Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 591–594
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Short communication

Volume learning algorithm significantly improved PLS model for predicting the estrogenic activity of xenoestrogens

Vasyl V. Kovalishyn [a,*], Vladyslav Kholodovych [a,b], Igor V. Tetko [a,c], William J. Welsh [b]

[a] *Institute of Bioorganic Chemistry and Petrochemistry, Kyiv, Murmanska 1, 02660, Ukraine*
[b] *University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, Piscataway, NJ, USA*
[c] *GSF – National Research Centre for Environment and Health, Institute for Bioinformatics, MIPS, Neuherberg, Germany*

Received 21 December 2006; accepted 12 March 2007
Available online 19 March 2007

Abstract

Volume learning algorithm (VLA) artificial neural network and partial least squares (PLS) methods were compared using the leave-one-out cross-validation procedure for prediction of relative potency of xenoestrogenic compounds to the estrogen receptor. Using Wilcoxon signed rank test we showed that VLA outperformed PLS by producing models with statistically superior results for a structurally diverse set of compounds comprising eight chemical families. Thus, CoMFA/VLA models are successful in prediction of the endocrine disrupting potential of environmental pollutants and can be effectively applied for testing of prospective chemicals prior their exposure to the environment.
© 2007 Elsevier Inc. All rights reserved.

Keywords: QSAR; CoMFA; Volume learning algorithm; Artificial neural network; PLS

1. Introduction

Among various 3D-QSAR approaches, comparative molecular field analysis [1,2] is undoubtedly the most popular method for an efficient modeling of the steric–electrostatic interactions of ligands. This technique is widely used in drug discovery, toxicology, environmental science, and materials science [3–5]. Numerous studies have shown that CoMFA protocols may be enhanced by implementing more accurate representations of the dataset, e.g., better alignment of the molecules or alternative choice of statistical method [6]. The commonly employed partial least squares (PLS) statistical analysis [7] achieves meaningful results if a linear correlation between the target activity and the CoMFA field variables exists; however, it is much less reliable in cases where this relationship is nonlinear. Although there are several studies describing nonlinear implementations of PLS [8–10], the particular form of nonlinearity in these applications is generally limited to quadratic terms and cross-terms of the input parameters. As a consequence, these nonlinear PLS models might lack accuracy in finding the proper relationship between the molecular structures and their activities. In an attempt to overcome this aforementioned drawback, new algorithms that implemented artificial neural networks (NN) have been proposed. A review on such approaches in QSAR studies can be found elsewhere [11,12]. Recently, we have introduced the volume learning algorithm (VLA) as an efficient tool for 3D-QSAR analysis [13]. The algorithm has been demonstrated to successfully correlate thousands molecular parameters representing electrostatic and steric properties of molecules with their biological activities in studies on cannabimimetic aminoalkyl indoles, *N*-benzylpiperidine analogs, etc. [13,14]. The current study, which extends our previous work [15], focuses on a series of estrogenic endocrine disrupting compounds (EDCs). We show that application of the VLA technique yields a significantly improved model compared with traditional CoMFA-PLS analysis and demonstrate that VLA can be applied to compounds with broad structural diversity.

2. Methods

2.1. Data set

Chemical structures and their normalized relative potency (RP) to an estrogen receptor were taken from our recently

* Corresponding author. Tel.: +380 44 5732592; fax: +380 44 5732552.
E-mail address: vkovalishyn@yahoo.com (V.V. Kovalishyn).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.03.005

Table 1
Observed and predicted activities (log(RP)) of compounds using PLS and VLA methods

Chemical name	Experimental values	VLA		PLS	
		LOO	Residuals	LOO	Residuals
17β-Estradiol	2.00	0.9	1.1	0.53	1.47
17β-Estradiol-3(β-D-glucuronide)	−0.50	−1.96	1.46	−1.15	0.65
17β-Estradiol-3-sulfate	−2.00	**−1.9**	**0.1**	−0.23	1.77
17α-Estradiol	0.72	**0.74**	**0.02**	1.18	**0.46**
Estriol	−0.20	0.54	0.74	1.03	1.23
Testosterone	−3.00	−1.99	1.01	−0.97	2.03
Androstenediol	−1.64	−2.76	1.12	−3.18	1.54
Dehydroepiandrosterone	−2.74	**−2.96**	**0.22**	−1.8	0.94
D-Norgestrel	−3.40	**−3.38**	**0.02**	−2.55	0.85
17α-Ethylnylstradiol	1.95	**1.68**	**0.27**	1.25	0.7
Mestranol	0.86	**1.11**	**0.25**	1.95	1.09
Diethylstilbestrol	1.87	**1.38**	**0.49**	1.17	0.7
Hexestrol	1.49	**1.79**	**0.3**	0.61	0.88
Dienestrol	1.40	0.11	1.29	0.67	0.73
Tamoxifen	−2.33	−1.32	1.01	−2.16	0.17
4-Hydroxytamoxifen	−2.14	−1.51	0.63	−2.52	**0.38**
α-Zearalenol	0.94	**0.67**	**0.27**	−0.53	1.47
β-Zearalenol	−1.18	−0.09	1.09	−0.28	0.9
α-Zearalanol (zeranol)	0.11	**0.29**	**0.18**	−0.44	0.55
β-Zearalanol	−0.34	**−0.79**	**0.45**	**0.09**	**0.43**
Coumestrol	−0.17	1.03	1.2	−0.89	0.72
Equol	−1.07	−1.58	0.51	**−0.78**	**0.29**
Daidzein	−2.89	−2.01	0.88	−1.79	1.1
Formononetin	−2.25	**−2.12**	**0.13**	−1.64	0.61
Genistein	−1.31	**−1.67**	**0.36**	−1.99	0.68
4-Nonylphenol	−2.66	−2.09	0.57	−1.89	0.77
4-Octylphenol	−2.52	−1.86	0.66	**−2.03**	**0.49**
4-*tert*-Octylphenol	−3.44	−2.14	1.3	−1.2	2.24
DDT	−4.52	−3.63	0.89	−3.08	1.44
o,*p*′-DDT	−3.96	**−4.17**	**0.21**	**−3.71**	**0.25**
o,*p*′-DDE	−4.40	**−4.25**	**0.15**	−2.95	1.45
2,3,7,8-Tetrachloro-dibenzo-*p*-dioxine	−0.59	**−0.58**	**0.01**	−4.25	3.66
4′-Chloro-4-biphenylol	−1.22	−0.38	0.84	−1.99	0.77
2′-Chloro-4-biphenylol	−2.43	**−2.27**	**0.16**	−1.74	0.69
2′,5′-Dichloro-4-biphenylol	−0.21	−1.78	1.57	**−0.67**	**0.46**
2′,4′,6′-Trichloro-4-biphenylol	0.00	−1.05	1.05	**−0.5**	**0.5**
2′,3′,4′,5′-Tetrachloro-4-biphenylol	−0.09	**−0.44**	**0.35**	**−0.45**	**0.36**
3,3′,5,5′-Tetrachloro-4,4′-biphenyldiol	−1.80	**−2.06**	**0.26**	**−1.49**	**0.31**
Bisphenol A	−2.30	−3.2	0.9	−4.17	1.87
Butylbenzyl-phthalate	−3.40	−1.1	2.3	−0.55	2.85
Estrone	0.98	0.4	0.58	**0.55**	**0.43**
Zearalenone	−0.59	**−0.77**	**0.18**	**−0.57**	**0.02**
Biochanin A	−2.04	**−2.09**	**0.05**	**−1.85**	**0.19**
Methoxychlor	−2.48	**−2.71**	**0.23**	**−2.84**	**0.36**
MAE			0.62		0.94

Values predicted with MAE < 0.5 log units are indicated in bold.

published paper [15]. RP was defined as 100 times the ratio of the concentration of 17β-estradiol (E2) giving 50% induction in β-galactosidase activity (EC50) and the EC50 of the tested compounds. Using this scheme, the RP of E2 equals 100 [15]. The initial data set comprised 44 compounds from 8 structurally diverse chemical families. This data set included 10 steroids, 5 synthetic estrogens, 2 antiestrogens, 5 lactones, 6 phytoestrogens, 3 alkylphenols, 11 organochlorines, and 2 ‘‘other’’ chemicals not belonging to any of these classes. The structures of the analyzed compounds and their activities are listed in Table 1.

2.2. Molecular modelling

Molecular structures for each of 44 compounds were modeled in Sybyl Version 7.1 (Tripos, Inc., St. Louis, MO) using the MMFF molecular mechanics force field and a conjugate gradient optimization method for energy minimization as described in our previous study [15]. Atomic partial charges were assigned according to the Gasteiger–Huckel procedure [16]. Briefly, molecular structures were constructed in Sybyl from a fragment database followed by energy-minimization to the putative low energy conformation. This

local minimum-energy structure was then subjected to further conformational search with respect to all rotatable (single) bonds in 10° increments and, after setting each torsion angle to its minimum-energy value, the molecule was energy minimized to the final global minimum-energy conformation. All calculations were performed on a Silicon Graphics Octane workstation running under the IRIX 6.5 operating system.

2.3. Field descriptor generation

Steric and electrostatic molecular fields were calculated using the standard methodology for CoMFA studies [15]. A three-dimensional grid with 2 Å nodes was generated to enclose preliminary aligned chemical structures, after which the steric (van der Waals) and electrostatic (Coulomb's Law) field descriptors were calculated for each molecule at all lattice points. To avoid the unfavorable influence of the sterically prohibited contacts between a probe atom and a molecule on the resulting field energies,the energy values were truncated to 30 kcal/mol. The CoMFA field descriptors for each individual molecule were then extracted for use as VLA input parameters.

2.4. Statistical analysis and model validation

The biological activity of the 44 compounds in the data set was correlated with the CoMFA generated steric and electrostatic fields using two statistical methods—partial least squares (PLS) regression [17] and the volume learning algorithm [13].

2.4.1. Partial least squares regression

PLS attempts to reduce the large number of steric–electrostatic descriptors to a few principal components (PCs) that are linear combinations of the original descriptors. The optimum number of PCs was determined by the leave-one-out (LOO) cross-validation procedure [17]. In this method, each compound was systematically excluded once from the training set, after which its activity was predicted by a model derived from the remaining compounds. After specifying an optimal number of PCs, the final PLS analysis was performed without cross-validation to generate a predictive QSAR model with a conventional correlation coefficient.

2.4.2. Volume learning algorithm

The volume-learning algorithm uses a recursive iterative application of a supervised feed-forward neural network (FFNN) together with an unsupervised self-organizing map (SOM) of Kohonen. The detailed description of the algorithm can be found elsewhere [13,14]. In brief, VLA partitions the input parameters classified into clusters by SOM and then employs the mean values of the clusters as an input for FFNN training. The supervised learning is achieved through the Associative Neural Network (ASNN) with one hidden layer [18]. During the refining stages of model construction, pruning algorithms [19,20] optimize the number of input parameters for ASNN training and estimate the statistical significance of clusters [13,14].

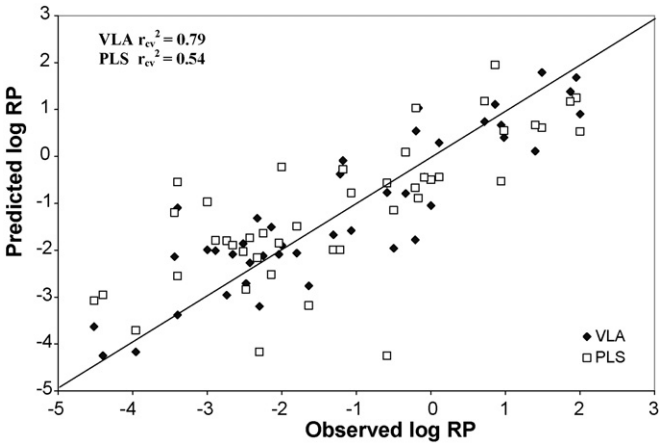

Fig. 1. Plot of CoMFA-predicted vs. observed values of log(RP) for the total data set of 44 compounds.

3. Results

The main goal of the current study was to compare and evaluate the proficiency of VLA and PLS for QSAR model generation. Values of the relative potency (RP) for the data set of compounds were converted to log(RP) values for building the 3D QSAR models.

The leave-one-out cross-validation procedure was used to test prediction ability of the methods. The mean absolute error (MAE) of the VLA and PLS predictions were 0.62 and 0.94, respectively (see Table 1). These results indicate that VLA provided higher prediction ability compared with the PLS approach (see Fig. 1). The Wilcoxon matched-pairs signed rank test [21] indicated a significant difference of MAE for both models at $p < 0.01$.

Analysis of the results revealed that the predicted log(PR) values were within 0.5 log unit absolute error for 22 of 44 molecules using VLA compared with only 14 molecules using PLS (Table 1). The VLA models also gave a better fit for the compounds in the training set. The absolute error exceeded 2 log units for only one molecule (butylbenzyl-phthalate) using VLA, but for four molecules (testosterone, 4-*tert*-octylphenol, 2,3,7,8-tetrachloro-dibenzo-*p*-dioxine and also butylbenzyl-phthalate) using PLS.

Although PLS is widely recognized as a reliable regression technique for linear models, it may provide low quality predictions for nonlinear models. In contrast, the VLA is inherently nonlinear which contributes to its higher prediction ability.

4. Conclusion

Our results confirmed that the appropriate choice of machine learning method for generation of 3D QSAR models is a crucial part of the analysis. In particular, nonlinear modeling approaches are preferable when nonlinear relationships exist between the dependent variables (target values) and the independent variables (calculated descriptors).

In the present application, the results indicate that the VLA models were significantly better than the corresponding PLS

models. In view of its inherent nonlinear nature, we expect that VLA will often lead to models with improved prediction ability compared with PLS. The current study suggests that VLA may serve as an efficient pre-screening technique for many applications, such as the identification of possible hazardous materials and environmental pollutants in risk assessment and the identification of new leads in drug discovery.

Acknowledgements

This work partially supported by ''Virtual Computational Chemistry Laboratory INTAS-INFO 00-0363'' grant. Support for this work has also been provided by the USEPA-funded Environmental Bioinformatics and Computational Toxicology Center (ebCTC), under STAR Grant number GAD R 832721-010. This work has not been reviewed by and does not represent the opinions of the funding agency.

References

[1] R.D. Cramer, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (COMFA). 1. Effect of shape on binding of steroids to carrier proteins, J. Am. Chem. Soc. 110 (1988) 5959–5967.
[2] R.D. Cramer, S.A. DePriest, D.E. Patterson, P. Hecht, The developing practice of comparative molecular field analysis, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, ESCOM, Leiden, 1993, pp. 443–485.
[3] A. Golbraikh, P. Bernard, J.R. Chretien, Validation of protein-based alignment in 3D quantitative structure–activity relationships with CoMFA models, Eur. J. Med. Chem. 35 (2000) 123–136.
[4] A. Palomer, J. Pascual, F. Cabre, M.L. Garcia, D. Mauleon, Derivation of pharmacophore and CoMFA models for leukotriene D(4) receptor antagonists of the quinolinyl(bridged)aryl series, J. Med. Chem. 43 (2000) 392–400.
[5] S.X. Zhang, J. Feng, S.C. Kuo, A. Brossi, E. Hamel, A. Tropsha, K.H. Lee, Antitumor agents. 199. Three-dimensional quantitative structure–activity relationship study of the colchicine binding site ligands using comparative molecular field analysis, J. Med. Chem. 43 (2000) 167–176.
[6] J. Polanski, R. Gieleciak, A. Bak, The comparative molecular surface analysis (COMSA) – A Nongrid 3D QSAR Method by a Neural Network and PLS – System: predicting pK_n values of benzoic and alcanoic acids, J. Chem. Inf. Comput. Sci. 42 (2002) 184–191.
[7] S. Wold, E. Johansson, M. Cocci, PLS—partial least squares projection to latent structures, in: H. Kubini (Ed.), 3D QSAR in Drug Design, ESKOM, Leiden, 1993, pp. 523–563.
[8] S. Wold, Non-linear partial least squares modelling II. Splin inner relation, Chemometr. Intell. Lab. Syst. 14 (1992) 71–84.
[9] A. Berglund, S. Wold, INLR, implicit non-linear latent variable regression, J. Chemom. 11 (1997) 141–156.
[10] A. Berglund, S. Wold, A serial extension of multiblock PLS, J. Chemom. 13 (1999) 461–471.
[11] S. Anzali, G. Barnickel, M. Krug, J. Sadowski, M. Wagener, J. Gasteiger, J. Polanski, The comparison of geometric and electronic properties of molecular surfaces by neural networks: application to the analysis of corticosteroid-binding globulin activity of steroids, J. Comput. Aided Mol. Des. 10 (1996) 521–534.
[12] S. Anzali, J. Gasteiger, U. Holzgrabe, J. Polanski, J. Sadowski, A. Teckentrup, M. Wagener, The use of self-organizing neural networks in drug design, Perspect. Drug Discovery Des. 9–11 (1998) 273–299.
[13] I.V. Tetko, V.V. Kovalishyn, D.J. Livingstone, Volume learning algorithm artificial neural networks for 3D QSAR studies, J. Med. Chem. 44 (2001) 2411–2420.
[14] V.V. Kovalishyn, I.V. Tetko, A.I. Luik, J.R. Chretien, D.J. Livingstone, Application of a neural network using the volume learning algorithm for quantitative study of the three dimensional structure–activity relationships of chemical compounds, Rus. J. Bioorg. Chem. 27 (2001) 267–277.
[15] S.J. Yu, S.M. Keenan, W. Tong, W.J. Welsh, Influence of the structural diversity of data sets on the statistical quality of three-dimentional quantitative structure–activity relationship (3D-QSAR) models: predicting the estrogenic activity of xenoestrogens, Chem. Res. Toxicol. 15 (2002) 1229–1234.
[16] J. Gasteiger, M. Marsill, Iterative partial equalization of orbital electronegativity a rapid access to atomic charges, Tetrahedron 36 (22) (1980) 3219–3228.
[17] S. Wold, C. Albano, W.J.I. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, M. Sjostrom, Multivariate data analysis in chemistry, in: B. Kowalski (Ed.), Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, The Netherlands, 1984 .
[18] I.V. Tetko, Neural network studies. 4. Introduction to associative neural networks, J. Chem. Inf. Comput. Sci. 42 (2002) 717–728. http://www.vcclab.org.
[19] I.V. Tetko, A.E.P. Villa, D.J. Livingstone, Neural network studies. 2. Variable selection, J. Chem. Inf. Comput. Sci. 36 (1996) 794–803.
[20] V.V. Kovalishyn, I.V. Tetko, A.I. Luik, V.V. Kholodovych, A.E.P. Villa, D.J. Livingstone, Neural network studies. 3. Variable selection in the cascade-correlation learning architecture, J. Chem. Inf. Comput. Sci. 38 (1998) 651–659.
[21] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C. The Art of Scientific Computing, Cambridge University Press, Cambridge, 2002.