Biochemistry and Biophysics Reports xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Biochemistry and Biophysics Reports

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

Gastric cancer biomarkers; A systems biology approach

Mohammad Saberi Anvar[a], Zarrin Minuchehr[a,*], Mohsen Shahlaei[b], Samira Kheitan[a]

[a] Department of Systems Biotechnology, National Institute of Genetic Engineering and Biotechnology, Tehran, Iran
[b] Nano Drug Delivery Research Center, Kermanshah University of Medical Sciences, Kermanshah, Iran

ARTICLE INFO

Keywords:
Gastric cancer
Systems biology
Neurotrophin signaling pathway
HNF4A
TAF1
TP53

ABSTRACT

Gastric cancer is one of the most fatal cancers in the world. Many efforts in recent years have attempted to find effective proteins in gastric cancer. By using a comprehensive list of proteins involved in gastric cancer, scientists were able to retrieve interaction information. The study of protein-protein interaction networks through systems biology based analysis provides appropriate strategies to discover candidate proteins and key biological pathways.

In this study, we investigated dominant functional themes and centrality parameters including betweenness as well as the degree of each topological clusters and expressionally active sub-networks in the resulted network. The results of functional analysis on gene sets showed that neurotrophin signaling pathway, cell cycle and nucleotide excision possess the strongest enrichment signals. According to the computed centrality parameters, HNF4A, TAF1 and TP53 manifested as the most significant nodes in the interaction network of the engaged proteins in gastric cancer. This study also demonstrates pathways and proteins that are applicable as diagnostic markers and therapeutic targets for future attempts to overcome gastric cancer.

1. Introduction

Gastric cancer is the third cause of death by cancer and the fifth most common cancer worldwide [1]. Like most other types of cancer, in addition to genetic factors non-genetic factors such as smoking, alcohol consumption, poor diet, physical inactivity, viral infections and stress increase the risk of being affected by this type of cancer [2,3]. Furthermore, the role of *H. pylori* infection in gastric cancer has been proven [4,5].

Numerous studies investigate the causes and genetic factors involved in gastric cancer, where effective proteins have been identified in the cancer's pathogenesis, and most often the expression level of receptor tyrosine-protein kinase erbB-2 (ERBB2) which increases its levels in gastric cancer [6,7]. Likewise, cellular tumor antigen p53 involved predominantly in cell division regulation and apoptosis induction, mutates in most cancers [8–10]. As such, Gastrokine (GKN1) reducing the expression of gastrin-CCKBR signaling pathway is capable of preventing gastric cancer [11–14]. This protein can also prevent the invasion of cancer cells into other tissues through inactivation of NF-kappaB pathway [15].

In addition to the aforementioned proteins, other biological molecules involved in cancer were detected, including miR-145 which prevents the tumor formation through a vitamin D3-dependent pathway, and its expression level decreases in gastric cancer cells [16].

As intracellular operator units, proteins interact with other molecules for their function in the cell. Disease or health condition of an organism can be determined by such interactions [17]. Deployment of interactions between proteins and their related networks remains a determinative method in biological cell studies. Investigating and constructing such networks improves our knowledge of physiological mechanisms in disease and health conditions [18,19].

In high throughput based methods applied to identify the potential treatment or diagnosis targets, only genes or proteins with significant expressional changes are applicable; a single criteria cannot be an indicator, because proteins such as various types of kinases have not shown significant expressional changes while their participation in a variety of cancers is certified [20]. On the other hand, it has been proven that elevated levels of varied proteins is not due to cancer, rather a result of increased physiological requirements [21].

Defining a threshold for output data in these methods would result in excluding unchanged level proteins from the study and considering less effective proteins. In this study, we have tried to pass through these problems using a new perspective based on recent systems biology methods. We hypothesized that the involved proteins in cancer are concentrated in limited numbers of cellular signaling pathways which lead to subsequent cellular changes.

To date, the only curing method for gastric cancer therapy is surgery, where chemotherapy has very limited effect, if any, lowering the

* Corresponding author.
E-mail address: minuchehr@nigeb.ac.ir (Z. Minuchehr).

https://doi.org/10.1016/j.bbrep.2018.01.001
Received 30 September 2016; Received in revised form 12 November 2017; Accepted 3 January 2018

quality of life. An urgent need for alternative curing strategies leads us to study the protein-protein interaction networks through systems biology-based approaches as an appropriate methodology to discover candidate proteins and key biological pathways in this mortal disease, which claims over 700,000 deaths each year [22]. Identifying these proteins and pathways according to the proposed systems biology enables us to introduce potential therapeutic targets as well as key diagnostic markers for gastric cancer.

2. Materials and methods

2.1. Collecting proteins involved in gastric cancer

PubMed database (www.ncbi.nlm.nih.gov/pubmed) search was performed using "gastric cancer" keyword, limited to the article title and excluding case reports. Validated introduced proteins were extracted from published articles since 2014.

2.2. Interaction network construction

Interaction information of introduced proteins in literature was gathered from well-known protein-protein interaction databases including Reactome [23,24], KEGG [25], BIND [26], CCSB [27], DIP [28], GRID [29], HPRD [30], IntAct [31], MINT [32] and MDC [33] using MiMI algorithm in Cytoscape 2.8.3 platform [34]. Information from different sources was retrieved and merged in MiMI repository based on an intelligent integration strategy [35].

2.3. Identification of sub-networks

In order to identify topologically highly dense areas in the network, clusters were determined using Cytoscape plugins including AllegroMCODE and clusterMaker which is based on Community Clustering (GLay) and Markov Clustering (MCL) algorithms. The clustering algorithm of AllegroMCODE plugin is Molecular COmplex DEtection (MCODE), based on node weighting according to the local neighborhood density. This algorithm performs in three steps including vertex weighting, molecular complex prediction and post-processing to filter or add proteins to the predicted complex [36]. MCODE parameters included Degree cutoff: 2, Node score cutoff: 0.5, K-score: 5 and Max. Depth: 100.

GLay by dynamically linking highly optimized C functions JAVA program, provides assorted collection of versatile community structure algorithms and graph layout functions for network clustering and structured visualization [37].

The Markov Cluster (MCL) defines a sequence of matrices by alternation of two operators on a generating matrix. It is basically all that is needed for the clustering of graphs, but it is useful to distinguish between the algorithm and the algebraic process employed by the algorithm. This algorithm performs subnet recognition based on network simulation [38].

2.4. Annotating network with microarray data

Arrayexpress database (www.ebi.ac.uk/arrayexpress) was searched using multiple criteria such as Affymetrix HGU133 plus platform, RNA assay, and simultaneous access to both cancer and non-cancer samples. Expression data of the series GSE19826 were applied for annotating the network, performing quality control and pre-processing using ArrayAnalysis modules (http://www.arrayanalysis.org) [39]. Normalizing statistical analysis, multiple-testing corrections on *p*-value and generating annotations using GEO2R software (www.ncbi.nlm.nih.gov/geo/geo2r). Were performed using imported expression data into the network, jActiveModules 3.1 detected active expression sub-networks [40,41].

2.5. Gene set analysis

The Database for Annotation, Visualization and Integrated Discovery, DAVID (https://david.ncifcrf.gov) is one of the most efficient online tools to organize and annotate heterogeneous data from high-throughput techniques such as microarray. This program includes 68 gene enrichment tools representing the gene sets in four different modules including Annotation tools, GOchart, KEGGchart and Domainchart [42–44]. Gene Ontology (GO) terms and pathways of detected sub-networks and clusters were retrieved using DAVID functional annotation tool.

2.6. Computing centrality parameters

In order to identify hubs, centrality parameters including Betweenness and Degree for each node were calculated with CentiScaPe 2.1 in the network and sub-networks were detected by jActiveModules, AllegroMCODE, GLay and MCL. The Degree index shows the number of directly connected edges to each node. The Role of a node in the linking with the rest of the network nodes is evaluated by the Betweenness index [45].

3. Results

3.1. Collecting the proteins involved in gastric cancer

We retrieved a total of 3500 articles based on the defined criteria from PubMed, which over 600 recently published articles. Sixty articles referred to 72 different proteins involved in the pathogenesis of gastric cancer (Supplementary Table 1).

3.2. Drawing interactive network and determining sub-networks

MiMI algorithm has the ability to search multiple databases. In addition to displaying interactions (edges) between seeds (protein input) and first neighbors, it can show the identified edges between first neighbors of different seeds. The primary interactive network with 1673 nodes (proteins) and 21,548 edges (interactions) was created using 72 seed proteins obtained from our bibliographic data with Cytoscape software and its MiMI plugin. We used AllegroMCODE (Table 1), GLay and MCL plugins (see Supplementary Table 2) to identify high-density areas of our constructed network, which can be sets of protein acting as a complex within the cell, and this calculation resulted in 4 subnetworks.

3.3. Loading microarray data in the network

With regard to the designated filters, we obtained only two datasets from ArrayExpress database, including E-GEOD-19826 and E-TABM-424. E-TABM-424 datasets were excluded because of the low number of samples (a tumor sample and a normal sample). After qualitative analysis of E-GEOD-19826 datasets with ArrayAnalysis, we found 7 slides lacking the necessary criteria for our statistical analyses (GSM495053 due to the paint stains on the slide, and GSM495051, GSM495057, GSM495063, GSM495071, GSM495072, and GSM495073 due to the contamination with other tissue cells), and hence they were excluded

Table 1
Sub-networks created by AllegroMCODE.

Cluster name	Score	Nodes	Edges
1	27.323	254	6940
2	7.571	182	1378
3	5.48	98	537
4	4.456	296	1319