



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	2
2	Методологија и резултати	3
2.1	Прикупљање протеина за одговарајуће болести	3
2.2	Конструкција мреже протеина и њене тополошке карактеристике . . .	4
2.3	Проналажење функционалних модула у "PPI" мрежи	8
2.4	Анализа прекомерне репререзентације скупа гена	11
2.5	Интеракције између лекова и циљаних протеина	16
3	Закључак	20

1 Увод

Тема семинарског рада је поновити поступак који је објављен у раду Мрежно-медицински приступ за идентификовање генетских повезаности аденома паратироидне жлезде са кардиоваскуларним 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 Iqbal Khan and Romana Ishrat published in Briefings 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 ("РТА") 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 "РТА" 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 коморбидитета и повећати ефикасност неких лекова.

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

Напомена: За све кодове коју су писани било у "R" или "Python" програмском језику, потребно је подесити радни директоријум на директоријум пројекта.

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 упоређујући са референтним резултатима из већ поменутог рада.

2.1 Прикупљање протеина за одговарајуће болести

Протеини који су повезани са болестима "РТА", "CVD" и "T2D" преузети су са "DisGeNet"[2] базе података, која садржи компилацију 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 фолдеру "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.

Broj gena	Naši podaci	Podaci prezentovani u radu
CVD	1756	1607
T2D	3034	2803
PTA	125	116

Слика 1: Упоредивање броја гена између података које смо ми преузели и оригиналних података

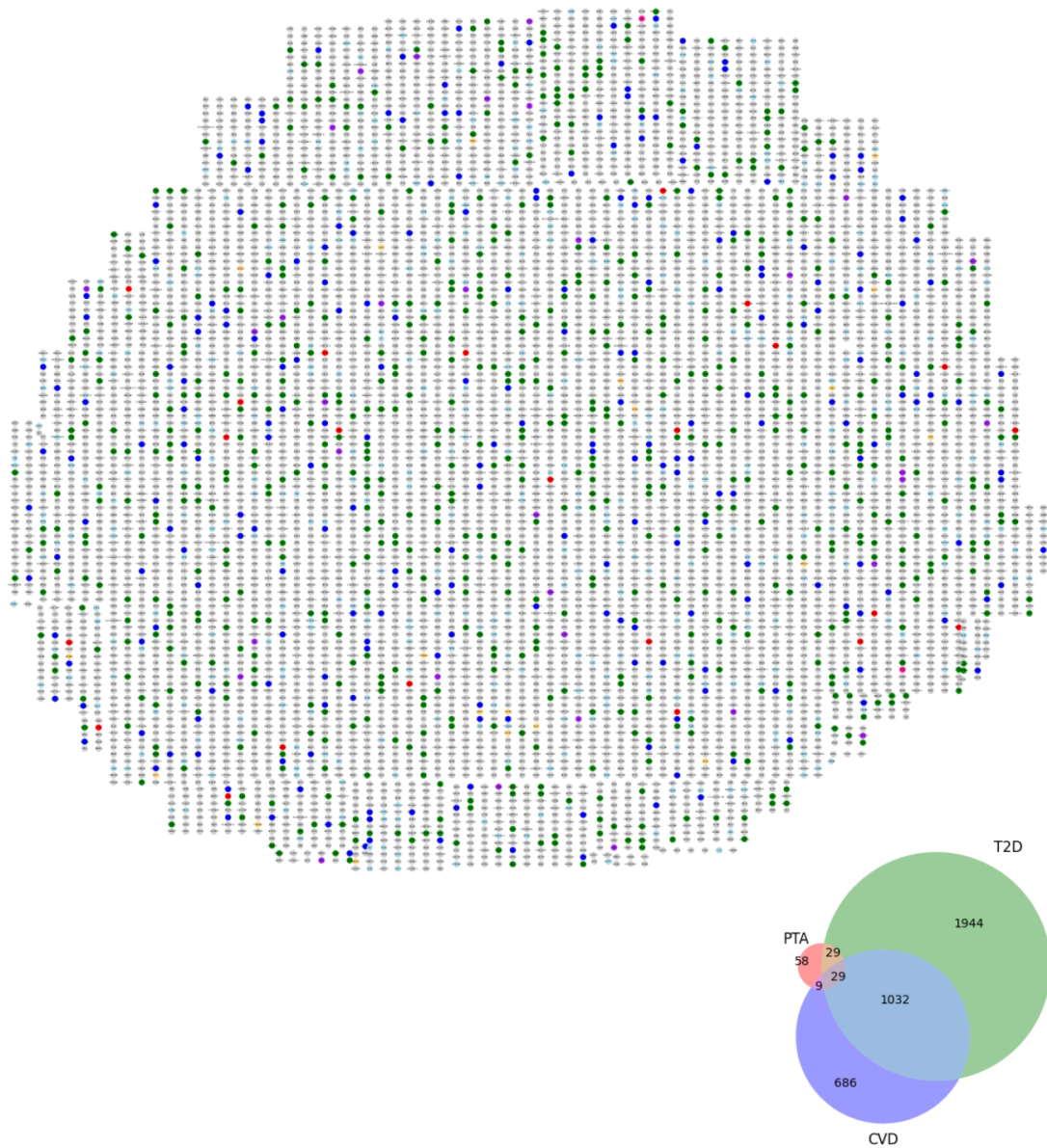
2.2 Конструкција мреже протеина и њене тополошке карактеристике

"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, идентификацију кључних протеина, проучавању механизма болести, итд.

У нашем случају мрежу смо конструисали над протеинима који су заједнички за све три болести (33 протеина) и њихових првих суседа. Скрипта која проналази заједничке протеине дата је у "scripts/common_genes.py", а резултати у "files/common_genes.xlsx". 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.

Ми смо конструисали мрежу користећи само "BioGRID"[3] базу, јер пружа много једноставнији програмерски интерфејс (енг. API), док пешке преузимања података преко веб сервиса не би било оствариво (потребно за сваки од 33 протеина пронаћи суседе, и онда за све те суседе који 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 фолдеру "files/construction_data".

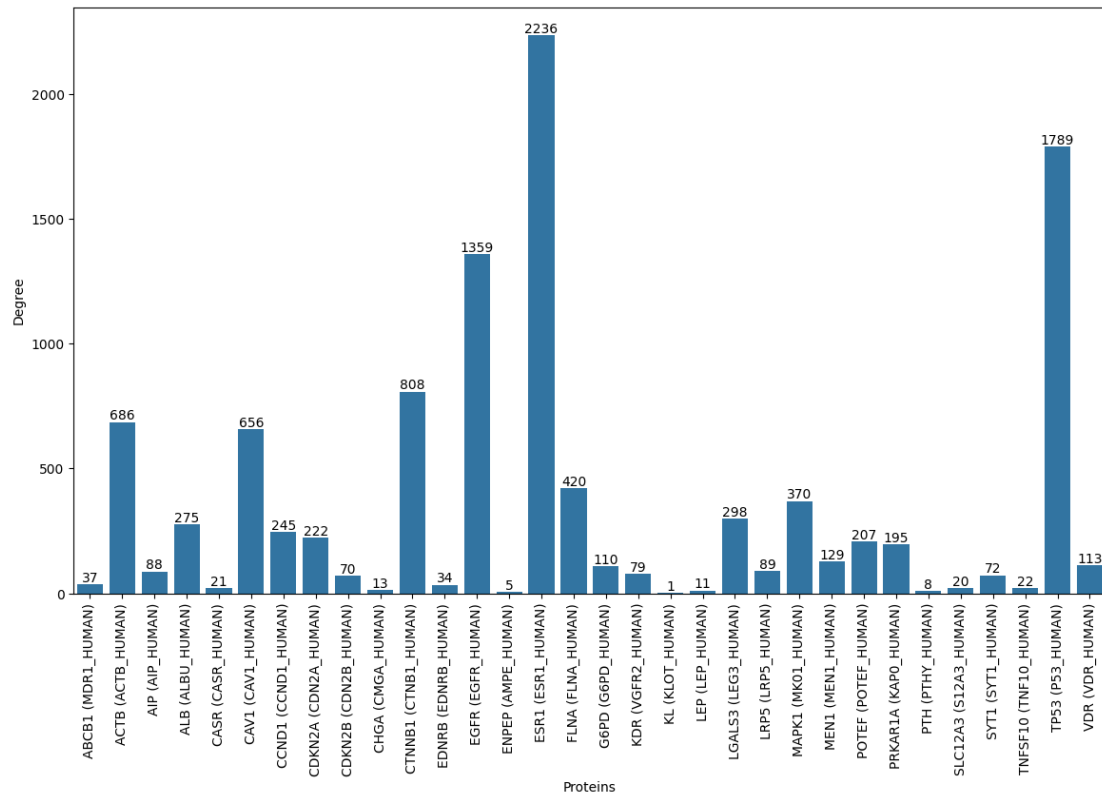
Код се може пронаћи у "network_construction" и "scripts/biogrid_download" (скрипта за преузимање података), излаз кода дат је фајлу "files/network.sif". Пошто је за визуализацију коришћен "Cytoscape"[5] алат, фајл има формат ".sif".



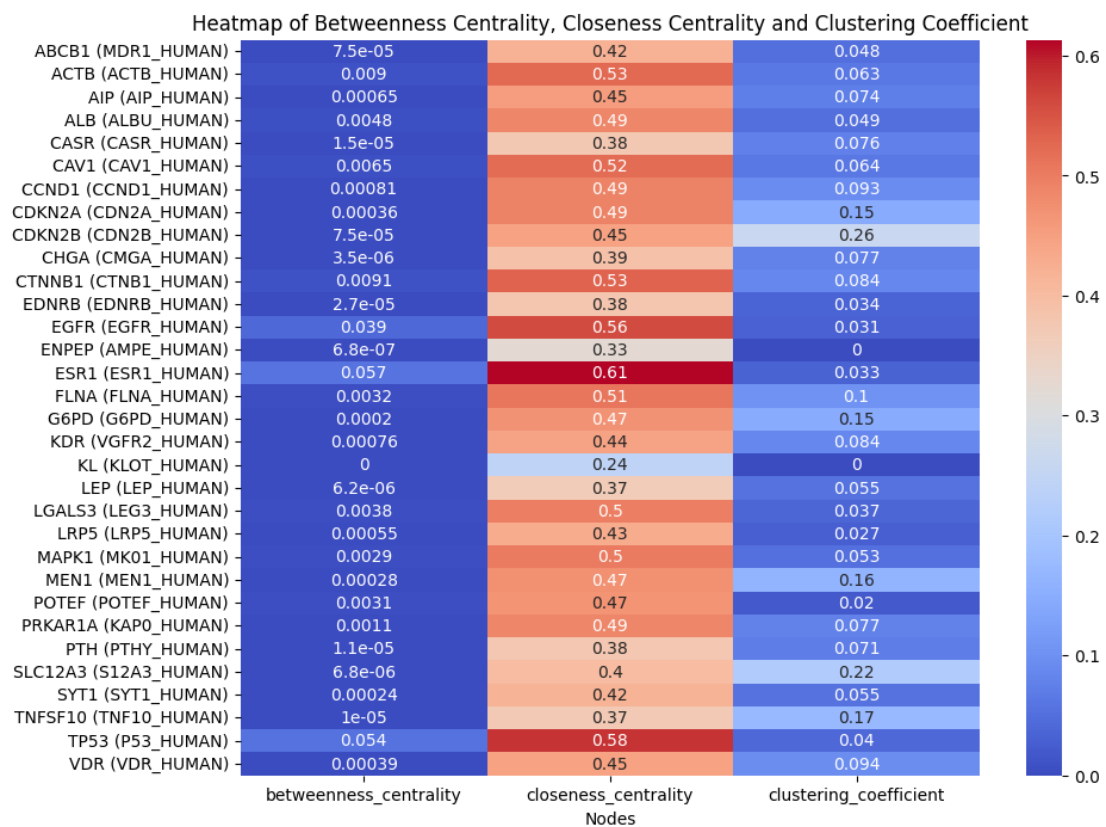
Слика 2: Мрежа интеракција између протеина, као и Венеов дијаграм болести и протеина. Боје у мрежи одговарају бојама скупова на дијаграму.

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 статистике: блискост (енг. closeness centrality), коефицијент кластеровања (енг. clustering coefficient), посредништво (енг. betweenness centrality) и степен чвора (енг. node degree). Код се може пронаћи у фајлу "network_topological_properties.ipynb".



Слика 3: Степен сваког од 33 гена у мрежи



Слика 4: 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), предикције 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 опционално постпроцесирање за филтрирање или додавање протеина у добијене комплексе по одређеним критеријумима повезаности.

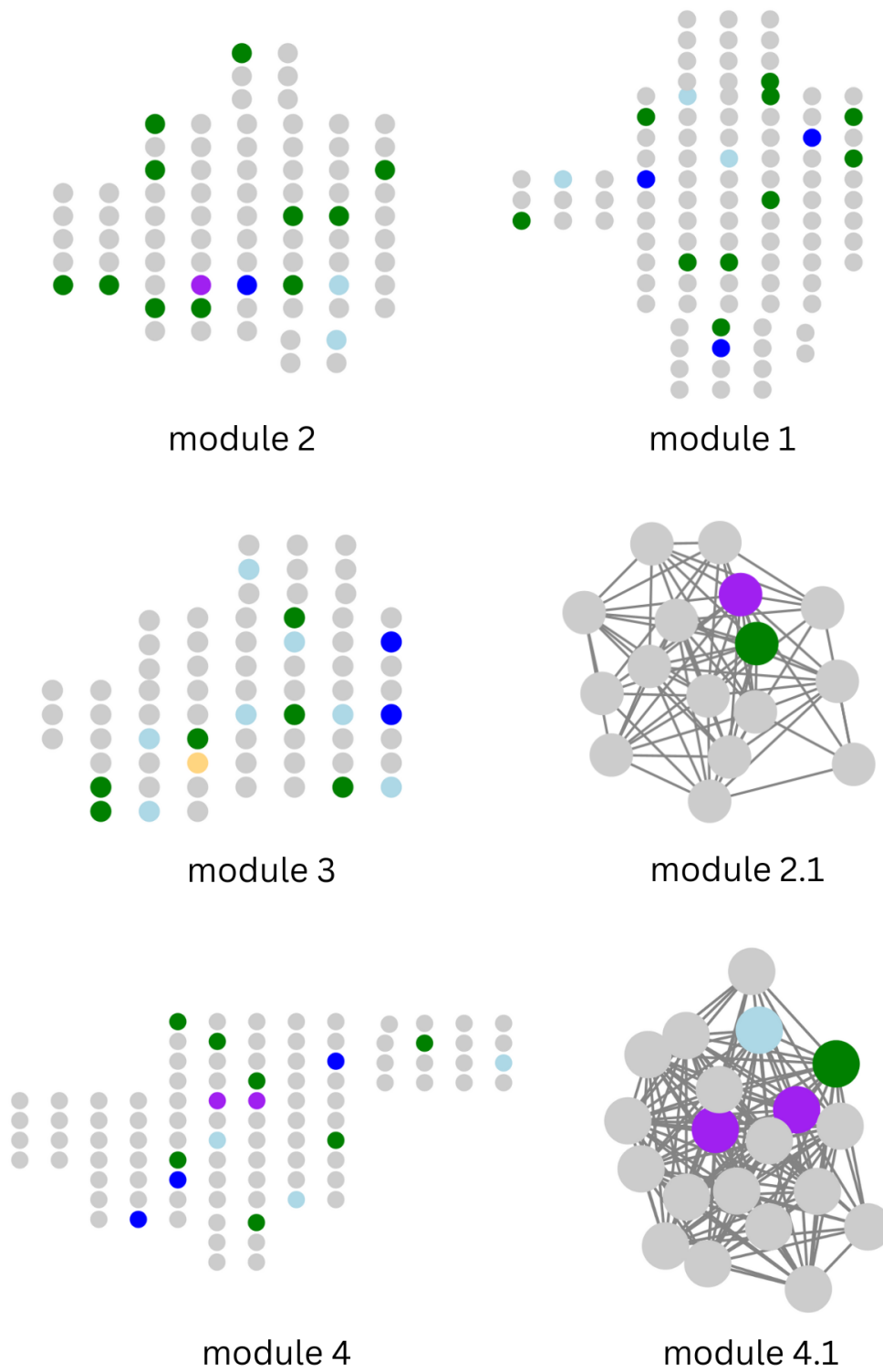
Мрежа интерагујућих молекула може се интуитивно моделирати као граф, где су чворови молекули, а гране молекуларне интеракције. Ако су познати временски путеви или информације о сигнализацији ћелија, могуће је креирати усмерени граф са стрелицама које представљају правац хемијске акције или информационог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

- Прва фаза, понедисрање 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 пунњења (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 оцјењују 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. Код за "MCODE" алгоритам се налази у фајлу "mcode.ipynb".

Прво смо "MCODE" покренули на целој оригиналној мрежи 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" а њихова визуализација у "modules_visualization.cys".



Слика 5: Неки од модула које смо добили из "MCODE"алгоритма

2.4 Анализа прекомерне репрезентације скупа гена

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

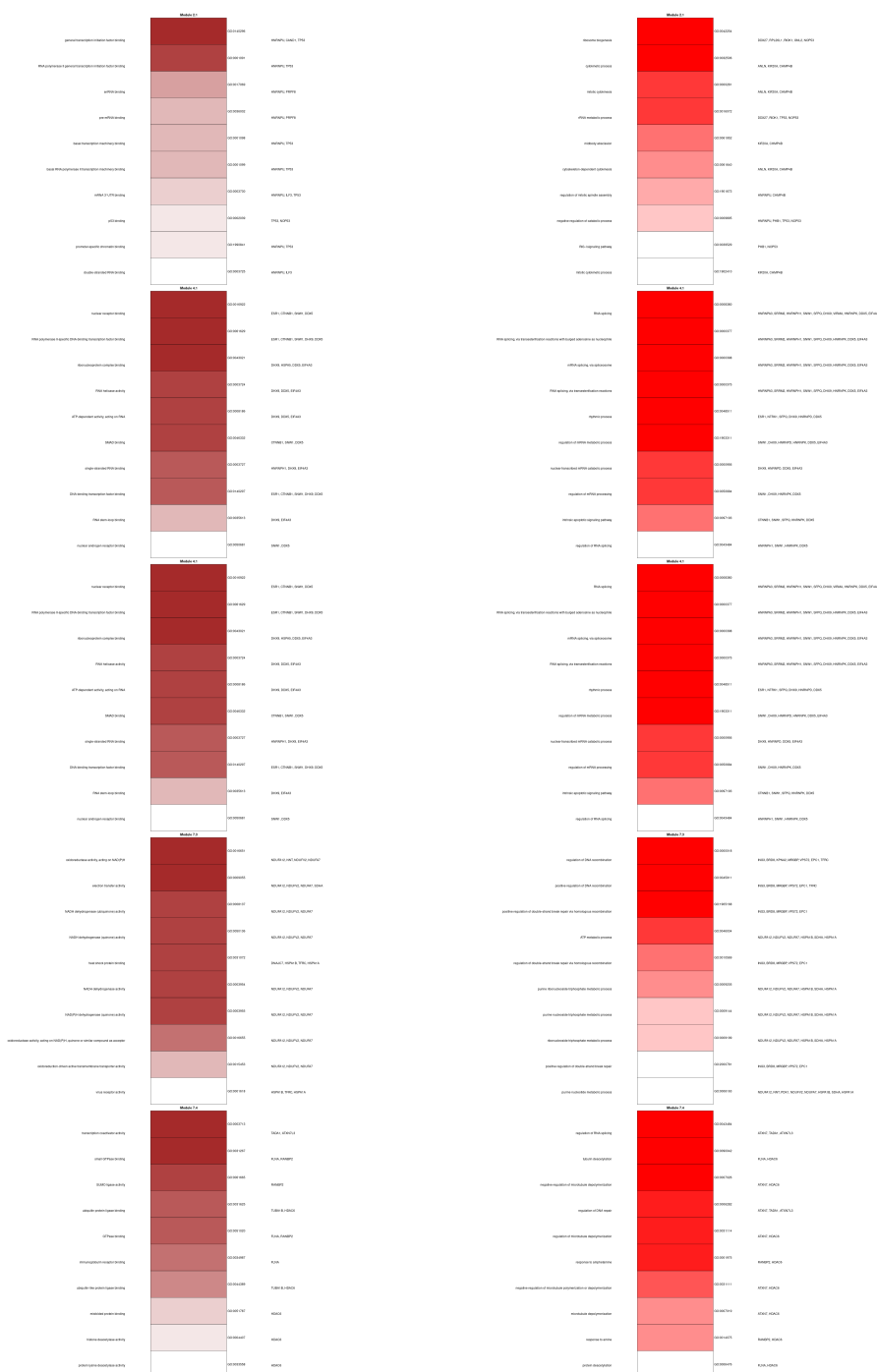
"GO-enrichment"[7] анализа је одрађена коришћењем "R" програмског језика и пакета "ClusterProfiler"[8] и "org.Hs.eg.db"[9]. Користили смо "enrichGO" функцију, при чему је узет "Benjamini-Hochberg" метод за прилагођавање п-вредности. Такође, функцију смо позивали два пута једном за биолошке процесе а други пут за молекуларну функцију (вредност "ont" параметра постављена на "BF" и "MF"). За визуализацију смо користили топлотне мапе.

Анализа обogaћивања путања (енг. enrich pathway) одрађена је исто у "R" језику уз помоћ пакета "ReactomePA"[10] и "ClusterProfiler". И овде смо користили 'Benjamini-Hochberg' метод, а за визуализацију је коришћен бипартитиван граф.

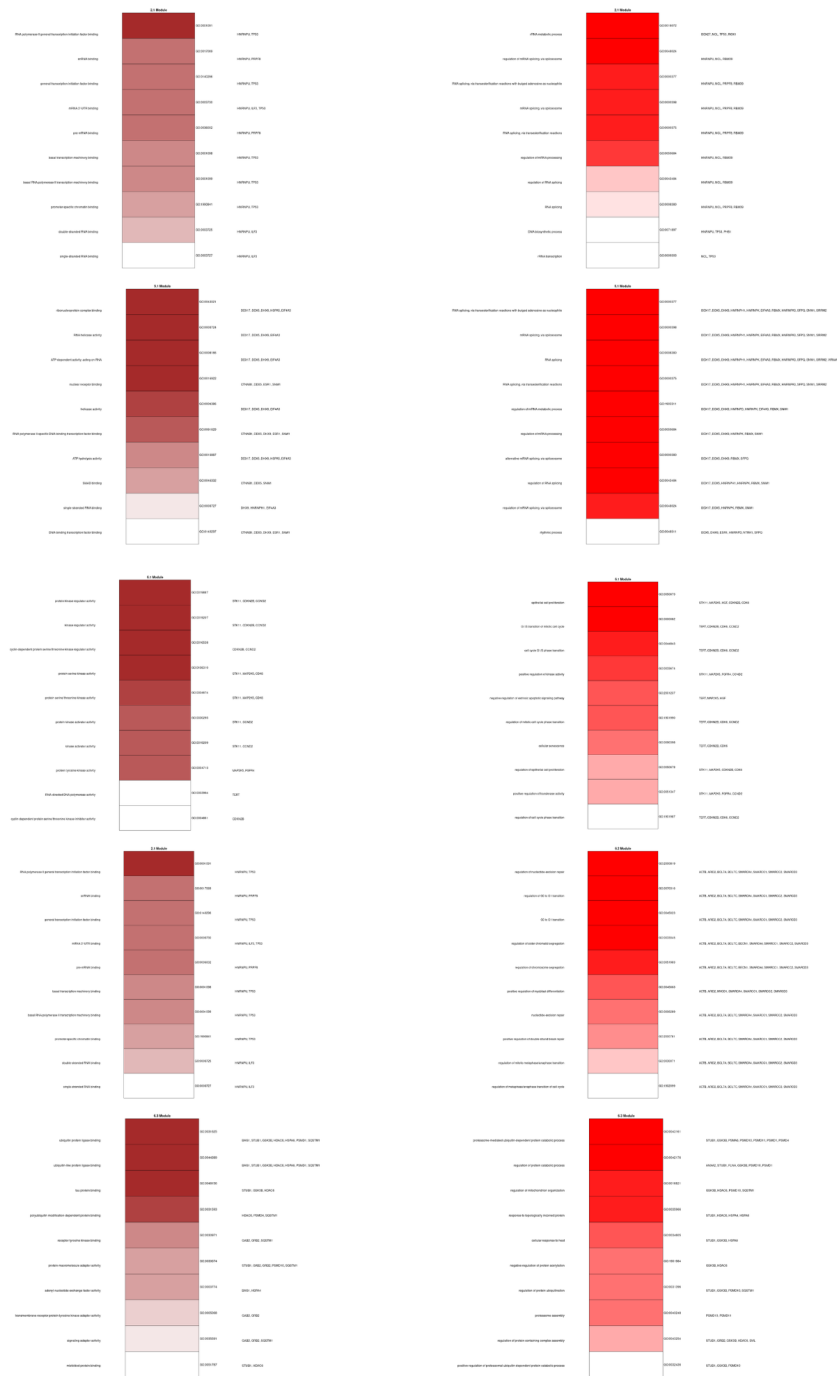
У обе анализе узели смо у разматрање само оне чланове чији је ниво значајности испод 0.1 (п-вредност је постављен као аргумент на 0.1). Минимална и максимална величина (за сваки модул) поставена је на 2-400 протеина. Кодови се могу пронаћи у фајлу "enrich_go_analysis.R" и "enrich_pathway_analysis.R".

Анализа је рађена на основу модула (из рада и 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.

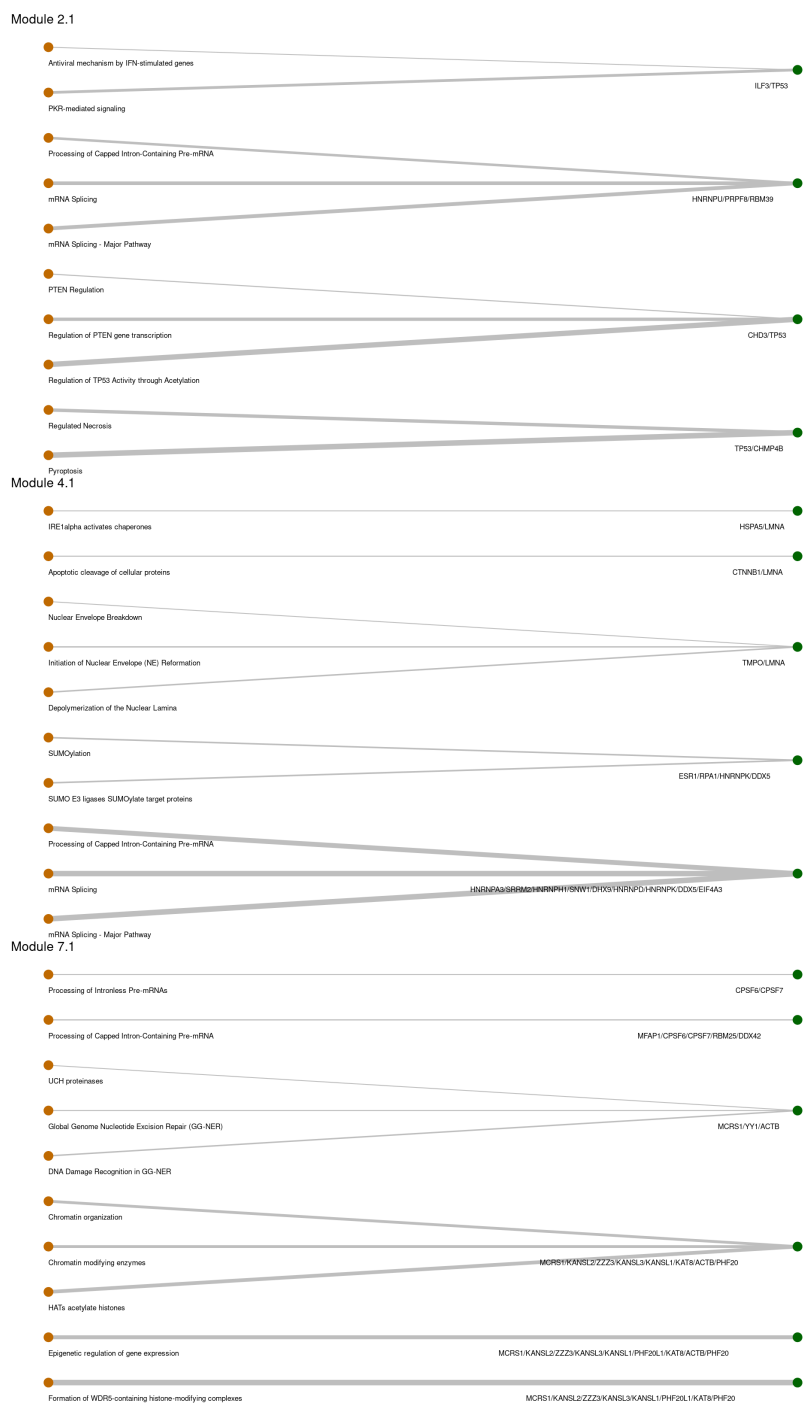
Сви резултати који су добијени овим методата се могу пронаћи у "files/enrich_analysis" директоријуму, при чему директоријуми који садрже суфикс "our", представља резултате које су добијени на нашим модулима, док остали представљају резултате добијене на модулима из оригиналног рада.



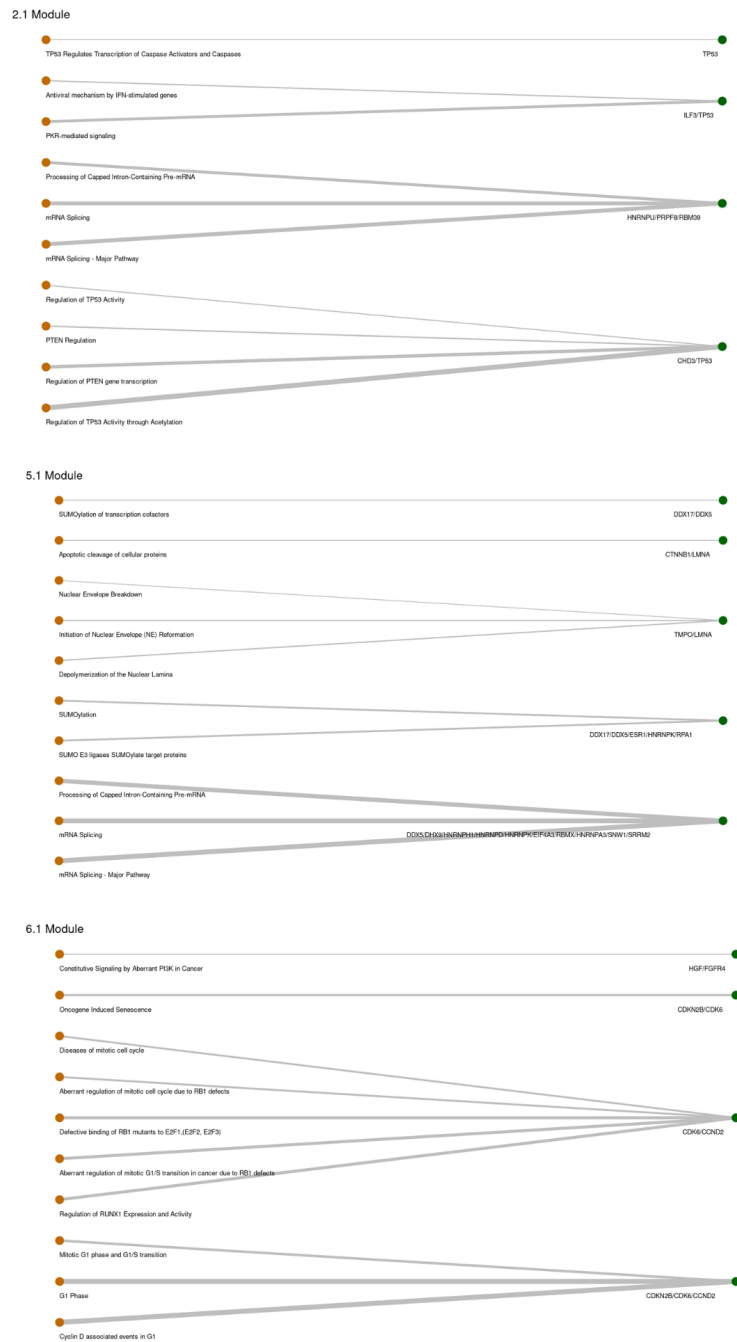
Слика 6: Део резултата који су добијени приликом анализе обogaћивања, са леве стране су дати дијаграми за молекуларне процесе док са десне се налазе дијаграми биолошких процеса. Тамнија нијансе представљају већи ниво значајности. Сlike већег квалитета се могу пронаћи у већ поменутом директоријуму.



Слика 7: Део резултата који добијамо када над њиховим модулима покренемо "enrichGO" анализу.



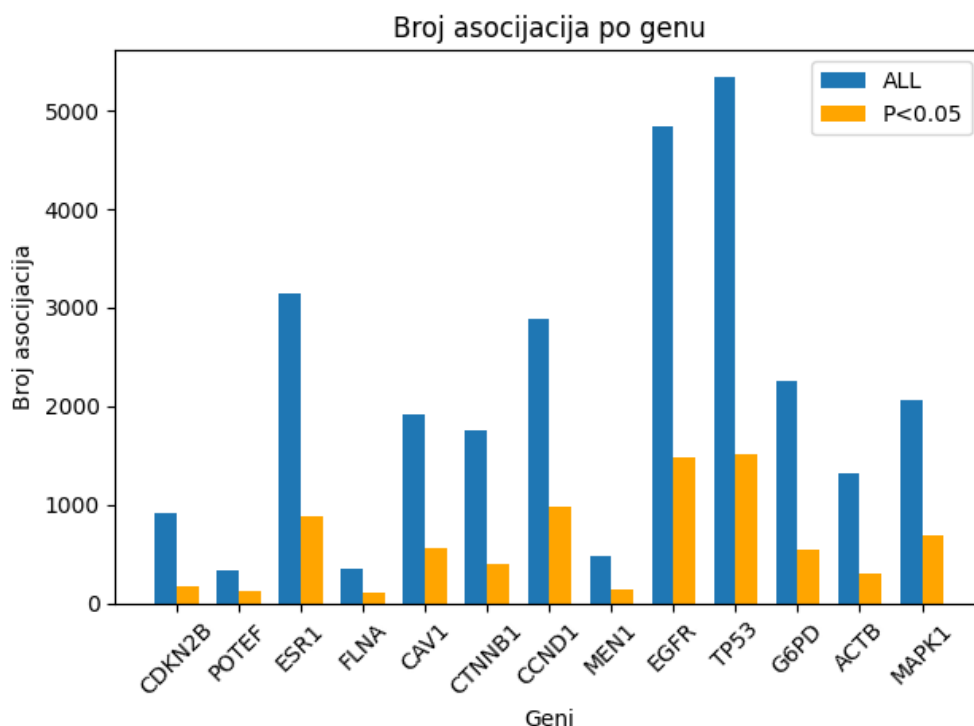
Слика 8: Део резултата који су добијени приликом анализе обогаћивања путања. Дебље ивице представљају већи ниво значајности (односно мању п-вредност) и обрнуто. Чворови на левој страни (обојени наранџасто) представљају функције, док се на десној страни налазе скупови гена/протеина (обојени зелено).



Слика 9: Део резултата који добијамо када над њиховим модулима покренемо "enrichPathway" анализу.

2.5 Интеракције између лекова и циљаних протеина

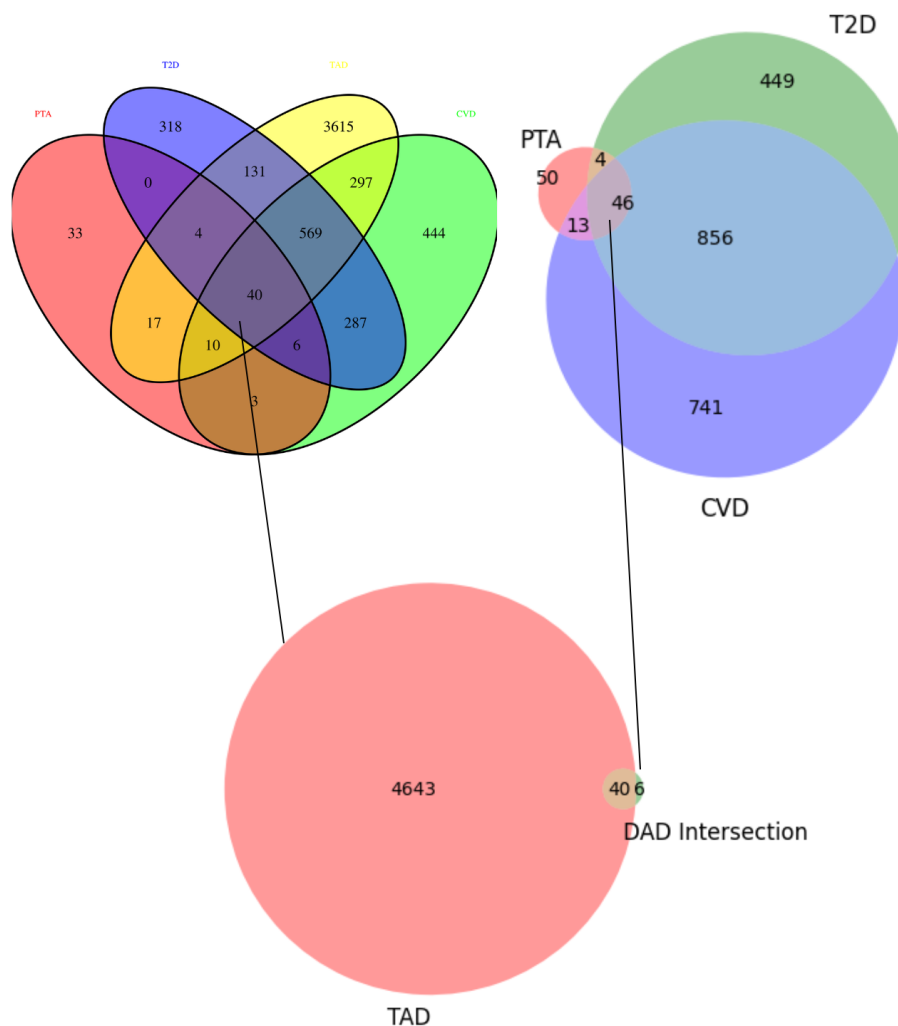
Податке за овај део анализе преузели смо са "COREMINE"[11] базе података. Прво смо преузели податке за сваки од 13 протеина (њихови централни протеини, слично би могло да се уради и са централним протеинима које смо ми добили) који се налазе у значајним модулима а који су заједнички за све три болести, преузети подаци се могу пронаћи у "files/drugs/gene_drug". Из тих података смо исфилтрирали 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 10: Број лекова повезаних са одређеним 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".

Код за овај део области дат је у "drug_target_interactions.ipynb" и у "scripts/venn_four_sets.R" том делу смо одрадили цртање Венових дијаграма, да би видели колики су пресеци између скупова лекова за болести и за одређене гене.

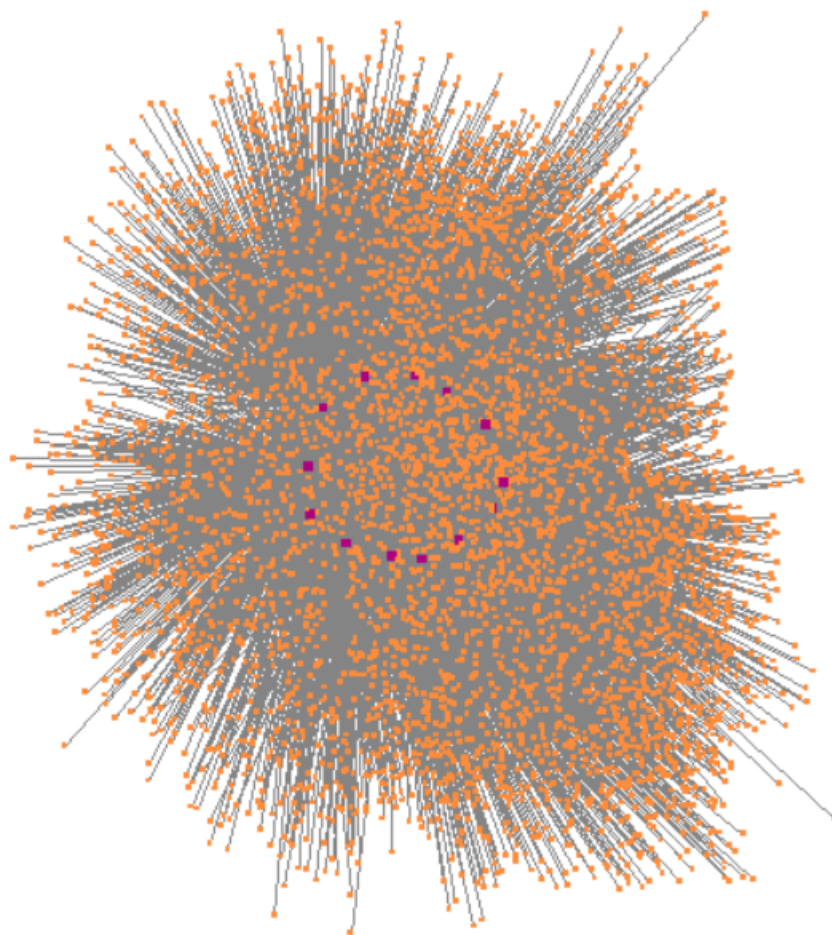


Слика 11: Венеови дијаграми лекова

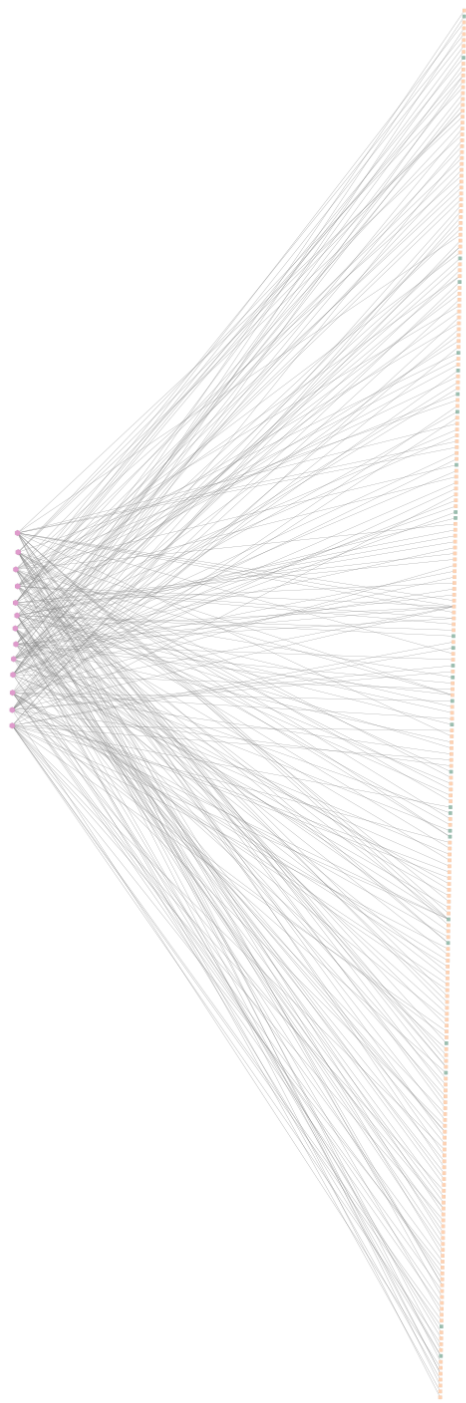
У истим фајловима се налази код који прави бипартитиван граф између гена и између одговарајућих лекова. За визуелизацију мрежа смо опет кориситили "Cytoscape" алат.

Такође 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).

Резултати формирања ових мрежа могу се пронаћи у "files/sign_drugs_genes_mapping.sif" и у "files/gene_most_sign_drugs.sif", а визуелизација 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 12: Сваки од 13 значајних протеина (љубичасти чворови) је мапиран у њему одговарајући значајан лек (наранџасти чворови)



Слика 13: Протеини и њихових 20 најзначајнијих лекова задати у хијерхијској 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 оригиналном 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, идентификовали 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 Ikbāl 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 варијанте 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 визуелизацију 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.es.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/>