*Journal of the Korean Chemical Society*
**2009**, *Vol. 53, No. 6*
Printed in the Republic of Korea

# SVM 방법을 이용한 hERG 이온 채널 저해제 예측모델 개발

강신문 · 김한조 · 오원석 · 김선영 · 노경태† · 남기엽*
(사)분자설계연구소 신약개발실
†연세대학교 생명공학과
(접수 2009. 9. 29; 수정 2009. 10. 29; 게재확정 2009. 11. 3)

# Development of Classification Model for hERG Ion Channel Inhibitors Using SVM Method

**Sinmoon Gang, Hanjo Kim, Wonseok Oh, Sunyoung Kim, Kyoung Tai No†, and Ky-Youb Nam***
*Drug Discovery Division, Bioinformatics and Molecular Design Research Center, 134 Sinchon-dong, Seodaemun-gu, Seoul, Korea*
*†Department of Biotechnology, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul, Korea*
(Received September 29, 2009; Revised October 29, 2009; Accepted November 3, 2009)

**요약.** 흡수, 분포, 대사, 배설 특성 및 독성을 예측하기 위한 효과적인 툴을 개발하는 것은 신약개발의 초기단계에서 NCE(new chemical entity)에 대한 가장 중요한 업무 중의 하나이다. 최근에 이런 시도중의 하나로서 ADME/T(absorption, distribution, metabolism, excretion, toxicity)관련 성질들의 예측에 support vector machine(SVM)을 이용하고 있다. 그리고 SVM은 ADME/T 성질들을 정확하게 예측하는데 많이 사용 되고 있다. 그러나 SVM 모델링에 두 가지 문제가 있다. 특성 선택(feature selection) 과 매개변수 설정(parameter setting)은 여전히 해결해야 할 과제이다. 이 두 가지 문제들은 SVM 분류의 효율성과 정확도에 결정적인 영향을 끼친다. 특히 특성 선택과 최적화된 SVM 변수의 설정은 서로 영향을 주기 때문에 동시에 다루어져야 한다. 여기서 우리는 genetic algorithm(GA) – 특성 선택에 사용 – 과 grid search(GS) method– 변수최적화에 사용 – 두 가지를 통합하는 효과적인 해결책을 제시하였다. ADME/T관련 성질 중 하나인 심장부정맥을 야기시키는 hERG 이온채널 저해제 분류 모델이 여기서 제안된 GA-GS-SVM을 위해 할당되고 테스트 되었다. 1891개의 화합물을 가지는 트레이닝 셋으로 단일 모델 3개, 앙상블 모델 3개, 총 6개의 모델을 만들었고 175개의 외부 데이터를 테스트 셋으로 사용하여 검증하였다. 데이터의 불균형 문제를 해결하기 위하여 GA-GS-SVM 단일 모델에 의한 예측 정확도와 GA-GS-SVM 앙상블 모델 예측 정확도를 비교하였으며, 앙상블모델을 사용하여 예측의 정확도를 높일 수 있었다.

**주제어:** ADME/T, hERG저해제, SVM, 유전자알고리듬, 분류모델

**ABSTRACT.** Developing effective tools for predicting absorption, distribution, metabolism, excretion properties and toxicity (ADME/T) of new chemical entities in the early stage of drug design is one of the most important tasks in drug discovery and development today. As one of these attempts, support vector machines (SVM) has recently been exploited for the prediction of ADME/T related properties. However, two problems in SVM modeling, i.e. feature selection and parameters setting, are still far from solved. The two problems have been shown to be crucial to the efficiency and accuracy of SVM classification. In particular, the feature selection and optimal SVM parameters setting influence each other, which indicates that they should be dealt with simultaneously. In this account, we present an integrated practical solution, in which genetic-based algorithm (GA) is used for feature selection and grid search (GS) method for parameters optimization. hERG

ion-channel inhibitor classification models of ADME/T related properties has been built for assessing and testing the proposed GA-GS-SVM. We generated 6 different models that are 3 different single models and 3 different ensemble models using training set - 1891 compounds and validated with external test set - 175 compounds. We compared single model with ensemble model to solve data imbalance problems. It was able to improve accuracy of prediction to use ensemble model.

**Keywords:** ADME/T, hERG inhibitor, SVM, Genetic algorithm, Classification model

## 서론

신약개발에서 높은 활성을 가진 대부분의 화합물들이 좋지 않은 흡수, 분배, 대사, 배설 특성들과 독성(ADME/T)때문에 임상실험에서 통과되지 못하였다. 심지어 의약시장에서 판매되다가 sertindole, grepafloxacin, terfenadine등과 같은 약물들은 QT연장증후군[심전도(Electrocardiogram)에서 심장의 Q파와 T파 사이의 간격이 늘어지는 현상]을 유발하고 심장부정맥을 야기시켜 시장에서 철수 하였다.[1-3] 그래서 좋지 않은 ADME/T 특성들을 가지는 이런 화합물들을 화학합성이나 임상실험 전에 제거하기 위해 초기 디자인 단계에서 화합물들의 ADME/T 특성들을 예측하는 방법이 많이 사용되고 있다.

선도 화합물의 hERG 저해에 의한 독성평가는 신약개발과정에서 가장 중요한 ADME/T 성질들 중 하나이다. hERG는 심근세포에서 칼륨이온의 출입을 담당하는 이온채널로 심장과 신경조직에서 나타나는 활동전압의 변화를 조절하는 역할을 한다. hERG 이온채널은 심장의 다른 이온채널들을 타깃으로하는 약물들에게 중요하다. hERG 저해에 대한 실험적 측정은 많은 시간과 비용이 소요된다. 따라서 고속탐색법(high throughput screening, HTS)에 적합하지 않으며, 비용이 적게 들고 보다 빠른 대안적인 방법은 컴퓨터 예측 모델이다. 지금까지 hERG 저해예측에 대하여 Homology모델링에 의한 단백질 구조를 이용한 방법, Pharmacophore method 또는 QSAR(Quantitative structure-activity relationship)을 이용한 방법들이 사용되었으며,[4-7] 이러한 시도중의 하나로서 통계학적 기계학습방법인support vector machine(SVM)은 최근에 ADME/T 관련 특성들을 예측하기 위해 사용되고 있다. 몇몇의 지난 연구에서 보여주듯이 SVM은 ADME/T 특성들을 정확하게 예측하는 가장 좋은 방법들 중 하나이다.[8] 그러나 SVM 모델링에서 특성 선택 과 매개변수 설정, 두 가지 문제는 여전히 해결해야 할 과제이다. 이 두 가지는 SVM 분류의 효율성과 정확도에 결정적인 영향을 미친다. 지금까지 genetic algorithm (GA),[9] recursive feature eliminations (RFE),[10] simulated annealing (SA) 접근[11] 등을 포함하는 몇 가지 방법들이 특성 selection을 위해 제안되었다. 변수최적화(Parameters optimization)을 위해서는 전통적인 Grid-search(GS) 알고리듬이 사용 되었다.[12]

우리는 이 연구를 통하여, 심장부정맥을 야기시키는 hERG 이온 채널 저해제 분류 모델에 대하여 GA-GS-SVM 알고리듬을 사용하여 데이터 불균형에 의한 예측의 부정확성을 보여주고, 이를 해결하기 위하여 도입한 GA-GS-SVM 앙상블모델에 의한 결과를 비교하여 정확도를 향상시켰으며 175개의 hERG 외부 테스트 셋을 통하여 이를 검증하였다.

## 이론 (방법)

### Support Vector Machine (SVM)

SVM 이론의 자세한 기술은 문헌에서 찾을 수 있다.[13] 여기서는 SVM의 기본 아이디어를 짧게 요약할 것이다.

SVM에서 각각의 object는 N차원 공간에서 한 점에 대응하는 N 정수의 벡터 $x_i$로 기술되어 진다. *Fig.* 1은 간단한 하이퍼평면에 의해 두 집단을 분류하는 것에 대한 설명이다. 활성그룹에 있는 object들은 각각 $y_i = +1$로 할당이 되고, 비활성그

*Corresponding Author: 남기엽
(사)분자설계연구소 신약개발실, 서울시 서대문구 신촌동 134 연세대학교 연세공학원 B138A
Tel: +82-2-393-9550, Fax: +82-2-393-9554
E-mail: kyn@bmdrc.org

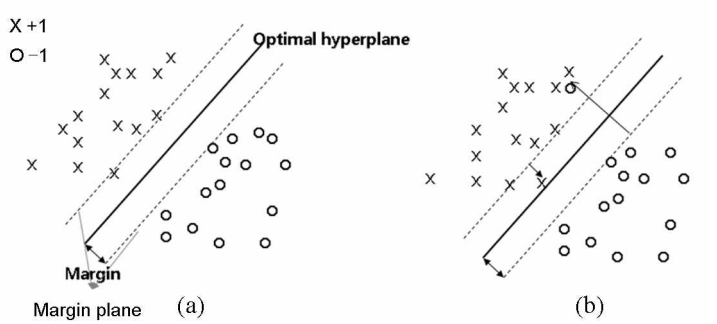


*Fig.* 1. SVM classifies two groups that were maximized margin between the O and X by hyper-plane. (a) an example was fully classified by a optimal linear hyperplane. (b) an example was not linear and completely broken.

룹에 있는 object들은 $y_i = -1$로 할당된다. 선형에서(*Fig.* 1(a)), 이 object들은 다음과 같이 분류된다.

$$w \cdot x_i + b \geq +1 \; for \; y_i = +1 \; (class\ 1) \tag{1}$$

$$w \cdot x_i + b \geq -1 \; for \; y_i = -1 \; (class\ 2) \tag{2}$$

여기서 $w$는 하이퍼평면에 대한 노말 벡터이고, $b$는 스칼라 값이다.

SVM은 다음의 최적화문제의 해를 구함으로서 최대마진을 가지는 최적의 분리 가능한 하이퍼평면을 찾는다.

$$\underset{w,\,b}{Max} \; \frac{2}{\|w\|} \; subject\ to \; y_i(w \cdot x_i + b) - 1 \geq 0 \tag{3}$$

여기서 $\frac{2}{\|w\|}$는 마진(margin)이다.

위 식은 라그랑쥐 승수 방법으로 해를 구할 수 있다. 결국 분류결정함수가 다음과 같이 얻어진다.

$$f(x) = sgn(w \cdot x_i + b) = sgn\left[\sum_{j=1}^{n} \alpha_i y_i (x \cdot x_i) + b\right] \tag{4}$$

이 식은 하이퍼평면이 점들을 두 개의 그룹으로 정확히 분리할 수 없는 선형 비분리 경우로 확장되어질 수 있다. (*Fig.* 1(b))에서의 경우 우리는 음이 아닌 변수 $\xi \geq 0$, $i = 1, 2, 3, \ldots, m$를 소개할 수 있다.

$$w \cdot x_i + b \geq +1 - \xi_i \; for \; y_i = +1 \tag{5}$$

$$w \cdot x_i + b \geq -1 + \xi_i \; for \; y_i = -1 \tag{6}$$

이것의 목적은 트레이닝 에러의 수를 최소화하는 하이퍼평면을 찾는 것으로 constraint violation을 가능한 적게 유지시키는 것이다. 방정식은 다음과 같이 된다.

$$\underset{w,\,b}{Max} \; \frac{2}{\|w\|} + C\sum_{i=1}^{m} \xi_i \; subject\ to \; y_i(w \cdot x_i + b) - 1 + \xi_i \geq 0 \tag{7}$$

## Cross validation

교차 검증 이론은 Seymour Geisser에 의해 고안되어 졌다.[14] 교차 검증은 샘플들이 위험하거나 비용이 많이 들고 수집하기 힘든 데이터 일 때 특히 유용하다. 교차검증을 위해 일반적으로 많이 사용되는 것은 K겹 교차이며, K겹 교차 검증에서, 원본 샘플은 k 개의 서브샘플로 나누어진다. 이 k 개의 서브 샘플들 중에서 하나의 서브샘플이 모델을 테스트 하기 위한 검증 데이터로서 남는다. 그리고 나머지 k-1 개의 서브샘플들은 트레이닝 데이터로 사용된다. 교차 검증은 k개의 샘플들이 정확히 한번씩 검증 데이터로 사용될 때까지 k번 반복되고, 여기서 나온 결과들을 평균하여 단일 평가 값을 산출한다. 이 방법의 장점은 모든 서브샘플들이 트레이닝과 검증을 위해 한번씩 사용이 되고 각각의 서브샘플들은 정확히 한번씩 검증 데이터로 이용이 된다는 것이다. *Fig.* 2와 같이 데이터 서브샘플을 4등분하고, 3등분을 트레이닝에 사용하고, 남은 1등분을 테스트 서브샘플로 사용하는, 5겹 교차검증은 속도와 정확도 면에서 아주 좋은 결과를 보여 주기 때문에 우리는 5겹 교차검증을 사용하여 각각의 예측 모델을 선정하였다.[12,15]

**Parameters optimization using the grid search (GS) method**

그리드 검색 방법은 정해진 범위 내에서 적당한 간격의 이산 값을 시도하여 최적의 매개변수를 찾는 방법이다. Radial Basis Function(RBF) 커널을 이용하여 SVM을 수행하기 위해서는 두 개의 매개변수 ($C$, $\gamma$)가 필요하다. C는 SVM의 penalty 매개변수이며, $\gamma$는 커널 매개변수이다. 그리드 검색에서는 기본적으로 ($C$, $\gamma$)의 짝을 교차 검증에 사용하며, 교차 검증의 정확도가 가장 큰 값이 선택 되어진다. 지수적으로 순차 증가하는 ($C$, $\gamma$) 값이 알맞은 매개변수를 찾는 좋은 방법이었다. (예를 들어 $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}, \gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$). 복소공배법이나 몬테카를로 방법 등 컴퓨터의 비용을 줄일 수 있는 보다 효율적인 방법들이 있지만 다음의 두 가지 면에서 그리드 검색이 선택되어 졌다.

첫 번째로 자기 발견적 방법이나 근사값을 이용한 매개변수 검색방법은 정확도를 확신할 수 없다. 두 번째로 매개변수의 수가 2개일 경우는 컴퓨터 계산능력의 발달로 인해 다른 방법에 비해 크게 느리지 않다. 또한 독립적인 ($C$, $\gamma$)값을 사용함으로써 병렬계산 코드를 쉽게 작성 할 수 있다.

**Genetic algorithm (GA) for the feature selection**

GA는 가장 인기 있는 최적화 알고리듬으로 생물학적 체계에서 자연선택과 유전자의 다윈 진화론에 대한 직접적 유추에 기반을 두고 있다.[16] GA는 데이터 마이닝과 최적화 같은 다양한 문제들에 성공적으로 적용 가능하다. 최근에는 SVM 모델링에서 특성 selection을 위해 사용되어지고 있다.

유전자 알고리듬은 일반적으로 염색체라고 부르는 유전자형을 사용하여 표현되어 지는 개체들의 집단을 써서 동작한다. 각 개체의 염색체는 새로운 세대를 생성하기 위해 돌연변이와 교차변이와 같은 연산들에 의해 조정된다. 개체의 질을 평가하기 위해 fitness 함수를 사용한다. 더 좋은 질을 가지는 개체들이 다음 세대에서 살아남거나 재생된다. 적합한 암호화 체계가 각각의 개체의 염색체를 암호화 하는데 필요하다. 보통 암호화 체계로 이진 문자열을 사용한다.

여기서는 특성(0: 선택 안됨, 1: 선택)를 표현하는 각 비트를 가지는 이진 문자열을 염색체를 표현하는데 사용한다. GA는 무작위로 생성된 이진 문자열의 개체군으로 적용된다. 각 문자열의 정확도는 다음과 같이 결정된다.

$$fitness = W_A \times \text{SVM\_accuracy} + W_F \times N_F \quad (8)$$

여기서 $W_A$는 SVM 분류 정확도 가중치, $N_F$는 선택된 특성의 수, $W_A$는 특성 수의 가중치이다. 5겹 교차타당성이 SVM_accuracy에 사용되어졌다. $W_A$와 $W_F$는 상대적인 중요도에 기반하여 조정되어질 수 있다.

룰렛 휠 선택 알고리즘이 자식을 생성하기 위한 교차변이를 위해 염색체를 선택하는데 사용되어 졌다. 교차위치는 무작위로 만들어지고 교차변이 비율은 적당히 조정되어 질 수 있다. 돌연변이는 허용되고 그 비율 역시 적당히 조정되어 질 수 있다.

**System architectures and implementation details**

GA-GS-SVM은 특성 선택과 parameter 최적화를 동시에 수행한다. 시스템 구조는 *Fig.* 2에서 보여주고 있다.

(1) 데이터셋 수집: 알려진 hERG 이온채널 저해제 특성을 가지는 화합물들에 대한 구조-활성 정보 수집

(2) 표현자계산과 클래스 라벨 할당: PreADMET 2.0[18]을 이용하여 1차원, 2차원 구조가 가지는 화합물의 표현자들를 계산한다. 화합물들을 상대적인 활성값에 따라 두개의 클래스로 구분한다. 첫번째 클래스는 화합물들이 hERG저해 활성이 이쓴 화합물에 대하

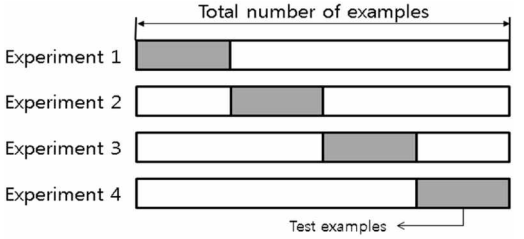


*Fig.* 2. Cross-validation may be used in the sub-sample, for example.

여 +1을 할당하고, 활성이 없는 화합물들에 대해 두번째 클래스에는 -1을 할당한다.

(3) 특성 전처리: 특성 전처리의 목적은 명백히 나쁜 표현자들을 제거하고 표현자들간의 중복과 겹치는 부분을 제거하는 것이다. 여기서는 다음과 같은 descriptor들이 제거된다. 1) 너무 많은 0 값을 갖는 표현자들 (> 90%) 2) 너무 작은 표준편차를 가지는 표현자들 (< 5%) 3) 다른 표현자들과 높은 연관성을 가지는 표현자들(연관계수 > 90%)

(4) GA-GS를 사용한 특성 선택과 매개변수 최적화: 우선 초기 개체군으로 표시하는 이진 문자열의 셋은 무작위로 생성된다. 각 개체에 대해 fitness 함수가 계산 되어진다. 이 과정에서 매개변수 (C, γ)는 그리드 검색 방법에 의해 최적화 된다. 그리고 종료 조건을 검사한다. 종료조건을 만족하지 않으면 교차와 돌연변이가 새로운 개체군을 만들기 위해 수행된다. 이 과정은 최종조건이 만족될 때까지 반복된다.

종료조건은 생성된 세대가 200번에 이르거나 fitness 값이 지난 10번의 세대동안 증가하지 않는 경우이다. 교차 비율은 0.8로 정해졌으며 돌연변이 비율은 0.05이다. 프로그램 언어는 C#과 .NET 3.5 Framework을 사용하였고 svm은 libsvm을 사용하였다.[19]

### hERG 이온 채널 데이터 셋

hERG 이온 채널 구조 및 활성에 대한 데이터들은 PubChem bioassay database[17]에서 추출하였으며, 총 활성화합물 250개, 비활성 1703개에서 금속을 포함한 화합물을 제거하여, 활성 247개와 비활성 1645개, 총 1892개의 화합물로 이루어져 있다. PreADMET2.0[18]을 이용하여 전체 화합물의 표현자(descriptor) 1000여개를 계산하였다.[18] 이 표현자들은 특성 선 처리를 다음과 같이 수행하였다. 1) 너무 많은 0 값(> 90%)을 갖는 표현자들, 2) 너무 작은 표준편차(< 5%)를 가지는 표현자들, 3) 다른 표현자들과 높은 연관관계(연관계수 > 90%)를 가지는 표현자들을 제거하였다. 이러한 일련의 과정을 통하여 최종 130여개로 줄여졌다.

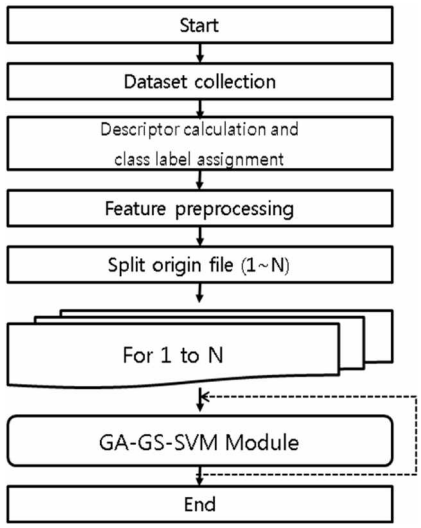


*Fig.* 3. Schematic process for developing the SVM models SVM.

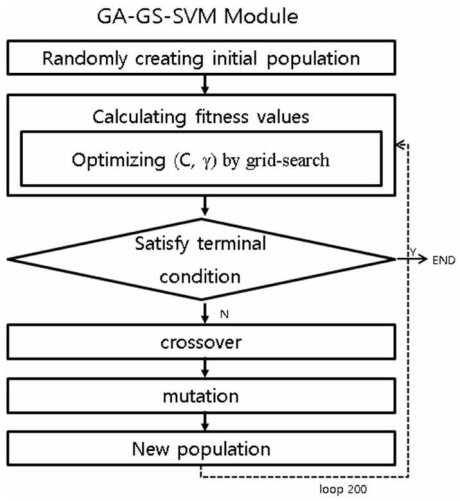


*Fig.* 4. Flowchart of the GA-GS-SVM program.

### hERG 이온 채널 테스트 셋

추출된 174개의 외부 테스트 셋(HEK Cell)은 67개의 활성 화합물과 107개의 비활성 화합물로 이루어져 있다.[20] 이 테스트 셋은 PreADMET2.0을 이용하여 화합물의 표현자 1000여 개를 계산하였다. 이 중에서 트레이닝 셋에서 모델을 만들 때 사용 되어진 표현자만 남기고 나머지는 제거 하여 사용하였다.

### Ensemble Method

Ensemble 방법은 분류기들의 셋을 만들고 예측 결과에 가중치를 두어 분류를 하는 방법이다. 단일GA-GS-SVM이 하나의 트레이닝 셋과 하나의 테스트 셋을 이용하는 반면 앙상블 방법은 여러 개의 트레이닝 셋과 여러 개의 테스트 셋을 이용한다. 따라서 더 많은 수의 데이터가 모델을 만드는데 참여함으로서 정확도를 더욱 높일 수 있다. 우리는 9개의 트레이닝 셋과 9개의 테스트 셋을 만들고 각각에 대해서 9개의 모델을 만들었다. 각 9개의 모델에 대해 예측을 하고 그 중에서 5개 이상이 활성을 나타내면 활성, 4개 이하가 활성이면 결과는 비활성으로 예측을 하도록 하였다. hERG 데이터 셋의 화합물 수는 활성 247개, 비활성 1644개이다. 여러 개의 트레이닝 셋을 만듦으로써 특히 비활성 화합물들이 모델을 만드는데 더 많이 참여하여 결과적으로 비활성 데이터의 예측 정확도가 더욱 높아졌다. 활성 역시 9개의 모델로 예측을 한 경우가 1개의 모델만을 사용한 경우보다 예측정확도가 더 높게 나왔다. 그러나 앙상블 방법을 사용하였을 경우에는 하나의 모델만을 만들 경우보다 시간이 더 많이 걸리게 되는 단점이 있다.

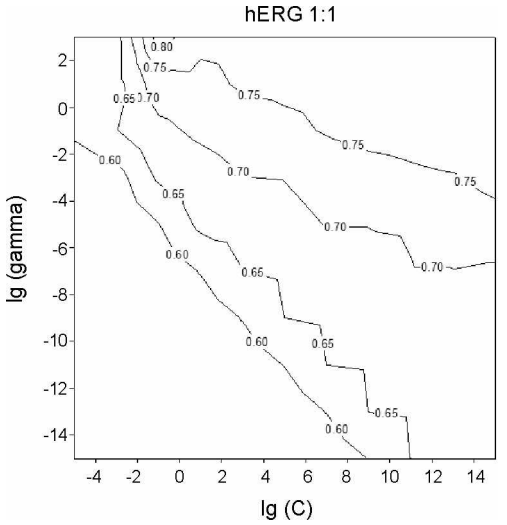


*Fig.* 5. The contour plot of predictability that depends on (C, γ).

## 결과 및 토의

### hERG Prediction Result - 1891 compounds training set

*Fig.* 5는 단일 모델 중 활성과 비활성의 개수를 1대 1로 하여 만든 모델의 결과 값으로 감마와 $C$ 값의 행렬을 이용하여 예측도에 대한 등고선을 그린 것이다. γ와 $C$ 매개변수는 로그 값을 사용하였다. ($C$, γ)에 대하여 그리드방법을 통하여, $C$값이 0.5일 때, γ이 8일 때 최적의 예측도 82.3의 예측도를 얻었다.

*Table* 1은 활성과 비활성의 비율을 1대 1로 하여 트레이닝 셋을 뽑은 결과이다. 비활성의 경우 전체 데이터 셋에서 트레이닝에 참여한 개수가 적

*Table* 1. Predictability of balanced model that was trained with 1:1 ratio of active *vs.* inactive

| | | | Predicted: Active | Predicted: Inactive | Accuracy[e] | Sensitivity[f] | Specificity[g] | Kappa[h] | MCC[i] |
|---|---|---|---|---|---|---|---|---|---|
| Observed | Training | Active | 134[A] | 64[B] | 0.67 | 0.68 | 0.67 | 0.34 | 0.34 |
| | | Inactive | 66[C] | 132[D] | | | | | |
| | Test | Active | 28 | 21 | 0.65 | 0.57 | 0.65 | 0.04 | 0.08 |
| | | Inactive | 501 | 936 | | | | | |

A : True Negative, B: False Positive, C: False Negative, D: True Positive

[e]Accuracy $\frac{A+D}{A+B+C+D}$ [f]Specificity $\frac{A}{A+B}$ [g]Sensitivity $\frac{D}{C+D}$

[h]Kappa $\frac{Accuracy-E}{1-E}$ $E=\frac{(A+C)(A+B)+(B+D)(C+D)}{(A+B+C+D)}$

[i]MCC (Matthews correlation coefficient) $\frac{AD-BC}{\sqrt{(A+B)(A+C)(B+D)(C+D)}}$

*Table* 2. Predictability of unbalanced model that trained with 1:2 ratio of active *vs.* inactive

| | | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Active** | **Inactive** | **Accuracy** | **Sensitivity** | **Specificity** | **Kappa** | **MCC** |
| Observed | Training | **Active** | 76 | 123 | 0.72 | 0.38 | 0.89 | 0.31 | 0.33 |
| | | **Inactive** | 42 | 351 | | | | | |
| | Test | **Active** | 24 | 24 | 0.89 | 0.50 | 0.90 | 0.21 | 0.24 |
| | | **Inactive** | 118 | 1124 | | | | | |

*Table* 3. Predictability of unbalanced model that was trained with 1:3 ratio of active *vs.* inactive

| | | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Active** | **Inactive** | **Accuracy** | **Sensitivity** | **Specificity** | **Kappa** | **MCC** |
| Observed | Training | **Active** | 48 | 150 | 0.80 | 0.24 | 0.99 | 0.30 | 0.39 |
| | | **Inactive** | 7 | 583 | | | | | |
| | Test | **Active** | 22 | 27 | 0.96 | 0.45 | 0.98 | 0.48 | 0.48 |
| | | **Inactive** | 17 | 1028 | | | | | |

*Table* 4. Average predictability of balanced ensemble model that was trained with 1:1 ratio of active *vs.* inactive by GAGSSVM

| | | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Active** | **Inactive** | **Accuracy** | **Sensitivity** | **Specificity** | **Kappa** | **MCC** |
| Observed | Training | **Active** | 154 | 43 | 0.77 | 0.78 | 0.76 | 0.50 | 0.54 |
| | | **Inactive** | 48 | 149 | | | | | |
| | Test | **Active** | 39 | 11 | 0.63 | 0.78 | 0.63 | 0.61 | 0.15 |
| | | **Inactive** | 533 | 904 | | | | | |

기 때문에 상대적으로 예측 정확도가 낮았다. 그러나 활성과비활성의 비율이 비슷하기 때문에 활성의 예측 정확도는 트레이닝에서 68%, 비활성의 예측 정확도는 트레이닝에서 67%로 비슷하게 나왔다. *Table* 2는 활성에 대해 비활성의 비율을 1대 2로 하여 모델을 만든 것이다. 트레이닝에 참여한 개수가 많아지고 전체 비활성 데이터 셋에 비해 모델을 만드는 데 참여한 화합물의 비율이 높아짐으로서 비활성의 예측 정확도는 89%로 좋아졌다. 그러나 트레이닝에 참여한 활성 화합물의 비율이 낮아짐으로써 활성의 예측정확도가 38%로 *Table* 1에 비해 낮아졌다. *Table* 3은 활성에 대해 비활성의 비율을 1대 3으로 하여 결과는 뽑은 것이다. 트레이닝에 참여한 비활성 화합물의 수가 더욱 많아짐으로써 비활성의 예측정확도가 트레이닝에서 99%, 테스트에서 98%로 아주 높게 나왔으나 활성의 예측정확도는 트레이닝에서 24%, test에서 45%로 다른 데이터 셋에 비해 낮게 나왔다. 결과에서, 트레이닝에 참여한 데이터의 수가 많을수록 예측정확도가 더 높게 나온 것을 알 수 있으며, 활성과 비활성의 비율이 다를 때 낮은 비율을 가지는 데이터 셋의 예측정확도가 활성과 비활성의 비율을 같이 했을 때 보다 낮아지는 것을 볼 수 있다.

트레이닝 데이터의 불균형은 예측도 저하에 심각한 영향을 줄 수 있으며, 이를 극복하기 위하여 앙상블모델 방법을 도입하였다.[21] *Table* 4부터 6까지는 앙상블 방법을 이용하여 결과를 도출 한 것이다. 9개의 서브 셋을 만들어 각각의 모델을 만들고 5개 이상의 모델이 활성으로 예측하면 활성, 4개 이하의 모델이 활성으로 예측하면 비활성으로 최종 결과를 뽑았다. *Table* 4는 활성과 비활성의 비율을 1대 1로 한 것이다. 예측 정확도는 트레이

*Table* 5. Average predictability of unbalanced ensemblemodel that was trained with 1:2 ratio of active *vs.* inactive by GAGSSVM

| | | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Active** | **Inactive** | **Accuracy** | **Sensitivity** | **Specificity** | **Kappa** | **MCC** |
| Observed | Training | **Active** | 72 | 127 | 0.74 | 0.36 | 0.94 | 0.61 | 0.38 |
| | | **Inactive** | 25 | 368 | | | | | |
| | Test | **Active** | 19 | 30 | 0.89 | 0.39 | 0.91 | 0.87 | 0.19 |
| | | **Inactive** | 111 | 1131 | | | | | |

*Table* 6. Average predictability of unbalanced ensemblemodel that was trained with 1:3 ratio of active *vs.* inactive by GAGSSVM

| | | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Active** | **Inactive** | **Accuracy** | **Sensitivity** | **Specificity** | **Kappa** | **MCC** |
| Observed | Training | **Active** | 46 | 152 | 0.79 | 0.23 | 0.98 | 0.28 | 0.36 |
| | | **Inactive** | 11 | 579 | | | | | |
| | Test | **Active** | 18 | 31 | 0.94 | 0.37 | 0.97 | 0.344 | 0.34 |
| | | **Inactive** | 30 | 1016 | | | | | |

*Table* 7. Predictability of external validation set for hERG toxicity with balanced ensemble model (1 *vs.* 1)

| | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Active** | **Inactive** | **Accuracy** | **Sensitivity** | **Specificity** | **Kappa** | **MCC** |
| Observed | **Active** | 59 | 8 | 0.63 | 0.88 | 0.47 | 0.30605 | 0.359107 |
| | **Inactive** | 57 | 50 | | | | | |

닝이 77%, 테스트가 63% 나왔다. *Table* 1과 비교해 테스트가 65%에서 63%로 낮아졌지만 활성의 정확도가 58%에서 78%로 상당히 높아졌다. 전체 정확도가 낮아진 것은 비활성의 화합물 수가 활성의 화합물 수보다 약 28배 많기 때문이다. *Table* 5는 활성과 비활성의 비율을 1대 2로 한 것인데 비율이 1대 1인 경우보다 테스트 셋의 예측 정확도가 63%에서 89%로 더욱 향상되었다. 활성의 예측 정확도가 78%에서 39%로 낮아졌지만 비활성의 예측정확도가 63%에서 91%로 높아졌고, 비활성의 화합물 수가 24배 많기 때문에 전체 예측정확도가 크게 상승하였다. *Table* 6은 활성과 비활성의 비율을 1대 3으로 한 것이다. 트레이닝 셋의 경우 비율이 1대 2인 경우보다 활성의 예측 정확도가 36%에서 23%로 더욱 낮아졌지만 테스트 셋의 경우 39%에서 37%로 비슷한 값을 보였다. 비활성의 예측 정확도의 경우는 94%, 91%에서 98%, 97%로 더욱 높아졌다. 결과에서 보듯이 앙상블 모델 방법을 사용하였을 경우 단일 모델보다 예측정확도가 활성과 비활성 모두 높게 나왔으나 활성과 비활성의 화합물 비율을 다르게 했을 경우는 여전히 낮은 비율의 활성 화합물 셋의 예측 정확도가 비활성 화합물 셋의 예측 정확도 보다 낮게 나왔다.

### hERG Prediction Result - 174 compounds external test set

*Table* 7은 활성과 비활성의 비율을 1대 1로 하고 앙상블 방법을 도입하여 만든 모델을 이용하여 174개의 외부 테스트 셋을 예측한 결과이다. 활성의 예측 정확도는 88%, 비활성의 예측 정확도는 47%이다. 전체 예측 정확도는 63%이다. *Table* 8는 활성에 대해 비활성의 비율을 1대 2로 하고 앙상

*Table* 8. Predictability of external validation set for hERG toxicity with unbalanced ensemble model (1 *vs.* 2)

| | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Active** | **Inactive** | **Accuracy** | **Sensitivity** | **Specificity** | **Kappa** | **MCC** |
| Observed | **Active** | 9 | 58 | 0.67 | 0.13 | 1.00 | 0.16026 | 0.295144 |
| | **Inactive** | 0 | 107 | | | | | |

*Table* 9. Predictability of external validation set for hERG toxicity with unbalanced ensemble model (1 *vs.* 3)

| | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Active** | **Inactive** | **Accuracy** | **Sensitivity** | **Specificity** | **Kappa** | **MCC** |
| Observed | **Active** | 40 | 27 | 0.82 | 0.60 | 0.95 | 0.58626 | 0.61152 |
| | **Inactive** | 5 | 102 | | | | | |

블 방법을 도입하여 만든 모델을 이용하여 테스트 셋을 예측한 결과이다. 모델에서 비활성의 개수가 *Table* 4에 비해 많아졌기 때문에 비활성의 예측 정확도가 *Table* 7의 47%에서 100%로 높아졌다. 활성의 예측정확도는 13%로 낮아졌으며 전체 예측정확도는 67%이다. *Table* 9은 활성에 대해 비활성의 비율을 1대 3으로 하고 앙상블 방법을 도입하여 만든 모델을 이용하여 테스트 셋을 예측한 결과이다. 비활성의 예측 정확도가 95%로 *Table* 8의 100%에 비해 다소 낮아졌으나, 활성의 예측정확도는 60%로 *Table* 8의 25%보다 상당히 높게 나왔다. 전체 예측 정확도는 82%로 나왔다. 모델에서의 결과와 같이 테스트 셋의 결과 역시 트레이닝에 참여한 데이터의 수가 많을수록 예측 정확도가 더 높게 나온 것을 알 수 있으며, 활성과 비활성의 비율이 다를 때 낮은 비율을 가지는 데이터 셋의 예측정확도가 활성과 비활성의 비율을 같이 했을 때 보다 낮아지는 것을 볼 수 있다.

## 결론

우리는 SVM modeling에서 특성 선택과 매개변수 최적화를 동시에 수행하는 스키마를 제안하였다. 이 스키마에는 유전자 알고리즘이 특성 선택에 사용 되었고, GS 방법이 매개변수 최적화를 위한 방법으로 선택되었다. 이전의 SVM 방법에 비해 GA-GS-SVM은 전체 예측 결과가 더 좋게 나왔으며, 입력 특성도 더 적은 수를 사용하였다. 하지만 매개변수가 고정된 SVM방법에 비해 속도는 매우 느리다. 이러한 단점을 극복하기 위하여, 향후에 GA방법을 특성선택뿐만 아니라 매개변수 최적화 방법에도 도입할 예정이다. GA-GS-SVM은 hERG 이온채널 저해제 데이터 1891개 화합물에 대해 최대 96%(*Table* 3 활성 대 비활성 데이터의 비율을 1:3로 했을 때의 모델 예측도)의 예측 정확도를 보여주었으며 174개의 테스트 셋에 대해서는 최대 82%(*Table* 9 활성 대 비활성 데이터의 비율을 1:3로 하고, 앙상블방법을 도입하여 만든 모델에 대한 hERG 외부 테스트 셋의 예측도)의 예측 정확도를 보여주었다. 본 논문에서는 다양한 표현자들 중에서 효과적인 표현자를 선택하기 위한 방법으로 유전자 알고리즘을 도입하여 특성 선택을 수행하였고, 매개변수를 최적화하기 위하여 GS방법을 사용하여 모델을 최적화 하였다. 또한 hERG 특성을 예측하는 것과 같은 복잡한 자연현상을 설명하는데 있어서 단일모델로 설명하는 것은 쉽지 않기에 앙상블모델을 도입하였으며, 특히 모델의 활성, 비활성 데이터의 불균형 정도가 심할 경우에는 데이터의 개수가 다른 여러 가지 학습모델을 제안하고, 외부데이터 셋을 통하여 예측률을 비교하여 가장 알맞은 예측모델을 선택하는 일련의 방법을 제시하고 있다.

**Acknowledgments.** 본 연구는 지식경제부 및 정

보통신연구진흥원(2009년은 한국산업기술평가관리원)의 IT핵심기술개발사업[2008-F-029-01, 사이버컴퓨팅 기반 e-Organ 시스템 개발] 사업의 일환으로 수행하였음

## REFERENCES


1. Abbott, G. W.; Sesti, F.; Splawski, I.; Buck, M. E.; Lehmann, M. H.; Timothy, K. W.; Keating, M. T.; Goldstein, S. A. *Cell* **1999**, *97*, 175-87.
2. Fermini, B.; Fossa, A. A. Nat. Rev. Drug Discovery **2003**, *2*, 439-47.
3. Keating, M. T.; Sanguinetti, M. C. *Cell* **2001**, *104*, 569-80.
4. Pearlstein, R.; Vaz, R.; Rampe, D. *J. Med. Chem.* **2003**, *46*, 2017-2022.
5. Aronov, A. M. *Drug Discovery Today* **2005**, *10*, 149-155.
6. Recanatini, M.; Poluzzi, E.; Masetti, M.; Cavalli, A.; De Ponti, F. *Med. Res. Rev.* **2005**, *25*, 133-166.
7. Mitcheson, J. S.; Chen, J.; Lin, M.; Culberson, C.; Sanguinetti, M. C. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12329-12333.
8. Li, Q.; Jorgensen, F. S.; Oprea, T.; Brunak, S.; Tabboureau, O. *Mol. Pharm.* **2008**, *5*(1), 117-127.
9. Lucasius, C. B.; Kateman, G. Chemometr. *Intell. Lab.* **1993**, *19*, 1-33.
10. Guyon, I.; Weston, J.; Barnhil, S.; Vapnik, V. *Mach. Learn.* **2002**, *46*, 389-422.
11. Sutter, J. M.; Kalivas, J. H. *Microchem. J.* **1993**, *47*, 60-66.
12. Hsu, C. W.; Chang, C.C.; Lin, C. J. http://www.csie.ntu.edu.tw//~cjlin/papers/guide/guide.pdf. **2003**.
13. Vapnik, V. Statistical Learning Theory; *Wiley: New York*, USA., **1998**.
14. Seymour G. *J. of the Am. Stat. Ass.* **1975**, *70*, 350.
15. Xue, Y.; li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630-1638.
16. Davis, L. handbook of genetic algorithms Van Nostrand Reinhold New York, USA., **1991**.
17. BMDRC, PreADMET 2.0; Seoul, Korea, **2007**, http://preadmet.bmdrc.org.
18. PubChem bioassay database (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=376)
19. Chang, C. C.; Lin, C. J. LIBSVM: A library for support vector machines. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm, **2001**.
20. Li, Q.; Jorgensen, F. S.; Oprea, T.; Brunak, S.; Taboureau, O. *Mol. Pharm.* **2008**, *5*(1), 117-127.
21. Kang. P. Cho. S. *Lecture Notes in Computer Science* Springer Berlin, Germany, **2006**, *4232*, 837-846.