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

12 January 2023

BIOINFORMATICS AND NETWORK MEDICINE

Putative disease gene identification and drug repurposing for Stomach Neoplasm

Silvia del Piano, Svenja Jedhoff, Marie Picquet

GROUP 10

ABSTRACT

Stomach neoplasms are nowadays, together with other cancers, still one of the major causes of death worldwide. Due to this reason, the objective of this study is to identify genes that may be connected to this family of cancers, to improve our understanding of this disease and its treatment. This work utilises network medicine techniques on data gathered from the biological databases BioGRID and DisGeNET to build a protein-protein-interaction network with genes which are known to be associated with stomach neoplasm. Afterwards, to predict putative disease genes, 5-fold cross-validation is used for a performance evaluation of three algorithms: DIAMOnD, DiaBLE, and heat-diffusion. The putative disease genes identified show a higher degree, betweenness, and closeness centrality than average, strongly suggesting a central role in the network. Using the databases of GO and KEGG, an enrichment analysis is executed with the web-platform EnrichR on the first 200 putative disease genes provided by the best performing algorithm. However, its results suggest that the identified putative disease genes are not involved in the disease. Finally, the drug repurposing possibility is investigated, using the first 20 putative disease genes. The drugs Selumetinib, Tanespimy and Trametinib could be identified. According to ClinicalTrials.gov, Selumetinib has not been involved in studies regarding stomach neoplasms yet. Nonetheless, it was found that the drug Tanespimycin has been in clinical trials but no results have been published and that the drug Trametinib is involved in two ongoing clinical trials for the disease of stomach neoplasm.

INTRODUCTION

Stomach neoplasm [1.a] (or stomach cancer) is a family of diseases that includes all types of cancer that can start in the stomach. This condition affects especially people that are 60 years old or older and of non-Caucasian ethnicity [1.b]. It's more diffused in East Asia, Eastern Europe, and South and Central America. Risk factors comprehend the typical ones for all cancers, like smoking or obesity, and specific ones for this disease, like the infection of Helicobacter pylori. Even if a lot of progress has been made in treating cancer, it's still one of the major causes of premature death.

The objective of this work is to provide a list of genes that could be connected to this illness, so that it may be used for further biological and medical research to improve our understanding of these cancers and hopefully better their treatment. In this purpose, we also searched for drugs that affect these identified genes and state-of-the-art studies on them.

To do this we used network medicine techniques on data gathered from reliable biological databases. Network medicine [2] is a relatively new field that exploits network theory and biology. Networks are often used to represent complex interactions, and studies show that a disease is rarely the consequence of an abnormality of a single gene. On the contrary, several factors that include interactions among genes, proteins, molecules and the environment contribute in different ways to the pathophenotype. In our study we first used the PPI (protein-protein interaction) network in BioGRID to build the human interactome. This is a graph where proteins are the nodes and an edge corresponds to an interaction between two proteins. We also used gene-disease association data from DisGeNET and checked for disease genes in the interactome. Afterwards, we performed the analysis of basic

metrics on this final network and then ran DIAMOnD, DiaBLE, and Cytoscape's diffusion algorithm to predict new putative disease genes for stomach cancer. After using 5-fold cross-validation to evaluate which algorithm performed best, we did the enrichment analysis on the list of the first 200 putative genes provided by said algorithm. Finally, after obtaining a new ranking of these genes, we selected the first 20. Then we searched for drugs that are associated with these 20 genes using DGIdb (Drug Gene Interaction Database) and selected the first 3 drugs. To conclude our work, we then looked on ClinicalTrials.gov database to see in which studies these drugs were involved to verify if there was a possibility for drug repurposing.

MATERIALS AND METHODS

As previously stated, the PPI network was downloaded from the BioGRID database [3] (Version 4.4.215, downloaded: 30.11.2022). Since only human-human and physical interactions are considered, the data was filtered with the help of Cytoscape[4] (Version 3.9.1), a software specifically designed to work with biological networks. The largest connected component (LCC) was extracted with the help of the App Cyfinder [5] (Version 2.1.1) in Cytoscape, such that the used interactome contains 19711 unique nodes and 799771 edges without self-loops. The Gene-Disease associations (GDA) for the stomach neoplasm were derived from DisGeNet [6] (downloaded: 30.11.2022) and they contain 820 genes, 784 of which are present in the interactome (table 1). The largest connected component of the disease genes in the interactome results having 713 nodes and 6832 edges.
In table 2 some significant network metrics computed for the 50 nodes with the highest ratio of betweenness to node degree of the LCC are shown. In addition to the name of the gene and the node's degree, its betweenness centrality, eigenvector centrality, and closeness centrality are also reported. The three of them are all path-based measures. Betweenness centrality indicates how many shortest paths of any two nodes pass through that node, eigenvector centrality is computed considering the quantity and quality of the connections to the node, and closeness centrality is the average of all the shortest paths from the node to every other node.

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
Stomach Neoplasm	C0038356	D013274	820	784	713

Table 2 Main network metrics of disease LCC genes

Ranking	Gene name	Degree	Betweenness	Eigenvector Centrality	Closeness Centrality	ratio Betw./Degree
1	KISS1	2	0.002809	0.002527	0.338725	0.0014
2	HTRA1	3	0.002809	0.000632	0.294946	9e-04
3	IL6R	5	0.002853	0.003915	0.351431	6e-04
4	RECK	2	0.001126	0.001792	0.331009	6e-04
5	AGT	5	0.002727	0.008501	0.384449	5e-04
6	TNFSF9	6	0.003112	0.001274	0.317715	5e-04
7	SOSTDC1	6	0.002958	0.003578	0.340507	5e-04
8	UGT1A10	6	0.002878	0.003682	0.341979	5e-04
9	HNRNPL	98	0.044457	0.088266	0.493416	5e-04
10	HSPA5	111	0.047527	0.124551	0.511127	4e-04
11	SFRP1	3	0.001248	0.004227	0.357609	4e-04
12	MYC	209	0.083532	0.2051	0.553225	4e-04
13	EGFR	123	0.048233	0.113294	0.513338	4e-04
14	ELAVL1	120	0.045753	0.103992	0.500000	4e-04
15	APOE	20	0.007418	0.024768	0.411561	4e-04
16	TRIM59	9	0.003185	0.01194	0.381974	4e-04
17	MGAT5	5	0.001675	0.003928	0.349877	3e-04
18	TP53	186	0.06215	0.192999	0.546431	3e-04
19	ARL6IP5	15	0.004827	0.01408	0.392719	3e-04
20	ERBB2	75	0.023875	0.06592	0.482712	3e-04
21	SRC	65	0.019672	0.063995	0.471836	3e-04
22	SOX2	111	0.032527	0.095319	0.487671	3e-04
23	PTCH1	24	0.007026	0.020411	0.408023	3e-04

24	ESR1	148	0.042651	0.159422	0.523915	3e-04
25	ECM1	10	0.002661	0.011703	0.390351	3e-04
26	DES	11	0.002881	0.013832	0.380139	3e-04
27	KRAS	79	0.020657	0.067764	0.468421	3e-04
28	CAV1	76	0.01977	0.058859	0.466274	3e-04
29	APEX1	77	0.02002	0.077702	0.470899	3e-04
30	FGFR2	14	0.003582	0.007828	0.369103	3e-04
31	CCNA2	24	0.006133	0.028064	0.403399	3e-04
32	MUC1	17	0.004324	0.022805	0.414918	3e-04
33	DDR1	12	0.003036	0.010419	0.378925	3e-04
34	ZIC1	7	0.001766	0.003291	0.338242	3e-04
35	ADRB2	30	0.00754	0.028625	0.41907	3e-04
36	FBLN1	9	0.00221	0.006191	0.364567	2e-04
37	ECHS1	12	0.002925	0.010801	0.368721	2e-04
38	LGALS3	35	0.008453	0.030554	0.425837	2e-04
39	HSPA8	118	0.028462	0.144137	0.511862	2e-04
40	WIF1	7	0.001663	0.006965	0.358329	2e-04
41	CD81	30	0.007112	0.032003	0.414677	2e-04
42	ASPH	22	0.005181	0.026646	0.429692	2e-04
43	FYN	40	0.009378	0.033141	0.435741	2e-04
44	PDGFRA	46	0.010667	0.037869	0.438424	2e-04
45	CTSL	17	0.00393	0.017835	0.405698	2e-04
46	EP300	124	0.028552	0.140789	0.511862	2e-04
47	NEIL1	13	0.002946	0.012225	0.364567	2e-04
48	PRKCA	35	0.007756	0.04034	0.434942	2e-04
49	RELA	79	0.017469	0.097687	0.476892	2e-04
50	FN1	94	0.020759	0.102585	0.485014	2e-04

After having computed all these metrics, we proceeded with the identification of putative disease genes. To do so, three algorithms were considered: DIAMOnD [7], DiaBLE [8], and the diffusion-based algorithm implemented in Cytoscape [9]. DIAMOnD is based on the idea of measuring the connectivity significance of each gene with an associated p-value, which is computed for a gene in a network of N nodes with s_0 seed genes as follows:

$$p\text{-value}(k, k_s) = \sum_{k_i = k_s}^{k} p(k, k_i), \quad \text{with} \quad p(k, k_s) = \frac{\frac{s_0 N - s_0}{k_s k - k_s}}{\frac{N}{k}},$$

where $p(k, k_s)$ is the probability that a gene with k connections has k_s connections to seed genes. This measure evaluates if the gene has more connections to seed genes than expected. Given a network, DIAMOnD computes the p-value for all the genes connected to the set of disease genes, and the one that has the lowest value is added to the set. The procedure can continue until the entire network is covered. DiaBLE is a variation of DIAMOnD: instead of considering the whole network as a fixed universe to compute the p-value, it defines a local expanding universe. This consists of the current disease genes set, the candidate genes (the ones that have at least one connection to a disease gene), and their first neighbours. The DIAMOnD implementation we used is the one provided by Ghiassian, Menche and Barabasi in Python [10], and we modified it, as previously explained, to run DiaBLE. Finally, the heat-diffusion algorithm is equivalent to a random walk with restart but with the following assumptions: no restart, undirected edges on the graph, time steps that approach zero in length, and not running the algorithm until an equilibrium is reached, but until a short amount of time (defined as a parameter) has passed. As values for the time parameters of the diffusion-based algorithms, we used t = 0.002, 0.005, 0.01.

To evaluate the performance of the algorithms we used 5-fold cross-validation [11] done with R [12] in RStudio [13] with the packages ggplot2 [14] and dplyr [15]. This method consists in dividing the disease genes' set into five different sets of equal size (we have four subsets of 156 genes and one of 157 genes). The algorithm is trained on a training set made by four of the five subsets, and the last subset is used as a test set. The procedure is repeated five times, changing at each iteration the training and test sets. The metrics that we analysed are precision, recall, and F1-score, and for each of them we measured the average value and the standard deviation over the five different sets. The metrics were computed for five different samples sizes N, with N $\in \{50, 0.1n, 0.25n, 0.5n, n\}$, where $n = 713$ is the number of known disease genes. When we try to predict new disease genes, there are four possible outcomes, and their frequency is described by the said metrics:

	Disease gene	Not disease gene
Identified as disease gene	True Positive (TP)	False Positive (FP)
Identified as not disease gene	False Negative (FN)	True Negative (TN)

- $precision = \frac{TP}{TP + FP}$, tells us which proportion of the genes identified as disease ones are actually disease genes
- $recall = \frac{TP}{TP + FN}$, tells us which proportion of disease genes was actually identified
- $F1 - score = 2 * \frac{precision * recall}{precision + recall}$, it's the harmonic mean of precision and recall

For the following analysis, the putative disease genes are the genes computed by the best-performing algorithm. In order to further understand their functional implications, we performed the enrichment analysis using the web-based platform EnrichR [16]. This process is useful to understand if a set of genes is involved in a particular biological process or molecular function. To achieve this goal, statistical methods, like the hypergeometric test, are used to evaluate if the given genes are over-represented in a certain pathway or Gene Ontology (GO) [17] category. EnrichR uses a large variety of datasets, but for the purpose of our study, we were only interested in the GO and the Kyoto Encyclopedia of Genes and Genomes (KEGG) datasets [18]. First, the disease genes initially obtained from DisGeNET, and then the first 200 putative disease genes of the list obtained at the previous step, were used as input. For both of them, a table for each of the three GO domains Biological Process, Cellular Components, Molecular Functions and a table for the KEGG pathways were returned. The genes of the first 10 terms of each table resulting from the analysis of the disease genes were compared with a short Python program with the genes of the first 10 terms of the homonymous table resulting from the putative disease genes, to see if common genes could be found.

Lastly, to check the opportunity of drug repurposing, the first 20 putative disease genes with the highest ranking resulting from the analysis performed until now, were selected to check for associated drugs via DGIdb [19]. The associated drugs were ranked by the number of drugs they are associated with. Since not all genes could be found in the Drug-Gene-Interaction database, the top 20 genes that are represented are chosen.

RESULTS AND DISCUSSION

In the following section the obtained results are presented and discussed. In the beginning the best algorithm for the used network is chosen to consecutively perform the enrichment analysis. After that the possibility of drug repurposing is discussed.

As previously stated, DIAMOnD, DiaBLE, and the heat-diffusion algorithm were compared using 5-fold cross-validation, considering precision, recall, and F1-score for different samples of size N.

In figure 1, the results of the performance evaluation for the three algorithms are shown. The dots represent the mean over the values computed with the five subsets, and the error bars each have the length of the corresponding standard deviation. For the heat-diffusion algorithm, only the value for $t = 0.005$ is considered for simplicity, since the different values for t only change the size of the values but not the overall behaviour as shown in figure 3 in the appendix. As can be seen, the DIAMOnD and DiaBLE algorithms performed exactly the same for each sample size. Both algorithms chose the same disease genes in the exactly same order despite computing different p-values, so the values of the performance metrics are the same. For increasing sample size N, the average value of the F1-score is staying the same, where the variation gets smaller as expected. The average values of the precision are decreasing with a growing sample size, where the values of the recall are rising. The standard deviation for the precision is shrinking drastically with increasing N, where the variation in the recall does not change. The diffusion algorithm had difficulties in predicting any putative disease genes, which are real disease genes, so for a small sample size N the number of true positives is zero. Since the performance metrics have the number of the true positives in the numerator, the mean of the metrics are zero as well as the standard deviation. For an increasing sample size, the values of all three metrics for the diffusion algorithm are getting larger with increasing variation, which can be seen from the size of the error bars.

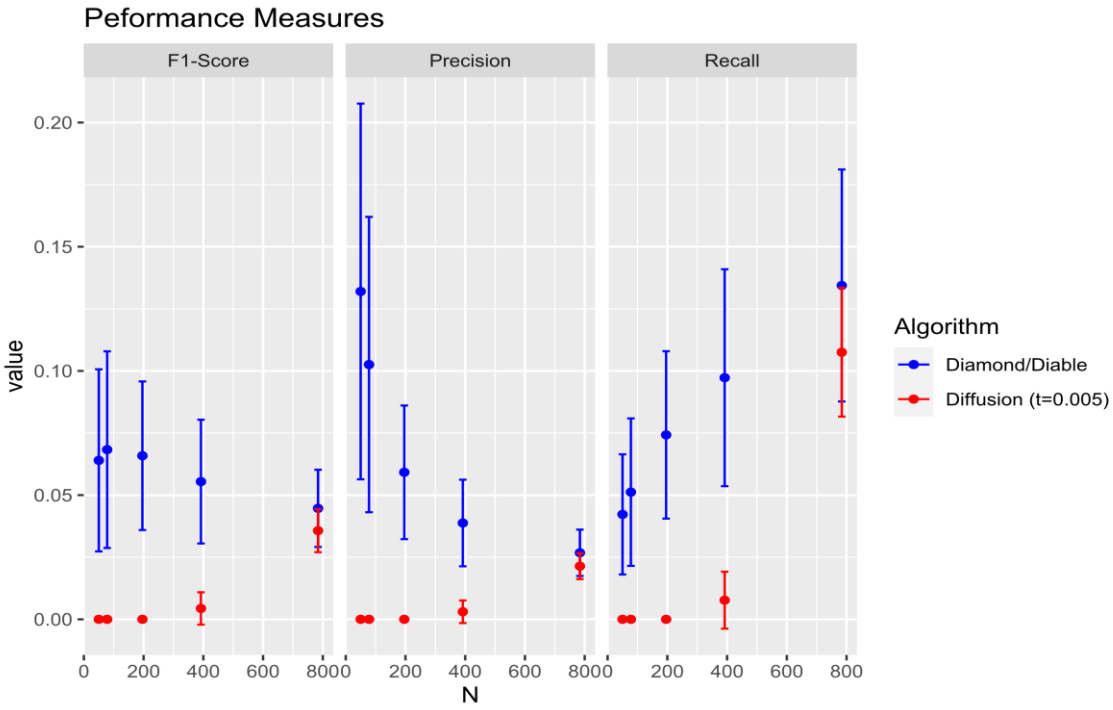

Figure 1 Performance Evaluation of the Algorithms. Sample size N vs. the mean values of performance measures as dots, variation as error bars.

Since the diffusion algorithm is not able to predict any true positives for small sample sizes, the DIAMOnD and DiaBLE algorithms are performing better. The performance measures for DIAMOnD and DiaBLE are exactly the same, so for computing the putative disease genes from the whole data for further investigation, both of the algorithms can be chosen. In the following, the results of the DiaBLE algorithm are used, since the algorithm uses a more accurate method to compute the p-values for choosing the next disease gene and generally performs better than DIAMOnD, as shown in [8].

The putative disease genes predicted with the DiaBLE algorithm are displayed in figure 2:

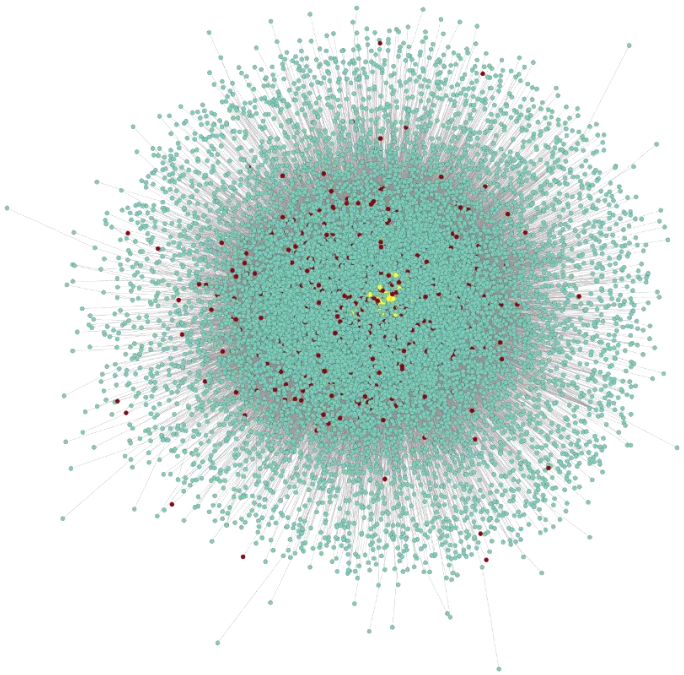

Figure 2 LCC (Disease genes in red, putative disease genes in yellow)

As can be seen, the putative disease genes (shown in yellow) are all in the centre of the network. Computing the average degree, betweenness and closeness centrality of the whole network and of the putative disease genes we can gather interesting observations.

Table 3 Mean values for different metrics (whole network vs only putative disease genes)

	Degree	Betweenness	Closeness Centrality	ratio Betw./Degree
Whole network	81.12913	0.00009	0.364361	4.376162e-07
Putative disease genes	564.87500	0.00068	0.445434	9.157499e-07

In fact, the putative disease genes show higher values in the metrics than average: this means they are central actors in the biological processes of the organism. They could more easily influence or be influenced by other genes, possibly leading to the disease. Furthermore, a high degree strongly suggests that the putative genes are more likely to be hubs: crucial to the functioning of the network. The higher betweenness centrality value shows that the detected putative disease nodes are on average more likely to be on the communication paths between other nodes, correlatively on the path between disease nodes. However, further research is needed on those matters in order to make proper conclusions.

Unfortunately, when performing the enrichment analysis, while comparing the results of disease genes and putative disease genes, no common genes or pathways were found. This suggests that in reality, these putative disease genes are likely not involved in the illness. As a consequence, in biological and medical studies, genes that have been found to be more likely connected to stomach neoplasms should be prioritised.

Nonetheless, for what concerns drug repurposing, the drugs Selumetinib, Tanespimy and Trametinib could each be associated with 5 different genes of the 20 putative genes we considered. For Selumetinib no study could be found regarding stomach neoplasms in the database of ClinicalTrials.gov [20]. The drug Tanespimycin, instead, was tested in a study in 2005 for different types of cancer, but unfortunately the results were not published [21]. For the drug Trametinib, there are two active studies found, where one of the studies started in 2015 with a lot of different drugs and various types of cancer and the other one started at the end of 2022 and is focused on the drug and gastric cancer specifically. For complementary research on this topic, it would be interesting to look at the results of these trials.

In conclusion, the putative disease genes predicted by the DiaBLE algorithm were found most likely to be central actors in the biological biological network, with higher values for degree, betweenness, eigenvector centrality and closeness centrality than the average of the whole graph. This shows that these genes could be important in the development of stomach cancer. However, the enrichment analysis suggested that they are not really involved in the biological processes and pathways that may lead to this disease. Nonetheless, encouraging prospects in treating stomach neoplasms could come from the ongoing clinical trials regarding Trametinib and from the results of the not yet published trial involving Tanespimycin.

We hope that this work can help pursue research on this dreadful disease, by aiding the prioritisation of genes to be studied and by supporting the medical trials on the aforementioned drugs.

AUTHOR CONTRIBUTIONS

Data Gathering: S.J., M.P.; algorithm implementation: S.J.; algorithm validation: S.J.; enrichment analysis: S.d.P., M.P.; writing: (Abstract: M.P., S.J.; Introduction: S.d.P.; Materials and Methods: S.J., M.P., S.d.P.; Results and Discussion: S.J., M.P., S.d.P.), review: J.S., M.P., S.d.P.

REFERENCES

[1] American Cancer Society:

[a] https://www.cancer.org/cancer/stomach-cancer/about/what-is-stomach-cancer.html
[b] https://www.cancer.org/cancer/stomach-cancer/causes-risks-prevention/risk-factors.html

[2] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011 Jan;12(1):56-68. doi:10.1038/nrg2918

[3] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D535-9. doi: 10.1093/nar/gkj109. PMID: 16381927; PMCID: PMC1347471.

[4] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramag D, … Ideker T. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Research. 2003;13(11), 2498–2504.

[5] https://apps.cytoscape.org/apps/cyfinder

[6] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong FI. The DisGeNET knowledge platform for disease genomics. 2019 update. Nucl. Acids Res. (2019) doi:10.1093/nar/gkz1021

[7] Ghiassian SD, Menche J, Barabási AL. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015 Apr 8;11(4):e1004120. doi: 10.1371/journal.pcbi.1004120. PMID: 25853560; PMCID: PMC4390154.

[8] Petti M, Bizzarri D, Verrienti A, Falcone R, Farina L. Connectivity Significance for Disease Gene Prioritization in an Expanding Universe. IEEE/ACM Trans Comput Biol Bioinform. 2020 Nov-Dec;17(6):2155-2161. doi: 10.1109/TCBB.2019.2938512. Epub 2020 Dec 8. PMID: 31484130.

[9] Carlin DE, Demchak B, Pratt D, Sage E, Ideker T. Network propagation in the cytoscape cyberinfrastructure. PLoS Comput Biol. 2017;13(10) :e1005598. https://doi.org/10.1371/journal.pcbi.1005598

[10] https://github.com/dinaghiassian/DIAMOnD

[11] Mitchell TM. Machine Learning. McGraw-Hill Science/Engineering/Math. 1997 March 1; ISBN: 0070428077.

[12] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[13] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[14] Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[15] Wickham H, François R, Henry L and Müller K (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7 https://CRAN.R-project.org/package=dplyr

[16] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics. 2013; 128(14)*.

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. 2016; gkw377

Xie Z, Bailey A, Kuleshov MV, Clarke DJB., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, & Ma'ayan A. Gene set knowledge discovery with Enrichr. Current Protocols, 1, e90. 2021. doi: 10.1002/cpz1.90

[17] Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. Jan 2021;49(D1):D325-D334.

[18] Kanehisa M, Goto S; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000). DOI: 10.1002/pro.3715 PMID: 31441146

Kanehisa M; Toward understanding the origin and evolution of cellular organisms. Protein Sci. 28, 1947-1951 (2019). DOI: 10.1002/pro.3715 PMID: 31441146

Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M; KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 51, D587-D592 (2023). DOI: 10.1093/nar/gkac963 PMID: 36300620

[19] Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, Griffith M, Griffith OL, Wagner AH, Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D1144–D1151, https://doi.org/10.1093/nar/gkaa1084

[20] https://clinicaltrials.gov/

[21] https://clinicaltrials.gov/ct2/show/NCT00004065?id=NCT00004065+OR+NCT02465060+OR+NCT04454476&draw=2&rank=3&load=cart

APPENDIX

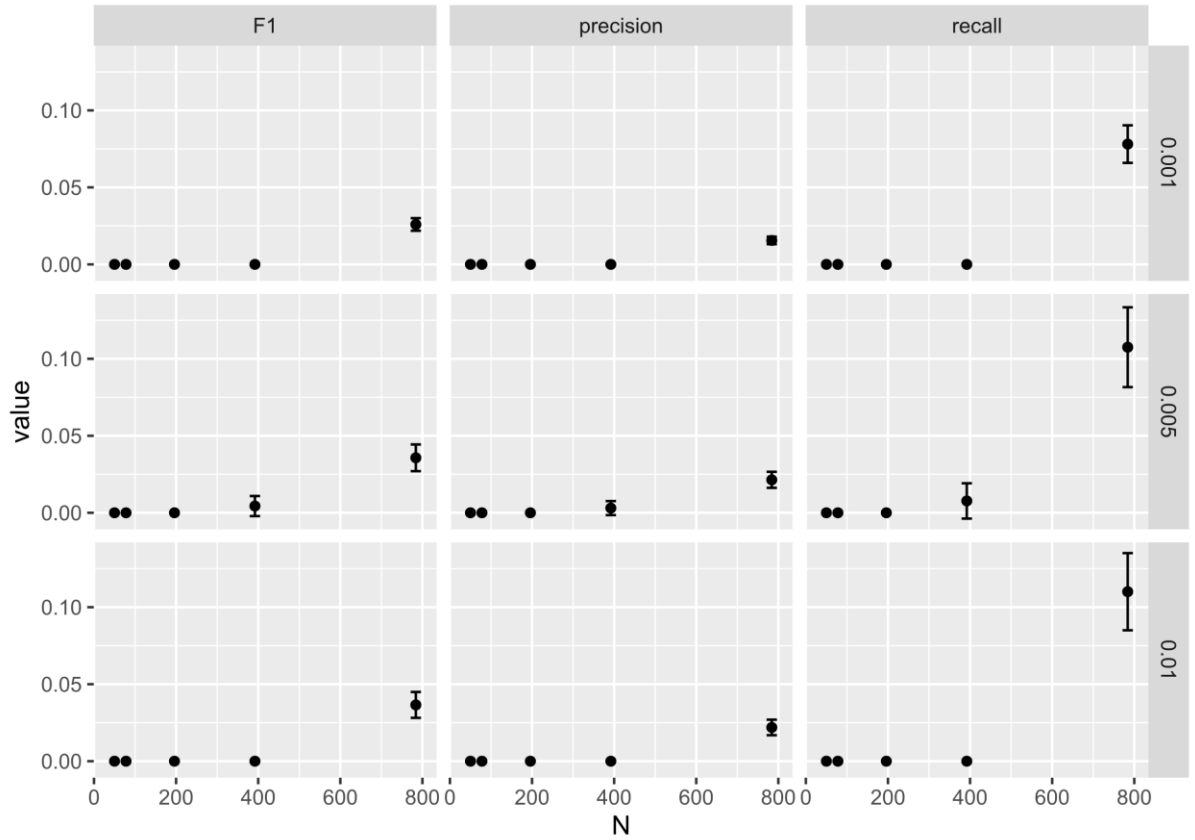

Figure 3 Performance Evaluation of the heat-diffusion algorithm