New Predictor Model for Classification Anti-breast cancer Compounds According to Multiple Parameters

Alejandra Aguirre-Crespo[1], Angel G. Arguello-Chan[1], Coraima I. Estrada-Domínguez[1], Francisco Aguirre-Crespo[1] and Francisco J. Prado-Prado [1*]

[1] *Biomedical Sciences Department, Health Sciences Division, University of Quintana Roo, UQROO, 77039, Mexico.*

Abstract. According to Global Health in 2013, it was estimated that there were 508 000 women deaths in the world in the year 2011 caused by breast cancer. Even though cancer can be treated with different treatments for example: immunotherapy, radiotherapy and chemotherapy surgical operation, this disease continues being a severe medical problem. For that reason it has to be found another methods for cancer treatment. The discovery of new drugs with better activity and less toxicity for the treatment of Breast Cancer is a goal of the major importance. In this sense, theoretical models as QSAR can be useful to discover new anti-breast cancer drugs. For this reason, we developed a new multi-parameter-QSAR (mp-QSAR) model to discover new drugs. However, almost all the computational models known focus in only one target or receptor. In this work, Breast cancer type 1 susceptibility protein, ATP-binding cassette sub-family G member 2, Human breast cancer cell lines, Peroxisome proliferator-activated receptor gamma/nuclear receptor coactivator 3, nuclear receptor coactivator 3 and STE20-related kinase adapter protein alpha were used as receptor inputs in the model. A linear technique like Linear Discriminant Analysis (LDA) is our statistical analysis, and we compared with others models to seek alternative multi-target models for inhibitors of some of these receptors. In so doing, we used as input Topological Indices, in specific Wiener, Barabasi and Harary indices calculated by Dragon software. These operators quantify the deviations of the structure of one drug from the expected values for all drugs assayed in different boundary conditions or parameters (type of receptor, type of assay, type of target, target mapping). The best model correctly classifies as active compounds 84.00 % and non-active compounds (99.06 %) in the training series. Overall training performance was 95.91%. Validation of the model was carried out by means of external predicting series. Overall predictability performance was 95.52%. By the first time, the present work reports the attempts to calculate within unified framework probabilities of new anti-Breast cancer drugs.

Keywords: Breast cancer; Multi-target receptors; Barabasi Topological Indices, Wierner Topological Indices, Harary Topological Indices, Breast cancer receptors.

***corresponding author:** PRADO PRADO, F; (fenol1@hotmail.com or fprado@uqroo.edu.mx), Biomedical Sciences Department, Health Sciences Division; University of Quintana Roo (UQROO).

1. INTRODUCTION

According to Global Health in 2013, it was estimated that there were 508 000 women deaths in the world in the year 2011 caused by this cancer. It has been found that breast cancer disease has led to an internationally increase in mortality. Even though cancer can be treated with different treatments for example: immunotherapy, radiotherapy and chemotherapy surgical operation, this disease continues being a severe medical problem. For that reason it has to be found another methods for cancer treatment. Between these diverse use of treatment, chemotherapy is the one that has cause the most impact in cancer (1).

As said before, the field in the research of better anti-cancer chemotherapies is getting bigger as time passes, as breast cancer is recognized as one of the main cancers that are causing feminine mortality. Chemoinfromatics is an area that has been widely used in these researches, making it as one of the most important fields. Basing on Chemoinformatics, there exists a multi-target model (mt) where by a big and heterogeneous database of several compounds is designed. In this model, the compounds are categorized as being active or inactive. Form these molecules, a portion of them were obtain in order to calculate there was a positive relationship and if it contributed to the blockage of breast cancer (2).

In this study, some proteins that were found to be related to breast cancer and used in this QSAR are Breast cancer type 1 susceptibility protein, Peroxisome proliferator-activated receptor gamma, STE20-related kinase adapter protein alpha, ATP-binding cassette sub-family G member 2, Human breast cancer cell lines, and Nuclear receptor coactivator 3 (3).

Some enzymes related to breast cancer were found to be useful in the treatment of this disease. If the expression of breast cancer susceptibility gene 1(BRCA1) is lost, breast cancer starts to progress. It has also been found that this gene product initiates cancerous cell migration. Gene expression analyses demonstrate that several of these proteomic hits are differentially expressed between early and advanced stage EOC thus suggesting clinical relevance of these proteins to disease progression (3).

The design of new enzyme inhibitors for the treatment of anticancer creates a main objective. From our point of view, QSAR techniques may be very helpful in this case. Unfortunately, some QSAR techniques predict new outcomes only for one specific assay. We can avoid this by developing a new Multi-target/Multiplexing QSAR models. These approaches are useful to process very large collections of compounds assayed against multiple molecular or cellular targets under different assay conditions (c_j) as is the case of ChEMBL (4, 5). This phase is significant for the Cheminformatics` future. QSAR models can foretell the results of the assay of different drugs for multiple targets. Nevertheless, QSAR models cannot foretell diverse results for a given sequence of targets when changed under a set of definite assay conditions for each target. Luckily, the new QSAR is not only useful for different targets but also to different multiplexing assay conditions (c_j) for all targets.

A diverse topological indices (TIs) of molecular graphs (G) can be used to speed up the procedure of codification of the molecular arrangement of drugs in Cheminformatics studies. Topological Indices (TIs), also charge transfer indices and their different variants, can be consider as useful in the field of Cheminformatics. For example, in the study of the HIV-1 RT inhibitory activity of thiazolidinones with different TIs, have been found to be accurate in predicting the activity of these same compounds (6). By the first time we used TI molecular descriptors developed by Wiener, Barabasi and Harary indices to develop one multi-target/multiplexing QSAR model for inhibitors of 6 different enzymes.

MATERIALS AND METHODS

1.1. Computational methods.

Techniques in Cheminformatics are capable to predict new drugs only for one specific illness, organism, assay, etc. In this work we evade this problem developing a new Multi-target QSAR model. These methods are powerful when we need to process very large collections of compounds or drugs assayed against multiple molecular or cellular targets in the different assay conditions or in different targets; as in the case of ChEMBL. A general data set composed of more than 17,000 drugs was downloaded from the public database ChEMBL (4, 5). This dataset includes drug number (N_d) = 16750 drugs and/or organic compounds previously assayed in different multi target assay conditions (c_j). Every drug evaluated in different r_t receptors were assigned to 1 out of 2 possible activity classes: active (C = 1) or non-active compounds (C = 0). One compound may lead to 1 or more statistical cases because it may give different outcomes (statistical cases) for alternative biological assays carried out in diverse sets of multiple conditions. In this work, we defined cj according to the ontology r_t => (a_u, c_j, r_t, t_e, s_x). The different conditions that may change in the dataset are: different: receptors (r_t), biological assays (a_u), molecular or cellular targets (t_e), or standard type of activity measure (s_x). Notably, multi-target QSAR models are able to predict the results of the assay of different drugs for multiple targets. Fortunately, the new class of mt-QSAR models applies not only to different targets but also to different multi target assay conditions (c_j) for all targets.. The different steps to develop our model are first, we calculate the molecular descriptors using D_i of a given i^{th} compound using one or more software for the generation of molecular descriptors. Next, we expand the raw dataset of molecular descriptors adding new variables with the form of Box-Jenkis Operators or moving averages $\Delta D_{ij} = D_i - \langle D_{ij} \rangle$. We can use both classes of descriptors to formulate linear and no-linear models.

Where, S_{ij} is a numerical score of the biological activity of the i^{th} compound measured under the j^{th} assay defined by the set of conditions c_j. In these models, the average $\langle D_{ij} \rangle = \langle D_i(c_j) \rangle$, used to calculate ΔD_{ij} values, is the average of the D_i for different compounds and do not runs over a time domain but over a set of molecular descriptors that obey a given limit condition c_j. (7-17). Last, we upload the input values in order in the Statistic or Machine Learning software to run different algorithms and seek different linear and non-linear models. In this work, we are going to use STATISTICA (7).

In total we analyzed N > 17000 statistical cases, which each one have been assayed in at least one out of the assays, Number (N_a) = 3 possible assays. For each one of these assays the dataset studied presents, for each drug, at least one out of Number Standart Types (N_s) = 17 standard types of biological activity measures in turn carried out in at least one out of Number Receptor N_r = 6 receptor. The values are reported in ChEMBL with three different levels of Curation Number (N_c) = 3 (expert, intermediate, or auto-curation level).

1.2 *Theoretical model*

In order to search the high-throughput mt-QSAR model we used the linear discrimant analysis (LDA) module of the software package STASTICA 6.0 (18). The model developed presented the general form.

$$\begin{aligned} S_i(m_j) &= b_0 + b_1 \cdot p(a_u) \cdot p(c_l) \cdot {}^{std}\mu_5^i + \sum_{j=2}^{4} b_j \cdot \Delta\mu_5^i(m_j) \qquad (1) \\ &= b_0 + b_1 \cdot p(a_u) \cdot p(c_l) \cdot {}^{std}\mu_5^i + \sum_{j=2}^{4} b_j \cdot \left({}^{std}\mu_5^i - \left\langle {}^{std}\mu_5^i(m_j) \right\rangle \right) \end{aligned}$$

Where, $S(m_j) = S(d_i, a_u, c_j, r_t, t_e, s_x)$ is a real-valued variable that scores the propensity of the drug to be active in multi target assays of the drug depending on the conditions selected cj. The statistical parameters used to corroborate the model were: Number of cases (N), Canonical Regression coefficient (R_c), Chi-square statistic (χ^2), and error level (p-level); which have to be < 0.05 (19).

2. RESULTS AND DISCUSSION

2.1. *Linear Multi-target model of drug-breast cancer receptor interactio*n.

The outcome of multi target breast cancer receptor inhibition depend both on drug structure and the set of assay conditions selected (c_j) (20). In this work, we report the first mt-QSAR model capable of predict whether a drug with a determined molecular structure may give or not a positive result in different multi target receptors r_t. The best mt-QSAR model found was the following:

$$\begin{aligned} S_{ij} = &-0.24 \cdot \langle D_1 \rangle - 0.003 \cdot \langle D_2(s_x) \rangle - 0.0033 \cdot \Delta.\ \{\ D_3(a_u)\ \} \qquad (2) \\ &+ 0.0023 \cdot \{\Delta D_4(t_e)\} - 3.84 \cdot \{\Delta D_5(t_e)\} + 0.38 \cdot \{\Delta D_6(t_e)\} \\ &+ 0.0087 \cdot \{\Delta D_7(r_t)\} - 13.22 \end{aligned}$$

$$N = 11548 \quad R_c = 0.787 \quad \chi^2 = 3667 \quad p-\text{level} < 0.001$$

The best model correctly classifies 667 out of 794 active compounds (84.00 %) and 2971 out of 2999 non-active compounds (99.06 %) in the training series. Overall training performance was 95.91%. Validation of the model was carried out by means of external predicting series, the validation correctly classifies 6083 out of 6148 non-active compounds (>99%) and 1325 out of 1607 active compounds (82.45%). Overall predictability performance was 95.52%, **see Table 1**. Where $S_{ij} = S(r_t, a_u, s_x, t_e)$ is a real-valued variable that scores the propensity of the drug to be active in multitarget pharmacological assays of the drug d_i carried out on the conditions The statistical parameters for the above equation in training are: Number of cases used to train the model (N), Canonical Regression Coefficient (R_c), Sensitivity (S_n), Specificity (S_p), and Accuracy (A_c) (7). The probability cut-off for this LDA model is ${}^{i}p_1(r_j) > 0.5 \Rightarrow C_i(c_j) = 1$. It means that the ith drug (d_i) predicted by the model with probability > 0.5 are expected to inhibit the enzyme present in the jth assays carry out under the given set of receptors r_j.

Table 1. Results of the classification model

Drug-Enzyme interaction	Training series			Statistical Parameter
	Positive	Negative	%	
Negative	28	2971	99.07	Sensitivity
Positive	667	127	84.01	Specificity
			95.91	Accuracy
Drug-Enzyme interaction	External validation series			Statistical Parameter
	Positive	Negative	%	
Negative	65	6083	98.9	Sensitivity
Positive	1325	282	82.45	Specificity
			95.5	Accuracy

[a] Sensitivity = Recall = True Positive/(True Positive + False Negative);
Specificity =True Negative / (True Negative + False Positive);
Accuracy = Ac = True Total / Total = (True Positive + True Negative)/Total

This linear equation presented good results both in training and external validation series with overall Accuracy in training series above 90%. The values of accuracy higher than 75% are acceptable for LDA models; according to previous reports (21-30). The reader should be aware that N here is not number of compounds but number of statistical cases. One compound may lead to 1 or more statistical cases because it may give different outcomes for alternative biological assays carried out in different organisms with different enzymes as targets (31). We used a big data from ChEMBL database, only using anti-breast cancer drugs and the model have good results. It is the first work on breast cancer in the mt-QSAR is used within the model using different enzymes, assays, and types of proteins.

3. CONCLUSION

It was possible to seek excellent predictors for DPIs using as input structural parameters of drugs and proteins. Theoretic multi-target QSAR models based on LDA and TI descriptors may become a useful tool in this sense. In this work, we developed a new LDA model using the Dragon descriptors, with a large data base using about 11000 different drugs obtained from the ChEMBL database. Is the first time that a mt-QSAR model is developed to study compounds with anticancer activity, study the main enzymes involved in breast cancer is paramount. In this sense, the model developed can help find more effective and less toxic drugs.

4. ACKNOWLEDGMENTS

The authors thank the sponsorship DCS-UQROO PROFOCIE (**P/PROFOCIE-2014-23MSU0140Z-09 DCS)** and sponsorships for research project PRODEP-Fortalecimiento Cuerpos Académicos 2012.

REFERENCES

1. Hajmohamad Ebrahim Ketabforoosh S, Amini M, Vosooghi M, Shafiee A, Azizi E, Kobarfard F. Synthesis, evaluation of anticancer activity and QSAR study of heterocyclic esters of caffeic Acid. Iranian journal of pharmaceutical research : IJPR. 2013 Fall;12(4):705-19.
2. Abifadel M, Pakradouni J, Collin M, Samson-Bouma ME, Varret M, Rabes JP, et al. Strategies for proprotein convertase subtilisin kexin 9 modulation: a perspective on recent patents. Expert opinion on therapeutic patents. [Research Support, Non-U.S. Gov't
Review]. 2010 Nov;20(11):1547-71.
3. Apostoli AJ1, Roche JM2, Schneider MM3, SenGupta SK4, Di Lena MA5, Rubino RE6, et al. Opposing roles for mammary epithelial-specific PPARγ signaling and activation during breast tumour progression.

4. Riera-Fernandez I, Martin-Romalde R, Prado-Prado FJ, Escobar M, Munteanu CR, Concu R, et al. From QSAR models of Drugs to Complex Networks: State-of-Art Review and Introduction of New Markov-Spectral Moments Indices. Curr Top Med Chem. 2012 Feb 14.

5. Prado-Prado F, Garcia-Mera X, Escobar M, Sobarzo-Sanchez E, Yanez M, Riera-Fernandez P, et al. 2D MI-DRAGON: a new predictor for protein-ligands interactions and theoretic-experimental studies of US FDA drug-target network, oxoisoaporphine inhibitors for MAO-A and human parasite proteins. Eur J Med Chem. [Research Support, Non-U.S. Gov't]. 2011 Dec;46(12):5838-51.

6. Prabhakar YS, Rawal RK, Gupta MK, Solomon VR, Katti SB. Topological descriptors in modeling the HIV inhibitory activity of 2-aryl-3-pyridyl-thiazolidin-4-ones. Comb Chem High Throughput Screen. 2005 Aug;8(5):431-7.

7. Hill T, Lewicki P. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining. Tulsa: StatSoft; 2006

8. Box GEP, Jenkins GM. Time series analysis: Holden-Day; 1970.

9. Luan F, Cordeiro MN, Alonso N, Garcia-Mera X, Caamano O, Romero-Duran FJ, et al. TOPS-MODE model of multiplexing neuroprotective effects of drugs and experimental-theoretic study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases. Bioorg Med Chem. 2013 Apr 1;21(7):1870-9.

10. Tenorio-Borroto E, Garcia-Mera X, Penuelas-Rivas CG, Vasquez-Chagoyan JC, Prado-Prado FJ, Castanedo N, et al. Entropy Model For Multiplex Drug-Target Interaction Endpoints Of Drug Immunotoxicity. Curr Top Med Chem. 2013 Jul 24.

11. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. A ligand-based approach for the in silico discovery of multi-target inhibitors for proteins associated with HIV infection. Molecular bioSystems. [Research Support, Non-U.S. Gov't]. 2012 Aug;8(8):2188-96.

12. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Chemoinformatics in anti-cancer chemotherapy: Multi-target QSAR model for the in silico discovery of anti-breast cancer agents. Eur J Pharm Sci. Netherlands: 2012 Elsevier B.V; 2012. p. 273-9.

13. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. In Silico Discovery and Virtual Screening of Multi-Target Inhibitors for Proteins in Mycobacterium tuberculosis. Comb Chem High Throughput Screen. Netherlands2012. p. 666-73.

14. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Predicting multiple ecotoxicological profiles in agrochemical fungicides: a multi-species chemoinformatic approach. Ecotoxicol Environ Saf. United States: 2012 Elsevier Inc; 2012. p. 308-13.

15. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. Bioorg Med Chem. England: 2012 Elsevier Ltd; 2012. p. 4848-55.

16. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. Bioorg Med Chem. [Research Support, Non-U.S. Gov't]. 2011 Nov 1;19(21):6239-44.

17. Gonzalez-Diaz H, Dea-Ayuela MA, Perez-Montoto LG, Prado-Prado FJ, Aguero-Chapin G, Bolas-Fernandez F, et al. QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new Leishmania infantum protein. Molecular diversity. [Research Support, Non-U.S. Gov't]. 2010 May;14(2):349-69.

18. StatSoft.Inc. STATISTICA (data analysis software system), version 6.0, www.statsoft.com.Statsoft, Inc. 6.0 ed2002.

19. Van Waterbeemd H. Chemometric methods in molecular design. Manhnhold R, Krogsgaard-Larsen P, Timmerman H, Van Waterbeemd H, editors. New York: Wiley-VCH; 1995.

20. Gerets HH, Dhalluin S, Atienzar FA. Multiplexing cell viability assays. Methods Mol Biol. 2011;740:91-101.

21. Patankar SJ, Jurs PC. Classification of inhibitors of protein tyrosine phosphatase 1B using molecular structure based descriptors. J Chem Inf Comput Sci. 2003 May-Jun;43(3):885-99.

22. Garcia-Garcia A, Galvez J, de Julian-Ortiz JV, Garcia-Domenech R, Munoz C, Guna R, et al. New agents active against Mycobacterium avium complex selected by molecular topology: a virtual screening method. J Antimicrob Chemother. 2004 Jan;53(1):65-73.

23. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castanedo N, et al. Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic. Bioorg Med Chem. 2005 Feb 15;13(4):1005-20.

24. Marrero-Ponce Y, Machado-Tugores Y, Pereira DM, Escario JA, Barrio AG, Nogal-Ruiz JJ, et al. A computer-based approach to the rational discovery of new trichomonacidal drugs by atom-type linear indices. Curr Drug Discov Technol. 2005 Dec;2(4):245-65.

25. Casanola-Martin GM, Marrero-Ponce Y, Khan MT, Ather A, Sultan S, Torrens F, et al. TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: evaluation of different classification model combinations using bond-based linear indices. Bioorg Med Chem. 2007 Feb 1;15(3):1483-503.

26. Casanola-Martin GM, Marrero-Ponce Y, Tareq Hassan Khan M, Torrens F, Perez-Gimenez F, Rescigno A. Atom- and bond-based 2D TOMOCOMD-CARDD approach and ligand-based virtual screening for the drug discovery of new tyrosinase inhibitors. J Biomol Screen. 2008 Dec;13(10):1014-24.

27. Casanola-Martin GM, Marrero-Ponce Y, Khan MT, Khan SB, Torrens F, Perez-Jimenez F, et al. Bond-based 2D quadratic fingerprints in QSAR studies: virtual and in vitro tyrosinase inhibitory activity elucidation. Chemical biology & drug design. [Research Support, Non-U.S. Gov't]. 2010 Dec;76(6):538-45.

28. Rodriguez-Soca Y, Munteanu CR, Dorado J, Pazos A, Prado-Prado FJ, Gonzalez-Diaz H. Trypano-PPI: A Web Server for Prediction of Unique Targets in Trypanosome Proteome by using Electrostatic Parameters of Protein-protein Interactions. Journal of Proteome Research. 2010 FEB 2010;9(2):1182-90.

29. Gonzalez-Diaz H, Muino L, Anadon AM, Romaris F, Prado-Prado FJ, Munteanu CR, et al. MISS-Prot: web server for self/non-self discrimination of protein residue networks in parasites; theory and experiments in Fasciola peptides and Anisakis allergens. Molecular Biosystems. 2011 2011;7(6):1938-55.

30. Riera-Fernandez P, Munteanu CR, Escobar M, Prado-Prado F, Martin-Romalde R, Pereira D, et al. New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks. Journal of theoretical biology. [Research Support, Non-U.S. Gov't]. 2012 Jan 21;293:174-88.

31. Martinez-Romero M, Vazquez-Naya JM, Rabunal JR, Pita-Fernandez S, Macenlle R, Castro-Alvarino J, et al. Artificial intelligence techniques for colorectal cancer drug metabolism: ontology and complex network. Curr Drug Metab. 2010 May;11(4):347-68.