Rodríguez-Bazaga, A.; Vellido, A. Network community cluster-based analysis for the identification of potential leukemia drug targets. A: International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization. "Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization: proceedings of the 13th International Workshop, WSOM+ 2019: Barcelona, Spain, June 26-28, 2019". Berlín: Springer, 2019, p. 314-323.

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-19642-4_31

Network Community Cluster-Based Analysis for the Identification of Potential Leukemia Drug Targets

Adrián Bazaga[1,2], Alfredo Vellido[3]

[1] Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
[2] STORM Therapeutics Ltd, Cambridge, CB22 3AT, UK
[3] Computer Science Department, Intelligent Data Science and Artificial Intelligence (IDEAI) Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
ar989@cam.ac.uk, avellido@cs.upc.edu

Abstract. Leukemia is a hematologic cancer which develops in blood tissue and causes rapid generation of not mature and abnormal-shaped white blood cells. It is one of the most prominent causes of death in both men and women for which there is currently not an effective treatment. For this reason, several therapeutical strategies to determine potentially relevant genetic factors are currently under development, as targeted therapies promise to be both more effective and less toxic than current chemotherapy. In this paper, we present a network community cluster-based analysis for the identification of potential gene drug targets for Acute Lymphoblastic Leukemia and Acute Myeloid Leukemia.

Keywords: Network analysis, Community clusters, Drug discovery, Therapeutic targets, Leukemia

1 Introduction

Leukemia is known to be a group of cancers that usually begin in the bone marrow and result in a high number of abnormal white blood cells. Although its prevalence is low, the chance of surviving is one of the lowest among cancer diseases. Moreover, the population in developed countries is aging incrementally, and taking into account that older people have a higher risk of developing Leukemia, a steady increase of cases is to be expected.

Leukemia involves complex genetic factors, and identifying which ones are relevant to treat the disease can make a difference between life or death. The treatments for and reactions of each type of Leukemia may vary considerably [1]. The probability of survival can be increased by methods that allow to identify types of Leukemia accurately, as well as by the use of computational methods for the discovery of relevant targets in the genome that might be of interest for drug development [2]. In this paper, we will focus on the task of discovering promising target genes for the treatment of Acute Lymphoblastic Leukemia (ALL) [3] and Acute Myeloid Leukemia (AML) [4] types, which represent almost half of the totality of Leukemia cases.

The main objective of this paper can be stated as follows: given the gene-interaction network related to ALL and AML, where some genes are known to be targets of certain drugs approved by the U.S. Food and Drug Administration (FDA), as they are highly significant for the disease at hand, and some others are not (or not known to be), we aim to assess if the topological structure of the sub-networks related to the genes belonging to the two different classes, namely target and non-target, have any specificity in terms of statistical properties; we also want to analyze if, by using network community cluster detection techniques, it is possible to find potential drug targets.

The rest of the paper is structured as follows: Section 2 introduces the basic techniques employed in our analysis, as well as the data used in the study. Then, Section 3 describes and discusses in detail the experimental findings, while Section 4 concludes the paper and points to issues deserving further research.

2 Materials and methods

2.1 Graph-theoretical centralities

Degree centrality. This paper focuses on graph community cluster analysis [5]. For this, some graph centrality measures must be defined first. The degree centrality is defined as the number of edges going into or out of a node.

Betweenness centrality quantifies the number of times a node is in the way along the shortest path between two other nodes. For a node v, it is defined as

$$B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{1}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those shortest paths that pass through v.

Closeness centrality of a node is defined as the average length of the shortest path between the node and all other graph nodes. A variant that accounts for the possibility of having a not connected graph is known as *harmonic centrality*, which is the one used in this paper. For a node v, it is defined as

$$HC(v) = \sum_{y \neq v} \frac{1}{d(y, v)}, \tag{2}$$

where $d(y, v)$ is the distance between nodes y and v, imposing that $\frac{1}{d(y,v)} = 0$ if there is no path from y to v.

PageRank centrality. The PageRank centrality [6], which is a variant of the Eigenvector centrality [7], was originally defined for a scenario where a user surfs the web by clicking links. The PageRank value of a website is an estimation of the probability that the user is on a web page at a given moment. Generalizing from webpages to network nodes, three elements determe the PageRank of a node: the number of incoming edges, the number of outgoing edges of the linking nodes, and the PageRank of the linking nodes.

2.2 Random graph null hypothesis: the Erdös-Rényi model

The Erdös-Rényi graph [8] is a random network model where edges are connected independently between each pair of nodes with a probability p that follows a Bernoulli distribution, thus they have no community (cluster) structures. This model has been widely used as a null hypothesis to find patterns in the topology and community structure of real networks [9]. For this reason, in this work we use it to study the particular properties of gene-interaction networks.

2.3 Community finding algorithms

Walktrap algorithm. The Walktrap algorithm [10] is a community detection algorithm that uses a distance metric based on performing random walks and uses a hierarchical agglomerative clustering method. Formally, if two nodes i and j belong to the same community, the probability to reach a third node k belonging to the same community by means of a random walk, should differ minimally for i and j. Then, the distance between two nodes i and j is constructed by summing these differences over all nodes, with a correction for each node degree.

Infomap algorithm. The Infomap algorithm represents the community-cluster structure of a graph by means of a two-level nomenclature based on Huffman coding [11]. It defines the problem of finding the optimal clustering of a graph as finding a description of minimum information using random walks on the graph. Moreover, the algorithm objective function is to maximize the so-called Minimum Description Length [12].

2.4 Data gathering & building the gene-interaction network

The data in our experiments is divided over different sources. First, we searched for which FDA-approved drugs are currently used to treat ALL and AML types of Leukemia. This was obtained from the U.S. National Cancer Institute [13].

On the other hand, from the Drug-Gene Interaction Database [14], the genes that are targeted by a given drug, which total 197, were obtained. Then, in order to obtain negative samples (non-target genes), we queried human gene identifiers from HumanMine, a biological database developed by the University of Cambridge [15, 16]. From a pool of 62,906 genes, 197 were randomly sampled, constrained to only those genes with at least one known interaction with another

gene, and that are not known to be a target of any disease. This was done to obtained a balanced dataset for analysis.

After that, we built the gene-interaction network of both the 197 target and non-target genes. To do so, for each gene we queried the BioGRID database [17] for the genes interacting directly with each of the genes in our dataset. Consequently, for each gene we have its direct interacting neighborhood, where some interacting gene neighbors may be shared among different genes, thus leading to a connected graph. This resulted in a network comprised of 12,761 nodes and 72,634 edges (see Table 1 for summary information).

Network	Nodes	Edges
Target Network	11966	50512
Non-target Network	11966	22122
Full	12761	72634

Table 1: Network composition

3 Experimental Results & Discussion

3.1 Description of the network structure

The illustration of the full network is provided in Figure 1, with the node sizes drawn according to their PageRank [6, 7] value. These values are linearly related to the dimension of the vertex, where a greater PageRank value corresponds to a greater dimension of the node. In Figure 1, blue nodes are the non-target genes and the red ones are the target genes, that are given by approved FDA drugs for ALL and AML. Also, orange nodes are genes that interact with either target genes, non-target genes, or both, as is the case when they have common neighbors. These orange nodes are not known to be targets or non-target genes, and thus we refer to them as unknown genes from now on.

In order to further characterize our network, we calculate two widely known graph metrics: the transitivity (also known as clustering coefficient), and the diameter. We measure these metrics for the full network, the target genes sub-network and the non-target genes sub-network. The transitivity, T, of a graph, G, is based on the relative number of triangles in the graph, compared to the total number of connected triples of nodes. Formally, transitivity T of a graph G is calculated as

$$T(G) = \frac{3 \times \text{number of triangles in the graph}}{\text{number of connected triples in the graph}}, \tag{3}$$

where the factor of three in the numerator takes into account the fact that each triangle contributes to three different connected triples in the graph, one centered at each node of the triangle.

The diameter of a graph is the length of the longest shortest path between any two graph vertices, u and v, where $d(u, v)$ is a graph distance, that is, the largest

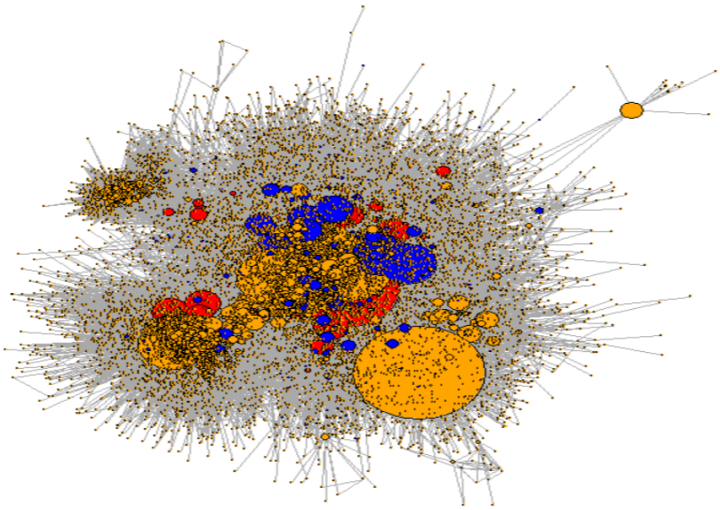

Fig. 1: Visualization of the gene-gene interaction network where node sizes are relative to their PageRank value

number of vertices which must be traversed in order to travel from one vertex to another, without taking into account loops or backtracking paths. Thus, if the shortest path between the two farthest nodes in the graph has a length of 3, the diameter of the graph will have a value of 3.

Network	Transitivity	Diameter
Target Network	0.1365	10
Non-target Network	0.1436	10
Full	0.1346	10

Table 2: Transitivity and diameter metrics of the full graph and target/non-target subgraphs

In Table 2, we show the transitivity and diameter metrics of the full network, as well as for the target and non-target sub-networks. As can be seen, the diameter in the three scenarios is the same, thus the longest shortest path is the same. Also, the transitivity of the non-target sub-network is higher than the one of the full and target networks, thus indicating that this sub-network has a higher density of connections between nodes. Consequently, in the non-target sub-network, the number of shared interacting genes between the non-target genes is higher than the corresponding one in the other networks.

In order to study the topology of the graph, we assessed its clustering coefficient as a measure of characterization of the graph topology, with respect to how

it would behave in a random scenario, using the the Erdös–Rényi model as our null hypothesis. The null hypothesis we set is such that the clustering coefficient of our network is not significantly different to that of a random model, given a 0.95 of confidence ($\alpha = 0.05$). Consequently, the hypothesis we want to verify is that the clustering coefficient of our network, and hence, its topology, is due to its specific nature and not due to a random behaviour. In order to reject or accept the null hypothesis, we compute the p-value by dividing the number of times the random model has a higher value than our network, divided by the total number of times we carry out the experiment. For this study, we build 30 random graph as experimental set and compute the clustering coefficient of each one.

Moreover, given the size of the graph, and in order to speed up the computation, we carried out an optimization to calculate the metrics exactly but without scanning all the graph, bounding the values of the clustering coefficient in the null hypothesis, C_{NH}, below C_{NH}^{min} and above C_{NH}^{max}, exploring only a subset of M nodes of the network. The value of the lower bound C_{NH}^{min} is calculated as $\frac{1}{N} \cdot \sum_{i=1}^{M} C_i$, and the value of the upper bound C_{NH}^{max} as $\frac{1}{N} \cdot \sum_{i=1}^{M} C_i + 1 - \frac{M}{N}$.

Then, with the previous formulae, after calculating the clustering coefficient C_i for only the first M nodes in the network produced by the null hypothesis, we can compare it with the bounds, and assume that if $C_{NH}^{min} \geq$ C, then $C_{NH} \geq$ C, also if $C_{NH}^{max} <$ C then $C_{NH} <$ C.

Furthermore, by allowing for a certain degree of error in evaluating C (the clustering coefficient in our network) and C_{NH}, we can further optimize the calculation by means of a Monte Carlo procedure. To do so, we order the graph by doing a uniform permutation of the vertices, and then calculate the clustering coefficient only for the first M vertices. The value of M we use is based on the fact that even when $M << N$, we can get a good estimation of the clustering coefficient [18], such as $100M/N = 10\%$, and with that premise we can solve as $M = \lceil \frac{0.1*N}{100} \rceil$.

Network	Erdös–Rényi (random model)
Target Network	0.03
Non-target Network	0

Table 3: Statistical marginal significance (p-value) after the Monte Carlo procedure on the clustering coefficient. The selected confidence value is 0.95 ($\alpha = 0.05$).

As shown in Table 3, the p-value with respect to the random network null hypothesis is lower than the significance level $\alpha = 0.05$, leading to significant evidence for the rejection of the null hypothesis. This means that, as our network clustering coefficient is a particular characteristic of it, our network topology can be seen as having specific, non-random, characteristics.

3.2 Community analysis

Since we are dealing with a graph with different node categories, it is worth investigating if a community detection algorithm is able to detect these underlying clusters. In order to build the communities, we used the previously described Infomap [11] and Walktrap [10] algorithms, taken from the R package *igraph*. Notice that the genes that are not known to be targets or non-targets have been excluded from the community analysis, as the main idea in this part of the analysis is to verify if target (or non-target) genes are similar enough, and, due to the interactions between them and their local topology in the network, they can be grouped in pure communities or clusters. Here, we define a pure community or cluster as a group formed of only one type of genes, either targets or non-targets. Hence, an impure community would be formed of a mixture of targets and non-targets.

After running the community finding algorithms, we measured the goodness of the communities found by the two different chosen algorithms. For this, we rely on different quality measures: Triangle Partition Ratio (TPR), expansion, conductance and modularity. TPR is the fraction of nodes within a cluster that belongs to a triad; thus a higher value translates into a clustering with a higher quality. Expansion is the number of nodes leaving the cluster; thus, a lower value means a clustering of better quality. Conductance is the fraction of total edge volume that points outside the cluster; thus, a lower value is better. Modularity is the difference between the number of edges in the cluster and the expected number of edges of a random graph with the same degree distribution; thus, a higher value is better.

Algorithm	TPR	Expansion	Conductance	Modularity	Communities
Walktrap	**0.763**	**1.216**	**0.188**	**0.6**	318
Infomap	0.68	6.906	0.394	0.56	370

Table 4: Quality measures of community structure.

The quality metrics for each of the algorithms are summarized in Table 4. From these results, it is clear that the Walktrap algorithm yields the best values for TPR, expansion, modularity and conductance. Consequently, by relying exclusively on the values of these heuristic quality metrics to judge the goodness of the community finding algorithm performance, we can say that the Walktrap algorithm produces the best segmentation of the graph into different communities. We then proceeded to analyze in more detail such communities.

In Figure 2, we show the percentage of genes per category (target and non-target) of each community given by the result of the Walktrap community detection algorithm. From this result, we can say that there is a densely connected cluster of target genes and a densely connected cluster of non-target genes. The result given by the community finding algorithm is very interesting as, since it is able to find pure communities of targets, it means that the target genes

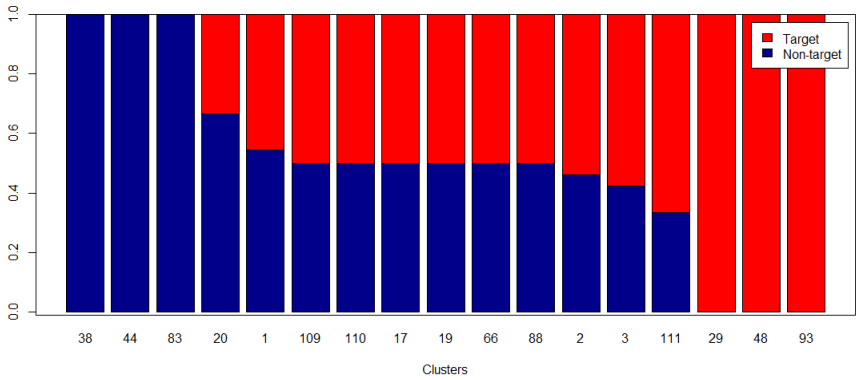

Fig. 2: Communities composition based on the Walktrap algorithm.

have a particularly characteristic topology in the network, thus leading to the conclusion that it is possible to unveil new targets by means of their structure on the network (e.g., by taking into account their graph-theoretical centrality). Then, unknown genes that have characteristics similar to the target genes may be potential targets.

3.3 Analysis of the network based on centrality measures

After ensuring that the topology of the two types of genes in our network is characteristic, we studied the network using the graph-theoretical centrality measures described in section 2. Their values for the full network and the target and non-target sub-networks are summarized in Table 5.

Network	Degree	Betweenness	Closeness
Target	0.0022	0.00021	0.29
Non-target	0.0021	0.00023	0.28
Full	0.0020	0.0002	0.29

Table 5: Average values of the centrality measures in the three networks

Then, given the previous insights, we analyzed the genes with the highest values for each of the graph-theoretical based centralities, as those are the ones showing clearer predominant centrality values in the whole network. After that, we took the genes that were shortlisted in the three sets. Thus we took the genes that are among the top with respect to their centrality values in all the centrality measures at the same time.

In Figure 3 the set of genes that have the highest values in all the centralities is illustrated. As can be seen, this set is comprised of only target and unknown genes at the same time, pointing out that these unknown genes may be potential candidates as drug targets for AML and ALL. These genes are HSP90AA1,

TRIM25, ELAVL1, APP, MCM2, CUL3, HSPA8, XPO1, EGLN3, UBC and NXF1.

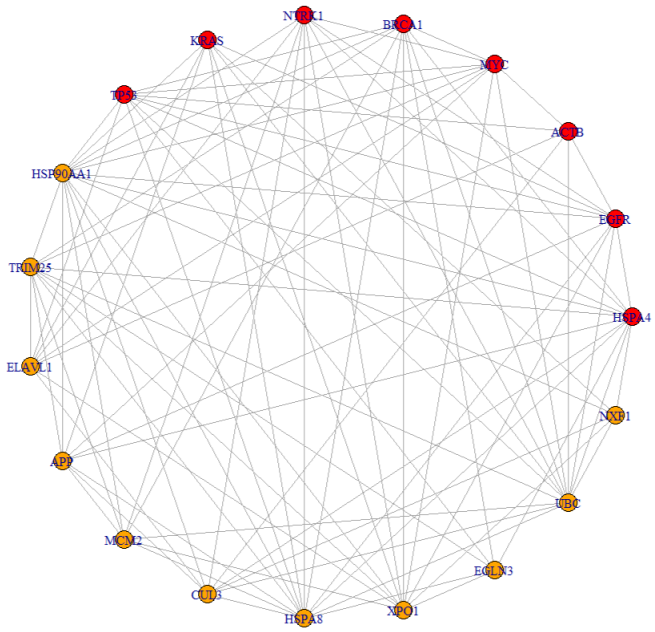

Fig. 3: Set of top shortlisted genes with the highest values in all centrality measures: degree, closeness and betweenness.

4 Conclusions

This brief paper has reported network analysis and community cluster detection for the identification of potential drug targets for ALL and AML types of Leukemia. We have described in detail how we built a gene-interaction network for the genes targeted by currently FDA approved drugs for ALL and AML, and for non-target genes obtained from HumanMine, a publicly available biological database.

The non-randomness (and therefore the topological specificity) of the network has been asserted. Furthermore, by using the Walktrap and Infomap community detection algorithms, it has been shown that both target and non-target genes can be segmented in pure groups. Finally, by analyzing several graph-theoretical centrality measures, and taking the genes that hold the highest values of these measures in the network, we were able to identify a set of potential drug target candidates for ALL and AML.

This study could be extended to the complete Leukemia spectrum, including Chronic Lymphocytic Leukemia, chronic Myelogenous Leukemia, Hairy Cell Leukemia, Mast Cell Leukemia and Meningeal Leukemia. In addition, it would be interesting to analyze in more detail the specificities of those genes that were found to be potential targets, by carrying out a gene set enrichment analysis to check if, statistically, they have significant pathways affected and their relation to ALL and AML Leukemias.

References

1. Pui, C.H., Evans, W.E.: Treatment of acute lymphoblastic leukemia. New Engl J Med 354(2), 166–178 (2006)
2. Arakawa, H., et al.: Identification and characterization of the arp1 gene, a target for the human acute leukemia all1 gene. PNAS 95(8), 4573–4578 (1998)
3. Pui, C.H.: Acute lymphoblastic leukemia. Springer (2011)
4. Lowenberg, B., Downing, J.R., Burnett, A.: Acute myeloid leukemia. New Engl J Med 341(14), 1051–1062 (1999)
5. Fortunato, S.: Community detection in graphs. Physics Reports 486(3-5), 75–174 (2010)
6. Franceschet, M.: PageRank. Commun ACM 54(6), 92 (jun 2011)
7. Bonacich, P., Lloyd, P.: Eigenvector-like measures of centrality for asymmetric relations. Soc Networks 23(3), 191 – 201 (2001)
8. Erdős, P., Rényi, A.: On the evolution of random graphs. In: Publication of the Mathematical Institute of the Hungarian Academy of Sciences. pp. 17–61 (1960)
9. Yates, P.D., Mukhopadhyay, N.D.: An inferential framework for biological network hypothesis tests. BMC Bioinformatics 14(1), 94 (mar 2013)
10. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: Computer and Information Sciences - ISCIS 2005, pp. 284–293. Springer Berlin Heidelberg (2005)
11. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. PNAS 105(4), 1118–1123 (jan 2008)
12. Rissanen, J.: Modeling by shortest data description. Automatica 14(5), 465 – 471 (1978)
13. Institute, N.C.: Drugs approved for leukemia (2019), https://www.cancer.gov/about-cancer/treatment/drugs/leukemia
14. McDonnell Genome Institute, W.U.S.o.M.: The drug-gene interaction database (2019), http://www.dgidb.org/
15. Smith, R.N., et al.: InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics 28(23), 3163–3165 (2012)
16. Kalderimis, A., et al.: InterMine: extensive web services for modern biology. Nucleic Acids Res 42(W1), W468–W472 (apr 2014)
17. Chatr-aryamontri, A., et al.: The biogrid interaction database: 2017 update. Nucleic Acids Research 45(D1), D369–D379 (2017)
18. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. nature 393(6684), 440 (1998)