https://dx.doi.org/10.21577/0103-5053.20200112

J. Braz. Chem. Soc., Vol. 31, No. 10, 2114-2119, 2020
Printed in Brazil - ©2020 Sociedade Brasileira de Química

Virtual Screening of Secondary Metabolites of the Family Velloziaceae J. Agardh with Potential Antimicrobial Activity

Anderson A. V. Pinheiro,[a] Renata P. C. Barros,[a] Edileuza B. de Assis,[a] Mayara S. Maia,[a] Diego I. A. F. de Araújo,[a] Kaio A. Sales,[a] Luciana Scotti,[a,b] Josean F. Tavares,[a] Marcus T. Scotti[a] and Marcelo S. da Silva*,[a]

[a] *Programa de Pós-Graduação em Produtos Naturais e Sintéticos Bioativos,
Universidade Federal da Paraíba, 58051-900 João Pessoa-PB, Brazil*

[b] *Gestão de Ensino e Pesquisa, Hospital Universitário Lauro Wanderley,
Universidade Federal da Paraíba, 58050-585 João Pessoa-PB, Brazil*

The objective of this work was to carry out a bibliographic survey of secondary metabolites isolated from the Velloziaceae family, creating a bank of compounds. After the bank was created, four prediction models for potentially active compounds against pathogenic microorganisms (*Candida albicans*, *Escherichia coli*, *Pseudomonas aeruginosa* and *Salmonella* sp.) were obtained trying to identify which metabolites would be more active against the strains. Four sets of compounds with known activity for microorganisms were selected for the construction of predictive models from the CHEMBL database. Another bank with 163 unique molecules isolated from the Velloziaceae family was built. The Volsurf+ v.1.0.7 software obtained the molecular descriptors and Knime 3.5 generated the *in silico* model. The performances of the internal and external tests were also analyzed. The study contributed through the virtual screening of a bank of metabolites to select several compounds with potential antimicrobial activity, highlighting the biflavonoid amentoflavone which showed potential activity against the four strains.

Keywords: amentoflavone, virtual screening, antibacterial activity, antifungal activity, Velloziaceae

Introduction

Natural products have been used historically for the treatment of various diseases, where medicinal plants act as an important resource for the recovery, cure and prevention of numerous diseases.[1] Thus, their use as a target for the discovery and/or obtaining of new drugs, whether in their entire form or in isolated compounds, is currently emphasized and, in data, it is observed that more than 70% of a total of 1562 new drugs approved by the Food and Drug Administration (FDA, 1981-2014) are of natural origin.[2,3]

The Velloziaceae family is native and not endemic to Brazil, where it currently comprises five genera (*Acanthochlamys*, *Barbacenia*, *Barbaceniopsis*, *Vellozia* and *Xerophyta*) and about 274 species,[4,5] inhabiting arid, rocky and elevated places.[6] The vast majority of species are distributed in Neotropical America (*Barbacenia*, *Barbaceniopsis* and *Vellozia*), others occur in Africa, Madagascar and the Arabian Peninsula (*Xerophyta* and *Vellozia*) and one in China (*Acanthochlamys*).[7] As for ethnopharmacological use, the aerial parts of some species of the family are used as anti-inflammatory, anti-rheumatic, treatment of bruises and bone fractures (topical use) and infections.[8,9] Besides this information, studies of phytochemicals of several species of the family are observed in the literature, as well as limited *in silico* and pharmacological studies of compounds used.

Bacterial resistance to more traditional antimicrobials is one of the biggest and most considerable obstacles to public health, where, according to the World Health Organization (WHO), *Escherichia coli*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, *Neisseria gonorrhoeae*, *Chlamydia trachomatis* and *Treponema pallidum*, are notorious examples of microorganisms that have been showing resistance to antimicrobials. Thus, there is a need for governments to encourage the development of new low-cost antibiotics adapted to the global need.[10-12]

*e-mail: marcelosobral@ltf.ufpb.br

In this perspective of obtaining compounds and envisioning their pharmacological potential, the use of computational methods in order to carry out the virtual screening of bioactive substances has been widely used. The search consists in selecting compounds with the computer axis from data in a database with a large number of molecules for diseases and contributing to the advancement in the planning of medicines, reduction of time, costs and animals in research.[13-15]

Thus, in this study a bibliographic survey of secondary metabolites isolated from the Velloziaceae family was carried out, creating a bank of compounds. After the bank was created, four prediction models for potentially active compounds against pathogenic microorganisms (*Candida albicans*, *Escherichia coli*, *Pseudomonas aeruginosa* and *Salmonella* sp.) were obtained trying to identify which metabolites would be more active against the strains.

Methodology

Computational chemistry

Database

From the ChEMBL database, four sets of chemical structures with known activity for microorganisms were selected: *Candida albicans*, *Escherichia coli*, *Pseudomonas aeruginosa* and *Salmonella* sp., for building predictive models. The details of each set are described in Table 1. The compounds were classified from pMIC_{50} ($-\log\text{MIC}_{50}$) (where pMIC_{50} is the planktonic minimum inhibitory concentration or 50%); emphasizing that the MIC_{50} represents the minimum concentration necessary for a 50% inhibition of the studied microorganisms. Another database of isolated molecules of the Velloziaceae family was built from a literature review of this family, with a total of 196 botanical occurrences and 163 unique molecules.

SMILES codes were used for all structures as input data for Marvin.[16] Standardizer software[17] was also used, which converts the various chemical structures into personalized canonical representations. This standardization is of paramount importance to create libraries of consistent compounds, in addition to obtaining the structures in canonical forms, adding hydrogens, flavoring, generating the 3D and saving the compounds in SDF format.

Volsurf descriptors

Molecular descriptors were used to predict biological and physicochemical properties of the molecules in the four databases. The calculation of the descriptors was generated when the molecules were transformed into a molecular representation that allows mathematical treatment.

The Volsurf+ v.1.0.7 software[18] has the ability to calculate 128 molecular descriptors, using molecular interaction fields (MIFs) through N1 probes (nitrogen-hydrogen starch hydrogen bond donor), O (hydrogen bond acceptor), OH (water) and DRY (hydrophobic probe) and also calculation of non-MIF-derived descriptors.

Prediction model

Knime 3.5 software[19] was used to perform the analyses and generate the model *in silico*. The banks of molecules with the calculated descriptors were imported from the Dragon software,[20] and for each one, the data were divided using a "Partitioning" tool with the option of "Stratified sample", separating in Training and Testing, representing 80 and 20% of all compounds, respectively, where they were randomly selected, but maintaining the same proportion of active and inactive substances, in both databases.

For internal validation, cross-validation was used, where 10 stratified groups were selected, randomly selected, but distributed according to the activity variable in all validation groups. With the selected descriptors, the model was generated using the training set applying the random forest (RF) which is an algorithm for building decision trees,[21] used in WEKA.[22] 100 forests and 1 random seed were the selected parameters for build the RF models.

The performance of the models' internal and external tests were analyzed for sensitivity (true positive rate, that is, the active rate), specificity (true negative rate, that is, the inactive rate) and accuracy (general predictability). In addition, the sensitivity and specificity of the receiver

Table 1. ChEMBL databases

Database against microorganisms	Total chemical structures	Active molecule	Inactive molecule	ChEMBL ID
Candida albicans	10436	($\text{pMIC}_{50} \geq 4.46$)	($\text{pIC}_{50} < 4.46$)	ChEMBL366
Escherichia coli	982	486 ($\text{pIC}_{50} \geq 5.00$)	496 ($\text{pIC}_{50} < 5.00$)	ChEMBL354
Pseudomonas aeruginosa	10693	($\text{pIC}_{50} \geq 5.00$)	143 ($\text{pIC}_{50} < 5.00$)	ChEMBL348
Salmonella enterica	316	129 ($\text{pIC}_{50} \geq 5.00$)	187 ($\text{pIC}_{50} < 5.00$)	ChEMBL613762

pMIC_{50}: planktonic minimum inhibitory concentration or 50%; pIC_{50}: planktonic inhibitory concentration or 50%.

operating characteristic (ROC) curve was used to describe the true performance of the model, with more clarity than precision.

The model was also analyzed by the Matthews' coefficient,[23] a way of globally evaluating the model from the results obtained from the confusion matrix. The Matthews' correlation coefficient (MCC) is, in essence, a correlation coefficient between observed and predictive binary classifications. It results in a value between −1 and +1, where a coefficient of +1 represents a perfect forecast, 0 is nothing more than a random forecast and −1 indicates total disagreement between forecast and observation.

Matthews' correlation coefficient can be calculated from the following formula:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (1)$$

In this equation, TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives.

The applicability domain based on Euclidean distances was also used in order to signal compounds in the test set for which predictions may be unreliable. Similarity measurements are used to define the model's applicability domain based on Euclidean distances between all training, test and virtual screening compounds. The distance of a compound from a test compound to its closest neighbor in the training set is compared to the predefined limit of applicability domain, if the similarity is beyond that limit, the prediction is considered unreliable.[24]

Results and Discussion

The secondary metabolite data set was composed of a total of 196 botanical occurrences and 163 different chemical compounds, from 34 species of the Velloziaceae family (genera *Vellozia*, *Acanthochlamys* and *Barbacenia*). It was identified that although several species make up the family, few have phytochemical and/or pharmacological studies, predominantly the isolation of diterpenes (109), flavonoids (21), triterpenes (21), steroids/glycosylated steroids (3), biflavonoids (2), other classes (7). This data set is available in SistematX.[25]

The generated models obtained excellent performances, with an accuracy greater than 75%. What also corroborates with these data are the high indexes of the MCC, thus informing the good prediction rate of the models (Table 2).

Looking at the values of the ROCs curves of the models, we see that they all have a high probability of selecting truly positive compounds, that is, with a low probability of classifying inactive compounds as active. The area under the curve is greater than 0.83, remembering that a perfect model has an area under the curve equal to 1 (Figure 1).

For the models of *C. albicans*, *E. coli* and *P. aeruginosa*, only the flavonoid kaempferol 3-*O*-(3",6"-di-*O*-*E*-*p*-coumaroyl)-β-*D*-glucopyranoside was outside the scope of application. Among the remaining 162 molecules that remained within the domain, 86 were classified as likely to be active ranging between 51 and 76% in the *C. albicans* model, 26 in the *E. coli* model with a probability between 50 and 78% and only 10 molecules in the model of *P. aeruginosa* with probability varying between 52 and 62%. The molecules with the greatest potential to be active for these models are described in Table 3.

Some studies have reported the use of quantitative structure-activity relationship (QSAR) models to select molecules with potential antimicrobial activity. Trush *et al.*[26] used three types of classification models; the random forest (WEKA-RF), k-nearest neighbors and associative neural networks to select potent inhibitors against *C. albicans*. In cross-validation, the models achieved a corresponding predictive rate of 81-90%. The experimental results confirmed the predictive power of the models with the selection of the compound 1,3-oxazol-4-yl (triphenyl) phosphonium. The same predictive ability was also observed in the study by Hodyna *et al.*,[27] where they used models identical to the previous study. The results of the

Table 2. Summary of cross-validation results and model tests using the random forest (RF) algorithm

Model		Specificity	Sensitivity	Accuracy	PPV	NPV	MCC
Candida albicans	external test	0.90	0.75	0.84	0.81	0.85	0.71
	validation	0.89	0.72	0.82	0.80	0.84	0.70
Escherichia coli	external test	0.78	0.78	0.78	0.77	0.78	0.64
	validation	0.77	0.72	0.75	0.75	0.74	0.60
Pseudomonas aeruginosa	external test	0.87	0.72	0.80	0.82	0.80	0.67
	validation	0.86	0.72	0.80	0.81	0.80	0.67
Salmonella enterica	external test	0.87	0.75	0.82	0.80	0.83	0.70
	validation	0.87	0.73	0.81	0.80	0.82	0.68

PPV: positive predictive value; NPV: negative predictive value; MCC: Matthews' correlation coefficient.

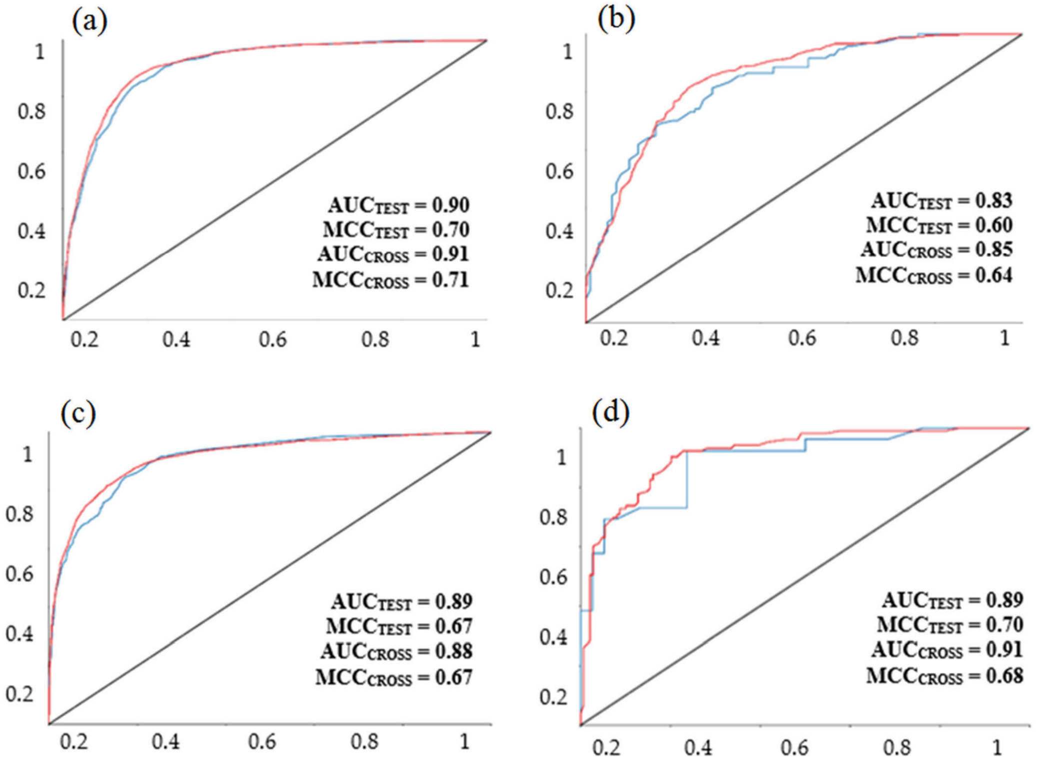

Figure 1. ROC curve of each model. True positives *versus* false positives, generated for the selected RF models for cross-validation and test sets: (a) *Candida albicans*; (b) *Escherichia coli*; (c) *Pseudomonas aeruginosa* and (d) *Salmonella enterica*. AUC = value of the area under the curve; MCC = Matthews' correlation coefficient.

Table 3. Secondary metabolites selected with the highest probability of active potential for the *C. albicans*, *E. coli*, *P. aeruginosa* and *S. enterica*

Model	Probability of active potential	Compound name
C. albicans	0.76	velloquercetin
	0.75	betulonic acid
	0.75	velloquercetin 3,5,3'-trimethyl ether
	0.74	5,4'-dihydroxy-3,6,7,3'-tetramethoxy-8-*C*-methylflavone
	0.72	20(*R*)-hydroxydammar-24-en-3-one
E. coli	0.78	(4a*R*,5*S*,6*R*,8a*R*)-5-[2-(2,5-dihydro-5-methoxy-2-oxofuran-3-yl)ethyl]-3,4,4a,5,6,7,8,8a-octahydro-5,6,8a-trimethylnaphthalene-1-carboxylic acid
	0.76	3-oxo-17-carboxy-3,18-*seco*-barbacenic acid
	0.74	amentoflavone
	0.73	euscaphic acid
	0.72	3',8"-biisokaempferide
P. aeruginosa	0.62	3-oxo-17-carboxy-3,18-*seco*-barbacenic acid
	0.61	3',8"-biisokaempferide
	0.57	7β,8,14β-trihydroxy-15-isopimare-18-oic acid
	0.56	amentoflavone
	0.55	betulonic acid
S. enterica	0.71	5,3',4'-trihydroxy-3,6,7-trimethoxy-8-*C*-methylflavone
	0.71	3,5,7,30,40-pentahydroxy-6-prenylflavonol
	0.70	amentoflavone
	0.67	3',4',5,7-tetrahydroxy-3,6-dimethoxy-8-methylflavone
	0.65	3',8"-biisokaempferide

5-fold cross-validation resulted in 80% prediction accuracy identifying the best compounds based on imidazolium ionic liquids and experimentally validated.

Cho *et al.*[28] constructed six models using energy relationship descriptors against *E. coli*, *S. aureus* and *C. albicans* using the MIC and minimum bactericidal concentration (MBC) values for each species. The predictability of the models was estimated by obtaining R^2 = 0.90 and 0.93 (R^2 = determination coefficient) for MIC and MBC of *E. coli*, respectively, R^2 = 0.91 and 0.94 for MIC and MBC of *S. aureus*, R^2 = 0.89 and 0.80 for *C. albicans*. According to the authors,[28] the QSAR models will support a reliable, economical, fast and safe evaluation as a supplementary method of experimental testing.

Although the number of studies with the use of QSAR classificatory models is increasing, further studies are needed with the application of these methodologies that can identify potential molecules and assist experimental tests.

In the *S. enterica* model, five molecules were outside the applicability domain, in addition to the flavonoid kaempferol 3-*O*-(3”,6”-di-*O*-*E*-*p*-coumaroyl)-β-D-glucopyranoside, isorhamnetin 3-*O*-(3”,6”-di-*O*-*E*-*p*-coumaroyl)-β-D-glucopyranoside, tetracosanoic acid, palmitic acid and heptacosan-1-ol. Of the molecules that remained within the domain, 18 obtained a probability of active potential greater than 50%, varying up to 71%. The chemical compounds with the best probability are also described in Table 3.

The flavonoid amentoflavone had a probability of being active for all models, despite not being represented in the table for the *C. albicans* model, it had a probability of an active potential of 0.65 for this model.

Conclusions

Through the *in silico* tools used in this work, it was possible to generate a model bank to virtually track isolated compounds from the Velloziaceae family with probable antimicrobial potential. The models for *C. albicans* and *E. coli* were the ones that presented compounds with the highest probability of activity.

For *C. albicans*, the model selected thirty-one molecules with a potential activity greater than 60%, twenty-nine molecules with a probability greater than 50% for *E. coli*, eleven molecules with a probability greater than 52% for *P. aeruginosa* and nineteen molecules with a probability of 50% for *Salmonella* sp.

Biflavonoid amentoflavone was the only compound to be likely to be active for all four models with a considerable percentage, with a potential probability of 65, 74, 56 and 70% for *C. albicans*, *E. coli*, *P. aeruginosa* and *Salmonella* sp., respectively.

The present study contributed, through the virtual screening of a bank of secondary metabolites, to select several proposed compounds with potential antimicrobial activity, especially biflavonoid amentoflavone and, in the future, assist biological testing in discovering potential drug candidates.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil (CAPES), finance code 001 and Instituto Nacional de Ciência e Tecnologia (INCT), RENNOFITO.

Author Contributions

Pinheiro, Barros, Assis, Scotti, Tavares and Silva were resposible for the conceptualization, data curation, formal analysis, investigation and supervision; Silva, Tavares, Scotti and Scotti for the funding acquisition and resources; Pinheiro, Barros, Assis, Araújo, Sales, Maia, Scotti and Tavares for the methodology; Pinheiro, Scotti, Tavares and Silva for the project administration; Barros, Maia, Scotti and Scotti for the software; Pinheiro, Barros, Assis, Maia, Scotti, Scotti, Tavares and Silva for the validation; Pinheiro, Assis, Barros, Sales, Araújo, Maia, Scotti, Scotti, Tavares and Silva for the visualization; Pinheiro, Barros and Assis for writing original draft; Pinheiro, Assis, Barros, Sales, Araújo, Maia, Scotti, Scotti, Tavares and Silva for writing review and editing.

References

1. Brown, K.; *J. Ethnobiol.* **2016**, *36*, 861.
2. Tewari, D.; Mocan, A.; Parvanov, E. D.; Sah, A. N.; Nabavi, S. M.; Huminiecki, L.; Ma, Z. F.; Lee, Y. Y.; Horbańczuk, J. O.; Atanasov, A. G.; *Front. Pharmacol.* **2017**, *8*, 518.
3. Newman, D.; Cragg, G. M.; *J. Nat. Prod.* **2016**, *79*, 629.
4. Mello-Silva, R.; *Velloziaceae in Lista de Espécies da Flora do Brasil*; Jardim Botânico do Rio de Janeiro: Rio de Janeiro, Brazil, 2015. Available at http://floradobrasil.jbrj.gov.br/jabot/floradobrasil/FB245, accessed in June 2020.
5. Mello-Silva, R.; *Rodriguésia* **2018**, *69*, 259.
6. Paula-Souza, J.; Brandão, M. G. L.; *História das Plantas Medicinais e Úteis do Brasil*, vol. 1, 1st ed.; Fino Traço: Belo Horizonte, Brazil, 2016.
7. Mello-Silva, R.; Santos, D. Y. A. C.; Salatino, M. L. F.; Motta, L. B.; Cattai, M. B.; Sasaki, D.; Lovo, J.; Pitai, P. B.; Rocini, C.; Rodrigues, C. D. N.; Zarrei, M.; Chase, M. W.; *Ann. Bot.* **2011**, *108*, 87.
8. Souza, C. D.; Felfili, J. M.; *Acta Bot. Bras.* **2006**, *20*, 135.
9. Messias, M. C. T. B.; Menegatto, M. F.; Prado, A. C. C.; Santos, B. R.; Guimarães, M. F. M.; *Rev. Bras. Plant. Med.* **2015**, *17*, 76.
10. Rocha, C.; Reynolds, N. D.; Simons, M. P.; *Rev. Peru. Med. Exp. Salud Publica* **2015**, *32*, 139.

11. http://www.who.int/mediacentre/factsheets/fs194/en/, accessed in June 2020.
12. Souza, H. D. S.; de Sousa, R. P. F.; Lira, B. F.; Vilela, R. F.; Borges, N. H. P. B.; de Siqueira-Júnior, J. P.; Lima, E. O.; Jardim, J. U. G.; da Silva, G. A. T.; Barbosa-Filho, J. M.; de Athayde-Filho, P. F.; *J. Braz. Chem. Soc.* **2019**, *30*, 188.
13. Hoque, I.; Chatterjee, A.; Bhattacharya, S.; Biswas, R.; *Int. J. Adv. Res. Biol. Sci.* **2017**, *4*, 60.
14. Das, P. S.; Saha, P. A.; *World J. Pharm. Pharm. Sci.* **2017**, *6*, 279.
15. Piccirillo, E.; Amaral, A. T.; *Quim. Nova* **2018**, *41*, 662.
16. ChemAxon; *MarvinSketch*, v.18.10.0; ChemAxon, Cambridge Innovation Center, Cambridge, USA, 2018. Available at www.chemaxon.com, accessed in June 2020.
17. *Standardizer Software*, version for Win32; Cambridge Innovation Center, Cambridge, USA, 2017. Available at www.chemaxon.com, accessed in June 2020.
18. Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B.; *J. Mol. Struct.* **2000**, *503*, 17.
19. *Knime*, 3.5.3; Konstanz Information Miner Copyright, Germany, 2003-2017. Available at https://www.knime.com/community/enalos-nodes, accessed in June 2020.
20. *Dragon*, 7.0; Kode Chemoinformatics, Italy, 2016.
21. Quilan, J. R.; *C4.5: Programs for Machine Learning*, vol. 1, 1st ed.; Morgan Kaufmann Publishers: San Mateo, California, USA, 1993.
22. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H.; *SIGKDD Explor.* **2009**, *11*, 10.
23. Matthews, B. W.; *Biochim. Biophys. Acta* **1975**, *405*, 442.
24. Zhang, A.; Golbraikh, S.; Oloff, H.; Kohn, A.; Tropsha, J.; *J. Chem. Inf. Model.* **2006**, *46*, 1984.
25. Scotti, M. T.; Herrera-Acevedo, C.; Oliveira, T. B.; Costa, R. P. O.; Santos, S. Y. K. O.; Rodrigues, R. P.; Scotti, L.; Da-Costa, F. B.; *Molecules* **2018**, *23*, 103.
26. Trush, M. M.; Kovalishyn, V.; Ocheretniuk, A. D.; Kobzar, O. L.; Kachaeva, M. V.; Brovarets, V. S.; Metelytsia, L. O.; *Curr. Drug Discovery Technol.* **2019**, *16*, 204.
27. Hodyna, D.; Kovalishyn, V.; Rogalsky, S.; Blagodatnyi, V.; Metelytsia, L.; *Curr. Drug Discovery Technol.* **2016**, *13*, 109.
28. Cho, C.-W.; Park, J.-S.; Stolte, S.; Yun, Y.-S.; *J. Hazard. Mater.* **2016**, *311*, 168.

Submitted: February 3, 2020
Published online: June 10, 2020