Journal of Theoretical Biology. 2013; 317:331-337

Random Forest classification based on star graph topological indices for antioxidant proteins

Enrique Fernández-Blanco, Vanessa Aguiar-Pulido, Cristian Robert Munteanu, Julian Dorado

University of A Coruña, ICT Dept., Facultad de Informática, Campus de Elviña s/n, 15071 A Coruña, Spain

Abstract

Aging and life quality is an important research topic nowadays in areas such as life sciences, chemistry, pharmacology, etc. People live longer, and, thus, they want to spend that extra time with a better quality of life. At this regard, there exists a tiny subset of molecules in nature, named antioxidant proteins that may influence the aging process. However, testing every single protein in order to identify its properties is quite expensive and inefficient. For this reason, this work proposes a model, in which the primary structure of the protein is represented using complex network graphs that can be used to reduce the number of proteins to be tested for antioxidant biological activity. The graph obtained as a representation will help us describe the complex system by using topological indices. More specifically, in this work, Randić's Star Networks have been used as well as the associated indices, calculated with the S2SNet tool. In order to simulate the existing proportion of antioxidant proteins in nature, a dataset containing 1999 proteins, of which 324 are antioxidant proteins, was created. Using this data as input, Star Graph Topological Indices were calculated with the S2SNet tool. These indices were then used as input to several classification techniques. Among the techniques utilised, the Random Forest has shown the best performance, achieving a score of 94% correctly classified instances. Although the target class (antioxidant proteins) represents a tiny subset inside the dataset, the proposed model is able to achieve a percentage of 81.8% correctly classified instances for this class, with a precision of 81.3%.

Highlights

► This work presents an automatic antioxidant protein detection method. ► The new method uses graphical information processing theory which has never previously used in this kind of problem. ► The results can be qualified as notable compared with the state of the art.

Keywords

Multi-target QSAR; Star Graph; Topological índices; Antioxidant protein

1. Introduction

Life expectancy is increasing every year, especially in developed societies. Nowadays, in these countries, it is not strange to find some people that are near one hundred years, when 20 years ago this was quite rare. For example, in Spain, life expectancy at birth has increased from 73 years in 1975 to more than 81 in 2011 (OECD, 2011). In this context, it is obvious that people may want to spend the biggest part of their life in optimum health conditions. In order to achieve this objective, finding some mechanism that delays aging (Cevenini et al., 2010; de Magalhães, 2010, 2011, 2012; Freitas and de Magalhães, 2012; Harman, 1981; Hayflick, 2000) is necessary. Several important works have proposed specific relationships between genes or proteins and aging (Aledo et al., 2011, 2012; de Magalhães et al., 2009; Freitas et al., 2011; Gomes et al., 2011; Li et al., 2010).

More research focused on antioxidant molecules may be useful for this purpose, since, for example, oxidative stress is one of the risk factors of colorectal carcinogenesis. In inflammatory reactions the activated leucocytes produce mutagenic and mitogenic free radicals, hereby promoting tumour formation. In addition, obesity, hyperlipidemia and hyperinsulinemia increase the energy supply of epithelial cells, thus leading to deregulation of the mitochondrial electron transport chain. Finally, the latter leads to increased free radical production, causing troubles in cell cycle regulation, mutations, and unrestricted proliferation of damaged cells (Regöly-Mérei et al., 2007).

Unfortunately, the number of molecules that have antioxidant properties in nature is quite low. Therefore, developing models that help to detect molecules with antioxidant properties would be very helpful. On this basis, the main objective of this paper will be to develop models that, on one hand, will reduce the number of molecules for tests in different trials and, on the other hand, to increase the success rates when molecules are tested looking for these properties.

In order to achieve this, the authors have used Quantitative Structure Activity Relationships (QSARs) (Devillers and Balaban, 1999). QSARs are based on Graph Theory, one of the most common techniques used in protein analysis. Using this technique, macromolecular descriptors, named topological indexes (TIs), are calculated for its later analysis. This branch of mathematical chemistry has become an intense area of research, generating new information regarding DNA/proteins by representing them as graphs and obtaining the corresponding TIs in order to analyse the resulting complex networks (Agüero-Chapin et al., 2006; Bielińska-Wa-z et al., 2007; Munteanu et al., 2010; Randić and Balaban, 2003). In order to perform these analyses, the TIs are then processed by a classification technique such as Support Vector Machines (SVMs) (Vapnik, 1995), Artificial Neural Networks (ANNs) (Rivero et al., 2011), Random Space Classifiers (Skurichina and Duin, 2002), Linear Discriminant Analysis (LDA), etc, abstracting general properties for future molecules that have not been already tested. Many examples involving QSAR can be found in literature (González-Díaz et al., 2006, 2007a, 2010; Prado-Prado et al., 2008; Riera-Fernández et al., 2012) regarding protein folding kinetics (Chou, 1990), enzyme-catalyzed reactions (Chou, 1989; Chou and Forsen, 1980; Chou and Liu, 1981; Kuzmic et al., 1992), inhibition kinetics of processive nucleic acid polymerases and nucleases (Althaus et al., 1993a, 1993b, 1994, 1996; Chou et al., 1994), DNA sequence analysis (Qi et al., 2007), anti-sense strands base frequencies (Chou et al., 1996), analysis of codon usage (Chou and Zhang, 1992; Zhang and Chou, 1994), Cancer prediction (Aguiar-Pulido et al., 2012), as well as complex network systems investigations (Diao et al., 2007; Gonzalez-Diaz et al., 2007b, 2008).

In this work, the authors propose the first non-antioxidant/antioxidant protein classification model based on embedded/ non-embedded Star Graph TIs including the trace of connectivity matrices, Harary number, Wiener index, Gutman index, Schultz index, Moreau-Broto indices, Balaban distance connectivity index, Kier–Hall connectivity indices and Randić connectivity index. This information is then used as input to several classification techniques, obtaining the best results when the Random Forest technique is used.

2. Materials and methods

The description of the methodology followed in this work is presented in Fig. 1. The input data is represented by the amino acid sequences (primary structure) antioxidant and non-antioxidant proteins in FASTA format. By using the S2SNet tool (Munteanu et al., 2009), the sequences of amino acids are transformed into Star Graphs and the corresponding topological indices are calculated. The resulting numbers that characterised each graph (that is, a protein graphical representation) are then used in Weka (Hall et al., 2009a) to find the best QSAR classification model. The final model is used to predict antioxidant activity for new amino acid sequences.

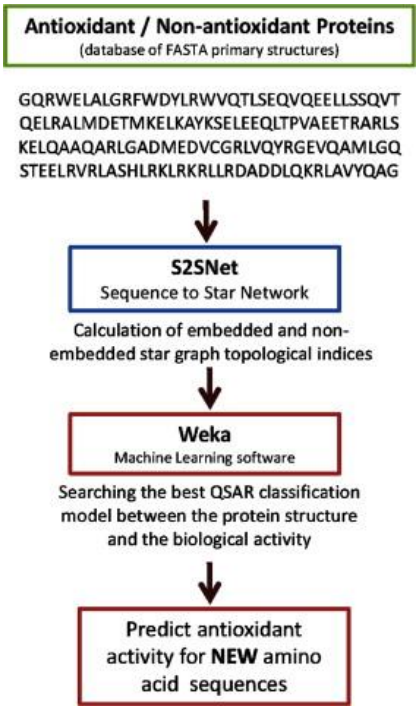

Fig. 1. Flowchart of building QSAR classification models for protein antioxidant activity prediction.

2.1. Protein set

This work is based on datasets extracted from several protein databases. The sets of protein primary sequences are represented by 324 proteins with antioxidant activity and 1675 proteins without. The antioxidant protein FASTA sequences (positive group) have been downloaded from the Protein Databank (Berman et al., 2000), the "Antioxidant activity" list obtained with the "Molecular Function Browser" in the "Advanced Search Interface". The negative group was constructed using the PISCES CulledPDB (Wang and Dunbrack, 2003) list of proteins with identity less than 20%, resolution of 1.6 Å and *R*-factor 0.25 (non-antioxidant proteins included, but any other possible biological function). Identity is the degree of correspondence between two sequences and a value of 25% or higher implies similarity of function. The sequence identities for PDB sequences have been determined using Combinatorial Extension (CE) structural alignment (Shindyalov and Bourne, 1998). The PIECES server (http://dunbrack.fccc.edu/PISCES.php) used a Z-score of 3.5 as the threshold to accept possible evolutionary relationships. PISCES' alignments are local, so that two proteins that share a common domain with sequence identity above the threshold are not both included in the output lists. Both lists have not been post-filtered for any source organism.

2.2. Star Graph topological indices

Each protein was transformed into a Star Graph, where the amino acids are the vertices (nodes), connected in a specific sequence by the peptide bonds. The Star Graph is a special type of tree with *N* vertices where one has got *N*-1 degrees of freedom and the remaining *N*-1 vertices have got one single degree of freedom (Harary, 1969). Each of the 20 possible branches ("rays") of the star contains the same amino acid type and the star centre is a non-amino acid vertex. This way, the following information of the protein primary structure is encoded into the Star Graph connectivity: amino acid type, sequence and frequency.

A protein can be represented by diverse forms of graphs, which can be associated with distinct distance matrices. The best method to construct a standard Star Graph is described subsequently: each amino acid/vertex holds the position in the original sequence and the branches are labelled by alphabetical order of the three-letter amino acid code (Randić et al., 2007). The graph is embedded if the initial sequence connectivity in the protein chain is included. Fig. 2 presents the embedded/ non-embedded Star Graphs of PRPS1 using the alphabetical order of one-letter amino acid code.

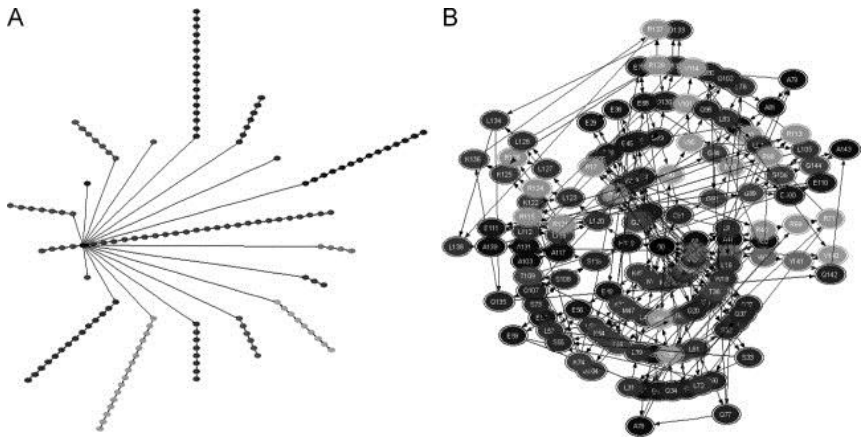

Fig. 2. The non-embedded (A) and embedded (B) Star Graphs for 1BZ4, chain A.

Graphs are compared using the corresponding connectivity matrix, distance matrix and degree matrix. In the case of the embedded graph, the connectivity matrices in the sequence and in the Star Graph are combined. These matrices and the normalized ones are the basis of the TIs calculation.

The conversion of the amino acid sequences into Star Graph TIs was performed by using the Sequence to the Star Networks (S2SNet) application, developed by our group. S2SNet is based on wxPython (Rappin and Dunn, 2006) for the GUI application and has *Graphviz* (Koutsofios and North, 1993) as a graphics back-end. The present calculations are characterized by embedded and non-embedded TIs, no weights, Markov normalization and power of matrices/indices (*n*) up to 5. The results file contains the following TIs (Todeschini and Consonni, 2002):

Trace of the *n* connectivity matrices (Tr_n):

$$Tr_n = \sum_i (Mn)ii, \quad (1)$$

where *n*=0 – power limit, *M*=graph connectivity matrix (*i*□*i* dimension); *ii*=*i*th diagonal element;

Harary number (H):

$$H = \sum_{i<j} m_{ij}/d_{ij}, \tag{2}$$

where d_{ij} are the elements of the distance matrix and m_{ij} are the elements of the M connectivity matrix;

Wiener index (W):

$$W = \sum_{i<j} d_{ij}, \tag{3}$$

Gutman topological index (S_6):

$$S_6 = \sum_{ij} \deg_i \times \deg j/d_{ij}, \tag{4}$$

where deg_i are the elements of the degree matrix;

Schultz topological index (non-trivial part) (S):

$$S = \sum_{i<j} (\deg_i + \deg_j) \times d_{ij}, \tag{5}$$

Balaban distance connectivity index (J):

$$J = (edges - nodes + 2) \times \sum_{i<j} m_{ij} \times \mathrm{sqrt}(\sum_k d_{ik} \times \sum_k d_{kj}), \tag{6}$$

where *nodes*+1=AA numbers/node number in the Star Graph+origin, $\Sigma_k d_{ik}$ is the node distance degree;

Kier–Hall connectivity indices (^nX):

$${}_0X = \sum_i 1/\mathrm{sqrt}(\deg_i), \tag{7}$$

$${}_2X = \sum_{i<j<k} m_{ij} \times m_{jk}/\mathrm{sqrt}(\deg_i \times \deg_j \times \deg_k), \tag{8}$$

$${}_3X = \sum_{i<j<k<m} m_{ij} \times m_{jk} \times m_{km}/\mathrm{sqrt}(\deg_i \times \deg_j \times \deg_k \times \deg_m), \tag{9}$$

$${}_4X = \sum_{i<j<k<m<o} m_{ij} \times m_{jk} \times m_{km} \times m_{mo}/\mathrm{sqrt}(\deg_i \times \deg_j \times \deg_k \times \deg_m \times \deg_o), \tag{10}$$

$${}_5X = \sum_{i<j<k<m<o<q} m_{ij} \times m_{jk} \times m_{km} \times m_{mo} \times m_{oq}/\mathrm{sqrt}(\deg_i \times \deg_j \times \deg_k \times \deg_m \times \deg_o \times \deg_q), \tag{11}$$

Randic connectivity index (^1X):

$${}_1X = \sum_{ij} m_{ij}/\mathrm{sqrt}(\deg_i \times \deg_j), \tag{12}$$

These TIs and other derivate ones will be used in the next step to construct an antioxidant/ non-antioxidant classification model using machine learning methods.

2.3. Random Forest

Random Forest was first proposed by Breiman, (2001). This technique combines many decision trees to make a prediction, giving as output the class that is the mode of the classes output by individual trees. Thus, this technique can be considered an "ensemble learning" technique, since it uses multiple models to obtain better predictive performance. These decision trees are constructed by means of bagging classification trees (Breiman, 1996), where each tree is constructed independently based on a random sample and a majority vote of the trees is taken as prediction. Random Forest adds an extra random layer to bagging. Normally, decision trees are built from a random sample and nodes are split by the best among a subset of predictors randomly chosen at that node.

The main advantage of Random Forest over other techniques such as Artificial Neural Networks, Support Vector Machines, Linear Discriminant Analysis, etc. is the robustness of this technique regarding solution overfitting, tending to converge always when the number of trees is large.

The typical Random Forest algorithm is composed of three steps:

- Get *n* random samples from the original dataset to use them as tree seeds.
- For each seed, grow a non-pruned tree, and for each node, randomly choose *m* predictors and the best split among those.
- Execute the different prediction trees and select as prediction the most voted one.

It may be highlighted that this technique is quite efficient because, when constructing the trees, the pruning phase has been deleted and the search is performed over a small set. This simplification can give the idea that a single tree may have better performance, but it was empirically proved that Random Forest overcomes the performance of CART single tree predictors (Chipman et al., 1998).

3. Results

The dataset used in this paper is composed of 1999 protein sequences, from which 324 have proved to have antioxidant activity (positive group). The remaining 1675 proteins (negative group) are sequences from the CulledPDB server with identity less than 20%, without antioxidant biological activity. These protein sequences have been processed with the S2SNet application (Munteanu et al., 2009) in order to obtain the different topological indexes used in this study. Specifically, from each sequence 42 attributes are extracted from the embedded/non-embedded Star Graph.

The series of topological indices for each protein have been used to find the best antioxidant classification model with Machine Learning methods included in Weka (Hall et al., 2009b). In order to extract more general conclusions from this study, the authors have tested the different classification techniques using 10-fold cross-validation (McLachlan et al., 2004). 10-fold cross-validation is the most common among the *k*-fold cross-validation family and its objective is to minimize the influence of the randomness in creating the training and test sets for a specific classification technique.

The objective of this work is to select the technique with the highest classification score, having a good precision value, due to the nature of the problem. The first approach considered was to use linear regression, but the results showed that it was impossible to achieve good classification scores with this technique.

Table 1 shows the results of the different classification models tested, those that obtained the best scores, considering all the attributes extracted from the Star Graph, that is, 42 attributes. The algorithms used in the tests are those implemented in the Weka Machine Learning framework. This table shows, for each model, the classification scores obtained for the different classes, as well as the global classification percentages, the precision values for the target class (antioxidant proteins), the ROC values and the number of attributes that were considered.

Table 1. Performance of the classification methods considering all the attributes.

Technique	% Antiox	% Non antiox	% Global	Precision antiox (%)	Global precision (%)	ROC
Naive bayes	97.5	49.1	57.0	27.1	87.4	0.78
MLP	22.8	97.5	85.4	63.8	83.0	0,874
K-star	86.7	94.3	93.1	74.7	93.7	0,971
JRip	64.8	96.1	91.0	76.1	90.6	0.814
Random tree	81.8	95.0	92.8	75.9	93.1	0.884
Random Forest	84	96.7	94.6	82.9	94.6	0.954

The Random Forest technique seems to be the best option because it achieves a percentage of 94.6% correctly classified instances. In addition, it is interesting to note that, for the antioxidant class, it achieves a percentage of 84% correctly classified instances. This model achieves a precision of 82.9%, which is the highest among the tested machine learning methods.

In order to reduce the noise and to improve the classification scores, the data used as input has been divided into three subsets depending on the nature of the attributes:

- A subset named *Sh*, which includes the attributes related with the entropy of the embedded and non-embedded Graph.
- A subset named *Tr*, which includes the attributes related with the traces of the embedded and non-embedded Graph.
- And a subset named *X*, which includes the attributes related with the polygon indexes to represent the subspaces in the graph.

Table 2 shows the result of this division. It should be highlighted that not all of the original attributes have been included in one of these three subsets; more specifically, some attributes regarding the general shape of the graphs were not included in any of these subsets.

Table 2. Attributes subsets for the tests.

Subset Name	Attributes	
	Non-embedded graph	Embedded graph
Sh	Sh0,Sh1, Sh2, Sh3, Sh4, Sh5	eSh0,eSh1, eSh2, eSh3, eSh4, eSh5
Tr	Tr0, Tr2, Tr4	eTr0, eTr2, eTr3,eTr4,eTr5
X	X0, X1R, X2, X3, X4, X5	eX0, eX1R, eX2, eX3, eX4, eX5
Remaining	H, W, S6, S, J	eH, eW, eS6, eS, eJ

The different methods were then tested using each of these subsets as well as their combination, in order to find the best possible one. Results of these tests are shown in Tables 3 and 4. These results show that Random Forest can still be considered adequate to solve the problem proposed in this work and that there is nearly no difference between using the *X* subset as input and all of the attributes. Regarding classification scores, this technique achieves 82.1% of correctly classified instances for the target class (that is, the antioxidant class) with a precision of 80.4% considering the 12 attributes part of the *X* subset, compared to 84% of correctly classified instances with a precision of 82.9% when all the attributes were considered (that is, 42 attributes). Therefore, it is very likely that some of these attributes may give little extra information. Reducing the number of attributes considered as input may be interesting, improving even the performance or precision of the model.

Table 3. Results obtained using the different subsets as input, considering 12 attributes.

Technique	% antiox	% non antiox	% global	Precision antiOx (%)	Global precision (%)	ROC
Naive bayes	95.7	56.3	62.7	29.8	87.4	0.79
MLP	38.6	95.5	86.2	62.2	84.6	0.851
K-star	51.5	95.2	88.1	67.3	87.2	0.926
JRip	47.2	98.6	90.0	86.4	89.9	0.726
Random tree	80.9	94.2	92.0	73.0	92.5	0.875
Random Forest	79.3	94.4	91.9	73.2	92.3	0.913
Naive bayes	74.0	57.3	60.1	74.7	60.1	0.797
MLP	0	100	83.8	0	83.8	0.644
K-star	82.1	94.0	92.0	72.5	92.6	0.961
JRip	63.9	97.0	91.6	80.2	91.2	0.815
Random tree	79.0	94.3	91.8	72.9	92.2	0.867
Random Forest	79.9	96.1	93.5	79.9	93.5	0.95
Naive bayes	77.5	55.8	59.3	25.3	81.8	0.772
MLP	0	100	83.8	0	83.8	0.644
K-star	77.2	94.2	90.6	70.7	90.7	0.946
JRip	67.0	96.7	91.9	79.8	91.5	0.840
Random tree	82.1	94.9	92.8	75.6	93.1	0.885
Random Forest	82.1	96.1	93.8	80.4	93.9	0.948

Table 4. Results obtained using combinations of the different subsets as input, considering 20 attributes.

Technique	% antiox	% non antiox	% global	Precision antiox (%)	Global precision (%)	ROC
Naive bayes	96.0	57.0	63.3	30.1	87.5	0.807
MLP	16.4	98.2	84.9	63.9	82.3	0.867
K-star	84.3	93.9	92.3	72.8	93.0	0.967
JRip	65.7	97.0	91.9	81.0	91.6	0.843
Random tree	82.4	94.9	92.8	75.6	93.1	0.886
Random Forest	81.8	96.5	94.1	81.8	94.1	0.947
Naive bayes	80.6	55.5	59.5	25.9	82.7	0.783
MLP	38.9	95.2	86.1	61.2	84.5	0.877
K-star	78.4	94.4	91.8	73.2	92.1	0.957
JRip	65.1	96.8	91.7	79.9	91.3	0.836
Random tree	81.8	95.3	93.1	77.3	93.3	0.886
Random forest	81.2	95.7	93.3	78.5	93.4	0.952
Naive bayes	78.4	54.2	58.1	24.9	81.8	0.792
MLP	0	100	83.8	0	83.8	0.644
K-star	86.4	93.7	92.5	72.5	93.3	0.97
JRip	68.2	96.5	91.9	78.9	91.6	0.846
Random tree	81.8	94.7	92.6	74.9	92.6	0.882
Random forest	83.6	96.9	94.7	83.9	94.7	0.951

After analysing the results shown above, it seems that Random Forest is the best and most robust classification model. As it was previously mentioned, the subsets *Sh*, *Tr* and *X* contain the properties of the embedded and non-embedded graph. Therefore, in order to try to reduce the number of input attributes, the authors have tested the Random Forest in more depth, distinguishing between the properties of both types of graph. Results regarding this are shown in Table 5, as well as the number of attributes used as input to the method.

Table 5. Scores obtained by the Random Forest method for each input dataset tested.

Subset	% antiox	% non antiox	% global	Precision antiox (%)	Global precision (%)	ROC	Number attributes
Sh	79.3	94.4	91.9	73.2	92.3	0.913	12
Sh-embedded	79.0	94.1	91.6	72.1	92	0.897	6
Sh-non-embedded	75.0	94.6	91.4	73.0	91.5	0.906	6
Tr	79.9	96.1	93.5	79.9	93.5	0.95	8
Tr-embedded	**81.8**	**96.4**	**94.0**	**81.3**	**94.0**	**0.954**	**5**
TR-non-embedded	79.9	94.0	91.7	72.1	92.2	0.903	3
X	82.1	96.1	93.8	80.4	93.9	0.948	12
X-embedded	82.4	95.7	92.5	78.8	93.7	0.938	6
X-non-embedded	79.9	95.2	92.7	76.2	92.9	0.926	6
Sh and Tr	81.8	96.5	94.1	81.8	94.1	0.947	20
Sh- and Tr-embedded	81.2	96.0	93.6	79.7	93.6	0.946	11
Sh- and Tr-non-embedded	79.6	95.5	92.9	77.5	93.0	0.927	9
Sh and X	81.2	95.7	93.3	78.5	93.4	0.952	24
Sh- and X-embedded	80.2	95.1	92.7	76.0	92.9	0.947	12
Sh- and X-non-embedded	79.6	95.5	92.9	77.5	93.0	0.927	12
Tr and X	83.6	96.9	94.7	83.9	94.7	0.951	20
Tr- and X-embedded	**83.6**	**96.8**	**94.6**	**83.6**	**94.7**	**0.958**	**11**
Tr- and X-non-embedded	80.2	95.5	93.0	77.4	93.1	0.935	9
All	**84**	**96.7**	**94.6**	**82.9**	**94.6**	**0.954**	**42**
All-embedded	82.1	96.8	94.4	83.1	94.4	0.954	22
All-non-embedded	81.2	95.6	93.2	78.0	93.4	0.934	20

Again, results show that Random Forest is able to achieve better classification scores and similar precision values considering less attributes as input; in this case, taking only into consideration those included in the *Tr* subset (which contains only the values of the embedded graph). By adding the embedded attributes of the *X* subset, results are somehow better. However, this implies doubling the number of attributes used as input to the model. Thus, these results confirm that the rest of the attributes seem to add very little information or may even introduce noise inducing worse classification scores. If the ROC value is checked, it can be observed that the same ROC values are obtained when using the *Tr*-embedded dataset and the dataset containing all the attributes. The ROC curve for the *Tr*-embedded dataset is shown in Fig. 3.

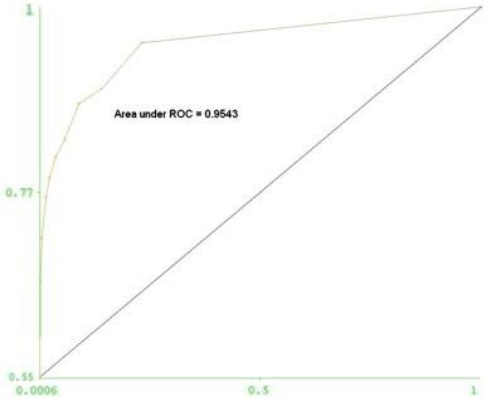

Fig. 3. ROC curve plot for the best classification method and the dataset containing the smallest number of attributes.

4. Discussion

This study proposes a model designed to identify proteins that have antioxidant activity by using Star Graph TIs obtained from protein amino acid sequences. The proposed model, based on only five attributes extracted from the embedded graph, shows good predictive capacity, achieving 94% of correctly classified instances. It is also important to highlight that, even though the non-antioxidant class was not the target class of this study, the model achieves a score of 81.8% correctly classified instances with good precision (81.3%).

Antioxidant proteins are very important molecules in pharmacology today. It can be concluded from this study that this model may help reducing the number of proteins to be tested in antioxidant research, being very probable that the selected proteins have antioxidant properties.

Acknowledgements

Vanessa Aguiar-Pulido and Cristian R. Munteanu acknowledge the funding support for a research position by the “Plan I2C” and an “Isidro Parga Pondal” Program both from Xunta de Galicia, Spain (supported by the European Social Fund). The authors also want to thank the support from different proyects that has funded part of this research (CN 2011/034, CN2012/127, 10SIN105004PR, O9SIN010105PR and TIN-2009-07707).

References

Agüero-Chapin, G., Gonzalez-Diaz, H., Molina, R., Varona-Santos, J., Uriarte, E., Gonzalez-Diaz, Y., 2006. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L. FEBS Lett. 580, 723–730.

Aguiar-Pulido, V., Munteanu, C.R., Seoane, J.A., Fernández-Blanco, E., Pérez-Montoto, L.G., González-Díaz, H., Dorado, J., 2012. Naïve Bayes QSDR classification based on spiral-graph Shannon entropies for protein biomarkers in human colon cancer. Mol. Biosyst. 8, 1716–1722.

Aledo, J.C., Li, Y., de Magalhães, J.P., Ruiz-Camacho, M., Perez-Claros, J.A., 2011. Mitochondrially encoded methionine is inversely related to longevity in mammals. Aging Cell 10, 198–207.

Aledo, J.C., Valverde, H., de Magalhães, J.P., 2012. Mutational bias plays an important role in shaping longevity-related amino acid content in Mammalian mtDNA-encoded proteins. J. Mol. Evol. 74, 332–341.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993a. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548–6554.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993b. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J. Biol. Chem. 268, 6119–6124.

Althaus, I.W., Chou, J.J., Gonzales, A.J., LeMay, R.J., Deibel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., et al., 1994. Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. Experientia 50, 23–28.

Althaus, I.W., Chou, K.C., Lemay, R.J., Franks, K.M., Deibel, M.R., Kezdy, F.J., Resnick, L., Busso, M.E., So, A.G., Downey, K.M., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1996. The benzylthio-pyrimidine U-31,355, a potent inhibitor of HIV-1 reverse transcriptase. Biochem. Pharmacol. 51, 743–750.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242.

Bielinska-Wa-z, D., Nowak, W., Wa-z, P., Nandyc, A., Clark, T., 2007. Distributionmoments of 2D-graphs as descriptors of DNAsequences. Chem. Phys. Lett. 443,408–413.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random Forest. Mach. Learn. 45, 5–32.

Cevenini, E., Bellavista, E., Tieri, P., Castellani, G., Lescai, F., Francesconi, M., Mishto, M., Santoro, A., Valensin, S., Salvioli, S., Capri, M., Zaikin, A., Monti, D., de Magalhães, J.P., Franceschi, C., 2010. Systems biology and longevity: an emerging approach to identify innovative anti-aging targets and strategies. Curr. Pharm. Des. 16, 802–813.

Chipman, H.A., George, E.I., McCulloch, R.E., 1998. An introduction to Classification and Regression Tree (CART) analysis. J. Am. Stat. Assoc., 935–948.

Chou, K.C., 1989. Graphical rules in steady and non-steady enzyme kinetics. J. Biol. Chem. 264, 12074–12079.

Chou, K.C., 1990. Review: applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys. Chem. 35, 1–24.

Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem. J. 187, 829–835.

Chou, K.C., Kezdy, F.J., Reusser, F., 1994. Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal. Biochem. 221, 217–230.

Chou, K.C., Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. J. Theor. Biol. 91, 637–654.

Chou, K.C., Zhang, C.T., 1992. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retroviruses 8, 1967–1976.

Chou, K.C., Zhang, C.T., Elrod, D.W., 1996. Do ‘‘antisense proteins’’ exist? J. Protein Chem. 15, 59–61.

de Magalhães, J.P., Curado, J., Church, G.M., 2009. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics 25, 875–881.

de Magalhães, J.P., Finch, C.E., Janssens, G., 2010. Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. Ageing Res. Rev. 9, 315–323.

de Magalhães, J.P., 2011. The biology of ageing: a primer. In: I., S.-H. (Ed.), An Introduction to Gerontology. Cambridge University Press, Cambridge, UK, pp. 21–47.

de Magalhães, J.P., Wuttke, D., Wood, S.H., Plank, M., Vora, C., 2012. Genome-environment interactions that modulate aging: powerful targets for drug discovery. Pharmacol. Rev. 64, 88–101.

Devillers, J., Balaban, A.T., 1999. Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach, The Netherlands.

Diao, Y., Li, M., Feng, Z., Yin, J., Pan, Y., 2007. The community structure of human cellular signaling network. J. Theor. Biol. 247, 608–615.

Freitas, A.A., de Magalhães, J.P., 2012. A review and appraisal of the DNA damage theory of ageing. Mutat. Res. 728, 12–22.

Freitas, A.A., Vasieva, O., de Magalhães, J.P., 2011. A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related. BMC Genomics. 12, 27

Gomes, N.M., Ryder, O.A., Houck, M.L., Charter, S.J., Walker, W., Forsyth, N.R., Austad, S.N., Venditti, C., Pagel, M., Shay, J.W., Wright, W.E., 2011. Comparative biology of

mammalian telomeres: hypotheses on ancestral states and the roles of telomeres in longevity determination. Aging Cell 10, 761–768.
González-Díaz, H., Bonet, I., Terán, C., de Clercq, E., Bello, R., García, M., Santana, L., Uriarte, E., 2007a. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. Eur. J. Med. Chem. 42, 580–585.
Gonzalez-Diaz, H., Gonzalez-Diaz, Y., Santana, L., Ubeira, F.M., Uriarte, E., 2008. Proteomics, networks and connectivity indices. Proteomics 8, 750–778.
González-Díaz, H., Sanchez-Gonzalez, A., Gonzalez-Diaz, Y., 2006. 3D –QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. J. Inorg. Biochem. 100, 1290–1297.
González-Díaz, H., Vilar, S., Rivero, D., Fernández-Blanco, E., Porto, A., Munteanu, C.R., 2010. QSPR Models for Cerebral Cortex Co-Activation Networks, Topological Indices for Medicinal Chemistry, Biology, Parasitology, and Social Networks. Research Signpost.
González-Díaz, H., Vilar, S., Santana, L., Uriarte, E., 2007b. Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices. Curr. Top Med. Chem. 7, 1025–1039.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.A., 2009a. The WEKA data mining software: an update. SIGKDD Explor., 11.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009b. The WEKA data mining software: an update. SIGKDD Explor., 11.
Harary, F., 1969. Graph Theory, Reading, MA.
Harman, D., 1981. The aging process. Proc. Natl. Acad. Sci. U.S.A. 78, 7124–7128.
Hayflick, L., 2000. The future of ageing. Nature 408, 267–269.
Koutsofios, E., North, S.C., 1993. Drawing Graphs with Dot. AT&T Bell Laboratories, Murray Hill, NJ, USA.
Kuzmic, P., Ng, K.Y., Heath, T.D., 1992. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. Anal. Biochem. 200, 68–73.
Li, Y.H., Dong, M.Q., Guo, Z., 2010. Systematic analysis and prediction of longevity genes in Caenorhabditis elegans. Mech. Ageing Dev. 131, 700–709.
McLachlan, G.J., Do, K.-A., Ambroise, C., 2004. Analyzing Microarray Gene Expression Data. Wiley.
Munteanu, C.R., Fernandez-Blanco, E., Seoane, J.A., Izquierdo-Novo, P., Rodriguez-Fernandez, J.A., Prieto-Gonzalez, J.M., Rabunal, J.R., Pazos, A., 2010. Drug discovery and design for complex diseases through QSAR computational methods. Curr. Pharm. Design 16, 2640–2655.
Munteanu, C.R., Magalhães, A.L., Uriarte, E., González-Díaz, H., 2009. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J. Theor. Biol. 257, 303–311.
OECD, 2011. <http://stats.oecd.org/index.aspx?DataSetCode=HEALTH_STATS>.
Prado-Prado, F.J., González-Díaz, H., Martinez de la Vega, O., Ubeira, F.M., Chou, K.C., 2008. Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorg. Med. Chem. 16, 5871–5880.
Qi, X.Q., Wen, J., Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. J. Theor. Biol. 249, 681–690.
Randic, M., Balaban, A.T., 2003. On a four-dimensional representation of DNA primary sequences. J. Chem. Inf. Model. 43, 532–539.
Randic, M., Zupan, J., Vikic-Topic, D., 2007. On representation of proteins by star-like graphs. J. Mol. Graph. Model, 290–305.
Rappin, N., Dunn, R., 2006. wxPython in Action. Manning Publications Co., Greenwich, CT.
Regöly-Mérei, A., Bereczky, M., Arató, G., Telek, G., Pallai, Z., Lugasi, A., Antal, M., 2007. Nutritional and antioxidant status of colorectal cancer patients. Orv. Hetil. 148, 1505–1509.
Riera-Fernández, I., Martín-Romalde, R., Prado-Prado, F., Escobar, M., Munteanu, C., Concu, R., Duardo-Sanchez, A., González-Díaz, H., 2012. From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markov-spectral moments indices. Curr. Top. Med. Chem. 8, 927–960.
Rivero, D., Fernandez-Blanco, E., Dorado, J., Pazos, A., 2011. Using recurrent ANNs for the detection of epileptic seizures in EEG signals. Evolutionary Computation (CEC), 2011 IEEE Congress on IEEE, pp. 587–592.
Shindyalov, I.N., Bourne, P.E., 1998. Protein structure alignment by incremental combinatorial extension of the optimum path. Protein Eng. 11, 739–747.
Skurichina, M., Duin, R.P.W., 2002. Bagging, boosting and the random subspace method for linear classifiers. Pattern Anal. Appl. 5, 121–135.

Todeschini, R., Consonni, V., 2002. Handbook of Molecular Descriptors. Wiley-VCH.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory.
Wang, G., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591.
Zhang, C.T., Chou, K.C., 1994. Analysis of codon usage in 1562 E. coli protein coding sequences. J. Mol. Biol. 238, 1–8.