농약과학회지 제9권 제2호(2005)
The Korean Journal of Pesticide Science
Vol. 9, No. 2. pp.140~145(2005)

Two Class Approximation of TLB (Tomato Late Blight) Activity Data

Hoh Gyu Hahn · Ashek Ali MD · Seung Joo Cho*

Life Science Division, Korea Institute of Science and Technology, Seoul 130-650, Korea

Abstract : Quantitative Structure Activity Relationship (QSAR) assumes the relatedness between physical property and biological activity. However, activity data measured at single concentration such as percent activity have not been used extensively for modeling purpose. This probably comes from the fact that these values are qualitative instead of quantitative. To utilize percent activity data for molecular modeling, we classified the whole data into two classes. One class represents the active while the other signifies the inactive. The percent activity data of β-Ketoacetoanilides measured for TLB (Tomato Late Blight) were investigated. CoMFA (Comparative Molecular Field Analysis) was used as a discriminant function. Using CoMFA provides 3D (three dimensional) information, which is crucial for chemical insight. It can also serve as a predictive model. The resultant model classified the given data correctly (98%). When LOO (leave-one-out) crossvalidation procedure was applied, the classification accuracy was 69%. Therefore two class approximation of percent activity data with CoMFA can be utilized to understand the relationship between chemical structure and biological activity and design subsequent chemical analogs. (Received February 28, 2005; accepted June 24, 2005)

Key Words : QSAR; CoMFA; Percent Activity; Tomato Late Blight.

Introduction

Quantitative Structure Activity Relationship (QSAR) usually assumes the relatedness between physical property and biological activity (Leach et al, 2001). Biological activity data such as IC50, EC50, LD50 and equilibrium constants are usually used as target variables. These values are actually representative values, having information of biological activity over several concentrations. On the other hand, in some situations such as high throughput screening (HTS), activity data are measured just once (one concentration). One of the purposes of these experiments is to decide which compounds are suitable to investigate further. These values are expected to have large errors due to non-repeated measurements. Although data appears quantitative, these values possess only qualitative information. For example, 100% activity may mean micromolar, nanomolar, picomolar, or even femtomolar activity (such as in IC50). Likewise, 0% activity may mean either this compound is completely irrelevant to a certain receptor of interest, or the activity is just not strong enough to be detected at that given concentration. Furthermore all experimental data have some errors. When measuring the activity small difference does not have much meaning because of the large experimental measurement error. For example, both 100% and 99% imply high activity without much confidence that 100% actually indicates higher activity than 99%. Although this kind of experiments is crude, it is still expensive. Clearly there is a need to get some information from this kind of data. So we applied two-class approximation to handle this kind of qualitative data. This approximation concurs experimental point of view. At the early stage of the drug development, the activities are screened for many compounds using quick and dirty biological activity tests. From this screening, we decide whether a certain compound is active or not. If this compound is active, it might be considered further. Therefore our main objective is to analyze the data for modeling, i.e., for prediction of activity for untested compounds. Here we roughly divided the whole set of compounds into two

* Corresponding author (Tel: +82-2-958-5134, E-mail : chosj@kist.re.kr)

categories, either being active or inactive, setting 50% as a classification criterion. If the activity of a compound is higher than 50%, it has class membership value of 1, which means it belongs to the active class. Likewise if the activity is lower than 50%, the compound membership value is −1, which is inactive. The data set was chosen because they look congeneric and have a wide span of percent activities. The structures of ß-ketoacetoanilide derivatives are shown in Figure 1 and the percent activity data are listed in Table 1 (Hahn *et al*, 2004).

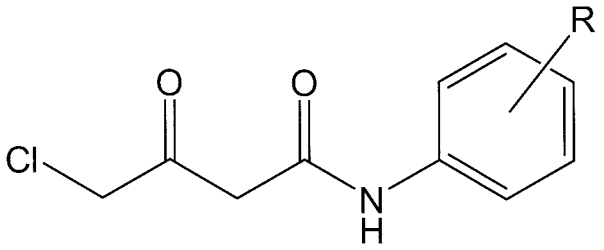


Fig. 1. The structure of ß-ketoacetoanilide derivatives. See Table 1 for substitutents.

Methods

Quantitative structure-activity relationships (QSARs) are important tools to understand why the active compounds exhibit certain biochemical activities. One of the most widely used tools in 3D QSAR study is comparative molecular filed analysis (CoMFA) (Cramer et al, 1988). CoMFA is based on the assumption that changes in the biological activity correlate with changes in the steric and electrostatic fields of molecules. While this approach has been widely accepted and scientifically feasible, it is not without problems.

Both potential functions are very steep near the van der Waals surface of the molecule, causing rapid changes, and requiring the use of cut-off values. So changes in orientation of the superimposed molecules, relative to the calculation grid, can cause significant changes in CoMFA results. In addition, a scaling factor is applied to the steric field, so both fields can be used in the same PLS analysis. Here CoMFA is used for classification rather than for regression. The molecular modeling software "SYBYL 7.0" was used for three-dimensional structure generation and molecular modeling studies. The molecular geometry of each compound was first minimized using a standard Tripos molecular mechanics force field with a 0.005 kcal/mol energy gradient convergence criterion and their charge were calculated by the Gasteiger-Huckel methods.

The most important requirement for 3D-QSAR techniques (CoMFA) is that the 3D structures of the molecules to be analyzed by aligned according to a suitable conformational template, which is assumed to adopt a "bioactive conformation". In the present study, since the structural information on these inhibitor-protein complexes are not available; the conformation of the molecules was obtained from systematic conformational search procedures.

All the rotatable bonds on the analogues were searched by rotating from 0 to 330 by 30 increments and the lowest energy conformer was selected. In case of some compounds the lowest energy conformer was not taken to keep structural consistency with previous compounds.

The molecule in the dataset (Table 1) were aligned using the database RMSD fitting option available in SYBYL using compound 11 as a template (the simplest compound). The resultant superimposed structures are shown in Figure 2.

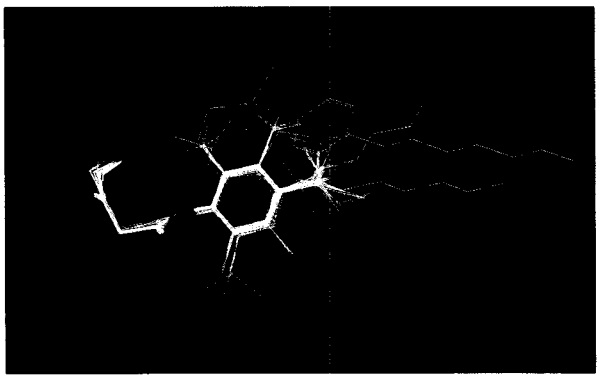


Fig. 2. Superposition of ß-ketoacetoanilide derivatives.

Sterically favored areas (contribution level of 85%) are represented by green polyhedra. Sterically disfavored areas (contribution level of 15%) are represented by yellow polyhedra (A). Positive charged favored areas (contribution level of 90%) are represented by blue polyhedra. Negatively charged favored areas (contribution level of 15%) are represented by blue polyhedra (B). The standard molecule is shown in the maps.

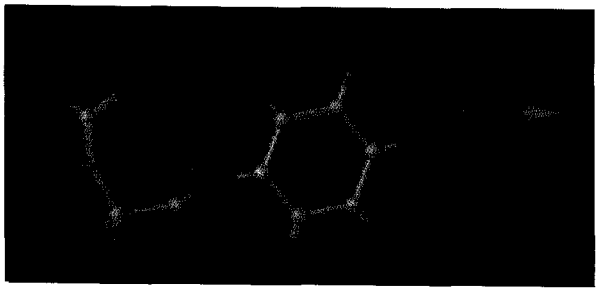
Table 1. The Structure, Percent Activity of ß-Ketoacetoanilide Derivatives and their Classes

Compound name	Percent activity	Class a	CoMFA Class b	CoMFA class(LOO) c
$C_6H_4(4\text{-}CO_2C_2H_5)$	99	1	1	1
$C_6H_3(3,5\text{-di F})$	99	1	1	-1
$C_6H_4(3\text{-}CO_2C_2H_5)$	96	1	1	-1
$C_6H_3(3,4\text{-di CH3})$	95	1	1	1
$C_6H_3(3,5\text{-di Cl})$	90	1	1	1
$C_6H(2,3,5,6\text{-tetra Cl})$	90	1	1	1
$C_6H_4(4\text{-}COCH_3)$	90	1	1	1
$C_6H_4(4\text{-}CH_2CH_2CH_3)$	89	1	1	-1
$C_6H_4(3\text{-Cl, 4-CN})$	89	1	1	-1
$C_6H_4(4\text{-}CH(CH_3)_2)$	88	1	1	1
$C_6H_4(4\text{-}O(CH_2)_3CH_3)$	88	1	1	1
$C_6H_4(4\text{-}OC_6H_5)$	84	1	1	-1
$C_6H_3(3,4\text{-di }CH_3)$	84	1	1	1
$C_6H_4(4\text{-CH2CH3})$	84	1	1	1
$C_6H_3(3\text{-Cl, 4-}OCH_3)$	84	1	1	1
$C_6H_2(2,4,6\text{-tri }CH_3)$	84	1	1	-1
$C_6H_2(2,6\text{-di Br, 4-OCF3})$	84	1	1	1
$C_6H_4(4\text{-}OCH((CH_3)_2)$	84	1	1	1
$C_6H_3(3\text{-}CH_2CH_2CH_2(4\text{-}))$	84	1	1	1
$C_6H_3(3\text{-Br, 4-}CH_3)$	84	1	1	1
$C_6H_4(4\text{-Cl})$	81	1	1	1
$C_6H_4(4\text{-}OCF_3)$	81	1	1	1
$C_6H_4(4\text{-}CH_3)$	81	1	1	-1
$C_6H_4(4\text{-}C(CH_3)_3)$	81	1	1	1
$C_6H_4(4\text{-}O(CH_2)4CH_3)$	81	1	1	1
$C_6H_4(4\text{-Br})$	78	1	1	1
$C_6H_3(2,5\text{-di Cl})$	78	1	-1	-1
$C_6H_3(2\text{-}CH_2CH_2CH_2CH_2(3\text{-}))$	78	1	1	-1
$C_6H_4(4\text{-}O(CH_2)4CH_3)$	78	1	1	1
$C_6H_4(4\text{-}OCH_3)$	69	1	1	1
$C_6H_2(2,6\text{-di Cl, 4-}CF_3)$	69	1	1	1
$C_6H_3(3\text{-}OCH_2CH_2O(4\text{-}))$	69	1	1	1
$C_6H_2(2,4,6\text{-Tri Cl})$	63	1	1	1
$C_6H_2(2,6\text{-di Cl, 4-}OCF_3)$	63	1	1	1
$C_6H_4(4\text{-}CO_2CH_3)$	63	1	1	1
$C_6H_4(4\text{-}CHF_2)$	56	1	1	-1
$C_6H_3(3\text{-F, 4-}OCH_3)$	56	1	1	1
$C_6H_4(2\text{-}CH_3, 5\text{-}CH_3)$	56	1	1	-1
$C_6H_4(4\text{-}OCH_2CH_3)$	50	?	?	?
$C_6H_4(4\text{-}(CH_2)3CH_3)$	44	-1	-1	1
C_6H_5	44	-1	-1	1
$C_6H_4(4\text{-}OC_6H_4(4\text{-Cl}))$	44	-1	-1	-1
$C_6H_3(2,4\text{-di Cl})$	44	-1	-1	1
$C_6H_4(4\text{-I})$	44	-1	-1	1
$C_6H_4(4\text{-}NO_2)$	38	-1	-1	-1
$C_6H_3(2,6\text{-di }CH_2CH_3)$	31	-1	-1	-1

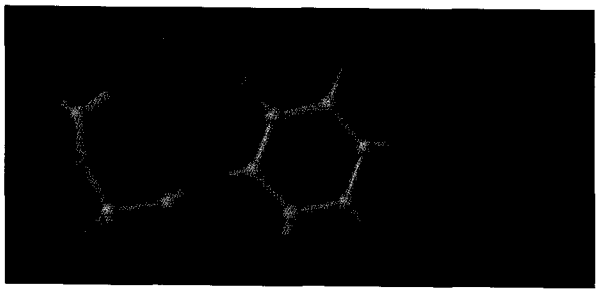
Table 1. continued

Compound name	Percent activity	Class a	CoMFA Class b	CoMFA class(LOO) c
C_6H_3(2,4-di CH_3)	31	-1	-1	-1
C_6H_3(2-CH_3, 4-Br)	31	-1	-1	-1
C_6H_3(2-CH_3, 4-Cl)	31	-1	-1	-1
C_6H_4(4-CN)	25	-1	-1	1
C_6H_3(2-CH_3, 4-OCH_3)	25	-1	-1	1
C_6H_4(3-OCH_3)	25	-1	-1	-1
C_6H_4(2-CH_3)	25	-1	-1	-1
C_6H_4(2-Cl)	25	-1	-1	-1
C_6H_3(2,3-di Cl)	25	-1	-1	-1
C_6H_4(4-F)	25	-1	-1	-1
C_6H_4(3-CN)	25	-1	-1	1
C_6H_4(2-CH_2CH_3)	25	-1	-1	-1
C_6H_3(3,5-di CF_3)	19	-1	-1	1
C_6H_2(2,6-di Br, 4-F)	19	-1	-1	1
C_6H_4(2-NO_2)	19	-1	-1	-1
C_6H_4(2-F)	13	-1	-1	-1
C_6H_4(2-Br)	13	-1	-1	-1
C_6H_3(2,4-di OCH_3)	13	-1	-1	-1
C_6H_3(2-NO_2, 4-Cl)	13	-1	-1	1
C_6H_3(3-F, 4-CH_3)	13	-1	-1	1
C_6H_2(2,4,6-tri $C(CH_3)_3$)	13	-1	-1	1
C_6H_4(2-CF_3)	6	-1	-1	-1
C6H(3,5-di F)	6	-1	1	1
C_6H_4(3-F)	6	-1	-1	-1
C_6H_3(2-F, 5-Cl)	6	-1	-1	-1
C_6H_2(2,3,4-tri F)	6	-1	-1	-1
C_6H_3(3,5-di OCH_3)	6	-1	-1	1
C_6H_4(4-C_6H_{13}(n-)$_3$)	6	-1	-1	-1
C_6H_4(4-$C_{10}H_{21}$(n-))	6	-1	-1	1
C_6H_3(2-Br, 4-CH_3)	3	-1	-1	-1
C_6H_4(3-CH_3)	0	-1	-1	1
C_6H_4(4-CF_3)	0	-1	-1	-1
C_6H_2(2,4,6-tri F)	0	-1	-1	-1
C_6H_3(2-OCH_3, 4-CH_3)	0	-1	-1	-1
C_6H_3(2-F, 5-CH_3)	0	-1	-1	1
C_6H_3(2,5-di F)	0	-1	-1	-1
C_6H_3(2-F, 4-NO_2)	0	-1	-1	-1
C_6H_3(2-Cl,4-NO_2)	0	-1	-1	-1
C_6H_3(2-CH_3, 5-Cl)	0	-1	-1	-1
C_6H_3(2-OCH_3, 5-CH_3)	0	-1	-1	-1
C_6H_3(2-NO_2, 4-CH_3)	0	-1	-1	-1
C_6H_3(3-OCH_2O(4-))	0	-1	-1	-1
C_6H_4(4-C_7H_{15}(n-))	0	-1	-1	-1

a1 means active, -1, inactive; bCoMFA Prediction; cCoMFA Prediction with LOO(Leave one out) Procedure.



(A) Steric Map



(B) Electrostatic Map

Figure 3. CoMFA steric STDDEV*COEFF contour plots. Sterically favored areas (contribution level of 85%) are represented by green polyhedra. Sterically disfavored areas (contribution level of 15%) are represented by yellow polyhedra (A). Positive charged favored areas (contribution level of 90%) are represented by blue polyhedra. Negatively charged favored areas (contribution level of 15%) are represented by blue polyhedra (B). The standard molecule is shown in the maps.

Results and Discussion

In CoMFA analysis the compounds listed in the table 1 yielded a model with a q^2 value of 0.142 in an optimum component number of 6. The non-crossvalidated PLS analysis with the optimum number of components determined by the cross-validated analysis, has given a r^2 of 0.739, F value of 38.763 and an estimated standard error of 0.264. The CoMFA using the class membership value in Table 1 gave results which are real numbers. Here we assigned the value into 1 when the outcome of CoMFA is positive, otherwise, -1. Therefore CoMFA is used as a discriminant function rather than regression in this study. The results are summarized in Table 2. Like the previous reports, CoMFA map shows that when substituents are attached at 4-position of the benzene ring, the activity is enhanced. Unfortunately, the complexity of resultant map prevents the easy explanation why some compounds are more active than others. However using this CoMFA model, we could correctly classify 98% of the compounds in the data set. When LOO (leave one out) procedure was applied to test predictability of this model, 69% of the compounds were classified correctly (Table 2). Therefore the classification power for this model for given dataset is reasonably good. While predictive power of this model is somewhat lower it is still statistically very significant (Pearson's chi square is 30.89 and p-value is 2.73×10^{-8}) (Fienberg, S. E. et al, 1985). Therefore this model can be utilized to design new compounds for TLB. This study demonstrates that two-class approximation of percent activity data with CoMFA can be utilized to understand the relationship between chemical structure and biological activity and design subsequent chemical analogs.

Reference

Leach, A. R. (2001) The Use of Molecular Modeling and chemoinformatics to Discover and Design New Molecules. pp.640~726, Molecular Modeling Addison Longman Limited.

Cramer, R. D., D. E. Petterson and J. D. Bunce (1988) Comparative Molecular Field Analysis (CoMFA). 1.

Table 2. The predicted and experimental values

	Active (predicted)	Inactive (predicted)	Sum
Active (experiment)	37(27)	1(11)	38
Inactive(experiment)	1(17)	49(33)	50
Sum	38(44)	50(44)	88

Data in parentheses are CoMFA results with LOO crossvalidation. The explanatorary power is very good. However internal prediction shows that the data may not be very stable for further prediction.

Effect of Shape on Binding of Steroids to Carrier Proteins J. Am. Chem. Soc. 110:5959~5967.

Fienberg, S. E. (1985) Inference for Contingency Table pp 70-114 The Analysis of Cross-Classified Categorical Data MIT press.

Hahn, H. -G., K. D. Nam, S. Bae, B. S. Yang, S. -W. Lee and K. Y. Cho (2004) Synthesis of combinatorial library of b-ketoacetoanilide chlorides and their antifungal activity against main plant pathogens Korean J. of Pesticide Sci. 8(1):8~15.

토마토 역병균 항균 활성 데이터의 이분번 근사모델링

한호규, Ashek Ali MD, 조승주*(한국과학기술연구원)

요약 : 정량적 구조 활성관계 모델링은 물리적인 성질과 생물학적 활성이 관계 있다는 것을 전제로 한다. 그러나, 퍼센트 활성과 같은 데이터들은 모델링에 많이 활용되지 않았다. 이것의 중요한 이유중의 하나는 이러한 값들이 정량적이 아니고 정성적인 데에 있다. 본 연구에서는 분자모델링에 퍼센트활성 데이터를 활용하기 위하여 데이터 값들을 2개의 계층으로 분류하고 CoMFA(비교분자장)를 판별함수로 활용하였다. 즉, 베타-케토아세트아닐라이드 유도체들의 토마토 역병균에 대한 항균력 시험의 퍼센트 활성 데이터를, 한 계층은 활성이 있는 것, 다른 계층은 활성이 없는 것으로 나누었다. 특히, CoMFA를 활용함으로써 화학적인 이해에 중요한 3차원적인 정보를 얻을 수 있었다. 이 모델은 주어진 데이타를 98%의 정확도로 설명하였으며, LOO 검증을 해본 결과 예측력은 약 69% 정도였다. 이 결과는 활성 데이터를 근사적으로 2개의 계급으로 나누고 CoMFA를 활용하는 방식이 구조활성관계를 이해하고 화합물 유도체를 합성하는데 활용될 수 있음을 보여준다.

색인어 : 구조활성관계(QSAR), 비교분자장분석(CoMFA), 퍼센트 활성, 토마토 역병.

*Corresponding author(Tel : +82-2-958-5134, E-mail : chosj@kist.re.kr)