Journal of Theoretical Biology 293 (2012) 174–188

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

New Markov–Shannon Entropy models to assess connectivity quality in complex networks: From molecular to cellular pathway, Parasite–Host, Neural, Industry, and Legal–Social networks

Pablo Riera-Fernández [a], Cristian R. Munteanu [b], Manuel Escobar [c], Francisco Prado-Prado [c], Raquel Martín-Romalde [a], David Pereira [b], Karen Villalba [b], Aliuska Duardo-Sánchez [d], Humberto González-Díaz [a,*]

[a] Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela (USC), 15782 Santiago de Compostela, Spain
[b] Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, 15071 A Coruña, Spain
[c] Department of Organic Chemistry, Faculty of Pharmacy, USC, 15782 Santiago de Compostela, Spain
[d] Department of Special Public Law, Financial and Tributary Law Area, Faculty of Law, USC, 15782 Santiago de Compostela, Spain

ARTICLE INFO

Article history:
Received 22 July 2011
Received in revised form
9 October 2011
Accepted 14 October 2011
Available online 25 October 2011

Keywords:
Shannon Entropy
Markov Chains
Metabolic Pathways
Host–Parasite networks
Brain Cortex network

ABSTRACT

Graph and Complex Network theory is expanding its application to different levels of matter organization such as molecular, biological, technological, and social networks. A network is a set of items, usually called *nodes*, with connections between them, which are called *links* or *edges*. There are many different experimental and/or theoretical methods to assign node–node links depending on the type of network we want to construct. Unfortunately, the use of a method for experimental reevaluation of the entire network is very expensive in terms of time and resources; thus the development of cheaper theoretical methods is of major importance. In addition, different methods to link nodes in the same type of network are not totally accurate in such a way that they do not always coincide. In this sense, the development of computational methods useful to evaluate connectivity quality in complex networks (*a posteriori* of network assemble) is a goal of major interest. In this work, we report for the first time a new method to calculate numerical quality scores $S(L_{ij})$ for network links L_{ij} (connectivity) based on the Markov–Shannon Entropy indices of order k-th (θ_k) for network nodes. The algorithm may be summarized as follows: (i) first, the $\theta_k(j)$ values are calculated for all j-th nodes in a complex network already constructed; (ii) A Linear Discriminant Analysis (LDA) is used to seek a linear equation that discriminates connected or linked ($L_{ij}=1$) pairs of nodes experimentally confirmed from non-linked ones ($L_{ij}=0$); (iii) the new model is validated with external series of pairs of nodes; (iv) the equation obtained is used to re-evaluate the connectivity quality of the network, connecting/disconnecting nodes based on the quality scores calculated with the new connectivity function. This method was used to study different types of large networks. The linear models obtained produced the following results in terms of overall accuracy for network reconstruction: Metabolic networks (72.3%), Parasite–Host networks (93.3%), CoCoMac brain cortex co-activation network (89.6%), NW Spain fasciolosis spreading network (97.2%), Spanish financial law network (89.9%) and World trade network for Intelligent & Active Food Packaging (92.8%). In order to seek these models, we studied an average of 55,388 pairs of nodes in each model and a total of 332,326 pairs of nodes in all models. Finally, this method was used to solve a more complicated problem. A model was developed to score the connectivity quality in the Drug–Target network of US FDA approved drugs. In this last model the θ_k values were calculated for three types of molecular networks representing different levels of organization: drug molecular graphs (atom–atom bonds), protein residue networks (amino acid interactions), and drug–target network (compound–protein binding). The overall accuracy of this model was 76.3%. This work opens a new door to the computational reevaluation of network connectivity quality (collation) for complex systems in molecular, biomedical, technological, and legal–social sciences as well as in world trade and industry.

1. Introduction

Graph and Complex Network theory is expanding its application to different levels of matter organization such as molecular, biological,

* Corresponding author. Tel.: +34 981 167 000; fax: +34 981 167 160.
E-mail address: gonzalezdiazh@yahoo.es (H. González-Díaz).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.10.016

technological, and social networks (Bornholdt and Schuster, 2003; Boccaletti et al., 2006; Dehmer and Emmert-Streib, 2009). A network is a set of items, usually called *nodes*, with connections between them, which are called *links* or *edges* (Newman, 2003). The nodes can be atoms, molecules, proteins, nucleic acids, drugs, cells, organisms, parasites, people, words, laws, computers, or any other part of a real system. The edges or links are relationships between the nodes such as chemical bonds, physical interactions, metabolic pathways, pharmacological action, law recurrence, or social ties.

There are many different experimental and/or theoretical methods to assign node–node links depending on the type of network we want to construct. Unfortunately, many of these methods are expensive in terms of time or resources. In addition, different methods to link nodes in the same type of network are not totally accurate in such a way that they do not always coincide. For instance, Modha and Singh, in their work 'Network architecture of the long-distance pathways in the macaque brain' (Modha and Singh, 2010) studied the information contained in the 'Collation of Connectivity data on the Macaque brain' (CoCoMac) neuroinformatic database in order to construct the most comprehensive long-distance network of the Macaque brain. This database contains 410 anatomical tracing studies, 10,681 connectivity relations and 16,712 mapping relations, and after collation of all connections, a final network of 383 brain regions and 6602 long-distance brain connections that travel through the brain's white matter were obtained. However, to construct this network, the authors had to solve problems related with the multiplicity of brain maps, divergent nomenclature, boundary uncertainty, different resolutions depending on the work studied. In this context, the development of fast and cheap computational methods in order to collate connectivity information becomes a goal of major importance.

One possible solution to this problem is the use of Quantitative Structure–Activity/Property Relationships (QSAR/QSPR) models, which have been traditionally studied in the field of chemoinformatics and are used to predict the biological activity of drugs (QSAR) or physicochemical properties of organic compounds (QSPR) using as input structural parameters of the system under study (Puzyn et al., 2010). In the case of global studies (properties of full system) these parameters are Topological Indices (TIs) derived from the graphical representation of the system (molecule, etc.). On the other hand, we can use node centralities or local TIs of a sub-graph if we want to predict a local property of part of the system (local chemical reactivity, biotransformation of a toxicophore group in a drug, etc.). Currently, the use of QSPR-like models in which the inputs are graph parameters is not limited to the study of molecules and has been extended to other complex systems (González-Díaz and Munteanu, 2010).

Specifically, Shannon entropy is one of the most useful parameters used as input in QSAR/QSPR studies to quantify structural information of molecular graphs (Dehmer et al., 2009). In all the above-mentioned cases, Shannon entropy parameters can be used to quantify structural information locally (nodes, edges, paths, clusters, etc.) and/or globally (full graph). In fact, we have used Markov Chain (MC) to calculate Shannon entropies locally or globally within a graph considering all possible branches at different topological distances. The information is quantified in terms of $\theta_k(j)$ values, which are called the Markov–Shannon entropy node centralities of order kth for all jth states (nodes) of a MC associated to the system. This MC is expressed by a Markov or Stochastic matrix ($\mathbf{\Pi}_1$) and represented by a graph of the studied system. The elements of $\mathbf{\Pi}_1$ are the probabilities ${}^1p_{ij}$ with which the ith and jth nodes connect each other (there is a physical or functional tie, link, or relationship) within a graph. Using Chapman–Kolmogorov equations it is straightforward to realize the way to calculate $\theta_k(j)$ values for all nodes in a graph. We can use these values directly or sum some of them to obtain total or local entropies (see Section 2). Our group has introduced the software called MARCH-INSIDE (Markovian Chemicals In Silico Design), which has become a very useful tool for QSAR/QSPR studies (Gonzalez-Diaz et al., 2010). This software can calculate 1D (sequence), 2D (connectivity in the plane) and 3D (connectivity in the space) MC parameters, including $\theta_k(j)$ values, for many molecular systems. MARCH-INSIDE is able to characterize small molecules (drugs, metabolites, organic compounds), biopolymers (gene sequence, proteins sequence or 3D structure, and RNA secondary structure) and artificial polymers but can perform a limited manage of other complex networks. It happens because MARCH-INSIDE can read, transform into Markov matrix, represent as graph, and calculate entropies for molecular formats (.mol or SMILE .txt files for drugs, .pdb for proteins, or .ct files for RNAs) but it is unable to upload formats of Complex Networks (.mat, .net, .dat, .gml, etc.).

In this work, we use for the first time QSPR-like models able to assess the quality of the connectivity of new complex networks assembled with information obtained from many sources not totally accurate. The idea is to seek a QSPR-like model that use as input the

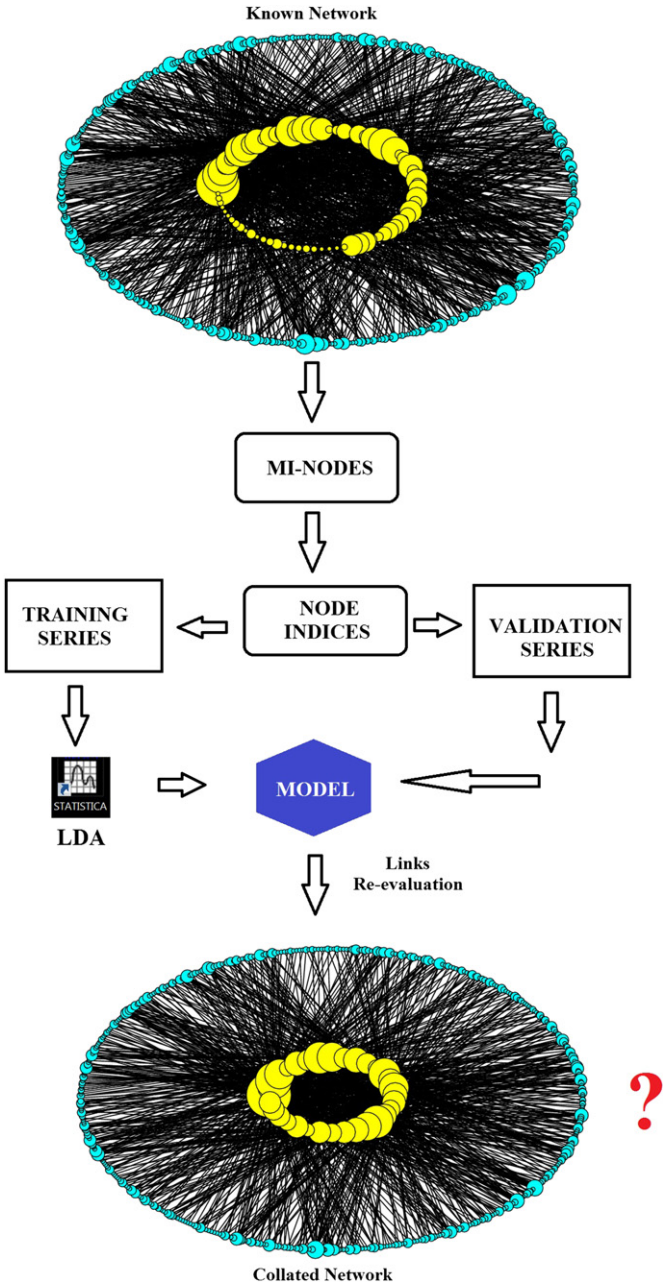

Fig. 1. General workflow used in this work.