Master Degree: Data Science
a.y. 2022/2023

12 January 2023

BIOINFORMATICS AND NETWORK MEDICINE

Putative disease gene identification and drug repurposing for Acute Myelocytic Leukemia

Cruoglio Antonella, Mascolo Davide, Napoli Mario

GROUP 03

ABSTRACT

In this work, we explore different methodologies to study protein-protein interactions and its' role in the disease gene association. The aim is to find the most useful algorithm to extract correct genes to trace disease existence. The disease of interest is the Acute Myelocytic Leukemia, a malignant disorder of hematopoietic stem and progenitor cells, characterized by accumulation of immature blasts in the bone marrow and peripheral blood of affected patients. Different algorithms were considered: DIAMOnD (a disease module detection), DiaBLE and heat diffusion. The different methods were compared through 5-fold Cross Validation, using some evaluation metrics such as Precision, Recall, F1 score. As a final result, we found out that the most powerful algorithm is DiaBLE with further enrichment analysis. Also clustering algorithms (MCL and Louvain) were applied to the network in order to identify disease modules, that have been validated via enriched analysis. The second task of this work refers to drug repurposing, in order to identify existing drugs that can be used for this disease, as they are associated with the putative genes found. All the analysis were performed using Python language and Cytoscape software.

INTRODUCTION

Acute myeloid leukemia (AML) is a heterogeneous disease characterized by the accumulation of immature myeloid progenitor cells in the bone marrow, compromising of normal blood cell production and ultimately resulting in bone marrow failure. Response to chemotherapy treatment in patients with AML is wide-ranging, and there are no adequate biomarkers to predict their clinical outcome. Standard induction therapy leads to complete remission in approximately 50% to 75% of patients, depending on prognostic factors, such as age or the presence of certain gene or chromosomal changes. Despite favorable primary response rates, only approximately 20% to 30% of the patients enjoy long-term disease survival. Protein-protein interaction (PPI) networks seek to characterize the flow of information within the cell and the organism in order to understand the functional relevance of expressed proteins. Analysis of PPI networks can help understand mechanisms involved in diseased states, and orient research strategies into biomarkers or therapeutic targets [1].

MATERIALS AND METHODS

PPI and GDA data gathering and interactome reconstruction

Protein-protein interaction networks are mathematical representations of the physical contacts between proteins in the cell. In this work, Human Protein-Protein Interaction network was built using the database BioGRID (The Biological General Repository for Interaction Datasets). From the original downloaded dataset, all non-human interactions were filtered out and only "physical" interactions were kept. Moreover, the redundant self-loops have been eliminated from the network. We obtained a network with 19724 nodes and 766017 edges. From this graph we isolated the largest connected component, composed by 19724 nodes. In order to find the seed genes for the Acute Myeloid Leukemia, we used DisGeNET, one of the largest and comprehensive repositories of human gene-disease associations (GDAs) currently available. 2709 of these genes were present in the human interactome. On the disease LCC we computed some centrality metrics: Node degree, Betweenness centrality, Eigenvector centrality, Closeness centrality, ratio Betweenness/Node degree. In Table 2. is possible to observe the first 50 disease genes in the disease LCC ordered for ratio Betweenness/Node degree from higher to lower.

Putative disease genes identification algorithms

In this section we used different algorithms in order to identify putative disease genes. Disease genes are not randomly scattered within these protein-protein interaction networks, but agglomerate in specific regions, suggesting the existence of specific disease modules for each disease. The identification of these modules is the first step towards elucidating the biological mechanisms of a disease or for a targeted search of drug targets. The algorithm **DIAMOnD** seeks to find genes connected to a set of seed genes based on the significance of that connectivity. It is a well-established method that provides network context to our set of putative leukemia genes and allows us to expand the set in a reliable manner. **DiaBLE** algorithm is a modified version of DIAMOnD: instead of taking the whole interactome as the background model, DiaBLE considers as gene universe the smallest local expansion of the current seeds set at each iteration step. The third algorithm we used is **Heat Diffusion**, based on the concept of network propagation. Propagation provides a robust estimate of the distance between groups of nodes. Network propagation uses the network of interactions to find new genes that are most relevant to a well-understood set of genes. In our case, if we know that mutations in a particular protein cause a disease, we can also hypothesize that mutations in related proteins may also cause that disease. Thus we can find related proteins by examining the network of interactions. In order to validate the computational performances, we used a 5-fold cross validation. We split the disease genes set into 5 subsets; each time, one subset is used as probe set and the remaining four subsets as training set.

Optional Task - Putative disease genes identification

Cellular components associated with a specific disease phenotype show a tendency to cluster in the same network neighbourhood. The identification of these neighborhoods, or disease modules, is therefore a prerequisite of a detailed investigation of a particular pathophenotype. All cellular components that belong to the same topological, functional or disease module have a high likelihood of being involved in the same disease. These methods start with identifying the disease modules and inspecting their members as potential disease genes. In this section **Markov Clustering** and **Louvain** algorithm are applied. MCL Markov Clustering simulates flow diffusion in a graph. The idea is that random walks between two nodes that belong to the same group are more frequent than between two nodes belonging to different groups. The Louvain algorithm is a heuristic method based on modularity optimization. The algorithm works in 2 steps. On the first step it assigns every node to be in its own community and then for each node it tries to find the maximum positive modularity gain by moving each node to all of its neighbor communities. If no positive gain is achieved the node remains in its original community.

Best algorithm choice and putative disease gene identification

Using the best performing algorithm, we tried to predict new putative disease genes using all known GDAs as seed genes. Over the first 200 genes obtained the **Enrichment Analysis** was performed, using Enrichr. Enrichment Analysis is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched groups of genes. In particular, we can identify GO categories and/or biological pathways that are enriched in a gene list more than would be expected by chance. The first technique used, **Gene Ontology (GO)**, introduced the concept of associating a collection of genes with a functional biological term in a systematic way. Given a genes list, through the GO technique we can observe which bin has been enriched for the input gene list. Another used technique is called **Pathway Analysis**, it consists in the identification or the construction of a pathway starting from a set of proteins. More in detail a pathway is a process involved in a cell or in a tissue. Watching proteins' interactions, the pathways analysis let to identify the diseases most statistically related to the triggered proteins chains.

Drug repurposing

Drug repurposing is a strategy for identifying new uses for approved or investigational drugs that are outside the scope of the original medical indication. After selecting the first 20 putative disease genes using the DiaBLE algorithm, we associated such 20 genes to drugs using DGIdb. Starting from the drugs associated with the most of the above 20 genes down we compiled a ranking and used this to check on https://clinicaltrials.gov if there are clinical trials using each of the drug for the Acute Myelocitic Leukemia. The drugs most associated were:

- BLEOMYCIN, mostly used to treat cancer. This includes testicular cancer, ovarian cancer, and Hodgkin's disease, and less commonly non-Hodgkin's disease;

- TRETINOIN is a medication used for the treatment of acne and acute promyelocytic leukemia, a subtype of AML.
- THALIDOMIDE, used as a first-line treatment in multiple myeloma in combination with dexamethasone or with melphalan and prednisone to treat acute episodes of erythema nodosum leprosum and for maintenance therapy.
- CISPLATIN, a chemotherapy medication used to treat a number of cancers. These include testicular cancer, ovarian cancer, cervical cancer, breast cancer, bladder cancer, head and neck cancer, esophageal cancer, lung cancer, mesothelioma, brain tumors and neuroblastoma.

For these, we found 57 studies for TRETINOIN and 1 study for THALIDOMIDE; zero for the other two mentioned.

RESULTS AND DISCUSSION

Table 1: Summary of GDAs and basic network data

Disease Name	UMLS disease ID	MeSH disease class	Number of associated genes	Number of genes present in the interactome	LCC size of the disease interactome
Leukemia, Myelocytic, Acute	C0023467	C04	3111	2709	2570

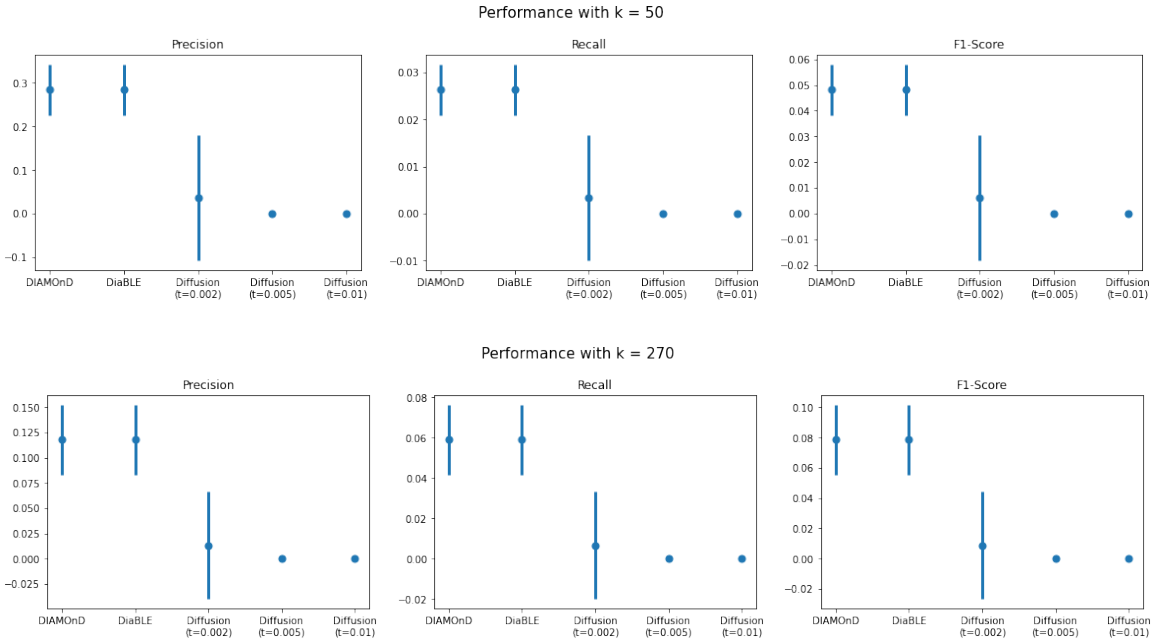
Table 2: Main network metrics of disease LCC genes

Ranking	Gene Name	Degree	Betweenness	Eigenvector	Closeness	Ratio Betweennes:Degree
1	CALCA	2	0.000779	0.000042	0.265585	0.000389
2	GHRHR	2	0.000779	0.000195	0.293098	0.000389
3	IL17A	3	0.000783	0.000350	0.312416	0.000261
4	IL23R	3	0.000779	0.000978	0.329697	0.000260
5	KIR2DS2	4	0.000797	0.000238	0.303628	0.000199
6	FABP4	4	0.000791	0.001313	0.333593	0.000198
7	CORT	4	0.000781	0.000114	0.285921	0.000195
8	KCNJ5	4	0.000780	0.000414	0.312189	0.000195
9	KITLG	4	0.000780	0.000557	0.319012	0.000195
10	LTA	5	0.000820	0.000961	0.330801	0.000164
11	CCL19	5	0.000785	0.000580	0.314790	0.000157
12	CD200R1	5	0.000783	0.000117	0.296104	0.000157
13	CXCL8	6	0.000880	0.001651	0.346881	0.000147
14	APP	388	0.056303	0.069970	0.500682	0.000145
15	DDX58	311	0.044013	0.054742	0.479112	0.000142
16	SETBP1	6	0.000829	0.001800	0.352691	0.000138
17	CD96	6	0.000792	0.000989	0.334549	0.000132
18	CD48	6	0.000790	0.001323	0.338739	0.000132
19	GGT1	6	0.000785	0.000450	0.312645	0.000131
20	DCAF6	12	0.001569	0.004116	0.361475	0.000131
21	EFNA5	16	0.001932	0.002246	0.357004	0.000121
22	PDCD1LG2	7	0.000797	0.000449	0.318814	0.000114
23	SFRP2	7	0.000786	0.001323	0.334680	0.000112
24	TRIM34	7	0.000785	0.001023	0.330588	0.000112
25	GNMT	7	0.000780	0.001975	0.332342	0.000111
26	PLXNA2	11	0.001192	0.001339	0.340175	0.000108
27	HNRNPH1	303	0.031675	0.073259	0.485541	0.000105
28	NEAT1	8	0.000813	0.003117	0.360460	0.000102

29	KCNIP3	8	0.000805	0.001425	0.337228	0.000101
30	ISCA1	8	0.000797	0.001210	0.335773	0.000100
31	GYPB	8	0.000789	0.000969	0.328349	0.000099
32	NMU	18	0.001667	0.001100	0.349239	0.000093
33	NRM	11	0.000978	0.000401	0.316068	0.000089
34	TP53	481	0.040349	0.136334	0.518048	0.000084
35	ERMP1	22	0.001823	0.004522	0.380988	0.000083
36	CKMT2	10	0.000820	0.001535	0.327595	0.000082
37	ALDH1A1	10	0.000812	0.002984	0.349714	0.000081
38	PLAU	10	0.000811	0.002351	0.358049	0.000081
39	MYC	547	0.043530	0.156813	0.531664	0.000080
40	RARG	10	0.000791	0.001755	0.330886	0.000079
41	WLS	16	0.001263	0.003563	0.371511	0.000079
42	APEX1	247	0.019147	0.066002	0.480546	0.000078
43	SFRP1	4	0.000303	0.002481	0.351052	0.000076
44	MPC2	9	0.000676	0.001805	0.349809	0.000075
45	EGFR	342	0.024981	0.089071	0.499805	0.000073
46	PTPRD	30	0.002142	0.004803	0.380762	0.000071
47	MUC1	46	0.003270	0.014832	0.425754	0.000071
48	DPH1	13	0.000913	0.003981	0.368157	0.000070
49	HSP90AA1	343	0.023387	0.101778	0.498254	0.000068
50	WIPI2	12	0.000806	0.003069	0.356459	0.000067

Performance Comparison

The different algorithms were compared using 5-Fold Cross Validation on the basis of Precision, Recall and F1-Score. The figure below shows the error bars for each method, for different number of selected genes. We can notice that the performances of DIAMOnD and DiaBLE are better than the ones of the Diffusion Algorithm for all cases. According to the paper [3], the DiaBLE algorithm provides more biological meaningful results compared to DIAMOnD. For this reason, even if the two methods have quite the same performance, we decide to use DiaBLE as the best algorithm in order to extract new putative genes.



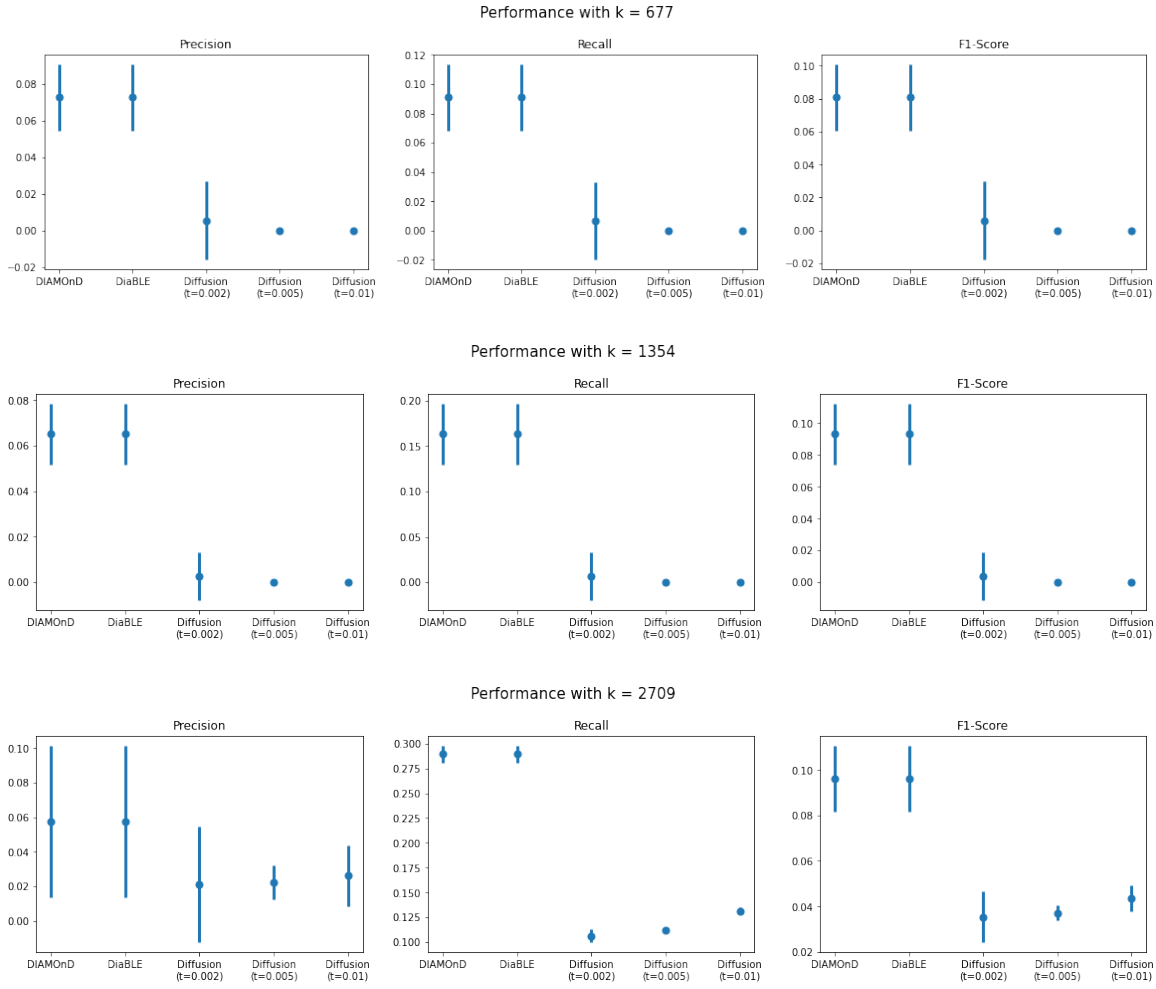


Figure 2: Performance Comparison between different algorithms for different numbers of selected genes

Enrichment Analysis

In the figures below they are shown the results of the Enrichment Analysis for the first 200 putative genes found through DiaBLE. The bar plots represent the level of significance using the p-value: the closer to zero, the more significant the particular GO term associated with the group of genes is or in the case of the pathway analysis the more probably the disease is. GO analysis suggested putative genes are significantly enriched in the mRNA splicing. The result showed putative genes were enriched in pathways in spliceosome, ubiquitin mediated proteolysis, DNA replication. In addition, we can observe the presence of some diseases like amyotrophic lateral sclerosis, chronic myeloid leukemia and coronavirus disease. By consulting existing literature, we found out that spliceosome mutations are encountered in about 50% of secondary acute myeloid leukemia cases [4]. As far as ubiquitination is concerned, it is one of the post-translational modifications and the ubiquitin-like proteins play a critical role in various cellular processes, including autophagy, cell-cycle control, DNA repair, signal transduction, and transcription. Also, the importance of Ubls in AML is increasing, with the growing research defining the effect of Ubls in AML. Numerous studies have actively reported that AML-related mutated proteins are linked to Ub and Ubls. The current review discusses the roles of proteins associated with protein ubiquitination, modifications by Ubls in AML, and substrates that can be applied for therapeutic targets in AML [5].

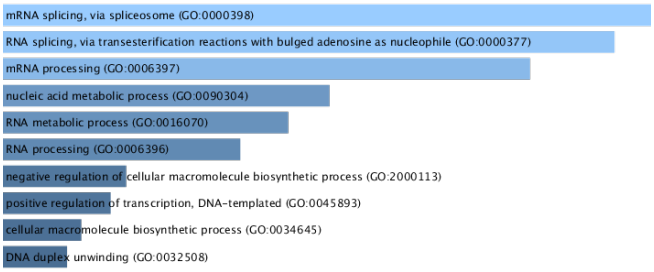


Figure 3: GO Biological Proces

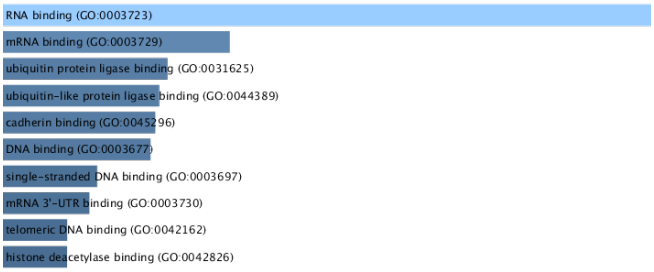


Figure 4: GO Molecular Function

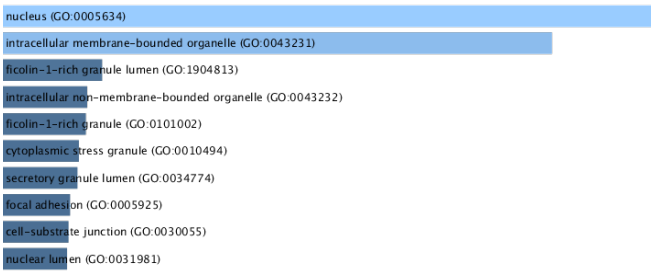


Figure 5: GO Cellular Component

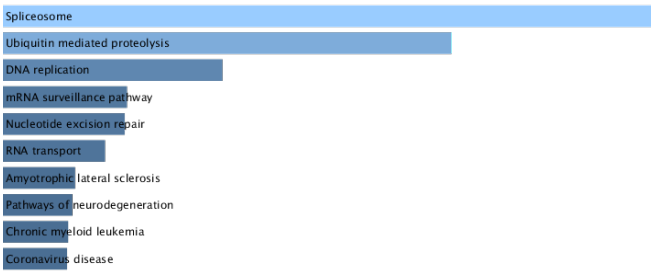


Figure 6: KEGG Pathways

Putative disease genes identification with MCL and Louvain

MCL and Louvain algorithms were performed in order to identify putative disease modules. We tried to find the best inflation parameter for Markov Cluster algorithm, doing a grid search and, using an inflation parameter equal to 1.8, we obtained 1749 communities, with a modularity of 0.85. On the other hand, the Louvain algorithm found only 8 groups. We considered only the communities with a number of nodes $s : 10 \leq s \leq 1000$. After filtering these communities, we performed the **hypergeometric test** to find putative disease modules. For the first algorithm, 11 communities were selected as putative disease module, as the corresponding p-value from the hypergeometric test was below 0.05. From the Louvain communties, only one was selected as putative disease module. On this one was performed the enrichment analysis, excluding the already known disease genes. Given the number of modules found through MCL, we don't show here the enrichment analysis results for this algorithm.

Table 3: Communities tagged as putative disease module for MCL

Size	Known disease genes
69	18
120	30
10	4
29	12
54	15
21	7
32	11
52	14
12	6
20	9
37	17

Table 4: Communities tagged as
putative disease module for Louvain

Size	Known disease genes
38	15

Enrichment Analysis of Louvain Putative Disease Module:

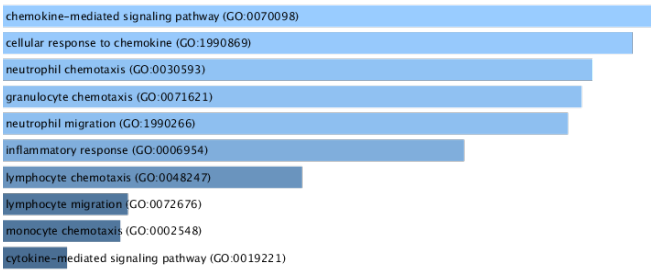


Figure 7: GO Biological Process

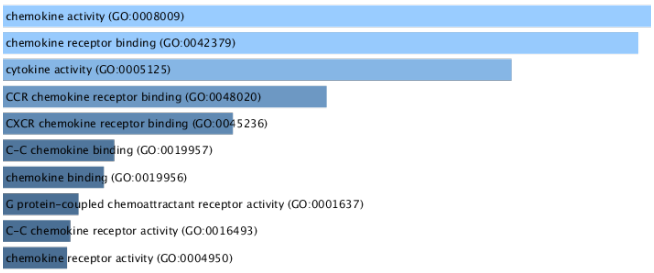


Figure 8: GO Molecular Function

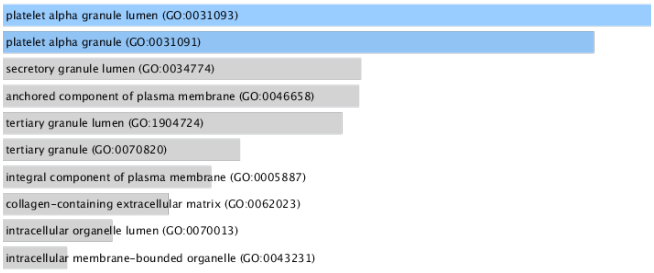


Figure 9: GO Cellular Component

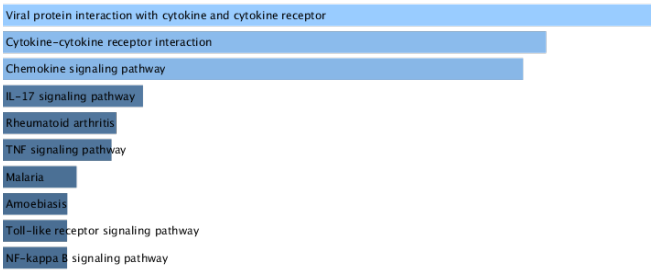


Figure 10: KEGG Pathways

GO analysis suggested putative genes are significantly enriched in the chemokine-mediated and cytokine-mediated signaling pathway. Abnormalities in the cytokine signaling pathways are characteristic of all forms of leukemia. In normal hematopoietic cells, cytokines stimulate proliferation, differentiation, self-renewal, survival, and functional activation. In leukemic cells, however, these pathways are exploited to serve critical parts of the malignancy program. Research results suggest that various cytokines released by the leukemic cells regulate the differentiation and growth of the malignant cells [10].

AUTHOR CONTRIBUTIONS

Here is a description of the contribution to the project of each author. D.M.: data gathering; A.C., D.M., M.N.: algorithm implementation; A.C., D.M., M.N.: optional task; A.C., D.M.: cross-validation; A.C., D.M., M.N.: writing-original draft preparation, A.C., M.N.: writing-review editing,

REFERENCES

[1] Rendon-Rodriguez, Juan Jose et al. “Interaction network of proteins associated with unfavorable prognosis in acute myeloid leukemia.” Acta biochimica Polonica vol. 67,4 (2020): 475-483. doi:10.18388/abp.2020$_5$094

[2] Ghiassian, Susan Dina, Jörg Menche, and Albert-László Barabási. ”A DIseASe MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome.” PLoS computational biology 11.4 (2015): e1004120.

[3] M. Petti, D. Bizzarri, A. Verrienti, R. Falcone and L. Farina, ”Connectivity Significance for Disease Gene Prioritization in an Expanding Universe,” in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 6, pp. 2155-2161, 1 Nov.-Dec. 2020, doi: 10.1109/TCBB.2019.2938512.

[4] Lachowiez, Curtis A et al. “Impact of splicing mutations in acute myeloid leukemia treated with hypomethylating agents combined with venetoclax.” Blood advances vol. 5,8 (2021): 2173-2183. doi:10.1182/bloodadvances.2020004173

[5] Park SS, Baek KH. Acute Myeloid Leukemia-Related Proteins Modified by Ubiquitin and Ubiquitin-like Proteins. Int J Mol Sci. 2022 Jan 3;23(1):514. doi: 10.3390/ijms23010514. PMID: 35008940; PMCID: PMC8745615.

[6] Wikipedia contributors. ”Thalidomide.” Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 28 Dec. 2022. Web. 12 Jan. 2023.

[7] Brandt JP, Gerriets V. Bleomycin. [Updated 2022 Aug 29]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK555895/

[8] ”Cisplatin”. The American Society of Health-System Pharmacists. Archived from the original on 21 December 2016. Retrieved 8 December 2016.

[9] ”Tretinoin”. The American Society of Health-System Pharmacists. Archived from the original on 30 November 2016. Retrieved 8 December 2016.

[10] Omid Karimdadi Sariani, Sara Eghbalpour, Elahe Kazemi, Kimia Rafiei Buzhani, Farhad Zaker, Pathogenic and therapeutic roles of cytokines in acute myeloid leukemia, https://doi.org/10.1016/j.cyto.2021.15550