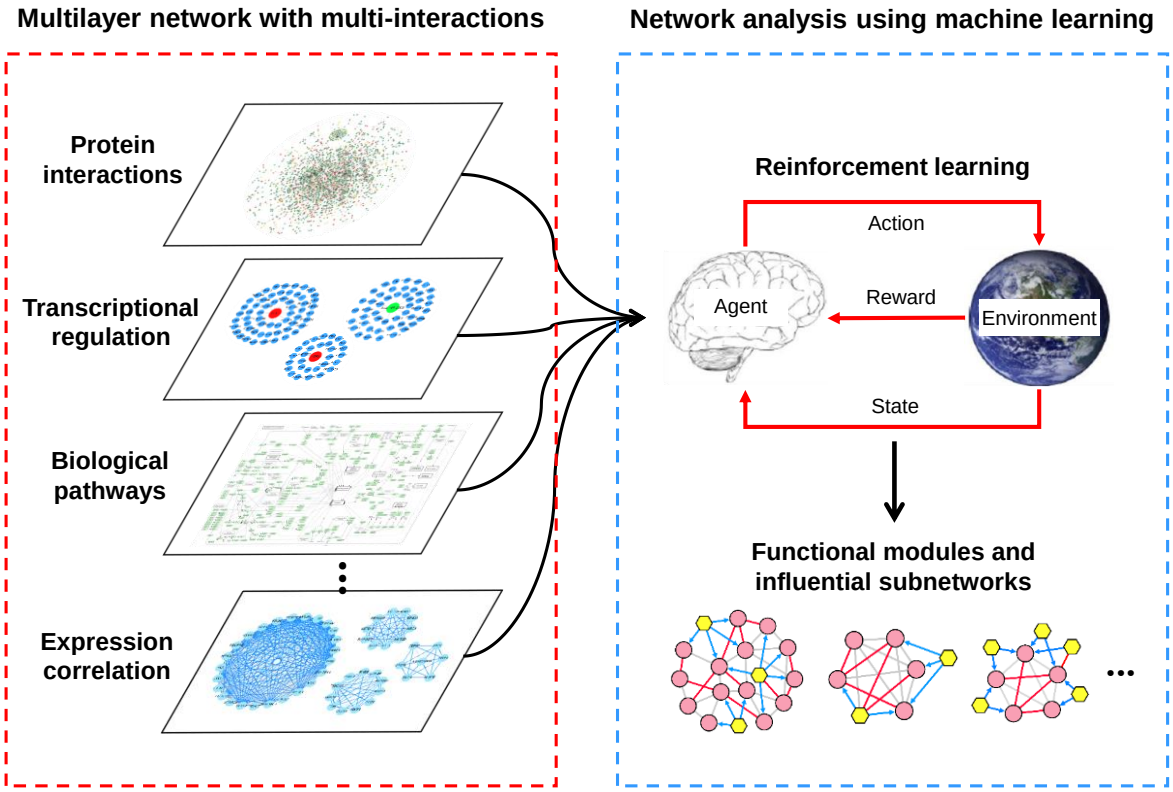
User Manual of CREAM

Contents

1 Introduction

CREAM (Correlation-based Regulatory Events and Modules) is a multi-interaction network framework for gene expression analysis. It enables users to construct multi-interaction networks from gene expression data using desired interactions, identifying functional modules within these networks, and extracting prognostic subnetworks for specific clinical outcomes. This system contains two executable programs: 'CREAM Module Detection.jar' for network construction and module detection and 'CREAM Survival Analysis.jar' for prognostic subnetwork identification. The source code and the complied tool can be downloaded at https://github.com/free1234hm/CREAM.



2 Preliminaries

- To use CREAM a version of Java 1.5 or later must be installed. If Java 1.5 or later is not currently installed, then it can be downloaded from http://www.java.com.
- CREAM can be executed by double-clicking on 'CREAM Module Detection.jar' or 'CREAM Survival Analysis.jar', or from a command line change to the CREAM directory and then type: java -mx1024M -jar CREAM Module Detection.jar (or CREAM Survival Analysis.jar). If CREAM reports 'Out of Memory Error', users can increase the -mx parameter.

3 CREAM Module Detection.jar

3.1 Main Interface

The main interface has three parts: ‘Load Data’, ‘Set Parameters’, and ‘Search Gene Modules’.

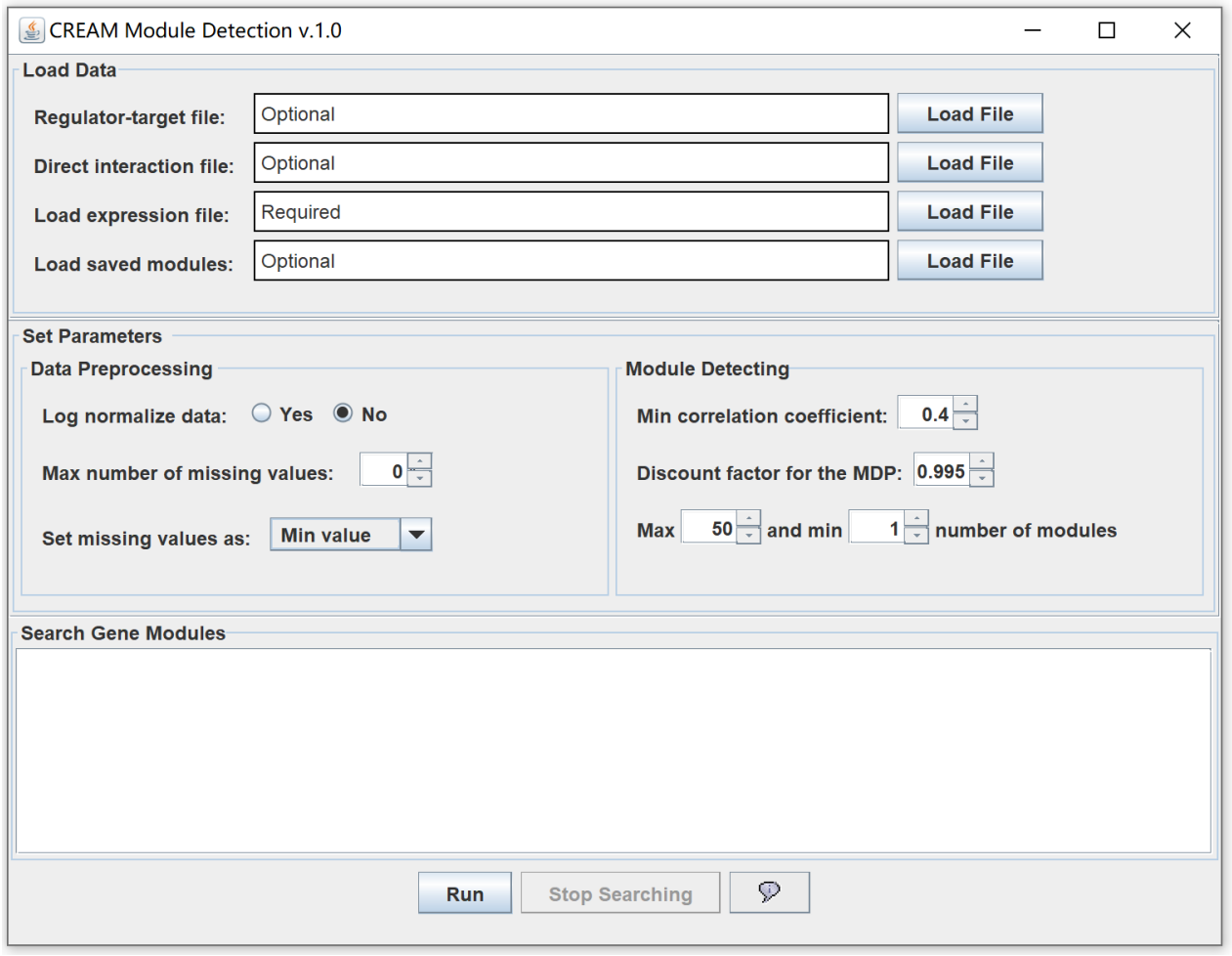


Fig. 1 The main interface of CREAM Module Detection

3.1.1 Regulator-target files

Load established regulator-target pairs, such as transcription factor (TF)-target and microRNA (miRNA)-target pairs. The regulator-target files should be in a three-column format: the first column contains regulators, the second column target genes, and the third column confidence scores from -1 to 1. A score greater than zero represents that the regulator-target pair shows positive regulation, while a score lower than zero represents that the regulator-target pair shows negative regulation. The first row is a header row with user-defined column names (Fig. 2).

TF	Gene	Input
AATF	MYC	1
ABL1	BAX	1
ABL1	CCND2	1
ABL1	CDKN1A	1
ABL1	CSF1	1
ABL1	FOXO3	1
ABL1	JUN	1

Fig. 2 A sample of regulator-target file in three-column format when viewed in Microsoft Excel.

CREAM allows users to customize multi-interaction networks with desired interactions, offering the flexibility to upload personal regulator-target data or utilize curated public databases, encompassing TF–target databases hTFtarget[1] and TRRUST v2.0[2], and miRNA–target databases miRDB[3] and miRTarBase[4] (in the path 'Data\Files for Module Detection\1_Regulator-target file').

3.1.2 Direct interaction files

Load direct interaction files, such as protein–protein interactions (PPIs) and gene functional relationships. The direct interaction files also use a three-column format as described in 3.1.1. A sample direct interaction file is shown in Fig. 3. CREAM allows users to upload personal interaction data or utilize curated public databases, encompassing PPI databases HuRI[5] and STRING[6], and protein family databases InterPro[7] and Pfam[8] (in the path 'Data\Files for Module Detection\2_Direct interaction file').

Protein	Protein	Input
EPHB6	MTARC1	1
NAA11	MTARC1	1
AQP6	MARCHF2	1
APOL2	MARCHF3	1
FATE1	MARCHF3	1
ANKS6	MARCHF5	1
AQP6	MARCHF5	1
BIK	MARCHF5	1
BNIP1	MARCHF5	1
CGRRF1	MARCHF5	1
CREB3	MARCHF5	1
ERGIC3	MARCHF5	1
FATE1	MARCHF5	1

Fig. 3 A sample of regulator-target file in three-column format when viewed in Microsoft Excel.

3.1.3 Load expression file

The 'Load expression file' field is used to Import a gene/protein expression data (Fig. 4). The first column is gene names, and the remaining columns contain the expression values in each sample. If an expression value is missing, then the field should be left empty. The first row of the data contains column headers. Several examples of expression data are included in the path 'Data\Files for Module Detection\ 4_Expression data'.

Ensembl_II	TCGA-AA-	TCGA-AA-	TCGA-A6-	TCGA-A6-	TCGA-AA-	TCGA-CK-	TCGA-AA-	TCGA-AA-	TCGA-AU-	TCGA-QG
A1CF	0.92999	1.066218	2.564578	2.242078	1.208603	0.159212	2.189218	0.061351	0.889145	1.183308
A2M	4.529061	4.870139	5.687503	6.662382	5.405043	5.591557	3.445681	5.108467	4.960791	3.854576
A4GALT	2.662687	1.147233	1.104459	2.482948	1.853063	1.383715	0.828363	2.687003	2.098147	1.040627
AAAS	3.833031	3.628329	3.198295	3.443551	4.099529	3.6168	3.774873	3.734673	3.361572	3.774655
AACS	2.145268	2.235108	2.06301	1.917419	2.63819	1.72618	2.344681	1.980136	2.0168	2.221353
AADAT	0.901021	2.117209	2.33651	1.226698	1.584483	3.170597	2.184566	1.468223	2.000801	2.486626
AAGAB	3.384627	4.911665	4.380784	3.860188	3.745479	3.901575	4.193553	4.471359	3.947197	3.8725
AAK1	0.891756	0.938476	2.48266	1.607293	0.550033	1.147723	1.93236	1.247271	1.246245	1.121613
AAMDC	2.905587	2.386935	3.092975	2.742403	2.883851	2.447846	2.444698	3.620468	1.942847	2.423223
AAMP	6.108965	5.606741	5.733432	5.76106	5.797197	5.758501	5.858539	6.147721	5.74096	5.935313
AAR2	4.62933	3.917359	5.064283	4.874728	4.248512	3.939791	4.888974	4.012086	3.845083	3.927044
AARS2	2.320018	2.451048	3.10118	2.902352	3.031804	3.035382	2.926849	3.194291	2.81012	3.086282
AARSD1	0.78421	0.60167	0.864706	0.742812	0.932212	0.947192	1.105827	0.979237	1.248157	1.35347
AASDH	1.038219	1.847106	2.895586	1.67843	1.136062	1.625119	2.044143	1.792878	1.646203	1.880203
AASDHPPT	2.52936	2.220441	3.44982	2.819377	1.674377	3.527966	2.598485	2.700203	2.565926	2.825342
AATF	3.834741	3.709372	4.429679	4.393276	4.616406	4.195666	4.284369	4.555975	4.052856	4.076867
ABAT	1.619443	0.532441	2.368285	3.182818	2.357015	0.972122	2.26131	0.342083	0.986267	1.378474
ABCA1	1.230966	1.153374	2.143139	1.730441	0.891639	2.322669	1.32176	1.31	1.863681	1.358856
ABCA2	2.71114	4.202821	1.586837	3.293604	2.069393	3.293773	4.021893	3.098194	2.880384	3.48955

Fig. 4 A sample of expression data file when viewed in Microsoft Excel.

3.1.4 Load saved modules

The 'Load saved modules' field allows the user to specify a file containing a set of saved gene modules (Fig. 5), thus saving time if the gene modules of an expression data have already been detected. The saved module files contain at least 3 columns: the first column contains genes, the second column module names, and the third column indicates the correlations between genes. If all genes in a specific module are positively correlated, the values in the third column will be 1. If a module contains two negatively correlated gene subsets, they are represented as 1 and -1 respectively. Several examples of saved module files are included in the path 'Data\Files for Module Detection\ 5_Saved module file'.

Gene	Module	Correlation	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	PCA8
AARSD1	Module_1	1	-2.87454	-1.28459	2.215632	-3.338	-0.67496	-1.74119	-2.91016	2.086842
ABCA7	Module_1	1	1.261271	-3.04789	-1.55607	-2.25275	-2.68476	-2.56174	-0.6391	0.23436
ABR	Module_1	1	0.282974	-3.05526	-1.31856	-4.16938	-1.33879	0.589838	-0.7513	0.915965
ACADVL	Module_1	1	1.518815	-0.29755	-2.8492	-3.7508	-2.00649	9.86E-04	-0.23224	3.043774
ACAP3	Module_1	1	2.102768	-3.68385	-0.13942	-2.14491	-2.28447	-2.02586	0.321794	2.434402
ACBD4	Module_1	1	0.55945	-1.13054	1.390641	-1.89181	-4.02158	-2.19076	-1.88379	0.015443
ACCS	Module_1	1	0.008412	-2.45555	2.152759	-3.25687	-2.29349	-0.61687	-1.55215	1.261047
ADAMTS1	Module_1	1	1.491846	-2.82921	0.746017	-2.77182	-3.54513	-0.97599	-2.69628	0.402332
ADCY4	Module_1	1	0.853392	-4.79114	-1.09348	-0.26687	-2.55595	0.557648	-1.59544	1.85302
ADCY6	Module_1	-1	0.902775	-4.3333	0.52923	-5.00637	-1.67677	-0.40937	-0.03272	1.404401
ADM5	Module_1	-1	2.391278	-3.14078	-2.03112	0.338904	-1.90138	-1.72279	-2.16556	2.485522
AGAP6	Module_1	-1	-1.19568	-2.46131	2.217016	-3.98068	-2.16711	-2.386	-2.03733	3.206904
AGER	Module_1	-1	-0.01759	-2.5068	-0.81549	-2.08482	-3.91732	-1.07157	-1.65232	4.170771
AKAP17A	Module_1	-1	0.400148	-2.44081	2.338096	-2.79191	-2.14225	-2.84758	-0.62578	1.648463
AKR1C1	Module_1	-1	-0.39046	-0.26334	0.312892	-0.84468	-0.88432	-1.13488	0.057864	-1.48286
AKR7L	Module_1	-1	-0.1902	-1.14945	-0.75627	-3.48223	-2.78468	-2.25223	-0.6585	0.479894
ALKBH6	Module_1	-1	-2.78468	-1.10404	1.891781	-3.05114	-3.69975	-1.52282	-1.45756	2.025009
ALOX12	Module_1	-1	-1.06751	-3.05878	-1.11583	-2.73698	-3.07635	-0.93534	-0.82178	0.835769
ALS2CL	Module_1	-1	1.072246	-2.13804	1.360604	-4.38696	-3.22443	-1.02233	-1.11365	0.605726

Fig. 5 An example of saved module file

3.1.5 Set Parameters

Assuming that the expression vector of a gene is $\{v_1, v_2, \dots, v_n\}$.

- **Log normalize data**—transforms the vector to $\{\log_2 v_1, \dots, \log_2 v_n\}$.
- **Maximum number of missing values**—a gene will be filtered if the number of missing values in

all samples exceeds this parameter.

- **Set missing value as**—the missing value of a gene can be set as ‘min value’ (default), ‘mean value’ or ‘zero’.
- **Min correlation coefficient**—when calculating the probability of correlation between gene expression patterns, we define a minimum correlation coefficient threshold (0.4 by default) to increase the distinction between low and high correlation coefficients. When the absolute correlation coefficient between two genes is lower than the minimum threshold, the probability of a correlation between them is defined as 0 ($|r| \in [-0.4,0.4]$).
- **Discount factor for the MDP**—the discount factor for Markov decision process (0.995 by default) is used to limit overfitting. The closer this parameter is to 1, the greater the number of modules generated.
- **Max and min number of modules**—the maximum (50 as default) and minimum (1 as default) number of final modules. Users can obtain a specific number of modules by setting them to the same value.

3.1.6 Results

The text box here displays the running progress of CREAM after pressing the ‘Run’ button. Pressing the ‘Stop Searching’ Button forces CREAM to stop running. CREAM then proceeds to the display interface.

3.2 Display Interface

3.2.1 Display Modules

After the module detection process executes, the main output window appears. Fig. 6A displays the detailed module information. Users can select an item in the module table to display two sets of genes positively and negatively correlated with the selected expression pattern. Users can convert the gene tables to gene expression heatmaps by clicking the ‘HeatMap’ button (Fig. 6B).

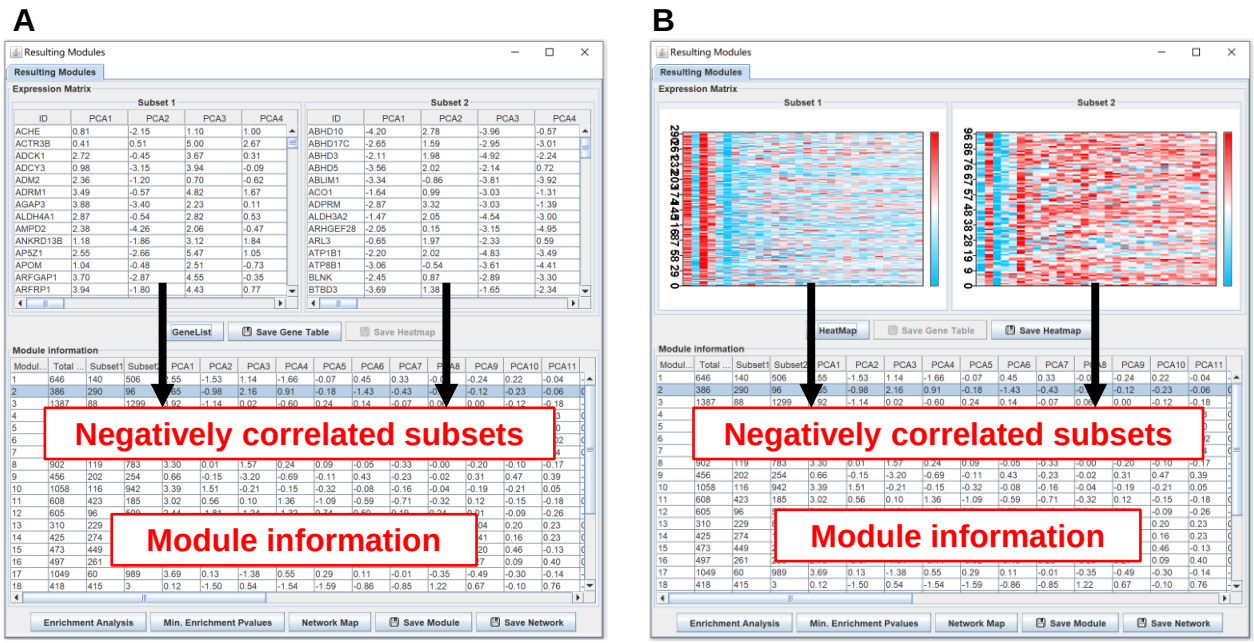
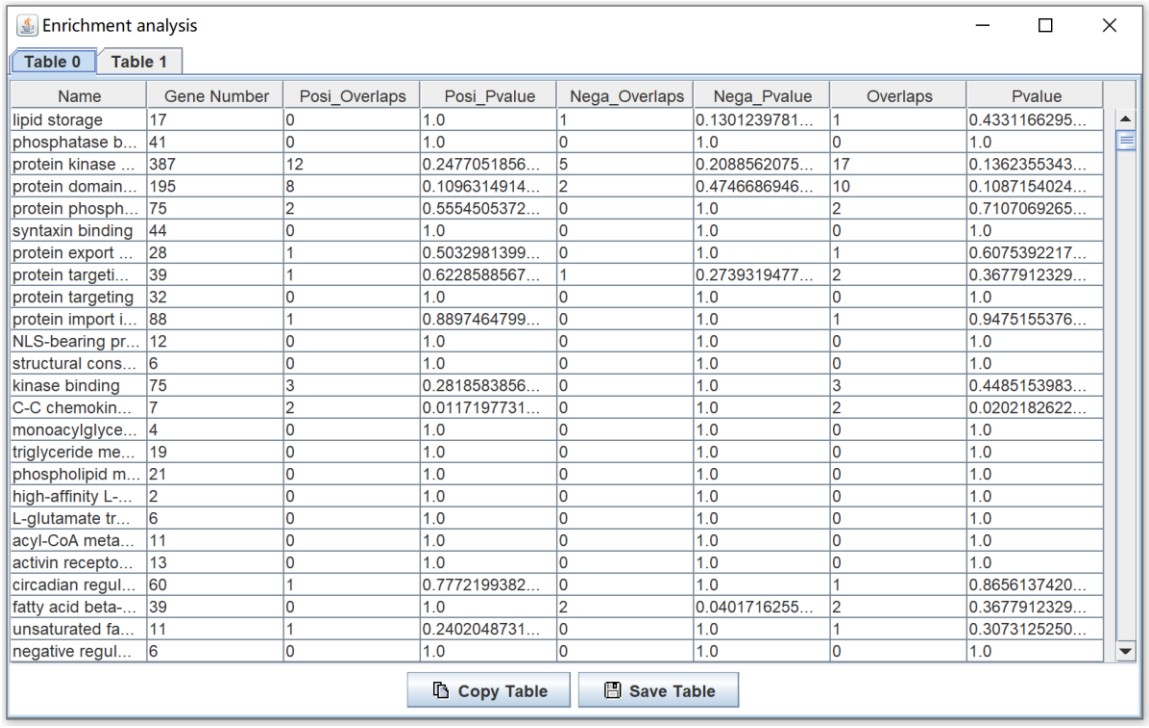


Fig. 6 An example of the module information.

3.2.2 Enrichment Analysis

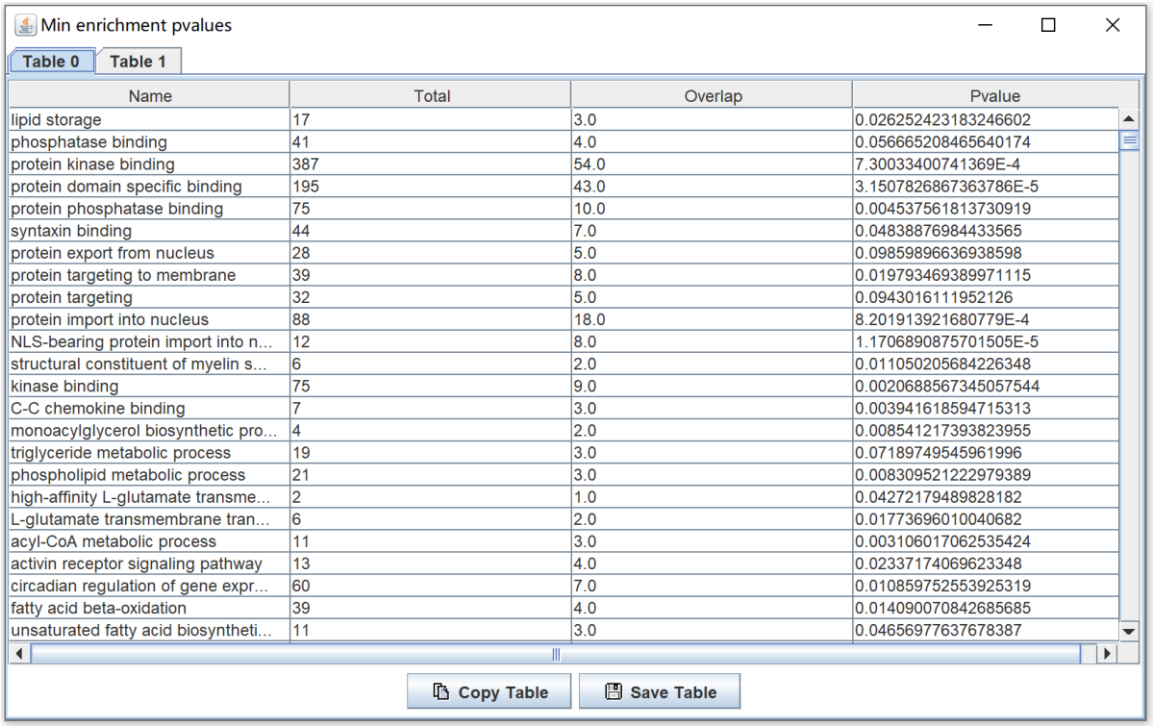
User can select an item in the module table and press the 'Enrichment Analysis' button to check the KEGG pathways and Gene Ontology (GO) terms significantly enriched in the selected module (Fig. 7).



Name	Gene Number	Posi_Overlaps	Posi_Pvalue	Nega_Overlaps	Nega_Pvalue	Overlaps	Pvalue
lipid storage	17	0	1.0	1	0.1301239781...	1	0.4331166295...
phosphatase b...	41	0	1.0	0	1.0	0	1.0
protein kinase ...	387	12	0.2477051856...	5	0.2088562075...	17	0.1362355343...
protein domain...	195	8	0.1096314914...	2	0.4746686946...	10	0.1087154024...
protein phosph...	75	2	0.5554505372...	0	1.0	2	0.7107069265...
syntaxin binding	44	0	1.0	0	1.0	0	1.0
protein export ...	28	1	0.5032981399...	0	1.0	1	0.6075392217...
protein targeti...	39	1	0.6228588567...	1	0.2739319477...	2	0.3677912329...
protein targeting	32	0	1.0	0	1.0	0	1.0
protein import i...	88	1	0.8897464799...	0	1.0	1	0.9475155376...
NLS-bearing pr...	12	0	1.0	0	1.0	0	1.0
structural cons...	6	0	1.0	0	1.0	0	1.0
kinase binding	75	3	0.2818583856...	0	1.0	3	0.4485153983...
C-C chemokin...	7	2	0.0117197731...	0	1.0	2	0.0202182622...
monoacylglyce...	4	0	1.0	0	1.0	0	1.0
triglyceride me...	19	0	1.0	0	1.0	0	1.0
phospholipid m...	21	0	1.0	0	1.0	0	1.0
high-affinity L-...	2	0	1.0	0	1.0	0	1.0
L-glutamate tr...	6	0	1.0	0	1.0	0	1.0
acyl-CoA meta...	11	0	1.0	0	1.0	0	1.0
activin recepto...	13	0	1.0	0	1.0	0	1.0
circadian regul...	60	1	0.7772199382...	0	1.0	1	0.8656137420...
fatty acid beta-...	39	0	1.0	2	0.0401716255...	2	0.3677912329...
unsaturated fa...	11	1	0.2402048731...	0	1.0	1	0.3073125250...
negative regul...	6	0	1.0	0	1.0	0	1.0

Fig. 7 An example of enrichment analysis

Users can also press the 'Min. Enrichment Pvalues' button to summarize the enrichment results of all modules, obtaining the minimum p-value for each pathway or GO term (Fig. 8).



Name	Total	Overlap	Pvalue
lipid storage	17	3.0	0.026252423183246602
phosphatase binding	41	4.0	0.056665208465640174
protein kinase binding	387	54.0	7.30033400741369E-4
protein domain specific binding	195	43.0	3.1507826867363786E-5
protein phosphatase binding	75	10.0	0.004537561813730919
syntaxin binding	44	7.0	0.04838876984433565
protein export from nucleus	28	5.0	0.09859896636938598
protein targeting to membrane	39	8.0	0.019793469389971115
protein targeting	32	5.0	0.0943016111952126
protein import into nucleus	88	18.0	8.201913921680779E-4
NLS-bearing protein import into n...	12	8.0	1.1706890875701505E-5
structural constituent of myelin s...	6	2.0	0.011050205684226348
kinase binding	75	9.0	0.0020688567345057544
C-C chemokine binding	7	3.0	0.003941618594715313
monoacylglycerol biosynthetic pro...	4	2.0	0.008541217393823955
triglyceride metabolic process	19	3.0	0.07189749545961996
phospholipid metabolic process	21	3.0	0.008309521222979389
high-affinity L-glutamate transme...	2	1.0	0.04272179489828182
L-glutamate transmembrane tran...	6	2.0	0.01773696010040682
acyl-CoA metabolic process	11	3.0	0.003106017062535424
activin receptor signaling pathway	13	4.0	0.02337174069623348
circadian regulation of gene expr...	60	7.0	0.010859752553925319
fatty acid beta-oxidation	39	4.0	0.014090070842685685
unsaturated fatty acid biosyntheti...	11	3.0	0.04656977637678387

Fig. 8 An example of 'Min. Enrichment Pvalues'

3.2.3 Network Map

User can select an item in the module table and press the 'Network Map' button to show the muti-interaction network associated with the selected module (Fig. 9). In the network map, regulators are

shown as green octagons, genes as pink rectangles, regulator–gene interactions as black arrows, and direct interactions as blue lines. Users can filter regulators based on degree of connectivity (Fig. 10), select a regulator in the right table to display its targets (Fig. 11), and elect a gene in the right table to display its interacting partners (Fig. 12).

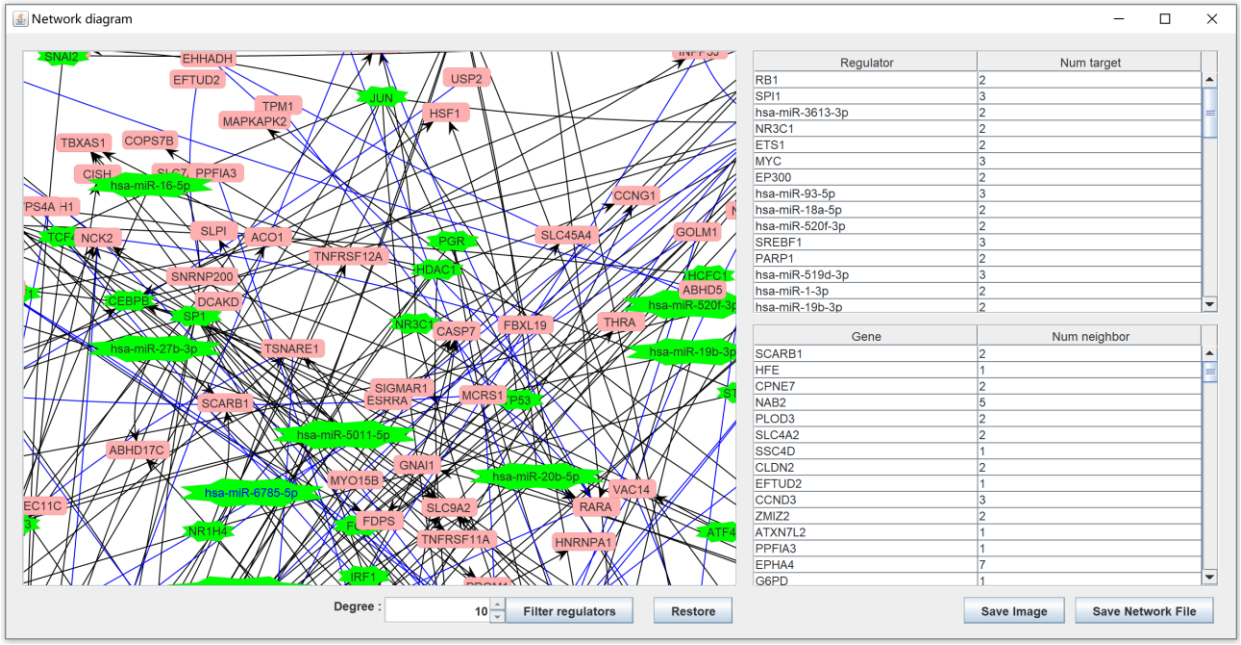


Fig. 9 An example of network map.

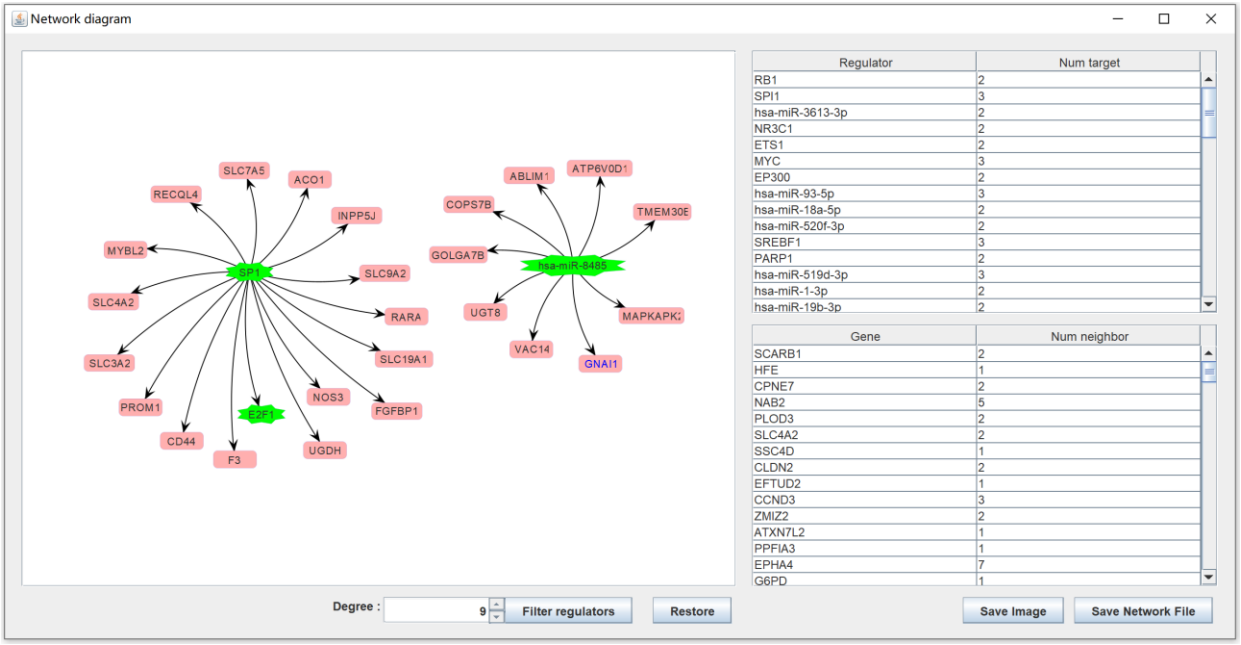


Fig. 10 An example of pressing the ‘Filter regulators’ button.

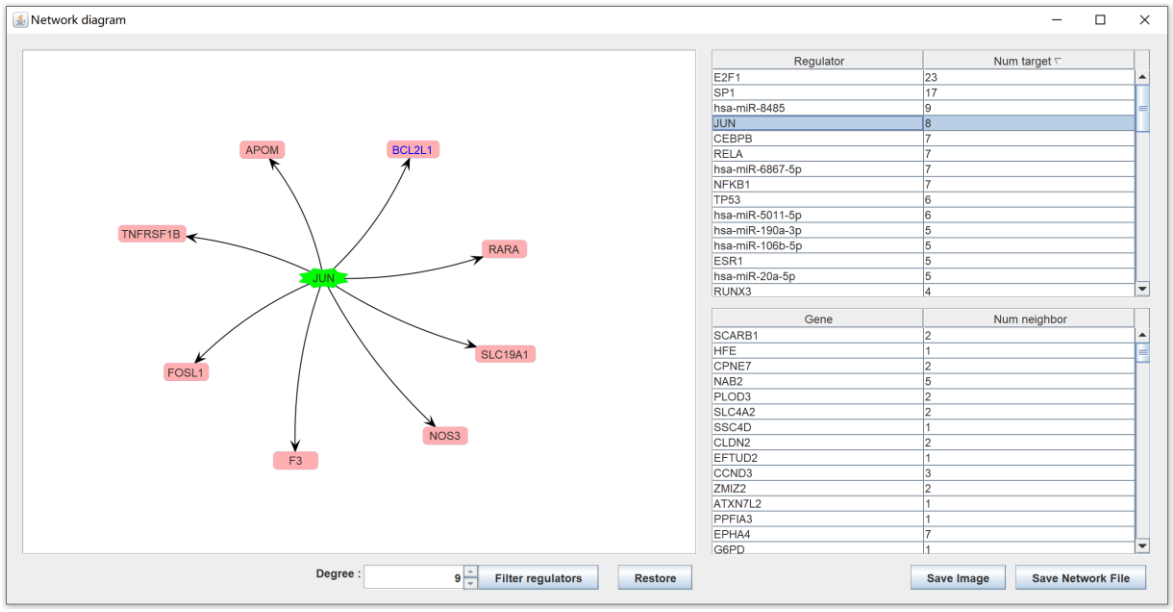


Fig. 11 Example of selecting regulator JUN.

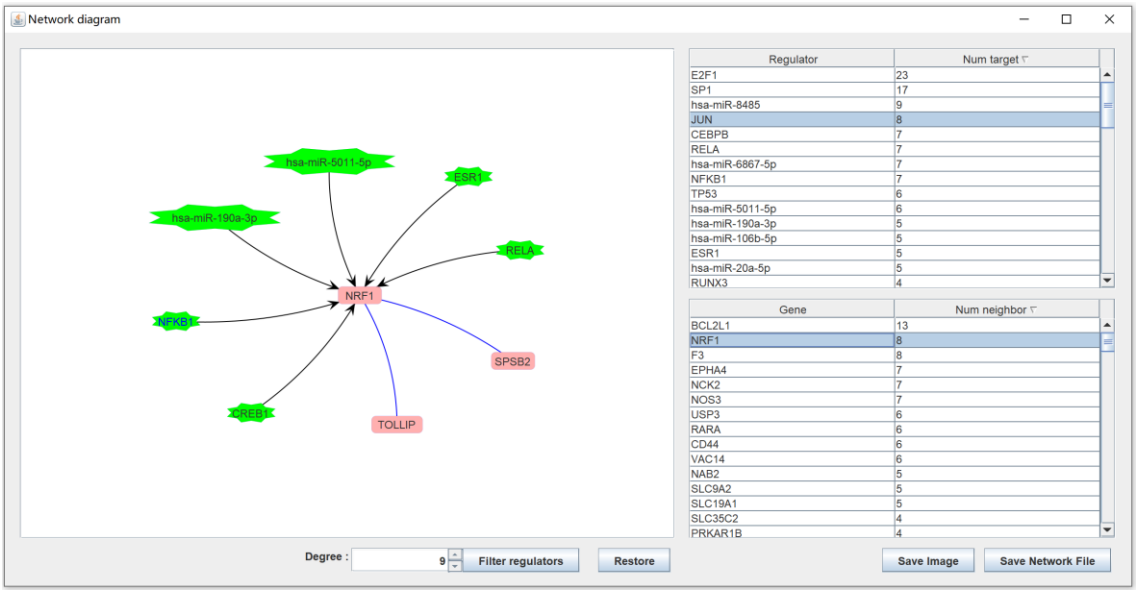
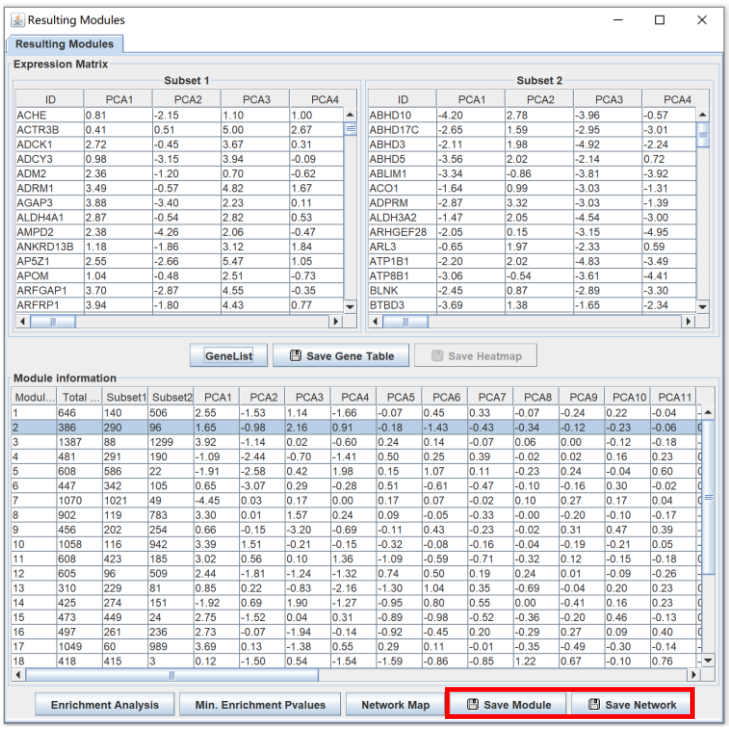


Fig. 12 Example of selecting gene NRF1.

3.2.4 Save Module and Save Network

Users can press the ‘Save Module’ button to store their identified gene modules (Fig. 13). Notably, users can upload a saved module file using the 'Load Saved Modules' button to save time the next time they analyze that data. Additionally, users can press the ‘Save Network’ button to integrate all modules into a genome-wide multi-interaction network.



Saved Module File

Gene	Module	Correlation	PCA1	PCA2	PCA3
AARSD1	Module_1	1	-2.87454	-1.28459	2.215632
ABCA7	Module_1	1	1.261271	-3.04789	-1.55607
ABR	Module_1	1	0.282974	-3.05526	-1.31856
ACADVL	Module_1	1	1.518815	-0.29755	-2.8492
ACAP3	Module_1	1	2.102768	-3.68385	-0.13942
ACBD4	Module_1	1	0.55945	-1.13054	1.390641
ACCS	Module_1	1	0.008412	-2.45555	2.152759

Saved Network File

ID1	ID2	Type
ACOT11	SUOX	direct
ACTN4	TRIM3	direct
ACTN4	C14ORF11	direct
ACVRL1	NR1H2	direct
ACY1	BCAS2	direct
ARAF	YWHAE	direct
ARAF	KRAS	direct
BAG6	DNAJB11	direct
CASP9	BIRC5	direct
CSNK1D	TRIM3	direct
CSNK1E	TRIM3	direct
ERBB2	PTK6	direct

Fig. 13 Examples of saved modules and networks

4 CREAM Survival Analysis.jar

4.1 Main Interface

'CREAM Survival Analysis.jar' is designed to identify prognostic subnetworks associated with specific clinical outcomes by integrating molecular networks, gene expression data, and patient clinical data. Its main interface consists of two parts: 'Load Data' and 'Analysis' (Fig. 14).

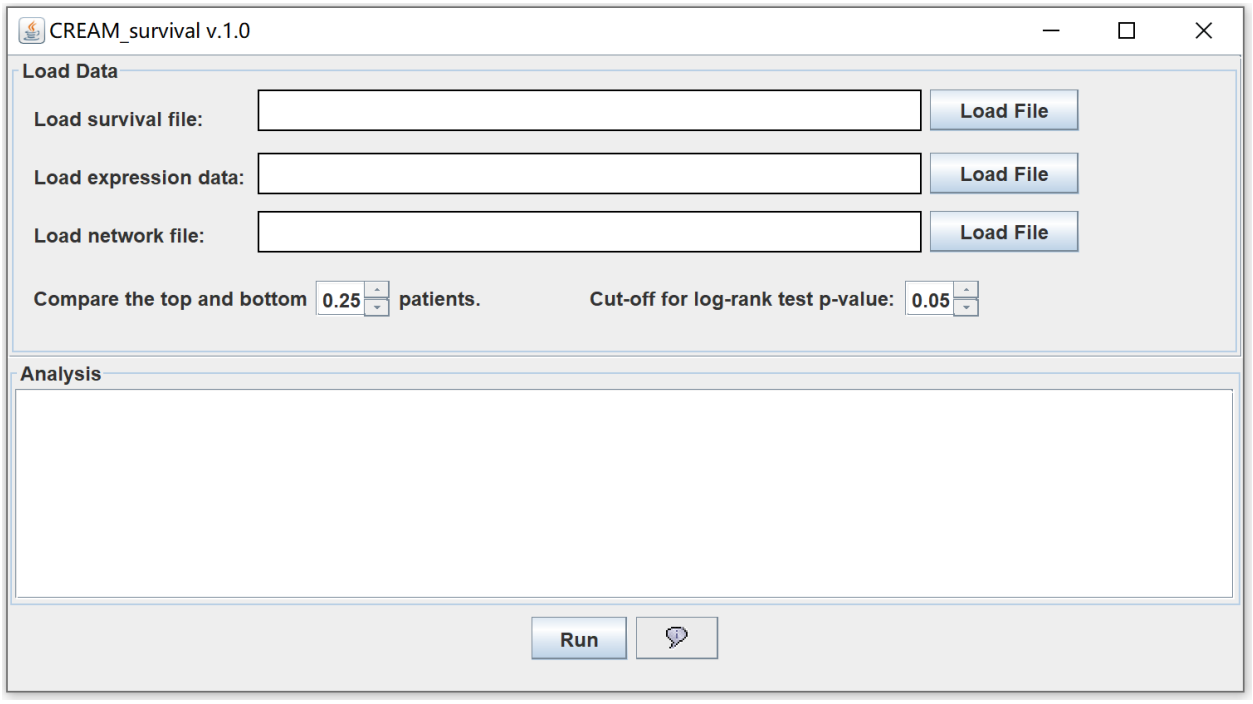


Fig. 14 The main interface of CREAM Survival Analysis.jar

4.1.1 Load survival file

Load patient clinical data such as overall survival, disease recurrence, and tumor metastasis. The file

should be in a three-column format (Fig. 15): patient ID, patient state (the value 1 represents the occurrence of death, recurrence, or metastasis, the value 0 indicates that the event has not occurred yet), and time. The first row is a header row with user-defined column names. A sample survival file is included in the path 'Data\Files for Survival Analysis\Survival file'.

sample	OS	OS.time
TCGA-AA-3492-01A	1	1
TCGA-AA-3492-11A	1	1
TCGA-G4-6626-01A	1	1
TCGA-AA-3525-01A	0	1
TCGA-AA-3525-11A	0	1
TCGA-AY-6196-01A	0	6
TCGA-AM-5820-01A	0	14
TCGA-F4-6854-01A	0	16
TCGA-AA-A02O-11A	0	28
TCGA-AA-A02O-01A	0	28
TCGA-AM-5821-01A	0	28
TCGA-AY-4071-01A	1	29
TCGA-AA-A00R-01A	0	30
TCGA-AA-3543-01A	0	30
TCGA-AA-3818-01A	1	30

Fig. 15 An example of survival file

4.1.2 Load expression file

The 'Load expression file' field is used to import patient expression data. The data format is the same as 'CREAM Module Detection' (see section 3.1.3). Notably, the patient IDs in the first row of the data should match the patient IDs in the first column of the survival file.

4.1.3 Load network file

The 'Load network file' field is used to import network information. The file should be in a three-column format: the first two columns represent interacting molecules and the third column specifies their interaction type (Fig. 16). A sample network file is included in the path 'Data\Files for Survival Analysis\Network file'.

ID1	ID2	Type
SCOC	SIKE1	direct
SCOC	C21ORF91	direct
SCOC	ABCE1	co-expression
SCOC	AIMP1	co-expression
SCOC	ANAPC10	co-expression
SCOC	ARFIP1	co-expression
SCOC	ASNSD1	co-expression
SCOC	BET1	co-expression

Fig. 16 An example of network file

4.1.4 Survival analysis parameters

CREAM utilizes the log-rank test to assess differences in clinical outcomes between the top and bottom portions (1/4 as default) of ranked patients. Additionally, it allows users to customize the p-value cutoff, with a default setting of 0.05, for this test.

4.1.5 Analysis

The text box here displays the running progress of network-based survival analysis after pressing the 'Run' button.

4.2 Display Interface

4.2.1 Survival Table

The table in Fig. 17 is an example of survival analysis results: the first column lists genes, the second column presents their respective log-rank p-values, the third column displays the number of interacting partners for each gene, and the fourth column specifies the type of interaction (e.g., co-expression, direct interaction, etc.).

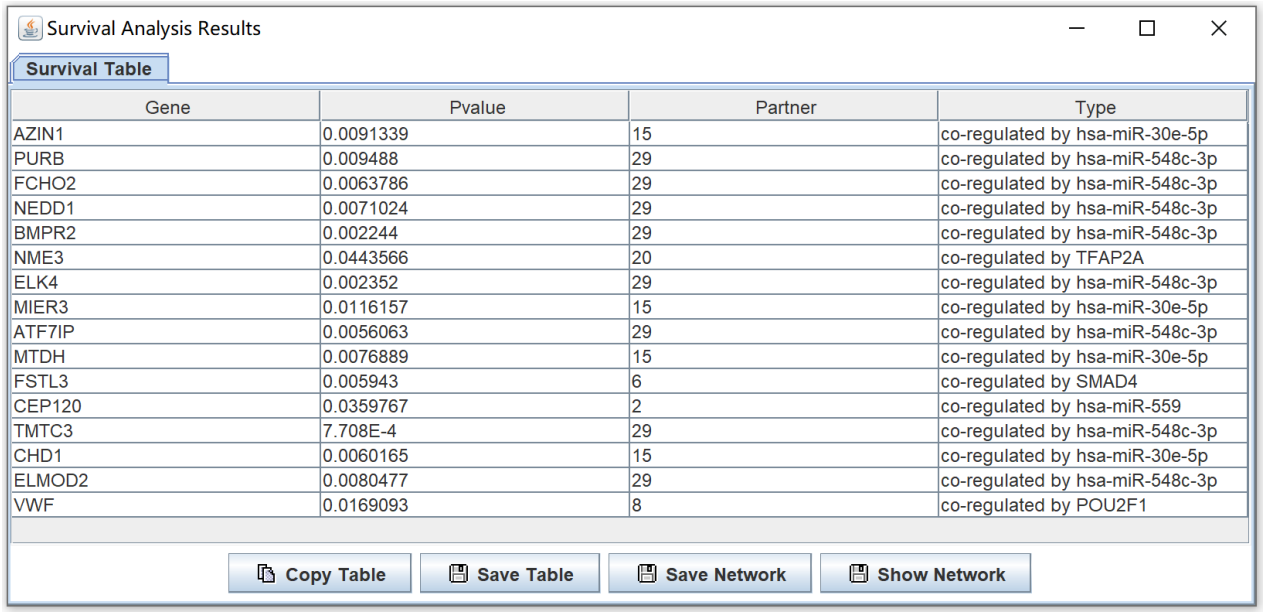


Fig. 17 An example of survival analysis results

4.2.2 Show Network

User can select an item in the 'Survival Table' and press the 'Show Network' button to view the regulators and genes linked to the selected gene (Fig. 18A). In the network map, regulators are shown as green octagons, genes as pink rectangles, regulator–gene interactions as black arrows, and direct interactions as blue lines. By pressing the 'Key nodes' button, users can further view the subnetworks of significantly survival genes with p-values below a user-defined cutoff (Fig. 18B).

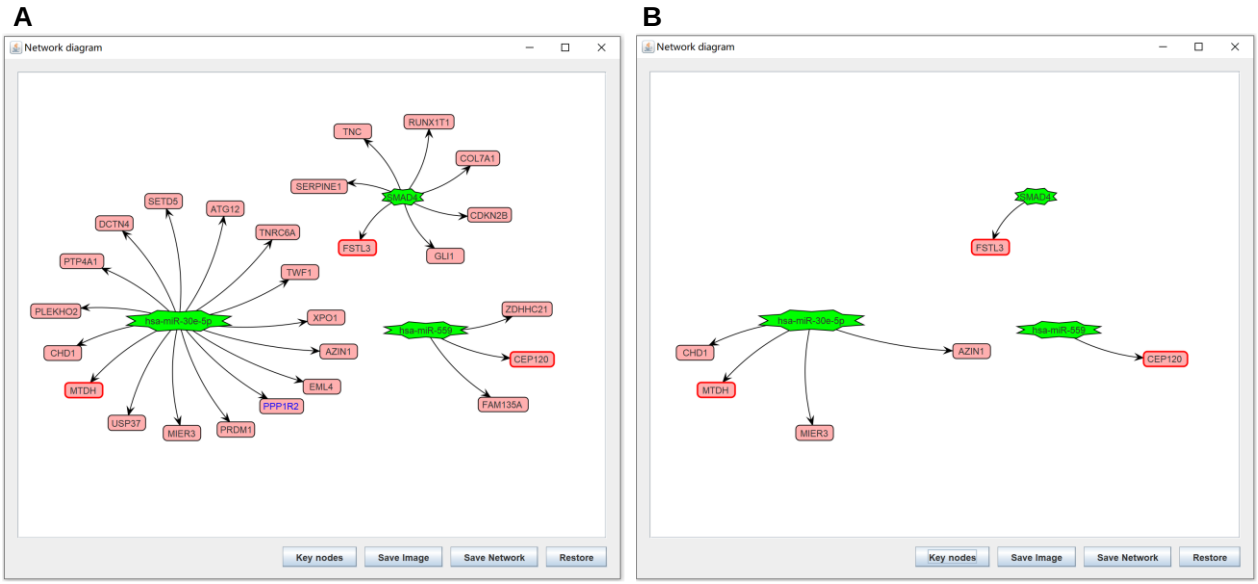


Fig. 18 Examples of pressing the ‘Show Network’ button and the ‘Key nodes’ button

Reference

1. Zhang, Q. et al. hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. *Genomics Proteomics Bioinformatics* **18**, 120-128 (2020).
2. Han, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* **46**, D380-D386 (2018).
3. Chen, Y. & Wang, X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res* **48**, D127-D131 (2020).
4. Huang, H.Y. et al. miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* **50**, D222-D230 (2022).
5. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402-408 (2020).
6. Szklarczyk, D. et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* **49**, D605-D612 (2021).
7. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic Acids Res* **51**, D418-D427 (2023).
8. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-D419 (2021).