Bioinformatics 2018-2019

Date: 21 01 2019

Project report

|  |
| --- |
| Bioinformatics@Data Science A.Y. 2018-2019  Network Biology Project, Part 2  Vigèr Durand Azimedem Tsafack1 Pratuat Amatya1 Malick Alexandre Ngorovitch Sarr1  1Group no. 2 Abstract The Chagas disease is a parasitic disease caused by the protist Trypanosoma cruzi. Mainly found in South America, It is spread mostly by insects known as Triatominae, or "kissing bugs". In this project, we will be carring out network analysis of the human genes involved in the disease. The relevant set of genes was provided by the instructor and previously analysed. We will be performing modules discovery by applying clustering methods and then carry on with a enrichment analysis on the putative disease modules. |

Introductions

Network Measures for Seed Genes , Interaction and Union Interactomes

Explain briefly the methods you followed to get the information about the seed genes and add the related table (see table format below).

With the results of the various interactomes previously obtained, we calculated the global measures of the various networks. Indeed the global measures include:

* Number of nodes, which will in our case represent the unique genes present in our interactomes
* Number of edges, which are the connection shared between the interactome’s genes
* Number of connected components, which corresponds to the a group of highly connected nodes forming a group.
* Number of isolated nodes, which corresponds to the nodes with degree 0. This should technically be 0 due to the previous preprocessing of the data.
* Average path length, corresponds to the average shortest path between two genes. The smaller it is, the easier it is for two proteins to communicate between each other.
* Average degree, is the average number of connection a nodes has with another
* Average clustering coefficient, shows how easy it is to split the graph into clusters
* Network diameter, which is the maximum eccentricity of any vertex in the graph (or in other work the longuest shortest path)
* Network radius, which is the minimum eccentricity
* Centralization, which refers to the overall cohesion or integration of the graph. The higher the number, the more chances there are that the graph is centralized around particular points or sets of points.

Table 2.1 shows the various Global measures for the Seed Gene Interactome (SGI) and the Union interactome U. We unfortunately could no manage to extract the global measures for the intersection interactome (I) since the number of nodes were below 20.

In order to have a better undestanding between the connection involved with the genes, we decided to focus our attention to the LCC for each interactomes. For each LCC, we calculated the various local properties which are:

* Node degree, which shows the number of connection attached to a particular node
* Betweenness centrality, which shows the amount of influence one node has over the information flow of the network. Ergo, proteins having a high betweeness will have an important function in the interaction within the network[1]
* Eigenvector centrality, measures the influence of a gene within the network. The idea is to rank nodes that are connected to high valued nodes. Thus, proteins with high eigenvalues will interact with key proteins withing a network. [2]
* Closeness centrality, refers to the ability of the network nodes that are able to spread information very efficiently.[3]
* ratio Betweenness/Node degree. The higher the ratio the more influence a particular node has.

Table 2.2 shows the LCC for both intersection and union interactomes. Unfortunately we did not have a large network for the intersection dataset, analysing it is almost irrelevant. However we have a decent amount of nodes and interaction for the union interactome. Table 2.3 and 2.4, show the local measures for the LCC for intersection and union interactomes respectively. The union interactome results were limited to the nodes with degree > 20; for the full table, please consult the files attached to this report.

Table 2.1 Global Measures for SGI and U interactomes

|  |  |  |
| --- | --- | --- |
|  | **SGI** | **U** |
| **Number of nodes** | 16 | 1918 |
| **Number of links** | 49 | 3428 |
| **Number of connected components** | 1 | 1 |
| **Number of isolated nodes** | 0 | 0 |
| **Average path length** | 1.975 | 3.085068943 |
| **Average degree** | 6.125 | 3.57455683 |
| **Average clustering coefficient** | 0.482738095 | 0.242292188 |
| **Network diameter** | 4 | 6 |
| **Network radius** | 2 | 3 |
| **Degree centralization** | 0.219047619 | 0.373067913 |

Table 2.2 Global Measures for Largest Connected Component(LCC) in I and U interactomes

|  |  |  |
| --- | --- | --- |
|  | **I Largest component** | **U Largest component** |
| **Number of nodes** | 3 | 1918 |
| **Number of links** | 3 | 3428 |
| **Number of connected components** | 1 | 1 |
| **Number of isolated nodes** | 0 | 0 |
| **Average path length** | 1.333333333 | 3.085068943 |
| **Average degree** | 2 | 3.57455683 |
| **Average clustering coefficient** | 0 | 0.242292188 |
| **Network diameter** | 2 | 6 |
| **Network radius** | 1 | 3 |
| **Degree centralization** | 3 | 0.373067913 |

Table 2.3 Local measures for LCC of intersection interactomes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Node degree** | **Betweenness centrality** | **Eigenvector centrality** | **Closeness centrality** | **Ratio Betweenness/Node degree** |
| **P01579** | 4 | 1 | 0.816496581 | 1 | 0.25 |
| **Q9HD26** | 1 | 0 | 0.40824829 | 0.666666667 | 0 |
| **P38484** | 1 | 0 | 0.40824829 | 0.666666667 | 0 |

Table 2.4 Local measures for LCC of Union Interactome with Node degree > 20

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Node degree** | **Betweenness centrality** | **Eigenvector centrality** | **Closeness centrality** | **Ratio Betweenness/Node degree** |
| **P01137** | 718 | 0.446893672 | 0.485641434 | 0.521774633 | 0.000622415 |
| **P01375** | 407 | 0.237042136 | 0.266185208 | 0.496760819 | 0.000582413 |
| **P01584** | 341 | 0.169315991 | 0.234361464 | 0.443544655 | 0.000496528 |
| **P13500** | 299 | 0.136747384 | 0.232447695 | 0.47591857 | 0.000457349 |
| **P02778** | 263 | 0.108167961 | 0.18121054 | 0.435088516 | 0.000411285 |
| **P04440** | 234 | 0.141423942 | 0.075908108 | 0.366539197 | 0.000604376 |
| **P05231** | 228 | 0.105139971 | 0.197149518 | 0.479849812 | 0.00046114 |
| **P29460** | 199 | 0.073900231 | 0.126891761 | 0.385481601 | 0.000371358 |
| **P01374** | 144 | 0.052433474 | 0.089590549 | 0.423459245 | 0.000364121 |
| **P24394** | 140 | 0.05003749 | 0.101291994 | 0.392747388 | 0.000357411 |
| **P01579** | 125 | 0.054432467 | 0.074814351 | 0.372377622 | 0.00043546 |
| **Q13651** | 81 | 0.032329937 | 0.05049627 | 0.386491935 | 0.000399135 |
| **P01920** | 72 | 0.014990937 | 0.03089841 | 0.350585223 | 0.000208207 |
| **P51681** | 71 | 0.019229837 | 0.040852377 | 0.368653846 | 0.000270843 |
| **P09960** | 59 | 0.045770602 | 0.00462324 | 0.324750127 | 0.000775773 |
| **Q07325** | 49 | 0.017072755 | 0.044622349 | 0.402392947 | 0.000348424 |
| **P82251** | 28 | 0.024895621 | 0.001248384 | 0.300187911 | 0.000889129 |

Clustering Methods for disease discovery

Explain briefly the methods you followed to get the information about the interaction data and add the related table. Refer clearly to different files (i.e. when necessary.

The whole idea behind seperating diseases into clusters, is based on the assusption that the proteins involved in a particular disease will tend to interact more with each other. This means that their component should theoretically form within the same cluster in the network. Thus, a disease module will correspond to a a localized region of connections between disease-related protein[4]. In this section we applied 2 clustering methods namely the Markov Clustering and the Louvain Clustering to the LCC for the intersection and union interactomes .

The Markov clustering algorithm (MCL) simulates random walks on the underlying interaction network, by alternating two operations: expansion, and inflation. The exact mathematical steps can be find in various literature but on the higher level, the main idea is to preserve and focus on flows where the information tranfered between nodes is high and disregard flows where the information transfered is weak. Clusters are identified by alternating expansion and inflation until the graph is partitioned into subsets so that there are no longer paths between these subsets. [5]

The Louvain clustering basically maximizes a modularity score for each cluster, where the modularity quantifies the quality of an assignment of nodes to clusters by evaluating how much more densely connected the nodes within a cluster are, compared to how connected they would be in a random network. [6]

After applying the clustering methods to our LCCs we recorded the modules with the number of nodes highter the 10 in which the seed genes are statistically over represented with (p<0.05) by applying a hypergeometric test in order to distinguis the putative disease modules. Unfortunately due to the lack of data the intersection interactomes turned out empty. The union interactome, represented in table 3.1, resulted with 5 of such clusters falling within the above parameters, 4 of which came from the louvain clustering. The full list of all genes is available in the files attached to this report.

Table 3.1 Putative disease module for U-interactome

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Clustering Algorithm** | **Module ID** | **Number of Seed Genes** | **Number of Genes** | **List of Seed Genes** | **List of All Genes** | **P-Value** | **Id** |
| Louvain | 9 | 1 | 36 | {'Q13651'} | {'Q9Y5Q3'…… 'Q9BXN2'} | 0.956588505 | 10 |
| Louvain | 11 | 1 | 29 | {'P82251'} | {'Q16288', ……. 'Q7Z2H8'} | 0.970920243 | 12 |
| Louvain | 12 | 1 | 35 | {'P24394'} | {'P43629', ……. 'Q96HE9'} | 0.958773299 | 13 |
| Louvain | 13 | 1 | 17 | {'P51681'} | {'P51681', ……'Q9H4G1'} | 0.98958876 | 14 |
| MCL | 9 | 1 | 29 | {'P82251'} | {'Q16288', …… 'Q7Z2H8'} | 0.970920243 | 24 |

Figure 3.1 Union interactome LCC clusters - MCL

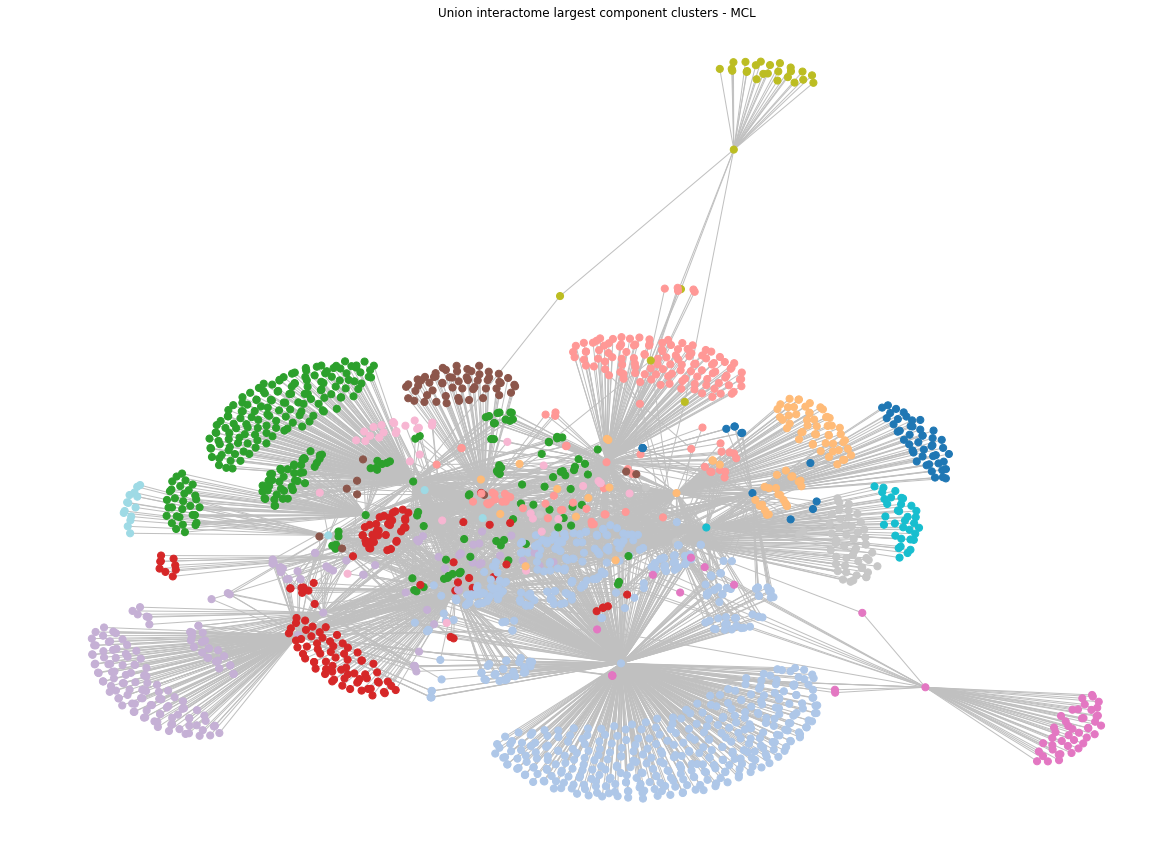
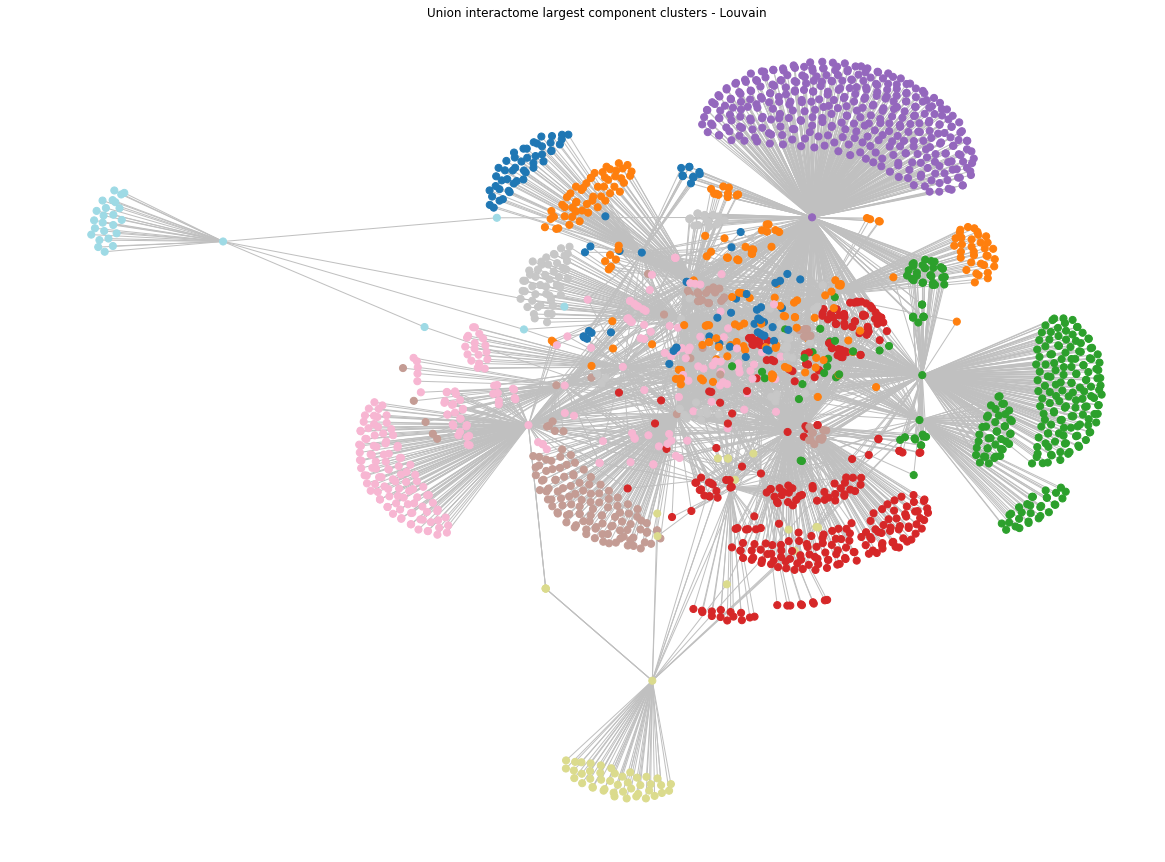


Figure 3.2 Union interactome LCC clusters - Louvain



References (if any)

[1] BR Jasny. (2015) A network approach to finding disease modules | Science.

[2] RS Wang. (2018) Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications. Brigham and Women's Hospital, Harvard Medical School, Boston,

[3] Brian Levinstein (2011) Network Centrality Using Eigenvectors*.* Wolfram. CA

[4] Kathy Macropol

[5] James Vlasblom and Shoshana J Wodak (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs

[6] Neo4j. The Louvain algorithm