Математички факултет, Универзитет у Београду

Семинарски рад из предмета: Истраживање података 2

Аутори:
Марко Никитовић
Бошко Андрић

Предметни наставник:
Проф. др Ненад Митић

Школска година 2023/2024

Садржај

1 Увод

Тема семинарског рада је поновити поступак који је објављен у раду Мрежно-медицински приступ за идентификовање генетских повезаности аденома паратироидне жлезде са кадриоваскуларним болестима и дијабетесом типа 2 објављеног у Информисање о Функционалној Геномици (енг. Network-medicine approach for the identification of genetic association of parathyroid adenoma with cardiovascular disease and type-2 diabetes, Nikhat Imam, Aftab Alam, Mohd Faizan Siddiqui, Akhtar Veg, Sadik Bay, Md. Jawed Ikbal Khan and Romana Ishrat published in Briefngs in Functional Genomics, 2023, 22, pp.250-262)[1]. Фокусираћемо се на методе приказане у поменутом раду, те ћемо покушати поновити њихове резултате користећи сличне приступе.

Садржај рада се ослања на чињеницу да се аденом паратироидне жлезде ("PTA") повремено јавља заједно са кардиоваскуларним болестима ("CVD") и са тип 2 дијабетесом ("T2D"). Даље се у раду наводи да тренутна студија предлаже упоређивање протеина повезаних са "PTA" са преклапајућим протеинима "CVD" и "T2D" како би се утврдила повезаност болести. Циљањем централних протеина који су повезани са ове три болести може довести до боље клиничке слике у случају истовремене присутности и повећати ефикасност неких лекова.

Даље у тексту ћемо се пре свега бавити програмерским делом овог рада, слабије се осврћући на медицинске и биолошке детаље.

Напомена: За све кодове коју су писани било у "R" или "Python" програмском језику, потребно је подесити радни директоријум на директоријум пројекта (погледати достављено упутство). Такође све путање дате су релативно у односу на почетни директоријум.

2 Методологија и резултати

У овом делу ћемо по целинама, онако како су задате у оригиналном раду, разматрати имплементације одређених поступака као и добијене резултате, при чему их упоредујући са референтним резултатима из већ поменутог рада.

2.1 Прикупљање протеина за одговарајуће болести

Протеини који су повезани са болестима "PTA", "CVD" и "T2D" преузети су са "DisGeNet"[2] базе података, која представља збирку података из разних других база, при чему смо водили рачуна да преузимамо само оне податке који су референцирани бар једном ("PubRef" ≥ 1, коришћен филтер на самом сајту базе). Подаци се могу пронаћи у директоријуму "files/disease_genes".

Подаци које смо добили нису идентични али су приближни подацима који су аутори рада оставили и користили у свом раду. На даље, кад год будемо имали нека одступања у подацима, користићемо оригиналне податке како бисмо имали упоредиве резултате.

Број гена	Наши подаци	Подаци коришћени у раду
"CVD"	1756	1607
"T2D"	3034	2803
"PTA"	125	116

Табела 1: Табела која упоређује податке које смо ми презуели са оргиналним подацима

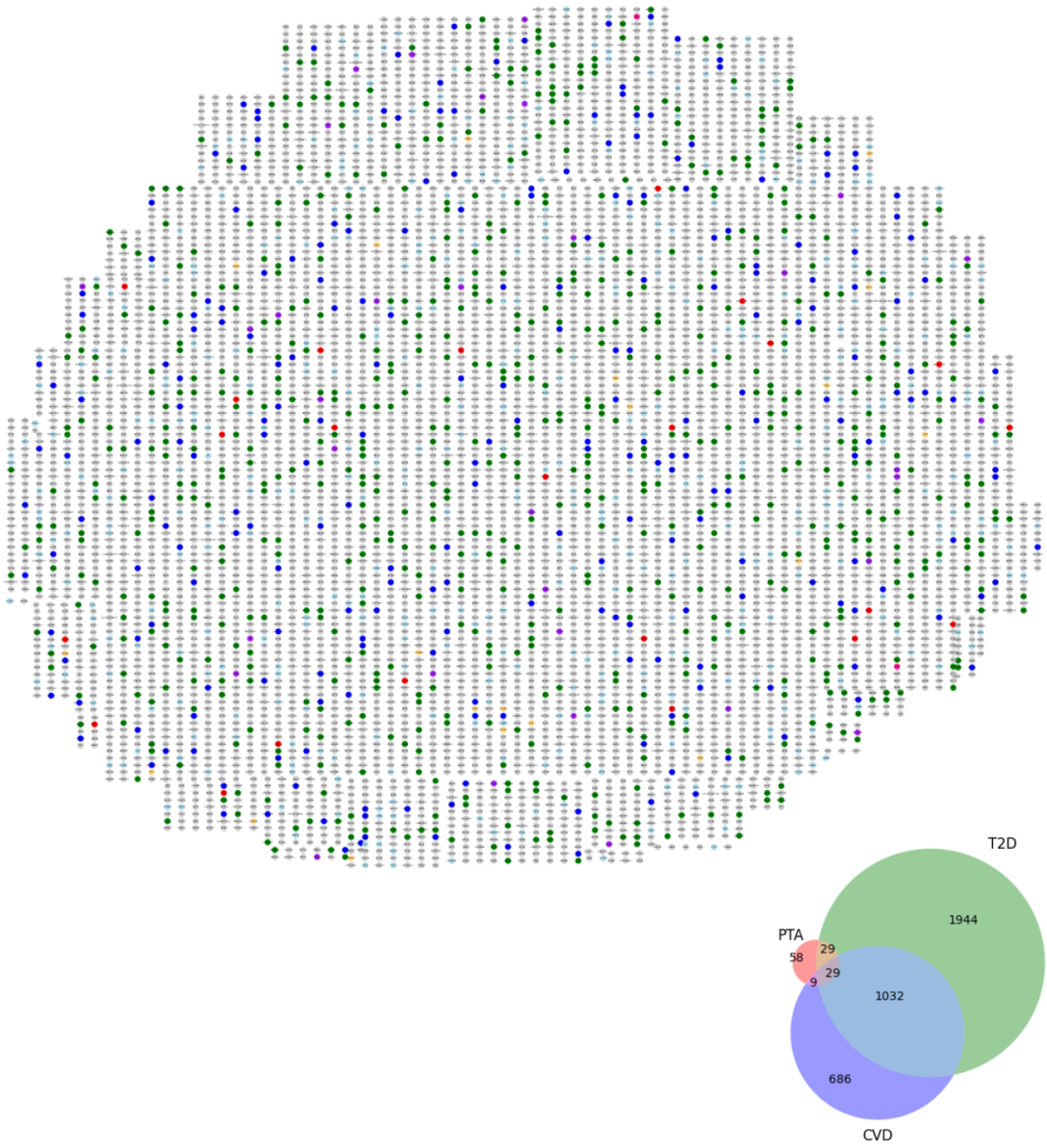
2.2 Конструкција мреже протеина и њене тополошке карактеристике

"PPI" (енг. protein-protein interaction, протеин-протеин интеракција) мрежа описује како протеини између себе интерагују, односно граф ће садржати грану за свака два протеина која интерагују. "PPI" мрежа нам даје бољи увид у: разумевању биолошких функција, одређивање кључних протеина, проучавању механизама болести, итд.

У нашем случају мрежу смо конструисали над протеинима који су заједнички за све три болести (33 протеина из рада) и њихових првих суседа. Скрипта која проналази заједничке протеине дата је у "scripts/common_genes.py", а резултати у "files/common_genes.xlsx" (покренута над њиховим подацима). Када скрипту покренемо над подацима које смо ми преузели добијамо листу од 29 протеина који се могу пронаћи у "files/our_common_genes.xlsx" и која је доста слична њиховој листи. Надаље ћемо радити са њиховим списком од 33 протеина да би испратили поступак до краја. Скрипту покрећемо тако што поставимо путању до директоријума (променљива "directory_path", у нашем случају њена вредност је "files/disease_genes/") и имена датотека (променљива "file_names").

Ми смо конструисали мрежу користећи само "BioGRID"[3] базу, јер пружа много једноставнији програмерски интерфејс (енг. API), док ручно преузимање података преко веб сервиса не би било остварив (потребно за сваки од 33 протеина пронаћи суседе, и онда за сваког од суседа, којих има пар хиљада, проверити да ли између њих постоји грана, и убрзо долази до броја провера који се мери у милионима). Слично би се могло одрадити и за "IntAct"[4], док би сам алгоритам конструкције остао исти. Приликом преузимања интеракције означене са "genetic" или оне чији су експерименти типа "Co-localization", "Genetic interference" "Synthetic Rescue" "Synthetic Growth Defect" и "Synthetic Lethality" су избачене. Преузети подаци се налазе у директоријуму "files/construction_data".

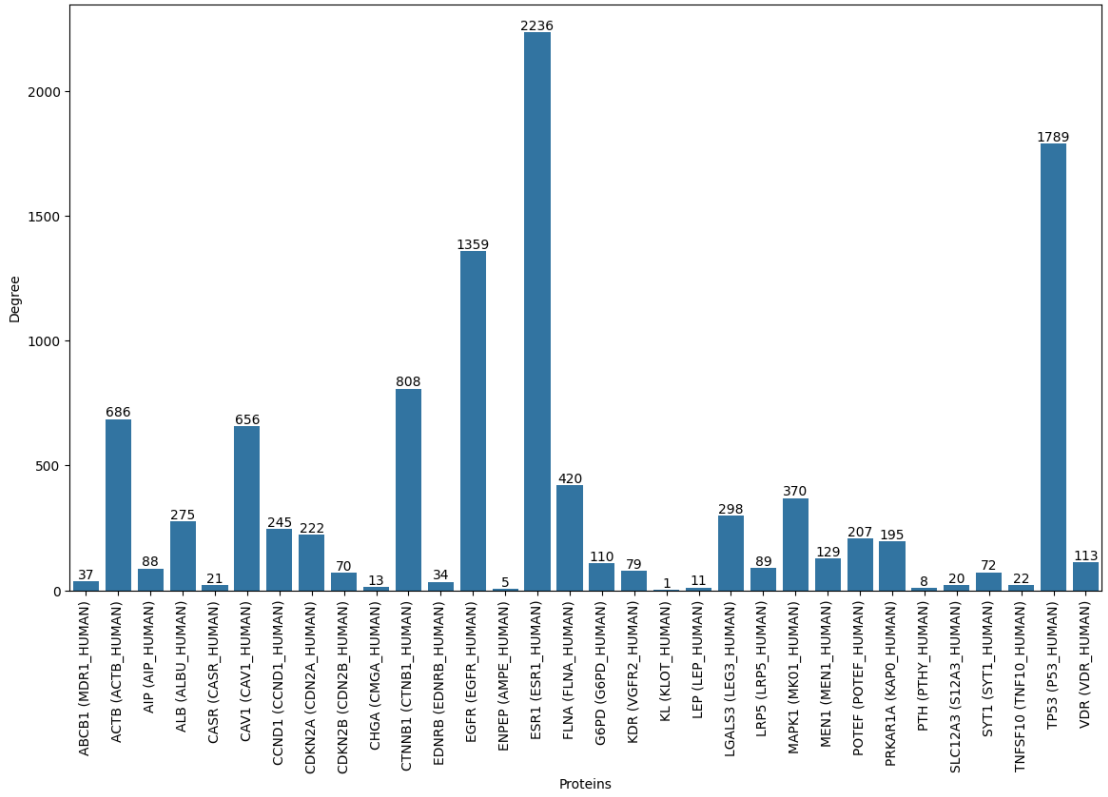
Код се може пронаћи у "network_construction" и "scripts/biogrid_download" (скрипта за преузимање података уз помоћ њиховог веб сервиса [12]), излаз кода дат је датотеци "files/network.sif". Пошто је за приказивање коришћен "Cytoscape"[5] алат, датотека има формат ".sif".



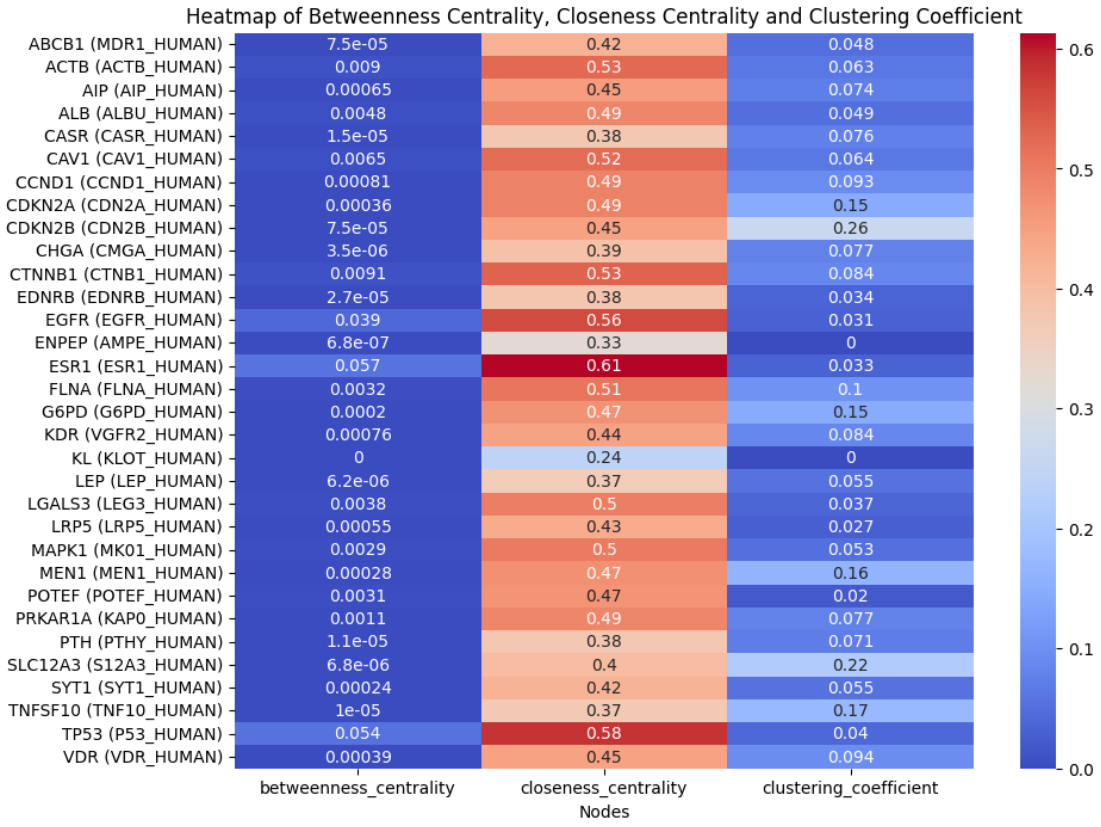
Слика 1: Мрежа интеракција између протеина, као и Венов дијаграм болести и протеина. Боје у мрежи одговарају бојама скупова на дијаграму.

Приликом конструкције водили смо рачуна да избацимо дупле гране и петље. Добили смо мрежу која има 7075 чворова и 192650 грана (нешто мање него њихови резултати, што је и у реду пошто смо податке преузимали само са једне базе). Коначна мрежа може се пронаћи у "output_newtork.cys".

За сваки од 33 протеина смо израчунали следеће тополошке карактеристике: блискост (енг. closeness centrality), коефицијент кластеровања (енг. clustering coefficient), посредништво (енг. betweenness centrality) и степен чвора (енг. node degree). Карактеристике су рачунате над оригиналном мрежом. Код се може пронаћи у датотеци "network_topological_properties.ipynb".



Слика 2: Степен сваког од 33 гена у мрежи



Слика 3: Тополошке карактеристике за већ поменути скуп гена

2.3 Проналажење функционалних модула у "PPI" мрежи

Функционални модули у "PPI" мрежи се траже из више разлога: терапијске стратегије (можемо да гађамо цео модул а не неки конкретан протеин леком), предвиђање функције протеина, разумевање биолошких процес итд.
Међу кластерима и густим подмрежама би се могли тражити функционални модули. Такође, постоји велика вероватноћа да уколико протеин интерагује са одређеном групом протеина, да ће они имати сличну функцију у организму, иако између њих не постоји директна интеракција ипак има смисла да припадају истим модулима. У овом делу смо за проналажење функционалних модула користили "MCODE" алгоритам [6].

"MCODE" алгоритам ради у три фазе: тежинско одређивање чворова, предвиђање комплекса и опционо постпроцесирање за филтрирање или додавање протеина у добијене комплексе по одређеним критеријумима повезаности.

Мрежа интерагујућих молекула може се интуитивно моделирати као граф, где су чворови молекули, а гране молекуларне интеракције. Ако су познати временски путеви или информације о сигнализацији ћелија, могуће је креирати усмерени граф са стрелицама које представљају правац хемијске акције или информационог тока, у супротном се користи неусмерени граф.

- Прва фаза, пондерисање чворова, тежи све чворове на основу локалне густине мреже користећи највиши k-језгро суседства чвора. K-језгро је граф минималног степена k. Коефицијент кластеровања језгра чвора дефинише се као густина највишег k-језгра непосредног суседства чвора, укључујући и сам чвор. Коначна тежина чвора је производ коефицијента кластеровања језгра и највишег нивоа k-језгра у суседству чвора.

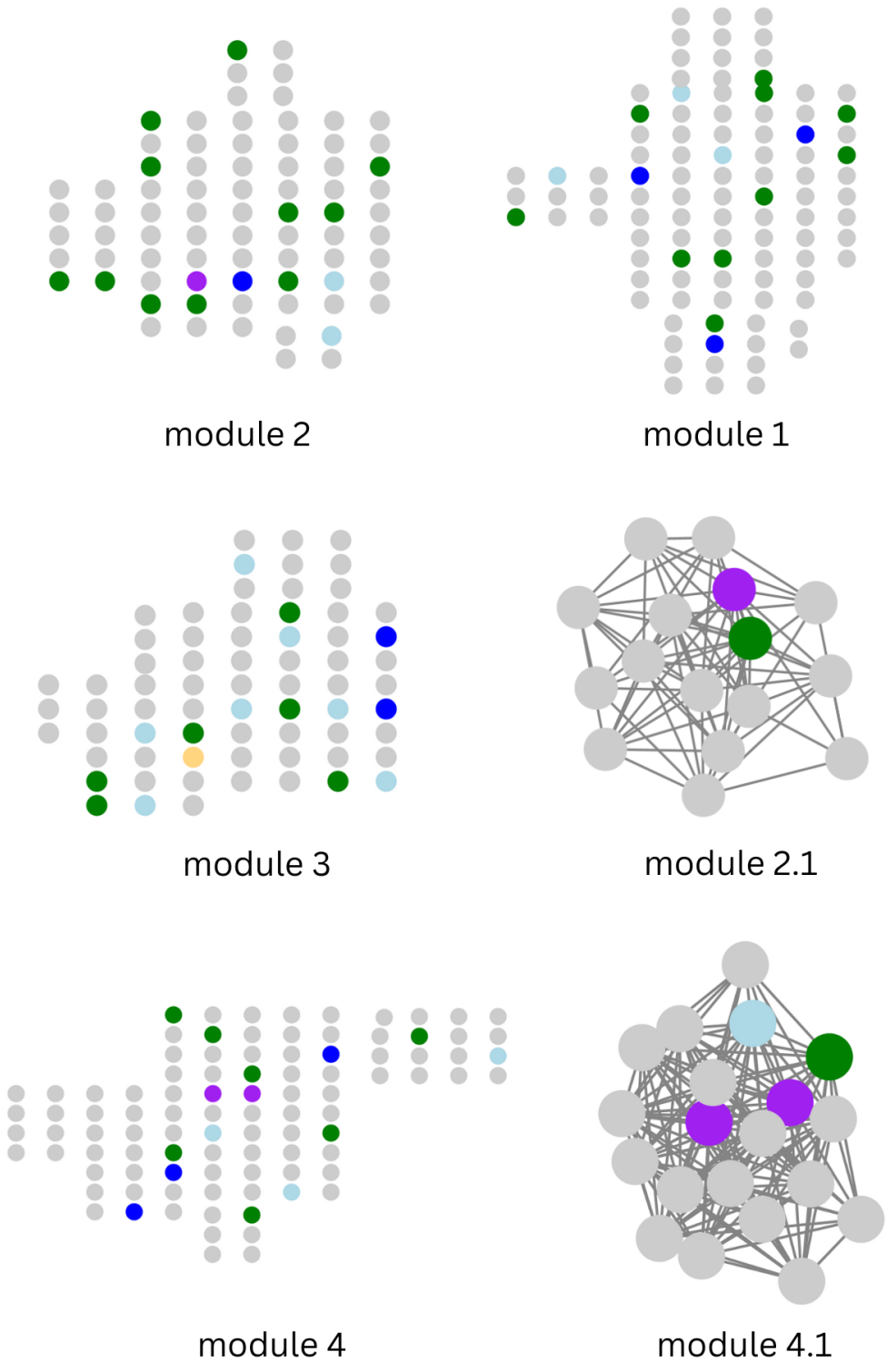
- Друга фаза, предвиђање комплекса, узима као улаз тежински одређен граф и почиње са чвором највеће тежине као семеном комплекса. Комплекс се гради рекурзивно укључујући чворове чија је тежина изнад задатог прага у односу на тежину почетног чвора. Процес се понавља док се не могу додати нови чворови.

- Трећа фаза је постпроцесирање. Комплекси се филтрирају ако не садрже најмање 2-језгро. Алгоритам може радити са опцијом пуњења (енг. fluff) која повећава величину комплекса додавањем суседних чворова ако је њихова густина већа од задатог прага. Опција одсецање (енг. haircut) уклања чворове који су повезани са комплексом само једном граном.

Комплекси се оцењују и рангирају на основу густине подграфа и броја чворова у комплексу. "MCODE" такође може радити у усмереном режиму где је почетни чвор унапред одређен. У овом режиму, "MCODE" предвиђа само један комплекс коме припада дати чвор.

Ми нисмо користили опцију пуњења јер је показано да она не утиче на значајно побољшање резултата, а за праг тежине смо користили вредност 0.2. Код за "MCODE" алгоритам се налази у датотеци "mcode.ipynb" [13].

Прво смо "MCODE" покренули на целој оргиналној мрежи ("files/original_network") и узели 15 највећих кластера. Затим смо задржали само оне који садрже протеине који су заједнички за све болести. Након тога, за сваки од тих модула смо поново покренули "MCODE" алгоритам и добили 11 подмодула (садрже већ поменуте протеине) као и следећи списак централних (енг. hub) протеина: "TP53, ESR1, CTNNB1, ACTB, CDKN2B, CAV1, FLNA, MEN1, CDKN2A, EGFR, ALB". Наш списак није исти али се добрим делом преклапа са њиховим. Сви резултати из "MCODE" анализе могу се пронаћи у "files/mcode_results" : "mcode_clusters_images" директоријум за слике, "sif_files_for_clusters" директоријум који садржи мрежу за сваки од модула и "auxilary_files" директоријум са међурезултатима који садрже списак гена за сваки од модула . Приказивање модула дато је у "modules_visualization.cys".



Слика 4: Неки од модула које смо добили из "MCODE" алгоритма

2.4 Анализа прекомерне репререзентације скупа гена

Анализа прекомерне репрезентације скупа гена (енг. Gene set over-representation Analysis, "GSORA") је биоинформатичка техника која се користи за проналажење скупова гена који су статистички значајно обогаћени у одређеном скупу података. У питању је Фишеров тест тачне вероватноће заснован на хипергеометријској расподели, који обједињује најизраженије биолошке функције добијене из протеина модула.

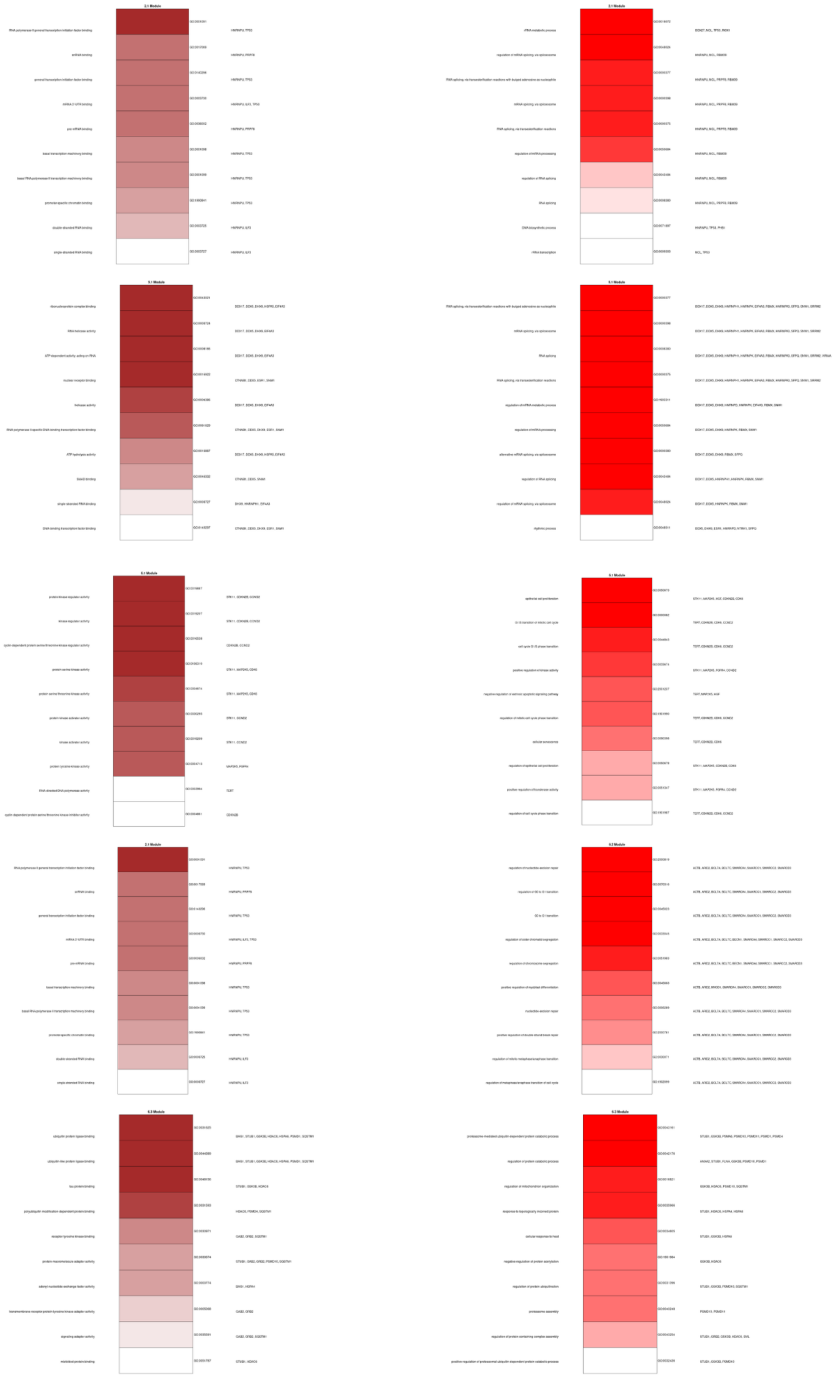
"GO-enrichment"[7] анализа је одрађена коришћењем "R" програмског језика и пакета "ClusterProfiler"[8] и "org.Hs.eg.db"[9]. Користили смо "enrichGO" функцију, при чему је узет "Benjamini-Hochberg" метод за прилагођавање п-вредности. Такође, функцију смо позивали два пута, једном за биолошке процесе а други пут за молекуларну функцију (вредност "ont" параметра постављена на "BF" и "MF"). Ниво значајности је постављен на 0.1 ("pvalueCutoff" аргумент). Од осталих параметара поставили смо "keyType=SYMBOL" за стандардно означавање протеина и "OrgDb=org.Hs.eg.db" за постављање већ поменуте базе података. За приказивање смо користили топлотне мапе.

Анализа обогаћивања путања (енг. enrich pathway) одрађена је исто у "R" језику уз помоћ пакета "ReactomePA"[10] і "ClusterProfiler". И овде смо користили 'Benjamini-Hochberg' метод, а за приказивање је коришћен бипартитиван граф. Минимална и максимална величина (за сваки модул) поставена је на 2-400 протеина. Такође је ниво значајности постављен на 0.1 ("pvalueCutoff"), док постављање "organism" на "human" означавамо да се анализа врши за човека.

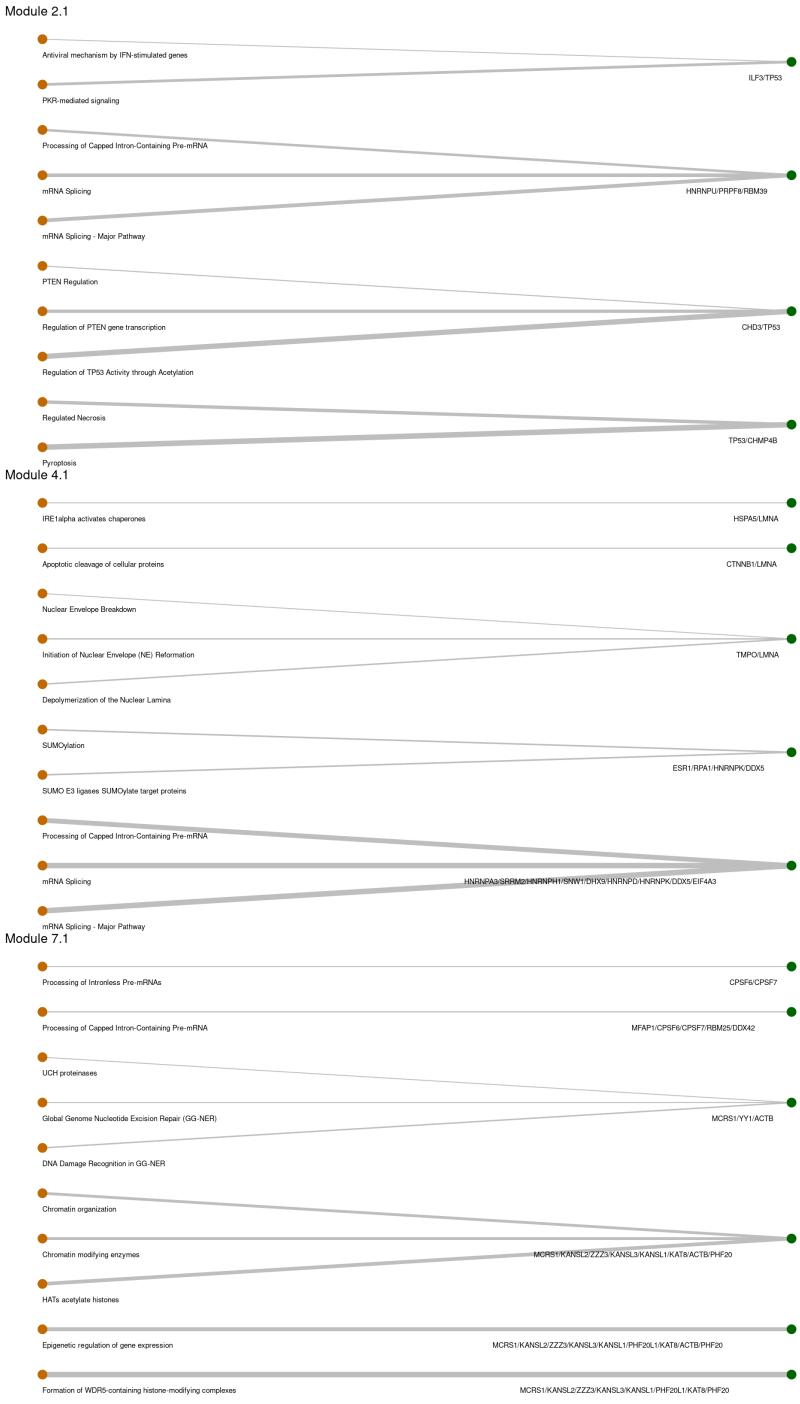
Анализа је рађена на основу модула (из рада и из оних које смо ми добили) који су добијени као излаз "MCODE" алгоритма и резултати које смо добили су упоредиви са њиховим резултатима узимајући у обзир да се база од тренутка објављивања рада мењала. Кодови се могу пронаћи у датотеци "enrich_go_analysis.R" и "enrich_pathway_analysis.R".

Сви резултати који су добијени овим методата се могу пронаћи у "files/enrich_analysis" директоријуму, они укључују слике (директоријуми "images") и пропратне ".csv" (директоријуми "csv_files") датотеке са резултатима анализе. При чему директоријуми који садрже суфикс "our", представља резултате које су добијени на нашим модулима, док остали представљају резултате добијене на модулима из оргиналног рада. Улазни модули над којима се покрећу анализе се налазе у "files/enrich_analysis/our_modules.txt" (наши) и "files/enrich_analysis/original_modules.txt" (њихови). Само извршање кода се своди на читање једног по једног модула из датотеке и позивање "generate_heatmaps" у случају "enrichGO" анализе, односно "generate_pathway_analysis" у случају "enrich pathway" при чему шаљемо одогварајуће листе гена.

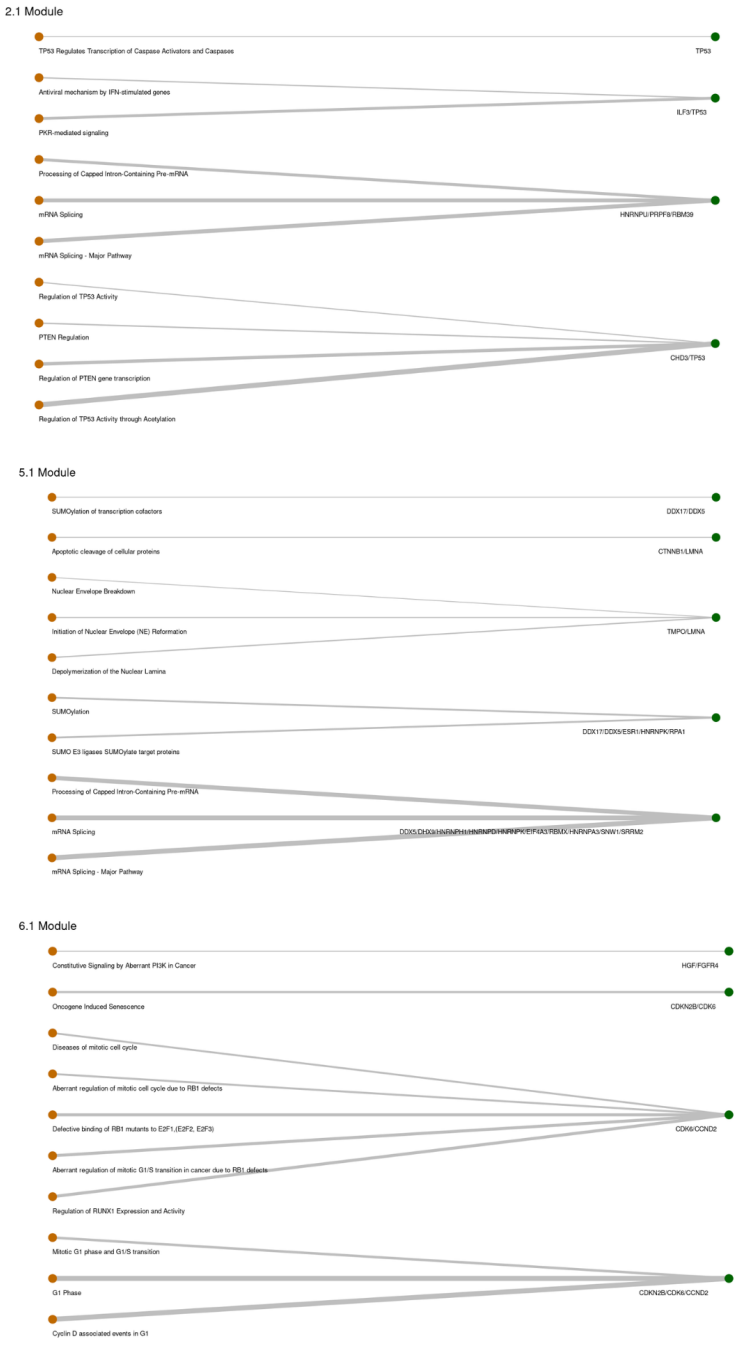
Слика 5: Део резултата који су добијени приликом анализе обогаћивања, са леве стране су дати дијаграми за молекуларне процесе док са десне се налазе дијаграми биолошких процеса. Тамнија нијансе представљају већи ниво значајности. Слике већег квалитета се могу пронаћи у већ поменутом директоријуму.



Слика 6: Део резултата који добијамо када над њиховим модулима покренемо "enrichGO" анализу.



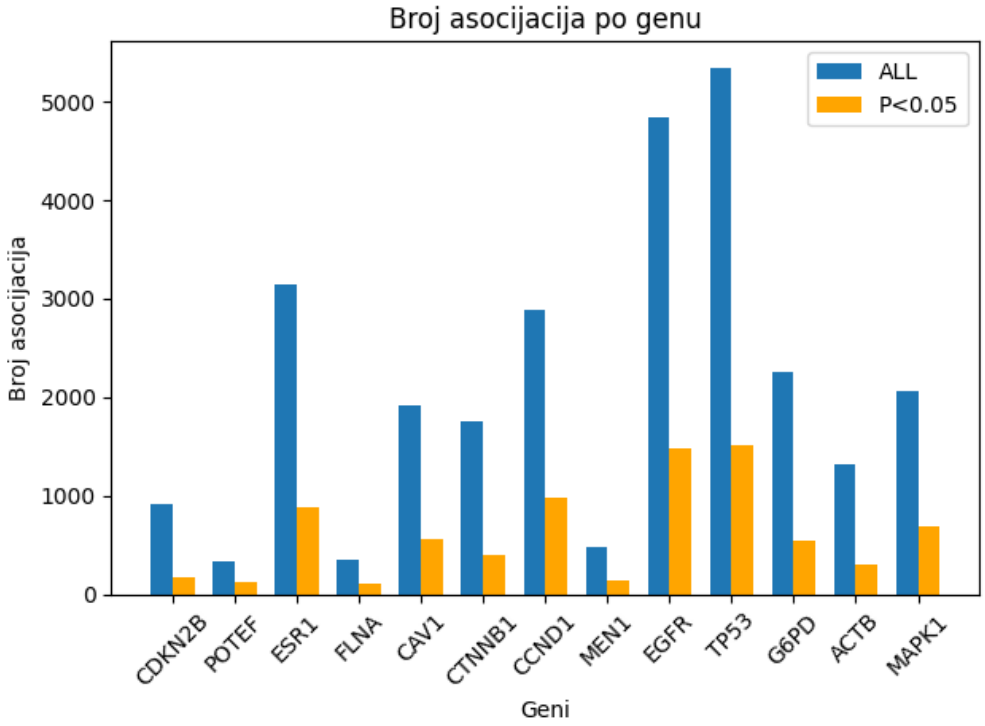
Слика 7: Део резултата који су добијени приликом анализе обогаћивања путања. Дебље ивице представљају већи ниво значајности (односно мању п-вредност) и обрнуто. Чворови на левој страни (обојени наранџасто) представљају функције, док се на десној страни налазе скупови гена/протеина (обојени зелено). Покренуто на нашим модулима.



Слика 8: Део резултата који добијамо када над њиховим модулима покренемо "enrichPathway" анализу.

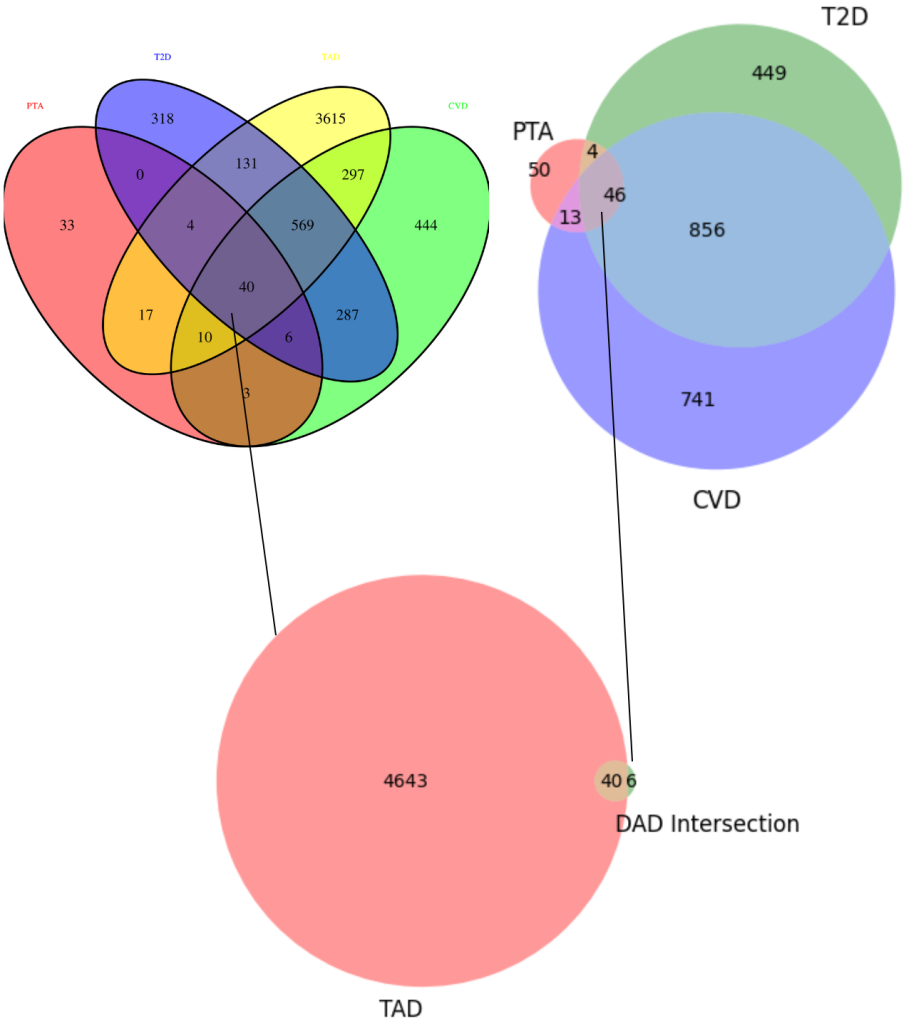
2.5 Интеракције између лекова и циљаних протеина

Податке за овај део анализе преузели смо са "COREMINE"[11] базе података. Прво смо преузели податке за сваки од 13 протеина (њихови централни протеини, слично би могло да се уради и са централним протеинима које смо ми добили) који се налазе у значајним модулима а који су заједнички за све три болести, преузети подаци се могу пронаћи у "files/drugs/gene_drug". Из тих података смо издвојили само оне где чији је ниво значајности испод 0.05. У раду су ови лекови означавати са "TAD" (енг. target associated drugs), тако ћемо их и ми даље називати.



Слика 9: Број лекова повезаних са одрђеним протеином и њихов однос са број оних чија је п-вредност < 0.05

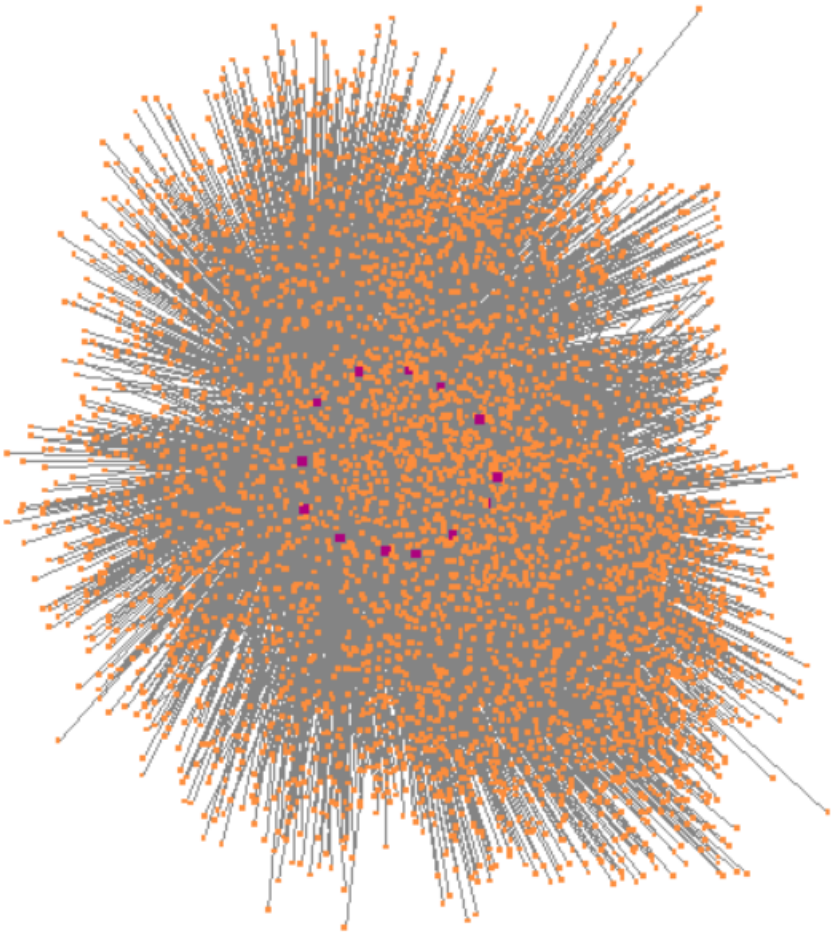
Такође смо са исте базе података преузели лекове за сваку од болести (енг. "DAD"). Ти подаци се могу пронаћи на "files/drugs/disease_drug" и они су нам били улазни подаци за овај део анализе. Код за овај део области дат је у "drug_target_interactions.ipynb" и у "scripts/venn_four_sets.R".
У овом делу смо одрадили цртање Венових дијаграма, да би видели колики су пресеци између скупова лекова за болести и за одређене гене.



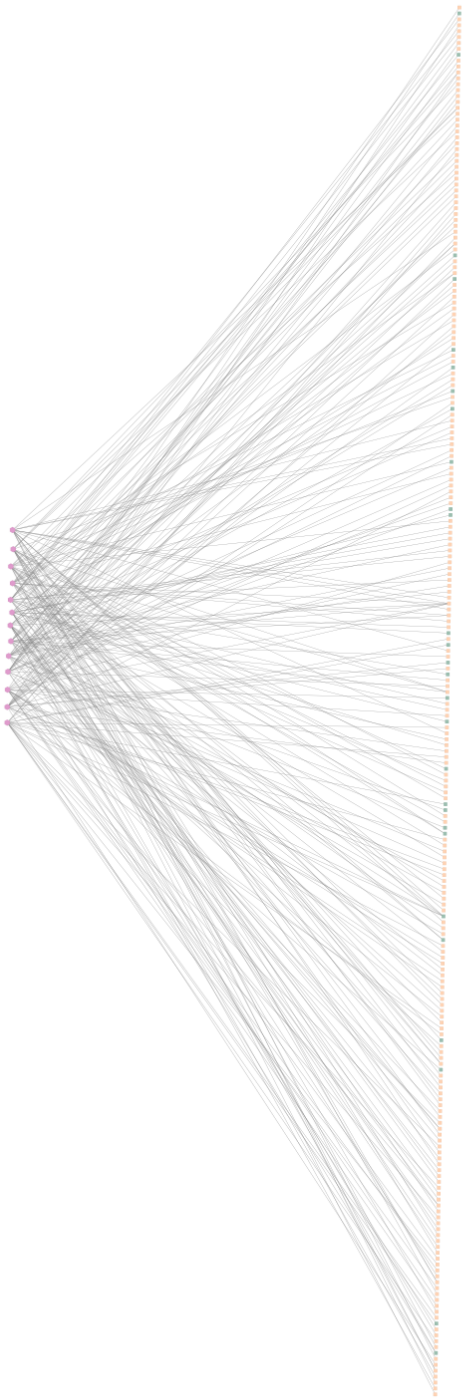
Слика 10: Венови дијаграми лекова

У истим датотекама се налази код који прави бипартитиван граф између гена и између одговарајућих лекова. Приликом истог проласка кроз податке смо за сваки од гена пронашли 20 најзначајних лекова и формирали хијерархијску мрежу, при чему смо посебним бојама бојили лекове који су заједнички за "TAD" и "DAD" (у нашем случају то је скуп од 40 лекова). За приказивање мрежа смо опет користили "Cytoscape" алат.

Резултати формирања ових мрежа могу се пронаћи у "files/sign_drugs_genes_-mapping.sif" и у "files/gene_most_sign_drugs.sif", а приказивање мрежа у "visualization_drugs.cys". Можемо приметити када упоређујемо са резултатима изложеним у раду, да наши одступају и по броју лекова за сваки од скупова а и саме мреже и дијаграми, али да су приближни њиховим.



Слика 11: Сваки од 13 значајних протеина (љубичасти чворови) је мапиран у њему одговарајући значајан лек (наранџасти чворови)



Слика 12: Протеини и њихових 20 најзначајнијих лекова задати у хијеархијској мрежи, зеленом бојом су представљени чворови који припадају "TAD" и "DAD"

3 Закључак

Успешно смо поновили поступак описан у одабраном научном раду, чиме смо потврдили његове налазе и методологију. Наша анализа је показала доследност резултата са изворном студијом, што указује на ваљаност и поузданост коришћених метода. Током овог процеса, стекли смо дубље разумевање предметне теме и методолошких приступа који су примењени.

Иако су наши резултати у великој мери били у складу са оригиналним, уочили смо неколико мањих разлика које би могле бити последица разлике у полазним подацима, као и да се базе које садрже те податке временом мењају. Зато смо често у наставку поступка користили резултате добијене у раду, да би нам за конкретан корак били исти улазни подаци, а самим тим и могућност поређења резултата као и контрола ваљаности поступка.

Кроз детаљну анализу и понављање поступка описаног у оригиналном раду, имајући у виду да нам је ово први семинарски рад оваквог типа, научили смо како се приступа научној литератури, како се идентификују кључни елементи методологије, и како се врши анализа одређеног научног рада.

Литература

[1] Nikhat Imam, Aftab Alam, Mohd Faizan Siddiqui, Akhtar Veg, Sadik Bay, Md. Jawed Ikbal Khan and Romana Ishrat, Network-medicine approach for the identification of genetic association of parathyroid adenoma with cardiovascular disease and type-2 diabetes

[2] "DisGeNet" - платформа која садржи јавно доступну колекцију гена и вараијанте повезане са људским болестима, https://www.disgenet.org/

[3] "BioGRID" - база података биомедицинских интеракција,
https://thebiogrid.org/

[4] "IntAct" - база података и алати за анализу молекуларних интеракција,
https://www.ebi.ac.uk/intact/home

[5] "Cytoscape" - софтвер за приказивање комплексних мрежа,
https://cytoscape.org/

[6] Bader, G.D., Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2 (2003). https://doi.org/10.1186/1471-2105-4-2

[7] "GO-enrichment" - анализа обогаћивања,
https://geneontology.org/docs/go-enrichment-analysis/

[8] "ClusterProfiler" - пакет у програмском језику "R" који служи за спровођење анализе обогаћивања,
https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html

[9] "org.Hs.es.db" - пакет који садржи базу података коју смо користили,
https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html

[10] "ReactomePA" - још један пакет у програмском језику "R" који имплементира анализу обогаћивања,
https://bioconductor.org/packages/release/bioc/html/ReactomePA.html

[11] "COREMINE" - платформа са које смо преузели податке о лековима,
https://www.coremine.com/

[12] "BioGRID REST Service" - веб сервис који омогућава програмерски приступ и преузимање података са "BioGRID" базе података,
https://wiki.thebiogrid.org/doku.php/biogridrest

[13] "BaderLab MCODE" - "Java" имплементација "MCODE" алгоритма,
https://baderlab.org/Software/MCODE

[14] "Uniprot" - база података која садржи пресликавање између различитих "ID"типова протеина,
https://www.uniprot.org

[15] "UniProtKB/Swiss-Prot" - база протеинских секвенци која пружа прецизне и свеобухватне информације о протеинима
https://www.expasy.org/resources/uniprotkb-swiss-prot