Универзитет „Св. Кирил и Методиј“
Факултет за информатички науки и компјутерско инженерство
Институт за интелигентни системи
Скопје

Кире Триводалиев

ФУНКЦИОНАЛНА АНОТАЦИЈА ВО ПРОТЕИНСКИ ИНТЕРАКЦИСКИ МРЕЖИ

– докторска дисертација –

Ментор: проф. д-р Љупчо Коцарев

Скопје, 2014

Комисија за оценa:

проф. д-р Данчо Давчев, претседател
Факултет за информатички науки и компјутерско инженерство - Скопје

проф. д-р Љупчо Коцарев, ментор
Факултет за информатички науки и компјутерско инженерство - Скопје

вон. проф. д-р Ана Мадевска Богданова
Факултет за информатички науки и компјутерско инженерство - Скопје

вон. проф. д-р Андреа Кулаков
Факултет за информатички науки и компјутерско инженерство - Скопје

вон. проф. д-р Слободан Калајџиски
Факултет за информатички науки и компјутерско инженерство - Скопје

Kire Trivodaliev

Functional annotation in protein interaction networks

ABSTRACT: *Complex networks have recently become the focus of research in many fields. Their structure reveals crucial information for the nodes, how they connect and share information. In this thesis protein interaction networks are analyzed as complex networks from two aspects, how to extract knowledge about the functions of an unknown protein from the network in a direct manner and how to utilize the network's functional modular structure to achieve the same goal. Different graph representations for the protein interaction network are proposed, each having different level of complexity and different inclusion of the annotation information within the graph and each of these is explored as to what the benefits and the drawbacks are when it is used in the functional annotation process. The first research direction is based on the hypothesis that the simultaneous activity of sometimes functionally diverse functional agents comprises higher level processes in different regions of the protein interaction network. In line with this a functional neighborhood is defined and constructed by using random walks on the protein interaction graph. Modularity is utilized via clustering in the protein interaction graph with the purpose of obtaining functionally enriched protein groups that can later be used in the functional annotation of an unknown protein. The experiments are performed using a purified and reliable Saccharomyces cerevisiae protein interaction network, which is then used to generate the different graph representations. We evaluate results in regards of biological validity and function prediction performance. Our results indicate that the new complex graph representations improve the prediction process and the proposed algorithms are better or in line with the up to date best algorithms as referenced in the literature.*

Кире Триводалиев

Функционална анотација во протеински интеракциски мрежи

РЕЗИМЕ: *Комплексните мрежи од неодамна преминаа во фокусот на истражувањата во многу области. Нивната структура открива суштински информации за јазлите, како тие се поврзуваат и делат информации. Во рамки на оваа докторска дисертација се анализираат протеинските интеракциски мрежи како комплексни мрежи од два аспекти. Како на директен начин да се извлече знаење од мрежата за функциите на непознат протеин и како да се искористи функционалната модуларна структура на мрежата за истата таа цел. Предложени различни граф репрезентации за протеинската интеракциска мрежа, секоја со различно ниво на комплексност и различно вклучување на информацијата за функциите присутни во интеракциската мрежа и испитани се кои се добрите и лошите страни на секоја од нив кога се користат со различни алгоритми во процесот на функционална анотација. За првата насока на истражување на анотацијата хипотезата е дека симултаната активност на понекогаш функционално дивергентни функционални агенти може да биде дел од некои процеси на повисоко ниво во различни делови од протеинската интеракциска мрежа. За таа цел се дефинира и конструира функционално соседство на протеинот преку користење на случајни изминувања на графот. Модуларноста е искористена преку процесот на кластерирање во рамки на графот со цел да се добијат функционално богати групи кои можат да бидат искористени за анотирање на еден прашален протеин. Експериментите се вршени врз доверливо и обработено множество за интеракциската мрежа на лебен квасец врз основа на кое се добиваат различните репрезентации. Резултатите се евалуираат од аспект на биолошка валидност и перформанси на предвидувањето. Предложените алгоритми се подобри или споредливи со најдобрите алгоритми референцирани во литературата.*

СОДРЖИНА

ЛИСТА НА СЛИКИ

1

ВОВЕД

Ервин Шредингер во неговата книга од 1944 година, Што е животот?, го опишува животот како систем со способност за предавање на ентропијата [1]. Кажано со едноставни зборови, ако системот не е во можност да одржи ред и да ја предаде ентропијата во својата околина, истиот ќе постигне состојба на рамнотежа и ќе угине. Гледано во обратна насока, ако системот може да ја пренесе ентропијата од внатрешноста на системот кон надворешноста, процес со кој ќе ја одржува својата подреденост, истиот ќе продолжи да живее. Ваквиот опис на животот е многу јасен и концизен. Меѓутоа, самите механизми со кои се остваруваат задачите на животот се со огромна комплексност.

Во ваквата сложена „машинерија“ она што ние го перципираме како биолошки живот го има протеинот на своето најниско ниво, односно целокупноста на секое биолошко суштество е овозможена од интеракцијата на „единечните“ функции на

секој негов составен протеин. Протеините претставуваат синџири од аминокиселини кои имаат суштинска улога во огромен број на клеточни механизми. Со скорешните достигнувања во полето на секвенцирачките техники за еден протеин може да се определи точната низа на аминокиселини кои го формираат. Меѓутоа, откривањето на постоењето на еден протеин, или секвенцирањето на аминокиселините на постоечки протеин не е доволно за определување на функцијата на истиот. Токму поради тоа една од најактуелните области на истражување на пресметковната билогија е расветлувањето на протеинските функции. До денес постојат многу техники користени за in silico предвидување на функцијата на еден протеин, со поврзување на неколку извори на податоци вклучувајќи секвенца, структура, експресија на гени, класификација на домеин и фамилија, и протеински интеракции. Истражувањата на функцијата на протеините во последните неколку години укажуваат дека количеството на информации содржано во една мрежа од протеински интеракции далеку ги надминува сите останати извори и користењето на нивната топологија и структура преку соодветни техники води кон најдобри резултати.

Функција на протеин е генерички опис на активноста на протеинот. Функцијата на протеинот може да се движи од опис на неговата молекуларната активност на атомично ниво до објаснување на болестите со кој е асоциран истиот. Со цел да се опише функционалноста на протеините со општа терминологија, различни групи и конзорциуми развиваат контролирани речници т.е. онтологии. Од сите постоечки како стандард се издвојува Gene Ontology (GO) [2] бидејќи претставува најпрецизен и најкомплетен речник на протеински функции. GO има три различни хиерархии на термини (фрази) кои ги опишуваат *молекуларната функција*, *клеточната локација* и *биолошките процеси* за еден протеин. Анотациите за молекуларната функција на протеинот ги специфицираат особеностите на протеинот на компонентно ниво (индивидуални или групи од генски продукти). Биолошките процеси претставуваат серија на настани или молекуларни функции и анотациите за оваа хиерархија ги опишуваат функциите на протеинот на ниво на клетка. Анотациите за клеточната локација ја дефинираат локализацијата на протеинот во подклеточни структури или во макромолекуларни комплекси.

Иако дел од протеините имаат индивидуални задачи, најголемиот дел интерактираат (соработуваат) со други протеини со цел да се организираат и изведат механизмите вклучени во клеточната структура и функција. Една протеинска интеракциска мрежа може да се состои од физички интеракции, комбинација на физички и генетски интеракции, или корелации помеѓу различни профили на експресија на микро-решетки изразени како интеракции. Според последното, протеинска интеракција е широк термин кој опишува било каков тип на асоцијација помеѓу протеини. Протеинските интеракции може да се претстават во форма на граф, во кој јазлите се поистоветуваат со протеините, а врските се однесуваат на интеракциите помеѓу протеините. Вообичаено интеракциите помеѓу протеините се моделираат двонасочно, со што мрежата на протеински интеракции се претставува како ненасочен граф. Јазлите во ваквите мрежи не се поврзуваат случајно [3][4]. Специјални особини, како што се дистрибуција на степен на поврзаност [5][6], ефект на мали светови [7], модуларни подструктури [7][8][9], и асортативно поврзување [3] се само дел од атрибутите кои ги опишуваат мрежите од реалниот свет. Во конструирањето на алгоритми и техники за добивање знаење за функцијата на протеините од протеинските интеракциски мрежи најмногу се користи особината на асортативно поврзување која укажува на фактот дека кај мрежите од реалниот свет сличните ентитети имаат тенденција меѓусебно да се поврзуваат. Ваквата особина важи и за биолошките мрежи [4] во насока на постоење на интеракции помеѓу протеини кои имаат слични или исти функционалности [10] [11] [12] [13] [14] [15] [16] [17] [18] [19]. Дополнително асортативното поврзување кај биолошките мрежи може да набљудува во форма на неповрзани, но слични мрежни фрагменти. Ова е поради постоењето на редундантни елементи во биолошките мрежи кои им овозможуваат да останат робусни и покрај пертурбациите во нивната средина [14].

Иако претставува наједноставен можен модел за протеинските интеракциски мрежи, ненасоченият граф се покажува како богат извор на информации за анализи на високо ниво и ниво на геном [20] [21] [22] [23]. Многубројните истражувања овозможуваат значајни резултати за евалуацијата на врската болест/гени [24], идентификување на значајни протеини [25], и предвидување на протеински интеракции [26][27] и протеински функции [28]. Неодамнешните

напредоци во експерименталните техники дозволуваат детекција на стотици протеински интеракции со еден единствен експеримент [29]. Зачувувањето на информацијата за протеинските интеракции од различни експерименти, и повеќе организми во еден ресурс резултира во јавно достапни бази на податоци за протеински интеракции чиј број и големина постојано се зголемува [30] [31] [32] [33] [34] [35]. Покрај големиот број на недоверливи интеракции во нивната содржина [20], овие бази се богат извор за проучување на функционалноста на протеините во групи, и формулирање на работниот тек на операциите извршени од протеините.

Главната цел на оваа докторска дисертација ќе биде дефинирање и реализација на алгоритми за функционална анотација на протеини во рамки на протеинските интеракциски мрежи. Ќе биде направена евалуација на постоечките податоци за протеинските мрежи од аспект на нивната доверливост и корисност во предвидувањето на функцијата на протеините. Ќе бидат предложени алгоритми и техники кои ќе ги користат информациите за топологијата и структурата на протеинските интеракциски мрежи, како и нивните специфични особености, како и техники за извлекување на дополнително знаење од постоечките функционални онтологии. Ќе биде предложена една генерална рамка на систем за функционална анотација со сите етапи од податоци до евалуација и тестирање. На крај ќе бидат прикажани резултатите од таквиот систем и развиените алгоритми.

2

ПРЕГЛЕД НА ПОДАТОЦИ ЗА ПРОТЕИНСКИ ИНТЕРАКЦИСКИ МРЕЖИ

Протеините никогаш не функционираат самостојно, туку влегуваат во интеракции со други протеини за да ја извршат својата функција. Интеракциите помеѓу нив можат да се класифицираат според различни критериуми, но на највисоко ниво, тие се делат на генетски и физички интеракции. Генетските интеракции се случуваат кога мутациите на еден ген предизвикуваат промена на однесувањето на друг ген, но во суштина, ваквите интеракции не спаѓаат во типичните карактеристики на еден протеом. Физичките интеракции се оние кои се поврзани со процесот во рамки на кој секој протеин ја извршува својата функција [36]. Два протеини остваруваат физичка интеракција ако и двата истовремено влегуваат во следниве три функционални категории: метаболитички и сигнални патеки, морфогенетски патеки и протеински комплекси [33].

Науката која ги третира протеините и нивните структурни и функционални особини во последно време доживува експанзија и интересот за нивно

проучување се повеќе се зголемува. Како причина за ова првенствено треба да се спомне развојот на техниката и пред се биохемијата кои промовираат се пософистицирани техники и методи за анализа на протеомите. Така, се зголемува и брзината на продукција на информации за биомолекулите, како и за асоцијациите меѓу нив. Оттука неопходно е да се воведат стандардни спецификации во вид на добро организирани бази на податоци кои би ги менаџирале сите овие податоци и деталите за нив, би ги категоризирале и би овозможиле лесен и интуитивен пристап до нив, како и нивна анализа.

Солиден е бројот на биохемиски методи за детектирање на протеинските интеракции, а квалитетот на секој од нив се оценува според мерките сензитивност и специфичност. Висока сензитивност значи дека методот открива голем дел од интеракциите кои реално се случуваат, а висока специфичност значи дека голем дел од детектираните интеракции навистина и реално се случуваат. Стандардни биолошки методи за оваа намена се: ко-имунопреципитација, бимолекуларна флуоресцентна комплементација (BiFC), трансфер на енергија со флуоресцентна резонанса (FRET), двохибриден метод кај квасец, affinity electrophoresis, tandem affinity purification (TAP), хемиско вкрстување, Strep-protein interaction experiment (STREP), дуална поларизациска интерферометрија (DPI) и многу други. Во поново време достапни се и методи со висок капацитет на детекција на интеракции кои резултираат со забрзано растење на интеракциските множества. Основни претставници на ваквите методи се: двојно-хибриден метод кај квасец, корелирана експресија на mRNA, протеински микро-решетки (protein microarrays), синтетичка смртност и методи кои детектираат протеински комплекси, како масна спектрометриска анализа на афинитетно исчистени мултипротеински комплекси (mass spectrometric analysis of affinity purified multi-protein complexes). Подетален преглед на овие методи е даден во [29].

2.1 Бази на податоци за протеински интеракции

Денес се достапни повеќе бази на податоци кои ги систематизираат протеинските интеракции добиени по експериментален пат. Генерално, ваквите бази можат да се групираат во две групи: бази кои складираат физички интеракции помеѓу два

протеини, и бази кои складираат податоци за протеински комплекси. Нивната цел не е само да ги обединат податоците за интеракциите добиени со различни методи, туку и оние објавени во научни трудови од областа на молекуларната биологија. Исто така, нивна задача е и да ја проверат нивната веродостојност преку определување кои од нив се потврдени во повеќе експерименти. Покрај тоа, некои од нив вметнуваат и податоци за интеракции предвидени преку некој пресметковен алгоритам. Дел од овие бази имаат соодветен интерфејс за пребарување на протеински интеракции и визуелизација на мрежите. Во продолжение ќе бидат претставени и накратко опишани ваквите најзначајни и најкористени податочни каталози.

Две интеракциски множества кои служат како основа за градење на останатите и чиј квалитет често се испитува преку различни биоинформатички алгоритми, се множествата на Ito et al. [37] и Uetz et al. [38]. Тие со идентичен метод, двохибриден метод кај квасец , но во независни експерименти под различни експериментални услови, добиле две различни интеракциски множества на протеомот на пивскиот квасец и истите имаат многу мало преклопување од околу 16.3% за множеството на Ito односно 13.6% за множеството на Uetz, со тоа што и множествата на протеини кои се застапени во двете интеракциски мрежи се преклопуваат во не толку голем дел од само околу 40% [39]. Тоа е и основниот недостаток на сите методи кои даваат обемни резултати со еден експеримент меѓу кои спаѓа и двохибридниот метод: голем дел од откриените интеракции се погрешни (false-positive) и реално не постојат. Сите останати достапни и систематизирани бази на податоци ги сумираат резултатите добиени од овие две множества како и податоци од многу други експерименти и методи со цел да се добие интеракциска мрежа која ќе биде најдобра слика на реалноста.

Database of Interacting Proteins (DIP) [34] [40] е база чија цел е интеграција на разнообразните експериментални резултати од биохемиските анализи на протеинските интеракции во една единствена, лесно достапна база на податоци. На почетокот базата содржела само информации за меѓусебна интеракција на два протеини, но со тек на време воведени се и податоци за протеински комплекси. Во базата се чуваат податоци за протеините (референца до некоја од базите

SWISSPROT, GENBANK или сл., суперфамилија, организам во кој се среќава, краток опис на неговата функција во клетката итн.), интеракциите меѓу нив (локациите во рамки на аминокиселинската секвенца на протеинот каде се случува поврзувањето, изворот на експериментални податоци и податоци за индивидуалните експерименти, методот кој се користел за одредување на интеракцијата итн.), а каде што е возможно дадени се и податоци за топологијата на молекуларниот комплекс. За интеракциите се чуваат уште и информации за бројот на експерименти со кои е потврдена интеракцијата и број на публикации каде таа е спомната.

Малото преклопување на множествата на интеракции кои се добиваат со различните методи ја доведува во прашање веродостојноста на добиените интеракции и ова е основниот проблем кој DIP се обидува да го реши. Затоа DIP воведува мануелно испитување на квалитетот врз кое ја базира веродостојноста на каталогизираните интеракции. Во моментов DIP содржи 26453 протеин кои припаѓаат на 649 организми и меѓу кои се забележани 76844 интеракции. Најцелосно покриени организми во DIP се: лебниот квасец (Saccharomyces cerevisiae), винската мушичка (Drosophyla melanogaster) и бактеријата Escherichia coli.

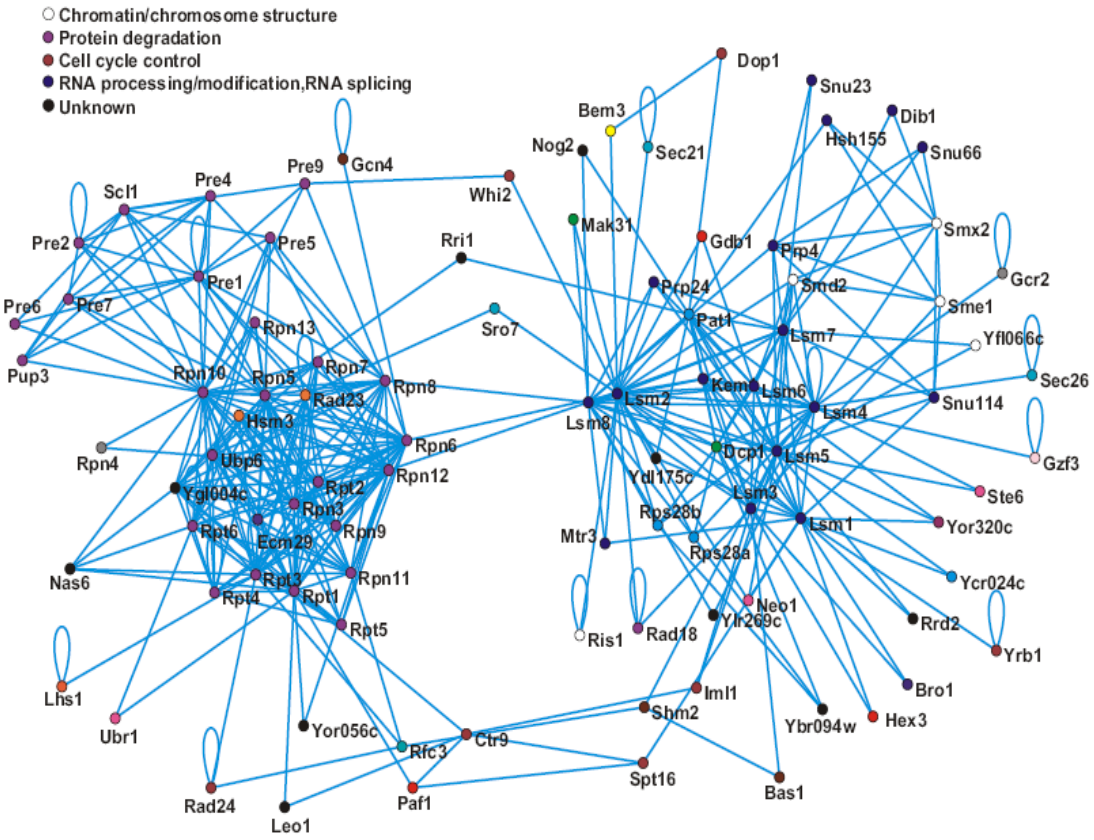
BIND (Biomolecular Interaction Network Database) [30] е уште една спецификација за три основни податочни типови: биомолекуларни интеракции, молекулски комплекси и молекулски патеки. BIND се базира на податоците во NCBI (National Center for Biotechnology Information), па атрибутите кои ги имаат објектите се наследени од таа база. Според авторите на BIND, секоја спецификација на протеинско податочно множество мора да ги исполнува следниве критериуми: да ги опишува сите детали на биолошките податоци, од едноставни бинарни интеракции до големи молекулски комплекси и мрежи на интеракции; да складира податоци за протеини, DNA, RNA и други молекули заедно со деталите и интеракциите меѓу нив; да биде лесно пресметлива и разбирлива, лесно да може да се пристапи до податоците и да биде независна од платформа и база на податоци. BIND користи специјален јазик за спецификација,

ASN.1, кој се користи и од NCBI за опис на нивните податоци во базите GenBank, PubMed и тн.

Од сите објекти кои се чуваат во BIND, најважни се интеракциите и тоа за повеќе организми, за кои, стандардно, се чува текстуален опис на интеракцијата, локација на интеракцијата во рамки на клетката, условите за изведување на експериментот во кој била откриена интеракцијата, положбата на врската во рамки на протеинската секвенца, хемиските акции кои се предизвикани од оваа интеракција и тн. Податоците складирани во BIND се акумулирани од најразлични извори и научни трудови.

Munich Information center for Protein Sequences (MIPS) [32] е истражувачки центар кој креира и одржува разновидни генерички бази на податоци на геноми на систематски начин. Во однос на проблематиката на протеински интеракции, постојат два типа бази на податоци задолжени за складирање на интеракциите: база со физички интеракции помеѓу два протеини и бази за протеински комплекси. Бројот на организмите чиј протеом е зачуван во некоја од MIPS базите е голем. MIPS Comprehensive Yeast Genome Database (CYGD) е база која презентира информации за молекуларната структура и функционалната мрежа на лебниот квасец (Saccharomyces cerevisiae), кој претставува најдетално проучен модел за функционални интеракции меѓу еукариотските организми. Оваа база е надградена со MPact, база на протеински интеракции за протеомот на квасецот која е искористена во бројни анализи на протеинските интеракциски мрежи и која се смета за златен стандард поради квалитетот и опсежноста. MPact е базирана на интеграција на податоци добиени со индивидуални биохемиски експерименти издадени во научни трудови, како и напредни биохемиски методи кои даваат обемни резултати како двохибридната анализата. За да се надмине грешката која ја дава двохибридниот метод, секоја од добиените интеракции засебно се евалуира од страна на експерти пред да се вметне во базата, од каде произлегува и големата доверливост на MPact. Затоа MPact многу често е користена и во системите за предвидување на протеински интеракции и протеински функции.

На Слика 2.1. визуелно е претставена MIPS мрежата на протеински интеракции. Дека оваа база чува информации и за протеинските комплекси во рамки на мрежата, може да се согледа и од сликата: протеините кои припаѓаат на ист функционален домен се обоени со иста боја, а биолошките функции на засебните модули се дадени во легендата.



Слика 2.1 Визуелен поглед на MIPS мрежата на протеински интеракции

General repository for Interaction Datasets (BioGRID) [41] е база на интеракции за протеоми на повеќе организми, во која имињата на интеракциските компоненти се валидирани со базата SGD на геномот на лебниот квасец (Saccharomyces Genome Database). Оттука оваа база при предвидување на протеинските функции се поврзува токму со SGD базата каде се анотирани протеините по функција со GO термини. Базата содржи само најосновни атрибути како учесници во асоцијацијата, експериментален систем со кој таа е откриена и изворот на податокот. Голем дел од информациите се преземени од BIND и MIPS, а некои се преземени од извори кои не се експертно верифицирани, како базите креирани од

Uetz, Ito и др. Интеракциите кои ги каталогизира се генски или физички интеракции меѓу протеини и ги има вкупно 506961 помеѓу 54566 протеини од сите организми. Повеќе од половина од овие интеракции припаѓаат токму на протеомот на лебниот квасец.

2.2 Бази на податоци за анотација на протеини

При предвидување на протеинска функција преку анализа на интеракциски мрежи, многу е важна доверливоста на базата од каде се црпат информациите за функциите на протеините. **Yeast Protein Database (YPD)** [42] е база со опширни податоци за протеомот на лебниот квасец. Таа е прва база која го опфаќа комплетниот протеом на еден организам и дневно се обновува. Во YPD се содржат пресметаните карактеристики на еден протеин како молекуларна тежина и изоелектрични точки, негова локализација во рамки на клетката, негови модификации, опис на познатите функции, мутирани фенотипи, сличности со други протеини и други атрибути за околу 6000 протеини. Податоците во оваа база се базирани врз анализа на генските секвенци, како и детален преглед на научна литература од молекуларна биологија. Најважен податок, секако, е анотацијата на протеинот со определени функции. Моментално базата содржи 6021 запис, секој од нив класифициран според критериумите: на кој ген од шеснаесетте гени на квасецот е продукт, која е неговата функциска категорија, молекуларна средина, локализација во молекулата и тн.

SGD (Saccharomyces Genome Database) [43] е база во која се складирани информации за гените и нивните продукти (протеините), како и анотација на нивните функции, повторно за лебниот квасец. Информациите кои SGD ги обезбедува се со цел да се овозможи наоѓање врски помеѓу генските продукти. Дополнително, SGD ги анотира гените според GO, што претставува структурирана, јасна и општо прифатена репрезентација на биолошкото знаење. Значи SGD објавува информации со кои секој генски продукт на лебниот квасец го поврзува со GO термин во некоја од трите GO хиерархии: функција, биолошки процес во кој учествува и на која клеточна компонента припаѓа. При

анотирањето, SGD се придржува до следниве принципи: анотирањето на гените е врз основа на информации од литературата, асоцираниот GO термин мора да биде на најспецифично можно ниво во GO хиерархијата и на секоја анотација и се придружува цитат од литературата, како и код за евиденција кој овозможува проценка и на нивото на доверливост на асоцијацијата. Генерално, кодот за евиденција укажува на тоа дали асоцијацијата со GO терминот е добиена со анализа на мутирачки фенотипи, сличност на секвенци, физички интеракции, генетски интеракции и тн. Постои и 'unknown' анотација, која се доделува на протеини или генски продукти кои биле анализирани, но за нив не е најдена релевантна функција.

2.3 Gene Ontology (GO) – унификација на термините кои ги опишуваат протеинските функции

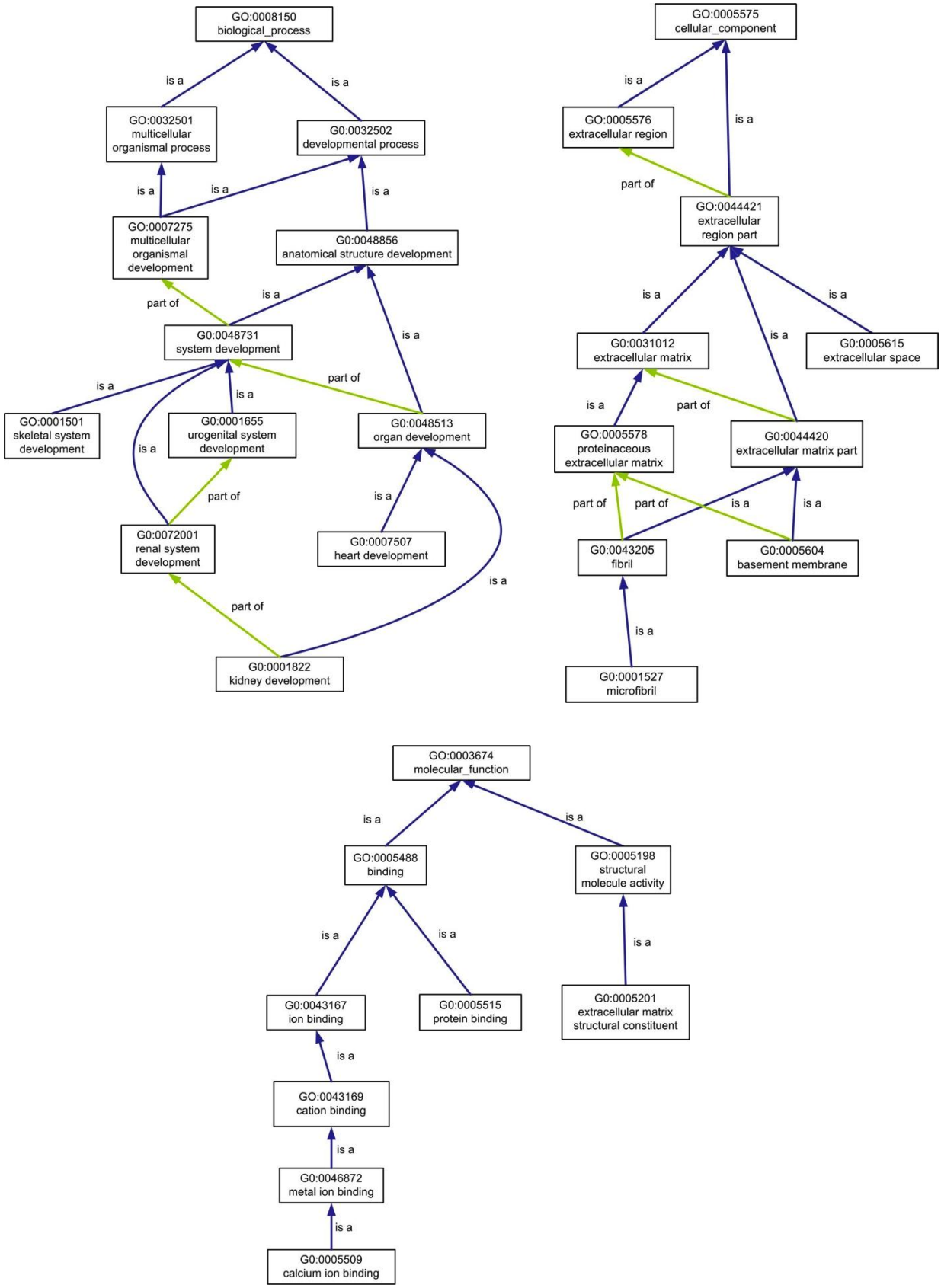
Дефинирањето на унифициран начин за обележување и анотација на генските продукти, од кои од најголем интерес во овој труд се самите протеини, е сериозно прашање од кое зависи успешноста и меѓусебната компатибилност на методите за функционална анотација на протеини. Всушност, целта на еден ваков систем е организирање на знаењето преку усвојување на еден унифициран начин на обележување и именување на генските функции, со кои потоа ќе се анотираат самите генски продукти. Системот за именување треба да обезбеди можност за широка покриеност на што е можно поголем број на функционални феномени (со оглед на големиот број на организми и енормната големина на нивните протеоми), стандардизиран формат (за полесно разбирање од страна на машините), хиерархиска структура (бидејќи можните протеински функции варираат од многу специфични до многу генерални), различни категории за различни нивоа на апстракција, можност за доделување на повеќе функции на еден продукт и динамичка природа и скалабилност (можност за дополнување со нови ознаки) [44]. Од многу предложени шеми за именување на протеинските функции, како Enzyme Classification (EC), EcoCyc, SubtiList, TIGRFAM, KEGG, WIT и други, денес најреферентна е Gene Ontology (GO), која е прифатена како

најгенерална и која именува протеини на голем број организми, а не само на некои, како што е најчесто случај.

GO се состои од три структурирани контролирани речници (онтологии), што овозможуваат опишување на генските продукти на три различни нивоа на апстракција (Слика 2.2). Првата онтологија содржи термини кои означуваат клеточни компоненти (cellular component), како јадро или ендоплазматичен ретикулум или било која друга клеточна органела, или пак дел од непосредната вонклеточна околина. Генските продукти се анотираат со овие термини за да се означи нивна асоцијација или припадност кон некоја од компонентите на клетката. Втората онтологија дефинира термини за означување на биолошките процеси во кои учествува генскиот продукт (biological process). Третата, пак, е речник на термини за молекуларните функции на генските продукти (molecular function). Додека генскиот продукт учествува во низа на настани кои се дел од некој биолошки процес, неговата молекуларна функција се согледува токму во неговата улога кога се случуваат овие настани.

Секој GO термин има единствен нумерички идентификатор во облик: GO:xxxxxxx, по кој следи и биолошкото име на терминот, како и припадноста во една од трите онтологии. GO е дизајниран да биде независен од животинските видови, и вклучува термини карактеристични и за прокариотските и за еукариотските организми.

GO претставува насочен ацикличен граф (directed acyclic graph, DAG) чии јазли се GO терминте. Структурата на GO онтологиите е хиерархиска, т.е. секој термин има свој родител - термин кој на поопшто ниво опишува, на пример, даден биолошки процес (во онтологијата на биолошки процеси), со тоа што еден родител може да има повеќе потомци, но и обратно.



Слика 2.2 Трите онтологии на Gene Ontology

За некој генски продукт може да се знае општиот биолошки процес во кој учествува (класа), а доколку е подетално испитан, може да се знае и специфично во кој подпроцес учествува (подкласа). Базите на протеински анотации како што е на пример SGD се стремат анотацијата на протеините да ја прават со што е можно поспецифични термини. Доколку еден протеин е анотиран со даден термин, се подразбира дека може да е анотиран и со родителите на тој термин.

Со други зборови, односите меѓу GO термините може да бидат:
- *is_a,* однос меѓу даден термин и неговиот родител, како класа и суперкласа.
- *part_of,* означува однос помеѓу два термини, при што доколку првиот термин е дел од вториот, но вториот не мора задолжително да го содржи првиот (на пр. во даден биолошки процес може но и не мора да се случи некој подпроцес, но доколку се случи, тоа ќе е исклучиво во рамки на тој процес).
- *regulates, positively_regulates, negatively_regulates,* означува интеракција помеѓу даден биолошки процес кој на некој начин регулира друг биолошки процес или молекуларна функција.

Мора да се разграничи дека GO е база исклучиво на можните атрибути на генските продукти, но не и на самите продукти. Генските продукти како протеините се сместуваат во други бази, како претходно споменатите YPD и SGD во кои можат да им бидат доделени анотации според различни критериуми. SGD како таква база е најпогодна за протеинска анотација, бидејќи анотацијата ја прави по критериумите: припадност на клеточна компонента, учество во биолошки процес и молекуларна функција, односно користи GO термини за анотација, со што анотирањето е многу јасно и унифицирано.

2.4. Доверливост на базите за протеински интеракции

Еден од најголемите проблеми со кој се соочуваат базите на протеински интеракции е ниското ниво на доверливост на протеински интеракции. Во сите бази на протеински интеракции, во помала или поголема мера, се запишани

интеракции кои реално не се случуваат во протеомот. Според [29], помалку од 1% од сите познати интеракции се потврдени со повеќе од еден метод. Голем број на фактори влијаат нивото на шум во базите на протеински интеракции да е многу високо. На пример, голем број од интеракциите се одредени со помош на високо-пропусните методи за одредување на протеинските интеракции, чија што пак основен недостаток е големиот број на лажно позитивно идентификувани интеракции (false-positive) [29][45]. Често се среќаваат и погрешни интеракции кои произлегуваат од директни експерименти за одредување на поединечни интеракции, бидејќи многу од нив се поставени во базата од страна на компјутерски софтвер кој ги анализира публикациите каде тие се објавени, а ваквиот софтвер е подложен на грешки. Секако, најдоверливи се оние бази на податоци кои се мануелно анализирани и одобрени од човечки експерт. Но, од друга страна, ваквите бази растат многу бавно [33].

Иницијалната задача во определувањето на доверливоста на ППИ е проценатa на покриеноста односно кој е процентот од ППИ податоците што е откриен, и преоценката на прецизноста односно кој е процентот на точни ППИ податоци. Deane et al. [46] и Mrowka at al. [47] прават проценка на ратата на лажно позитивни ППИ податоци со споредување на статистичките параметри на ППИ податоците добиени со високо-перформансни експерименти со оние на високо-квалитетно референтно ППИ податочно множество. И во двата случаи добиената рата на лажно позитивни податоци е значајна. Во [29] направена е евалуација на интерактомите добиени со повеќе високо-пропусни методи со цел да се откријат преклопувања и комплементарности на податоците помеѓу различни бази. Квантитативна мерка за зависноста на ефикасноста на предвидувањето на протеинската функција од базата која се користи е дадена во [48]. Доверливоста на податочните множества се испитува според тоа колкав е коефициентот на корелација за експресија на гени на податоците. Исто така, квалитетот се оценува и според точноста при предвидување на протеинските функции преку интеракциската мрежа. Во [48] споредба се прави помеѓу податочни множества кои содржат физички интеракции меѓу протеините како: DIP, MIPS, и множествата на Uetz и Ito, и множества на протеински комплекси како TAP и HMS-PCI (бази на молекуларни комплекси на пивски квасец добиени со TAP

методот и масовна спектрометриска идентификација на протеински комплекси соодветно) и повторно MIPS базата на комплекси. Добиените резултати се во полза на MIPS множеството на комплекси како најверодостоен модел на протеински интеракции и најдоверлива мрежа за предвидување на протеински функции.

Секундарните методи вклучуваат евристика во определувањето на доверливоста на интеракциите. Ravasz et al. [49] заклучува дека протеините кои интерактираат формираат густо поврзани протеински кластери во ППИ мрежите. Користејќи ја оваа особина, за дадена ППИ мрежа и една ППИ, Goldberg et all [50] го брои бројот на протеини со кои и двата интеракциски партнери имаат врска и истото го определува како мерка за доверливост на интеракциите во мрежата. Saito at al. [51] [52] [53], од друга страна, го користат фактот дека постојат т.н. „лепливи“ протеини кои имаат интеракции со многу други протеини. Меѓутоа, само дел од овие интеракции имаат биолошко значење. Се дефинира мерка за генералност на ППИ како бројот на протеини кои имаат директна интеракција со протеините од една ППИ, така што висока генералност е индикатор за лажно позитивна протеинска интеракција, предизвикана од лепливи протеини. Chen et al. [54] ја подобрува мерката за генералност со додавање на влијанието на далечните соседи на протеините кои се во интеракција. Најпрво се применува генералноста за да се пресметаат тежините на врските во ППИ мрежата, така што подоверливите врски добиваат поголема тежина. Потоа се пресметува производот на тежини помеѓу два протеини, и се нарекува Интеракциска Доверливост преку Алтернативен Пат (ИДАП) вредност, и се избира патот со максимална ИДАП вредност како најсоодветен помеѓу двата протеини. Максималната ИДАП вредност за два протеини ја претставува мерката на доверливост за интеракцијата помеѓу тие протеини. Предноста на ваквиот пристап е што доверливоста може да се пресмета за протеински парови кои немаат директна врска. Како резултат оваа мерка може да се искористи за определување и на лажно позитивни (ниски ИДАП вредности за директно поврзани парови) и на лажно негативни (високи ИДАП вредности за неповрзани парови) протеинско – протеински интеракции. Сличен пристап користи и [55] каде шумот во мрежата се наоѓа преку мерката h-доверливост позајмена од методот за анализа на асоцијации во податочно рударење. Со таа

мерка се одредува веројатноста за постоење на една врска во мрежата. Интересно за овој пристап е што може да ја одреди веројатноста за појавување и на непостоечки врски во оригиналната мрежа, овозможувајќи дополнување на мрежа која е некомплетна, со што се решава уште еден проблем кој обично се јавува кај податоците за протеинска интеракција. Графот ревидиран со овој метод на тестирање на неговата доверливост исто така дава подобра прогноза за протеинската функција, отколку оригиналниот граф.

Податоците користени во рамки на оваа докторска дисертација се подетално објаснети во рамки на поглавјето 5.1.

3

ПРЕГЛЕД НА ПРИСТАПИ ЗА ФУНКЦИОНАЛНА АНОТАЦИЈА НА ПРОТЕИНИ

Првобитните пристапи кон предвидување на функцијата на протеин биле експериментални и вообичаено биле фокусирани на специфичен ген или протеин, или на мало множество од протеини чија природа им наложува да формираат определени групирања, како што се на пример протеинските комплекси. Овие пристапи вклучуваат исклучување на гени, целно насочени мутации и инхибиција на генска експресија. Без да навлегуваме во деталите ќе кажеме дека сите овие пристапи се ниско-пропусни поради огромните ресурси и труд што треба да се вложи, од експериментален и човечки аспект, за да се анализира само еден ген или протеин. Дури и определени иницијативи од поголеми размери за експериментално анотирање се покажале како неадекватни во анотирањето на нетривијален дел од протеините што стануваат достапни како резултат на екстремно брзиот развој на технологијата за секвенционирање на геномот. Сето ова довело до постојано зголемување на разликите помеѓу откриените секвенци и откриените функции за новите, дотогаш непознати протеини. Овој тренд е

мотивот за почетоците на развојот на пресметковни техники за предвидување на функција на протеин, коишто користат различни типови на експериментални високо-пропусни податоци, како што се протеински и геномски секвенци, податоци за генска експресија, филогенетски профили и протеински интеракциски мрежи. Во краткиот период од нешто повеќе од една декада во кој оваа проблематика е актуелна објавени се неколку стотици трудови кои ја третираат истата.

3.1 Класични методи

Класичните методи за функционална анотација на протеини, очекувано, се засноваат на поодамна стекнатите сознанија за протеините. Една од најпроучуваните карактеристики на протеините е неговата структура, која може да биде примарна, секундарна, терциерна или квартерна. Покрај над педесетгодишната традиција на проучување на примарната протеинска структура, не е чудно што пионерските техники во предвидувањето на протеинската функција ги користат токму тие информации. И покрај нивната рудиментираност сепак овие методи имаат огромно историско значење и сеуште широка распространетост и нивниот развој претставува одраз на напредокот во молекуларно-биолошките сознанија и пристапите кон анализата на протеините.

Примарно развиениот начин за функционална анотација на еден протеин е наоѓање на негов хомолог и трансфер на неговите функции и се базира на претпоставката дека протеини кои имаат слични аминокиселински секвенци, имаат и слични функции [56]. Воглавно постојат две стратегии што се користат во овој пристап. Првата стратегија е базирана на глобално и локално порамнување на секвенци [57] [58] [59] [60] [61], а втората на пронаоѓање на мотифи од секвенците [62][63][64]. Најпопуларни методи во рамки на оваа група се оние што ја определуваат функцијата на еден непознат протеин преку барање на слични протеини во јавно достапни бази, при што сличноста на секвенците е базирана на порамнувањето добиено со користење на BLAST [59] или FASTA [65]. Анотациите на протеините со позначајна сличност се користат за предвидување на функцијата на непознатиот протеин. Меѓутоа, се покажало дека протеините

што дивергирале од некој заеднички генски предок можат да имаат иста функција, а да немаат видлива сличност на секвенците [66]. Методите базирани на профили од секвенци како што е PSI-BLAST [67] имаат висока осетливост во детекцијата на далечни хомолози. Во детекцијата на блиски и далечни хомолози се користат и мотифите и шаблоните од секвенци. Овој пристап дава како многу повисока осетливост, така и многу повисока прецизност во однос на методите базирани на порамнување бидејќи за голем број на протеински фамилии се откриени голем број на функционални мотифи и шаблони. И покрај ова, пристапите базирани на сличност на примарна секвенца не секогаш се адекватни за идентификување на функцијата на нови, непознати протеини.

Кога примарните методи не овозможуваат високо ниво на доверливост можат да се применат пресметковни методи кои ја користат тридимензионалната структура на протеините за определување на анотациите на непознат протеин. Ова се должи на фактот дека протеинските 3Д структури се менуваат многу помалку од секвенцата во процесот на еволуција. Најразлични аспекти на структурните податоци, како што се извиткувањето во просторот, обликот на активните места, интеракцијата со лигандите и другите молекули, можат да дадат увид во можните функции на непознат протеин. Следствено, постојат различни категории на методи за функционална анотација базирана на структура. Методи кои ја користат информацијата за извиткувањето се зависни од алгоритми за глобално и локално структурно порамнување [68] [69] [70] [71]. Глобалните и локалните сличности во обликот на протеините укажуваат на функционални сличности и се корисни за определувањето на функциите на непознат протеин. Исто така, развиени се неколку методи за откривање на површински „џебови“ и „шуплини“ во протеинската структура со чија помош се откриваат потенцијалните активни места (или места на врзување) и нивните аминокиселински остатоци [72] [73] [74] [75] [76]. Овој пристап е посебно корисен при предвидувањето на ензимски функции. Детекцијата на слични локални геометрии на функционално значајни остатоци имплицира постоење на слични функции дури и кај далечно поврзани протеини [77]. Достапноста на ко-кристалните структури на протеин-лиганд, протеин-протеин и протеин-ДНК/РНК комплексите овозможува карактеризирање на детални интеракции на атомско ниво. Анализите на овие структури

овозможуваат подобар увид во принципите според кои се одвиваат меѓумолекуларните интеракции значајни за функцијата на протеините и се искористени во нов пристап за функционална анотација на непознати протеини [78]. Со примената на техники за молекуларна динамика и симулации за ориентирањето на молекулите се отвораат нови хоризонти во разбирањето на молекуларното движење и интеракциите вклучени во остварувањето на определена функција. Со ваквите симулации се добиени мноштво податоци за анализа на функциите од аспект на деталните механизми присутни на атомично ниво [79] [80] [81] [82] [83]. Комбинацијата на повеќе различни пристапи базирани на различни структурни карактеристики можат да дадат многу полезна методологија за функционална анотација на непознати протеини. Сепак, недостапноста на структурни податоци со висока резолуција за непознатиот протеин или неговите хомолози сеуште претставува огромно ограничување во оваа методологија.

До неодамна методите базирани на секвенца и структура коишто ја користат хомологијата помеѓу протеините беа доминантни во функционалната анотација на непознати протеини. Меѓутоа, овие методи страдаат од ограничувањата наметнати од недостигот на адекватни информации за хомологни протеини. Дополнително, истите се неуспешни кога не можат да воспостават хомологни врски за целниот непознат протеин [84]. Иако сличноста во секвенцата е во корелација со функционалната сличност, постојат исклучоци и во двете крајности на скалата за сличност [85][86]. Функционалната анотација базирана на структура е со ограничен домен поради достапноста само на ограничен број на структури и извиткувања во базите на податоци. Сите овие фактори допринесуваат кон развој на пристапи за пресметковна функционална анотација коишто освен мерките за сличност можат да користат и дополнителни значајни карактеристики од секвенцата и структурата. Пристапите базирани на машинско учење се покажуваат како посебно корисни во предвидувањето на различни функционални аспекти кај протеините. Главната предност на овие методи е во тоа што истите го пресликуваат проблемот на функционална анотација во проблем на генерирање на класификациски модели [87]. Методите базирани на машинско учење ги употребуваат податоците за секвенцата и/или структурата претставени како

трансформирана и поразбирлива информација во облик на вектор од карактеристики. Се покажува дека со користење на класификатори може да се направи функционална анотација на непознат протеин без користење на информација на хомологија [88]. Слично на ова, се развиваат и сè повеќе методи со машинско учење кои ја користат 3Д структурата како основа за функционалната анотација [89].

Со оглед на тоа што протеините се синтетизираат преку комплексниот процес на транскрипција и транслација на гените кои ги носи ДНК, логично е што нивната функција како и нивните карактеристики на некој начин може да се каже дека се закодирани во тие гени. Со развојот на методите за секвенционирање на гени и распространувањето на базите на податоци кои содржат генски секвенци, како GenBank, NCBI и други, се појавиле и методите за анотација кои својата идеја ја базираат на претпоставки за генските секвенци [90] [91] [92]. Првиот тип на методи е проширување на методот за директна анотација преку препишување на функциите помеѓу два протеини кои се хомологни по структура. Порамнувањето на генски секвенци од геноми на различни организми може да резултира со наоѓање на ортологни гени кои потекнуваат од ист предок, па и покрај тоа што во текот на еволуцијата се разделиле и припаѓаат на различни организми, сепак синтетизираат протеини со исти функции. Втората хипотеза која е основа за развој на ваквите методи е дека протеините чии што соодветни гени се блиску едни до други, влегуваат во меѓусебна интеракција, па следствено имаат и заеднички функции. Третата хипотеза која оправдува дел од методите базирани на генска секвенца е дека постојат гени кои се дел од еден геном и кои во рамки на друг геном се измешани за да креираат нов ген. Притоа, ваквите гени делат заеднички функционалности, па и соодветните протеини најверојатно ќе имаат слични функционални анотации.

Филогенетиката е наука која се занимава со проучување на еволуциските врски помеѓу организмите. Во процесот на специјација, односно еволуирање на организмите од еден вид во друг, протеините присвојуваат или отфрлаат некоја своја карактеристика или функција, но голем дел од нив не се променуваат. Тоа значи дека важни заклучоци за протеинската функција може да се донесат и од

филогенетските податоци. Прв тип на филогенетски податоци се филогенетските профили. Тоа се бинарни вектори за секој ген со должина еднаква на бројот на геноми кои се разгледуваат и секој елемент прима вредност 1 ако дадениот ген има свој хомолог во геномот кој одговара на тој индекс во векторот, или 0 во спротивен случај. Методите кои се базираат на анализа на филогенетските профили претежно вршат споредба меѓу нив. Доколку некои два гена имаат заедничка функција или влегуваат во интеракција, тие исто така би биле наследени во повеќе геноми на различни, но сродни организми [93] [94] [95] [96] [97]. Филогенетските стебла се структура која носи повеќе информации за филогенетската поврзаност бидејќи ги инкорпорира точно информациите за хиерархијата при еволуцијата. Јазлите на филогенетските стебла претставуваат организми и секој организам е поврзан со неговите претходници и следбеници. Најчесто филогенетските стебла се добиваат од филогенетските профили со користење на статистички техники и хиерархиско кластерирање. Методите кои користат филогенетски стебла користат техники на податочно рударење и машинско учење [98] [99]. Постојат и методи за предвидување на протеинска функција кои донесуваат заклучоци со анализа на знаењето здружено од филогенетските профили и филогенетските стебла [100][101].

Експресија на гени е процес при кој со квантитативни мерки се мери првата, транскрипциона, фаза при синтетизирањето на протеин. Имено, во оваа фаза еден ген се транскриптира во mRNA, за потоа во втората фаза азотните бази од секвенцата на mRNA да се препишат во соодветна аминокиселинска секвенца. Експресијата на гени најчесто се мери преку изведување на експериментот на микро-решетки (microarrays), каде на квадратен стаклен чип е впишана матрица со точки, по една точка за секој ген кој се испитува. Точната процедура за изведбата на овој експеримент е надвор од доменот на овој текст, но важно е да се знае дека експресијата на ген претставува нумеричка вредност за тоа колку генот е активен при синтетизирањето на еден протеин во условите во кои се изведува експериментот и таа вредност се добива според интензитетот на боите добиени на чипот за соодветниот ген. Вообичаено експериментот се прави за голем број на гени наеднаш, со што се добива нивната симултана активност и експресија под дадени услови. Бидејќи резултатите на крај се добиваат во матрица,

пресметковната анализа на овие резултати е лесно изводлива. Резултатите од засебните експерименти се чуваат во репозиториуми со јавна достапност. Искористувањето на генската експресија во предвидувањето на протеинската функција се должи на тоа што резултатите даваат кои гени се активни во даден момент при процесот на транскрипција, односно кои од нив во дадените услови (услови каде се симулира некоја болест на пример) симултано синтетизираат протеини. Тоа индицира дека добиените протеини би имале слична функција. Резултатите од генската експресија можат да се кластерираат во кластери на гени со слични профили на експресија [102] [103] [104], да се искористат за моделирање на класификатор од типот на SVM, невронска мрежа или Баесов класификатор [105] [106] [107], или да се анализираат во временски рамки со цел да се доделат функции на протеините [108] [109] [110] [111].

Методите за функционална анотација базирани на рударење на текст се разликуваат од останатите по тоа што не користат податочни структури добиени со анализа на биолошки податоци. Ваквиот пристап подразбира парсирање на текст и извлекување на метаподатоци од публикации во литературата која ја опишува функционалноста на целниот протеин, процеси кои самите по себе се тешки задачи. Тие ги користат огромните репозиториуми на разни публикации, трудови, книги и текстови од областа на молекуларната биологија кои се објавени низ годините. Во сите овие трудови се наоѓаат информации за гени, болести, протеини и сл., од кои со внимателно спроведена анализа можат да се извлечат податоци за протеинската функција. Денес ваквите трудови се складирани на едно место во бази на податоци од кои најтипичен пример е MEDLINE. За обработка на текстовите се користат техники од машинско учење, како извлекување на информации, рударење на текст, обработка на природни јазици, пребарување на клучни зборови и тн [112] [113].

3.2 Методи базирани на протеински интеракциски мрежи

Протеински интеракциски мрежи се огромен извор на податоци за протеинската функција. Ниту еден протеин никогаш својата функција не ја извршува изолирано, туку во рамки на група од протеини со кои соработува за да се случат многу процеси, како на клеточно така и на повисоко биолошко ниво. Функцијата на протеинот и неговата улога во организмот може да биде прецизно дефинирана преку тополошките карактеристики на мрежата на која и припаѓа [114]. Физичката интеракцијата помеѓу протеините се случува со одредена цел, и затоа може да се размислува за одредување на функцијата на даден протеин со анализа на функцијата на протеините со кои влегува во интеракција. Уште еден аргумент во прилог на користењето на мрежи на протеинска функција е тоа што податоците за ваквите мрежи се добиваат директно од експериментални методи. Исто така, секоја мрежа може да се разгледува како граф, и оттука да се развие математички пристап за негова анализа кој лесно може да се имплементира во пресметковен алгоритам. При дизајнирањето на ваквите алгоритми, мора да се внимава дека сепак може да се случи не сите функционални категории на кои припаѓаат партнерите кои влегуваат во интеракција да се преклопуваат, туку само некои, па оттука анализата на оваа комплексна мрежа станува уште поголем предизвик кој сеуште не е најидеално решен.

Постоечките алгоритми за предвидување на протеинска функција од мрежите на протеинска интеракција глобално може да се поделат на две категории: директни методи и методи базирани на кластерирање [28]. Во првата категорија спаѓаат методи базирани на анализа на непосредно или подалечно соседство и методи базирани на анализа на целата мрежа како граф и нејзините карактеристики. Втората категорија се методи кои ги анализираат засебно кластерите од протеини во рамки на мрежата на протеински интеракции како засебни функционални модули. Како посебна категорија можат да се спомнат и методите кои користејќи техники за асоцијативни правила од податочно рударење наоѓаат шаблони кои често се појавуваат во податоците. Овие методи ги учат шемите на анотација во рамки на ППИ мрежата и го анотираат непознатиот протеин преку шемите

пронајдени помеѓу неговите соседи. Техниките како веројатностни суфикс дрва и рударење на корелации [115] и анотација базирана на порамнување на шеми на повторување [116] спаѓаат во оваа категорија.

3.2.1 Директни методи

Наједноставен и најинтуитивен пристап при анализата на мрежите на протеински интеракции би било разгледувањето на непосредните соседи на даден неанотиран протеин. На неанотираниот протеин му се препишуваат функциите кои се најдоминантни меѓу неговите соседи. Недостаток на овој метод е тоа што два соседни протеини можат да припаѓаат во различни функционални категории иако имаат меѓусебна интеракција. На пример, доколку функционалната анотација се прави според GO, можно е два соседни протеини да имаат иста анотација за клеточната компонента на која и припаѓаат, но да не им се поклопуваат молекуларната или биолошката функција. Затоа, овој пристап често дава погрешен заклучок за функционалноста на неанотираниот протеин.

Во [117] за прв пат е предложена идејата за користење на податоците од интеракциските мрежи за анотирање на протеините. Тука најпрво е дадена анализа на интеракциска мрежа од 2709 интеракции помеѓу 2039 протеини од протеомот на лебниот квасец со цел да се утврди која е врската помеѓу функциите и интеракциите меѓу протеините. Заклучено е дека протеините се групираат во своевидни функционални кластери кои и физички во интеракциската мрежа се воглавно групирани во подмрежи, што само докажува дека анализата на интеракциската мрежа може да се покаже како корисна. Во анализираната мрежа, 65% од интеракциите се случуваат помеѓу протеини со барем една заедничка функција, а 78% од интеракциите се случуваат помеѓу протеини кои се лоцирани во иста клеточна компонента или органела. Како метод за предвидување на протеинската функција предложено е броење на појавувањата на функциите кај протеините во непосредно соседство на неанотираниот протеинот, и тој се анотира со трите функции со најголема фреквенција. Заради тоа, овој алгоритам е наречен neighbor counting. Точноста на ваквото предвидување е 72%, при што за

точно анотиран протеин се зема оној за кој е точно предвидена барем една функција.

Сепак, во истиот труд со анализа на мрежата е забележано дека има и голем број на интеракции помеѓу протеини кои не припаѓаат на исти, туку на сродни функционални групи, како на пример протеините кои учествуваат во превиткувањето на протеините (protein folding) и оние кои учествуваат во протеинска транслокација или протеините кои учествуваат во мембранска фузија со оние вклучени во везикуларен транспорт. Исто така, постојат и интеракции помеѓу протеини кои не припаѓаат на исти, туку на соседни клеточни компоненти, на пример голем број на интеракции се случуваат помеѓу протеините во јадрото и цитоплазмата на клетката. Тоа е основната причина за не толку големата прецизност на алгоритамот и се согледува потребата од друг алгоритам кој ќе ја земе во предвид структурата на целата мрежа.

Обид да се надмине овој проблем со разгледување на целата мрежа наместо само директните соседи на некој протеин е даден во [118]. Таму се разгледуваат функциите на протеините што се директно поврзани со даденият, но и оние на растојание од максимум *n* од него, т.е. се разгледува *n*-соседството на протеинот и се бројат појавувањата на секоја функција. Потоа за секоја од функциите се пресметува нејзината важност со помош на χ^2 тест, при што секоја функција добива оценка во зависност од фреквенцијата на појавување, но и очекувањето таа да се појави во *n*-соседството врз основа на фреквенцијата на нејзино појавување во целата мрежа. Таа оценка се пресметува со формулата (3.1), при што $n_i(j)$ е бројот на протеини кои се во интеракција со протеинот *i* и ја поседуваат функцијата *j,* а $e_i(j)$ е очекувањето колку пати таа функција би требало да се појави во *n*-соседството на протеинот.

$$S_i(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)} \tag{3.1}$$

Најдобри резултати и точност од 72.7%, 63.3% и 52.7% при предвидување на клеточна локализација, молекуларна функција односно биохемиска функција соодветно се добива за *n = 1* или *n = 2*. Недостаток на овој пристап е тоа што еднакво ги третира сите протеини во *n*-соседството на неанотираниот протеинот,

иако со оддалечување од него веројатноста за функционално преклопување со далечниот сосед се намалува. Во понатамошниот текст овој алгоритам ќе биде референциран како χ^2 neighborhood.

Овој проблем е адресиран во [119], каде најпрво е направена анализа на повеќе податочни множества за да се докаже два протеини кои индиректно влегуваат во интеракција често имаат иста функција (само 2% од протеините имаат заедничка функција исклучиво со протеини од нивното најблиско соседство). Потоа се моделира тежинска функција со која се доделуваат различни тежини на соседите од прво и второ ниво според претпоставената функционална сличност која се проценува со формула изведена според локалната топологија на интеракциската мрежа. Како и во претходните примери, повторно се пресметува оценка за секоја функција која ја имаат соседите од прво и второ ниво на неанотираниот протеин и оваа оценка зависи од фактичката и очекуваната фреквенција на појавување на функцијата во тоа соседство, но сега оваа фреквенција се множи и со тежинскиот фактор на протеинот. Точноста која се постигнува со овој алгоритам достигнува до над 90% за минимален одзив. Меѓу другото, оваа студија докажува уште и дека најголема веројатност за функционална сличност имаат протеините кои се соседи и на прво и на второ ниво.

Еден од основните недостатоци на алгоритмите базирани на анализа на соседството на неанотираниот протеинот, е погрешното решавање на проблемот во случај кога даден протеин нема доволен број на партнери, или има партнери кои не се анотирани. Конзистентно и веродостојно решение исто така не е можно доколку соседите на протеинот се комплетно различно анотирани. Често информацијата за функционалните карактеристики на протеинот е скриена во целиот граф. Затоа постојат неколку пристапи кои се обидуваат да го опишат целиот граф со една глобална функција која математички ќе изразува некоја глобална карактеристика на графот [36].

Во [120] предложен е алгоритам за глобална оптимизација кој функционалните класи на протеинот ги доделува со минимизирање на бројот на интеракции во мрежата помеѓу протеини од различни функционални категории. За секој од

некласифицираните протеини во мрежата се доделува оценка за секоја конфигурација на функции со која може да се анотира тој протеин. Таа оценка зависи од бројот на интеракциски партнери на протеинот кои ја имаат дадената функциска конфигурација. Математички кажано, целта е за протеинот *i* да се минимизира вредноста *E* дадена со формулата (3.2), каде што *J* е матрицата на соседство за мрежата, $\delta(\sigma_i, \sigma_j)$ добива вредност 1 ако протеините *i* и *j* имаат иста функција, а 0 во спротивно, а $h_i(\sigma_i)$ е бројот на сите соседи на *i* анотирани со функцијата σ_i.

$$E = -\sum_{i,j} J_{i,j}\delta(\sigma_i, \sigma_j) - \sum_i h_i(\sigma_i) \tag{3.2}$$

Во предвид не се земаат само соседните анотирани протеини, туку и соседните некласифицирани протеини кои потенцијално би ја имале таа функција во зависност од нивните соседи. Оттука овој проблем станува проблем на глобална оптимизација, чие алгоритамско решение има голема комплексност. Понекогаш оптимални можат да бидат повеќе функционални конфигурации. За решавање на проблемот се користи техниката на симулирано калење (simulated annealing). Овој алгоритам достигнува точност и до 94% за протеините кои имаат над 7 интеракциски партнери.

Во [121] претходно опишаниот пристап е пресметковно оптимизиран така што во една итерација се пресметуваат оценките за повеќе функции. Освен подобрување на времето на извршување кај овој модифициран и побрз пристап за глобална оптимизација (MFGO), подобрувањата во прецизноста не се забележителни.

Во методот опишан во [122] се конструира еден вид на Хопфилдова невронска мрежа за секоја функција која се разгледува. Улога на влезни неврони имаат анотираните протеини, и ним им се доделува вредност +1 доколку ја имаат функцијата која во моментот се разгледува, а -1 во обратен случај. Неанотираните протеини иницијално имаат вредност 0. Мрежата на протеински интеракции која се разгледува е тежинска, а тежините на врските се добиени од информации од генска експресија. Ако со s_j се бележи вредноста на даден протеин при анотација со некоја функција, тогаш на неанотираниот протеин му се доделува вредност s_i така што да се минимизира сумата дадена со формулата (3.3), при што *U* е

множеството на соседи на протеинот *I,* а w_{ij} е тежината на врската помеѓу двата протеини. Постигната е точност од 93.6% прецизност и 63.7% одзив.

$$E = -\frac{1}{2}\sum_{i=1}^{n}\sum_{s_j \in U} w_{ij} s_i s_j \tag{3.3}$$

Многу значајна придобивка на полето на предвидување на протеинска функција е системот за функционален тек (Functional Flow) предложен во [123]. Во овој труд се прави аналогија на мрежата на протеински интеракции со мрежа на цевки низ кои тече воден тек. Секој анотиран јазел од мрежата може да се замисли како извор на т.н. функционален тек за функцијата со која е анотиран, а неанотиран јазел би претставувал „одлив“. Секоја „цевка“ односно врска во мрежата има одреден капацитет т.е. тежина на врската. Тежините на врските во мрежата на протеински интеракции се доделуваат според бројот и веродостојноста на биохемиски експерименти кои го потврдуваат постоењето на интеракцијата. Симулирањето на ширењето на функционалниот тек низ мрежата трае неколку итерации, токму онолку колку да се опфати блиското соседство на изворот, бидејќи колку се оди подалеку од изворот во мрежата, толку е помало неговото влијание. Неанотираниот протеин се анотира со дадената функција во зависност од оценката која ја добива според количеството на функционален тек кој ќе стигне до него земајќи ги во предвид топологијата на мрежата и капацитетот и бројот на врските преку кои е поврзан со изворот. За разлика од другите пристапи, овде во предвид се зема не само најблиското соседство на изворот, туку и подалечната околина до одредена граница, и притоа влијанието на анотираниот протеин градуирано опаѓа со оддалечување од него, сосема потпирајќи се на топологијата на мрежата. Затоа овој пристап дава резултати многу подобри од претходните методи. Покрај тоа, трудот [123] е уште значаен и поради компаративната евалуација на повеќе познати методи.

Баесов пристап при одредување на протеинската функција со помош на Маркови случајни полиња (Markov Random Field, MRF) е предложен најпрво во [124]. MRF е техника која дава веројатносен модел за симулирање на меѓусебното влијание на случајните променливи преку систем на соседство [28]. Ова може да се прилагоди за проблемот на анотација ако за случајна променлива се земе протеинот, а неговата состојба се неговите сигурни функционални анотации. За секој

неанотиран протеин се одредува колкава е веројатноста тој да има некоја функција. Тоа пак зависи од веројатноста два протеини да се во интеракција, а таа веројатност е голема ако тие делат една функција, помала ако и двата не се анотирани со таа функција, а најмала ако едниот од нив е анотиран со таа функција, а другиот не е. Нека со X е означен векторот со должина еднаква на бројот на протеини во мрежата и прима вредност 1 доколку еден протеин е анотиран со функцијата која е од интерес, и 0 во спротивно. Најпрво се одредува априори веројатноста за X да има одредена конфигурација од нули и единици во зависност од мрежата на протеински интеракции, и е дадена со формулите (3.4) и (3.5).

$$\Pr(X \mid \theta) = \frac{\exp(-U(x))}{\sum_{x} \exp(-U(x))} \tag{3.4}$$

$$U(x) = -\alpha N_1 - \beta N_{10} - \gamma N_{11} - N_{00} \tag{3.5}$$

Во дадените формули, со $U(x)$ е обележана т.н. потенцијална функција, а целата равенка означува дека случајната променлива има Гибсова распределба. $\sum_{x} \exp(-U(x))$ е нормализирачка константа која се добива со сумирање по сите можни конфигурации, а $\theta = (\alpha, \beta, \gamma)$ е множеството на параметри кои се проценуваат со методот на проценка на максимална веродостојност (maximum likelihood estimation) во зависност од бројот на интеракциски парови од протеини анотирани со иста функција во мрежата. Веројатноста протеинот i да има некоја функција $\Pr(X_i = 1 \mid X_{[-i]}, \theta)$, се пресметува со равенка од која може да се заклучи дека тоа зависи од анотацијата на сите останати протеини $X_{[-i]}$, вклучувајќи ги и оние на кои на почеток не им се знае функцијата. Тоа би значело дека за да се анотира целото множество на неанотирани протеини, треба итеративно за секој протеин да се пресмета веројатноста $\Pr(X_i = 1 \mid X_{[-i]}, \theta)$, сè додека не се добијат анотации за секој протеин и конфигурацијата на анотации не конвергира. На почеток, анотациите на неанотираните протеини се земаат да бидат некои случајни вредности. Резултатите кои се добиваат со овој пристап се супериорни во однос на сите претходно опишани методи.

Концептот на MRF е искористен и во [125] каде се претпоставува дека во дадена мрежа на влијанија, состојбата на една случајна променлива се зема да биде независна од сите други случајни променливи ако се дадени состојбите на нејзините директни соседи. Алгоритамот исто така се базира на хипотезата дека бројот на соседи на еден протеин анотирани со дадена функција е случајна променлива со биномна распределба. Исто така, овој труд применува и пропагација на веродостојноста, со цел при анотација на даден протеин во предвид да се земат и неговите неанотирани протеини кои во меѓувреме добиле своја анотација.

Методот опишан во [115], пак, го воведува поимот анотациска секвенца како секвенца на функционални категории кои редоследно се појавуваат во неколку еден по друг прстенесто поврзани протеини. Потоа се формира Марков модел за анотациски секвенци со различна должина. Неанотираниот протеин се наоѓа на опашката на анотирана секвенца од протеини, па тој протеин се анотира со последната функција на најверојатната анотациска секвенца.

Конечно, во [126] е предложен алгоритам кој мрежата на протеински интеракции ја надополнува со тежини на врските кои зависат од бројот на заеднички соседи што ги имаат двата протеини кои влегуваат во интеракција. Оваа мерка за сличност се нарекува коефициент на тополошко преклопување. На секој неанотиран протеин му се доделуваат функциите кои ги имаат неговите најслични соседи, што значи дека овој метод е аналоген на методот за броење на најчесто појавуваните функции кај соседите [117], со тоа што е подобрен со користење на тежински граф.

3.2.2 Методи базирани на кластерирање

Во протеинските интеракциски мрежи постојат региони кои се карактеризираат со тоа што протеините во рамки на еден модул се густо поврзани помеѓу себе, а многу ретко се поврзани со протеините од остатокот од мрежата. Тоа е и главниот мотив за постоењето на алгоритми чија цел е наоѓање на ваквите протеински

кластери. Тие претпоставуваат дека овие региони претставуваат функционални модули, во чии рамки се согледува како протеините се групираат за извршување на некоја конкретна биолошка функција. Според [9], постојат два типа на молекуларни модули: протеински комплекси (групи на протеини кои влегуваат во меѓусебна интеракција во исто време и на исто место формирајќи мултимолекуларна структура која извршува некоја функција) и функционални модули (протеини кои учествуваат во некоја клеточна функција, но меѓусебно се поврзуваат во различна фаза од клеточниот циклус). Сумарно, кластерирањето на протеинските мрежи е корисно за разјаснување на мрежната структура и релациите помеѓу нејзините компоненти, одредување на кардиналните функции на кластер од повеќе протеини, и конечно, преку нив, одредување на функциите на неанотирани членови на кластерот [127].

Диференцирањето на молекуларните модули се прави преку методи за кластерирање на графови, каде мора да се дефинира мерка на сличност помеѓу протеините. Голем дел од методите успешно ги диференцираат овие кластери, но поретки се публикациите кои одат еден чекор понапред и вака добиените информации ги користат за предвидување на функциите на неанотирани протеини во рамки на кластерот. Најчесто тие имаат за цел само да ја докажат модуларноста на мрежата или фокусирана подмрежа на протеински интеракции, вршат само биолошка евалуација во кој клеточен процес учествуваат протеините од кластерот или им доделуваат генерални функционални класи на целите кластери. На пример, [9] идентификува протеински комплекси и функционални модули кои ги категоризира во следниве функционални категории: регулација на транскрипција, контрола на клеточен циклус, процесирање на RNA и транспорт на протеини. [8] врши кластерирање на мала подмрежа од протеини кои учествуваат во сигнални процеси, а на анотираните кластери им доделува генерална анотација во која сигнална патека учествуваат.

Дали кластерирањето ќе се смета за успешно може да се оцени на повеќе начини. Според еден од критериумите, кластерирањето се оценува како добро доколку добиените региони имаат густа поврзаност во рамки на кластерот, а низок степен на поврзаност со протеините од остатокот од мрежата [128]. Исто така, постојат и

алатки кои обезбедуваат граф кој може да се смета за репер и за кој кластерите се однапред познати. Потоа, тој се кластерира со помош на алгоритмите за кластерирање, по што се споредуваат вистинските и добиените кластери. Еден ваков тест е понуден во [129], каде за репер се креира граф кој по својата структура ја пресликува хетерогеноста на степенот на јазлите и големината на кластерите кај реалните мрежи. Дел од алгоритмите [8] [9] [128] се евалуираат преку нивната способност да ги реконструираат експериментално биолошки добиени протеински комплекси или функционални модули, за кои се достапни каталози во некои од базите на протеински интеракции, како на пример MIPS. За потребите на алгоритмите за функционална анотација на протеини, најкорисно би било кластерирањето да се оценува според тоа колку секој од добиените кластери е функционално хомоген.

Сосема различно прашање е како од добиените кластери да се одредат функциите на секој протеин во рамки на кластерот. Оваа втора фаза е слична кај сите алгоритми кои проблемот на анотација го решаваат со кластерирање на графови, а разликата меѓу нив е првата фаза, односно начинот на кој го вршат кластерирањето [28]. Наједноставен начин за одредување на функциите на неанотиран протеин во рамки на модул е тој да се анотира со оние функции кои ги делат мнозинството анотирани протеини во модулот. Алтернативен начин е за секоја функција да се пресмета вредност на збогатување во рамки на еден модул. Доколку таа вредност е под некоја прагова вредност, тогаш на секој неанотиран протеин во модулот му се доделува соодветната функција [28].

Ако како критериум се земат иницијалните податоци кои се користат при одредување на модулите, тогаш методите базирани на кластерирање можат да се поделат во две поголеми класи: методи кои ги користат исклучиво карактеристиките на мрежната топологија со цел да ја декомпонираат мрежата во подмрежи, и методи кои користат и дополнителни информации, како податоци од генска експресија [28]. Според начинот на работа на алгоритамот, кластерирањето може да биде базирано на растојание или базирано на граф. Кај првиот тип на алгоритми се користат класични техники за ненадгледувано кластерирање, при

што треба да се дефинира функција на растојание помеѓу протеините, додека кај вториот тип во предвид се зема самата топологија на графот [127].

Во [8] извршено е кластерирање на мрежа од протеини кои учествуваат во некоја сигнална патека, за потоа да може да утврди кои од нив точно на која сигнална патека припаѓаат. Кластерирањето се врши со хиерархиско агломеративно кластерирање врз матрица на асоцијации помеѓу протеини која е формирана врз основа на најкратката патека помеѓу нив.

Системот MCODE (Molecular Complex Detection Algorithm), имплементиран во [130] има задача да детектира густо поврзани региони во големи протеински интеракциски мрежи, за кои постои веројатност да претставуваат молекуларни комплекси. Алгоритамот поминува низ три фази. Правата фаза е одредување на тежина на јазел во графот која се базира на коефициентот на кластерирање на соседството на јазелот како мерка за густината на поврзаност на јазлите во графот. Притоа, како тежина се зема коефициентот на кластерирање на најгусто поврзаниот подграф од соседството на протеинот. Во вториот чекор се избира иницијално јазелот со максимална тежина и рекурзивно се изминува сега веќе тежинскиот граф нанадвор, и се креира молекуларен комплекс од оние јазли кои ќе се изминат и кои имаат тежина поголема од некој праг. Кога веќе нема јазли за додавање во првиот предвиден молекуларен комплекс, се зема јазелот со најголемата тежина од неизминатите јазли и се повторува постапката. Со менување на праговата вредност може да се добијат помали или поголеми модули. Последниот чекор е само постпроцесирање. Како евалуација на овој метод е предложена оценка за преклопувањето на предвиден модул со протеински комплекс од каталогот на протеински комплекси MIPS и прецизноста е 79% за сензитивност од 31%.

Во [131] се дефинира поимот припадност на кластер на еден јазел во некој кластер како количник помеѓу бројот на јазли помеѓу јазелот и јазлите содржани во кластерот и средната вредност на припадноста на јазлите во кластерот. Иницијално кластерот има само еден јазел, за во секоја итерација да се додаваат нови јазли се додека вредноста на припадноста на јазлите не надмине одредена

гранична вредност. NetworkBlast [132] е алатка во која секој подграф од протеини од мрежата на протеински интеракции се смета за кандидат за функционален модул. Секој подграф – кандидат добива оценка според односот на веродостојноста тој да биде нагоден на претходно креиран модел на протеински комплекс, и веројатноста дека врските во него се појавиле случајно. Потоа се користи алчен алгоритам за да се одредат модулите кои имаат најголема оценка. Алгоритамот предложен во [133] открива подграфови со n јазли кај кои потребно е да се избришат најмалку $n/2$ јазли за да се наруши нивната поврзаност.

Марков кластер алгоритамот (Markov Cluster - MCL) [134] симулира тек на граф преку пресметка на последователните степени на матрицата на соседство. Во секоја итерација, се применува чекор на инфлација за да се засили контрастот помеѓу регионите со силен или слаб тек во графот. Процесот конвергира кон поделба на графот со множество од региони со висок проток (кластерите) подели меѓусебно со граници каде нема проток.

Кластерирање со пребарување на ограничено соседство (Restricted Neighborhood Search Clustering - RNSC) [135], е алгоритам кој врши локално пребарување на просторот на решенија за да минимизира функција на цена на чинење, пресметана според броевите на врските внатре во кластерите и помеѓу кластерите. Почнувајќи од некое иницијално случајно решение, RNSC итеративно поместува јазел од еден кластер во друг ако тоа поместување на намалува вкупната цена на чинење.

Супер парамагнетното кластерирање (Super Paramagnetic Clustering - SPC) [136] е алгоритам за хиерархиско кластерирање инспириран од аналогија со физичките карактеристики на феромагнетен модел подложен на флуктуации при ненулеви температури. Најпрво, SPC со секој јазел од графот придружува спин. Спиновите што припаѓаат на високо поврзан регион флуктуираат на корелиран начин, па јазлите со корелирани спинови се сместуваат во ист кластер. Со зголемување на температурата системот станува помалку стабилен и кластерите стануваат помали.

Друг тип на методи за наоѓање на протеински комплекси се оние кои започнуваат со делумно познати протеински комплекси, на кои потоа се обидуваат да присоединат други протеини од мрежата. Пример за таков алгоритам е Complexpander [137], кој за дадено множество на клучни протеини кои припаѓаат на еден комплекс, генерира листа на кандидат протеини, рангирани според веројатност да припаѓаат на комплексот. Оваа веројатност е добиена од веројатноста да постои патека од стабилни интеракции помеѓу протеинот и некој член на комплексот. Во оваа група на алгоритми спаѓаат и алгоритмите кои кластерирањето го вршат врз база на принципот на заеднички соседи. Имено, според [138], ако два протеин имаат значително поголем број на заеднички соседи од случајната шанса за заеднички соседи, имаат поголема веројатност да се функционално поврзани [138]. Врз основа на ова, за секој пар на протеини се пресметува *p*-вредност која ја означува веројатноста два протеини да имаат некој специфициран број на заеднички соседи, со претпоставка дека распределбата на врските помеѓу протеините е биномна. Таа вредност се смета за растојание помеѓу тие два протеините и е основната мерка според која потоа се врши хиерархиско кластерирање на протеините. Заклучено е дека најголем дел од така добиените кластери имаат протеини анотирани со исти или слични функции. Друг репрезентативен пример од овој тип на алгоритми е алгоритамот PRODISTIN [139], чија идеја лежи врз хипотезата дека растојанието помеѓу протеините пресметано со формулата на Чекановски-Дајс (Czekanovski-Dice) која во предвид ги зема нивните заеднички соседи, го пресликува и функционалното растојание меѓу нив. Добиените кластери се функционално конзистентни, а алгоритамот работи и врз мрежи со вештачки генериран шум.

Покрај споменатите методи, денес теоријата на графови нуди широка палета на методи за кластерирање на графови [140], за кои е вредно да се испита дали би биле поволни за функционална анотација во протеински интеракциски мрежи. Дел од нив веќе се ставени во функција на детектирање на функционални модули во мрежите на протеински интеракции, иако не се конкретно искористени за предвидување на протеинската функција. Меѓу нив, најзначајни се [141], каде кластерирањето се врши според методот на средишност на врски во графот (edge-beetweenness), [7] и [128] каде истиот тој метод е прилагоден за тежински граф

добиен со користење на податоци од генска експресија. Техника за филтрирање на комплетно поврзани графови е имплементирана во [142]. [143] ги одредува функционалните модули преку сопствените вектори на Лапласовата матрица, а во [144] се испитуваат карактеристиките на модулите добиени со мултирезолуциски пристап.

Покрај сите претходно споменати техники за одредување на протеински комплекси и функционални модули, треба да се спомнат уште и асоцијативните правила како техника од податочно рударење која овозможува наоѓање на шаблони на податоци кои се повторуваат. При примена кај мрежите на протеински интеракции, тие шаблони ќе бидат всушност идентичните или сличните подграфови кои се јавуваат во повеќе податочни множества [116].

Често податоците од мрежите на протеински интеракции не се анализираат самостојно, туку заедно со други податоци, како мрежи на генски интеракции, интеракции на коекспресија или повеќе мрежи на протеински интеракции се анализираат истовремено. Предложени се и методи во кои се вклучени комбинации помеѓу мрежите на протеински интеракции со профили од експресија на гени. Овие податоци најчесто се додаваат како тежини на врските на мрежата на протеински интеракции, а се користат и кај било која категорија на методи за предвидување на протеински интеракции [122] [123] [128].

Методите користени во рамки на оваа докторска дисертација се подетално објаснети во рамки на главите 4.3, 5.2 и 5.3.

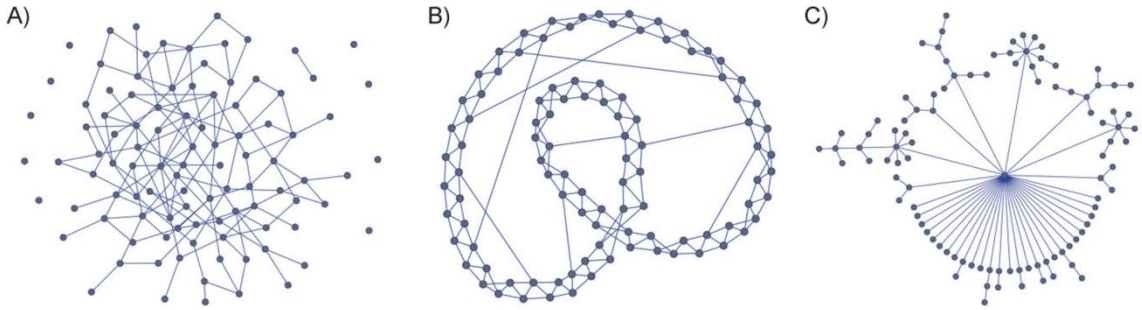
4

РЕШЕНИЈА ЗА ФУНКЦИОНАЛНА АНОТАЦИЈА

Од досега изложеното јасно се наметнуваат неколку проблеми кои треба да бидат надминати за да се постигне крајната цел, а тоа е функционалната анотација на непознат протеин. Во рамки на оваа глава ќе бидат изложени решенија и концепти за решавање на овие проблеми почнувајќи од анализата на протеинските интеракциски мрежи и оправданоста за нивно користење, преку дефинирање на начините за репрезентација на протеинските интеракциски мрежи, до извлекувањето на потребното знаење од дефинираните репрезентации.

4.1 Анализа на протеинските интеракциски мрежи и оправданост за нивно користење

Мрежите на протеински интеракции лесно може да се сместат во групата на комплексни мрежи, но за да се проучат нивните карактеристики кои можат да донесат важни информации и да откријат некои скриени регуларности и хиерархии во графот [145], потребно е да се изврши нивна статистичка и математичка анализа. Овие информации се важни не само за одредување на врската помеѓу структурата на мрежите и функцијата на протеините, туку пред сè за формирање на модел за еволутивниот развој на протеомот на еден организам, бидејќи сите протеини во една фамилија настанале од заеднички предок преку процесите на дупликација и мутација на гени, и според тоа, мрежата на протеински интеракции е отпечаток од целата историја на еволуција на геномот [44].



Слика 4.1 Примери на модели на мрежи:A) Erdös-Rényi случаен граф; B) мрежа на мал свет; C) мрежа со слободен раст

Erdös-Rényi (ER) случајните графови се првите модели на случајни графови. Кај овие графови, врските помеѓу јазлите се додаваат униформно, по случаен избор со истата веројатност, p [146]. Овој модел е детално проучен и најголем дел од неговите математички карактеристики се добро разбрани [147]. Поради оваа причина, истиот е стандарден модел наспроти кој се споредуваат податоците, иако не се очекува да се добие добро поклопување. Поради тоа што ER графовите, за разлика од протеинските интеракциски мрежи, имаат Поасонова распределба на степенот и мали коефициенти на кластерирање се разгледуваат други модели на мрежи. Кај „воопштените случајни графови“ врските со одбираат случајно како

кај ER графовите, меѓутоа распределбата на степенот на јазлите е ограничена така да одговара на онаа присутна во податоците [148] [149]. Мрежите на „мали-светови“ претставуваат еден вид на интеграција помеѓу правилноста (регуларноста) и случајноста бидејќи истите претставуваат правилни прстенести решетки со мал број на случајно преповрзани врски, па според тоа имаат мали дијаметри и големи коефициенти на кластерирање [150]. За интеракциската мрежа на лебниот квасец просечното минимално растојание помеѓу јазлите изнесува 4.16, додека пак максималното растојание помеѓу два протеини, т.е. дијаметарот на мрежата е 17.

Мрежите со слободен раст (scale-free) вклучуваат дополнително ограничување, а тоа е дека распределбата на степенот на јазлите следи степенски закон (power-law) [151] [152]. Веројатноста случајно избран јазел да има степен на поврзаност k е даден со формулата (4.1) во која γ е параметар на степенската распределба и го одредува наклонот на кривата. Тоа значи дека мрежата е доста нехомогена и постојат голем број на јазли со релативно мал степен, и мал број на јазли кои имаат голем број на врски.

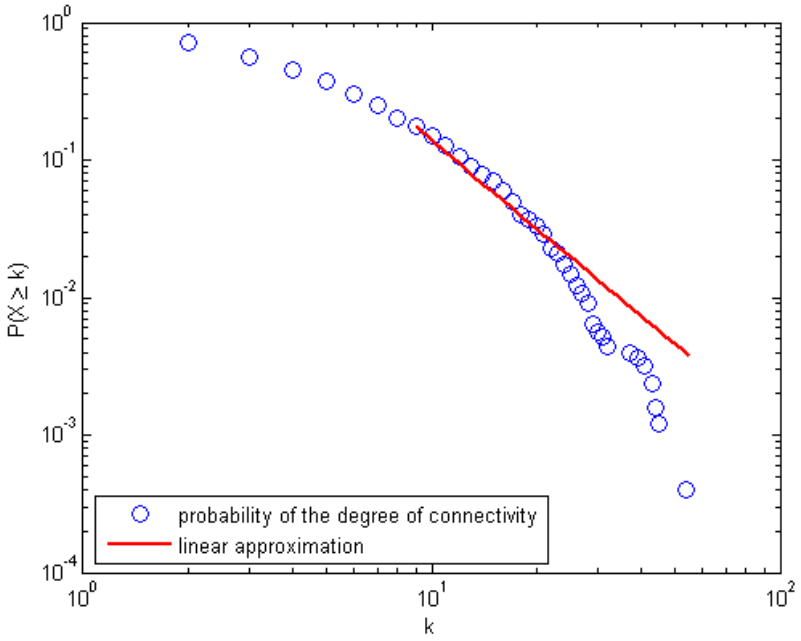
$$P(k) \sim k^{-\gamma} \tag{4.1}$$

Предложени се голем број на варијанти на модели на мрежи со слободен раст [8] [9] [153], од кои најзначајни се оние што се базирани на биолошки мотивираните принципи на генска дупликација и мутација [44] [154] [155] [156]. На сликата 4.1 се прикажани примери на мрежните модели.

Дупликацијата и мутацијата на гените, како основни механизми на еволуција на протеомот, во рамки на моделите на мрежи со слободен раст и степенска распределба на степенот на јазлите се користат за моделирање на приоритетното поврзување. Имено, сите ново настанати јазли кои се приклучуваат на мрежата со поголема веројатност ќе се поврзат со јазли кои веќе имаат голем број на врски, отколку со јазли кои имаат мал број на врски. При процесот на размножување еден организам на својот потомок може да му предаде ген кој е идентична копија на друг ген, процес наречен дупликација. Двата гена на почеток имаат идентични функции. Процесот на дивергенција пак означува мутација на генот, што значи

одредена промена во неговите функционалности. На ниво на протеом, тоа значи дека двата гена продуцираат протеини кои се доволно различни за да имаат различни множества на протеини со кои стапуваат во интеракција, а сепак доволно слични за тие множества да имаат значително голем пресек. Кажано со речникот на математиката, процесот на дупликација е моделиран на следниов начин: случајно се избира јазел *i* од графот и се креира нов јазел *i'* кој е поврзан со сите соседи на *i,* а со самиот јазел *i* се поврзува со веројатност *p.* При моделирање на процесот на дивергенција пак, за секој од јазлите *j* кои се соседи на *i* и *i'* се одбира по случаен избор една од врските (*i*, *j*) или (*i'*, *j'*) и се отстранува со веројатност *q* [44]. Дополнително, моделот [154] дава можност и за вметнување на нови врски во мрежата. Од емпириските податоци за било која мрежа на интеракции може да се одредат параметрите *p* и *q,* така што механизмите на дупликација и дивергенција да ги опфатат сите феномени кои придонесуваат за добиената топологија. Веројатноста со која степенот на еден јазел се зголемува при додавањето на нов јазел во мрежата е линеарно право пропорционална со степенот на поврзаност на јазелот што е суштината на концептот на приоритетно поврзување.

За да се донесе заклучок за типот на распределба на степенот на поврзаност, треба од емпириските податоци да се формира хистограм на бројот на јазли со даден степен на поврзаност и тој да се разгледува на двојно логаритамски график (што се добива со логаритмирање на равенката 4.1), како што е прикажано на Слика 4.2. Можноста тој да се апроксимира со линеарна функција е потребен услов за да се постави хипотезата дека распределбата на степенот на поврзаност на јазлите следи степенска функција.



Слика 4.2 Функција на распределба на степенот на поврзаност на јазел кај мрежа на протеински интеракции на лебен квасец

Методот за определување на параметарот на дискретна степенската распределба за дадено доволно големо множество на емпириски податоци се базира на методот на максимална веродостојност. Со анализа на мрежата на протеински интеракции на лебниот квасец, каде за секој протеин е познат бројот на соседи, се добива дека параметарот на степенската распределба изнесува 3.07. Оваа вредност е надвор од типичниот опсег на вредности помеѓу 2 и 3, па затоа можеме да кажеме дека степенот на јазлите кај протеинските интеракции опаѓа според некој приближен степенов закон. Како дополнително испитување може да се изврши и т.н. тест за квалитет на поклопување (goodness-of-fit). Имено, треба да се генерира множество на податоци изведени директно од степенска распределба со добиениот параметар и да се пресмета растојанието помеѓу синтетички генерираните и емпириските податоци. Сепак, оваа анализа е надвор од доменот на овој труд.

Како алтернативен модел на мрежата на протеински интеракции и против теза на тезата за степенска распределба на степенот на поврзаност на јазлите, во [157] е предложен геометриски граф. Мрежата е опишана преку мотифите кои ги содржи, односно најзастапените типови на графлети во неа, каде под поимот графлет се

подразбира микро граф од 3, 4 или 5 јазли поврзани меѓусебно на сите можни начини. Со споредба на фреквенцијата на различните графлети кај различни типови на мрежи, заклучено е дека мрежата на протеински интеракции има повеќе сличности со геометриски граф, отколку со мрежа со слободен раст. Геометриски граф со радиус *r* е граф за кој важи дека секој јазел е поврзан со оние јазли со кои во одреден метрички простор е оддалечен на растојание не поголемо од *r* по Евклидова норма.

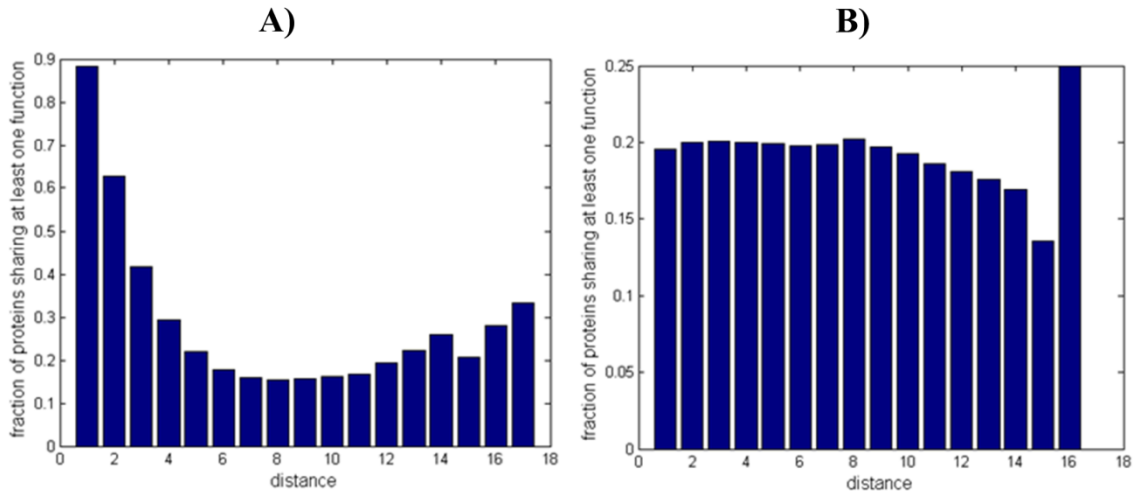
Друга разгледувана статистика за интеракциската мрежа на лебниот квасец е коефициентот на кластерирање на јазлите како мерка за локалната кохезивност на мрежата [145]. По дефиниција, за даден јазел коефициентот на кластерирање е број на негови соседи кои се директно поврзани и меѓу себе. За протеинот *i,* коефициентот на кластерирање C_i е даден со равенката (4.2) при што k_i е бројот на соседи на јазелот i , а e_i е број на врски меѓу нив.

$$C_i = \frac{2e^i}{k_i(k_i - 1)} \tag{4.2}$$

За разгледуваната мрежа просечниот коефициентот на кластерирање изнесува 0.3375, што е релативно голема вредност споредено со просечниот коефициент на кластерирање на случајна мрежа каде тој е од ред на големина 10^{-4} [153].

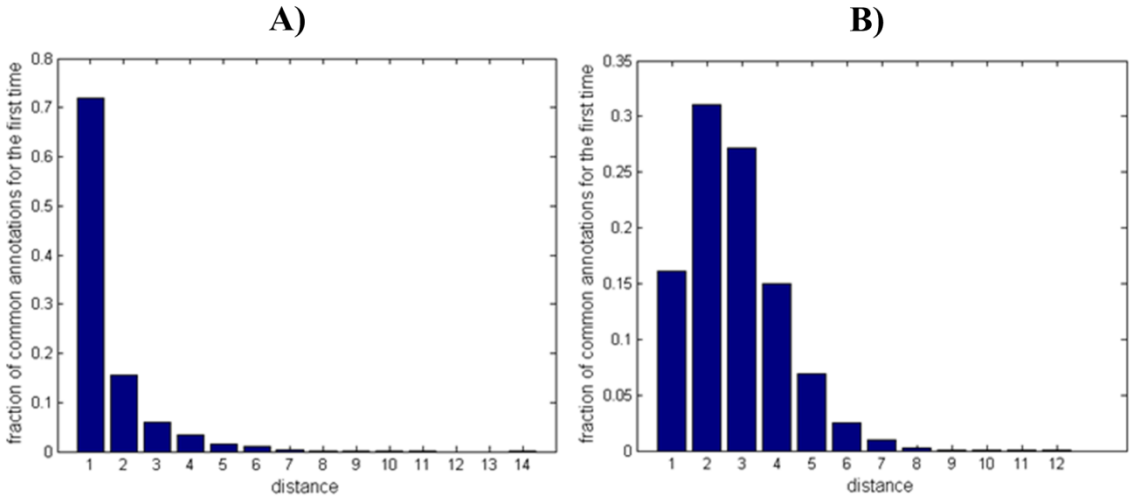
За мрежите на протеинска интеракција да можат да се искористат за предвидување на протеинска функција, нивната анализа мора да се прошири во правец на наоѓање на корелација помеѓу растојанието помеѓу два протеини во графот и нивната функционална анотација. Најпрво, направена е анализа колку од протеините во мрежата делат барем една заедничка функција со протеините со кои се наоѓаат на растојание *k.* Како што може да се заклучи од Слика 4.3А, приближно 90% од протеините кои се директни соседи делат најмалку една заедничка функција. Како што растојанието се зголемува, драстично се намалува бројот на протеини со барем една заедничка анотација. За споредба, на Слика 4.3В даден е и хистограм со идентична анализа и за мрежа со иста структура, но кај која анотациите на јазлите се случајно доделени. Очигледно е дека за ваква мрежа нема никаква корелираност помеѓу растојанието и функционалната анотација на два протеини, па бројот на протеини кои делат најмалку една

заедничка функција е релативно униформен, без разлика на растојанието меѓу протеините.

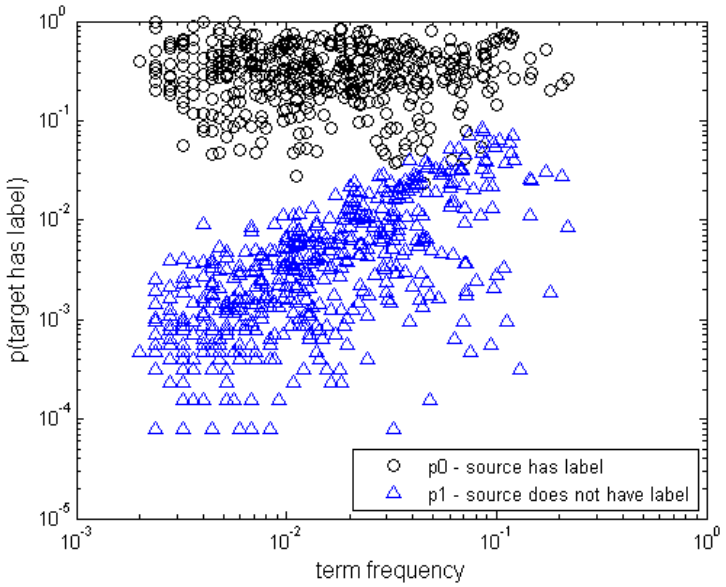


Слика 4.3 Процент на протеини кои делат барем една заедничка функција во зависност од меѓусебното растојание за А) мрежа на протеински интеракции, В) случајно анотирана мрежа

Од голема важност е за даден протеин да се определи растојанието до најблискиот протеин со кој делат заедничка функција. Хистограм за процентот на функции со кои е анотиран еден протеин кои по прв пат се среќаваат кај протеин на растојание *k* од дадениот протеин, е даден на Слика 4.4А. Над 70% од функциите протеините ги делат со своите директни соседи. Повторно, кај мрежа со јазли анотирани по случаен избор (Слика 4.4В), ваков заклучок не може да се изведе. Оваа анализа е многу важна, бидејќи врз неа се базираат идеите на голем дел од алгоритмите за функционална анотација на протеини.



Слика 4.4 Процент на анотации на еден протеин во зависност од растојанието до друг протеин кај кој се појавуваат, а кој е најблизок до дадениот протеин, кај А) мрежа на протеински интеракции, В) случајно анотирана мрежа



Слика 4.5 Веројатност целниот протеин да биде анотиран со дадена функција ако неговиот директен сосед, изворен протеин, е или не е анотиран со таа функција, во зависност од застапеноста на таа функција во целата мрежа.

На Слика 4.5 е даден график на кој е прикажано како зависи веројатноста p_o протеинот наречен целен протеин (target) да биде анотиран со дадена функција ако негов директен сосед, изворен протеин (source) е анотиран со таа функција, од застапеноста на таа функција во целата мрежа. Веројатноста p_1 пак, е веројатноста целниот протеин да е анотиран со таа функција, ако изворниот протеин не е. Анализата е направена за сите 888 функции во мрежата (од податочното множество врз кое се вршени експериментите во овој труд). Доколку изворниот протеин е анотиран со дадена функција, тогаш најчесто веројатноста и целниот протеин да е анотиран со таа функција е поголема отколку ако изворниот протеин не е анотиран со таа функција, без разлика на фреквенцијата на појавување на дадената функција во мрежата.

4.2 Репрезентација на протеинските интеракциски мрежи

Како што беше претходно изложено протеинските интеракциски мрежи имаат особини на комплексни мрежи, па следствено на тоа најсоодветниот начин за нивна манипулација би бил истите да бидат претставени како графови. Во рамки

на оваа докторска дисертација се воведуваат неколку различни граф репрезентации на протеинската интеракциска мрежа, при што секоја од овие репрезентации ги претставува информациите содржани во податоците на различно ниво на апстракција. Целта на ова е да се испита нивото на детали коешто е доволно за ефективно да се изврши целосниот процес на функционална анотација во рамки на протеинските интеракциски мрежи. Пред да бидат објаснети треба да се напомене дека сите различни графови кои произлегуваат од протеинските интеракциски мрежи се ненасочени поради фактот дека и самите интеракции се ненасочени. Во продолжение се дадени дефинициите на различните репрезентации почнувајќи од онаа со најниска, па се до онаа со најголема комплексност.

4.2.1 Едноставни графови

Најосновната дефиниција за граф репрезентација на протеинската интеракциска мрежа е преку *едноставни графови* со $G_1=(V,E)$ каде јазлите на графот $i,j \in V$ соодветствуваат на протеини, а врските $(i, j) \in E$ соодветствуваат на интеракциите помеѓу „протеините“ i и j. Едноставните графови се нетежински. Кога се користат ваквите графови единствено топологијата на протеинската интеракциска мрежа се користи како информација во процесот на функционална анотација. За податоците користени во рамки на оваа докторска дисертација (глава 5.1) важи $|V|=2502$ и $|E|=6354$.

4.2.2 Тежински графови

Наједноставниот начин за збогатување на претходната репрезентација е да се додадат тежини на врски од E и со тоа да се дефинира *тежински граф* $G_2=(V,E,W)$ за протеинската интеракциска мрежа, каде W е матрица чиј елементи w_{ij} ги претставуваат тежините на врските $(i, j) \in E$. Тежините можат да се пресметаат на различни начини. Во суштина овие пресметки се засноваат на определување на семантичката сличност помеѓу два јазли во рамки на графот, односно два

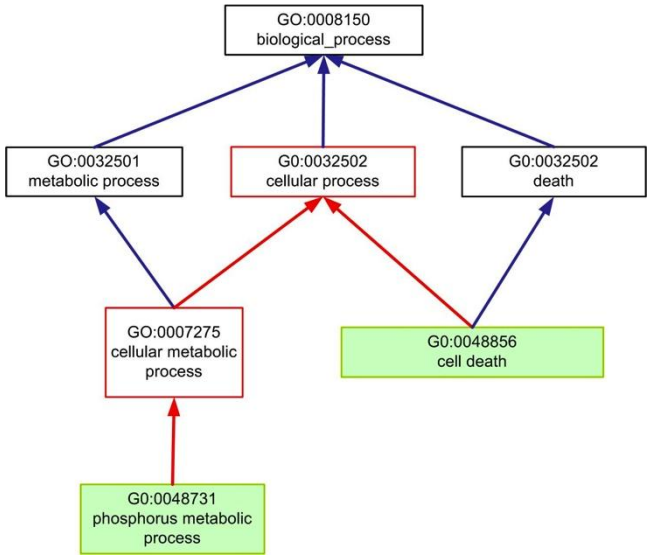
протеини кои влегуваат во интеракција во рамки на протеинската интеракциска мрежа.

4.2.2.1 Семантичка сличност во протеинска интеракциска мрежа

Семантичката сличност по дефиниција прави оцена на врската (корелацијата) помеѓу два објекти во однос на значењето кое тие го носат, при што семантичкото растојание е обратно пропорционално на сличноста. Важноста на семантичката сличност како истражувачка алатка доаѓа до израз благодарение на обемните истражувања направени во областа на обработката на природни јазици. Овде семантичката сличност се користи за решавање на многу суштински проблеми како разрешување на двосмисленост на зборови и текст, пребарување базирано на содржината, автоматско индексирање и поправање на грешки во текст итн. Токму поради овие причини најзначајните достигнувања се постигнати во лингвистичката обработка и поконкретно онтологијата WordNet [158] како најзначајна онтологија за англиски јазик.

Со развојот на биолошките онтологии се отвораат нови можности за споредба на биолошки ентитети. За ваквите ентитети, во којшто спаѓаат и протеините кои се од интерес на ова истражување, да можат да бидат споредени потребно е истите да бидат опишани со идентична онтологија. Ваквиот тип на споредба е семантичка сличност на протеини и ја проценува корелацијата помеѓу два протеини од аспект на сличноста во значењето што го носат анотациите на протеините што се споредуваат. Определувањето на сличноста помеѓу ентитети кон кои се придружени информации (се анотирани) организирани во рамки на определена онтологија претставена како ацикличен граф како што е WordNet, или во нашиот случај Gene Ontology, може да биде базирано на растојание во рамки на графот (врски во графот) и/или на информацискатa содржина (јазли во графот). Метриката за семантичка сличност претставува функција која враќа нумеричка вредност за споредбата на два ентитети претставени преку две множества термини со кои истите се анотирани и припаѓаат на определена онтологија.

Метриките базирани на врските во графот на онтологијата вредноста за сличноста ја определуваат врз основа на должината на најкратката патека помеѓу два јазли во графот или врз основа на просечната должина, доколку постојат повеќе патеки [159]. Дополнително вредноста може да се определи и преку растојанието помеѓу коренот на графот и најнискиот заеднички предок на двата јазли за кои ја пресметуваме сличноста [160]. На сликата 4.6 е даден еден дел од GO за кој треба да се определи сличноста помеѓу јазлите обележани со зелена боја (GO:0048731 и GO:0048856). Најкратката патека на овој пример има должина 3, па и семантичкото растојание може д се земе да биде еднакво на 3. На истиот пример ако ги земеме во предвид сите патеки помеѓу двата јазли, семантичкото растојание може да се пресмета како просек од нивните должини, односно може да биде еднаква на 4.5. Последната варијанта за вредноста на растојанието е ако се гледа најнискиот заеднички предок (GO:0032502) и должината на патеката до коренот, односно 1.



Слика 4.6 Пример за пресметување на растојание во Gene Ontology

За да можат вака пресметаните растојанија да бидат релевантни за соодветната биолошка онтологија мора да важи дека:(а) јазлите и врските во онтологијата се рамномерно распределени и (б) врските што се наоѓаат на исто ниво во онтологијата соодветствуваат на идентична семантичка сличност помеѓу термините. Направени се обиди за надминување (исполнување), како што е на пример дефинирање тежини на врските базирани на хиерархиската длабочина на

истите, или пак дефинирање на тип на врската и густина на јазли [161]. Меѓутоа, специфичноста на јазлите на иста хиерархиска длабочина не мора да биде иста, исто како што и врските на исто ниво не мора да одговараат на идентично семантичко растојание, што би значела дека ваквите стратегии не водат кон исполнување на поставените барања.

Од оваа класа на семантички метрики најпозната е онаа на Pekar и Staab [162] која се пресметува преку должината на најдолгиот пат помеѓу најнискиот заеднички предок и коренот и најдолгиот пат помеѓу секој од јазлите и најнискиот заеднички предок:

$$sim_{PS}(c_1, c_2) = \frac{d(c_a, koren)}{d(c_a, koren) + d(c_1, c_a) + d(c_2, c_a)} \tag{4.3}$$

каде $d(c_1, c_2)$ е должината (изразена преку број на врски) на најдолгата патека меѓу јазлите c_1 и c_2, а c_a е најнискиот заеднички предок јазлите. Ваквата метрика во [163] е применета врз GO.

Останати метрики кои припаѓаат на оваа класа во која се земаат во предвид само врските во графот се метриката дефинирана во [164] каде длабочината на секој јазел се зема во предвид со додавање на тежини, нетежинската метрика од [165] во која специфичноста на јазелот е дефинирана преку неговото растојание до најблискиот терминален јазел, како и онаа во [166] каде наместо самата GO се користи посебно конструирано функционално дрво добиено врз основа на фреквенцијата на различните термини во самото податочно множество.

Класата на семантички метрики базирани на јазли (термини) во графот пресметките ги заснова на самите карактеристики на јазлите, нивните потомци или предци во графот. Кај овие методи како мерка за специфичноста и информативноста на даден термин најчесто се користи информацискатa содржина на тој термин. Информацискатa содржина на јазелот (терминот) c во графот се дефинира преку:

$$IC(c) = -\log p(c) \tag{4.4}$$

каде $p(c)$ е веројатноста за појавување на терминот во дадено податочно множество:

$$p(c) = \frac{freq(c)}{N} \tag{4.5}$$

Информацискатa содржина како концепт може да се примени врз предците на два термини што се споредуваат и врз основа на тоа да се одреди количеството информација заедничко за истите. Метриките за сличност базирани на информациска содржина за разлика од оние базирани на врските во графот имаат помала чувствителност на променливо семантичко растојание и променлива густина на јазли, поради фактот дека информацискатa содржина претставува мерка за специфичност на даден термин (јазел) без оглед на неговата длабочина во графот.

Метриката на Ресник [167], базирана на информацискатa содржина, е првенствено креирана за WordNet и претставува една од најупотребуваните семантички метрики. Овде семантичкатa сличност помеѓу два јазли (термини) во графот на онтологијата се добива преку информацискатa содржина на најнискиот (најспецифичен) заеднички предок. Информацискатa содржина на еден термин е обратно пропорционална на неговото ниво во онтологијата поради тоа што поопштите термини се многу почести од останатите поспецифични термини. Тоа би значело дека коренот на една онтологија ќе има најмала, додека листовите ќе имаат најголема информациска содржина.

Веројатноста за појавување на даден термин c зависи од фреквентноста на тој термин и истата за GO може да се пресмета како:

$$freq(c) = anno(c) + \sum_{h \in deca(c)} freq(h) \tag{4.6}$$

каде *anno(c)* одговара на бројот на протеини во податочното множество кон кои c е придружен како анотација, а $deca(c)$ е множеството на деца на овој термин. Веројатноста за појавување на терминот c е дадена со:

$$p(c) = \frac{freq(c)}{N} = \frac{freq(c)}{freq(root)} \tag{4.7}$$

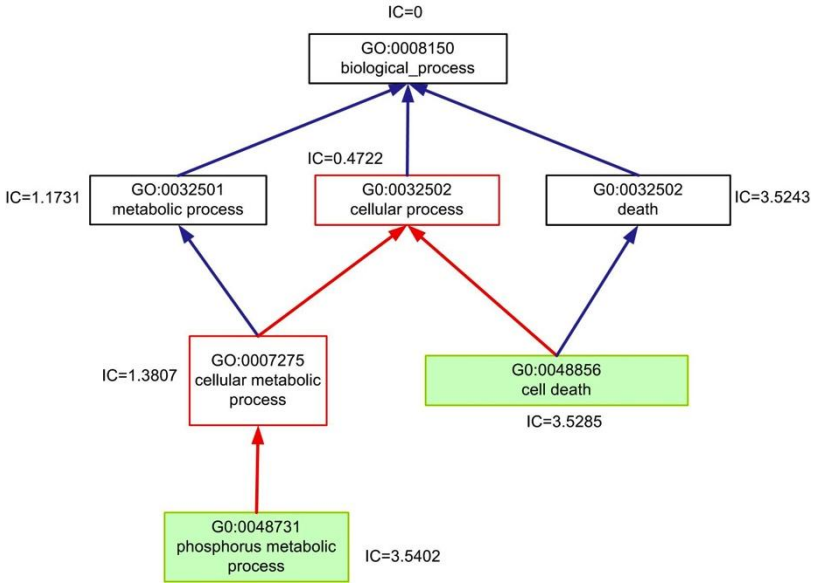
каде *freq(root)* е фреквентноста на коренот на онтологијата.

Сега можеме да ја пресметаме семантичката сличност според:

$$sim_{Resnik}(c_1, c_2) = max_{c \in S(c_1, c_2)}(-log\ p(c)) \quad (4.8)$$

каде $S(c_1, c_2)$ е множество на заеднички предци на c_1 и c_2. Значи, колку е поголема фреквенцијата на некој термин, толку истиот е погенерален. Вредностите за семантичката сличност припаѓаат на опсегот [0, ∞).

На сликата 4.7 е прикажан еден дел од GO при што кон секој јазел (термин) е придружена вредноста за неговата информациска содржина пресметана според дадено податочно множество (податочното множество користено во овој труд). Семантичката сличност на јазлите кои одговараат на термините GO:0048731 и GO:0048856 според Ресник е вредноста на информацискатa содржина на нивниот најспецифичен (најнизок) заеднички предок, GO:0032502 (бидејќи истиот има максимална вредност од сите можни заеднички предци на двата јазли).



Слика 4.7 Информациска содржина за дел од GO

Метриката на Ресник го инкорпорира значењето на најспецифичниот заеднички предок и соодветните растојанија на секој од термините до него. Информациската содржина кај оваа метрика зависи од статистиката на определено податочно множество и ја игнорира специфичната структура на GO, што може да се смета за нејзин недостаток. Иако постојат метрики кои го надминуваат овој проблем, сепак метриката на Ресник се покажува како најдобра во споредбата со други метрики

во најразлични области. Останати метрики во рамки на класата на семантички метрики базирани на информациска содржина кои треба да се споменат се метриката на Лин [168] и метриката на значајност [169] кои покрај метриката на Ресник беа исто така тестирани во рамки на овој труд. Добиените резултати во сите експериментални сценарија беа подобри при користење на метриката на Ресник, па токму поради тоа во овој текст ја презентираме само таа техника како претставник на соодветната класа.

Таканаречените хибридни метрики се дизајнирани за да го надминат претходно споменатиот недостиг кај метриката на Ресник, со тоа што дополнително ги земаат во предвид и врските што постојат во графот на онтологијата, односно неговата структура. Во оваа група на метрики спаѓаат метриките на Jiang и Conrath [170], на Wang [171] и метриката за наоѓање на најкраток пат [172]. Од оваа група како претставник е одбрана метриката на Wang која се покажа како најдобра во извршените експерименти.

Метриката на Wang ја пресметува семантичка сличност врз основа на семантичката вредност на еден термин која е претставена како сума од семантичките придонеси на сите предци на дадениот термин. Во рамки на оваа метрика секоја врска во графот на онтологијата добива определена тежина која зависи од типот на врската и истата треба да биде позитивна и помала од 1. На секој пар термин и негов предок му се придружува семантички придонес на предокот кон терминот. Придонесот се пресметува како производ од тежините на врските што ја сочинуваат најдобрата патека помеѓу терминот и неговиот предок (најдобра е онаа патека за која се добива максимален производ, односно го максимизира придонесот). Семантичката сличност за два термини се добива како количник помеѓу сумата на семантичките придонеси на заедничките предци на двата термини и вкупниот семантички придонес на секој од предците за секој од терминте посебно.

За формално да ја дефинираме оваа метрика за секој термин (јазел) c во графот на GO се дефинира посебен граф $DAG_c=(c,A_c,E_c)$, каде A_c е множество на сите предци на c (ова множество го вклучува и самиот термин), односно јазлите во DAG_c, а E_c

е множество на сите врски помеѓу терминот c и елементите на множеството A_c (неговите предци), односно врските во DAG_c. Под семантичка вредност на еден термин c ќе го подразбираме вкупниот придонес од сите термини од A_c. Бидејќи придонесот се добива како производ од вредности помали од 1, предците кои се поблиску до даден термин c ќе имаат поголем придонес во неговата семантичка вредност. За произволен јазел d од графот DAG_c (односно, $d \in A_c$) дефинираме функција $S_c(d)$ како:

$$\begin{cases} S_c(c) = 1, \text{ за } d = c \\ S_c(d) = max(w_e \times S_c(d') \mid d' \in deca(d)), \text{ за } d \neq c \end{cases} \tag{4.9}$$

каде w_e е тежината за семантички придонес на врската $e \in E_c$ која го поврзува d со некое негово дете d'. Оваа функција всушност го моделира семантичкиот придонес. Самиот термин c има семантички придонес 1. Оттука според дефиницијата за семантичка вредност дадена погоре истата можеме да ја пресметаме како:

$$SV(c) = \sum_{a \in A_c} S_c(a) \tag{4.10}$$

Сега формално можеме да го запишеме изразот за семантичка сличност според Wang. За два термини c_1 и c_2, и нивните соодветни графови $DAG_{c_1} = (c_1, A_{c_1}, E_{c_1})$ и $DAG_{c_2} = (c_2, A_{c_2}, E_{c_2})$, семантичката сличност ќе ја добиеме со:

$$sim_{Wang}(c_1, c_2) = \frac{\sum_{a \in A_{c_1} \cap A_{c_2}} S_{c_1}(a) + S_{c_2}(a)}{SV(c_1) + SV(c_2)} \tag{4.11}$$

каде $S_{c_1}(t)$ и $S_{c_2}(t)$ се семантичките придонеси на терминот t кој е заеднички предок на термините c_1 и c_2.

Со помош на претходно изложените метрики може да се определи сличност помеѓу два термини од GO. Нашата крајна цел е одредување на семантичка сличност помеѓу два протеини кои можат да бидат анотирани со повеќе од еден термин што значи треба да се определи начин на комбинирање на семантичките сличности што се добиваат за паровите термини придружени на двата протеини.

Нека се дадени две множества на термини $T_1 = \{ t_{11}, t_{12}, ..., t_{1m} \}$ и $T_2 = \{ t_{21}, t_{22}, ..., t_{2n} \}$ и истите да бидат придружени како анотации на два протеини p_1 и p_2, соодветно. Нека сличноста на два термини е определена со $sim(t_{1i}, t_{2j}), i \in [1,m], j \in [1,n]$. Можеме да дефинираме матрица на сличност за протеините p_1 и p_2, $SIM = [sim_{ij} = sim(t_{1i}, t_{2j})]$ со димензии $m \times n$. Постојат различни начини за искористување на оваа матрица за да се најде сличноста на парот протеини [173] [174] [175] [176], но ние овде ќе го изложиме само оној што се покажа како најдобар во нашите експерименти [177]. Сличноста на p_1 и p_2 ќе ја најдеме како максимум од просекот од максимални сличности по редици и просекот од максимални сличности по колони:

$$sim(p_1, p_2) = max \left(\frac{\sum_{i=1}^{m} max_{1 \le j \le n} sim(t_{1i}, t_{2j})}{m}, \frac{\sum_{j=1}^{n} max_{1 \le i \le m} sim(t_{1i}, t_{2j})}{n} \right) \quad (4.12)$$

4.2.2.2 Стратегии за доделување тежини во тежинскиот граф

Метриките за семантичка сличност претставуваат појдовна точка за дефинирање на тежините во графот за протеинската интеракциска мрежа. Во рамки на истражувањето покрај претходно изложените метрики, кои во позадина ја имаат GO, за определување на тежините беа искористени и едноставни несемантички мерки за корелација помеѓу два јазли во графот што ја претставува протеинската интеракциска мрежа. Таквата корелација можеме уште да ја наречеме анотациска корелација бидејќи се заснова на анотациите доделени на парот јазли (протеини). Да разгледаме два јазли i и j, $i, j \in V$ за кои постои врска $(i, j) \in E$ во графот G_2. Ако $T_i = \{ t_{i1}, t_{i2}, ..., t_{im} \}$ и $T_j = \{ t_{j1}, t_{j2}, ..., t_{jn} \}$ се множествата на термини придружени кон јазлите i и j, соодветно, тогаш несемантичката корелација помеѓу двата јазли ја пресметуваме како нормализиран Џакард (Jaccard) индекс:

$$J_{ij} = \frac{1}{2} \left(\frac{|T_i \cap T_j|}{|T_i|} + \frac{|T_i \cap T_j|}{|T_j|} \right) \quad (4.13)$$

Ваквата несемантичка метрика ја користиме за да ги евалуираме семантичките метрики од аспект на нивниот придонес во нашата крајна цел, функционалната анотација во протеинските интеракциски мрежи.

Сега тежините во графот G_2 можеме да ги дефинираме на еден од следните начини:

а) *Содржински базирани тежини*: содржински базирана пресметка на тежината е онаа што доделува тежина w_{ij}^1 на врската $(i, j) \in E$ со тоа што ги разгледува единствено термините („содржините“) придружени на јазлите i и j, притоа не земајќи ја во предвид нивната околина (структурата на графот). Притоа вредноста на тежината може да биде или семантичка или несемантичка, односно:

$$w_{ij}^1 = \frac{max([sim(i,j)]) - sim(i,j)}{max([sim(i,j)]) - min([sim(i,j)])} \quad \text{или} \quad w_{ij}^1 = J_{ij} \tag{4.14}$$

каде што семантичката сличност $sim(i,j)$ се пресметува според равенката (4.12) со користење на една од семантичките метрики: sim_{Wang} според равенка (4.11) или sim_{Resnik} според равенката (4.8), $max([sim(i,j)])$ и $min([sim(i,j)])$ се максималната и минималната вредност од сите можни семантички сличности, соодветно. На овој начин обезбедуваме тежините секогаш да добиваат вредности во опсегот [0,1].

б) *Структурно базирани тежини*: структурно базираната пресметка на тежини е онаа што во предвид го зема контекстот на јазлите i и j, но не и содржината на самите јазли, кога ја пресметува тежината w_{ij}^2 за врската $(i, j) \in E$. За да ја пресметаме w_{ij}^2 треба да определиме начин на пресликување на контекстот на i и j така што резултатот ќе ги содржи сите структурни информации за овие јазли. Структурната информација за графот G_2 е природно закодирана во неговата матрица на соседство $A=[a_{ij}]$, па искористувајќи го ова тежинската матрица $W^2 = [w_{ij}^2]$ можеме да ја дефинираме на следниот начин:

$$W^2 = W^1 \times A + A \times W^1 \tag{4.15}$$

каде $W^1 = [w_{ij}^1]$ е содржински базираната тежинска матрица. Бидејќи важи $a_{ij} = 0$, $\forall (i,j) \notin E$, за секоја w_{ij}^2 првиот дел од равенката (4.15) ја дава сумата на сите содржински базирани тежини помеѓу јазелот i и сите соседи на j, додека вториот дел е сумата на сите содржински базирани тежини помеѓу јазелот j и сите соседи на i. Од претходно направената анализа на протеинските интеракциски мрежи видовме дека во истите постојат мал број протеини кои влегуваат во интеракција со многу други протеини, што ќе значи дека ќе имаме хабови во графот што ја претставува таквата протеинска интеракциска мрежа. Равенката (4.15) ќе даде високи вредности за јазли што имаат висок степен и обратно, па поради тоа правиме усреднување со цел да го избегнеме ваквиот ефект на фаворизирање и ја добиваме равенката (4.16). Дополнително w_{ij}^2 се нормализирани за да бидат во истиот опсег како и w_{ij}^1.

$$W^2 = \frac{1}{2}(W^1 \times A^1 + A^2 \times W^1) \tag{4.16}$$

каде $A^1 = [a_{ij} / (\sum_{n=1}^{N} a_{nj})]$, $A^2 = [a_{ij} / (\sum_{n=1}^{N} a_{in})]$, и $N=|V|$.

в) *Хибридни тежини*: хибридните тежини прават комбинација од содржински и структурно базираните тежини, при што најприродниот начин е да се земе нивниот просек, со што и овие тежини добиваат вредности од истиот опсег:

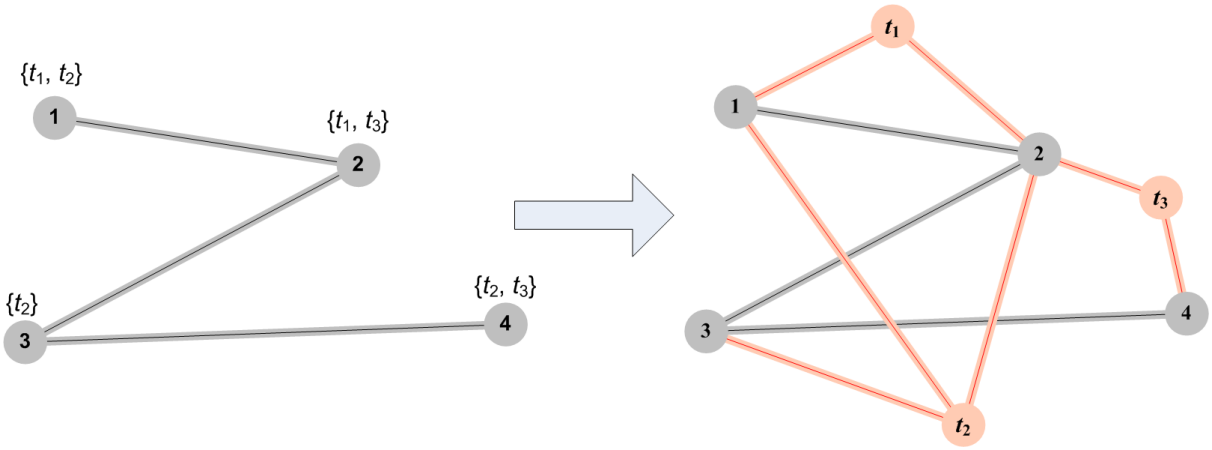
$$W = \left(W^1 + W^2\right)/2 \tag{4.17}$$

Во глава 6 ќе видиме како овие различни стратегии за доделување на тежини влијаат на резултатите за функционалната анотација.

4.2.3 Протеин-термин графови

Дефинираме $G_3=(V \cup T,\ E \cup E_t)$ како *протеин-термин граф* кај кој термините придружени кон протеините во рамки на протеинската интеракциска мрежа стануваат дел од нејзината репрезентација. Поконкретно, T е множеството од сите термини што се појавуваат во рамки на протеинската интеракциска мрежа и секој термин t_i е претставен како јазел во рамки на графот. E_t е множеството на врски (i,t_j) каде $i \in V$, $t_j \in T$ и терминот t_j е придружен кон протеинот i (i е анотиран со t_j)

во рамки на протеинската интеракциска мрежа. Оваа дефиниција на репрезентацијата и множеството на дополнителни врски E_t ги зема во предвид дополнителните врски единствено помеѓу јазлите протеини (V) и новите јазли термини (T), при што не постојат врски помеѓу два јазли термини. V и E се дефинирани идентично како во претходните репрезентации. Графот е нетежински.



Слика 4.8 Протеин-термин граф.

На сликата 4.8 е даден еден пример за добивање на протеин-термин граф. Термините придружени кон протеините и нивните врски се додаваат во графот. Сивите јазли 1, 2, 3 и 4 се протеини, а црвените јазли t_1, t_2 и t_3 се термини. Новите врски се означени со црвено.

На овој начин функционалните поврзаности помеѓу протеините во протеинската интеракциска мрежа се директно вклучени во нејзината граф репрезентација, па според тоа и во процесот на обработка на графот и функционална анотација. При креирање на протеин-термин графот за податоците користени во рамки на оваа докторска дисертација (глава 5.1) се добиваат вкупно 3390 јазли ($|V|$=2502, $|T|$=888) и 37869 врски ($|E|$=6354, $|E_t|$=31515).

4.2.4 Комплетно функционално поврзани графови

Комплетно функционално поврзаните графови се дефинирани со $G_4 = (V, E \cup E_f, W^f)$. Нека T_i и T_j се множествата од термини придружени на

јазлите i и j, соодветно, тогаш за врската (i, j) ќе важи $(i, j) \in E_f$ ако и само ако $(i, j) \notin E$ и $T_i \cap T_j \neq \varnothing$. W^f е тежинската матрица. Со други зборови ако два протеини во протеинската интеракциска мрежа делат (имаат заеднички) барем еден термин во графот се додава врска помеѓу нив дури и ако истите не влегуваат во интеракција, со што креираме еден вид на „лажни“ интеракции. Сепак, информацијата за „вистинските“ интеракции се зачувува преку тежинската матрица. Имено, на секоја врска и се доделува содржински базирана тежина w_{ij}^1, како што е дефинирана со равенката (4.14), и за секоја врска која претставува постоечка интеракција се додава дополнителна константа. Формално имаме:

$$w_{ij}^f = \frac{1}{2}(w_{ij}^1 + c_{ij}) \tag{4.17}$$

каде

$$c_{ij} = \begin{cases} 1, \text{ ако } (i, j) \in E \\ 0, \text{ инаку} \end{cases} \tag{4.18}$$

за секој пар $(i, j) \in E \cup E_f$. Константата ја земаме да биде 1 бидејќи тоа е максималната вредност за содржински базираната тежина w_{ij}^1 во случај на идентични термини во двата поврзани јазли. На овој начин обезбедуваме тежината на секоја врска што претставува вистинска интеракција да биде поголема (или во најлош можен случај еднаква) од тежината на врските што ги претставуваат лажните интеракции, но истовремено овозможуваме содржинската сличност во својот максимум да има ист ефект како и било која вистинска интеракција. Факторот на нормализација 1/2 се зема за да тежините во ваквата репрезентација припаѓаат во истиот опсег како и оние за тежинскиот граф G_2.

4.3 Пристапи за обработка на протеински интеракциски мрежи

Извршените анализи и дефинираните репрезентации на протеинската интеракциска мрежа во форма на граф од претходните поглавја ни даваат за право истата да ја третираме како било која друга комплексна мрежа. Во последните години комплексните мрежи се наоѓаат во фокусот на истражувањата на многу научни области. Нивната структура открива суштински информации за јазлите,

како истите се поврзуваат и како споделуваат информации. Доследно на поделбата дефинирана во глава 3 пристапите за обработка користени во рамки на овој труд можеме да ги поделиме на две суштински различни групи односно директни пристапи и пристапи базирани на кластерирање. Како што беше покажано директните пристапи се далеку подетално проучени кога станува збор за извлекување на функционална анотација. Кластерирањето како пристап во протеинските интеракциски мрежи е многу малку искористено и тоа воглавно во откривањето на протеински комплекси. Токму поради тоа во оваа докторска дисертација поголем акцент е поставен на пристапите базирани на кластерирање. Во оваа поглавје ќе се задржиме на општи концепти и на „филозофиите“ на пристапите избрани како решенија за обработка на протеинските интеракциски мрежи. Деталите за конкретните алгоритми ќе бидат објаснети во глава 5.

4.3.1 Директни пристапи

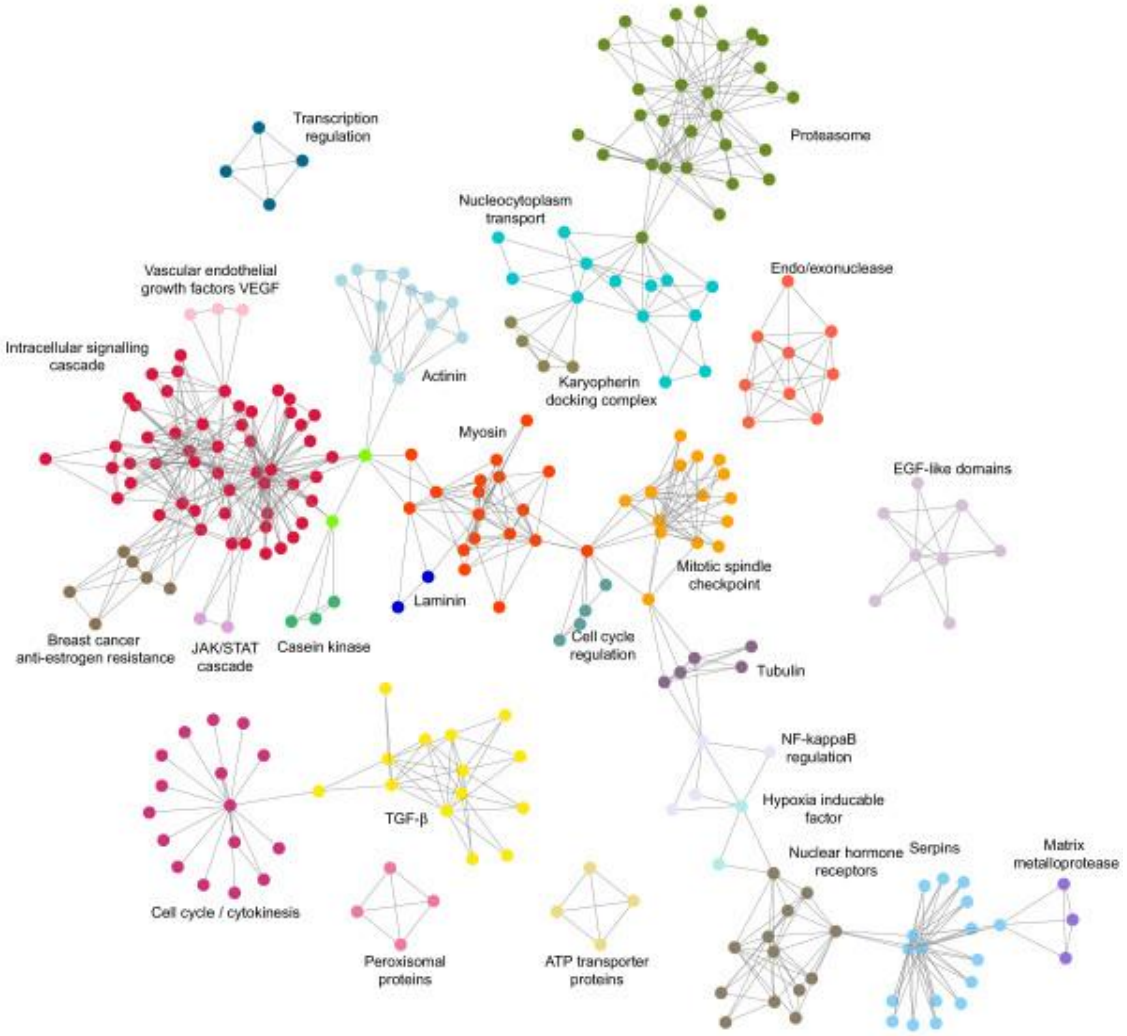
Како што беше претходно наведено еден од главните недостатоци на директните методи е нивната хипотеза дека протеините со слични функции се секогаш тополошки блиску во рамки на интеракциската мрежа. Тоа се должи на фактот дека ваквите методи ја искористуваат информацијата за своите соседи само до определено ниво, па оттука произлегува дека истите не можат да изведат функционална анотација на непознат протеин чиј директни соседи во интеракциската мрежа исто така се непознати, односно неанотирани. Хипотезата од која тргнуваме е дека симултаната активност на понекогаш функционално различни агенти опфаќа процеси на повисоко ниво во различни делови од протеинската интеракциска мрежа. Ваквата хипотеза е погенерална, бидејќи клика (clique) од протеини со слична функција може да се третира еквивалентно како множество од јазли што го набљудува истото функционално соседство. Анализата направена во главата 4.1 укажува дека протеини со слични функции можат да се појават и на поголеми растојанија во мрежата, што претставува дополнително оправдување за нашиот пристап кој протеинската интеракциска мрежа ја гледа глобално.

Методот на случајна патека (random walk) е концепт познат повеќе од 100 години и претставува математички формализам за траекторија која се состои од повеќе случајни последователни чекори. Има голема примена при моделирање на проблеми од многу области, како економија, физика и математика, а кај компјутерските науки е најприменуван како техника за сегментирање на слики и при анализа на мрежи. На пример, врз него се базира познатиот PageRank алгоритам кој Google го користи за пребарување на Интернет страници. Во теоријата на графови, методот на случајна патека може да се дефинира преку идејата за случаен пешак кој на почеток се наоѓа во еден јазел од графот. Во првиот чекор, по случаен пат се избира еден од соседите на првиот јазел и пешакот се придвижува кон него. Во наредниот чекор, постапката се повторува, со тоа што сега се избира еден од јазлите – соседи на јазелот на кој моментално се наоѓа пешакот, повторно по случаен пат. Токму овој метода е искористена како директен пристап за глобална анализа на протеинската интеракциска мрежа. Со помош на случајното изминување на графот кој ја претставува протеинската интеракциска мрежа имаме за цел да направиме сумаризација на еден вид на профил на соседство за некој непознат протеин, нешто што природно може да се извлече како информација од стабилната состојба на изминувањето која е укажува на меѓусебниот афинитет на јазлите во рамки на графот.

4.3.2 Пристапи базирани на кластерирање

Групирањето на протеините од протеинската интеракциска мрежа во множества коишто покажуваат поголема сличност помеѓу протеините во исто множество во однос на протеини од други различни множества претставува ефективен пристап кон разбирањето на врската помеѓу интеракцискиот контекст на еден протеин и неговите функции. Поради фактот дека биолошките функции можат да бидат извршени од страна на конкретна група од протеини, поделбата на протеинските интеракциски мрежи во природно групирани делови е од суштинско значење за испитувањето на евентуалните врски помеѓу функцијата и топологијата на мрежите или за откривање на скриено знаење содржано во истите.

На слика 4.9 е прикажан еден пример на протеинска интеракциска мрежа за протеомот на стаорец [178]. Идентификуваните кластери соодветствуваат на функционални групи, т.е. на протеини кои имаат исти или слични функции, за кои се очекува да бидат вклучени во истите процеси. Кластерите се лабелирани со најдоминантната функција или протеинска класа. Најголем дел од овие кластери се поврзани со рак или метастази, што индиректно укажува на огромното значење на детекцијата на кластери во протеинските интеракциски мрежи. Пристапите за обработка на протеинските интеракциски мрежи врз база на кластерирање на графовите што ги претставуваат истите генерално можат да се поделат на две поголеми групи според нивната историја во научните истражувања и тоа традиционални и современи пристапи.



Слика 4.9 Идентификувани кластери во протеинската интеракциска мрежа на стаорец при анализата на канцер [178]

Традиционалните пристапи се најшироко распространетите во литературата и се користат одамна и во скоро секоја можна истражувачка област, при што истите се релативно едноставни во своето функционирање. За разлика од нив современите пристапи се развиени во последните неколку години и воглавно произлегуваат од истражувањата направени во теоријата на комплексни мрежи и истите се високо ниво на комплексност. Во рамки на овие две класи постои и дополнителна поделба која што не е строга и е направена врз основа на она што е суштинска карактеристика на соодветните пристапи.

4.3.2.1 Традиционални пристапи

Хиерархиско кластерирање

Еден граф може да има хиерархиска структура, т.е. може да се појавуваат повеќе нивоа на групирање на јазлите, со тоа што во таквиот граф може да се увиде дека помалите кластери припаѓаат на поголеми кластери, коишто припаѓаат на уште поголеми кластери и така натаму. Во такви случаи може да се користат алгоритми за хиерархиско кластерирање, т.е. техники што ќе ја откријат повеќе-нивовската структура на графот. Почетната точка за било кој пристап за хиерархиско кластерирање е дефинирање на мерка за сличност помеѓу јазлите. После избирањето на мерка се пресметува сличноста за секој пар на јазли, без разлика дали истите се поврзани или не. Техниките за хиерархиско кластерирање имаат за цел да идентификуваат групи од јазли со висока сличност, и можат да се поделат на две категории. Класичниот пристап е да се започне од единечни кластери (секој јазел во посебен кластер) и итеративно да се спојуваат кластерите доколку нивната сличност е доволно висока и уште се нарекува агломеративен пристап. Обратниот процес на започнување од еден единствен кластер и негово итеративно делење започнува да се користи во поново време па истиот го разгледуваме во групата на современи пристапи (разделувачко кластерирање). За да може да се одвива процесот на спојување на кластерите потребно е да се дефинира мерка за сличност на кластери која може да биде најмалата, најголемата или просечната сличност помеѓу поединечните јазли од двата кластери во зависност од тоа дали

кластерирањето е со единечно, комплетно или просечно поврзување на кластерите, соодветно. Хиерархиското кластерирање ја има погодноста што не е потребно претходно знаење за бројот и големината на кластерите. Меѓутоа, истото не нуди начин за дискриминација помеѓу големиот број на кластери добиени со процедурата, ниту пак начин како да се избере оној или оние што подобро ја претставуваат структурата на графот. Дополнително еден од најголемите недостатоци на ваквиот пристап е лошото скалирање, посебно во случаи каде мерката за сличност е нетривијална и има незанемарлива цена на чинење од аспект на перформанси.

Партиционирачко кластерирање

Делбеното кластерирање опфаќа уште една многу популарна класа на методи за наоѓање на кластери во множество од податочни точки. Овде бројот на кластери , нека го земеме како *k*, се задава однапред. Податочните точки мора да бидат претставени во определен метрички простор, така што секој јазел се пресликува во точка и се дефинира метрика за растојание помеѓу парови од точки во просторот. Растојанието треба да биде мерка за различноста помеѓу јазлите. Целта е да се поделат точките во *k* кластери така што ќе се максимизира/минимизира определена функција на цена на чинење врз основа на растојанијата помеѓу точките и/или од точките до центроидите т.е. соодветно дефинирани позиции во просторот. Една од најпознатите техники за кластерирање, кластерирањето со *k*-средини, припаѓа на оваа група. Главниот недостаток на ваквото кластерирање е потребата од предефинирање на бројот на кластери и неможноста за негово добивање на друг начин. Исто така како проблем може да се гледа и пресликувањето во метричкиот простор, што за определени графови може да биде многу природно, додека за други вештачко што значи тешко толкување на добиените резултати.

Спектрално кластерирање

Спектралното кластерирање ги вклучува сите методи и техники кои дадено множество од објекти го делат на кластери врз основа на сопствените вектори на

матрицата на сличност (горно-триаголна матрица која ги содржи вредностите за сличноста помеѓу било кој пар објекти, добиени со помош на некоја метрика) или некоја друга матрица изведена од неа. Конкретно, објектите можат да бидат точки во некој метрички простор, или, како во нашиот случај, јазли на некој граф. Спектралното кластерирање со состои од трансформација на иницијалното множество од објекти во множество од точки во простор, чијшто координати се елементи на сопствените вектори, по што множеството од точки се кластерира со некоја традиционална метода, како што е кластерирањето со *k*-средини. Главната причина за промена на репрезентацијата на објектите преку користењето на сопствените вектори е тоа што со ваквата претстава кластерирачките карактеристики на иницијалните објекти стануваат поочигледни. На овој начин, спектралното кластерирање е во можност да раздели податочни точки кои не би можеле да бидат разграничени доколку директно се примени кластерирање со *k*-средини (има тенденција да продуцира конвексни множества од точки).

4.3.2.2 Современи пристапи

Разделувачко кластерирање

Еден едноставен начин за идентификување на кластери во еден граф е да се откријат врските што поврзуваат јазли од различни кластери и истите да се отстранат, така што кластерите би биле меѓусебно неповрзани. Ова е филозофијата на разделувачките алгоритми. Кај овие алгоритми од суштинско значење е да се најде определена карактеристика на меѓу-кластерските врски којашто ќе овозможи истите да бидат идентификувани. Раздвојувачките алгоритми не воведуваат значителен концептуален напредок во однос на традиционалните техники и како што беше претходно кажано истите едноставно вршат хиерархиско кластерирање на графот. Главната разлика од традиционалното радвојувачко хиерархиско кластерирање е во тоа што овде се отстрануваат меѓу-кластерски врски наместо да се отстрануваат врски помеѓу парови од јазли со ниска сличност. Притоа, не можеме со сигурност да гарантираме дека меѓу-кластерските врски поврзуваат јазли со ниска сличност. Во

определени случаи наместо единечна врска, може да се отстранат јазли (заедно со сите нивни врски) или цели подграфови.

Кластерирање базирано на модуларност

Голем број на алгоритми можат да идентификуваат определено множество од групирања. Идеално би се добило само едно групирање и само неколку, иако постојат и техники, како што е хиерархиското кластерирање, што произведуваат голем број на можни групирања. Тоа не значи дека сите пронајдени групирања се подеднакво добри. Поради тоа потребно е да се има определен квантитативен критериум за процена на квалитетот на определено групирање во графот. Функција на квалитет е функција којашто доделува определена нумеричка вредност на секое групирање во графот. На овој начин може да се изврши рангирање на групирањата врз основа нивната вредност доделена од функцијата за квалитет. Групирањата со висока вредност се сметаат за „добри“, што значи дека групирањето со највисока вредност по дефиниција е најдобро. Сепак, треба да се има во предвид дека прашањето кога определено групирање е подобро од друго е „лошо поставено“ (не задоволува дека има единствено решение кое се менува континуално со промената на почетните услови) и одговорот зависи од специфичниот концепт на кластер и/или применетата функција на квалитет.

Најпопуларната функција на квалитет е модуларноста на Newman and Girvan [179]. Модуларноста се базира на идејата дека за случаен граф не се очекува да поседува структура на кластери, па можното постоење на кластери се открива со споредбата помеѓу реалната густина на врски во подграф и густината што би се очекувала да постои во рамки на подграфот доколку јазлите на графот се поврзани независно од структурата на кластерите. Оваа очекувана густина на врски зависи од избраниот нулти модел т.е. копија на оригиналниот граф со задржување на одредени структурни карактеристики, но без структура на кластери. Врз основа на ова модуларноста може да се пресмета како:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j) \tag{4.19}$$

каде што сумата се извршува за сите парови на јазли, A е матрицата на соседство, m е вкупниот број на врски во графот и P_{ij} го претставува очекуваниот број на врски помеѓу јазлите i и j во нултиот модел. Функцијата δ (уште се нарекува и Кронекер делта симбол) е еднаква на 1 ако јазлите i и j се во истиот кластер ($C_i = C_j$), а во секој друг случај е еднаква на 0. Изборот на нултиот модел во принцип е произволен и постојат повеќе можности. Сепак поради импликациите на распределбата на степенот на јазлите врз структурата и функцијата на реалните мрежи, пожелно е да се избере нулти модел со идентична распределба на степенот како и оригиналниот граф. Стандардниот нулти модел за модуларноста наложува да очекуваната секвенца од степени (после усреднувањето над сите можни конфигурации на моделот) се совпаѓа со вистинската секвенца од степени на графот. Во ваквиот нулти модел, еден јазел може да биде поврзан со било кој друг јазел од графот и веројатноста дека јазлите i и j, со степени k_i и k_j, се поврзани, може лесно да се пресмета. Всушност, за да се формира врска помеѓу i и j треба да се поврзат две полу-врски што припаѓаат на i и j. Веројатноста p_i за случајно да се избере полу-врска која припаѓа на i е $k_i/2m$, бидејќи има k_i полу-врски што му припаѓаат на i од вкупно $2m$. Сега, веројатноста за врска помеѓу i и j е дадена со производот p_ip_j, бидејќи врските се поставуваат независно една од друга. Резултатот е $k_ik_j/4m^2$, што дава $P_{ij} = 2mp_ip_j = k_ik_j/2m$ за очекуваниот број на врски помеѓу i и j. Па крајниот израз за модуларноста ќе биде:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_ik_j}{2m}\right)\delta(C_i, C_j) \qquad (4.20)$$

По претпоставка, високи вредности за модуларноста укажуваат на добро групирање (иако тоа не мора секогаш да биде случај). Според тоа, групирањето кое соодветствува на максималната вредност на модуларноста за даден граф треба да биде најдобро, или во најмала рака многу добро. Токму ова е главната мотивација за максимизација на модуларноста на што се темелат пристапите за кластерирање базирани на модуларност. Исцрпната оптимизација на Q е невозможна поради огромниот број на начини на кои може да се групираат јазлите на еден граф, дури и кога графот е релативно мал. Токму поради тоа овие алгоритми за кластерирање всушност претставуваат алгоритми за наоѓање апроксимација на максимумот на модуларноста во разумно време.

Мултирезолуциско кластерирање

Оптимизацијата на модуларноста има резолуциска граница што може да ги спречи алгоритмите базирани на модуларност во детекцијата на кластери коишто се споредбено мали во однос на графот како целина, дури и кога се добро дефинирани како што е тоа примерот со кликите (cliques). Значи, ако групирањето со максимална модуларност вклучува кластери со вкупен степен од ред на големина $\sqrt{m}$ (каде m е вкупниот број на врски во графот) или помал, не можеме со сигурност однапред да кажеме дали кластерите претставуваат единствени целини или се добиени со комбинација на помали целини којшто се слабо меѓусебно поврзани. Овој проблем на резолуцијата има големо влијание во практичните апликации. Реалните графови коишто поседуваат определена структура на кластери вообичаено содржат кластери кои се многу различни во својата големина, па голем број на мали кластери можат да останат неоткриени. Модуларноста е осетлива дури и на единечни врски. Многу реални графови, како што претходно видовме за оние во биологијата, се реконструираат преку различни експерименти и може да се добија лажно позитивни врски, што би значело дека ако се случи два мали подграфови да бидат поврзани со неколку лажни врски, алгоритмите базирани на модуларности ќе ги сместат во ист кластер, изведувајќи врска помеѓу ентитети што во реалноста може да се комплетно различни. Резолуциската граница потекнува од самата дефиниција на модуларноста, поконкретно од нејзиниот нулти модел. Слабата страна на нултиот модел е имплицитната претпоставка дека секој јазел може да се поврзе со секој друг јазел, што имплицира дека секој дел од графот знае за сè останато.

Во суштина резолуциската граница имплицира дека обичната отимизација води до груб опис на структурата на кластери во графот во ред на големина којашто нема никаква врска со големината на самите кластери во графот, информација која a priori не ни е достапна. Во отсуство на ваквата информација, кластерирачкиот метод треба да биде во можност да ги испита сите можни размери (редови на големина) за да се осигура дека на крај ќе ги идентификува

сите релевантни кластери. Мултирезолуциското кластерирање е базирано токму на овој принцип. Ваквиот пристап подразбира проширување на модуларноста со определен параметар којшто може да се менува и уште се нарекува параметар на резолуција. Со помош на овој променлив резолуциски параметар можат да се подесат карактеристичните големини на кластерите што треба да се откријат.

Кластерирање базирано на статистичко изведување

Статистичкото изведување има за цел да изведува карактеристики за податочни множества, почнувајќи од множество на набљудувања и хипотези за моделот. Кога податочното множество е граф, тогаш моделот, базиран на хипотезите за тоа како јазлите се поврзани меѓусебно, мора да се поклопи со вистинската топологија на графот. Техниките за кластерирање базирани на статистичко изведување се обидуваат да го најдат најдоброто поклопување на модел со графот, при што моделот претпоставува дека јазлите имаат определена класификација базирана на нивните шаблони на поврзување.

Селекцијата на модел има за цел наоѓање на модели кои истовремено се едноставни и добри во опишувањето на системот/процесот. Еден базичен пример за селекција на модел е поклопувањето на криви. Не постои еднозначен рецепт според кој се селектира модел, а наместо тоа постојат најразлични евристики со кои се прави нивна процена. Модуларната структура на еден граф може да се разгледува како компресиран опис на графот со цел да се апроксимира целосната информација содржана во неговата матрица на соседство. Токму ова е основната идеја на оваа класа на алгоритми, да се опише графот со помалку информации од оние што се содржани во матрицата на соседство.

Преклопувачко кластерирање

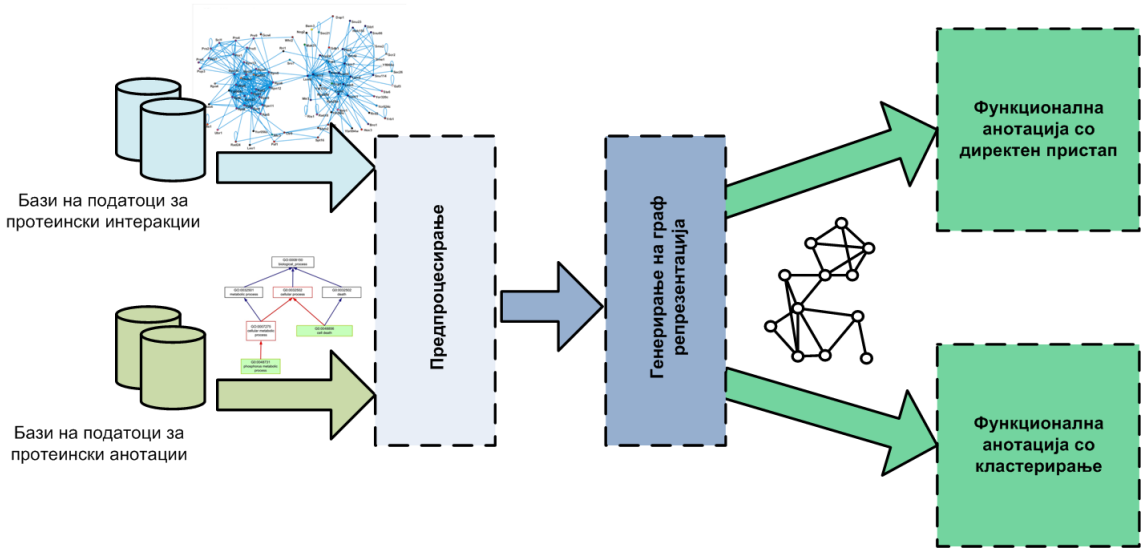
Групирањето на јазли во графот има недостаток од аспект на неговата некомпатибилност со постоењето на преклопувачки кластери т.е. ситуации во кои јазлите реално припаѓаат на повеќе од еден кластер. За ваквото преклопување се знае дека се појавува на интерфејсот помеѓу кластерите, но може да биде силно

изразено и во целиот граф. Во овие ситуации групирањето на јазлите е спорно бидејќи наметнува несакани ограничувања врз проблемот на детекција на кластери. Во нашиот конкретен случај врските во рамки на графовите со кои ја претставуваме протеинската интеракциска мрежа често соодветствуваа на еден конкретен тип на интеракција во мрежата, па според тоа истите вообичаено припаѓаат на еден единствен кластер. Доколку ги групираме врските во графот резултатот ќе зависи единствено од јазлите со кои е дефинирана врската. Тргнувајќи од ова кластерите ги дефинираме како групирање на врски наместо групирање на јазли. Врските кои припаѓаат на еден јазел можат да бидат доделени во различни кластери, па во таа смисла јазлите можат да бидат членови на различните кластери.

5

СИСТЕМ ЗА ФУНКЦИОНАЛНА АНОТАЦИЈА

Архитектурата на системот за функционална анотација во протеинските интеракциски мрежи го подразбира вкупното множество на чекори, од моментот на прибирање на влезни податоци, па се до моментот на користење на ново стекнатото знаење за протеинот на биолошки ефективен начин. Слика 5.1 дава еден генерализиран шаблон за ваквиот систем. Во рамки на оваа докторска дисертација се развиени решенија за функционална анотација со директен пристап и функционална анотација со кластерирање, кои се различни како во својата суштина така и во начинот на вршење на функционалната анотација на непознат протеин и подоцна негова анотација. Поради тоа во оваа поедноставена архитектура ваквите решенија се прикажани како два паралелни подсистеми кои се независни еден од друг и како такви можат и да се користат.



Слика 5.1 Поедноставена архитектура на систем за функционална анотација

Информациите потребни на подсистемите за функционална анотација се преземаат од два типа на бази на податоци. Во првиот тип се сместени податоци за самите интеракциски мрежи, додека пак во вториот се сместени информации за функциите на протеини кои веќе се функционално анотирани по биохемиски пат. Преземените информации не можат веднаш да се користат, туку претходно треба да се обработат. Ова е особено важно ако се знае дека базите на податоци не се унифицирани, па често не само протеините, туку и нивните функции се обележани со различни имиња во различни бази. Затоа, често е неопходна конверзија од една во друга конвенција на именување. Обработката исто така подразбира отстранување на недоверливите врски во мрежата ако е тоа возможно, обединување на повеќе податочни множества итн. Деталите за постоечките бази, проблемите кои се јавуваат и обработката на податоците беа изложени во глава 2. Деталите за податоците користени во рамки на овој труд како и нивната обработка се дадени подолу во поглавјето 5.1.

По обработката, она што се добива е интеракциска мрежа во која врските се прочистени и доверливи и секој протеин е придружен со соодветните анотации (множество на термини од некоја база на податоци за протеински анотации) и сите анотации во мрежата се меѓусебно унифицирани. Во следниот чекор од ваквата мрежа се генерираат соодветните граф репрезентации на начин како што

беше изложено во поглавјето 4.2. Како потсетување, нотациите кои ќе ги користиме во понатамошниот тек од оваа глава се следните:

- едноставен нетежински граф: $G_1(V,E)$
- тежински граф: $G_2(V,E,W)$
- протеин-термин граф: $G_3(V_3,E_3), V_3 = V \cup T, E_3 = E \cup E_T$
- комплетно функционално поврзан граф: $G_4(V,E_4,W^f), E_4 = E \cup E_f$

Ваквите графови сега можат да се преземат од соодветните подсистеми за функционална анотација и да бидат обработени зависно од типот на подсистемот. Архитектурата и начинот на работа на овие подсистеми, заедно со конкретно применетите алгоритми се во детали изложени во поглавјата 5.2 и 5.3. Резултатите од примената на овој систем се дадени во глава 6.

5.1 Агрегација и предпроцесирање на податоците за протеински интеракциски мрежи

Шумот кој се јавува кај мрежите на протеински интеракции може значително да ја промени точноста на методите за предвидување на протеинска функција. Релевантна евалуација на предложените методи е возможна само доколку избраното податочно множество нуди доволно висока веродостојност. Методите разгледувани во овој труд се тестирани и евалуирани врз податочно множество за интерактомот на лебниот квасец (Saccharomyces cerevisiae). Податочното множество претставува компилација од интеракции запишани во барем една од базите DIP, MIPS, BIND и BioGRID и се добива како резултат од агрегацијата и филтрирањето на податоците публикувани во трудови кои се сметаат за златен стандард за протеинските интеракциски мрежи. Во продолжение е дадена процедурата по чекори, според која е генерирано податочното множество:

1. Од базите за протеински податоци се превземаат податоците како што се наведени во [22] [37] [38][180] [181] [182]. Се врши нивно комбинирање и се отстрануваат дупликатите (за дупликати се сметаат оние интеракции кои се дефинирани помеѓу ист пар протеини и имаат идентична поткрепа за нивното

постоење, ако за истиот пар во различните множества постои различна поткрепа одделните записи не се отстрануваат).

2. Се креира податочно множество од протеински интеракции потврдени во литературата со преземање на сите такви протеински интеракции од BioGRID со исклучок на оние наведени во чекор 1. Од ова множество се отстрануваат сите интеракции што се поткрепени единствено со докази добиени од ко-локализација или ко-фракционирање и се отстрануваат сите РНК-протеин интеракции.
3. Податочните множества од чекорите 1 и 2 секое посебно се дели на две нови множества со интеракции и тоа едно за оние кои се единечно валидирани (ЕВ) и едно за оние кои се повеќекратно валидирани (ПВ) врз основа на следниот критериум: Една интеракција А – В се смета за ПВ ако податочното множество ја содржи реципрочната интеракција В – А или ако податочното множество ја содржи А – В барем уште еднаш, но со различна публикација или различен код за експериментална поткрепа. По дефиниција, ако интеракцијата не е ПВ ја сметаме за ЕВ.
4. Двете ЕВ множества од претходниот чекор се спојуваат во единствено податочно множество. Ваквото интегрирано ЕВ множество повторно се евалуира како во чекорот 3 и се дели на две нови ЕВ и ПВ множества.
5. ПВ множествата од чекорите 3 и 4 се спојуваат за да се добие финалното множество од протеински интеракции.

На вака добиеното податочно множество од протеински интеракции му се доделуваат анотации врз основа на поврзувањата направени помеѓу идентификаторите (ORF) на протеините во рамки на множеството и дефинираните анотации за тие идентификатори во рамки на SGD базата на податоци. Анотациите во рамки на SGD се според GO. На ваквото податочно множество му се врши дополнителна обработка:

1. Од сите функционални анотации се отстрануваат тривијалните, како на пример: “unknown cellular compartment”, “unknown molecular function” и “unknown biological process” (било каква анотација која нема значење на конкретен термин од GO).

2. Се пресметуваат нови анотации на секој протеин според принципот на транзитивно затворање (transitive closure), така што на секој протеин со функционална ознака *v* му се припишуваат и сите ознаки *u* со кои *v* е во isA релација во GO хиерархијата. На овој начин секој протеин ги добива и ознаките кои имаат поопшто значење во GO хиерархијата.
3. Од множеството се отстрануваат функционалните ознаки кои се среќаваат повеќе од 300 пати, бидејќи овие многу фреквентни анотации се најчесто многу општи поими во GO хиерархијата и не носат значајна информација.

Финалното множество кое се добива е високо доверливо и се состои од 2502 протеини меѓу кои се забележани 12708 интеракции, а се анотирани со вкупно 888 функционални ознаки. Мрежата на протеински интеракции не претставува поврзан граф, туку се состои од повеќе компоненти, од кои најголемата содржи 2146 протеини. Ваквото множество е користено во сите експерименти извршени во рамки на оваа докторска дисертација, односно за добивање на соодветните граф репрезентации кои претставуваат влез во подсистемите за функционална анотација (деталите за различните графови и нивните параметри се дадени во поглавјето 4.2).

5.2 Подсистем за функционална анотација со директен пристап

Сите директни методи за предвидување на протеинска функција поминуваат низ исти етапи во процесот на анотација. Како прв чекор, треба да се одреди еден или повеќе неанотирани или прашални протеини, чија анотација ќе биде цел на алгоритамот. Токму ова множество на прашални протеини е еден од основните влезни параметри на голем дел од алгоритмите. Потоа, се извршува соодветниот алгоритам врз непосредното или поширокото соседство на прашалниот протеин, или пак целиот граф. Она што се добива како резултат на алгоритамот обично е скала со оценки (кои можат да бидат веројатности или некоја друга форма на проценка) за погодноста на секоја од функциите да биде доделена на прашалниот протеин.

Методот кој што е предложен во овој труд како влезен параметар има еден прашален протеин, q, чија функционална анотација се зема како непозната. Тргнувајќи од тој протеин, со помош на алгоритамот за случајно изминување на граф, попознат како алгоритам на случајна патека (random walk), се формира профил на соседството на тој протеин, каде под профил се подразбира оценка за влијанието на секој протеин врз функционалноста на прашалниот протеин. Ако со терминот функционално соседство се означат протеините кои, според својот оценка, се блиску до прашалниот протеин, ќе се забележи дека ова функционално соседство има многу поголем домен од само најблиските соседи на протеинот. Како последен чекор од алгоритамот е доделување на оценки на расположливите функции, во зависност од оценките на протеините од функционалното соседство на прашалниот протеин, кои се анотирани со тие функции.

Нека почетниот јазел на случајната патека на графот е обележан со v_0, а позицијата на случајниот пешак во моментот t е обележан со v_t. Всушност, v_t е случајна променлива која може да прима вредности од 1 до $|V|$. Нека $k(i)$ е функција која го дава степенот на јазелот i. Нека со P_t е обележан векторот на веројатности дека случајниот пешак ќе се најде на секој од јазлите во графот, односно веројатностната распределба за случајната променлива v_t. Математички кажано, за P важи $P_t(i) = \text{Prob}(v_t = i)$. Почетната веројатносна распределба P_0, во општ случај, е униформна за сите јазли од графот, бидејќи почетната позиција на случајниот пешак може да биде било која. Ако секој од јазлите во графот се смета за состојба во Марков процес, тогаш нека со $M = [\, p_{ij} \,]_{i,j \in V}$ е обележана матрицата на веројатности на транзиција од една во друга состојба. Во зависност од тоа каков е графот со кој ја репрезентираме протеинската интеракциска мрежа дефиницијата на оваа матрица ќе биде различна. Кога графот е нетежински (G_1 или G_3) во секој нареден чекор случајниот пешак случајно го одбира следниот јазел од листата на соседи на јазелот на кој моментално се наоѓа со униформна веројатност, додека за тежинскиот граф веројатноста за преминување на некој од соседите на тековниот јазел е еднаква на тежината на врската помеѓу тие два јазли. Веројатноста за премин ќе се пресметува во зависност од типот на графот:

- За G_1: $$p_{ij} = \begin{cases} 1/k(i), ако (i,j) \in E \\ 0, во спротивно \end{cases} \quad (5.1)$$

- За G_2: $$p_{ij} = w_{ij} \quad (5.2)$$

- За G_3: $$p_{ij} = \begin{cases} 1/k(i), ако (i,j) \in E_3 \\ 0, во спротивно \end{cases} \quad (5.3)$$

- За G_4: $$p_{ij} = w_{ij}^f \quad (5.4)$$

Земајќи го во предвид претходно кажаното, веројатносната распределба P_t во моментот t може да се пресмета со рекурентната равенка дадена со:

$$P_t = M^T P_{t-1} \quad (5.5)$$

Доколку со P_0 се обележи почетната веројатносна распределба пешакот да се наоѓа на некој јазел од графот, тогаш веројатносната распределба во моментот t може да се пресмета уште и со:

$$P_t = (M^T)^t P_o \quad (5.6)$$

Во општ случај, во секој момент од времето, распределбата на случајната променлива v_t е различна. Тоа значи дека веројатноста случајниот пешак да се најде на определен јазел се менува во секој момент од времето. Но, математички може да се докаже дека распределбата P_t на случајната променлива v_t кај небипартитен граф (каква што всушност е мрежата на протеински интеракции) се стреми кон стационарна ако $t \to \infty$. Со други зборови, после извесен број на чекори t, $P_t = P_{t+1}$. Токму стационарната состојба на векторот на распределбата е од интерес за алгоритамот за функционална анотација.

Една важна модификација на методот за случајна патека кај граф е можноста за повторен почеток (random walk with restart). Имено, во оваа варијанта на алгоритамот додадена е можноста случајниот пешак во одреден момент да се телепортира назад, на почетниот јазел. Всушност, во секој момент од времето t, случајниот пешак може да се телепортира на почетниот јазел со веројатност c или пак да продолжи по некоја од врските на јазелот v_t со веројатност 1-c. Во овој

случај, равенката за векторот на веројатносна распределба P_t ќе добие нов облик, како во формулата:

$$P_t = (1-c)M^T P_{t-1} + cr \tag{5.7}$$

каде r е вектор за кој важи дека елементот на позиција v_0 има вредност единица, а сите останати елементи се нули.

За да може ваквиот алгоритам да се искористи за функционална анотација потребно е да се направат определени адаптации. Најпрво, како почетна позиција v_0 на случајниот пешак се зема јазелот од графот кој го означува неанотираниот, односно прашалниот протеин q. Истовремено, се пресметува и матрицата M. Дополнителен влезен параметар е почетниот вектор на веројатносна распределба на v_t, P_0. Во овој случај, со оглед на тоа што точно е позната конкретната почетна позиција на случајниот пешак, $P(v_0 = q) = 1$. Јасно е дека веројатностите за сите останати состојби на случајниот пешак во почетниот момент се еднакви на 0. Конечно, како влезен параметар треба да се проследи и веројатноста за телепортирање на почетната позиција, c.

Клучниот излезен параметар од алгоритамот е векторот на веројатносна распределба P_t во неговата стационарна состојба. Имено, овој вектор може да се интерпретира како мерка за блискоста на почетниот јазел до сите останати јазли во мрежата. Во проблематиката на предвидување на протеинските функции, овој вектор е мерка за афинитетот на неанотираниот протеин кон сите други јазли во мрежата, изведена единствено врз основа на структурата на мрежата. Токму овој вектор претставува функционален профил на соседство на прашалниот протеин.

Секако дека со итеративниот алгоритам за методот на случајна патека на граф, кој во псевдо код сумарно е даден на Слика 5.2, не може да се стигне до конечната стационарна состојба на векторот на веројатносна распределба, каде $P_t = P_{t+1}$. Заради тоа, се одбира прагова вредност ε, а критериумот за конвергенција на векторот P_t се смета дека е задоволен ако е задоволено неравенството во формулата (5.8). Во овој труд беше експериментирано со вредности за ε од ред на големина од 10^{-12} до 10^{-5}.

$$| P_t - P_{t+1} | \leq \varepsilon \quad (5.8)$$

Влез: граф репрезентација на протеински интеракции G_i, i=1,..,4;
почетен јазел q;
веројатност c за телепортирање на почетниот јазел ;
Излез: вектор на стационарна распределба P_t

Иницијализација: матрица на веројатност на транзиција M;
почетен вектор на распределби P_0

Се додека P_t не конвергира
$$P_t = (1-c)M^T P_{t-1} + cr$$

Слика 5.2 Псевдо код за алгоритамот за случајна патека во граф

Важен елемент при извршувањето на алгоритамот за случајна патека во граф со повторен почеток е изборот на вредноста за веројатност за телепортирање c. Со намалување на вредноста на c се намалува и дијаметарот на соседство на прашалниот јазел кое се зема во предвид, па јазлите на поголемо растојание од него ќе имаат мала и скоро незначителна вредност во профилот на соседство. Со зголемување на c се дава поголемо значење и на овие пооддалечени протеини. Изборот на с влијае и на брзината на конвергенција на алгоритамот, така што поголемо c значи и побавно движење кон стационарната состојба.

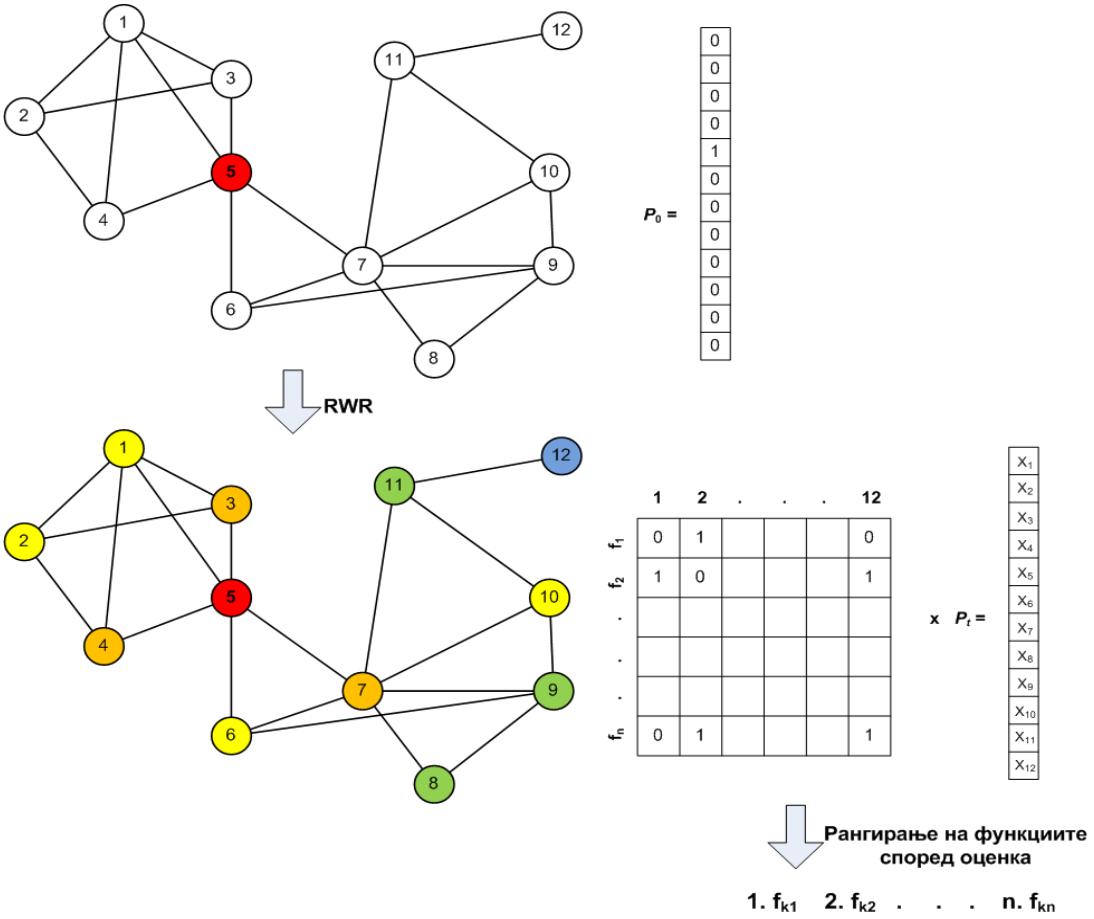
Втората фаза на овој алгоритам е како од профилот на соседство за прашалниот протеин да се извлечат протеинските функции кои тој би ги имал. Предложени се и тестирани два можни пристапи за решавање на овој проблем. Притоа овие пристапи важат кога како влез на алгоритамот за случајна патека се пропушти еден од G_1, G_2 или G_4. Поради својата специфична природа графот G_3 мора да се третира поинаку.

Првиот, поедноставен и поинтуитивен пристап, го разгледува комплетниот вектор на стабилна распределба, односно секој јазел во графот придонесува во крајната оцена преку своите анотации земени со фактор кој одговара на стабилната распределба. Оценката на секоја функција се пресметува според формулата (5.9), а потоа добиените оценки се нормализираат во опсег од 0 до 1.

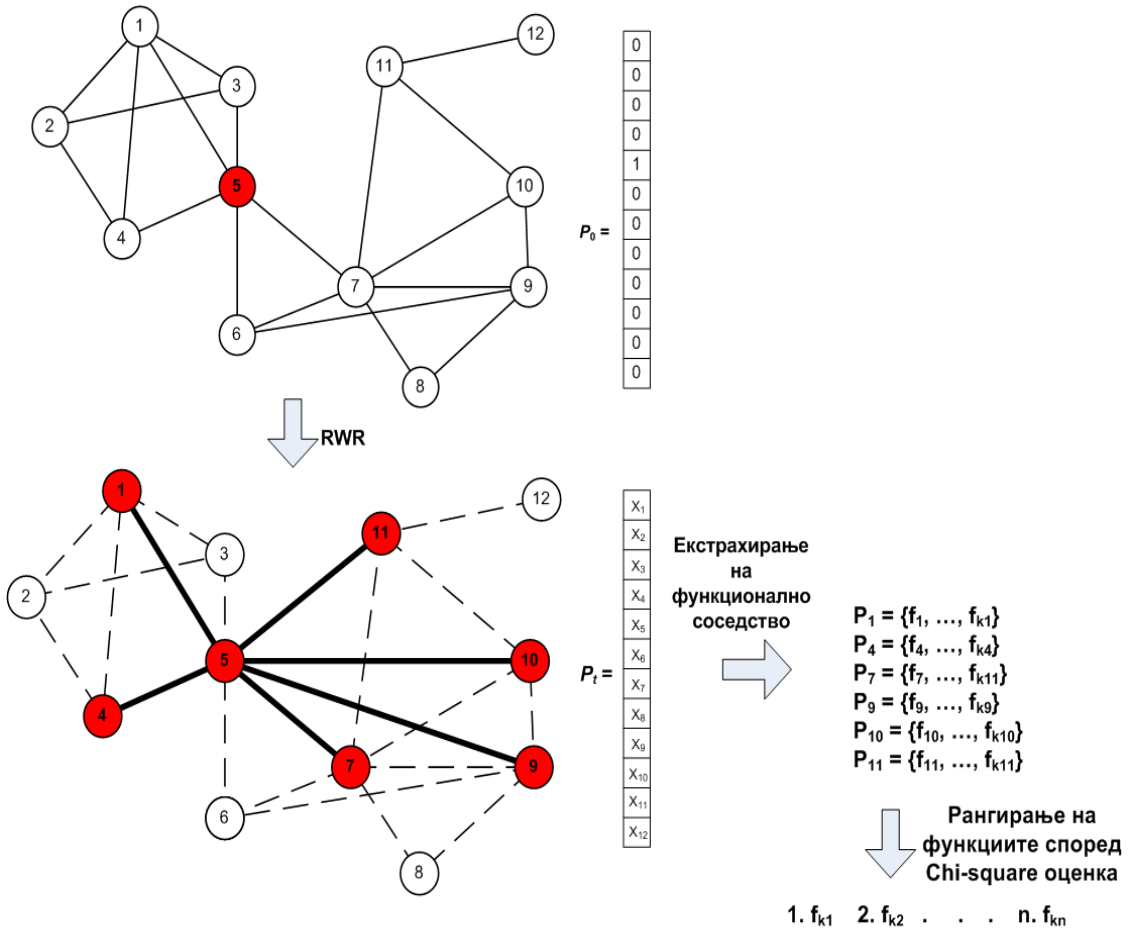
$$f(j)_{j\in F} = \sum_{i\in V} z_{ij} P_t(i) \tag{5.9}$$

каде $P_t(i)$ е вредноста на јазелот i во стабилната состојба на векторот на распределба, F го претставува целокупното множество на функции (термини) во графот, а матрицата $Z = [\, z_{ij} \,]_{i\in V, j\in F}$ ја содржи информацијата за функциите (терминте) што им се придружени на јазлите во графот и важи $z_{ij} = 1$ ако на јазелот i му е придружена функцијата j.

Целокупниот процес за предвидување на протеинска функција, од специфицирање на влезни променливи до добивање на конечниот резултат според првиот пристап на оценување на функциите со проста сума, сликовито е прикажан на слика 5.3.



Слика 5.3. Функционална анотација со користење на алгоритамот за случајна патека кога во предвид се земаат сите протеини



Слика 5.4 Функционална анотација со користење на алгоритамот за случајна патека со ограничено функционално соседство

Вториот пристап и доделува важност на секоја функција со помош на χ^2 тест. Најпрво, од функционалниот профил на соседство на прашалниот јазел се избира подмножество од јазли со највисоки оценки. Бројот на избрани јазли е променлив во зависност од поставената прагова вредност за оценката на јазелот. На тој начин се формира ограничено функционално соседство на јазли кои имаат најголемо влијание врз функционалната анотација на прашалниот јазел. Потоа, секоја функција добива оценка во зависност од фреквенцијата на нејзино појавување во функционалното соседство, но и очекувањето таа да се појави во ограниченото функционално соседство врз основа на фреквенцијата на нејзино појавување во целиот граф. Формулата (5.10) ја дава равенката за овој χ^2 тест. При тоа, $n(j)$ е бројот на јазли кои се припаѓаат во ограниченото функционално соседство на прашалниот јазел и ја поседуваат функцијата j, додека $e(j)$ е математичкото

очекувањето колку пати таа функција би требало да се јави во функционалното соседство на протеинот.

$$f(j)_{j\in F} = \frac{(n(j) - e(j))^2}{e(j)} \tag{5.10}$$

На Слика 5.4. прикажано е како се одвива предвидувањето на протеинската функција според вториот пристап.

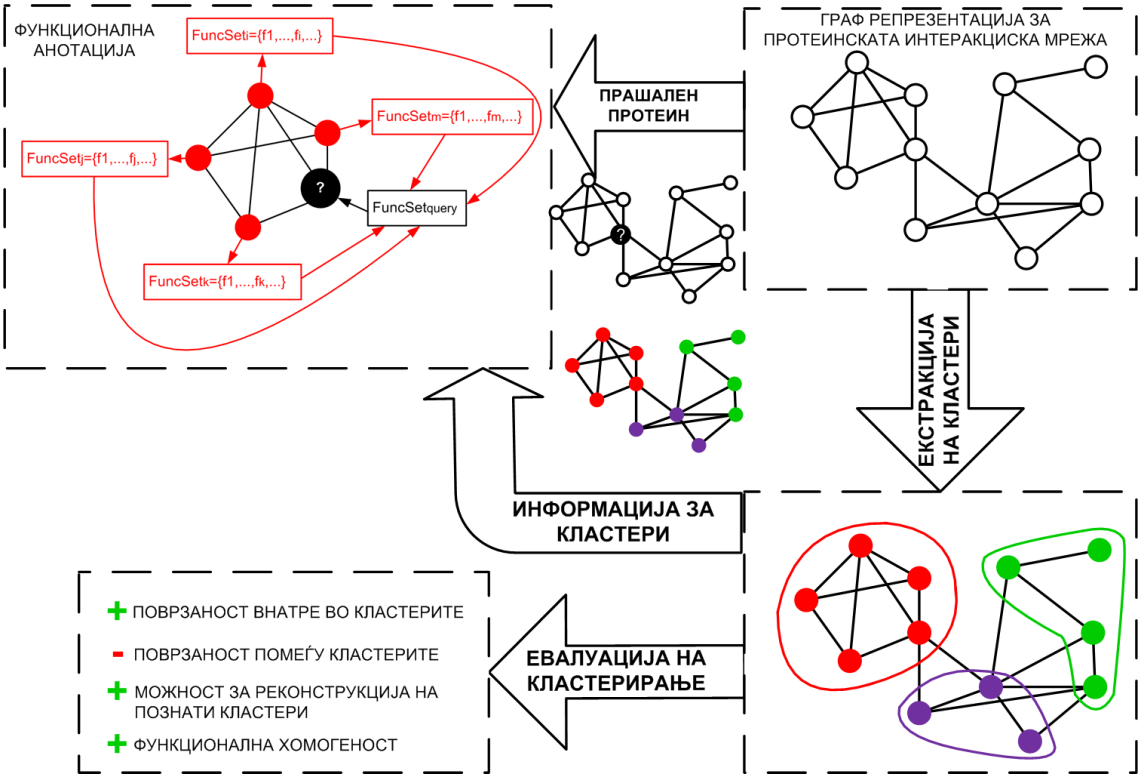
Кога работиме со графот G_3 треба да имаме на ум дека дел од јазлите на истиот се всушност термините (функциите) со кои јазлите (протеините) се анотирани. Всушност овде не е потребно да прават никакви оценки за тоа колку некој термин би соодветствувал за прашалниот протеин затоа што врската на различните термини и прашалниот јазел е директно содржана во информацијата за стационарната распределба на случајното изминување. Односно, во овој случај важи:

$$f(j)_{j\in F} = P_t(j) \tag{5.11}$$

5.3 Подсистем за функционална анотација со кластерирање

Без оглед на тоа која техника на кластерирање ќе биде искористена архитектурата на овој подсистем е составена од три делови како што е прикажано на слика 5.5.
Првиот дел или првиот чекор е да се подели графот на протеинската интеракциска мрежа на кластери преку примена на некоја техника за кластерирање која ќе ги искористи информациите содржани во соодветниот граф. Во овој дел без никакво ограничување може да се примени било која техника од претходно изложените во поглавјето 4.3.2. Изборот на техниката исто така никако не влијае на следните чекори. Вториот чекор е евалуација на кластерите и колку е добра структурата на кластерите и тоа од аспект на можност за реконструкција на познати кластери и функционална хомогеност на кластерите. Високата поврзаност внатре во кластерот и малата поврзаност помеѓу различните кластери е инхерентно својство на самите техники за кластерирање. Функционалната анотација на непознат

протеин врз основа на добиените кластери е задачата на третиот чекор од овој подсистем.



Слика 5.5 Архитектура на подсистемот за функционална анотација со кластерирање

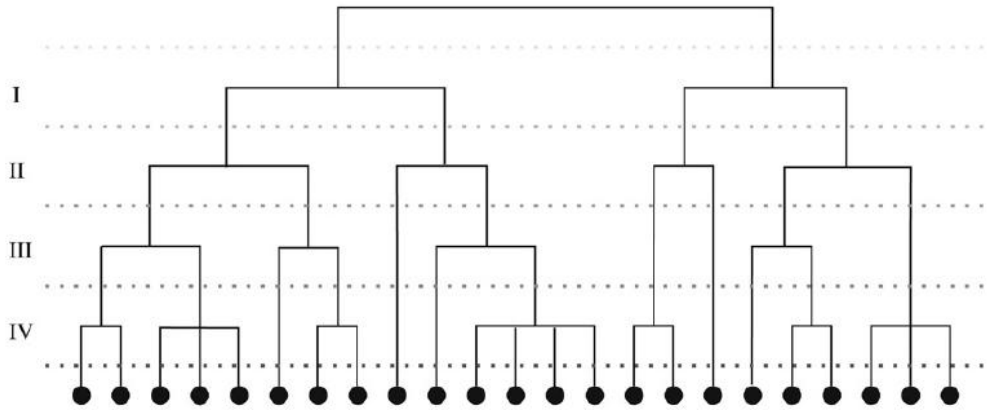
5.3.1 Екстракција на кластери

Во ова поглавје ќе ги дадеме деталите за алгоритмите за кластерирање имплементирани во рамки на оваа докторска дисертација, по еден претставник на секоја од класите како што беа дефинирани во поглавјето 4.3.2. Притоа треба да се напомене дека ниту еден од овие алгоритми досега во литературата не е применет во функционална анотација во протеински интеракциски мрежи. Дел од алгоритмите како филозофија се општо познати и применувани во многу области (поглавјата 5.3.1.1-5.3.1.4) меѓутоа деталите и конкретната имплементација за протеинските интеракциски мрежи се овде се оригинални. Дел од алгоритмите се издвоени како најдобри во литературата за кластерирање во комплексни мрежи (поглавја 5.3.1.5-5.3.1.8) меѓутоа никогаш не се применети за функционална

анотација во протеински интеракциски мрежи во сценарија како што се предложени овде. Предложен е и еден алгоритам (поглавје 5.3.1.9) кој претставува еден вид комбинација помеѓу овие две групи, односно алгоритам кој користи мерка за квалитет на кластерирање дефинирана според концепти од првата група и оптимизација на таа мерка според она што постои во втората група.

5.3.1.1 Агломеративно хиерархиско кластерирање

Агломеративното хиерархиско кластерирање започнува со тоа што секој јазел (протеин) се доделува во одделен кластер. Потоа во итеративен процес, парот од кластери кои се меѓусебно најслични се спојува и формира еден нов кластер. Процес трае сè до моментот на добивање на еден единствен кластер. Хиерархиската структура на финалниот кластер подразбира дека иницијално креираните кластери се „вгнездени“ во оние кои се креирани во следните фази од алгоритамот, при што големината на кластерите содржани во финалниот кластер може да биде променлива. Ваквата структура вообичаено се претставува како дендограм (слика 5.6). Пресекот направен на определено ниво од ваквото дрво дава едно можно кластерирање.



Слика 5.6 Илустрација на дендрограм за агломеративно хиерархиско кластерирање

За да можеме да го имплементираме агломеративното хиерархиско кластерирање потребно е графот да го претставиме во некој метрички простор за да можеме да ги дефинираме поимите на сличност и растојание на кластерите. За таа цел ќе го искористиме алгоритамот за случајно изминување на патека како што беше

дефиниран по претходното поглавје со равенките (5.1)-(5.4) и (5.7) и псевдо кодот даден на слика 5.2. Разликата во однос на претходната примена на овој алгоритам е во тоа што овде го применуваме за секој јазел *i* од графот, односно добиваме по еден стационарен вектор P_t^i за секој можен почетен јазел. Ваквите вектори укажуваат на „сличноста“ помеѓу јазлите, односно нивниот меѓусебен афинитет. Она што е потребно за агломеративното хиерархиско кластерирање е мерка за растојание, па имајќи во предвид дека вредностите во стационарниот вектор на распределба (или сличностите) не можат да бидат поголеми од 1 ја дефинираме следната мерка за растојание:

$$d'(i,j)=1-P_t^i(j) \tag{5.12}$$

Бидејќи графот на протеинската интеракциска мрежа е нерегуларен ќе важи $d'(i,j)\neq d'(j,i)$, па затоа дефинираме ново растојание:

$$d(i,j)=\frac{d'(i,j)+d'(j,i)}{2} \tag{5.13}$$

Врз основа на ова растојание помеѓу јазлите можеме да го дефинираме растојанието помеѓу два кластери *A* и *B*:

$$D(A,B)=\frac{1}{|A||B|}\sum_{i\in A}\sum_{j\in B}d(i,j) \tag{5.14}$$

Врз основа на ова растојание кластерирањето се врши според алгоритамот даден на слика 5.7.

Последниот чекор во хиерархиското кластерирање е да се одреди нивото на кое ќе се направи сечењето при што множеството кластери дефинирани на тоа ниво е крајниот резултат од кластерирањето. Овде е применет алгоритамот предложен во [183]. Алгоритамот имплементира адаптивен, итеративен процес на декомпозиција и комбинација на кластери и запира кога бројот на кластери станува стабилен. Започнува со земање на неколку големи кластери дефинирани на 90% од длабочината на дендограмот. Се анализираат длабочините (или нивоата) на спојување на секој кластер со цел да се добијат карактеристични шаблони на флуктуација што би укажувале на постоење на „под-кластерска“ структура и кластерите кај кои се појавуваат ваквите шаблони рекурзивно се

делат. За да се избегне прекумерно делење, многу малите кластери се спојуваат со нивните соседни поголеми кластери.

Влез: граф репрезентација на протеински интеракции G_i, i=1,..,4;

Излез: структура на кластери за агломеративно кластерирање

Иницијализација: ниво на кластерирањето $l = 0$;
секој јазел k поставен во посебен кластер C_k^o;
пресметка на $d(i,j), \forall i,j \in V$;

Се додека $count(C^l) > 1$ повторувај [C^l *е множество на кластери на ниво* l]
$Min = \min_{i \neq j} D(C_i^l, C_j^l)$;
Ако $D(C_A^l, C_B^l) = Min$ и $\exists a,b(a \in C_A^l \wedge b \in C_B^l \wedge (a,b) \in E)$
[*дополнителен услов е новиот кластер да биде поврзан*]
$C_{A,B}^l = спој(C_A^l, C_B^l)$; [*спојувањето става маркер на кластерите што се спојуваат*]
l++; [*зголеми го нивото*]
$C^l \rightarrow C^{l+1}$ [*зголеми го нивото на кластерите без маркер од тековното ниво*]

Слика 5.7 Псевдо код за алгоритамот за агломеративно хиерархиско кластерирање во граф

Предноста кај агломеративното хиерархиско кластерирање е во добивање на подреденост кај ентитетите што се кластерираат и променливата големина на кластерите кои може да се креираат. Ова може да се искористи за лесно индексирање и пребарување на ентитетите. Главен недостаток кај ова кластерирање е неможноста за префрлање на ентитетите од еден кластер во друг.

5.3.1.2 Кластерирање со k-медоиди

Алгоритамот за кластерирање со k-медоиди е една од верзиите на алгоритамот за кластерирање со k-средини, со тоа што овде прототипот (медоидот) на секој кластер е еден од ентитетите што припаѓаат на кластерот, за разлика од алгоритамот со k-средини каде прототипот (центроидот) е усреднување на ентитети што припаѓаат на кластерот. Кластерирањето со k-медоиди може да се

смета за дискретна верзија на кластерирањето со *k*-средини. Ваквата промена значи дека се зголемува отпорноста на шум што се манифестира како ентитети кои се многу различни (оддалечени) од останатите ентитети кои се кластерираат (outliers).

Бидејќи алгоритамот работи на принцип на споредување на јазлите кои ги кластерира потребно е да дефинираме мерка за растојание помеѓу јазлите. Можеме да ја користиме истата мерка која ја дефиниравме кај агломеративното хиерархиско кластерирање, но сепак за да го покажеме влијанието на различните репрезентации на граф кои се користат ќе дефинираме нова мерка. За да го направиме тоа најпрво треба да го дефинираме поимот на најкратка патека. Под најкратка патека помеѓу два јазли i и j, $SP(i, j)$ ќе го подразбираме множество на врски (m, n) такви што почнувајќи од јазелот i и движејќи се по секоја од овие врски можеме да стигнеме до јазелот j, и притоа важи дека цената на чинење на патеката е минимална. Ако земеме дека секоја врска има определен отпор што треба да се совлада, при што за тежинските графови тој отпор е обратно пропорционален со тежините на врските (врската со максимална вредност на тежината има отпорност 0), а кај нетежинските графови можеме да го земеме да биде еднаков на 1 (максималната вредност на тежините кај тежинскиот граф), тогаш цената на чинење е право пропорционална на вкупниот отпор на патеката или поинаку кажано најевтина е патеката со најмала вкупна отпорност. Најкратката патека ќе го дефинира растојанието (што е исто што и вкупниот отпор на патеката) кое ќе го користиме за кластерирањето со *k*-медоиди, при што неговата пресметка се разликува во зависност од тоа која граф репрезентација ја користиме:

- За G_1 и G_3: $$d(i, j) = \left|SP(i, j)\right| \tag{5.15}$$
- За G_2: $$d(i, j) = \sum_{(m,n)\in SP(i,j)} (1 - w_{mn}) \tag{5.16}$$
- За G_4: $$d(i, j) = \sum_{(m,n)\in SP(i,j)} (1 - w_{mn}^{f}) \tag{5.17}$$

Во пресметката за тежинските графови (5.16) и (5.17) поради тоа што тежината на врската е одраз на сличноста на јазлите што ги поврзува мора да се земе со негативен предзнак, односно да се одземе од нејзината теоретски максимална вредност (во нашиот случај тоа е 1).

На слика 5.8 е даден алгоритамот за кластерирање на граф со *k*-медоиди. Еден од главните недостатоци на ваквиот алгоритам е секако изборот на бројот на кластери, вредност што мора да се зададе однапред и единствен начин за нејзино приближно определување е со емпириско тестирање.

Влез: граф репрезентација на протеински интеракции G_i, i=1,..,4;
број на кластери k;

Излез: содржина на кластери C_k;

Иницијализација: Случаен избор на k јазли за медоиди M_k на кластерите;
Пресметка на матрица на растојанија;

Повторувај
секој јазел i додели го на кластер C_j ако важи
$\arg\min_j d(i, M_j)$ [*најблискиот кластер*]
пресметај ги новите медоиди M_k
$\arg\min_{M_k} \sum_{i \in C_k} d(M_k, i)$
[*за медоид избери го оној јазел што е на најмало вкупно растојание од останатите во кластерот*]
додека не настане стационарна состојба;
[*сите медоиди останат исти*]

Слика 5.8 Псевдо код за алгоритамот за кластерирање на граф со k-медоиди

5.3.1.3 Спектрално кластерирање

Првиот чекор во спектралното кластерирање е трансформација на иницијалното податочно множество во множество од точки во n-димензионален простор, чијшто координати се елементите на n-те избрани сопствени вектори. Оваа промена во репрезентацијата на податоците ги засилува карактеристиките на кластерите правејќи ги поизразени. Откако ќе биде извршена трансформацијата

може да се примени било кој класичен алгоритам за кластерирање, како што е на пример кластерирањето со *k*-средини.

Го воведуваме поимот на матрица на сличност S која е симетрична матрица и се однесува на сличноста на јазлите во рамки на графот. Во зависност од тоа за каков граф станува збор дефиницијата на матрицата на сличност се разликува, односно за нетежинските графови G_1 и G_3 таа ќе биде еднаква на матрицата на соседство A нормализирана по колони, а за тежинските графови G_2 и G_4 ќе биде еднаква на тежинските матрици W и W^f, соодветно. Со вака дефинираната матрица на сличност можеме да ја дефинираме Лапласовата матрица L на графот според формулата:

$$L = D - S \tag{5.18}$$

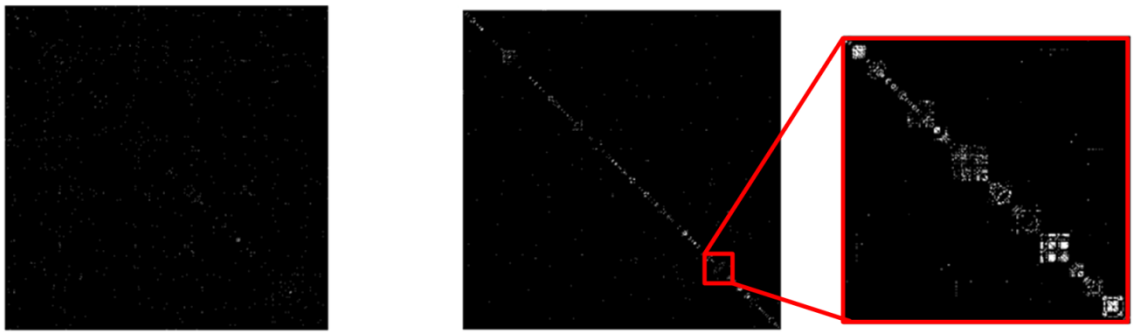
каде D е дијагонална матрица чијшто елемент на главната дијагонала d_{ii} е еднаков на степенот на јазелот i во графот. Потоа ги пресметуваме првите k сопствени вектори $u_1,\ldots,u_k$ дефинирани со

$$Lu = \lambda Du \tag{5.19}$$

Главна карактеристика на Лапласовата матрица на графот е фактот дека бројот k на нулеви сопствени вредности е еднаков на бројот на поврзани компоненти во графот. Ако колоните и редиците во Лапласовата матрица се преуредат на тој начин што јазлите кои припаѓаат на една поврзана компонента се во соседни редици и колони, тогаш во матрицата се добиваат блокови наредени по дијагоналата, така што голем дел од елементите во еден блок имаат вредност различна од нула, а сите останати елементи од секоја редица и колона се еднакви на нула (јазелот е поврзан само со јазли од сопствената компонента). Ваквата Лапласова матрица ќе има k сопствени вектори кои ги репрезентираат кластерите и чии ненулеви компоненти ќе одговараат на индексите на јазлите кои припаѓаат на соодветниот кластер. Ако таквите сопствени вектори се постават како колони во една $U^{|V|\text{x}k}$ матрица, секој ред ќе одговара на еден јазел и ќе има само една ненулева вредност: на позицијата од сопствениот вектор за поврзаната компонента на која тој јазел припаѓа.

Ако графот има само една поврзана компонента, тогаш Лапласовата матрица ќе има само една нулева сопствена вредност (како што е во нашиот случај). Нека бројот на кластери на кои треба да се поделат јазлите во графот биде еднаков на *k*. Земајќи ги *k*-те сопствени вектори кои соодветствуваат на *k*-те сопствени вредности најблиски до 0, и трансформирајќи ги јазлите од графот во *k*-димензионалниот простор што тие го формираат, сите јазли кои припаѓаат на ист кластер ќе се наоѓаат блиску еден до друг во ваквиот простор. Новиот *k*-димензионален простор е претставен со матрицата *U* и секој јазел претставува една редица во оваа матрица. Овие |*V*| „точки“ можеме да ги кластерираме со стандарден алгоритам за кластерирање со *k*-средини.

Доколку се преуреди матрицата на сличност така што јазлите од ист кластер да се најдат во соседни редици и колони ќе се види блоковска структура по дијагоналата на матрицата на сличност. Сепак, блоковите нема да се чисти, односно во некои редици и колони ќе постојат елементи со ненулеви вредност надвор од блокот, и тие ќе ги репрезентираат врските помеѓу јазли кои, според ова кластерирање, припаѓаат на различни кластери. На сликата 5.9А даден е оригиналниот изглед на матрицата на сличност. После кластерирањето и преуредувањето, матрицата на сличност изгледа како на слика 5.9В. Тука јасно може да се види блоковската структура по дијагоналата на матрицата, што јасно ја отсликува поделбата на јазлите на кластери.



Слика 5.9 Визуелен приказ на матрица на сличност А) пред кластерирање и В) после преуредување според кластерирање базирано на спектрална анализа

На слика 5.10 е даден псевдо кодот за алгоритамот. Променлив параметар на овој алгоритам е бројот на сопствени вредности k од кој директно зависи бројот на кластери кои ќе се добијат, а тој може да се добие емпириски, преку испитување на квалитетот на кластерирањето со различен број на сопствени вектори. Алтернативен пристап е бројот на кластери да не биде однапред определен, туку да се постави некоја гранична вредност ε за сопствените вредности на Лапласовата матрица. Потоа, во предвид се земаат само оние сопствени вредности кои не ја надминуваат таа гранична вредност, а бројот на вакви сопствени вектори го дава и бројот на кластери. Граничната вредност ε може да биде поставена и за максимална дозволена разлика помеѓу две соседни по големина сопствени вредности. Во ваков случај, во игра влегуваат најмалите k сопствени вредности, за кои важи дека $\lambda_i - \lambda_{i-1} < \varepsilon$, $i < k$ и $\lambda_{k+1} - \lambda_k \geq \varepsilon$.

Влез: граф репрезентација на протеински интеракции G_i, i=1,..,4;
број на кластери k;

Излез: содржина на кластери C_k;

Иницијализација: Матрица на сличност S;
Дијагонална матрица D (d_{ii} = степен на i)

Пресметка на ненормализирана Лапласова матрица $L=D-S$

Пресметка на првите k сопствени вектори $u_1,..., u_k$ за сопствениот проблем $Lu = \lambda Du$

Конструкција на матрица $U \in R^{|V| \times k}$ која ги содржи векторите $u_1,..., u_k$ како колони

Нека $y_i \in R^k$ е векторот што соодветствува на i-тата редица од матрицата U

Кластерирање на точките $(y_i)_{i=1,...,|V|}$ во R^k со алгоритамот со k-средини во кластери $C_1, ..., C_k$

Слика 5.10 Псевдо код за спектрално кластерирање

5.3.1.4 Кластерирање базирано на средишност на врски

Идејата за средишност на врски потекнува од [7] и е проширување на поимот средишност на јазли воведен уште многу пред тоа. Средишност на јазел е оценка

за тоа колку еден јазел во мрежата зазема централно место и кое е неговото целокупно влијание во мрежата. Средишност на јазелот *i* претставува број на најкратки патеки помеѓу други јазли во мрежата SP (како што беа дефинирани во 5.3.1.2), што поминуваат низ тој јазел. Средишност на врска по аналогија се дефинира како број на најкратки патеки помеѓу два јазли, кои поминуваат низ таа врска.

Интересна карактеристика на врските во рамки на еден граф е тоа што оние врски кои се наоѓаат помеѓу кластери, т.е. кои поврзуваат две точки од различни кластери во мрежата имаат поголема средишност отколку оние кои поврзуваат јазли од ист кластер. Ова произлегува од фактот што низ врските кои се наоѓаат помеѓу кластерите поминуваат многу најкратки патеки, односно сите оние помеѓу јазел од првиот и јазел од вториот кластер. Со бришење на врските со најголема средишност, после одреден број на итерации ќе се добие неповрзан граф, односно графот ќе се раздели на повеќе компоненти кои можат да се третираат како различни кластери. На слика 5.11 е даден псевдо кодот за алгоритамот за кластерирање базирано на средишност на врски (EdgeBetweenness).

Влез: граф репрезентација на протеински интеракции G_i, i=1,..,4;
број на врски што треба да се избришат d;

Излез: поврзани компоненти од графот C_k;

Додека $d > 0$ повторувај

Пресметај средишност на сите врски;
[*Пресметај ги сите најкратки патеки* $SP(i, j), \forall i, j \in V$]

Отстрани ја врската со најголема средишност;
[*Отстрани ја врската* (i, j) *која се јавува најмногу во* $\bigcup_{i, j \in V, i \neq j} SP(i, j)$]

$d = d - 1$;

Слика 5.11 Псевдо код за кластерирање базирано на средишност на врски

Влезен параметар кој може да се менува е бројот на врски кои треба да се избришат. Се разбира, не може директно и однапред да се предвиди бројот на врски кои треба да се избришат за да се добијат доволно добри кластери, па овој број се добива емпириски. Вкупната комплексноста на алгоритамот е $O(|E|^2|V|)$, каде $|V|$ е број на јазли, а $|E|$ е број на врски во графот. При тоа, средишноста на секоја врска повторно се пресметува после секоја итерација, бидејќи со бришење на една врска, автоматски се менуваат и најкратките патеки помеѓу јазлите. Стратегијата за пресметување на средишноста на врските само еднаш и тоа на почетокот на алгоритамот, а потоа директно бришење на оние со најголема средишност би била лоша во случај кога помеѓу два кластери постои повеќе од една врска. Во таквиот граф на почетокот голема е веројатноста дека само една од нив ќе има голема средишност, што значи дека останатите нема да бидат избришани и кластерите нема да се раздвојат.

Специјален случај на алгоритамот имам кога истиот работи со граф репрезентацијата $G_3(V_3, E_3)$. Имено овде имаме различни типови на врски, па нивното отстранување не е еднозначно како кај останатите репрезентации. По дефиниција за овој граф важи $E_3 = E \cup E_T$, односно вкупното множество на врски претставува унија од врските што постојат помеѓу пар од јазли протеин-протеин (E) и врските што постојат помеѓу пар јазли протеин-термин (E_T). Јазлите термини во ваквиот граф имаат многу висок степен и за очекување е дека во ваквиот пристап средишноста на врските во E_T ќе биде голема. Меѓутоа отстранувањето на врска од множеството E_T може да се толкува како „де-анотирање“ на соодветниот јазел протеин од парот дефиниран со отстранетата врска, односно губење на информацијата потребна во чекорот што следува после кластерирањето (определување на функција на прашален протеин врз основа на добиените кластери). Токму поради тоа кога работиме со G_3 чекорот од алгоритамот кој прави отстранување на врската со најголема средишност се дополнува со условот врската да биде од типот протеин-протеин.

5.3.1.5 Кластерирање со алчна оптимизација на модуларноста

Оптимизација на модуларноста е NP комплетен проблем и неговата временска комплексност расте полиномно во зависност од големината на графот, што би значело дека е практично невозможно да се најде точно решение на проблемот. Во рамки на овој докторски труд се прифатени и адаптирани два алгоритми за кластерирање кои се базираат на алчна оптимизација на модуларноста.

Алгоритамот "Fast Community" (FC) [184] се базира на алчна техника која ја максимизира функцијата на модуларност дефинирана со равенката (4.20) и заради прегледност овде повторно дефинирана со (5.20):

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j) \tag{5.20}$$

За поедноставно објаснување на алгоритамот ќе воведеме две нови величини:

$$e_{vw} = \frac{1}{2m}\sum_{ij} A_{ij}\delta(C_i, v)\delta(C_j, w) \tag{5.21}$$

што го претставува делот од врските во графот што ги поврзуваат јазлите од кластерот *i* со јазлите од кластерот *j*, и:

$$a_v = \frac{1}{2m}\sum_{i} k_i\delta(C_i, v) \tag{5.22}$$

што го претставува делот од краеви на врски што се закачени на јазли од кластерот *i*. Ако го искористиме фактот дека делта симболот можеме да го презапишеме како $\delta(C_i, C_j) = \sum_v \delta(C_i, v)\delta(C_j, v)$, тогаш за модуларноста ќе имаме:

$$\begin{aligned} Q &= \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\sum_v \delta(C_i, v)\delta(C_j, v) \\ &= \sum_v\left[\frac{1}{2m}\sum_{ij} A_{ij}\delta(C_i, v)\delta(C_j, v) - \frac{1}{2m}\sum_i k_i\delta(C_i, v)\frac{1}{2m}\sum_j k_j\delta(C_j, v)\right] \\ &= \sum_v (e_{vv} - a_v^2) \end{aligned} \tag{5.23}$$

Операциите во рамки на алгоритамот вклучуваат наоѓање на промените на *Q* што би се случиле при спојувањето на секој пар од кластери, при што се избира најголемата од нив и се извршува соодветното спојување. За таа цел во рамки на

алгоритамот се чува матрица со вредности на промените ΔQ_{ij} што настануваат при спојување на кластерите *i* и *j*. Притоа, поради фактот дека спојувањето на два кластери што немаат меѓусебна врска никогаш нема да предизвика промена во *Q*, треба да се чуваат само ΔQ_{ij} за оние парови *i, j* што се поврзани со една или повеќе врски. Добрите перформанси на алгоритамот од аспект на зачувување на меморија и време се должат на користење на ефикасни податочни структури за чување на потребните информации и тоа:

- Ретка (sparse) матрица која ги содржи вредностите ΔQ_{ij} за секој пар на кластери *i, j* кој има барем една меѓусебна врска. Секоја редица од матрицата се чува на два начини: како балансирано бинарно дрво (елементите можат да се најдат или вметнат во $O(\log n)$ време) и како макс-хип дрво (најголемиот елемент може да се најде во константно време).
- Макс-хип дрво *H* кое го содржи најголемиот елемент на секоја редица од матрицата ΔQ_{ij} заедно со лабелите *i, j* за соодветните парови на кластери.
- Обична низа со елементи a_i.

Како што беше претходно кажано алгоритамот започнува со тоа што секој јазел се поставува во посебен кластер и во тој случај ќе имаме дека $e_{ij} = 1/2m$ ако кластерите *i* и *j* се поврзани и нула ако не се и $a_i = k_i / 2m$. Според тоа иницијализацијата се сведува на

$$\Delta Q_{ij} = \begin{cases} 1/2m - k_i k_j / (2m)^2, & \text{ако } i \text{ и } j \text{ се поврзани} \\ 0 & \text{, инаку} \end{cases} \tag{5.24}$$

и

$$a_i = \frac{k_i}{2m} \tag{5.25}$$

за секое *i*. Одделните параметри на овие пресметки имаат различно значење ако се разгледуваат различните граф репрезентации, односно дефиницијата на k_i и m е различна за нетежински и тежински граф. Кај нетежински граф важат k_i го претставува степенот на јазелот *i*, а m е вкупниот број на врски во рамки на графот. Кај тежинскиот граф важи

$$k_i = \sum_j w_{ij} \tag{5.26}$$

односно она што е степен на јазелот кај нетежинскиот граф кај тежинскиот е збир на тежините на врските што излегуваат од јазелот, и

$$m = \frac{1}{2}\sum_{ij} w_{ij} \tag{5.27}$$

односно вкупниот број на врски кај нетежинскиот граф се заменува со вкупната тежина на сите врски кај тежинскиот граф. Факторот 1/2 мора да се земе во предвид бидејќи врските се двонасочни.

Врз основа на претходното псевдо кодот на алгоритамот е дефиниран на слика 5.12.

Влез: граф репрезентација на протеински интеракции G_i, i=1,..,4;

Излез: структура на кластери

Иницијализација: Пресметка на ΔQ_{ij} и a_i (5.24) и (5.25)
Полнење на H со најголемите елементи од секоја редица на ΔQ
Секој јазел во посебен кластер

Додека $БројКластери > 1$ повторувај

Најди $\max(\Delta Q_{ij})$ од H
Спој (i, j)
Ажурирај (ΔQ), Ажурирај (H), Ажурирај (a_i)
$Q = +\max(\Delta Q_{ij})$

Слика 5.12 Псевдо код за Fast Community (FC) кластерирање

Во рамки на алгоритамот треба да се внимава на ажурирањето на матрицата ΔQ. Ако претпоставиме дека сме ги споиле кластерите i и j, и новиот кластер има индекс j, тогаш треба целосно да се отстранат i-тата редица и колона, а j-тата редица и колона треба соодветно да се ажурираат. Притоа можни се следните случаи:

- Ако кластерот k бил поврзан и со кластерот i и со кластерот j:

$$\Delta Q'_{jk} = \Delta Q_{ik} + \Delta Q_{jk} \tag{5.28}$$

- Ако кластерот k бил поврзан со кластерот i, но не и со кластерот j;

$$\Delta Q'_{jk} = \Delta Q_{ik} - 2a_j a_k \quad (5.29)$$

- Ако кластерот *k* бил поврзан со кластерот *j*, но не и со кластерот *i*;

$$\Delta Q'_{jk} = \Delta Q_{jk} - 2a_i a_k \quad (5.30)$$

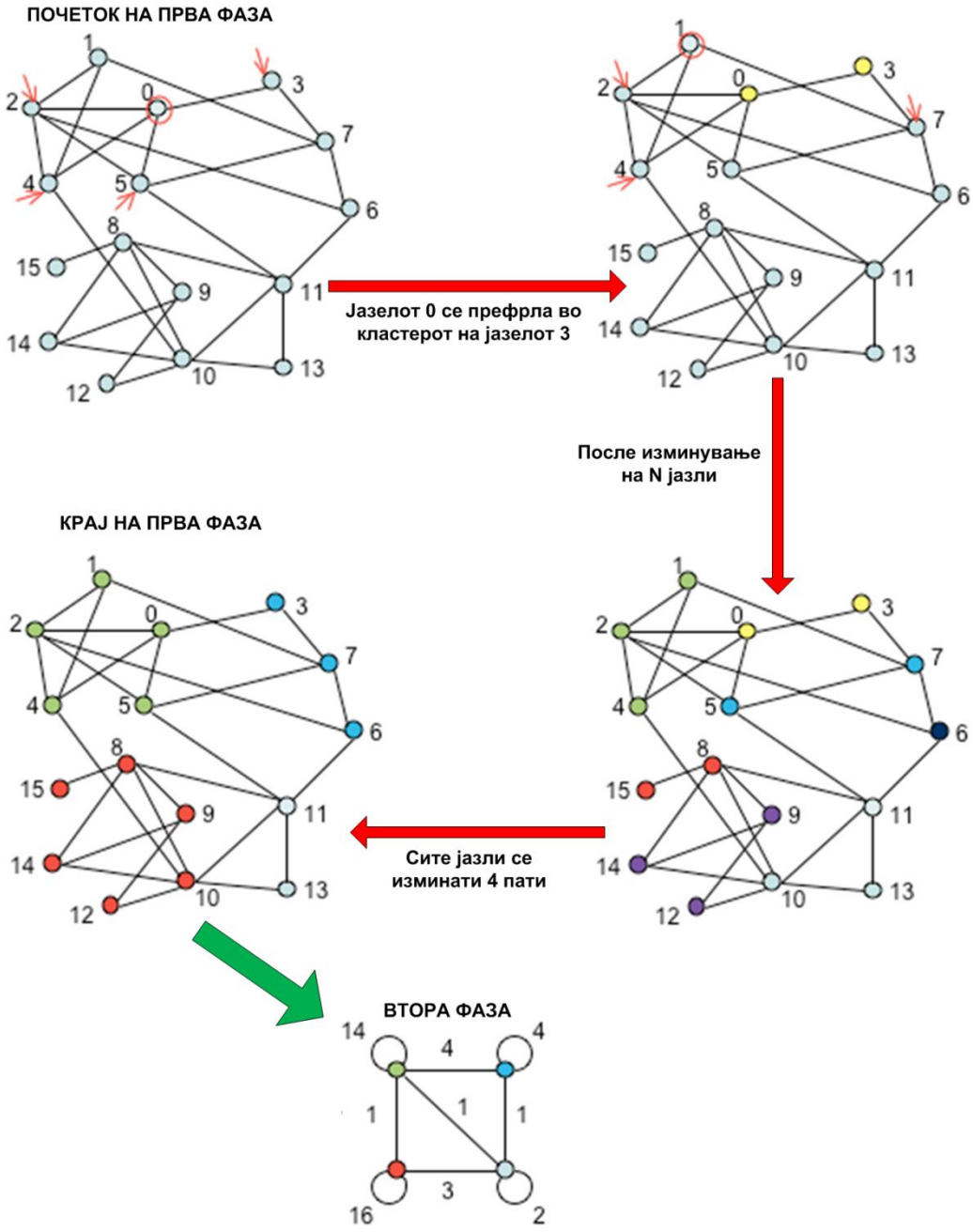
Алгоритамот предложен од Blondel et. al (BGLL) [185] користи поинаква алчна техника базирана на „суперјазли“ за репрезентација на кластерите и пресметка на модуларноста. Овој алгоритам е подобар во оптимизацијата на модуларноста од FC, но има ограничувања од аспект на големите мемориски барања. И во рамки на овој алгоритам важат аналогиите помеѓу нетежински и тежински граф кои беа претходно изложени. Заради поедноставување на терминологијата во рамки на овој алгоритам графот ќе го гледаме од обратен аспект, односно како тежински граф, со тоа што за нетежинските граф репрезентации врз основа на претходното ќе сметаме дека секоја врска има единечна тежина па повторно важат сите претходни дефиниции.

Алгоритамот е поделен на две фази. На почетокот секој јазел се поставува во посебен кластер. Потоа, за секој јазел *i* се разгледуваат сите негови соседи *j* така што се пресметуваат разликите во модуларноста што би настанале доколку *i* го извадиме од неговиот кластер и го додадеме на кластерот каде припаѓа *j*. Јазелот *i* ќе се премести во кластерот на оној сосед за кој добивката во модуларноста е најголема (тоа подразбира добивката да биде позитивна). Доколку ниту една од пресметаните разлики не е позитивна јазелот останува во кластерот каде тековно се наоѓа. Ваквиот процес се повторува итеративно и секвенцијално за сите јазли, односно откако ќе се изминат сите јазли повторно се започнува од „првиот“ и постапката продолжува сè додека не настане стабилна состојба, односно во текот на едно изминување ниту еден јазел да не го промени својот кластер и со ова се достигнува некој локален максимум на модуларноста.

Дел од ефикасноста на алгоритамот се должи на ефикасната пресметка на промената на модуларноста што настанува при преместување на изолиран јазел *i* во кластер C_j:

$$\Delta Q_j = \left[\frac{S_{in} + k_{i,in}}{2m} - \left(\frac{S_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{S_{in}}{2m} - \left(\frac{S_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] \quad (5.31)$$

каде S_{in} е сумата на тежините на врските што се наоѓаат внатре во кластерот C, S_{tot} е сумата на тежините на врските што излегуваат од кластерот C, k_i е сумата на тежините на врските кои излегуваат од јазелот i, $k_{i,in}$ е сумата на тежините на врските помеѓу i и јазлите во C_j и m е сумата на тежините на сите врски во графот.



Слика 5.13 Сликовит приказ на едно изминување на BGLL алгоритамот

Втората фаза на алгоритамот се состои од градење на нов граф чијшто јазли ги претставуваат кластерите пронајдени во првата фаза. Секоја врска помеѓу два нови „суперјазли“ добива тежина еднаква на сумата од тежините на врските што постоеле помеѓу кластерите заменети со новите јазли. Врските помеѓу јазли од ист кластер од првата фаза се трансформираат во јамка на „суперјазелот“ за тој кластер, со тежина еднаква на сумата од тежините на соодветните врски. Откако ќе биде завршена втората фаза повторно се применува првата фаза врз новодобиениот граф. На слика 5.13 е прикажана пример комбинација од фазите.

Ако за едно изминување се смета комбинацијата на двете фази, тогаш со секое изминување бројот на кластери се намалува, па најголем дел од пресметувачкото време се троши на првата фаза. Изминувањата се применуваат итеративно се додека не се возможни повеќе промени и се добие максимална вредност за модуларноста. Висината на хиерархијата на кластери која се добива е еднаква на бројот на направени изминувања.

Влез: граф репрезентација на протеински интеракции G_i, i=1,..,4;

Излез: структура на кластери

Иницијализација: Секој јазел во посебен кластер

Додека *ИмаПромениПомеѓуИзминувања* повторувај
    i = случаен јазел; N = 0;
    Додека $N < |V|$
        За секој j = Сосед(i)
            Пресметај ΔQ_j;
        Ако $\max[\Delta Q_j] > 0$
            N = 0;
            Префрли(i, C_{jmax});
        Инаку
            N++;
        i = СледенЈазел(i);
    $G_i \leftarrow$ Трансформирај(G_i) [*втора фаза*]

[*прва фаза*]

Слика 5.14 Псевдо код за BGLL алгоритамот

5.3.1.5 Мултирезолуциско кластерирање

Како што беше претходно изложено во поглавјето 4.3.2.2 кај мултирезолуциското кластерирање се дефинира резолуциски параметар во рамки на функцијата за квалитет на кластерирање. Во рамки на овој труд е прифатен и имплементиран пристап дефиниран во [186]. Со цел да се дефинира резолуциски параметар на задоволителен начин кластерите се разгледуваат не од комбинаторен аспект, туку од аспект на нивното динамичко однесување. За определен тек што се случува во рамки на некој граф се очекува да биде „заробен“ подолго време во рамки на добро формиран кластер пред да успее да избега. Клучната идеја е да се мери квалитетот од аспект на стабилност на кластерите придружени на стационарен Марков процес кој се моделира како случајно изминување на графот. Резултантната функција на квалитет за детектирање на модули на различни нивоа се дефинира со:

$$Q = (1-t) + \frac{1}{2m}\sum_{i,j}\left[tA_{i,j} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \qquad (5.32)$$

каде t го претставува временскиот параметар на случајното изминување. Оваа равенка е еквивалентна со модуларноста кога параметарот t е еднаков на 1. Оптимизацијата на оваа ф-ункција на квалитет ја правиме идентично како што во претходно поглавје беше дефинирано за BGLL алгоритамот со определени измени во деловите каде обликот на функцијата на квалитет е од значење. Беа извршени експерименти за временскиот параметар во опсег од 1 до 10 (како што е предложено во [187]) и најдобрите резултати беа добиени кога параметарот има вредност 5. Ваквиот алгоритам (со t = 5) понатаму ќе го нарекуваме TimeBGLL.

5.3.1.6 Кластерирање со мапи од случајни патеки

Најгенералната дефиниција на кластер е дека кластерот е група од јазли кои се меѓусебно тесно поврзани. Меѓутоа, од аспект на пропагација на информација може да се предложи и друга дефиниција: Кластер е група од јазли кај која

веројатноста дека информацијата ќе биде заробена е поголема од веројатноста дека истата ќе се рашири. Имајќи во предвид дека случајното изминување на граф е најфундаменталниот модел на пропагација на информацијата, структурата на кластерите може да се детектира преку наоѓање на локалната структура која го заробува случајниот пешак. Ова е принципот што се користи и кај мултирезолуциското кластерирање.

Infomap [188] методот кој е имплементиран во овој систем ги детектира кластерите според дефиницијата базирана на пропагација на информацијата користејќи ја филозофијата на MDL - Minimum Description Length (минимална должина на опис) принципот. Основната идеја на MDL е дека било каква закономерност во податоците може да се искористи за да се компресира нивната должина. Ако можеме да најдеме начин за да ја закодираме патеката на случајниот пешак низ графот и ако кластерската структура ја земеме како законитост во графот, истата можеме да ја детектираме преку наоѓање на поделбата што ќе овозможи минимална должина на описот на патеката.

Наједноставниот начин за еднозначно да ја опишеме случајната патека низ графот би бил на секој јазел да му доделиме недвосмислен код со цел да се избегне повеќезначноста, и должината на описот би станала пократка ако на почесто посетуваните јазли им доделиме пократок код, а на поретко посетуваните релативно подолг код, што е всушност методата на Хафманово кодирање. Меѓутоа, доделувањето на уникатен код на секој јазел во графот ќе биде многу неефикасно ако графот е голем и движењето на случајниот пешак често заглавува во некој мала област – кластер од јазли. Подобра стратегија за кодирање би била поделба на јазлите во кластери и користење на дво-нивовско кодирање. Првото ниво од кодот го опишува кластерот на кој јазелот припаѓа, а второто го разграничува конкретниот јазел од останатите јазли во истиот кластер. Во оваа стратегија, кодот на кластерот (прво ниво) треба да се земе во предвид само тогаш кога случајниот пешак преминува од еден во друг кластер, додека пак движењето внатре во кластерот може уникатно да се опише преку земање во предвид на второто ниво од кодот. Дополнително, на секој кластер треба да му се додели излезен код и истиот се зема во предвид кога случајниот пешак излегува од

кластерот и тоа се прави со цел да можат да се разграничат првото и второто ниво од кодот. Цената на користење на дво-нивовски код ќе биде многу исплатлива ако графот има изразена структура на кластери која добро се детектира, бидејќи во тој случај кодовите од второто ниво ќе станат многу пократки, а кодовите од првото ниво нема да се користат често, што во крајна линија ќе влијае на намалувањето на вкупната должина на описот на случајната патека. Според ова, најдоброто кластерирање на графот ќе биде она кое ја минимизира просечната должина на описот на случајната патека, при користење на стратегија за кодирање како претходно опишаната.

Откако ќе биде определена поделбата на кластери *M*, лесно може да се пресмета веројатноста за користење на секој код и равенката на мапа *L*(*M*), којашто е дефинирана како теоретски минимум за просечната должина описот, може да се дефинира преку Шеноновата теорема за кодирање како

$$L(M) = q_{\curvearrowright} H(Q) + \sum_{i=1}^{C} p^i_{\circlearrowright} H(P^i) \tag{5.33}$$

каде i е индексот на кластерот, α е индексот на јазелот и C е бројот на кластери; $q_{\curvearrowright} \equiv \sum_{i=1}^{C} q^i_{\curvearrowright}$ е вкупната веројатност за користење на код од прво ниво каде $q^i_{\curvearrowright}$ е веројатноста за користење на код од прво ниво за кластер i; $p^i_{\circlearrowright} \equiv q^i_{\curvearrowright} + \sum_{\alpha \in i} p_\alpha$ е веројатноста за користење на код од второ ниво и излезен код за кластерот i; и p_α е веројатноста јазелот α да биде посетен, што е еднакво на веројатноста за користење на код од второ ниво за α. $H(Q)$ е просечната должина на описот за која допринесува првото ниво од кодот:

$$H(Q) = -\sum_{i=1}^{C} \frac{q^i_{\curvearrowright}}{q_{\curvearrowright}} \log \frac{q^i_{\curvearrowright}}{q_{\curvearrowright}} \tag{5.34}$$

додека $H(P^i)$ е должината на описот за која допринесува второто ниво од кодот за кластерот i:

$$\begin{aligned} H(P^i) = &- \frac{q^i_{\curvearrowright}}{q^i_{\curvearrowright} + \sum_{\alpha \in i} p^i_\alpha} \cdot \log \frac{q^i_{\curvearrowright}}{q^i_{\curvearrowright} + \sum_{\alpha \in i} p^i_\alpha} \\ &- \sum_{\alpha \in i} \frac{p^i_\alpha}{q^i_{\curvearrowright} + \sum_{\beta \in i} p^i_\beta} \cdot \log \frac{p^i_\alpha}{q^i_{\curvearrowright} + \sum_{\beta \in i} p^i_\beta} \end{aligned} \tag{5.35}$$

каде p_α^i е еднакво на p_α кога α припаѓа на *i*, или е нула во секој друг случај. Веројатноста $q_{\curvearrowright}^i$ е вклучена во равенката (5.35) за да го претстави придонесот на излезните кодови за кластерот *i* и за дадена структура на кластери *M* може да се пресмета според следната равенка:

$$q_{\curvearrowright}^i = \sum_{\alpha \in i} \sum_{\beta \notin i} p_\alpha \frac{A_{\alpha\beta}}{k_\alpha} \tag{5.36}$$

каде $A_{\alpha\beta}$ е елемент од матрицата на соседство, а $k_\alpha = \sum_\beta A_{\alpha\beta}$. Должината на описот се мери во нити ако логаритмите во претходните равенки се земат со основа 2.

Структурата на кластерите може да се детектира преку наоѓање на поделбата на јазлите која ја минимизира равенката на мапа *L*(*M*) од (5.33). Во конкретниот алгоритам е избрана алчна стратегија како онаа опишана кај BGLL алгоритамот.

5.3.1.7 Кластерирање на врски

Во излагањето во поглавјето 4.3.2.2 беше кажано дека пристапот за определување на преклопувачки кластери во рамки на графот се базира на идејата за кластерирање на врски наместо на јазли. Ако одиме уште еден чекор понапред ваквата идеја можеме да ја преведеме во процедура која графот за кој сакаме да најдеме преклопувачки кластери најпрво го трансформира во граф во кој едноставно кажано врските стануваат јазли, а јазлите врски, и потоа применуваме алгоритам за кластерирање на јазли.

Го применуваме пристапот за трансформација на нетежински граф предложен во [189] кој со едноставни измени може да се примени и кај тежински графови. Без губење на генералноста можеме да претпоставиме дека работиме со дефиницијата $G_1(V,E)$ за нетежински граф од јазли. Методата најпрво го трансформира G_1 во нетежински граф на врски $L_1(G_1)$ и потоа користи динамика на случајно изминување на графот за да ја измери функцијата на квалитет на кластерирањето. Во принцип може да се примени било кој алгоритам за кластерирање на јазли. Меѓутоа бидејќи оптимизацијата на модуларноста е поврзана со однесувањето на

случајни пешаци на графот (како што веќе беше изложено во претходните алгоритми) и конструкцијата на $L_1(G_1)$ не ја менува динамиката на случајните пешаци, најлогично беше да се примени пристап за оптимизација на модуларноста за да се најдат кластерите на графот на врски $L_1(G_1)$. Конкретно беше искористен пристапот за максимизација на модуларноста предложен кај BGLL алгоритамот.

Конверзијата на графот од јазли во врски се прави на следниот начин: најпрво графот на јазли се претставува со користење на матрицата на инцидентност (incidence matrix) $B_{|V|\text{x}|E|}$, во која $B_{i\alpha}$ е еднакво 1 ако врската α е поврзана со јазелот i и 0 во секој друг случај. Матрицата B може да се гледа како матрица на соседство на некој бипартитен граф. Графот на врски се конструира со проекција на бипартитниот граф со тоа што сите јазли од еден тип се земаат за јазли на проектираниот граф. Помеѓу два јазли во проектираниот граф се додава врска доколку двата јазли имаат барем еден заеднички јазел од другиот тип во рамки на оригиналниот бипартитен граф, што резултира во матрица на соседство $C_{|E|\text{x}|E|}$ за графот на врски $L_1(G_1)$, чиј елементи се дефинирани со:

$$C_{\alpha\beta} = \sum_i B_{\alpha i} B_{i\beta} (1 - \delta_{\alpha\beta}) \tag{5.37}$$

каде $\delta_{\alpha\beta}$ е Кронекер делта симболот. Со пресметка на матрицата на соседство како во равенката (5.37) на јазлите кои имаат многу висок степен, хабовите, им се дава преголема важност во рамки на гафот на врски, па затоа користиме нормализација за да го избегнеме тој ефект и $C_{\alpha\beta}$ го пресметуваме како:

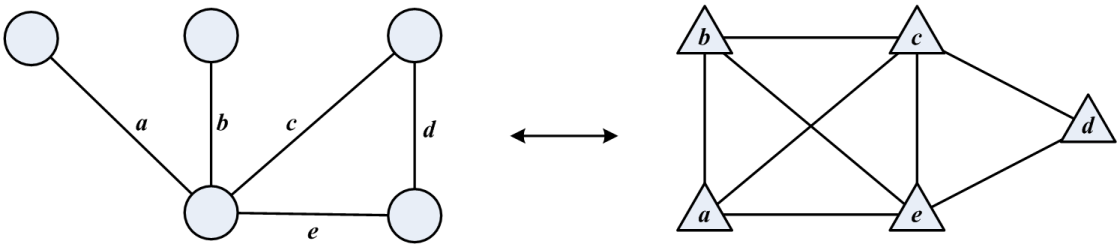
$$C_{\alpha\beta} = \sum_{i,k_i>1} \frac{B_{\alpha i} B_{i\beta}}{k_i - 1} (1 - \delta_{\alpha\beta}) \tag{5.38}$$

каде k_i е степенот на јазелот i.

Кога работиме со тежински граф од јазли, пример $G_2(V,E,W)$, воведуваме уште една, тежинска, матрица на инцидентност $\tilde{B}$, за која важи $\tilde{B}_{\alpha j} = w_\alpha$ ако врската α е инцидентна на јазелот j и има тежина w_α. Секој јазел i има јачина s_i, дефинирана како сума од тежините на сите негови инцидентни јазли. Како и во случајот на нетежински граф и овде ја пресметуваме нормализираната матрица на соседство за, сега тежинскиот граф на врски $L_2(G_2)$ преку формулата:

$$\tilde{C}_{\alpha\beta} = \sum_{i,k_i>1} \frac{\tilde{B}_{\alpha i} B_{i\beta}}{s_i - w_\beta} (1 - \delta_{\alpha\beta}) \tag{5.39}$$

На слика 5.15 е даден визуелен приказ на трансформацијата на графот на јазли (од лево) во граф на врски (од десно). Врските *a*, *b*, *c*, *d* и *e* од графот на јазли се пресликуваат во јазлите *a*, *b*, *c*, *d* и *e* во графот на врски, соодветно. Овој алгоритам во понатамошниот текст ќе го нарекуваме EdgeCluster.



Слика 5.15 Визуелен приказ на трансформација на граф на јазли во граф на врски

5.3.1.8 Кластерирање базирано на хомогеност на кластери

Во рамки на оваа докторска дисертација е развиен нов алгоритам за кластерирање базиран на хомогеност на кластери. Во рамки на овој пристап функцијата за квалитетот на кластерирањето ја дефинираме како:

$$Q = \frac{BCV}{WCV} \tag{5.40}$$

каде *BCV* го дефинираме како мерка за различност на добиените кластери, а *WCV* како мерка за хомогеност на кластерите. За да можеме да ги формулираме *BCV* и *WCV* најпрво ќе дефинираме мерка за растојание во рамки на графот. За таа цел ќе го искористиме стационарниот вектор на случајно изминување на графот P_t^i почнувајќи од јазел *i* дефиниран според равенките (5.1)-(5.4) и (5.7). Користејќи ја процедурата дефинирана во поглавјето 5.3.1.1 и равенките (5.12)-(5.14) ги дефинираме растојанијата $d(i, j)$ како растојание помеѓу два јазли *i* и *j* во графот и $D(A, B)$ како растојание помеѓу два кластери *A* и *B*. Во суштина она што ќе го добиеме е определена мерка за одбивност помеѓу јазлите, што произлегува од фактот дека стационарната распределба на случајното изминување укажува на определен афинитет на јазлите еден кон друг. Во случаите кога работиме со тежинските репрезентации ова е уште поизразено бидејќи афинитетот е

дополнително воден од семантичката сличност помеѓу јазлите, а одбивноста од нивната различност. Дефинираме

$$BCV = \frac{2}{N_c(N_c - 1)} \sum_{i=1}^{N_C - 1} \sum_{j=i+1}^{N_C} D(C_i, C_j) \tag{5.41}$$

каде N_C е вкупниот број на кластери. Вака дефинираната мерка за различност на кластерите може да има максимална вредност 1, што би значело дека растојанието помеѓу било кој пар од кластери е максимално, т.е. еднакво на 1 (теоретски тоа би значело дека двата кластери имаат дијаметрално спротивни содржини пр. сите јазли во A се бели, а сите јазли во B се црни). Мерката за хомогеност на еден кластер C_i ја дефинираме со

$$WCV_{C_i} = \frac{2}{|C_i| \cdot (|C_i| + 1)} \sum_{v,w \in C_i} d(v, w) \tag{5.42}$$

Целта на кластерирањето е да го оптимизира односот помеѓу различноста и хомогеноста. Применуваме оптимизација како онаа дефинирана кај BGLL алгоритамот. Значи на почетокот сите јазли ги сместуваме во посебни кластери и итеративно ги изменуваме јазлите проверувајќи дали нивното приклучување кон кластер на некој сосед ќе го подобри *BCV*/*WCV* односот. За побрза конвергенција јазлите ги изминуваме по редослед дефиниран според нивниот степен (број на врски на јазелот на нетежински граф или вкупна тежина на врски на јазелот за тежински граф) почнувајќи од најголемиот кон најмалиот. За да ги дефинираме тежините во графот што се добива со замена на кластерите со нови јазли после извршувањето на првото изминување, односно откако ќе се стабилизираат кластерите (нема поместувања што можат да го подобрат *BCV*/*WCV* односот) ги дефинираме следните величини:

$$WWC_k = \sum_{i,j \in C_k} a_{ij} \ , \ WBC_{kl} = \sum_{i \in C_k} \sum_{j \in C_l} a_{ij} \tag{5.43}$$

каде WWC_k го претставува вкупниот збир на тежини на врските во рамки на кластерот C_k, WBC_{kl} го претставува вкупниот збир на тежини на врските што постојат помеѓу јазли од кластерите C_l и C_k, и a_{ij} е елемент на матрицата на соседство, која за тежинскиот граф е идентична со тежинската матрица (A=W), а за нетежинскиот важи дека a_{ij}=1 ако постои врската (i,j) и 0 во секој друг случај. Елементите на новата матрица на соседство ќе бидат дефинирани со:

$$b_{ij} = \begin{cases} WWC_i, i = j \text{ (јамка на нов јазел)} \\ WBC_{ij}, i \neq j \text{ (врска помеѓу нови јазли)} \end{cases} \quad (5.44)$$

Сега на новиот граф ја применуваме истата постапка од претходно и итерираме сè додека се возможни подобрувања, односно можно е да се направи барем едно разместување во првата фаза од оптимизацијата. Овој алгоритам во понатамошниот текст ќе го референцираме како HomogeneityOptimization или скратено HO.

5.3.2 Евалуација на кластерирање

Евалуацијата на кластерирањето работи како независен дел од целиот процес на функционална анотација врз база на кластерирање, односно истата не е никако условена од тоа кој алгоритам ќе биде искористен за добивање на кластерите. Валидацијата на кластерите беше извршена со користење на синтетички бенчмарк граф дефиниран како во [129] со цел да се споредат различните алгоритми за кластерирање применети во системот. Синтетичкиот граф се генерира според параметрите на нашите податоци. Се тргнува од претпоставката дека и распределбата на степенот на јазлите и големината на кластерите е степенска со коефициенти γ и β, соодветно. Бројот на јазли во мрежата е N, а просечниот степен е $\langle k \rangle$.Постапката на градење се состои од следните чекори:

1. На секој јазел му се доделува степен што се добива од степенска распределба со експонент γ. Екстремните вредности на распределбата $k_{\min}$ и $k_{\max}$ се избираат така што просечниот степен е $\langle k \rangle$. Се користи конфигурацискиот модел за поврзување на јазлите за да се одржи нивната секвенца на степени. Конфигурацискиот модел подразбира доделување на „полу-врски“ на секој јазел, при што бројот на полу-врски одговара на степенот на јазелот. Случајно се избираат парови од вакви полу-врски и се поврзуваат соодветните јазли.
2. Секој јазел дели $(1-\mu)$ од своите врски со останатите јазли од неговиот кластер и μ врски со останатите јазли во мрежата; μ е параметарот на мешање.

3. Големините на кластерите се добиваат од степенска распределба со експонент β, така што вкупната големина на сите кластери е еднаква на вкупниот број N на јазли во графот. Минималните и максималните големини на кластери $s_{\min}$ и $s_{\max}$ се избираат така да биде задоволено: $s_{\min} > k_{\min}$ и $s_{\max} > k_{\max}$. Ова обезбедува дека било кој јазел со произволен степен ќе биде вклучен во барем еден кластер.
4. На почетокот сите јазли се „бездомници“ т.е. не припаѓаат на ниту еден кластер. Во првата итерација даден јазел се доделува во случајно избран кластер; ако големината на кластерот го надминува внатрешниот степен на јазелот (бројот на соседи на јазелот што се наоѓаат во истиот кластер), тогаш јазелот се приклучува на кластерот, ако не останува бездомник. Со последователни итерации доделуваме јазел бездомник на случајно избран кластер: ако кластерот е пополнет (достигнува проектирана големина), се исфрла случајно избран јазел од кластерот и тој јазел станува бездомник. Процедурата запира кога нема повеќе јазли бездомници.
5. За да се исполни условот за бројот (делот) на соседи внатре во кластерот, изразен со параметарот на мешање μ, се извршуваат неколку чекори на преповрзување, така што степените на јазлите ќе останат исти и ќе се промени само односот помеѓу внатрешниот и надворешниот степен кога има потреба за тоа.

Значи резултантниот граф од ваквата постапка има определена структура од кластери и евалуацијата продолжува во насока на споредба на кластерите добиени со примена на различните алгоритми за кластерирање врз ваквиот граф со неговите a priori познати кластери. A priori кластерите на синтетичкиот граф уште ќе ги нарекуваме „вистински“, а кластерите добиени со алгоритамот што го евалуираме „пронајдени“. Споредбата се прави преку Нормализираната Здружена Информација [190].

Нека со X ја означиме случајната променлива даден случајно избран јазел од графот да припаѓа на „вистински“ кластер и дефинираме $P(X = a) = P(X_a)$ како веројатност избраниот јазелот да припаѓа на кластерот a, и нека со Y ја означиме

случајната променлива даден случајно избран јазел од графот на припаѓа на „пронајден“ кластер и дефинираме $P(Y=b)=P(Y_b)$ како веројатност избраниот јазел да припаѓа на кластерот *b*. Поклопувањето помеѓу вистинското и пронајденото кластерирање може да се измери со помош на заедничката информација на овие две случајни променливи:

$$I(X,Y)=\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}P(X_i,Y_j)\cdot\log\frac{P(X_i,Y_j)}{P(X_i)\cdot P(Y_j)} \tag{5.45}$$

каде бројот на вистински кластери е означен со C_A и бројот на пронајдени кластери е означен со C_B, а $P(X,Y)$ е здружената веројатност на двете случајни променливи. Од дефиницијата на заедничката информација може да се покаже дека важи $I(X,Y)\le[H(X)+H(Y)]/2$, каде $H(X)$ и $H(Y)$ се ентропиите на случајните променливи *X* и *Y*, соодветно, и во конкретниот случај се дефинирани со $H(X)=-\sum_{i=1}^{C_A}P(X_i)\log P(X_i)$ и $H(Y)=-\sum_{j=1}^{C_B}P(Y_j)\cdot\log P(Y_j)$. Врз основа на ова се дефинира нормализирана заедничка информација преку:

$$NMI(X,Y)=\frac{2\cdot I(X,Y)}{H(X)+H(Y)} \tag{5.46}$$

Токму ваквата нормализирана вредност на заедничката информација ја користиме за евалуација на кластерите. Нормализираната здружена информација е еднаква на 1 ако пронајдените и вистинските кластери се идентични и 0 ако се целосно независни. За да може да се пресмета сè што е потребно е да се пресметаат соодветните веројатности $P(X_a)$, $P(Y_b)$ и $P(X_a,Y_b)$. Ќе дефинираме матрица на забуна *M*, во која редиците соодветствуваат на „вистинските“, а колоните соодветствуваат на „пронајдените“ кластери. Елементот на *M*, M_{ij} е бројот на јазли од вистинскиот кластер *i* што се појавуваат во пронајдениот кластер *j*. Веројатностите ќе бидат еднакви на соодветните релативни фреквенции односно $P(X_a)=M_{a+}/M$, $P(Y_b)=M_{+b}/M$ и $P(X_a,Y_b)=M_{ab}/M$, каде сумата на *a*-тата редица од матрицата *M* е означена со M_{a+}, сумата на *b*-тата колона е означена со M_{+b}, а *M* е вкупниот број на јазли во графот. M_{a+} всушност одговара на бројот на јазли што припаѓаат во „вистинскиот“ кластер *a*, додека M_{+b} е бројот на јазли што припаѓаат во „пронајдениот“ кластер *b*. Врз основа на овие дефиниции и (5.46) нормализираната заедничка информација го добива обликот:

$$NMI(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} M_{ij}\log\left(\frac{M_{ij}M}{M_{i+}M_{+j}}\right)}{\sum_{i=1}^{C_A} M_{i+}\log\left(\frac{M_{i+}}{M}\right)+\sum_{j=1}^{C_B} M_{+j}\log\left(\frac{M_{+j}}{M}\right)} \quad (5.47)$$

Бидејќи алгоритамот со кластерирање на врски продуцира преклопувачки кластери, за да го споредиме со останатите алгоритми малку ја менуваме процедурата. Односно, наместо синтетичкиот граф да го генерираме со параметрите на оригиналниот граф, го генерираме според параметрите на графот на врски (оној што се добива со трансформација на оригиналниот граф). Кога се врши споредбата за „пронајдени“ кластери ќе ги сметаме оние добиени со кластерирањето на графот на врски. Со ваквото поставување повторно ја пресметуваме нормализираната заедничка информација како во (5.47).

Покрај евалуацијата од аспект на можноста за репродуцирање на претходно познати кластери, кластерирање го испитуваме и од аспект на биолошка валидност на добиените кластери. Ова е потребно бидејќи различните алгоритми што се користат продуцираат кластери што имаат различна големина и структура и за истите треба да ја испитаме биолошката релевантност, или со други зборови да потврдиме дека структурата на кластерите не е случајна. Ако кластерот е биолошки релевантен протеините што припаѓаат на истиот кластер имаат слични биолошки функции [5]. Според тоа, функционалната хомогеност на кластерот ќе биде индикатор за неговата биолошка валидност. Најголем дел од методите за пресметка на функционална хомогеност вклучуваат некој облик на пресметка на P-вредноста. Меѓутоа ваквата пресметка подразбира дека треба однапред да бидат познати вистинските кластери во графот, што не е секогаш случај.

Многу природен пристап кон испитување на функционалната хомогеност е преку пресметувањето на ентропијата на функциите присутни во рамки на кластерот. Ентропијата се пресметува преку фреквентноста на појавување на даден термин (функција) во рамки на кластерот и е дефинирана со:

$$H = -\sum_{\text{сите } i} F_i \log F_i \quad (5.48)$$

каде

$$F_i = \frac{T_i}{\sum_i^n T_i} \tag{5.49}$$

каде F_i е фреквенцијата на појавување на терминот i, дефинирана со (5.49), T_i е бројот на појавувања на тој термин во рамки на кластерот и n е бројот на уникатни термини присутни во кластерот. Ако јазлите во еден кластер имаат конзистентни термини вредноста на функционалната ентропија ќе биде мала, и би била нула ако сите јазли имаат само еден термин.

5.3.3 Функционална анотација

Постојат неколку различни методи во литературата за доделување на термини на даден прашален протеин откако ќе бидат определени кластерите. Секој од овие методи се базира на пресметка на определена вредност за термините кои се присутни (се доделени на протеини) во кластерот каде што се наоѓа прашалниот протеин, при што на прашалниот протеин му се доделуваат оние термини кои имаат вредност поголема или помала од некој предефиниран праг во зависност од тоа како се пресметуваат вредностите. Во рамки на овој труд се применети три различни пресметки за вредноста на термините и тоа хипергеометриска P-вредност, χ-квадратна статистика и фреквентност на термини во рамки на кластерот.

Хипергеометриска P-вредност се пресметува со:

$$P_t = \sum_{i=n_t}^{C} \frac{\binom{T}{i}\binom{N-T}{N-i}}{\binom{N}{C}} \tag{5.50}$$

каде N е бројот на јазли во графот со кој е претставена протеинската интеракциска мрежа, T е бројот на јазли во графот на кои им е доделен терминот t, C е големината на кластерот и n_t е бројот на јазли ви кластерот на кои им е доделен терминот t. Терминте за кои се добива P-вредност помала од некој праг му се доделуваат како анотации на прашалниот протеин.

Вредноста на χ-квадратната статистика за терминот t е дефинирана со:

$$\chi_t^2 = \frac{(n_t - e_t)^2}{e_t} \tag{5.51}$$

каде n_t го има истото значење како кај претходната вредност, а e_t е очекуваниот број на јазли во кластерот на кои им е доделен терминот *t*. Очекуваниот број се пресметува со користење на едноставна пропорција $e_t = (T/N)*C$, при *T*, *N* и *C* со исти значења како во претходната вредност.

Наједноставниот и најинтуитивниот пристап за пресметка на вредноста е секој термин да се рангира според неговата фреквенција на појавување како анотација на јазлите во рамки на кластерот. Овој пристап е изведен од добро познатиот алгоритам на мнозинска одлука [117], каде на јазелот му се доделуваат најчестите термини од неговите соседи. Нашата дефиниција го проширува соседството од јазли не само на директните соседи туку на сите јазли во кластерот во кој припаѓа прашалниот протеин, *K*:

$$s(j)_{j \in T_K} = \sum_{i \in K} z_{ij} \tag{5.52}$$

каде T_K е множеството од термини присутни во кластерот *K*, и

$$\boldsymbol{Z}_{ij} = \begin{cases} 1, \text{ ако i-тиот јазел од K го има j-тиот термин од } T_K \\ 0, \text{ инаку} \end{cases} \tag{5.53}$$

Овде треба да напоменеме дека кога работиме со протеин-термин графот G_3, дефиницијата на одредени величини што се користат во пресметката на вредностите треба да се променат. Имено, велиме дека терминот *t* е присутен во рамки на кластерот ако соодветниот јазел-термин *t* припаѓа на тој кластер. Вкупниот број на јазли во графот соодветствува со вкупниот број на јазли-протеини, големината на кластерот соодветствува со бројот на јазли-протеини во кластерот, бројот на јазли во графот на кои им е доделен терминот *t* соодветствува на степенот на јазелот-термин *t* и бројот на јазли во кластерот на кои им е доделен терминот *t* соодветствува на бројот на врски помеѓу јазелот-термин *t* и јазлите-протеини што припаѓаат на кластерот. За вредноста на фреквентноста T_K претставува множество од јазли-термини и z_{ij} е дефинирано со:

$$\boldsymbol{Z}_{ij} = \begin{cases} 1, \text{ ако i-тиот јазел-протеин од K има врска со j-тиот јазел-термин од } T_K \\ 0, \text{ инаку} \end{cases} \tag{5.54}$$

6

РЕЗУЛТАТИ И ДИСКУСИЈА

Евалуација на функционалната анотација во протеинските интеракциски мрежи претставува предизвик поради фактот што не постојат усогласени и стандардизирани мерки, ниту пак некаков бенчмарк во однос на кој би се правела оценатa за квалитетот на користените пристапи. Првиот проблем е во тоа што протеините во мрежата можат да бидат анотирани со повеќе функции. Тоа го отежнува дефинирањето на стандардните поими за евалуација во т.н. матрица на забуна:

- True Positive (TP): вистински позитивни анотации
- True Negative (TN): вистински негативни анотации
- False Positive (FP): лажно позитивни анотации
- False Negative (FN): лажно негативни анотации

Пристапите кои се користат за функционална анотација на некој прашален протеин враќаат листа на предвидени функции за истиот. Проблемот е тоа што

дел од тие функции се точни, а дел не. Прашањето кое се поставува е како ќе се определи кои се вистинските позитивни анотации (TP). Еден одговор би бил за TP да се изберат оние предвидувања кои погодиле барем една од можните функции на протеинот. Проблемот овде е очигледен. Имено, ако на сите протеини им се доделат сите функции кои постојат тогаш сите предвидувања ќе бидат TP што значи ваквиот пристап е премногу лабав и неупотреблив. Сосема спротивно на ова би било ако за TP ги сметаме само оние предвидувања што точно ги погодуваат сите функции на протеинот. Секако и ваквиот пристап е неупотреблив затоа што е премногу строг и нема да даде вистинска слика за квалитетот на алгоритамот кој на пример греши само една функција, иако ги погодува сите останати. Во рамки на евалуацијата во овој труд секоја функција се разгледува посебно во однос на секој протеин. Односно секоја можна функција ја сместуваме во една од четирите класи кога го задоволува соодветниот услов:

- TP: Кога функцијата е предвидена и е дел од вистинските функции на прашалниот протеин;
- TN: Кога функцијата не е предвидена и не е дел од вистинските функции на прашалниот протеин;
- FP: Кога функцијата е предвидена, но не е дел од вистинските функции на прашалниот протеин;
- FN: Кога функцијата не е предвидена, но е дел од вистинските функции на прашалниот протеин.

За да можат да се пресметаат овие вредности овде се користи *leave-one-out* (изостави-еден) методата со која во рамки на едно извршување на алгоритамот (или фиксна низа од алгоритми) за функционална анотација само еден протеин ја има улогата на прашален протеин. Алгоритамот се извршува онолку пати колку што има протеини во мрежата, така да секој протеин во точно едно извршување биде прашален. Прашалниот протеин кај *leave-one-out* методата се смета за неанотиран и целта е да му се доделат функции. Следниот проблем кој треба да се реши се импликациите од претпоставката дека прашалниот протеин не е анотиран. Имено, во зависност од тоа каква репрезентација на графот користиме ваквата претпоставка може да влијае врз конструкцијата на графот, со тоа и на самите методи кои потоа се користат за да се предвидат функциите. Единствено

кај едноставниот нетежински граф (G_1) не се потребни никакви промени бидејќи неговата конструкција не е никако засегната од содржината на јазлите во графот. За разлика од нив тежинските графови (G_2,G_4) мора да се видоизменат бидејќи непостоењето на термини придружени кон прашалниот јазел (протеин) ќе значи дека тежините на врските што излегуваат од него повеќе не можат да бидат дефинирани со соодветните формули (вака дефинирани сите тежини би биле 0). Во ваков случај сите врски што излегуваат од прашалниот протеин мора да добијат структурно базирани тежини (што значи тежините ќе зависат од соседството на прашалниот протеин), а сè останато останува исто како претходно. За протеин-термин репрезентацијата (G_3) претпоставката за неанотиран прашален протеин значи дека сите врски на прашалниот јазел-протеин кон јазли-термини треба да бидат избришани. Евентуалните промени кои настануваат поради претпоставката за неанотиран прашален протеин дополнително ќе значат дека соодветниот алгоритам за предвидување треба да се изврши од почеток (ова е значајно кога предвидувањето е базирано на кластерирање, односно за секој прашален протеин ќе треба да се изградат кластерите од почеток).

Откако ќе биде применет соодветниот алгоритам за функционална анотација добиваме листа на функции што се можни решенија на проблемот и треба да се пресмета нивниот ранг. Во претходната глава прикажавме различни пристапи за пресметка на рангот на функциите како кај директниот метод, така и кај методите базирани на кластерирање. Од направените експерименти и добиените резултати увидовме дека пресметката според (5.10) (χ^2 статистика на ограничено соседство) кај директниот метод, и пресметката според (5.53),(5.54) (фреквенција на појавување во кластер) кај методите базирани на кластерирање се подобри од останатите, па во продолжение на оваа глава сите резултати се однесуваат на користење токму на овие пресметки. Откако ќе се пресмета рангот на можните функции потребно е соодветните вредности да се нормализираат во опсег [0,1]. За прашалниот протеин се предвидуваат сите функции кои имаат ранг поголем од некој претходно дефиниран праг ω. На пример, за $\omega = 0$, на прашалниот протеин му се доделуваат сите најдени можни функции. Врз основа на направените предвидување се пресметуваат соодветните вредности во матрицата на забуна (во зависност од тоа кој услов е исполнет една од четирите класи ја зголемува својата

вредност за 1, со тоа што на почетокот сите вредности се 0). Финалната матрица на забуна за даден праг ω претставува збир од матриците на забуна добиени за секој прашален протеин посебно. Прагот ω го менуваме со чекор 0.1, во интервалот од 0 до 1, и за секоја прагова вредност се пресметува посебна матрица на забуна.

Врз основа на вредностите добиени во матрицата на забуна се пресметуваат стандардните статистички мерки на сензитивност и специфичност. Сензитивноста (уште се нарекува стапка на вистински позитивни примероци) на еден алгоритам го претставува делот на точно предвидени функции од сите функции кои би требало да ги има еден протеин и е дефинирана со:

$$Sensitivity\ (TruePositiveRate) = \frac{TP}{TP + FN} \tag{6.1}$$

Специфичноста (уште се нарекува стапка на вистински негативни примероци) го претставува делот на функции кои не се предвидени за еден протеинот од сите функции кои навистина не му припаѓаат на протеинот и е дефинирана со:

$$Specificity(TrueNegativeRate) = \frac{TN}{TN + FP} \tag{6.2}$$

Од аспект на испитување на перформансите на алгоритмите за функционална анотација (и генерално на алгоритмите кои вршат некакво предвидување) многу поинтересна е стапката на лажно позитивни примероци (6.3) (погрешно предвидени функции во однос на вкупен број на функции кои не треба протеинот да ги има) бидејќи таа претставува мерка за грешката што ја прави алгоритамот и влијае на доверливоста на истиот.

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{6.3}$$

Ако се исцртаат како координатни парови, сензитивноста (ординатa) и стапката на лажно позитивни примероци (апциса) се добива ROC (Receiver Operating Characteristic) кривата [191]. Алгоритамот за функционална анотација можеме да го гледаме како класификатор кај кој инстанците што треба да се класифицираат се функциите и треба да одлучи дали функцијата (примерокот) му припаѓа на прашалниот протеин (позитивен примерок) или не му припаѓа (негативен примерок). ROC кривата се исцртува така што се генерира по една точка во

просторот (дефиниран со стапката на лажно позитивни примероци и сензитивноста) над целиот опсег за прагот на класификација (прагот во нашиот случај е ω). Ако класификаторот одлуката дали примерокот е позитивен или негативен ја прави случајно тогаш ROC кривата ќе се поклопува со $x=y$ правата и тоа е најлошиот можен случај, односно за еден класификатор се очекува неговата ROC крива да се наоѓа над $x=y$ правата. Површината под ROC кривата (Area Under Curve, во понатамошниот текст ќе се користи AUC како кратенка) на еден класификатор е еквивалентна со веројатноста дека класификаторот некој случајно избран позитивен примерок ќе го рангира повисоко од некој случајно избран негативен примерок. Од самата дефиниција на ROC кривата се гледа дека максималната AUC вредност што може еден класификатор да ја добие е 1, а за да класификаторот биде подобар од некој друг случаен класификатор треба AUC да биде поголемо од 0.5. Поголемите AUC вредности укажуваат дека предвидувањето има подобри перформанси. Меѓутоа во нашите анализи многу значајна е и стапката на лажно позитивни примероци (во понатамошниот текст стапка на грешка) затоа што покрај зголемувањето на можноста за погодување сакаме да ја намалиме и можноста за грешка, па затоа во дискусијата на резултатите ги разгледуваме двата аспекти. Направени се експерименти за сите можни комбинации на репрезентација на протеинската интеракциска мрежа и предложените пристапи за функционална анотација.

Најпрво да ги разгледаме резултатите добиени од функционалната анотација со директниот метод. Како што беше изложено во поглавјето 6.1 овој метод има три променливи параметри, и тоа се големината на функционалното соседство, веројатноста за телепортирање c, граничната вредност за конвергенција ε. Беа направени експерименти за вредности на ε во опсегот $[10^{-5},10^{-15}]$ и се покажа дека вредноста на овој параметар нема влијание врз перформансите на предвидување и единствено влијае на брзината на конвергенција. За разлика од ε, веројатноста за телепортирање c има големо влијание врз крајниот резултат, што може да се забележи од резултатите од експериментите изведени за вредности $0.1 \leq c \leq 0.9$. За големината на функционалното соседството беа направени експерименти со вредности од 5 до 200. Овде само ќе напоменеме дека ако во овој пристап функцијата ја определуваме врз основа на комплетниот граф резултатите

укажуваат дека се добиваат многу големи стапки на лажно позитивни предвидувања и како што наведовме погоре тоа е неприфатливо за ваквиот алгоритам. Во табела 1 се најдобрите добиени резултати за ваквиот пристап каде c=0.1. Прикажаните резултати се однесуваат на тежински граф со семантичка метрика на Ресник.

Директна метода за тежински граф (Ресник) со евалуација на функции со комплетен граф												
	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
***c* = 0.1**	**sensitivity**	0.9135	0.7178	0.6416	0.5781	0.5193	0.4590	0.4021	0.3430	0.2865	0.2172	**0.8384**
	fpr	0.7776	0.0358	0.0187	0.0118	0.0082	0.0058	0.0040	0.0027	0.0019	0.0013	

Табела 1. Најдобри резултати од директната метода за тежински граф (Ресник) со евалуација на функции со комплетен граф (се добиваат за* c = *0.1)

Во табела 2 се прикажани резултатите добиени со користење на директниот пристап при што евалуацијата на можните функции се врши во рамки на ограничено функционално соседство. Резултатите се однесуваат на репрезентација со едноставен нетежински граф. Од прикажаните резултати се гледа дека AUC вредностите драстично не се менуваат и може да се забележи дека скоро за сите експерименти при ω = 0 се добива задоволителен одзив кој зависи од големината на функционалното соседство и се движи над 80%, а во некои случаи и над 90%. И стапката на грешка се менува во зависност од големината на функционалното соседство, односно поголемото соседство води кон повеќе лажно позитивни функции. Она што дополнително е евидентно е дека резултатот не зависи од големината на параметарот c што би значело дека за нетежинскиот граф функционалното соседство не се менува значително при промена на веројатноста за телепортирање (функционалните соседи се наоѓаат на мали растојанија во графот). Можеби најинтересното во овие резултати е фактот што при зголемувањето на прагот ω од само 0.1, одзивот драстично се намалува на околу 40-50%. Ова води до многу важниот заклучок дека во функционалното соседство на еден протеин добиено со методот на случајна патека се наоѓаат најмалку 80% од функциите со кои е анотиран, но тие добиваат слаби оценки доколку се евалуираат со χ^2 тест. Тоа значи дека во иднина треба да се работи кон наоѓање на поефикасен начин за рангирање на овие функции што би требало значително да ги подобри резултатите.

Треба да напоменеме дека беа направени експерименти со различни големини на соседство, но заради прегледност овде се прикажани само резултатите за големина на соседство 10, 50 и 100. Она што го добивме како резултат укажува дека перформансите на методата се оптимални за големина на соседство од 50 протеини. Ваквиот резултат во принцип е многу осетлив од самите податоци кои се обработуваат односно од конкретната протеинска интеракциска мрежа. Тоа значи дека за други влезни податоци големината на соседство од 50 може да не води кон најдобри перформанси. Ова е уште една од насоките за иден развој, односно наоѓање на начин за автоматско определување на големината на функционалното соседство.

Директна метода за нетежински граф со евалуација на функции со функционално соседство													
	#соседи	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
c = 0.1	**10**	**sens**	0,8090	0.4165	0.2873	0.2140	0.1596	0.1203	0.0941	0.0764	0.0595	0.0456	**0.8388**
		fpr	0.1321	0.0657	0.0486	0.0375	0.0287	0.0204	0.0158	0.0125	0.0072	0.0011	
	50	**sens**	0.9120	0.4498	0.3381	0.2643	0.2090	0.1623	0.1288	0.0973	0.0710	0.0496	**0.8492**
		fpr	0.2730	0.0724	0.0525	0.0388	0.0304	0.0220	0.0167	0.0126	0.0076	0.0013	
	100	**sens**	0.9419	0.4682	0.3640	0.2913	0.2320	0.1838	0.1462	0.1097	0.0746	0.0477	**0.8504**
		fpr	0.3874	0.0800	0.0561	0.0415	0.0309	0.0230	0.0172	0.0124	0.0079	0.0015	
c = 0.3	**10**	**sens**	0.8104	0.4179	0.2901	0.2163	0.1596	0.1211	0.0959	0.0776	0.0609	0.0469	**0.8400**
		fpr	0.1310	0.0655	0.0489	0.0374	0.0287	0.0204	0.0158	0.0125	0.0072	0.0010	
	50	**sens**	0.9132	0.4557	0.3419	0.2639	0.2098	0.1635	0.1272	0.0957	0.0696	0.0489	**0.8504**
		fpr	0.2734	0.0724	0.0522	0.0387	0.0302	0.0219	0.0166	0.0126	0.0075	0.0013	
	100	**sens**	0.9425	0.4703	0.3648	0.2882	0.2309	0.1832	0.1437	0.1088	0.0756	0.0501	**0.8342**
		fpr	0.3865	0.0794	0.0555	0.0409	0.0304	0.0227	0.0169	0.0122	0.0078	0.0015	
c = 0.5	**10**	**sens**	0.8104	0.4187	0.2884	0.2151	0.1598	0.1211	0.0956	0.0780	0.0611	0.0471	**0.8401**
		fpr	0.1307	0.0654	0.0485	0.0375	0.0287	0.0204	0.0157	0.0125	0.0072	0.0010	
	50	**sens**	0.9132	0.4557	0.3422	0.2627	0.2069	0.1613	0.1247	0.0930	0.0689	0.0495	**0.8500**
		fpr	0.2747	0.0722	0.0521	0.0386	0.0301	0.0218	0.0166	0.0125	0.0075	0.0012	
	100	**sens**	0.9424	0.4708	0.3615	0.2878	0.2307	0.1819	0.1442	0.1071	0.0754	0.0500	**0.8335**
		fpr	0.3892	0.0794	0.0553	0.0407	0.0304	0.0227	0.0169	0.0121	0.0078	0.0014	
c = 0.7	**10**	**sens**	0.8116	0.4192	0.2882	0.2151	0.1594	0.1211	0.0955	0.0780	0.0611	0.0471	**0.8407**
		fpr	0.1306	0.0654	0.0485	0.0375	0.0287	0.0204	0.0158	0.0125	0.0072	0.0010	
	50	**sens**	0.9145	0.4541	0.3416	0.2621	0.2063	0.1607	0.1244	0.0925	0.0691	0.0486	**0.8501**
		fpr	0.2759	0.0722	0.0521	0.0386	0.0301	0.0219	0.0166	0.0125	0.0074	0.0012	
	100	**sens**	0.9424	0.4710	0.3615	0.2875	0.2291	0.1810	0.1444	0.1059	0.0737	0.0489	**0.8330**
		fpr	0.3913	0.0792	0.0552	0.0406	0.0303	0.0227	0.0169	0.0121	0.0078	0.0015	
c = 0.9	**10**	**sens**	0.8104	0.4190	0.2882	0.2152	0.1594	0.1211	0.0952	0.0783	0.0612	0.0472	**0.8402**
		fpr	0.1301	0.0654	0.0485	0.0374	0.0287	0.0204	0.0158	0.0125	0.0072	0.0010	
	50	**sens**	0.9145	0.4538	0.3410	0.2609	0.2049	0.1584	0.1234	0.0926	0.0690	0.0480	**0.8496**
		fpr	0.2772	0.0722	0.0520	0.0386	0.0301	0.0218	0.0165	0.0125	0.0074	0.0012	
	100	**sens**	0.9420	0.4728	0.3603	0.2863	0.2288	0.1792	0.1433	0.1051	0.0722	0.0485	**0.8327**
		fpr	0.3926	0.0794	0.0553	0.0406	0.0303	0.0227	0.0169	0.0121	0.0078	0.0015	

Табела 2. Резултати од примена на директната метода врз нетежински граф со евалуација на функции со функционално соседство

Во табелите 3, 4 и 5 се прикажани резултатите добиени при користењето на тежински граф и различните метрики и стратегии за определување на тежините.

Директна метода за тежински граф (Џакард) со евалуација на функции со функционално соседство													
	c =	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
содржинска	**0.1**	**sens**	0.8424	0.6832	0.6157	0.5563	0.4949	0.4473	0.3856	0.3388	0.2685	0.1958	**0.8961**
		fpr	0.1103	0.0373	0.0194	0.0123	0.0083	0.0057	0.0039	0.0027	0.0019	0.0013	
	0.5	**sens**	0.8635	0.6971	0.6322	0.5742	0.5309	0.4604	0.3954	0.3413	0.2939	0.2379	**0.9099**
		fpr	0.1046	0.0229	0.0147	0.0107	0.0083	0.0065	0.0055	0.0044	0.0032	0.0018	
	0.9	**sens**	0.8612	0.6899	0.6294	0.5686	0.5245	0.4616	0.3846	0.3376	0.2925	0.2462	**0.9053**
		fpr	0.1173	0.0273	0.0169	0.0109	0.0086	0.0067	0.0057	0.0051	0.0047	0.0044	
структурна	**0.1**	**sens**	0.8537	0.6998	0.6282	0.5651	0.5004	0.4492	0.3872	0.3396	0.2686	0.1958	**0.9033**
		fpr	0.1084	0.0344	0.0183	0.0114	0.0081	0.0057	0.0039	0.0027	0.0019	0.0013	
	0.5	**sens**	0.8745	0.7133	0.6451	0.5872	0.5383	0.4652	0.3981	0.3425	0.2938	0.2379	**0.9167**
		fpr	0.1022	0.0220	0.0138	0.0101	0.0081	0.0065	0.0055	0.0044	0.0032	0.0018	
	0.9	**sens**	0.8732	0.7044	0.6424	0.5812	0.5315	0.4672	0.3872	0.3382	0.2925	0.2462	**0.9132**
		fpr	0.1121	0.0251	0.0156	0.0104	0.0083	0.0067	0.0057	0.0051	0.0047	0.0044	
хибридна	**0.1**	**sens**	0.8567	0.7058	0.6332	0.5701	0.5032	0.4501	0.3881	0.3399	0.2692	0.1961	**0.9054**
		fpr	0.1079	0.0329	0.0173	0.0107	0.0079	0.0057	0.0039	0.0027	0.0019	0.0013	
	0.5	**sens**	0.8783	0.7198	0.6488	0.5902	0.5407	0.4668	0.3985	0.3430	0.2944	0.2382	**0.9192**
		fpr	0.1014	0.0212	0.0128	0.0097	0.0079	0.0065	0.0055	0.0044	0.0032	0.0018	
	0.9	**sens**	0.8777	0.7098	0.6433	0.5812	0.5335	0.4672	0.3879	0.3387	0.2931	0.2465	**0.9160**
		fpr	0.1114	0.0241	0.0144	0.0099	0.0080	0.0067	0.0057	0.0051	0.0047	0.0044	

Табела 3. Резултати од директна метода врз тежински граф со Џакард метрика

Директна метода за тежински граф (Ресник) со евалуација на функции со функционално соседство													
	c =	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
содржинска	**0.1**	**sens**	0.8735	0.7178	0.6416	0.5781	0.5193	0.4590	0.4021	0.3430	0.2865	0.2172	**0.9139**
		fpr	0.1076	0.0358	0.0187	0.0118	0.0082	0.0056	0.0039	0.0027	0.0019	0.0013	
	0.5	**sens**	0.8935	0.6734	0.6255	0.5796	0.5405	0.5002	0.4573	0.4125	0.3590	0.2872	**0.9249**
		fpr	0.1003	0.0213	0.0134	0.0102	0.0081	0.0065	0.0055	0.0044	0.0032	0.0018	
	0.9	**sens**	0.8923	0.6450	0.6111	0.5760	0.5425	0.5048	0.4454	0.4171	0.3385	0.2745	**0.9212**
		fpr	0.1028	0.0256	0.0154	0.0106	0.0086	0.0067	0.0057	0.0051	0.0047	0.0044	
структурна	**0.1**	**sens**	0.8822	0.7305	0.6537	0.5914	0.5264	0.4641	0.4101	0.3455	0.2877	0.2175	**0.9200**
		fpr	0.1042	0.0324	0.0166	0.0103	0.0081	0.0056	0.0039	0.0027	0.0019	0.0013	
	0.5	**sens**	0.9081	0.6955	0.6418	0.5901	0.5405	0.5058	0.4632	0.4148	0.3596	0.2873	**0.9337**
		fpr	0.0983	0.0198	0.0122	0.0097	0.0080	0.0065	0.0055	0.0044	0.0032	0.0018	
	0.9	**sens**	0.9064	0.6627	0.6275	0.5882	0.5425	0.5093	0.4501	0.4193	0.3391	0.2749	**0.9299**
		fpr	0.1002	0.0232	0.0141	0.0098	0.0083	0.0067	0.0057	0.0051	0.0047	0.0044	
хибридна	**0.1**	**sens**	0.8844	0.7328	0.6551	0.5931	0.5275	0.4653	0.4110	0.3459	0.2879	0.2174	**0.9214**
		fpr	0.1037	0.0313	0.0161	0.0101	0.0080	0.0056	0.0039	0.0027	0.0019	0.0013	
	0.5	**sens**	0.9097	0.6983	0.6445	0.5915	0.5417	0.5064	0.4635	0.4148	0.3596	0.2873	**0.9348**
		fpr	0.0981	0.0191	0.0119	0.0096	0.0080	0.0065	0.0055	0.0044	0.0032	0.0018	
	0.9	**sens**	0.9077	0.6644	0.6301	0.5897	0.5432	0.5101	0.4503	0.4193	0.3391	0.2749	**0.9308**
		fpr	0.1001	0.0217	0.0138	0.0097	0.0083	0.0067	0.0057	0.0051	0.0047	0.0044	

Табела 4. Резултати од директна метода врз тежински граф со метрика на Ресник

Директна метода за тежински граф (Ванг) со евалуација на функции со функционално соседство													
	c =	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
содржинска	**0.1**	**sens**	0.8839	0.7404	0.6534	0.6113	0.5410	0.4743	0.4081	0.3474	0.2815	0.2042	**0.9158**
		fpr	0.1259	0.0482	0.0239	0.0147	0.0096	0.0065	0.0044	0.0029	0.0020	0.0013	
	0.5	**sens**	0.8942	0.7214	0.6632	0.6169	0.5717	0.5227	0.4700	0.4147	0.3552	0.2774	**0.9239**
		fpr	0.1213	0.0246	0.0176	0.0126	0.0095	0.0074	0.0060	0.0048	0.0034	0.0018	
	0.9	**sens**	0.8918	0.6799	0.6519	0.6134	0.5756	0.5289	0.4797	0.4310	0.3846	0.3369	**0.9198**
		fpr	0.1218	0.0256	0.0172	0.0128	0.0099	0.0076	0.0062	0.0054	0.0049	0.0046	
структурна	**0.1**	**sens**	0.8925	0.7521	0.6652	0.6229	0.5497	0.4806	0.4388	0.3499	0.2822	0.2045	**0.9218**
		fpr	0.1229	0.0458	0.0218	0.0132	0.0091	0.0065	0.0044	0.0029	0.0020	0.0013	
	0.5	**sens**	0.9117	0.7372	0.6748	0.6238	0.5813	0.5291	0.4732	0.4161	0.3561	0.2776	**0.9341**
		fpr	0.1188	0.0226	0.0163	0.0118	0.0091	0.0074	0.0060	0.0048	0.0034	0.0018	
	0.9	**sens**	0.9077	0.6962	0.6637	0.6231	0.5845	0.5361	0.4818	0.4325	0.3852	0.3371	**0.9293**
		fpr	0.1198	0.0234	0.0156	0.0119	0.0095	0.0076	0.0062	0.0054	0.0049	0.0046	
хибридна	**0.1**	**sens**	0.8940	0.7553	0.6678	0.6241	0.5509	0.4819	0.4395	0.3505	0.2822	0.2045	**0.9231**
		fpr	0.1221	0.0446	0.0201	0.0124	0.0090	0.0065	0.0044	0.0029	0.0020	0.0013	
	0.5	**sens**	0.9126	0.7391	0.6761	0.6254	0.5830	0.5298	0.4738	0.4161	0.3561	0.2776	**0.9351**
		fpr	0.1175	0.0213	0.0147	0.0103	0.0090	0.0074	0.0060	0.0048	0.0034	0.0018	
	0.9	**sens**	0.9091	0.6986	0.6651	0.6252	0.5858	0.5368	0.4823	0.4325	0.3852	0.3371	**0.9304**
		fpr	0.1186	0.0223	0.0150	0.0116	0.0093	0.0076	0.0062	0.0054	0.0049	0.0046	

Табела 5. Резултати од директна метода врз тежински граф и метрика на Ванг

Во табелите се прикажани резултатите што се добиваат со секоја од предложените стратегии за доделување тежини во графот. Резултатите покажуваат јасно подобрување во однос на нетежинскиот граф како од аспект на сензитивност и одржување на истата без драстични падови, така и од аспект на значително намалување на стапката на грешка што е можеби и позначајно (за ω=0 се движи околу 10%). Ова се должи на тоа што благодарение на тежините алгоритамот многу подобро го одредува функционалното соседство. Прикажани се единствено резултатите за вредноста на *c* поставена на 0.1, 0.5 и 0.9 заради прегледност и затоа што овие се доволни за да се воочи „тенденцијата“. Перформансите се најдобри кога *c* = 0.5. Толкувањето на овој резултат се сведува на постоењето на еден вид на шум кога *c* е помало (се разгледува поголем дел од графот и во функционалното соседство влегуваат и протеини кои носат функции што предизвикуваат грешки во предвидувањето) и некомплетноста на функционалното соседство кога е поголемо.

Семантичките метрики базирани на GO даваат подобри резултати споредено со корелациската метрика на Џакард, но сепак резултатите не се драстично во полза на првите. Ова може да биде индикатор на тоа дека кога се бараат сличности помеѓу јазлите во графот треба да се конструира хибридна метрика која ќе ги земе во предвид и семантичката сличност според GO и корелацијата помеѓу јазлите. Од аспект на двете GO метрики резултатите се речиси идентични со тоа што метриката на Ванг има поголема сензитивност, а метриката на Ресник пониска стапка на грешка. Од алгоритамски аспект како што претходно кажавме ќе го префериравме вториот случај, па од овие резултати би донеле заклучок дека подобро е да се користат GO метриките базирани на информациска содржина (Ресник), иако не треба целосно да се исклучат ниту хибридните метрики (Ванг). Подетална дискусија за тоа која е предноста на високата сензитивност, а која на ниската стапка на грешка ќе дадеме во анализата на резултатите од функционалната анотација базирана на кластерирање.

Различните стратегии за доделување на тежините по самата своја дефиниција се на некој начин адитивни па затоа и добиените резултати се очекувани, па како најдобра стратегија се издвојува хибридната, а по неа следуваат структурната и содржински базираната. Од табелите може да се забележи дека кога праговата вредност ќе надмине 0.5 резултатите се стабилизираат гледано од аспект на различни стратегии и различни вредности на c. Тоа се должи на фактот што доминантните функции во функционалното соседство на прашалниот протеин, без оглед на пристапот, се повторуваат бидејќи потекнуваат од протеини што се во неговото тополошко соседство и токму на нив се должи повторливоста на резултатите. Оваа е во склад и со резултатите кои се добиваат со други директни методи кои работат со тополошки соседства во графот, а кои накратко ќе ги објасниме подолу и ќе ги споредиме со нашиот пристап.

Во табелата 6 се прикажани резултатите кои се добиваат кога како репрезентација на протеинската мрежа се користат протеин-термин графот и комплетно функционално поврзаниот граф. Резултатите се прикажани само за најдобриот случај (за c = 0.1, соседство со големина 50 и хибридни тежини добиени со метриката на Ресник) заради прегледност и заради тоа што различните

комбинации на параметри се однесуваат идентично како во претходните експерименти, па овие резултати е доволно репрезентативни за да можеме да ја направиме дискусијата.

Директна метода за протеин-термин и комплетно функционално поврзан граф со евалуација на функции со функционално соседство												
	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
протеин-термин	**sens**	0.9591	0.7812	0.6893	0.6075	0.5336	0.4602	0.3949	0.3313	0.2630	0.1912	**0.8672**
	fpr	0.8459	0.0890	0.0324	0.0218	0.0206	0.0136	0.0046	0.0031	0.0021	0.0013	
комплетно функционално поврзан	**sens**	0.9215	0.7478	0.6352	0.5312	0.4457	0.3809	0.3281	0.2819	0.2247	0.1603	**0.8816**
	fpr	0.4776	0.0761	0.0315	0.0210	0.0180	0.0112	0.0042	0.0029	0.0020	0.0011	

Табела 6. Резултати од директната метода врз протеин-термин и комплетно функционално поврзан граф

Како што се гледа од табелата и двете граф репрезентации имаат висока сензитивност, но и многу висока стапка на грешка (кај протеин-термин графот од дури 84%). Ова не е неочекувано со оглед на тоа дека овие репрезентации се многу карактеристични по својата природа. Протеин-термин графот има јазли-термини кои се врзуваат со сите јазли-протеини за кои се доделени, што значи дека јазлите-термини ќе имаат многу висок степен во ваквата репрезентација, што пак од друга страна значи дека случајниот пешак (кој го дефинира нашиот метод) со многу поголема веројатност ќе ги посетува ваквите јазли и истите ќе имаат високи афинитети кон прашалниот протеин. Ваквиот случај се сведува (помалку или повеќе) на пристапот за евалуација на функции врз основа на целиот граф, па заради тоа и резултатите се слични. Значи потребно е да се измине поголем дел од графот (c = 0.1) за да можат да се покријат поголем дел од функциите на прашалниот протеин (висока сензитивност), но барајќи ги нив се наоѓаат и многу други (висока стапка на грешка). Истото важи и за комплетно функционално поврзаниот граф каде функционалните соседства се многу блиски до тополошките и повторно по цена на побогата анотација се прават повеќе грешки.

Со цел да се направи споредба на предложениот директен метод за функционална анотација со некој од претходно развиените системи разгледани се два алгоритми. Овие два алгоритми се избрани затоа што се исти по природа со предложениот

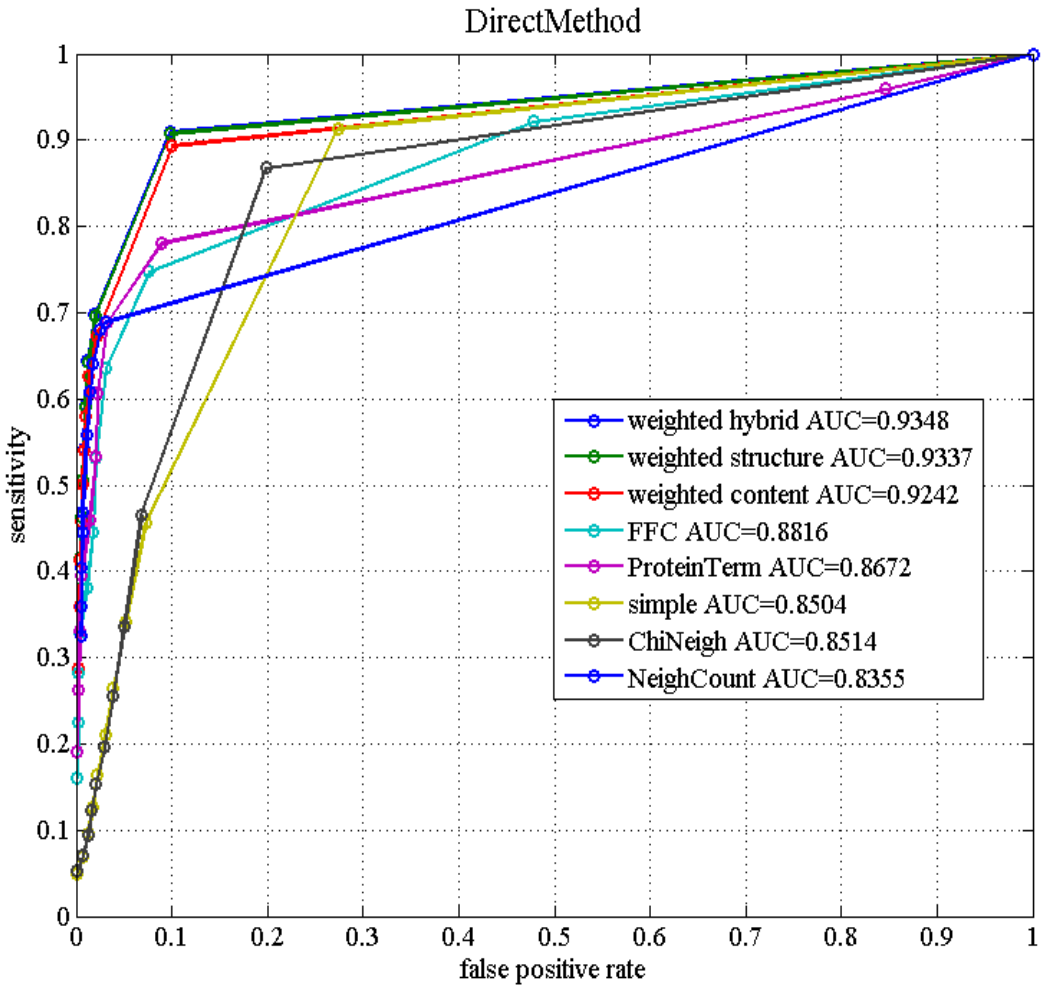
односно ги користат само информациите содржани во протеинската интеракциска мрежа и се обидуваат преку некоја дефиниција на соседство да предвидат функција на прашален протеин. Станува збор за пионерскиот алгоритам во оваа област, наречен neighborhood counting во соседство и опишан во [117], кој ги брои појавувањата на секоја функција меѓу директните соседи на прашалниот протеин, како и пософистицираниот алгоритам предложен во [118] во кој се разгледуваат функциите во соседство со одреден радиус на прашалниот протеин и секоја од нив се оценува со χ^2 тест, и кој во поглавјето 3.2.1 беше наречен χ^2 neighborhood. Во табела 7 се дадени резултатите од предвидувањето на овие два алгоритми за нашата протеинска интеракциска мрежа. За вториот од нив, радиусот на соседството кое се разгледува за прашалниот протеин е 2.

Neighborhood counting и χ2 neighborhood алгоритми												
	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
Neighborhood counting	**sens**	0.6899	0.6803	0.6411	0.6077	0.5590	0.4690	0.4457	0.4033	0.3595	0.3255	**0.8355**
	fpr	0.0306	0.0252	0.0179	0.0146	0.0109	0.0065	0.0062	0.0056	0.0053	0.0052	
χ2 neighborhood	**sens**	0.8674	0.4649	0.3369	0.2551	0.1964	0.1540	0.1225	0.0949	0.0716	0.0524	**0.8514**
	fpr	0.1983	0.0679	0.0500	0.0388	0.0294	0.0208	0.0161	0.0127	0.0073	0.0011	

Табела 7. Резултати добиени од споредбените директни алгоритми за функционална анотација

Додека методот neighborhood counting има многу мал одзив од 68.99% за ω=0, методот χ^2 neighborhood има одзив од 86.74%, но тоа е на сметка на стапката за грешка, која изнесува 19.83%. Со зголемување на ω одзивот се намалува на само 46.49%. Ова однесување е слично резултатите од нашиот метод за нетежински граф со евалуација на функциите со χ^2 тест во функционалното соседство (табела 2). Евидентно е дека нашиот метод има многу подобри перформанси.

На слика 6.1 е прикажан график за ROC кривите од најдобрите добиени резултати за секоја од различните граф репрезентации за протеинската интеракциска мрежа и двата споредбени методи.



Слика 6.1 ROC криви за најдобрите резултати од директниот метод за функционална анотација

Пред да започнеме со евалуацијата на методите за кластерирање потребно е да ги дефинираме влезните параметри за алгоритмите кои имаат потреба од тоа. Имено за класичните методи дефинирани во поглавјата 5.3.1.1-5.3.1.4 има потреба од дефинирање на влезен параметар и тоа за *k*-медоиди и спектралното кластерирање потребно е да се зададе бројот на кластери, додека за алгоритамот со средишност на врски е потребно да се специфицира колку од врските ќе бидат отстранети. Овие вредности во нашите експерименти беа изведени емпириски и тоа врз основа на ефектот на овие параметри врз перформансите на функционалната анотација.

За бројот на врски што треба да бидат отстранети кај EdgeBetweenness алгоритамот беа тестирани вредности во опсег од 500 до 2000. Она што го

забележавме е дека квалитетот на функционалната анотација во рамки на овој опсег достигнува врв за вредности околу 1000 (може да се избере било која вредност ±10), при што налево и надесно оваа вредност опаѓа. Бројот на отстранети врски кај овој алгоритам влијае на бројот на кластери што се добиваат па според тоа лево од „оптималната“ вредност има помалку кластери кои следствено на тоа се поголеми па имаат и повеќе функции внатре во кластерот со кои можат да анотираат некој прашален протеин. Ова е добро од аспект на сензитивноста, но од аспект на стапката на грешка се случува истото што го имавме и кај директната метода со евалуација на функции од комплетниот граф, односно добиваме големи вредности на стапката на грешка бидејќи на прашалниот протеин му се доделуваат многу функции што не треба да ги има. Од друга страна, десно од „оптималната“ вредност го имаме обратниот случај односно повеќе помали кластери со кои се добиваат ниски стапки на грешка, но тоа е по цена на сензитивноста која се намалува. Во табела 8 се дадени пример на резултатите од три можни вредности, едната „оптимална“ (1000), една лево од неа (800) и другата десно од неа (1400). Оваа анализа се однесува на нетежинскиот граф меѓутоа заклучоците важат без разлика каква репрезентација ќе се искористи. Значи резултатите за EdgeBetweenness се однесуваат на 1000 отстранети врски.

Ефект на број на отстранети врски врз перформансите на функционалната анотација												
#отстранети врски	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
800	**sens.**	0.7712	0.6720	0.5476	0.4631	0.3761	0.3049	0.2561	0.2038	0.1547	0.1188	0.8477
	fpr	0.1571	0.0525	0.0327	0.0148	0.0091	0.0056	0.0039	0.0028	0.0016	0.0011	
1000	**sens.**	0.7578	0.6693	0.5492	0.4753	0.3957	0.3266	0.2859	0.2459	0.1845	0.1445	0.8515
	fpr	0.1093	0.0456	0.0221	0.0136	0.0083	0.0051	0.0037	0.0027	0.0016	0.0011	
1400	**sens.**	0.7090	0.6651	0.5761	0.5131	0.4448	0.3741	0.3304	0.2872	0.2237	0.1623	0.8383
	fpr	0.0619	0.0355	0.0180	0.0118	0.0078	0.0046	0.0035	0.0025	0.0017	0.0012	

Табела 8. Резултати од функционална анотација со EdgeBetweenness при различен број на отстранети врски

Кога станува за предефинирање на бројот на кластери за *k*-медоиди и спектралното кластерирање доволно е да се евалуира еден од алгоритмите и добиениот „оптимален“ број на кластери да се искористи и за другата метода

бидејќи се работи за истата протеинска интеракциска мрежа чија модуларна структура е идентична без оглед на применетиот алгоритам. И овде важи истата дискусија како погоре околу добивањето на „оптималната“ вредност и што се случува лево и десно од неа. Во табела 9 се дадени резултатите направени од емпириската анализа за бројот на кластери кај спектралното кластерирање. Според ова, сите понатамошни резултати за спектрално кластерирање и k-медоиди се однесуваат на 150 кластери.

Ефект на број на кластери (спектрално кластерирање) врз перформансите на функционалната анотација												
#кластери	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
100	**sens.**	0.7517	0.6701	0.5566	0.4712	0.4075	0.3375	0.2921	0.2429	0.1841	0.1404	0.8447
	fpr	0.1258	0.0526	0.0254	0.0156	0.0099	0.0061	0.0043	0.0030	0.0018	0.0011	
150	**sens.**	0.7490	0.6709	0.5725	0.5052	0.4323	0.3619	0.3166	0.2687	0.2108	0.1597	0.8481
	fpr	0.1108	0.0439	0.0213	0.0141	0.0089	0.0055	0.0039	0.0028	0.0019	0.0012	
200	**sens.**	0.7234	0.6783	0.5914	0.5294	0.4630	0.3859	0.3484	0.3053	0.2405	0.1870	0.8450
	fpr	0.0667	0.0382	0.0185	0.0116	0.0077	0.0048	0.0036	0.0027	0.0016	0.0012	

Табела 9. Резултати од функционална анотација со спектрално кластерирање при различен број на кластери

Евалуацијата на методите кои вршат функционална анотација врз база на кластерирање во графот од протеинската интеракциска мрежа ја започнуваме со евалуација на самите алгоритми за кластерирање. Ваквата евалуација ја правиме според постапката која беше образложена во поглавјето 5.3.2. Во табела 10 се дадени подредени пресметаните вредности за нормализираната здружена информација како мерка за способноста на алгоритмите да репродуцираат познати кластери во рамки на еден граф. Покрај алгоритмите кои беа опишани во поглавјето 5.3.1 пресметани се и соодветните вредности за нашата протеинска интеракциска мрежа за алгоритми за кластерирање кои во литературата се користени за функционална анотација во протеински интеракциски мрежи, односно за MCL [134], RNSC [135], MCODE [130] и SPC [136]. Може да се види дека споредено со претходно користените алгоритми во литературата алгоритмите во овој труд работат или многу подобро или спордливо добро. Овие вредности се индикативни и за општите перформанси на алгоритмите во процесот на функционална анотација. Односно, „рангирањето“ на алгоритмите според способноста да предвидат функција на непознат протеин го следи ваквиот редослед.

Вредности на NMI	
алгоритам	**NMI**
Infomap	0.9916
TimeBGLL	0.9062
EdgeCluster	0.8732
BGLL	0.8514
FC	0.8230
EdgeBetweenness	0.6981
HO	0.5283
MCL	0.4979
Spectral	0.4733
RNSC	0.4562
Aglomerative	0.4125
K-medoids	0.3057
MCODE	0.2360
SPC	0.2147

Табела 10. Вредности на нормализираната здружена информација за алгоритмите за кластерирање

Разликите кои се јавуваат во NMI вредностите пред сè се должат на бројот и големината на кластерите која соодветните алгоритми ја враќаат. Од тој аспект најдетални кластери (најмногу на број и најмали по просечна големина на кластер) дава Infomap алгоритамот што е одразено во неговата екстремно висока NMI вредност. Како што опаѓаат NMI вредностите така опаѓа и бројот на кластери и истите се зголемуваат. На пример односот на бројот на добиени кластери од Infomap и групата на алгоритми што следува после него (BGLL, TimeBGLL, EdgeCluster и FC) е отприлика 2.5 : 1. Класичните алгоритми очекувано имаат помали NMI вредности од едноставна причина што кај нив кластерирањето се одвива „круто“ (со предефиниран број на кластери). HomogeneityOptimization алгоритамот кој е предложен овде е полош од најдобрите меѓутоа подобар од споредбените алгоритми. Основната причина за неговата вредност е што истиот има тенденција да гради големи кластери, што би значело дека просторот за негово подобрување би бил во изнаоѓање на начин за вметнување на структурата на кластерите во функцијата за квалитет, или дефинирање на нова функција за квалитет како онаа за модуларноста, со тоа што нултиот модел треба да ги зема во предвид и анотациите придружени кон јазлите. Овие проблеми се планирани да бидат решавани во идни проширувања на овој труд.

Во следниот чекор кластерирањата ги евалуираме од аспект на хомогеноста на кластерите кои се добиваат. Овде напоменуваме дека при пресметката на хомогеноста ги земаме само кластерите кои имаат барем три јазли (протеини). Во табела 11 се дадени пресметаните вредности за алгоритмите во комбинација со секоја можна репрезентација на графот на протеинската интеракциска мрежа. Притоа за тежинскиот граф, резултатите се однесуваат на тежини добиени преку метриката на Ресник (подолу ќе покажеме дека и овде метриките од аспект на перформанси се однесуваат како кај директните методи).

Ентропија на кластерирањето за секој алгоритам и секоја репрезентација						
Репрезентација / Алгоритам	едноставен граф	тежински граф (содржина)	тежински граф (структура)	тежински граф (хибриден)	протеин-термин граф	комплетно функционално поврзан граф
Infomap	0.2528	0.3034	0.3018	0.3002	0.3156	0.5361
TimeBGLL	0.3064	0.3381	0.3271	0.3213	0.5832	0.5783
EdgeCluster	0.2953	0.3294	0.3216	0.3172	0.5716	0.6713
BGLL	0.2707	0.3113	0.3027	0.2993	0.5613	0.6472
FC	0.2807	0.3121	0.3042	0.3001	0.5589	0.6452
EdgeBetweenness	0.3012	0.3078	0.3044	0.3021	0.3544	0.7012
HO	0.4536	0.4027	0.3999	0.3929	0.6234	0.5241
Spectral	0.3053	0.3037	0.3022	0.3012	0.3765	0.5799
K-medoids	0.5569	0.5207	0.5153	0.5076	0.5997	0.6143
Aglomerative	0.5364	0.5112	0.5082	0.4996	0.7821	0.5974

Табела 11. Вредности за ентропиите за секој од користените алгоритми со секоја од граф репрезентациите

Земајќи ја во предвид дефиницијата на мерката за ентропија, пониските вредности на ентропија би значеле алгоритам кој е построг во идентификувањето на функционално кохерентни кластери. Втор и поинтересен аспект на ентропијата, а поврзан со истражувањето во овој труд е корелацијата на вредностите на ентропијата и резултатите од функционалната анотација со користење на алгоритми за кластерирање. Имено, колку е помала ентропијата на еден алгоритам, толку опфатноста на просечниот кластер е помала. Опфатноста на еден кластер овде е дефинирана како односот помеѓу бројот на термини (функции) присутни во кластерот и бројот на термини присутни во целиот граф.

Кластерите со помала опфатност водат кон правење на помал број на грешки во процесот на доделување на функции на прашалниот протеин, но лошата страна на ваквите кластери е што кај нив може да недостасуваат функциите потребни за точна анотација. Од аспект на дефинициите искористени за валидација на анотацијата (матрицата на забуна) ова би значело дека ниски вредности на ентропијата водат кон пониски вредности на лажно позитивните примероци (FP), но и повисоки вредности за лажно негативните примероци (FN). За високи вредности на ентропијата важи обратното.

Пред да ги видиме комплетните резултати од функционалната анотација кај секој од алгоритмите ќе го покажеме ефектот на различните мерки врз основа на кои се добиваат тежините во тежинскиот граф. Заради прегледност „однесувањето“ на метриките ќе го покажеме на резултатите кои се добиваат при различни метрики за Infomap алгоритамот затоа што однесувањето е идентично и за сите останати алгоритми. Во табелите 12-14 се дадени резултатите од функционалната анотација со користење на Infomap алгоритамот и тежински граф со тежини определени со секоја од метриките.

Infomap и Џакард метрика												
тежина	**ω =**	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
содржинска	**sens.**	0.7757	0.6818	0.5901	0.5203	0.4549	0.3913	0.3604	0.3133	0.2562	0.1996	**0.873**
	fpr	0.0547	0.0309	0.0163	0.0115	0.0071	0.0042	0.0034	0.0024	0.0017	0.0013	
структурна	**sens.**	0.8034	0.7123	0.6301	0.5563	0.491	0.4204	0.3805	0.3342	0.2633	0.2136	**0.8886**
	fpr	0.0536	0.0283	0.0133	0.0109	0.0069	0.0038	0.0031	0.0023	0.0012	0.0011	
хибридна	**sens.**	0.8104	0.7254	0.6457	0.561	0.4951	0.4234	0.3831	0.3363	0.2651	0.2240	**0.8928**
	fpr	0.0531	0.0277	0.0124	0.0098	0.0064	0.0038	0.0031	0.0021	0.0012	0.0010	

Табела 12. Резултати од функционална анотација со Infomap врз тежински граф со Џакард метрика

Infomap и Ресник метрика												
тежина	**ω =**	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
содржинска	**sens.**	0.8053	0.6588	0.5834	0.5257	0.4645	0.4311	0.4223	0.3845	0.3213	0.2489	**0.8882**
	fpr	0.0502	0.0295	0.0150	0.0110	0.0069	0.0042	0.0034	0.0024	0.0017	0.0013	
структурна	**sens.**	0.8368	0.6959	0.6268	0.5592	0.4932	0.4610	0.4456	0.4065	0.3291	0.2630	**0.9056**
	fpr	0.0499	0.0265	0.0117	0.0105	0.0068	0.0038	0.0031	0.0023	0.0012	0.0011	
хибридна	**sens.**	0.8417	0.7034	0.6414	0.5623	0.4961	0.4630	0.4481	0.4081	0.3303	0.2731	**0.9086**
	fpr	0.0499	0.0247	0.0115	0.0097	0.0065	0.0038	0.0031	0.0021	0.0012	0.0010	

Табела 13. Резултати од функционална анотација со Infomap врз тежински граф со Ресник метрика

Infomap и Ванг метрика												
тежина	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
содржинска	**sens.**	0.8064	0.7061	0.6211	0.5630	0.4957	0.4536	0.4350	0.3867	0.3175	0.2391	**0.8862**
	fpr	0.0714	0.0326	0.0192	0.0134	0.0083	0.0051	0.0039	0.0028	0.0019	0.0013	
структурна	**sens.**	0.8406	0.7362	0.6598	0.5929	0.5340	0.4843	0.4556	0.4078	0.3256	0.2533	**0.9053**
	fpr	0.0702	0.0289	0.0158	0.0126	0.0079	0.0047	0.0036	0.0027	0.0014	0.0011	
хибридна	**sens.**	0.8447	0.7447	0.6730	0.5962	0.5374	0.4864	0.4584	0.4094	0.3268	0.2634	**0.9082**
	fpr	0.0692	0.0278	0.0143	0.0104	0.0075	0.0047	0.0036	0.0025	0.0014	0.0010	

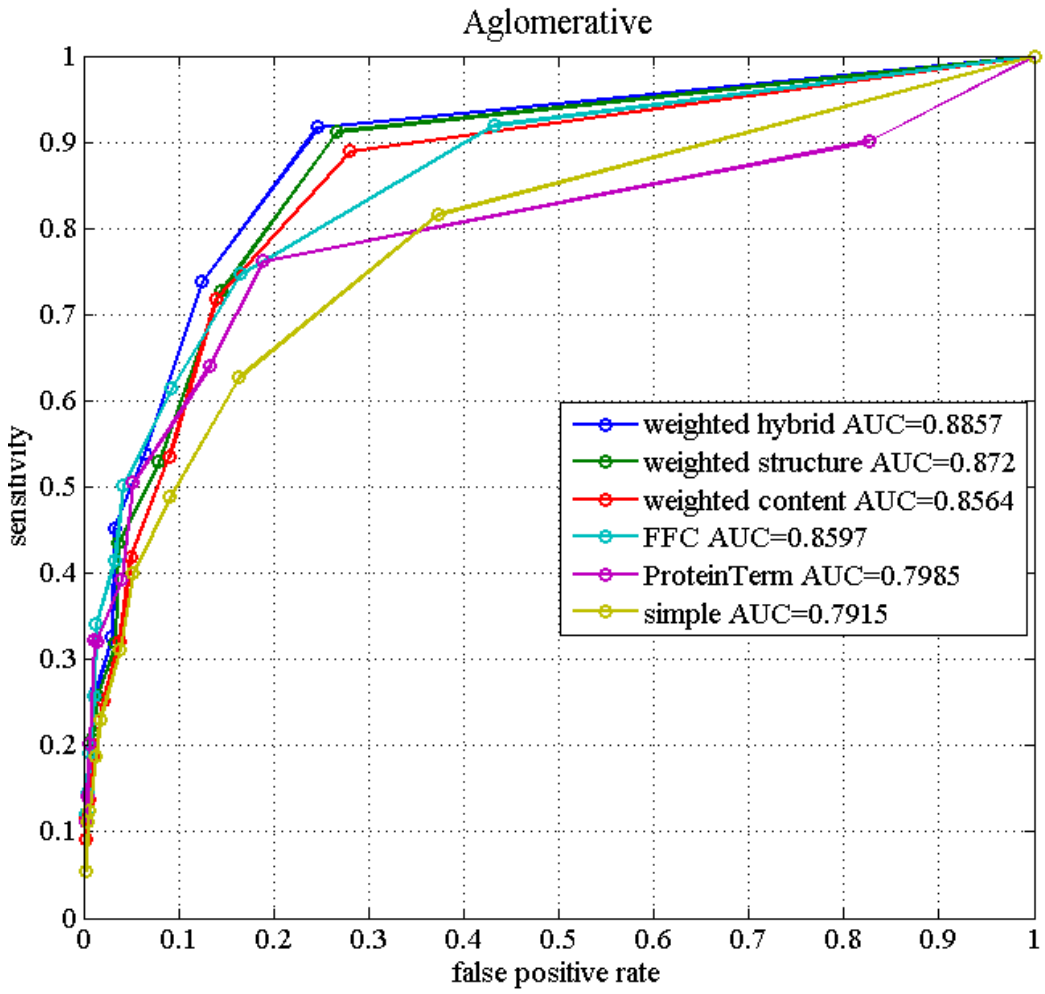
Табела 14. Резултати од функционална анотација со Infomap врз тежински граф со Ванг метрика

И овде исто како кај директниот метод доаѓаме до истиот заклучок, дека семантичките метрики базирани на GO даваат подобри резултати од корелациската метрика, а метриката на Ресник може да се смета за нијанса подобра од метриката на Ванг. Секако и овде користењето на било која од овие метрики не смее да се исклучи целосно. Во преостанатиот дел од оваа глава резултатите за тежинските графови ќе подразбираат користење на метрика на Ресник за добивање на тежините.

Во продолжение во табелите 15-24 се дадени резултатите добиени од функционалната анотација со секој од алгоритмите за кластерирање во детали објаснети во поглавјето 5.3.1, дополнително со секоја табела е придружен и график на кој се прикажани ROC кривите како визуелен приказ на податоците. Притоа во табелите и графиците е користена следната нотација: simple за едноставниот нетежински граф, weighted content за тежински граф со содржински тежини, weighted structure за тежински граф со структурни тежини, weighted hybrid за тежински граф со хибридни тежини, FFC (FullFunctionalConnected) за комплетно функционално поврзан граф и ProteinTerm за протеин-термин графот. AUC вредностите прикажани на графиците до секоја од репрезентациите се однесуваат на соодветните површини под ROC кривите.

Функционална анотација со агломеративно кластерирање												
тип граф	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
Simple	Sens.	0.8166	0.6280	0.4886	0.3997	0.3123	0.2302	0.1878	0.1244	0.1117	0.0546	0.7915
	fpr	0.3736	0.1634	0.0912	0.0524	0.0381	0.0174	0.0131	0.0062	0.0044	0.0022	
Weighted Content	Sens.	0.8896	0.7177	0.5351	0.4182	0.3204	0.2514	0.1916	0.1369	0.1156	0.0909	0.8564
	fpr	0.2799	0.139	0.0899	0.0495	0.0373	0.0204	0.0111	0.0052	0.0032	0.0019	
Weighted Structure	Sens.	0.9124	0.7272	0.5292	0.4352	0.3181	0.2578	0.2074	0.1428	0.1192	0.0909	0.8720
	fpr	0.2667	0.1443	0.0793	0.0364	0.0337	0.0144	0.0100	0.0048	0.0032	0.0019	
Weighted Hybrid	Sens.	0.9176	0.7387	0.5396	0.4526	0.3266	0.2622	0.2074	0.1428	0.1192	0.0909	0.8857
	fpr	0.246	0.1252	0.0643	0.0326	0.0300	0.0126	0.0100	0.0048	0.0032	0.0019	
Protein-Term	Sens.	0.9012	0.7623	0.6398	0.5054	0.3931	0.3205	0.3234	0.2022	0.1406	0.1112	0.7985
	fpr	0.8259	0.1890	0.1324	0.0518	0.0406	0.0136	0.0116	0.0063	0.0041	0.0022	
FFC	Sens.	0.9203	0.7478	0.6152	0.5012	0.4157	0.3409	0.2581	0.1919	0.1447	0.1203	0.8597
	fpr	0.4326	0.1654	0.0923	0.0421	0.0323	0.0128	0.0112	0.0055	0.0041	0.0022	

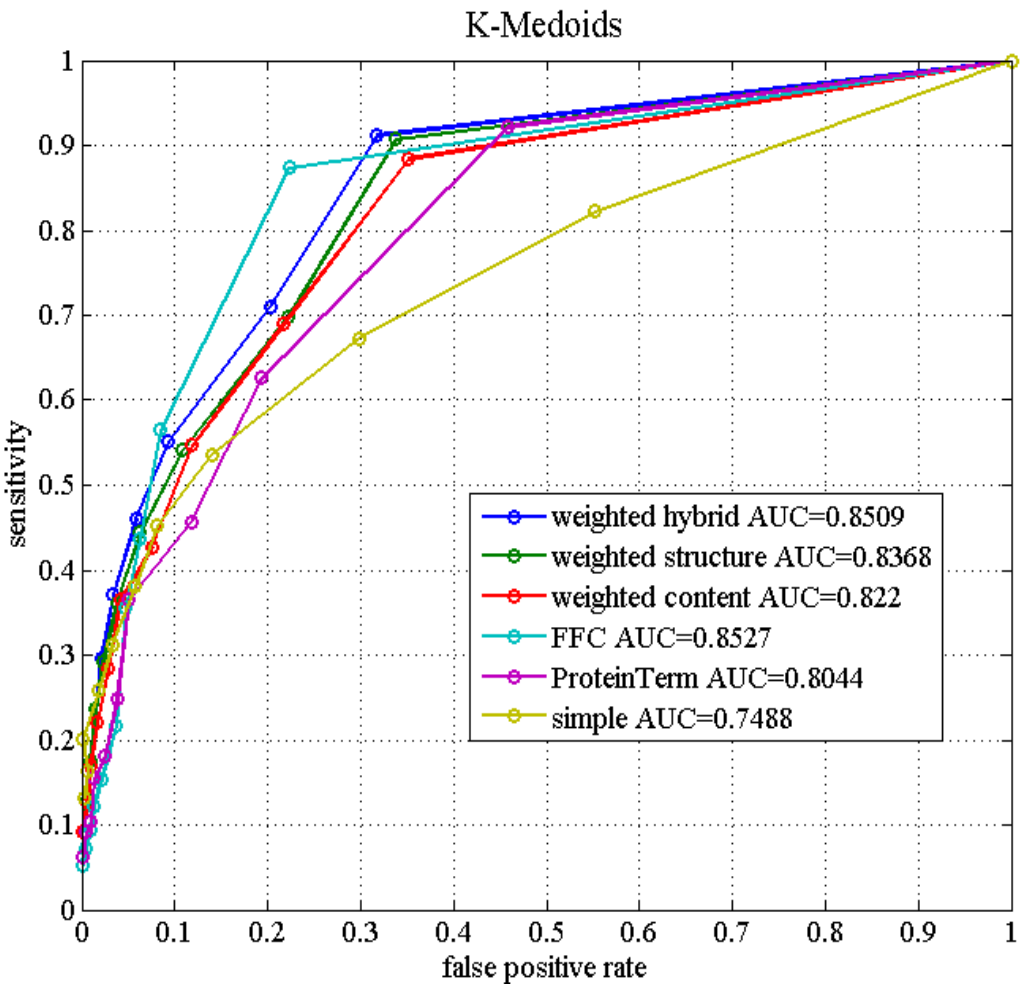
Табела 15. Резултати од функционална анотација со користење на агломеративно кластерирање



Слика 6.2 ROC криви за функционална анотација со агломеративно кластерирање за секоја граф репрезентација

Функционална анотација со *k*-медоиди кластерирање												
тип граф	**ω =**	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
Simple	**Sens.**	0.8221	0.6729	0.5366	0.4533	0.3802	0.3126	0.2586	0.2017	0.1632	0.1325	**0.7488**
	fpr	0.5514	0.2980	0.1398	0.0813	0.0572	0.0325	0.0182	0.00112	0.0063	0.0022	
Weighted Content	**Sens.**	0.8845	0.6893	0.5475	0.4268	0.3646	0.2848	0.2205	0.1684	0.1257	0.0928	**0.8220**
	fpr	0.3514	0.2164	0.1188	0.0753	0.0411	0.0285	0.0163	0.0093	0.0059	0.0017	
Weighted Structure	**Sens.**	0.9073	0.6988	0.5416	0.4438	0.3623	0.2912	0.2363	0.1743	0.1293	0.0928	**0.8368**
	fpr	0.3382	0.2217	0.1082	0.0622	0.0375	0.0225	0.0152	0.0089	0.0059	0.0017	
Weighted Hybrid	**Sens.**	0.9125	0.7103	0.552	0.4612	0.3708	0.2956	0.2363	0.1743	0.1293	0.0928	**0.8509**
	fpr	0.3175	0.2026	0.0932	0.0584	0.0338	0.0207	0.0152	0.0089	0.0059	0.0017	
Protein-Term	**Sens.**	0.9225	0.6268	0.4573	0.3662	0.2488	0.1810	0.1553	0.1032	0.0919	0.0633	**0.8044**
	fpr	0.4591	0.1932	0.1175	0.0503	0.0382	0.0254	0.0133	0.0095	0.0047	0.0012	
FFC	**Sens.**	0.8739	0.5649	0.4369	0.3551	0.2164	0.1540	0.1225	0.0949	0.0716	0.0524	**0.8527**
	fpr	0.2234	0.0845	0.0623	0.0474	0.0366	0.0212	0.0122	0.0087	0.0043	0.0011	

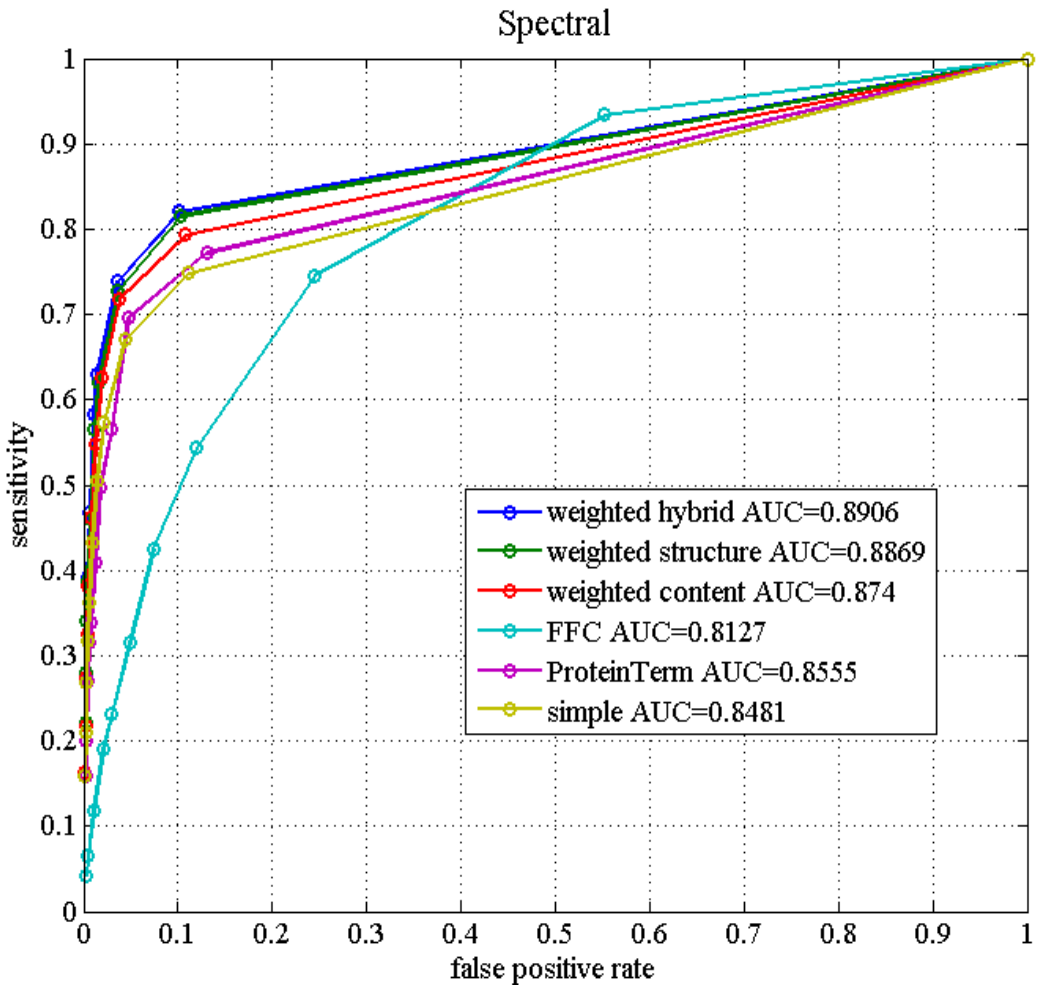
Табела 16. Резултати од функционална анотација со користење на k-медоиди кластерирање



Слика 6.3 ROC криви за функционална анотација со кластерирање со k-медоиди за секоја граф репрезентација

Функционална анотација со спектрално кластерирање												
тип граф	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
Simple	**Sens.**	0.7490	0.6709	0.5725	0.5052	0.4323	0.3619	0.3166	0.2687	0.2108	0.1597	**0.8481**
	fpr	0.1108	0.0439	0.0213	0.0141	0.0089	0.0055	0.0039	0.0028	0.0019	0.0012	
Weighted Content	**Sens.**	0.7932	0.7185	0.6260	0.5487	0.4619	0.3811	0.3244	0.2735	0.2187	0.1633	**0.8740**
	fpr	0.1083	0.0382	0.0193	0.0123	0.0079	0.0049	0.0036	0.0026	0.0017	0.0011	
Weighted Structure	**Sens.**	0.8160	0.7280	0.6201	0.5657	0.4596	0.3875	0.3402	0.2794	0.2223	0.1633	**0.8869**
	fpr	0.1031	0.0364	0.0155	0.0115	0.0067	0.0041	0.0025	0.0022	0.0017	0.0011	
Weighted Hybrid	**Sens.**	0.8212	0.7395	0.6305	0.5831	0.4681	0.3919	0.3402	0.2794	0.2223	0.1633	**0.8906**
	fpr	0.1018	0.0358	0.0137	0.0106	0.0064	0.0040	0.0025	0.0022	0.0017	0.0011	
Protein-Term	**Sens.**	0.7733	0.6968	0.5660	0.4967	0.4099	0.3384	0.3149	0.2704	0.2008	0.1588	**0.8555**
	fpr	0.1310	0.0473	0.0298	0.0172	0.0118	0.0076	0.0056	0.0043	0.0031	0.0017	
FFC	**Sens.**	0.9338	0.7458	0.5445	0.4257	0.3153	0.2310	0.1901	0.1191	0.0660	0.0427	**0.8127**
	fpr	0.5524	0.2448	0.1199	0.0738	0.0485	0.0300	0.0204	0.0103	0.0050	0.0031	

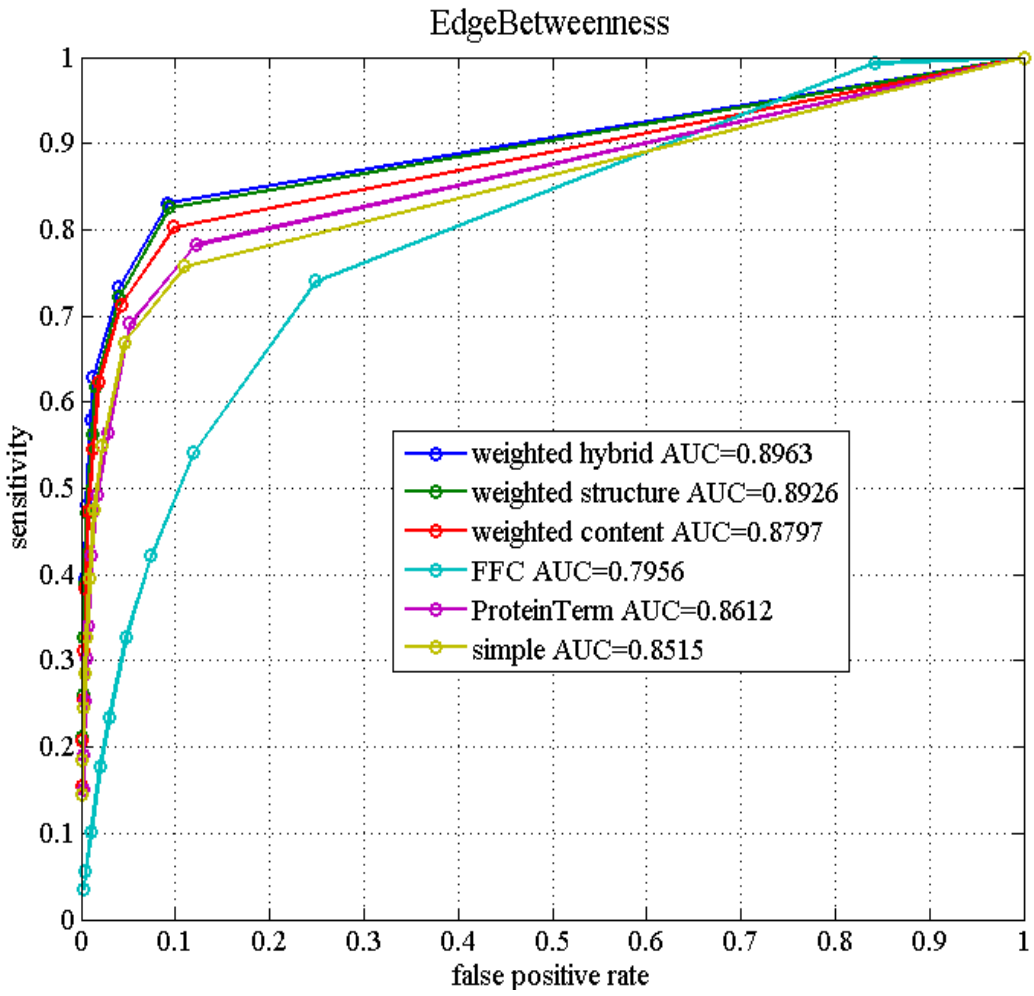
Табела 17. Резултати од функционална анотација со користење на спектрално кластерирање



Слика 6.4 ROC криви за функционална анотација со спектрално кластерирање со секоја граф репрезентација

Функционална анотација со кластерирање според средишност на врски												
тип граф	**ω =**	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
Simple	**Sens.**	0.7578	0.6693	0.5492	0.4753	0.3957	0.3266	0.2859	0.2459	0.1845	0.1445	**0.8515**
	fpr	0.1093	0.0456	0.0221	0.0136	0.0083	0.0051	0.0037	0.0027	0.0016	0.0011	
Weighted Content	**Sens.**	0.8031	0.7132	0.6238	0.5449	0.4743	0.3842	0.3114	0.2557	0.2086	0.1550	**0.8797**
	fpr	0.0987	0.0421	0.0184	0.0122	0.0068	0.0045	0.0032	0.0024	0.0014	0.0011	
Weighted Structure	**Sens.**	0.8259	0.7227	0.6179	0.5619	0.4720	0.3906	0.3272	0.2616	0.2122	0.1550	**0.8926**
	fpr	0.0935	0.0403	0.0146	0.0114	0.0056	0.0037	0.0021	0.0020	0.0014	0.0011	
Weighted Hybrid	**Sens.**	0.8311	0.7342	0.6283	0.5793	0.4805	0.3950	0.3272	0.2616	0.2122	0.1550	**0.8963**
	fpr	0.0922	0.0397	0.0128	0.0105	0.0053	0.0036	0.0021	0.002	0.0014	0.0011	
Protein-Term	**Sens.**	0.7832	0.6915	0.5638	0.4929	0.4223	0.3415	0.3019	0.2526	0.1907	0.1505	**0.8612**
	fpr	0.1214	0.0512	0.0289	0.0171	0.0107	0.0072	0.0052	0.0041	0.0028	0.0017	
FFC	**Sens.**	0.9937	0.7405	0.5423	0.4219	0.3277	0.2341	0.1771	0.1013	0.0559	0.0344	**0.7956**
	fpr	0.8428	0.2487	0.1190	0.0737	0.0474	0.0296	0.0200	0.0101	0.0047	0.0031	

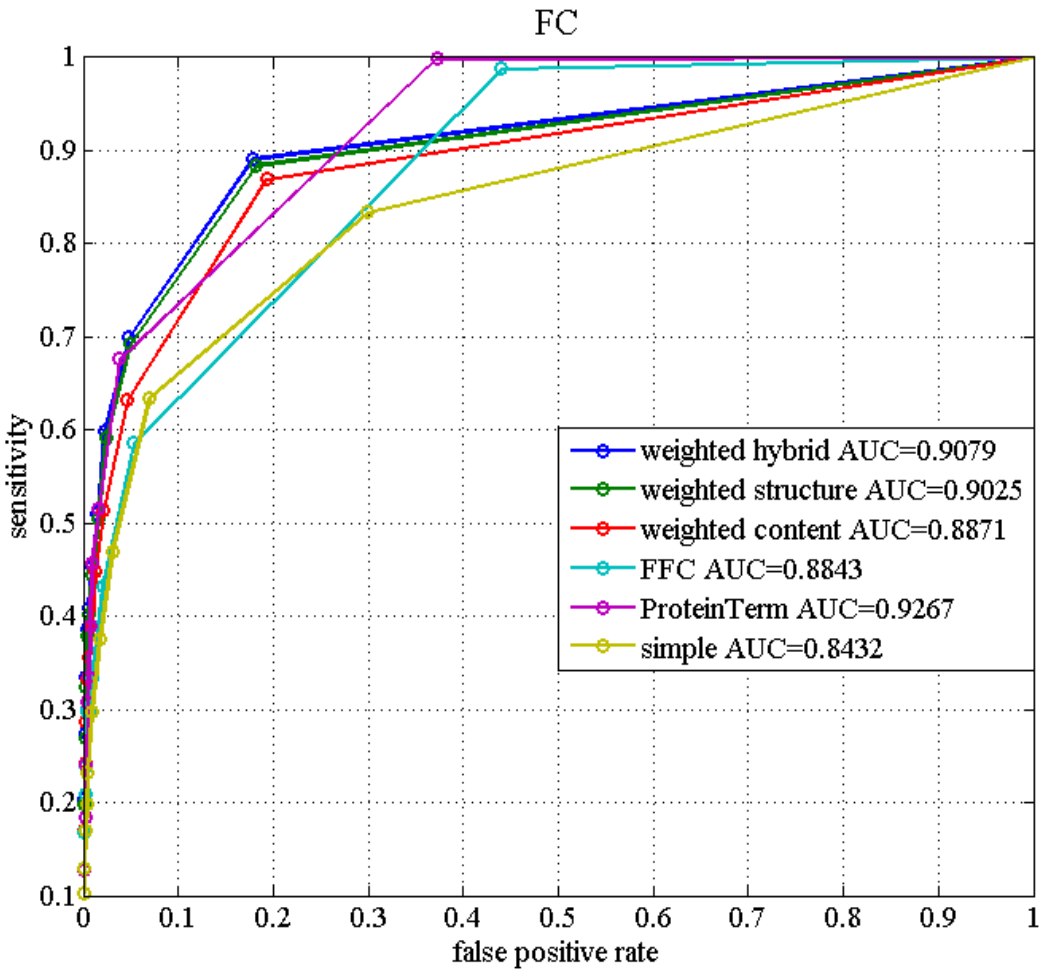
Табела 18. Резултати од функционална анотација со користење на кластерирање според средишност на врски



Слика 6.5 ROC криви за функционална анотација со кластерирање според средишност на врски со секоја граф репрезентација

Функционална анотација со FC												
тип граф	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
Simple	Sens.	0.8343	0.6346	0.4694	0.3760	0.2972	0.2325	0.1986	0.1692	0.1287	0.1023	0.8432
	fpr	0.2995	0.0694	0.0304	0.0170	0.0092	0.0050	0.0035	0.0025	0.0014	0.0010	
Weighted Content	Sens.	0.8693	0.6328	0.5136	0.4477	0.3896	0.3561	0.3326	0.2862	0.2420	0.1691	0.8871
	fpr	0.1929	0.0466	0.0212	0.0129	0.0086	0.0058	0.0043	0.0028	0.0020	0.0012	
Weighted Structure	Sens.	0.8847	0.6934	0.5911	0.5044	0.4451	0.4023	0.3796	0.3246	0.2691	0.1986	0.9025
	fpr	0.1825	0.0501	0.0237	0.0155	0.0084	0.0055	0.0040	0.0026	0.0019	0.0010	
Weighted Hybrid	Sens.	0.8919	0.7004	0.5980	0.5106	0.4526	0.4093	0.3859	0.3343	0.2751	0.2040	0.9079
	fpr	0.1779	0.0481	0.0222	0.0147	0.0087	0.0056	0.0041	0.0027	0.0018	0.0010	
Protein-Term	Sens.	0.9997	0.6763	0.5156	0.4566	0.3894	0.3346	0.3079	0.2402	0.1833	0.1274	0.9267
	fpr	0.3723	0.0381	0.0166	0.0090	0.0069	0.0052	0.0046	0.0030	0.0022	0.0011	
FFC	Sens.	0.9877	0.5858	0.4315	0.3330	0.2982	0.3017	0.2975	0.2390	0.2093	0.1676	0.8843
	fpr	0.4398	0.0532	0.0205	0.0090	0.0058	0.0046	0.0038	0.0020	0.0018	0.0011	

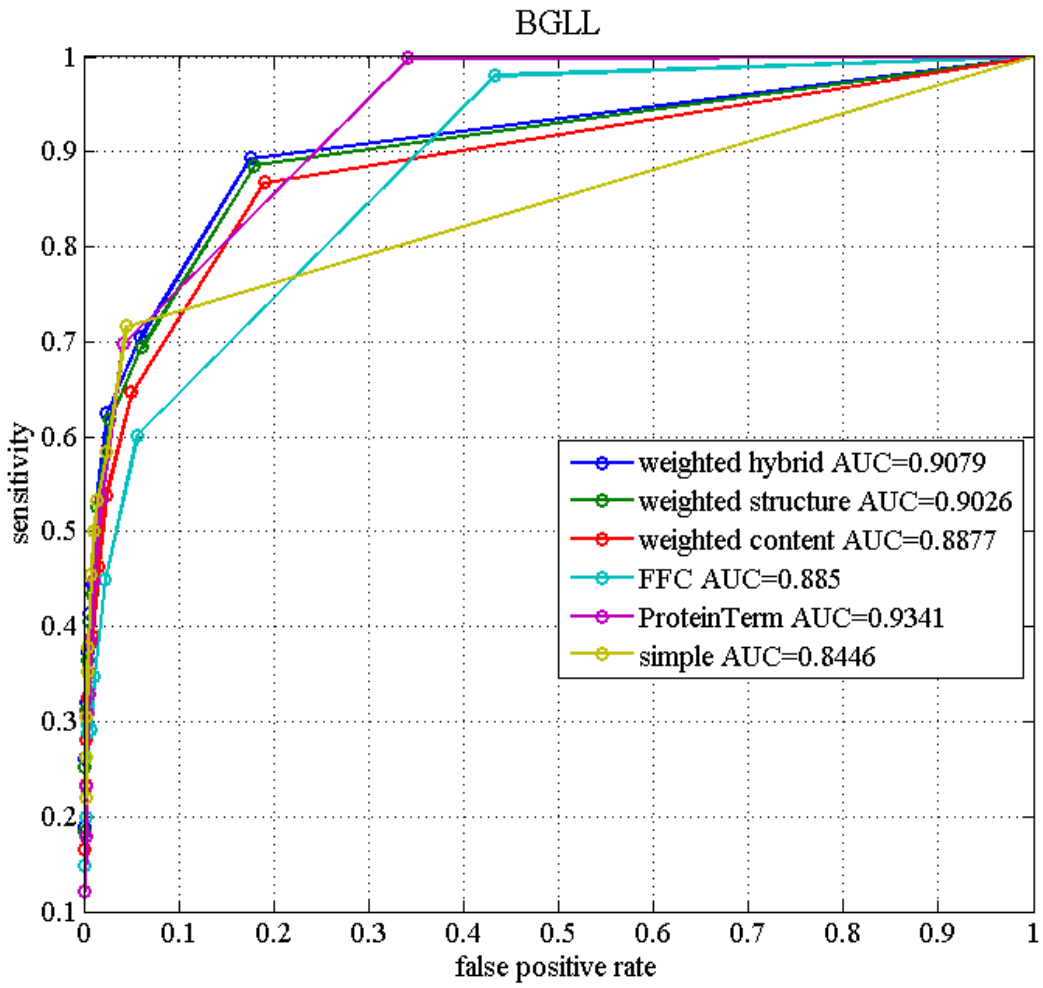
Табела 19. Резултати од функционална анотација со користење на FC алгоритамот



Слика 6.6 ROC криви за функционална анотација со кластерирање со FC алгоритамот за секоја граф репрезентација

Функционална анотација со BGLL												
тип граф	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
Simple	Sens.	0.7166	0.5843	0.5338	0.4999	0.4542	0.3776	0.3523	0.3049	0.2619	0.2199	0.8446
	fpr	0.0447	0.0238	0.0142	0.0106	0.0075	0.0041	0.0034	0.0027	0.0021	0.0019	
Weighted Content	Sens.	0.8681	0.6462	0.5388	0.4633	0.3894	0.351	0.3246	0.2808	0.2341	0.1653	0.8877
	fpr	0.1899	0.0499	0.0233	0.0149	0.0096	0.0061	0.0044	0.0028	0.002	0.0013	
Weighted Structure	Sens.	0.886	0.6945	0.6176	0.5264	0.4345	0.4059	0.3654	0.3122	0.253	0.1842	0.9026
	fpr	0.1794	0.061	0.0271	0.0139	0.0094	0.0058	0.004	0.0023	0.0015	0.001	
Weighted Hybrid	Sens.	0.8929	0.7041	0.6256	0.5332	0.4442	0.4145	0.3737	0.3201	0.2607	0.1901	0.9079
	fpr	0.1753	0.0593	0.0244	0.0138	0.0096	0.0058	0.0039	0.0023	0.0015	0.001	
Protein-Term	Sens.	0.9992	0.6978	0.5343	0.4478	0.3837	0.3292	0.3079	0.232	0.1784	0.1204	0.9341
	fpr	0.3416	0.041	0.0187	0.0104	0.0073	0.0057	0.0046	0.003	0.0021	0.0011	
FFC	Sens.	0.9807	0.6018	0.4495	0.3479	0.2917	0.2963	0.2884	0.2324	0.2002	0.148	0.8850
	fpr	0.4328	0.0561	0.0227	0.0109	0.0067	0.0049	0.0038	0.002	0.0018	0.0011	

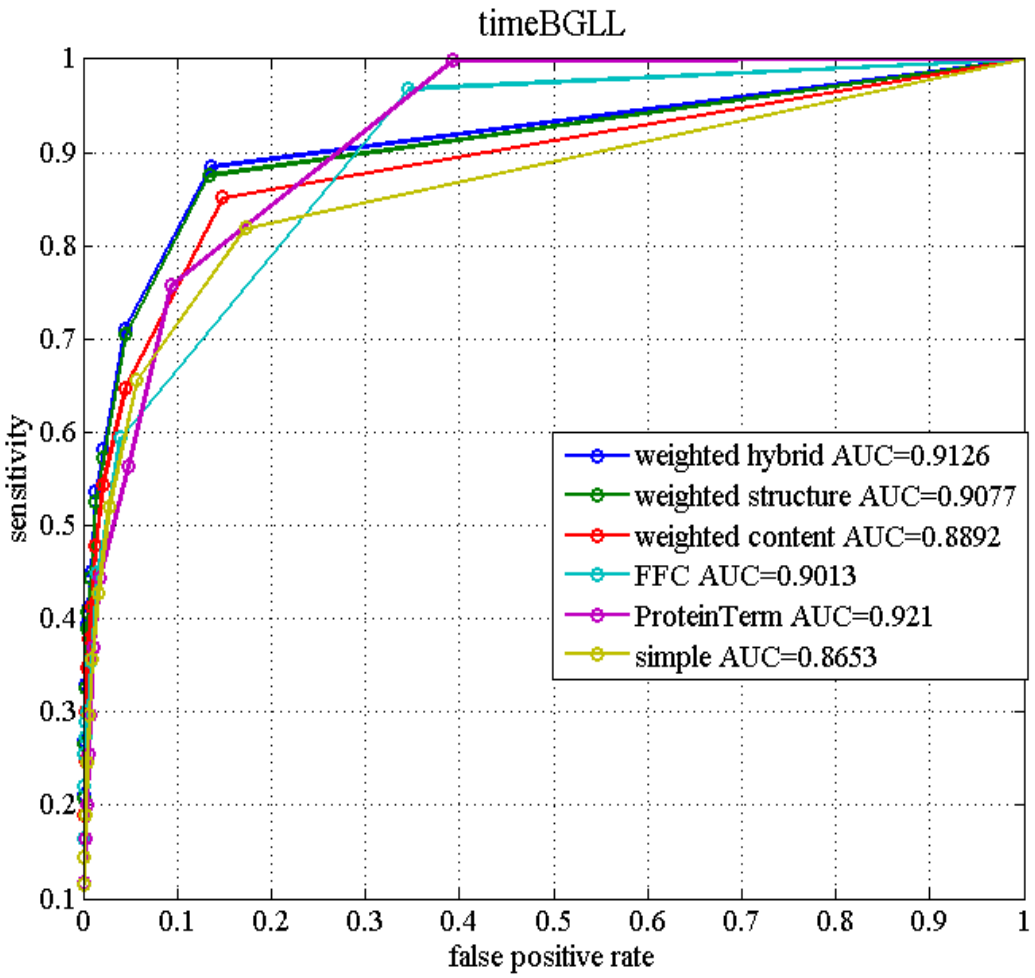
Табела 20. Резултати од функционална анотација со користење на BGLL алгоритамот



Слика 6.7 ROC криви за функционална анотација со кластерирање со алгоритамот BGLL со секоја граф репрезентација

Функционална анотација со timeBGLL												
тип граф	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
Simple	Sens.	0.8185	0.6562	0.5205	0.428	0.3565	0.2972	0.2453	0.1904	0.1446	0.1153	0.8653
	fpr	0.1724	0.0570	0.0275	0.0156	0.0094	0.0061	0.0041	0.0026	0.0016	0.0011	
Weighted Content	Sens.	0.8520	0.6464	0.5432	0.4777	0.4123	0.3792	0.3471	0.3000	0.2487	0.1894	0.8892
	fpr	0.1482	0.0440	0.0208	0.0135	0.0085	0.0056	0.0042	0.0028	0.0018	0.0012	
Weighted Structure	Sens.	0.8762	0.7053	0.5728	0.5263	0.4443	0.4077	0.3898	0.3252	0.2661	0.2088	0.9077
	fpr	0.1337	0.0451	0.0214	0.0128	0.0078	0.0048	0.0039	0.0023	0.0014	0.0011	
Weighted Hybrid	Sens.	0.8853	0.7113	0.5814	0.537	0.4516	0.4129	0.3942	0.3289	0.2696	0.2125	0.9126
	fpr	0.1356	0.0443	0.0216	0.0132	0.0085	0.0051	0.0042	0.0024	0.0015	0.0011	
Protein-Term	Sens.	0.9995	0.7582	0.5635	0.4433	0.3690	0.2971	0.2544	0.2012	0.1639	0.1180	0.9210
	fpr	0.3927	0.0934	0.0474	0.0171	0.0112	0.0071	0.0054	0.0033	0.0018	0.0011	
FFC	Sens.	0.9687	0.5950	0.4501	0.3548	0.3011	0.289	0.2734	0.2554	0.2198	0.1636	0.9013
	fpr	0.3459	0.0395	0.0151	0.0077	0.005	0.0031	0.0021	0.0014	0.0010	0.0007	

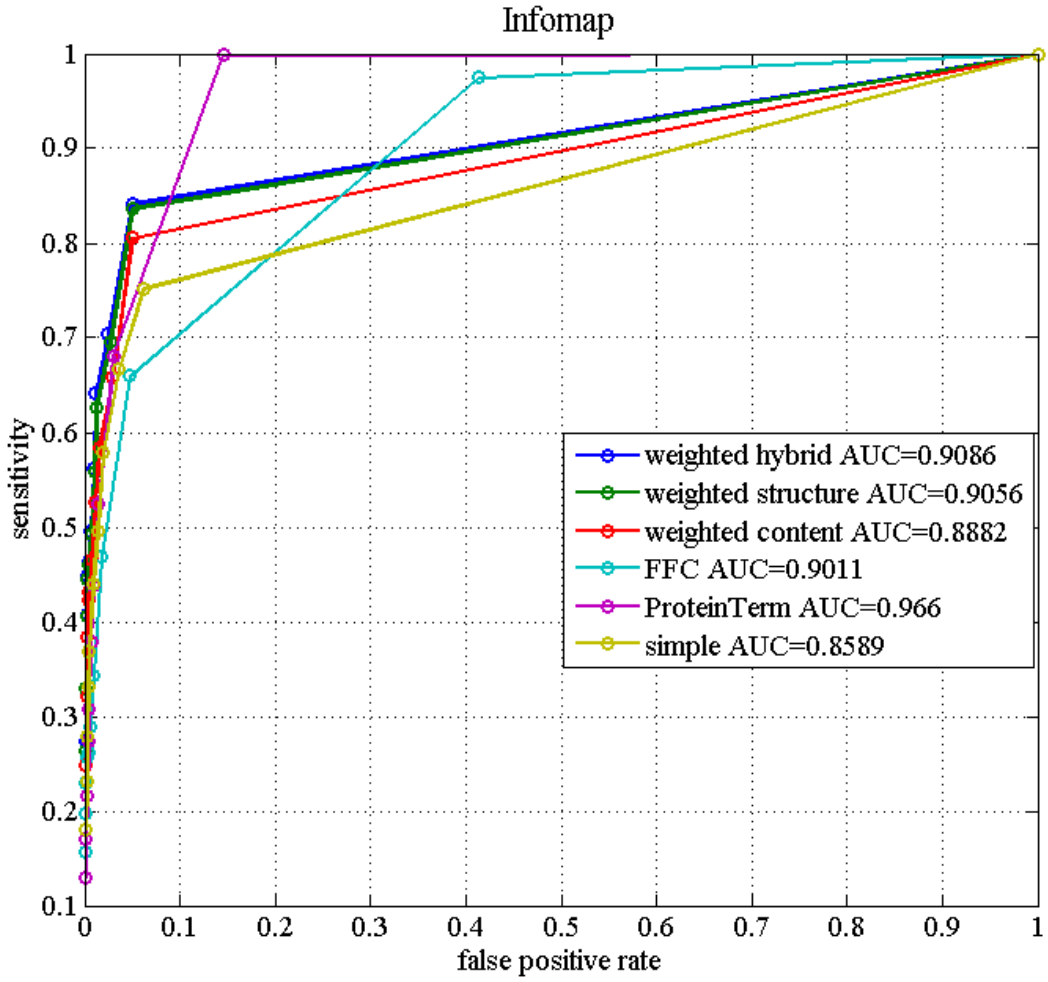
Табела 21. Резултати од функционална анотација со користење на timeBGLL алгоритамот



Слика 6.8 ROC криви за функционална анотација со кластерирање со timeBGLL алгоритамот за секоја граф репрезентација

Функционална анотација со Infomap												
тип граф	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
Simple	Sens.	0.7523	0.6676	0.5789	0.4952	0.4397	0.3691	0.3307	0.2796	0.2306	0.1810	0.8589
	fpr	0.0614	0.0358	0.0191	0.0133	0.0083	0.0048	0.0039	0.0028	0.0019	0.0014	
Weighted Content	Sens.	0.8053	0.6588	0.5834	0.5257	0.4645	0.4311	0.4223	0.3845	0.3213	0.2489	0.8882
	fpr	0.0502	0.0295	0.0150	0.0110	0.0069	0.0042	0.0034	0.0024	0.0017	0.0013	
Weighted Structure	Sens.	0.8368	0.6959	0.6268	0.5592	0.4932	0.4610	0.4456	0.4065	0.3291	0.2630	0.9056
	fpr	0.0499	0.0265	0.0117	0.0105	0.0068	0.0038	0.0031	0.0023	0.0012	0.0011	
Weighted Hybrid	Sens.	0.8417	0.7034	0.6414	0.5623	0.4961	0.4630	0.4481	0.4081	0.3303	0.2731	0.9086
	fpr	0.0499	0.0247	0.0115	0.0097	0.0065	0.0038	0.0031	0.0021	0.0012	0.0010	
Protein-Term	Sens.	0.9999	0.6804	0.5249	0.4375	0.3780	0.3076	0.2733	0.2154	0.1709	0.1295	0.9660
	fpr	0.1456	0.0303	0.0143	0.0094	0.0066	0.0045	0.0035	0.0022	0.0016	0.0011	
FFC	Sens.	0.9750	0.6598	0.4688	0.3424	0.2894	0.2626	0.2572	0.2298	0.1971	0.1562	0.9011
	fpr	0.4131	0.0472	0.0178	0.0093	0.0060	0.0035	0.0025	0.0016	0.0012	0.0010	

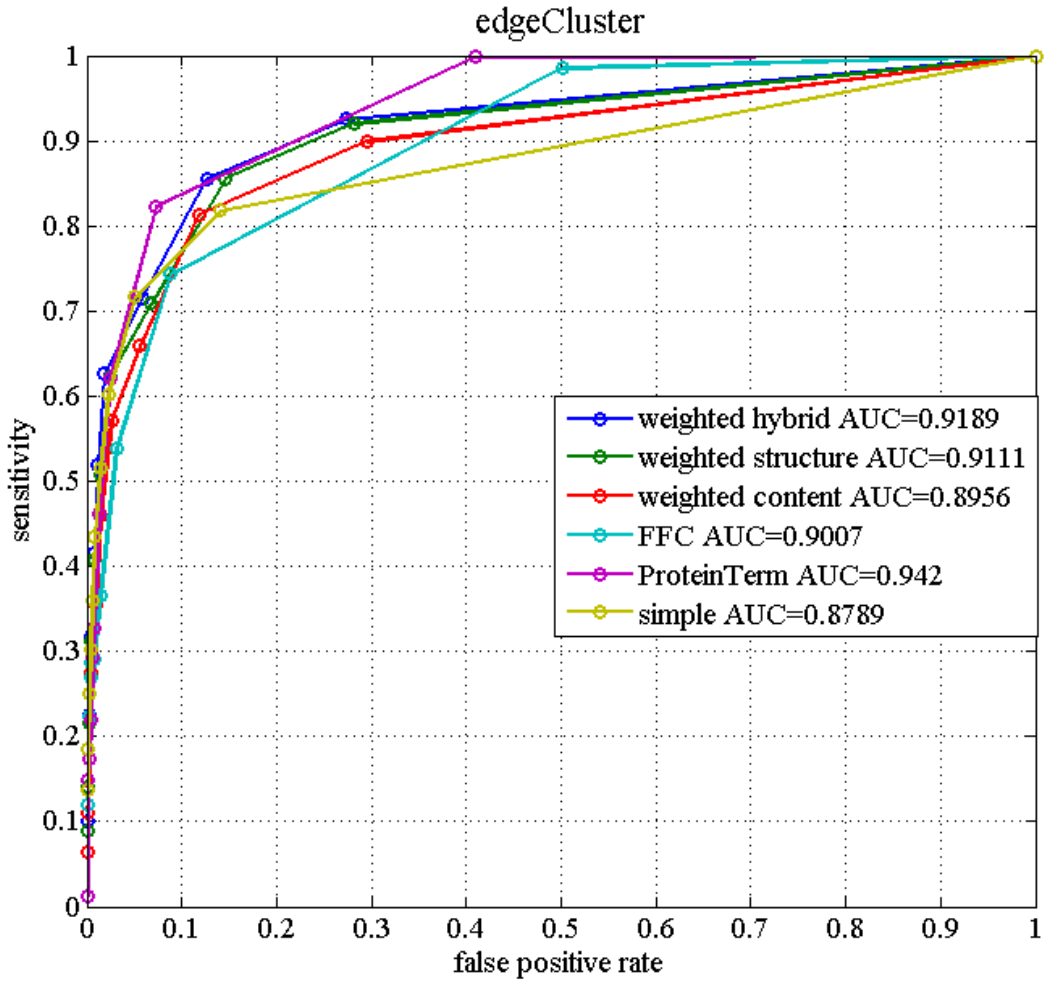
Табела 22. Резултати од функционална анотација со користење на Infomap алгоритамот



Слика 6.9 ROC криви за функционална анотација со кластерирање со Infomap алгоритамот за секоја граф репрезентација

Функционална анотација со кластерирање на врски												
тип граф	**ω =**	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	**AUC**
Simple	**Sens.**	0.8184	0.7163	0.6021	0.5154	0.4340	0.3604	0.3028	0.2513	0.1854	0.1374	**0.8789**
	fpr	0.1403	0.0503	0.0228	0.0137	0.0085	0.0053	0.0036	0.0025	0.0016	0.0011	
Weighted Content	**Sens.**	0.8999	0.8134	0.6589	0.5701	0.4594	0.3537	0.2763	0.1735	0.1096	0.0639	**0.8956**
	fpr	0.2950	0.1180	0.0555	0.0262	0.0153	0.0086	0.0048	0.0029	0.0014	0.0007	
Weighted Structure	**Sens.**	0.9207	0.8559	0.7092	0.6225	0.5082	0.4078	0.3118	0.2166	0.1402	0.0900	**0.9111**
	fpr	0.2818	0.1457	0.0675	0.0240	0.0153	0.0079	0.0045	0.0026	0.0013	0.0007	
Weighted Hybrid	**Sens.**	0.9261	0.8560	0.7150	0.6272	0.5192	0.4165	0.3180	0.2259	0.1490	0.1002	**0.9189**
	fpr	0.2731	0.1264	0.0581	0.0179	0.0112	0.0078	0.0043	0.0025	0.0012	0.0007	
Protein-Term	**Sens.**	0.9997	0.8233	0.6203	0.4623	0.3272	0.2918	0.2191	0.1744	0.1491	0.0118	**0.9420**
	fpr	0.4104	0.0732	0.0242	0.0128	0.0072	0.0056	0.0045	0.0028	0.0014	0.0009	
FFC	**Sens.**	0.9859	0.7429	0.5387	0.3661	0.2910	0.2879	0.2695	0.2239	0.1846	0.1201	**0.9007**
	fpr	0.5009	0.0881	0.0310	0.0137	0.0072	0.0049	0.0037	0.0018	0.0010	0.0009	

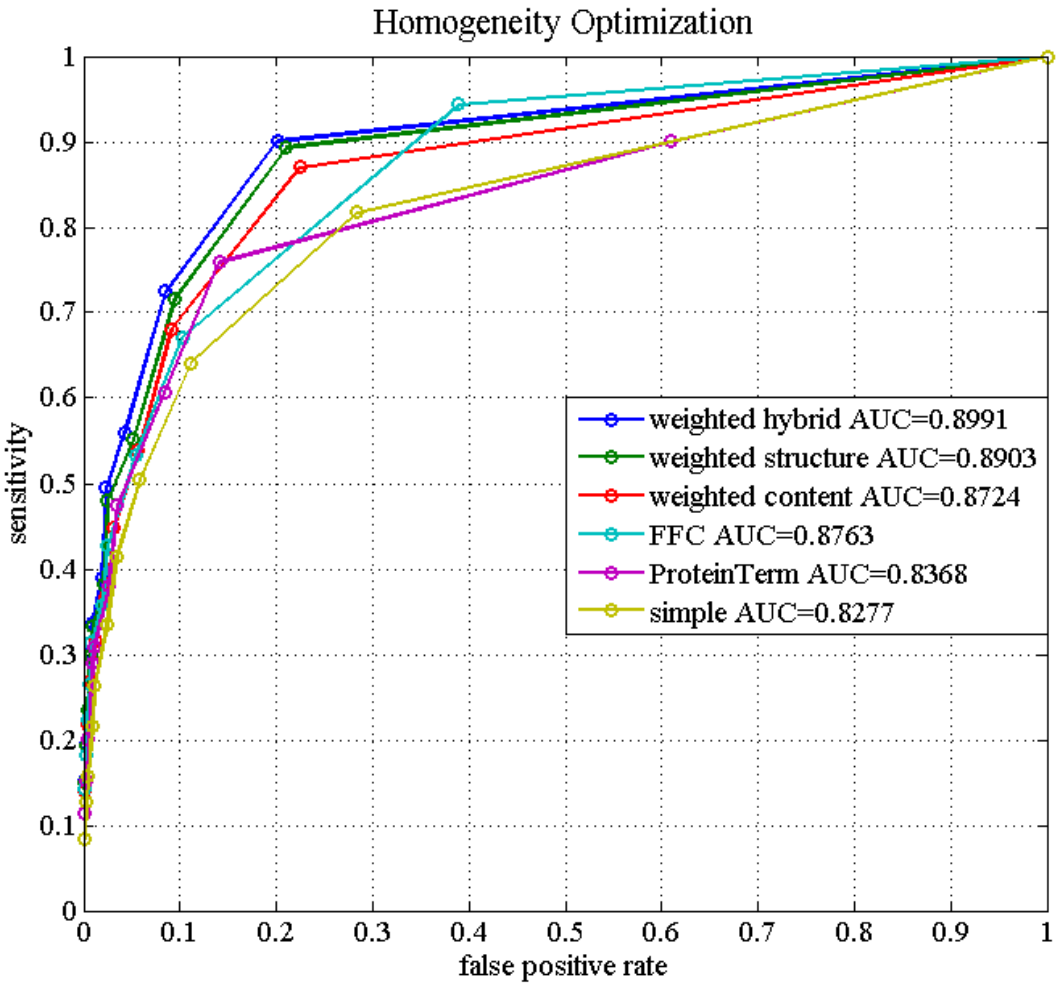
Табела 23. Резултати од функционална анотација со користење на алгоритамот за кластерирање на врски



Слика 6.10 ROC криви за функционална анотација со кластерирање на врски за секоја граф репрезентација

Функционална анотација со оптимизација на хомогеноста												
тип граф	ω =	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	AUC
Simple	Sens.	0.8174	0.6411	0.5046	0.4138	0.3344	0.2637	0.2165	0.1574	0.1281	0.0849	0.8277
	fpr	0.2831	0.1112	0.0583	0.0340	0.0237	0.0117	0.0086	0.0044	0.0030	0.0016	
Weighted Content	Sens.	0.8707	0.6810	0.5391	0.4479	0.3663	0.3153	0.2693	0.2184	0.1821	0.1401	0.8724
	fpr	0.2241	0.0905	0.0563	0.0315	0.0229	0.013	0.0076	0.0040	0.0025	0.0015	
Weighted Structure	Sens.	0.8942	0.7162	0.5517	0.4807	0.3812	0.3327	0.2986	0.2340	0.1926	0.1498	0.8903
	fpr	0.2103	0.0942	0.0513	0.0246	0.0207	0.0096	0.0069	0.0035	0.0023	0.0015	
Weighted Hybrid	Sens.	0.9013	0.7250	0.5585	0.4948	0.3891	0.3375	0.3008	0.2358	0.1944	0.1517	0.8991
	fpr	0.2009	0.0846	0.0419	0.0229	0.0192	0.0088	0.0071	0.0036	0.0023	0.0015	
Protein-Term	Sens.	0.9002	0.7602	0.6065	0.4743	0.3810	0.3088	0.2889	0.2017	0.1522	0.1146	0.8368
	fpr	0.6094	0.1412	0.0839	0.0344	0.0259	0.0103	0.0085	0.0048	0.0029	0.0016	
FFC	Sens.	0.9443	0.6714	0.5336	0.4280	0.3584	0.3149	0.2657	0.2236	0.1822	0.1419	0.8763
	fpr	0.3891	0.1026	0.0537	0.0249	0.0186	0.0079	0.0066	0.0034	0.0025	0.0014	

Табела 24. Резултати од функционална анотација со користење на оптимизација на хомогеноста



Слика 6.11 ROC криви за функционална анотација со кластерирање со оптимизација на хомогеноста за секоја граф репрезентација

Од прикажаните резултати можеме да го видиме она што претходно беше кажано за влијанието на вредностите на ентропијата. Како што се очекуваше покомплексните репрезентации (G_3 или протеин-термин графот и G_4 или комплетно функционално поврзаниот граф - FFC) имаат повисоки вредности за ентропијата што имплицитно ја зголемува сензитивноста и стапката на грешка (преку зголемување на FP вредностите и намалување на FN вредностите). Спротивното важи за поедноставните репрезентации (G_1 или едноставниот нетежински граф и G_2 или тежинскиот граф).

Во табела 25 се прикажани сумарните вредности за добиените AUC за секој алгоритам за кластерирање во комбинација со секоја граф репрезентација за протеинската интеракциска мрежа. Како што се гледа од табелата највисока просечна вредност за AUC има алгоритамот за кластерирање на врски кој продуцира преклопувачки кластери. Овој резултат е во согласност со добро познатиот факт дека протеинските интеракциски мрежи имаат многу мултифункционални протеини коишто извршуваат повеќе функции и се очекува истите да влегуваат во специфични интеракции со различни множества од партнери, симултано или не, во зависност од функцијата која се извршува.

Вредности на AUC за функционална анотација со секој алгоритам и секоја репрезентација							
Репрезентација / Алгоритам	едноставен граф	тежински граф (содржина)	тежински граф (структура)	тежински граф (хибриден)	протеин-термин граф	комплетно функционално поврзан граф	AVG
Aglomerative	0.7523	0.8053	0.8368	0.8417	0.9999	0.9750	**0.8439**
K-Medoids	0.8185	0.8520	0.8762	0.8853	0.9995	0.9687	**0.8192**
Spectral	0.8184	0.8999	0.9207	0.9261	0.9997	0.9859	**0.8613**
EdgeBetweenness	0.7166	0.8681	0.886	0.8929	0.9992	0.9807	**0.8628**
FC	0.8343	0.8693	0.8847	0.8919	0.9997	0.9877	**0.8919**
BGLL	0.8166	0.8896	0.9124	0.9176	0.9012	0.9203	**0.8936**
TimeBGLL	0.8221	0.8845	0.9073	0.9125	0.9225	0.8739	**0.8995**
Infomap	0.7578	0.8031	0.8259	0.8311	0.7832	0.9937	**0.9047**
EdgeCluster	0.7490	0.7932	0.8160	0.8212	0.7733	0.9338	**0.9078**
HO	0.8174	0.8707	0.8942	0.9013	0.9002	0.9443	**0.8671**
AVG	**0.8358**	**0.8752**	**0.8908**	**0.8978**	**0.8846**	**0.8669**	

Табела 25. Сумарна табела за AUC вредностите од функционалната анотација со комбинацијата на секој алгоритам (заедно со просекот по алгоритам) со секоја граф репрезентација (заедно со просекот по репрезентација)

Од истата табела може да се види дека најлоши резултати се добиваат од алгоритмите кои прават круто кластерирање (агломеративното и *k*-медоиди кластерирањето), иако и тие резултати не се драстично полоши од пософистицираните алгоритми. Алгоритамот кој врши оптимизација на хомогеноста дава резултати кои се споредливи и блиски до најдобрите алгоритми, што со оглед на неговата едноставност е значителен резултат. Од аспект на различните репрезентации кои се користат за претставување на протеинската интеракциска мрежа се гледа дека секоја од ново предложените репрезентации даваат подобри резултати од едноставната нетежинска граф претстава, со тоа што тежинскиот граф во кој тежините се задаваат со хибридна стратегија има највисока просечна вредност за AUC. Подобрувањето на перформансите при користење на комплетно функционално поврзаниот граф (FFC) укажува на тоа дека во рамки на протеинската интеракциска мрежа всушност недостасуваат дел од вистинските интеракции што се случуваат помеѓу парови од протеини.

Вредности на sensitivity при ω=0 за функционална анотација со секој алгоритам и секоја репрезентација							
Репрезентација / Алгоритам	едноставен граф	тежински граф (содржина)	тежински граф (структура)	тежински граф (хибриден)	протеин-термин граф	комплетно функционално поврзан граф	AVG
Aglomerative	0.8166	0.8896	0.9124	0.9176	0.9012	0.9203	**0.8929**
K-Medoids	0.8221	0.8845	0.9073	0.9125	0.9225	0.8739	**0.8871**
Spectral	0.749	0.7932	0.816	0.8212	0.7733	0.9338	**0.8144**
EdgeBetweenness	0.7578	0.8031	0.8259	0.8311	0.7832	0.9937	**0.8324**
FC	0.8343	0.8693	0.8847	0.8919	0.9997	0.9877	**0.9112**
BGLL	0.7166	0.8681	0.886	0.8929	0.9992	0.9807	**0.8905**
TimeBGLL	0.8185	0.852	0.8762	0.8853	0.9995	0.9687	**0.9000**
Infomap	0.7523	0.8053	0.8368	0.8417	0.9999	0.975	**0.8685**
EdgeCluster	0.8184	0.8999	0.9207	0.9261	0.9997	0.9859	**0.9251**
HO	0.8174	0.8707	0.8942	0.9013	0.9002	0.9443	**0.8880**
AVG	**0.7903**	**0.8535**	**0.8760**	**0.8821**	**0.9278**	**0.9564**	

Табела 26. Сумарна табела за вредностите на сензитивноста при праг ω=0 од функционалната анотација со комбинацијата на секој алгоритам (заедно со просекот по алгоритам) со секоја граф репрезентација (заедно со просекот по репрезентација)

Ако ја погледнеме табела 26 во која се прикажани вредностите на сензитивноста при праг ω=0 (што би значело во случајот кога процесот на доделување на функции на прашалниот протеин е најмалку ограничен) она што можеме да го

видиме во резултатите е всушност колку добиените кластери се богати со функционални термини врз основа на кои се врши функционалната анотација или според она што претходно го дефиниравме кај ентропијата на кластерирањето, колку добиените кластери се опфатни. Од аспект на различните алгоритми повторно најдобар е алгоритамот кој кластерира врски (преклопувачки кластери) меѓутоа она што е различно во однос на претходната табела е однесувањето на останатите алгоритми. Имено наједноставните алгоритми кој прават круто кластерирање (агломеративното и *k*-медоиди) имаат доста високи вредности за сензитивноста што укажува дека истите продуцираат кластери со поголема опфатност за разлика од на пример алгоритамот кој е второ рангиран според просечната AUC, односно Infomap, кој има доста мали вредности за сензитивноста. Од аспект на различните репрезентации разликата се забележува кај комплексните репрезентации протеин-термин графот и комплетно функционално поврзаниот граф кои продуцираат кластери со поголема опфатност од обичните тежински графови, каде повторно хибридно определените тежини даваат подобри резултати. Меѓутоа ваквите резултати не можат да се гледаат изолирано без да се земат во предвид стапките на грешка (табела 27) кои се продуцираат, и кои можеме да ги гледаме како еден вид на цена која се плаќа за да се постигне определена сензитивност.

Вредности на fpr при ω=0 за функционална анотација со секој алгоритам и секоја репрезентација							
Репрезентација / Алгоритам	едноставен граф	тежински граф (содржина)	тежински граф (структура)	тежински граф (хибриден)	протеин-термин граф	комплетно функционално поврзан граф	AVG
Aglomerative	0.3736	0.2799	0.2667	0.246	0.8259	0.4326	**0.4041**
K-Medoids	0.5514	0.3514	0.3382	0.3175	0.4591	0.2234	**0.3735**
Spectral	0.1108	0.1083	0.1031	0.1018	0.131	0.5524	**0.1845**
EdgeBetweenness	0.1093	0.0987	0.0935	0.0922	0.1214	0.8428	**0.2263**
FC	0.2995	0.1929	0.1825	0.1779	0.3723	0.4398	**0.2774**
BGLL	0.0447	0.1899	0.1794	0.1753	0.3416	0.4328	**0.2272**
TimeBGLL	0.1724	0.1482	0.1337	0.1356	0.3927	0.3459	**0.2214**
Infomap	0.0614	0.0502	0.0499	0.0499	0.1456	0.4131	**0.1283**
EdgeCluster	0.1403	0.295	0.2818	0.2731	0.4104	0.5009	**0.3169**
HO	0.2831	0.2241	0.2103	0.2009	0.6094	0.3891	**0.3194**
AVG	**0.2146**	**0.1938**	**0.1839**	**0.1770**	**0.3809**	**0.4572**	

Табела 27. Сумарна табела за вредностите на стапката на грешка при праг ω=0 од функционалната анотација со комбинацијата на секој алгоритам (заедно со просекот по алгоритам) со секоја граф репрезентација (заедно со просекот по репрезентација)

Она што може да се види од табела 27 е дека вредностите на стапката на грешка се однесуваат обратнопропорционално на сензитивноста, што е за очекување бидејќи кластерирањата кои продуцираат кластери со помала опфатност се построги во носењето на одлуките. Кажано со други зборови се добиваат кластери кои се посиромашни од аспект на диверзитет на функционални термини што води кон многу прецизна, но некомплетна функционална анотација. За разлика од ваквите кластерирања оние кои имаат повисока сензитивност можат поцелосно да го анотираат прашалниот протеин бидејќи множеството на потенцијални анотации е побогато, но и со поголем шум што води кон повисоки вредности на стапките на грешка. Во контекст на ова „најпрецизен“ или најстрог е алгоритамот Infomap, а најмалку прецизни се агломеративното и *k*-медоиди кластерирањето. Алгоритамот со оптимизација на хомогеноста дава полоши резултати од најдобрите што се должи пред сè на лошото справување со репрезентација со протеин-термин граф и ова е дел на кој треба во иднина да се работи за да може да се подобри. Од аспект на различните репрезентации најпрецизна анотација се добива со користење на тежинските графови, додека протеин-термин графот и комплетно функционално поврзаниот граф го имаат проблемот на анотациско множество со висок шум (дури и за Infomap алгоритамот кој вообичаено има многу низок шум). Ова укажува дека при користењето на овие репрезентации потребно е да се обмисли поинаква стратегија за доделување на функционални термини на прашалниот термин која ќе биде поригорозна од аспект на тоа кој термин од потенцијалните може да му се додели на протеинот, а кој не.

Од аспект на комплексноста јасно е дека протеин-термин графот (G_3) и комплетно функционално поврзаниот граф (G_4) се далеку покомплексни и ваквата пресметковна комплексност треба да биде земена во предвид кога се одлучува која репрезентација на протеинската интеракциска мрежа ќе биде избрана. Функционалната анотација на ниво на цела мрежа би била многу непрактична доколку се користат G_2, G_3 или G_4 репрезентациите бидејќи алгоритамот за кластерирање треба да се извршува за секој можен прашален протеин. Од друга страна во сценарио во кое треба да се определи пошироко множество на можни анотации (функционални термини) за еден (или неколку) протеин(и) би имало

многу голем бенефит од ваквите збогатени граф репрезентации на протеинската интеракциска мрежа.

Како последна забелешка треба да се укаже на уште еден потенцијален проблем во процесот на функционална анотација, без разлика дали се користи директна метода или метода базирана на кластерирање. Имено станува збора за комплетноста на податоците. Направена е процена дека комплетната *S. cerevisiae* протеинска интеракциска мрежа има помеѓу 37800 и 75500 протеински интеракции [192]. Моментално постојат помеѓу 55000 и 60000 интеракции содржани во јавно достапни бази на податоци за *S. cerevisiae*, што значи дека постојат потенцијално непознати региони од протеинската интеракциска мрежа што може да објаснување за високите стапки на грешка и ниските сензитивности за кои зборувавме претходно.

7

ЗАКЛУЧОК И ИДНА РАБОТА

Во рамки на оваа докторска дисертација беа предложени различни пристапи за решавање на проблемот на функционална анотација и тоа директен пристап и неколку различни пристапи базирани на кластерирање. Дополнително беа предложени и различни репрезентации за протеинската интеракциска мрежа. Бидејќи ваквите мрежи содржат информација не само за интеракциите, туку и за постоечките анотации на протеините, различните репрезентации кои ги предлагаме го збогатуваат процесот на предвидување на функции на непознат протеин во мрежата преку вклучување на овие информации во самите алгоритми. Беа испитани различни метрики за искористување на врските помеѓу различните функционални термини во мрежата и заклучивме дека дури и наједноставната корелациска метрика за пар од протеини го подобрува целокупниот процес, додека користењето на семантички метрики базирани на онтолошка структура дава најдобри резултати. Ваквите метрики беа комбинирани на различни начини во рамки на граф репрезентациите на протеинската интеракциска мрежа за да се добијат тежини на врските во рамки на графот и се покажа дека најдобри

перформанси се добиваат кога таквата комбинација е хибридна односно кога метриката се гледа не само од аспект на содржината на два соседни протеини туку и од аспект на нивниот контекст во мрежата. Истиот пристап за определување на тежини беше искористен при дополнителното збогатување на графот преку додавање на вештачки врски со цел да се долови добро познатиот факт дека во моментот протеинските интеракциски мрежи сеуште не се целосно откриени со помош на експерименталните методи за нивно конструирање. Ваквата репрезентација е многу комплексна и пресметковно скапа, но потенцијалот за откривање на ново знаење е значително зголемен. Нашите резултати покажаа дека подеднакво информативна, но и подеднакво комплексна е и репрезентацијата каде генерираме граф во кој секој функционален термин придружен на некој протеин станува јазел и асоцијацијата помеѓу протеините и термините се претставува со додавање на врска помеѓу секој пар. Во иднина можни правци на истражување би биле наоѓање на соодветен начин за комбинирање на корелациската и семантичката метрика во едно бидејќи резултатите укажуваат дека и двете имаат значително влијание врз перформансите.

Новиот метод за директна функционална анотација е базиран на општа хипотеза, поради фактот што една клика (clique) од протеини со слични функции може да се третира еквивалентно како множество од јазли што гледаат исто функционално соседство. Предложениот метод беше детално анализиран како од аспект на неговите параметри така и од аспект на неговото однесување во комбинацијата со различните метрики и стратегии за доделување на тежини. Добиените резултати се многу добри од аспект на сензитивноста на алгоритамот и прифатливи од аспект на стапката на грешка, при што збогатените репрезентации значајно ги подобруваат целокупните перформанси. Овој алгоритам беше спореден со други два алгоритми кои му се слични по природа при што се покажа значително подобар. Резултатите од директниот метод се на ниво на најдобрите резултати од методите базирани на кластерирање. Идни правци на истражување и подобрување на овој алгоритам се автоматското определување на оптималното функционално соседство, како и имплементација на надгледувана варијанта на алгоритамот.

Комплексните протеински интеракциски мрежи откриваат определени карактеристики на граф кои можат да се анализираат од аспект на функционални модули придружени кон биолошката функција која ја вршат. Во овој дел од истражувањата ја истражувавме моќта на алгоритмите за кластерирање со цел откривање на овие структури. Беа имплементирани и адаптирани моментално најдобрите (според литературата) алгоритми за кластерирање на комплексни мрежи и беа прилагодени кон нашиот систем со цел да ја видиме нивната способност за функционална анотација. Исто така беа развиени и нови алгоритми за кластерирање во протеинските интеракциски мрежи, при што алгоритамот за оптимизација на хомогеноста по своите перформанси е споредлив со најдобрите. Резултатите од експериментите го валидираат пристапот со користење на збогатени репрезентации. Дури и за наједноставното проширување т.е. различните тежински графови за протеинската интеракциска мрежа се добиваат значајни подобрувања на резултатите за функционалната анотација. Во иднина овде има огромен потенцијал за подобрување на алгоритамот за оптимизација на хомогеноста при што може да се дефинираат пософистицирани мерки за пресметка на сличноста/различноста на различните кластери, како и евентуалната можност за дефинирање на нулти модел за мрежата при што во предвид ќе бидат земени и анотациите кои им се придружени на секој од протеините.

Генерално доколку е потребно да се изврши анотација на мрежно ниво, се препорачува користењето на репрезентацијата со тежински граф, додека при проучувањето на еден протеин, или мала група на протеини, треба да се изврши или со комплетно функционално поврзаните графови или со протеин-термин графовите. Од аспект на избирање на соодветен алгоритам за функционална анотација треба да се има во предвид што е приоритетно во процесот. Доколку приоритетот е добивање на мало множество на точни функции за прашалниот протеин треба да се искористи некој од алгоритмите кои имаат ниска стапка на грешка (Infomap, спектрално кластерирање), додека за случајот кога откривањето на сите можни функции е од поголемо значење треба да се искористи некој од алгоритмите кои имаат висока сензитивност (EdgeCluster, timeBGLL, HO). Директниот метод би бил добар и од обата аспекти со тоа што за истиот претходно треба да се пронајдат оптималните параметри.

8

РЕФЕРЕНЦИ

8.1 Листа на објавени трудови во областа во кои кандидатот е (ко)автор

1. **K. Trivodaliev**, I. Cingovska, S. Kalajdziski, D. Davcev, "Protein Function Prediction Based on Neighborhood Profiles**",** *in proceedings of The ICT Innovations* 2009, *Springer*, Ohrid, Macedonia, September 28 – 30, 2009, p.125
2. S. Kalajdziski, B. Pepik, I. Ivanoska, G. Mirceva, **K. Trivodaliev**, D. Davcev, "Automated Structural Classification of Proteins by Using Decision Trees and Structural Protein Features**",** *in proceedings of The ICT Innovations* 2009, *Springer*, Ohrid, Macedonia, September 28 – 30, 2009, p.135
3. I. Ivanoska, G. Mirceva, **K. Trivodaliev**, S. Kalajdziski - "Hierarchical Protein Classification based on Gene Ontology and Decision Trees", *in web Proceedings of the 2nd ICT Innovations Conference*, Ohrid, Macedonia, 12-15 September 2010

4. I. Cingovska, A. Bogojeska, **K. Trivodaliev**, S. Kalajdziski, "Protein Function Prediction by clustering of Protein- Protein Interaction Network", *in proceedings of The ICT Innovations* 2011, *Advances in Intelligent and Soft Computing, Springer*, Volume 150, pp 39-49, 2012
5. **K. Trivodaliev**, I. Cingovska, S. Kalajdziski, "Protein Function Prediction by Spectral Clustering of Protein Interaction Network", *Database Theory and Application, Bio-Science and Bio-Technology, Communications in Computer and Information Science, Springer*, Volume 258, pp 108-117, 2011
6. I. Ivanoska, **K. Trivodaliev**, S. Kalajdziski, "Protein Function prediction using Semantic similarity metrics AND Random walk algorithm", *in proceedings of the The 9th Conference for Informatics and Information Technology (CIIT 2012)*
7. C. Atanasovska, **K. Trivodaliev**, S. Kalajdziski, "Determination of protein functional groups using the Bond Energy Algorithm", in *the 4th ICT Innovations Conference 2012*, Ohrid, Macedonia
8. I. Ivanoska, **K. Trivodaliev**, S. Kalajdziski, "Protein Function Prediction Using Semantic Driven K-Medoids Clustering Algorithm", *International Journal of Machine Learning & Computing,* Vol. 4 Issue 1, p52, 2014
9. **K. Trivodaliev**, A. Bogojeska, L. Kocarev, "Exploring Functional Prediction in Protein Interaction Networks via Clustering Methods", *Plos One* (IF:3.73, 5yr IF: 4.244) (in revision)

8.2 Листа на користени трудови во истражувањето

[1] Schrödinger, E., "What is life?: The physical aspect of the living cell.", The University Press, Cambridge, 1944

[2] Consortium, Gene Ontology, "The GO database and informatics resource", Nucl Acids Res 2004, 32:D258-D261

[3] Newman, MEJ, "Assortative Mixing in Networks", Phys Rev Lett, 2002, 89(20).

[4] Newman, MEJ, " Mixing patterns in networks", Phys Rev E, 2003, 67(2).

[5] Barabasi and Oltvai, "Network Biology: Understanding the Cell's Functional Organization", Nat Rev Genet, 2004, 5(2):101-13.

[6] Guelzim et al., "Topological and causal structure of the yeast transcriptional regulatory network", Nature Genetics, 2002, 31:60-63.

[7] Girvan and Newman, "Community structure in social and biological networks", PNAS, 2002, 99(12):7821-7826.

[8] Rives and Galitski, "Modular organization of cellular networks", PNAS, 2003, 100:1128-1133.

[9] Spirin and Mirny, "Protein complexes and functional modules in molecular networks", PNAS, 2003, 100:12123-12128.

[10] Lee et al., “Transcriptional Regulatory Networks in Saccharomyces cerevisiae”, Science, 2002, 298(5594):799- 804.
[11] Milo et al., “Network Motifs: Simple Building Blocks of Complex Networks”, Science, 2002, 298:824-827.
[12] Orr et al., “Network motifs in the transcriptional regulation network of Escherichia coli”, Nat Genet, 2002, 31:64-68.
[13] Berg and Lässig, “Local graph alignment and motif search in biological networks”, PNAS, 2004, 101:14689-14694
[14] Koyuturk et al., “An efficient algorithm for detecting frequent subgraphs in biological networks” Bioinformatics, 2004, 20(Suppl 1):i200-i207
[15] Kitano, “Biological Robustness”, Nat Genet, 2004, 5:826-838
[16] Tong et al., “Global Mapping of the Yeast Genetic Interaction Network”, Science, 2004, 808-813
[17] Yu et al., “Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs”, Genome Res, 2004, 14(6):1107-1118
[18] Sharan et al., “Conserved patterns of protein interaction in multiple species”, PNAS, 2005, 102(6):1974-1979
[19] Pandey et al., “Functional annotation of regulatory pathways”, Bioinformatics, 2007, 23(13):i377-i386
[20] Hart et al., “How complete are current yeast and human protein-interaction networks?”, Genome Biol, 2006, 7(11):120
[21] Chen, Y. and Xu, D., “Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae”, Nucl Acids Res, 2004, 32(21):6414-24
[22] Gavin et al., “Proteome survey reveals modularity of the yeast cell machinery”, Nature, 2006, 440:631-636
[23] Uetz, P. and Finley, R., “From protein networks to biological systems”, FEBS Letters, 2005, 579(8):1821-1827
[24] Ideker, T. and Sharan, R., “Protein networks in disease”, Genome Res, 2008, 18:644-52
[25] He and Zhang, “Why Do Hubs Tend to Be Essential in Protein Networks?”, PLoS Genet, 2006, 2(6):e88
[26] Aytuna et al., “Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces”, Bioinformatics, 2005, 21(12):2850-55
[27] Jansen et al., “A Bayesian networks approach for predicting protein-protein interactions from genomic data”, Science, 2003, 302:449-453
[28] Sharan et al., “Network-based prediction of protein function” Mol Sys Bio, 2007, 3:88
[29] von Mering et al., “Comparative assessment of large-scale data sets of protein-protein interactions”, Nature, 2002, 417:399-403
[30] Bader et al., “BIND: the Biomolecular Interaction Network Database”, Nucl Acids Res, 2003, 31(1):248-250
[31] Hermjakob et al., “IntAct: an open source molecular interaction database”, Nucl Acids Res, 2004, 32:D452-D455
[32] Pagel et al., “The MIPS mammalian protein-protein interaction database”, ioinformatics, 2005, 21(6):832-834
[33] Xenarios and Eisenberg, “Protein interaction databases”, Curr Opin Biotechnol, 2001, 12(4):334-339
[34] Xenarios et al., “DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions”, Nucl Acids Res, 2002, 30(1):303-305
[35] Zanzoni et al., “MINT: a Molecular INTeraction database”, FEBS Letters, 2001, 513(1):135-140
[36] Pandey G., Kumar V., Steinbach M., “Computational Approaches for Protien Function Prediction: A Survey”, Technical Report, Department of Computer Science and Engineering, University of Minnesota, 2006

[37] Ito et al., " A comprehensive two-hybrid analysis to explore the yeast protein interactome", Genetics, vol. 98, no. 8, 2001
[38] Uetz et al.," A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae", Nature, 2000
[39] Nakaya, A., Yoshizawa, A. C., Goto, S., Kanehisa, M., "Indirect Relations in Yeast Protein Interactome", Genome Informatics 13, 2002
[40] Salwinski et al., "The Database of Interacting Proteins", Nucleic Acids Res. 2004 January 1
[41] Breitkreutz, B. J., Stark, C., Tyers, M., "The GRID: The General Repository for Interaction Datasets", Genome Biology, 2003
[42] Payne, W. E. and Garrels, J. I., "Yeast Protein Database (YPD): a database for the complete proteome of Saccharomyces cerevisiae", Nucleic Acids Research, 1997, Vol.25, No.1
[43] Dwight et al., "Saccharomyces Genome Database (SGD) provides secondary gene annotation using Gene Ontology (GO)", Nucleic Acids Research, 2002, Vol.30, No.1
[44] Vasquez, A., Flammini, A., Maritan, A., Vespignani, A.,"Modelling of Protein Interaction Networks", ComPlexUs (2003); 1:38-44
[45] Bader and Hogue, "Analyzing yeast protein-protein interaction data obtained from different sources", Nat Biotechnol, 2002, 20:991-997
[46] Deane et al., "Protein interactions: two methods for assessment of the reliability of high throughput observations", Mol Cell Proteomics, 2002, 1(5):349-356
[47] Mrowka, R. et al., "Is there a bias in proteome research?", Genome Res, 2001, 11:1971-1973
[48] Deng, M., Sun, F., Chen, T.,"Assesment of Reliability of Protein-Protein Interactions and Protein Function Prediction", Pacific Symposium on Biocomputing 8; pp. 140-151, 2003
[49] Ravasz, E. et al., "Hierarchical organization of modularity in metabolic networks", Science, 2002, 297:1551-1555
[50] Goldberg and Roth, "Assessing experimentally derived interactions in a small world", PNAS, 2003, 100(8):4372-4376
[51] Saito et al., "Construction of reliable protein-protein interaction networks with a new interaction generality measure", Bioinformatics, 2002a, 19:756-763
[52] Saito et al., "Interaction generality, a measurement to assess the reliability of a protein-protein interaction", Nucleic Acids Res, 2002b, 30(5):1163-1168
[53] Saito et al., "Construction of reliable protein-protein interaction networks with a new interaction generality measure", Bioinformatics, 2003, 19(6): 756-763
[54] Chen et al., "Increasing confidence of protein interactomes using network topological metrics", Bioinformatics, 2006a, 22(16):1998-2004
[55] Pandey, G. and Kumar, V., "Incorporating functional inter-relationships into algorithms for protein function prediction", ISMB/ECCB Special Interest Group meeting on Automated Function Prediction, 2007
[56] Bork, P. and Koonin, E.V., "Predicting functions from protein sequences—where are the bottlenecks?", Nat Genet, 1998(18), 313–318
[57] Needleman, SB., Wunsch, CD., "A general method applicable to the search for similarities in the amino acid sequences of two proteins", J. Mol. Biol., 1970(48), 443-453.
[58] Smith, TF., Waterman, MS., "Identification of common molecular subsequences", J. Mol. Biol., 1981(147), 195-197.
[59] Altschul, S.F., et al., "Basic local alignment search tool", J. Mol. Biol., 1990(215), 403–410.
[60] Pearson, W. R., "Effective protein sequence comparison", Methods Enzymol., 1996(266), 227–258.
[61] Sturrock, S. S. and Collins, J. F., "MPsrch version 1.3", Biocomputing Research Unit, University of Edinburgh, Edinburgh, UK, 1993.
[62] Bairoch, A. et al., "The PROSITE database, its status in 1995", Nucleic Acids Res., 1995(24), 189–196.

[63] Henikoff, S. and Henikoff, J. G., "Protein family classification based on searching a database of blocks", Genomics, 1994(19), 97–107.
[64] Attwood, T. K., et al., "PRINTS – A database of protein motif fingerprints", Nucleic Acids Res., 1994(22), 3590–3596.
[65] Pearson, W. R., Lipman, D. J., "Improved tools for biological sequence comparison", Proc Natl Acad Sci USA, 1988(85), 2444-2448.
[66] Benner, S.A., et al., "Functional inferences from reconstructed evolutionary biology involving rectified databases – an evolutionarily grounded approach to functional genomics", Res Microbiol, 2000(151), 97-106.
[67] Altschul, S.F., et al., "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", Nucleic Acids Res., 1997(25), 3389–3402.
[68] Holm L., Sander C., "Protein structure comparison by alignment of distance matrices", J Mol Biol, 1993(233), 123-138.
[69] Madej, T., Gibrat, JF., Bryant, SH., "Threading a database of protein cores", Proteins, 1995(23), 356–69.
[70] Orengo, CA., Taylor, WR., "SSAP: sequential structure alignment program for protein structure comparison", Methods Enzymol, 1996(266), 617-635.
[71] Harrison A., Pearl F., Sillitoe I., Slidel T., Mott R., Thornton J. M., Orengo C., "Recognising the fold of a protein structure", Bioinformatics, 2003(19), 1748-1759.
[72] Capra, JA., et al., "Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure", PLoS Comput Biol, 2009, 5(12), e1000585.
[73] Najmanovich, R., et al., "Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites", Bioinformatics, 2008, 24(16), 105-11.
[74] Gold, ND., Jackson, RM., "Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships", J Mol Biol., 2006, 3, 355(5), 1112-24.
[75] Chang, DTH., Weng, YZ., Lin, JH. et al., "Protemot: prediction of protein binding sites with automatically extracted geometrical templates" *Nucleic Acids Res*, 2006, 34, W303–9.
[76] Wass, MN., et al., "3DLigandSite: predicting ligand-binding sites using similar structures", Nucleic Acids Res, 2010, 38, W469–W473.
[77] Torrance, J.W., et al.., "Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families", J Mol Biol, 2005, 347, 565-581.
[78] Kinoshita, K., Kono, H. and Yura, K., "Prediction of Molecular Interactions from 3D-Structures: From Small Ligands to Large Protein Complexes", In: *Prediction of Protein Structures, Functions, and Interactions (ed J. M. Bujnicki), John Wiley & Sons, Ltd*, Chichester, UK, 2008
[79] Glazer, DS., Radmer, RJ., Altman, RB., "Improving structure-based function prediction using molecular dynamics", Structure, 2009, 17, 919–929.
[80] Dodson, GG.,Lane, DP.,Verma, CS., "Molecular simulations of protein dynamics: new windows on mechanisms in biology", EMBO Rep, 2008, 9, 144–150.
[81] Pierri, C.L., Parisi, G., Porcelli, V., "Computational approaches for protein function prediction: A combined strategy from multiple sequence alignment to molecular docking-based virtual screening", Biochimica et Biophysica Acta - Proteins and Proteomics, 2010, 1804 (9), 1695-1712.
[82] Favia, A. D. and Nobeli, I., "Using Chemical Structure to Infer Biological Function", In: *Computational Approaches in Cheminformatics and Bioinformatics (eds R. Guha and A. Bender), John Wiley & Sons, Inc*., Hoboken, NJ, USA, 2011.
[83] Chang, DT., Oyang, YJ., Lin, JH., "MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm", Nucleic Acids Res, 2005, 33, W233–W238.
[84] Whisstock, JC., Lesk, AM., "Prediction of protein function from protein sequence and structure", Q Rev Biophys, 2003, 36, 307-40.
[85] Galperin, MY., Walker, DR., Koonin, EV., "Analogous enzymes: independent inventions in enzyme evolution", Genome Res, 1998, 8, 779-90.

[86] Rost, B., "Enzyme function less conserved than anticipated", J Mol Biol, 2002, 318, 595-608.
[87] Tan, P.N., Steinbach, M., Kumar, V., "Introduction to Data Mining", Addison-Wesley, 2005.
[88] Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y., Chen, Y., "Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity", Proteomics, 2006, 6, 4023-4037.
[89] Al-Shahib, A., Breitling, R., Gilbert, D. R., "Predicting protein function by machine learning on amino acid sequences – a critical evaluation", BMC Genomics, 2007, 8, 78.
[90] Marcotte, E. M. 2004. Practical computational approaches to inferring protein function. *Drug Discovery Today: BIOSILICO 2*, 1, 24–29.
[91] Marcotte, E. M., et al. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science 285*, 5428, 751–753.
[92] Marcotte, E. M., et al. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature 402*, 6757, 83–86.
[93] Pellegrini, M., et al., "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles", Proc Natl Acad Sci U.S.A. 96, 1999, 8, 4285–4288.
[94] Marcotte, E. M., et al., "Localizing proteins in the cell from their phylogenetic profiles", Proc Natl Acad Sci U.S.A. 97, 2000, 22, 12115–12120.
[95] Liberles, D. A., et al., "The use of phylogenetic profiles for gene predictions", Current Genomics 3, 2002, 3, 131–137.
[96] Enault, F., et al., "Phydbac "Gene Function Predictor" : a gene annotation tool based on genomic context analysis", BMC Bioinformatics 6, 2005, 247.
[97] Zheng, Y., Roberts, R. J., and Kasif, S., "Genomic functional annotation using co-evolution profiles of gene clusters", Genome Biology 3, 2002, 11, research0060.1–0060.9.
[98] Sjolander, K., "Bayesian evolutionary tree estimation", In Proc. Computing in the Genome Era, 1997.
[99] Engelhardt, B. E., et al., "Protein molecular function prediction by bayesian phylogenomics", PLoS Comput Biol. 1, 2005, 5, e45.
[100] Narra, K. and Liao, L., "Use of extended phylogenetic profiles with E-values and support vector machines for protein family classification", International Journal of Computer and Information Sciences 6, 2005, 1.
[101] Vert, J.-P., "A tree kernel to analyze phylogenetic profiles", Bioinformatics 18, 2002, Suppl 1, S276–S284.
[102] Swift, S., et al., "Consensus clustering and functional interpretation of gene-expression data", Genome Biology 5, 2004, 11, R94.
[103] Bryan, K., et al., "Biclustering of expression data using simulated annealing", In Proc. 18th IEEE Symposium on Computer-Based Medical Systems (CBMS)., 2005, 383–388.
[104] Liu, J., et al., "Gene ontology friendly biclustering of expression profiles", In Proc. IEEE Computational Systems Bioinformatics Conference (CSB)., 2004, 436–447.
[105] Brown, M. P., et al., "Knowledge based analysis of microarray gene expression data by using support vector machines", Proc Natl Acad Sci 97, 2000, 1, 262–267.
[106] Mateos, A., et al., "Systematic learning of gene functional classes from dna array expression data by using multilayer perceptrons", Genome Research 12, 2002, 11, 1703–1715.
[107] Mukherjee, S., "Classifying microarray data using support vector machines", In A Practical Approach to Microarray Data Analysis, 2003, D. P. Berrar, W. Dubitzky, and M. Granzow, Eds. Kluwer Academic Publishers, Chapter 9, 166–185.
[108] Moller-Levet, C. S., et al., "Clustering of gene expression time-series data", Tech. report, 2003, Department of Computer Science, University of Rostock, Germany.
[109] Hvidsten, T., et al., "Predicting gene function from gene expressions and ontologies", In Proc. Pacific Symposium on Biocomputing (PSB)., 2001, 299–310.

[110] Laegreid, A., Hvidsten, T. R., Midelfart, H., Komorowski, J., and Sandvik, A. K., "Predicting gene ontology biological process from temporal gene expression patterns", Genome Research 13, 2003, 5, 965–979.
[111] Deng, X. and Ali, H. H., "A hidden markov model for gene function prediction from sequential expression data", In Proc. CSB., 2004, 670–671.
[112] Izumitani et al., "Assigning Gene Ontology Categories (GO) to Yeast Genes Using Text-Based Supervised Learning Methods", CSB2004, 2004, 503-504
[113] Asako et al., "Automatic extraction of gene/protein biological functions from biomedical text", Bioinformatics, 2005, 21(7):1227-1236
[114] Fraser, A. G., Marcotte, E. M., "A Probabilistic View of gene Function", Nature Genetics, Vol. 36, Number 6, 2004
[115] Kirac et al., "Annotating proteins by mining protein interaction networks", Bioinformatics, 2006, 22:e260-e270
[116] Kirac and Ozsoyoglu, "Protein Function Prediction based on Patterns in Biological Networks", RECOMB 08, 2008
[117] Schwikowski, B., Uetz, P., Fields, S., "A Network of Protein-Protein Interaction in Yeast", Nature Biotechnology, Vol. 18, 2000
[118] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T., "Assessment of Prediction Accuracy of Protein Function from Protein-Protein Interaction Data", Yeast, 18: 523-531. 2001
[119] Chua, H. N., Sung, W.-K., Wong, L., "Exploiting Indirect Neighbors and Topological Weight to Predict Protein Function from Protein-Protein Interactions", Bioinformatics, Vol. 22, No. 13, pp. 1623-1630, 2006
[120] Vasquez, A., Flammini, A., Maritan, A., Vespignani, A., "Global Protein Function Prediction from Protein-Protein Interaction Networks", Nature Biotechnology, Vol. 21, Number 6, 2003
[121] Sun, S., Zhao, Y., Jiao, Y., Yin, Y., Cai, L., Zhang, Y., Lu, H., Chena, R., and Bu, D., "Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm", FEBS Letters 580, 7, 1891–1896, 2006
[122] Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., "Whole-Genome Annotation by Using Evidence Integration in Functional-Linkage Networks", PNAS, Vol. 101, No. 9, pp. 2888-2893, 2004
[123] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M., "Whole-Proteome Prediction of Protein Function via Graph-Theoretic Analysis of Interaction Maps", Bioinformatics, Vol. 21, Suppl. 1, pp. 302-310, 2005
[124] Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F., "Prediction of Protein Function Using Protein-Protein Interaction Data", Journal of Computational Biology, Vol. 10, No. 6, pp. 947-960, 2002
[125] Letovsky, S., Kasif, S., "Predicting Protein Function from Protein/Protein Interaction Data: a Probabilistic Approach", Bioinformatics, Vol.19 (2003) 197-204
[126] Freschi, V., "A Graph-Based Semi-Supervised Algorithm for Protein Function Prediction from Interaction Maps", Proc. of Third International Conference Learning and Intelligent Optimization, Machine Learning and Intelligent Optimization Workshop, LNCS 5851, pp. 249-258, 2009.
[127] Lin, C., Cho, Y., Hwang, W., Pei, P., Zhang, A., "Clustering Methods in Protein-Protein Interaction Networks", Knowledge Discovery in Bioinformatics: Techniques, Methods and Application, Chapter. 1, John Wiley and Sons. Inc, 2006
[128] Chen, J., Yuan, B., "Detecting Functional Modules in the Yeast Protein-Protein Interaction Network", Bioinformatics, Vol. 22, No. 18, pp. 2283-2290, 2006
[129] Lancichinetti, A., Fortunato, S., Radicchi, F., "Benchmark Graphs for testing Community Detection Algorithms", Physical Review E78, 046110, 2008
[130] Bader, G. D., Hogue, C. WV., "An automated method for finding molecular complexes in large protein interaction networks", BMC Bioinformatics 2003, 4:2, 2003

[131] Altaf-Al-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., Kanaya, S., "Development and Implementation of an Algorithm for Detection of Protein Complexes in Large Interaction Networks", BMC Bioinformatics 7:207, 2006
[132] Sharan, R., Ideker, T., Kelley, B., Shamir, R. Karp, RM., "Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data", Computational Biology 12: 835-846, 2005
[133] Przulj, N., Wigle, D. A., Jurisica, I., "Functional Topology in a Network of Protein Interactions", Bioinformatics, Vol. 20, No. 3, pp. 340-348, 2004
[134] Enright, AJ, Dongen, SV, Ouzounis, CA, "An efficient algorithm for large-scale detection of protein families", Nucleic Acids Res, 30(7):1575-84, 2002
[135] King, AD, Przulj, N, Jurisica, I, "Protein complex prediction via cost-based clustering", Bioinformatics, 20(17):3013-20, 2004
[136] Blatt M, Wiseman S, Domany E: Superparamagnetic clustering of data. Phys Rev Lett, 76(18):3251-3254, 1996.
[137] Asthana, S., King, O. D., Gibbons, F. D., Roth, F. P., "Predicting Protein Complex Membership Using Probabilistic Network Reliability", Genome Research, 14:1170-1175, 2004
[138] Samanta, M. P., Liang, S., "Predicting protein functions from redundancies in large-scale protein interaction networks", PNAS, October 28, Vol. 100, No. 22, 2003
[139] Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guénoche, A., Jacq, B., "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network", Genome Biology 5:R6, 2003
[140] Fortunato, S., "Community Detection in Graphs", Physics Reports 486, pp. 75-174, 2010
[141] Dunn, R., Dudbridge, F., Sanderson, C. M., "The Use of Edge-Betweenness Clustering to Investigate Biological Function in Protein Interaction Networks", BMC Bioinformatics 6:39, 2005
[142] Adamchek, B., Palla, G., Farkas, I.J., Derenyi, I., Viscek, T., "CFinder: Locating Cliques and Overalpping Modules in Biological Networks", Bioinformatics 22, pp. 1021-1023, 2006
[143] Sen, T. Z., Kloczkowski, A., Jernigan, R. L., "Functional Clustering of Yeast Proteins from the Protein-Protein Interaction Network", BMC Bioinformatics 7:355, 2006
[144] Lewis, A. C. F., Jones, N. S., Porter, M. A., Deane, C. M., "The Function of Communities in Protein-Protein Interaction Networks at Multiple Scales", BMC Systems Biology 4:100, 2010
[145] V. Colizza, A. Flammini, A. Maritan, A. Vespignani; *Characterization and Modelling of Protein-Protein Interaction Networks*; Physica A 352 (2005) 1-27
[146] Erdös P, Rényi A.,"On random graphs", Publicationes Mathematicae, 1959, 6: 290–7.
[147] Bollobas B., "Random Graphs", Academic, London, 1985.
[148] Newman, M.E.J., Strogatz, S.H., Watts, D.J., "Random graphs with arbitrary degree distributions and their applications", Phys Rev E 64: 026118–1, 2001.
[149] Aiello, W., Chung, F., Lu, L., "A random graph model for power law graphs", Exp Math 10: 53–66, 2001.
[150] Watts, D.J., Strogatz, S.H., "Collective dynamics of 'small-world' networks", Nature 393: 440–2, 1998.
[151] Barabási A-L., Albert R., "Emergence of scaling in random networks", Science 286: 509–12, 1999.
[152] Li, L., Alderson, D., Tanaka, R., Doyle, J.C., et al., "Towards a theory of scale-free graphs: Definition, properties, and implications", Internet Math 4: 431–523, 2005.
[153] Wu, X.-R., Zhu, Y., Li, Y., "Analyzing Protein Interaction Networks via Random Graph Model", International Journal of Information Technology Vol.11, No. 8 (2005)
[154] Sole, R.V., Pastor-Satorras, R., Smith, E., Kepler, T.B., "A Model of Large-Scale Proteome Evolution", Adv. Complex Systems 5 (2002) 43.
[155] Pastor-Satorras, R., Smith, E., Sole, R.V., "Evolving protein interaction networks through gene duplication", J Theor Biol 222: 199–210, 2003.

[156] Wagner A., "How the global structure of protein interaction networks evolves", Proc Biol Sci 270: 457–66, 2003.
[157] Przulj, N., Corneil, D.G., Jurisica, I., "Modeling Interactome: Scale-Free or Geometric", Bioinformatics (2004) 20(18):3508-3515.
[158] Miller, G. A., "WordNet: A Lexical Database for English", Communications of the ACM Vol. 38, No. 11: 39-41, 1995.
[159] Rada, R., Mili, H., Bicknell, E., Blettner, M., "Development and application of a metric on semantic nets", IEEE Transaction on Systems, Man, and Cybernetics, volume 1, 1989.
[160] Wu, Z., Palmer, M. S., "Verb semantics and lexical selection", Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics, pp 133-138, 1994.
[161] Richardson, R., Smeaton, A. F., Murphy, J., "Using WordNet as a knowledge base for measuring semantic similarity between words", Technical Report CA-1294, Dublin, Ireland, 1994.
[162] Pekar, V., Staab, S., "Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision", Proceedings of the 19th international conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics, pp. 1–7, 2002.
[163] Yu, H., Gao, L., Tu, K., Guo, Z., "Broadly predicting specific gene functions with expression similarity and taxonomy similarity", Gene, Volume 352, pp. 75–81, 2005.
[164] Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., "A knowledge-based clustering algorithm driven by gene ontology", Journal of Biopharmaceutical Statistics 14, pp: 687–700, 2004.
[165] Wu, X., Zhu, L., Guo, J., Zhang, D.Y., Lin, K., "Prediction of yeast protein protein interaction network: insights from the gene ontology and annotations", Nucleic Acids Research 34, pp. 2137–2150, 2006.
[166] Pozo, A. D., Pazos, F., & Valencia, A., "Definining functional distances over GeneOntology", BMC Bioinformatics 2008, 9:50, 2008.
[167] Resnik, P.,"Using information content to evaluate semantic similarity", Proceedings of the IJCAI05, pp. 448 –453, 1995.
[168] Lin, D.,"An information-theoretic definition of similarity", Proceedings of the 15th Int. Conf. on Machine Learning, 1998.
[169] Schlicker, A., Domingues, F., Rahnenfuhrer, J., Lengauer, T., "A new measure for functional similarity of gene products based onGene Ontology", BMC Bioinformatics 2006, 7:302, 2006.
[170] Jiang, J., Conrath, D.W., "Semantic Similarity based on corpus and lexical taxonomy", Proc. Of10th Int. Conf. COLING, 1997.
[171] Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F., "A new method to measure the semantic similarity of GO term", Bioinformatics 2007, 23(10):1274-1281, 2007.
[172] Shen, Y., Zhang, S., Wong, H.S., and Zhang, L., "A new method for measuring the semantic similarity on Gene Ontology", in Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 533–538, 2010.
[173] Sevilla, JL, Segura, V, Podhorski, A, Guruceaga, E, Mato, JM, Martinez-Cruz, LA, Corrales, FJ, Rubio, A, "Correlation between Gene Expression and GO Semantic Similarity", IEEE/ACM Transactions on Computational Biology and Bioinformatics 2005.
[174] Couto, F, Silva, M, Coutinho, P, "Measuring Semantic Similarity between Gene Ontology Terms", Data & Knowledge Engineering 2007, 61**:**137-152.
[175] Schlicker, A, Domingues, FS, Rahnenführer, J, Lengauer, T, "A new measure for functional similarity of gene products based on Gene Ontology", BMC Bioinformatics, **7**(302), 2006
[176] Azuaje, F, Wang, H, Bodenreider, O, "Ontology-driven similarity approaches to supporting gene functional assessment", Proceedings of the ISMB2005 SIG meeting on Bio-ontologies 2005.
[177] Ovaska, K., Laakso, M., Hautaniemi, S., "Fast Gene Ontology based clustering for microarray experiments", BioData Mining 2008, 1:11.

[178] Jonsson, P.F., Cavanna, T., Zicha, D., Bates, P.A., "Cluster analysis of networks generated through homology: Automatic identification of important protein communities involved in cancer metastasis", BMC Bioinf. 7 (2006) 2.

[179] Newman, M.E.J., and Girvan, M., "Finding and evaluating community structure in networks," Phys. Rev. E, vol. 69, no. 2, p. 026113, 2004

[180] Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., "Global Landscape of Protein Complexes in the Yeast Saccharomyces cerevisiae", Nature, Vol. 440, pp. 637-643, 2006

[181] Ito et al., "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins", Proc Natl Acad Sci U S A, 97(3):1143-7, 2000.

[182] Ho, Y., et al., "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry", Nature, 415(6868):180-3, 2002.

[183] Langfelder, P., Zhang, B., Horvath, S., "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R", Bioinformatics 24(5):719-720, 2008.

[184] Clauset, A., Newman, M.E.J., and Moore, C. , "Finding community structure in very large networks", Physical Review E, vol. 70, no. 6, pp. 066 111+, Dec. 2004.

[185] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E., "Fast unfolding of communities in large networks", Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, pp. P10 008+, Oct. 2008.

[186] Lambiotte, R., "Multi-Scale Modularity in Complex Networks", Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on, 546-553, 2010

[187] Lambiotte, R., Delvenne, J-C., Barahona, M., "Laplacian dynamics and multiscale modular structure in networks", arXiv:0812.1770, 2009

[188] Rosvall, M. and Bergstrom, C.T., "Maps of random walks on complex networks reveal community structure", Proceedings of the National Academy of Sciences, vol. 105, no. 4, pp. 1118–1123, 2008

[189] Evans, T.S., and Lambiotte, R., "Line graphs, link partitions, and overlapping communities", Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), vol. 80, no. 1, pp. 016 105+, 2009.

[190] Fred, A. and Jain, A., "Robust data clustering", in Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 2, 2003, pp. II-128-II-133 vol.2

[191] Fawcett, T., "An introduction to ROC analysis", Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[192] Hart, GT, Ramani, AK, Marcotte, EM, "How complete are current yeast and human protein interaction networks?", Genome Biol 7: 120, 2006.