Journal of the Korea Convergence Society
Vol. 10. No. 3, pp. 23-30, 2019

ISSN 2233-4890
https://doi.org/10.15207/JKCS.2019.10.3.023

당뇨병 치료제 후보약물 정보를 이용한 기계 학습 모델과 주요 분자표현자 도출

남궁윤[1], 김창욱[2*], 이창준[3]
[1]연세대학교 융합기술경영공학과 박사과정, [2]연세대학교 산업공학과 교수, [3](주)닷매틱스 응용과학자

A machine learning model for the derivation of major molecular descriptor using candidate drug information of diabetes treatment

Youn Namgoong[1], Chang Ouk Kim[2*], Chang Joon Lee[3]
[1]Ph.D Student, Department of Convergence Technology & Management Engineering, Yonsei University
[2]Professor, Department of Industrial Engineering, Yonsei University
[3]Application Scientist, Dotmatics, Co., Ltd

요 약 본 연구는 당뇨병 치료제 후보약물 정보를 이용하여 항당뇨에 영향을 미치는 물질구조를 발견하는데 목적이 있다. 정량적구조 활성관계를 이용한 기계 학습 모델을 만들고 부분최소자승 알고리즘을 통해 실험데이터 별로 결정계수를 파악한 후 변수중요도척도를 활용하여 주요 분자표현자를 도출하였다. 연구 결과, 후보약물 구조정보를 반영한 molecular access system fingerprint 데이터로 분석한 결과가 *in vitro* 데이터를 이용한 분석 결과보다 설명력이 높았으며, 항당뇨에 영향을 미치는 주요 분자표현자 역시 다양하게 도출할 수 있었다. 제안된 항당뇨 예측 및 주요인자 분석 방법을 활용한다면 유사한 과정을 반복 실험하는 기존 신약개발 방식과는 달리, 많은 비용과 시간이 소요되는 후보물질 스크리닝 (screening) 기간을 최소화하고, 신약개발 탐색기간도 단축하는 계기가 될 수 있을 것으로 기대한다.

주제어 : 후보약물 정보, 부분최소자승, 변수중요도척도, 주요 분자표현자 도출, 항당뇨 예측

Abstract The purpose of this study is to find out the structure of the substance that affects antidiabetic using the candidate drug information for diabetes treatment. A quantitative structure activity relationship model based on machine learning method was constructed and major molecular descriptors were determined for each experimental data variables from coefficient values using a partial least squares algorithm. The results of the analysis of the molecular access system fingerprint data reflecting the candidate drug structure information were higher than those of the *in vitro* data analysis in terms of goodness-of-fit, and the major molecular expression factors affecting the antidiabetic effect were also variously derived. If the proposed method is applied to the new drug development environment, it is possible to reduce the cost for conducting candidate screening experiment and to shorten the search time for new drug development.

Key Words : Candidate drug information, Partial least square, Variable importance in projection, Finding major molecular descriptor, Antidiabetic prediction

*Corresponding Author : Chang Ouk Kim(kimco@yonsei.ac.kr)
Received January 28, 2019 Revised March 8, 2019
Accepted March 20, 2019 Published March 28, 2019

1. 서론

제약산업은 국민의 생명과 건강을 책임지는 미래 성장 산업분야의 성격을 가지고 있다. 특히 신약개발 기간은 여러 단계를 거쳐 평균 13.7년이 소요되며 전임상까지의 개발 성공률은 3%[1]로 성공률이 매우 낮은 고위험 산업의 특징을 가지고 있다. 그 중 탐색기간은 2 ~ 4년[2]이 소요되며, Fig. 1과 같이 이 기간에는 후보물질에 대한 모델링과 합성, *in vitro* (시험관 내 실험) 및 *in vivo* (동물실험) 등이 이루어지고 있으며 신약개발은 탐색과 개발 단계를 거쳐 상용화 단계로 구성되어 있다.

Fig. 1. New drug exploration phase

우리나라 당뇨병 환자는 사회경제가 발달하고 생활이 서구화 되면서 급속히 늘어나고 있다. 2018년 건강보험통계연보 보도자료에 따르면 2010년부터 2017년까지 만성질환 진료현황에서 당뇨병 진료인원은 286만 여명으로 평균 5.1%의 꾸준한 증가세를 보이고 있으며, 특히 2017년에는 전년 대비 5.9%의 증가율과 22,238억 원의 진료비가 발생하고 있다[3].

본 연구는 식(1)과 같이 화합물 구조 (molecular structure)와 활성 (activity) 간의 관계를 데이터를 이용해서 설명하는 정량적구조 활성관계(quantitative structure activity relationship: QSAR) 접근 방식을 채택한다. 구체적으로 본 연구에서는 식 (1)의 관계함수 f를 기계 학습 (machine learning) 분야에서 활용되고 있는 부분최소자승법 (partial least square: PLS)과 변수중요도척도(PLS-variable importance in projection: PLS-VIP) 알고리즘을 사용하여 모델링하는 방법을 제안한다. 제안된 모델을 통해 당뇨병 치료제 유효물질 정보에 대한 주요 분자표현자를 도출하고 항당뇨에 미치는 물질 구조에 대해 예측하여 많은 비용과 시간이 소요되는 후보물질 스크리닝 (screening) 기간 최소화와 신약개발 탐색기간 단축을 목적으로 한다.

$$\text{activity} = f(\text{molecular structure}) + \text{error} \quad (1)$$

2. 관련 연구

2.1 당뇨병

건강보험심사평가원과 국민건강보험공단에서 공동 발간한 2018년 건강보험통계연보에 따르면 당뇨병은 2017년 기준 우리나라 만성질환 진료현황 2위[3]로 많은 사회적 비용이 발생하고 있다.

당뇨병은 인슐린 수용체의 수준 감소 또는 인슐린의 생리 기능 이상으로 인한 환자의 소변에서 발견되는 포도당 과다 혈당증이고, 라이프 스타일의 변화와 기대 수명의 증가로 인해 당뇨병 환자수가 꾸준히 늘고 있다[4]. 증가된 당뇨병 환자의 삶의 질 향상 연구에 따르면 사회적 프로그램 도입과 문제원인에 맞는 구체적인 방안을 모색하는 보건의료정책 수립의 노력이 필요해 보인다[5]. 또한 모바일 휴먼코칭 헬스케어 서비스를 당뇨병 환자의 혈당계 측정에 활용한 결과, 자기건강관리 능력을 향상시키고 치료의 접근성과 비용적인 측면에서도 일부 효과가 확인되었다[6]. 당뇨병 치료제 연구개발에 있어 글루코키나이즈 활성제는 간 내 포도당 생산 및 인슐린 분비 억제에 대한 효과로 잠재적인 치료제로써 연구되고 있다[7]. 당뇨병 예방 및 관리를 위한 곡물 연구 분야에서는 당뇨를 유발시킨 쥐를 대상으로 상엽추출물을 첨가한 보리면 (보리 99.8% 함유)을 섭취시켜 혈당조절능력의 상승효과를 검증하고 간 기능 개선 및 지질 강하효과도 나타났다[8].

2.2 QSAR

QSAR은 구조가 비슷한 화합물은 활성이 비슷하다는 가정 하에, 화합물이 활성에 중요한 영향을 미치는 상관관계를 찾아서 모델을 만들고 개발된 모델을 이용해서 공통된 패턴을 찾아 화합물의 활성을 미리 예측하는데 사용된다[9]. 당뇨병 치료제로써의 역할을 할 수 있는 다양한 생리활성 화합물질들에 대해 Sim[10]은 QSAR를 통한 연구가 새로운 유도체를 합성하는데 효율적이라고 제안하고 있다. 제도적인 측면에서도 EU의 REACH제도 도입에 따라 각종 화합물질에 대한 독성 및 활성 정보 확보를 위해 화학물질의 분자구조 정보를 기반으로 독성 및 활성을 예측[11]하는 QSAR에 대한 연구가 활발히 진행되고 있다. 선도물질 최적화 관점에서는 QSAR 모델링과 동물의 생리학적인 특성을 바탕으로 약동학적 프로파일을 예측하는 PBPK (physiologically based pharmaco

kinetics) 모델링은 기여[12]할 수 있다고 소개하고 있다. 매년 늘어나는 방대한 신규물질들의 위해성 평가는 상당한 비용과 시간을 필요로 하여, OECD는 회원국을 중심으로 QSAR 사용을 의무화하고 있다[13]. 사례 연구에서 살펴 본 바와 같이 QSAR은 구조와 활성간의 상관관계를 통계학적 방법을 이용해서 연구하는 것을 의미한다.

2.3 Fingerprint

분자를 표현하는 방식 중 molecular fingerprint는 분자 구조 및 특성을 비트 (bit) 문자열로 표현한 것으로 전산적인 효율성과 활성 화합물의 검출 효과, 직관적인 디자인으로 유사성 검색에 활용되고 있다.

고성능 하부구조 검색에서도 구조 검색은 molecular 데이터베이스의 중요한 기능 중 하나이며, fingerprint 기반 검색 방법은 확인 절차의 호출 횟수를 줄여 검색 성능을 향상시키는 데 사용된다[14]. 기계 학습 방법에 의한 HIV-1 프로테아제 억제 분류에서도 분자 구조는 molecular access system (MACCS) fingerprint 와 pubchem fingerprint를 포함한 표현자와 CORINA symphony (managing and profiling molecular datasets)가 계산한 물리화학적 표현자로 특정 지어졌다. 특히 MACCS fingerprint로 개발 된 모델 C3a는 학습 집합과 검증 집합에서 가장 높은 정확도를 보였다[15]. 이렇게 fingerprint는 컴퓨터 보조 약물 설계 기술의 중요한 요소이다. 화합물 데이터베이스를 구조적 연결성 fingerprint를 사용해 검색한다면, 활성 화합물과 비활성 화합물을 빠르게 구별하여 유용 할 확률이 낮은 분자를 신속하게 제거할 수 있어 매우 효과적이다[16]. 예측 모델의 개발을 위해 수행 된 QSAR 제안은 fingerprint 또는 descriptor 행렬을 해당 알고리즘의 입력 데이터로 사용하는 것을 기반으로 한다[17].

2.4 PLS와 PLS-VIP

PLS는 주성분 분석 및 다중회귀를 일반화한 다변량 회귀 모델로써 독립변수들의 선형결합을 통해 잠재변수 (latent variable)를 만들고 잠재변수와 종속변수 간의 상관관계를 최대화하여 잠재변수로부터 종속변수 예측을 목표로 하고 있다[18]. PLS는 제조 공정에서 좋은 성과를 보이고 있는데, 구체적으로 화학공정에서 연속 회분 슬러지 반응기 품질 예측은 공정의 변수가 매우 많아서 그들 간에 심한 상관관계가 큰 경우, PLS는 다른 방법에 비해 좋은 성능을 보이는 것으로 증명되었다[19]. 유기물의 인화점 예측을 위한 PLS와 SVM (support vector machine)의 비교에 따르면 PLS는 다중공선성 문제를 효과적으로 처리하여 예측력[20]이 뛰어나 넓은 분야에서 사용되고 있다. 주성분 회귀 및 PLS 회귀와 같은 다변량 회귀분석은 자연과학을 포함한 다양한 분야에서도 활용되고 있다[21].

PLS-VIP는 PLS 결과를 이용하여 종속변수에 영향을 미치는 주요 독립변수를 찾아내는데 활용된다. 반도체 인화 공정의 코팅두께 예측모델 개발에서 PLS-VIP를 이용한 변수선택은 전체 독립변수의 30%만을 사용해서도 전체변수를 사용했을 때와 유사한 예측 오차를 갖는다는 것을 보여주었다[22]. 불량 분말식품 비파괴검사 기술 개발을 위해서도 PLS-VIP를 이용하여 예측모델 개발에 기여도가 높은 라만 스펙트럼을 선정 후 이 스펙트럼을 이용하여 새로운 예측모델을 개발하였다[23]. 또한 산수유 원산지 판별법 개발을 위한 원산지 판별 마커 대사체 탐색법에서도 PLS-DA (discriminant analysis)와 VIP 분석을 통해 유의성 있는 변수를 추출하였다[24]. 기계 학습을 활용한 학습데이터 구축 표준화 방안에 대한 연구에서는 다중 기계 학습 구조 모델링과 $+a$ 기계 학습 구조 모델링이 다양한 분야의 기계 학습 성능 개선 방법으로 활용되며, 특히 의약분야는 의사결정나무, 공학분야는 SVM이 빈번히 사용되고 있다[25]. 최근 들어 사물인터넷 환경에서 제품 불량 예측을 위해 기계 학습 모델에 대한 연구가 활발히 이루어지고 있다[26].

3. 연구방법

3.1 실험 데이터 소개

본 연구에서는 당뇨병 치료제 합성화합물 개발을 위해 실험용 쥐에 혈당 개선 효과를 측정한 *in vivo* 실험 결과 값과 화합물의 구조 및 구조에 따른 다양한 물리화학적 특성을 가지는 화합물 데이터를 확보하였으며[27-30], 총 세 가지 실험을 진행했다. 첫 번째 실험에서 독립변수는 *in vitro* 실험 값과 계산 값 (데이터집합)이며 자세한 설명은 Table 1과 같다. 두 번째와 세 번째 실험에서는 화합물 데이터 구조를 SMILES (simplified molecular input line entry system)[31]를 이용하여 변환하고 독립변수로 설정했다. 구체적으로 두 번째 실험에서는

SMILES 형식을 881비트로 이진화한 pubchem fingerprint를 독립변수로 사용했으며, 세 번째 실험에서는 166비트로 이진화한 MACCS fingerprint를 독립변수로 사용했다. 세 실험 모두에서 종속변수 (Y)는 oral glucose tolerance test (OGTT)로 실험용 쥐에 화합물을 경구 투여하여 혈당 강하 활성평가를 실험한 *in vivo* 값으로 설정했다. 본 연구에서 fingerprint 형식으로 변환하기 위해서 과학적 의사결정 지원을 위한 분석 솔루션인 Dotmatics Vortex[32] 프로그램을 사용했다.

Table 1. Description of variables

Variable	Description
Y	OGTT_inhibition(*in vivo* assay)
X1	EC50_UM
X2	-LOG(EC 50)
X3	EMAX
X4	Glocose uptake
X5	Caco2
X6	MDCK
X7	Molecular_weigh
X8	No_total_atoms
X9	No_total_bonds
X10	No_rotatable_bond
X11	No_rings
X12	No_aromatic_ring
X13	No_H_bond_acceptions
X14	No_H_bond_donors
X15	Topological_PSA
X16	2D_VDW_surface
X17	2D_VDW_volume
X18	Polarizability_miler
X19	SKlogP_value
X20	Pure_water_solubility_mg_L
X21	AlogP98_value
X22	C ratio
X23	N ratio
X24	NO ratio
X25	Hetero ratio
X26	Halogen ratio
X27	Number of rings(size 3)
X28	Number of rings(size 4)
X29	Number of rings(size 5)
X30	Number of rings(size 6)
X31	Surface tension
X32	Density
X33	Polarizability
X34	LogD(ph = 7.40)
X35	Pe(jejunum), 10^-4cm/s

3.2 연구 방법

실험데이터 별로 변수 간 다중공선성이 존재하기 때문에 본 연구에서는 PLS를 이용하여 종속변수인 OGTT 실험 값을 예측한다. 독립변수는 실험 별로 다르게 설정했다. 구체적으로 첫 번째 실험에서는 *in vitro* 데이터집합, 두 번째 실험에는 881 비트로 이진화한 pubchem fingerprint, 세 번째 실험에서는 166 비트로 이진화한 MACCS fingerprint를 사용하였다. 각 실험 별로 PLS 모델을 학습하고 검증 데이터를 이용하여 모델의 적합성인 결정계수 (R^2)를 측정했다. 또한 학습된 PLS 모델로부터 독립변수의 중요도를 측정하기 위해 PLS-VIP를 실험 별로 실행했다. 모든 실험에서 모델의 학습과 검증은 R 프로그램으로 진행하였다.

PLS는 독립변수들이 선형결합을 통해 잠재변수를 만들면서 차원을 축소시키고 만들어진 잠재변수와 종속변수 간의 상관관계를 최대화 하여 종속변수를 추정한다. 차원이 축소된 잠재변수와 종속변수의 관계식은 (2), (3)과 같다[9].

$$X = TP^T + E = \sum_{a=1}^{l} t_a P_a^t + error \quad (2)$$

$$Y = UQ^T + F = \sum_{a=1}^{l} u_a q_a^T + error \quad (3)$$

식 (2)에서 P는 독립변수 X 의 선형결합으로 구성된 잠재변수의 계수 행렬을 의미하며, 로딩 (loading) 벡터라고 부른다. T는 스코어 (score) 벡터의 행렬을 의미하며, 각 스코어 벡터는 잠재변수로 구성된 공간에서의 데이터를 의미한다. 즉 독립변수 공간에서의 데이터와 잠재변수 공간에서의 데이터는 식 (2)와 같이 선형결합 관계로 규정된다. 식 (2)에서 E는 오차 행렬로 l개의 잠재변수 P_a^t $(a=1,...,l)$로 독립변수 데이터를 설명하지 못하는 잔차를 표현한다. 잠재변수 개수가 많아질수록 잔차량 (error)은 적어진다. 한편 식(3)은 종속변수 Y를 잠재변수 공간에서 표현한 식이다. 이 식에서 Q는 종속변수의 차원을 축소한 잠재변수를 표현한 행렬이고 U는 축소된 잠재변수 공간에서의 데이터인 스코어 행렬을 의미하며 F는 잔차를 의미한다.

PLS에서는 두 잠재변수 공간에서의 데이터인 스코어 T와 U의 상관관계를 식 (4)과 같이 선형식으로 표현한다[9].

$$U = TB^t + R \quad (4)$$

식 (4)에서 B는 T와 U의 회귀계수를 의미하며, 잔차를 최소로 만드는 최소자승법을 이용하여 도출한다. R은 선형회귀 식으로 설명 못하는 잔차를 의미한다. PLS 모델을 만든 후 다음으로 잠재변수 개수 l를 결정해야 한다. 잠재변수 수가 많아지면 과적합 (overfitting)이 발생할 가능성이 높다. 본 연구에서는 통계모델 검증에서 가장 널리 사용되는 교차 검증 (cross validation)을 사용하는데, 이 방법은 학습 데이터를 여러 개의 그룹들로 나눈 후에 그 그룹들 중에서 하나를 제외시킨 데이터 세트로 모델을 학습시키고 나머지 한 개의 그룹으로 Y 변수의 실제 값과 예측 값의 차이를 검증한다. 교차 검증 그룹 수를 k로 설정하면 총 k번의 모델 학습과 검증이 가능하다. 모든 실험에서 얻어진 검증 오차를 합산하여 그 모델의 예측력을 측정할 수 있다. 본 연구에서는 10차 교차 검증 (k=10)을 통해 PLS 모델의 예측력을 측정하였으며, 90% 이상의 설명력을 갖는 3개의 잠재변수를 찾아냈다.

PLS 알고리즘을 세 가지 실험에 적용한 결과, *in vivo* 실험데이터 (종속변수)와 MACCS fingerprint 데이터 (독립변수)의 결정계수 R^2가 가장 높게 나왔으며, 이를 근거로 *in vivo* 실험데이터에 영향을 미치는 주요 분자표현자를 찾는 과정을 PLS-VIP로 실행했다. PLS-VIP는 회귀 결과를 바탕으로 독립변수가 종속변수에 미치는 영향을 산정하는 방법이다. PLS-VIP는 식 (5)와 같이 정의되며 유의미한 독립변수를 판별하는데 있어 일반적으로 VIP 값이 1.0 이상일 경우를 기준으로 한다[9].

$$VIP_j = \sqrt{\frac{p\sum_{a=1}^{l} R^2(y,t_a)(w_{aj}/\|w_a\|)^2}{\sum_{a=1}^{l} R^2(y,t_a)}} \quad (5)$$

식 (5)에서 p는 독립변수의 개수, l은 잠재변수의 개수를 의미하며, w_{aj}는 PLS 회귀에서 a번째 잠재변수 내 j번째 독립변수의 기여도인 로딩 값을 의미하며, $\|w_a\|$는 잠재변수 a의 총 로딩을 의미한다. 또한 $R^2(y,t_a)$는 잠재변수 a가 종속변수 y에 대한 설명력을 의미한다. 따라서 VIP_j는 j번째 독립변수가 모든 잠재변수에 기여한 정도를 잠재변수의 설명력을 가중하여 산정한 값이다.

4. 연구 결과

세 가지 실험데이터들을 학습 집합과 검증 집합으로 구분하여 PLS를 통해 산출한 결정계수 R^2는 Table 2와 같다.

Table 2. R^2 result for the three experiments

Independent variable		Average R^2	
		Training data	Test data
in vitro dataset		0.612	0.537
Fingerprint	Pubchem	0.41	0.248
	MACCS	0.788	0.569

PLS 검증 결과, MACCS fingerprint 데이터를 이용한 실험의 통계적 설명력 (R^2 = 0.569)이 기존 *in vitro* 데이터집합을 이용한 실험의 설명력 (R^2 = 0.537) 보다 높았다. 그러나 pubchem fingerprint 데이터를 이용한 실험에서는 검증 설명력이 *in vitro* 데이터집합 보다 떨어짐을 알 수 있었다. 이는 pubchem fingerprint 데이터가 *in vivo* 실험결과를 설명하는데 부족하다는 증거다. 이 결과를 바탕으로 본 실험에서는 독립변수인 *in vitro* 데이터 집합 중에서 종속변수인 *in vivo* 실험결과를 설명하는데 중요한 독립변수를 PLS-VIP를 이용해 찾아냈다. 그 결과 VIP score가 1.0을 넘는 변수는 총 6개가 존재했으며, 해당 변수에 대한 기술은 Table 3과 같다.

Table 3. Major *in vitro* variables

Variable		Description
X3	EMAX	Degree of protein activity
X5	Caco2	Caco2 cells
X7	Molecular weight	Molecular weight
X15	Topological_PSA	Topological polarity surface acrea
X16	2D_VDW_surface	2D van der waals surface area
X17	2D_VDW_volume	2D van der waals compound volume

한편, PLS-VIP를 MACCS fingerprint 데이터에 적용한 결과, VIP score가 1.0을 상회하는 주요 분자표현자는 총 44개가 도출되어서 기존 *in vitro* 데이터집합 보다 훨씬 많은 인자를 찾아 낼 수 있었다. 그 중 VIP score가 2.0을 넘는 9개의 주요 분자표현자는 Table 4에 기술되어 있다.

Table 4. Major MACCS variable

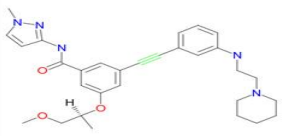
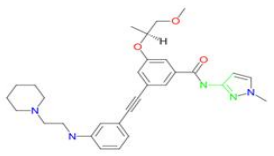
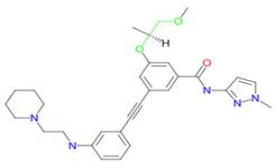
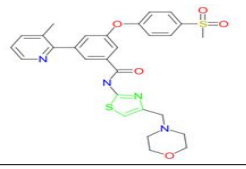
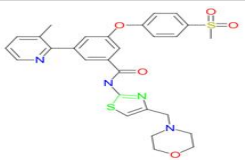
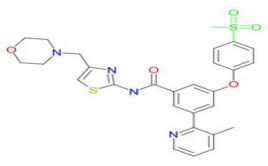
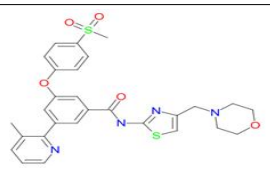
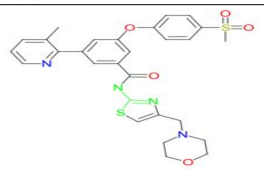
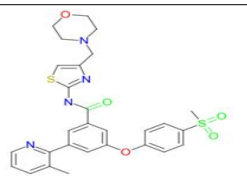
Molecular descriptor		Description (Structure)
FP17	CTC	
FP38	NC(C)N	
FP72	OAAO	
FP36	S Heterocycle	
FP47	SAN	
FP81	SA(A)A	
FP88	S	
FP106	QA(Q)Q	
FP136	O=A>1	

Table 4에서 주요 분자표현자를 살펴보면 FP36 (S (황)을 포함한 링 형 구조)과 FP47 (황과 질소 사이에 하나의 원자가 존재하는 형태), FP81 (황 주위에 임의의 원자 3개가 연결된 형태)과 FP88 (황 원자가 존재하는 형태)은 동일한 구조 형태로 나타나는 것을 볼 수 있고, 그 구조 중 황을 포함한 사이클 구조는 FP47의 형태를 포함하는 것으로 S 주위에 3개 이상의 원소를 가지고 있음을 보여준다.

통계적으로 MACCS fingerprint 주요 분자표현자 상위 9개의 변수와 종속변수의 상관관계는 0.7 이상으로, 이는 각 변수들이 종속변수에 의미 있는 영향을 준다고 볼 수 있다. 또한 종속변수와의 경향성을 보면 FP17 (탄소와 탄소간의 3중 결합), FP38 (탄소 주변에 질소 2개와 탄소 1개), FP72 (산소 2개 사이에 2개 원자 존재)는 비례하는 형태이고 FP36, FP47, FP81, FP88, FP106, FP136은 반비례하는 형태를 보였다. 위와 같은 경향을 이용하여 당뇨병 치료제 후보물질을 도출할 때 FP17, FP38, FP72 구조를 우선 선택하고 FP36, FP47, FP81, FP88, FP106, FP136은 포함하지 않는 구조를 찾게 되면 좀 더 빠르게 분자표현자를 필터링 할 수 있다.

결과적으로 기존 QSAR 연구에서는 SVM과 PLS 등을 이용하여 화합물의 특성을 예측하는 연구는 수행되었으나, MACCS와 pubchem 핑거프린트 데이터와 PLS-VIP를 이용하여 주요 분자표현자를 도출하는 연구는 수행되지 않았다. 따라서 제안 방법에 의해 분자표현자들을 도출하게 되면 후보약물을 찾는 과정에서 긍정적인 효과를 가져 올 수 있을 것으로 기대한다.

5. 결론

본 연구는 항당뇨 화합물 구조와 활성간의 관계를 PLS 기법을 이용해서 도출하였고, 변수 간 다중공선성이 존재할 때 변수선택 예측성능이 우수한 PLS-VIP 알고리즘을 통해 항당뇨에 영향을 미치는 주요 인자를 확인해 보았다. 연구 결과, 화합물 구조를 표현하는 MACCS fingerprint 방식의 검증데이터가 통계적으로 설명력(R^2 = 0.569)이 기존 실험데이터 설명력(R^2 = 0.537)보다 높았다. 이러한 실험 결과를 근거로 주요변수를 확인 해 본 결과, 항당뇨 결과에 비례적인 경향을 보이는 FP17, FP38, FP72와 반비례적 경향을 보이는 FP36, FP47,

FP81, FP86, FP106, FP136가 주요 분자표현자로 도출하였다. 도출된 분자표현자들을 기반으로 당뇨병 치료제 후보약물을 개발한다면 동물에서의 항당뇨 효과를 사전에 예측할 수 있어 신약 개발 비용과 시간을 단축할 수 있을 것으로 예상한다. 물론 최종 약으로 출시되기까지는 물리화학적 검증 및 임상시험을 거쳐야 하지만 제안 방법을 올바르게 이해하고 활용할 수 있다면 국내 제약 업계가 글로벌 제약 경쟁력을 갖출 수 있는 초석이 될 수 있을 것으로 기대한다.

REFERENCES

[1] Ministry of Health & Welfare. (2018). *Conducted R&D consulting support project for innovative new drugs*. Ministry of Health & Welfare. http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=345392&page=1

[2] Ministry of Health & Welfare. (1998). *A share of robots in new drug development*. Ministry of Health & Welfare. http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=345392&page=1

[3] Ministry of Health & Welfare. (2018). *Annual health insurance statistical report 2017*. Ministry of Health & Welfare. http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=346196&page=1

[4] Y. J. Jung. (2008). *Current trends in diabetes medical treatment*. Master thesis. Mokpo University, Mokpo.

[5] J. H. Lim & C. H. Oh. (2013). Medical care utilization status and quality of life in diabetes mellitus patients. *The journal of Digital Policy & Management. 11(10)*, 609-618.

[6] M. J. Lee & H. K. Kang. (2017). Effects of mobile based-healthcare service using human coaching to the self-care of diabetes. *Convergence Society for SMB. 7(4)*, 83-89.

[7] W. E. Chong. (2016). *Preclinical evaluation of a new glucokinase activator, YH10561 as a therapeutic drug candidate for type II diabetes mellitus*. Doctoral dissertation. Seoul National University. Seoul.

[8] C. M. Park & H. S. Yoon. (2018). Effects of barley noodles contained mulberry leave extracts on blood glucose regulation in diabetic mice. *Journal of the Korea Convergence Society, 9(8),* 101-108.
DOI: 10.1520/JKCS.2018.9.8.101

[9] H. G. Shin. (2017). *New drug development QSAR model using computer aided*. http://www.ibric.org/myboard/skin/news1/print.php?Board=news&id=279383

[10] M. J. Sim, (2007), *QSAR study of biologically active compounds with biologically activities for diabetes medicine,* Master thesis. Yonsei University, Seoul.

[11] D. W. Kim, S. C. Lee, M. J. Kim, E. J. Lee & C. K. Yoo. (2016). Development of QSAR model based on the key molecular descriptors selection and computational toxicology for prediction of toxicity of PCBs. *Korean Chem.Eng.Res, 54(5),* 621-629.

[12] J. J. Hyeon, M. H. Park, S. H. Shin & Y. G. Shin, (2015). Novel lead optimization strategy using quantitative structure-activity relationship and physiologically-based pharmacokinetics modeling. *The pharmaceutical society of korea, 59(4),* 151-157.

[13] G. H. Kim, K. G, Lyu, Y. J. Kim & H. C. Kim. (2008). A survey on quantitative structure-activity relationship(QSAR) models. *Korean Institute of Information Scientists and Engineers, 35(1A)*, 43-44.

[14] M. Kratochvíl, J. Vondrášek & J. Galgonek. (2018). Sachem: a chemical cartridge for high performance substructure search. *Journal of Cheminformatics, 10(1).* 1-11.
DOI: 10.1186/s13321-018-0282-y

[15] Y. Li, Y. Tian, Q. Qin & A. Yan. (2018). Classification of HIV 1 protease inhibitors by machine learning methods. *ACS OMEGA, 3(11),* 15837-15849.
DOI 10.1021/acsomega.8b01843

[16] K. Rataj, W. Czarnecki, S. Podlewska, A. Pocha & A. Bojarski, (2018). Substructural connectivity fingerptint and extreme entropy machines－A new method of compound representation and analysis. *Molecules, 23(6).*
DOI 10.3390/molecules23061242

[17] L. Ruiz, M. Neito, (2018). A new data representation based on relative measurements and fingerprint patterns for the development of QSAR regression models. *Chemometrics and Intelligent Laboratory Systems, 176*, 53-65.
DOI: 10.1016/j.chemolab.2018.03.007

[18] H. Abdi. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics, 2(1),* 97-106.
DOI: 10.1002/wics.51

[19] C. K. Yoo, (2002). *Statistical method for quality prediction of continuous batch sludge reactor (SBR),*

Chemical Engineering and Materials Research Information Center, www.cheric.org

[20] C. J. Lee, J. W. Ko & G. B. Lee. (2010). Comparison of partial least squares and support vector machine for the flash point prediction of organic compounds. *Korean Chem.Eng.Res, 48(6),* 717-724.

[21] B-H. Mevik, R. Wehrens, (2007). The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software, 18(2).* 1-24.

[22] H. S. Lee, Y. R. Lee, C. H. Jun & J. H. Hong, (2010). A prediction model for coating thickness based on PLS model and variable selection. *The Korean Statistical Society, 23(2),* 295-304.

[23] S. D. Lee, S. T. Lohuni, B. K. Cho, M. S. Kim & S. H. Lee, (2014). Development of nondestructive detection method for adulterated powder products using raman spectroscopy and partial least squares regression. *Journal of the Korean Society for Nondestructive Testing, 34(4),* 283-289.
DOI: 10.7779/JKSNT.2014.34.4.283

[24] J. Y. Leem, (2016). Discrimination model of cultivation area of corni fructus using a GC-MS based metabolomics approach. *Analytical Science & Technology, 29(1),* 1-9.
DOI: 10.5806/AST.2016.29.1.1

[25] H. H. Lee, S. H. Chung, E. J & E. J. Choi (2016). A case study on machine learning applications and performance improvement in learning algorithm. *The Society of Digital Policy & management, 14(2),* 245-258.

[26] J. H. Ku, (2017). A study on the machine learning model for product faulty prediction in internet of things environment, *Convergence Society for SMB, 7(1),* 55-60.

[27] K. J. Park. (2012). Identification of YH-GKA, a novel benzamide glucokinase activator as therapeutic candidate for type 2 diabetes mellitus. *Archives of Pharmacal Research, 35(12),* 2029-2033.
DOI: 10.1007/s12272-012-1201-9

[28] K. J. Park et al. (2013). Discovery of a novel phenylethyl benzamide glucokinase activator for the treatment of type 2 diabetes mellitus. *Bioorganic & Medicinal Chemistry Letters, 23(2),* 537 - 542.
DOI: 10.1016/j.bmcl.2012.11.018

[29] K. J. Park et al. (2014). Discovery of 3-(4-methanesulfonylphenoxy)-N-[1-(2-methoxy-ethoxymethyl)-1H-pyrazol-3-yl]-5-(3-methylpyridin-2-yl)-benzamide as a novel glucokinase activator (GKA) for the treatment of type 2 diabetes mellitus. *Bioorganic & Medicinal Chemistry, 22(7),* 2280-2293.
DOI: 10.1016/j.bmc.2014.02.009

[30] K. J. Park, B.M. Lee, K. H. Hyun, T. Han, D.H. Lee & H. H. Choi. (2015). Design and synthesis of acetylenyl benzamide derivatives as novel glucokinase activators for the treatment of T2DM. *ACS Medicinal Chemistry Letter, 6(3),* 296-301.
DOI: 10.1021/ml5004712

[31] WIKIPEDIA. (2018). *Simplified molecular-input line-entry system*, WIKIPEDIA, https://www.wikipedia.org.

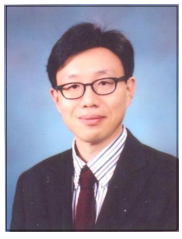
[32] Dotmatics. (2018). *Intuitive and versatile scientific data visualization and analysis,* Dotmatics, https://dotmaticscom/products/vortex

남 궁 윤(Youn Namgoong) [정회원]

- 2014년 2월 : 연세대학교 산업정보경영학과 (공학석사)
- 2018년 6월 : 연세대학교 융합기술경영학과 (박사수료)
- 현재 : ㈜유한양행 IT팀 부장
- 관심분야 : GxP, 제약데이터사이언스

- E-Mail : ngy@yuhan.co.kr

김 창 욱(Chang Ouk Kim) [정회원]



- 1990년 2월 : 고려대학교 산업공학과 (공학석사)
- 1996년 5월 : Purdue University (공학박사)
- 현재 : 연세대학교 공과 대학 산업공학과 교수

- 관심분야 : 인공지능, 제조데이터사이언스
- E-Mail : kimco@yonsei.ac.kr

이 창 준(Chang Joon Lee) [정회원]

- 2002년 8월 : 숭실대학교 화학과 (이학석사)
- 2009년 2월 : 연세대학교 생명공학. (공학박사)
- 현재 : ㈜Dotmatics
- 관심분야 : 딥러닝, 라이프사이언스

- E-Mail : chang-joon.lee@dotmatics.com