Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry 14 (2006) 1199–1206

Bioorganic & Medicinal Chemistry

QSAR analysis of phenolic antioxidants using MOLMAP descriptors of local properties

Sunil Gupta, Susan Matthew, Pedro M. Abreu and João Aires-de-Sousa*

REQUIMTE, CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Received 6 June 2005; revised 13 September 2005; accepted 20 September 2005
Available online 17 October 2005

Abstract—Molecular maps of atom-level properties (MOLMAPs) were developed to represent the diversity of chemical bonds existing in a molecule. Chemical reactivity, being related to the ability for bond breaking and bond making, is primarily determined by the properties of bonds available in a molecule. In order to use physicochemical properties of individual bonds for an entire molecule, and at the same time having a fixed-length molecular representation, all the bonds of a molecule are mapped into a fixed-size 2D self-organizing map (MOLMAP). This article illustrates the application of MOLMAP descriptors to QSAR, with a study of the radical scavenging activity of 47 naturally occurring phenolic antioxidants. Counterpropagation neural networks (CPG NNs) were trained with MOLMAP descriptors selected using genetic algorithms to predict antioxidant activity. The model was subsequently validated by the leave-one-out (LOO) procedure obtaining a q^2 of 0.71. Random Forests were grown with the entire set of MOLMAP descriptors giving 70% of correct classifications as potent, active or inactive in a LOO experiment. Interpretations of both models in terms of discriminant variables were concordant and allowed identifying bonds and substructures that are mostly responsible for antioxidant activity. This work shows how MOLMAPs can be used for data mining of structural and biological activity data, leading to the extraction of relationships between local properties and activity.

1. Introduction

The widely distributed phenolics in higher plants, such as flavonoids, catechols, and derivatives of gallic acid, are considered dietary antioxidants. Besides antioxidant activity, phenolic molecules elicit several interesting and varied biological responses ranging from antimicrobial to immunomodulatory activities. Sergediene et al. studied prooxidant toxicity of polyphenolic antioxidants using enthalpies of single electron oxidation (quantum mechanical calculation) and attributed cytotoxicity to the ease of oxidation and lipophilicity.[1] Verma and Hansch found that different biological activities of caffeic acid derivatives are dependent on hydrophobicity or molar refractivity with a bilinear correlation.[2]

The structural requirements for radical scavenging activity are reported mostly for tannins, flavonoids, phenolic acids, and lignins. Yokozawa et al. reported that tannins are more active scavengers against diphenyl picrylhydrazyl (DPPH) radical than flavonoids. The galloyl groups enhance the activity of tannins, while the number and position of hydroxyl groups are important for the radical scavenging activity of flavonoids. The methoxylation or glycosylation of a free hydroxyl group decreased or abolished the activity of flavonoids.[3] Matsuda et al. studied the structural requirements of flavonoids for inhibition of protein glycation and reported that methylation or glucosylation of the 4-hydroxyl group decreases the activity of flavonols, flavones, and isoflavones, while methylation or glucosylation of 3-hydroxyl or 7-hydroxyl increases the activity. The flavonoids with strong inhibition of protein glycation correlated with scavenging activity for DPPH and superoxide anion radicals.[4] In another study employing different oxidants including DPPH, the presence of a catechol group was found to result in high antioxidant capacity amongst flavonoids possessing different basic structures but the same hydroxylation pattern. The flavone kaempferol, in spite of bearing no catechol group, presented a high antioxidant activity against some oxidants because of the presence of both a 2,3-double bond and a 3-hydroxyl group.[5] Heim et al. concluded that for flavonoids, multiple hydroxyl groups confer antioxidant

Keywords: Local descriptors; Chemical reactivity; Neural networks; Antioxidant activity.

* Corresponding author. Tel.: +351 21 2948300; fax: +351 21 2948550; e-mail addresses: sunil.gupta@dq.fct.unl.pt; jas@fct.unl.pt

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.09.047

activity while methoxy groups introduce unfavorable steric effects and increase lipophilicity. A double bond and carbonyl function in the heterocycle or polymerization of nuclear structure increases activity by affording a stable flavonoid radical through conjugation and electron delocalization.[6] Amic et al. built a QSAR model for the DPPH antiradical activity of 29 flavonoids using descriptors that encode the position of phenolic hydroxyl groups.[7] Saroka and Cisowski investigated the hydrogen peroxide scavenging activity of water soluble phenolic acid derivatives and found that the strongest activity is exhibited by molecules with three hydroxyl groups bonded to the aromatic ring in an *ortho* position relative to each other, followed by molecules with two hydroxyl groups bonded to an aromatic ring in *ortho* position. The molecules with two hydroxyl groups bonded to the aromatic ring in *meta* position were next, followed by molecules with one hydroxyl group exhibiting the lowest antioxidant activity.[8] Dizhbite et al. studied the DPPH radical scavenging activity of lignins and proposed that non-etherified hydroxyl phenolic group, *ortho* OMe groups, and the double bond between outermost carbon atoms in the side chain increase scavenger activity.[9]

The scavenging of various free radicals has been explained using structural features of antioxidants. Tyrakowska et al. reported that the trolox equivalent antioxidant capacity (TEAC) of 4-hydroxybenzoates is not determined by the tendency of the molecule to donate an electron but by its ability to donate a hydrogen atom.[10] Lien et al. used the heat of formation, E_{LUMO}, E_{HOMO}, and number of hydroxyl groups to develop a model for phenolic antioxidants and found that the number and location of hydroxyl groups govern the TEAC of the flavonoid ring system.[11] Zhang et al. reported that the bond dissociation energy (BDE) of O–H correlates well with the logarithm of the peroxy radical scavenging rate constant for phenolic antioxidants. Although the O–H charge difference and E_{HOMO} can determine the O–H BDE for simple phenols, they are invalid when the phenols possess intramolecular hydrogen bonds.[12] Zhang and Wang used the density function theory to characterize peroxyl radical scavenging by coumarins and thiaflavins. They reported that H-atom transfer is the preferred mechanism of scavenging and can be measured by O–H bond dissociation enthalpy.[13] Cheng et al. reported a correlation between hydroxyl radical scavenging activity and OH bond strength, electron-donating ability, enthalpy of single electron transfer, and spin distribution of phenoxyl radicals after H-abstraction.[14] Sadeghipour et al. evaluated the ability of polyphenols for inhibition of peroxynitrite-induced nitration of tyrosine using heat of formation and found the flavonoids with 3,4-hydroxyl substructures to be the most effective.[15]

Chemical reactivity, being related to the ability for bond breaking and bond making, is primarily determined by the properties of bonds available in a molecule. Gasteiger and co-workers[16] have proposed seven empirical physicochemical properties of chemical bonds for modeling chemical reactivity: difference of sigma electronegativity between the two atoms of the bond (DENSIG—$\Delta\chi_{\sigma}$), difference of total atomic charge (DQTOT—Δq_{tot}), difference of pi atomic charge (DQPI—Δq_{π}), mean bond polarizability (BPOLARIZ—α_{b}), bond dissociation energy (BDE), resonance stabilization (STABRS—$R\pm$), and bond polarity (SQIT—Q_{σ}). In order to use all that information for an entire molecule, and at the same time having a fixed-length representation, we mapped all the bonds of a molecule into a fixed-length 2D self-organizing map.

A self-organizing map (SOM) must be trained beforehand with a diversity of bonds from different structures (each bond described by the seven bond properties calculated by PETRA[17]). Then all the bonds of one molecule are submitted to the trained SOM, and the pattern of activated neurons is a map of the reactivity features of that molecule (MOLMAP)—a fingerprint of the bonds available in that structure. Such MOLMAP (molecular maps of atom-level properties) descriptors can be directly used in QSAR studies related to chemical reactivity, in situations involving different types of reaction sites in a single data set, more than one reaction site in a single structure, or unknown reaction sites. In this paper, we illustrate the application of MOLMAP descriptors to the prediction of DPPH radical scavenging activity of naturally occurring phenolic antioxidants.

Being rapid, simple, and independent of sample polarity, the DPPH method is very convenient for the quick screening of samples for radical scavenging activity.[18] The free radical scavenging activity was measured as the concentration required for inhibiting DPPH radical formation by 50%. Models were developed for prediction of radical scavenging activity with counterpropagation neural networks (CPG NN)[19] and Random Forests.[20]

2. Conclusion

MOLMAP descriptors encode local aspects of a chemical structure, exclusively on the basis of physicochemical properties, in a fixed-length code. The resolution of the code can be adjusted to the universe of the data set to investigate. MOLMAP descriptors are easily correlated with local structural features, without requiring the explicit detection of substructures. Their use by machine learning techniques such as neural networks or Random Forests for QSAR applications can lead to the identification of structural features responsible for activity. Exploring a data set consisting of the antioxidant activity of 47 natural products related to cinnamic acid and benzoic acid, MOLMAP descriptors were selected by genetic algorithms and used to train CPG NNs that yielded a LOO q^2 value of 0.712 between predicted and experimental IC_{50} value. Random Forests were grown with the entire set of MOLMAP descriptors giving 70% of correct classifications as potent, active or inactive in a LOO experiment. Interpretation of the models allowed for the identification of the 3,4-hydroxyl substitution pattern of aromatic rings as a typical motif contributing to high activity, and 3-OMe, 4-OH as a motif leading to mid-range activity. This work illustrates how the new MOLMAP approach can be used for data mining of structural and biological activity data, and for the extraction of relationships between local properties and activity.

3. Methods

The experiments here described required two major steps, the generation of the descriptors and the development of predictive models. Generation of the descriptors, the so-called MOLMAPs, was obtained by a Kohonen self-organizing map.[19] For training this map, each object of the training set was a chemical bond, represented by seven empirical physicochemical properties. This training set comprised bonds from a diversity of structures. Once trained, the map was used to obtain molecular descriptors—all the bonds of one molecule were submitted to the map and the resulting pattern of activated neurons (the MOLMAP) was the descriptor of the molecule. The second step consisted of establishing relationships between MOLMAPs and radical scavenging activity. The details are explained below.

3.1. Data set

A set of 47 natural products, evaluated by DPPH assay, was compiled from the literature (Table 1). Some of these molecules are related to cinnamic acid and were isolated and characterized at our department,[21] and were evaluated for antioxidant activity by the DPPH assay. The antioxidant activity was reported as IC_{50} (micromolar concentration required to inhibit the DPPH radical by 50%). The IC_{50} values were converted to micromoles to make the activity data homogenous and directly comparable. Since the radical scavenging activity varied by orders of magnitude, it was transformed into natural log values. The molecules were classified as potent if they exhibited $IC_{50} < 3.5$ (17 molecules); active if IC_{50} was between 3.5 and 5.5 (19 molecules) and inactive if IC_{50} was 5.5 or more (11 molecules).

3.2. Training of a Kohonen self-organizing map with bonds

Kohonen self-organizing maps (SOM) can be used for the reduction of multidimensional objects to 2D.[19] In this study, we used SOMs to reduce to 2D the dimension of chemical bonds, represented by seven empirical physicochemical properties calculated by PETRA 3.20[17]—bond dissociation energy (BDE), resonance stabilization (STABRS—$R\pm$), difference of sigma electronegativity between the two atoms of the bond (DENSIG—$\Delta\chi_\sigma$), difference of total atomic charge (DQTOT—Δq_{tot}), difference of pi atomic charge (DQPI—Δq_π), mean bond polarizability (BPOLARIZ—α_b), and bond polarity (SQIT—Q_σ). The values for each property were linearly scaled between 0.1 and 0.9. As some properties depend on the orientation of the bond, each bond was represented twice (as A–B and B–A). In order to focus on hydroxyl groups attached to aromatic systems (mainly responsible for reactivity) only bonds were considered that include an oxygen atom belonging to an aromatic hydroxyl group.

SOMs learn by unsupervised training, revealing similarities between objects (bonds). A Kohonen SOM consists of a grid of so-called neurons, each containing as many elements (weights) as there are input variables. Here, the input variables are the seven properties of bonds. Before the training starts, the weights take random values. During the training, each individual bond is mapped into the neuron that contains the most similar weights compared to its properties. This is the central neuron, or winning neuron—Figure 1. It is said that the winning neuron was activated by the bond, and its weights are then adjusted to make them even more similar to the properties of the presented bond. Not only the winning neuron has its weights adjusted, but also the neurons in its neighborhood. The extent of adjustment depends, however, on the topological distance to the winning neuron—the closer a neuron is to the central neuron the larger is the adjustment of its weights. The objects of the training set are iteratively fed to the map, the weights corrected, and the training is stopped when a pre-defined number of cycles are attained. A trained Kohonen SOM will reveal similarities in the objects of a data set in the sense that similar objects (similar bonds) are mapped into the same or closely adjacent neurons. A self-organizing map with 12×12 neurons was trained with 428 bonds extracted from the 47 molecular structures. The initial learning span was set at five and the network was trained over 50 epochs. SOMs were implemented with in-house developed software based on JATOON Java applets.[22,23]

3.3. Molecular descriptors (molecular MOLMAPs)

A representation of the set of bonds existing in a molecule can be obtained by mapping the bonds of that molecule on the SOM previously trained with a diversity of bonds. The pattern of activated neurons (Fig. 2) can be interpreted as a fingerprint of the reactivity of the molecule, and it was used as a molecular descriptor (MOLMAP). For numerical processing, each neuron got a value equal to the number of times it was activated by bonds of the molecule. The map was then transformed into a vector by concatenation of columns resulting in a fixed-length ($12 \times 12 = 144$) MOLMAP descriptor for each molecule. In order to account for the relationship between similarity of bonds and proximity in the map, a value of 0.3 was added to each neuron multiplied by the number of times a neighbor was activated by a bond (Fig. 2). In the following sections, we call each of these neurons a component of the MOLMAP, and we follow the terminology according to which an activated neuron contributes to its neighboring components. Figure 2 illustrates the set of bonds mapped on the SOM previously trained with a diversity of bonds and the generation of a MOLMAP for nepetoidin B (**A14**). The top left neuron of the MOLMAP is defined as component (1, 1) and the bottom right as (12, 12).

Table 1. Experimental ln(IC_{50}) values and predictions by counterpropagation neural networks (CPG NNs) and random forests (RF)

ID	Molecule	CPG	RF	Exptl.	Ref.
A01	Piceatannol	5.70	4.66	5.64	28
A02	*trans*-Resveratrol	5.34	4.76	6.03	28
A03	Scirpusin A	5.72	3.59	5.11	28
A04	Olivil	4.09	4.03	4.64	28
A05	(−)-Carinol	4.22	4.40	3.78	28
A06	(+)-Cycloolivil	3.24	3.91	4.34	28
A07	2-Hydroxy-6-methoxybenzoic acid	7.22	6.13	7.08	28
A08	Caffeic acid	3.24	3.02	2.95	29
A09	Caftaric acid	3.13	3.03	3.02	29
A10	Chlorogenic acid	3.11	3.07	2.94	29
A11	Cyanarin	2.38	2.64	2.41	29
A12	Echinacoside	2.79	2.86	1.89	29
A13	Cichoric acid	2.29	2.73	2.15	29
A14	Nepetoidin B	2.48	2.97	0.96	30
A15	Gallic acid	3.07	4.35	1.55	30
A16	Rosmarinic acid	2.73	2.97	1.65	30
A17	Curcumin	4.19	4.84	1.03	31
A18	Demethoxy curcumin	3.58	3.75	3.67	31
A19	Bisdemethoxy curcumin	3.46	2.48	5.73	31
A20	Rosmarinic methyl ester	2.56	2.52	2.57	32
A21	Dihydroferulic acid	4.80	4.65	4.34	33
A22	Ferulic acid	3.92	4.87	4.74	33
A23	Sinapic acid	3.63	4.85	4.35	33
A24	Dihydrosinapic acid	3.73	4.31	3.79	33
A25	Vanillic acid	7.12	5.25	5.52	33
A26	*p*-Hydroxycinnamic acid	6.62	5.86	7.66	33
A27	Sargachromenol	3.85	4.13	2.55	34
A28	Sargachromenol methyl ester	3.87	4.05	2.76	34
A29	Sargahydroquinoic acid	3.63	3.40	1.81	34
A30	Sargahydroquinoic methyl ester	3.24	2.85	2.85	34
A31	6-*O*-Acetyl-martynoside	5.17	5.17	5.30	35
A32	Wiedemannioside B	5.10	5.17	5.30	35
A33	Wiedemannioside C	5.05	5.03	5.30	35
A34	Wiedemannioside D	4.88	5.03	5.30	35
A35	Wiedemannioside E	5.12	5.03	5.30	35
A36	Acetoside/verbascoside	2.14	1.95	4.14	35
A37	Martynoside	5.17	5.19	5.30	35
A38	Methyl 4-*O*-β-D-glucopyranosylcaffeate	3.56	4.69	4.79	21
A39	1-*O*-Caffeyl-β-D-glucopyranoside	2.89	2.82	3.71	21
A40	Glucosylcinnamate derivative	7.13	5.23	7.16	21
A41	Glucosylcaffeate derivative	4.13	5.00	4.18	21
A42	β-D-Glucopyranosyl 4-hydroxybenzoate	7.06	6.49	7.86	21
A43	Glucosylbenzoate derivative	7.16	6.85	7.40	21
A44	2-Phenylethyl-β-D-glucopyranoside	7.00	6.66	7.94	21
A45	5-*O*-Caffeyl quinic acid	3.26	3.20	2.52	3
A46	Yunnaneic acid C	3.30	4.46	2.62	3
A47	Bergenin	4.78	3.55	6.21	3

3.4. Selection of descriptors by genetic algorithm

For the selection of MOLMAP descriptors (or components), evolution of a population was simulated using genetic algorithms.[24] Each individual of the population represented a subset of components and was defined by a chromosome of binary values. At the beginning of the evolution, the chromosomes were assigned random values. In each generation, half of the individuals mated (the fittest individuals), and the other half died. The chromosomes of the offspring resulted from crossover of their parents' chromosomes, followed by mutation. The fitness (scoring) of each chromosome was evaluated by the ability of the corresponding subset of descriptors to predict antioxidant activity with counterpropagation neural networks. In these experiments, the CPG NN was trained with the whole data set and predictions were obtained for all the molecules. The individuals with higher scores were allowed to mate. A population size of 50 individuals was allowed to evolve over 100 generations and the chromosome of the best individual was used for model development.

3.5. Model development with CPG NNs

CPG neural networks were used to model the relationship between the MOLMAP descriptors of antioxidants and the corresponding IC_{50} values. The 25 MOLMAP

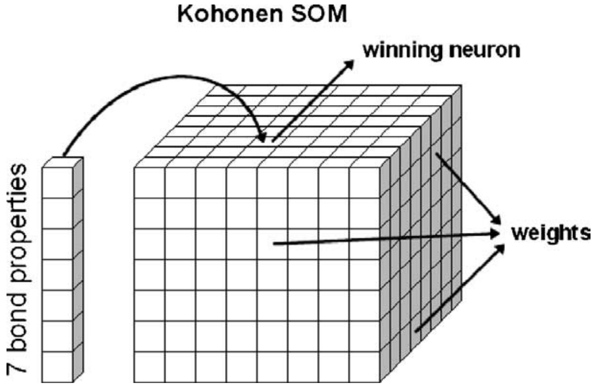


Figure 1. Representation of the input of a bond to a Kohonen self-organizing map and activation of the winning neuron. This is the SOM on which generation of the MOLMAP descriptors is based.

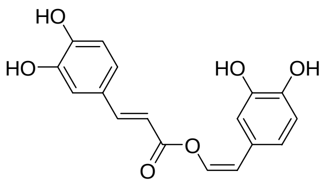
descriptors selected by genetic algorithms were used to train CPG NNs of size 10 × 10 over 50 epochs with an initial learning span of two. Leave-one-out (LOO) method was used to predict the IC_{50} values because the data set was too small to be divided into training and test sets. A CPG NN consists of two layers of neurons. The first layer is a Kohonen map, and this is responsible for choosing the winning neuron. It stores information concerning the input data. The second layer stores information concerning output (IC_{50} value). Note that this map is used for a completely different purpose than the Kohonen map mentioned in Sections 3.2 and 3.3. During the training, each time a winning neuron is chosen, its weights at both the input and output layers are adjusted to make them more similar to the presented data. The trained network can then make predictions for the IC_{50} value of one molecule when the descriptors are submitted as input—the winning neuron is chosen and the correspondent value in the output layer is taken as the predicted IC_{50} value. We averaged the predicted IC_{50} values from 10 random leave-one-out experiments to develop quantitative structure–activity relationships because different runs of the training procedure yield different networks (due to random initialization of weights and order of presentation of objects).

3.6. Model development with Random Forests[20,25]

A Random Forest is an ensemble of unpruned regression trees created by using bootstrap samples of the training data and random subsets of variables to define the best split at each node. It is a high-dimensional nonparametric method that works well on large numbers of variables. It has been shown that the method is extremely accurate in a variety of applications. Additionally, the method quantifies the importance of a variable by the increase in standard error when the values of the variable are randomly permuted. In this study, Random Forests were grown with the R program version 2.0.1[26] using the randomForest library.[27] They were used to predict the IC_{50} values of antioxidants from the MOLMAP of the molecules without previous selection of MOLMAP components. A LOO procedure was followed as for the CPG NNs, with 47 forests grown on the basis of 46 molecules.

4. Results and discussion

CPG NNs were trained to predict IC_{50} values on the basis of 25 MOLMAP descriptors selected by genetic algorithms. Each molecule was thus submitted to the CPG NN in the form of a 25-dimension vector. The IC_{50} values predicted by counterpropagation networks correlated with the experimentally observed IC_{50} values exhibiting a q^2 of 0.712 ($n = 47$) and RMS error of



a

0	0	0	0	0	0	1	0	0	1	0	0
1	1	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	1	0	0
1	0	1	0	1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	1	0	0
0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0

b

0.6	0.6	0.3	0.0	0.0	0.6	1.3	0.6	0.6	1.3	0.6	0.3
1.3	1.3	0.3	0.0	0.0	0.3	0.3	0.3	0.6	1.3	0.6	0.3
0.6	0.6	0.3	0.0	0.0	0.0	0.3	0.3	0.9	0.6	0.6	0.3
0.3	0.6	0.3	0.6	0.3	0.3	0.3	1.0	0.6	1.0	0.3	0.3
1.0	0.6	1.0	0.9	1.3	0.6	0.3	0.3	0.6	0.3	0.3	0.3
0.3	0.6	0.3	0.9	1.3	0.6	0.0	0.0	0.0	0.0	0.0	0.3
0.0	0.0	0.0	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.6	0.3	0.3	0.0
0.0	0.6	0.6	0.6	0.0	0.0	0.3	1.0	0.6	1.0	0.3	0.0
0.0	0.6	2.0	0.6	0.0	0.0	0.3	0.3	0.6	0.3	0.3	0.0
0.0	0.6	0.6	0.6	0.0	0.3	0.3	0.3	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.6	1.3	0.6	0.3	0.3	0.3	0.0

Figure 2. (a) Pattern of neurons activated by bonds of molecule **A14** on the 12 × 12 self-organizing map. (b) Numerical transformation of the pattern into a MOLMAP.

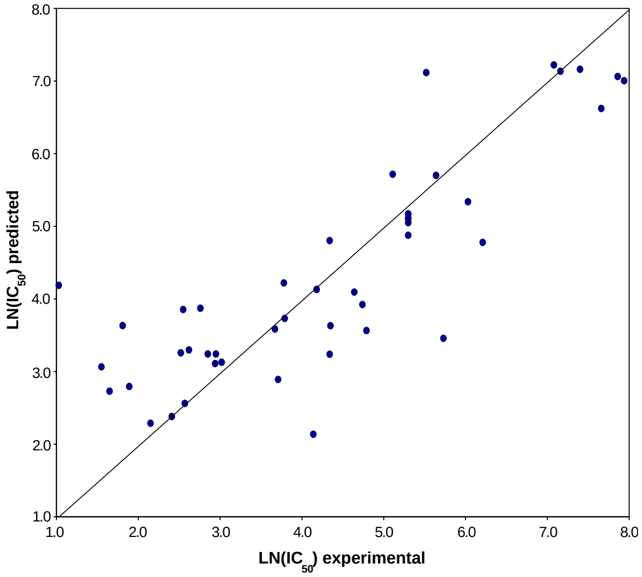


Figure 3. The correlation of experimental LN (IC_{50}) values with predictions by CPG NNs using leave-one-out procedure.

1.001 (Table 1 and Fig. 3). The omission of **A17** (squared-residual > three standard deviations) and **A19** (squared-residual > two standard deviations) resulted in q^2 of 0.790 ($n = 45$) and RMS error of 0.842. The molecules with squared residuals within one standard deviation result in q^2 of 0.838 ($n = 43$) between the predicted and experimental IC_{50} values (RMSE = 0.756).

CPG NNs trained with MOLMAP descriptors not only make predictions but can also highlight relationships between structural features and activity. Inspection of a CPG NN trained with all the objects (Fig. 4) revealed that structurally similar molecules were mapped together in one neuron or as a cluster. In Figure 4, the neurons with a low output (low IC_{50} values) are colored in red and neurons corresponding to high IC_{50} values are colored in blue. The molecules **A08**, **A09**, **A10**, **A39**, and **A45** predicted to be potent were mapped into the same neuron at the top left. Also **A12**, **A16**, **A20**, and **A36** predicted to be potent were appearing in a single neuron. Seven of these nine molecules were experimentally observed to be potent and two active (**A36** and **A39**). Both neurons exhibited a particularly high weight at the seventh layer that corresponds to component (3, 10) of the MOLMAPs. It was observed that this component is activated by C–O bonds of an OH group at *meta* position in the aromatic ring and adjacent to another hydroxyl group at *para* position. The component (3, 10) of the MOLMAPs was activated by C–O bonds from 12 molecules—two active (**A36** and **A39**) and 10 potent compounds.

Molecules **A17**, **A18**, **A22**, **A23**, and **A24** were clustered together (at the mid-bottom of the CPG NN surface in Fig. 4) and predicted to be active. Four of the five compounds were correctly predicted to be active, while one is potent (**A17**). The first and second layers of the neurons in this region had high values. These layers correspond to components (1, 1) and (1, 6) of the MOLMAPs, respectively. Significantly, both layers are linked to the same structural feature—a hydroxyl group at *para* position in the aromatic ring and adjacent to a methoxy group. Component (1, 1) was activated by C–O bonds at *para* position from six molecules (**A15**, **A17**, **A18**, and **A22–24**). All the six molecules exhibited $IC_{50} < 4.75$ and were predicted to be active. Except for **A15**, these compounds have a hydroxyl group at *para* position and a methoxy group on the adjacent position. Component (1, 6) corresponds to O–H bonds at *para* position and adjacent to a methoxy group. It was activated by O–H bonds from eight molecules, all of them active.

The molecules **A07**, **A26**, **A40**, **A42**, **A43**, and **A44** predicted inactive were clustered into the dark blue neurons. All the six molecules were correctly predicted as inactive. Here, a clear common structural feature could not be found.

All the 144 MOLMAP components were submitted to Random Forests as molecular descriptors, and models were developed to estimate antioxidant activity. In contrast to CPG NNs, no previous selection of variables was performed, as Random Forests can safely work with a large number of variables and have their intrinsic variable selection approach. With the LOO procedure, the IC_{50} values were predicted exhibiting q^2 of 0.495 ($n = 47$) and RMS error of 1.324 (Table 1). The omission of **A17** and **A19** (squared-residual > three standard deviations) resulted in q^2 of 0.639 ($n = 45$) and RMS error of 1.129. The subset of molecules with squared-residuals within one standard deviation yielded q^2 of 0.789 ($n = 42$) between the predicted and experimental IC_{50} values (RMSE = 0.947). Although the results are quantitatively inferior to those obtained by the CPG NNs, they were qualitatively reasonable, with more than

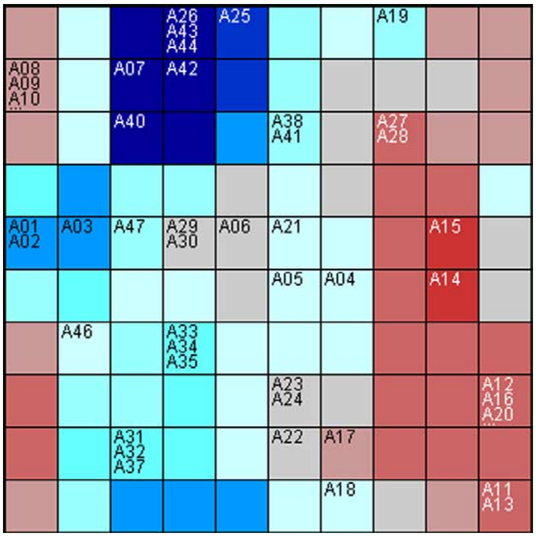


Figure 4. Representation of the output layer of the CPG NN, with neurons predicting high activity in red and those predicting inactivity in blue. The compounds **A08**, **A09**, **A10**, **A39**, and **A45** predicted to be potent were mapped into the same neuron, and the same happened with **A12**, **A16**, **A20**, and **A36**.

70% of molecules correctly classified as potent, active or inactive. Amongst the remaining 30%, the molecules were classified in adjacent classes and only one inactive molecule **A19** was wrongly classified as potent.

Random Forests assess the importance of variables by measuring the increase in mean standard error when a variable is randomly permuted. The results went in line with the interpretation of CPG NNs. Indeed, component (3, 10) of the MOLMAPs was within the 10 most important variables (in 144), as well as two neighbours of component (1, 1)–component (2, 1) was the most important and component (2, 12) was ranked as fifth. C–O bonds from 32 molecules contributed to component (2, 1), 94% of which exhibit $IC_{50} < 5.5$. The third most important component was activated by O–H bonds from 12 molecules, at the *para* position in an aromatic ring. Significantly, 7 of these 12 bonds also activated component (1, 6) that has revealed importance from the analysis of CPG NNs. As mentioned in the methodology, each bond is included twice with opposite orientations. These seven bonds activate component (1, 6) when oriented as ‘H–O’ and activate component (9, 5) when oriented as ‘O–H’. The second most important component—(9, 10)—received contributions from C–O bonds belonging to 19 molecules, 89% of which had an $IC_{50} < 4.8$.

When ranked according to increasing IC_{50} values predicted by CPG NNs, the first nine molecules (**A11-16**, **A20**, **A36**, and **A39**) possessed C–O bonds at the *meta* position (adjacent to a hydroxyl group at the *para* posi-

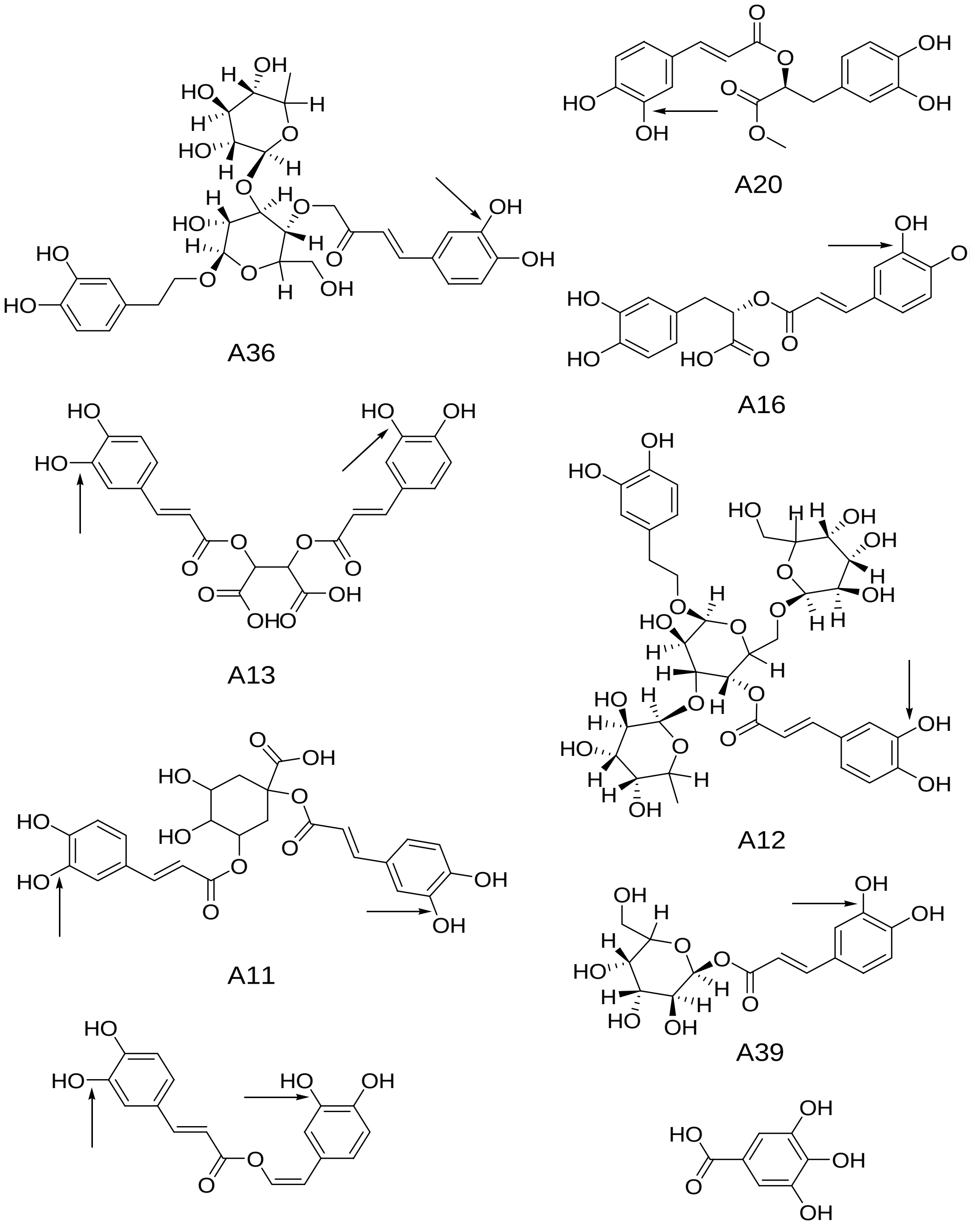


Figure 5. The nine molecules with the lowest predicted IC_{50} values by the CPG NNs, with bonds activating MOLMAP component (3, 10) highlighted by arrows.

tion) and activated MOLMAP component (3,10), except for **A15** (Fig. 5). This result is in agreement with the existing knowledge that the presence of catechol groups contributes to high antioxidant activity.[5,8,15] On the other hand, the molecules predicted inactive had bonds activating components scattered all over the MOLMAPs.

Curcumin (**A17**) was predicted active although it is potent and bisdemethoxy curcumin (**A19**) was predicted potent although it is inactive. These two problematic cases can be rationalized in terms of data available for model generation. Molecule **A17** was predicted active because molecules **A18** and **A22** with similar structural features (thus activating neighbouring neurons) were active—this is apparent in Figure 4. Compound **A19** exhibits hydroxyl groups on two adjacent aromatic carbon atoms at *para* and *meta* positions, which was perceived as a factor contributing to potent activity. However, **A19** is inactive. If an error in the experimental data is excluded, this result points toward factors responsible for the inhibition of the antioxidant activity that are not encoded by the used MOLMAP descriptors.

Acknowledgments

S. Gupta and S. Mathew acknowledge Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for financial support under postdoctoral grants (SFRH/BPD/14475/2003, SFRH/BPD/8570/2002). The authors thank Molecular Networks GmbH for providing access to PETRA software package. The assistance of Qing-You Zhang and Gonçalo Carrera in the implementation of MOLMAP descriptors is gratefully acknowledged.

References and notes

1. Sergediene, E.; Jonsson, K.; Szymusiak, H.; Tyrakowska, B.; Rietjens, I. M. C. M.; Cenas, N. *FEBS Lett.* **1999**, *462*, 392.
2. Verma, R. P.; Hansch, C. *Chembiochem.* **2004**, *5*, 1188.
3. Yokozawa, T.; Chen, C. P.; Dong, E.; Tanaka, T.; Nonaka, G. I.; Nishioka, I. *Biochem. Pharmacol.* **1998**, *56*, 213.
4. Matsuda, H.; Wang, T.; Managi, H.; Yoshikawa, M. *Bioorg. Med. Chem.* **2003**, *11*, 5317.
5. Silva, M. M.; Santos, M. R.; Caroço, G.; Rocha, R.; Justino, G.; Mira, L. *Free Radical Res.* **2002**, *36*, 1219.
6. Heim, K. E.; Tagliaferro, A. R.; Bobilya, D. J. *J. Nutr. Biochem.* **2002**, *13*, 572.
7. Amic, D.; Davidovic-Amic, D.; Beslo, D.; Trinajstic, N. *Croat. Chem. Acta* **2003**, *76*, 55.
8. Saroka, Z.; Cisowski, W. *Food Chem. Toxicol.* **2003**, *41*, 753.
9. Dizhbite, T.; Telysheva, G.; Jurkjane, V.; Viesturs, U. *Bioresource Technol.* **2004**, *95*, 309.
10. Tyrakowska, B.; Soffers, A. E. M. F. S.; Szymusiak, H.; Boeren, S.; Boersma, M.; Lemanska, K.; Vervoort, J.; Rietjens, I. M. C. M. *Free Radical Biol. Med.* **1999**, *27*, 1427.
11. Lien, E. J.; Ren, S.; Bui, H. H.; Wang, R. *Free Radical Biol. Med.* **1999**, *26*, 285.
12. Zhang, H. Y.; Sun, Y. M.; Zhang, G. Q.; Chen, D. Z. *Quant. Struct. Act. Rel.* **2000**, *19*, 375.
13. Zhang, H. Y.; Wang, L. F. *J. Mol. Struct. (Theochem)* **2004**, *673*, 199.
14. Cheng, Z.; Ren, J.; Li, Y.; Chang, W.; Chen, Z. *Bioorg. Med. Chem.* **2002**, *10*, 4067.
15. Sadeghipour, M.; Terreux, R.; Phipps, J. *Toxicol. In vitro* **2005**, *19*, 155.
16. Simon, V.; Gasteiger, J.; Zupan, J. *J. Am. Chem. Soc.* **1993**, *115*, 9148.
17. <http://www2.chemie.uni-erlangen.de/software/petra/>
18. Koleva, I. I.; van Beek, T. A.; Linssen, J. P. H.; Groot, A.; Evstatieva, L. N. *Phytochem. Anal.* **2002**, *13*, 8.
19. Zupan, J.; Gasteiger, J.. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.
20. Breiman, L. *Mach. Learn.* **2001**, *45*, 5.
21. Braham, H.; Mighri, Z.; Ben-Jannet, H.; Matthew, S.; Abreu, P. M. *J. Nat. Prod.* **2005**, *68*, 517.
22. Aires-de-Sousa, J. *Chemometr. Intell. Lab. Syst.* **2002**, *61*, 167.
23. <http://www.dq.fct.unl.pt/staff/jas/jatoon/>
24. Homeyer, A. V. Evolutionary Algorithms and their Applications in Chemistry. In *Handbook of Chemoinformatics*; Gasteiger, J., Engel, J., Eds.; Wiley-VCH: New York, 2003; Vol. 3, p 1239.
25. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947.
26. R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>
27. Fortran original by Leo Breiman, Adele Cutler, R port by Andy Liaw and Matthew Wiener. 2004. <http://www.stat.berkeley.edu/users/breiman//>
28. Lee, S. K.; Mbwambo, Z. H.; Chung, H.; Luyengi, L.; Gamez, E. J. C.; Mehta, R. G.; Kinghorn, A. D.; Pezzuto, J. M. *Comb. Chem. High Throughput Screen.* **1998**, *1*, 35.
29. Pellati, F.; Benvenuti, S.; Magro, L.; Melegari, M.; Soragni, F. *J. Pharm. Biomed. Anal.* **2004**, *35*, 289–301.
30. Grayer, R. J.; Eckert, M. R.; Veitch, N. C.; Kite, G. C.; Marin, P. D.; Kokubun, T.; Simmonds, M. S. J.; Paton, A. *J. Phytochem.* **2003**, *64*, 519–528.
31. Song, E. K.; Cho, H.; Kim, J. S.; Kim, N. Y.; An, N. H.; Kim, J. A. *Planta Med.* **2001**, *67*, 876–877.
32. Parejo, I.; Caprai, E.; Bastida, J.; Viladomat, F.; Jáuregui, O.; Codina, C. *J. Ethnopharmacol.* **2004**, *94*, 175–184.
33. Shimoji, Y.; Tamura, Y.; Nakamura, Y.; Nanda, K.; Nishidai, S.; Nishikawa, Y.; Ishihara, N.; Uenakai, K.; Ohigashi, H. *J. Agric. Food Chem.* **2002**, *50*, 6501–6503.
34. Pérez, C. A. L.; Arciniegas, A.; Ramírez Apan, M. T.; Villaseñor, J. L.; Romo de Vivar, A. *Planta Med.* **2002**, *68*, 645–647.
35. Abougazar, H.; Bedir, E.; Khan, I. A.; Calis, I.; Wiedemanniosides, A.-E. *Planta Med.* **2003**, *69*, 814–819.