Journal of Theoretical Biology 317 (2013) 331–337

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

# 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*

## 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.

## ARTICLE INFO

*Article history:*
Received 9 July 2012
Received in revised form
17 September 2012
Accepted 2 October 2012
Available online 29 October 2012

*Keywords:*
Multi-target QSAR
Star Graph
Topological indices
Antioxidant protein

## 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%.



## 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

* Corresponding author at: University of A Coruña, ICT Dept., Facultad de Informática, Campus de Elviña s/n, 15071 A Coruña, Spain. Tel.: +34 981 167 000; fax: +34 981 167 160.

*E-mail addresses:* efernandez@udc.es (E. Fernández-Blanco), vaguiar@udc.es (V. Aguiar-Pulido), muntisa@gmail.com (C.R. Munteanu), julian@udc.es (J. Dorado).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.10.006

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.

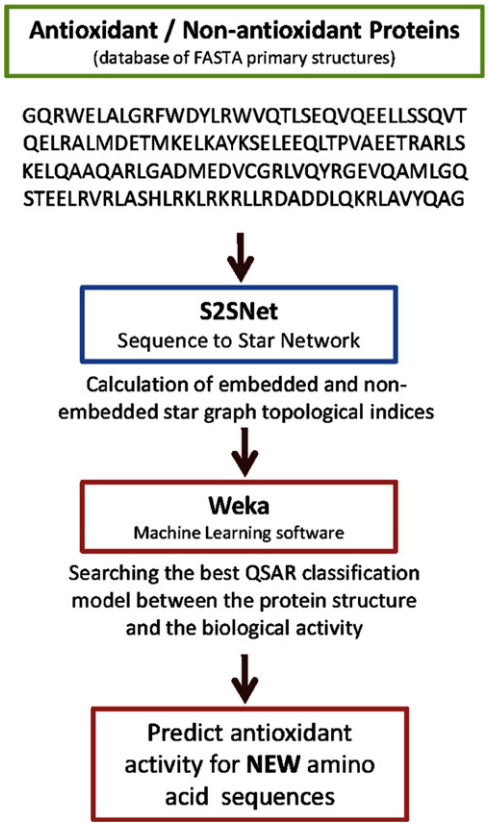


**Fig. 1.** Flowchart of building QSAR classification models for protein antioxidant activity prediction.

### 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")