Supplementary Information
Topological patterns in microRNA-Gene regulatory network: Studies in colorectal and breast cancer

Debarka Sengupta and Sanghamitra Bandyopadhyay*

February 11, 2013

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India.

* Corresponding Author: Sanghamitra Bandyopadhyay, email: sanghami@isical.ac.in

Contents

List of Tables

1 Terminologies Related to Network Topology

In the present work, a directed network of miRNAs and TFs are built across human genome. A directed graph (or digraph) G is a collection of a set of vertices V and a set of edges E. This, formally, is often denoted as $G = (V, E)$. For a digraph E is a set of ordered pairs of vertices. These orderings are often realized by arrows or directed edges.

1.1 Density

Density of a directed graph G is equal to the proportion of possible arcs present in the digraph. This is defined as follows:

$$\Delta(G) = \frac{|E|}{|V|(|V| - 1)}$$

Here, $|E|$ indicated the number of arcs present in the network and $|V|$ indicated the number of vertices. It is clear that the denominator of the above expression indicates the possible number of arcs. It is clear that graph density may range from [0,1].

1.2 Betweenness centrality of nodes

This measures the centrality of a node in the corresponding network. Numerically, this is the count of the shortest paths between all possible pair of vertices in the network that pass through the subject node. Formally, for any vertex v it is defined as follows:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Here, $\sigma_{st}(v)$ is the number shortest paths from node s to t that pass through v whereas, σ_{st} is the number of shortest paths from s to t.

1.3 Clustering coefficient

In a directed G, clustering coefficient of a node v is defined as follows:

$$C(v) = \frac{e_v}{k_v(k_v - 1)}$$

Here, e_v is the number of connected pairs between all neighbors of v whereas, k_n is the number of neighbors of v. Clustering coefficient of all nodes can be averaged to obtain the clustering coefficient of the concerned network. Clustering coefficient of a node and the average clustering coefficient of a network both may range between [0,1].

1.4 Strongly Connected component (SCC)

A directed graph is called strongly connected if each vertex is reachable from every other vertices. Strongly Connected Components (SCC) of a digraph G are its maximal strongly connected subgraphs. SCC containing a single node is trivial where as more than one nodes are non-trivial.

1.5 Graph diameter and radius

Graph diameter $D(G)$ is the length of the longest shortest path (i.e., the longest graph geodesic) between any two vertices in the graph.

$$\max_{u,v} d(u, v)$$

where $d(u, v)$ denotes length of the shortest path between u and v. On the other hand graph radius $R(G)$ is defined as the minimum graph eccentricity of any vertex in a graph.

$$R(G) = \min_{v} \max_{s} d(v, s)$$

1.6 Characteristic path length

Average shortest path length is known as the characteristic path length. This give the expected distance between two connected nodes.

1.7 Average number of neighbors

The average number of neighbors indicates the expected size of the neighborhood of any connected vertex. This is obtained by averaging on the individual neighborhood sizes of all the vertices in a graph.

2 Shortest Routes of Regulations Within Transcription Factors

Many possible regulatory links are found within the TF pairs of the breast and colorectal cancer specific DSCS modules. The shortest directed routes are furnished in Table 1.

3 Sources of Transcription Factor Information

Transcription factors that are implicated in the breast and colorectal cancers are manually curated from literature. The associated PUBMED IDs/ DOI are furnished fin the Table 2.

4

breast cancer related shortest routes

AP1	MIR-136	E2F1	
AP1	MIR-206	ETS1	
AP1	MIR-34A	MYB	
AP1	MIR-204	POU2F2	MYC
AP1	MIR-203	RUNX2	
E2F1	MYC		
ER	MIR-136	E2F1	
ER	MIR-155	ETS1	
ER	MIR-155	MYB	
ER	MIR-202	POU2F2	MYC
ER	MIR-155	MYB	MIR-203 RUNX2
ETS1	MYB	MIR-136	E2F1
ETS1	MYB		
ETS1	MYB	MIR-210	POU2F2 MYC
ETS1	MYB	MIR-203	RUNX2
MYB	MIR-136	E2F1	
MYB	MIR-206	ETS1	
MYB	LET-7A	POU2F2	MYC
MYB	MIR-203	RUNX2	
P53	MIR-136	E2F1	
P53	MIR-206	ETS1	
P53	MIR-34C-5P	MYB	
P53	MIR-204	POU2F2	MYC
P53	MIR-204	RUNX2	
PAX5	MIR-136	E2F1	
PAX5	MIR-206	ETS1	
PAX5	MIR-34A	MYB	
PAX5	MIR-202	POU2F2	MYC
PAX5	MIR-203	RUNX2	
POU2F2	MYC		
WT1	MIR-136	E2F1	
WT1	MIR-206	ETS1	
WT1	MYB		
WT1	LET-7A	POU2F2	MYC
WT1	MIR-203	RUNX2	

colorectal cancer related shortest routes

EGFR	MIR-21	KLF5	MYC	MIR-106B	POU3F2	MIR-222	ETS2
EGFR	MIR-21	KLF5	MYC	MIR-34A	KLF4		
EGFR	MIR-21	KLF5					
EGFR	MIR-21	KLF5	MYC				
EGFR	MIR-21	KLF5	MYC	MIR-106B	POU3F2	MIR-181B	PROX1
EGFR	MIR-21	KLF5	MYC	MIR-106B	STAT3		
EGFR	MIR-21	KLF5	MYC	MIR-106B	TCF4		
GATA	MIR-7	EGFR					
GATA	MIR-222	ETS2					
GATA	MIR-29A	KLF4					
GATA	MIR-145	KLF5					
GATA	MIR-34B	MYC					
GATA	MIR-181B	PROX1					
GATA	MIR-181B	STAT3					
GATA	MIR-29A	TCF4					
KLF5	MYC	MIR-93	POU3F2	MIR-7	EGFR		
KLF5	MYC	MIR-93	POU3F2	MIR-222	ETS2		
KLF5	MYC	MIR-29A	KLF4				
KLF5	MYC						
KLF5	MYC	MIR-93	STAT3	MIR-181B	PROX1		
KLF5	MYC	MIR-93	STAT3				
KLF5	MYC	MIR-29A	TCF4				
MYC	MIR-141	POU3F2	MIR-7	EGFR			
MYC	MIR-141	POU3F2	MIR-222	ETS2			
MYC	MIR-29A	KLF4					
MYC	MIR-141	KLF5					
MYC	MIR-93	STAT3	MIR-181B	PROX1			
MYC	MIR-93	STAT3					
MYC	MIR-141	TCF4					
P53	MIR-7	EGFR					
P53	MIR-130B	POU3F2	MIR-222	ETS2			
P53	MIR-135B	KLF4					
P53	MIR-141	KLF5					
P53	MIR-145	MYC					
P53	MIR-181B	PROX1					
P53	MIR-106B	STAT3					
P53	MIR-130B	TCF4					
STAT3	MIR-181B	POU3F2	MIR-7	EGFR			
STAT3	MIR-181B	POU3F2	MIR-222	ETS2			
STAT3	MYC	MIR-29A	KLF4				
STAT3	MYC	MIR-141	KLF5				
STAT3	MYC						
STAT3	MIR-181B	PROX1					
STAT3	MYC	MIR-29A	TCF4				

Table 1: Shortest routes among deregulated TFs are enumerated for breast and colorectal cancer.

TFs related to breast cancer	Source	TFs related to colorectal cancer	Source
p53	doi: 10.1677/erc.0.005027	TCF4	PMID: 21983179
Myc	doi: 10.1677/erc.0.005027	PROX1	PMID: 21970873
myb	PMID: 21953443	TNF	PMID: 21945666
wt1	PMID: 21710692	P53	PMID: 21927023
ER	PMID: 18312651	EGFR	PMID: 21924350
RUNX2	PMID: 15665096	KLF5	PMID: 21885866
ETS1	PMID: 20392592	MYC	PMID: 21851818
AP1	PMID: 17637753	STAT3	PMID: 21840932
E2F1	PMID: 21573702	KLF4	PMID: 21814594
PAX5	PMID: 21710692	GATA	PMID: 21779441
POU2F2	PMID: 19424592	ETS2	PMID: 21763315

Table 2: Manually curated transcription factors related to breast and colorectal cancer with sources.