ORIGINAL PAPER

COMPUTATIONAL MODELLING OF THE 5α –REDUCTASE INHIBITORS BASED ON THE MIA-SAR APPROACH AND DESIGN OF NEW COMPOUNDS

FAEZEH AMRI[1] , ALI NIAZI[2*], ATISA YAZDANIPOUR[2], TAHEREH MOMENI-ISFAHANI[1]

Manuscript received: 30.10.2019; Accepted paper: 26.11.2019; Published online: 30.12.2019.

Abstract. *5α –reductase 2 is an interesting pharmaceutical target for the treatment of several diseases, including prostate cancer, benign prostatic hyperplasia, male pattern baldness and hirsutism. Quantitative structure-activity relationship (QSAR) analysis has been carried out for the prediction of inhibitory activity of a set of 4-X-17-Y-4-Azaandrost-3-ones as 5α –reductase 2 inhibitors. Bi-dimensional images were applied to calculate some pixels and partial least squares (PLS) algorithm was applied to QSAR modelling of 5α –reductase inhibitors. In this paper, we surveyed the effect of variable selection by application of genetic algorithms (GAs) for the PLS model. The GAs is very helpful in the variable selection in modeling and selecting the subset of pixels with the low prediction error. Pre-processing methods such as orthogonal pixel correction (OPC) were also used to provide the suitable input for modeling. These models were applied to the prediction of the molecules inhibition, which were not in the modeling procedure. The resulted model showed a high predictive ability with the root mean square error of prediction (RMSEP) of 0.52, 0.35 and 0.94 for PLS, GA-PLS and OPC-GA-PLS models respectively. Furthermore, the proposed QSAR model with the OPC-GA-PLS method was developed to predict the inhibitory activity of the new compounds.*

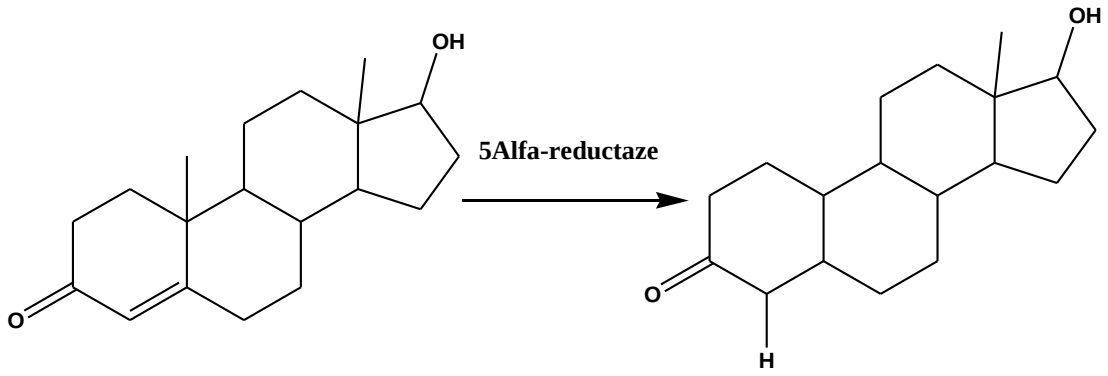
Keywords: *Quantitative structure-activity relationship, Multivariate image processing, Orthogonal pixel correction, Genetic Algorithms, Partial least squares, 5α –reductase.*

1. INTRODUCTION

One of the most important steroid hormones is testosterone that plays androgenic and anabolic roles in the human body. This hormone is responsible for development of libido and initiation of spermatogenesis at puberty, growth of male genitalia, building and maintaining muscle mass. Hormonopoiesis starts in the fetal period and increased in males at puberty. Normal circulating levels of androgens is necessary for the maintenance of structural function, growth and integrity of the prostate tissue [1]. A more active metabolite of testosterone is dihydrotestosterone that has various effects such as water differentiation and growth of prostate gland, male external genitalia and pubertal growth of facial and body hair and is effective in several human diseases like acne, hirsutism, male pattern baldness, benign

[1] Islamic Azad University, Arak Branch, Department of Chemistry, Arak, Iran.
[2] Islamic Azad University, Central Tehran Branch, Department of Chemistry, Tehran, Iran.
E-mail: Ali.Niazi@gmail.com.

prostatic hyperplasia (BPH) and prostate cancer (PCa) [2]. Testosterone (T) is converted to dihydrotestosterone (DHT) while enzyme 5α –reductase (5αR) is catalyzing this reduction (Fig. 1). So 5αR regulates DHT rate in the body. The 5α–reductase distributes widely in the brain, spinal cord, liver and skin.



testostrone(I) **dihydrotestostrone(DHT)**

Figure 1. 5α –reductase catalyzed conversion of testosterone to dihydrotestosterone.

Two dihydronicotinamide adeninedinuclecetioide phosphate (NADPH) - dependent isozymes of 5α -Reductase are known: Type 1 and Type 2. These types are different in pH optima so that the type I isozymes is active at pH 6–9 while type II is active at pH 5.5. The two isozymes also differ in sensitivity to inhibitors and distribution in the body. 5αR-1 can be found generally in hair follicles, sebaceous glands of the skin and liver, but 5aR-2 is found mostly in the seminal vesicles, prostate, genital skin, liver and epididymis [3]. The other difference is in the location of the gene structure while type I is located at 5p15 and type II is located at 2p22 although they had same gene structure.

As regards these enzymes play a basic function in many endocrine diseases such as prostate cancer, men alopecia, Hirsutism in women and Benign prostatic hyperplasia (BPH), inhibiting them leads to the treatment of such diseases. Many steroidal and non-steroidal inhibitors have reported to block the 5α-reductase as their performance depends on the amide ring which reacts with the active site in 5αR.

Most drugs currently used for the treatment of the mentioned diseases have side effects [4-6]. So designing a new 5α -Reductase inhibitor is inevitable. Using empirical methods for molecule design contains time-consuming steps such as screening, synthesis and unrational derivation, while may exhibit unsuccessful results [7]. So, in recent years, computer-aided drug discovery (CADD) is taken into consider and been widely used in medical chemistry. Computer-aided drug design uses computational approaches to discover, develop, and analyze drugs and similar biologically active molecules. One of these methods is Quantitative structure-activity relationship (QSAR).

QSAR is one of the most important applications of Chemometrics that gives useful information for the design of new compounds acting on a specific target and with desired properties. Nowadays, the QSAR technique was applied in pharmaceutical chemistry, toxicology, drug design, geology and remote sensing extensively [8-11]. QSAR-method reports a mathematical relationship between the chemical structure of the compounds and the physical, chemical or biological properties of them. Then surveying the reaction between ligand and receptor, designs novel molecule. Biological effects of compounds with similar physico-chemical properties are similar. For drug design, QSAR method makes a correlation between structural properties of potential drug candidates and their potency in inhibiting a specific biological function. Compared to different QSAR methods, multivariate image

analysis (MIA-QSAR) provided a rapid analysis and results as reliable as the most sophisticated methodologies available today, concurrently, it is inexpensive and facile to handle and predict any modeled response for a congeneric series of chemical structures without 3D alignment and conformational analysis. MIA-QSAR method was demonstrated by Esbensen and Geladi [12] and after that was used by researcher [11, 13-19]. In this technique, 2D images of pixel are indicative of topo-chemical properties of the compounds and build a model between these descriptors and y-block consisted of independent variables. MIA-QSAR is a non-invasive analysis, which avoids inessential time and costly procedures while handle a lot of information. Purpose of MIA-QSAR technique is to correlate several columns of independent variables to one column dependent variable, y.

In MIA-QSAR approach, various coordinate of pixels in the molecular drawing show structural changes and these changes were used for illustrating variance in bioactivity for a congeneric group in drug-like compounds. In modeling step, substitution pattern along with the congeneric series of compounds use and prediction of the bioactivities of similar compounds can be feasible. MIA-QSAR method Include the following steps, respectively:

- Drawing of molecule structures, image producing and alignment of them
- Denoising and then unfolding of images to a two-way array and descriptors generation
- Feature selection and regression modeling

One of the most important steps mentioned, is modeling. Different methods can be used for it such as multiple linear regression (MLR) [20, 21], partial least squares (PLS) [22] and artificial neural network (ANN). MLR has been applied to QSAR investigations widely, although it provides relatively poor accuracy. Moreover, MLR is successful when the amount of rows exceeds the number of columns. ANN exhibits quite enough accuracy in most cases, but there is a probability of overfitting the training data and may not be able to extrapolate suitable information of data consequently.

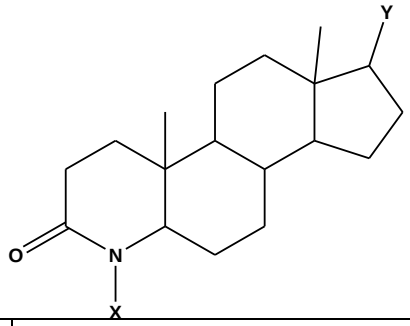
The partial least squares (PLS) regression analysis is the most commonly used method for this goal and is based on the factor analysis that chemically applied by Joreskog and Wold [23]. The PLS transformed matrices X and Y simultaneously to a sum of latent variable [24]. Applications of PLS have been used by several groups [9]. The success rate of modeling depends on the correct selection of molecular descriptors. So the method of variable selection is so much effective and should show the most information on activity variations and less collinearity among them. According to the satisfied results in pervious works by GA [11, 19, 22, 25] and also, lot numbers of descriptors in MIA, we deduced that Genetic Algorithms (GAs) can respond to this need [21]. GAs is a stochastic procedure that explained by Leardi and his associate [20]. This algorithm has simulated theory of evolution that believes the best genome has more chance of surviving and transferring by reproducing. GAs utilizes different fitness criterion and genetic functions [26, 27]. It was proved to preprocessing before PLS regression removes unwanted information and provides appropriate input for PLS, accordingly improve model quality. In 1998, Wold et al. introduced orthogonal signal correction (OSC) as preprocessing [28] .OSC was discussed in previous studies [29-32]. The main idea of OSC (in this study orthogonal pixel correction) is to delete extraneous and systematic variance from the matrix X which are unrelated and orthogonal to the matrix Y indeed structured noise in X [25].

The aim of this study is to generate a MIA-QSAR model for 4-X-17-Y-4-Azaandrost-3-ones that are 5α-reductase inhibitor Type 2 and subsequently predict their IC_{50} with OPC - GA-PLS model and finally, designing the new compounds based on this model. The half maximal inhibitory concentration (IC_{50}) is one of the parameters used to explain a measure of the potency of a substance in inhibiting a specific biological or biochemical function showing the minimum molar concentration leading to 50% of the enzyme inhibition.

2. MATERIALS AND METHODS

The IC_{50} Data of 4-X-17-Y-4-Azaandrost-3-ones for inhibitory were received from the literature [33]. The structure of each compound and its corresponding IC_{50} are presented in Table 1.

Table 1. Chemical structure and IC50 Data of 4-X-17-Y-4-Azaandrost-3-ones.



Compound	X	Y	Log1/c
1	H	OH	6.60
2	Me	OH	6.82
3*	H	COMe	7.30
4	Me	COMe	7.64
5	H	CHOHMe	7.35
6*	Me	CHOHMe	7.72
7	Me	$CH(Me)CH_2OH$	8.42
8	H	COO-	7.60
9	Me	COO-	7.43
10	Me	COOMe	8.12
11*	Me	$CONH_2$	7.77
12	Me	CONHMe	7.42
13	Me	$CONHC_2H_5$	7.92
14	H	$CONHC_2H_5$	8.20
15	ME	$CONHC_2H_5$	9.25
16	H	CONH(CH2)7Me	9.89
17	Me	$CONH(CH_2)_7Me$	9.68
18	Me	$CONH(CH_2)_8CH=CH(CH_2)_7Me$	7.70
19	Me	$CONH(CH_2)_2OH$	7.03
20	Me	$CONHCH_2CH(OMe)_2$	8.34
21	H	$CON(C_2H_5)_2$	7.39
22*	Me	$CON(C_2H_5)_2$	7.96
23	H	$CON(CHMe_2)_2$	7.96
24*	Me	$CON(CHMe_2)_2$	8.46
25	Me	$CON(C_8H_{17})_2$	7.48
26	Me	CO-morpholino	7.57
27	Me	$CH_2COOC_2H_5$	8.17
28	Me	$CH_2COO(C_2H_5)_2$	8.10
29	Me	$=CHCON(C_2H_5)_2$	8.60
30	Me	CHMeCOO-	8.77
31	Me	CHMeCOOMe	7.66
32	Me	$CHMeCON(C_2H_5)_2$	7.89
33	Me	$CHMe(CH_2)_2COO-$	8.44
34	Me	$CHMe(CH_2)_2CON(C_2H_5)_2$	8.47
35	Me	CHMeCN	8.12
36	H	$HOCHMeC_2H_5$	8.92
37*	Me	$COCHMeC_2H_5$	8.72
38	Me	NHCOMe	6.57

**Compounds selected for test set*

The data set is partitioned into the parts of the training and prediction set according to the Kennard-Stones algorithm [34, 35]. This algorithm, as one of the best procedure of building training and prediction sets in QSAR way, covers the total space occupied by the original data set.

2.1. HARDWARE AND SOFTWARE

The calculations were carried out with the Asus personal (8 GB RAM) equipping with the Windows 10 operating system and MATLAB (Version 10.0). The molecular structures were drawn using ChemOffice package (Version 2010). Kennard–Stones program was written in MATLAB according to the algorithm [34, 35].

2.2. PREPARATION OF CHEMICAL IMAGES AND DATA MATRIX

In the MIA-QSAR modeling, independent variables or descriptors which are the pixels of the 2D or 3D images, are correlated with the dependent variables or the specific property. The structures of each compound in Table 1, were drawn in same size with the ChemOffice first and next were turned into bitmaps in 100 × 74 pixels windows, then built congruent the common structural scaffold of all structures so that, they were fixed at the 30 × 70 coordinate that is visible in Fig. 2. This pixel was chosen as a reference in the alignment step. Each 2D image was read and transformed to double array in MATLAB. In this matrix, 0 and 756 respectively, correspond to black and white pixels according to RGB color composition. Alignment of the 38 images gave a three-way array of 38×100×74 dimension. In unfolding step, this array changed to a two-way array (matrix) of 38×22200 dimension. The columns that are exactly same to each other were eliminated.

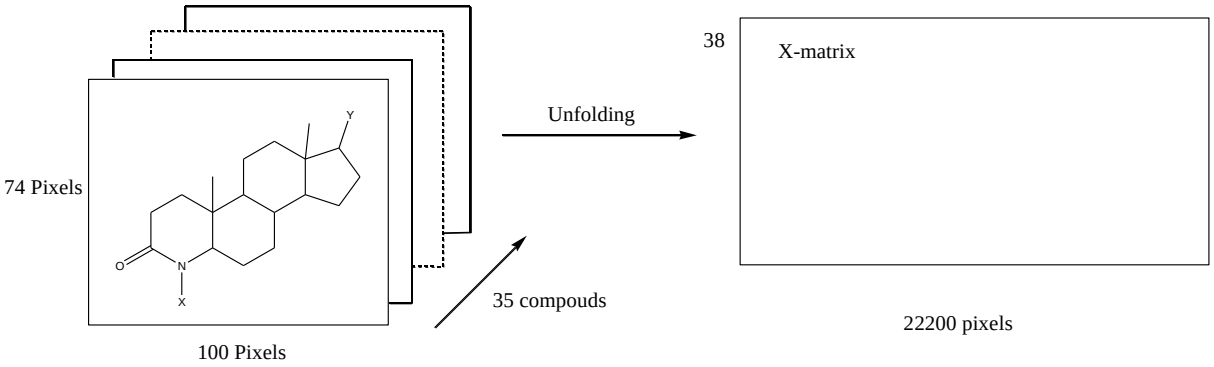


Figure 2. 2D images and unfolding step of the 38 chemical structures to give the X-matrix. The arrow in structure indicates the coordinate of a pixel in common among the whole series of compounds, used in the 2D alignment step.

3. RESULTS AND DISCUSSION

3.1. PRINCIPAL COMPONENT ANALYSIS OF THE DATA SET

For the purpose of initial data analysis, a Chemometrics tool, namely principal component analysis (PCA) was done on 38 compounds. The aim of PCA contains surveying

distribution in chemical space, studying clusters and exploring outliers. We founded from the PCA results that 72.09% of the overall variances related to three PCs as follows: PC1=35.85%, PC2=23.27% and PC3=12.97% (Fig. 3). As regards the most of the variables can be accounted for the first three PCs, their score plot is a reliable presentation of the spatial distribution of the points in the data set. It is not seen clear clustering between compounds in Fig. 3 which is very important in the generation of efficient QSAR models.

For regression analysis, the data set was divided into two classes, a training set (30 data) and a prediction set (8 data) according to Kennard-Stones algorithm. As shown in Fig. 3, these 2 groups were selected balancedly from the whole of the space of the principal components. The Kennard–Stone algorithm is a method for choosing a set of compounds from an original data set. That way, the first two molecules are selected from two distant points and the third molecule is selected the farthest from the first two molecules, etc.

3.2. PLS ANALYSIS

By exerting Partial least-square (PLS) modeling, which is a potent multivariate statistical tool, the relationship between the matrix of pixel as latent variable and activity matrix is developed. For this purpose, the Kennard - Stone algorithm is used to sort data into training and prediction sets and then the PLS model was built. In order to determinate the number of optimum latent variables in modeling, cross-validation (leave-one-out) method was employed. This method works based on minimum in prediction error variance or RMSECV, considering that RMSECV for the model is not significantly greater than the minimum RMSECV. In this way, only one compound at a time was deleted and the remaining of the training set was used in modeling, finally, activity of the eliminated compound was predicted by using this model. In the next steps, this was repeated for the other compounds. RMSEP values correspond the optimum number of factors which is shown in Table 2.

Table 2. Observation and calculation values of log(1/IC_{50}) using PLS, GA-PLS and OPC-GA-PLS models.

Number of compounds (Table 1)	Observation activity	PLS		GA-PLS		OPC-GA-PLS	
		Predicted	Error (%)	Predicted	Error (%)	Predicted	Error (%)
3	7.30	7.65	4.79	7.27	-0.41	7.39	1.32
6	7.72	7.45	-3.50	7.78	0.78	7.56	-2.07
11	7.77	7.60	-2.19	7.50	-3.47	7.67	-1.29
14	8.20	9.23	12.56	8.94	9.02	8.82	7.56
16	9.89	9.50	-3.94	9.44	-4.55	9.82	-0.71
22	7.96	7.38	-7.29	7.76	-2.51	7.48	-6.03
24	8.46	7.82	-1.76	9.19	8.63	8.36	-1.18
37	8.72	8.88	1.81	9.02	1.58	8.91	2.18
LVs		20		9		6	
Q^2		0.582		0.812		0.908	
R^2		0.682		0.842		0.926	
RMSEP		0.521		0.353		0.941	
RSEP (%)		0.401		0.182		0.130	

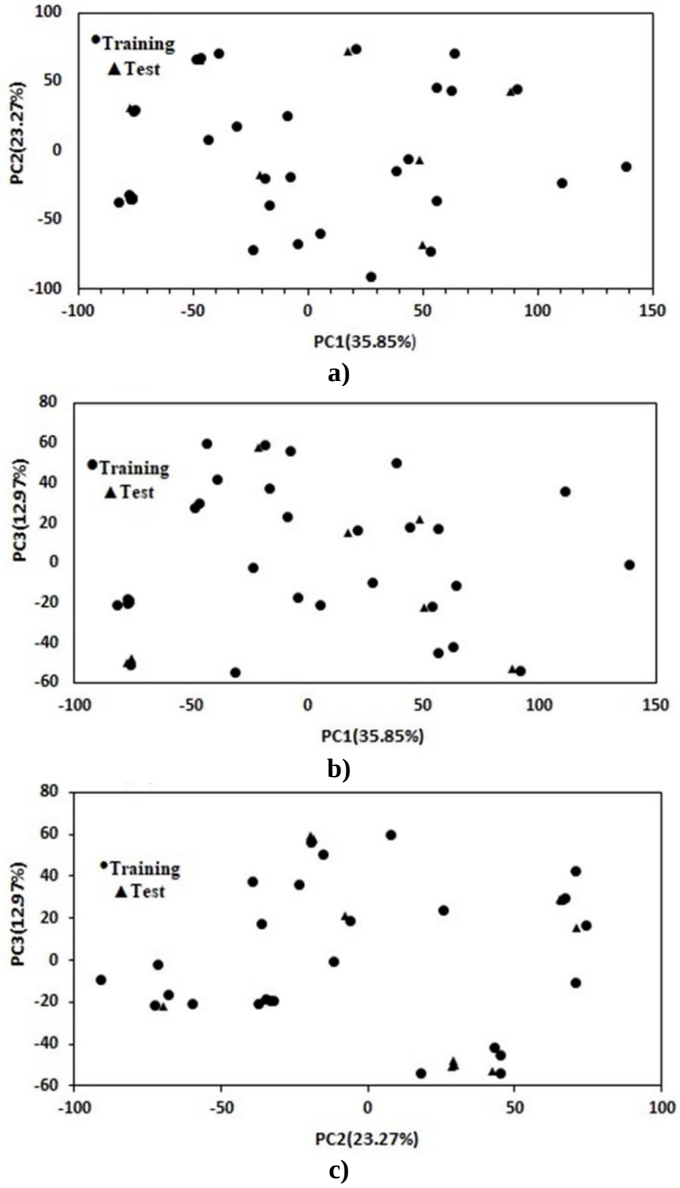


Figure 3. Principal components analysis of the 2D image descriptors for the data set; (A) PC1 versus PC2, (B) PC1 versus PC3 and (C) PC2 versus PC3.

3.3. GA-PLS MODELING

The more appropriate and effective the descriptor, the model becomes more precise, thus variable selection leads to robust model. In this study, the GA was run for descriptor selection. One of the most important characteristics of GA is that it investigates many possible solutions simultaneously that each one explores different regions in space of input variable [22]. The GA produces many random sets of variables defined by a chromosome. The number of genes or variables in each chromosome is equal to the number of descriptors. Each subset of chromosome is checked by its fitness to predict inhibitory activity values. After running the GAs, pixels were selected based on the best prediction of IC_{50}. The parameters of GAs used in this study are as follows: the probability of mutation 1 and 90% for crossover, number of runs are 100 and window size for smoothing is 3. 2565 pixels were selected among 22200 pixels by GAs method and then modeled by PLS. Table 2 shows

RMSEP values for this model. The present study appears the GAs can be a good method for pixel selection in image analysis. The range of selected pixel descriptors by GA is displayed in Fig. 4. Seeing this figure, it is obvious that the maximum structural effects are in a and b regions and these parts have a greater impact on the 5α –reductase inhibiton activity. Because that these parts were alterations by different different groups and these groups are effective on the 5α –reductase inhibition.

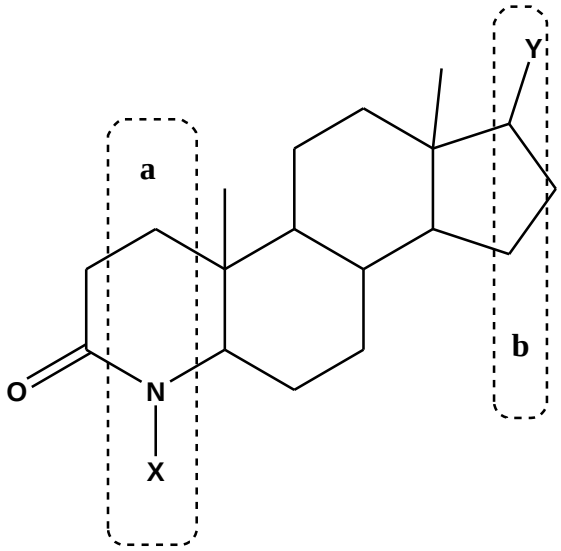


Figure 4. Selected regions by genetic algorithms.

3.4. OSC-GA-PLS MODELING

In order to eliminate systematic variation from the response pixel that is unrelated or orthogonal, to the property matrix *Y* and extraction of important information regarding the data, OPC algorithms were used. After preprocessing, GA was applied as descriptor selection and optimized variation and selection of the fitness values, and then calibration model provided by PLS regression. Table 2 shows the optimum number of RMSEP values. In order to estimate the models applied in this study (PLS, GA-PLS and OPC-GA-PLS), the inhibitory activities were tested with a set of tests. The results are seen in Table 2. It can be found that the OPC treated data give substantially lower RMSEP values than two other models.

3.5. MODEL VALIDATION AND PREDICTION OF INHIBITORY ACTIVITY

For the evaluation of the predictive ability of different models (PLS, GA-PLS and OPC-GA-PLS), the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) and cross validation coefficient (Q^2 and R^2) can be used:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_{pre}-y_{exp})^2}{n}} \quad (1)$$

$$RSEP(\%) = 100 \times \sqrt{\frac{\sum_{i=1}^{n}(y_{pre}-y_{exp})^2}{\sum(y_{exp})^2}} \quad (2)$$

$$Q^2_{abs} = 1 - \frac{\sum_Y (Y_{exp} - Y_{LOO})^2}{\sum_Y (Y_{exp} - \bar{Y}_{exp})^2} \quad (3)$$

$$R^2_{abs} = 1 - \frac{\sum_Y (Y_{exp} - Y_{pred})^2}{\sum_Y (Y_{exp} - \bar{Y}_{exp})^2} \quad (4)$$

For this object, prediction of inhibitory activity of 8 molecules (their structures are given in Table 1) was investigated, the statistical parameters were estimated (see Table 2) that showed the good statistical qualities.

4. MOLECULAR DESIGN

One of the drug design methods is computational approaches. As an application of the proposed method, we used OPC-GA-PLS model to predict the inhibitory activity of four new compounds on which biological tests were not performed yet. Finasteride is a potent type 2 selective 5α-reductase inhibitor having a promising therapeutic potential for the treatment of dependent diseases (Log1/c = 8.5). Based on this IC_{50} value, we designed new compounds. The inhibitory activity of tthese four compounds which were manually designed calculated by this proposed method and can be seen in Table 3.

Table 3. Chemical structures with the observed values of the inhibitory activity for the 5α –reductase.

No.	X	Y	Observed
1	H	$CONHCH(CH_3)_2$	8.29
2	CH_3	$CHCH_3(CH2)_2CONH(C_2H_5)$	6.78
3	H	$CONH(CH_2)_8CH{=}CH(CH_2)_7CH_3$	7.28
4	CH_3	$CHCON(CH_3)_2$	6.40

5. CONCLUSION

A satisfactory and robust MIA-QSAR/OPC-GA-PLS model was developed for 5α –reductase inhibitors, providing a simple 2D image-based approach for measurement of inhibitory. The rresults of the table showed the power of pixels in the prediction of inhibitory activity of 4-X-17-Y-4-Azaandrost-3-ones. The QSAR model developed in this study can provide a useful tool to predict the activity of new compounds and also to design new compounds with high activity.

Acknowledgements: *The authors gratefully acknowledge the support of this work by the Islamic Azad University, Arak Branch.*

REFERENCES

[1] Aggarwal, S. et al., *Steroids*, **75**(2), 109, 2010.
[2] Kumar, R. et al., *Med. Chem. Res.*, **22**(10), 4568, 2013.
[3] Fan, G.-j. et al., *Bioorganic Med. Chem. Lett.*, **11**(17), 2361, 2001.
[4] Faragalla, J. et al., *J. Mol. Graph. Model.*, **22**(1), 83, 2003.
[5] Frye, S. V., *Curr. Pharm. Design*, **2**(1), 59, 1996.
[6] Rasmusson, G. H. et al., *J. Med. Chem.*, **29**(11), 2298, 1986.
[7] Freitas, M. P., *Curr. Comput. Aided Drug Des.*, **3**(4), 235, 2007.
[8] Cui, W. Yan, X., *Chemom. Intell. Lab. Syst.*, **98**(2), 130, 2009.
[9] Goodarzi, M. Freitas, M. P., *Chemom. Intell. Lab. Syst.*, **96**(1), 59, 2009.
[10] Ioele, G. et al., *Talanta*, **79**(5), 1418, 2009.
[11] Sarkhosh, M. et al., *Chemom. Intell. Lab. Syst.*, **139**, 168, 2014.
[12] Geladi, P. Esbensen, K., *J. Chemom.*, **3**(2), 419, 1989.
[13] Akrami, A. Niazi, A., *Polycycl. Aromat. Comp.*, **37**(5), 442, 2017.
[14] Asadollahi-Baboli, M., *SAR QSAR Environ. Res.*, **24**(12), 1041, 2013.
[15] Duarte, M. et al., *SAR QSAR Environ. Res.*, **26**(3), 205, 2015.
[16] Goodarzi, M. de Freitas, M. P., *Mol. Simulat.*, **36**(4), 267, 2010.
[17] Khorshidi, N., Sarkhosh, M. Niazi, A., *J. Sci. Innov. Res.*, **3**(2), 189, 2014.
[18] Nunes, C. A. Freitas, M. P., *LWT-Food Sci. Technol.*, **51**(2), 405, 2013.
[19] Veyseh, S. et al., *J. Chil. Chem. Soc.*, **60**(3), 2985, 2015.
[20] Narasimhan, B. et al., *Bioorganic Med. Chem. Lett.*, **17**(21), 5836, 2007.
[21] Zareh, M. K., *Acta Chim. Slov.*, **50**, 259, 2003.
[22] Niazi, A., Bozorghi, S. J. Shargh, D. N., *Turk. J. Chem.*, **30**(5), 619, 2006.
[23] Jöreskog, K. G. Wold, H., *North-Holland, Amsterdam*, **68**, 108, 1982.
[24] Niazi, A. Ghasemi, N., *Annali di Chimica*, **97**(9), 845, 2007.
[25] Niazi, A. Azizi, A., *Turk. J. Chem.*, **32**(2), 217, 2008.
[26] Niazi, A. Leardi, R., *J. Chemom.*, **26**(6), 345, 2012.
[27] Leardi, R., in *Genetic algorithms in molecular modeling*, Elsevier, pp. 67, 1996.
[28] Wold, S. et al., *Chemom. Intell. Lab. Syst.*, **44**(1-2), 175, 1998.
[29] Guan, X. Liu, J., *Int. J. Pept. Res. Ther.*, 1, 2018.
[30] Heidari, A., *Insight Pharm. Res.*, **1**(1), 2016.
[31] Long, H. X. et al., *J. Chin. Chem. Soc.*, **57**(3A), 417, 2010.
[32] Pauli, E. D., Bruns, R. E. Scarminio, I. S., *Anal. Methods*, **8**(41), 7537, 2016.
[33] Kurup, A., Garg, R. Hansch, C., *Chem. Rev.*, **100**(3), 909, 2000.
[34] Daszykowski, M., Walczak, B. Massart, D., *Anal. Chim. Acta*, **468**(1), 91, 2002.
[35] Kennard, R. W. Stone, L. A., *Technometrics*, **11**(1), 137, 1969.