DEPENDENCE BETWEEN DETERMINATION COEFFICIENT AND NUMBER OF REGRESSORS: A CASE STUDY ON RETENTION TIMES OF MYCOTOXINS

SORANA D. BOBLOACĂ[a,*], LORENTZ JÄNTSCHI[b,c], RADU E. SESTRAŞ[c,*]

ABSTRACT. In the present work, a dependence analysis of the determination coefficient and the number of regressors was carried out on a set of 65 mycotoxins. The simple and multiple regression techniques were used to identify the linear relationship between retention times and molecular descriptors calculated with HyperChem from the optimized 3D structures. The highest number of regressors to be used in investigating the retention times as function of mycotoxins investigated properties must be equal to 5. As far as the dependence between number of regressors and determination coefficient was concerned, the analysis revealed that the best relationship is linear if the cutoff is set at 4 or 5 regressors while exponential for more than 5 regressors. The results must be verified on other classes of compounds and other dependent or independent variables in order to be extrapolated.

Keywords: *mycotoxins; retention time; regression analysis.*

INTRODUCTION

Mycotoxin, a secondary metabolite produced by a fungus [1,2], is produced according to the intrinsic and extrinsic environmental conditions [3,4]. Aflatoxins, discovered in later 1950's and early 1960's [5], produced by species of Aspergillus (*Aspergillus flavus*, *Aspergillus parasiticus*), are known to act at the DNA level (e.g. gene point mutations, deletions and insertions, recombination, rearrangements and amplifications) as carcinogenetic substances [6,7]. Trichothecenes, produced by species as *Fusarium*, *Myrothecium*, *Trichoderma*, *Trichothecium*, *Cephalosporium*, *Verticimonosporium*, and *Stachybotrys*, proved to be detrimental to the neurological system [8,9]. Roquefortin, produced by fungi from Penicillium genus [10,11], proved to be toxic for animals [12,13]. Ochratoxin, mycotoxins produced by some *Aspergillus* (e.g. *Aspergillus*

[a] *„Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 13 Emil Isac, RO-400023 Cluj-Napoca, sbolboaca@umfcluj.ro*

[b] *Technical University of Cluj-Napoca, 28 Memorandumului, RO-400114 Cluj-Napoca, Romania, lori@academicdirect.org*

[c] *University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 3-5 Mănăştur Street, RO-400372 Cluj-Napoca, rsentras@usamvcluj.ro*

ochraceus) and *Penicillium* species (e.g. *Penicillium viridicatum*) is accumulated in the meat of animals and is known to be a human carcinogen [14,15,16]. Thus, mycotoxins proved to have important effects on animal and human health [17,18,19,20]. Moreover, a series of researches have been conducted to assess the cost and efficacy of public health interventions carried out in order to reduce the mycotoxin-induce human diseases [21,22,23].

The present research was aimed to obtain the dependence of determination coefficient of retention times as function of the number of independent variables, as properties obtained from the structure of a sample of mycotoxin. Starting with these results, the best dependence between determination coefficients and the number of regressors was investigated.

RESULTS AND DISUSSIONS

The search of the maximum determination between retention times and some properties of the studied mycotoxins revealed the following:

The distribution of the steps where an increase in determination coefficient value was observed, is: 5 (1 independent variable) - 3 (2 independent variables) - 7 (3 independent variables) - 4 (4 independent variables) - 10 (5 independent variables) - 5 (6 independent variables) - 3 (7 independent variables) - 2 (8 independent variables) - 6 (9 independent variables) - 9 (10 independent variables).

The descending classification of the range (defined as the difference between maximum and minimum value) of the determination coefficient according to the same number of independent variable was as follow: ▪ 0.3909 (1 variable); ▪ 0.0977 (5 variables); ▪ 0.0647 (2 variables); ▪ 0.0605 (3 variables); ▪ 0.0517 (4 variables); ▪ 0.0239 (6 variables); ▪ 0.0132 (9 variables); ▪ 0.0108 (8 variables); ▪ 0.0102 (7 variables); and ▪ 0.0083 (10 variables). Note that all models were statistically significant at a significance level of 5%.

The summary of the results obtained in the best regression when the dependence between retention times and properties of the studied mycotoxins was investigated is presented in Table 1.

The matrix of properties used as independent variables in the regressions models when the highest value of the determination coefficient was obtained is shown in Table 2.

The dependence between the number of regressors (independent variables) and determination coefficient was as follows:

- Case $x \leq 4$, linear (Figure 1): $\hat{y} = 6.797 \cdot 10^{-2} * x + 0.3270$

$$(r^2 = 0.9991,\ r^2_{adjusted} = 0.9987;\ F = 2299) \qquad Eq(1)$$

where x = number of independent variables; ŷ = estimated determination coefficient.

- Case $4 < x \leq 10$, exponential (Figure 2): $\hat{y} = 0.8173 - 0.7972*\exp(-x/2.6772)$
  $(r^2 = 0.9998;\ r^2_{adjusted} = 0.99965;\ F = 7044)$ Eq(2)

Table 1. Summary of linear regressions: mycotoxins dataset

x	r^2	t	p
1	0.3935	6.39	$1.13 \cdot 10^{-8}$
2	0.4635	7.32	$3.02 \cdot 10^{-10}$
3	0.5343	8.37	$5.17 \cdot 10^{-12}$
4	0.5964	9.42	$9.88 \cdot 10^{-14}$
5	0.6943	11.58	$4.04 \cdot 10^{-17}$
6	0.7318	12.58	$1.61 \cdot 10^{-18}$
7	0.7599	13.43	$1.32 \cdot 10^{-19}$
8	0.7766	13.95	$3.57 \cdot 10^{-20}$
9	0.7898	14.38	$1.41 \cdot 10^{-20}$
10	0.7982	14.61	$1.03 \cdot 10^{-20}$

x = number of independent variable; r^2 = determination coefficient;
t = t-value associated to r^2; p = p-value associated to t-value

Table 2. Matrix of mycotoxins properties used in the models with highest determination coefficients

x	LogP	InvHydE	LnPol	LnRefr	Vol	SAG	LnMW	negte	dm	negsae	sce	Invnegte	Invnegsae	Invnegsbe	Lnnegsae	Lnnegsbe	Lnsce	Lnnegsee	**Total**
1	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
2	0	0	**1**	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**
3	0	0	0	0	0	0	**1**	0	0	0	**1**	0	**1**	0	0	0	0	0	**3**
4	0	**1**	**1**	0	**1**	**1**	0	0	0	0	0	0	0	0	0	0	0	0	**4**
5	0	**1**	**1**	0	**1**	0	0	**1**	0	0	0	0	0	0	0	0	0	**1**	**5**
6	0	**1**	**1**	0	**1**	0	0	0	0	**1**	0	0	0	0	0	**1**	**1**	0	**6**
7	0	**1**	**1**	0	**1**	0	0	0	**1**	**1**	0	0	0	0	0	**1**	**1**	0	**7**
8	**1**	**1**	**1**	0	**1**	0	0	**1**	**1**	0	0	0	0	**1**	0	0	**1**	0	**8**
9	**1**	**1**	**1**	**1**	**1**	0	0	**1**	**1**	0	0	0	0	**1**	0	0	**1**	0	**9**
10	**1**	**1**	**1**	**1**	**1**	0	0	0	**1**	0	0	**1**	0	0	**1**	**1**	0	**1**	**10**
Total	**3**	**7**	**9**	**2**	**8**	**1**	**1**	**3**	**4**	**2**	**1**	**1**	**1**	**2**	**1**	**3**	**4**	**2**	**55**

LogP = octanol-water partition coefficient (log scale);
HydE = hydration energy; Pol = polarizability; Refr = refractivity;
Vol = volume; SAG = surface area grid; MW = molecular weight;
te = total-energy; dm = dipole moment; sae = scf-atom-energy;
sce = scf-core-energy; sbe = scf-binding-energy;
see = scf-electronic-energy; ln = natural logarithm; inv = inverse value

The regression between retention times and five mycotoxin properties also provided a linear dependence between determination coefficient and the number of regressors but the performance of the model decreased with 0.005 in terms of determination coefficient of the overall model:

- $r^2 = 0.9941$, $r^2_{adjusted} = 0.9921$; F = 504 Eq(3)

The correlation between determination coefficients and the number of regressors, presented in terms of correlation coefficient, t-value associated to correlation coefficient and its p-value, are shown in Table 3.

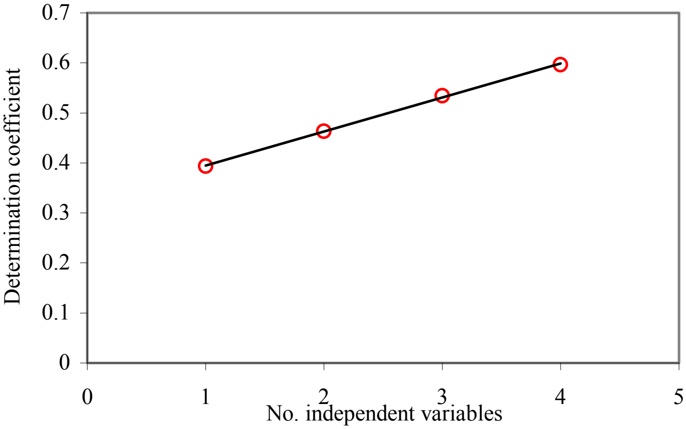

Figure 1. The dependence between r^2 and the number of independent variables for $1 \leq x \leq 4$

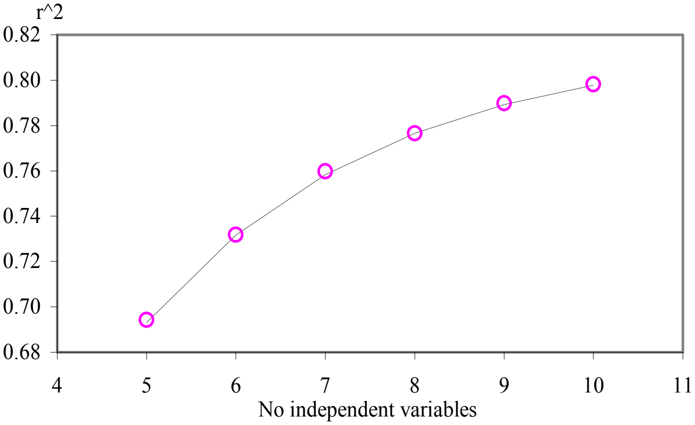

Figure 2. The dependence between r^2 and the number of independent variables for $4 < x \leq 10$

Table 3. Summary of linear relation between determination coefficient and the number of independent variables

x	r	t-value	p-value
3	> 0.9999	282.72	$1.13\cdot10^{-3}$
4	0.9996	47.95	$2.17\cdot10^{-4}$
5	0.9970	22.45	$9.68\cdot10^{-5}$
6	0.9964	23.54	$9.66\cdot10^{-6}$
7	0.9907	16.31	$7.90\cdot10^{-6}$
8	0.9808	12.31	$8.75\cdot10^{-6}$
9	0.9690	10.38	$8.37\cdot10^{-6}$
10	0.9557	9.19	$7.96\cdot10^{-6}$

x = number of regressors; r = correlation coefficient;
t-value = Student statistics associated to r;
p-value = significance associated to Student statistics

The present study was aimed to answer to the following questions: "How many regressors (independent variables) are needed to obtain a certain determination of retention times as a function of mycotoxins properties?" and "Which is the type of relation between the determination coefficient and the number of regressors able to estimate the mycotoxins retention times?". The answers are as follows.

All linear regression models, able to explain the retention times as function of investigated mycotoxins properties, where selected such as to accomplish the Hawkins principles [24]: highest correlation coefficient, highest Fisher parameter, lowest standard error of the estimate, and smallest possible number of significant parameters ($n = 5\cdot v$, where n = sample size, v = number of variables in the model).

As expected, the highest determination coefficient was obtained by a linear regression model with 10 independent variables. The number of acceptable models, identified for a given number of independent variables, varied, for the studied mycotoxins, from 2 to 9 (e.g., for v=8; 2 models while for v=10; 9 models). Naturally, the number of regressions increases with the number of independent variables as well as the possible combinations of variables; in the present experiment, we limited the investigation at 10 independent variables. Within such limits, no linear dependence between the number of independent variables and the number of significant models could be identified on linking the retention times with mycotoxin selected properties.

The results are as follows:

- The most important increase in determination coefficient is observed when one independent variable is used, followed by the case when 5 and respectively 2 independent variables were used.

- The highest difference is observed when the best 4-regressors model was compared to the best 5-regressors model. The observed difference in determination coefficient was of 0.0978. This observation suggest that the highest number of regressors to be used in investigation of the retention times as function of mycotoxin properties must not exceed 5. The use of more than 6 regressors is useless and will not bring a significant increasing of the determination coefficient in explaining the retention times as function of mycotoxin properties.
- The most frequent mycotoxin property able to explain the retention times proved to be the polarizability (it appears in all models with one exception, the model with 3 independent variables), closely followed by the molecular volume (it appears in all models with two exceptions, the model with one independent variable and the model with three independent variables) and the inverse of hydration energy (it appears in all model with more than three variables) (Table 2).

The relationship between retention times and structural properties of mycotoxins has previously been studied. A set of 15 toxins with the trichothecene nucleus was investigated by applying a ComMFA ($r^2 = 0.921$) [25]. Khosrokhavar et al. [26] had identified, on a set of 67 mycotoxins, a multiple linear regression model with a determination of 0.931 that included the following descriptors: the octanol/water partition coefficient, electronic energy, dipole length, and the LUMO energy. It is no doubt that the models reported in literature are better than models presented in this paper, in terms of determination coefficient. However, the identification of the linear relationship between the retention times and the investigated properties was performed in order to answer to the research questions.

As far as dependence between the number of regressors and determination coefficient was concerned, the performed analysis revealed the following:

- The identified linear model (presented in Eq(1)) as well as the exponential model (presented in Eq(2)) proved to be statistically significant ($F > 2200$).
- The linear relationship was identified if the applied cutoff was set at 4 regressors. For this regression, 99.91% of the variation of determination coefficient proved to be explained by a linear relationship with the number of regressors (Eq(1)). The cutoff set to 5 regressors also led to a statistically significant linear regression between the number of regressors and determination coefficient. In this case, 99.41% of variation in determination coefficient could be explained (Eq(3)). Although the models presented in Eq(1) and Eq(3) are statistically significant, the F-statistics is almost 5 times smaller if the cutoff is set at 5 regressors compared to a cutoff set at 4 descriptors.

- The best relationship between the determination coefficient and regressors proved to be of exponential type. Almost 100% from variation of determination coefficient could be explained, with a number of regressors between $4 < x \leq 10$ (Eq(2)).

As far as the results presented in Table 3 are concerned, the minimum value of the significance was obtained for $1 \leq x \leq 7$ ($p = 7.90 \cdot 10^{-6}$). This means that the maximum number of regressors able to explain the relationship between retention times and investigated mycotoxin properties must not exceed 7.

The present study was aimed to provide answers to two main research questions: “How many regressors (independent variables) are needed to obtain a certain determination of retention times as a function of mycotoxin properties?” and “Which is the type of relation between the determination coefficient and the number of regressors in case of linear regressions?”. Valid answers were identified and the aim of the research was reached. However, a new questions arise: “Is the cutoff maintained as far as the number of regressors is concerned dependent by the sample size?”, “The exponential relationship between the determination coefficient and the number of regressors is valid only for the investigated set of structures or it is obeyed to any sets of compounds?” Current research in our laboratory is aimed to answer to these new questions.

CONCLUSIONS

In the present study, the dependence of maximum value of determination coefficient of the number of regressors, as well as the type of this relationship were investigated on a set of 65 mycotoxin properties. The most frequent mycotoxin properties able to explain the retention times proved to be the polarizability, closely followed by the molecular volume and inverse of hydration energy. The highest number of regressors to be used in investigation of the retention times as function of mycotoxin properties needed do not exceed 5. As far as the dependence between number of regressors and determination coefficient was concerned linear, the cutoff was set at 4, or > 4 for an exponential dependence. The obtained results must be verified on other classes of compounds and other dependent or independent variables in order to be extrapolated.

EXPERIMENTAL SECTION

A sample of 65 mycotoxins was included in the present study. The structures of the mycotoxins were downloaded from PubChem. The name of the investigated mycotoxins and the PubChem compound identifier (CID) is listed in Table 4.

A home made *.php program was used to transform the *.sdf files in *.hin files. Another *.php program was used to prepare the structures for modeling. This program used the HyperChem, release 8, software to prepare the mycotoxins for modeling. A three-step protocol was implemented in this respect: ▪ obtain the 3D structure of mycotoxins (1st step); define the molecular mechanics model (AMBER) and optimize the molecular structure (AM1 method [27], Polak-Ribiere algorithm [28]) (2nd step); ▪ minimize the energy (3rd step).

Table 4. Mycotoxins and treir retention times

	Compound	CID	t_R (min)
Aflatoxins and their precursors			
1	Aflatoxicol I	104744	12.45
2	Aflatoxin B_1	14403	11.50
3	Aflatoxin B_2	2724360	10.33
4	Aflatoxin B_2 α	23648	6.60
5	Aflatoxin G_1	14421	10.16
6	Aflatoxin G_2	2724362	8.97
7	Aflatoxin G_2α	105095	5.00
8	Aflatoxin M_1	23236	7.21
9	Austocystin A	115307	21.57
10	Averufin	131171	25.65
11	5-Methoxysterigmatocystin	5379210	18.02
12	Demethyl-diacetyl-sterigmatocystin	151544	17.70
13	Methoxysterigmatocystin	5351257	15.03
14	Sterigmatocystin	5280389	18.91
15	Norsolorinic acid	25102	31.08
16	Parasiticol	90905	10.73
Trichothecenes			
1	Nivalenol	31829	1.27
2	Fusarenone X	31763	2.35
3	Deoxynivalenol	430147	1.54
4	3-Acetyldeoxynivalenol	104759	5.21
5	15-*O*-Acetyl-4-deoxynivalenol	16218854	5.10
6	Scirpentriol	73495	1.82
7	15-Acetoxyscirpenol	105023	7.40
8	Diacetoxyscirpenol	422111	11.28
9	3α-Acetyldiacetoxyscirpenol	11969469	15.56
10	Neosolaniol	72650	3.19
11	HT-2 Toxin	73050	13.69
12	T-2 Toxin	44575871	17.06
13	Acetyl-T-2 toxin	3034766	21.12
14	Trichodermin	20806	16.13
15	Trichodermol	114680	9.69
16	7-α-Hydroxytrichodermol	127380	2.59
17	Verrucarol	5660	2.89
18	4,15-Diacetylverrucarol	6451372	14.15
19	Trichothecin	260779	16.29
20	Trichothecolone	107974	3.63
21	Trichoverrol A	6440574	10.16

	Compound	CID	t_R(min)
Roquefortines			
1	Agroclavine-1	73484	17.00
2	Auranthine	130919	10.51
3	Aurantiamine	375594	10.49
4	Chanoclavine-I	5281381	8.59
5	Costaclavine	160462	17.00
6	Cyclopenin	73525	11.60
7	Cyclopenol	101201	6.20
8	Cyclopeptin	n.a.	12.05
9	Dihydroergotamine	10531	18.60
10	Elymoclavine	11051	5.34
11	Epoxyagroclavine-1	133945	10.00
12	Ergocristine	31116	25.10
13	Ergotamine	3251	19.60
14	Fumigaclavine C	173878	21.40
15	Marcfortine A	433181	19.59
16	Marcfortine B	n.a.	17.39
17	Meleagrin	6437857	18.90
18	Oxaline	6438440	21.60
19	Pyroclavine	3083578	14.81
20	Roquefortine C	21608802	20.50
21	Roquefortine D	n.a.	6.09
22	Rugulovasine A and B	115153	8.43
23	Secoclavine	n.a.	20.40
24	α-Ergocryptin	134551	19.20
Ochratoxins			
1	Ochratoxin α	107911	5.60
2	Ochratoxin B-methyl ester	609664	20.00
3	Ochratoxin B-ethyl ester	609665	19.41
4	Ochratoxin α-methyl ester	n.a.	16.16

n.a. = not available in PubChem (these mycotoxins were drawn with HyperChem v. 8.0).

The following properties were evaluated by HyperChem to be the regressors for modeling the retention times (experimental values were taken from ref. [29]) of the investigated mycotoxins: negative value of the total-energy (abbreviated as *negte*), dipole moment (*dm*), negative value of the scf-atom-energy (*negsae*), negative value of the scf-binding-energy (*negsbe*), scf-core-energy (*sce*), negative value of the scf-electronic-energy (*negsee*), heat-of-formation (*hf*), molecular weight (*MW*), surface area (*SA*), surface area grid (*SAG*), volume (*Vol*), hydration energy (HydE), octanol-water partition coefficient (*LogP*), refractivity (*Refr*) and polarizability (*Pol*). The corresponding abbreviations are listed at the bottom of Table 2.

Since the expression of dependence of the retention times vs. one ore more mycotoxin properties could be inversed or logarithmical, the search was extended by including the following regressors: LogP, InvlogP, HydE, InvHydE, Invhf, Pol, LnPol, LnRefr, Vol, InvVol, LnVol, SAG, MW, InvMW, LnMW, negte, dm, negsae, negsbe, sce, negsee, Invnegte, Invdm, Invnegsae, Invnegsbe, Invsce, Invnegsee, Lnnegte, Lndm, Lnnegsae, Lnnegsbe, Lnsce, and Lnnegsee (where `ln` is for natural logarithm and `inv` for inverse value).

A home made *.php program was developed in order to identify the best regression (the main criterion used was r^2 = max, where r^2 is determination coefficient) for every number of regressors (abbreviated as *x*, $1 \leq x \leq 10$). Other objective of the current research was to find the model of the dependence between the determination coefficient and the number of regressors. The SlideWrite software was used in order to identify non-linear relationships (a significance level of 5% was used).

ACKNOWLEDGEMENTS

This paper is dedicated to Prof. Simion GOCAN on the occasion of his 80th birthday.

REFERENCES

1. N.W. Turner, S. Subrahmanyam, S.A. Piletsky, *Anal. Chim. Acta,* **2009**, *632(2)*, 168.
2. J.L. Richard, *Int. J. Food Microbiol.,* **2007**, *119(1-2)*, 3.
3. H.S. Hussein, J. M. Brasel, *Toxicology,* **2001**, *167(2)*, 101.
4. M. Reverberi, A. Ricelli, S. Zjalic, A.A. Fabbri, C. Fanelli, *Applied Microbiology and Biotechnology,* **2010**, *87(3)*, 899.
5. W.P. Blount, *Journal of the British Turkey Federation,* **1961**, *9*, 55.
6. M.C. Lancaster, F.P. Jenkins, J.M. Philp, *Nature,* **1961**, *192*, 1095.
7. F.G. Peers, C.A. Linsell, *Ann. Nutr. Aliment.,* **1977**, *31*, 1005.
8. E. Karunasena, M.D. Larrañaga, J.S. Simoni, D.R. Douglas, D.C. Straus, *Mycopathologia,* **2010**, *170(6)*, 377.
9. B.B. Jarvis, J.D. Miller, *Appl. Microbiol. Biotechnol.,* **2005**, *66(4)*, 367.

10. A.A. El-Banna, J.I. Pitt, L. Leistner, *Systematic and Applied Microbiology,* **1987**, *10(1)*, 42.
11. M. Kokkonen, M. Jestoi, A. Rizzo, *International Journal of Food Microbiology,* **2005**, *99(2)*, 207.
12. D.L. Arnold, P.M. Scott, P.F. McGuire, *Food and Cosmetics Toxicology,* **1978**, *16(4)*, 369.
13. I.M.L.D. Storm, J.L. Sørensen, R.R. Rasmussen, K.F. Nielsen, U. Thrane, *Stewart Postharvest Review,* **2008**, *4(6)*, 1.
14. O.O.M. Iheshiulor, B.O. Esonu, O.K. Chuwuka, A.A. Omede, I.C. Okoli, I.P. Ogbuewu, *Asian Journal of Animal Sciences,* **2011**, *5(1)*, 19.
15. R.R. Paterson, N. Lima, *EXS,* **2010**, *100*, 31.
16. A. Pfohl-Leszkowicz, *Arhiv za Higijenu Rada i Toksikologiju,* **2009**, *60(4)*, 465.
17. A.K. Bingham, T.D. Phillips, J.E. Bauer, *J. Am. Vet. Med. Assoc.,* **2003**, *222(5)*, 591.
18. R. Krska, *Analytical and Bioanalytical Chemistry,* **2009**, *395(5)*, 1203.
19. G.H. Degen, *Journal of Veterinary Pharmacology and Therapeutics,* **2009**, *32 (SUPPL. 1)*, pp. 28.
20. C. Raghavender, B. Reddy, *World Mycotoxin Journal,* **2009**, *2(1)*, 23.
21. P. Khlangwiset, F. Wu, *Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess.,* **2010**, *27(7)*, 998.
22. F. Wu, P. Khlangwiset, *Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess.,* **2010**, *27(5)*, 658.
23. D.L. Park, T.C. Troxell, *Adv. Exp. Med. Biol.,* **2002**, *504*, 277.
24. D.M. Hawkins, *J. Chem. Inf. Comput. Sci.,* **2004**, *44(1)*, 1.
25. W.E. Steinmetz, C.B. Rodarte, A. Lin, *European Journal of Medicinal Chemistry,* **2009**, *44(11)*, 4485.
26. R. Khosrokhavar, J.B. Ghasemi, F. Shiri, *International Journal of Molecular Sciences,* **2010**, *11(9)*, 3052.
27. M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, *J. Am. Chem. Soc.,* **1985**, *107*, 3902.
28. B. Polak, G. Ribiere, *Rev. Fr. Imform. Rech. Oper.,* **1969**, *16*, 35.
29. K.F. Nielsen, J. Smedsgaard, *J. Chromatog. A,* **2003**, *1002*, 111.