Is Molecular Alignment an Indispensable Requirement in the MIA-QSAR Method?

Stephen J. Barigye* and Matheus P. Freitas

For a decade, the multivariate image analysis applied to quantitative structure–activity relationship (MIA-QSAR) approach has been successfully used in the modeling of several chemical and biological properties of chemical compounds. However, the key pitfall of this method has been its exclusive applicability to congeneric datasets due to the prerequisite of aligning the chemical images with respect to the basic molecular scaffold. The present report aims to explore the use of the 2D-discrete Fourier transform (2D-DFT) as a means of opening way to the modeling, for the first time, of structurally diverse noncongruent chemical images. The usability of the 2D-DFT in QSAR modeling of noncongruent chemical compounds is assessed using a structurally diverse dataset of 100 compounds, with reported inhibitory activity against MCF-7 human breast cancer cell line. An analysis of the statistical parameters of the built regression models validates their robustness and high predictive power. Additionally, a comparison of the results obtained with the 2D-DFT MIA-QSAR approach with those of the DRAGON molecular descriptors is performed, revealing superior performance for the former. This result represents a milestone in the MIA-QSAR context, as it opens way for the possibility of screening for new molecular entities with the desired chemical or therapeutic utility. © 2015 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23992

Introduction

The advancement of existing computational tools and/or methods to enhance their utility in molecular modeling constitutes one of the fundamental tasks of theoretical and computational chemists. For example, several well-known topological indices have been extended to incorporate information on the spatial arrangement (conformation) of chemical structures, many atom-based indices have been generalized to consider n-tuple relations of atoms, and some 3D-QSAR methods have been extrapolated to include information on the ligand–receptor interactions [1–6]. Additionally, quantum mechanics derived descriptors have been revolutionized with the introduction of the density functional theory allowing for a more wholesome description of the electronic structure.[7]

A decade ago, an innovative method motivated by the notion that images of chemical structures contain useful chemical information was introduced.[8] This method, which considers image pixels as molecular descriptors, is denominated as MIA-QSAR (acronym for Multivariate Image Analysis-Quantitative Structure Activity Relationship) and is based on the reasoning that there exists a close relationship between the chemical, physicochemical, and biological behavior of compounds in a congeneric chemical dataset and the nature of the substituents linked to the basic molecular scaffold. The MIA-QSAR method may well be considered as a success as it has found applications in the modeling of a numerous bioactivities ranging from antimalarials, glycogen synthase kinase 3 (GSK-3) inhibitors, anticancer, HIV reverse transcriptase inhibitors, anti-inflammatory activity, phosphodiesterase type 5 (PDE-5) inhibitors, antifungals, farnesyltransferase inhibitors to antischistosomal activity, among others.[8–23] This method has also been applied in the modeling of physicochemical properties, herbicide phytotoxic and soil sorption profiles, and chemical shifts of compounds.[24–28]

Recent advances in the MIA-QSAR strategy to improve its usability have included the incorporation of color schemes defined to integrate important chemical information for instance the atomic electronegativity according to Pauling's scale and the modification of the atomic sizes in the images in accordance to the atoms' Van der Waals radii with the aim of codifying information on the steric features of the atoms. Several studies have revealed that these schemes in fact do enhance the modeling capacity of the MIA-QSAR models, in addition to permitting greater interpretation of the information codified.[22–24,27–29]

However, this success story is not exempt of pitfalls. The MIA-QSAR approach, just like all alignment-based methods, has been up to the moment applicable exclusively to datasets of congeneric chemical images, and thus precluding the possibility of performing virtual screening tasks for novel lead compounds with this method. Additionally, while the MIA-QSAR method follows a much simpler manual alignment rule relative to other alignment-based techniques like COMFA, COMSIA, SOMFA, the fact that the alignment is manually performed relative to a common pixel coordinate of the basic scaffold adds

S. J. Barigye, M. P. Freitas
Department of Chemistry, Federal University of Lavras, P.O. Box 3037, Lavras, Minas Gerais 37200-000, Brazil
E-mail: sjbarigye@gmail.com
Contract grant sponsors: Barigye, S. J. and Freitas, M.P. acknowledge financial support from CNPq and FAPEMIG (to S.J.B. and M.P.F.)

subjectivity to this method, in the sense that proper alignment basically depends on visual accuracy.

In a previous report, we proposed the 2D-discrete Fourier transform (2D-DFT) as a means of creating a common base for congruent chemical structural images.[30] The 2D-DFT converts images of chemical structures into magnitude spectra in the frequency domain and thus creating a common base for constructing a Multivariate Image (MVI). Experimental studies revealed that with the 2D-DFT, the manual alignment procedure could be precluded and thus constituting an important benefit in the MIA-QSAR context. However, 2D-DFT is in fact useful in the analysis of images of objects not only of different orientation, but of varying sizes and shapes. Extrapolating this understanding to the MIA-QSAR perspective, the present report aims to explore the use of the 2D-DFT approach as a means of opening way to the modeling, for the first time, of structurally diverse noncongruent chemical images. The rationale behind this hypothesis is that as the 2D-DFT decomposes an image into its constituent linear functions, congruence becomes a function of these constituents in magnitude spectra and thus superposition of the chemical images with respect to a particular basic scaffold is not necessary. This should in turn allow for the QSAR modeling of structurally diverse datasets.

Materials and Methods

Definition of the 2D-discrete Fourier transform

Fourier transform (FT), whose origin is traced way back in 1822 with the seminal work of the French mathematician Jean Baptiste Joseph Fourier,[31] is probably one of the most relevant transforms used in modern signal and image processing. The core postulate of FT is that functions (or phenomena) may be represented in the frequency domain as linear combinations of trigonometric sine and cosine functions of varying periodicity. These functions are weighted by coefficients of different magnitudes denominated as Fourier coefficients. In the context of images, these are simply an arrangement of discrete linear spatial functions of different orientation and periodicity (in space). Viewed from this perspective, the FT permits obtaining a frequency domain representation of images, denominated as the *Fourier spectrum*. The 2D-DFT for an image function $f(m, n)$ is defined by the following expression:

$$F(u,v)=\frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(m,n)\exp[-i2\pi(um/M+vn/N)] \quad (1)$$

where u and v denote the spatial frequencies, M and N the number of points sampled in the R^2 space. For a given spectrum, it follows that the magnitudes of the Fourier coefficients are related to the intensity of the frequencies in the spatial domain. The magnitude matrix computed for an image function $f(m, n)$ is as an $M \times N$ image, whose coordinates are the frequencies (u, v) and the corresponding intensities the pixel values.

Chemical dataset for 2D-DFT-based MIA-QSAR modeling

To assess the usability of the 2D-DFT in QSAR modeling of noncongruent chemical compounds, a structurally diverse dataset comprised of 100 compounds reported to be active against the MCF-7 human breast cancer cell line was constructed, through a painstaking search of literature.[32–40] The MCF-7 cell line is a popular estrogen receptor positive control cell line, widely used for *in vitro* studies in breast cancer research. Due to the disparity in the units for the IC_{50} values and in some cases improper conversions, these were first regularized and then converted to pIC_{50} values. Figure 1 shows the chemical structures of the dataset used in this study (see Supporting Information SI1 for identity of the substituents). As may be noted, these structures are not entirely superimposable, and thus modeling using the classical MIA-QSAR approach would not be possible.

The chemical structures were drawn using the ChemBioDraw program and transferred to a work space in the Windows Paint program whose dimension was kept the same for all structures. The chemical structure images were saved as TIFF image files and posteriorly converted into magnitude spectra using the fast Fourier transformation algorithm available in the MATLAB program.[41] The magnitude spectra obtained for the chemical dataset were then used to construct an $l \times m \times n$ dimensional MVI, where l is the number of magnitude spectra (instances) and $m \times n$ the sample points of a magnitude spectra. The MVI was later unfolded to a two-way $l \times (m \times n)$ data matrix. Figure 2 is an illustration of the 2D-DFT MIA-QSAR work flow.

Bearing in mind that the magnitude spectra are symmetrical, half of the variables [i.e., $(m \times n)/2$] are used. Additionally, all the zero variance variables were excluded as these are considered as portions of communality in all the chemical structures of the dataset. The construction of the MVI as well as the unfolding procedure were performed using the MATLAB program.[41]

Dataset splitting and model building

The chemical dataset was split into training and test sets using the cluster analysis method. First, hierarchical cluster analysis (HCA) was performed using the squared Euclidean distance and Ward's algorithm as the dissimilarity measure and linkage rule, respectively. Using the amalgamation schedule, the distance corresponding to the steepest ascent was identified and used to determine the optimum number of clusters. Posteriorly, *k*-means cluster analysis (*k*-MCA) was performed, using the number of clusters determined with the HCA method. For each cluster, the compounds were ordered according to their pIC50 values, and the compounds selected to span the cluster's activity range, with 75 and 25% constituting the training and test sets, respectively (see Supporting Information SI2 and SI3 for HCA dendogram and the clusters' membership). This approach guarantees representativity in the training and test sets in terms of the structural characteristics of the chemical compounds and their quantitative bioactivity.

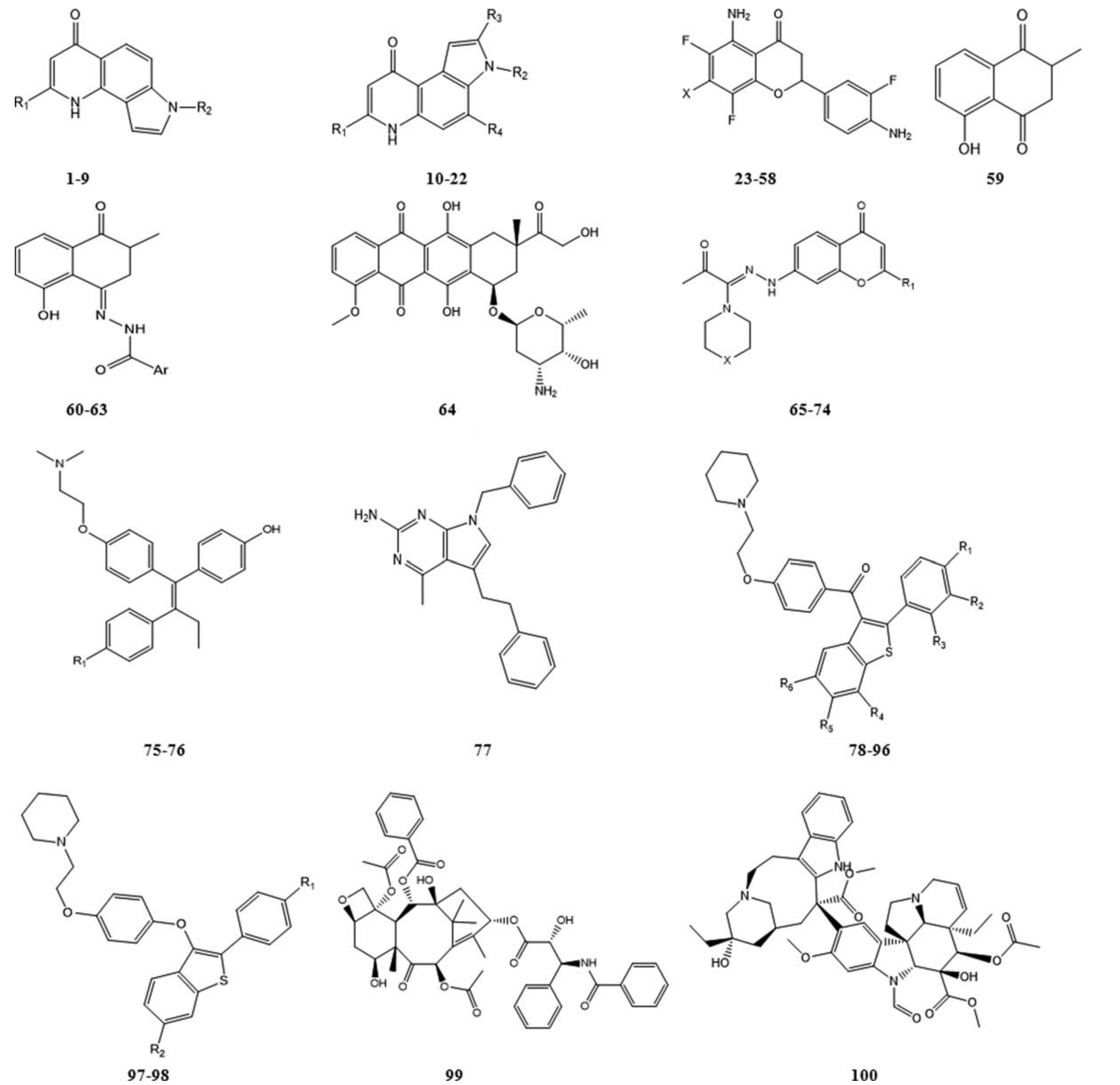

Figure 1. Molecular structure scaffolds of compounds that constitute the dataset used to build models for the MCF-7 cells inhibitory activity.

Multiple linear regression (MLR)-based models were built for the MCF-7 cells inhibitory activity, using the MOBYDIGS software which allows for the exploration of optimal models using the genetic algorithm (GA).[42] Given the fact that the MOBYDIGS software allows for modeling with a maximum of 2000 variables, a rank-based feature selection filter denominated as the Differential Shannon's Entropy[43] was applied to the magnitude spectra data matrix and the accepted number of variables retained. This dimensionality reduction procedure was performed using the IMMAN software.[43] The following

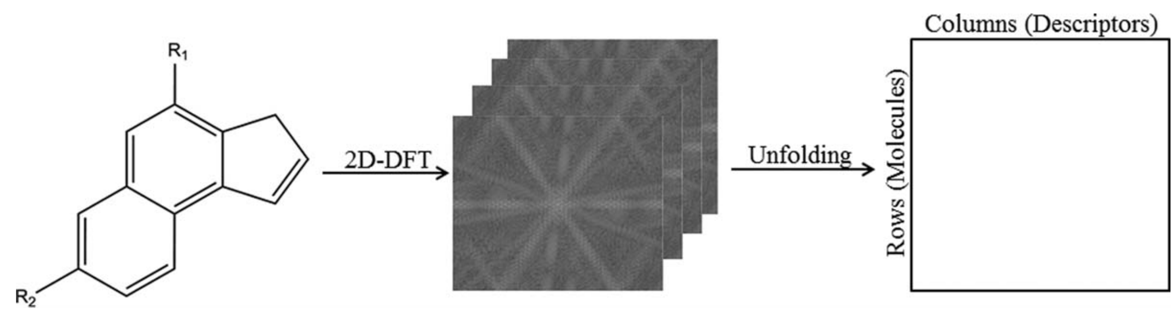

Figure 2. Illustration of the 2D-DFT MIA-QSAR workflow (note that for simplicity a congeneric series is considered).

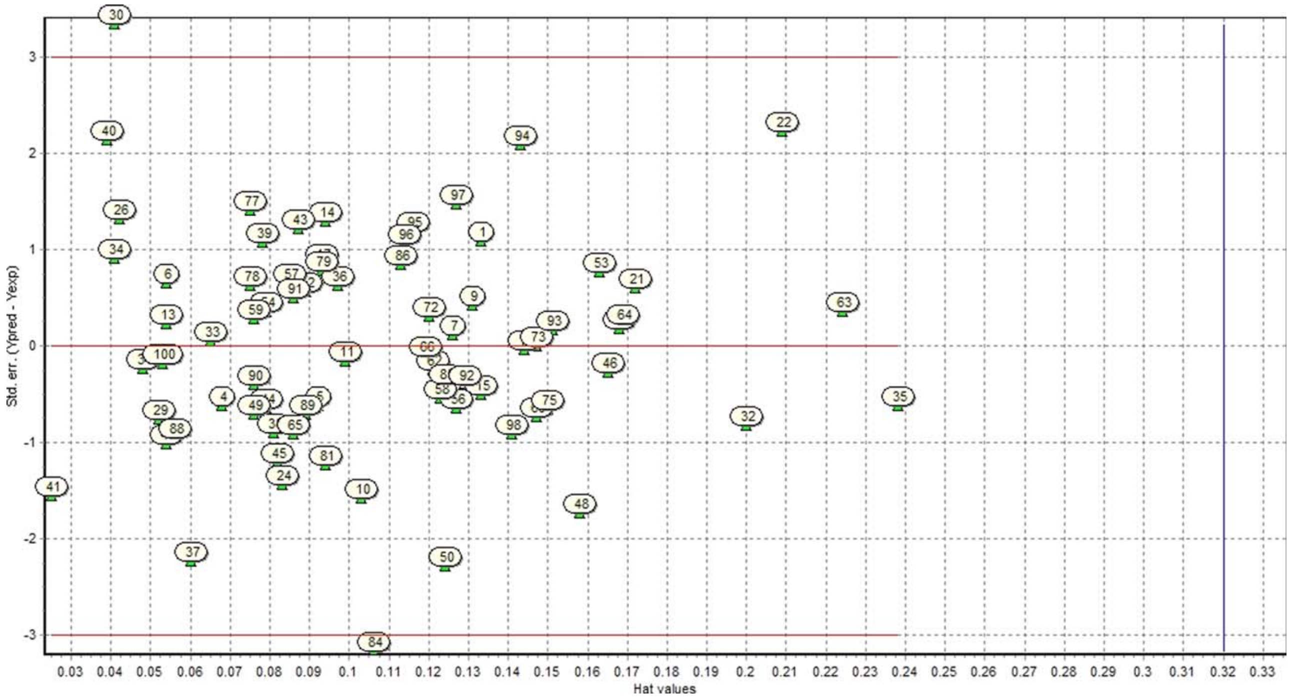

Figure 3. William's plot for the 7 variable 2D-DFT-based MIA-QSAR model.

configurations were used for the GA-based exploration of MLR models: the leave one out cross validation parameter (Q^2loo) was considered as the optimization function, the population size and the reproduction/mutation trade-off (T) were set at 100 and 0.5, respectively. The MLR-GA approach generates numerous models, and thus the selection of the best model was performed considering the squared correlation coefficient (R^2), the standard deviation (s), the Fisher's ratio (F) and the Q^2loo parameters. Posteriorly, the selected models were vigorously validated using the validation techniques bootstrapping (Q^2 boot) and external validation [i.e., the correlation coefficient for the test set (Q^2ext), the modified correlation coefficient with respect to the origin (Q^2_0ext) and the slope (k) were considered] to evaluate their robustness and earnest predictive power, respectively, with the latter performed over a set of compounds not used in the model building.[44] A 10-fold Y-randomization procedure was performed to check for fortuitous correlation. It follows that stable prediction models exhibit low intercept values [i.e., a(R^2) and a(Q^2)], while the high values indicate high propensity to this phenomenon. Additionally, leverage and outlier diagnosis was performed using the William plot.[45]

Results and Discussions

QSAR modeling of the MCF-7 cells inhibitory activity

The best 3–7 variable MLR-based models for the MCF-7 cells inhibitory activity, determined according to the quality of the respective statistical parameters were retained. The key advantage of MLR as a statistical technique is its simplicity. Figure 3 shows the William's plot for the 7 variable model.

As can be observed compounds 30 and 84 in the training set exhibited outlying behavior and their removal improved the models' performance. Likewise, compound 76 in the test set exhibited atypical behavior and was thus excluded. Table 1 shows the model equations as well as the corresponding statistical parameters (for variables constituting the models 2–6, see Supporting Information, SI4).

As can be observed from Table 1, the obtained models are generally robust and possess high predictive power evidenced by the high Q^2loo, Q^2boot, and Q^2ext values, in addition to the minimal difference between these parameters and relative to the R^2adj. Moreover, the small intercept values [a(R^2), a(Q^2)] suggest that the obtained models are not prone to chance correlation. On the whole, the 7 variable model [eq. (2)] presents the best results although the 6 and 5 variable models [eqs. (3) and (4)] possess superior Q^2ext values, respectively. Table 2 shows the experimental and predicted values for the models 2–6.

The low residual values, even with a lower degree of freedom, suggest satisfactory behavior for the obtained models. These results demonstrate that the application of the 2D-DFT to the MIA-QSAR approach permits adequate stratification of information on chemical topology in the sense that structural patterns of congruence (or similarity) are duly identified, permitting the "filtering" the patterns of dissimilarity which are in theory considered to be responsible for the variation in the inhibitory activity of these structurally diverse chemical compounds.

Comparison of 2D-DFT MIA-QSAR approach with DRAGON descriptors

With the aim of gaining greater insight on the performance of the 2D-DFT MIA-QSAR approach relative to other alignment

Table 1. Statistical parameters and equations for models for the MCF-7 cells inhibitory activity of dataset constructed in this study.[a]

MDs	N	R^2	Q^2loo	Q^2boot	Q^2ext	Q^2_oext	k	a(R^2)	a(Q^2)	R^2adj	s	F	Model	Eq.
2D-DFT MIA-QSAR	7	89.36	86.89	86.09	74.72	65.55	1.01	0.057	−0.195	88.22	0.52	78.02	pIC50 = 11.2243 (±0.6473) − 0.0097 (±0.0022) **X21** + 0.0131 (±0.0026) **X63** + 0.0119 (±0.0024) **X110** − 0.0082 (±0.0025) **X141** − 0.0250 (±0.0022) **X172** − 0.0103 (±0.0021) **X190** + 0.0147 (±0.0021) **X239**	**(2)**
	6	87.13	84.52	83.78	82.96	78.73	0.98	0.044	−0.171	85.96	0.57	74.47	pIC50 = 10.4155 (±0.8178) + 0.0150 (±0.0026) **X110** − 0.0076 (±0.0020) **X121** − 0.0125 (± 0.0027) **X141** − 0.0228 (± 0.0023) **X172** − 0.0104 (±0.0023) **X190** + 0.0268 (±0.0030) **X212**	**(3)**
	5	84.43	81.86	81.28	79.51	67.93	1.00	0.028	−0.157	83.27	0.62	72.66	pIC50 = 9.9739 (±0.8833) + 0.0128 (±0.0027) **X110** − 0.0127 (±0.0029) **X141** − 0.0235 (±0.0025) **X172** − 0.0095 (±0.0025) **X190** + 0.0228 (±0.0031) **X212**	**(4)**
	4	81.01	78.37	77.99	71.77	46.64	1.01	0.017	−0.134	79.90	0.68	72.54	pIC50 = 10.2705 (±0.9645) + 0.0128 (±0.0030) **X110** − 0.0156 (±0.0031) **X141** − 0.0283 (±0.0024) **X172** + 0.0207 (±0.0033) **X212**	**(5)**
	3	78.68	76.43	76.41	70.16	58.29	0.98	0.001	−0.118	77.76	0.72	86.10	pIC50 = 9.8523 (±1.0014) − 0.0247(±0.0028) **X172** − 0.0111 (±0.0027) **X190** + 0.0255 (±0.0031) **X212**	**(6)**
DRAGON	7	86.27	83.67	82.48	81.95	67.48	1.02	0.052	−0.224	84.79	0.61	58.36	pIC50 = 16.0496 (±1.6990) − 0.6660 (±0.0989) **ESpm13d** − 1.2843 (±0.2285) **Mor16e** + 1.5736 (±0.2057) **L3e** − 4.0314 (±1.0124) **E3s** − 1.6251 (±0.4411) **nArCOOH** + 0.4715 (±0.0477) **F03[C-S]** + 0.6445 (±0.1378) **F06[O-F]**	**(7)**
	6	83.41	80.6	79.32	81.15	67.78	1.05	0.046	−0.172	81.9	0.66	55.29	pIC50 = 16.0359 (±1.8537) − 0.6647 (±0.1079) **ESpm13d** − 1.2818 (±0.2493) **Mor16e** + 1.5140 (±0.2238) **L3e** − 3.9381 (±1.1043) **E3s** + 0.4646 (±0.0521) **F03[C-S]** + 0.6692 (±0.1501) **F06[O-F]**	**(8)**
	5	81.00	78.48	77.41	65.80	45.40	1.05	0.027	−0.187	79.58	0.70	57.11	pIC50 = 15.8323 (±1.9671) − 0.6479 (±0.1145) **ESpm13d** − 1.4606 (±0.2609) **Mor16e** + 1.0385 (±0.1739) **L3e** − 1.6789 (±0.5097) **nArCOOH** + 0.4616 (±0.0553) **F03[C-S]**	**(9)**
	4	78.21	75.48	75.02	28.78	38.38	1.05	0.019	−0.139	76.93	0.75	61.02	pIC50 = 11.7525(±2.1199) − 0.4780 (±0.1343) **ESpm12d** − 1.6086 (±0.2885) **Mor16u** + 9.8295 (±1.8635) **E3v** + 2.4428 (±0.3059) **B10[C-S]**	**(10)**
	3	74.81	71.22	70.96	56.78	45.89	1.05	0.008	−0.112	73.72	0.80	68.32	pIC50 = 2.0042 (±0.8118) + 5.5638 (±1.2818) **HOMA** − 1.5048 (±0.2494) **Mor16u** + 2.8180 (±0.2919) **F08[O-S]**	**(11)**

[a] $n_{training} = 73$, $n_{test} = 24$.

Table 2. Experimental and predicted inhibitory activity values for the MLR models 2–6.

ID	Status	Exp.	Pred[eq. (2)]	Pred[eq. (3)]	Pred[eq. (4)]	Pred[eq. (5)]	Pred[eq. (6)]
1	Training	4.00	4.51	4.11	4.29	4.43	4.51
2	Test	4.00	3.47	3.98	3.94	3.98	4.51
3	Test	4.52	5.58	5.61	5.65	5.73	5.68
4	Training	5.40	5.09	4.75	4.87	4.52	5.07
5	Training	5.40	5.13	4.99	5.01	4.55	5.57
6	Training	4.22	4.64	3.79	4.16	4.30	4.07
7	Training	4.16	4.13	4.55	4.78	4.85	4.86
8	Test	6.00	5.32	5.14	5.21	4.96	5.64
9	Training	4.30	4.6	4.89	5.00	4.55	5.59
10	Training	6.10	5.3	5.23	5.36	5.14	5.00
11	Training	5.70	5.71	5.93	5.77	5.76	5.02
12	Training	5.16	4.98	4.65	4.93	4.69	4.82
13	Training	5.30	5.47	4.95	5.10	5.01	4.78
14	Training	4.30	5.05	4.85	5.07	5.10	4.91
15	Training	4.30	4.06	4.16	4.63	4.27	4.53
16	Test	4.30	4.6	4.00	4.18	4.37	4.30
17	Training	4.82	5.32	5.52	5.59	5.55	5.09
18	Training	5.16	5.00	5.16	5.26	5.11	5.39
19	Test	5.10	5.00	4.19	4.41	4.28	5.50
20	Training	4.30	4.45	4.66	5.04	4.85	5.23
21	Training	4.30	4.50	3.93	4.19	4.22	3.68
22	Training	4.30	5.38	5.23	5.35	5.76	4.79
23	Test	7.40	7.68	7.77	7.55	7.60	7.37
24	Training	7.89	7.22	7.14	6.97	6.85	6.84
25	Test	6.89	6.81	6.75	6.62	6.54	6.58
26	Training	5.06	5.93	6.26	6.08	6.08	6.06
27	Test	5.00	5.05	5.43	5.16	5.43	5.51
28	Test	6.68	7.71	7.92	7.56	7.47	7.23
29	Training	6.8	6.40	6.16	6.22	6.23	6.08
30[a]	Training	5.00	–	–	–	–	–
31	Training	7.00	6.91	6.34	6.31	6.40	6.27
32	Training	5.85	5.65	6.62	6.51	6.23	6.51
33	Training	6.72	6.80	6.71	6.65	6.89	6.93
34	Training	5.60	6.12	6.11	6.26	6.52	6.58
35	Training	5.00	4.67	5.27	5.49	5.77	6.22
36	Training	6.34	6.70	6.29	6.59	7.14	6.42
37	Training	8.11	6.85	7.19	6.50	6.77	6.67
38	Training	7.41	6.92	6.41	6.32	6.72	6.22
39	Training	5.85	6.62	7.21	7.01	6.78	6.63
40	Training	5.00	6.34	6.23	6.00	6.08	5.82
41	Training	7.21	6.34	6.04	5.69	5.69	5.74
42	Test	7.59	7.34	7.70	7.31	7.08	7.14
43	Training	6.89	7.62	7.51	7.15	7.22	7.48
44	Training	6.80	6.44	7.08	6.72	7.00	6.88
45	Training	7.00	6.29	7.21	6.65	6.63	7.07
46	Training	5.00	4.73	4.65	4.51	4.53	5.03
47	Training	7.59	6.96	7.30	6.71	6.90	6.62
48	Training	7.72	6.90	7.07	6.70	6.03	6.63
49	Training	7.54	7.15	6.74	6.64	6.37	6.49
50	Training	7.34	5.91	6.25	6.20	6.23	5.88
51	Test	5.92	5.72	5.83	5.90	5.47	6.09
52	Training	5.15	5.46	5.01	5.06	4.90	5.72
53	Training	5.00	5.22	5.14	5.07	5.31	5.89
54	Training	5.00	5.23	5.06	5.11	4.69	5.54
55	Test	7.59	8.56	8.56	8.70	8.26	9.28
56	Training	7.14	6.84	7.38	7.12	7.45	6.21
57	Training	6.47	6.94	6.65	6.30	5.86	6.19
58	Training	5.51	5.23	5.75	5.59	5.99	5.49
59	Training	5.26	5.43	5.72	5.79	6.08	5.58
60	Test	5.21	5.57	5.22	5.18	5.24	5.25
61	Test	5.57	6.02	5.59	5.62	5.63	5.41
62	Training	5.28	5.19	5.13	5.03	5.10	5.43
63	Training	5.55	5.84	4.98	4.96	5.08	4.98
64	Training	6.51	6.82	7.03	6.39	6.24	5.81
65	Training	5.23	4.71	5.27	5.14	5.52	5.11
66	Training	4.65	4.58	5.04	5.07	5.43	5.32

(Continued)

Table 2. (*Continued*)

ID	Status	Exp.	Pred[eq. (2)]	Pred[eq. (3)]	Pred[eq. (4)]	Pred[eq. (5)]	Pred[eq. (6)]
67	Test	4.25	4.03	4.35	4.46	4.87	4.80
68	Test	4.87	4.57	4.76	4.83	5.28	4.73
69	Training	4.67	4.11	4.58	4.66	5.10	4.72
70	Test	5.56	4.61	5.99	5.55	6.00	5.61
71	Test	5.20	5.14	5.95	5.63	5.93	5.63
72	Training	5.06	5.17	4.90	4.92	5.34	5.78
73	Training	4.24	4.21	4.66	4.62	5.10	4.42
74	Test	5.47	4.80	5.75	5.52	5.88	5.13
75	Training	6.32	6.03	5.39	5.06	5.45	5.16
76[a]	Test	8.92	–	–	–	–	–
77	Training	4.97	5.80	5.61	6.09	6.18	5.81
78	Training	6.52	6.87	7.00	6.83	6.78	6.48
79	Training	6.60	7.04	6.97	7.01	7.27	6.68
80	Test	6.00	6.48	6.06	5.91	5.79	6.17
81	Training	6.52	5.93	6.68	6.71	6.41	6.85
82	Test	9.70	7.31	8.25	7.44	7.19	7.82
83	Test	7.46	7.93	7.99	8.01	6.96	9.02
84[a]	Training	9.00	–	–	–	–	–
85	Training	7.70	7.38	6.86	7.26	7.35	7.97
86	Training	8.16	8.55	8.41	8.65	9.06	8.12
87	Test	8.64	8.64	8.86	9.01	8.73	9.00
88	Training	8.50	7.90	8.23	8.12	7.87	7.87
89	Training	9.16	8.66	8.23	8.72	8.57	8.72
90	Training	8.60	8.31	8.54	8.74	8.57	8.39
91	Training	8.30	8.57	8.58	8.63	8.16	8.43
92	Training	9.52	9.22	8.61	9.20	9.49	8.72
93	Training	7.30	7.32	7.39	7.55	7.00	8.47
94	Training	6.49	7.62	7.90	7.96	7.39	8.55
95	Training	8.52	9.15	8.57	9.10	9.43	8.50
96	Training	8.00	8.57	8.09	8.13	7.73	8.69
97	Training	9.40	10.25	9.77	9.94	9.52	9.50
98	Training	10.30	9.75	9.80	9.68	9.32	9.29
99	Test	8.16	7.08	8.13	7.51	6.83	6.69
100	Training	7.70	7.56	8.00	8.01	8.06	7.70

[a] Outlier compounds.

free approaches, the 0D–3D molecular descriptors implemented in the DRAGON software[46] were computed for the constructed chemical dataset and posteriorly used in the modeling of the MCF-7 cells inhibitory activity. Prior to modeling, prefilters for constant and highly correlated variables (for pair correlation coefficients = 1.0) were applied yielding 1612 variables, with which 3–7 variable MLR-based QSAR models were built. To ensure comparativity, the same compounds used in the training and test sets with the 2D-DFT MIA-QSAR approach were used. Likewise, the best models were selected considering the squared correlation coefficient (R^2), the standard deviation (s) and the Fisher's ratio (F) and the Q^2loo parameters. An analysis of the William plot for the 7-variable model revealed outlying behavior for compounds 30 and 79, with the former being equally identified in the 2D-DFT MIA-QSAR models as an outlier. The removal of these compounds improved the performance of the DRAGON models and they were thus ultimately excluded. Despite the high residual value for compound 76 in the test set, it did not amount to outlying behavior. However, for an earnest comparison, it was not included in the test series. Table 1 shows the statistical parameters and equations of the best models obtained with the DRAGON molecular descriptors. With the exception of model 9 which is characterized by a low Q^2ext value (28.78), the rest of the DRAGON models depict good behavior evidenced by the quality of the respective statistical parameters, and therefore, the usefulness of the DRAGON MDs in modeling the MCF-7 cells inhibitory activity of the constructed dataset is demonstrated.

As for the comparison with the 2D-DFT MIA-QSAR approach, it is interesting to note that generally superior performance is obtained with the 2D-DFT MIA-QSAR models relative to the DRAGON descriptor-based models, with the exception of the external validation parameter (Q^2ext) for the 7 variable model for the latter, where a higher value is obtained (see Table 1). It should be noted that the DRAGON software is a compilation of a diverse pool of MDs defined following a wide range of concepts and mathematical formalisms, representing over 50 years of huge research efforts. This result validates the 2D-DFT MIA-QSAR approach as an important tool to count on in the modeling of properties/bioactivities of structurally diverse chemical compounds.

Conclusions

The usability of the 2D-DFT approach in modeling structurally diverse datasets has been demonstrated. This approach is based on the transformation of chemical images into magnitude spectra, which are subsequently used to construct the MVI instead of the original images. The pixel values of the magnitude spectra are used as the chemical structure descriptors. This result represents a milestone in the MIA-QSAR context, as it opens way possibility for the first time of the screening for new molecular entities with the desired chemical or therapeutic utility. Future tasks include the initiative to exploit the inverse FT to identify the portions or functional groups responsible for the considered bioactivity. Additionally, the applicability of other transforms commonly used in digital image processing for example, the Walsh-Hadamard, Wavelets, and Discrete Cosine transforms in modeling structurally diverse datasets will be evaluated.

Keywords: multivariate image · multivariate image analysis · quantitative structure activity relationship · 2D-discrete Fourier transform · MCF-7 cells

How to cite this article: S. J. Barigye, M. P. Freitas. *J. Comput. Chem.* **2015**, *36*, 1748–1755. DOI: 10.1002/jcc.23992

Additional Supporting Information may be found in the online version of this article.

[1] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Vol. *1*; Wiley-VCH: Weinheim, **2009**; p. 1265.
[2] S. J. Barigye, Y. Marrero-Ponce, V. Alfonso-Reguera, F. Pérez-Giménez, *Chem. Phys. Lett.* **2013**, *570*, 147.
[3] C. R Garcia-Jacas, Y. Marrero-Ponce, S. J Barigye, J. R Valdes-Martini, O. M Rivera-Borroto, J. Olivero-Verbel, *Curr. Drug Metab.* **2014**, *15*, 441.
[4] S. J. Barigye, Y. Marrero-Ponce, Y. M. López, O. M. Santiago, F. Torrens, R. G. Domenech, J. Galvez, *SAR & QSAR Environ. Res*. **2013**, *24*, 3.
[5] S. J. Barigye, Y. Marrero-Ponce, Y. Martínez-López, F. Torrens, L. M. Artiles-Martínez, R. W. Pino-Urias, O. Martínez-Santiago, *J. Comput. Chem*. **2013**, *34*, 259.
[6] O. Martínez-Santiago, Y. Marrero-Ponce, R. Millán-Cabrera, S. J. Barigye, Y. Martínez-López, L. M. Artiles-Martínez, J. O. Guerra de León, F. Perez-Giménez, F. Torrens, *MATCH Commun. Math. Comput. Chem.* **2015**, *73*, 397.
[7] T. Puzyn, J. Leszczynski, M. T. Cronin, Recent Advances in QSAR Studies: Methods and Applications, Vol. *8*; Springer Science and Business Media: Heidelberg, **2010**; p. 428.
[8] M. P. Freitas, S. D. Brown, J. A. Martins, *J. Mol. Struct*. **2005**, *738*, 149.
[9] M. P. Freitas, *Org. Biomol. Chem*. **2006**, *4*, 1154.
[10] M. Freitas, *Med. Chem. Res*. **2007**, *16*, 461.
[11] J. E. Antunes, M. P. Freitas, R. Rittner, *Eur. J. Med. Chem*. **2008**, *43*, 1632.
[12] M. P. Freitas, E. F. F. da Cunha, T. C. Ramalho, M. Goodarzi, *Curr. Comput. Aided Drug Des*. **2008**, *4*, 273.
[13] M. Goodarzi, M. P. Freitas, *QSAR Comb. Sci*. **2008**, *27*, 1092.
[14] M. Bitencourt, M. P. Freitas, *Med. Chem*. **2009**, *5*, 79.
[15] M. Goodarzi, M. P. de Freitas, *Mol. Simul*. **2009**, *36*, 267.
[16] M. Goodarzi, M. P. Freitas, *Chemometr. Intell. Lab. Syst*. **2009**, *96*, 59.
[17] G. R. Lloret, Á. Cunha Neto, R. Rittner, M. Bitencourt, M. P. Freitas, N. S. Aquino, *J. Phys. Org. Chem*. **2009**, *22*, 1188.
[18] M. Goodarzi, M. P. Freitas, *Eur. J. Med. Chem*. **2010**, *45*, 1352.
[19] M. Goodarzi, M. P. Freitas, *Med. Chem*. **2011**, *7*, 645.
[20] J. M. Silla, C. A. Nunes, R. A. Cormanich, M. C. Guerreiro, T. C. Ramalho, M. P. Freitas, *Chemometr. Intell. Lab. Syst*. **2011**, *108*, 146.
[21] M. Bitencourt, M. P. Freitas, R. Rittner, *Arch. Pharm*. **2012**, *345*, 723.
[22] M. H. Duarte, S. J. Barigye, M. P. Freitas, *Comb. Chem. High Throughput Screen*. **2015**, *18*, 208.
[23] M. Duarte, S. Barigye, E. da Mota, M. Freitas, *SAR QSAR Environ. Res*. **2015**, *26*, 205.
[24] C. A. Nunes, M. P. Freitas, *LWT Food Sci. Technol*. **2013**, *51*, 405.
[25] M. Goodarzi, M. P. Freitas, T. C. Ramalho, *Spectrochim. Acta Part A* **2009**, *74*, 563.
[26] M. R. Freitas, S. V. B. G. Matias, R. L. G. Macedo, M. P. Freitas, N. Venturin, *J. Agric. Food. Chem*. **2013**, *61*, 8499.
[27] M. Freitas, M. Freitas, R. G. Macedo, *Bull. Environ. Contam. Toxicol*. **2014**, *93*, 489.
[28] M. R. Freitas, S. J. Barigye, M. P. Freitas, *RSC Adv*. **2015**, *5*, 7547.
[29] M. C. Guimarães, E. G. da Mota, D. G. Silva, M. P. Freitas, *Chemometr. Intell. Lab. Syst*. **2014**, *134*, 53.
[30] S. J. Barigye, M. P. Freitas, *Chemometr. Intell. Lab. Syst*. **2015**, *143*, 79.
[31] J. B. J. Fourier, Théorie analytique de la chaleur; Chez Firmin Didot, père et fils: Paris, France, **1822**.
[32] T. Akama, H. Ishida, U. Kimura, K. Gomi, H. Saito, *J. Med. Chem*. **1998**, *41*, 2056.
[33] M. G. Ferlin, G. Chiarelotto, V. Gasparotto, L. Dalla Via, V. Pezzi, L. Barzon, G. Palù, I. Castagliuolo, *J. Med. Chem*. **2005**, *48*, 3417.
[34] [34] T. S. Dowers, Z.-H. Qin, G. R. Thatcher, J. L. Bolton, *Chem. Res. Toxicol*. **2006**, *19*, 1125.
[35] V. Gasparotto, I. Castagliuolo, G. Chiarelotto, V. Pezzi, D. Montanaro, P. Brun, G. Palù, G. Viola, M. G. Ferlin, *J. Med. Chem*. **2006**, *49*, 1910.
[36] A. Gangjee, J. Yu, J. E. Copper, C. D. Smith, *J. Med. Chem*. **2007**, *50*, 3290.
[37] S. Mukherjee, S. Nagar, S. Mullick, A. Mukherjee, A. Saha, *J. Mol. Graph. Model*. **2008**, *26*, 884.
[38] M. N. Abu-Aisheh, M. S. Mustafa, M. M. El-Abadelah, R. G. Naffa, S. I. Ismail, M. A. Zihlif, M. O. Taha, M. S. Mubarak, *Eur. J. Med. Chem*. **2012**, *54*, 65.
[39] P. Dandawate, E. Khan, S. Padhye, H. Gaba, S. Sinha, J. Deshpande, K. V. Swamy, M. Khetmalas, A. Ahmad, F. H. Sarkar, *Bioorg. Med. Chem. Lett*. **2012**, *22*, 3104.
[40] D. P. Kishore, A. Rana, U. K. Jain, P. M. Rao, *Asian J. Chem*. **2013**, *25*, 10588.
[41] MATLAB 7.10.0 ed.; The MathWorks Inc., Natick, Massachusetts, **2010**.
[42] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, MOBYDIGS 1.0, Milano, Italy, **2005**.
[43] R. W. P. Urias, S. J. Barigye, Y. Marrero-Ponce, C. R. García-Jacas, J. R. Valdes-Martiní, F. Perez-Gimenez, *Mol. Divers*. **2015**, *19*, 305.
[44] K. Roy, I. Mitra, S. Kar, P. K. Ojha, R. N. Das, H. Kabir, *J. Chem. Inf. Model*. **2012**, *52*, 396.
[45] K. Roy, S. Kar, P. Ambure, *Chemometr. Intell. Lab. Syst*. **2015**, *145*, 22.
[46] R. Todeschini, V. Consonni, M. Pavan, DRAGON Software 2.1, Milano Chemometric and QSAR Research Group, Milano, Italy, **2002**.

Received: 6 April 2015
Revised: 18 May 2015
Accepted: 7 June 2015
Published online on 29 June 2015