Complexity and Fitness of Methylated DNA sites and RNAs in the context of Breast Cancer Tissues

June 27, 2023

1 Aim of the study

Highlighting the nestedness structure of the bipartite network representing correlations between methylated DNA sited and RNAs.

2 Methods and Data

2.1 Nestedness

Nested structures appear in the nature such that in ecological and socio-economic systems. In ecological systems, for instance, there is a nested distribution of species habitats. Where as some species only live in particular locations, some others live in a large variety of environments including those particular locations. In the context of network theory, a perfectly nested structure is such that for any pair of node (i, j) with $k_i < k_j$ their respective degree (number of neighbors), the set of nodes i is interacting with is included in the set of nodes j is interacting with. The adjacency matrix of such a network shows a triangular shape when columns and rows are sorted accordingly to the degree ordering of respective nodes. There are many metrics (other than degrees) permitting to infer the nested structure of a given network. Nestedness can appear in unipartite and bipartite networks. In the context of economical systems, especially the world trade network seen as a country-product bipartite network, recent research shows the existence of nestedness, see a complete review in [**mariani19**]. In the context of biological systems nestedness structures have also been studied. In [**cantor17**] different biological scales are considered. However, it seems that DNA-RNA interactions have not been studied yet. We choose the metric called fitness-complexity introduced in [**tacchella12**] to investigate the nestedness structure of DNA-RNA bipartite network. Fitness and complexity is a non-linear and iterative algorithm permitting to infer the nestedness in bipartite network such as economical networks. In this context, the fitness is a country quality, and complexity is a product quality. In our approach we try to make an analogy where RNAs play the role of products and DNAs the role of countries. Here are the economical definitions of fitness and complexity.

Fitness and Complexity metrics

$$\begin{cases} F_d^{(n)} = \sum_r B_{dr} Q_r^{(n-1)} \\ Q_r^{(n)} = \frac{1}{\sum_d B_{dr}(1-F_c^{(n-1)})} \end{cases} \tag{1}$$

with B_{dr} an element of the bi-adjacency matrix representing the correlation (positive or negative) interaction between a Methylated DNA d and an RNA r, $F^{(n)}$ and $Q^{(n)}$ the Fitness and Complexity vector measured at iteration n.

Product's complexity

2.2 Data

We used an open access cancer dataset from GDC Data Portal. From raw data consisting in methylated DNA and RNA's beta-values for up to 841 tumorous and normal breast tissues, we built a Matrix of Pearson Correlation Coefficients between Methylated DNA and RNA from these beta-values. A network representing the Pearson correlation coefficients between pairs (M-DNA,RNA) consists in a bipartite graph with coefficients as weight of the links containing. We also project DNAs into spacial clusters of length 40K base-pairs, this turns $N_{DNA} = 364285$ DNAs into a set of $N_{cl} = 93690$ clusters. The nature of correlations are either positive or negative.

We propose to measure complexity of RNAs from the complexity-fitness metric proposed in [**tacchella12**].

3 Results

Table 1: Top 10 DNAs and RNAs in terms of fitness and complexity scores, 715 RNAs and 1980 Clusters, with $\rho_c = 0.7$.

RANK	CLUSTER ID	Fitness Score	RNA	Complexity Score
1	Cluster 63264	0.0525	TNMD	0.0041
2	Cluster 66370	0.0291	LIPE	0.0041
3	Cluster 73992	0.0196	ASPA	0.0041
4	Cluster 34909	0.0195	KCNIP2	0.0041
5	Cluster 18410	0.0141	CD36	0.0041
6	Cluster 93686	0.0140	RDH5	0.0041
7	Cluster 83354	0.0129	ALDH1L1	0.0041
8	Cluster 55962	0.0121	GLYAT	0.0041
9	Cluster 11394	0.0099	PLIN1	0.0041
10	Cluster 5042	0.0083	GPD1	0.0041

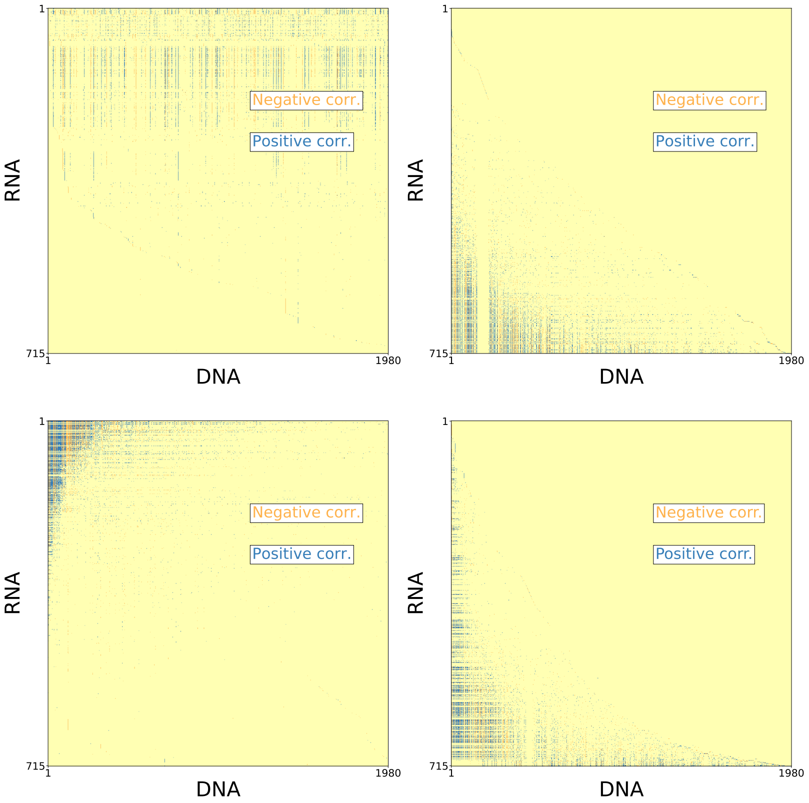

Figure 1: Binary bi-adjacency matrix representing the interactions between Clustered DNAs and RNAs involved in the whole data set. Matrix entries related to positive correlations are in blue, and negative in orange, finally empty cells are in bright yellow. Raw matrix (top left), Fitness and Complexity based reorganized matrix (top right), Degree based reorganized matrix (bottom left) and matrix with cols and rows reorganized from BINMATNEST algorithm. Here we have considered all Pearson correlation such that $\rho_c = 0.7$.

Table 2: Last 10 DNAs and RNAs in terms of fitness and complexity scores, 715 RNAs and 1980 Clusters, with $\rho_c = 0.7$.

RANK	CLUSTER ID	Fitness Score ($\times 10^{-5}$)	RNA	Complexity Score ($\times 10^{-5}$)
1	Cluster 86765	0.2	FOXC1	0.6
2	Cluster 86562	0.2	IL2RG	1
3	Cluster 81797	0.2	RGMA	1
4	Cluster 81555	0.2	SNX20	1
5	Cluster 81499	0.2	TBX21	1
6	Cluster 76088	0.2	PTPN7	1
7	Cluster 72652	0.2	SIT1	1
8	Cluster 71468	0.2	STAC	1.1
9	Cluster 70729	0.2	SLAMF6	1.1
10	Cluster 69709	0.2	BTLA	1.2

Table 3: Nestedness scores of the network obtained from $\rho_c = 0.7$. The NODF (BINMATNEST) score goes from 0 (form 100) for non nested network to 1 (to 0) for highly nested network.

NODF	Temperature
0.7	0.3

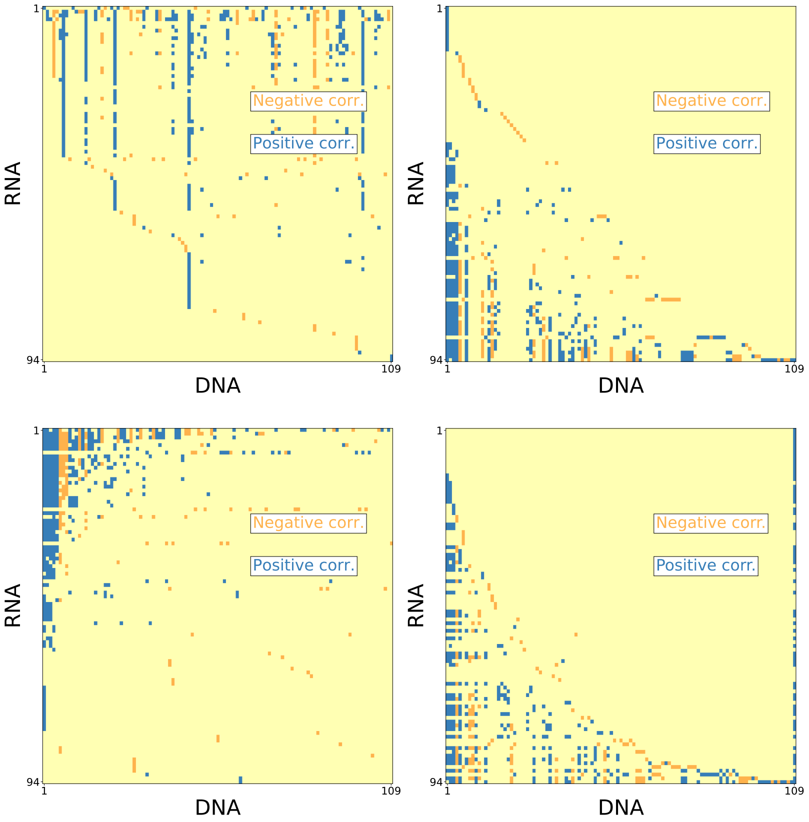

Figure 2: Binary bi-adjacency matrix representing the interactions between Clustered DNAs and RNAs involved in the whole data set. Matrix entries related to positive correlations are in blue, and negative in orange, finally empty cells are in bright yellow. Raw matrix (top left), Fitness and Complexity based reorganized matrix (top right), Degree based reorganized matrix (bottom left) and matrix with cols and rows reorganized from BINMATNEST algorithm. Here we have considered all Pearson correlation such that $\rho_c = 0.8$.

Table 4: Top 10 DNAs and RNAs in terms of fitness and complexity scores, 94 RNAs and 109 Clusters, with $\rho_c = 0.8$.

RANK	CLUSTER ID	Fitness Score	RNA	Complexity Score
1	Cluster 34909	0.2464	GZMK	0.0247
2	Cluster 83354	0.0800	CD96	0.0247
3	Cluster 18410	0.0712	EOMES	0.0247
4	Cluster 5042	0.0676	GPR171	0.0247
5	Cluster 70240	0.0489	TRBV19	0.0247
6	Cluster 80727	0.0411	TRAV8-3	0.0247
7	Cluster 11394	0.0339	TRAV12-3	0.0247
8	Cluster 24330	0.0222	TRBV2	0.0247
9	Cluster 34745	0.0213	TRBV3-1	0.0247
10	Cluster 50492	0.0196	LINC00861	0.0247

Table 5: Last 10 DNAs and RNAs in terms of fitness and complexity scores, 94 RNAs and 109 Clusters, with $\rho_c = 0.8$.

RANK	CLUSTER ID	Fitness Score	RNA	Complexity Score
1	Cluster 91366	0.0002	SNX20	0.0005
2	Cluster 89519	0.0002	PTPN7	0.0006
3	Cluster 86376	0.0002	IL2RG	0.0006
4	Cluster 66076	0.0002	SIT1	0.0009
5	Cluster 55494	0.0002	SIRPG	0.0009
6	Cluster 54291	0.0002	SLA2	0.001
7	Cluster 46347	0.0002	FOXC1	0.0011
8	Cluster 33213	0.0002	TBX21	0.0012
9	Cluster 32767	0.0002	SLAMF6	0.0012
10	Cluster 17458	0.0002	SP140	0.0013

Table 6: Nestedness scores of the network obtained from $\rho_c = 0.8$. The NODF (BINMATNEST) score goes from 0 (form 100) for non nested network to 1 (to 0) for highly nested network.

NODF	Temperature
0.8	0.9