Application of KNN algorithm in determining the total antioxidant capacity of flavonoid-containing foods

Estela GuardadoYordi[1,2], Raúl Koelig[3], Yailé Caballero Mota[3], Maria João Matos[2], Lourdes Santana[2], Eugenio Uriarte[2] and Enrique Molina[1,2]

[1] Universidad de Camagüey Ignacio Agramonte Loynaz, Facultad de Química, Cuba, estela.guardado@reduc.edu.cu

[2] Universidad de Santiago de Compostela, Facultad de Farmacia, España

[3] Universidad de Camagüey Ignacio Agramonte Loynaz, Facultad de Informática, Cuba

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 Neighbors*) 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 ($TEAC_{exp}$), iv) probability of clastogenicity and clastogenicity classification by *Quantitative Structure-Activity Relationship* (QSAR) method and e) total polyphenol value (TP_{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 ($ORAC_{exp}$) value, expressed in μmol 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). $ORAC_{exp}$ and TP_{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).

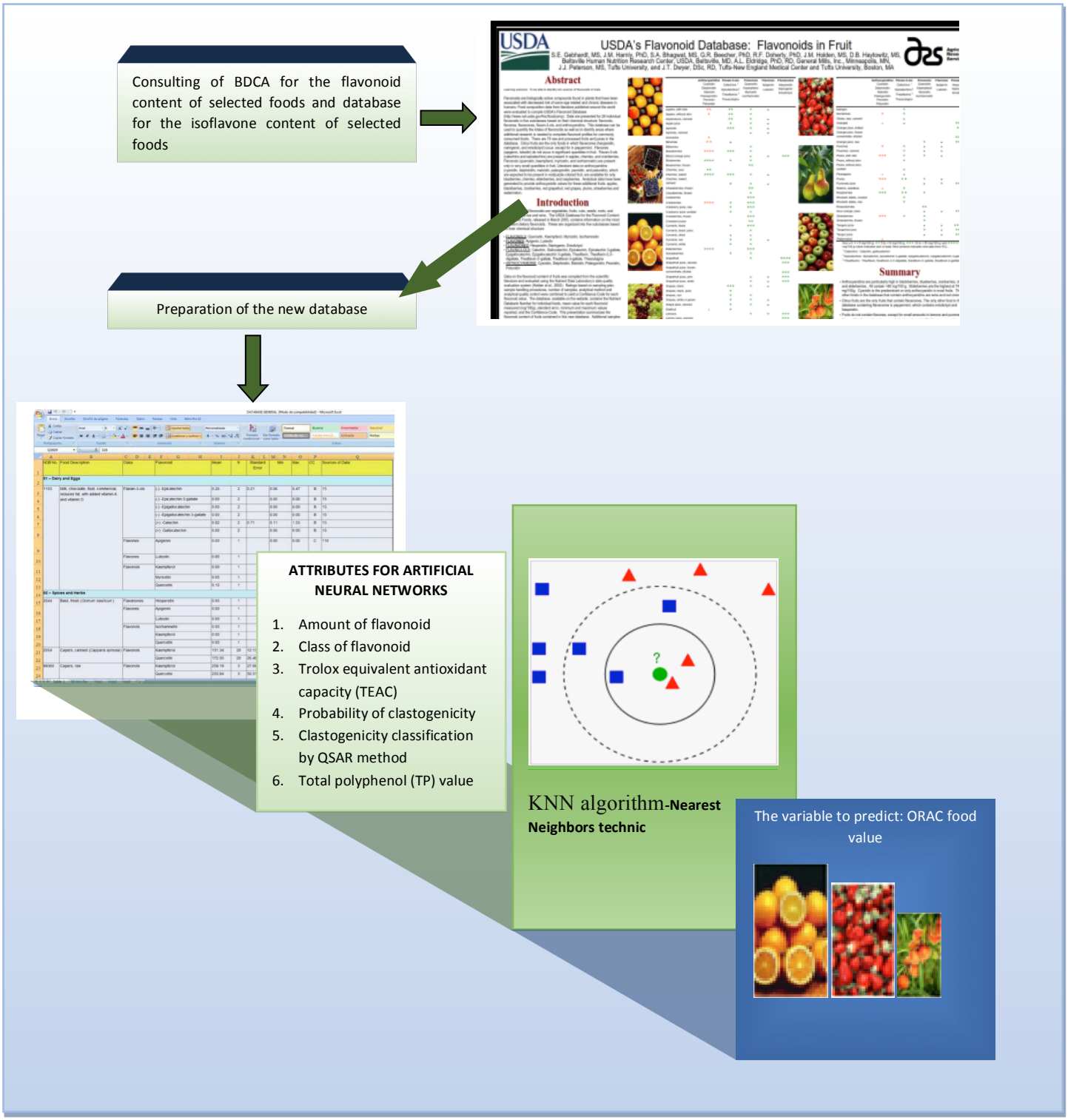

Figure 1. General scheme of the applied methodology.

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.

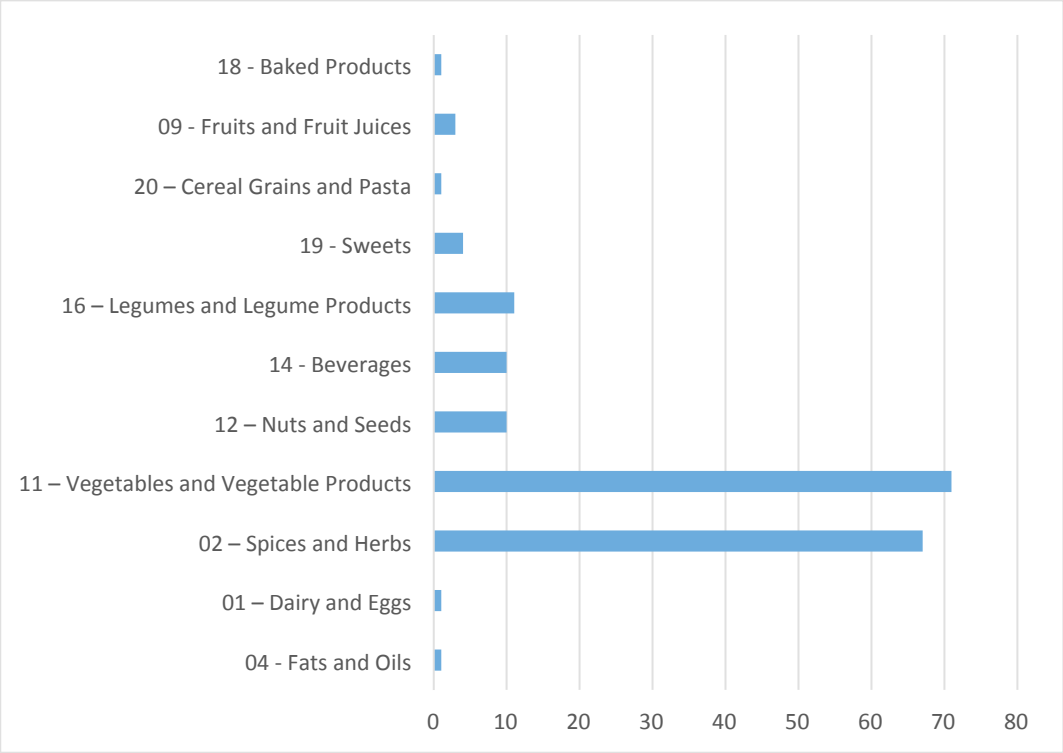

Figure 2. Quantities of aliments (divided by alimentary groups) presented in the database.

Table 1. Examples of flavonoids presented in the database.

Names	SMILE	Examples of sources (NDB No.)*
(-)-Epicatechin 3-gallate	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	Apples, Fuji, raw, with skin (NDB No., 97066)
(+)-Catechin	OC1CC2=C(O)C=C(O)C=C2OC1C3=CC=C(O)C(=C3)O	Bananas, raw (*Musa acuminata Colla*) (NDB No., 9040)
Hesperetin	O=C(CC(C3=CC(O)=C(OC)C=C3)O2)C1=C2C=C(O)C=C1O	Juice, orange, raw (NDB No., 9206)
Naringenin	OC1=CC=C(C=C1)C2CC(=O)C3=C(O2)C=C(O)C=C3O	Melons, honeydew, raw (*Cucumis melo*) (NDB No., 9184)
Apigenin	O=C(C=C(C3=CC=C(O)C=C3)O2)C1=C2C=C(O)C=C1O	Pineapple, raw, all varieties (*Ananas comosus*) (NDB No., 9266)

Luteolin	O=C(C=C(C3=CC(O)=C(O)C=C3)O2)C1=C2C=C(O)C=C1O	Pomegranates, raw (*Punica granatum*) (NDB No., 9286)
Kaempferol	O=C(C(O)=C(C3=CC=C(O)C=C3)O2)C1=C2C=C(O)C=C1O	Broccoli, cooked, boiled, drained, without salt (NDB No., 11091)
Quercetin	O=C(C(O)=C(C3=CC(O)=C(O)C=C3)O2)C1=C2C=C(O)C=C1O	Mushrooms, white, raw (*Agaricus bisporus*) (NDB No., 11260)
Myricetin	O=C(C(O)=C(C3=CC(O)=C(O)C(O)=C3)O2)C1=C2C=C(O)C=C1O	Potatoes, red, flesh and skin, raw (*Solanum tuberosum*) (NDB No., 11355)

* 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.

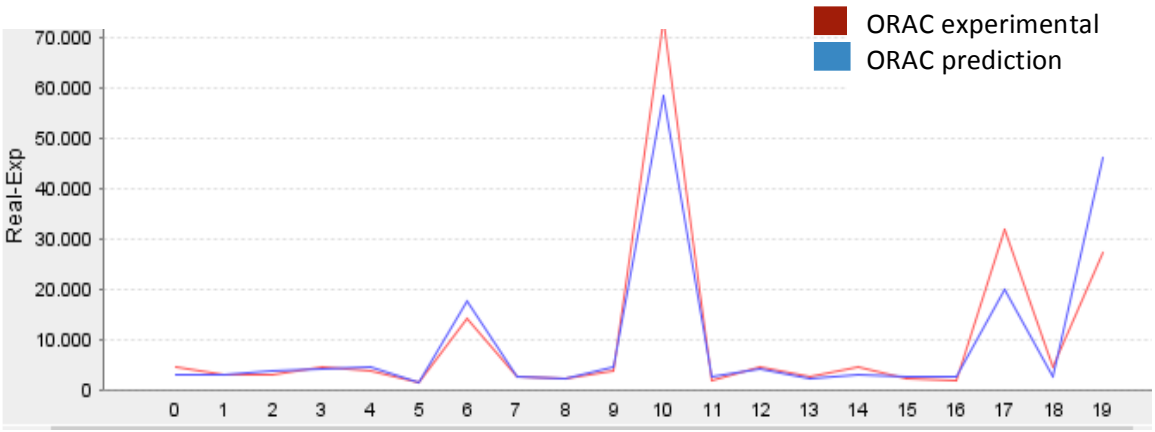

Figure 3. Comparative study of conjunct # 4 ORAC values (experimental and prediction). Results obtained using KNN – software PROCONS.

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

1. Holdena JM, Bhagwata SA, Haytowitza DB, Gebhardta SE, Dwyerb JT, Petersonb J, et al. Development of a database of critically evaluated flavonoids data: application of USDA's data quality evaluation system. Journal of Food Composition and Analysis 2005;18:829-44.

2. Bhagwat S, Haytowitz DB, Holden JM. USDA Database for the Flavonoid Content of Selected Foods, Release 3.1. NDL Web site: http://www.ars.usda.gov/Services/docs.htm?docid=62312012.

3. U.S. Department of Agriculture ARS. USDA Database for the Isoflavone Content of Selected Foods. Release 2.0. http://www.ars.usda.gov/Services/docs.htm?docid=63822008.

4. Harnly JM, Doherty RF, Beecher GR, Holden JM, Haytowitz DB, Bhagwat S, et al. Flavonoid Content of U.S. Fruits, Vegetables, and Nuts. Journal of Agriculture and Food Chemistry 2006;54:9966-77.

5. Kay CD. The future of flavonoid research. British Journal of Nutricion 2010;104:91-5.

6. Robards K, Antolovich M. Analytical chemistry of fruit bioflavonoids. A review. Analyst 1997;122:11-34.

7. Halliwell B. Oxidative stress, nutrition and health. Experimental strategies for optimization of nutritional antioxidant intake in humans. Free Radical Research 1996;25:57-74.

8. Chun O, Floegel A, Chung S, Chung C, Song W, Koo S. Estimation of antioxidant intakes from diet and supplements in US adults. Journal of Nutrition 2010;140:317-24.

9. Institute of Medicine of the Nation al Academies. Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids. Washington, D.C.: National Academy Press; 2000.

10. FAO. Retos sobre la composicion de alimentos. International Network of Food Data Systems (INFOODS); 2014 [cited 2015 May]; Available from: http//www.fao.org/infoods/infoods/retos.

11. Haytowitz DB, Bhagwat S, Holden JM. Sources of variability in the flavonoid content of foods. Procedia Food Science 2013;2:46-51.

12. Bhagwat S, Haytowitz DB, Wasswa-Kintu SI, Holden JM. USDA develops a database for flavonoids to assess dietary intakes. Procedia Food Science 2013;2 :81-6.

13. Trujillano J, March J, Sorribas A. Aproximación metodológica al uso de redes neuronales artificiales para la predicción de resultados en medicina. Medicina clinica 2004;122(15).

14. U.S. Department of agriculture ARS. USDA National Nutrient Database for Standard Reference. Nutrient Data Laboratory Home Page; 2009 [cited 2010 April 22]; Available from: http://www.ars.usda.gov/nutrientdata.

15. Filiberto Y, Bello R, Caballero Y, Frias M. A method to build similarity relations into extended Rough Set Theory. 10th International Conference on Intelligent Systems and Applications ISDA 2010. Cairo, Egypt. 2010. IEEE Catalog Number CFP 10384 CDR, ISBN 978-1-4244-8135-4. Thomson Reuters.

16. Awika JM, Rooney LW, Wu X, Prior RL, Cisneros-Zevallos L. Screening methods to measure antioxidant activity of sorghum (Sorghum bicolor) and sorghum products. Journal of Agriculture and Food Chemistry 2003;51:6657-62.

17. Prior RL, Hoang H, Gu L, Wu X, Bacchocca M, Howard L, et al. Assays for hydrophilic and lipophilic antioxidant capacity (oxygen radical absorbance capacity (ORAC) of plasma and other biological and food samples. Journal of Agriculture and Food Chemistry 2003;51:3273-9.