}}\right)^2}{k}} \quad (18)$$

$$where\ \overline{S_{S_{i,j},Qs}}<S_{S_{i,j},Qs}$$

제안한 수식의 의미는 단순 합이나 평균을 사용하게 되면 소수의 이상치 때문에 대표 유사도 값이 편향될 수 있다. 따라서 평균을 기준으로 하여 평균값 이상의 값에 대해서 논리곱을 사용하여 표준편차를 가중하였다. 이는 평균보다 높은 유사도 점수들에 대해서 높은 민감도로 분석을 할 수 있다. 즉, 평균 이상의 유사도에 보다 높은 반응성을 반영하여 민감도를 높이는 것이다. 위 수식에 의해 Qs와 E_1사이의 반응성은 0.5이고 Qs와 E_2사이의 반응성은 0.75이다. $p_{E_i,Qs}$ 값의 범위는 식(19)와 같다.

$$\overline{S_{E_i,Qs}}<\ p_{E_i,Qs}<\ \frac{1}{4}\left(\sum_{i=1}^{k}\left(S_{S_{i,k},Qs}-\overline{S_{E_i,Qs}}\right)+\sqrt{\frac{\sum_{t=1}^{k}\left(S_{S_{i,k},Qs}-\overline{S_{E_i,Qs}}\right)}{k}}\right)^2 \quad (19)$$

위 과정을 반복 수행하여 질의 기질과 해당 효소와 반응하는 기질들에 대한 유사도를 계산하고, 이를 기반으로 반응성을 측정하여 가장 높은 반응성을 갖는 효소를 제시한다.

3.3 실험 환경 및 결과

유사도 모델을 활용한 질의 기질에 대한 효소 반응성을 예측하기 위해, 먼저 획득한 2,531개의 효소 및 6,069개 기질 정보로부터 유일한 기질 정보 1,685개를 추출하였다. 이를 질의 기질(true set)로 구성하였다. 다음으로 6,069개의 기질에 대한 물리화학적 정보를 구축하기 위해, PaDEL descriptor를 활용하여 각 기질 당 1,444개의 속성 정보를 추출하였다. 다음으로 표 1에서 선별한 유사도 모델을 Java 언어로 구현하였다. 이를 바탕으로 질의 기질 1,685개에 대해 0.1부터 1까지 반응성 임계값(threshold)를 지정하여 민감도를 분석하였다.

하나의 질의 기질에 대한 반응성 계산에 있어서 입력되어지는 1,444개의 물리화학적 정보는 다양한 범위의 실수 값이기 때문에, 일관성 있는 처리를 위해 질의 기질과 유사도 계산 대상 기질들에 대한 정규화를 수행하였다. 정규화 된 물리화학적 속성 1,444개를 대상으로 유사도 계산을 수행하였다. 13개 유사도 모델을 사용한 계산과 반응성 결과가 다양한 값의 범위를 갖기 때문에 각 반응성 값을 모델별로 정규화 하였다.

결과에 대한 분석을 위해 민감도와 AUC 값을 사용하였다. 민감도는 정확하게 예측된 결과 중에서 실제 결과 또한 사실일 경우를 나타내는 변수로 모델 구축과 입력 값에 대한 품질 보증을 위한 필수요소이다. 본 연구에서는 결과 값에 대한 임계 값을 결정하는데 사용된다. 두 개체간의 유사도에 대해서 임계 값을 만족하는 데이터로 민감도를 분석했을 때, 민감도가 감소하는 부분은 효소와 기질간의 예측에 대한 신뢰도가 감소하는 것으로 판단할 수 있다. 따라서 임계값에 대한 민감도를 가시화하고 민감도가 감소하는 임계 값을 찾는다. 이는 이후 분석을 위한 임계 값을 선정하는데 중요한 요소가 될 수 있다. 실험결과는 그림 1과 같다. 민감도에 있어 Pearson χ^2 similarity가 가장 우수 하였으며, Jeffreys similarity, 다음으로 Taneja similarity 순이었다.

AUC 값은 통계분석이나 머신러닝의 결과를 평가하기 위해 많이 쓰이는 기준으로 민감도(True Positive Rate)를 y축으로 하고, 특이도(Specificity, False Negative Rate)를 x축으로 하는 ROC(: Receiver Operating Characteristic curve) 그래프의 밑 면적을 나타낸다. 이는 민감도와 특이도의 상관관계를 확인할 수 있어서 결과 값의 성능을 평가하는데 유용한 방법이다. 평가 값은 0부터 1사이의 값을 가지고, 값이 1에 가까울수록 민감도가 증가하고 특이도가 감소하기 때문에 결과 값의 성능이 좋다고 판단할 수 있다.

유사도 모델에 대한 AUC는 그림 2와 같다. 가장 좋은 성능을 보이는 모델은 Simpson coefficient 모델이었고, 다음으로 Sorensen similarity, Cosine similarity 순으로 나왔다.

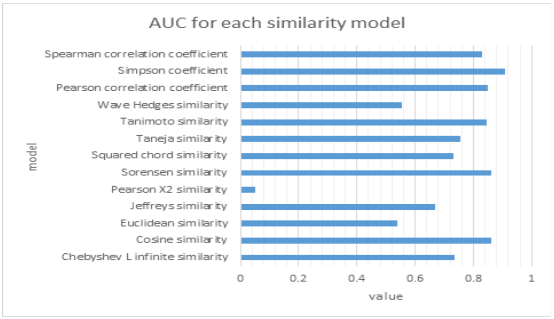


그림 2. 유사도 모델들의 AUC
Fig. 2 AUC for each similarity model

민감도와 AUC 모두를 고려할 때, Simpson coefficient가 가장 좋을 결과를 생성했고 다음으로 Pearson χ^2 similarity, Pearson correlation coefficient 순이었다.

Ⅳ. 효소 반응성 예측 시스템 구축

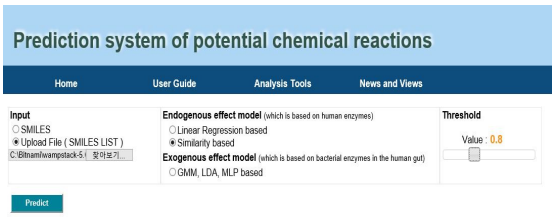
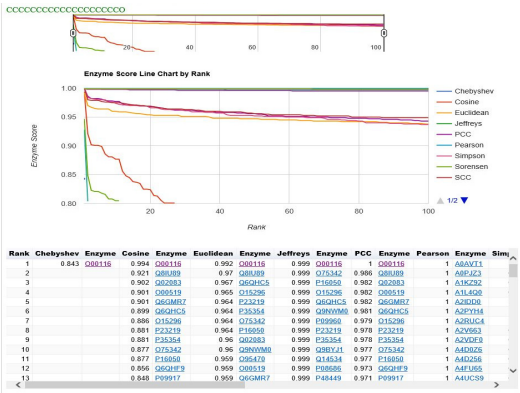


그림 3. 유사도 모델 기반 효소 반응성 예측 시스템
Fig. 3 Enzyme reaction prediction based on similarity model

웹을 기반으로 제안한 효소 반응성 예측 시스템을 구축하였다. URL은 http://rxn.jnu.ac.kr이다. 시스템

구축을 위해 2장에서 제시한 HMDB와 BRENDA 데이터를 획득하여 데이터베이스를 구축하였고, 표1에서 제시한 모델들을 구현하였다. 사용방법은 그림 3과 같은 화면에서 하나의 질의 기질의 SMILES 값을 입력하거나 여러 기질들의 SMILES로 구성된 파일을 입력하고 임계치를 지정해야 한다. 지정된 임계치 이상의 유사도를 갖는 효소들의 정보를 그림 4와 같이 확인할 수 있다. 사용자의 편의성을 위해 그림 4와 같이 구글차트를 활용하여 제시한다. 이때 기준은 반응성 값에 따라 효소명과 반응성 값을 내림차순으로 제시한다.



Rank	Chebyshev	Enzyme	Cosine	Enzyme	Euclidean	Enzyme	Jeffreys	Enzyme	PCC	Enzyme	Pearson	Enzyme	Sim
1	0.843	Q00116	0.994	Q00116	0.992	Q00116	0.999	Q00116	1	Q00116	1	A8AVT1	
2			0.921	Q8IU89	0.97	Q8IU89	0.999	Q75342	0.986	Q8IU89	1	A8PJZ3	
3			0.902	Q02083	0.967	Q6QHC5	0.999	P16050	0.982	Q02083	1	A1KZ92	
4			0.901	Q00519	0.965	O15296	0.999	O15296	0.982	Q00519	1	A1L4Q8	
5			0.901	Q6GMR7	0.964	P23219	0.999	Q6QHC5	0.982	Q6GMR7	1	A2IDD0	
6			0.899	Q6QHC5	0.964	P35354	0.999	Q9NWM0	0.981	Q6QHC5	1	A2PYH4	
7			0.886	O15296	0.964	O75342	0.999	P00960	0.979	O15296	1	A2RUC4	
8			0.881	P23219	0.964	P16050	0.999	P23219	0.978	P23219	1	A2V663	
9			0.881	P35354	0.96	Q02083	0.999	P35354	0.978	P35354	1	A2VDF8	
10			0.877	O75342	0.96	Q8NWM0	0.999	Q8EYJ1	0.977	O75342	1	A4D0Z6	
11			0.877	P16050	0.959	O95470	0.999	Q14534	0.977	P16050	1	A4D256	
12			0.856	Q6QHF9	0.959	O00519	0.999	P08686	0.973	Q6QHF9	1	A4F1U6	
13			0.848	P09917	0.959	Q6GMR7	0.999	P48449	0.971	P09917	1	A4UCS9	

그림 4. 효소 반응성 예측 결과
Fig. 4 Enzyme reaction prediction result

V. 결 론

본 연구는 천연물과 약용 후보물질이 체내에 유입되었을 때, 어떠한 효소와 대사반응을 나타낼지 유사도 모델을 통해 예측을 하였다. 그리고 어떤 유사도 모델이 예측에 적합한지를 분석하였다. 총 13가지의 상이한 유사도 모델을 통해 분석을 하였다. 결과 값에서 민감도와 AUC를 기반으로 비교 분석하였다. 그 결과 천연물, 약용 후보물질과 체내의 효소 사이의 반응성을 예측하는데 있어서 Simpson coefficient 모델이 가장 좋은 성능을 보여주었다. 이는 이후 천연물이나 약용 후보물질들을 통해 신약을 개발하는데 있어서 시간과 비용의 단축에 기여를 할 수 있다.

향후 연구에서는 인간의 내장에 존재하는 미생물들의 대사 데이터를 수집, 분석하여 본 연구를 확장 시키고자 한다. 장에 존재하는 미생물들의 대사 반응이 인체에 미치는 영향에 대해 규명을 하는 것은, 인체에 유용물질의 발굴에 있어서 도움이 되기 때문이다.

감사의 글

본 연구는 미래창조과학부 및 한국연구재단의 (재)유전자동의보감사업단 (NRF-2015M3A9C4075820) 연구비 지원에 의해 수행되었습니다.

References

[1] A. Tarca, V. Jarey, X. Chen, R. Romero, and S. Drăghici, "Machine Learning and Its Applications to Biology," *J. of Public Library of Science(PLOS) Computational Biology*, vol. 3, issue 6, 2007, pp. 953-963.

[2] K. Park, D. Kim, S. Ha, and D. Lee, "Predicting pharmacodynamic drug-drug interactions through signaling propagation interference on protein-protein interaction networks," *J. of Public Library of Science(PLOS) ONE*, vol. 10, no. 10, 2015, pp. 1-13.

[3] H. Ceong and C. Park, "Enzyme Metabolite Analysis Using Data Mining," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 10, 2016, pp. 969-982.

[4] G. Jim and H. Lee, "The Developement of Liver cancer Vital Sign Information Prediction System using Aptamer Protein Biochip," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 6, no. 6, 2011, pp. 965-971.

[5] S. Yoon and G. Kim, "Personal Biometric Identification based on ECG Features," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 10, no. 4, 2015, pp. 521-526.

[6] Y. Kim, W. Kim and M. Jo, "Learning System for Big Data Analysis based on the Raspberry

Pi Board," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 4, 2016, pp. 433-439.

[7] D. Wishart, T. Jewison, A. Guo, M. Wilson, C. Knox, Y. Liu, and S. Bouatra, "HMDB 3.0 – The Human Metabolome Database in 2013," *Nucleic Acids Research*, vol. 41, issue D1, 2013, pp. D801-D807.

[8] S. Placzek, I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack, and D. Schomburg, "BRENDA in 2017: new perspectives and new tools in BRENDA," *Nucleic Acids Research*, vol. 45, issue D1, 2017, pp. D380-D388.

[9] V. Monev, "Introduction to Similarity Searching in Chemistry," *Comunication in Mathematical and in Computer Chemistry*, vol. 51, no. 51, 2004, pp. 7-38.

[10] D. Ellis, J. F. Hines, and P. Willett, "Measuring the degree of similarity between objects in text retrieval systems," *Perspectives in Information Management*, vol. 3, no. 2, 1993, pp. 128-149.

[11] S. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *Int. J. of Mathematical Models and Methods in Applied Sciences*, vol. 1, issue. 4, 2007, pp. 300-307.

[12] J. Holliday, C. Hu, and P. Willett, "Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings," *Combinatorial Chemistry & High Throughput Screening*, vol. 5, issue 2, 2002, pp. 155-166.

[13] C. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *J. of Computational Chemistry*, vol. 32, issue 7, May 2011, pp. 1466-1474.

[14] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition Society*, vol. 30, no. 7, 1997, pp. 1145-1159.

저자 소개

오주성(Joo-Seong Oh)

2011년 목포대학교 컴퓨터교육과 졸업(공학사)
2016년 목포대학교 대학원 컴퓨터공학과 졸업(공학석사)
2016년 ~현재 전남대학교 대학원 생물과학·생명기술학과 박사과정
※ 관심분야 : 기계학습, 생물정보학, 분자생물학

나도균(Dok-Yun Na)

2000년 고려대학교 생명과학부 졸업(이학사)
2002년 고려대학교 생명공학원 졸업(이학석사)
2008년 KAIST 대학원 바이오및뇌공학과 졸업(공학박사)
2013년 ~현재 : 중앙대학교 융합공학부 의료공학전공 교수
※ 관심분야 : 합성생물학, 시스템생물학

박춘구(Chun-Goo Park)

2000년 인하대학교 전자계산공학과 졸업(공학사)
2002년 광주과학기술원 정보통신공학과 졸업(공학석사)
2010년 펜실베니아주립대학교 대학원 생물학과 졸업(이학박사)
2013년~현재 : 전남대학교 생명과학기술학부 조교수
※ 관심분야 : 생물정보학, 의료정보학

정희택(Hyi-Thaek Ceong)

1992년 전남대학교 전산통계학과 졸업(이학사)
1995년 전남학교 대학원 전산통계학과 졸업(이학석사)
1999년 전남대학교 대학원 전산통계학과 졸업(이학박사)
1999년 ~현재 : 전남대학교 멀티미디어전공 교수
※ 관심분야 : 데이터마이닝, 생물정보학, 기계학습