Abstract

Flavonoids are bioactive compounds that can display antioxidant activity. Their must important source is the vegetal kingdom. Their composition in different foods is compiled into several databases organized by USDA. This information enabled the creation of a data record that was used in the work to predict the total antioxidant capacity of food by the oxygen radical absorbance capacity (ORAC) method, using algorithms of artificial intelligence. K-Nearest Neighbors (KNN) was used. The attributes were: a) amount of flavonoid, b) class of flavonoid, c) Trolox equivalent antioxidant capacity (TEAC) value, d) probability of clastogenicity and clastogenicity classification by Quantitative Structure-Activity Relationship (QSAR) method and e) total polyphenol (TP) value. The selected variable to predict was the ORAC value. For the prediction, a cross-validation method was used. For the KNN algorithm, the optimal K value was 3, making clear the importance of the similarity between objects for the success of the results. It was concluded the successful use of the KNN algorithm to predict the antioxidant capacity in the studied food groups.

Keywords: Flavonoid, Oxygen radical absorbance capacity (ORAC) method, Artificial intelligence, K-Nearest Neighbors (KNN) algorithm.
**Introduction**

In the recent years, database including information about the emerging food composition database were created (1-3). These BDCA are centred in the composition of bioactive substances (non-nutrients) including flavonoids. Flavonoids are present in several sources in the vegetal kingdom and display a large range of biological properties. It is already proved their benefits for health. Therefore, their study is a topic of interest (2, 3). The most important activity is related to their antioxidant capacity (1, 4-6). A substance with antioxidant capacity, even in small amounts comparing to the substrate, is able to decrease the oxidation of that substrate (7). The antioxidant activity is correlated to the prevention of chronic diseases of high prevalence in different countries (8).

The food composition database of flavonoids has huge chemical information due to the structural diversity of the compounds included on it. This database provides researchers with new values on the flavonoid content of many more foods in order to better as certain the impact of flavonoid consumption on various chronic diseases (2, 3).

This project was developed taking into account the possibility of generating predictive information related to the information found in the food composition database. In particular, we were looking for a tool to predict the antioxidant capacity of food containing different compounds with flavonoid scaffold (dietary exogenous antioxidants). This project was focus on the idea that a dietary antioxidant is a substance in foods that significantly decreases the adverse effects of reactive species, such as reactive oxygen and nitrogen species, on normal physiological function in humans (9).

The data regarding the composition of food is complex and extensive (10). It is hard to process all the information regarding the different assays presented in literature. This variability transforms this study in a complex system. However, the processing of the information is still performed by classic statistical methodologies (11, 12). When the problem is complex and mediated by non-lineal behaviours, it could be studied either by a multivariate perspective or by using artificial intelligence technics (13). In particular, the artificial neuronal networks (ANN) are able to develop a predictive model that automatically includes relationships between the analysed variables with no necessity of included them in the model.

In the biomedical field, several unidirectional supervised networks were used, specially based on the Multi Layered Perceptron (MLP) (13). However, as far as we know, these technics were never been used related to food composition database. The current work is centred on the development of an artificial intelligence algorithm that allow the prediction of the total antioxidant capacity of the food, based on quantitative information, topologic-structural and the bioactivity of flavonoids.

**Methods**

In **Figure 1** it is shown a general scheme of the methodology used in this study. It is divided in the next steps:

1. **Conformation of the register of the data related to the food composition.**
The information obtained in different food composition database (i.e. database for the flavonoid content of selected foods and database for the isoflavone content of selected foods) (2, 3) was used to prepare the register of the data related to the composition of flavonoids in different foods.

2. **Procedure for the prediction using artificial intelligence algorithms**

To obtain the training set and test set it was used the cross validation of k10 iterations. The KNN (K-Nearest Neighbours) was used. This algorithm is implemented in the PROCONS software version 4.0.

The attributes were: i) amount of flavonoid (mean), ii) class of flavonoid, iii) trolox equivalent antioxidant capacity value \((\text{TEAC}_{\text{exp}})\), iv) probability of clastogenicity and clastogenicity classification by Quantitative Structure-Activity Relationship (QSAR) method and e) total polyphenol value \((\text{TP}_{\text{exp}})\). These experimental parameters were taken from the scientific literature. A different weight was assigned to each attribute. It was realized manually and using the Particle Swarm Optimization (PSO) method, implemented in the PROCONS software (PSO+RST (Rougt Set Theoryc)) (15).

The variable selected to predict was the oxygen radical absorbance capacity \((\text{ORAC}_{\text{exp}})\) value, expressed in \(\mu\text{mol TE/100 g}\). ORAC was selected because it is considered to be the preferable methodology to evaluate the antioxidant capacity due to its biological relevance to the in vivo antioxidant efficacy (16). \(\text{ORAC}_{\text{exp}}\) and \(\text{TP}_{\text{exp}}\) (mg GAE/100 g) for each substrate were found in the literature. The analytical method developed by Prior et al was used as the reference method for select published sources (17).
Figure 1. General scheme of the applied methodology.

Consulting of BDCA for the flavonoid content of selected foods and database for the isoflavone content of selected foods

Preparation of the new database

ATTRIBUTES FOR ARTIFICIAL NEURAL NETWORKS
1. Amount of flavonoid
2. Class of flavonoid
3. Trolox equivalent antioxidant capacity (TEAC)
4. Probability of clastogenicity
5. Clastogenicity classification by QSAR method
6. Total polyphenol (TP) value

KNN algorithm - Nearest Neighbors technique

The variable to predict: ORAC food value
Results and Discussion

The studied food was divided in 11 groups (Figure 2). The vegetables, vegetable and spices, and herbs, are the groups with more flavonoid-containing food: 39% and 37%, respectively.

The monomeric dietary flavonoids present in the studied data are from the chemical subclasses: flavonols, flavones, flavanones, flavan-3-ols (Table 1). Flavonoids from the anthocyanidin subclass can be found in several aliments. However, they were not included in this study due to structural that invalided the application of TOPSMODE approach.

![Figure 2. Quantities of aliments (divided by alimentary groups) presented in the database.](image)

Table 1. Examples of flavonoids presented in the database.

<table>
<thead>
<tr>
<th>Names</th>
<th>SMILE</th>
<th>Examples of sources (NDB No.)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-)-Epicatechin 3-gallate</td>
<td>C1C(C(OC2=CC(=CC(=C21)O)O)C3=CC(=C(C=C3)O)O)OC(=O)C4=CC(=C(C(=C4)O)O)O</td>
<td>Apples, Fuji, raw, with skin (NDB No., 97066)</td>
</tr>
<tr>
<td>(+)-Catechin</td>
<td>OC1CC2=C(O)C=C(O)C=C2OC1C3=CC(C(=C3)O)C=O</td>
<td>Bananas, raw (<em>Musa acuminata Colla</em>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(NDB No., 9040)</td>
</tr>
<tr>
<td>Hesperetin</td>
<td>O=C(CC(C3=CC(=C(O)=C(O)C=C3)O2)C1=C2C=C(O)C=C1O</td>
<td>Juice, orange, raw (NDB No., 9206)</td>
</tr>
<tr>
<td>Naringenin</td>
<td>OC1=CC=C(C=C1)C2CC(=O)C3=C(O2)C=C(O)C=C3O</td>
<td>Melons, honeydew, raw (<em>Cucumis melo</em>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(NDB No., 9184)</td>
</tr>
<tr>
<td>Apigenin</td>
<td>O=C(C=C=C3=CC(=C(O)C=C3)O2)C1=C2C=C(O)C=C1O</td>
<td>Pineapple, raw, all varieties (<em>Ananas comosus</em>)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(NDB No., 9266)</td>
</tr>
<tr>
<td>Flavonoid</td>
<td>Chemical Structure</td>
<td>Food Description</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Luteolin</td>
<td>O=C(C=C(C3=CC(O)=C(O)C=C3)O2)C1=C2C=C(O)C=C1O</td>
<td>Pomegranates, raw (<em>Punica granatum</em>) (NDB No., 9286)</td>
</tr>
<tr>
<td>Kaempferol</td>
<td>O=C(C(O)=C(C3=CC(C(O)C=C3)O2)C1=C2C=C(O)C=C1O</td>
<td>Broccoli, cooked, boiled, drained, without salt (NDB No., 11091)</td>
</tr>
<tr>
<td>Quercetin</td>
<td>O=C(C(O)=C(C3=CC(O)=C(O)C=C3)O2)C1=C2C=C(O)C=C1O</td>
<td>Mushrooms, white, raw (<em>Agaricus bisporus</em>) (NDB No., 11260)</td>
</tr>
<tr>
<td>Myricetin</td>
<td>O=C(C(O)=C(C3=CC(O)=C(O)C=C3)O2)C1=C2C=C(O)C=C1O</td>
<td>Potatoes, red, flesh and skin, raw (<em>Solanum tuberosum</em>) (NDB No., 11355)</td>
</tr>
</tbody>
</table>

* Bhagwat S, Haytowitz DB, Holden JM (2012)

Figure 3 shows the obtained prediction by the KNN algorithm for the conjunct # 4. X represents the number of rows in the database, in which everyone has an ORAC value represented in Y. In the graphic it is possible to notice the similarity between the predicted and the obtained ORAC values.

![Figure 3](image)

**Figure 3.** Comparative study of conjunct # 4 ORAC values (experimental and prediction). Results obtained using KNN – software PROCONS.

**Conclusions**

The best results were obtained when the calculation of weight and similarity were included in the algorithms. Using KNN, the optimum k value was 3, making evident the importance of the similarity between objects for the good predictive results.

It was concluded the importance of the use of KNN technic for the prediction of the antioxidant activity en different alimentary groups. This algorithm can be used, in future work, to identify the responsible features for the relationship between quantity of flavonoids, topologic-structural information and alimentary matrix. It will be further studied the relationship between antioxidant capacity of the food and the composition in flavonoids of a complex alimentary matrix.

**Acknowledgments**
The authors thank the partial financial support of University of Santiago de Compostela, University of Camagüey Ignacio Agramonte Loynaz and Galician Plan of research, innovation and growth 2011-2015 (Plan I2C). The authors also thank Y. Filiberto et al. for the access to the PROCONS software.

References


