Computers in Biology and Medicine 97 (2018) 113–123

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Complex network theory for the identification and assessment of candidate protein targets

Ken McGarry [a,*], Sharon McDonald [b]

[a] Faculty of Health Sciences and Well-being, University of Sunderland, City Campus, Sunderland, SR1 3SD, UK
[b] Faculty of Computer Science, University of Sunderland, St Peters Campus, Sunderland, SR6 ODD, UK

ARTICLE INFO

Keywords:
Complex network theory
Link-clustering
Protein interactions
Ontologies

ABSTRACT

In this work we use complex network theory to provide a statistical model of the connectivity patterns of human proteins and their interaction partners. Our intention is to identify important proteins that may be predisposed to be potential candidates as drug targets for therapeutic interventions. Target proteins usually have more interaction partners than non-target proteins, but there are no hard-and-fast rules for defining the actual number of interactions. We devise a statistical measure for identifying hub proteins, we score our target proteins with gene ontology annotations. The important druggable protein targets are likely to have similar biological functions that can be assessed for their potential therapeutic value. Our system provides a statistical analysis of the local and distant neighborhood protein interactions of the potential targets using complex network measures. This approach builds a more accurate model of drug-to-target activity and therefore the likely impact on treating diseases. We integrate high quality protein interaction data from the HINT database and disease associated proteins from the DrugTarget database. Other sources include biological knowledge from Gene Ontology and drug information from DrugBank. The problem is a very challenging one since the data is highly imbalanced between target proteins and the more numerous nontargets. We use undersampling on the training data and build Random Forest classifier models which are used to identify previously unclassified target proteins. We validate and corroborate these findings from the available literature.

1. Introduction

Protein interactions play a key role in the majority of activities occurring in the cell and participate in communications between cells [24]. The connectivity patterns of the interacting proteins can be modeled by complex network theory (graph theory) which can provide a statistical explanation of these activities and processes [21]. Integrating clustering methods with complex networks has enabled further insights, revealing the modular nature of proteins [28]. Proteins are often cooperate in modules and may be shared between several different cellular activities. Those proteins with a large number of verified interactions are classed as hub proteins. If they are implicated in one disease it is possible they may be participating in other disorders [23]. It should be noted that high connectivity (degree) or hubness does not necessarily imply that a given protein is important in some way with respect to disease. In this work we investigate the degree of protein connectivity patterns and also the location of a proteins position in the local network with respect to its predisposition to be a drug target (see Fig. 1).

The majority of disease causing genes are generally implicated with a single or small number of disorders although there are striking exceptions. The tumor suppressor gene TP53 appears to be involved with up to ten related diseases [12]. This gives credence to the disease network theory which is providing a new insights regarding how diseases occur [5]. Some diseases are more difficult to resolve, often a module of cooperating proteins can compensate for malfunctions of individual proteins. Consequently, making the identification of the faulty biological process more difficult to identify [17]. The idea of structural motifs may be a good candidate to help resolve the challenges such as cellular organization [3]. We can improve our knowledge and understanding of the mechanisms of disease based on a better understanding of protein targets and non-targets and may suggest alternative therapeutic interventions [13, 22].

However, any potential for a protein to be drug target implies it must possess a particular shape that can bind/interact with drug-like molecules i.e. it must contain a binding site. Recent research has investigated the role of the types of proteins such as G-protein coupled receptors, ion

* Corresponding author.
E-mail address: ken.mcgarry@sunderland.ac.uk (K. McGarry).

https://doi.org/10.1016/j.compbiomed.2018.04.015
Received 17 February 2018; Received in revised form 21 April 2018; Accepted 21 April 2018

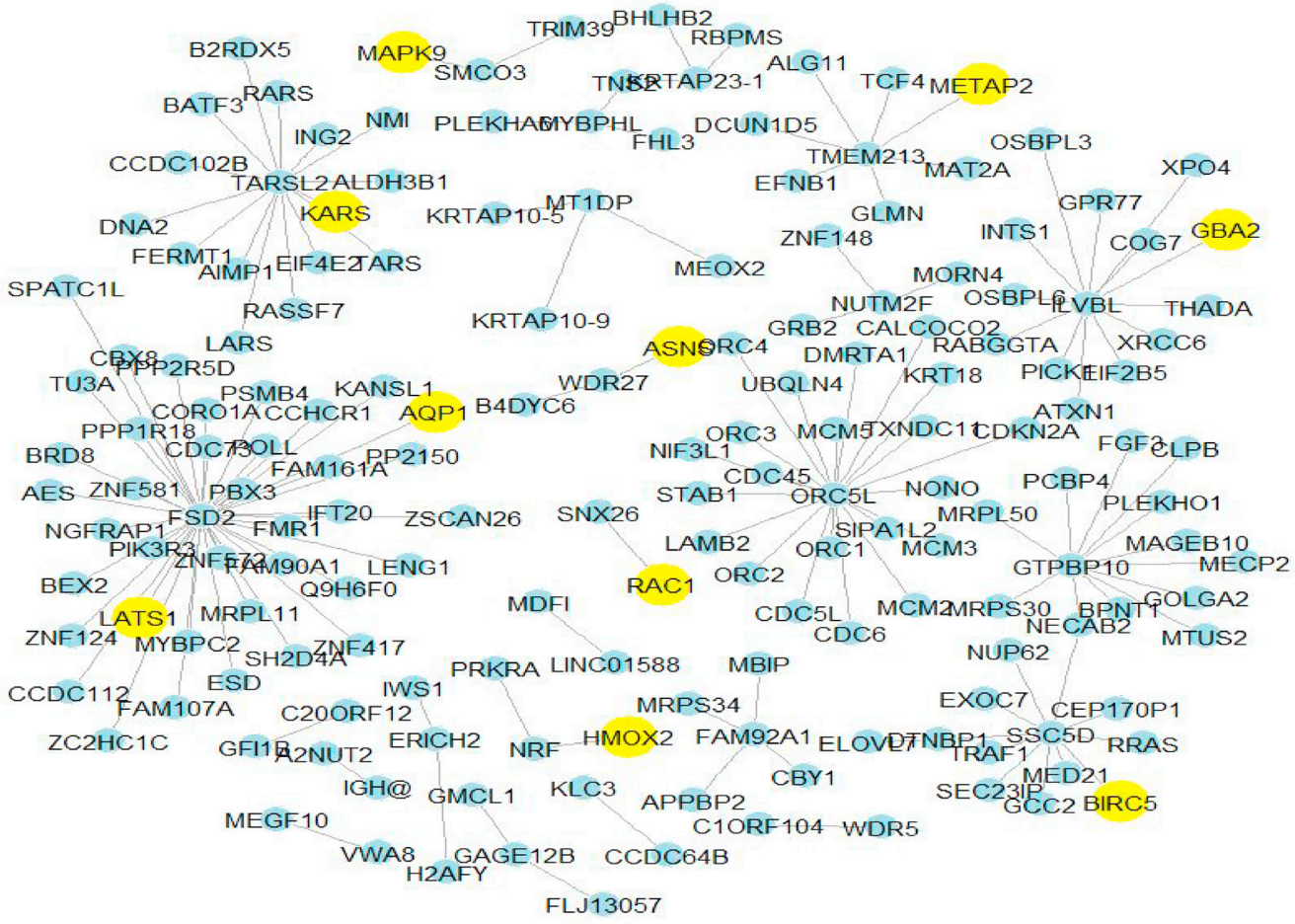

Fig. 1. Small fraction of the protein network with drug targets colored yellow and slightly larger in size, non-targets are colored light blue. However, based on their connectivity patterns their biological and complex network statistics some of the non-targets may be prove to be viable drug targets. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

channels and kinases [7]. This work determined that a proteins relationship to the membrane and it's hydrophobicity may play an important role. However, it does beg the question, how many potential protein targets are out there? [25]. One analysis suggested there may be between 2000 and 3000 proteins that are potentially druggable candidates [26]. Another approach was able to identify 668 proteins that are currently not drug targets but that have target-like potential [2]. Some proteins may be completely undruggable, while others can only be perturbed by targeting their network neighborhood proteins. Currently, complete knowledge of the proteome and interaction targets is some years away from completion [4].

1.1. Related work

The technique developed by Yu et al, considers the problem as one of module distance estimation with the understanding that the human interactome is still incomplete and with all the uncertainty inherent [31]. Yu's ultimate goal was concerned with repositioning drugs for different diseases. The modules are composed of drug-protein pairs and all are involved with cancer specific functions. The disease module distance metric was able to identify several candidate drugs. The MBiRW method developed by Luo et al uses a bi-random walk to measure similarity of drugs and diseases [20]. MBiRW uses novel similarity measures and is well validated against gold standard data but lacks target information and biologically relevant information. The CommWalker algorithm devised by Luecken uses a random walk approach to sample the proteins assigned to functional modules [19]. For robustness, the modules are formed by three different link analysis procedures and an average walk will produce a goodness of fit value. The walks are terminated when they have approach a critical value. At each step the functional GO annotation is averaged out to calculate the module homogeneity, scores are then combined to enable each module to be ranked on its biological plausibility.

The closest work to ours tackles the challenges and opportunities of integrating biological knowledge in the form of annotations from gene ontology (GO). For example, Hsing *et al* used GO to build classifiers to identify hub proteins which are highly connected proteins with many interaction partners [14]. However, the classifiers performed badly on some proteins through lack of suitable annotations. Work by Zhang *et al* explored the issues of identifying protein interaction partners through use of GO terms [32]. Support Vector Machine classifiers were constructed on the GO annotated PPI data and good accuracy was achieved on predicting the likely interaction partners. Research by Fu *et al* explored the likelihood that intrinsic disorder proteins will form highly interconnected hubs and potentially drug targets [11]. Again, the usefulness of GO was employed to annotate and analyze the relationships.