Journal of Theoretical Biology 257 (2009) 303–311

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices

Cristian Robert Munteanu [a], Alexandre L. Magalhães [a], Eugenio Uriarte [b], Humberto González-Díaz [b,*]

[a] REQUIMTE/Faculty of Science, Chemistry Department, University of Porto, 4169-007, Portugal
[b] Unit of Bioinformatics & Connectivity Analysis (UBICA), Institute of Industrial Pharmacy, and Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Spain

ARTICLE INFO

Article history:
Received 22 July 2008
Received in revised form
11 November 2008
Accepted 22 November 2008
Available online 6 December 2008

Keywords:
Input-coded multi-target QPDR
Star graph
Cancer theoretical model
Clinical proteomics
GDA method

ABSTRACT

The cancer diagnostic is a complex process and, sometimes, the specific markers can interfere or produce negative results. Thus, new simple and fast theoretical models are required. One option is the complex network graphs theory that permits us to describe any real system, from the small molecules to the complex genetic, neural or social networks by transforming real properties in topological indices. This work converts the protein primary structure data in specific Randic's star networks topological indices using the new sequence to star networks (S2SNet) application. A set of 1054 proteins were selected from previous works and contains proteins related or not with two types of cancer, human breast cancer (HBC) and human colon cancer (HCC). The general discriminant analysis method generates an input-coded multi-target classification model with the training/predicting set accuracies of 90.0% for the forward stepwise model type. In addition, a protein subset was modified by single amino acid mutations with higher log-odds PAM250 values and tested with the new classification if can be related with HBC or HCC. In conclusion, we shown that, using simple input data such is the primary protein sequence and the simples linear analysis, it is possible to obtain accurate classification models that can predict if a new protein related with two types of cancer. These results promote the use of the S2SNet in clinical proteomics.

1. Introduction

Cancer is a leading cause of death worldwide, accounted for around 13% of all deaths in 2007 (WHO, 2008). Two of the leading types of cancer are the human breast cancer (HBC) and the human colon cancer (HCC). The estimated new cancer cases and deaths in US for 2008 shows that HBC will affect 26% of the women (15% will die) and HCC will involve 10% of the men/women (8% men and 9% women will die) (Jemal et al., 2008). Therefore, simple and fast theoretical method can be very useful in the detection of cancer diseases.

The actual work will use the protein quantitative proteome-disease relationship (QPDR) (Ferino et al., 2008), similar to quantitative structure-activity relationship (QSAR) (Devillers and Balaban, 1999). QPDR is one of the widely used analyse for predicting the protein properties and, in the present study, is using the macromolecular descriptors, named topological indices (TIs), obtained with the graph theory. The branch of mathematical chemistry dedicated to encode the DNA/protein information in graph representations by the use of the TIs has become an intense research area (Agüero-Chapin et al., 2006; Bielinska-Waz et al., 2007; Liao and Wang, 2004; Liao and Ding, 2005; Randic, 2000; Randic and Basak, 2001; Randic and Balaban, 2003; Randic et al., 2000). The graphic approaches of the biological systems study can provide useful insights in QSAR studies (González-Díaz et al., 2006, 2007c; Prado-Prado et al., 2008), 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, b; Althaus et al., 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) and in complicated network systems investigations (Diao et al., 2007; Gonzalez-Diaz et al., 2007a, 2008). Recently, the "cellular automaton image" (Wolfram, 1984, 2002) has also been applied to study hepatitis B viral infections (Xiao et al., 2006a), HBV virus gene missense mutation (Xiao et al., 2005b), and visual analysis of SARS–CoV (Gao et al., 2006; Wang et al., 2005), as well as representing complicated biological sequences (Xiao et al., 2005a) and helping

* Corresponding author. Tel.: +34 981563100; fax: +34 981594912.
E-mail addresses: muntisa@gmail.com (C.R. Munteanu), almagalh@fc.up.pt (A.L. Magalhães), eugenio.uriarte@usc.es (E. Uriarte), humberto.gonzales@usc.es, humbertogd@gmail.com (H. González-Díaz).

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.11.017

to identify protein attributes (Xiao and Chou, 2007; Xiao et al., 2006b). We have chosen the TIs for these QPDR models based on the previous work results with similar QSAR/QPDR models. Even if the TIs cannot be always interpreted, they demonstrate to encode the information that permits to create accurate QSAR/QPDR models.

Other interesting fields to apply the graph theory are the oncology and clinical proteomics. A classification model for discriminating prostate cancer patients from control group with connectivity indices where constructed by González-Díaz et al. (2007b). Vilar's group designed a QSAR model for alignment-free prediction of HBC biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks (Vilar et al., 2008).

The actual work is proposing a new cancer/non-cancer classification model based on protein embedded/non-embedded star graph TIs such are 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 Randic connectivity index. This classification can predict two types of cancer: HBC and HCC. The primary protein sequence is transformed in connectivity star graph's TIs that are used by a statistical linear method in order to construct an input-coded multi-target classification model.

2. Materials and methods

2.1. Protein set

Two sets of protein primary sequences are used: a set of 189 HBC/HCC cancer proteins (Sjoblom et al., 2006) and 865 non-cancer proteins (Dobson and Doig, 2005; Dobson et al., 2004). The list of cancer-related proteins in our work is the same with the list obtained by the Sjoblom group after the experimental analysis of 13,023 genes in 11 breast and 11 colorectal cancers.

2.2. Star graph TIs

Each protein sequence was transformed in 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 case of trees 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.

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 the following: each amino acid/vertex holds the position in the original sequence and the branches are labelled by alphabetical order of the 3-letter amino acid code (Randic et al., 2007). The graph is embedded if the initial sequence connectivity in the protein chain is included. Figs. 1A and B present the non-embedded/embedded star graphs of PRPS1 using the alphabetical order of one-letter amino acid code. Thus, the primary structure of protein chains are transformed in the correspondent Star graphs invariant TIs. The resulted graphs are not depending on the three-dimensional structure or the shape of the protein.

The comparison of the graphs is made by using the corresponding connectivity matrix, distance matrix and degree matrix. The matrices of the connectivity in the sequence and in the star graph are combined in the case of the embedded graph. These matrices and the normalized ones are the base of the TIs calculation.

The conversion of the amino acid sequences in star graph TIs was made by using sequence to star networks (S2SNet) application,

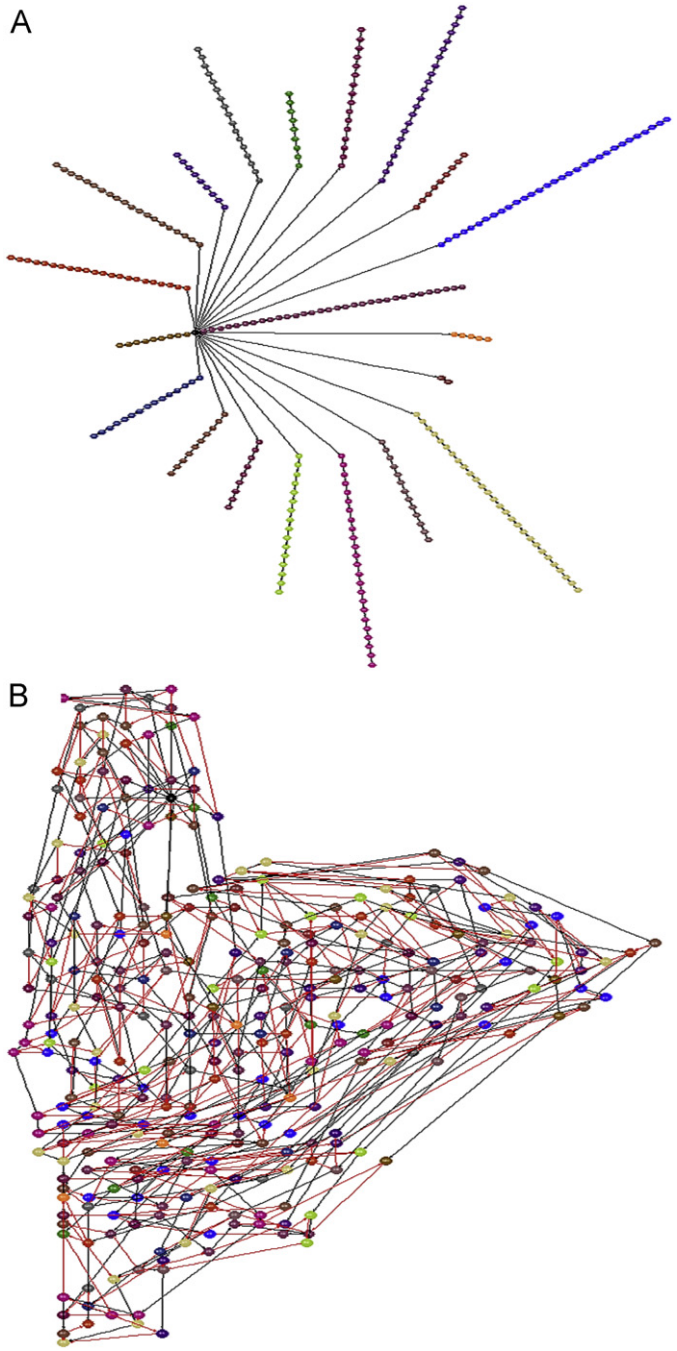

Fig. 1. (A) The non-embedded star graphs for PRPS1 and (B) the embedded star graphs for PRPS1.

developed by our group (Munteanu and Gonzáles-Diáz, 2008). 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) or the spectral moments:

$$\mathrm{tr}_n = \sum_i (M^n)_{ii}, \quad (1)$$