Application of Multi-Resolution Partitioning of Interaction Networks to the Study of Complex Disease

Malte D. Luecken

Trinity College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2016

In memory of Heino Lücken,
from whom I learned perseverance and determination.
You are sorely missed.

Abstract

Large-scale gene expression studies are widely used to identify genes that are differentially expressed between phenotypes relevant to disease. Often thousands of differentially expressed genes (DEGs) are found using this type of analysis, which complicates the interpretation of the data. In this project we treat DEGs as windows into the biological processes that underlie disease. In order to find these processes, we put DEGs into the context in which they perform their functions – through the interactions of their protein products.

Protein-protein interactions can provide biological context to DEGs in the form of functional modules. These modules are groups of proteins that together perform cellular functions. In this thesis we have refined a functional module detection process that consists of two steps. Firstly, community detection methods are applied to protein interaction networks (PINs) to detect groups of interacting proteins, and secondly, the biological coherence of the proteins grouped together is evaluated to select communities that represent potential functional modules.

Two features that are central to this work are the detection of modules at different scales of network organization, and CommWalker, a module evaluation method that we developed which is able to detect signals of poorly-studied functions. By integrating these methods into our functional module detection process, we were able to obtain a good coverage of potential functional modules. Testing for enrichment of DEGs on these functional modules can uncover biological processes that are involved in the contrasted phenotypes and merit further investigation.

We have applied our pipeline to find differentially regulated functions between hypoxic and normoxic breast cancer cell lines, and between M1 and M2 macrophages. Our results generate biological hypotheses of cellular functions that are differentially regulated in the investigated phenotypes, and proteins that are involved in these functions. We were able to validate several proteins in enriched modules which did not correspond to DEGs that were input into the pipeline, which suggests our methodology can reveal new biological insight.

Author Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in text.

Malte D. Luecken
4th October 2016

Acknowledgements

I would like to thank my supervisors Charlotte, Gesine, and Matt, for their guidance, support, and tireless attempts to improve my scientific writing. I feel very lucky to have had the opportunity to learn from such knowledgeable and inspiring people.

Thanks are also due to all the people that helped with the work in this thesis. Sean Mason and Andrea Crosby ensured that the biological focus was never lacking along the way. The biological insight provided by Francesca Buffa and Fernando Martinez; the support from Vincent Traag, Mason Porter, Caleb Webber, Jiye Shi, and Sebastian Kelm; the proofreading by Florian Klimm; and the other members of the Oxford Protein Informatics Group who were a sounding board for my ideas played an important role in this project. Particularly, I would like to mention Luis Ospina, a good friend and a great intellectual sparring partner who taught me to think like a statistician.

I would also like to express my gratitude to all the people who made the past four years a memorable experience and provided motivation and support. The OPIG family, Beverley and her cookies that brought people together, floorball friends, DTC friends, Nikola Vlahov, Lien Davidson, Olivia Viessmann, and other college friends had a profound impact on my time at Oxford. My main sources of motivation and support have always been my family and closest friends. Idan and Dave have both had to endure complaining and always came by to share important moments or provide a distraction when needed. The incredible and enduring support from my family, and the example provided by my grandparents, are the reason I am where I am. Lastly, I am indepted to the most important person in my life. Nayab, you have been there to support me in the lows, and share the highs of this ride – a ride that only marks the beginning of our journey together. You are my person.

Contents

List of Abbreviations **xv**

1 Introduction **1**
1.1 Overview . . . 1
1.1.1 Cellular functions are modular . . . 1
1.1.2 Functional module detection . . . 2
1.1.3 The developed methodology . . . 5
1.2 Biological applications . . . 8
1.2.1 Breast cancer hypoxia . . . 8
1.2.2 Macrophage classification . . . 10
1.3 Thesis outline . . . 13

2 Background **15**
2.1 Protein-protein interactions . . . 16
2.1.1 Measuring protein-protein interactions . . . 16
2.1.2 Data quality . . . 19
2.1.3 Protein-protein interaction databases . . . 21
2.1.4 Network description of protein-protein interaction data . . . 24
2.2 Networks . . . 25
2.2.1 Network summary statistics . . . 26
2.2.2 Network structure . . . 28
2.3 Communities in networks . . . 29
2.3.1 Resolution . . . 31
2.3.2 Non-overlapping community detection . . . 31
2.3.3 Overlapping community detection . . . 37
2.3.4 Comparing community detection methods . . . 43
2.4 Functional annotations . . . 44
2.4.1 Ontologies and ontological concepts . . . 45
2.4.2 The Gene Ontology . . . 46
2.4.3 Other functional annotation resources . . . 49
2.5 Functional homogeneity . . . 51
2.5.1 Functional enrichment . . . 51

2.5.2 Functional similarity 53
2.6 Gene expression data 58
2.6.1 Co-expression analysis 59
2.6.2 Differential gene expression 60
2.7 Summary 61

3 Data sets and processing **65**
3.1 Network data 65
3.1.1 Protein-protein interaction data 65
3.1.2 Network partitions 69
3.2 Functional annotations 70
3.2.1 Gene Ontology 71
3.2.2 Human Phenotype 73
3.2.3 Functional homogeneity 74
3.3 Gene expression data 75
3.3.1 Co-expression analysis 75
3.3.2 Biological applications 76

4 Functional homogeneity evaluation **79**
4.1 Introduction 79
4.2 Testing GO annotations 81
4.2.1 Methods 81
4.2.2 Results 82
4.3 Testing functional enrichment 82
4.3.1 Methods 84
4.3.2 Results 84
4.4 Selecting a semantic similarity measure 86
4.4.1 Methods 87
4.4.2 Results 95
4.4.3 Discussion 100
4.5 Testing Human Phenotype annotations 102
4.5.1 Methods 102
4.5.2 Results 103
4.6 Discussion and conclusions 105

5 Attempting to identify functional modules in protein interaction networks 107
5.1 Introduction . . . 107
5.2 Evaluating communities using functional homogeneity . . . 109
5.3 Criteria for the evaluation of community detection methods . . . 110
5.4 Louvain community detection . . . 112
5.4.1 Configuration model . . . 112
5.4.2 Constant Potts model . . . 118
5.5 Overlapping community detection . . . 121
5.5.1 BigCLAM . . . 122
5.5.2 Link clustering . . . 124
5.6 PIN comparison . . . 128
5.7 Discussion and conclusions . . . 130

6 CommWalker 135
6.1 Introduction . . . 135
6.2 Annotation bias . . . 137
6.3 CommWalker . . . 143
6.4 CommWalker module analysis . . . 148
6.4.1 Thresholding . . . 148
6.4.2 The effect of CommWalker on evaluation results . . . 150
6.4.3 Coverage of functionally significant modules . . . 152
6.4.4 Module statistics . . . 152
6.5 CommWalker module validation . . . 158
6.5.1 Gene co-expression validation . . . 159
6.5.2 Case studies . . . 163
6.6 Discussion and conclusions . . . 164

7 Biological applications 167
7.1 Introduction . . . 167
7.2 Methodology . . . 169
7.2.1 Functional module detection . . . 169
7.2.2 Overlaying DEGs . . . 171
7.2.3 Consensus clustering of enriched modules . . . 172
7.2.4 Module prioritization . . . 173
7.3 Enriched modules in biological applications . . . 174
7.3.1 Visualization . . . 174
7.3.2 Hypoxia . . . 175
7.3.3 Macrophage differentiation . . . 179
7.4 Retrospective implications for module detection . . . 184
7.5 Discussion and conclusions . . . 186

8 Conclusions and future work **189**

Appendices

A Link Edgehop Clustering **197**
A.1 Introduction 197
A.2 Data sets and processing 198
A.2.1 Networks 198
A.2.2 Gene Ontology annotations 200
A.3 Ledgehop methodology 203
A.4 Comparing network partitions: Normalized Mutual Information . . 205
A.5 Testing Ledgehop 207
A.5.1 Differences between Ledgehop and link clustering 208
A.5.2 Capturing known community structure 210
A.6 Biological applications 214
A.7 Discussion and conclusions 216

B Data Sets **221**
B.1 Pre-processing protein-protein interaction data 221
B.2 Microarray probe set mapping 222
B.3 Network statistics 224

C Disconnected communities in the Louvain algorithm **227**

D CommWalker Results **231**
D.1 T-Value stability on BioGrid-AP 231
D.2 Coverage of Accepted modules 232
D.3 Module Statistics 233
D.4 Computational Module Validation 258

Bibliography **261**

List of Abbreviations

PIN . . . . . Protein interaction network

DEG . . . . . Differentially expressed gene

IFN-γ . . . . . Interferon-γ – a cytokine that induces M1 macrophage polarization

IL-4/10/12/13 Interleukin 4/10/12/13 – cytokines associated with different macrophage polarizations

A-type . . . . . Associations – a protein-protein interaction characterized by biochemical binding

P-type . . . . . Physical associations – a protein-protein interaction characterized by biophysical attraction

TAP-MS . . . Tandem affinity purification with mass spectrometry – an experimental technique to measure A-type interactions

CPM . . . . . Constant Potts model – a null model for Modularity Maximization

DAG . . . . . Directed acyclic graph

IC . . . . . (Shannon) Information Content

GO . . . . . Gene Ontology

BP . . . . . Biological process GO sub-ontology

MF . . . . . Molecular function GO sub-ontology

CC . . . . . Cellular component GO sub-ontology

HPO . . . . . Human Phenotype Ontology

HP . . . . . Human Phenotype

OMIM . . . . . Online Mendelian Inheritance in Man database

FDR . . . . . False discovery rate – the fraction of significant tests that actually follow the null hypothesis

FWER . . . . Family-wise error rate – the probability of any test in a family of tests being false

1 Introduction

Contents

1.1 Overview 1
1.1.1 Cellular functions are modular 1
1.1.2 Functional module detection 2
1.1.3 The developed methodology 5
1.2 Biological applications 8
1.2.1 Breast cancer hypoxia 8
1.2.2 Macrophage classification 10
1.3 Thesis outline 13

1.1 Overview

1.1.1 Cellular functions are modular

Cells, the fundamental building blocks of life, are complex systems of biological molecules. Through the interplay of these molecules, cells perform functions. In humans for example, macrophages move around (chemotaxis) and engulf foreign bodies such as bacteria to digest them (phagocytosis) [1]; neurons transmit sensory signals via action potentials [2]; and cardiomyocytes form muscle tissue that contracts to pump blood around the body [3]. While early molecular biology was focused on finding individual molecules responsible for cellular functions, it is now thought

that cellular functions are performed by functional modules of molecules [4–6].

A functional module is a group of interacting molecules (proteins, DNA, RNA, and small molecules) that perform a distinct cellular function [5]. Modules have been shaped by the evolutionary process to be robust to perturbations and adaptive to changing environments, and thus may incorporate more than the essential components to perform a particular function [6,7]. In this aspect functional modules differ from pathways, which are chains of molecular interactions within a cell, resulting in the cell performing a function.

The modular level of organization of biological molecules is thought to represent the link between the genes that are expressed in a cell (genotype), and the traits of this cell (phenotype) [6,8–11]. Thus, it is of particular interest in the study of disease, where the activation or malfunctioning of specific molecular mechanisms cause unwanted phenotypes. Groups of disease-related genes can be mapped to disease phenotypes via functional modules [12, 13].

1.1.2 Functional module detection

At the end of the last century the ongoing efforts of the Human Genome Project [14] created an environment that promoted the development of technologies which were able to quickly generate large quantities of biological data at low cost [15]. The development of these high-throughput technologies for gene expression (eg. [16,17]), protein-protein interactions (eg. [18–20]), and protein function (eg. [21]) triggered the development of methods to systematically detect functional modules using the available large-scale, so-called "omics" data sets.

As mentioned in Section 1.1.1 functional modules are groups of biological molecules that interact to perform cellular functions. Thus, the detection of functional modules can be approached from two angles: via the physical interaction of molecules (eg. [22–28]; topological clustering), and via functional relatedness (eg. [29–31]; biological clustering).

Due to their availability in large-scale data sets, human and yeast protein-protein interactions have been the most common basis for functional module detection [32].

Recent efforts to map the human RNA-RNA interaction space will further expand available molecular interaction resources [33, 34].

In general, topological clustering is performed on protein-protein interaction data that are converted into a network representation, so-called protein interaction networks (PINs; see Section 2.1.4). Groups of proteins that interact more with each other than with the rest of the PIN are detected using network-based clustering techniques called community detection methods (see Section 2.3). While there appears to be a consensus that modules are represented by highly interacting sets of proteins in PINs [35], some methods look for distinct modules with sparse connections between them (non-overlapping modules, eg. [24–26, 28]; see Section 2.3.2), and others allow proteins to belong to multiple modules (overlapping modules; eg. [22, 23, 27]; see Section 2.3.3).

Biological clustering is performed based on data from which a functional relatedness can be inferred. For example, similar expression patterns of genes across samples are thought to link to involvement in a common biological process [36] and can thus be used to define functional modules of their protein products (eg. [30, 31, 37]). Likewise, genes that show similar mutation patterns across cancer samples have been used to detect functional modules in cancer [29].

While these two approaches have been successfully used to detect groups of proteins or genes that resemble functional modules, combining topological and biological clustering is likely to provide a more comprehensive view of the functional organization of molecular networks [38]. Indeed, it has been shown that integrating PINs, genetic networks describing functional links between genes (based on eg. epistasis), and gene regulatory networks improves our ability to capture biological functions compared to any of these data sets individually [39].

The most common integration of topological and biological clustering approaches is the superposition of gene expression data onto PINs. Methods that integrate these data sets can be subdivided into those that find active subnetworks (eg. [40–42]) and methods that perform community detection on protein-protein interactions weighted by expression data (eg. [43]). Active subnetworks are regions of PINs

where the associated genes exhibit strong changes in expression under perturbations such as disease [38]. Due to the focus on disease perturbations, active subnetworks do not necessarily correspond to modules with a discrete function, but may instead resemble disease modules which incorporate many functions affected in disease phenotypes [13, 44]. An active subnetwork approach has also been applied to cancer mutation data sets to detect subnetworks associated with highly mutated genes [45].

While functional annotations are commonly used to validate topological clustering techniques for PINs (eg. [22, 23, 25, 27]), they can also be integrated into the module detection process. This integration has been performed by using the similarity of the functional annotations of interacting proteins to weight interactions in PINs [46]; by selecting the largest set of connected proteins that share a functional annotation [47]; and by selecting PIN communities which are evaluated as functionally homogeneous at multiple clustering resolutions [48] (see also Section 2.5 for functional homogeneity, and Section 2.3.1 for the concept of resolution).

Data integration from more than two sources to detect functional modules has also been performed. Biological clustering based on gene co-expression and drug response data has been combined with pathway data integrated into a PIN to find cancer-related functional modules [49]. Furthermore, protein interaction, gene co-expression, transcription factor binding, and micro-RNA networks have been combined into a multilayer network to detect consensus modules across data types [50].

The described methods give only a small overview of topological clustering, biological clustering, and how topological and biological clustering has been combined to detect functional modules to date. A more extensive review of data integration for functional modules detection can be found in [38]. In future, it is likely that additional data sources such as experimental measurements of how protein-protein interactions change under perturbation [51] will be used to improve functional module detection.

One of the main challenges in the field of functional module detection is module validation. Often many hundreds of modules are predicted, and experimental

validation is tedious and slow. Thus, especially in early module detection approaches based solely on topological clustering adequate validation is often rudimentary or entirely lacking. Publications commonly follow the pattern of presenting a developed community detection method; showing its application to a particular biological network (possibly augmented with other biological data on the nodes); and qualitatively describing some of the modules found. Alternatively functional enrichment (see Section 2.5.1) is used to confirm the quality of the modules. This publication blueprint complicates the comparison of detected modules and methods, although comparative studies do exist [27].

In this work, we have put particular focus on the evaluation of predicted modules. We show that current methods of module evaluation are subject to bias and propose a novel framework to evaluate communities proposed as modules. This community evaluation framework is based on a biological clustering approach. In so doing, we integrate topological and biological clustering techniques to predict functional modules.

1.1.3 The developed methodology

Using a functional module detection pipeline, we have developed a computational methodology to find cellular functions that may be involved in a phenotype of interest. This methodology has the potential for uncovering molecular mechanisms underlying disease. It is an advancement over standard approaches in that it can highlight disease processes even when these processes are poorly studied or occur at different biological scales.

Differential gene expression studies capture how the expression levels of genes change between samples that exhibit two distinct phenotypes (see Section 2.6.2). Our methodology (Figure 1.1) uses differentially expressed genes (DEGs) from these data, maps them to their protein products, which in turn are mapped to functional modules. Each module that is enriched for DEGs in this way describes a potential function that is involved in the phenotype and is thus a biological hypothesis that can be experimentally validated. To make this experimental validation feasible we

focused on modules of size 6 - 35 as recommended by collaborators at UCB Pharma. The lower boundary of six proteins was used to reduce the likelihood that detected modules represent small protein complexes, and the upper size limit of 35 proteins ensures communities proposed as modules are viable to be experimentally tested.

Given that functional modules are used to map DEGs to differentially regulated functions, obtaining a good coverage of functional modules on protein space is of central importance for our methodology. Thus, two features were implemented to optimize our ability to detect functional modules: topological and biological clustering was separated into two steps (see Section 1.1.2 and Figure 1.1 which indicates the separation of functional similarity calculation and community detection), and modules were detected at multiple scales.

It has been argued that the combination of topological and biological clustering represents the most promising approach for functional module detection (see Section 1.1.2). By clearly separating the two steps as in [48] we were able to optimize each step individually for optimal detection of functional modules. Furthermore, as gene expression data is used to introduce phenotype-relevant information into the methodology at a later stage, this type of data was excluded from the module detection process to avoid circularity.

The idea of mapping disease-related proteins to modules is not new [13, 44]. One distinguishing factor of our approach is the use of multi-resolution module detection. Functional modules can be found at different scales. For example, a module in a molecular network may take the form of a connected set of proteins all related to the ribosome. This module will contain further substructures such as a group of ribosome subunit proteins, and groups of precursors to and processors of the ribosomal subunits [48]. As the optimal scale at which functional modules should be generated cannot be known a priori, we specifically set up our functional module detection approach to detect modules at different scales (see Section 2.3.1).

A second novelty in our module detection pipeline is the CommWalker community evaluation framework (see Chapter 6). This biological clustering framework was developed to deal with poorly studied proteins and their lack of detailed annotation.

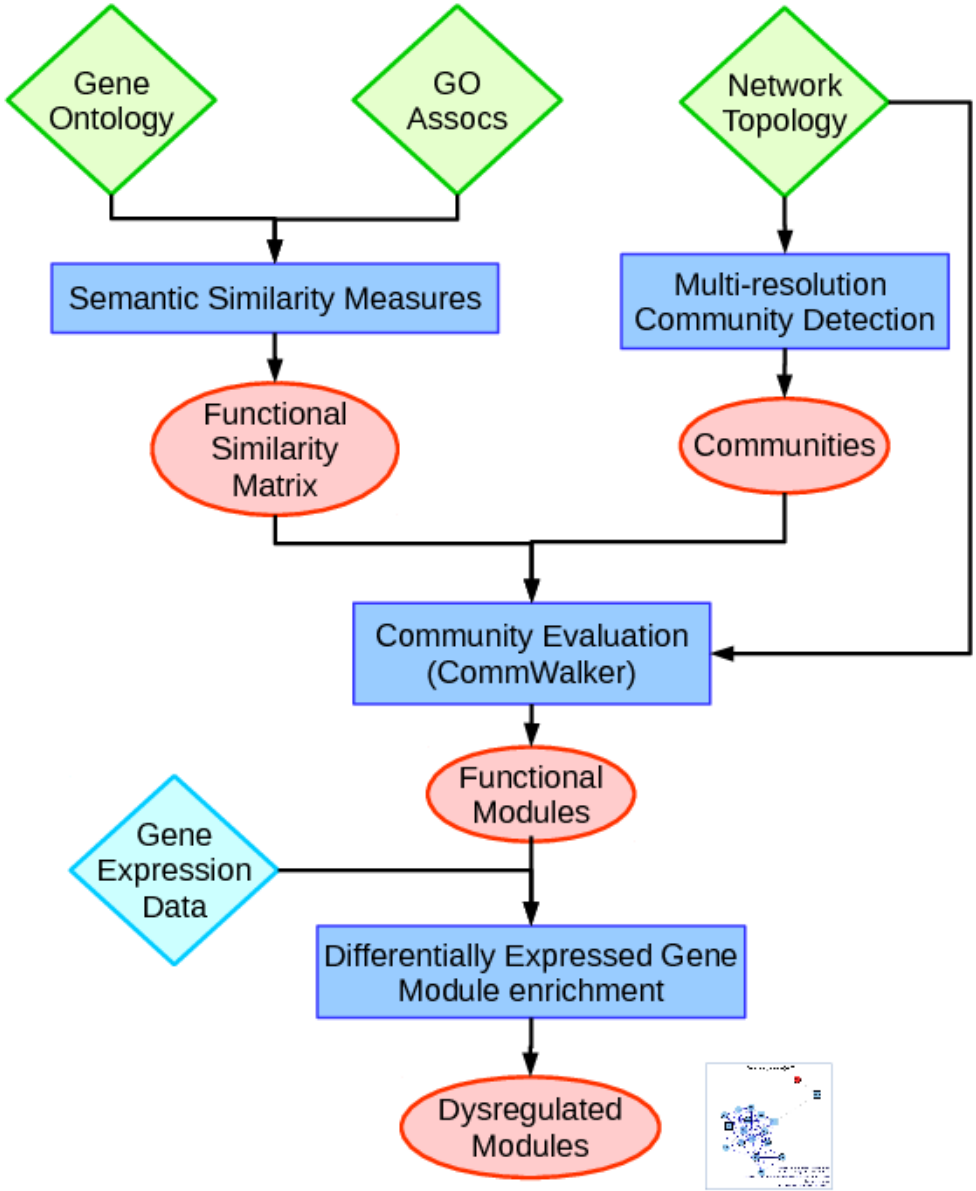

Figure 1.1: Pipeline flowchart Schematic of the developed pipeline to detect differentially regulated cellular functions represented as dysregulated functional modules. The rhombuses indicate the data sources, the blue boxes indicate implemented bioinformatics tools, and the red ellipses represent processed data. Data sources are split into green rhombuses representing data sources that are consistent for any application of the pipeline (The Gene Ontology (GO) structure and annotations, and a PIN), and the cyan rhombus indicating the data source that introduces disease-application-specific information into the analysis. The image next to the dysregulated modules ellipsis shows an example module output. In the pipeline a PIN is clustered into communities at multiple resolutions, and the GO is used to calculate the similarity of any two proteins in this PIN. Using our novel community evaluation framework CommWalker, we retain communities of functionally coherent sets of proteins based on the GO functional similarity scores. These communities represent our set of functional modules which are tested for enrichment of differentially expressed genes (DEGs). Functional modules enriched for DEGs from the gene expression data set are dysregulated modules that represent differentially regulated cellular functions.

Its application has the effect that we can detect signals of dysregulated cellular functions even when these functions are poorly studied. These two features have been integrated to produce a generally applicable tool that can help to uncover the mechanistic basis of disease phenotypes.

Our pipeline can be viewed as an extension to methods such as Gene Set Enrichment Analysis (GSEA) [52] which evaluates biological functions implicated in differential gene expression data. GSEA takes a pre-defined set of proteins which is linked to a particular biological function and evaluates the significance of the combined expression level of the associated genes to assess whether they are differentially regulated in the phenotype of interest. Using functional modules, our methodology can generate the required pre-defined protein sets. This extension is especially powerful for the investigation of disease phenotypes where implicated cellular functions are poorly-studied.

The purpose of the developed pipeline is to find differentially regulated modules. These modules may be similar to active subnetworks in PINs (eg. [40–42]) or other disease modules [13, 44] (cf. Section 1.1.2). The difference between these approaches and ours lies in the definition of functional modules. As discussed in Section 1.1.2 methods that cluster PIN nodes based on differential expression represent active regions of PINs related to disease, rather than modules with distinct cellular functions. Thus, a clear functional interpretation of differential gene expression studies may not be straightforward using these approaches.

1.2 Biological applications

Our pipeline (Figure 1.1) was applied to find differentially regulated functional modules in two biological applications. Here, we give a short overview of the biological phenotypes to which this methodology was applied.

1.2.1 Breast cancer hypoxia

Cancer is a disease that is caused by somatic mutations in cellular DNA that lead to quickly proliferating and mutating cells. Aggressive cancers crowd out normal

cells in a tissue due to a strong competitive advantage. These cancers can spread to parts of the body away from the primary tumour site through blood vessels, a process which is known as metastasis. Mortality rates increase when cancers become metastatic. Hypoxia in cancers has been linked with increased aggressiveness, increased rate of metastasis, and an increased mortality rate [53].

Hypoxia is a low-oxygen cellular microenvironment, which is commonly delimited at an oxygen diffusion pressure of 10 mm Hg [54]. Such an environment can arise due to distance from blood vessels, diffusive barriers between the tumour and the blood vessel, or an irregular supply of oxygen from tumour blood vessels [53, 54]. It has been suggested that all invasive tumours grow under hypoxic conditions [53].

To cope with the low oxygen environment, cells switch from aerobic respiration to anaerobic respiration. Anaerobic respiration is a considerably less-efficient form of metabolism (2 versus 36 ATP molecules per glucose) that does not require oxygen to break down glucose into ATP, the molecular energy source of the cell. This form of metabolism creates an acidic extracellular environment via the release of lactic acid. A cellular environment that is low in vital resources, and is acidic, gives rise to a mutation pressure which selects for cancer cells with continuous anaerobic respiration that are acid resistant [53]. In contrast, normal cells which cannot quickly adapt to such an environment die. Thus, hypoxic tumours are more invasive than normal (normoxic) tumours by performing different functions (anaerobic respiration, recruiting stem cells for metastasis, ion transport to counteract acidic environment).

On a molecular level, a hypoxic extracellular environment causes changes in gene expression that lead to the observed phenotypes. The main mechanism by which these phenotypes arise is through the deregulation of one of the two subunits of the Hypoxia-inducible Factor (HIF) complex in the presence of oxygen [54]. The genes that express this subunit are down-regulated by the presence of oxygen, but in low-oxygen conditions the HIF complex can form. This complex has been shown to target over 1500 proteins contolling metastasis, angiogenesis (the formation of blood vessels), and other phenotypes observed in aggressive tumours [54].

While many of the differentially regulated functions in hypoxic and normoxic tumours are known, especially in well-studied breast cancer, understanding the molecular mechanisms underlying these functions is an ongoing field of research. There are drugs targeting two of the three commonly diagnosed breast cancer subtypes. The third subtype, known as triple-negative, is more difficult to target possibly due to hypoxic conditions creating drug-resistance in tumours. Drugs targeting HIF have shown promising results in mouse models and are currently in development [54]. Due to the large number of differentially regulated functions in hypoxic and normoxic conditions, breast cancer hypoxia is an ideal system to evaluate our pipeline.

Using our methodology on gene expression data from a breast cancer cell line cultivated under hypoxic and normoxic conditions, we have found differentially regulated modules that describe several functions that have been reported to be linked specifically with hypoxic tumours. For example, we found modules related to metastatic activity [55], initiation of transcription (see HIF promoting transcription of over 1500 genes), deacidification of the intracellular environment, and autophagy, a process that promotes cell survival under environmental stresses [55, 56].

1.2.2 Macrophage classification

Macrophages are cells of the mononuclear phagocyte system that form part of the innate immune system. Whilst a primary activity is phagocytosis, they are some of the most heterogeneous cells in the human body, where specific function, and thus phenotype, is dependent upon their particular tissue or organ localization and local microenvironment. They play key roles in microbial defence, tissue homeostasis and remodelling, wound repair, and initiation of immune response [1, 57]. This plasticity and phenotypic diversity has led to the recognition that macrophages can function pathogenically and are associated with a multitude of diseases, including fibrosis, cancer, obesity, and various inflammatory diseases, making them interesting therapeutic targets [58].

To grasp the functional diversity observed in macrophages, phenotypical classifications are used to characterize macrophages sub-types. Macrophages have been broadly classified into M1 and M2 phenotypes based on their response to Interferon-γ (IFN-γ), or the cytokines Interleukin-4 (IL-4) and Interleukin-13 (IL-13) respectively [1]. A further prominent distinction between macrophage sub-types are the expression levels of IL-10 and IL-12 [59]. M1 macrophages are characterized by low IL-10 and high IL-12 expression levels, while M2 macrophages exhibit the opposite profile of these cytokines. Further characterization has shown that M1 macrophages secrete pro-inflammatory cytokines and have increased phagocytic activity, while M2 macrophages tend to function in wound healing and clearing parasites. Furthermore, M1 macrophages have been found to be tumouricidal, while tumour associated macrophages resemble an M2-activation state, inhibiting the immune response to the tumour and even aiding tumour progression. Thus, being able to affect the differentiation of macrophages into the observed phenotypes has the potential to enhance cancer treatments. More comprehensive reviews can be found in [59] and [60].

Since the initial M1/M2 classification, the means to phenotype and characterize cells have become more sophisticated with gene microarrays, one of the tools frequently used in this context. Data generated from such studies together with the additional *in vitro* and *in vivo* model research has led to the suggestion that the intial phenotype categories may simply be extreme cases on a continuous scale [1,59] or even that this linear classification system imposes limitations which may not be able to explain macrophage diversity adequately [60]. However, even if this should be the case, the distinct phenotypes observed in disease association strongly suggest that different pathways are activated in the two initial macrophage polarizations. This aspect is of interest irrespective of the broader classification question.

These pathways are chains of molecular interactions within a cell, resulting in the cell performing a function, such as inhibiting tumour growth by phagocytosis. A biological pathway is thus an underlying cellular mechanism causing the observed phenotype if the proteins in the pathway are sufficiently abundant that the function

can be performed (activated pathway). It is ultimately these pathways that are targeted by drugs that may affect properties such as macrophage phenotype.

Developing the above classification system, especially regarding disease association, has proven difficult as macrophage experiments are generally performed on macrophages artificially created *in vitro* from monocytes (by exposure to IFN-γ for M1 macrophages, and IL-4 or IL-13 for M2 macrophages), which can be easily isolated from blood [60]. In contrast *in vivo* macrophages are typically organized as a type of tissue within a tissue. They are interwoven to such an extent that it is very difficult to extract *in vivo* macrophages [1]. It is also due to this complication that, despite the increased insight that has been gained regarding macrophage function, knowledge is still limited with respect to molecular pathways that are activated to cause the observed macrophage phenotypes.

In [61], differential gene expression analysis has been applied to this classification problem. DEGs were clustered based on their expression patterns and these clusters were investigated at different phases of differentiation from monocytes for overrepresentation of Gene Ontology annotations (see Section 2.4) which indicate the function of the gene products. The results showed that expression of cell cycle genes was different between M1 and M2 macrophages. However, inferring experimentally testable hypotheses from these conclusions is difficult as connections between DEGs and the associated phenotype remain vague, especially for biological functions that are less well captured in these annotations. Functional context between the discovered DEGs in the form of pathway data or dysregulated biological modules may provide new insights and has the potential to clarify the classification problem.

We have applied our methodology to the data set from [61] to recover functional modules of proteins that are differentially abundant between M1 and M2 macrophages. The detected differentially regulated modules describe phagocytosis, inflammation, differential antigen presentation, and chemotaxis (see Chapter 7), which are known to be differentially associated with the M1 and M2 activation state [62–64]. Less studied differences between M1 and M2 macrophages concerning the epigenetic regulation of transcription were also found (see Section A.6).

1.3 Thesis outline

In this thesis we describe the development of the methodology introduced in Section 1.1.3 (Figure 1.1) and its application to the biological questions described in Section 1.2.

Chapter 2 is devoted to describing the data types and introducing concepts usedin the remainder of the thesis. The data sources and how they are processed is elaborated on in Chapter 3.

Chapters 4-7 focus on the selection and development of the computational tools shown in the four blue boxes in Figure 1.1 respectively.

In Chapter 4 we present an exploratory analysis of functional homogeneity evaluation ("Semantic similarity measures" tool in Figure 1.1), and an exploratory analysis of community detection methods for the topological clustering component of the module detection pipeline is presented in Chapter 5 ("Multi-resolution community detection" tools from Figure 1.1).

Based on the results shown in Chapters 4 and 5 we discovered that the detection of functional modules was complicated by biases in the functional annotation data especially affecting poorly-studied proteins. Thus, we developed a framework for biological clustering that counteracts this bias. This framework, CommWalker, is introduced in Chapter 6 ("Community Evaluation" tool; see Figure 1.1).

Using our functional module detection method optimized as discussed in Chapters 4–6, we applied our methodology to two biological problems: macrophage differentiation, and hypoxia in breast cancer tumours (Chapter 7; "Differentially Expressed Gene Module Enrichment" tool in Figure 1.1).

Finally, the conclusions drawn from this project and work that leads on from our results is detailed in Chapter 8.

The Appendices contain unfinished work a novel topological clustering method (Appedix A), and additional information on data sets and supporting information for the CommWalker framework (Appendices B and D respectively).

2 Background

Contents

2.1 Protein-protein interactions 16
2.1.1 Measuring protein-protein interactions 16
2.1.2 Data quality 19
2.1.3 Protein-protein interaction databases 21
2.1.4 Network description of protein-protein interaction data 24
2.2 Networks 25
2.2.1 Network summary statistics 26
2.2.2 Network structure 28
2.3 Communities in networks 29
2.3.1 Resolution 31
2.3.2 Non-overlapping community detection 31
2.3.3 Overlapping community detection 37
2.3.4 Comparing community detection methods 43
2.4 Functional annotations 44
2.4.1 Ontologies and ontological concepts 45
2.4.2 The Gene Ontology 46
2.4.3 Other functional annotation resources 49
2.5 Functional homogeneity 51
2.5.1 Functional enrichment 51
2.5.2 Functional similarity 53
2.6 Gene expression data 58
2.6.1 Co-expression analysis 59
2.6.2 Differential gene expression 60
2.7 Summary 61

In this chapter we introduce the data sources and explain the concepts that

are used in the remainder of this thesis.

We start by describing protein-protein interactions, the data source underlying our topological clustering approach, and then go on to elaborate on the methods used to analyse these data. The biological clustering approach we implemented is based on functional annotations. These are described in Section 2.4, before mathematical methods of using these data are explained in Section 2.5. Finally, we briefly describe gene expression data and how we used these data in Section 2.6.

2.1 Protein-protein interactions

Protein-protein interactions are the basis of most biological processes, and have been shown to be perturbed in disease [65, 66]. Thus, they represent a key component of how functions are mediated at a molecular level.

As discussed in Section 1.1.3 our methodology relies on functional modules to provide context between differentially expressed genes. These modules are commonly detected based on protein-protein interactions (see Section 1.1.2). Here we describe this data type.

2.1.1 Measuring protein-protein interactions

Protein-protein interactions can be divided into two categories which are defined by how they are measured [67, 68]. These two types are interchangeably referred to as binary and complex [68, 69], or "physical associations" (P-type) and "associations" (A-type) [48,70]. The two main experimental techniques that define these interaction types are yeast two-hybrid [15, 19, 20] for P-type, and tandem affinity purification with mass spectrometry (TAP-MS) [18, 71, 72] for A-type interactions.

The yeast two-hybrid approach is built on transcription factors that consist of two domains: a DNA-binding domain and a domain that activates transciption. So-called "bait" proteins are fused to the binding domain and introduced into a yeast cell nucleus where they are exposed to a "prey" protein which is similarly fused to the activation domain. If the two proteins interact then the activation domain activates the transcription of a reporter gene which is adjacent to the binding site

of the DNA-binding domain hybrid protein. Using this method protein-protein interactions are tested in a one-by-one manner (binary). A strength of this technique is that it can also identify transient interactions where proteins are sufficiently close due to electrostatic attraction (physical association).

In contrast, TAP-MS detects interactions based on biochemical binding of proteins, which leads to a more stable interaction than electrostatic attraction alone. In this experimental technique "bait" proteins are fused to a TAP tag, which binds to an affinity column. The "bait" protein is expressed in a yeast cell and exposed to other proteins with which it can form complexes. After washing contaminants off the complexes while bound to the affinity column, the proteins are cleaved off the column and identified by mass spectrometry [73]. TAP-MS thus identifies several protein-protein interactions (complex) in a single experiment, which must bind strongly to each other to be detected (association). Interactions within a detected complex can be reported in two ways: via the spoke model, and via the matrix model [74]. While the matrix model reports all complex proteins as interacting with each other, the more common spoke model only reports interactions with the "bait" protein [75]. Selection of the reporting model represents a choice between coverage and quality of reported interaction data.

Yeast two-hybrid and TAP-MS have contrasting limitations and benefits. While yeast two-hybrid can detect transient interactions, it does so by assessing potential interactions which are not necessarily found in a physiological setting. In contrast, TAP-MS focuses on biochemical binding at the cost of transient interactions [67]. Furthermore, yeast two-hybrid interactions are detected in a soluble state in the nucleus, which is not necessarily the natural compartment for proteins and leads to a selection against insoluble proteins (however, two-hybrid type methods have been developed to address this limitation [15]). In contrast, TAP-MS has been shown to select for interactions between abundant proteins, while high-throughput yeast two-hybrid data sets report a similar number of interactions for proteins of different abundances [67]. Due to these individual limitations and advantages

the yeast two-hybrid and TAP-MS methods tend to detect interactions associated with different biological functions [67].

A limitation that is common to both experimental techniques, and indeed to most other techniques of determining protein-protein interactions, is that alternative splicing is not taken into account. Genes can code for more than one protein by different combinations of protein coding DNA regions (exons). These different proteins from the same gene are called isoforms and it is estimated that at least 47% of genes exhibit alternative splicing [76]. Protein-protein interactions tend to only be reported for the most common (dominant) isoform, due to the difficulty of introducing alternative splicing into the experimental techniques [77]. It has been shown that more interactions tend to be found for proteins which have isoforms [78], suggesting that other isoforms may contribute to the interaction measurements. While isoform-resolved interaction data sets have recently been acquired which address this limitation (eg. [77, 78]), the majority of interaction data is reported by simplifying the gene-to-protein relationship to a one-to-one relationship. Whether or not this simplified mapping occurs is evident from the identifiers used for proteins when protein interactions are reported. Databases that report interactions between Entrez Gene IDs use this simplified mapping, while UniProt identifiers (UniProt Knowledge Base Accession Numbers) allow for isoform-level interaction mapping (see Section 2.1.3 for protein interaction databases).

Here we have focused on yeast two-hybrid and TAP-MS to measure protein-protein interactions. These two techniques of measuring protein-protein interactions have been used to generate large-scale maps of the human interactome (eg. [79–81]). Other techniques such as Fluorescence Resonance Energy Transfer (FRET), co-immunoprecipitation, or Protein Fragment Complementation Assays (PCA) have also been used to detect protein-protein interactions, although often on a smaller scale [82]. Interactions detected by these techniques can be assigned to A-type and P-type categories as shown in Appedix B. These methods, yeast two-hybrid, TAP-MS, and others are reviewed in [67, 68, 75, 83, 84].

The so-called interactome that is measured by the above techniques represents a set of possible interactions which is limited by experimental methods. The proteins that are reported as interacting are not necessarily present in the same cell at the same time given the experimental set-up. Thus, it has been argued that there is no single human interactome, but instead a dynamically changing environment of protein-protein interactions which is not captured in the static reporting of interactions [85]. This issue is starting to be addressed by dynamic interaction measurements incorporating temporal gene expression data (eg. [86]; see Section 2.6 for gene expression data).

2.1.2 Data quality

2.1.2.1 Error rates and coverage

As mentioned in Section 2.1.1 experimental techniques for measuring protein-protein interactions are subject to limitations. These limitations and other sources of error give rise to interactions that are misreported as existing (false-positives), and those that are misreported as not existing (false-negatives). For example, in yeast two-hybrid assays proteins may misfold due to fusion with transcription factor or binding domains, leading to unspecific false-positive interactions. Similarly, in TAP-MS, TAP-tagging may cause misfolding, or interactions may be reported with proteins that do not interact with the "bait" protein, but instead with another "prey" that is in the complex. False-negative interactions may similarly arise from misfolding of fused or tagged proteins, or from the lack of post-translational modifications for human proteins in yeast. Further sources of false-positive and false-negative interactions are reviewed in [20, 67, 68, 87].

The false-positive error rates of these data sets have been estimated at 35 - 70% for yeast two-hybrid, and 35% for TAP-MS data [83, 88]. The high yeast two-hybrid error estimates have been disputed based on a more detailed statistical analysis that gives values of 17% and 21% for two high-throughput yeast two-hybrid data sets [89]. Comparatively lower error rates for yeast two-hybrid assays have been supported by a study on the yeast interactome [90]. In comparison, false-positive rates for

manual literature curation of experimental evidence from small-scale studies have been estimated to be between 2 and 9% [91], or even as high as 35% [82] by different studies. Generally it is assumed that literature curated protein-protein interactions represent the closest data set to a gold standard that is available [92].

False-negative error rates are more difficult to assess as these require data on the protein interaction space that was investigated by each data set, which is generally not reported. Independent estimates on yeast and human data have suggested that 90% of interactions are falsely reported as not present in any single high-throughput experiment [88, 90].

Due to the large variation in error rate estimates, the estimated size of the human interactome varies similarly. The total number of estimated interactions has been quoted as between 154,000 and 650,000 [88, 93], and it appears to increase with the availability of more data [90]. These esimates do not take into account protein isoforms or differently post-translationally modified proteins, which would increase the size of the system considerably [13]. Given these size estimates, the largest currently available human interactome data sets have a coverage of between 179% (for example due to oversaturation with false-positive interactions) and 42% (see Table 2.1 for interactome sizes).

2.1.2.2 Quality control

Due to the high error rates estimated for protein-protein interaction data sets (see Section 2.1.2.1), methods to improve the quality of these interaction data sets have been developed. The two most common quality control approaches are manual literature curation (eg. [91]) and multiple reporting (eg. [82, 88]).

In Section 2.1.2.1 we mentioned that manual literature curation is regarded as the gold-standard for protein-protein interaction quality control. The "depth" of this curation refers to the detail with which a protein-protein interaction is reported. While deep curation efforts produce high-quality interaction data, they involve curators carefully reading publications and are therefore slow.

In contrast, filtering protein-protein interaction data sets for interactions that have been reported by at least two independent studies is a quick computational method for quality control. This filtering has been shown to improve false-positive rates for interactions from 45% to 8.5% in a small sample of published data [82]. Indeed, this method is commonly used to obtain high-quality interaction data sets intended to represent ground-truth interactions (eg. [82, 89, 94]).

Other methods of improving the quality of interaction data sets revolve around assigning confidence scores to interactions and filtering based on these scores (eg. [95–97]).

The implementation of quality control measures represents a compromise between optimizing false-positive and false-negative rates. Increasing the false-negative rate has been shown to introduce biases into the data sets. Literature curation of small-scale studies creates data sets of proteins that have many interactions due to them being well-studied (inspection bias) [74, 79, 98]. Furthermore these data sets tend to include proteins that are more functionally similar than would be expected [99]. It has been suggested that this may be due to non-reporting of interactions that do not fit the biological hypothesis investigated in a study. Filtering interactions by multiple reporting has been argued to be subject to similar biases [74]. Due to the low coverage of protein space currently experimentally tested, interactions between well-studied proteins are selected for even in high-throughput data sets [74, 89].

2.1.3 Protein-protein interaction databases

While protein-protein interactions are measured using a variety of experimental techniques under different conditions, it is important to combine them as no individual study is able to give sufficient coverage of protein space [12]. Such combinations of protein-protein interaction data sets are stored in databases. The databases differ in their depth of curation, in the extent of the curation efforts (coverage), and in further implemented quality control measures. Due to the large number of databases for protein-protein interaction data which are continuously increasing, a comprehensive review is not feasible. Several important databases

are reviewed in [68, 75, 84]. Here we give a small overview of databases that store experimentally determined physical protein-protein interactions such as the A-type and P-type interactions discussed in Section 2.1.1.

The main protein-protein interaction databases are BioGrid [100,101], IntAct [102], HPRD [103, 104], and STRING [105].

BioGrid is a large interaction database (see Table 3.1) which allows for public submission of small-scale interaction data and submission of large scale data sets with little curation [101]. While yeast data is fully manually curated, human data quality is addressed by random manual re-curation of small interaction samples and themed curation efforts to increase the coverage of interaction data for specific applications. BioGrid stores both physical protein-protein interactions as described in Section 2.1.1, and genetic interactions based on evidence from gene knockout experiments. As most experimental techniques focus on detecting protein-protein interactions for only the dominant isoform (see Section 2.1.1), BioGrid reports proteins by Entrez Gene IDs simplifying the relationship between genes and proteins to one-to-one.

IntAct is another large interaction database, that focuses on high coverage. It has recently merged with the well-curated MINT database and curators from both teams are now focusing their curation efforts on specific datasets similar to BioGrid [102, 106,107]. As a member of the iMex consortium (of which BioGrid is an observer) it is a part of a large, deep protein-protein interaction curation effort to curate the entirety of protein-protein interaction publications [108]. Systematic re-curation of older database entries will gradually improve the data quality in IntAct. Interaction data from all iMex Consortium members have recently become available via the online tool, PSIQUIC. While this tool allows for simple queries, large scale downloads require individual downloads from all databases. Furthermore, distinguishing between inferred and experimentally determined interactions is currently difficult as non-iMex members can also be queried with this tool [108]. In IntAct, proteins are reported via UniProt protein identifiers, thus allowing for reporting of isoform-resolved protein-protein interactions (see Section 2.1.1).

Although the Human Protein Reference Database (HPRD) [103,104] has not been updated since 2010, it is still a commonly used source of protein-protein interaction data. This database was initially compiled from manually curated small-scale protein-protein interaction publications around *in vivo* experiments with little focus on high-throughput studies. Similar to BioGrid and IntAct, submission of interaction data via an online tool has increased the database size drastically from its initial size [104].

The STRING database [105] integrates interactions reported in other databases that perform primary literature curation, and includes other data sets which can be used to infer "functional linkages" between proteins (eg. [109]). This database assigns confidence scores to protein-protein interactions based on experimental evidence as well as other supporting data in a non-transparent fashion. Interactions can be thresholded using these scores to allow the user to decide on the compromise between quality and coverage of interactions. However, performing simple thresholding may not be the best way to use these confidence scores [110]. Furthermore, STRING is designed as a query interface for small interaction data sets. Obtaining the entire interactome requires a license.

A further protein-protein interaction database which is used in this project is the HINT database [69]. HINT is an interaction database that combines interaction data from several other databases including IntAct and BioGrid and implements rigorous quality control based on multiple reporting and manual curation (see Seciton 2.1.2.2). Interactions from small-scale studies are only included in HINT if they are reported by at least two publications and high-throughput studies are manually curated. Proteins are reported as gene identifiers as in BioGrid, and made available split into high-throughput and literature curated data sets, or A-type and P-type interactions (see Section 2.1.1).

To get an idea of the differences in coverage of the described databases the protein-protein interaction data set sizes retrieved at the time of writing (16th August 2016) mapped to Entrez gene IDs are shown in Table 2.1.

Table 2.1: Protein-protein interaction database Human data set sizes.

Network	Proteins	Interactions
BioGrid	20,899	275,501
IntAct	13,726	103,199
HPRD	9,270	36,918
STRING	20,457	-
HINT	12,227	73,317

Table 2.1: The data sets were obtained on the 16th of August 2016. The number of edges for the STRING database could not easily be obtained due to licensing issues, and it is unclear whether the number of nodes quoted represents different genes or includes isoforms. The HPRD data is equivalent to that in Table 3.1 (retrieved February 2014) due to its lack of recent updates.

2.1.4 Network description of protein-protein interaction data

Protein-protein interaction data are often represented as a network (Figure 2.1). In the network representation, nodes represent proteins (or dominant protein isoforms denoted by gene identifiers, see Section 2.1.1), and edges between nodes denote interactions reported between these proteins. While the edges in this network are commonly unweighted and undirected, edge weights can be added for example as confidence scores (see the STRING database in Section 2.1.3), or to represent protein similarities (see Section 1.1.2). Self-interactions and multiple edges are generally filtered out.

The network that remains after these processing steps is called a protein interaction network (PIN). In this PIN it may not be possible to reach every protein from every other one only by tracing the edges (it may not be connected). Connectedness is however a desirable quality for methods such as community detection that partition the network into smaller substructures (see Section 2.3). For this purpose the largest connected component of a PIN is commonly used. Using the largest connected component is important in this project as its use guarantees that there are connections between any two differentially expressed genes overlaid onto the network (see Section 1.1.3).

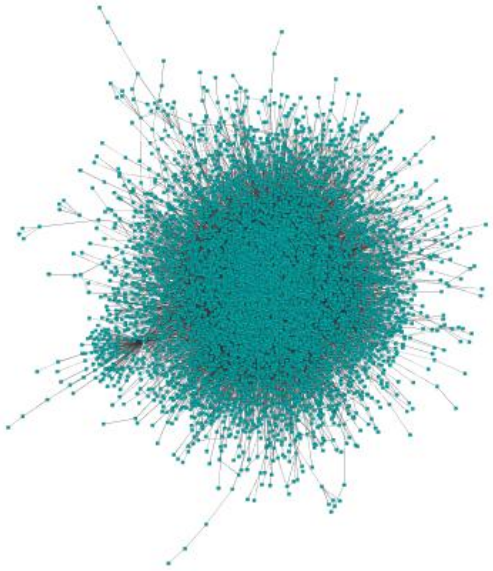

Figure 2.1: Network representation of a PIN from the HINT database Diagram of the largest connected component of HINT-P-14 [69] (Data set retrieved May 2013 and processed as described in Section 3.1.1). The network contains 7,869 nodes and 24,375 edges. Nodes represent proteins and edges are interactions between proteins. All edges are undirected and multiple edges and self-loops are filtered out. The network is too complex to detect substructures visually.

2.2 Networks

Networks, or graphs, are sets of nodes which are connected by edges. They are commonly referred to using the notation $G(V, E)$, where V represents the node set, E the edge set, and G the graph or network. Mathematically, a graph can be described via its *adjacency matrix* $A = A_{ij}$ for $i, j \in V$. For a graph of N nodes, A is an $N \times N$ matrix which has an entry of 1 if there is an edge between nodes i and j, and 0 otherwise. In the case of a network of undirected interactions (eg. PINs in Section 2.1.4) the respective adjacency matrix A is symmetric. Here, we refer to a graph as a mathematical object, and a network as a representation of a data set such as protein-protein interactions.

Recently, the concept of a graph has been extended to multilayer graphs. Multilayer graphs contain multiple layers of standard graph representations with interlayer edges connecting nodes [111]. In these graphs adjacency tensors or supra-adjacency matrices replace the described adjacency matrix. While such a graph representation has been used to integrate biological data sets for functional

module detection [50], we focus on single layer graphs due to the availability of a vast array of analysis methods.

In this section we give a short introduction to the network/graph terminology used to describe PINs. A more comprehensive review of network theory can be found in [112].

2.2.1 Network summary statistics

Networks can contain millions of nodes and billions of edges (eg. [113]). To generate insight from such complex networks it is necessary to describe the networks in a simple, summarized manner. Network summary statistics are used for this purpose.

One of the simplest descriptions of a network is its *density*. The density of a network is the fraction of possible edges that are realized in the network. It is calculated by the equation:

$$\text{Density} = \frac{2|E|}{N(N-1)}, \tag{2.1}$$

where $N = |V|$ denotes the number of nodes, and $|E|$ the number of edges. While the density puts the number of edges into the context of the network size, it does not convey information about the network structure. Network descriptions that consider the network structure include the degree distribution, clustering coefficients, and shortest path length statistics, which are described as follows.

The nodes with which a node is connected are called its *neighbours*. The number of neighbours a node i has, is given by its *degree* k_i. The degree is calculated from the adjacency matrix A_{ij} by the equation:

$$k_i = \sum_j A_{ij}. \tag{2.2}$$

Using this concept networks can be described by the distribution of node degrees. The number of nodes with degree $k = K$ for all $K = 0, \ldots, |V| - 1$ is called the *degree distribution* of the network (Figure 2.2). Figure 2.2 shows a typical PIN degree distribution with a large proportion of nodes with a low

degree, and few nodes with a high degree. The degree distribution can in turn be summarized via its mean and variance.

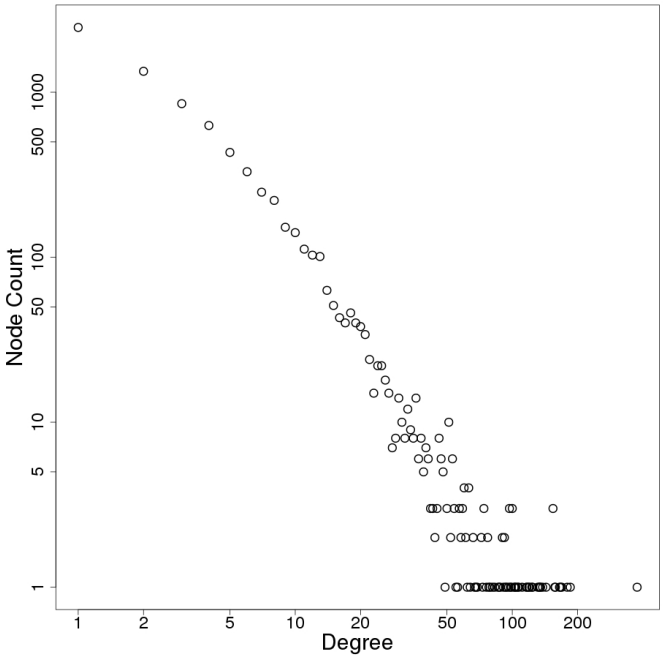

Figure 2.2: Degree distribution of HINT-P-14 The number of nodes with a given degree in a log-log plot for HINT-P-14 (see Section 3.1.1 for HINT-P-14). This is the degree distribution of the network shown in Figure 2.1. The plot depicts a characteristic heavy-tailed degree distribution for PINs showing there are many low degree nodes, but only few very high degree nodes. The highest degree node in HINT-P-14 is GRB2 (degree 376).

How closely nodes group together in a network can be assessed via global and local clustering coefficients. Similar to the density, these statistics assess the proportion of triangles that are realized in the network. The *global clustering coefficient* C_{global} is calculated by:

$$C_{global} = \frac{3 \times \text{Number of Triangles}}{\text{Number of Connected Node Triples}}. \tag{2.3}$$

In contrast, the *local clustering coefficient* is calculated for every node individually. It is defined as the fraction of a node's neighbours which are each others neighbours. The local clustering coefficient thereby evaluates how close to a clique, a complete subgraph, the neighbours of a node are. Taking the average of the local clustering coefficients in a network gives a measure of clustering which is different from the global clustering coefficient.

A *path* in a graph is a trajectory between connected nodes. The *length* of this path is the number of edges traversed in the trajectory. Thus, a *shortest path* between two nodes is a path between these nodes that traverses the fewest edges. Summary statistics based on shortest paths include the radius and the diameter of a network. These statistics are defined via node eccentricities. The *eccentricity* of node i is the maximum shortest path length between node i and any other node in the network. In turn, the *radius* and *diameter* of a network respectively denote the minimum and maximum node eccentricities.

A more comprehensive overview of network summary statistics can be found in [112].

2.2.2 Network structure

As mentioned in Section 2.2.1, network summary statistics can give insight into the structure of a network. For example, networks with comparatively small shortest path lengths and high clustering coefficients are thought to exhibit the small world property [114]. This property suggests that the network exhibits locally dense substructures with few long-range edges connecting these clusters. Likewise, networks with a degree distribution that resembles a power law have been modeled as emerging from a process where the network grows by new nodes preferentially attaching to other nodes with a high degree [115]. While PINs have been described as scale-free alluding to a power law degree distribution [13, 116], this conclusion is not sufficiently supported by the data [117–119]. Indeed common network models have not been able to reproduce the structure observed in PINs [119–121].

In this project meso-scale network structure is of particular interest. Meso-scale network structure describes patterns in network structure that exist between the global level of summary statistics and the level of nodes and edges [122]. This level of structure helps to break down complex networks into smaller chunks that can be separately analysed. Two examples of meso-scale structure are *community structure* and *core-periphery structure*. Core-periphery structure consists of a dense network core and a sparsely connected network periphery. Peripheral nodes tend

to be connected to the community core, but not to each other [122]. Community structure, which is ubiquitous in PINs (see Section 1.1.2), describes local modular clusters of nodes in networks. These two types of meso-scale structure are often found in combination in real-world networks [122, 123].

The structure of a network can be sampled using methods such as ego-networks (eg. [124]) or random walks (see Chapter 6). *Ego-networks* are subnetworks that are constructed by taking a node i, the neighbours of this node, and all edges between the nodes in this subset. Ego-networks can be extended past the immediate neighbourhood by including all nodes that are neighbours of the nodes in the initial ego-network and all edges between the nodes in the extended subset. The extension of an ego-network is denoted by the number of hops. A one-hop ego network includes only the immediate neighbours of a node, while a two-hop ego-network also includes indirect neighbours (neighbours of neighbours).

Random walks are dynamic processes on a network that trace network paths. In its most basic form a random walk is a process that starts at a node i, and moves to a node j chosen uniformly at random from the neighbours of node i. This process is repeated until a stop criterion is reached. Such a stop criterion can be determined based on the length of the random walk. Unless otherwise stated, in this thesis we generally refer to the *length of a random walk* as the number of unique nodes that have been traversed, including the start node. Other definitions focus on the number of edges (see also path length in Section 2.2.1), or allow for multiple counting of nodes that are traversed more than once.

2.3 Communities in networks

Community detecton methods attempt to uncover the modular structure of networks. By finding groups of nodes which interact more with each other than with the rest of the network, networks can be partitioned into connected substructures called *communities*. This partitioning is performed based on the topology of the network and edge weights.

The way in which a network is partitioned into communities depends on the definition of the term "community" (Figure 2.3). For example, communities can be defined as densely connected groups of nodes where each node is assigned to a single group with few edges between groups (non-overlapping communities). Alternatively, communities may overlap. The nature of this overlap can be defined differently between methods (Figure 2.3). The community definition used by a method is often incorporated into an objective function which can be optimized with respect to the network partition to detect communities.

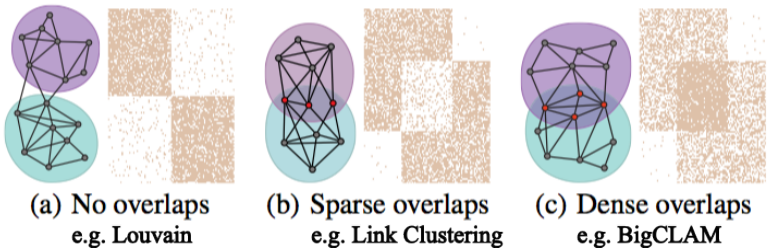

Figure 2.3: Types of community detection methods. The left-hand side of each of the three images represents a network view and the right-hand side the corresponding adjacency matrix. The darker the regions in the adjacency matrix are, the denser the corresponding network regions. a) The standard view of non-overlapping community detection as dense communities with less dense inter-community regions. b) Less dense community overlaps between dense clusters by eg. edge-based methods. c) Dense community overlaps between communities. (After [125] ©2017 Association for Computing Machinery, Inc. Reprinted by permission.)

As PINs are thought to have a modular structure that reflects the underlying molecular organization of biological functions (see Section 1.1.1), community detection methods are used to extract functional modules (see Section 1.1.2). In this section we introduce the three approaches to the community detection problem that were applied to PINs for functional module detection in this project. The selected approaches include non-overlapping methods (Modularity Maximization using the Louvain algorithm [126]), and overlapping community detection methods (link clustering [127] and BigCLAM [125]). The methods presented here were chosen as they represent characteristically different approaches to the community detection problem (see Figure 2.3) and have previously been applied to PINs [48, 125, 127]. A further selection criterion was speed. The methods used easily scale up to commonly available PIN sizes and are thus feasible for community detection at multiple scales (see Section 2.3.1). More extensive reviews can be found in [128–130].

2.3.1 Resolution

Functional modules can be found at multiple scales (see Section 1.1.3). In community detection, this notion of "scale" is called *resolution*. In the same way that a village is part of a county, which in turn is part of a country, communities exist at different resolutions (Figure 2.4). To detect hierarchical community structure, *resolution parameters* are incorporated into community detection methods.

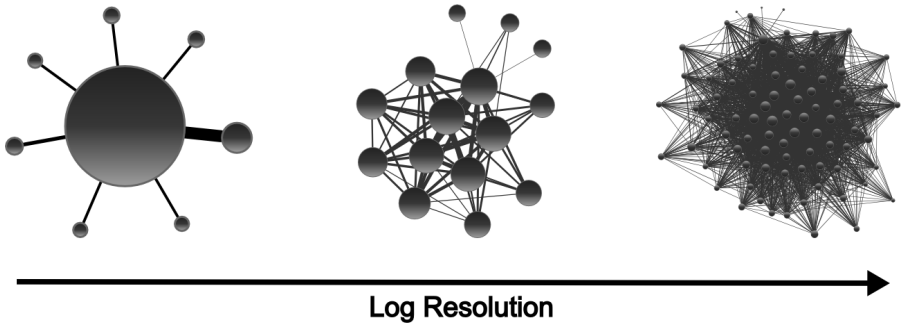

Figure 2.4: The effect of a resolution parameter in community detection The circles represent communities and the lines denote edges between them. The size of a circle, or a line, is proportional to the community size, or the number of edges between two communities, on a logarithmic scale. The three graph representations show network partitions generated at different resolutions by the configuration model Louvain community detection method implemented on HINT-P-14 (see community detection method in Section 2.3.2 and HINT-P-14 in Section 3.1.1). Different community decompositions are generated at different resolutions.

As shown in Figure 2.4, increasing the resolution parameter has a "zooming-in" effect on the network partition. At low resolution the entire network is partitioned into a single community, while all nodes are in their own communities at high resolution. In between these extremes, the network is partitioned into smaller communities as the resolution is increased.

In order to detect functional modules at different scales, the community detection methods used in this project either incorporate a resolution parameter, or include a parameter that may be used as a proxy for resolution.

2.3.2 Non-overlapping community detection

The traditional image of communities in networks are those of sparsely interconnected dense clusters shown in Figure 2.3(a) [128, 129]. While statistical methods of detecting such communities based around the stochastic block model [131] have

been developed since the 1980s, in 2002 community detection received renewed attention with the publication of a seminal computational technique, the Newman-Girvan algorithm. This algorithm iteratively removes edges from the network, based on the number of shortest paths between all nodes that traverse these edges, to leave disconnected subnetworks that represent communities [132] (see Section 2.2.1 for shortest paths).

The increased research interest into community detection methods following this publication led the authors to introduce the concept of Modularity to evaluate the quality of the detected community structure [133]. Modularity has since been used directly to detect communities using various optimization algorithms (eg. [27, 126, 134]). As Modularity Maximization has previously been used to detect functional modules in PINs [48], we focused on this method for non-overlapping functional module detection.

Modularity Q is a measure that describes the strength of the community structure in a given partition by evaluating how much more communities are connected than expected by a random background model [133]. It is given by the equation

$$Q = \sum_i (e_{ii} - a_i^2), \tag{2.4}$$

where e_{ii} denotes the fraction of edges in the network that are between nodes in community i, and a_i is the fraction of edges that connect to nodes in this community (Note $a_i = \sum_j e_{ij}$ where e_{ij} is the fraction of edges between communities i and j). By finding a network partition that maximizes this measure, Modularity can be used to partition networks (Modularity Maximization). The obtained partitions are both non-overlapping (a node is assigned to only one community) and complete (each node is partitioned into a community).

A multi-resolution adaptation of Modularity Maximization is the *Potts method* [135]. The objective function that is optimized in the Potts method (the energy function H_γ) is:

$$H_\gamma(\sigma_i, \ldots, \sigma_N) = \sum_{ij}(A_{ij} - \gamma p_{ij})\delta(\sigma_i, \sigma_j). \tag{2.5}$$

This equation is equivalent to Equation (2.4) for $\gamma = 1$ reframed in an integer linear programming approach [136]. In this equation, σ_i denotes the community membership of node i, A denotes the adjacency matrix, γ is the resolution parameter and p_{ij} denotes the probability of interaction between nodes i and j under a null model. As indicated by the Kronecker-Delta function $\delta(\sigma_i, \sigma_j)$, which is 1 if $\sigma_i = \sigma_j$ and 0 otherwise, the sum is taken over all nodes i and j which are in the same community. Thus, the interaction strength $A_{ij} - \gamma p_{ij}$ gives an indication of how well connected the two nodes i and j are, compared to what is expected under the null model p_{ij}.

The null, or background, model should thus summarize the expected structure of PINs. Pairs of nodes that are connected more than expected from this null model and the resolution factor γ increase the energy function H_γ, and should thus be assigned to the same community by the optimization algorithm. The lower the value of γ, the more likely it is that two nodes are assigned to the same community and larger communities will be created, and vice versa for high resolutions. This interplay of resolution parameter and null model shows that the choice of null model is central to the network partition obtained from multi-scale Modularity Maximization. In the absence of a null model that accurately describes the structure of PINs (see Section 2.2.2), two existing models are commonly used.

2.3.2.1 Configuration model

The null model implicit in the definition of Modularity (Equation (2.4)) is the Newman-Girvan null model, also often called the *configuration model* [133, 134]. This null model is given by the equation

$$p_{ij} = \frac{k_i k_j}{2m}, \quad \text{for } i \neq j, \tag{2.6}$$

where k_i denotes the degree of node i, and m is the total number of edges in the network as calculated by $\sum_{ij} A_{ij}/2$. Equation (2.4) can be recovered from Equation (2.5) with this null model, by including the factor $\frac{1}{2m}$ and using $\gamma = 1$ [112].

The Newman-Girvan null model is a commonly used background model (also for PINs [48]), which conserves the degree distribution of the network. This model evaluates the significance of an edge between two nodes based on the probability of there being an edge between two nodes of the given degrees under random rewiring. Although rare in sparse networks [112, pp.440-441], the model thus incorporates the possibility of self-loops and multiple edges, which are generally excluded in PINs (see Section 2.1.4). However, the conservation of the degree sequence of the network enforces similarity to the original network, thereby making the configuration model a plausible model of what an "expected" network would be.

As the configuration model uses node degrees to calculate expected interaction probabilities, network partitions by Modularity Maximization using this null model will be dominated by high-degree nodes (hubs). For example, in currently available PINs which contain over 100,000 edges, Potts method Modularity is maximized when two interacting degree two nodes are partitioned into the same community up to a resolution of $\gamma = 50,000$ (as $1 - (50,000)\frac{(2)(2)}{200,000} = 0$ by Equation (2.5)). This behaviour can lead to longer chains of degree two nodes in the same community. Including edges in a community negatively contributes to the overall Modularity only when the degrees of the interacting nodes are sufficiently high.

A further limitation of configuration model Modularity that was recently found is its implicit assumption that all communities are "statistically similar" [137]. By showing that network partitions using this method are equivalent to those found by a simplified stochastic block model (see Section 2.3.3.1), it was shown that Modularity assumes that the interaction probabilities within all communities in a network are the same after taking into account node degrees.

It should be noted that the term "configuration model" was initially used to refer to a generative network model in which the edges are randomly rewired between nodes with the degree distribution kept constant. In this model the degrees are

kept exactly constant for each realization, while the described Newman-Girvan null model uses probabilities based on which a graph can be generated with consistent expected degrees. These methods are regarded as near equivalent [129] and thus the term "configuration model" is used here to give credit to the earlier work [138].

2.3.2.2 Constant Potts model

Modularity Maximization can be subject to a resolution limit [139–142]. This resolution limit has the effect that large communities of random structure may be subdivided, and small communities compared to the network size may be merged with neighbouring ones, even in cases with clearly pre-defined modular structure [139]. The resolution limit thus introduces a source of error which is dependent on community size. While multi-resolution community detection can affect the community sizes at which this resolution limit affects the network partition, it cannot always circumvent the issue [141]. It has however been suggested that this phenomenon is dependent on the null model used, with the *Constant Potts Model (CPM)* not exhibiting this effect [140]. The CPM uses a complete graph as expectation which is given by the equation

$$p_{ij} = p = 1\,, \quad \text{for } i \neq j. \tag{2.7}$$

Using the CPM as null model for Modularity Maximization leads to a resolution parameter γ that controls the interaction probability between two nodes in the same way as a Bernoulli random graph (see Equation (2.5)). It can be shown that this null model directly controls the density of detected communities.

In Modularity Maximization nodes are assigned to the same community when this assignment increases the overall Modularity. By Equation (2.5) with the CPM (Equation (2.7)), a positive contribution to the overall Modularity is given when $A_{ij} - \gamma > 0$. As the Kronecker-delta function has the effect that the sum in Equation (2.5) goes over all node pairs in a community, a positive community contribution to Modularity can be written as:

$$|E_C| - \gamma \frac{N_C(N_C - 1)}{2} > 0.$$

Here N_C is the number of nodes in the community, $\frac{N_C(N_C-1)}{2}$ is the number of node pairs in the community, and $|E_C|$ is the number of edges between nodes in the community. This equation can be rearranged for γ to give

$$\gamma < \frac{2|E_C|}{N_C(N_C - 1)},$$

where the right hand side is equal to the density of the community (see also the density of a network in Equation (2.1)). The re-arranged equation shows that nodes are partitioned into the same community if the resulting community has a density greater than the resolution parameter γ. Thus, γ directly controls the density of communities detected by CPM Modularity Maximization. This density selection has the effect that node chains will not maximize the Modularity using the CPM in contrast to the configuration model, suggesting that the two null models select for different types of communities.

While the configuration model has previously been used to find functional modules in PINs [48], the validity of communities found using the CPM in the context of PINs requires further investigation. Furthermore, the absence of a resolution limit for Modularity Maximization with the CPM null model has been contested based on the analysis of simulated networks [141]. Recently, it has also been suggested that using constant null models such as the CPM in Modularity Maximization is equivalent to fitting a standard stochastic block model (see Section 2.3.3.1). This model assumes the network exhibits a Poisson degree distribution [137] which is not the case for PINs (see Figure 2.2).

2.3.2.3 Maximizing Modularity: the Louvain algorithm

Modularity Maximization is an NP-hard problem [136], which implies that finding a network partition that maximizes Modularity cannot be done both precisely and time-efficiently. Therefore, heuristic methods are used to find network partitions that approximately maximize Modularity.

The heuristic used to maximize the Modularity equation has an effect on the resulting network partitions. It has been argued that there can exist many structurally different network partitions that approximately maximize Modularity using the configuration model [142]. This effect is more pronounced, the more modular the network structure. Thus, a single network partition generated from a heuristic optimization of Modularity may not resemble the Modularity-optimal case. This effect is referred to as the *degeneracy problem* and can be addressed through the use of consensus partitioning using multiple community detection runs [50, 143]. As modular structure is arguably easier to detect the more modular the network is, a good optimization heuristic may also minimize the effect of the degeneracy problem.

Here, we use the Louvain algorithm to maximize Modularity [126] (Figure 2.5). This algorithm has been shown to outperform other optimization techniques for Modularity Maximization even without the use of consensus clustering [27, 144].

The Louvain algorithm performs Modularity Maximization by initially placing all nodes into separate communities. In the Modularity optimization step nodes are moved between communities by community mergers and separations that increase the Modularity of the partition. This process is performed with a random node ordering until no further node movements can increase the Modularity. In the subsequent community aggregation step, the communities are aggregated into nodes. Edges between communities become weighted edges between the new nodes so that the process can be repeated iteratively using a weighted adjacency matrix $W = w_{ij}A_{ij}$ with edge weights w_{ij}. These two steps are repeated until no node rearrangement step can be performed which would further maximize the Modularity of the partition [126].

2.3.3 Overlapping community detection

While Louvain Modularity Maximization detects non-overlapping communities, in real-world networks communities may overlap. For example, a protein can have multiple functions and thus be a member of multiple functional modules (see Section 4.2), or a person may have friends at work and at their sports club and

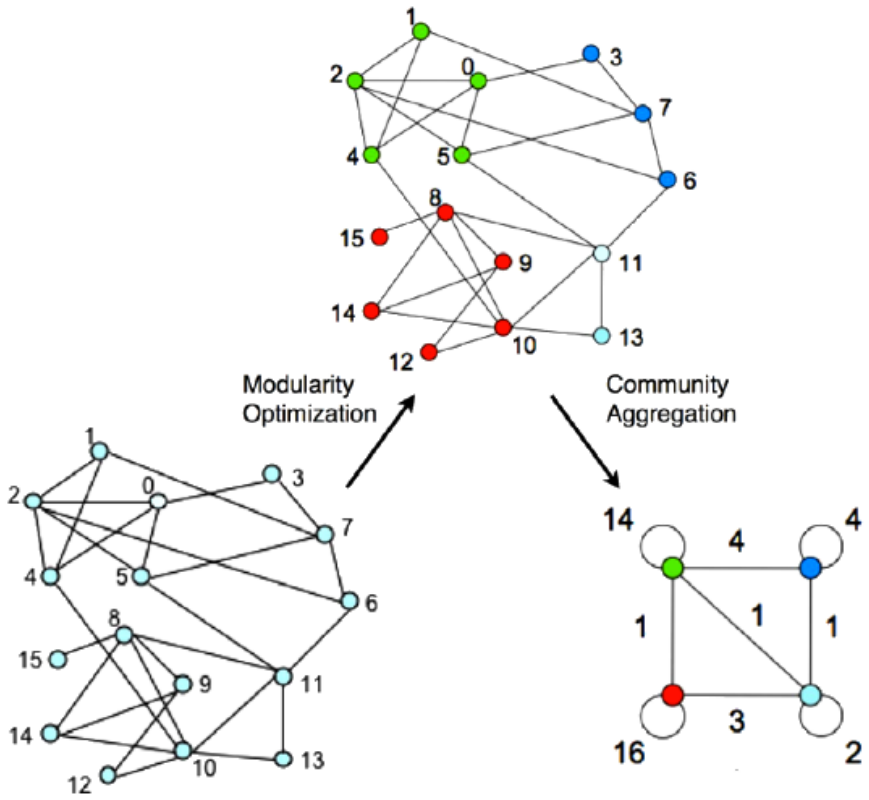

Figure 2.5: Schematic diagram of a single pass of the Louvain algorithm The Modularity optimization step rearranges single nodes into larger communities, and the community aggregation step subsequently aggregates these communities into individual nodes with weighted interactions. The displayed pass is repeated until no further community splits or mergers are beneficial. Node colour represents community affiliation, edge labels represent weights, and aggregate node labels represent numbers of intra-community edges. (After [126]; ©SISSA Medialab Srl.. Reproduced by permission of IOP Publishing. All rights reserved.)

thus belong to both friendship groups. Many community detection methods have been proposed that detect overlapping communities [129]. These methods can be categorized into those that detect sparsely overlapping communities and those that detect dense overlaps [145] (see Figure 2.3).

Here we introduce one community detection method that detects dense overlaps (BigCLAM [125]), and one method that detects sparse overlaps between communities (Link clustering [127]). These methods were chosen as different approaches to the community detection problem which are fast and have previously been applied to PINs.

2.3.3.1 BigCLAM

The Cluster Affiliation Model for Big Networks [125] (BigCLAM) is a fast approximation of the Affiliation Graph Model (AGM) [145], which in turn expands the concept of a stochastic block model [131] to overlapping communities.

Stochastic block models assume a probability of interaction for every node pair based on latent parameters θ for groups of nodes (blocks). These blocks are often viewed to represent communities although the model itself does not assume higher probabilities of edges within blocks than across blocks. Given the block membership of the nodes, the edges are assumed to be independent. Using such a model, the probability $P(G|\theta)$ that a network G is generated from a specific parameter set θ can be calculated. A parameter set $\hat{\theta}$ which maximizes this probability represents an optimal community partition of the network. The approach taken to maximize $P(G|\theta)$ determines the speed and precision of the community detection method. Optimization methods are reviewed in [139].

The AGM [145] models the interaction probability between two nodes u and v based on the set of communities these nodes share $C_{u,v}$, and the interaction probabilities p_k of nodes in these communities by the equation:

$$p(u,v) = 1 - \prod_{k \in C_{u,v}} (1 - p_k). \tag{2.8}$$

Here, the interaction probabilities p_k are similar to the interaction probabilities within blocks in a stochastic block model. As nodes can be assigned to multiple communities, the interaction probability is calculated over all communities that two nodes share.

This model is fitted to networks by Metropolis Monte-Carlo and gradient descent to obtain network partitions [145]. As this method does not scale well to large networks, BigCLAM was developed to approximate the model and address the scaling issue. The improved time-efficiency can be seen using our HINT-P-14 data set with 7,869 nodes and 24,375 edges (see HINT-P-14 in Section 3.1.1). While

AGM community detection took approximately a month to fit 1000 communities, BigCLAM did the same in minutes.

Instead of having a single parameter p_k for every community, BigCLAM assigns every node u an *interaction strength* F_{uk} in every community $k \in C$ [125] (see Figure 2.6). This interaction strength parametrizes how strongly node u is connected with other nodes in the community k. Each node u is thus assigned a $1 \times K$ vector of interaction strengths $\mathbf{F_u}$, where $K = |C|$ is the number of communities. In this model the probability that two nodes u and v interact $p(u, v)$ is then given by:

$$p(u, v) = 1 - \exp(-\mathbf{F_u} \cdot \mathbf{F_v}^T).$$

Here, $\mathbf{F_v}^T$ is the transpose of the vector of interaction strengths of node v. Nodes are modelled to be more likely to interact the stronger their interaction strengths in common communities are.

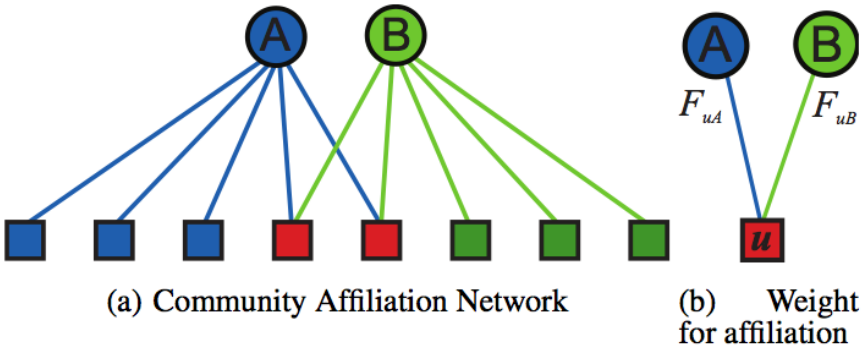

Figure 2.6: Schematic Diagram of BigCLAM. Boxes represent nodes, and circles represent communities. a) The community affiliation graph showing which node is a member of which community. Here red nodes are affiliated with both communities A and B, blue nodes are affiliated only with community A, and green only with community B. b) Schematic showing that nodes may have different interaction strength weights for different communities. Node u is depicted having an interaction strength F_{uA} in community A, and an independent interaction strength F_{uB} in community B. (After [125] ©2017 Association for Computing Machinery, Inc. Reprinted by permission.)

The interaction strengths, F_{uk}, are turned into discrete community memberships by thresholding. When F_{uk} is greater than a threshold, usually set to the background interaction probability in the network (the density), the node is regarded as affiliated with community k. The community affiliation matrix is thus determined by a $K \times N$ interaction strength matrix $\mathbf{F}$, where N is the number of nodes in the graph. The

network G is partitioned into overlapping communities in BigCLAM by finding the values of $\mathbf{F}$ which maximize the log likelihood function, $l(\mathbf{F}) = \log P(G|\mathbf{F})$ [125]:

$$l(\mathbf{F}) = \sum_{(u,v)\in E} \log(1 - \exp(-\mathbf{F_u}\mathbf{F_v}^T)) - \sum_{(u,v)\notin E} \mathbf{F_u}\mathbf{F_v}^T. \tag{2.9}$$

Here, $(u, v) \in E$ denotes all node pairs connected by an edge in the edge set E of the graph. The first term is a sum over the interaction likelihoods of all edges in the graph G, giving a positive contribution to the overall likelihood, while the second term penalizes for the node pairs with a non-zero interaction likelihood which are not present in the graph.

By generalizing the otherwise discrete community affiliation values into a matrix of continuous values $\mathbf{F} \in [0, 1]^{K\times K}$, BigCLAM can use a non-negative matrix factorization approach to optimize Equation (2.9) [125,146]. While the implemented optimization algorithm (backtracking line search) is not guaranteed to find a global optimum partition, investigations on synthetic networks show "good partitions" are found 98% of the time [125]. This imprecision in partitioning is a compromise made in favour of speed. BigCLAM has been shown to outperform other overlapping community detection methods for algorithmic efficiency, while still performing favourably in accuracy comparisons [125].

A limitation of the algorithm is that the number of communities to be fitted is not optimized for using BigCLAM. Instead it is proposed that 20% of the edge data can be used as a hold-out set to evaluate the quality of fit of the remaining 80% for a varying number of communities [125]. As a resolution parameter is not built into the method we use the number of communities as a proxy for resolution (see Section 3.1.2).

2.3.3.2 Link clustering

Link clustering is a hierarchical, edge-based community detection method. As such, rather than partitioning nodes into communities, link clustering generates edge communities. These edge communities can be turned into sparsely overlapping

node communities by assigning the nodes to which the clustered edges connect to communities [127].

Using this method edges are clustered based on edge similarities. These similarities are calculated for all edges that share a so-called "keystone node" k. The nodes not shared by the edges are referred to as "endpoint nodes". The edge similarity is calculated by the equation

$$s(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}. \tag{2.10}$$

Here e_{ik} denotes the edge between nodes i and k, and $n_+(i)$ is the *inclusive neighbour set* of node i defined as the set of neighbours of node i including the node itself. This similarity measure is the Jaccard index of the inclusive neighbour sets of the endpoint nodes and ranges between zero and one (Figure 2.7).

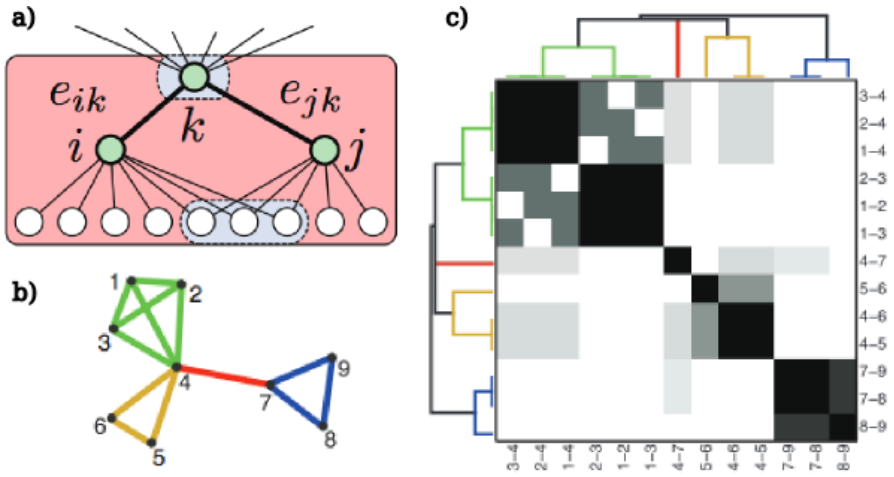

Figure 2.7: Schematic diagram of link clustering community detection. a) shows a schematic of how the similarity metric of link clustering is computed between the links e_{ik} and e_{jk}. Nodes within the dotted grey boundaries are overlapping between both neighbour sets $n_+(i)$ and $n_+(j)$, giving a similarity of $\frac{4}{12}$ for this edge comparison by Equation (2.10). b) shows an example link clustered network where edge communities are represented by different colours. The corresponding edge similarity matrix and dendrogram is shown in c). Darker regions in the matrix denote higher edge similarity with black being a similarity of 1, and white of 0. This edge similarity matrix exhibits blocks of similar edges which are clustered together to give the community structure shown in b). (Adapted by permission from Macmillan Publishers Ltd: Nature [127], copyright 2010).

All edges with a similarity value above a threshold s are clustered together to form edge communities, where s is the resolution parameter. This procedure can be interpreted as a cut through the dendrogram at height s in Figure 2.7c).

Link clustering uses local network information to calculate edge similarities. This local focus has the advantage that edge similarity values are quick to compute and thus link clustering scales well to available PIN sizes. In contrast, the edge similarities being calculated using only local information also means that less information is taken into account in the community detection process as compared to global methods such as Modularity Maximization. Given this consideration, it makes sense that link clustering has been shown to perform better on denser networks [127] where more information is available at a local level.

2.3.4 Comparing community detection methods

In this section we have described a few of the many community detection methods that differ in their definition of "community" and in their approach to community detection [128, 129]. To evaluate their performance, community detection methods can be compared in their application to different data sets. In order to evaluate successful partitioning, such comparisons require the knowledge of ground truth community assignments of the nodes in a network. This ground truth information can either be obtained from known communities in networks, or from generative network models with community structure.

Several data sets with known community structure have been used to compare or test community detection methods. A prominent example of such a network is the Zachary Karate Club, which split into two groups after a dispute between two leading members while the friendship network underlying the club was being studied [147]. Other real world networks with known community structure include an ecological network of dolphins which divided into two groups upon the disappearance of an individual [148], or the network of political committees and sub-committees in the house of representatives in the US [128]. Networks with covariates that are thought to define communities include 100 Facebook networks of US universities from September 2005 with data on major subjects, year of study, and so forth [149, 150], Youtube user groups [151, 152], or an Amazon co-purchasing network with product categrories [152].

A network model that has been suggested as a benchmark for community detection algorithms is the LFR graph [153]. This network model generates modular graphs with power law degree and community size distributions. LFR graphs are generated according to the configuration model with a mixing parameter controlling the proportion of edges from each node that link to other nodes in the same community. While this network model attempts to model characteristics observed in real-world networks, it lacks hierarchical community structure and clustering coefficients comparable to those observed in PINs and other networks [144].

In the case of PINs, neither ground truth modules nor a network model that adequately models PIN structures exist (see Section 2.2.2). In a yeast PIN study [23], protein complexes and pathway information were used as an approximation to functional modules to evaluate detected communities. In the absence of standard methods to compare community detection methods on PINs, functional annotations are used to evaluate detected communities. A key contribution of this thesis is to tackle the problematic nature of this approach. Functional annotations are discussed in the following section.

2.4 Functional annotations

The quality of a community is commonly assessed via functional annotations. These annotations describe protein characteristics in a structured way. Functional annotations are used to evaluate the biological coherence of the proteins grouped together in a community in order to evolve the mathematical concept of a community into the biological concept of a functional module. In this way functional annotations can be used as the basis for biological clustering in functional module detection (see Section 1.1.2).

In this section we give an overview of some important functional annotation resources with specific focus on the Gene Ontology [154].

2.4.1 Ontologies and ontological concepts

Functional annotations are terms that describe the characteristics of molecules such as proteins. These terms are generally organized in structured vocabularies called ontologies. An *ontology* is a hierarchical structure, a directed acyclic graph (DAG), that organizes terms by their *specificity* (Figure 2.8). The specificity of an annotation confers the precision with with a characteristic is understood. Terms higher up in the structure are general descriptions of protein characteristics and terms lower in the ontology describe more specific characteristics. Links between terms in an ontology describe hierarchical relationships directed from the general term (parent) to the specific term (child) (eg. Figure 2.8).

As relationships between terms are directed from general to specific, the *depth* in an ontology is loosely related to the specificity of a term. Here, the depth of a term denotes the number of edges that separate this term from the root node, the most general term in the ontology. As there can be multiple paths of different lengths to the root node, the depth of a term is not necessarily well-defined in ontologies. Thus, not every link in an ontology denotes the same increase in specificity (*ontology structure bias*). To better model the specificity of an annotation, the *Shannon Information Content (IC)* is used [155]. The IC quantifies the information conveyed by a term based how commonly it is associated with a molecule using the equation:

$$IC = -\log p. \tag{2.11}$$

Here, p denotes the probability of finding a specific annotation in an external information corpus, such as a list of genes with associated functional annotations or a PIN with associated annotations. While IC is not subject to ontology structure bias, it is affected by literature bias. *Literature bias* is the effect that terms which describe protein characteristics that are often experimentally tested are assigned a lower specificity, even if the tested characteristic would subjectively be viewed as specific [155]. IC-based specificities may change over time as the information corpus is expanded.

When a functional annotation from an ontology is associated with a protein, this term describes the most specific annotation that could be assigned to the protein by a particular experiment or publication. The full set of annotations that together describe the protein characteristic denoted by the specific annotation is contained in the ontology structure. This full annotation set can be obtained by tracing all paths from the specific annotation back to the root node against the direction of the links. The set of terms traversed in this tracing is the full set of annotations that describe the protein characteristic. We call this process *ancestral path mapping*.

As an example, using the subgraph of the Gene Ontology DAG shown in Figure 2.8, ancestral path mapping from the term "negative regulation of zinc ion import" gives the a full term set that further includes "negative regulation of zinc ion transmembrane transport", "regulation of zinc ion import", "regulation of zinc ion transmembrane transport", "negative regulation of zinc ion transport", "biological regulation", and "biological process".

2.4.2 The Gene Ontology

The Gene Ontology (GO) [154] is one of the most popular functional annotation resources [155]. The ontology is divided into three sub-ontologies: "biological process" (BP), "molecular function" (MF), and "cellular component" (CC). The BP sub-ontology describes the biological objective to which a protein or gene contributes; MF describes the biochemical activity of the protein; and CC denotes the place where a gene product is found [154]. For example, the gene HBA1 which encodes for a subunit of the hemoglobin complex is associated with the BP term "oxygen transport", the MF term "oxygen binding", and the CC term "hemoglobin complex". The relative sizes of the three sub-ontologies can be shown via the number of leaves in the DAG. In a version of the GO from August 2015 (see Section 3.2.1), the GO had 21,859 non-obsolete leaf terms which represent the most specific level of annotation. Of these, 11,409 were from the BP sub-ontology, 7,922 were from MF, and 2,528 belonged to the CC sub-ontology.

The annotations in these sub-ontologies are linked by relations such as “is a” and “part of”. While “is a” relations always link two annotations in the same sub-ontology, “part of” relations may cross over between sub-ontologies. Therefore, only “is a” relations are used for ancestral path mapping in the GO (see Section 2.4.1). The structure of the GO is demonstrated by a subgraph of the GO DAG in Figure 2.8.

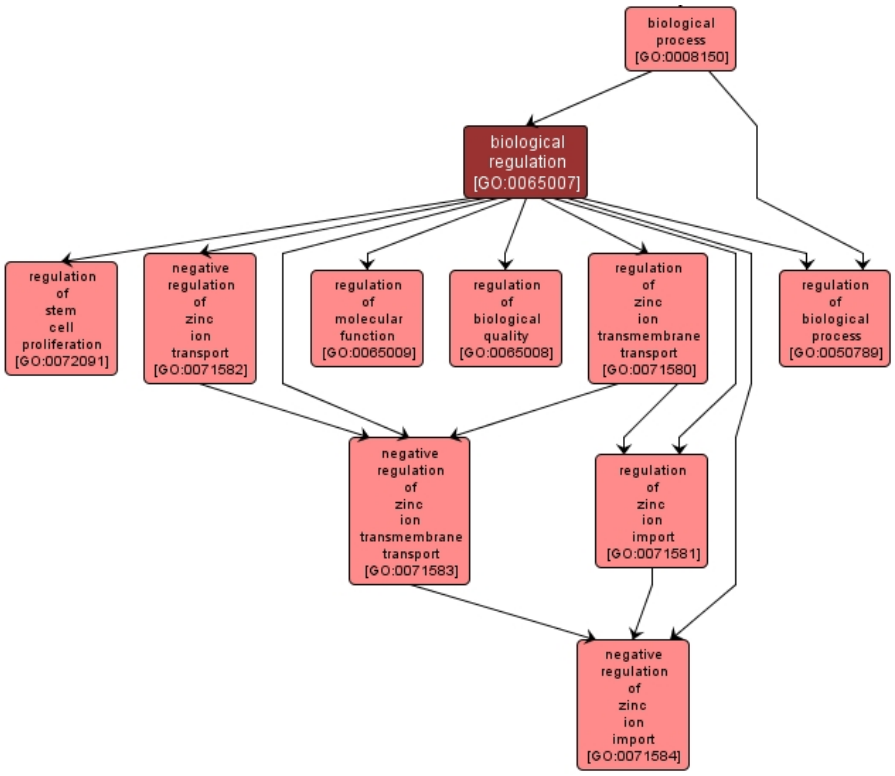

Figure 2.8: Subgraph of the GO DAG. This subgraph was induced by finding the neighbours of the “biological regulation” GO-term. Arrows denote “is a” relationships and point towards the more specific term. Often several paths between two GO-terms can be observed with a different number of links, showing that links do not represent a fixed increase in specificity (ontology structure bias). (Retrieved from yeastrc.org).

GO annotations are linked with genes or their protein products in gene association files available from the GO. These associations are generated from computational and experimental evidence, which are denoted by evidence codes (described in [156]). Each association represents a manual curation effort, except “IEA” evidence code associations which are generated computationally. Associations with an “IEA” evidence code are the most common. In a human gene association file obtained in July 2015 (see Section 3.2.1) over 30% of associations (146,398 of 486,639) were based on “IEA” evidence. These associations are mainly transferred from other annotation

databases from the same or related species. While studies have produced conflicting results on whether these annotations should be included or not, the increase in coverage given by their inclusion generally outweighs other arguments [157].

Other evidence codes that are of interest here include "IPI", "RCA", and "ND". "IPI" associations are inferred from protein interactions and should thus be excluded when evaluating communities of interacting proteins for reasons of circularity. The same argument holds for the "RCA" evidence code which denotes associations inferred from reviewed computational analysis that includes large scale data sets such as PINs. The "ND" evidence code denotes associations that were made based on lack of biological data available and thus associate proteins or genes only with the root BP, MF, or CC terms. "ND" associations do not necessarily conform to the same notion of functional annotation as associations from other evidence codes. The same argument is true of associations that have "NOT-qualifiers", which denote annotations that are not associated with a particular protein.

The GO is affected by both ontology structure bias and literature bias (see Section 2.4.1). As most terms have several paths of different lengths to the root term, not every link represents the same increase in specificity in the GO [155]. Similarly, due to high-throughput mapping of functional annotations that test for particular functions, there are overrepresented functional annotations in GO gene associations [158]. These biases affect depth and IC as measures of term specificity (see Section 2.4.1). Despite these biases it has been suggested that IC is the preferred method of capturing the specificity of annotations in the GO [157].

2.4.2.1 GO slims

A *GO slim* is a subset of the GO which typically contains general functional annotations at a low depth. Gene associations of more specific terms can be mapped to GO slim terms by tracing back "is a" relations towards the root node until a GO slim term is reached. This process may map a specific annotation to multiple GO slim terms (see Figure 2.8). The GO slim gene associations resulting

from this mapping are designed to give a broad overview of the characteristics of a protein at a glance [159].

GO slim sets can either be obtained directly from the GO, or generated computationally. A GO slim set should provide a collection of terms at a similar level of specificity. Using the approximations of specificity discussed in Section 2.4.1, term sets at a similar specificity can be found using computational algorithms as in [160].

2.4.3 Other functional annotation resources

Functional annotations can also be extracted from a variety of resources other than the GO, such as the MIPS FunCat [161], the Human Phenotype Ontology [162], and pathway annotation databases such as KEGG [163] or Reactome [164, 165]. These resources are briefly outlined below.

It should be noted that the GO transfers gene associations from other databases under the "IEA" evidence code. This transfer is performed via mappings from GO terms to the annotations in the respective databases. Such mappings exist for MIPS, Reactome, and KEGG. Thus, associations from these resources are likely to be contained in the GO.

2.4.3.1 MIPS

The MIPS database [166, 167] contains a functional annotation resource called the functional catalogue (FunCat). Funcat describes the cellular function of proteins and aims to be simple and intuitive [161]. It contains 1,069 functional categories with which human proteins can be annotated compared to the 11,409 biological process leaf terms (see Section 2.4.2) in the GO. While the simplicity of the ontology is viewed as a strength of FunCat [161], it also suggests that FunCat categories are coarse descriptions of protein functions. Such coarse descriptions may not be ideally suited to assess the homogeneity of proteins in a proposed module.

2.4.3.2 Human Phenotype Ontology

The Human Phenotype Ontology (HPO) describes human phenotypic abnormalities caused by disease [162]. Similar to the GO, the HPO is divided into three sub-ontologies: "mode of inheritance", "onset and clinical course", and "phenotypic abnormalities". The latter of these is typically used to describe disease phenotypes.

HPO disease annotations can be used as protein functional annotations via protein-disease associations which are in turn mapped to HPO terms. Thus, similarity of two human proteins based on HPO terms is a measure of how similar the effects of the diseases are to which these proteins are linked.

Gene or protein-disease associations can be obtained from ClinVar [168] or OMIM [169]. ClinVar contains information on risk loci which denote an increased risk of a disease linked with a genetic variant, and both databases contain associations for Mendelian diseases. Risk loci are regarded as low confidence associations as an increased risk does not denote a causative link. Furthermore, risk loci often lie outside defined genes on the DNA (intergenic) and therefore cannot be unambiguously mapped to a respective gene [170]. In contrast, Mendelian disease associations are viewed as causative. Mendelian disease traits are rare inherited genetic variants with a high likelihood of disease association to a single gene (high penetrance). In light of these considerations it makes sense that the HPO was developed around Mendelian diseases [171].

2.4.3.3 Pathway annotations

Pathway annotations, such as those provided by KEGG [163] and Reactome [164, 165], summarize detailed knowledge of cellular pathways that perform biological processes. The necessity to have detailed knowledge to assign a pathway annotation leads to comparatively low coverage. For example, at the time of writing (August 2016) Reactome contains 9,321 genes with pathway annotations while 18,265 genes are associated with GO annotations in the July 2015 data set (see Section 3.2.1).

2.5 Functional homogeneity

Methods that use functional annotations to calculate the biological coherence of proteins in a community are summarized by the term "functional homogeneity". There exist two main approaches for functional homogeneity calculation: functional enrichment and functional similarity. These approaches are described in this section.

2.5.1 Functional enrichment

Functional enrichment is a statistical method of assessing how significant the prevalence of a functional annotation is in a set of genes or proteins. While this method was initially used to evaluate the statistical significance of a set of genes obtained from the analysis of gene expression studies (eg. [172, 173]; see Section 2.6 for gene expression data), it has become the most popular method of community evaluation (eg. [22, 24, 25, 27, 49, 50, 174–176]).

There are several methods of calculating the enrichment of a functional annotation which are reviewed in [172]. Here we focus on how the hypergeometric distribution is used to assess the significance of a functional annotation k being associated with f_C proteins in a community where h_C proteins are functionally annotated, given a PIN with H annotated proteins where F proteins are associated with the annotation k. The enrichment of the term k is quantified by the p-value P_k which denotes the probability of observing a prevalence of at least f_C uniformly at random. This p-value is calculated by the equation:

$$P_k = \sum_{i=f_C}^{h_C} \frac{\binom{F}{i}\binom{H-F}{h_C-i}}{\binom{H}{h_C}}. \tag{2.12}$$

It should be noted that the number of annotated proteins in a community h_c is not necessarily the same as the number of proteins in a community N_C. As annotations can only be shared by proteins that are functionally annotated, enrichment is calculated over only these proteins.

Significant enrichment of an annotation in a group of proteins is determined when the p-value is below a significance threshold α, which is often set at 0.05. The

parameter α controls the Type-I error of the significance test. This error denotes the probability that significance was assumed when the annotation was in fact distributed according to the null hypothesis of uniformly distributed annotations.

When using functional enrichment for the evaluation of PIN communities, the presence of a significantly enriched annotation at a specific significance level (often 0.05) is interpreted as a signal that a community is biologically coherent. This is similar to using the p-value of the most enriched functional annotation as a score for the community quality.

2.5.1.1 Correcting for multiple testing

At a significance threshold $\alpha = 0.05$, it is accepted that every tested annotation has a 5% chance of being declared significantly enriched when it is actually uniformly distributed. Thus, when performing thousands of enrichment tests on the same set of proteins at this α, a considerable number of annotations will falsely be evaluated as significant. To counteract this effect, we can correct for multiple testing of related hypotheses (a family).

There are many ways in which to correct for multiple testing, which are reviewed in [177,178]. Here we briefly exlain two methods of controlling the family-wise error rate (FWER) and the false discovery rate (FDR) that are used in this project.

The FWER is the probability that at least one annotation was falsely evaluated as significant and the FDR denotes the expected proportion of annotations evaluated as significant that are in fact not significant. Multiple testing correction approaches that correct for these error rates are the Bonferroni correction and the Benjamini-Hochberg correction respectively.

To keep a constant FWER, the Bonferroni correction divides the intended FWER by the number of tests performed to obtain a multiple-testing corrected significance level α. The probability of any significantly enriched annotation at this significance level being false is given by the FWER used. Benjamini-Hochberg controls for the FDR using a stepwise multiple testing correction procedure. In this procedure the p-values are ranked and each assigned a separate significance level

given by $\alpha_i = \frac{\text{FDR} \times i}{m}$, where m is the number of tests, i is the rank of an individual test, and FDR is the pre-set acceptable false-discovery rate.

A further method of applying these multiple testing corrections is by directly adjusting the p-value to reflect a more stringent significance test [177]. For example, adjusted p-values using the Bonferroni correction are obtained by multiplying the p-value and the number of tests m, while keeping a maximum adjusted p-value of 1.

As controlling the FWER is a strong limitation of the significance evaluation, the Bonferroni correction is regarded as a conservative approach to multiple testing correction. In comparison, controlling for the FDR is a less conservative approach [177].

2.5.2 Functional similarity

The term "functional similarity" refers to the similarity of the cellular functions in which two biological molecules are involved. This similarity is quantified using *semantic similarity measures* which calculate the functional similarity of two proteins based on shared functional annotations (see Section 2.4).

Semantic similarity measures can be divided into pairwise measures that compute the similarity between functional annotations, and groupwise measures that compute the similarity between groups of annotations. As proteins are typically associated with groups of annotations, pairwise semantic similarity measures use mixing strategies to combine pairwise similarity scores for protein functional similarity computation. These mixing strategies entail using the maximum pairwise similarity (Max), the average of all pairwise similarities (Avg), or combinations of these strategies via the average of the highest similarity scores (Best Match Average - BMA) [157, 179].

A common feature in the calculation of both groupwise and pairwise semantic similarity is the use of a quantification of annotation specificity, such as the ontology depth or the IC (see Section 2.4.1). Terms or term sets are compared via their ancestral term sets obtained by ancestral path mapping (see Section 2.4.1). The

specificity of the shared ancestral annotations is typically used to quantify the functional similarity.

One of the oldest and most common semantic similarity measures [179], called the Resnik score [180], epitomizes this approach by calculating the pairwise similarity of two terms $\rho_{Res}(t_1, t_2)$ via the highest IC score of the shared ancestral annotation set (MIA = maximum informative ancestor). The Resnik score is given by the equation

$$\rho_{Res}(t_1, t_2) = IC(\text{MIA}),$$

and shown schematically in Figure 2.9.

Since this seminal work, many semantic similarity measures have been proposed (reviewed in [157, 179]). While the Resnik score is still shown to perform well in comparative studies (eg. [94, 157]), it has been suggested that groupwise measures that quantify annotation specificity via IC are best suited for protein functional similarity analysis [157]. Here we focus on three groupwise semantic similarity measures: the Pandey measure [181], simGIC [182], and simUI [183].

The Pandey measure [181] can be viewed as an extension of the Resnik score to groups of annotations (Figure 2.9). Instead of using the maximum IC of shared ancestral annotations, the Pandey measure takes all shared ancestral annotations into account. Similar to the calculation of the IC (see Equation (2.11)), the prevalence of the set of shared annotations in the ancestral annotation sets of all proteins in an external corpus (here, the PIN) is used to assess the similarity of the proteins by equation:

$$\rho_{pan}(S_i, S_j) = -\log_2 \left(\frac{G_{\Lambda(S_i,S_j)}}{G_r} \right). \tag{2.13}$$

Here, $G_{\Lambda(S_i,S_j)}$ denotes the number of proteins in the PIN which are annotated with the intersection set of the ancestral term sets S_i and S_j of nodes i and j respectively, and G_r represents the total number of annotated proteins in the network. Prior to the publication of this semantic similarity measure in [181], similar semantic similarity measures were used in [184, 185].

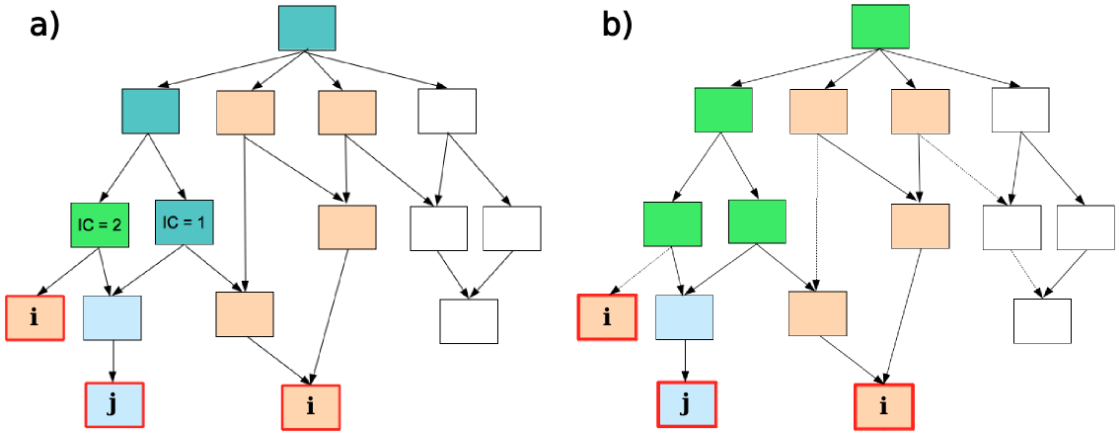

Figure 2.9: Schematic diagram of how a) the Resnik score with Max mixing strategy, and b) the Pandey measure are used to compute functional similarity. The directed acyclic graph represents an ontology, with the terms in red borders denoting the annotations associated with the two proteins i (orange) and j (light blue). The ancestral term sets S_i and S_j (see Equation (2.13)) are given by the orange, turquoise and light green boxes (S_i), and the light blue, turquoise and light green boxes (S_j). In a) the turquoise boxes indicate the overlap term data which are not used to compute the similarity via the Resnik score. Only the box in light green which is shown to have an IC of two is used to asses the similarity. Using this method the information content of a single functional annotation determines the similarity of two proteins. In b) the light green boxes show the data used to evaluate the similarity of proteins i and j via the Pandey measure. The probability of picking a protein from the designated PIN uniformly at random which is associated with all of these terms determines the similarity via this measure as calculated by Equation (2.13).

The simUI measure [183] uses the Jaccard index between the ancestral term sets, S_i and S_j, of two proteins, i and j, to compute their similarity by the equation:

$$\rho_{ui}(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}. \tag{2.14}$$

The simGIC semantic similarity measure [182] is an extension of simUI, which uses the cumulative information content of the ancestral term sets instead of their cardinality. This is given by the equation

$$\rho_{gic}(S_i, S_j) = \frac{\sum_{t \in \{S_i \cap S_j\}} IC(t)}{\sum_{t \in \{S_i \cup S_j\}} IC(t)}. \tag{2.15}$$

Here, $IC(t)$ denotes the IC of the term t.

While simUI and simGIC have performed well in comparative studies (eg. [94, 179]), the Pandey measure has previously been applied to functional module detection in PINs [48].

2.5.2.1 Annotation length bias

It has been shown that semantic similarity measures are affected by the number of functional annotations with which proteins are associated. By investigating the functional similarity of random groups of GO BP terms of different lengths it was found that the functional similarity tends to increase with increasing numbers of terms [186]. This effect, called *annotation length bias*, was observed for both groupwise and pairwise semantic similarity measures. As well-studied proteins tend to have more annotations, annotation length bias has the effect that well-studied proteins tend to have higher functional similarities than poorly-studied proteins (see Section 6.2).

2.5.2.2 Comparing semantic similarity measures

A large variety of semantic similarity measures exist [157, 179]. To help in the selection of an appropriate measure for a specific task, comparative studies have been carried out [94, 157]. As functional similarity is a concept that cannot be directly measured, comparative studies evaluate the performance of semantic similarity measures using approximations of this concept. Thus, semantic similarity measures can only be compared for a particular data set, limiting the generality of performance assessments.

Several data sets that are thought to relate to functional similarity have been used to evaluate semantic similarity measures. The tool CESSM [187] assesses the performance of semantic similarity measures via sequence, enzyme commission (EC), or Pfam domain data. Other data sets that have been used to approximate functional similarity include gene expression [157], protein-protein interactions [94], pathways [157], protein complexes [157], and subjective human expert judgement [188]. Evaluations of semantic similarity measures using different data sets are reviewed in [157].

The link between these data sets and functional similarity is not always straightforward. For example, while sequence similarity has been shown to be correlated with functional similarity across semantic similarity measures, this relationship does

not appear to be linear [157]. For comparisons based on protein-protein interaction data, results are confounded by annotation length bias (see Section 2.5.2.1). In these tests semantic similarity measures are evaluated in their ability to discern protein-protein interactions (positive reference set) from random protein pairs (random reference set) given that non-interactions are rarely published. In order to obtain a positive reference set of interactions that can be seen as ground-truth, interactions between well-annotated and well-studied proteins are used (eg. [94]). As argued in Section 2.5.2.1 such a set selects for high functional similarity scores relative to other interacting proteins. In contrast, random reference sets do not exhibit the same selection bias as the positive reference set. Thus, the evaluation rewards measures that find well-studied protein-protein interactions.

2.5.2.3 Community homogeneity from similarity scores

Semantic similarity measures are used to quantify the functional similarity of two proteins. These pairwise similarity scores must be combined to determine the functional homogeneity of a community. This combination is performed in two ways: by averaging all pairwise similarities (eg. [189]), or by averaging the similarities of all interacting proteins (eg. [48]).

It has been argued that using only interacting proteins corrects for higher similarity scores between interacting proteins compared to non-interacting proteins [48]. This disparity in similarity scores particularly affects larger communities that tend to be less dense. However, all proteins in a functional module are expected to contribute to the same cellular function whether they interact or not. Thus, the more stringent full pairwise averaging approach is preferred here. This functional homogeneity scoring approach is given by the equation

$$\mathrm{fh} = \frac{2}{|c|(|c|-1)} \sum_{u,v \in c;\, v>u} \rho_{SS}(u,v). \tag{2.16}$$

Here, u and v are proteins in the community c, and ρ_{SS} denotes the functional similarity score by semantic similarity measure SS.

2.6 Gene expression data

Gene expression refers to the process by which the information stored in a gene is used to produce a gene product such as a protein (by transcription and translation). Measurements of the rate of this expression process give a snapshot of the dynamical cellular environment. As gene products interact to perform cellular functions (see Section 1.1.1), their abundances can give insight into processes that are being performed in the cell.

It has been shown that protein abundance is determined primarily by the rate of transcription (DNA into mRNA) rather than translation (mRNA into protein) [190]. While protein abundances can be directly measured [73, 191], measurements of transcription are cheaper and thus more widespread. However, as measurements of these quantities are subject to experimental error, Pearson correlations between transcription and protein abundance data sets typically range from 0.2 to 0.5 depending on the measurement technique [76]. Thus, the two measurements are not equivalent. The two main experimental techniques for determining levels of transcription via mRNA abundance are DNA microarrays [17] and RNA-seq [192].

DNA microarrays use the binding of complementary single-stranded DNA sequences (DNA hybridization) to measure mRNA abundances. Short strips of single-stranded DNA (probes) are immobilized on an array and are exposed to fluorescently labelled cDNA that is reverse transcribed from cellular mRNA. The intensity of the fluorescence is used to infer probe mRNA concentrations. These probe concentrations can then be aligned to the genome to obtain measures of gene expression [17]. Alternatively, mRNA abundances can be measured directly via quantitative sequencing techniques. In RNA-seq experiments, mRNA are broken down into fragments of 50-500 base pairs and sequenced. The obtained sequences are aligned to the genome to give quantitative information on how often different regions of the genome were transcribed [192].

These two techniques have been shown to generate different yet correlated gene expression levels (Pearson correlation of $r = 0.67$ in [76]). This difference can have effects on the analysis of these data sets. For example, the structure

of gene regulatory networks inferred from expression data obtained by RNA-seq and microarray experiments was found to differ [193]. Due to a higher level of background noise in microarray experiments from effects such as unspecific hybridization (cross-hybridization), it has been suggested that RNA-seq is the more reliable technique [192]. Indeed, by comparison with direct measurements of protein abundances RNA-seq was found to better capture absolute expression values [76].

A further argument for RNA-seq is that of data reproducibility. As levels of gene expression are determined by a complex cellular environment, it is not expected that this environment can be perfectly recreated in replicated experiments. This circumstance emphasizes the need for replicates in the statistical analysis of gene expression data. The variation between replicates can be used to evaluate experimental techniques for gene expression measurement. Investigations on the consistency of microarray gene expression studies have shown that while results can be reproduced in replicated experiments, they can vary between microarray platforms and particularly between laboratories [194–196].

While RNA-seq has shown "high levels of reproducibility" [192], it is subject to transcript length bias. As mRNA are randomly fragmented into 50-500 base pair sequences, longer mRNA strands are expected to have more sequence reads. In order to prevent the transcript length from affecting the overall gene expression level, RNA-seq data are reported using units of RPKM which are obtained by dividing the number of reads by the transcipt length. However, due to the stronger sampling of long mRNAs, the expression levels of the corresponding genes have a lower variance. Statistical tests such as those for differential expression (see Section 2.6.2) thus have more power, the longer the transcript [197]. A similar bias that selects for highly expressed genes exists for both experimental techniques [17, 197].

2.6.1 Co-expression analysis

A popular use for gene expression data is gene co-expression analysis. Gene co-expression has been linked to functional relatedness [198–200]. While this link initially appears weak, reducing effect of noise in the analysis [199, 200], or

allowing for time delayed expression correlations [198] increases the signal strength. Therefore, gene co-expression can be used to evaluate functional similarity measures (see Section 2.5.2.2) or detect functional modules (see Section 1.1.2).

Gene co-expression scores are calculated by correlating the expression profiles of two genes. A gene's expression profile is a vector of its normalized expression scores across samples (see [201] for expression score normalization). Typically absolute Pearson correlation coefficients are used to quantify the level of co-expression of two genes (eg. [174]). The absolute values are used as both positive and negative correlations are of interest. A cellular function may involve a gene product suppressing the expression of another gene such that their expression levels are negatively correlated.

Co-expression analysis for functional module detection requires the assessment of a module co-expression score. Module co-expression can be quantified by the average of pairwise gene co-expression scores (eg. [189]). Thus, similar to functional homogeneity, module co-expression is calculated by Equation (2.16), replacing the functional similarity scores by co-expression scores.

2.6.2 Differential gene expression

Finding *differentially expressed genes (DEGs)* is the process of identifying genes that can distinguish between samples of two phenotypic categories such as "case" and "control", or "disease" and "healthy". For this purpose differential gene expression studies are set up to measure whole genome expression levels of samples that exhibit these phenotypes. A gene whose expression level changes significantly between the phenotype samples is thought to be involved in the investigated phenotype. Assessing the significance of the change in expression levels between the two sample categories is the subject of differential gene expression analysis.

Differential gene expression data sets are $N \times P$ matrices, where N is the number of samples, P is the number of genes, and $N \ll P$ (typically $N < 100$ and $P > 20,000$ for whole genome data). While gene expression levels are not independent of each other, the complexity of the data makes exhaustive testing

of sets of genes infeasible. Instead, a one-by-one approach is used to test the differential expression of each gene individually.

A simple measure for differential expression is *fold change*. The fold change denotes the ratio of mean expression scores between the phenotype sample sets. To assess the significance of the observed fold changes, statistical tests based on linear models are commonly used [202]. For example, two-sample t tests can be used to assess whether linear model coefficients fitted to the expression scores of the sample sets are significantly different. As the t test assumes normally distributed data, expression scores are commonly log-transformed to approximate this distribution. Statistical tests for differential gene expression analysis are reviewed in [202].

As tests for differential expression are performed individually for each of the P genes, there is a large multiple testing burden that must be corrected for (see Section 2.5.1.1). To mitigate the reduction in power that comes with multiple testing correction, genes that are unlikely to be differentially expressed can be filtered out based on small fold changes or low expression scores that tend to have low signal-to-noise ratios [202].

Differential gene expression analysis is a popular tool. Following the described procedure it is common that thousands of DEGs are found. Rather than detecting DEGs, it has been suggested that the core challenge is to put these genes into the correct context [17]. Addressing this challenge is the focus of this project.

2.7 Summary

In this chapter we have described the data sources and provided background information on the methods used to analyse these data in the remainder of this thesis.

The data which most of the work presented in this thesis is based on, are protein-protein interactions. In Section 2.1 we described the two main experimental techniques used to measure these data (yeast two-hybrid and TAP-MS) and detailed their limitations. These limitations have the effect that the data are noisy, and that currently available large-scale data sets that are collated in databases have limited coverage of the human interactome. Furthermore, we described how inspection

bias arises in these data sets due to selections made on which proteins to test for interactions.

Protein-protein interaction data are often represented as a network. Networks, statistics used to describe them, and different types of network structure were introduced in Section 2.2. The predominant network analysis tool used in this project is community detection. In Section 2.3 we gave a short overview of community detection methods that detect communities at different scales via the inclusion of a resolution parameter. These methods were divided into methods that generate non-overlapping communities, and methods that generate overlapping communities.

In our work we evaluated communities in PINs detected by different community detection methods. This evaluation was performed on the basis of functional annotations which are described in Section 2.4. In this section we gave a brief overview of several sources of functional annotation data with particular focus on the largest of these resources, the GO. We further elaborated on the organization of functional annotations into ontologies and explained how literature bias and ontology structure bias affects their use for evaluation purposes.

Methods of using annotations to evaluate the biological coherence of communities in PINs were summarized in Section 2.5. These methods vary in their use of the information contained in the ontology structure. While functional enrichment uses only the annotation sets to evaluate whether an annotation is significantly enriched, functional similarity measures relate sets of annotations to each other using the ontology structure. This increased use of the ontology structure has the effect that functional similarity methods can be subject to ontology structure bias or literature bias. Furthermore, these measures are affected by annotation length bias which describes how proteins associated with more annotations are evaluated as being more similar. We further discussed how the many available functional similarity measures are compared to assess which measure is best for a particular task.

The third important data source that is used in this project is gene expression. In Section 2.6 we gave a brief overview of the two main methods that are used to measure these data (microarray and RNA-seq). Furthermore, we elaborated on how

these data can be analysed to either infer functional similarity by co-expression, or to detect a signal for the molecular mechanisms that are involved in a disease phenotype.

3

Data sets and processing

Contents

3.1 Network data 65
3.1.1 Protein-protein interaction data 65
3.1.2 Network partitions 69
3.2 Functional annotations 70
3.2.1 Gene Ontology 71
3.2.2 Human Phenotype 73
3.2.3 Functional homogeneity 74
3.3 Gene expression data 75
3.3.1 Co-expression analysis 75
3.3.2 Biological applications 76

In this chapter we detail the data sets that were used in this work, how they were filtered for quality control, and how they were processed. The chapter is subdivided into sections on network data, functional annotations (GO and human phenotype), and gene expression data.

3.1 Network data

3.1.1 Protein-protein interaction data

Protein-protein interaction data are available from many online databases (see Section 2.1.3). Selecting an appropriate database for a particular investigation,

often depends on striking a balance between the quality of interactions and their coverage that best suits the problem at hand. When developing a generally applicable methodology that relies on PIN modular decomposition, good coverage is essential. However, high error rates may affect community detection methods used to detect functional modules. In order to select suitable PINs for our investigations, we compared PINs from several databases using summary statistics.

As discussed in Section 2.1.1 protein-protein interactions exist in two types: physical associations (P-type), and associations (A-type). Human protein-protein interaction data obtained from HPRD [103,104] (retrieved February 2014), IntAct [102] (retrieved February 2014), HINT [69] (A-type retrieved January 2014; P-type retrieved May 2013), and BioGrid [100] (retrieved January 2014) were split to reflect these types.

The division into A-type and P-type interactions is performed by the databases themselves in the case of HINT and IntAct. The HINT database denotes A-type interactions as "co-complex", and P-type interactions as "binary", while IntAct uses the MIPS categories for "physical association" and "association" [70]. The MIPS category "direct interaction" was associated with only 3.9% of interactions and was omitted.

While HINT and IntAct databases offer a division into A-type and P-type interactions, BioGrid is divided into "physical" and "genetic" interactions. This split reflects a broader difference between interactions for which direct evidence was reported (physical), and interactions inferred based on mutation experiments on the associated genes [100]. Here, only physical interactions are of interest. Physical interactions were split into A-type and P-type data based on experimental evidence codes associated with each interaction as shown in Table B.1 in Appendix B. As the BioGrid data sets are the most comprehensive among the data sources we used, an additional network of both A-type and P-type data from BioGrid was set up to give a maximum coverage PIN.

In contrast to the other databases, HPRD was not divided into two interaction types. As HPRD is a database of manually curated small-scale experiments which

reports its data as "binary interactions", it is classified entirely as a P-type data set. While this classification should not suggest that yeast two-hybrid-type experiments were performed to obtain the data, we used it to reflect that experimental data were reported in the same way as P-type data.

Further filtering was applied to the A-type and P-type data sets to include only interactions between human proteins, collapse multiple edges, and exclude self-interactions and any proteins not in the largest connected component of the network (see Section 2.1.4). The resulting PINs were compared using the network statistics discussed in Section 2.2.1 (Table 3.1).

Table 3.1: PIN network statistics.

Network	Nodes	Edges	Density	GC	ALC	Av. k	Var(k)
HPRD-14	9,270	36,918	0.00086	0.053	0.138	7.97	215
BioGrid-P-14	10,667	41,051	0.00072	0.015	0.104	7.70	613
HINT-P-14	7,869	24,375	0.00079	0.028	0.100	6.20	167
IntAct-P-14	11,068	41,428	0.00068	0.022	0.093	7.49	284
BioGrid-A-14	13,165	93,275	0.00108	0.017	**0.410**	14.17	**8,335**
HINT-A-14	2,720	6,532	**0.00177**	**0.073**	0.334	4.80	122
IntAct-A-14	6,399	21,190	0.00104	0.025	0.296	6.62	593
BioGrid-AP-14	**14,899**	**123,025**	0.00111	0.019	0.320	**16.51**	8099

Table 3.1: GC refers to the global clustering coefficient, ALC to the average local clustering coefficient, Av to the average, and k to the degree. Global and local clustering coefficients are calculated as explained in Section 2.2.1. The networks are split into three groups: the upper four PINs are P-type, the middle three are A-type, and the network at the bottom is a combined A and P type network. The highest values in each category are shown in bold. It can be seen that A-type networks tend to have higher local clustering coefficients and densities.

Table 3.1 shows that the P-type PINs tend to have more nodes than A-type networks, with the exception of BioGrid-A-14, and A-type networks tend to be more locally clustered than P-type networks.

A second consideration for the selection of PINs relates to the initial target application of this work (see macrophage differentiation described in Section 1.2.2). In order to assess the networks, we consider their coverage of several signalling proteins, called interleukins, that have previously been implicated in different macrophage polarizations. The results of whether the interleukins IL-4, IL-13, IL-10, or IL-12A/B are contained in the tested PINs is shown in Table 3.2.

Table 3.2: Interleukin Protein Interaction Network Membership.

Network	IL-4	IL-13	IL-10	IL-12A/B
HPRD-14	✓	✓	✓	✓/ ✓
BioGrid-P-14	✓	✓	✓	✗/ ✗
HINT-P-14	✓	✓	✓	✗/ ✗
IntAct-P-14	✓	✓	✗	✓/ ✓
BioGrid-A-14	✗	✗	✓	✗/ ✓
HINT-A-14	✗	✗	✓	✗/ ✗
IntAct-A-14	✓	✗	✗	✗/ ✗
BioGrid-AP-14	✓	✓	✓	✓/ ✓

Table 3.2: The networks are split into three groups: P-type PINs (upper), A-type PINs (middle) and a combined PIN (lower). A checkmark denotes the given interleukin gene is present in the network, while a cross denotes it is not. These interleukins were selected for their relevance to macrophage differentiation (see Section 1.2.2). IL-12 is a heterodimer composed of protein products from the two genes IL12A and IL12B which are reported individually in PINs. Network summary statistics for these networks are given in Table 3.1.

Table 3.2 shows that while P-type networks all show a good coverage of the interleukins previously linked with the different macrophage polarizations, none of the A-type networks do. The BioGrid-AP-14 data set is the only data set that both contains A-type interaction and sufficient coverage of the tested interleukins.

We selected two PINs with different characteristics to represent opposite ends of the compromise between the coverage and the quality of interactions (false-positive rate). This selection allows us to evaluate the effects of these characteristics in our investigations (see Chapter 5). Importantly, both selected PINs exhibit a good coverage of macrophage-relevant interleukins. BioGrid-AP-14 was chosen to represent a PIN with high coverage, but likely also a comparatively high false-positive error rate. In contrast, HINT-P-14 was selected as a high-quality PIN (see HINT quality control in Section 2.1.3). These PINs overlap in 7,606 of 7,869 possible nodes and 17,912 of 24,375 possible edges. Although many of the interactions in HPRD-14 were manually curated and it exhibits a better coverage of relevant interleukins, it was not selected as the high-quality interaction data set due to concerns with its lack of recent updates and its construction. HPRD is likely more affected by inspection bias as it is focused on small-scale studies around disease-related proteins (see Section 2.1.3).

Much of the research presented in this thesis was performed on updated versions of the HINT-P-14 and BioGrid-AP-14 networks (updates denoted HINT-P and BioGrid-AP; HINT-P retrieved August 2015; BioGrid-AP retrieved August 2015). These PINs were set up using the pre-processing procedure explained above. Network statistics of HINT-P and BioGrid-AP are shown in Table 3.3.

Table 3.3: Network statistics of HINT-P and BioGrid-AP.

Network	Nodes	Edges	Density	GC	ALC	Av. k	Var(k)
HINT-P	10,927	49,301	0.00083	0.034	0.099	9.02	329
BioGrid-AP	15,405	165,343	0.00139	0.055	0.130	21.47	2475

Table 3.3: Network summary statistics as in Table 3.1. GC refers to the global clustering coefficient, ALC to the average local clustering coefficient, Av to the average, and k to the degree. BioGrid-AP is the larger network with a higher density. The PINs overlap in 10,617/10,927 possible nodes, and 40,853/49,301 possible edges.

3.1.2 Network partitions

Much of the work in this thesis is done on communities derived from PINs. These communities were generated by partitioning the two selected PINs using four community detection methods (see Section 2.3). The methods used include two non-overlapping community detection methods (configuration model Modularity Maximization [126, 135], and Constant Potts model (CPM) Modularity Maximization [126, 140]) and two overlapping community detection methods (link clustering [127] and BigCLAM [125]). These methods were chosen as they have been previously used on PINs and are fast algorithms that scale well to currently available PIN sizes. Specifically, configuration model Modularity Maximization was chosen to reproduce and evaluate yeast PIN results [48] on human data in Chapter 5. As we argue that the results of this study may be confounded by the resolution limit affecting the chosen community detection method (see Section 5.1), we chose CPM Modularity Maximization which is not subject to this limitation (see Section 2.3.2.2).

PINs are partitioned into communities by the multi-resolution Modularity Maximization adaptation known as the Potts method using the Louvain algorithm (`https://launchpad.net/louvain` [126], retrieved May 2013; see Section 2.3.2.3)

which includes the resolution parameter γ. Network partitions were generated at 51 resolutions evenly spanning the interval $\gamma \in [10^{-1}, 10^{3}]$ for the configuration model, and $\gamma \in [10^{-4}, 10^{0}]$ for the CPM, on a logarithmic scale. Overlapping communities were generated using link clustering (`https://github.com/bagrow/linkcomm`, retrieved June 2014), which has an inbuilt resolution parameter S, at 121 resolutions in the interval $S \in [0, 0.6]$ in steps of $\Delta S = 0.005$. In contrast to the other three community detection methods, the free parameter in BigCLAM (`http://snap.stanford.edu`, retrieved June 2014) is not a resolution parameter, but instead the number of communities K to be fitted to the data. As increasing the resolution in a community detection method results in a "zooming-in" effect such that there are more, smaller communities (see Section 2.3.1), we used the parameter K as a proxy for resolution. BigCLAM communities were obtained at 101 values of K evening spanning the interval $K \in [1, 5001]$. Using this proxy, PIN partitions at neighbouring resolutions only differ in the 50 communities that are additionally fitted to the network data. When no further communities can be fitted to the network, the BigCLAM algorithm fits small communities of size 2 - 10 that are generally disconnected. The existence of such disconnected communities signals an overfitting of the network data.

Next to the number of communities fitted K, BigCLAM contains two further parameters α and β which parametrize the optimization algorithm. These parameters control the trade-off between the accuracy of the log likelihood maximum found and the speed of converging to a solution (see Section 2.3.3.1). We set $\alpha = \beta = 0.9$ based on a parameter sweep for BigCLAM on HINT-P-14. For values of $K \in [501, 3001]$ in steps of 500, the number of disconnected communities was counted at a range of α and β values between 0.0001 and 0.9. At $\alpha = \beta = 0.9$ no disconnected communities were found.

3.2 Functional annotations

The coherence of proposed functional modules is generally assessed using functional annotations (see Section 2.4). These annotations can be sourced from several

databases of which the most common is the GO (see Section 2.4.2). Following an assessment of functional annotation resources in Chapter 4, we use GO annotations in the majority of the work presented in this thesis. Here, we elaborate on the quality control and pre-processing steps that were undertaken to ensure the functional annotation data used is adequate for the problem at hand.

3.2.1 Gene Ontology

The GO (see Section 2.4.2) is an ontology that describes gene/protein characteristics. The GO ontology structure was retrieved from `http://www.geneontology.org` [154] in December 2013 for Chapter 5 results, and August 2015 for results presented in other chapters.

As discussed in Section 2.4.2, associations between proteins and GO annotations are obtained from different sources which are denoted by evidence codes (`http://www.geneontology.org` [154], associations retrieved November 2013 for Chapter 5 results, and July 2015 elsewhere). Several sources of GO associations are unsuitable for evaluating communities in networks and are thus excluded here. To prevent data circularity communities were not evaluated using associations inferred based on protein-protein interactions or computational analysis which may include PINs ("IPI" and "RCA" evidence codes). Further filtering was applied to exclude associations with "ND" evidence codes and "NOT"-qualifiers which represent uninformative or negative associations (see Section 2.4.2). Electronically inferred "IEA" associations were included in our analysis to ensure a good coverage over PIN proteins.

The GO is split into three sub-ontologies: "biological process" (BP), "molecular function" (MF), and "cellular component" (CC). While we used all three sub-ontologies to investigate network connectivity, we found annotations from the BP ontology to be best suited for module evaluation (see Section 4.2). This result is further supported by the widespread use of GO BP annotations to characterise protein function in the literature [24, 28, 32, 109, 157]. Thus, we focused on GO BP annotations for functional homogeneity calcluation.

Applying our filtering protocol to the GO BP term associations with human proteins left 185,713 of 187,473 associations (154,243 of 155,218 associations in the 2013 GO data sets). This reduction in associations was split between the filters in the following way:

- 0 (14) "RCA" evidence codes,
- 92 (86) "IPI" evidence codes,
- 656 (604) "ND" evidence codes, and
- 1012 (272) "NOT"-qualifiers.

The terms in brackets refer to the 2013 GO data set used in Chapter 5. We mapped the pre-processed set of GO BP associations to the selected HINT-P and BioGrid-AP data sets to give a the coverages of 89.65% and 84.03% respectively. The 2013 GO data was mapped to HINT-P-14 and BioGrid-AP-14 to give coverages of 88.4% and 79.7% respectively.

As these term associations represent only the most specific annotation that could be assigned to a protein in a particular publication, the full set of annotations associated with a protein was obtained by ancestral path mapping (see Section 2.4.1).

3.2.1.1 GO Slims

To investigate general trends related to GO annotations (see for example Section 4.2), we used GO slims. GO slims are subsets of the GO which only contain terms that broadly categorize the different functional annotations in the three GO sub-ontologies (see Section 2.4.2.1). Specific GO annotations associated with a protein can be mapped to these GO slim terms to give a general overview of the characteristics of the protein.

In Chapter 4, we used the generic GO slim [159] for this general overview. This GO slim contains 69 BP, 42 MF, and 34 CC annotations. The GO slimmer tool [203] was used to obtain associations, pre-processed as described in Section 3.2.1, that were automatically mapped to the relevant GO slim terms (`http://www.geneont`

ology.org, retrieved 21. May 2015). The coverage of the retrieved protein GO slim associations is shown in Table 3.4.

Table 3.4: GO slim coverage.

Network	BP	CC	MF
HINT-P-14	85.91	90.79	75.10
BioGrid-AP-14	77.99	83.65	68.45

Table 3.4: Coverage of GO slim term associations from the three GO ontology branches biological process (BP), cellular component (CC), and molecular function (MF) on proteins in HINT-P-14, and BioGrid-AP-14. Coverages are quoted as percentages.

An alternative to using the GO slims provided by the GO, is the computational generation of such reduced GO term sets. This approach is used and elaborated on in Appendix A.

3.2.2 Human Phenotype

To test alternative functional annotation data sets for module evaluation, we used a Human Phenotype Ontology (HPO) data set obtained from collaborators at UCB Pharma. The HPO [162] is a structured vocabulary of annotations which describe phenotypic properties of diseases (see Section 2.4.3.2). These annotations can be mapped to genes via genetic variants that are associated with a specific disease. As the proteins in our PINs are reported by gene identifiers using a simplified one-to-one correspondence between proteins and genes, the associations are thereby directly mapped to proteins. Using Human Phenotype (HP) annotations the functional homogeneity of a community is evaluated based on the similarity of disease symptoms of the diseases with which proteins in the community are linked.

The obtained HP annotation data set was generated by integrating genetic variant disease associations from ClinVar [168] (http://www.ncbi.nlm.nih.gov/clinvar/, retrieved August 2015) with disease descriptions by HP annotations from the HP ontology (http://human-phenotype-ontology.github.io/, retrieved August 2015). ClinVar associations are inferred on the basis of different types of analyses (see Section 2.4.3.2). To minimize the uncertainty introduced by the HP annotations, gene disease associations were filtered for high confidence associations from

Mendelian diseases. Mendelian disease traits are rare inherited genetic variants with a high likelihood of disease association to a single gene. To select for these associations several filters were applied to the ClinVar data set. Associations were selected for diseases with a pathogenic clinical status that are reported within the highly curated Online Mendelian Inheritance in Man database (OMIM) [169]. Furthermore only associations with origins denoted as germline, de novo, inherited, maternal, paternal, biparental or uniparental were included. Specifically, somatic mutations were excluded in this filtering process ($\approx 1.5\%$ of associations) which mainly report cancer associations that were not of interest to UCB Pharma.

Using this pre-processing protocol 2146 proteins in HINT-P and 2755 proteins in BioGrid-AP could be associated with HP annotations giving respective coverages of 19.64% and 17.88%.

Using the main "phenotypic abnormality" sub-ontology of the HPO that describes disease phenotypes, the disease identifiers were mapped to sets of HP annotations. These sets of HP annotations were completed by ancestral path mapping (see Section 2.4.1) to elucidate the full list of HP annotations associated with a particular protein. Using the full HP annotation sets, functional similarity between proteins based on disease association was calculated using the Pandey measure (see Section 2.5.2).

3.2.3 Functional homogeneity

In this thesis functional annotation sets associated with proteins were used to assess the functional homogeneity of communities proposed as modules. Functional homogeneity was computed via functional enrichment (see Section 2.5.1) and via semantic similarity measures that compute the pairwise functional similarities of proteins (see Section 2.5.2). The three semantic similarity measures used are simUI [183], simGIC [182], and the Pandey measure [181] (see Section 2.5.2). Pairwise functional similarity values were combined to give a community functional homogeneity score by Equation (2.16) unless otherwise stated (see Chapter 5 for alternative functional homogeneity calculation).

3.3 Gene expression data

We used gene expression data for two main purposes: as indpendent computational module validation in Chapter 6, and to introduce disease-application-relevant data into our pipeline in Chapter 7. These two applications require different types of gene expression data sets. In this section we detail the gene expression data used, and briefly describe processing steps that were implemented.

3.3.1 Co-expression analysis

In Chapter 6 systematic module validation was performed by assessing the level of co-expression of genes whose proteins were assigned to the same module. As modules can describe any cellular function, it is important that the expression data covers a wide range of tissues where genes can be co-expressed. Thus, the gene expression data to perform this co-expression analysis was sourced from the Genotype Tissue Expression (GTEx) project (Version 6, RPKM format, from `www.gtexportal.org/home/datasets`, retrieved Nov. 2015) [204]. These data comprise of over 8500 tissue-specific, whole genome RNA-Seq samples which were extracted postmortem from human donors and prepared according to the same protocol.

The data were retrieved in a processed format, in which expression values are reported per kilobase of transcript per million reads. As proteins in HINT-P and BioGrid-AP are reported by gene IDs using a simplified one-to-one correspondence between proteins and genes, Transcript IDs (Ensembl Gene IDs) could be mapped to the PIN gene IDs (Entrez IDs) using the Ensembl release 82 BiomaRt tool [205]. This mapping resulted in 147 out of the 23,230 Entrez IDs being assigned multiple Ensembl Gene IDs. Expression profiles that were mapped to the same gene ID were averaged.

The level of co-expression of two genes was calculated via the absolute value of the Pearson correlation coefficient of the genes' expression profiles (see Section 2.6.1). The calculated co-expression scores were then combined to a community co-expression score by Equation (2.16). The distribution of these co-expression scores

was used to compare communities that were evaluated as functionally significant by different methods.

3.3.2 Biological applications

In Chapter 7 our pipeline was tested on two biological applications: breast cancer hypoxia and macrophage differentiation (see Section 1.2). Information which is specific to the biological application is introduced into this pipeline via differential gene expression data (see Section 2.6.2). Such data sets contrast the expression levels of genes between samples from two phenotypes of interest.

3.3.2.1 Breast Cancer Hypoxia

Differential gene expression data from the MCF7 breast cancer cell line was obtained from Dr. Francesca Buffa at the Department of Oncology, University of Oxford. The gene expression data was generated using RNA-seq (see Section 2.6) on cells that were exposed to different conditions to generate normoxic and hypoxic cellular environments. Each condition was sampled twice to give a data set of four samples, and 64,233 transcripts.

Ensembl IDs which identify the transcripts were mapped to Entrez Gene IDs using the BiomaRt tool [205]. Differential expression in this data set was analysed using the GEO2R script provided by the Gene Expression Omnibus (GEO) maintained by the NCBI [206]. The GEO2R script uses a linear model with an empirical Bayes method as implemented in the Limma R package [207, 208] to calculate differential expression in a gene-by-gene approach. Using this method it is possible to assess the significance of differences in expression levels between phenotypic sample classes despite having only few replicates available. This is achieved by estimating the variance of sample class expression levels of individual genes using data across all genes in the data set [202].

Multiple testing of differential expression was corrected for by using Benjamini-Hochberg corrected p-values (see Section 2.5.1.1). Due to the few replicates that were obtained at each condition, only samples with zero expression levels were

filtered out. In order to use all available data, no further filtering of gene expression data (for example for low expression levels, see Section 2.6.2) was performed which could otherwise reduce the multiple testing burden.

Applying the described data processing protocol, 26,537 of 64,233 transcripts could be mapped to Entrez Gene IDs. Of these genes, 16,853 contained only non-zero expression values, which corresponded to 16,510 distinct Entrez Gene IDs. In cases where two Ensembl IDs mapped to the same Entrez ID, the more differentially expressed transcript was used.

The number of DEGs found at different false discovery rate (FDR) thresholds is shown in Table 3.5. DEGs that could not be mapped to nodes in BioGrid-AP include genes that code for non-coding RNAs or proteins whose interactions are difficult to determine systematically such as membrane proteins (see Section 2.1.1). The comparatively small numbers of DEGs in the hypoxia data set is likely due to the small number of replicates reducing the statistical power of the differential expression test. Under these circumstances an FDR threshold of 0.05 has to be used to obtain a sufficiently large sample of DEGs that can convey signals for differential regulation of specific functions.

As RNA-seq is affected by transcript length bias (see Section 2.6), the DEGs found may include more long proteins than expected under random sampling. Yet, as we have no reason to assume that proteins in functional modules should be of a similar length, a selection for longer proteins does not have to prevent modules from being enriched for DEGs. However, the small number of DEGs in the hypoxia data set may exacerbate the effect.

3.3.2.2 Macrophage Differentiation

Differential gene expression data for macrophage differentiation was obtained from the GEO series file GSE5099 [61]. Whole genome expression levels of macrophages cultivated from monocytes that were extracted from blood samples were measured by microarray experiment (see Section 2.6). M1 macrophages were obtained from unpolarized macrophages by treatment with IFN-γ, and M2 macrophages by

treatment with IL-4 (see Section 1.2.2). Each condition was replicated three times to give a data set of six samples and 44,928 transcripts.

As this data set was obtained via an Affymetrix microarray platform, gene probes were described by Affymetrix probe IDs. These probes vary in quality due to for example differences in binding specificity to the targeted mRNAs. Thus, when mapping probe IDs to Entrez Gene IDs we prioritized the more reliable probe in cases when several probe IDs mapped to the same Entrez ID. Mapping Affymetrix gene probe IDs to Entrez IDs is further elaborated on in Appendix B.2.

Differential expression of this data set was assessed as on the hypoxia data set (see Section 3.3.2.1) with the addition of a further filtering step. As microarray gene expression data tend to have a higher level of background noise compared to RNA-seq data (see Section 2.6), we excluded genes whose mean expression level across samples was below the 10% quantile of all expression level measurements. The p-values for differential expression were adjusted according to Benjamini-Hochberg multiple testing correction (see Section 2.5.1.1)

The number of DEGs at different FDR thresholds is shown in Table 3.5. For this data set we can use a more conservative FDR threshold of 0.01.

Table 3.5: Number of DEGs at different FDR thresholds.

Data Set	$<$ **0.05**	$<$ **0.01**	$<$ **0.005**	$<$ **0.001**
Hypoxia Full	848	2	0	0
Hypoxia BioGrid-AP	616	2	0	0
Macrophage Full	4,423	2,736	2,216	1,325
Macrophage BioGrid-AP	3,894	2,416	1,946	1,160

Table 3.5: $<$ 0.05 denotes an FDR threshold of 0.05. "Full" data sets refer to the total number of DEGs found in the expression data, and "BioGrid-AP" data sets to the number of DEGs that could be mapped to BioGrid-AP.

4

Functional homogeneity evaluation

Contents

4.1 Introduction 79
4.2 Testing GO annotations 81
4.2.1 Methods 81
4.2.2 Results 82
4.3 Testing functional enrichment 82
4.3.1 Methods 84
4.3.2 Results 84
4.4 Selecting a semantic similarity measure 86
4.4.1 Methods 87
4.4.2 Results 95
4.4.3 Discussion 100
4.5 Testing Human Phenotype annotations 102
4.5.1 Methods 102
4.5.2 Results 103
4.6 Discussion and conclusions 105

4.1 Introduction

A functional module is a group of interacting proteins that together perform one or more cellular functions (see Section 1.1.1). These modules are thought to represent an important level of organization in biology. We reviewed methods for functional module detection in Section 1.1.2. Generally, modules are identified

by performing community detection on PINs (eg. [23–26, 48]), or other networks of integrated biological data (eg. [38, 43, 46, 50]). While these approaches have been successful, not every community output by a community detection method represents a functional module. To assess which communities are potential functional modules, the similarity of the functions to which the proteins in a community contribute is evaluated. This similarity can be quantified by semantic similarity measures which are discussed in Section 2.5.2, or by functional enrichment (see Section 2.5.1). The aggregate similarity of proteins grouped into a community is referred to as its functional homogeneity.

Functional homogeneity evaluation is based on functional annotations. As discussed in Section 2.4 there are several types of functional annotations which can be used for module evaluation. General ontologies such as the Gene Ontology (GO) [154] and MIPS [166, 167] provide a structured vocabulary to describe the characteristics of proteins. Other resources such as Reactome [164, 165] or KEGG [163] contain pathway information. Alternatively, the similarity of diseases associated with the proteins can be used to quantify protein similarity via Human Phenotype Ontology (HPO) annotations [162]. While the GO is the most comprehensive of these resources, functional annotations such as disease associations can provide different perspectives on protein function that may be useful in evaluating the coherence of proposed modules.

In this chapter we performed an exploratory analysis to find the best method of quantifying the functional homogeneity for the evaluation of communities in PINs.

Starting with the most comprehensive functional annotation database, the GO, we tested the suitability of the different branches of GO annotations for module evaluation in Section 4.2. By investigating the distribution of GO annotations on PINs, we found that GO biological process annotations best describe the functional characteristics of proteins likely to be shared in functional modules.

The most common method of assessing the biological coherence of a community using functional annotations is functional enrichment (eg. [22, 24, 25, 27, 49, 50, 174–

176]). In Section 4.3 we tested functional enrichment and show that this method is not suitable as an evaluation method for communities in networks.

The alternative to functional enrichment is functional homogeneity calculation via semantic similarity measures that compute the functional similarity of proteins. To find the optimal semantic similarity measure for module evaluation, we compared semantic similarity measures on simulated PIN and functional annotation data in Section 4.4.

Finally, we tested whether other functional annotation resources, specifically the HPO, can provide added value in the evaluation of communities as functional modules (Section 4.5).

4.2 Testing GO annotations

We tested whether annotations from the GO [154], the largest annotation database, are suitable to evaluate communities as modules. Annotations that describe modules should be distributed on PINs in such a way that groups of proteins which interact strongly share annotations. Hence, functional annotations suitable for community evaluation should be predictive of protein-protein interactions. Not all proteins that perform a common function are expected to interact, however it is not unreasonable to assume that the more biological processes two proteins have in common, the more likely they are to interact. This assumption underlies the models used for community detection in methods such as AGM [145] or BigCLAM [125] (see Section 2.3.3.1).

4.2.1 Methods

To test whether GO annotations are suitable for module evaluation, we calculated the probability of two proteins interacting in a PIN given that they share k functional annotations. This probability was approximated by the relative frequency specified as $P((u, v) \in E|G_u \cap G_v = k)$, where E is the set of edges represented by node tuples, u and v are proteins/nodes, and G_u denotes the GO term set associated with node u. As GO annotations for the same general characteristic exist at different levels of specificiy, we used GO slim data to obtain a broad overview of different annotation

categories. By reducing the number of possible annotations to a few high-level categories, more interactions are observed at each value of GO-term overlap k. Thus, we can estimate the probability of interaction with higher confidence.

4.2.2 Results

The probabilities of interaction of two proteins with k GO term overlaps are shown in Figure 4.1.

Figure 4.1 shows that while all GO sub-ontology annotations are predictive of protein-protein interactions, BP and MF annotations show the strongest signal. BP annotations showing a strong signal is expected from previous studies (eg. [79]). Surprisingly, MF annotations exhibit stronger signal than CC annotations. It has been previously argued that there is no reason to expect that MF annotations predict protein-protein interaction [94], and indeed MF is the only sub-ontology whose annotations could not be linked to a conserved network structure between species [209].

Overall, BP annotations appear best-suited to evaluate the coherence of a module according to our results and the literature [24, 28, 32, 109, 157]. This sub-ontology aims to capture the "biological objective" towards which a protein works [154], and it is these objectives that define protein modules [5].

4.3 Testing functional enrichment

The most popular way of assessing the coherence of a module is by functional enrichment of GO annotations (eg. [22, 24, 25, 27, 49, 50, 174–176]). Evaluating whether a functional annotation is more frequent than expected at random is an approach used commonly to biologically interpret lists of differentially expressed genes (DEGs) [109, 172, 210] and has been implemented in several freely available tools [172]. While the concept of randomness that is implemented in these tools is consistent with its uses for sets of DEGs, we argue that it is not suitable for network module evaluation.

As discussed in Section 2.5.1, given a set of proteins with functional annotations, functional enrichment evaluates whether any particular annotation occurs more

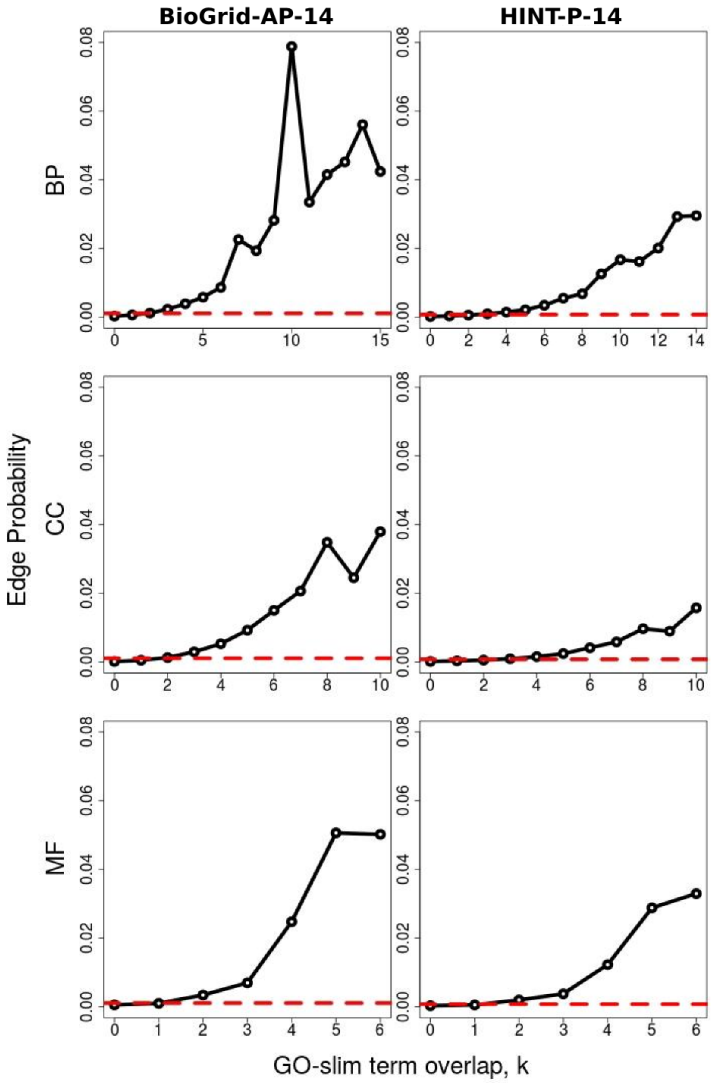

Figure 4.1: PIN description by GO annotations. The probability of interaction in a PIN was plotted versus the number of GO slim annotations shared k. Interaction probabilities were estimated based on relative frequencies calculated by dividing the number of proteins that interact and share k GO slim terms with the number of proteins that share k terms. The plots only contain data points where more than five interactions were observed. The red dashed lines show the density of the network, indicating the probability of interaction of two randomly selected proteins. "BP" refers to biological process annotations, "CC" to cellular component, and "MF" to molecular function.

frequently in this protein set than in a protein set of the same size picked uniformly at random. Thus, functional enrichment uses the null model that functional annotations are randomly distributed. As shown in Section 4.2 and again in Chapter 6 this null model does not hold for PINs. Proteins which interact are more likely to share functional annotations than those that do not (see Figure 4.1). Furthermore, community detection methods group proteins together based on

protein-protein interactions (see Section 2.3). Therefore, protein communities are expected to share more functional annotations than proteins selected uniformly at random. Any set of interacting proteins may be significantly enriched for a GO annotation under this null model.

4.3.1 Methods

To support the presented argument we investigated how functionally enriched random walks of different lengths on PINs are (see Section 2.2.2 for random walks). Here, the nodes visited by a random walk are used to represent a random community of interacting proteins. Each GO BP annotation associated with a protein in such a random community was tested for enrichment in the community by Equation (2.12). Multiple testing of the GO BP terms was corrected for using a Bonferroni correction, which represents a conservative approach to multiple testing (see Section 2.5.1.1). Rejection of the null hypothesis at significance levels α of 0.05, 0.01, and 0.001 were used to evaluate significant enrichment. If a random community was significantly enriched for any GO annotation by these significance thresholds, it was regarded as functionally coherent. 1000 random walks were performed from each network node to calculate the proportion of functionally coherent random communities in the PINs. This procedure was repeated for random walks of length 5 - 10, representing different random community sizes. The random walk length counts only the number of distinct nodes traversed (including the start node) to generate random communities of equal size.

4.3.2 Results

The proportion of functionally coherent random walks of different lengths is shown in Table 4.1.

Table 4.1 shows that at the commonly used significance level of $\alpha = 0.05$ up to $\approx 54\%$ of random communities of different sizes are evaluated as functionally coherent using functional enrichment. While it is likely that some of the random walks performed represent biologically meaningful communities, this fraction is

Table 4.1: Fraction of random communities that are significantly functionally enriched.

Network	Walk Length	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
HINT-P	5	0.414	0.278	0.149
	6	0.443	0.308	0.161
	7	0.473	0.329	0.174
	8	0.499	0.344	0.186
	9	0.521	0.359	0.194
	10	0.539	0.372	0.202
BioGrid-AP	5	0.373	0.220	0.099
	6	0.406	0.245	0.110
	7	0.435	0.268	0.119
	8	0.462	0.286	0.128
	9	0.487	0.303	0.137
	10	0.508	0.317	0.146

Table 4.1: GO BP term enrichment was assessed in random walks on HINT-P and BioGrid-AP. After Bonferroni correction for multiple testing of GO BP terms, a random walk was evaluated as functionally coherent if at least one GO BP term was enriched at the given significance levels, α. Using 1000 random walks per network node and length, the fraction of functionally coherent random walks was assessed.

unlikely to be as high as observed in the results. Furthermore, the fraction of random communities that are significantly functionally enriched can be seen to increase with community size. In this investigation the random walk communities sample the distribution of annotations between interacting proteins in the network (see also random walk PIN investigation in Section 6.2). The longer these random walks are, the better this distribution is sampled. Thus, the difference between the null model used in functional enrichment and the actual distribution of functional annotations on the network becomes clearer the larger the random walk communities.

While other methods of evaluating communities in networks may also assess some of these random walks as functionally significant, they are not subject to the same systematic bias evident using functional enrichment (see Chapters 5 and 6). The oversimplification of the distribution of functional annotations on PINs that is a central assumption in functional enrichment, renders this method unsuitable for evaluating communities in a network context. However, the use of functional enrichment for the labelling of communities with functional terms (eg. [211]) is not problematic as it does not represent an assessment of the community's quality.

4.4 Selecting a semantic similarity measure

As functional enrichment is unsuitable for functional homogeneity calculations of communities in PINs, we looked to semantic similarity measures to capture the coherence of proteins in a community. A large variety of semantic similarity measures exist (see Section 2.5.2). In this section we describe how the semantic similarity measure best suited to functional module evaluation was selected.

Semantic similarity measures use functional annotations to assess the similarity of proteins (see Section 2.5.2). As functional similarity cannot be experimentally measured, semantic similarity measures are compared using approximations of functional similarity. Thus, the results of comparative studies are limited to applications on particular data sets (see Section 2.5.2.2).

In this project we used semantic similarity to evaluate communities proposed as functional modules in PINs (module evaluation). To our knowledge a semantic similarity comparison based on functional module data does not exist, and only few comparisons between individual methods have been made using pathways or complexes [157]. To assess which semantic similarity measure is best suited for module evaluation we performed our own comparison of semantic similarity measures on simulated data.

It is impossible to compare all semantic similarity measures due to their sheer number. Thus the review in [157] focused on the characteristics that well-performing semantic similarity measures have in common. The authors concluded that measures which directly compare groups of annotations ("group-wise methods") are preferable to those which compare individual terms, and that measures based on information content (IC-based) tend to outperform others. Thus, for our comparison we focused on semantic similarity measures that incorporate these characteristics (see Section 2.5.2).

Despite having been successfully used to identify functional modules in the yeast PIN [48], to our knowledge the Pandey measure [181] does not appear in any comparative studies. We compared this group-wise, IC-based measure to others that performed well in comparative studies. One such measure that has been

consistently shown to outperform others across comparison methods [94, 157, 179] is simGIC [182]. It has however been suggested that this is due to it mostly being the only compared measure that is both IC-based and group-wise in the relevant studies [157]. We also considered simUI as one of the first group-wise measures to be used for functional similarity evaluation [179]. simUI compared favourably to more complex methods in a large-scale comparison [94].

In our comparison test we simulated PINs using a generative network model based on functional annotations. These annotations were extracted from a simple, conceived ontology. The simulated PIN was partitioned into communities using multi-resolution link clustering (see Section 2.3.3.2). Functional homogeneity evaluation of the obtained communities was performed based on the extracted annotations. In order to introduce noise into the functional annotation data, annotations were hidden from the semantic similarity measures based on a functional annotation model. The semantic similarity measures were scored by how well they evaluated network partitions based on hidden functional annotations in comparison to ground-truth evaluations using the full annotation set.

4.4.1 Methods

4.4.1.1 A generative network model

Commonly used generative network models have not been able to recreate the network structure observed in PINs (see Section 2.2.2). In order to carry out our tests, we extended a simple generative network model to incorporate noise on the edges and obscure functional annotations. These features were added to approximate the error rates in PINs, and the lack of functional annotation coverage and specificity, which are the main sources of error in the two types of data used in functional module detection.

As shown in Section 4.2 functional annotations describe PIN connectivity. Given two random proteins with functional annotations, the more annotations these two proteins share, the higher the probability of finding that they interact. This concept is implemented in the Affiliation Graph Model (AGM) [145] (see Section 2.3.3.1).

The AGM is designed to model the probability of two nodes interacting based on the number of communities they share. In order to generate network connectivity from ground truth functional annotations, we simplified the model and substituted the number of communities for the number of common functional labels in the AGM Equation (2.8) to give:

$$P(u,v) = c_p(1-(1-p)^{|G_u \cap G_v|}), \qquad \text{for} \quad u \neq v. \tag{4.1}$$

Here, $P(u,v)$ denotes the interaction probability between nodes u and v, p is the probability scaling factor that controls how the interaction probability scales with shared labels, and G_u is the label set associated with node u. The constant factor c_p in this model is used to regulate the expected number of edges in the network generation process.

The edge probabilities calculated by Equation (4.1) are based on ground truth label sets denoted by G_u. We generated these label sets by randomly sampling from the simple, conceived ontology shown in Figure 4.2.

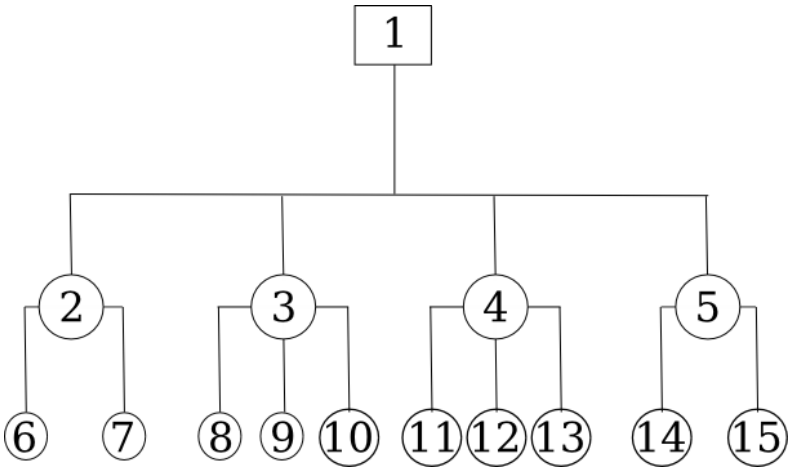

Figure 4.2: Conceived ontology tree for ground-truth labels. The ontology tree describes the relations between the functional labels used as ground-truth in our simulation study. Label 1 denotes the root term, labels 2 - 5 represent an intermediary level of organization, and labels 6 - 15 are the most specific annotations. In analogy with the GO, "is_a"-type relationships go downwards in the ontology.

In the case of PINs, although the most specific function a protein is associated with is not always known, it is this specific function that determines a protein's interactions. In analogy with the biological case, we sampled ground-truth labels

only from the most specific level of the ontology at labels 6 - 15. Given a number of nodes N, an expected number of edges $|E|$, a probability scaling factor p, and a distribution of labels per node D, the algorithm used to generate our simulated PIN-like networks performs the following steps:

1. For each node, Sample M labels from 6 - 15 with replacement, where M depends on the input distribution for the number of labels per node D.

2. Generate the full label sets via our ontology using ancestral path mapping.

3. Calculate the interaction probability matrix P by Equation (4.1).

4. Scale the interaction probabilities in matrix P to have the expected number of edges $|E|$.

5. Run the SourceGraph function (Algorithm 4.4.1) to generate a connected network.

For example, given a distribution D of 50% of nodes with 1 and 50% of nodes with 2 labels, a label distribution for $N = 4$ nodes (nodes A, B, C, and D) could be sampled as $G_A = \{6\}$, $G_B = \{11, 8\}$, $G_C = \{11, 9\}$, $G_D = \{12\}$. Based on our ontology tree (Figure 4.2) the full label sets for these nodes are $G_A = \{6, 2, 1\}$, $G_B = \{11, 8, 4, 3, 1\}$, $G_C = \{11, 9, 4, 3, 1\}$, $G_D = \{12, 4, 1\}$. For a probability scaling factor $p = 0.1$, the interaction probabilities are $P_1 = P[A, B] = P[A, C] = P[A, D] = 0.1c_p$, $P_2 = P[B, D] = P[C, D] = 0.19c_p$, and $P_3 = P[B, C] \approx 0.34c_p$ by Equation (4.1). For an expected number of edges $|E| = 3$, the probabilities are scaled to $P_1 \approx 0.29$, $P_2 \approx 0.56$, and $P_3 \approx 1.00$.

Sampling edges from the probability matrix between nodes P may result in the edge set $E = \{(B, C), (C, D), (B, D)\}$, which would not give a connected network. Therefore, we generate connected networks using the SourceGraph

function. The pseudocode to generate a connected network from the calculated interaction probabilities P is:

Algorithm 4.4.1: SourceGraph(N, P)

comment: sample(v,p): draw a sample from the elements of vector v with weights p
comment: graph(A): generate an undirected graph from the adjacency matrix A
comment: LCC(G): filter for the largest connected component of graph G
comment: V(G): the number of nodes in the graph G

repeat
  for $i \leftarrow 1$ **to** $N - 1$
  $\begin{cases} \textbf{for } j \leftarrow i + 1 \textbf{ to } N \\ \quad \textbf{do} \\ \{A[i, j] \leftarrow \text{sample}((0, 1), (1 - P[i, j], P[i, j])) \end{cases}$
  $G \leftarrow \text{graph}(A)$
  $G_{LCC} \leftarrow \text{LCC}(G)$
until $\frac{N - V(G_{LCC})}{N} \leq 0.01$

Using this algorithm, a connected network can be generated from specific ontology labels with an expected edge count $|E|$, and approximately N nodes. Here, we accepted any connected network where the number of nodes was within 1% of the target number N.

PINs are noisy (see Section 2.1.2.1). In this network model, noise is modelled by a simple approximation. As all nodes share at least the root annotation, $|G_u \cap G_v| \geq 1$ by construction (see Equation (4.1)). Thus, $|G_u \cap G_v| = 1$ is non-informative and describes proteins that interact "at random" with probability $p \times c_p$. We view these interactions as "noise", whereas interactions with $|G_u \cap G_v| > 1$ are viewed as "signal".

The relative size of the noise term is controlled by the probability scaling factor p from Equation (4.1). By this equation, the maximum signal is an interaction probability of c_p. Thus, the highest signal-to-noise ratio possible for any data set is $\frac{1}{p} : 1$. This optimal signal-to-noise ratio translates to a network where $\frac{1}{1+p}$ interactions represent signal, and $\frac{p}{1+p}$ result from the noise effect. Such a network has a false-positive rate of $\frac{p}{1+p}$. In our investigation p was set to 0.2, which corresponds to false-positive rates of $\approx 17 - 21\%$ observed in high quality protein-protein interaction data sets [79, 89].

4.4.1.2 Functional annotation model

The specific function of a protein that determines its interactions with other proteins is often not known. This circumstance was translated into our simulation setup by hiding node labels from the semantic similarity measures. In order to simulate the information a semantic similarity measure would have available when applied to PIN communities, a certain percentage of labels were hidden. Specifically, two label-hiding probabilities, $p_{H1} = P(\text{remove label at depth 2})$ and $p_{H2} = P(\text{remove label at depth 1}|\ \text{label at depth 2 was removed})$, were used to affect label hiding at the two depths in our ontology.

In our ontology (Figure 4.2) a node association with a label at a depth of two (6 - 15) defines the node's associations with depth one labels (2 - 5) and the root. Thus, a label at a depth of one can only be hidden if the corresponding depth two label was already hidden. This logic was applied in the label-hiding algorithm, shown in Algorithm 4.4.2.

Algorithm 4.4.2: LABELHIDER$(N, \text{labels}, p_{H1}, p_{H2})$

comment: labels: length N list containing the full label set for each node
comment: rand(): generate a random number in the interval [0,1]
for $i \leftarrow 1$ **to** N
do
  labelsLeft[i] = labels[i]
  for each lab $\in$ labelsLeft[i]
  do
    if (lab $\in [6:15]$ **and** rand() $< p_{H1}$)
    then
      Remove lab from labelsLeft[i]
      if (rand() $< p_{H2}$)
      then
        Remove parent lab from labelsLeft[i]

For example, given our network of four nodes with label sets $G_A = \{6, 2, 1\}$, $G_B = \{11, 8, 4, 3, 1\}$, $G_C = \{11, 9, 4, 3, 1\}$, $G_D = \{12, 4, 1\}$, we can hide labels with probabilities $p_{H1} = 0.5$ and $p_{H2} = 0.33$. Using the LabelHider function with these parameter sets an estimated 50% of the depth two labels of $[6, 11, 8, 11, 9, 12]$ would be hidden. If the most specific labels of nodes A and B are selected ($[6, 11, 8]$),

this leaves the depth one labels $[2, 4, 3]$ which can be selected for removal with the probability $p_{H2} = 0.33$. At this probability, on average one of the three labels will be removed. If the label "4" is chosen, this leaves the label sets $G_A = \{2, 1\}$, $G_B = \{3, 1\}$, $G_C = \{11, 9, 4, 3, 1\}$, $G_D = \{12, 4, 1\}$ for functional similarity evaluation.

By hiding node labels semantic similarity measures can be compared in their ability to evaluate communities in simulated networks given only partial label information in contrast to ground-truth community evaluation based on full label information.

4.4.1.3 The comparison test

Using our network model to generate a PIN-like network, and the functional annotation model to simulate the data available to semantic similarity measures in module evaluation, we developed a comparison test for semantic similarity measures.

Given a parameter set, 100 networks were generated by our network model. Each of these networks was generated for independently sampled node label sets. As the network model generates overlapping communities (see AGM in Section 2.3.3.1), the networks were partitioned into communities using link clustering [127] (see Section 2.3.3.2) for 251 resolutions evenly spanning the interval S $\in [0; 0.5]$.

After label hiding, the pairwise functional similarities of all proteins in non-trivial communities (size > 2) was calculated using the three semantic similarity measures. To calculate the quality of a network partition generated at a particular resolution, the functional similarity scores in non-trivial communities at the same resolution were used to compute the network partition functional homogeneity score using the equation:

$$\mathrm{fh}_{SS,r} = \frac{1}{|C_r|} \sum_{c \in C_r} \frac{2}{|c|(|c|-1)} \sum_{u,v \in c;\, v>u} \rho_{SS}(u, v). \tag{4.2}$$

Here, $\mathrm{fh}_{SS,r}$ is the network partition functional homogeneity score at a resolution r by a functional similarity measure SS. The set of communities of size > 2 at resolution r is denoted by C_r, u and v are nodes in a community c in this set, and $\rho_{SS}(u, v)$ is the semantic similarity of the two nodes by measure SS.

Equation (4.2) was derived by taking the averag of all community functional homogeneity scores (Equation (2.16)) at a single resolution. To assess how well communities are evaluated, each community was weighted equally regardless of size in this calculation. This processing leads to profiles of network partition functional homogeneity across resolutions for the three semantic similarity measures on each generated network (Figure 4.3).

The ground-truth data these functional homogeneity profiles were compared to is the full node label set used to calculate the initial edge probabilities for network generation. Given that no labels were hidden in this set, the average number of labels shared between all node-pairs in a community denotes how similar the nodes in this community are. Using this average overlap score as a ground-truth functional homogeneity, ground-truth network partition functional homogeneity profiles were calculated by Equation (4.2). These profiles are shown in Figure 4.3 for a network generated at a label number distribution of 70%-30% for one and two labels respectively, $p_{H1} = 0.33$, and $p_{H1} = 0.2$.

The comparison in Figure 4.3 was quantified using the Pearson correlation between the individual semantic similarity profiles and the ground-truth profile. The semantic similarity measures were then compared for a specific parameter set using the distribution of the Pearson correlation coefficients of the 100 networks generated. Taking the mean of the each Pearson correlation coefficient distribution we obtained our test statistic describing the performance of a semantic similarity measure in our simulation test for a specific parameter set. Summarizing this process, the test statistic is calculated by the equation:

$$\text{Test statistic}(SS) = \frac{1}{100}\sum_{i=1}^{100} corr(\mathbf{FH}_{SS}, \mathbf{FH}_{GT}), \tag{4.3}$$

where $\mathbf{FH}_{SS}$ denotes the vector with the elements $\text{fh}_{SS,r}$ representing the network partition functional homogeneity scores at each resolution, r, calculated by Equation (4.2). Here, $corr()$ denotes the Pearson correlation coefficient, SS represents a given semantic similarity measure, and GT is the ground truth evaluation measure based on full label overlaps.

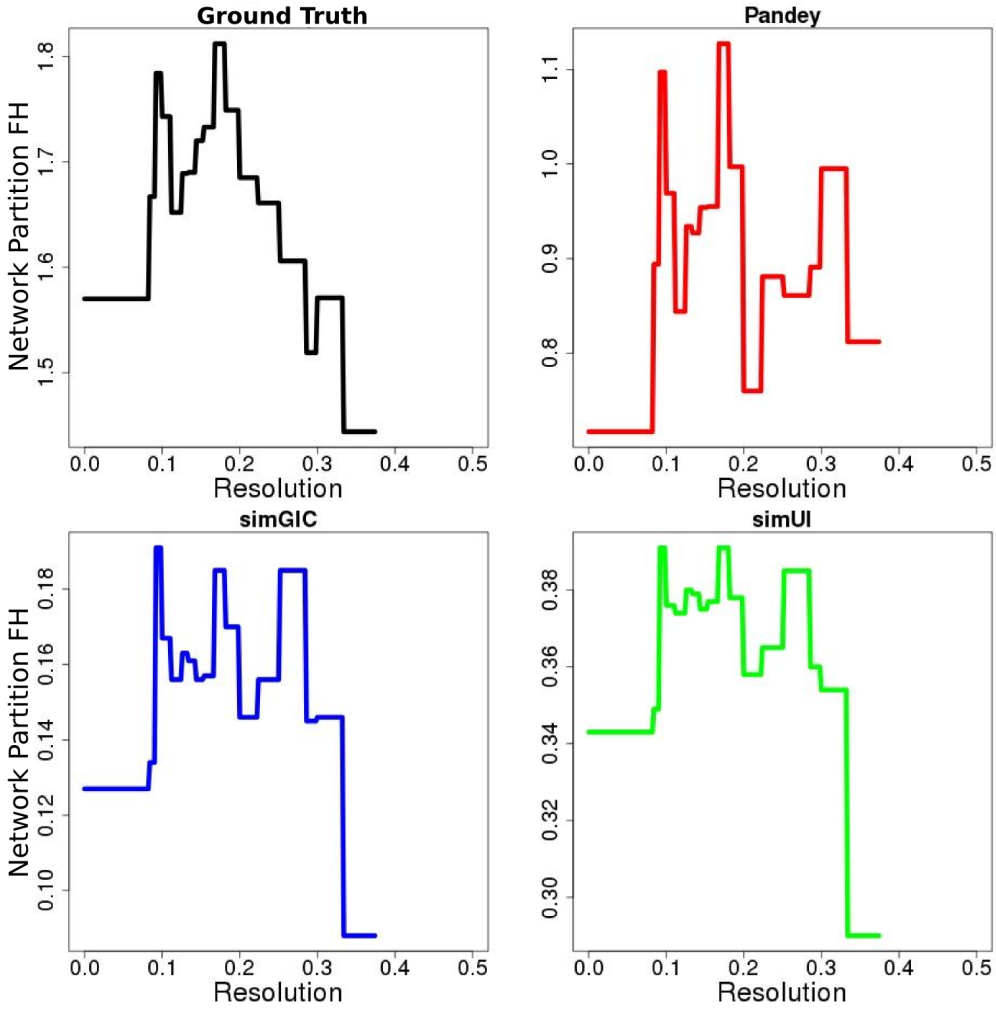

Figure 4.3: Network partition functional homogeneity profiles. Network partition functional homogeneity profiles for the ground truth evaluation measure, the Pandey measure, simGIC, and simUI. The network was generated with a target of 500 nodes, an expected edge number of 998, a 70%-30% proportion of nodes with one and two labels, and a probability scaling factor $p = 0.2$ by our generative network model. The three semantic similarity measures were used to compute the network partition functional homogeneity profiles by Equation (4.2) using labels preprocessed by label-hiding Algorithm 4.4.2 with parameters $p_{H1} = 0.33$, and $p_{H1} = 0.2$. Link clustering was used to partition the network at resolution increments of 0.002 in the interval $S \in [0; 0.5]$. Pearson correlation coefficients of these profiles were used to quantify the success of a semantic similarity measure in the context of module evaluation.

4.4.2 Results

Using the models and the test described in Section 4.4.1 we evaluated how simUI [183], simGIC [182], and the Pandey measure [181] compare in the functional homogeneity evaluation of communities in networks.

As many of the parameters in our models cannot easily be fitted using published results, we performed the comparison under a range of parameter settings to assess the reliance of our results on the parameter inputs. The parameters which were varied in this investigation are the number of labels per node and the label hiding probabilities p_{H1} and p_{H2}. Here, the target number of nodes, N, was set to 500. For a network of 500 nodes to be connected, at least 499 edges are required. Thus, the minimum possible density is ≈ 0.004, which is approximately five times as large as standard PIN densities (see Table 3.1). In order to reproduce PIN-like densities on a connected network that allows for sufficient edges to reproduce a modular structure at least 4000 nodes are required. This network size was found to be unfeasible for performing parameter sweeps of the above parameters given the computing power available.

To allow for the generation of coarse modular network structure on 500 nodes the expected edge number was set to twice the minimum number of edges required for a connected network. Due to random sampling of edges and taking the largest connected component of the generated networks, the actual number of nodes and edges varied between networks.

In December 2013, the GO contained 22,483 leaf terms in the ontology structure, spread across the three sub-ontologies. In comparison, the most GO-terms associated with a single protein in the November 2013 human protein association file was 147 (see Section 3.2.1). While these 147 represent the most specific annotations associated with the protein in a particular publication, it is unlikely that all of these annotations are leaf terms. Even in the scenario, where these 147 do all represent distinct leaf term annotations, the maximum number of leaf term annotations assigned to a protein is still $< 1\%$ of all leaf term annotations. As our ontology tree only contains ten leaves, it is impossible to translate this proportion directly.

Instead, the proportion of nodes with one and two labels was varied at 60%-40%, 70%-30%, and 80%-20% respectively. Two labels were allowed as most proteins are not limited to a single function.

Given that our ontology tree only contains three levels of specificity of which only two contain information relevant to node interactions, hiding labels on a single level already represents a loss of relevant information of 50%. Hiding labels across two levels can thus be interpreted as a lack of node annotation in analogy to PINs. The proportion of PIN proteins without functional annotations for the PINs used throughout this project is on average $\approx 15\%$ (see Sections 3.2.1 and 3.2.1). This percentage corresponds to the probability $p_{H1} \times p_{H2}$ in the functional annotation model. We thus performed the comparison test for p_{H1} values of 0.25, 0.33, and 0.5, and p_{H2} values of 0.1, 0.2, and 0.3. For these parameter values the expected proportion of nodes without relevant label associations ranges from 2.5% to 15%. Due to the complexity of the GO structure (difficulty of assigning depth to a term; different maximum depths across ontology; see Section 2.4.1), estimating p_{H1} from the GO is difficult. Thus, a broad range of p_{H1} values were tested.

The test statistic across parameter sets for the Pandey measure, simGIC and simUI, can be seen in Figures 4.4–4.6.

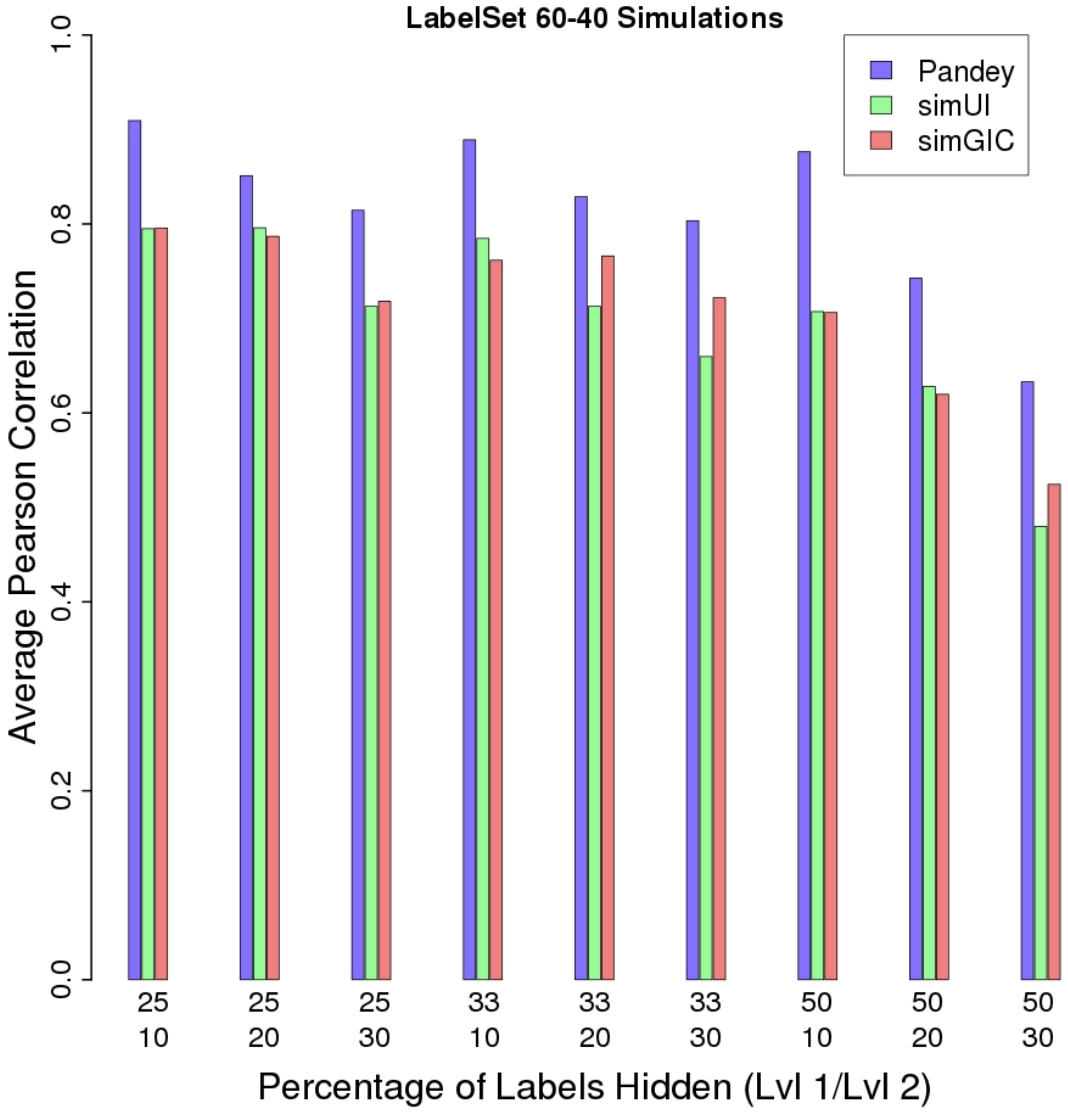

Figure 4.4: Semantic similarity test statistics for parameter sets with node label distributions of 60%-40%. Functional module simulation test statistics for the Pandey measure, simGIC, and simUI were calculated by Equation (4.3). The simulated networks were generated using the model described in Section 4.4.1 with a target node number of 500, an expected edge number of 998, a probability scaling factor of 0.2 and 60% of nodes with one label, and 40% with two. Labels were hidden using the label-hiding algorithm 4.4.2 at the parameters shown on the x-axis. The Pandey measure outperforms simGIC and simUI for all investigated parameter sets.

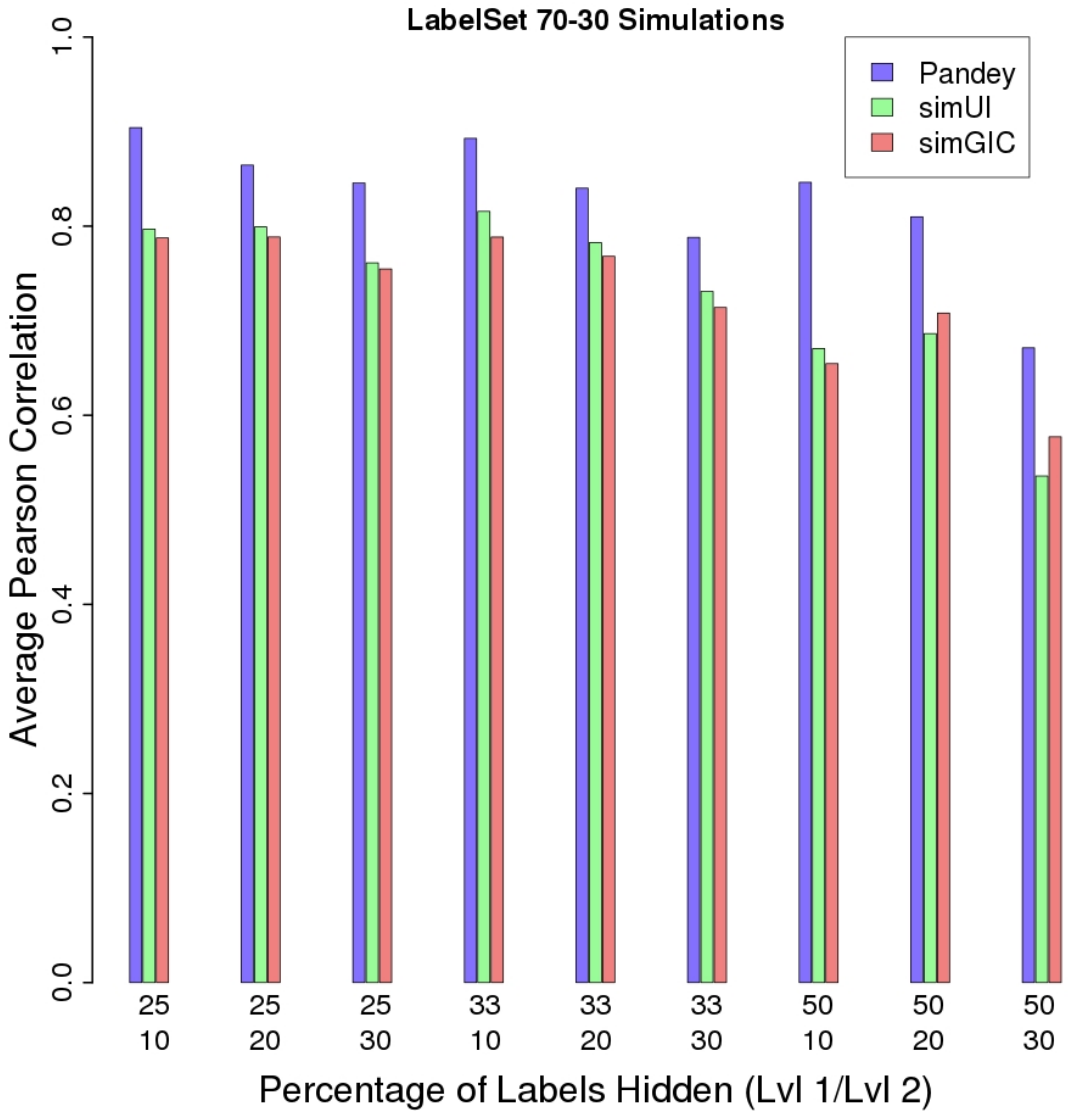

Figure 4.5: Semantic similarity test statistics for parameter sets with node label distributions of 70%-30%. Functional module simulation test statistics for the Pandey measure, simGIC, and simUI were calculated by Equation (4.3). The simulations were set up as in Figure 4.4 with 70% of nodes assigned one label and 30% two. The Pandey measure outperforms simGIC and simUI for all investigated parameter sets.

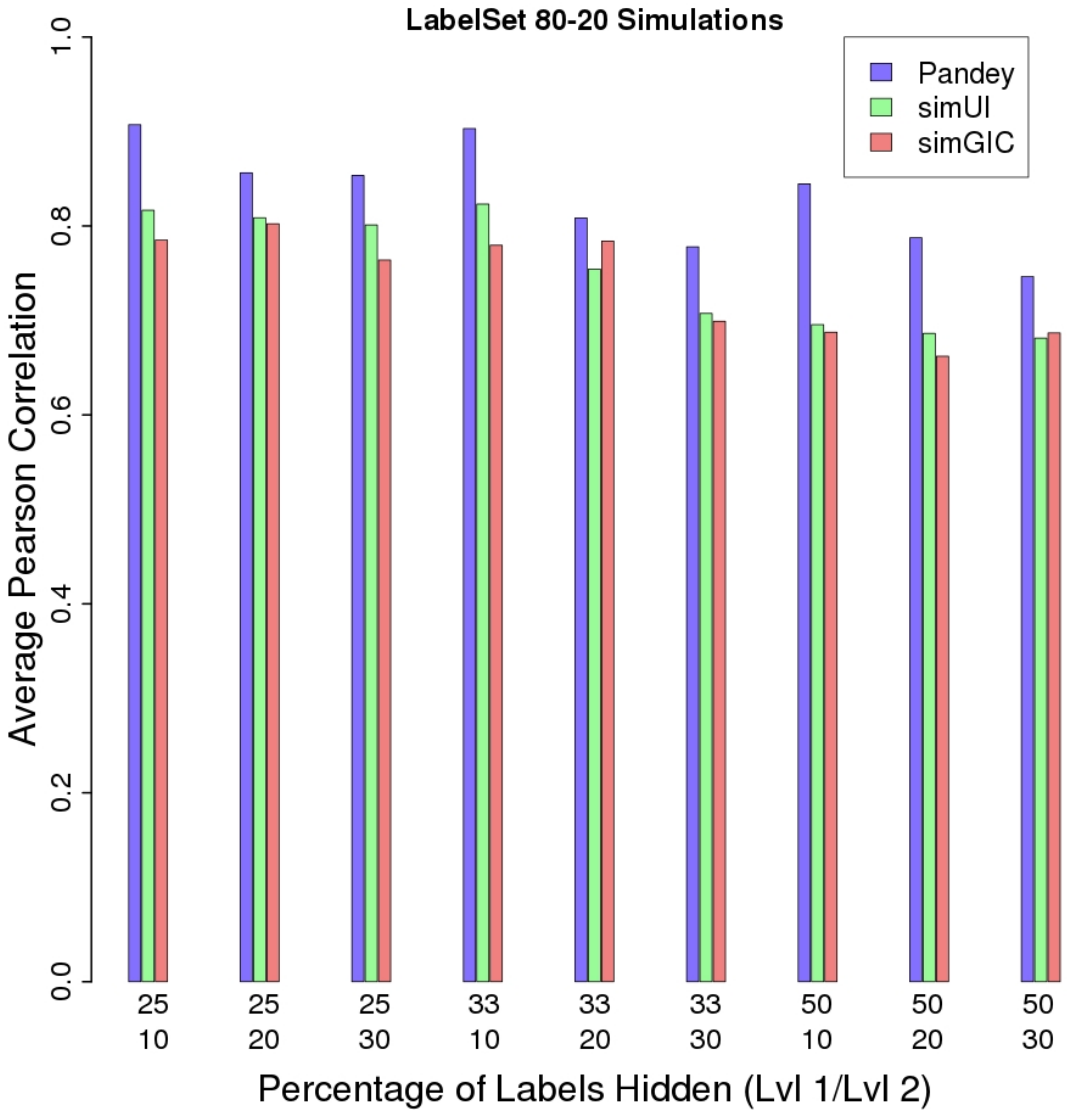

Figure 4.6: Semantic similarity test statistics for parameter sets with node label distributions of 80%-20%. Functional module simulation test statistics for the Pandey measure, simGIC, and simUI were calculated by Equation (4.3). The simulations were set up as in Figure 4.4 with 80% of nodes assigned one label and 20% two. The Pandey measure outperforms simGIC and simUI for all investigated parameter sets.

Figures 4.4–4.6 show that the Pandey measure consistently outperforms simUI and simGIC. In fact, the more difficult the simulated circumstances for community evaluation are, the more the Pandey measure stands out in the comparison. For example, the Pandey measure can be seen to outperform simUI and simGIC by the larest margin when an expected 50% of specific (depth 2) labels are hidden, and thus over 50% of the relevant label information is missing. Although coverage of annotations on proteins is difficult to assess, it is likely that even 50% is a

conservative estimate for the current coverage of specific functional annotations on proteins in PINs [158].

4.4.3 Discussion

In this section we presented a simulation test that assesses how well a semantic similarity measure captures the ground truth evaluation of communities given that it lacks information on the exact labels. The semantic similarity measures simUI [183], simGIC [182], and the Pandey measure [181] were compared in this simulated scenario. Despite simGIC comparing favourably to a range of semantic similarity measures in previous comparisons [94,157,179], we found that the Pandey measure outperforms simUI and simGIC across parameter sets in our simulations. To our knowledge, these simulations represent the first attempt to compare semantic similarity measures in the context of functional module evaluation.

There are several ways in which the simulations could have been extended to incorporate further factors that may affect functional module evaluation in PINs. These possible extensions are:

1. Nodes without annotation that are excluded in the functional homogeneity calculations could have been included by adding a p_{H3} parameter;

2. Random noise could have been introduced less evenly across the network. A skewed noise distribution is expected in PINs given not all proteins are equally well-studied;

3. Label-hiding could have been modified to generate a skewed distribution of labels on the network as observed in PINs (see Section 6.2).

While these extensions can make the simulation scenario more realistic, they also increase the complexity of the simulation. Extensions to the models in the above-mentioned way require extensive parameter sweeps which are not computationally feasible, especially given the multi-dimensional nature of proposed parameters 2 and 3.

Further possible sources of error in our simulations are the chosen community detection method, the label hiding method, and a possible loss of information in the aggregation of community functional homogeneity scores into a single test statistic. Link clustering was chosen as a fast algorithm that has been previously applied to PINs [127], but other approaches to the community detection problem are available (see Section 2.3). In rare cases our label hiding method can lead to a specific annotation without a parent annotation in the label set. As nodes can be assigned two labels, it can occur that a node is associated with two specific labels with the same intermediate label (for example 6 and 7 with intermediate label 2). In this scenario, label 2 can be hidden if label 7 is hidden when 6 is not. In our simulations this is expected to occur between 0.6 and 4.8 times in each 500 node graph and can thus be neglected. The test statistic calculation (see Equation (4.3)) is a possible source of error due to multiple averaging steps hiding the variance of community functional homogeneities. It may be the case that semantic similarity measures perform better in different circumstances depending on e.g. community size, or community label depth distributions. By averaging across these possible dependencies we test for global performance, instead of verifying individual community cases. As the same averaging procedure is applied to all investigated semantic similarity measures, the comparison is fair.

Despite efforts undertaken to simulate functional module evaluation accurately, it is possible that the aforementioned PIN characteristics which were not accounted for in our models, and other described sources of error, confounded the results of our investigation. As a consequence, the simulation should not be taken as hard evidence of a semantic similarity measure's superiority over others in the evaluation of functional modules. However, this caveat holds for any comparison of semantic similarity measures based on simulations or data approximations (which encompasses all the comparisons in [94, 157, 179]). The comparison performed can only suggest that the Pandey measure is the most suitable of the three measures tested in the context of module evaluation.

4.5 Testing Human Phenotype annotations

Despite the GO being the most comprehensive annotation database, it does not offer perfect coverage. Using additional functional annotation resources may thus improve our ability evaluate communities as functionally homogeneous. Functional annotations can also be extracted from sources such as MIPS [166,167], Reactome [164,165], KEGG [163], or the HPO [162] (described in Section 2.4.3). Pathway annotations, as available from Reactome or KEGG, and MIPS annotations are automatically mapped to the GO with "IEA" evidence codes and thus may not provide additional information (see Section 2.4.3). As such, HP associations are the best candidates for independent community evaluation. We investigated whether modules classified by Mendelian disease annotations mapped to the HPO could extend the list of functionally homogeneous communities by GO-based functional evaluation.

4.5.1 Methods

Less than 20% of protein in HINT-P and BioGrid-AP are associated with HP annotations (see Section 3.2.2). In order to use HP annotations to evaluate communities, they must have sufficient coverage on these communities. We assessed the extent of this issue for module evaluation by investigating the distribution of the number of annotated nodes in communities of size 6 - 35 in HINT-P and BioGrid-AP (Figure 4.7). Communities were generated by link clustering at 121 resolutions evenly spanning the interval $S \in [0, 0.6]$ (see Section 2.3.3.2). We specifically focussed on communities of size 6 - 35 as these represent size ranges where potential modules are unlikely to represent trivial associations or small complexes, but can still be feasibly experimentally tested (see Section 1.1.3).

To further assess whether HP annotations provide any added value in the evaluation of community functional coherence we investigated whether there were functionally homogeneous communities by HP functional annotations that were not evaluated as homogeneous by the equivalent GO BP annotation-based analysis in HINT-P. Functional homogeneity scores were calculated using the Pandey measure [181] with Equation (2.16). A community was evaluated as functionally

homogeneous if its functional homogeneity score exceeded that of the mean functional similarity of interacting proteins in the PIN.

4.5.2 Results

The distributions of the number of annotated nodes in communities of different sizes (Figure 4.7) shows that a substantial proportion of communities are not sufficiently annotated with HP terms. Indeed, over 50% of all communities of size 6 - 35 have less than two annotated nodes so that their functional homogeneity cannot be calculated. For communities whose functional homogeneity can be determined, few protein pairwise similarities determine the functional coherence of the whole community in most cases. This situation is akin to sampling from all pairwise protein similarities in a community and assuming the sample is representative of the entire community – an assumption which will not always hold true.

As Figure 4.7 shows that some communities exhibit good coverage, especially for the smaller community sizes, we investigated whether there are communities that are evaluated as functionally homogeneous by HP-based community evaluation that are not otherwise evaluated as homogeneous using GO-based analysis (Figure 4.8).

Figure 4.8a shows that up to approximately 105 communities of size 6 - 35 are only evaluated as functionally homogeneous using HP annotations. Taking only communities that contain at least three proteins associated with HP annotations, this number is still above 70. These figures represent approximately 55% of the total number of communities in this size range that are evaluated as functionally homogeneous by the HP-annotation-based analysis. While this result suggests that HP annotations can provide additional information for the evaluation of communities as modules, Figure 4.8b indicates that these values are likely the result of the poor coverage of HP annotations on the communities.

Figure 4.8b shows that communities that are evaluated as functionally homogeneous by only the HP-based analysis have a low proportion of annotated nodes. Thus, the communities that do exhibit a good annotation coverage in Figure 4.7 are

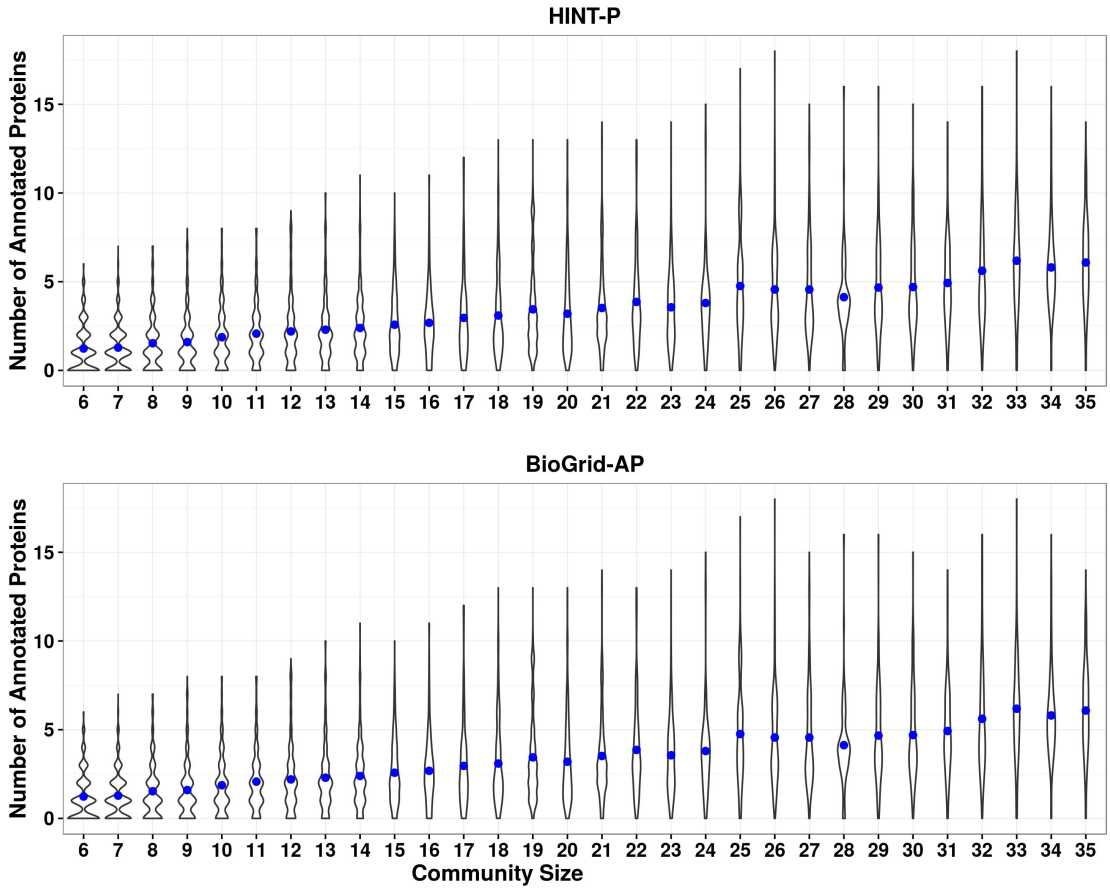

Figure 4.7: Number of HP annotated proteins in link clustering communities of size 6 - 35. Violin plots show the distribution of the number of proteins with HP annotations in link clustering communities across resolutions of sizes 6 - 35 in HINT-P and BioGrid-AP. The distribution of the number of annotated nodes at each community size is mirrored to give a violin plot. Although the data is discrete the plots are smoothed, generating the undulations specifically visible in small community sizes. The blue dots show the mean of each distribution. The mean number of annotated nodes in communities is $\approx \frac{1}{6}$ of the community size showing a poor coverage of HP annotations.

not the communities that are evaluated as functionally homogeneous by only HP-based community evaluation. Indeed, 12 of 3458 communities that are functionally homogeneous by only HP-based analysis have over 75% annotated nodes. These observations suggest that the functional homogeneity values obtained from the communities in the HP-only set may be unreliable due to few pairwise protein similarities determining the functional homogeneity of the whole community. Functionally homogeneous communities by HP-based analysis that do have an adequate coverage of functional annotations tend to be confirmed by GO-based analysis.

While there may be additional information in HP annotations that can be used

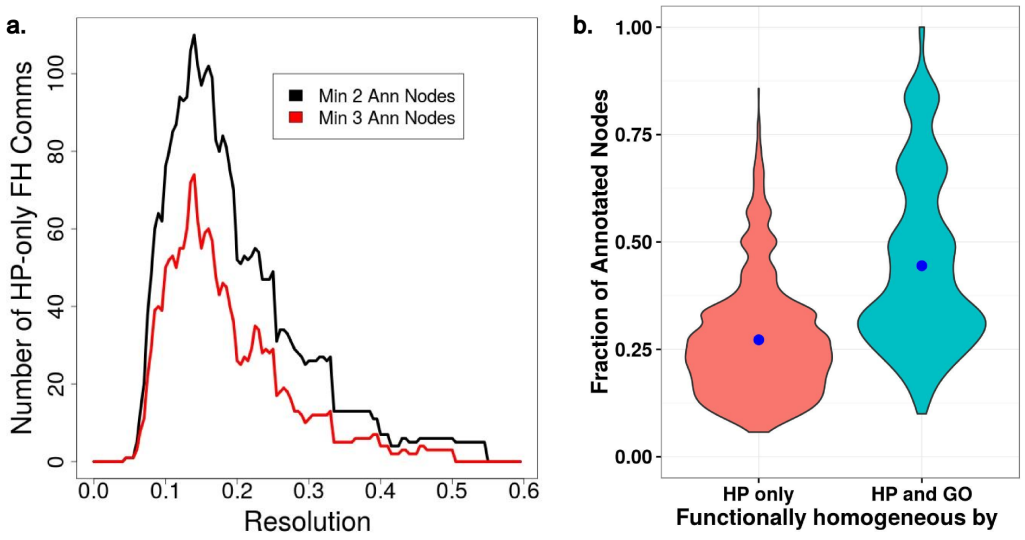

Figure 4.8: Comparison of HP-based and GO-based community analysis. a) Number of functionally homogeneous communities found by HP-based community evaluation that were not found by the GO-based analysis. The number of functionally homogeneous communities with at least 3 annotated nodes is shown in red. b) Violin plots of the distribution of the fraction of annotated nodes in communities evaluated as functionally homogeneous by HP community evaluation. The functionally homogeneous communities are split into two sets: communities that were only evaluated as homogeneous by HP-based analysis (HP only), and those that were evaluated as homogeneous by both HP-based and GO-based analyses (HP and GO). The functional homogeneity of link clustering communities on HINT-P was evaluated using the Pandey measure with HP annotations or GO annotations. A community was denoted as functionally homogeneous if its functional homogeneity score exceeded the average functional similarity of interacting proteins in HINT-P based on the respective functional annotations. While it appears that HP-based community evaluation enhances our ability to evaluate the functional homogeneity of communities, the distribution of annotated nodes in functionally homogeneous communities found only by HP-based analysis suggests their homogeneity scores may be unreliable due to a low coverage of annotated nodes. Communities with a higher proportion of annotated nodes tend to be found by GO-based community evaluation.

for community evaluation, this signal is difficult to detect due to the unreliability of calculated functional homogeneity scores from sparse annotation data. Overall, it appears that the low coverage of HP-term-associated proteins on PINs decreases its suitability for the evaluation of communities.

4.6 Discussion and conclusions

Evaluating the functional homogeneity of a community requires two components: a functional annotation resource and a method of using these functional annotations to compute a homogeneity score. In this chapter we performed an exploratory analysis to find the best combination of annotation data and functional homogene-

ity calculation to capture the biological coherence of communities proposed as functional modules.

The most popular method of community evaluation is using GO annotations with functional enrichment (eg. [22, 24, 25, 27, 49, 50, 174–176]). In a comparison of annotations from the three GO ontologies we found GO BP annotations to be most suitable for community evaluation, which corresponds to the consensus in the literature [24, 28, 32, 109, 157]. Surprisingly, MF annotations also showed a similarly strong signal. In contrast, functional enrichment was found to be unsuitable for the evaluation of communities in networks given its implied assumption that functional annotations are randomly distributed in PINs. While functional enrichment can be extended to address this issue as discussed in Chapter 6, currently semantic similarity represents the most appropriate approach for community evaluation. To determine which semantic similarity measure is best equipped to evaluate communities based on limited available functional annotations, we performed a simulation study. Contrasting three measures that had either been well-reviewed or previously used in the context of functional module evaluation, we determined that the Pandey measure is best suited to deal with the challenges of the evaluation of communities in PINs. Therefore, protein functional similarity was quantified by applying the Pandey measure to GO BP annotation sets in the work presented in Chapters 5 and 7.

As the GO does not offer perfect coverage of functional annotations, we investigated whether additional functional annotation resources could provide any added value for community evaluation. However, alternative annotation resources are either already incorporated into the GO (e.g. pathway annotations from KEGG or Reactome), or suffer from a low coverage on PINs resulting in unreliable community evaluation (HP annotations). Integration of GO BP with HP annotations for the evaluation of communities may be possible, however the potential contribution of HP annotations is unclear given the low coverage. It is therefore likely that the merits of such an integration will be outweighed by the effort required.

5

Attempting to identify functional modules in protein interaction networks

Contents

5.1 Introduction 107
5.2 Evaluating communities using functional homogeneity 109
5.3 Criteria for the evaluation of community detection methods 110
5.4 Louvain community detection 112
5.4.1 Configuration model 112
5.4.2 Constant Potts model 118
5.5 Overlapping community detection 121
5.5.1 BigCLAM 122
5.5.2 Link clustering 124
5.6 PIN comparison 128
5.7 Discussion and conclusions 130

5.1 Introduction

Functional modules represent an important level of organization in molecular biology (see Section 1.1.1). In this thesis we present a methodology that uses these modules to map DEGs to functions that are differentially regulated in disease phenotypes (see Section 1.1.3). To assess how best to detect functional modules, we performed an exploratory analysis of several functional module detection methods.

There exist a large variety of methods for functional module detection (reviewed in Section 1.1.2). Modules are commonly found by performing community detection on PINs [23–26, 48] or other networks of integrated biological data [38, 43, 46, 50]. An important consideration for the selection of community detection methods in this thesis is speed, and the ability to detect communities at multiple scales.

Functional modules exist at different scales of organization (see Section 1.1.3). Thus, we used community detection methods that take into account that communities can be found at multiple scales or resolutions by incorporating a resolution parameter (see Section 2.3.1).

While several studies have used multi-scale community detection to generate functional modules [37, 46, 48], to our knowledge only one study has previously investigated how these functional modules change across resolutions. Using configuration model Modularity Maximization to partition *S. Cerevisiae* PINs (see Section 2.3.2.1) it was concluded that there is no single resolution of interest, but instead cellular functions are best captured at different resolutions [48]. Given that configuration model Modularity Maximization is known to be subject to a resolution limit which bounds the community sizes that are correctly detected at any resolution [139–141] (see Section 2.3.2.2), this conclusion may be confounded by the community detection method used. In this chapter we thus attempt to reproduce this study on human data and assess the potential effect of the resolution limit in module detection. Further community detection methods are also evaluated.

In this chapter four different community detection were tested on two PINs at multiple resolutions. The initial aim of this investigation was to identify a community detection method and resolution (or narrow range of resolutions) that captures the functional organization of a human PIN. Finding no such "ideal" resolution or method we focused on the behaviour of the community detection methods that may limit their suitability for functional module detection. This focus was specifically centred around the detection of functionally homogeneous communities of sizes 6 - 35, which was suggested as an optimal size range for further investigation by collaborators at UCB Pharma (see Section 1.1.3).

Our exploratory analysis revealed several obstacles to functional module detection in PINs. These challenges are discussed in this chapter and addressed in further chapters of this thesis.

5.2 Evaluating communities using functional homogeneity

To compare the ability of community detection methods to partition PINs into functional modules, we must use a consistent evaluation measure. Following our functional homogeneity investigation in Chapter 4, we use GO BP annotations to characterize protein functional similarity with the Pandey measure in this chapter (see Section 3.2.1). This measure not only outperformed others in our simulation test in Chapter 4, it is also used in the multi-resolution yeast PIN study to which we compare our results [48].

Going from protein functional similarities to measures of community functional homogeneity can be done in two ways (see Section 2.5.2.3). While we opted to compute the homogeneity of a community by averaging all pairwise similarity scores in Chapter 4, we instead used the average similarities of interacting proteins in a community here. This method was chosen for consistency with the yeast PIN study whose conclusions we are evaluating on human data [48]. It has been argued that computing the functional homogeneity in this way takes into account that interacting proteins are more similar than non-interacting proteins (see also Figure 6.4 in Section 6.4.1). Thus, when community functional homogeneities are evaluated against the average similarity of interacting proteins, communities are not negatively impacted by containing fewer edges (having a lower density), which especially negatively affects large communities.

Using this method of computing the functional homogeneity, the quality of network partitions at different resolutions was calculated to assess whether there is an optimal resolution for partitioning a PIN. To evaluate the role of the resolution limit in this assessment, we investigated the network partition quality in community size bands. Here, the quality of the network partition at resolution r for the range

of community sizes B, was quantified by the average functional homogeneity of communities in the given size range, denoted by the network partition functional homogeneity score $\text{fh}_{B,r}$. Similar to Equation (4.2), this was calculated by:

$$\text{fh}_{B,r} = \frac{1}{|C_{B,r}|} \sum_{c \in C_{B,r}} \frac{1}{|E_c|} \sum_{u,v \in c; (u,v) \in E; v > u} \rho_{SS}(u, v). \tag{5.1}$$

Here, $C_{B,r}$ denotes the set of communities with a size in range B at resolution r, u and v are nodes in a community c in this set, and $\rho_{SS}(u, v)$ is the functional similarity score between the two nodes calculated as explained in Section 3.2.1. E denotes the set of edges in the PIN, and $|E_C|$ is the number of edges between nodes in the community. In the above equation each community is weighted equally regardless of community size to assess detected communities rather than network proportions.

In order to make network partition functional homogeneity scores informative, these scores were put into the context of background functional similarity scores. We used the mean of the functional similarity scores of interacting proteins (Int BG) and the mean of the functional similarity scores of all protein-pairs (Rand BG) to generate this context. Background functional similarity values for HINT-P-14 and BioGrid-AP-14 are shown in Table 5.1.

Table 5.1: Background functional similarity values.

Network	Int BG	Rand BG
HINT-P-14	5.8901 ± 3.8499	3.3237 ± 2.8212
BioGrid-AP-14	6.0565 ± 3.8032	3.0003 ± 2.7028

Table 5.1: The mean functional similarity score of interacting proteins is given in the Int BG column, and the mean functional similarity of all annotated protein-pairs in the PINs is denoted by Rand BG. Standard errors are quoted after the values.

5.3 Criteria for the evaluation of community detection methods

Each community detection method groups proteins in PINs according to its own notion of the concept of a community (see Section 2.3). For each method, at least some of these communities will be functionally homogeneous. A community

detection method that is able to capture the functional organization of the entire PIN should contain a large proportion of potential functional modules in the communities it generates. A signal for biologically meaningful community detection can thus be conveyed via the average functional homogeneity of communities in a network partition (see network partition functional homogeneity by Equation (5.1)).

Finding such a network partition that captures the functional organization of an entire human PIN would provide a robust basis for the pipeline developed in this project (see Section 1.1.3). As network partitions at neighbouring resolutions can be very similar, it is possible to generate a consensus network partition from network partitions in a narrow resolution range [143]. Finding such an ideal network partition is the central focus according to which community detection methods were evaluated. Specifically, two criteria were used for our evaluation.

Criterion 1: Optimal Resolution

The existence of a network partition (or a set of neighbouring partitions) that contains communities characterized by specific cellular functions.

Criterion 2: Proteins of Interest

As it is unlikely that the entire network is perfectly partitioned even at the optimal resolution identified according to the first criterion, it is important that proteins that are of interest to the biological application are assigned to potential functional modules. As mentioned in Section 3.1.1 the interleukins IL-4, IL-10, IL-12A, IL-12B, and IL-13 have previousy been linked with macrophage differentiation, the target biological application in this project. The second criterion for the evaluation of community detection methods concerns whether these proteins of interest are partitioned into communities of size 6 - 35, our designated size range (see Section 5.1).

In practice, Criterion 1 is fulfilled when the network partition functional homogeneity values peak in the same narrow resolution range for communities of all size ranges. Criterion 2 assesses whether a community detection method is suitable for the investigation of macrophage differentiation. Expanding on Criterion

2, we also investigated the coverage of proteins in communities of size 6 - 35 across PINs to assess the suitability of the generated network partitions beyond the biological application of macrophage differentiation. This coverage assessment represents a more general characteristic of suitable community detection methods.

5.4 Louvain community detection

In this section Modularity Maximization community detection as implemented in a modified Louvain algorithm (see Section 2.3.2.3 for Louvain algorithm, and Appendix C for the Louvain modification) is evaluated using the configuration and the Constant Potts null models. As this is an exploratory analysis each network partition was only performed once, although Lovain partitions may differ between runs. Unless otherwise stated the results presented here are consistent for HINT-P-14 and BioGrid-AP-14.

5.4.1 Configuration model

Community detection methods partition a network into groups of interacting nodes (see Section 2.3). Multi-resolution community detection methods incorporate a resolution parameter into this process which controls the scale of the network partition. At the lowest resolution the entire network is in a single community, while at other end of the resolution scale each node is in its own community. Meaningful network partitions exist between these extremes. To find the range of resolutions at which HINT-P-14 is partitioned into non-trivial communities by Louvain configuration model community detection, we looked at the number of communities of size ≥ 4 across resolutions (Figure 5.1), the total number of communities at each resolution (Figure 5.2), and investigated the community size distributions for each network partition (Figure 5.3).

Figure 5.1 shows that the number of communities of size ≥ 4 decreases from $\log \gamma \approx 2.3$. As the number of communities generally increases with resolution, a considerable proportion of communities of size ≥ 4 must be sub-divided into smaller components with fewer than four nodes from this resolution. A community size of

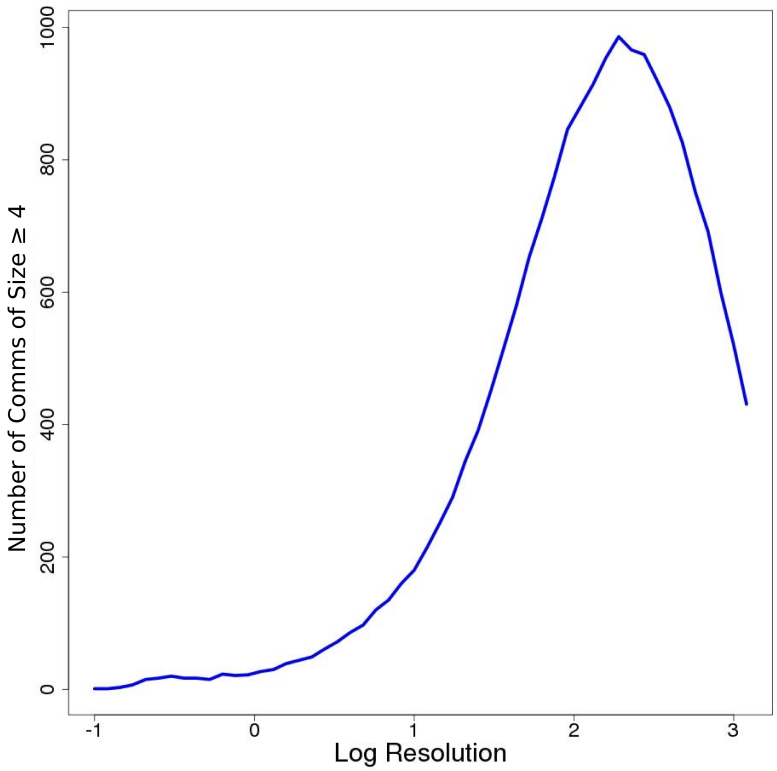

Figure 5.1: Number of communities of size ≥ 4 in HINT-P-14. Number of Communities of Size ≥ 4 versus the log of the resolution γ for a Louvain partition of the HINT-P-14 network using the configuration model. The plot shows a drop in the number of communities at $\log \gamma \approx 2.3$, indicating an over-partitioning of the network.

three or less represents a biologically trivial association of nodes. Thus, the network can be regarded as over-partitioned from this resolution.

While large communities are expected to be sub-partitioned into smaller ones as the resolution increases (see zooming in behaviour in Section 2.3.1), Figure 5.2 suggests that configuration model Louvain partitions do not always conform to this expectation. Figure 5.2 shows that the number of communities in HINT-P-14 and BioGrid-AP-14 does not monotonically increase with resolution, but instead exhibits local maxima at low resolutions. This observation suggests a re-partitioning effect is occurring where nodes are initially partitioned into small communities and subsequently re-partitioned into larger ones at a higher resolution. This effect is particularly pronounced for BioGrid-AP-14. Repeated configuration model Louvain runs were performed to confirm that the observed effect is not caused by the variability in the network partitions generated by the Louvain

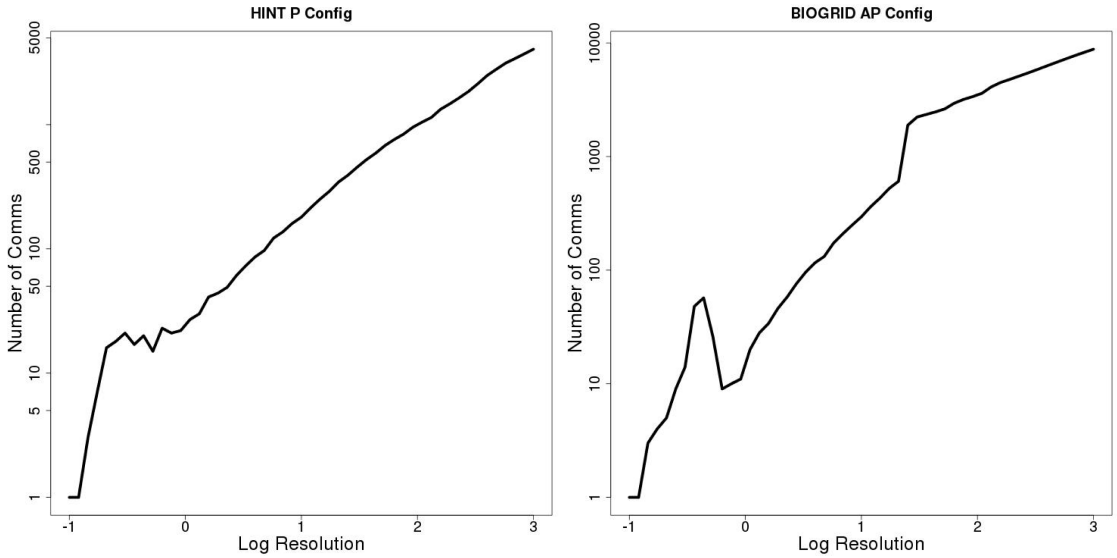

Figure 5.2: Re-partitioning effect for HINT-P-14 and BioGrid-AP-14 by configuration model Louvain. The number of communities of all sizes versus log resolution for Louvain runs using the configuration model (Config) on HINT-P-14 and BioGrid-AP-14. The number of communities does not increase monotonically, but exhibits a local maximum at lower resolutions. The small peak in community number quickly disappears suggesting the nodes in these communities were re-partitioned into larger ones.

algorithm (see Section 2.3.2.3). For a resolution range $\log \gamma > 0$, the resolution parameter behaves as expected.

The resolution range for well-behaved network partitioning identified from Figure 5.2 coincides with the lower bound for network partitioning set at $\log \gamma \approx 0$ based on the community size distribution plots in Figure 5.3. This figure shows that a community size similar to the size of the network is present in low resolution network partitions. The median community size decreases monotonically from a log resolution of approximately zero, which suggests the expected "zooming-in" behaviour.

According to the data shown in Figures 5.1–5.3 HINT-P-14 is partitioned into non-trivial communities by configuration model Louvain in the resolution range $0 < \log \gamma < 2.3$. An optimal resolution, or resolution range, that captures the functional organization of HINT-P-14 by criterion 1 (see Section 5.3) must lie within the above bounds. To identify such an optimal resolution range, we investigated the network partition functional homogeneity (see Equation (5.1)) for community size bands of 2 - 10, 11 - 30, 31 - 60, 61 - 100, and 101+ (Figure 5.4). These size

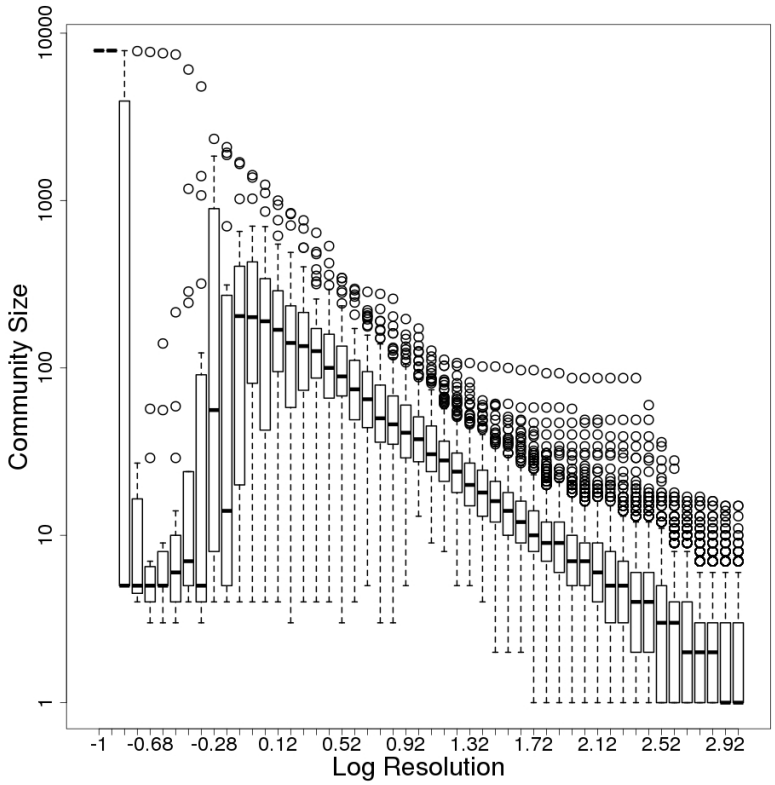

Figure 5.3: Community size distributions across resolutions. Boxplots of the distribution of community sizes at each resolution for Louvain partitioning using the configuration model on HINT-P-14. Lines in the centre of the boxplots are drawn at the median community size, while the boundaries of the box represent the 25% and 75% quantile range of the community sizes. The dotted lines represent the range of the distribution, with all values further than 1.5× the interquartile range from the median shown as outliers (dots). The graph shows that the network is not significantly partitioned until $\log \gamma \approx 0$ as there exists a community whose size is approximately the size of the network. For $\log \gamma > 0$ the median community size starts to monotonically decrease as expected.

bands were chosen to cover the full range of communities for which a functional homogeneity can be calculated based on the size distributions shown in Figure 5.3.

While Figure 5.4 shows functional homogeneity maxima at low resolution across community size bands, only the size band 2 - 10 exhibits a local maximum within the meaningful resolution range of $0 < \log \gamma < 2.3$. Furthermore, all network partition functional homogeneity maxima fall into resolution ranges where there are few communities in the respective size band. Few communities contributing to the observed functional homogeneity peaks suggests that configuration model Louvain does not capture the functional organization of the entire network at this resolution.

The disappearance of the high functional homogeneity signal in the size band 2

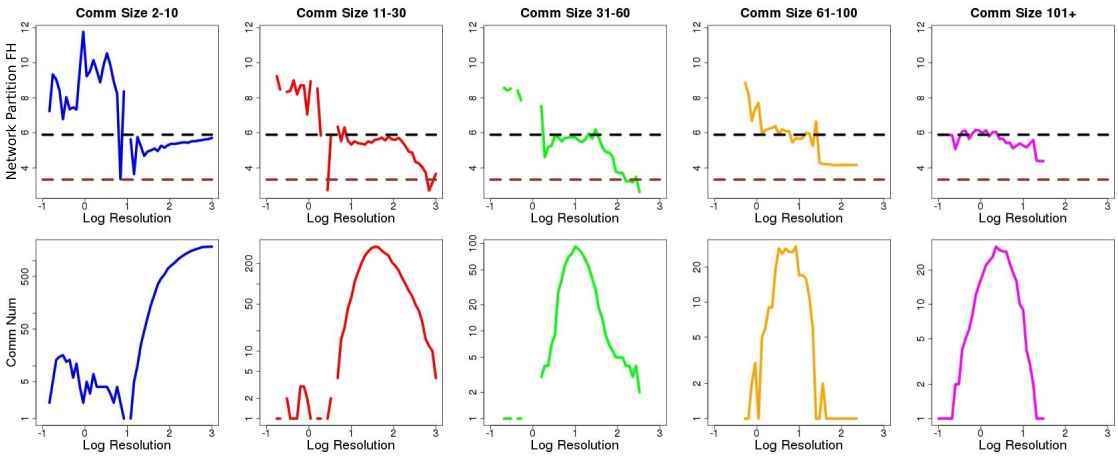

Figure 5.4: Configuration model Louvain network partition functional homogeneity in size bands versus the number of communities. Network partition functional homogeneity was calculated in community size bands by Equation (5.1) and plotted versus log resolution (upper plots). The lower plots show the number of communities in these community size bands. Partitions were obtained by multi-resolution Louvain runs using the configuration model on HINT-P-14. The functional homogeneity plots are put into perspective by comparison with the average functional similarity of interacting proteins in the PIN (black dotted line), and the average functional similarity of all node pairs (brown dotted line). The plot shows the initial high functional homogeneity scores result from only few communities.

- 10 at a resolution of $\log \gamma \approx 1$ suggests that the re-partitioning effect (Figure 5.2) is occurring up to this resolution. It appears that small, highly functionally homogeneous communities are "ripped off" the central community containing most of the PIN at low resolutions. These communities have been re-partitioned into larger communities at $\log \gamma \approx 1$, where there are no more communities in this size band and no communities of size 1 (see Figure 5.3).

At resolutions where large parts of the network are partitioned into certain size bands (the number of communities in the size bands peak), the network partition functional homogeneity is similar to the average functional similarity of interacting protein representing the background. As shown in Equation (5.1), the network partition functional homogeneity is computed by averaging community functional homogeneity scores. Thus, the question arises whether this averaging hides a large proportion of the network actually being partitioned into functionally homogeneous communities. We investigated this question by looking at the number of communities with a functional homogeneity higher than the average functional similarity of interacting proteins (functionally homogeneous communities), which is

similar to the previous analysis performed on yeast PINs [48] (Figure 5.5).

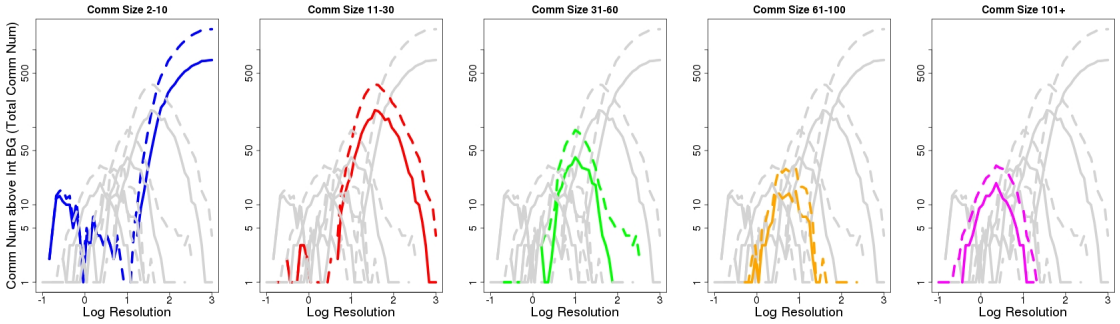

Figure 5.5: The number of functionally homogeneous configuration model Louvain communities across resolutions. Partitions were obtained by multi-resolution Louvain runs using the configuration model on HINT-P-14. Communities were evaluated as functionally homogeneous if their functional homogeneity was higher than the average functional similarity of interacting proteins in the PIN (Int BG, see Section 5.2). Dashed lines show the total number of communities in the given size range. Each small multiple figure highlights the plots for a specific range of community sizes with all other plots in grey. While the the number of functionally homogeneos communities peak where there are many communities in the respective size bands, they peak at different resolutions.

Figure 5.5 shows that the number of functionally homogeneous communities peaks at the same resolution where the number of communities in the respective size band peaks. Thus, a large proportion of the network is partitioned into potentially biologically relevant communities at these resolutions. However, Figure 5.5 also shows that the distance between the plots showing the number of communities in a respective size band (dashed line) and the number of functionally homogeneous communities in this size band (full line) increases at the resolutions where these maxima are found. Thus, the community detection algorithm is actually performing worse at these resolutions as a lower proportion of the partitioned communities are evaluated as functionally homogeneous. Optimal community detection performance (where the dashed and full lines are closest) across size bands generally corresponds to the network partition functional homogeneity maxima shown in Figure 5.4.

The results presented here suggest that the conclusion drawn from the multi-resolution yeast PIN analysis that biological functions are optimally partitioned at different resolutions [48] may be confounded by the resolution limit affecting configuration model Louvain community detection. As the number of functionally homogeneous communities peaks at different resolutions for each size band, the

community size affects the resolution at which the PIN is deemed as optimally partitioned. Given this observation, if it takes a comparatively large number of proteins to perform a particular cellular function, a functional module for this function would likely be found at lower resolution than a functional modules for a function that is performed by fewer proteins. This observation is an effect of the resolution limit affecting configuration model Louvain Modularity Maximization and is thus specific to this community detection method. It may be the case that other methods which are not affected by the resolution limit are able to optimally partition PINs at a single resolution.

5.4.2 Constant Potts model

The Constant Potts model (CPM) is a null model for Modularity Maximization that is not affected by the resolution limit [140] (see Section 2.3.2.2). Hence, CPM Louvain is a good candidate for a community detection method that can capture the functional organization of a PIN at a single resolution. To investigate how CPM Louvain partitions PINs we looked at the number of communities into which HINT-P-14 and BioGrid-AP-14 are partitioned at each resolution (Figure 5.6) and assessed the distributions of community sizes (Figure 5.7).

Figure 5.6 shows that the re-partitioning effect described in Section 5.4.1 is more pronounced for CPM Louvain than for configuration model Louvain (see Figure 5.2). This effect is especially visible for the BioGrid-AP-14 data set, where the number of communities decreases in the resolution range $-2.8 < \log\gamma < -1.6$. The expected "zooming-in" effect with increasing resolution only occurs for resolutions $\log\gamma > -2.7$ for HINT-P-14 and $\log\gamma > -1.6$ for BioGrid-AP-14.

Smaller communities tend to have higher densities. As the CPM directly selects for density via the resolution parameter (see Section 2.3.2.2), it also selects for small community size. This feature is particularly pronounced for BioGrid-AP-14 partitions (Figure 5.7). Figure 5.7 shows that 75% of the communities in BioGrid-AP-14 are of size 1 for $\log\gamma < -1.2$. Thus, a large proportion of the network is partitioned into trivial communities. While HINT-P-14 does not exhibit as strong

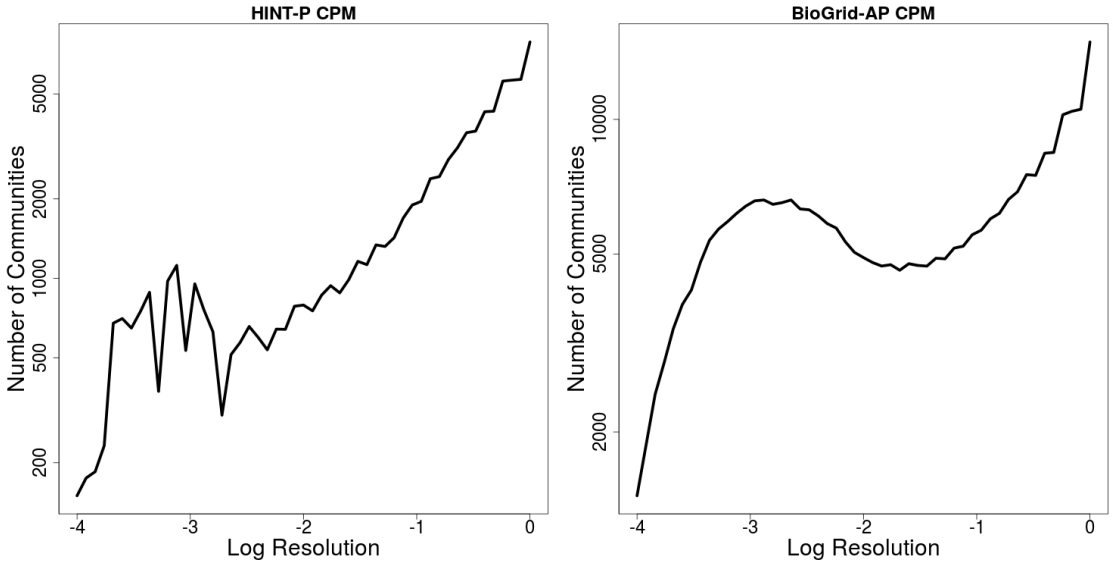

Figure 5.6: Re-partitioning effect for HINT-P-14 and BioGrid-AP-14 by CPM Louvain. The number of communities of all sizes versus log resolution for Louvain runs using the CPM on HINT-P-14 and BioGrid-AP-14. The number of communities does not increase monotonically, but exhibits a local maximum at lower resolutions. This effect is particularly pronounced for BioGrid-AP-14 where the number of communities is decreasing in the resolution range $-2.8 < \log\gamma < -1.6$.

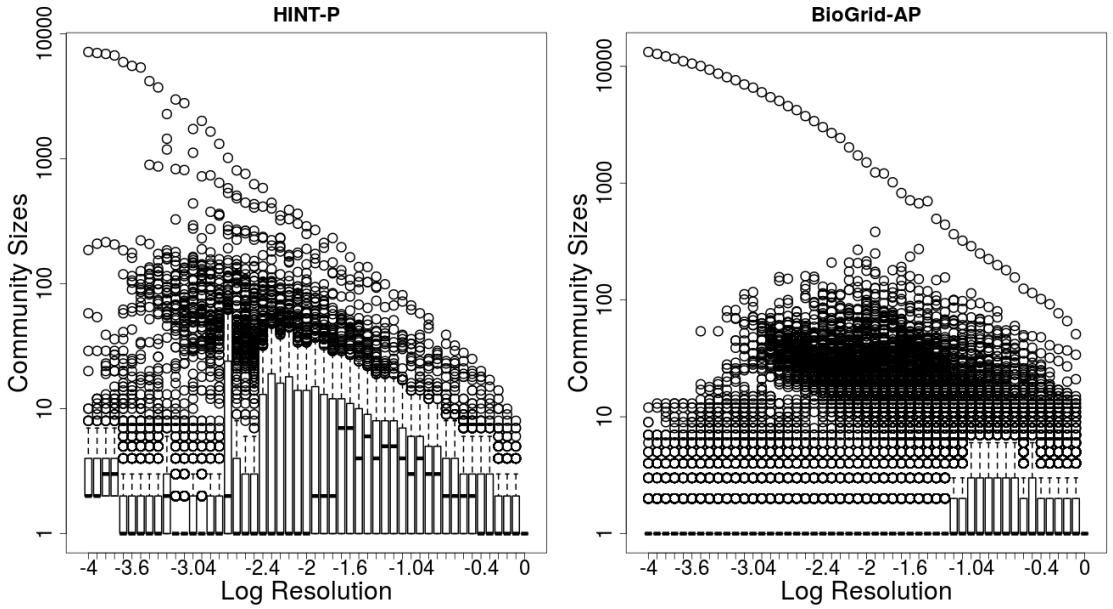

Figure 5.7: Louvain CPM community size distributions in HINT-P-14 and BioGrid-AP-14. Boxplots of the distribution of community sizes at different resolutions for CPM Louvain runs on HINT-P-14 and BioGrid-AP-14. The box represents the interquartile range of the distribution with the line denoting the median. The dotted lines represent the range of the distribution, with all values further than 1.5× the interquartile from the median shown as outliers (dots). The plots show that CPM Louvain selects for small communities. Especially in BioGrid-AP-14, a large proportion of the PIN is partitioned into single node communities (75% of communities are size 1 for $\log\gamma < -1.2$, given that the boxplots are flattened to a line).

a selection for single node communities, in no network partition are more than 50% of communities of size > 7. As functional modules are likely to contain at least six proteins, the community size ranges obtained by CPM Louvain are not ideal for functional module detection.

The large number of single node communities should not affect the ability of CPM Louvain community detection to capture the functional organization of a PIN. As PINs are incomplete, single node communities may arise due to incomplete coverage of functional modules in currently available PINs. We investigated how well CPM Louvain captures the functional organization of HINT-P-14 by network partition functional homogeneity evaluation after Equation (4.2) (Figure 5.8)

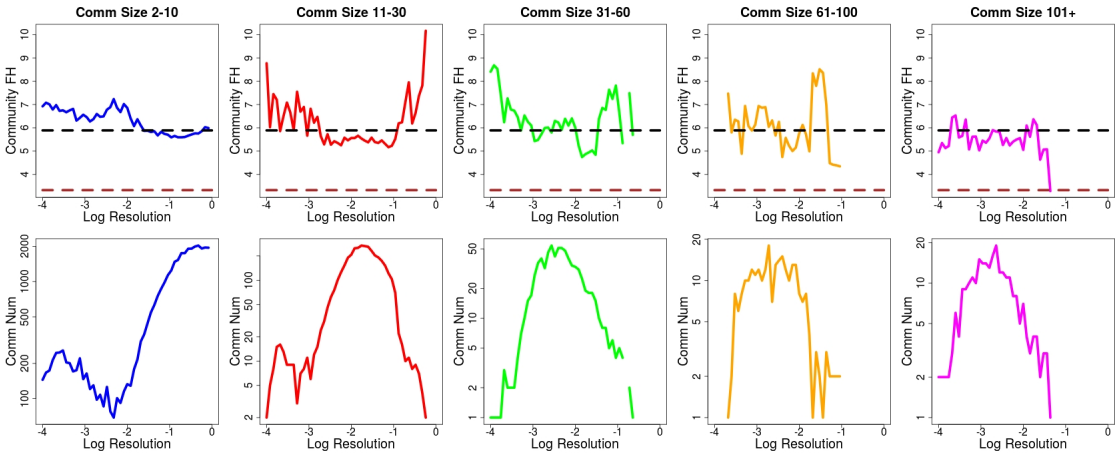

Figure 5.8: CPM Louvain network partition functional homogeneity in size bands versus the number of communities. Network partition functional homogeneity (Community FH) was calculated in community size bands by Equation (5.1) and plotted versus log resolution (upper plots). The lower plots show the number of communities in these community size bands. Partitions were obtained by multi-resolution Louvain runs using the CPM on HINT-P-14. The functional homogeneity plots are put into perspective by comparison with the average functional similarity of interacting proteins in the PIN (black dotted line), and the average functional similarity of all node pairs (brown dotted line). The graph shows that functional homogeneity maxima are found where there are few communities in the respective size bands.

The CPM network partition functional homogeneity plots in Figure 5.8 are similar to those obtained with the configuration model (see Figure 5.4). Functional homogeneity maxima tend to correspond to resolutions with few communities in the respective size band with the possible exception of size band 61 - 100, which exhibits a local maximum. Furthermore, local functional homogeneity maxima are

at different resolutions across size bands, thus our first criterion is not fulfilled by CPM Louvain community detection (see Section 5.3).

While further investigation into the number of functionally homogeneous communities similar to Figure 5.5 showed that CPM Louvain does find functionally homogeneous communities, the maximum number of functionally homogeneous communities is correlated with the number of communities in the respective size bands (data not shown). As the quality of community detection can be evaluated based on the proportion of communities that are functionally homogeneous, such maxima do not suggest good partitioning.

The selection of communities by a resolution parameter controlling for community density may not be affected by the resolution limit, however it does select for dense, small communities. From a biological perspective, we have no reason to assume that functional modules have similar densities. Finding such similar densities is especially unlikely for biologically meaningful communities in currently available PINs, as current PIN data sets are noisy and incomplete [88,89,93]. Hence, selecting for density may not be ideal to optimally partition a PIN at a single resolution.

5.5 Overlapping community detection

While the Louvain algorithm partitions proteins into non-overlapping communities, it seems counterintuitive that functional modules should not overlap. Proteins are often associated with multiple GO annotations. Indeed, many proteins have been shown to belong to multiple complexes [71]. Thus, overlapping community detection may allow for a more accurate representation of biological functions.

BigCLAM and link clustering are two community detection methods that scale well with network size and can thus generate overlapping network partitions, so-called network covers, for multiple resolutions in a reasonable time frame. While link clustering incorporates a resolution parameter, the number of communities fitted to the data K, is used as a resolution proxy for BigCLAM.

5.5.1 BigCLAM

In Section 4.2 we showed that proteins are more likely to interact the more functional annotations they share. This concept is built into the Affiliation Graph Model (AGM) [125], a model for community detection based on the stochastic block model (see Section 2.3.3.1). BigCLAM uses a fast approximation of the AGM to detect overlapping communities in networks.

We discussed the importance of proteins related to the intended biological application being partitioned into communities of size 6 - 35 that represent potential modules in Section 5.3. As BigCLAM uses the number of communities fitted to the data K as a resolution proxy, it is not certain that the entire network is fitted at every value of K. Thus, we investigated what proportion of the network was partitioned into communities of size > 2, > 5, > 9, and > 35 dependent on the number of communities fitted. Specific focus was put on whether the interleukins previously linked with macrophage differentiation were partitioned into communities in the above size bands (Figure 5.9).

Figure 5.9 shows that $\approx 40\%$ of proteins in BioGrid-AP-14 are never partitioned into communities of size > 5. Furthermore, although BioGrid-AP-14 contains all of the investigated interleukins (see Section 3.1.1), only IL-10 is partitioned into a community by BigCLAM. While the network partition functional homogeneity for BioGrid-AP-14 covers do peak in a narrow resolution range of $501 < K < 801$ (see Figure 5.13a), IL-10 is in a community of size 6 - 9 for a proxy resolution range $K > 1901$. Thus, BigCLAM may fulfil our first evaluation criterion for community detection methods, however it does not fulfill the second (see Section 5.3). Indeed, less than 50% of the PIN is affiliated with communities in the range $501 < K < 801$. In the case of HINT-P-14 covers, none of the interleukins are in communities of size > 5 at any resolution (data not shown).

Figures 5.9 and 5.13 further show that the number of communities fitted to the PIN is unsuitable as a resolution parameter. Although large communities appear to be fitted first, the first community fitted is of size 14. A single community of size 14 is not a representative low resolution community compared to the definition

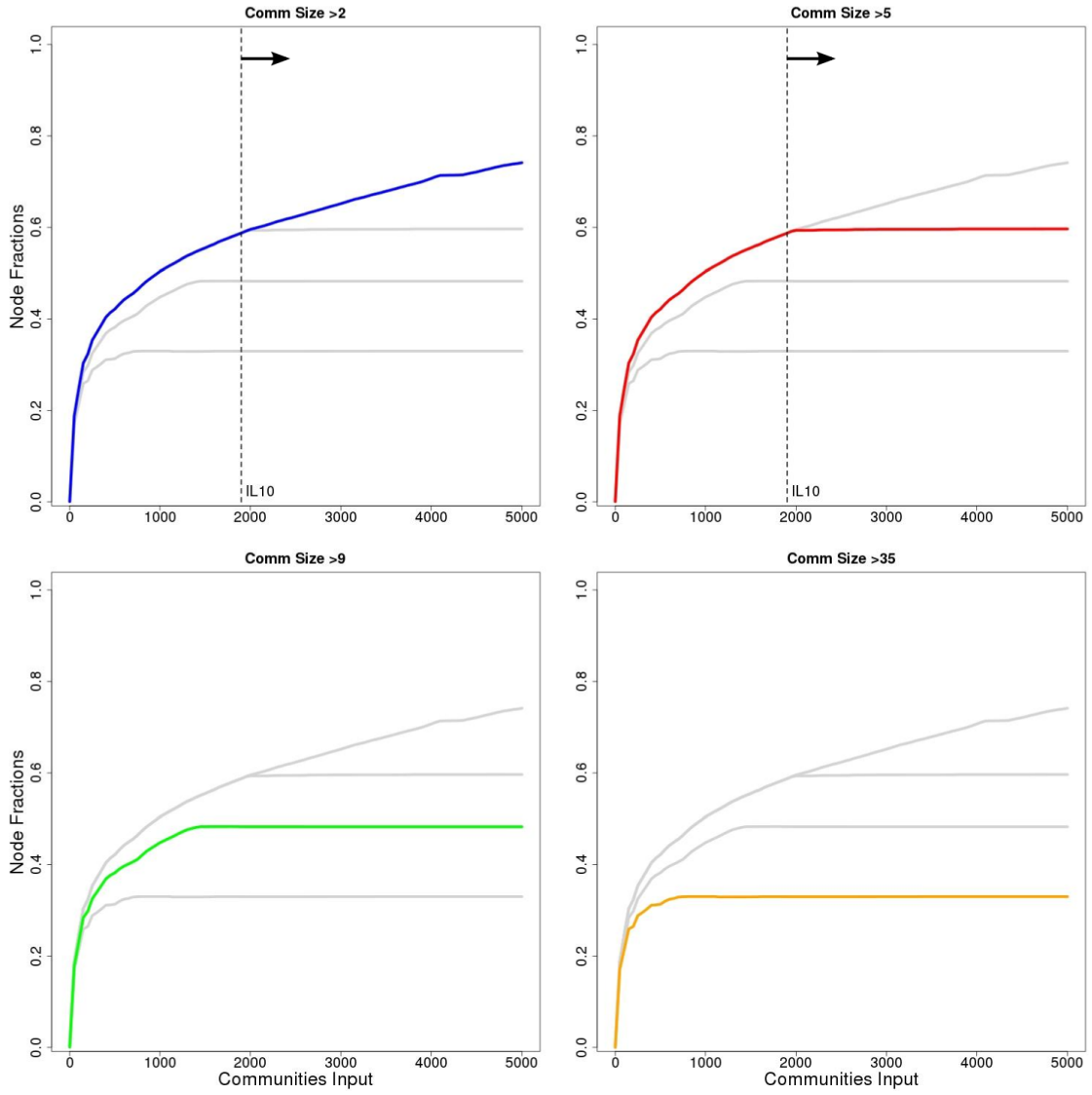

Figure 5.9: Proportion of nodes affiliated with BigCLAM communities of size > 2, > 5, > 9, and > 35. Plots showing the fraction of nodes in communities of size > 2, > 5, > 9, and > 35 versus the number of communities fitted by BigCLAM to BioGrid-AP-14. As nodes can be assigned to multiple communities, the community size of the largest community to which a node is assigned is used. Vertical dashed lines show when interleukins linked to macrophage differentiation are in the network partition defined by community affiliations above a certain size. The graphs show that only IL-10 is affiliated with a community at any resolution, and this community is smaller than size 10. The different length plateaus for the > 5, > 9, and > 35 graphs indicate that large communities are fitted first, and smallest tend to be fitted last.

of the resolution parameter in Louvain implementations. In addition, once fitted communities are not sub-partitioned at higher K values. Thus, BigCLAM does not in fact perform multi-scale partitioning of networks at different values of K. Rather, it may detect communities at multiple scales as overlapping communities at a single K value. In this set-up communities that would otherwise be detected at different resolutions are compressed into a single set of network covers. This precludes finding an optimal resolution for partitioning.

5.5.2 Link clustering

Proteins perform biological functions through their interactions. Thus, it is more likely the interaction that is characterized by a specific function rather than the protein, which can be characterized by multiple functions. As such, clustering interactions rather than proteins may be more suitable to capture individual functions in functionally homogeneous communities. Link clustering is a community detection method that implements this concept. We investigated whether link clustering can capture the functional organization of BioGrid-AP-14 via the network partition functional homogeneity in community size bands of 3 - 10, 11 - 30, 31 - 60, 61 - 100, and 101+ (Figure 5.10). The minimum investigated community size was set to three as a trivial link clustering community represents a single edge, which gives a community of size two.

Figure 5.10 suggests that BioGrid-AP-14 is optimally partitioned in a resolution range of $0.11 < S < 0.18$ across community size bands. In the community size bands of interest (community sizes 6 - 35) this range narrows to $0.15 < S < 0.18$, fulfilling our first criterion for the evaluation of community detection methods (see Section 5.3). Crucially, the functional homogeneity maxima are observed in a resolution range where the number of communities in the respective size bands also peak, suggesting a large proportion of the PIN is partitioned into functionally homogeneous communities.

To evaluate link clustering against our second evaluation criterion determining whether it is a suitable method for investigating macrophage differentiation (see

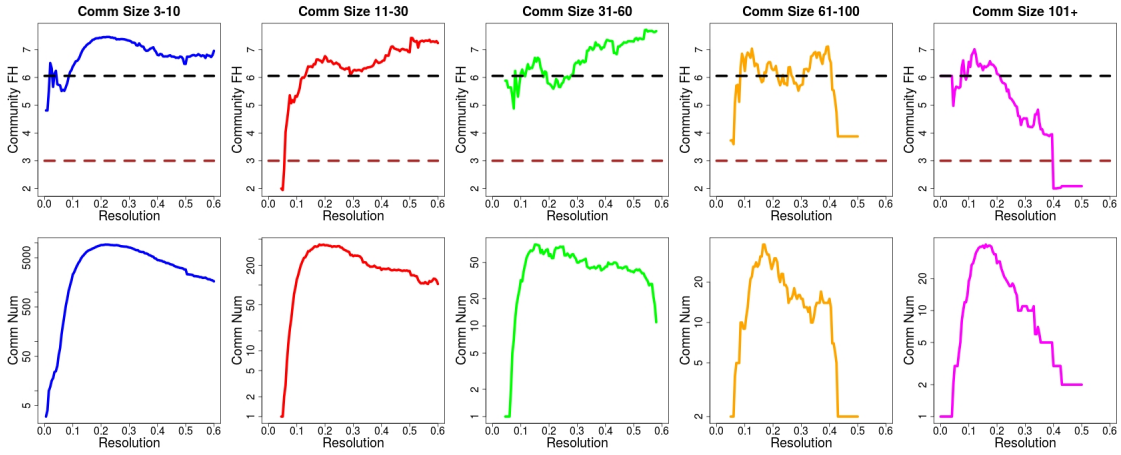

Figure 5.10: Link clustering network partition functional homogeneity in size bands versus the number of communities. Network partition functional homogeneity (Community FH) was calculated in community size bands by Equation (5.1) and plotted versus the resolution parameter (upper plots). The lower plots show the number of communities in these community size bands. Partitions were obtained by multi-resolution link clustering on BioGrid-AP-14. The functional homogeneity plots are put into perspective by comparison with the average functional similarity of interacting proteins in the PIN (black dotted line), and the average functional similarity of all node pairs (brown dotted line). Network partition functional homogeneity maxima across size bands can be seen in the resolution range $0.11 < S < 0.18$. Functional homogeneity maxima tend to be at the same resolution as maxima for the number of communities in the respective size bands.

Section 5.3), we looked at the fraction of proteins partitioned into non-trivial communities (size > 2), and communities that fall into our size range of interest for functional modules (size < 5, but not < 35) in Figure 5.11. As in Figure 5.9 we specifically noted when the interleukins linked with macrophage differentiation in the literature (see Section 3.1.1) are in the network partitions defined by these node community affiliations.

The difference between the red and yellow lines in Figure 5.11 show that $\approx 10\%$ of proteins are partitioned into communities of size 6 - 35 at any single resolution. Importantly, IL-4 and IL-10 are in this set for a resolution range of $0.16 < S < 0.195$. As this resolution range falls into our optimal resolution range identified above, link clustering satisfies both evaluation criteria for IL-4 and IL-10.

In contrast to these two interleukins, IL-12A, IL-12B, and IL-13, are not affiliated with communities in the size range of interest at any resolution. As IL-13 and IL-12A are degree one nodes in BioGrid-AP-14, it is not surprising that these proteins are already partitioned into trivial communities at low resolutions. Indeed

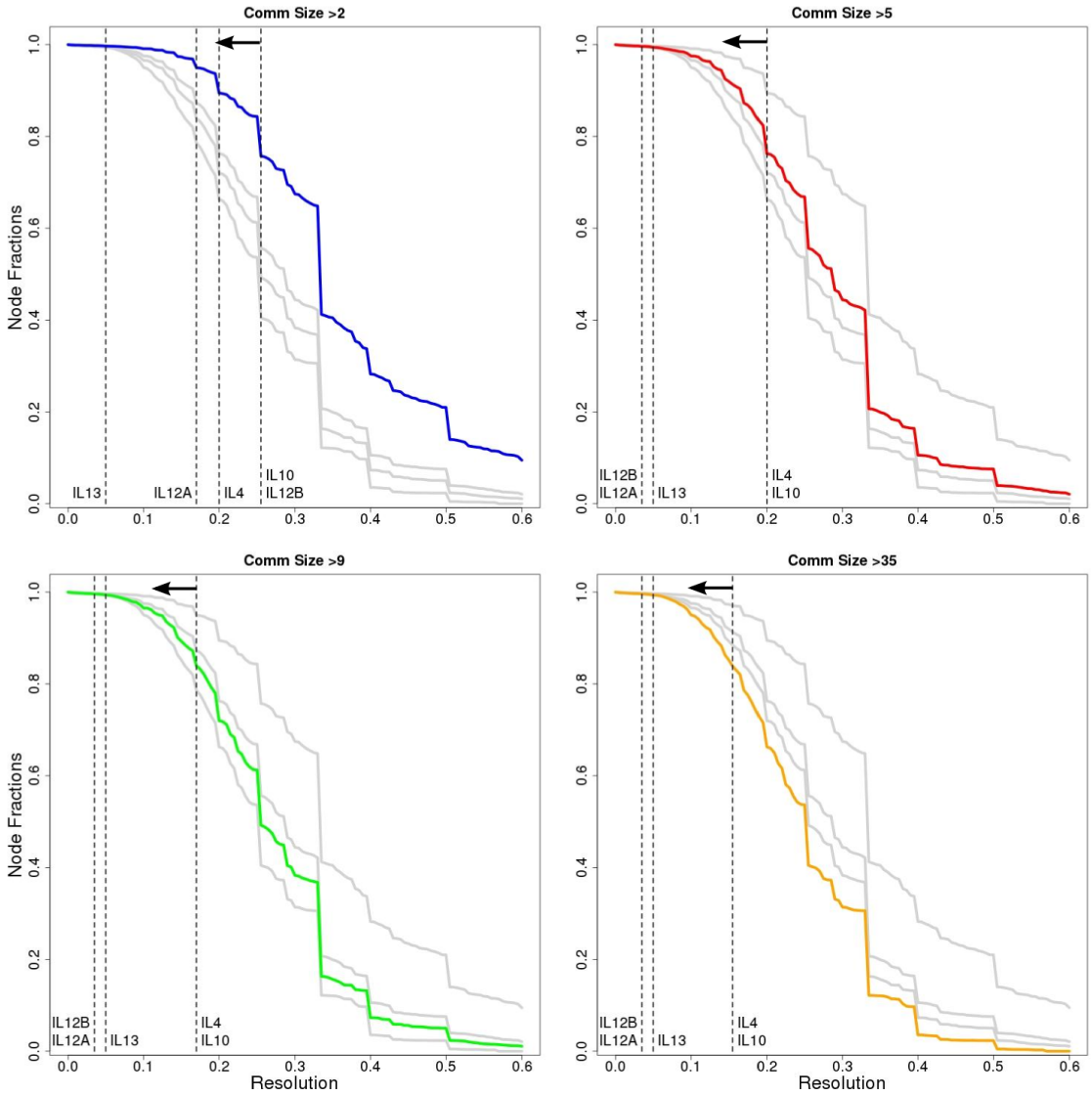

Figure 5.11: Proportion of nodes affiliated with link clustering communities of size > 2, > 5, > 9, and > 35. Plots showing the fraction of nodes in communities of size > 2, > 5, > 9, and > 35 versus resolution for BioGrid-AP-14 link clustering partitions. As nodes can be assigned to multiple communities, the maximum community size a node is assigned to is used. Vertical dashed lines show when interleukins linked to macrophage differentiation are no longer in the network partition defined by community affiliations above a certain size. The graphs show that IL-4 and IL-10 are in partitions in a relevant size range (6 - 35, Section 5.3) up to a resolution of 0.2. It appears the interleukins are in less dense regions of the PIN, as they are all in the first 25% of nodes partitioned into trivial communities of size 2.

all proteins of interest are among the first 25% of nodes to be assigned to only trivial communities of size 2 (see top left graph in Figure 5.11), which suggests they are in sparse regions of the PIN. As link clustering has been shown to perform better on dense networks [127] this circumstance may complicate the detection of functional modules relevant to macrophage differentiation.

A further complication was identified by closer inspection of the community to which both IL-4 and IL-10 are assigned in the optimal resolution range $0.16 < S < 0.195$. This community is a star-like community centred on the A2M gene with 7 or 11 nodes at different resolutions. Most of the proteins in the periphery do not appear functionally related. While the community is functionally homogeneous (functional homogeneity of 7.62 or 7.98, which is above the interaction background – see Section 5.2), a star-like community structure does not conform to what is expected for a functional module due to the lack of interactions between most proteins. However, the structure of a single community should not suggest that link clustering is unsuitable for our biological application.

5.5.2.1 Signal of core-periphery structure

Link clustering was found to satisfy the first criterion for the evaluation of community detection methods based on the existence of local functional homogeneity maxima in a narrow resolution band of $0.11 < S < 0.18$ (see Figure 5.10). However, the global functional homogeneity maxima for community size bands 11 - 30 and 31 - 60 are found at a higher resolution of $S \approx 0.55$. At this resolution it appears that communities with more than 60 proteins have been sub-partitioned into highly functionally homogeneous communities in the size range 11 - 60. This second resolution may be an expression of the core-periphery structure in the PIN (see Section 2.2.2).

Assuming a dense PIN core and a less dense periphery, communities in the periphery could be optimally partitioned at the same resolution when the dense core is clustered into a single community. At higher resolutions, the core would then be partitioned into functionally homogeneous components when the peripheral

communities are overpartitioned into small trivial communities. This model may explain the existence of the two functional homogeneity maxima for size bands 11 - 30 and 31 - 60. Furthermore, as core network proteins are likely to be well-studied given the number of reported interactions for a protein to be part of a densely connected core, they are also likely to be well functionally annotated. Such well-annotated proteins tend to have higher functional similarities to other proteins (see annotation bias study in Section 6.2), which may explain the higher functional homogeneity of the high resolution maxima.

5.6 PIN comparison

BioGrid-AP-14 and HINT-P-14 were chosen as characteristically different PINs representing two ends of the spectrum between prioritizing the quality of interactions (lower false-positive rate) and prioritizing maximum coverage (lower false-negative rate; see Section 3.1.1). This choice was made to test how the trade-off between these error rates affects community detection. For example it may be the case that structural differences between A-type and P-type data result in communities of data types, rather than functional relations. Using the previously presented results of the analysis of the four community detection methods, we investigated whether the different compromises made by the PIN data sets could be seen to affect the performance of the community detection methods.

Louvain community detection on HINT-P-14 and BioGrid-AP-14 showed similar results. The main differences between community detection on the network were twofold:

1. BioGrid-AP-14 exhibits a larger proportion of single node communities than HINT-P-14 by CPM Louvain

2. BioGrid-AP-14 shows a more pronounced peak at low resolutions for the number of communities (Figures 5.2 and 5.6)

While the large number of single node communities may arise due to the difficulty of partitioning PINs with false-positive interaction data, there are more nodes affiliated with non-trivial BioGrid-AP-14 communities than with non-trivial HINT-P-14 communities at every resolution due to the difference in PIN sizes (see Section 3.1.1). The more pronounced re-partitioning effect in BioGrid-AP-14 may be a feature of the false-positive rate. However, it is unclear whether this feature affects functional module detection. Indeed, HINT-P-14 exhibited no improved partitioning into functionally homogeneous communities despite alleged higher quality interaction data (see HINT-P-14 in Figure 5.4; BioGrid-AP-14 data not shown). As such, the benefits of higher coverage appear to outweigh those of high qualtiy interactions for non-overlapping community detection.

For overlapping community detection the advantage of the more comprehensive BioGrid-AP-14 was particularly visible for link clustering. While link clustering on BioGrid-AP-14 performed well against our evaluation criteria (see Section 5.3), the same analysis on HINT-P-14 did not (Figure 5.12). Figure 5.12 shows that network partition functional homogeneity values do not exceed the background functional similarity level across most size band and resolutions. Furthermore, similar to the Louvain results in Figures 5.4 and 5.8, resolutions at which the functional homogeneity peaks tend to exhibit few communities in the respective size bands.

Link clustering has been shown to perform better on denser networks [127]. Thus, it may be that the higher quality of network covers generated on BioGrid-AP-14 is due to the higher density in comparison to HINT-P-14. While the greater number of false-positive edges will contribute to this increased density, a smaller proportion of edges that are falsely reported as not present (false-negative rate) will also contribute. Thus, our results suggest that the false-negative rate of a PIN (higher coverage) affects overlapping community detection by link clustering more than the false-positive rate. This conclusion must however be confirmed on further PIN data sets.

Overall, our evaluation of four community detection methods for the generation of functional modules shows a preference for BioGrid-AP-14 over HINT-P-14. There

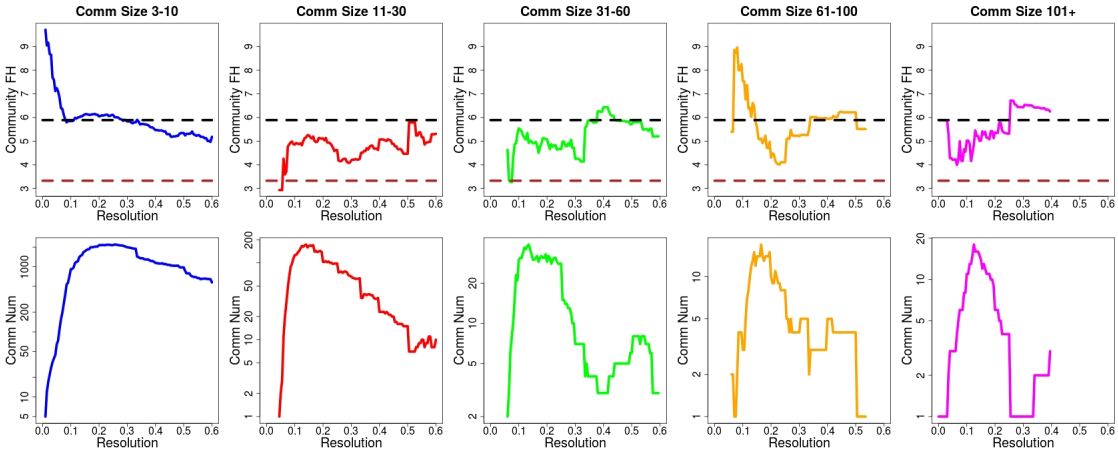

Figure 5.12: HINT-P-14 Link clustering network partition functional homogeneity in size bands versus the number of communities. Network partition functional homogeneity (Community FH) was calculated in community size bands by Equation (5.1) and plotted versus the resolution parameter (upper plots). The lower plots show the number of communities in these community size bands. Partitions were obtained by multi-resolution link clustering on HINT-P-14. The functional homogeneity plots are put into perspective by comparison with the average functional similarity of interacting proteins in the PIN (black dotted line), and the average functional similarity of all node pairs (brown dotted line). No functionally homogeneous resolution range across community size bands can be identified. Functional homogeneity maxima tend to correspond to resolutions where few communities are in the respective size bands.

appear to be no clearly visible negative consequences of integrating A-type and P-type data without quality control to generate a PIN with a large coverage.

5.7 Discussion and conclusions

In this chapter we evaluated the ability of four community detection methods to capture the functional organization of PINs. Specifically, the methods were compared against the benchmark of optimal partitioning at a single resolution (or a narrow resolution band), and the suitability of the network partitions at this resolution to be used to investigate macrophage differentiation. While one community detection method partially satisfied our evaluation criteria, we did not find any evidence that functional organization can be perfectly captured in a single network partition. Indeed, evidence for a core-periphery structure of PINs suggests at least two distinct resolutions are required for core and periperal cellular functions, confirming results obtained by multi-resolution partitioning of yeast PINs [48]. In this setting, all community detection methods may generate functional modules

across resolutions. Thus, our evaluation benchmark for community detection may be overly focused on an unrealistic, ideal scenario.

Next to highlighting the merits of link clustering, our investigations revealed problems with detecting functionally homogeneous communities in PINs. These problems fall into two categories: challenges concerning the use of community detection methods on PINs, and issues with currently used functional evaluation approaches.

The identified obstacles for multi-scale PIN community detection ranged from problems with the Louvain algorithm that led to disconnected communities which could be easily addressed (see Appendix C), to issues regarding the coverage of non-trivial communities on the networks.

While every protein has a function and should thus be a member of a functional module, CPM Louvain and BigCLAM fail to partition large parts of the investigated PINs into non-trivial communities. As currently available PINs are noisy and incomplete (see Section 2.1.2.1), it is not expected that all underlying functional modules can be found by community detection. Yet, assigning up to $\approx 40\%$ of nodes to degree one communities at resolutions where the rest of the network is non-trivially partitioned (CPM Louvain), or failing to assign even larger proportions of the network to communities (BigCLAM) represents a lack of functional module prediction at an unexpectedly large scale. Given that we intend to use functional modules to elucidate molecular mechanisms of poorly understood diseases, a good coverage of functional modules on PINs is of central importance.

While the omission of a large proportion of the network allowed BigCLAM to generate functionally homogeneous network partitions (see Figure 5.13a), CPM Louvain did not show a noticeable signal of biologically relevant network partitioning. Overall, our results suggest that the concept of overlapping communities better resembles modules in PINs. This conclusion is supported by link clustering successfully partitioning BioGrid-AP-14 in particular. Comparing this result to link clustering's performance on HINT-P-14, and its limited ability to meaningfully partition the sparsely connected interleukins, suggests that PIN density may play a role for such community detection methods.

The above limitations show where the community detection methods have potential for improvement. Building a resolution parameter into BigCLAM as the threshold value for nodes to be assigned to a community (see Section 2.3.3.1) may improve both the coverage of communities on the PIN, and allow for better control of the resolution. Likewise, link clustering may benefit from an extension which takes into account more than the first neighbour sets when computing edge similarities to improve performance in low density environments (see Section 2.3.3.2). Alternatively, the density and coverage of available PINs could be increased by integrating protein interaction data from different databases, or including other molecules such as RNA or metabolites. Data integration has been found to improve module detection results [39].

The issues encountered concerning the functional evaluation approaches were exemplified by the IL-4/IL-10 community found in link clustering partitions of BioGrid-AP-14. Despite being a star-community with partly unrelated peripheral proteins, this community was evaluated as highly functionally homogeneous (FH of 7.62 or 7.98 depending on resolution and size). This high score is a consequence of the functional homogeneity being calculated over interacting proteins (see Equation (4.2)). While this approach does account for the higher functional similarity of interacting proteins compared to non-interacting proteins [48], it also allows for star-like communities to be evaluated as functionally homogeneous despite peripheral nodes not being functionally similar. As modules are defined as proteins that interact to perform cellular processes (see Section 1.1.1), all proteins in a biologically meaningful community should have similar functional annotations rather than only those that interact. However, the drawback of the alternative functional homogeneity approach defined by Equation (4.2) is the introduction of a dependence of functional homogeneity on the size of a community as argued in Section 5.2. This size dependence is shown in Figure 5.13.

Figure 5.13 shows that functional homogeneity calculated by averaging all pairwise protein functional similarities causes larger communities to be evaluated as less functionally homogeneous. This size effect, shown here for BioGrid-AP-14

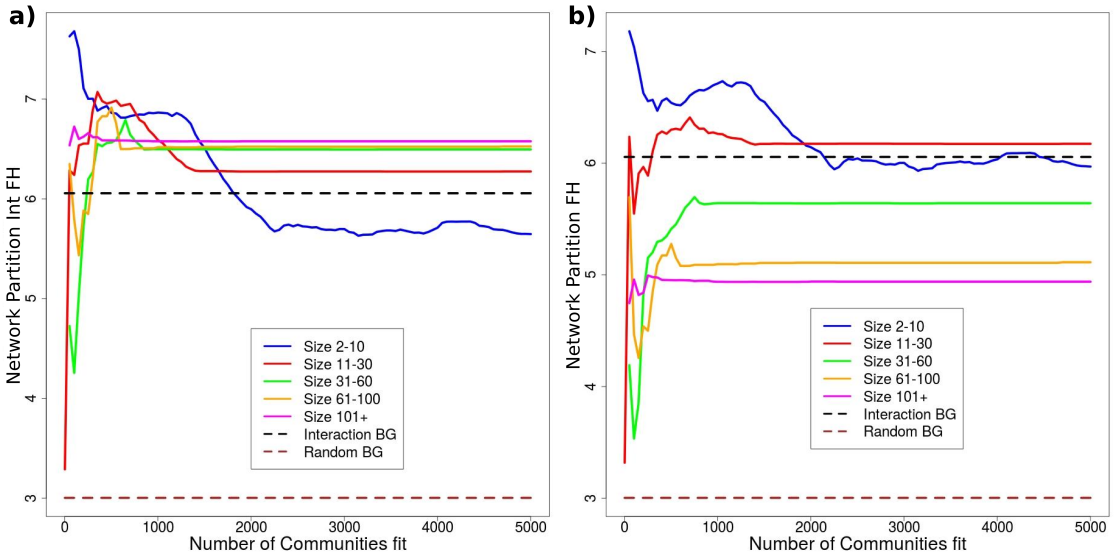

Figure 5.13: Comparison of functional homogeneity calculation approaches on BioGrid-AP-14 BigCLAM network partitions. Network partition functional homogeneity for BigCLAM BioGrid-AP-14 network covers was calculated in community size bands by a) Equation (5.1) and b) Equation (4.2). These equations incorporate: a) only interacting proteins (Int FH), or b) all protein pairs (FH), in the functional homogeneity calculation of a community. The functional homogeneity plots are put into perspective by comparison with the average functional similarity of interacting proteins in the PIN (black dotted line), and the average functional similarity of all node pairs (brown dotted line). Calculating functional homogeneity using all protein pairs introduces a size dependence into functional homogeneity. Larger communities tend to have lower functional homogeneity values.

BigCLAM, was observed across network partition data sets. Thus, when using the full comparison functional homogeneity approach it is necessary to relate community functional homogeneity values to those of other communities of the same size, rather than to a background similarity score.

A further issue discovered in the functional evaluation of communities are network partition functional homogeneity maxima at high resolutions where there were few communities (for example in Figure 5.8). This observation may be understood in light of inspection and annotation biases (see Sections 2.1.2.2 and 2.5.2.1). Well-studied proteins tend to have higher functional similarities with other proteins (see Annotation bias investigation in Section 6.2) and a higher degree [67, 79, 98]. When few communities are found in a specific size band at high resolutions, these are likely to be highly dense substructures as they are difficult to sub-partition despite the partition pressure. Such substructures can only exist in the PIN if all interactions connecting the member proteins are reported, suggesting these proteins must be

well-studied and are therefore evaluated as highly functionally homogeneous. Thus, global network partition functional homogeneity maxima may arise due to few communities of well-studied proteins rather than large proportions of functionally homogeneous communities (see global maxima leading to core-periphery argument in Section 5.5.2.1). Vice versa, this bias argument also has the consequence that poorly studied proteins may not be evaluated as significantly functionally homogeneous. Yet, it is exactly these poorly studied functions which hold the key to understanding biological problems such as that of macrophage differentiation.

The exploratory analysis performed in this chapter generated insights on how to proceed with the methodology, and on what must be improved in the functional module detection process. Based on these results we have developed the CommWalker framework, which is introduced in Chapter 6. CommWalker was conceived to evaluate communities in a size-dependent manner, overcoming the functional annotation bias causing well-studied proteins to be evaluated as more functionally similar to each other. Furthermore, an extension for link clustering that we developed to improve partitioning in less dense PIN environments is presented in Chapter A.

Partitioning PINs in a biologically meaningful way is difficult. The different characteristics of community detection methods and PINs make partitioning each network a unique problem. This uniqueness complicates the inference of conclusions on characteristics suitable for functional module detection. As such, we must emphasize that the conclusions drawn from this chapter may not be generally applicable.

Looking forward, the analysis in this chapter suggests that modules can be detected across resolutions. To further our project pipeline (see Section 1.1.3) it is thus important to consider communities generated across meaningful resolution ranges when identifying modules onto which differential gene expression data is projected. This result is implemented in Chapter 7 where the pipeline is applied to two biological data sets.

6 CommWalker

Contents

6.1 Introduction 135
6.2 Annotation bias 137
6.3 CommWalker 143
6.4 CommWalker module analysis 148
6.4.1 Thresholding 148
6.4.2 The effect of CommWalker on evaluation results 150
6.4.3 Coverage of functionally significant modules 152
6.4.4 Module statistics 152
6.5 CommWalker module validation 158
6.5.1 Gene co-expression validation 159
6.5.2 Case studies 163
6.6 Discussion and conclusions 164

This chapter closely follows a manuscript recently submitted for publication in Bioinformatics. The manuscript has been restructured and extended to incorporate more detailed explanations.

6.1 Introduction

There exists a plethora of approaches to functional module detection (see Section 1.1.2). In Chapter 5 we performed an exploratory study using a few popular methods to detect modules in PINs. We found that not only do these methods

not agree, but evaluating them is not straightforward.

In the absence of gold standard functional modules, functional annotations are commonly used to evaluate the homogeneity of proteins grouped in a module. While a significant amount of work has been done on how to assess the similarity of proteins (see Section 2.5.2), translating protein similarity to the evaluation of the functional homogeneity of protein modules has received less attention.

It has been well-documented that well-studied proteins tend to have more reported interactions than those which have received less research focus [67, 79, 98]. This phenomenon has been called "inspection bias" (see Section 2.1.2.2). Yet, while inspection bias is taken into account in assessing the quality of a reported interaction [69], or when predicting error rates in PINs [88], it is not considered that well-studied proteins may also be associated with more functional annotations ("annotation bias"). This consideration becomes important in combination with results showing that the number of annotations a protein is associated with affects semantic similarity measures which are used to assess protein functional similarity (annotation length bias; see Section 2.5.2.1).

Furthermore, using a semantic similarity measure specifically selected for module detection (see Chapter 4), we observed that larger communities were evaluated as less functionally homogeneous than smaller communities in our exploratory study (see Chapter 5). This size-effect biases module detection towards smaller communities.

These observations highlight the need for a framework that builds upon the work done on protein similarity assesssment and takes into account biases in the calculation of functional module homogeneity. We have developed such a framework, which we describe in this chapter.

We show the effect of annotation bias on module evaluation in PINs and detail a strategy of how this bias may be combated. This strategy was implemented in the CommWalker framework which is subsequently described.

We assessed whether CommWalker has the desired effect on module evaluation by comparing proposed modules accepted as functionally homogeneous by CommWalker to those accepted by conventional methods. To show the robustness of our framework,

this comparison was performed for proposed modules generated by four community detection methods on two PINs, using three semantic similarity measures. All 24 combinations of these inputs were used to show that CommWalker can accept modules of poorly-studied proteins, which conventional methods cannot.

Finally, CommWalker accepted modules were validated by systematic gene co-expression analysis, and more specifically using literature searches on case studies of individual modules.

6.2 Annotation bias

PINs are noisy and incomplete (see Section 2.1.2.1), and the extent to which a given protein has been studied affects its representation in the network. For example, well-studied proteins tend to have more reported interactions in PINs (inspection bias; see Section 2.1.2.2). Similarly, the better a protein is studied, the more functional annotations it is likely to have. Previous work has shown that the number of functional annotations affects semantic similarity measures (see annotation length bias in Section 2.5.2.1) which are used to determine protein functional similarities. Here we show how this effect impacts the evaluation of communities as functional modules.

The impact of research focus on module evaluation can be estimated by testing for correlation between the functional homogeneity of modules and how well-studied the proteins within these modules are. Performing this correlation study requires three definitions to be set:

What is a module? In the absence of gold-standard functional modules, community detection methods are used to obtain modules distributed across the network. Which community detection method should be used for our evaluation?

What is research focus? How can we quantify how well the proteins in a module are studied?

How should we quantify functional homogeneity? Functional homogeneity can be calculated directly by functional enrichment, or alternatively via protein functional similarity assesssment.

1. What is a module?

While it is generally accepted that network communities in PINs are related to functional modules [5,26], community detection methods tend to disagree with each other (eg. [27,144,212,213] and Chapter 5). Thus, to perform this investigation independently of a specific community detection method, we used short random walks to define proxy modules on the PINs. 10,000 length 3 random walks, and 10,000 length 6 random walks were performed from each node in the PINs to represent random proxy modules.

2. What is research focus?

There are several options for quantifying the research focus on proteins in these proxy modules: the number of functional annotations, the number of publications supporting protein-protein interactions, or more fundamentally whether the proteins have any associated functional annotations. Using the number of functional annotations to quantify research focus is based on the assumption that the better a protein is studied, the more is known about its function. Acknowledging that not all proteins are equally difficult to study, this assumption may still approximately hold. However, the number of annotations does not necessarily reflect how much is known about a protein. Due to ontology structure bias (see Section 2.4.1) in the GO, the specificity of a functional annotation is not reflected by its depth in the ontology. Yet, the depth of a term in the ontology determines the number of annotations in a term's ancestral set (see Section 2.4.1). Thus, the number of annotations is not an adequate approximation for research focus. In fact, as the number of functional annotations has already been shown to affect semantic similarity measures [186], the connection between the number of annotations and research focus is part of the topic of this investigation.

Alternatively, it may appear that the most direct way to measure research focus is via the number of publications. Publications can report evidence for a single observation or report high-throughput results supporting many interactions. Arguably, a small-scale study should carry a greater weight in determining research focus. Recent large-scale studies [80,81] have greatly increased the size of the known human interactome, such that a large proportion of interactions are supported by a very small number of publications. Therefore, focusing on small-scale studies may imply a lack of coverage on the network. Using a comprehensive text-mining approach on PubMed publications may be a more promising way of quantifying research focus by publications, but this is outside of the scope of this project.

Following the arguments presented, we used the binary label of whether a protein has any functional annotations or not to quantify research focus on modules (annotation fraction). For the length 3 random walk proxy modules, this gives only four possible values for the fraction of annotated proteins. To improve the sensitivity of this measure, we averaged the annotation fraction for all random walks (proxy modules) of the same size, started from the same node. Thus, the research focus on a node was defined as the average annotation fraction of 10,000 random walks started from this node. Using this quantification, the correlation test was performed on node "vicinities" defined by the proxy modules centred on the same node, rather than on individual modules.

3. How should we quantify functional homogeneity?

We quantified the functional homogeneity of proxy modules using simUI [183], simGIC [182], the Pandey measure [181], and functional enrichment. Despite having argued that the Pandey measure is the most suitable semantic similarity measure for evaluating modules in PINs in Chapter 4, we used a variety of approaches here to be able to draw more general conclusions. As research focus was quantified on node vicinities, functional homogeneity was averaged in the same way. In the case of simUI, simGIC, and the Pandey measure this was done by:

$$FH_i = \frac{1}{N_A} \sum_{c \in C_i} \sum_{(u,v) \in c} \rho_{SS}(u,v)\delta(u,v),$$

where the number of annotated nodes N_A is calculated by:

$$N_A = \sum_{c \in C_i} \sum_{(u,v) \in c} \delta(u,v).$$

Here, FH_i, denotes the functional homogeneity in the vicinity of node i and c is a proxy module in the set of all random walk proxy modules started from single node i, C_i. The term $\rho_{SS}(u,v)$ denotes the semantic similarity by measure SS between nodes u and v in proxy module c. The delta function is 1 when both u and v are annotated, and 0 otherwise.

Functional homogeneity calculation by functional enrichment was performed by taking the p-value of the most enriched GO BP term (see Section 2.5.1). As average functional homogeneities are computed over node vicinities, it is important that the p-values are comparable between proxy modules started from the same node. This comparability was ensured by adopting two measures. Firstly, we did not correct for multiple testing and secondly, random walk lengths were determined based on annotated nodes instead of traversed nodes. By omitting multiple-testing correction the p-values were not scaled by the number of terms annotated to proteins in a proxy module. While this prevents the p-value from being used as an evaluation of the statistical significance of a module, it instead allowed us to treat it as an enrichment score. The length of random walks was determined by annotated nodes instead of traversed nodes such that the range of enrichment scores for proxy modules were equal. If two proxy modules both contain six nodes, however only five of these are annotated in one of the modules, then the maximum number of nodes that can share an annotation differs in these proxy modules (five and six depending on the number of annotated nodes). Thus, by Equation (2.12) the minimum possible enrichment scores also differ. By terminating the random walks only after traversing a predetermined number of functionally annotated nodes, equal sample sizes are drawn from the set of annotated nodes, making enrichment scores comparable.

The annotation fraction and functional homogeneity scores for the node vicinities were ranked, and the Pearson correlation coefficient between the ranked variables was calculated - the so-called Spearman correlation coefficient. Under the null hypothesis of independence between the tested quantities, the standard errors for HINT-P and BioGrid-AP are 0.006, and 0.005 respectively by $\frac{0.67449}{\sqrt{(N-1)}}$ [214], where N is the number of nodes in the network. The correlation coefficients are shown in Table 6.1. All calculated correlation coefficients are significantly different from 0, representing the uncorrelated case. Differences between HINT-P and BioGrid-AP correlation scores may be due to HINT-P quality control increasing the effect of the bias (see Section 2.1.2.2) and a higher false-positive error rate in BioGrid-AP (see Section 3.1.1).

Table 6.1: Annotation bias investigation results.

Network	Functional Similarity	Walk Length 3	Walk Length 6
HINT-P	Pandey	0.414	0.572
	SimUI	0.200	0.202
	SimGIC	0.240	0.269
	Func Enrich	-0.473	-0.621
BioGrid-AP	Pandey	0.209	0.266
	SimUI	0.107	0.085
	SimGIC	0.134	0.104
	Func Enrich	-0.354	-0.409

Table 6.1: Spearman correlation coefficients between the functional homogeneity and annotation fraction score in node vicinities for four functional homogeneity measures. Negative coefficients are expected for functional enrichment, as more functionally homogeneous communities exhibit lower p-value scores in contrast to the semantic similarity measures. All correlations are significantly different from zero, under the null hypothesis of independence.

Our results show that proteins with fewer functionally annotated proteins in their vicinity tend to have a lower functional similarity to proteins in their vicinity. Furthermore, given that the correlation was performed on node vicinities, and that node vicinities of neighbouring nodes heavily overlap, the results suggest that there are regions in PINs where proteins have lower functional similarity scores, and regions of high functional similarity. An example of such regions can be seen in Figure 6.1. Figure 6.1 shows a subnetwork of HINT-P which exhibits a high functional similarity region in the top-right, and a low functional similarity region towards the bottom.

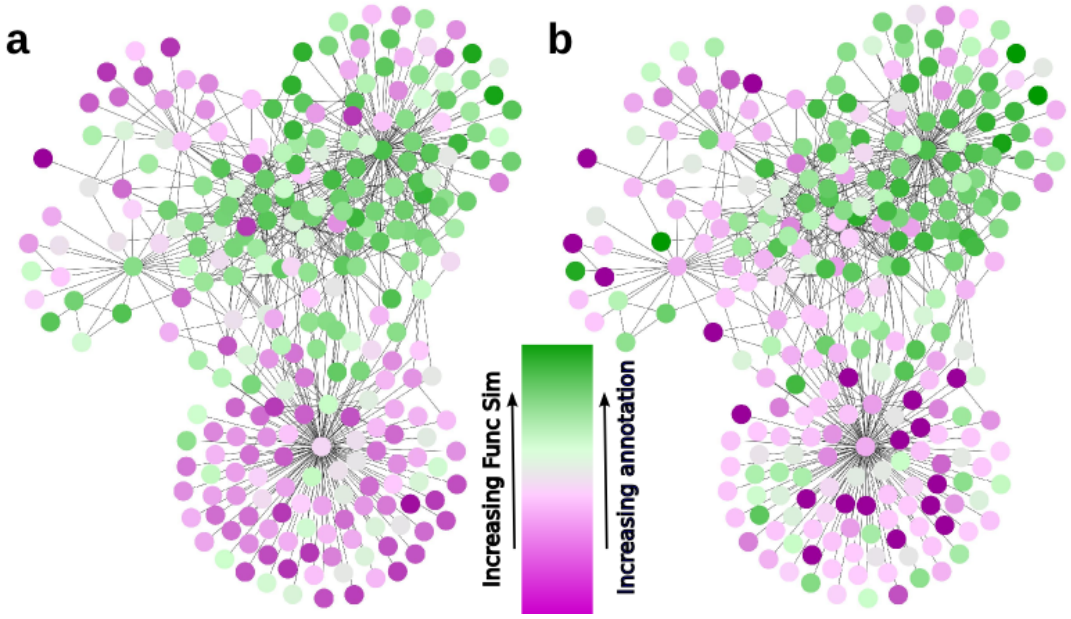

Figure 6.1: Semantic similarity and research focus correlation. The correlation between semantic similarity and research focus is shown on a subgraph of HINT-P, generated by taking all nodes connected to the gene FAT1 through at most two intermediary genes. (a) is coloured by the average functional homogeneity in size 3 proxy modules around the proteins, and (b) is coloured by the average fraction of annotated nodes (research focus) in these proxy modules. Regions of high functional homogeneity correlate with regions with strong research focus.

In regions of high functional similarity, proteins may be evaluated as functionally similar to random nodes in their local neighbourhood. Communities in these regions are therefore more likely to be evaluated as highly functionally homogeneous. Vice versa, due to annotation bias a biologically meaningful community in a low functional similarity region may be evaluated as less functionally homogeneous than a random community in a high functional similarity region. In this way, the heterogeneity of annotation in PINs biases module evaluation towards communities of well-studied proteins.

This effect can be counteracted by taking into account the local environment of a community in module evaluation. Specifically, a significance can be assigned to a community's functional homogeneity score based on the background functional similarity distribution of the community. We have developed the CommWalker framework to counteract the overestimation of the functional homogeneity of communities in well-studied network regions, while allowing for a rebalanced evaluation of modules in poorly-studied environments.

6.3 CommWalker

CommWalker is a module evaluation framework which can be used with any similarity measure defined between proteins. Using these measures, CommWalker calculates the functional significance of a community by relating its functional homogeneity score to a background functional similarity distribution from the community's local network environment. The CommWalker methodology is shown in Figure 6.2.

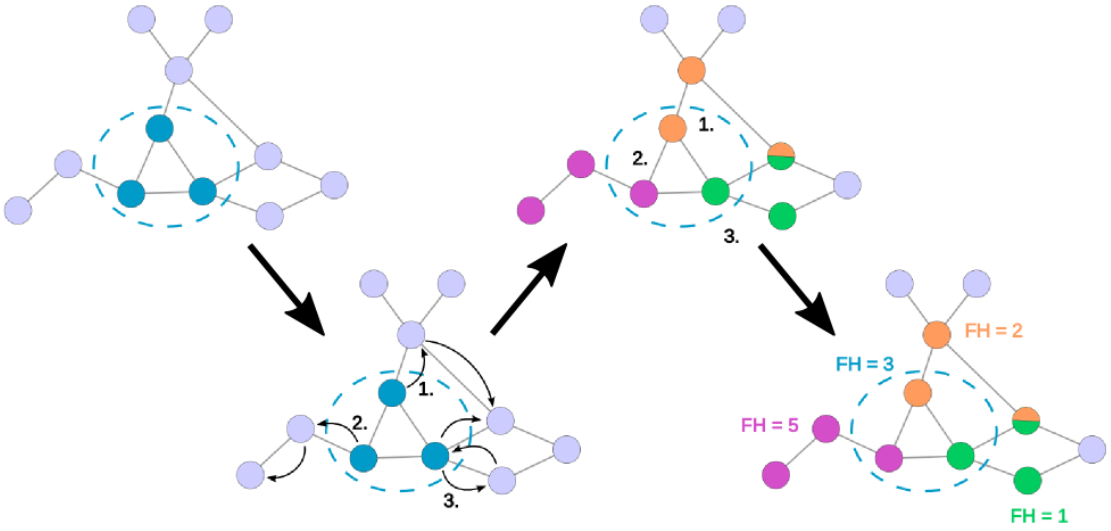

Figure 6.2: Schematic diagram of the methodology behind CommWalker. Random walks are started from the community nodes (dark blue nodes) to sample the local network area. These random walks are terminated upon visiting N_C nodes, where N_C denotes the number of nodes in the community (Here $N_C = 3$). The random walks represent proxy communities (orange, magenta, and green), whose functional homogeneity scores give the local context for the community's functional significance evaluation. The random walks may overlap as shown by the orange and green node. At a functional homogeneity score of 3, the T-value of the dark blue community is $\frac{1+1}{3+1} = 0.5$, as one proxy community has a higher functional homogeneity (Equation (6.1)).

Using short random walks, CommWalker probes the local network environment of a community to obtain a background functional homogeneity distribution. Random walks of length N_C (see Section 2.2.2) are started from each node in a community of N_C nodes to sample the network environment. These random walks can be loosely interpreted as proxy communities that represent an alternative choice of node grouping in this network environment. The functional homogeneity scores of the random walk proxy communities, as calculated by the chosen similarity measure, make up the background functional homogeneity distribution of the community.

This background distribution is used to assign a community functional significance score, called the tail-value or T-value T_C, by the equation:

$$T_C = \frac{m_C + 1}{M_C + 1}. \tag{6.1}$$

Here m_C denotes the number of random walks with a functional homogeneity higher than that of the community C, and M_C is the total number of random walks performed from community C to obtain the background distribution. The term M_C can be expressed in terms of the number of random walks started from each node in the community W_C, and the number of nodes in the community by:

$$M_C = W_C \times N_C.$$

The T-value is thus the fraction of the background functional homogeneity distribution in the upper tail as determined by the community functional homogeneity score (see Figure 6.3a).

Given a functional similarity measure, FH(), and a network partition, denoted "partition", the CommWalker framework is implemented algorithmically

in the following way.

Algorithm 6.3.1: CommWalker(partition, FH())

```
comment: W_C: Number of random walks from each node in the community
comment: FH(): Similarity measure to compute the functional homogeneity
comment: f(): computes W_C based on the community size (Equation (6.2))
for each comm ∈ partition
do { W_C ← f(|comm|)
     commFH ← FH(comm)

     M_C ← 0
     m_C ← 0

     for each node ∈ comm
       do
     do { for i ← 1 to W_C
            do
          { random walk from node
            randFH ← FH(random walk)
            M_C ← M_C + 1

          { if (randFH > commFH)
              then m_C ← m_C + 1

     T_C ← (m_C + 1)/(M_C + 1)
```

The stability of the significance score output by this algorithm depends on how well the local network environment was captured in the sampling process (Figure 6.3). To investigate when the environments of all communities are extracted adequately, we quantified the resampling of a community via the node score Z as follows.

For a community C of size N_C, each random walk started from the community visits N_C distinct nodes. If W_C random walks are started at each node in the community, a total of $Z = W_C N_C^2$ nodes are sampled by the random walks from community C. We call Z the node count for community C. To ensure that each community is sampled to a similar extent, the node count Z is kept near constant. Thus, the number of random walks per node for a community C (denoted f(|comm|) in the pseudocode above), can be calculated by the equation:

$$W_C = \frac{Z}{N_C^2}. \tag{6.2}$$

W_C is rounded up to the next integer so that the same number of random walks can be started at each node in the community, preventing a background sampling bias. Due to this rounding the actual number of nodes visited in the completed random walks from each community will tend to be slightly higher than Z. As a consequence of keeping Z near constant for all communities, more random walks are performed for smaller communities, as each random walk is shorter. Alternative ways to scale the number of random walks with community size were investigated and found to lead to T-value stability depending on community size given a preliminary analysis.

To find the node count value Z that gives the best trade-off between T-value stability and algorithm run-time, we randomly selected nine communities of sizes ≤ 35 from HINT-P (see expected module size ranges in Section 1.1.3) for repeated T-value measurement (Figure 6.3b). The stability of the T-value of a community was determined by running 100 repeats of CommWalker on the same community and taking the standard error of the resulting T-value samples.

The stability of the nine randomly selected communities was calculated at five different node counts. Figure 6.3b shows the trade-off between T-value stability and the node count, which is linearly related to the run time of the algorithm. Given these data, a node count of 100,000 was selected. As estimated based on the nine communities tested, T-values are taken with an associated standard error of ≈ 0.005 at this node count. While this stability calculation was repeated on BioGrid-AP to give similar results (Figure D.1 in Appendix D), it may vary between networks.

To further ensure a fast run time for CommWalker and a high stability of computed T-values, we implemented two filters in the algorithm.

The first filter is the community size that can be evaluated. A hard lower boundary for community size was set at three proteins, which is the minimum non-trivial community size. An upper boundary on the community size was set by the heuristic equation $\frac{N}{15} + 20$, where N is the number of nodes in the PIN.

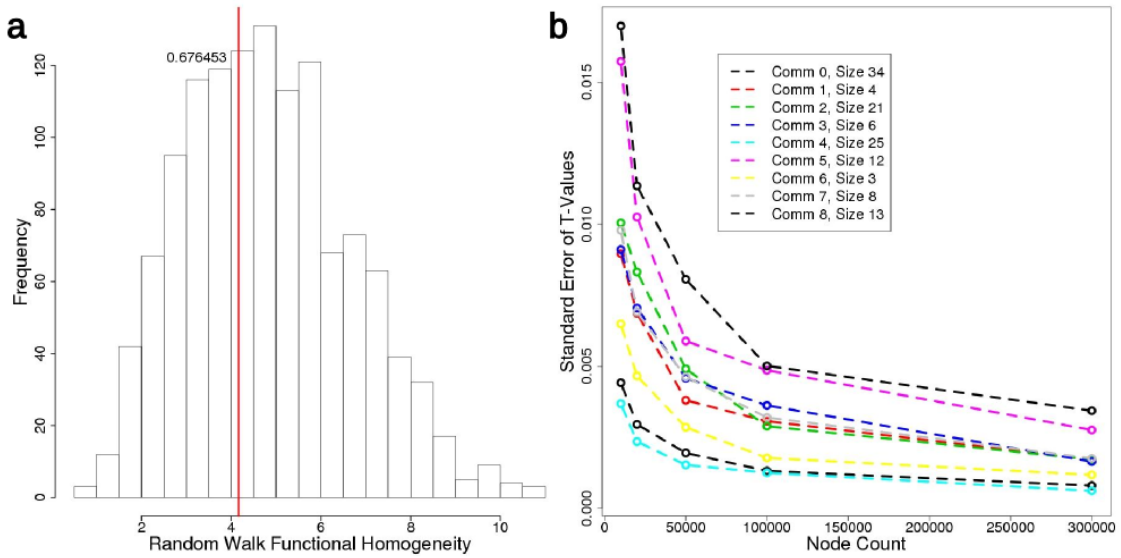

Figure 6.3: T-value calculation and stability. a) Background functional homogeneity distribution generated from 2,500 random walks for a community of size 6 in HINT-P. The red vertical line represents the community's functional homogeneity and its associated T-value of 0.676453. b) Standard errors of community T-values computed over 100 CommWalker runs on nine randomly selected communities of sizes ≤ 35 from HINT-P. The number of random walks started at each node is calculated from the node count by Equation (6.2). The trade-off between the stability of T-values and the run time, which is proportional to the node count, is deemed optimal at a node count of 100,000 based on this HINT-P sample. The random walk functional homogeneity distribution for the community denoted "Comm 3" at a node count of 10,000 is shown in a).

With currently available PINs often containing $> 10,000$ proteins, this boundary filters out communities that are unlikely to be of biological interest and are slow to evaluate, while allowing for meaningful community sizes in virus PINs with ≈ 100 nodes. The upper boundary filter can be switched off.

The second filter was implemented on the random walks. Random walk evaluation can present two problems. Firstly, walks may become trapped in a bottleneck in the network which prolongs the run time of the algorithm, and secondly it may be that no two proteins in a completed walk can be compared due to a lack of annotation. The first case is overcome by allowing a maximum of N_C^2 steps when attempting to reach N_C nodes, where N_C is the number of nodes in the community. For the second case we imposed a maximum number of random walk attempts. CommWalker aims to perform $W_C \times N_C$ successful random walks per community (see Equation (6.2)), where the total number of attempts allowed

to reach this target is twice this number. The number of times that the random walks are restarted per community is output by the software, and can be used to investigate confidence in a particular community T-value.

6.4 CommWalker module analysis

CommWalker is designed as a module evaluation framework which counteracts annotation bias to allow for a rebalanced module evaluation that does not disadvantage poorly-studied network regions. As such, the efficacy of this framework can be investigated via the communities CommWalker evaluates as functionally significant.

In this section we first show the extent to which module evaluation is changed by considering the local context via CommWalker. Subsequently, we demonstrate how the different prioritization of module characteristics by CommWalker affects how many proteins we can evaluate as being in functionally significant modules. Finally, we present summary statistics of communities evaluated as functionally significant by CommWalker versus conventional functional homogeneity evaluation, and show how proteins in modules accepted by these methods differ in their local environments.

In our analysis CommWalker community evaluation was compared to the established method of evaluating communities based solely on their functional homogeneity. These are not independent evaluation methods as CommWalker uses functional homogeneity to score its random walks. When functional homogeneity scoring and CommWalker were compared, we deliberately contrasted the use of functional homogeneity alone to evaluate communities, with including local network information in the calculation. In principle CommWalker can be used with any similarity measure between nodes in a network.

6.4.1 Thresholding

In order to compare CommWalker to functional homogeneity based on the communities these methods evaluate as functionally significant, the term functionally significant must be defined. We need a classification of communities into functionally-significant and non-functionally-significant categories based on a threshold. To

be able to compare the two methods, this threshold must be sufficiently similar in both methods, despite T-values and functional homogeneity scores not being directly comparable.

To fulfill this requirement we chose a T-value of 0.5, and the median functional similarity of interacting proteins in a PIN. At a T-value of 0.5 approximately half of the random walks have a functional homogeneity at least as high as that of the community. Similarly, the median of the functional similarities of interacting proteins is the largest value where at least half of the interacting proteins in the network have a functional similarity at least this high. While these thresholds are arbitrary, they are qualitatively similar.

How similar the two thresholds actually are also depends on the distribution of the functional similarity scores. As T-values are calculated based on community functional homogeneities, they are dependent on protein functional similarity averages by the definition of functional homogeneity in Equation (2.16). Thus, the more mean and median values of pairwise protein functional similarities differ, the more our chosen qualitatively similar thresholds do. The distribution of functional similarity scores for simUI, simGIC and the Pandey measure on proteins in HINT-P is shown in Figure 6.4. Due to the skewed nature of the simGIC score distribution, the mean functional similarity score is substantially higher than the median. Indeed, for HINT-P the mean functional similarity of interacting proteins by simGIC is 0.1008, while the median is 0.0597. This equates to a fractional difference of 40.8%. In comparison, the fractional differences on HINT-P for simUI and the Pandey measure are 20.6% and 5.4% respectively.

This effect will be counteracted by community functional homogeneity being calculated over interacting and non-interacting proteins in a community. When the functional homogeneity is compared to the median functional similarity of interacting proteins, non-interacting proteins having lower functional similarity scores (see Figure 6.4) will make this threshold appear more stringent. It is difficult to assess the size of this effect as a community represents a selection for interacting

protein pairs. Overall, it is likely that the functional homogeneity threshold is more lenient for simGIC, and slightly more lenient for simUI.

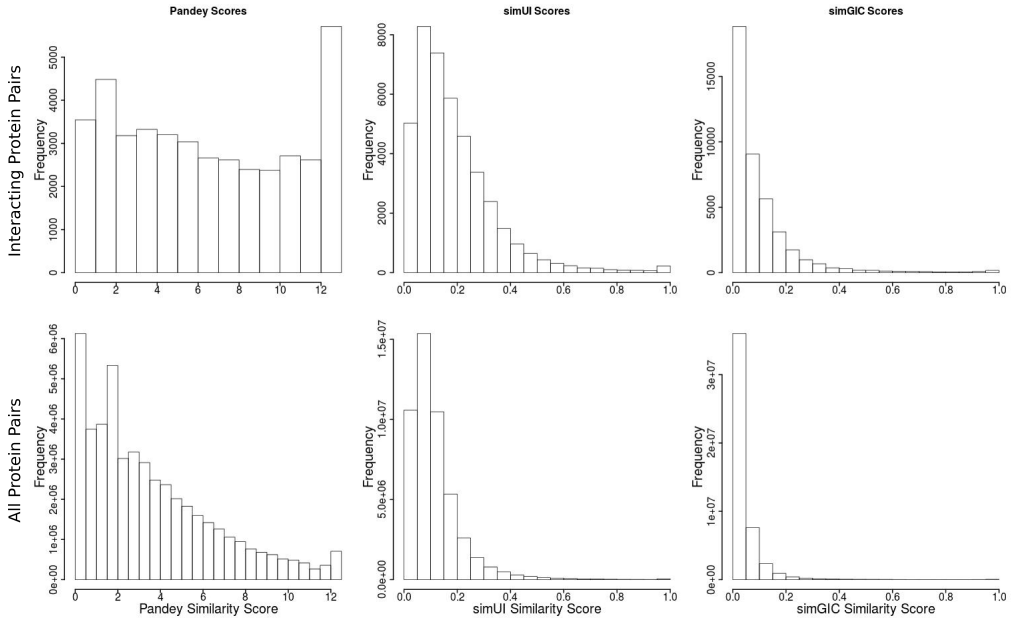

Figure 6.4: Functional similarity score distributions for HINT-P. Distribution of functional similarity scores for the Pandey measure, simUI, and simGIC between interacting protein pairs (top row), and all protein pairs (bottom row) in HINT-P. The scores clearly differ for interacting proteins compared to all proteins, with Pandey scores showing the strongest signal. Furthermore, the interacting protein score distributions show that the median score will be far lower than the mean score for simGIC.

6.4.2 The effect of CommWalker on evaluation results

To understand the extent to which CommWalker changes module evaluation results, we investigated the re-ordering that occurs when communities are ranked by CommWalker versus when they are ranked by functional homogeneity. For this analysis network partitions were screened to find the resolution at which there is a maximum number of proteins in functionally significant communities. Here functional significance was determined by a T-value threshold of 0.5, as calculated by CommWalker using the Pandey measure. For the network partition defined by this resolution, T-value and functional homogeneity scores were calculated for all communities using the Pandey measure. For a given T-value threshold the set of proteins in T-value-significant communities at this resolution was computed, and an equivalent functional homogeneity threshold was identified which gave the

Table 6.2: CommWalker community prioritization.

PIN	Comm Det	$T_C \leq 0.05$	$T_C \leq 0.1$	$T_C \leq 0.25$	$T_C \leq 0.5$
HINT-P	Link	0.631	0.724	0.839	0.948
	BigCLAM	0.630	0.791	0.873	0.964
	Config Mod	0.677	0.663	0.737	0.891
	CPM Mod	0.547	0.620	0.776	0.925
BioGrid-AP	Link	0.758	0.816	0.896	0.950
	BigCLAM	0.713	0.795	0.907	0.984
	Config Mod	0.730	0.761	0.848	0.950
	CPM Mod	0.648	0.727	0.816	0.933

Table 6.2: The fraction of proteins that are common to both the top T-value ranked communities and the top functional homogeneity ranked communities for all PIN and community detection (Comm Det) method combinations using the Pandey measure. The set of unique proteins in communities with a T-value below the given thresholds is compared to the set of proteins in the highest ranked functional homogeneity communities of the most similar size. The fraction is calculated by dividing by the smaller protein set.

most similar number of proteins in functional-homogeneity-significant communities. The overlap of these two protein sets was calculated to determine the similarity of the two evaluation approaches. This investigation was repeated for other PINs, community detection methods, and T-value thresholds (Table 6.2). Table 6.2 shows that while there is an overlap between the two evaluation methods, the ordering of communities according to T-value is different from that of functional homogeneity.

The overlap between the two protein sets at higher thresholds shows that CommWalker confirms a large proportion of the communities found by functional homogeneity. The communities evaluated as functionally significant by both methods are prime candidates for functional modules or protein complexes as they are verified by a two-pronged approach.

In contrast, especially at low T-value thresholds, the T-value and functional homogeneity protein sets are different. At a T-value threshold of 0.05, overlaps for the different networks and methods range from ≈55% to ≈76%. The different selection of communities by T-value and functional homogeneity evident from these results shows that CommWalker's approach of evaluating communities based on whether they stand out from their environment prioritizes different communities from the classical functional homogeneity approach.

6.4.3 Coverage of functionally significant modules

Following on from the previous investigation, the question arises as to whether the different prioritization of modules by CommWalker results in it being able to rebalance module evaluation to capture a larger proportion of a PIN. We addressed this question by counting the number of proteins in communities which are evaluated as functionally significant based on the qualitatively similar thresholds discussed in Section 6.4.1. As we are analysing network partitions that may be overlapping, we take into account only distinct proteins in this analysis in order to evaluate the network coverage accurately. The number of proteins in functionally significant communities by our thresholds for the BioGrid-AP link clustering Pandey measure data set are shown by the black and teal lines in Figure 6.5. This analysis was performed for all combinations of two PINs, four community detection methods and three functional similarity measures (Figures D.2–D.9 in Appendix D).

Using the Pandey measure there are consistently more unique proteins in CommWalker accepted communities than in those accepted by functional homogeneity. For simGIC this trend is reversed, which is expected given the distribution of the simGIC functional similarity scores. The distribution suggests that the functional homogeneity threshold is more lenient for simGIC, as argued in Section 6.4.1. The results for simUI are less clear, which again can be explained by the functional similarity score distributions. Taking the score distribution in Section 6.4.1 into consideration, CommWalker appears to accept more communities as functionally significant than functional homogeneity evaluation. This difference is more pronounced the more similar the chosen thresholds are.

6.4.4 Module statistics

We have shown that CommWalker module evaluation reprioritizes communities to give a greater coverage of functionally significant communities on PINs. The next step is to assess whether the improved coverage is the result of the intended increased sensitivity in poorly annotated network regions. To answer this question we identified characteristics of those communities which are evaluated as functionally significant

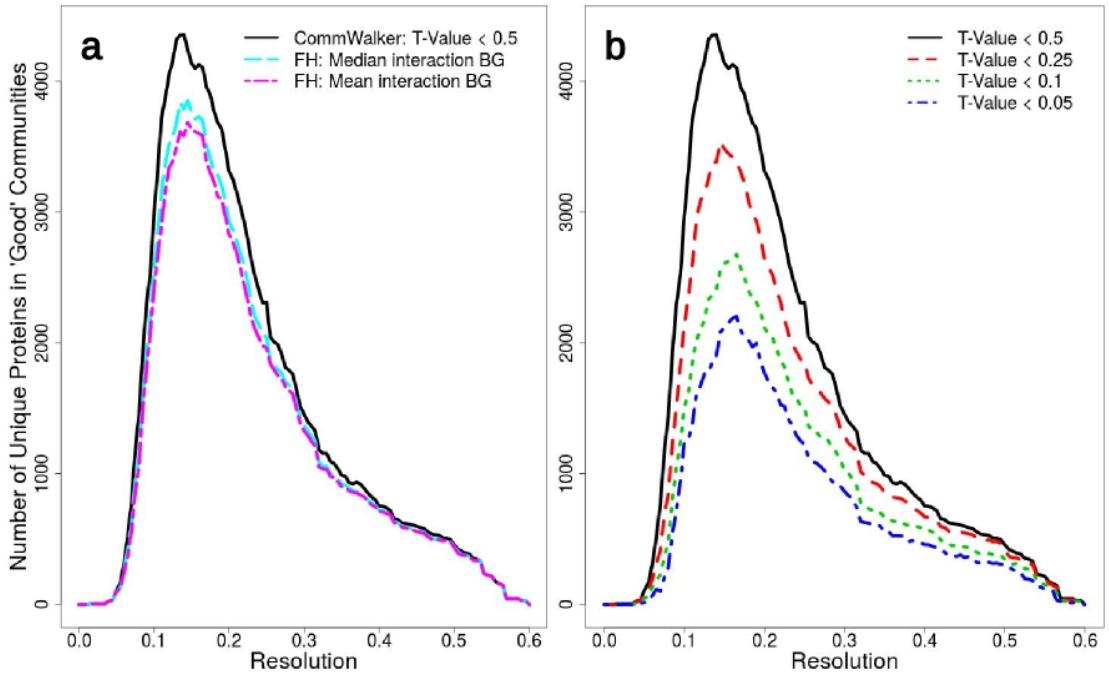

Figure 6.5: CommWalker protein count comparison. The number of distinct proteins evaluated as being in functionally significant communities of size 6 - 35. Functional homogeneity and CommWalker are compared using qualitatively similar thresholds in (a) and different T-value thresholds are compared in (b). The communities were generated by applying link clustering to BioGrid-AP. (a) shows that CommWalker detects a higher coverage of functionally significant communities than functional homogeneity at qualitatively similar thresholds (median functional similarity of interacting proteins and a T-value of 0.5). The effect of changing the thresholds can be seen in (b). At different T-value thresholds the plots exhibit a consistent maximum at a resolution of ≈ 0.145.

by one method but not the other. Using the qualitatively similar thresholds from Section 6.4.1 communities were divided into four sets: accepted by both methods; accepted only by CommWalker; accepted only by functional homogeneity; and rejected by both methods. Average network statistics for these community sets were calculated for both HINT-P and BioGrid-AP, all four community detection methods, and all three functional similarity measures. The results are shown in Figures 6.6 and 6.7 for the Pandey measure, and in Figures D.10–D.13 in Appendix D for simUI and simGIC.

The four community sets were analysed using three summary statistics: the average community size; the average level of annotation in the local environments (average random walk annotation fraction of random walks from the same community after Section 6.2); and the average level of annotation of the communities (average

community annotation fraction). The upper row of graphs in Figures 6.6, 6.7, and Figures D.10–D.13 show that communities accepted only by functional homogeneity are the smallest set for the Pandey measure and simUI, and communities accepted only by CommWalker are the smallest set using simGIC. In some data sets the smallest set becomes so small that the summary statistics calculated for this set are very variable, which affects our ability to interpret them. This difficulty can be observed in for example both of the Pandey measure Modularity Maximization data set analyses of the annotation levels on BioGrid-AP (bottom right quadrant Figure 6.6). The grey lines here are very variable due to a low sample size. Towards higher resolutions, where the proportion of communities only accepted by functional homogeneity increases, the plots become steadier and therefore more reliable to interpret. The same effect occurs in the BioGrid-AP BigCLAM annotation fraction graphs. Between 1 and 2,251 communities fitted to the data, there is a single community in the set of communities accepted by only functional homogeneity (grey line). A further jump in the grey lines can be seen at 2,751 fitted communities, when the number of communities jumps from three to seven. The small sample size is specifically an issue for the simGIC data shown in Appendix D.3. As a general rule, the community statistics are more robust at higher resolution due to a larger number of communities. So when network statistics become very variable, we focus our interpretation on results at high resolutions.

Taking small sample size effects into account, Figures 6.6, 6.7, and D.10–D.13 show that using functional homogeneity selects for smaller community sizes, for communities that have a higher proportion of annotated nodes, and for those that are in well-annotated environments. In contrast, CommWalker accepted communities show a broader distribution in these statistics. As functional modules should span PINs as every protein has a function, biologically relevant communities should span network regions and therefore also the investigated statistics. CommWalker communities' broad distribution is particularly visible for the average random walk annotation fraction statistic. Here, the blue lines representing communities only accepted by CommWalker tend to be more similar to the turquoise dashed lines for

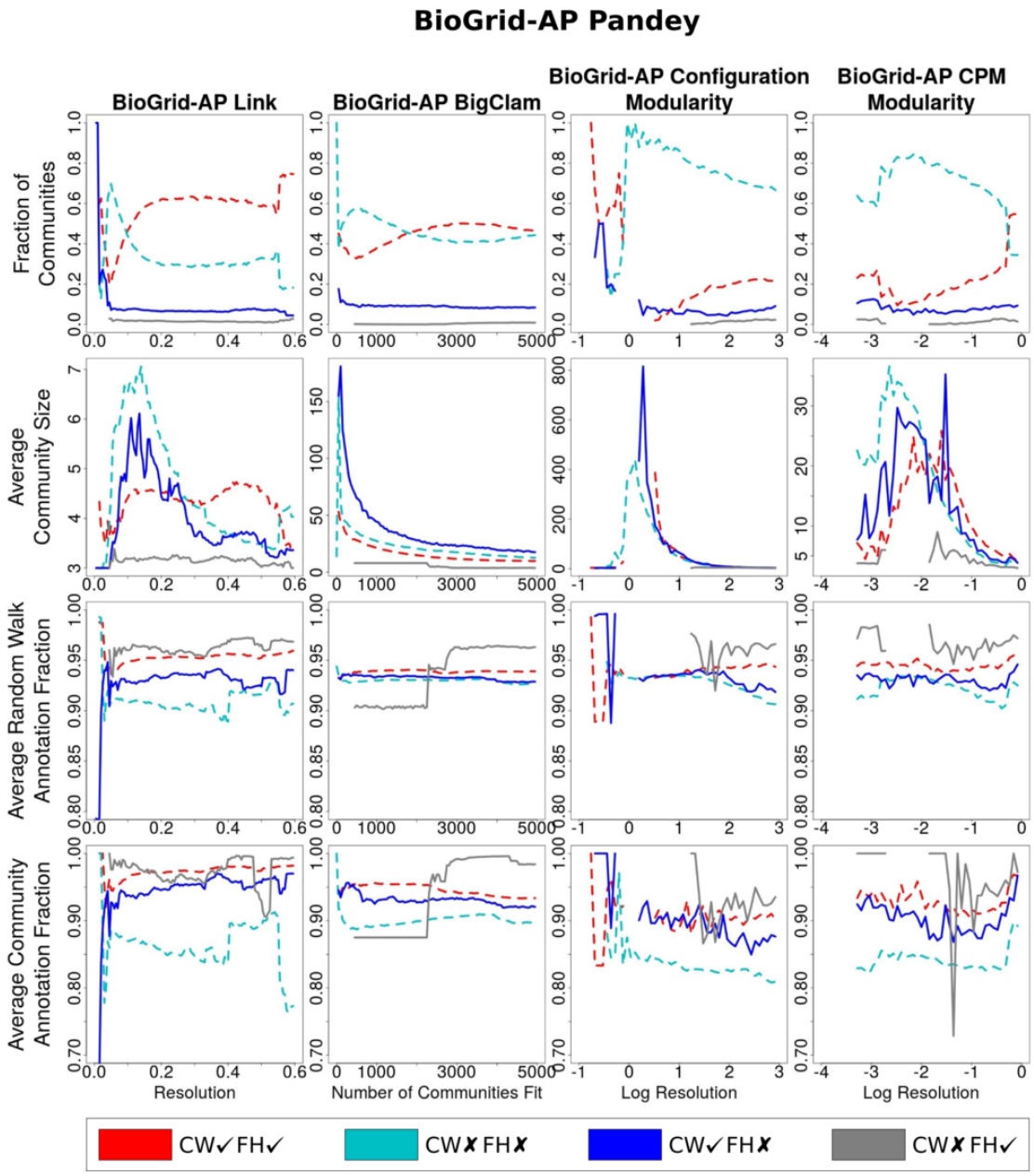

Figure 6.6: BioGrid-AP Pandey measure community summary statistics. Community statistics for BioGrid-AP communities generated by four multi-resolution community detection methods. Communities were divided into four groups depending on whether they were accepted or rejected by the qualitatively similar thresholds discussed in Section 6.4.1 using the Pandey measure. Communities only accepted by CommWalker or functional homogeneity thresholds are plotted as blue and grey lines respectively, and communities accepted or rejected by both methods are shown as red or turquoise dotted lines respectively. The fraction of communities that fall into these categories is shown in the top row, with the following rows showing the average community size, the average level of annotation of the community, and the average annotation level of the communities' environments. The data in the lower two rows of graphs are only shown for comparison fraction ranges of 0.6 - 1.0 and 0.75 - 1.0 to emphasize the comparison between the data sets. As communities accepted only by functional homogeneity (grey lines) are the smallest set, summary statistics calculated for this set can be very variable. Taking these effects into account, the data suggest that functional homogeneity selects for smaller communities in well-annotated environments which also have a high level of annotation themselves. In contrast, T-value significant communities tend to have a broad distribution in the investigated statistics as seen by the lines representing CommWalker accepted communities (red dashed line and blue line) and do not seem to favour certain community sizes.

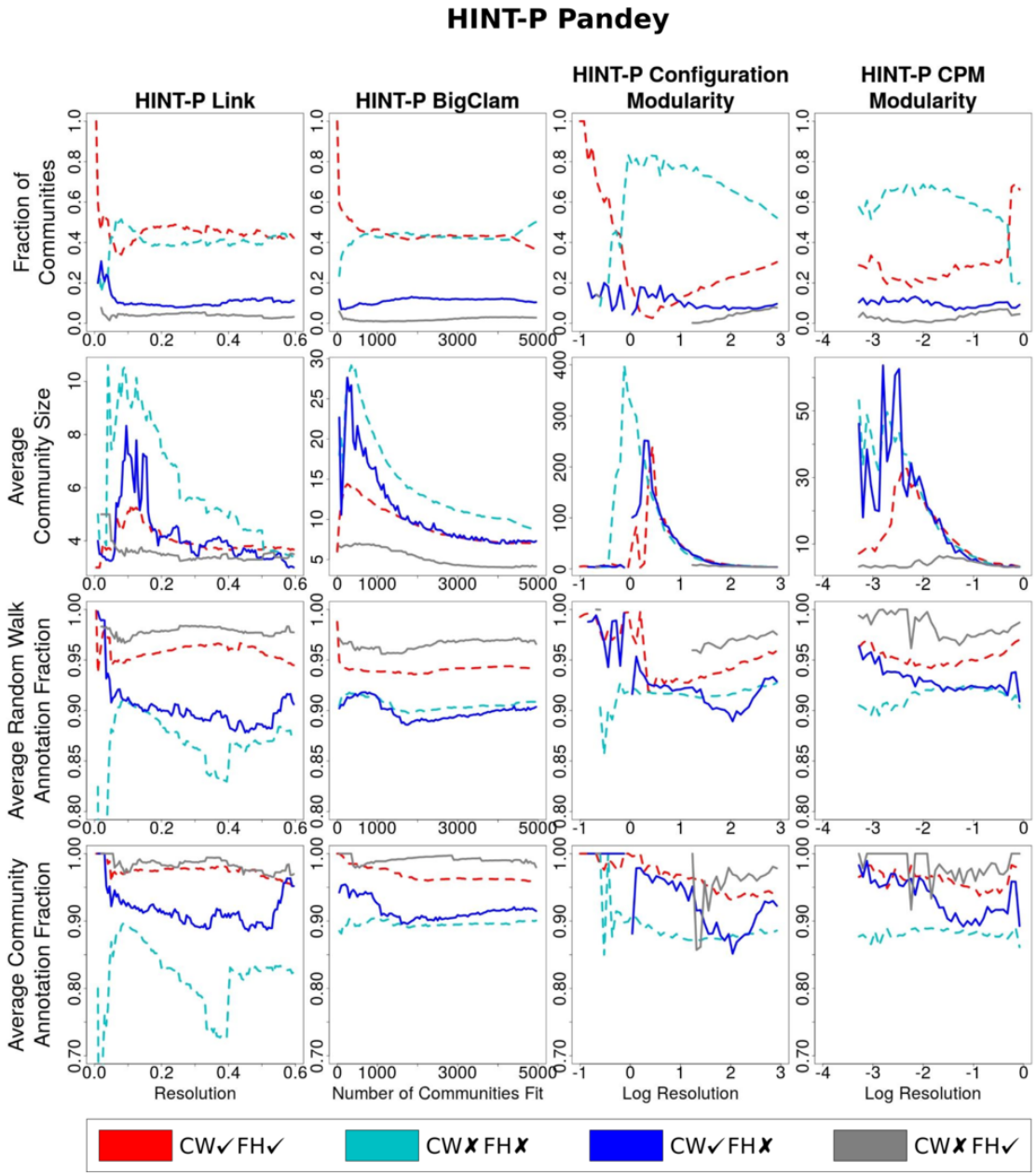

Figure 6.7: HINT-P Pandey measure community summary statistics. Community statistics for HINT-P communities generated by four multi-resolution community detection methods. The data shown was generated as in Figure 6.6 and suggests that functional homogeneity selects for smaller communities in well-annotated environments which also have a high level of annotation themselves. In contrast, T-value significant communities tend to have a broad distribution in the investigated statistics as seen by the lines representing CommWalker accepted communities (red dashed line and blue line) and do not seem to favour certain community sizes.

communities mutually rejected, than the red dashed lines for communities mutually accepted. Hence communities accepted by CommWalker (red dashed line and blue line) appear to show a larger spread in the random walk comparison fraction. As the turquoise line represents ranges of these statistics in which significant communities are difficult to detect, the results show that CommWalker can find communities that are significant even in ranges of the network annotation statistic in which communities tend to be rejected by functional homogeneity. CommWalker may detect functionally significant communities in regions of the network otherwise obscured by poor annotation coverage.

CommWalker's increased sensitivity in low functional similarity regions of PINs is further shown in Figure 6.8, which compares the distributions of proteins in functionally significant communities. For ease of visualization, non-overlapping community data were used from configuration model Modularity Maximization on HINT-P, in conjunction with the Pandey measure. In Figure 6.8 the proteins are ordered by their functional similarity with their "vicinity", measured using random walks as described in Section 6.2. Proteins towards the left have higher similarity with their vicinity and will thus tend to be better studied (see investigation in Section 6.2). On this layout we show the distribution of proteins in communities that were accepted as modules by both methods (Figure 6.8b), only by CommWalker (Figure 6.8c), or only by functional homogeneity (Figure 6.8d). Proteins in modules accepted by the standard functional homogeneity approach (Figure 6.8b,d) tend to be distributed towards the well-studied left side of the Figure. In contrast, modules accepted only by CommWalker (Figure 6.8c) have a broader distribution, reaching into the poorly-studied protein regions. Using non-overlapping community detection methods for both PINs and all three semantic similarity measures produced similar results (see Section D.3 in Appendix D). In light of these results, we can conclude that CommWalker is indeed sucessfully accepting modules in poorly-studied network regions, that are not prioritized by the conventional functional homogeneity approach.

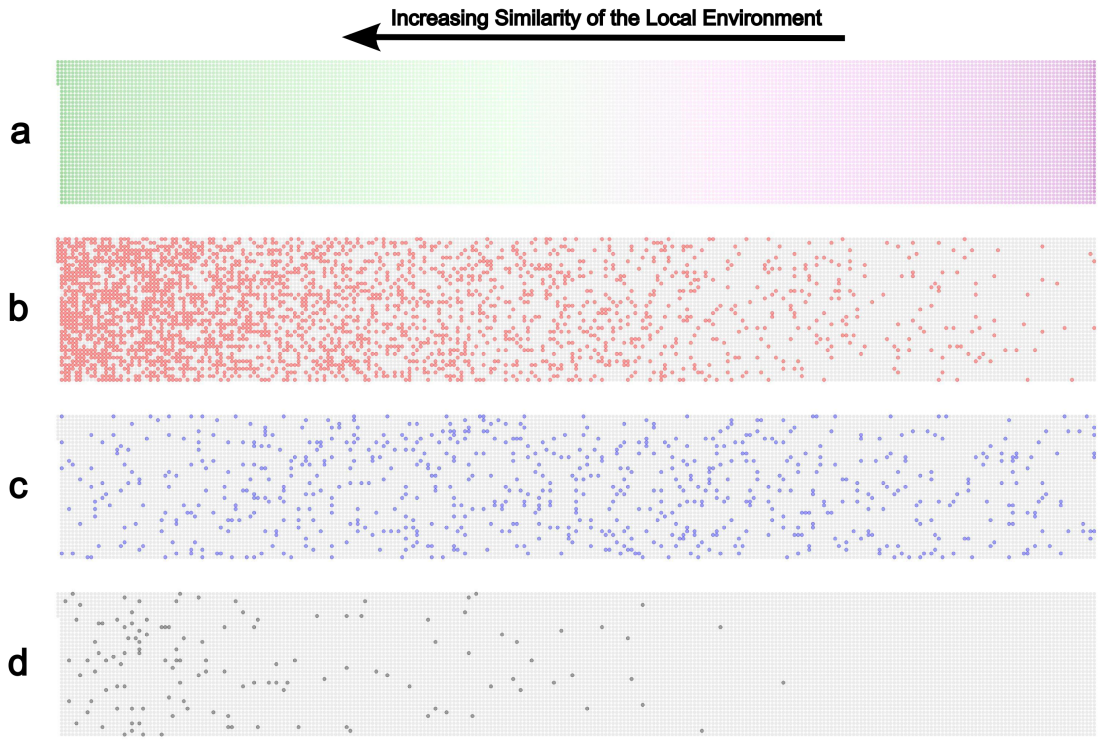

Figure 6.8: Local environment comparison of functionally significant community proteins. Nodes in HINT-P were ordered by their Pandey functional similarity with nodes in their vicinity as shown in (a). Nodes towards the left have higher similarity with their environment than those towards the right with the vertical dimension used purely for ease of visualization. Communities were generated by configuration model Modularity Maximization at the resolution where the maximum number of proteins are found in functionally significant communities by a T-value threshold of 0.5 (log resolution = 1.80, see Section D.2 in Appendix D). On this network layout the proteins in communities identified as functionally significant by CommWalker and functional homogeneity are shown in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d). The further left the coloured nodes are, the higher the functional similarity scores with their environment.

6.5 CommWalker module validation

CommWalker's greater sensitivity in poorly-studied network regions results from an increased leniency in module evaluation in these regions. Under the assumption that functional module detection should span PINs as every protein performs a function, this leniency is theoretically warranted. Practically however it may be that community detection fails in poorly-studied network regions, due to greater error rates in local network topology compared to well-studied regions. In this section we address this question by computational validation of modules which are accepted by CommWalker.

As we validate predictions from a framework that incorporates GO annotations with protein-protein interaction data, it is important to use a data source that is acquired independently to both these data types. This limitation rules out pathway annotations, which are based on protein-protein interactions and inform GO annotations. As human phenotype annotations have poor coverage on the PINs (see Section 4.5), we used gene expression data across a wide range of tissues for systematic validation. To support this validation individual communities were further investigated as case studies.

6.5.1 Gene co-expression validation

To systematically analyse the quality of modules accepted by CommWalker and functional homogeneity, we compared their levels of co-expression. Module co-expression measures the similarity of the expression profiles of genes, whose protein products are in the same module, across gene expression samples (see Section 3.3.1).

The relation between functional relatedness and gene co-expression is complex [190, 198–200] (see Section 2.6.1), especially given questions around the reproducibility of gene expression data [194–196]. Thus, to improve the reliability of our validation, we implemented three criteria:

- The gene expression data used to evaluate co-expression were chosen to give a good coverage of human tissues. To address doubts over data set reproducibility, we used a data set with large sample sizes obtained following the same experimental protocol (see Section 3.3.1).

- We do not expect all community detection methods to necessarily capture co-expression. Thus for our validation we chose the data set that appears to best capture gene co-expression as functional relatedness. This choice was made based on the assumption that communities accepted as modules by both CommWalker and functional homogeneity are our best approximation to true positive modules. Thus, the data set that has the highest level of module co-expression for communities accepted by both methods was chosen.

- We did not attempt to validate individual modules based on co-expression, but instead investigated the distribution of module co-expression scores for the accepted and rejected community sets by CommWalker and functional homogeneity.

To choose the data set that exhibits the highest level of co-expression, we compared the co-expression scores of the set of communities accepted by both CommWalker and functional homogeneity to what we would expect at random. Taking into account that interacting proteins are more likely to be co-expressed, we again defined the random background distribution for module co-expression via random walks.

Random background co-expression was sampled by performing 1000 short random walks of length six from each node in HINT-P and BioGrid-AP and computing their co-expression scores. These random walks represent random proxy communities in the lower module-relevant size range (see 6 - 35 in Section 1.1.3). To assess how similar the random walk co-expression is to the community set co-expression, we computed the fraction of the random walk co-expression scores that exceeded a threshold. This threshold was set at the 25% quantile of the community set co-expression score distribution. Using the null model that the community set co-expression scores do not differ from those of random walks, this fraction can be interpreted as a Type-I error. The Type-I error denotes the probability of rejecting the null hypothesis, given that the community set co-expression is in fact no different from random walk co-expression. The results of this investigation are shown in Table 6.3.

Based on the results shown in Table 6.3 we selected BioGrid-AP partitioned by link clustering using the Pandey measure to evaluate the resulting communities for gene co-expression validation. This selection was confirmed using a 10% quantile threshold for Type-I error calculation. The co-expression score distributions for the four community sets for the BioGrid-AP link clustering Pandey measure data set is shown in Figure 6.9.

Table 6.3: Type-I errors for data set selection for gene co-expression analysis.

PIN	Comm Det	Sem Sim	Acc	CW only	FH only	Rej
HINT-P	Link	Pandey	0.243	0.264	0.441	0.316
		simUI	0.249	0.297	0.301	0.334
		simGIC	0.247	0.316	0.349	0.331
	BigClam	Pandey	0.304	0.276	0.451	0.302
		simUI	0.267	0.315	0.293	0.324
		simGIC	0.271	0.301	0.372	0.318
	Config	Pandey	0.332	0.334	0.361	0.319
		simUI	0.298	0.331	0.360	0.335
		simGIC	0.302	0.315	0.334	0.337
	CPM	Pandey	0.312	0.332	0.403	0.315
		simUI	0.292	0.333	0.339	0.335
		simGIC	0.300	0.369	0.349	0.334
BioGrid-AP	Link	Pandey	**0.135**	0.247	0.401	0.356
		simUI	0.151	0.366	0.361	0.397
		simGIC	0.160	0.395	0.361	0.401
	BigClam	Pandey	0.312	0.303	0.444	0.331
		simUI	0.287	0.338	0.273	0.351
		simGIC	0.291	0.414	0.309	0.360
	Config	Pandey	0.374	0.386	0.548	0.403
		simUI	0.351	0.410	0.384	0.432
		simGIC	0.356	0.432	0.401	0.435
	CPM	Pandey	0.341	0.366	0.503	0.398
		simUI	0.342	0.401	0.389	0.415
		simGIC	0.347	0.474	0.391	0.418

Table 6.3: HINT-P and BioGrid-AP were partitioned using the four community detection methods: link clustering (Link), BigCLAM, configuration model Modularity Maximization (Config), and Constant Potts model Modularity Maximization (CPM). These communities were divided into four sets using qualitatively similar thresholds as described Section 6.4.1 (both accepted – Acc, only accepted by CommWalker – CW only, only accepted by functional homogeneity – FH only, and rejected by both methods – Rej). The presented fractional errors were obtained by computing the proportion of co-expression scores of length six random walks on the PINs that have a value higher than the 25% quantile of the distribution of the tested community set co-expression scores. The data show BioGrid-AP with link clustering and Pandey semantic similarity best captures gene co-expression.

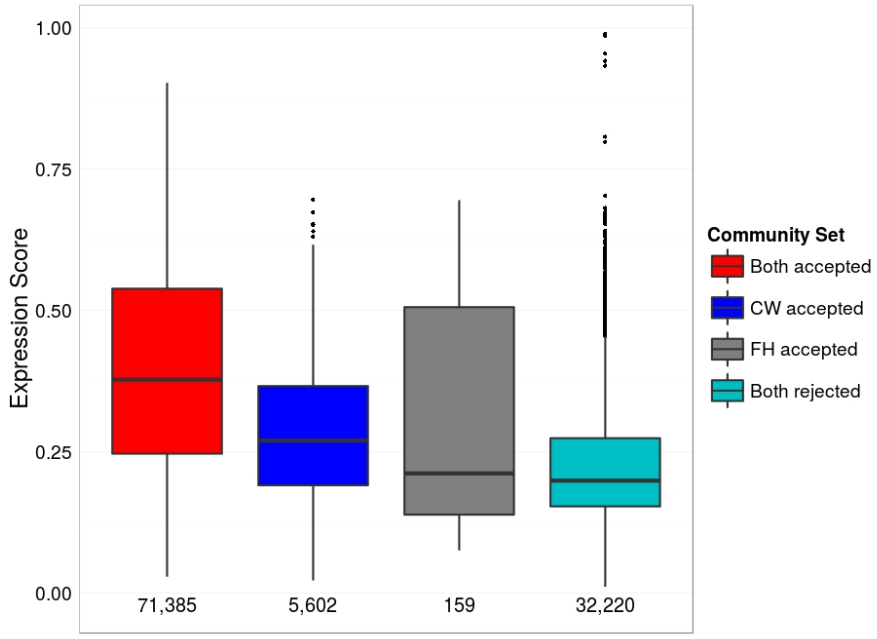

Figure 6.9: Comparison of community evaluation methods by gene co-expression. Link clustering was used to partition BioGrid-AP into communities at multiple resolutions. Using the Pandey measure, communities across all resolutions were divided into sets based on whether they were evaluated as functionally significant by both methods (red), by only CommWalker (blue), by only functional homogeneity (grey), or by neither method (turquoise). These sets are compared in a boxplot of the distributions of their community co-expression scores for communities of size 6 - 35. This size range allows us to exclude protein sets unlikely to be individual modules. The number of data points in each community set is shown under the respective boxplot. The number of communities only accepted by CommWalker is ≈ 35 times as large as the number of communities only accepted by functional homogeneity. As shown by the median values of the distributions, communities only accepted by CommWalker have higher coexpression scores than communities only accepted by functional homogeneity.

Figure 6.9 shows that modules accepted by only CommWalker in this data set exhibit a higher level of co-expression than those accepted only by functional homogeneity. Indeed, modules only accepted by functional homogeneity have a similar median co-expression level to those rejected by both methods. We confirmed that this result is not confounded by the community size distributions in the community sets by comparing the mean co-expression scores of communities of the same size between community sets (Figure D.25 in Appendix D). While Figure 6.9 does not provide conclusive evidence that all modules accepted by CommWalker are correct, it does suggest that CommWalker accepted modules are at least of a similar quality to modules accepted by commonly used functional homogeneity approaches.

6.5.2 Case studies

To complement the systematic validation of the CommWalker approach by gene co-expression analysis, we picked two modules to investigate more closely. For this purpose we defined two community sets: one strongly rejected by functional homogeneity, but clearly accepted by CommWalker, and vice versa. The largest proposed module in each set was chosen for further investigation.

Module candidates clearly favoured by CommWalker while rejected by functional homogeneity were defined by the arbitrary, but more stringent thresholds of a T-value < 0.25 and a functional homogeneity, FH < 5 (in contrast to the median semantic similarity of interacting proteins at 6.10552 for BioGrid-AP). The largest predicted module in this set contains the TRAPP proteins TRAPPC2, TRAPPC3L, TRAPPC4, TRAPPC6B, TRAPPC8, TRAPPC10, and TRAPPC12. These proteins are known to be members of the TRAPP complexes which are implicated in vesicular transport from the endoplasmic reticulum to the Golgi apparatus [215]. Despite its coherent functional description, it is only relatively poorly annotated at an average of 43.71 functional annotations per protein, compared to a mean of 89.85 functional annotions for proteins in BioGrid-AP (with a standard error of 33.91).

Similarly predicted modules by functional homogeneity that are strongly rejected by CommWalker were defined by a functional homogeneity FH > 6.5, and a T-value above 0.6. The largest community in this set is a star community centered around STAT3, further containing the genes LEP, CEP120, NFKBIZ, HES5, and IL22RA1. Most of these genes were found to be involved with signalling and/or regulating transcription, which explains their interaction with STAT3. However, we found no closer connection between the low degree proteins clustered together. The average number of GO BP annotations associated with the genes in this module is 257.17, which is significantly larger than the mean of 89.85 in BioGrid-AP. While this investigation cannot show that modules only accepted by functional homogeneity should not be considered, it does allow us to highlight the TRAPP module accepted by CommWalker. The contrasted community case studies suggest

that CommWalker can improve module evaluation both for poorly annotated, and well annotated communities.

6.6 Discussion and conclusions

As a complete picture of protein function with corresponding functional annotation is not available, it is important to consider the effect of the distribution of annotations on the network in module evaluation. In this chapter we have presented CommWalker, a module evaluation framework that takes this heterogeneity of annotation into account through local background sampling. To evaluate a module, CommWalker samples its local environment using random walks to put the community's functional homogeneity score into the correct context. CommWalker has two main advantages over currently available methods:

1. Functional evaluation of a community is focused on alternative choices of protein groupings that could have been made in the local network environment. Through this local context we can account for annotation, and therefore functional similarity, varying across the network.

2. The size of the community is taken into consideration when relating its functional homogeneity score to the background distribution. By forcing the random walk proxy communities to contain the same number of nodes as the initial community, the effect of the size-dependence of functional homogeneity scores observed in Chapter 5 is eliminated.

A third advantage of CommWalker lies in its reporting of significance scores. Functional homogeneity is generally quoted as a value calculated based on the semantic similarity measure used to obtain pairwise functional similarity scores. Here, we have used quantiles to make this score informative. However, a community T-value is immediately informative. A T-value of 0.67 tells us that approximately 67% of node groupings of the same size were evaluated as more functionally

homogeneous in the local network region. This advantage is especially valuable in stand-alone tools that evaluate individual communities.

Using three popular functional similarity measures on a variety of biological data sets, we have shown that CommWalker indeed allows for an increased sensitivity in poorly-studied network regions, without sacrificing its ability to evaluate well-studied communities. The greater coverage of functional modules on PINs as evaluated by CommWalker is specifically important in the context of this project. As we are attempting to develop a general pipeline to identify modules that are involved in poorly understood phenotypes, it is likely that it is exactly in poorly-studied network regions that we have to look.

While CommWalker will improve our ability to evaluate sparsely annotated protein communities, it is still important that some annotation exists. CommWalker can overcome lack of annotation to a certain extent, however it is only able to amplify an existing signal. As such, continuous improvement of the coverage of functional annotations is still key to understanding PIN functional organization. A further limitation to functional module discovery in poorly-studied regions of PINs are error rates. It is likely that protein-protein interaction false-positive error rates are not uniformly distributed across the network, but instead adversely affect poorly-studied proteins (eg. well-studied proteins found by multiple reporting are assumed to be true positive interactions [67,79]). These error rates affect community detection methods which are therefore more likely to fail to group proteins together meaningfully in these regions. For these reasons CommWalker should not be seen as a solution to lack of knowledge in certain areas of PINs, but instead an incremental development that allows us to look further than was previously possible. Overall CommWalker is a framework that builds on the work done on protein similarity evaluation and bridges the gap between protein similarities and module homogeneity.

As CommWalker can be used with any similarity measure defined between proteins, it is also capable of overcoming the issues of performing functional enrichment on network communities discussed in Section 4.3. Similar to the functional similarity approach, the frequency of a GO-term can be related to the

frequencies of this GO-term in random walks from the community to evaluate its significance. In this case the random walk background distribution acts as a null model for enrichment that takes into account the dependence structure between the distribution of functional annotations and protein-protein interactions.

A possible extension to CommWalker is to rein in the random walks in the background sampling process. The larger the community, the further into the network the random walks may sample – past what could be considered the local environment. This deep sampling may lead to a large community being related to the entire network. While local nodes are still sampled more frequently than distant nodes, this effect reduces the efficacy of the CommWalker approach. Thus, CommWalker may benefit from an augmented random walk process such as "random walk with restart" [216] which could ensure local sampling. Such a sampling method could be calibrated to sample only as far into the network as the average diameter of communities of this size is. As we are focusing on comparatively small communities that could be considered as functional modules (sizes 6 - 35, see Section 1.1.3), very long walks are unlikely to be an issue for the purpose of this project.

7 Biological applications

Contents

7.1 Introduction 167
7.2 Methodology 169
7.2.1 Functional module detection 169
7.2.2 Overlaying DEGs 171
7.2.3 Consensus clustering of enriched modules 172
7.2.4 Module prioritization 173
7.3 Enriched modules in biological applications 174
7.3.1 Visualization 174
7.3.2 Hypoxia 175
7.3.3 Macrophage differentiation 179
7.4 Retrospective implications for module detection 184
7.5 Discussion and conclusions 186

7.1 Introduction

Whole genome expression profiling is a powerful technique to uncover the molecular basis of disease phenotypes (see Section 2.6.2). Statistical tests are applied to such data sets to find genes that are significantly differentially expressed between two phenotypes of interest such as “disease” and “healthy”. Differentially expressed genes (DEGs) are thought to play causal roles in, or be symptomatic of, the investigated disease processes. Typically long lists of DEGs are obtained from these statistical

tests. Thus it has been suggested that the real challenge lies in the interpretation of DEGs rather than in their discovery [17].

Currently available approaches to interpret these lists of DEGs are twofold. Either the biological intuition of the experimentalist is used to focus on specific functions that are expected to be involved in the phenotype, or curated lists of genes involved in biological functions are investigated. Using biological intuition requires in-depth knowledge of the investigated phenotype that allows for subjective prioritization of certain DEGs. Curated lists of genes involved in specific functions as available from the Molecular Signature Database [52] can be used to assess differential expression of gene functions in a more objective, systematic way via tools such as Gene Set Enrichment Analysis (GSEA, see Section 1.1.3). However, even such curated lists require functions to be well-studied, which is not always the case for poorly understood disease phenotypes. Currently available methods that integrate PINs into this analysis for example by finding disease modules [13, 44], also rely on these curated gene lists to elucidate functional information (eg. by enrichment of pathway annotations).

We have developed a pipeline to objectively detect differential regulation of cellular functions in investigated phenotypes which is not limited to curated lists of proteins or genes. In this pipeline, DEGs are put into the context in which they affect the respective phenotype: through the interaction of their protein products. We map DEGs to cellular functions using functional modules detected in protein interaction networks (PINs) by the approach refined in Chapters 4–6. This approach integrates biological clustering using the CommWalker framework (Chapter 6) in conjunction with the Pandey measure on GO BP annotation sets (Chapter 4) with topological clustering at multiple resolutions (reviewed in Chapter 5). Functional modules that are enriched for DEGs represent differentially regulated cellular functions. This methodology can be seen as an extension of GSEA that is able to systemically propose the lists of functionally related proteins whose differential expression is assessed (see Section 1.1.3).

In this chapter we describe the developed pipeline in its current state and present results of its application to two biological problems: breast cancer hypoxia and macrophage differentiation. With this pipeline, differentially regulated functions that each correspond to an experimentally testable hypothesis can be detected. Using the functional modules found to be differentially regulated, we were able to retrospectively analyse the optimized module detection approach from Chapters 4–6. This analysis showed the importance of a multi-resolution approach to module detection and the benefits of using the CommWalker framework in this pipeline.

7.2 Methodology

In this section we describe our pipeline for the interpretation of differential gene expression data. This pipeline comprises of four steps:

1. The detection of functional modules,
2. Finding differentially regulated modules by overlaying DEGs,
3. Consensus clustering of highly similar enriched modules
4. Prioritizing those enriched modules most likely to be of interest.

By performing the first three steps our pipeline can objectively generate differentially regulated functional modules using only differential gene expression data as user input. Each of the generated DEG-enriched modules represent a biological hypothesis of a function, and candidate proteins involved in this function, that is causative or symptomatic of the investigated phenotype.

7.2.1 Functional module detection

In Chapter 5 we attempted to find an optimal community detection method and resolution to detect functional modules in PINs. Finding none, we concluded that all investigated community detection methods may detect functional modules across resolutions. Using the CommWalker module evaluation framework with the Pandey

measure (see Chapter 6 for CommWalker and Section 2.5.2 for the Pandey measure) we re-evaluated this conclusion by investigating the fraction of proteins in BioGrid-AP that were assigned to functionally significant communities based on a more stringent T-value threshold of 0.4, compared to 0.5 used in Chapter 6 (Figure 7.1).

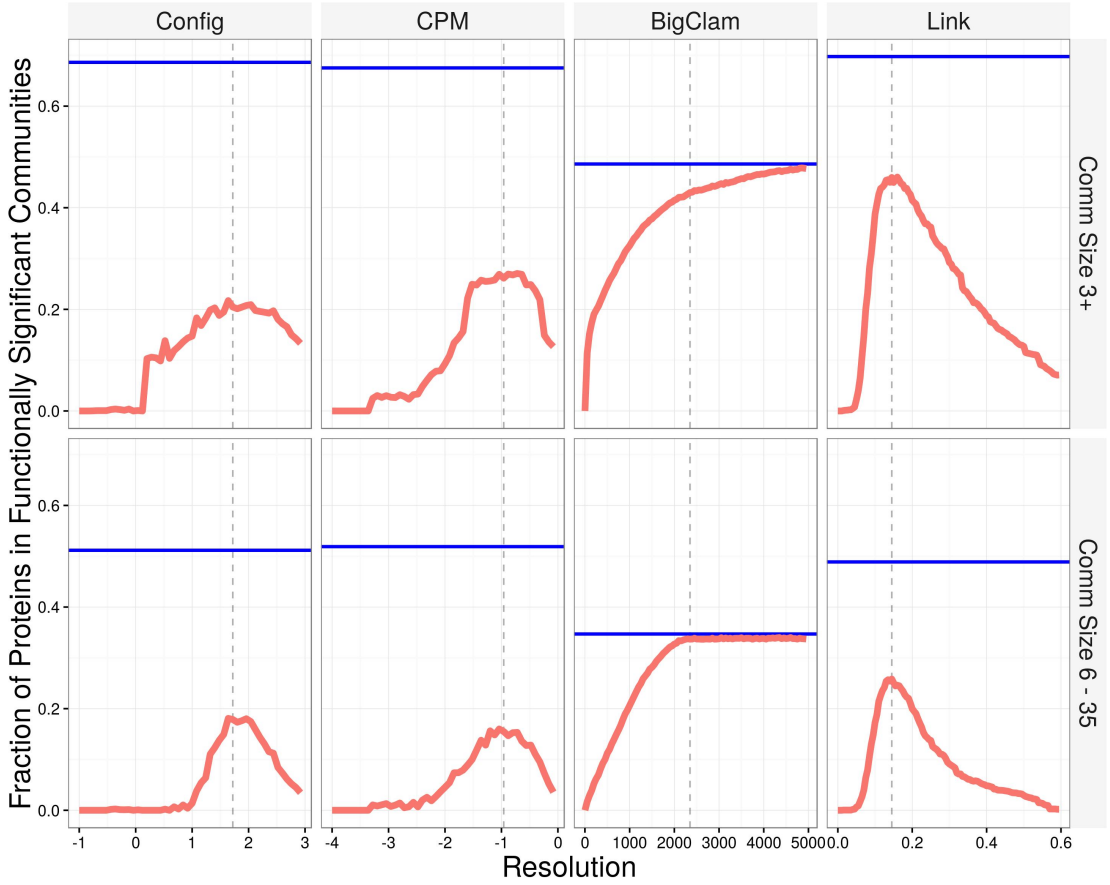

Figure 7.1: Fraction of BioGrid-AP proteins in functionally significant communities across resolutions. Functional significance was evaluated using CommWalker with the Pandey measure (see Chapter 6 for CommWalker and Section 2.5.2 for the Pandey measure) with a T-value threshold of 0.4. The fraction of proteins in BioGrid-AP that are assigned to functionally significant communities by the four investigated community detection methods (see Section 3.1.2) at each resolution are shown in red. The blue lines indicate the fraction of proteins that are assigned to functionally significant communities at any resolution for each community detection method and community size range. The vertical dashed lines show the resolution at which the network is partitioned such that the most proteins are in functionally significant communities for different T-value thresholds (see Section D.2 in the Appendix).

Figure 7.1 shows that while $\approx$ 68% of proteins are assigned to non-trivial functionally significant communities at some resolution, no single resolution network partition reaches this value. In our designated module size range of interest (6 - 35 proteins, see Section 1.1.3) this effect is further exacerbated. Based on these results

we looked for functional modules across all resolutions using the four community detection methods on BioGrid-AP (as described in Section 3.1.2). Only BioGrid-AP was used here as previous analysis suggested that having a high coverage of protein-protein interactions, was more important to module detection than high quality interaction data as provided by HINT-P (see Section 5.6).

After exploratory analysis on how to capture the functional homogeneity of a module of proteins (see Chapter 4) we used CommWalker with the Pandey measure using GO BP functional annotations (see Section 3.2.1) to perform community evaluation.

Previous work on multi-scale functional module detection in yeast used a functional homogeneity threshold to determine which communities represent functional modules [48]. This threshold was set at:

$$\mathrm{FH} > \mu_{int} + 0.3\sigma, \tag{7.1}$$

where FH denotes the functional homogeneity, μ_{int} is the average functional similarity of interacting proteins in the underlying PIN, and σ represents the standard deviation of this distribution. In analogy with Equation (7.1) we chose a T-value of 0.4 as a threshold for functional module detection. The reasoning for this choice is as follows. If the functional similarity scores of interacting proteins are approximated by a normal distribution, $\approx 38.2\%$ of protein pairs have a functional similarity that satisfies the threshold in Equation (7.1). Similarly, if these scores are approximated by a uniform distribution (which may be a better approximation given Figure 6.4), this percentage rises to $\approx$ 41.3%. These quantiles were used to qualitatively approximate functional homogeneity and T-value thresholds as in Section 6.4.1.

7.2.2 Overlaying DEGs

DEGs were obtained as described in Section 3.3.2. BioGrid-AP reports proteins by their gene identifiers. This reporting simplifies the gene to protein relationship to one-to-one. Therefore DEGs could easily be mapped onto the PIN (Figure 7.2).

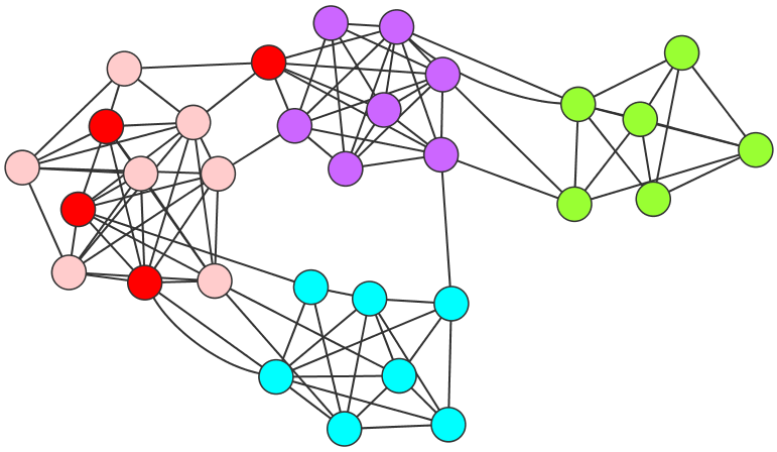

Figure 7.2: Schematic diagram of overlaying DEGs onto functional modules. Red nodes represent DEGs and nodes of other colours denote membership of different functional modules. The light red module on the left appears enriched for DEGs.

DEG enrichment was calculated after Equation (2.12). Communities detected at neighbouring resolutions may be similar or exactly the same. As these communities would be merged in the consensus clustering step (see Section 7.2.3), the dependence structure between the enrichment tests is difficult to determine. Thus, at this point of development we have chosen to omit multiple testing correction to keep a high sensitivity. DEG enrichment p-values were used to rank communities by enrichment and an arbitrary cut-off (p-value < 0.01) was used on the unclustered modules to select modules for further analysis. While we cannot assess the significance level at which these modules are enriched without multiple testing correction, ranking the "enriched" modules by uncorrected p-value does show which modules are more likely to be significantly enriched at a particular threshold. Furthermore, treating the enrichment p-values only as scores is reasonable given that there may be errors in the DEG assignment.

7.2.3 Consensus clustering of enriched modules

Network partitions generated at neighbouring resolutions can be highly similar. As modules were detected across resolutions, enriched modules can therefore also be similar or identical. Due to error rates in PINs (see Section 2.1.2.1) we cannot expect the node-community assignment to be completely free of error. Thus, we interpret similar modules as representative of the same underlying module with

node-community assignment errors. Given this consideration, it makes sense to merge similar communities by consensus clustering. This merging reduces the redundancy in our data set and speeds up the pipeline.

Consensus clustering was performed using the Jaccard index over the proteins in a module (see also Equations (2.10) or (2.14) for Jaccard index measures using different sets). Modules were regarded as similar when the Jaccard index of their member protein sets exceeded a value of 0.7. For two modules of six proteins, this is the case when five proteins are shared, or similarly 9 of 10, or 38 of 45. DEG-enriched modules that are similar by this threshold were merged and the highest enrichment p-value of the merged communities was recorded to represent the cluster. This decision ensured that strong signals for differential regulation are not overlooked due to the consensus clustering step when prioritizing module clusters (see Section 7.2.4).

7.2.4 Module prioritization

When large numbers of DEGs are found in differential gene expression studies (see macrophage differentiation data set in Table 3.5), it is likely that a considerable number of differentially regulated modules are found using this pipeline. To allow users of the pipeline to easily screen for interesting results, we implemented three methods to prioiritze differentially regulated modules: based on genes of interest, based on extremal fold change, and based on consensus fold change direction (see fold change in Section 2.6).

The first of these methods prioritizes DEG-enriched modules by their proximity to a particular protein of interest. It is not uncommon for differential gene expression studies to be performed with a particular biological hypothesis in mind. In cases where this hypothesis involves a particular gene, connections between this gene and enriched modules that are close to this gene in terms of network proximity (see shortest path length in Section 2.2.1) may show how this particular gene acts to cause differential regulation of a functional module.

The other two methods for module prioritization involve computing the average fold change of all genes or all DEGs in the module, and using the fraction of

genes with a positive fold change. The latter prioritizes modules where all genes are either up- or down-regulated.

7.3 Enriched modules in biological applications

In this section we present potentially differentially regulated functions and candidate proteins involved in the respective biological processes for two biological applications: breast cancer hypoxia and macrophage differentiation. In order to evaluate our pipeline we focus on well-studied functions due to the difficulty of validating novel biological hypotheses. Furthermore, we present modules with a strong enrichment signal to minimize the possibility of insignificant enrichment of DEGs (see Section 7.2.2). First we present our results for breast cancer hypoxia, which was chosen as a well-studied case and thus offers many known differentially regulated functions against which our pipeline can be tested. In comparison macrophage differentiation is less well-understood.

For practical reasons we increased the investigated module size range from our designated size range of 6 - 35 (see Section 1.1.3) to 10 - 45 and 15 - 45 for the hypoxia and macrophage data sets respectively. This decision was made as inspection of communities by experimental collaborators was difficult for smaller communities of sizes 6 and 7, which outnumbered modules of other sizes in preliminary runs. It appears that recognizing individual proteins, or understanding the modular context, helped in understanding community function. This increase in size range further considers that node-module assignments are also subject to error (see Section 7.2.3).

7.3.1 Visualization

Modules were visualized as a set of circles, with gene name labels with edges of the same colour between them (here blue). These gene names were mapped from Entrez Gene IDs (which identify proteins in BioGrid-AP, see Section 2.1.3), by a mapping from the HUGO Gene Nomenclature Committee (`http://www.genenames.org/cgi-bin/download` [217], retrieved April 2016). Grey edges were used to denote paths between the enriched module and a protein or gene of interest input into

the pipeline (which is shown as a red node). DEGs were highlighted by bold black contours around the node, and proteins with over 150 interactions (designated as hub proteins) are shown as squares. The displayed module was complemented with information on the average fold change of all module genes, and the fraction of fold changes that were positive. These statistics were also calculated for only the DEGs and shown in brackets. Furthermore, we performed module annotation by GO BP term enrichment (see Section 2.5.1) to label the module function. The bottom right corner of the display contains this additional information (see Figure 7.3 and Figure 7.7 with a protein of interest and grey edges).

7.3.2 Hypoxia

The breast cancer hypoxia data set was analysed using DEGs found at a FDR threshold of 0.05. The small number of replicates available in the data set did not permit a more stringent cutoff as the number of DEGs was already relatively small. In total this investigation produced 29 enriched modules with several differentially regulated functions. Modules representing known differentially regulated functions such as RNA-mediated regulation of translation [218], or chromatin silencing (personal communication with Dr. Francesca Buffa) were found. In Figures 7.3–7.6 we show four examples of enriched modules that represent differentially regulated functions.

The first of these modules is related to vesicle transport of matrix metalloproteinases (MMPs). The hypoxic tumour phenotype is strongly linked with aggressiveness and metastasis (see Section 1.2.1). A way in which these tumours invade neighbouring tissue is by the excretion of MMPs that attack the extracellular matrix [55]. Thus, vesicle transport of MMPs out of the cell is a function that is upregulated in hypoxia. Several vesicle transport modules were found among our enriched modules, such as the module shown in Figure 7.3. While we detected this module based on the differentially expressed genes SNAP23, SNAP29, STX7, and STX12, other genes in this module have previously been linked with tumour invasiveness, such as VAMP3, STX4, and other syntaxins (STXs)

and VAMPs [219]. We also found a further vesicle transport module containing the RAB gene family of which RAB25 and RAB11 have previously been linked with tumour invasiveness [220].

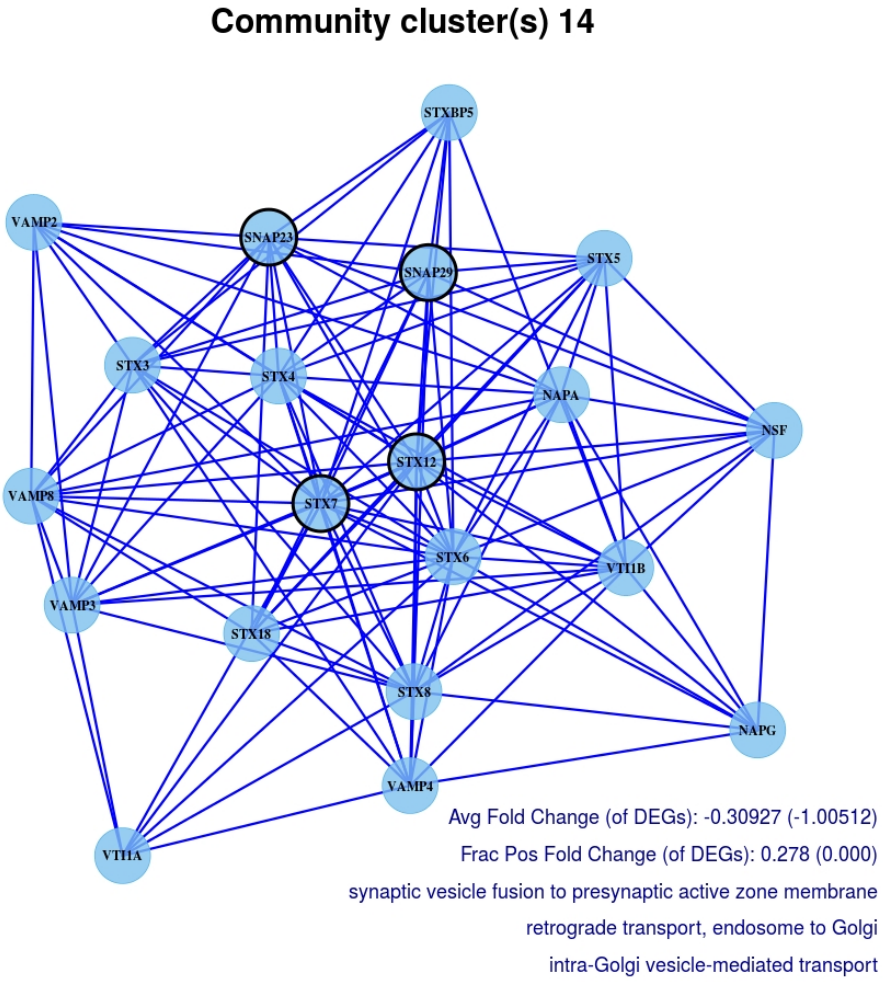

Figure 7.3: Matrix metalloproteinase vesicle transport module. A DEG-enriched module found by our pipeline using the breast cancer data set (see Section 3.3.2.1). The module is displayed as explained in Section 7.3.1. The module contains mainly genes, shown here by their HGNC gene names (see Section 7.3.1), that are upregulated in the hypoxic samples as shown by negative fold change displayed in the bottom right (Frac Pos Fold Change = Fraction of genes with a positive fold change). The last three lines on the bottom right are GO BP terms that are enriched in the genes in this module. These annotations indicate the vesicle transport function.

Figure 7.4 is a module of RNA polymerase I promoter genes. This module represents differential regulation of transcription in hypoxic versus normoxic cells. Other modules showing differential helicase activity, and transcriptional modulation by SMAD proteins were also found with the same overarching function. Indeed, the hypoxic microenvironment causes up-regulation in transcription as induced by the

HIF transcription factor (see Section 3.3.2.1). While there is no particular signal for gene up-regulation in the module shown in Figure 7.4, this is not necessarily expected as down-regulation of transcription inhibitors may similarly signal overall up-regulation of transcription. The differentially regulated initiation of transcription in hypoxic tumours is currently being investigated (personal communication with Dr. Francesca Buffa).

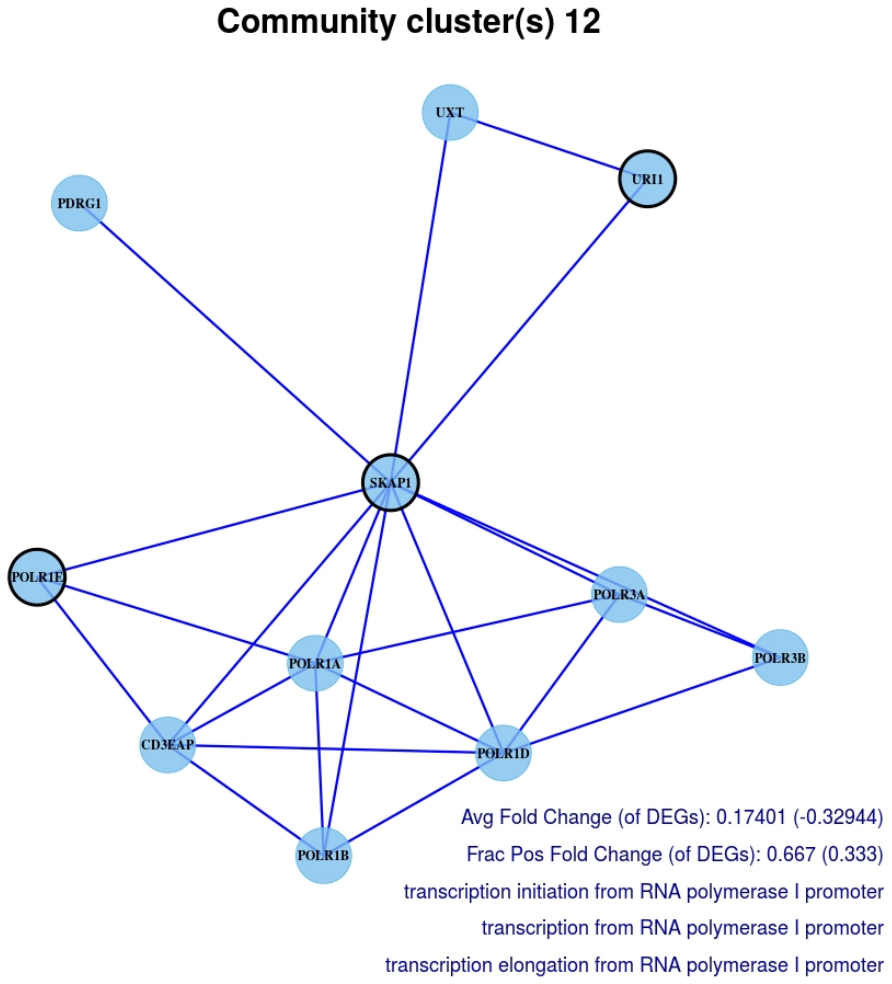

Figure 7.4: RNA polymerase I promoter module. This module is displayed as described in Section 7.3.1. The proteins in the module are subunits of RNA polymerase (POLR genes) and promoters of the complex. RNA polymerase I promoter acts to initiate transcription. This module shows no particular signal for up- or down-regulation.

A cellular response to hypoxia is the switch from aerobic to anaerobic respiration, which leads to the acidification of the extracellular environment (see Section 1.2.1). As hypoxic cells must be able to survive in these conditions, they must regulate their

intracellular pH. Given this consideration we can understand the enriched module of ATPases (Figure 7.5). ATPases act to deacidify the intracellular environment by transporting hydrogen ions across the cell membrane.

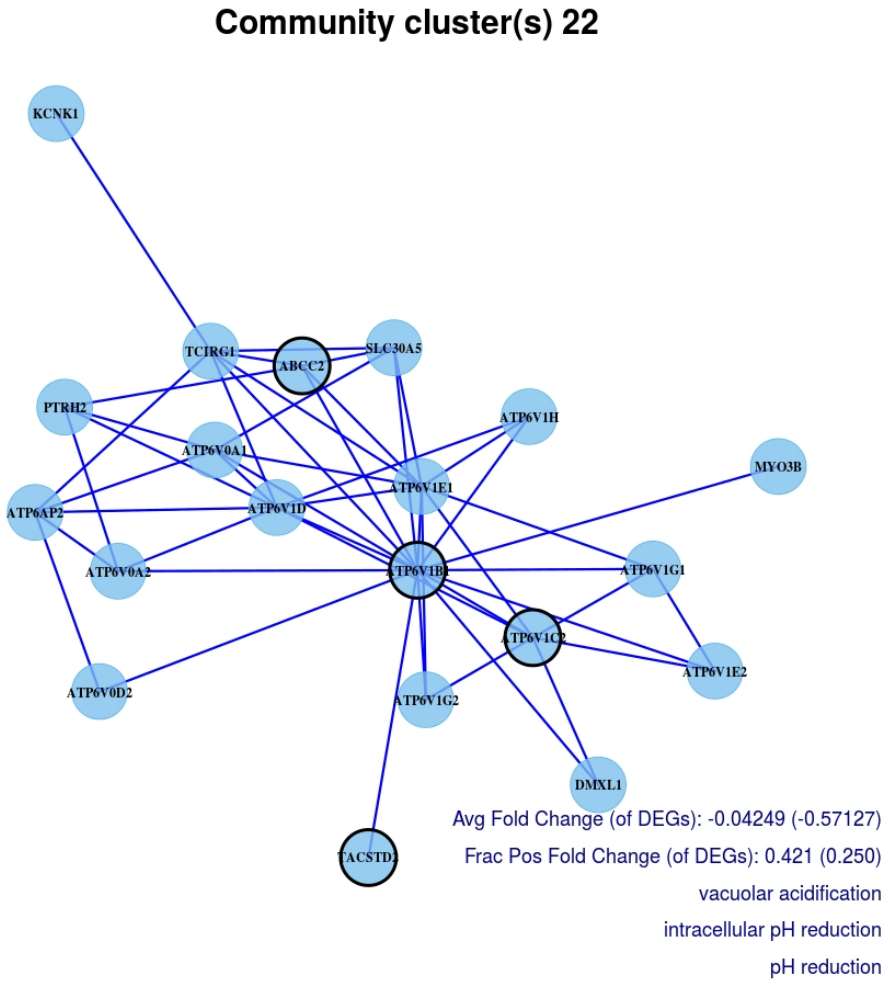

Figure 7.5: Deacidification module. This module is displayed as described in Section 7.3.1. The module contains many ATP6 genes, which are ATPases. ATPases transport H+ ions across the cell membrane to deacidify the intracellular environment. While the DEGs are up-regulated in hypoxia, the overall module shows no particular preference for up- or down-regulation.

A further differentially regulated function detected in our enriched modules is autophagy (Figure 7.6). Autophagy is a process by which cells dispose of damaged proteins and organelles thereby promoting their survival under microenvironmental stresses such as hypoxia [55, 56]. The relevance of this particular module is further shown by the module membership of ULK1 despite it not being a DEG. ULK1 is involved in hypoxia-induced autophagy [221] and has even been suggested as

a drug target for cancer therapy. This discovery suggests that our pipeline can infer important disease-related proteins that would otherwise not be found by differential expression analysis alone.

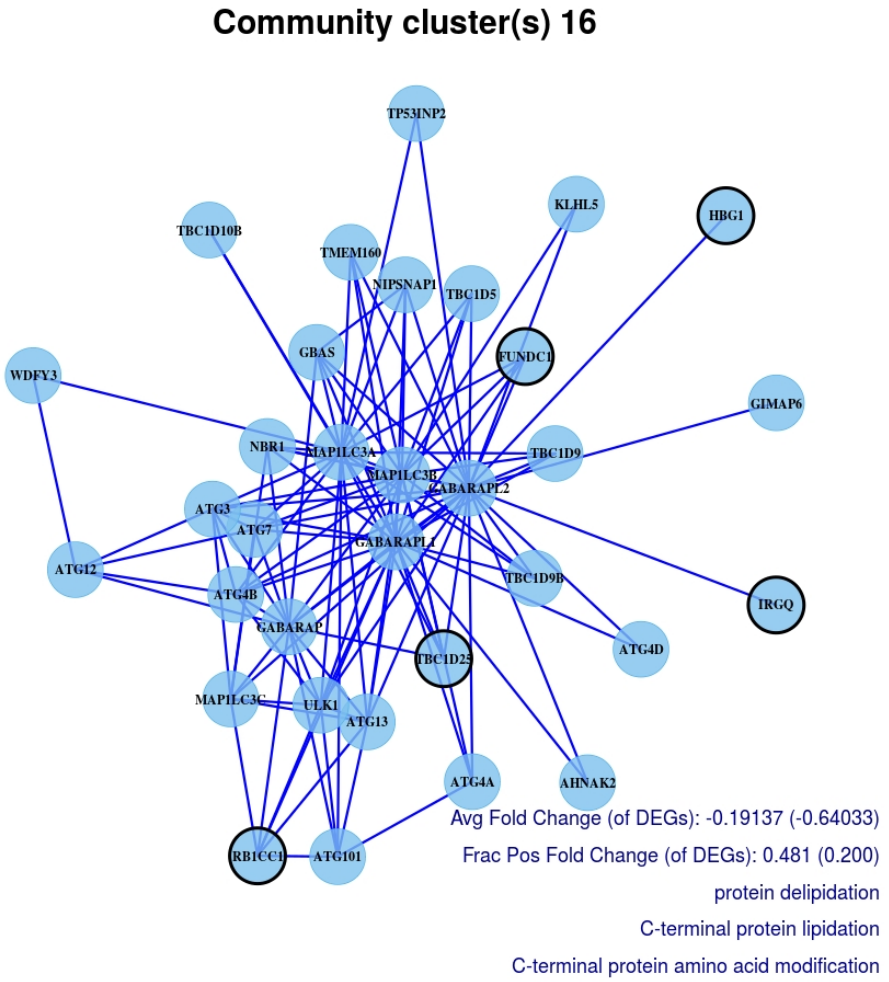

Figure 7.6: Autophagosome module. This module is displayed as described in Section 7.3.1. The module shows a slight up-regulation in hypoxic cells. While the enriched GO BP terms do not directly mention autophagy, the large proportion of Autophagy Related (ATG) and GABARAP genes shows this link. Autophagy is a process by which cells dispose of damaged organelles or proteins [55, 56]. As organelle membranes contain lipids, delipidation occurs when organelle membranes are destroyed.

7.3.3 Macrophage differentiation

In this project, the target biological application for our developed methodology was macrophage differentiation into M1 and M2 phenotypes (see Section 1.2.2). While the differences between macrophage polarizations have been studied, knowledge

of the molecular mechanisms that are differentially active is still limited. To test our pipeline we thus focussed on a few modules with functions that have been previously studied, as detailed below.

Enriched modules for the macrophage differentiation data set were generated using the list of DEGs found at a FDR threshold of 0.01. The comparatively large number of DEGs found in the macrophage data set allowed us to use this more stringent FDR threshold. Two proteins of interest (MSR1 and MRC1) were input into the pipeline to align with the research interest of collaborators at UCB Pharma.

The pipeline found 86 modules which represent differentially regulated functions between M1 and M2 macrophage phenotypes. A considerable number of these modules related to functions which are differentially active due to the macrophage treatment that stimulated differentiation into M1 and M2 phenotypes *in vitro* as described by the enriched GO BP terms "Regulation of response to IFN-γ", "Interferon mediated signalling pathway", and "Response to stimulus". While these functions are correctly detected, they are not interesting with regards to the macrophage phenotype but rather indicative of the experimental protocol. A further common function found in enriched modules is JAK/STAT signalling. Up-regulation of JAK signalling has been linked to inflammation [222], a central function attributed to the M1 phenotype. Yet, this process is also generally linked to interferon and interleukin signalling and may thus be a response to the macrophage treatment. Here we show three enriched modules that were found to be of interest to both collaborators at UCB Pharma and Dr. Fernando Martinez (the macrophage researcher who published the initial data set).

NF-$\kappa\beta$ transcription factors are important in inflammation and innate immunity. These transcription factors are secreted by tumour associated macrophages, which resemble an M2 activation state (see Section 1.2.2), to control the inflammatory response. The enriched module in Figure 7.7 describes an up-regulation of the NF-$\kappa\beta$ inhibitor A20 (shown here as TNFAIP3 and interacting proteins), and is thus up-regulated in M1 macrophages which are linked with increased inflammation [62] (M1 up-regulation is shown by positive fold change). Suppression of NF-$\kappa\beta$ via the

A20 complex is thought to occur via regulation of linear ubiquitination. Linear ubiquitination is a recently discovered regulatory signalling mechanism (specifically a post-translational modification) that is important in immune signalling [223, 224]. This signalling process has been linked with activating NF-$\kappa\beta$ signalling via TNF [224] and may thus explain further members of the detected module.

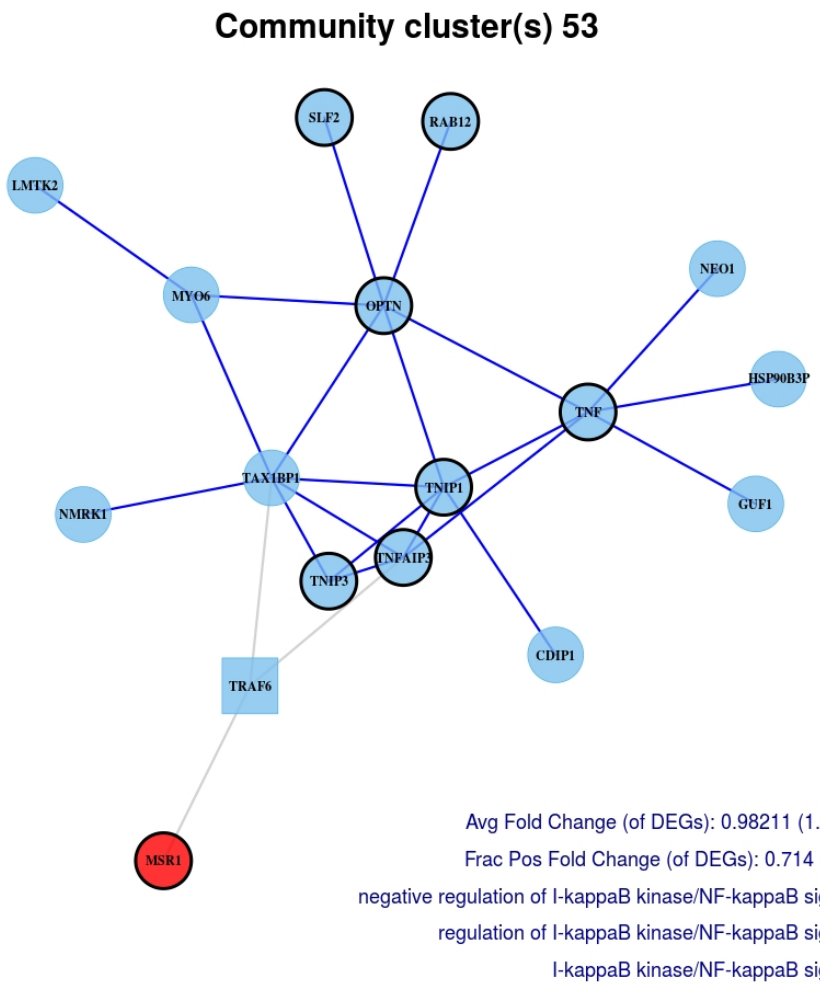

Figure 7.7: NF-$\kappa\beta$ signalling regulation module. This module is displayed as described in Section 7.3.1. The circle in red denotes the protein of interest MSR1 (macrophage scavenger receptor 1), which was input into the pipeline. Comparatively higher expression levels in M1 macrophages are denoted by a positive fold change. Negative regulation of NF-$\kappa\beta$ signalling is captured via the A20 complex shown here as TNFAIP3 with its interacting proteins TNIP1 and TNIP3. Up-regulation of the inhibition of NF-$\kappa\beta$ signalling in this module implies the signalling process itself is down-regulated in M1 macrophages.

Figure 7.8 shows an enriched module of immunoproteasome subunits. The immunoproteasome is a complex that creates small peptides for antigen presentation eliciting an immune response. Antigen presentation, which was described in several

enriched modules, has recently been found to be differentially regulated in alveolar M1 and M2 macrophages [63].

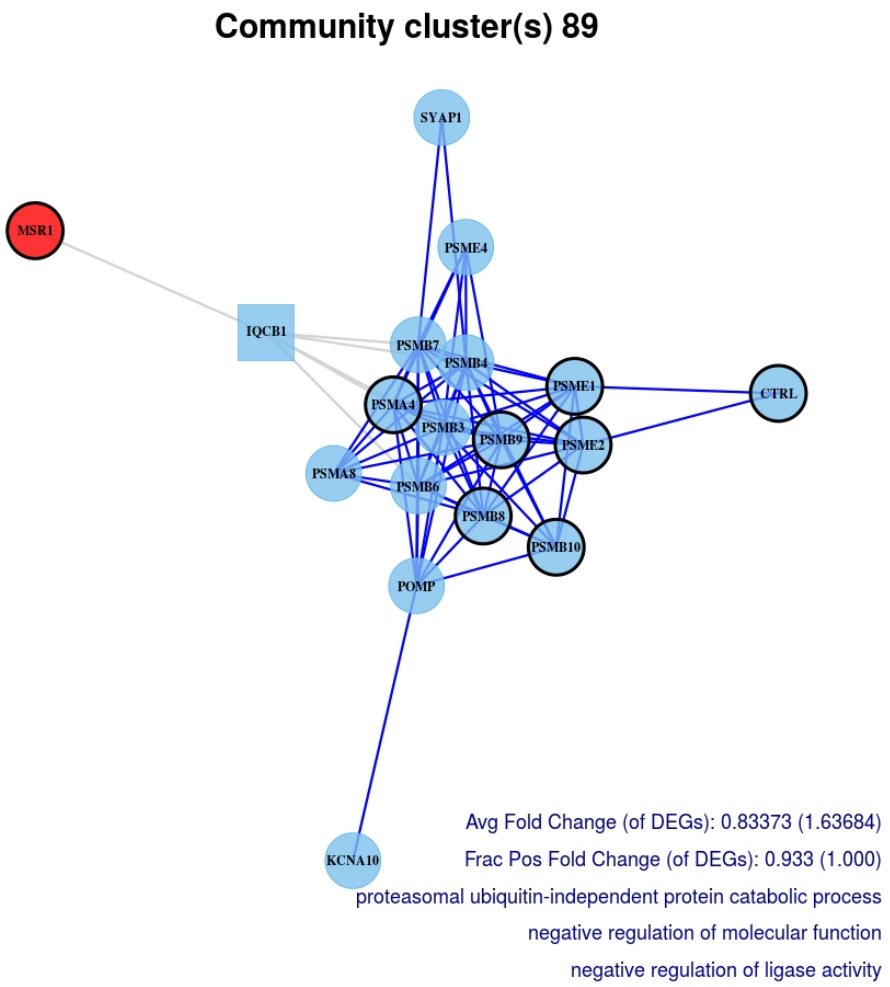

Figure 7.8: Immunoproteasome module. This module is displayed as described in Section 7.3.1 with the protein of interest (MSR1) shown as a red circle. The module shows immunoproteasome subunits (PSMs) that are up-regulated in M1 macrophages. The immunoproteasome cuts up antigenic proteins into small peptides (see protein catabolism in the enriched GO BP terms) to be presented on the cell surface. This antigen presentation starts an immune response.

A further function that was differentially regulated in our enriched modules is chemotaxis (Figure 7.9). Chemokines are signalling proteins that recruit immune cells in homeostasis and inflammation [59], two biological processes that have been linked with macrophage activity (see Section 1.2.2). A screen of 41 chemokines found several that are differentially linked to M1 or M2 chemotaxis [64]. Specifically, the chemokines CCL21 and CCL25 were reported to induce M1 chemotaxis. While

we did not find the genes expressing these chemokines to be differentially expressed, they are included in the enriched chemotaxis module which is up-regulated in M1 macrophages (Figure 7.9). As macrophage chemotaxis can be measured using xCelligence RTCA systems, this module can be experimentally validated without the need for gene manipulation (personal communication with Dr. Fernando Martinez).

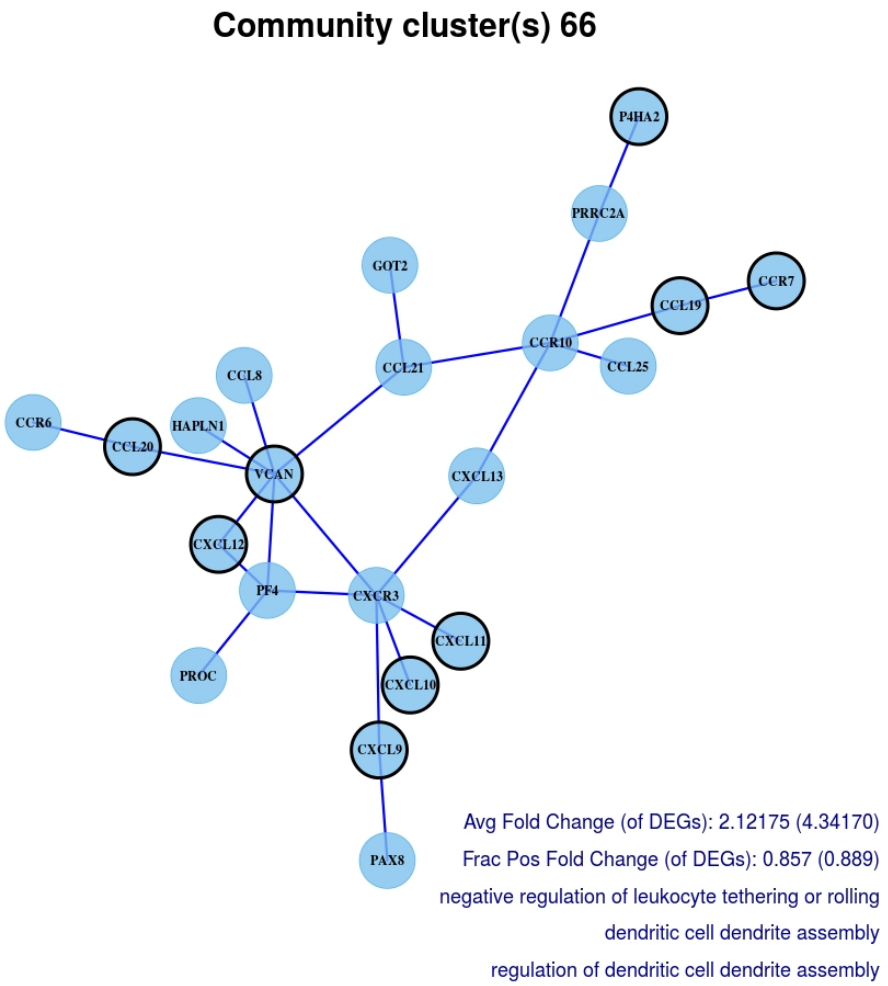

Figure 7.9: Chemotaxis module. The module is displayed as described in Section 7.3.1. The module contains chemokines (CCLs and CXCLs) and chemokine receptors (CCRs and CXCRs) that are up-regulated in M1 macrophages. The proteins in the module are up-regulated in M1 macrophages as indicated by the positive fold change.

7.4 Retrospective implications for module detection

If DEGs are interpreted as being representative of differentially regulated functions, then DEGs that cluster together should be able to define functional modules. This idea is used to detect functional modules in methods that look for active subnetworks (see Section 1.1.2). For our pipeline DEG enrichment can therefore be seen as a type of module validation. By design, this validation is biased toward functions that are differentially regulated in the investigated phenotypes and thus cannot be used to assess modules that were not found to be enriched for DEGs. Yet, enrichment for DEGs represents a further source of evidence that the module is a functionally coherent cluster of proteins. Based on these considerations we retrospectively analysed our module detection pipeline by investigating some characteristics of modules that were found to be enriched for DEGs.

The resolutions at which differentially regulated module communities in the hypoxia and macrophage data sets were found are shown in Figure 7.10. The distributions suggest that multi-resolution module detection is an important component of functional module detection. It is particularly visible in the CPM Louvain and the link clustering plots that not all differentially regulated modules would have been detected in a single resolution network partition. For example, the two communities that were combined for the deacidification module (Figure 7.5) were detected by CPM Louvain at a log resolution of -1.2 and -0.96. In contrast, the MMP vesicle transport module (Figure 7.3) is a consensus module of two CPM Louvain communities detected at log resolutions -0.48 and -0.4.

We further investigated whether evaluating communities via the CommWalker framework (see Chapter 6), rather than via the conventional functional homogeneity approach, affected the DEG-enriched modules found. Using the functional homogeneity threshold given by Equation (7.1) and a T-value threshold of 0.4 (see Section 7.2.1), we calculated the fraction of enriched module communities that are not evaluated as functionally significant using only functional homogeneity. For the hypoxia data set 34.66% and for the macrophage data set 17.50% of

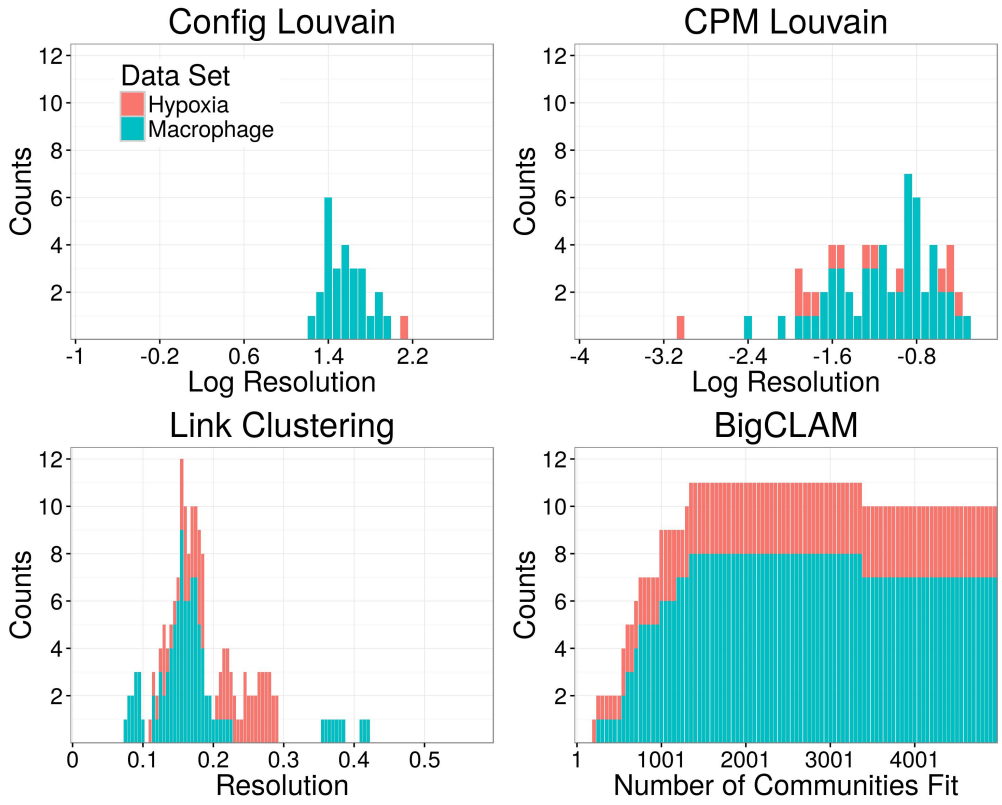

Figure 7.10: Resolutions at which differentially regulated modules in hypoxia and macrophage studies and were detected. Differential regulation was determined by enrichment of DEGs at a p-value < 0.01 (see Sections 3.3.2.1 and 3.3.2.2 for DEGs). Differentially regulated modules are consensus community clusters detected by several community detection methods (see Section 7.2). The distribution of the resolutions at which the communities were detected that contributed to these consensus community clusters are plotted for each community detection method. As the number of communities fit is used as a resolution proxy for BigCLAM (see Section 3.1.2) the set of detected communities at neighbouring resolutions tends to differ only in the 50 newly fitted communities, creating the observed increasing distribution. Relevant communities were detected at a range of resolutions by each community detection method.

communities are only evaluated as functionally significant by CommWalker. Using the mean functional similarity of interacting proteins as a more lenient functional similarity threshold, we still find 27.56% and 2.31% of communities only accepted by CommWalker in the hypoxia and macrophage data sets respectively.

The fraction of enriched module communities that were only evaluated as functionally significant by CommWalker translates to 37.93% and 40.70% of the enriched modules after consensus clustering for the hypoxia and macrophage data sets respectively. Thus, 11 of 29 enriched modules would not have been found without using CommWalker in the hypoxia data set, and 35 of 86 for the macrophage

data set. This group of modules includes the RNA polymerase promoter and autophagosome modules (Figures 7.4 and 7.6). These results show the power of the CommWalker framework in an applied setting.

7.5 Discussion and conclusions

In this chapter we have demonstrated how functions that are differentially regulated in contrasted cellular phenotypes can be identified by putting DEGs into the context of functional modules. Using our pipeline we were able to find several functional modules of proteins that are known to be involved in processes that are differentially active between hypoxic and normoxic cell lines, or M1 and M2 macrophages.

The developed pipeline, which objectively generates biological hypotheses from differential expression data sets, is an advancement over using biological intuition for refining lists of DEGs. As novel biology is difficult to evaluate computationally, we have focussed on well-studied functions. This prioritization has the effect that the results presented in this chapter may equally be found using curated lists of genes such as those found in the Molecular Signature Database (MSigDB) [52]. Yet results showing that $\approx 40\%$ of modules were only found due to integration of the CommWalker framework into our module detection method (see Section 7.4) suggest that less well-studied functions are also captured in our enriched modules. Fine-tuning of pipeline parametrization such as the T-value threshold, the DEG enrichment threshold, and the FDR threshold for determining which genes are differentially expressed may further improve our ability to investigate poorly studied functions. For example, VEGF-mediated angiogenesis was missing from our breast cancer hypoxia enriched modules. This absence was likely due to our parameter selection. VEGF was not sufficiently differentially expressed by our FDR threshold, and angiogenesis modules may have fallen through our functional significance filter.

An improvement over the MSigDB approach is the availability of candidate proteins that are involved in the differentially regulated functions. We have shown that our modules include proteins whose genes are known to be linked to the phenotype without being detected as DEGs in our data set (for example ULK1 or

VAMP3 in Section 7.3.2, and CCL21 or CCL25 in Section 7.3.3). Experimental validation of these modules, as discussed for the chemotaxis module (Figure 7.9), may uncover further proteins that play a role in the investigated phenotypes. An experimentally validated enriched module would show the power of this methodology in comparison to other approaches.

A notable absence in the enriched modules is that of the known hypoxia phenotypic marker HIF1, which revolves around transcription. DNA-binding interactions are not included in our underlying network and therefore such transcription factors are also missing in the set of functional modules. Instead, the pipeline represents a different perspective to the common analysis of transcription in gene expression studies. Our methodology detects the functions that are affected downstream of the initial signal. This focus allows the pipeline to complement current data analysis methods commonly used by experimental biologists that focus on the transcriptional response.

The enriched modules detected by our pipeline may be related to those obtained by active subnetwork methods (see Section 1.1.2). The difference between these approaches lies in their definition of a module (see Section 1.1.3). Active subnetworks are designed to select subnetworks of highly differentially expressed genes, yet our pipeline detected several enriched modules whose genes did not all exhibit significant differential expression but still captured known differentially regulated functions (eg. in the deacidification module in Figure 7.5). As we performed module validation by literature searching together with experimental collaborators, unfortunately a thorough comparison of the two approaches could not be performed given the time constraints.

A component of our pipeline that is not included in active subnetwork approaches is that of multi-resolution community detection. While our analysis suggested that multi-resolution partitioning is an important factor in the identification of enriched modules, the way modules across resolutions are reported can be improved. Many enriched modules generated by the pipeline represent the same differentially regulated function at different resolutions. Clustering methods can be used to

identify functional module clusters across resolutions to improve the reporting of results from the pipeline thereby making it easier to use.

Other ways in which the pipeline can be extended is by adequate correction for multiple testing in DEG enrichment, or by a more complex evaluation of differential expression altogether. Multiple testing corrections can be applied based on the expected fraction of communities that are consensus clustered into modules (see Section 7.2.3). For example, if 40% of communities are expected to be combined by consensus clustering, it can be estimated that 60% of the tests that are performed are not strongly dependent. A thorough handling of errors, which could include uncertainty in DEG assignment as well, may enable us to assess more subtle signals for differential regulation. A higher sensitivity in this evaluation can also be achieved by using the differential expression of all genes, instead of only those most differentially expressed. The GSEA method can be used in conjunction with our detected functional modules for this purpose.

8

Conclusions and future work

In this project we have developed a methodology to find differentially regulated functions between contrasted cellular phenotypes. This methodology exploits the mapping from genotype to phenotype provided by functional modules. Using functional modules, DEGs can be put into the context of the functions in which they are involved, that may ultimately affect the investigated phenotype.

We detected functional modules by a process that combines topological and biological clustering that was refined in Chapters 4–6. Following an exploratory study of the use of functional annotations to evaluate communities of proteins in PINs, we used the Pandey measure with GO BP annotations to assess the functional similarity of proteins (Chapter 4). These functional similarity scores were combined to assess the functional significance of communities using the CommWalker framework, which we developed to overcome a lack of detailed functional knowledge in certain PIN regions (Chapter 6). This biological clustering procedure was applied to subselect biologically meaningful communities from those detected by multi-resolution community detection methods (topological clustering) which we reviewed in Chapter 5. Furthermore, we started developing an improved topological clustering method called Link Edgehop Clustering (Appendix A), which however requires further testing.

Using our pipeline on differential gene expression data contrasting hypoxic and normoxic breast cancer cell lines, and M1 and M2 macrophages, we were able to recover functions known to be differentially regulated in these phenotypes (see Chapter 7). The functions were detected via DEG-enriched functional modules. These DEG-enriched modules each represent an experimentally testable hypothesis describing a differentially regulated function, and the proteins that are involved in this function. Using published data, we could validate several proteins in the enriched modules that did not correspond to DEGs in our input data set.

While several computational methods for interpreting differential gene expression data exist, the main contribution of this thesis lies in the use of multi-scale topological clustering and CommWalker biological clustering in this process. The detection of communities across resolutions attempts to mirror the hierarchical organization of cellular functions in PINs (see Section 1.1.3). Thus, by testing for DEG enrichment of potential modules across resolutions, the pipeline can automatically select the scale at which cellular functions are involved in the phenotype. The integration of CommWalker into this pipeline enables it to counteract the selection for well-studied functions when interpreting the differential expression data using functional modules. By taking into account that knowledge of protein function is unevenly distributed on PINs, CommWalker can evaluate communities as functionally significant even when they are poorly functionally annotated. Thus, we can test for DEG enrichment on a set of functional modules that represent a broad spectrum of cellular functions. The DEG-enriched modules detected using the breast cancer and macrophage gene expression data sets show that these features are important to the performance of our pipeline (see Section 7.4).

The work presented in this thesis shows a promising approach to interpreting differential gene expression data that can directly generate hypotheses on the investigated phenotype. As such, there is work that follows on from the results presented here. In future, we would like to quantify the advantgates of our methodology through in-depth comparisons with other approaches, expand several components of the pipeline, and investigate the limitations of others.

In particular, the limits of the CommWalker framework merit further investigation. While we have shown that CommWalker biological clustering counteracts annotation bias in functional homogeneity calculation, its robustness to decreasing levels of functional knowledge (both in specificity and in coverage) is yet to be investigated. As CommWalker can only amplify an existing functional homogeneity signal, there is a limit to the lack of information CommWalker can deal with. For example, low coverage can lead to the same two nodes being traversed in all random walks such that these have the same functional homogeneity and the community is assigned the lowest possible T-value regardless of the functional homogeneity score (see Equation (6.1)). If additionally the two proteins lack specific functional annotations (they are only associated with the BP root annotation), this functional homogeneity score could be 0, while the T-value is significant. This is not a correct assessment of the functional significance, it is merely an expression of a complete lack of knowledge of the network environment.

Further investigation could quantify the benefits of our pipeline by comparing the DEG-enriched modules for the breast cancer and macrophage data sets with active subnetworks detected using the same differential expression data. Furthermore, it would be interesting to test whether the differentially regulated functions found by our methodology are also found by disease module approaches that build a PIN between disease-related genes (here DEGs) and perform GO term enrichment on this disease network [13, 44]. In future we aim to apply the methodology to further differential expression data sets of phenotypes that are less well-studied with a specific focus on experimentally validating modules that may be of therapeutic relevance.

There are several possible extensions to the methods we use, as well as methods and data sets that were recently published, that may improve the components of the pipeline. Small improvements can be made to topological and biological clustering methods; the method of incorporating expression data can be refined; recently published semantic similarity measures can be tested and incorporated; and the pipeline can be amended to allow for future use of isoform-resolved protein-protein

interaction data. These extensions may improve the sensitivity of the methodology or allow it to depict the underlying biological processes more precisely.

Topological and biological clustering can be improved in the following two ways. Firstly, a resolution parameter could be built into BigCLAM by making the threshold that is used to turn the node-community interaction strengths F_{uk} (see Section 2.3.3.1) into discrete community memberships a variable. This community detection method is of particular interest as the analysis performed in Appendix D.2 shows that BigCLAM detects more functionally significant modules in our size range of interest than the other tested community detection methods. Secondly, CommWalker biological clustering can be extended by reining in the random walks as proposed in Section 6.6.

A further component improvement can be implemented in the way differential gene expression data is analysed in the pipeline. One of the drawbacks of one-by-one differential gene expression analysis is the inability to account for the dependence structure of gene expression levels. Yet, this approach is necessary given the high dimensionality of gene expression data sets (see Section 2.6.2). It has been suggested that the complexity of differential gene expression analysis can be reduced by incorporating biological knowledge to select relevant gene sets [202]. Functional modules could be used as these gene sets. Thus, instead of investigating the enrichment of DEGs on functional modules, assessing how well the genes in a functional module model differential expression would allow us to consider the dependence structure between genes in a biologically informed way. A simple application of this idea is to use functional modules as proposal gene sets for gene set enrichment analysis [52]. This approach may increase the power of our enrichment test.

Recent developments that can be integrated into the pipeline to refine the methodology include the TopoICSim measure [225] and isoform-resolved protein-protein interaction data (eg. [77, 78]). TopoICSim is a semantic similarity measure that was shown to compare favourably to simGIC and other semantic similarity measures on a range of tests [225]. Furthermore, it appears to be less affected by

annotation length bias than other semantic similarity measures. As independent comparisons of this measure have not been performed due to its recent publication (July 2016), it would be interesting to evaluate TopoICSim in our simulation test (see Section 4.4).

While comprehensive isoform-resolved protein-protein interaction data are currently not available (see Section 2.1.1), recently such data sets have been obtained for specific tissues (eg. the brain [77, 78]). Assuming this development continues, it may be beneficial to move to PINs which label proteins by UniProt identifiers which can describe interactions between protein isoforms. This type of protein-protein interaction data can be obtained from the IntAct database (see Section 2.1.3). Integrating RNA interactions into the underlying network data (eg. [33, 34]) may further improve the precision of the biological processes represented in the detected functional modules.

Our methodology provides an automated platform for the analysis of differential expression data that is specifically catered towards functions that are poorly understood. As such, this may be the right tool to use on complex diseases where conventional approaches have failed. With some luck, we can detect novel biology that is of therapeutic relevance.

Appendices

A

Link Edgehop Clustering

Contents

A.1 Introduction 197
A.2 Data sets and processing 198
A.2.1 Networks 198
A.2.2 Gene Ontology annotations 200
A.3 Ledgehop methodology 203
A.4 Comparing network partitions: Normalized Mutual Information 205
A.5 Testing Ledgehop 207
A.5.1 Differences between Ledgehop and link clustering 208
A.5.2 Capturing known community structure 210
A.6 Biological applications 214
A.7 Discussion and conclusions 216

A.1 Introduction

In this appendix we describe a novel methodology for community detection. This topological clustering method is based on link clustering and is called Link Edgehop Clustering (Ledgehop). The aim behind Ledgehop is to exploit the strong performance of link clustering for functional module identification whilst overcoming its relatively poorer performance when identifying modules in less dense networks (see Chapter 5).

Link clustering partitions a network into communities by clustering edges using an edge similarity score (see Section 2.3.3.2). The similarity of two edges that share a node (the "keystone node") is computed based on the overlap of the neighbourhoods of the nodes that are not shared by the edges (the "endpoint nodes"; see Equation (2.10)). Ledgehop extends this similarity measure by considering more of the information available in the local network area than just direct neighbours.

In order to evaluate Ledgehop, we compared the performance of Ledgehop and link clustering using networks with known community structure and networks with covariate data available that is thought to be related to community structure, including social networks and PINs clustered by GO slim annotations. Testing Ledgehop on these networks, we found that Ledgehop's performance is comparable to link clustering on dense social networks and better than link clustering on one of the two PINs. Further test cases are required to draw conclusion from these results. Applying Ledgehop to the biological application data sets from Chapter 7, we discovered enriched modules that are found exclusively by Ledgehop-based functional module detection.

A.2 Data sets and processing

Ledgehop and link clustering was compared using several network data sets. As knowledge of the underlying community structure is available in our social network data sets we used these in the first instance. Ledgehop was subsequently evaluated on PINs using GO term communities as an approximation of ground-truth. Here we briefly describe these data sets.

A.2.1 Networks

A.2.1.1 Social network data

Social network data sets are commonly used to evaluate community detection methods (see Section 2.3.4). We used the following social networks to evaluate Ledgehop community detection.

The Facebook100 data set is a set of 100 Facebook friendship networks from different US colleges and universities captured on the same day in September 2005 [149, 150] (`http://sociograph.blogspot.co.uk/2011/03/facebook100-data-and-parser-for-it.html`, retrieved April 2016). As friendship between two people is mutually agreed upon by so-called "friendship requests" in this online social networking platform, edges are undirected. The data set further contains covariate information on the nodes regarding gender, dormitory, major, second major, year of enrolment in the university, high school, and whether the node is a student of a member of faculty. The available covariate data does not have full coverage. While Facebook networks with numbers of nodes or edges similar to PINs exist in this data set, network densities tend to be considerably higher (up to 70-fold, see Table B.2 in the Appendix).

Social networks with known community structure were obtained from the Stanford Network Analysis Platform (SNAP, `snap.stanford.edu/data/`, retrieved April 2016). SNAP contains networks with known community structure that are considerably larger in size than for example the Zachary Karate Club network which is commonly used to evaluate community detection methods (see Section 2.3.4). These larger networks with known community structure represent a more stringent test for community detection. We used the smallest networks with known community structure that were available from SNAP to ensure that Ledgehop can be run at multiple resolutions in a reasonable time frame. These networks consist of an Amazon co-purchasing network [152], a DBLP collaboration network [152], and a Youtube friendship network [151, 152]. Here, ground-truth communities correspond to Amazon product categories for items often purchased together, publication venues for groups of computer scientists that publish together, and Youtube topic-based interest groups for friendship communities. These networks are larger than PINs with higher clustering coefficients, while also being considerably less dense. Network statistics for these networks are shown in Table B.2 in the Appendix.

A.2.1.2 Protein-protein interaction data

As Ledgehop was primarily developed to improve community detection performance on PINs, Ledgehop was also tested on the HINT-P and BioGrid-AP networks used in Chapters 4, 6, and 7. These PINs are described in Section 3.1.1.

A.2.2 Gene Ontology annotations

In Chapter 4 we investigated which functional annotations best capture the homogeneity of proteins grouped in the same community. Following this investigation we used GO BP terms to evaluate communities generated by Ledgehop and link clustering. Specifically, GO BP terms (obtained as described in Section 3.2.1) were used to cluster proteins into so-called label communities which were interpreted as an approximation to ground-truth modules in PINs.

GO BP terms that are associated with proteins vary in their specificity (see Section 2.4.2). To generate label communities based on these terms, it is necessary to have a categorization of GO BP terms which can be used to cluster proteins that share broad classes of functions rather than exact labels. Such a categorization is provided by GO slims (see Section 2.4.2.1).

GO slims can either be obtained directly from the GO, or generated using computational methods. Available GO slims provide a high level overview of protein characteristics using GO terms with low specificities. In contrast, computational methods can generate GO slims at different target specificities. In order to compare network partitions generated at different resolutions to suitable GO slim network decompositions, we used computational methods to generate GO slims for a range of target specificities.

Computational generation of GO slims generally relies on information content (IC, see Equation (2.11)) to evaluate the specificity of a term. To generate a GO slim at a target IC, algorithms are used to find sets of GO terms with specificities similar to the target IC that cover all paths from the leaves to the root node. Having this coverage ensures that any GO term can be mapped to a term in the GO slim by following their ancestral path (see ancestral path mappping in Section 2.4.1). GO

terms that cannot be mapped to the target specificity GO slim set, are mapped to the root node, which is included in the GO slim. Here, we used two computational methods to generate GO slims: the Alterovitz method and the threshold method.

Given a target specificity S measured via IC, the Alterovitz method [160] outputs the set of GO terms whose specificity is most similar to S while covering all possible paths from the leaves to the root. This selection is achieved by ranking all GO terms by the difference between their specificity and the target specificity S. The term whose specificity is closest to S is iteratively added to the slim set, under the condition that it is not an ancestral or descendant term of one already in this set. This method, which was used to produce the Gene Ontology Partition Database [160], may generate slim sets with GO terms of considerably higher or lower specificity than the target specificity S.

We also used our own conservative approach for GO slim generation, which we call the threshold method. Rather than including the GO terms whose specificity is most similar to the target specificity S, we use S as a threshold for GO term specificity. Starting at the leaves of the GO we trace each path to the root node until a term with an IC $< S$ is encountered. This term is added to the slim set and the process is continued until all leaf terms and paths have been considered. An example slim set generated by this method on a simplified ontology structure is shown in Figure A.1.

The number of GO BP terms in the GO slims at varying levels of IC target specificity is shown in Table A.1.

GO BP term associations were mapped to the GO slim terms using the Map2GOslim tool (`https://github.com/owlcollab/owltools` [203], retrieved May 2015). While the GO slim sets grow with higher target specificity, the number of associations mapped to the BP root term also increases. Proteins whose associated GO terms are mapped only to the root term are not included in the label communities, as the root term does not represent an informative label. Thus, the higher the specificity of the GO slim set, the fewer proteins are included in the label communities. For example, using the GO slim generated by the Alterovitz method at a target IC of 0.5, 858 of 13,134 annotated proteins are mapped only to

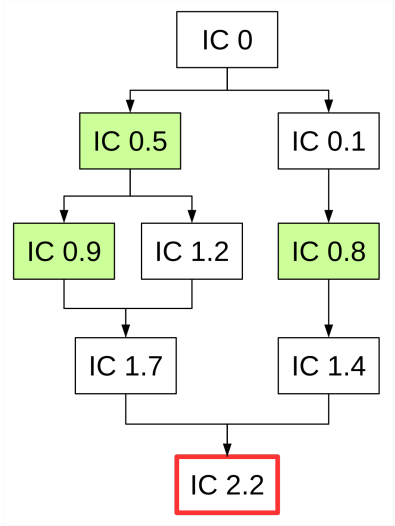

Figure A.1: Example slim set generated by the threshold method on a schematic ontology for a target IC of 1. Ontology terms are are shown with their IC specificities as labels. Starting from the leaf term in red, the threshold method traverses all paths from this leaf to the root node (IC of 0) until it hits a term with an IC below the target specificity of 1. This process gives the slim set shown in green. In contrast, the Alterovitz method with target IC of 1 would generate the slim set with the term ICs 0.9, 1.2 and 0.8.

Table A.1: The number of GO BP terms in the GO slims.

Target IC Specificity	Alterovitz	Threshold
0.5	44	9
1.0	95	21
1.5	204	74
2.0	334	115
2.5	545	198
3.0	713	311
3.5	1082	482
4.0	1407	721
4.5	1858	1057
5.0	2300	1512

Table A.1: GO slims were only calculated on the BP sub-ontology of the GO. "Alterovitz" refers to the Alterovitz method for GO slim generation [160] and "Threshold" refers to the threshold method we developed.

the BP root term. In contrast, using the IC 2.5 Alterovity slim set this number is 7,531 of 13,134 annotated proteins. Label communities for the IC 2.5 GO slim thus exclude 57.34% of proteins that are associated with GO BP terms.

A.3 Ledgehop methodology

Link clustering is a community detection method that clusters edges based on a local measure for edge similarity (see Section 2.3.3.2). While this method outperformed others in an exploratory analysis of community detection methods for functional module detection (see Chapter 5), our results also suggested a network density dependence of the performance of link clustering. This density dependence can be understood considering the local edge similarity measure:

$$s(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}. \tag{2.10}$$

Here, $n_+(i)$ denotes the inclusive neighbour set of node i (see Section 2.3.3.2), which is used to calculate the similarity of edge e_{ik} between nodes k and i, and edge e_{jk}. As denser networks contain more edges, more information is available at a local level for the edge similarity calculation. To address this limitation we developed Ledgehop community detection which takes more local network information into account in the edge similarity calculation (Figure A.2).

While link clustering uses node overlaps between the inclusive neighbour sets $n_+(i)$ of the so-called endpoint nodes i and j (see Section 2.3.3.2), Ledgehop also considers edges that link the two inclusive neighbour sets (see Figure A.2). Thus, the edge similarity score is extended from Equation (2.10) to:

$$s(e_{ki}, e_{kj}) = \frac{1 + \sum_{l \in n_+^{(-k)}(i)} |n_+^{(-k)}(l) \cap n_+^{(-k)}(j)|}{1 + |n_+^{(-k)}(i)||n_+^{(-k)}(j)|}. \tag{A.1}$$

Here, the similarity of edges e_{ki} and e_{kj} is calculated based on overlaps between the inclusive neighbour sets excluding the keystone node k, denoted $n_+^{(-k)}(i)$ for node i. Specifically, the sum adds the node overlaps between the inclusive neighbour set of endpoint node j without keystone node k, and the inclusive neighbour sets excluding k of all nodes l that are members of the inclusive neighbour set of endpoint node i excluding the keystone node k. Ledgehop thus considers nodes that are shared between the sets $n_+^{(-k)}(i)$ and $n_+^{(-k)}(j)$, and edges linking the nodes in these sets. Equation (A.1) doubly weights the shared nodes by counting the

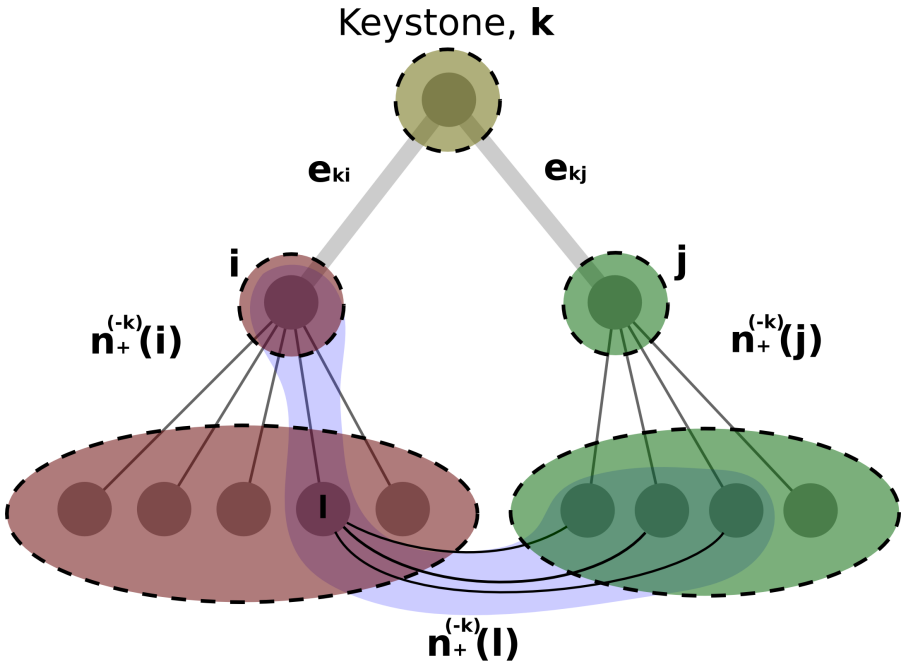

Figure A.2: Schematic diagram of information considered in Ledgehop edge similarity score. Ledgehop calculates the similarity of edges e_{ki} and e_{kj} by counting the number of edges between nodes in the inclusive neighbour sets of the endpoint nodes i and j without keystone node k (see Equation (A.1)). The inclusive neighbour sets of nodes i and j excluding k are denoted $n_+^{(-k)}(i)$ and $n_+^{(-k)}(j)$ and are shown in red and green respectively. The keystone node k is trivially part of both inclusive neighbour sets and thus excluded in the calculation. Node l is a node in the set $n_+^{(-k)}(i)$, with the inclusive neighbour set excluding k $n_+^{(-k)}(l)$ shown by the purple outline. While link clustering would assign the edges e_{ki} and e_{kj} a similarity score of 0 (see Equation (2.10)), the Ledgehop similarity of these edges is $\frac{1+3}{6\times5+1} = \frac{4}{31}$. The only node in the set $n_+^{(-k)}(i)$ that contributes to this score is node l, which has three links to nodes in the set $n_+^{(-k)}(j)$.

edges with each endpoint node once, and singly weights edges between nodes in these inclusive neighbour sets. Although the sum in Equation (A.1) goes over the inclusive neighbour set excluding k of only node i, the score is symmetric.

A notable difference between the link clustering and Ledgehop similarity scores is the exclusion of the keystone node k (see Equations (2.10) and (A.1)). The rationale behind this setup is the following. The keystone node k trivially belongs to both inclusive neighbour sets of the endpoint nodes. Allowing the k to affect the edge similarity would have had the effect that the inclusive neighbourhood of only one endpoint node can dominate the similarity score if many nodes in this set are connected to the keystone node. In this case a high similarity score can be achieved with any edge that is connected to this keystone node.

Calculating edge similarities by Equation (A.1) allows more network information

to be taken into account without extending the calculation past what could be considered the local network environment. Initial tests using link clustering edge similarity (Equation (2.10)) with two-hop inclusive neighbour sets illustrated the effect of using neighbour sets that extend past the local network environment (these neighbour sets can include 60% of a PIN). In this two-hop approach edge similarity scores were dominated by common connections with hub nodes, which led to a partition that resembled a core-periphery network decomposition (see Section 2.2.2). In order to avoid this situation in Ledgehop, we consider only edges between the defined neighbour sets from link clustering.

A.4 Comparing network partitions: Normalized Mutual Information

To evaluate the performance of Ledgehop and link clustering, network partitions generated by these methods were compared to network partitions based on node labels. These node labels were obtained from covariate data on the nodes such as major subjects of students in Facebook friendship networks or GO slim term associations in PINs (see Section A.2). Community detection network partitions and label-based network partitions were compared using an extension of normalized mutual information (NMI) to network covers (code from `https://sites.google.com/site/andrealancichinetti/mutual` [226], retrieved April 2016).

NMI calculates the similarity of two network partitions based on the amount of additional information given by a partition if the other partition is known. Here, information theory is used to quantify the information contained in a network partition (briefly reviewed in [129]) by interpreting the community assignments of two network partitions as random variables X and Y. For a network with N nodes, X is a vector of length N that contains the community assignment of each node in the network. The distribution of the variable X can be used to define the Shannon information entropy (see also Shannon information content in Section 2.4.1) of a partition $H(X)$ by the equation:

$$H(X) = -\sum_{x} P(X = x) \log P(X = x),$$

where $P(X)$ denotes the empirical probability distribution of node-community assignments calculated by $P(X = x) = |c_x|/N$. Here, $|c_x|$ is the number of nodes in community $c_x \in C_X$, where C_X denotes all communities in the network partition captured in X. Similarly, the conditional information entropy $H(X|Y)$, which quantifies the additional information provided by partition X given that partition Y is known, can be calculated by

$$H(X|Y) = -\sum_{x}\sum_{y} P(X = x, Y = y) \log P(X = x|Y = y).$$

Here, $P(X = x, Y = y)$ denotes the joint empirical probability distribution of variables X and Y, and P(X=x|Y=y) is the conditional empirical probability distribution of X given Y.

The NMI $I_{norm}(X, Y)$ is a measure that combines these quantities via the equation [227]:

$$I_{norm}(X, Y) = \frac{2(H(X) - H(X|Y))}{H(X) + H(Y)}. \tag{A.2}$$

When the network partitions, and thus the distributions of random variables X and Y are equal, $H(X|Y) = 0$ as there is no uncertainty about X when Y is known. In this scenario $H(X) = H(Y)$ and therefore $I_{norm} = 1$.

In the form shown in Equation (A.2), NMI assumes that each node is labelled with a single value determining its community assignment. In the case of overlapping communities this assumption does not hold. By allowing node-community assignments to be represented as vectors, NMI has been extended to network covers [226]. In this extension, overlapping communities in two network covers are matched based on the minimization of their conditional information entropy. This matching is used to define the overlapping information entropy between two network covers. The overlapping NMI extension is given by the equation [226]:

$$N(\mathbf{X}|\mathbf{Y}) = 1 - \frac{1}{2}[H(\mathbf{X}|\mathbf{Y})_{\text{norm}} + H(\mathbf{Y}|\mathbf{X})_{\text{norm}}], \tag{A.3}$$

where $H(\mathbf{X}|\mathbf{Y})_{\text{norm}}$ denotes the normalized overlapping conditional information entropy given by

$$H(\mathbf{X}|\mathbf{Y})_{\text{norm}} = \frac{1}{|C_X|}\sum_k \frac{\min_{l\in\{1,2...|C_Y|\}} H(X_k|Y_l)}{H(X_k)}.$$

Here, X_k denotes the binary vector of length N that represents the membership of each node in the community k, and $\mathbf{X}$ is the $|C_X| \times N$ matrix that represents the network cover with the community set C_X.

While this extension of NMI does not exactly recreate the classical NMI (Equation (A.2)) in the absence of community overlaps, it has become a popular method of comparing overlapping communities to covariate information (eg. [113, 144, 213]). When NMI is mentioned in this appendix, the extension of NMI to overlapping communities given by Equation (A.3) is referred to.

A.5 Testing Ledgehop

In this section we present the results of the comparison between Ledgehop and link clustering. The perfomance of the community detection methods was compared based on the NMI between the generated network partitions and label partitions from covariates or known community structure. Label network partitions were obtained by clustering together nodes with the same label or community assignment. Nodes without covariate information were excluded from these partitions. As label communities are generated based only on covariate information, these communities may be disconnected which indicates that the covariate is not a good predictor for community structure. Such a scenario would lead to low NMI scores for both methods. Importantly, Ledgehop and link clustering are always evaluated against the same data set. Thus, the comparison remains fair even when the covariate only conveys a weak signal for ground-truth community structure.

Community detection network partitions were generated at 60 resolutions spanning the range $s \in [0.005, 0.3]$ for the Facebook100 networks, and 121 resolutions in the range $s \in [0, 0.6]$ in increments of $\Delta s = 0.005$ for the SNAP networks and the two PINs. As communities of size two represent trivial single-edge communities, community detection and label-based network partitions were filtered to include only communities with at least three nodes.

A.5.1 Differences between Ledgehop and link clustering

Given that Ledgehop was extended from link clustering, we first investigated whether the methods generate different network partitions. This analysis was performed by matching each Ledgehop network partition with the most similar link clustering network partition across resolutions. Here, similarity was measured based on NMI. As a network partition that contains 2 communities can be matched with a network partition that contains 6,510 communities using this method (NMI of 0.451 for American75 network partitions at Ledgehop resolution s=0.3 and link clustering resolution s=0.005), we only considered matches where

$$\frac{||C_{\text{Link}}| - |C_{\text{Ledge}}||}{|C_{\text{Ledge}}|} \leq 0.5. \tag{A.4}$$

Here, $|C_{\text{Link}}|$ denotes the number of communities in the link clustering network partition. The NMI values of the matched network partitions are shown in Figure A.3.

The results presented in Figure A.3 can be put into context based on a previous comparison of community detection methods using NMI [213]. In this comparison, the average NMI across the 100 Facebook networks between network partitions from seven community detection methods ranged between 0.1 and 0.5 (see Figure A.4 reproduced from [213]). Considering these values, the results show that while network partitions at low resolutions may be identical, Ledgehop and link clustering differ as much as other common community detection methods. Especially at high resolutions, many matched network partition NMI scores are below 0.2. The two methods are however more similar than link clustering is to the other tested methods

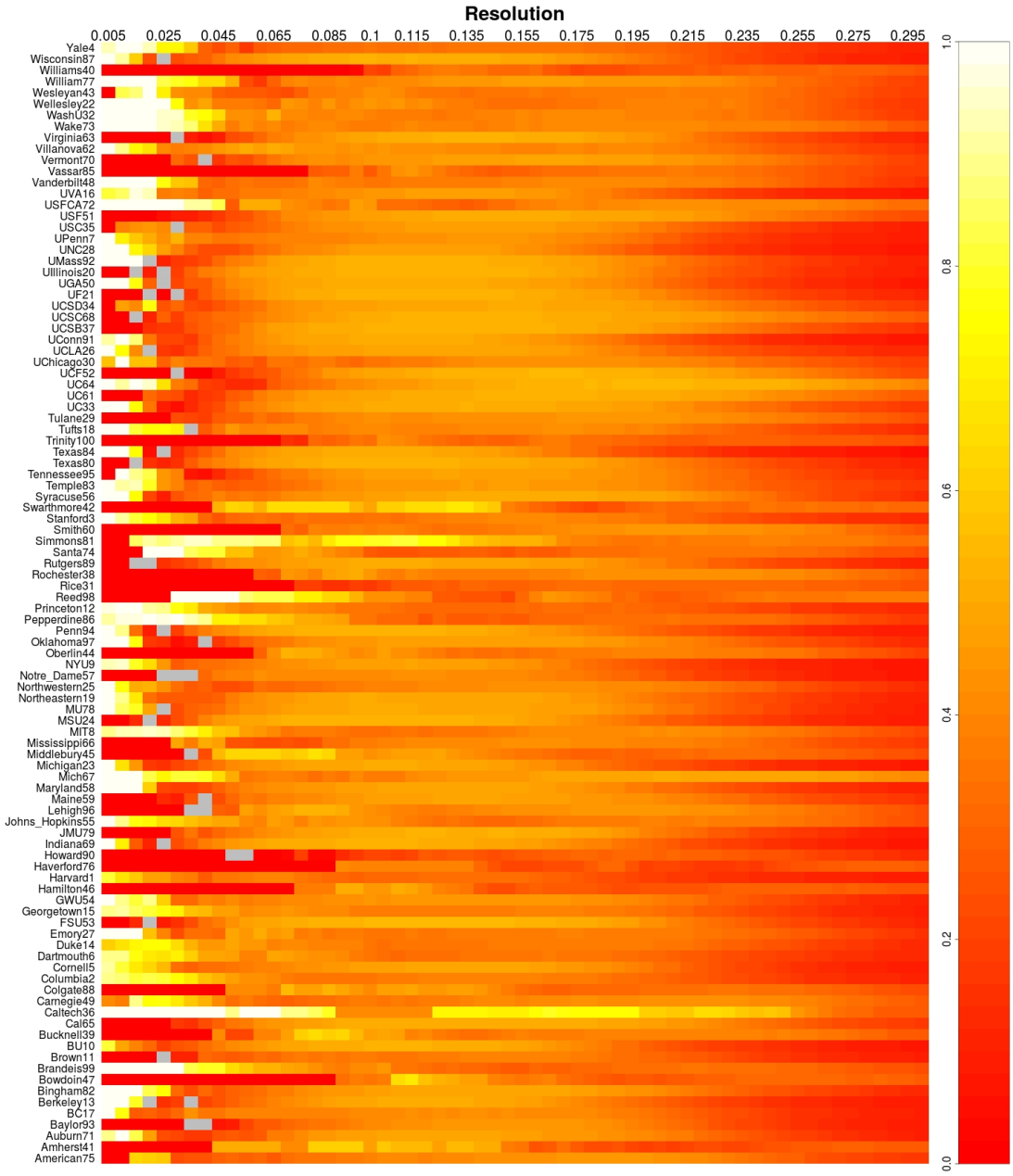

Figure A.3: Facebook100 Ledgehop link clustering similarity analysis. The vertical axis shows data for the different Facebook networks, and the horizontal axis represents Ledgehop resolution s at which network partitions were generated. For each Facebook100 network, Ledgehop network partitions at each resolution were matched with the most similar link clustering network partition based on NMI. Only link clustering partitions which fulfil Equation (A.4) were considered in this matching. The NMI values are shown on a colour scale from red to white, where red indicates unrelated, and white indicates similar partitions. Network partitions where no match could be found that satisfies Equation (A.4) are shown in grey. While partitions at low resolution can be identical, network partitions at higher resolutions are less similar. Indeed other common community detection methods show as much similarity as link clustering and Ledgehop especially at high resolutions (see Figure A.4 reproduced from [213]). However, link clustering is more similar to Ledgehop than it is to other investigated community detection methods.

(Copra [228], Ganxis [229], Infomap [230], Infomap single [231], Louvain [126], Oslom [232]; see Figure A.4).

A.5.2 Capturing known community structure

A.5.2.1 Social networks

To evaluate the ability of a community detection method to capture the label-based network structure, we used the maximum NMI between label and community detection network partitions across resolutions to score the performance of the method. This NMI-based performance metric has previously been used on the Facebook100 data set to benchmark community detection methods, including link clustering [213]. Figure A.4 (reproduced from [213]) shows the results of this analysis using NMI values averaged across the 100 networks. These results indicate that community detection network partitions and label community partitions tend to have an NMI between 0 and 0.1. Based on these data and other analyses, the authors in [213] concluded that while covariates correlate with topological community structure they cannot be used to recreate this structure. These results allow us to put the results of our NMI-based performance metric into the context of other community detection methods.

The results of our performance analysis of Ledgehop and link clustering for each network in the Facebook100 data set is shown in Figure A.5. Figure A.5 suggests that Ledgehop performs comparable to link clustering on the dense Facebook networks. For example, while link clustering appears to better capture label communities in the FSU53, Carnegie49, and Vermont70 networks, Ledgehop outperforms link clustering on Smith60, Bowdoin47, and Haverford76. Overall it appears that including more information for local edge similarity calculation has not improved the partitions relative to the covariates on these dense networks.

The results of the NMI analysis on the SNAP networks is shown in Table A.2. Similar to the relatively dense Facebook100 networks, it appears that Ledgehop community detection does not outperform link clustering on the large SNAP networks either. Despite all SNAP networks exhibiting a comparatively low density,

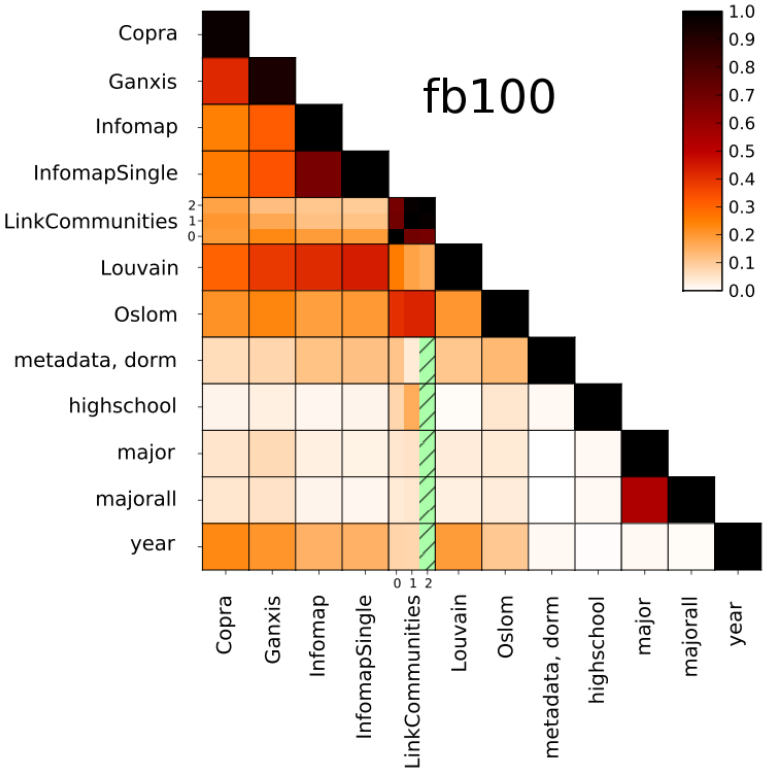

Figure A.4: Comparison of community detection methods and covariate information on the Facebook100 data set. Network partitions from seven community detection methods (upper seven rows) and five covariates (bottom five rows) were compared based on the NMI of the partitions. The tiles were coloured by the average NMI values for this comparison across the 100 Facebook networks in the Facebook100 data set. The darker the colour, the more similar the network partitions. Label communities were obtained by clustering nodes with the same label, using the connected components of these communities, and excluding components with less than three nodes. To maximize the score, only nodes present in both the community detection and the covariate network partitions were used to calculate the NMI. When less than 10% overlap was found between a network partition and a label partition, the NMI was not computed and shown as a green box with black lines (see "LinkCommunities" at resolution 2, denoting $S = 0.75$). Details of the community detection implementations can be found in [213]. The covariates are described in Section A.2.1.1 with "Majorall" representing the combination of major and second major covariates. (Figure reproduced from [213]).

our results show that Ledgehop and link clustering both perform comparativelly well on the Amazon and DBLP network but relatively poorly on the Youtube network. This difference may be due to higher average local clustering coefficients (ALCs) of these networks (see Table B.2). Indeed, in the Facebook100 data set neither method performs well on the Northeastern19, Duke14, and Texas80 networks which have ALCs of 0.252 and below, while both methods have NMI performance scores above 0.25 for several covariates on the Simmonds81, Hamilton46, and Swarthmore42 networks which have ALCs of 0.299 and above. Alternatively, it may be the case that

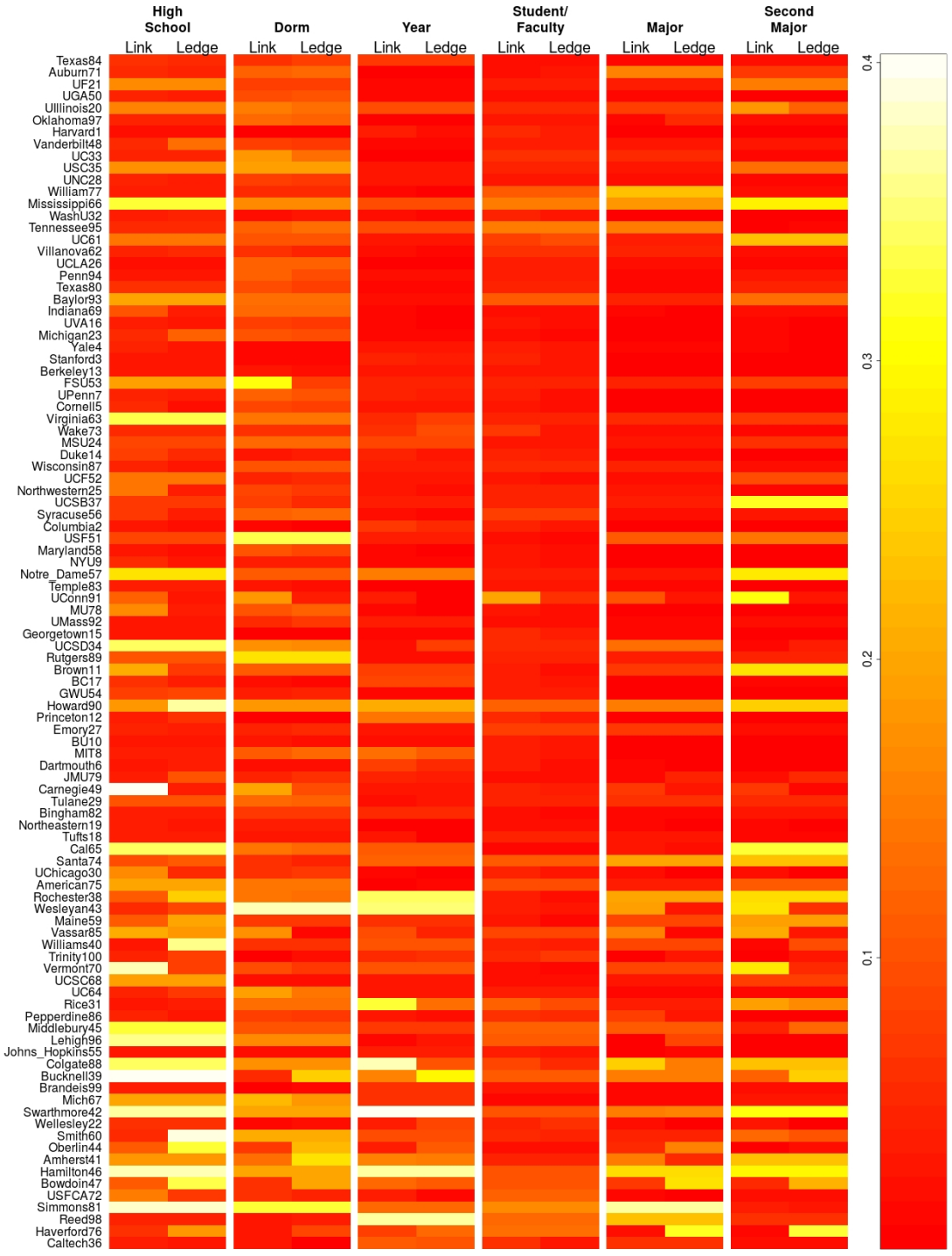

Figure A.5: Facebook100 NMI analysis results. The NMI (Equation (A.3)) of covariate-based label communities and network partitions by link clustering (Link) and Ledgehop (Ledge) was computed for each network in the Facebook100 data set. For each network and community detection method only the maximum NMI obtained across network partitions generated at different resolutions is shown. Low NMI scores (red) denote that network partitions from a specific community detection method do not approximate the label network partitions well. Ledgehop performs comparable to link clustering on this data set and thus does not confer an advantage.

the covariate information are better predictors of community structure in networks where we observe a high ALC. A more detailed analysis must be performed to assess the significance of the link between ALCs and community detection performance.

Table A.2: SNAP NMI analysis results.

Data Set	Link	Ledgehop
Amazon	0.2932	0.3234
DBLP	0.2054	0.1985
Youtube	0.0559	0.0463

Table A.2: NMI values were calculated as in Figure A.5. Ledgehop and link clustering perform similarly on these networks.

A.5.2.2 PINs

The community structure of social networks is different from that of PINs or other biological networks [233]. As Ledgehop was developed to improve topological clustering of PINs, we investigated how well it captures GO slim label communities. While GO slim terms do not define ground-truth functional modules, they do provide a signal for community structure (see GO slim analysis in Section 4.2).

GO slim label communities were generated by the Alterovitz and threshold methods (see Section A.2.2) at 10 target IC specificities evenly spanning the range $IC \in [0.5, 5]$ (Figure A.6).

Figure A.6 shows that Ledgehop communities better capture GO slim communities in BioGrid-AP especiallly at higher specificities. The lower NMI scores for low specificity GO slim partitions may be due to the existence of several disconnected node clusters in these label communities. As the specificity increases, the label communities will get smaller and thus more likely to be connected.

As proteins that are better studied will have more specific annotations, the label communities for high specificity GO slims are likely to contain well-studied proteins that are in the densely connected core of the PIN (see Section 2.2.2 for core-periphery structure). Thus the results of this analysis may only suggest that Ledgehop performs better on well-studied protein communities.

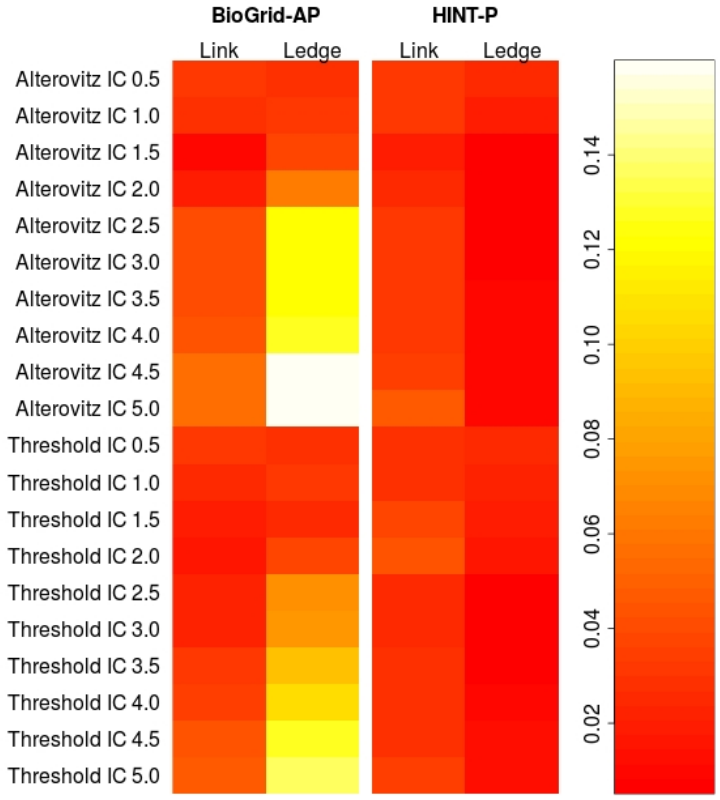

Figure A.6: PIN NMI analysis results. NMI analysis was performed as in Figure A.5 with label communities obtained based on protein GO slim associations. GO slims were generated computationally at several target IC specificites (see Section A.2.2). Proteins without GO slim annotations or with only BP root annotation were excluded from the label communities. In comparison with link clustering Ledgehop network partitions better capture the label communities on BioGrid-AP especially at higher specificities.

The poor performance of both methods on HINT-P may be due to HINT-P quality control, which may filter out edges necessary to recover the label communities by edge-based community detection.

A.6 Biological applications

As Ledgehop appears to improve topological partitioning of BioGrid-AP, we used Ledgehop community detection in our pipeline (see Section 7.2) to test whether it generates potential functional modules that can be used to map DEGs to differentially regulated functions. Using the same protocol that generated the

results in Section 7.3, we investigated whether Ledgehop communities can be used to find enriched modules of DEGs in the hypoxia (see Section 3.3.2.1) and macrophage (see Section 3.3.2.2) differential expression data sets.

In the hypoxia data set a Ledgehop enriched module was found that describes RNA-mediated modification of translation, which is a known differentially regulated function between hypoxic and normoxic cells [218]. Furthermore, the RNA polymerase I promoter module (Figure 7.4) was recovered as a Ledgehop community. As these differentially regulated functions, although not the exact modules, were also found by other community detection methods, Ledgehop does not appear to convey additional information for the hypoxia investigation.

Applying Ledgehop communities to the macrophage data set, we found several enriched modules describing differentially regulated functions related to how M1 and M2 macrophages were induced. Modules with the enriched GO terms "response to IFN-γ" or "response to stimulus" were identified both by Ledgehop and other community detection methods. Known macrophage-related functions such as phagocytosis/endocytosis and cytokine production (see Section 1.2.2) were also found to be differentially regulated in Ledgehop modules as well as link clustering modules. While endocytosis (a process that includes phagocytosis) has been predominantly attributed to M1 macrophages (see Section 1.2.2), it has been suggested that this may be due to a confirmation bias arising from endocytosis mainly being studied in M1 macrophages (personal communication with Dr. Fernando Martinez). Indeed, most of the detected enriched modules find endocytosis to be up-regulated in M2 macrophages. The module with the strongest signal for M2 up-regulation of endocytosis (by average fold change) is found exclusively by Ledgehop community detection.

A further enriched module that was exclusively found by Ledgehop is shown in Figure A.7. This module describes differential epigenetic regulation of transcription between M1 and M2 macrophages. While enriched modules from other community detection methods also contain histone deacetylases (HDACs) and histone acetyltransferases (KATs), this enriched module is the only one that shows

a signal for M2 up-regulation. Specifically HDAC2, HDAC3, and HDAC6 have been described as having anti-inflammatory functions in macrophages [234], a phenotype that has been attributed to M2 polarization (see Section 1.2.2). Of these HDACs, HDAC2 and HDAC3 are included in our module. Differential epigenetic regulation of transcription between macrophage polarizations is currently being studied and different HDACs have been shown to be related to different macrophage polarizations [234] (also supported by unpublished data from personal communication with Dr. Fernando Martinez).

As a further evaluation of Ledgehop community detection, we contrasted Ledgehop and link clustering based on their ability to detect functional modules that are enriched for DEGs from our two gene expression data sets. From this analysis, it appears that Ledgehop represents a new approach to functional module detection rather than an improvement on link clustering; Ledgehop cannot replace link clustering or vice versa. In the hypoxia and macrophage data sets 1 and 6 enriched modules were detected exclusively by Ledgehop. In contrast, 12 and 33 enriched modules were detected exclusively by link clustering respectively.

While link clustering plays a bigger role than Ledgehop in our module detection pipeline, Ledgehop does generate functional modules that would not otherwise be found. In the case of individual modules found by each method that describe the same functions, experimental validation is necessary to determine which module best represents the differentially regulated function.

A.7 Discussion and conclusions

In this appendix we described Ledgehop, a community detection method that was developed by extending the link clustering edge similarity measure to include information beyond direct neighbours of endpoint nodes. The motivation behind the development of this method was to address the network-density-dependent performance of link clustering (see Chapter 5). The NMI-based analysis performed in this appendix shows no evidence that this issue has been resolved. Specifically, the performance of link clustering and Ledgehop on HINT-P, the less dense of the

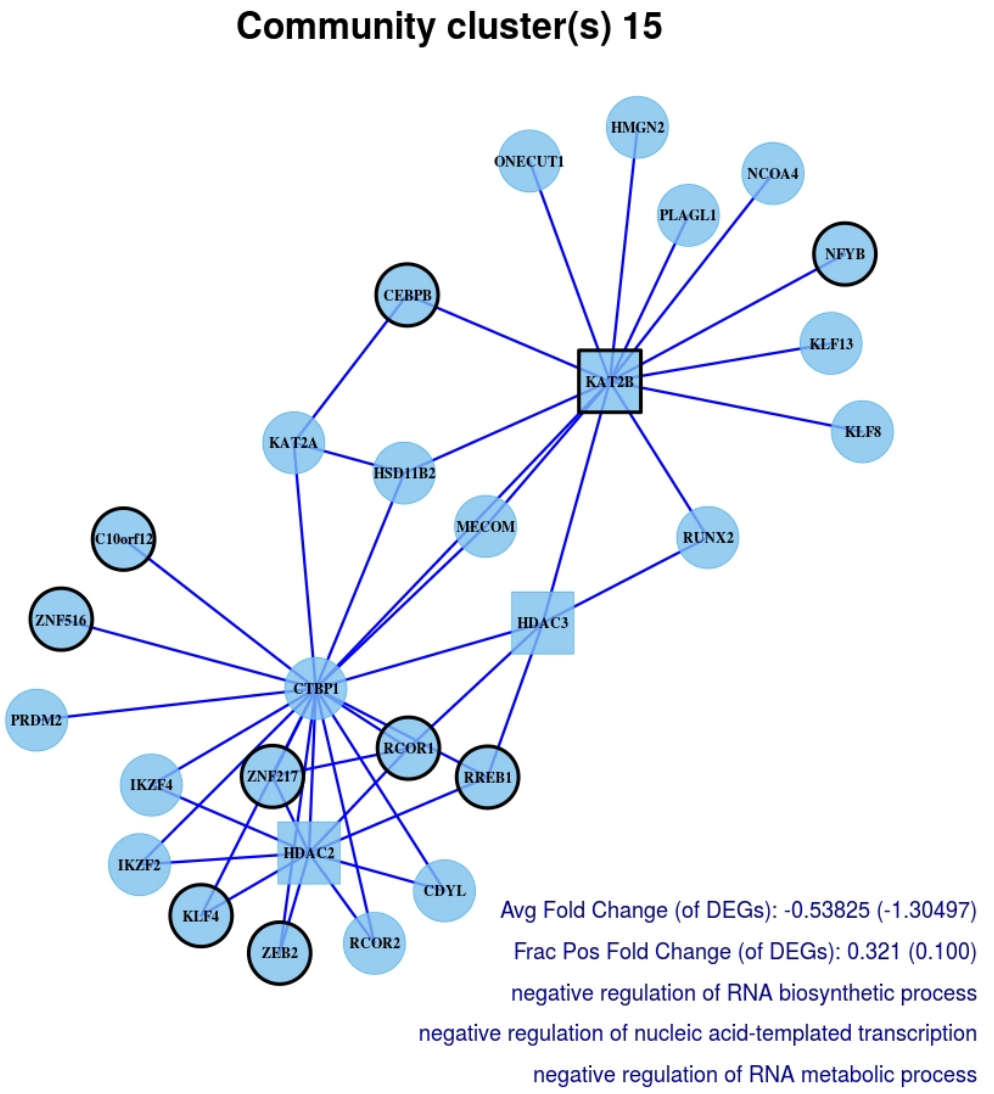

Figure A.7: Epigenetic regulation module. This module is displayed as described in Section 7.3.1. The DEG-enriched module was found by our pipeline using the macrophage differentiation gene expression data set (see Section 3.3.2.2). The module contains mainly genes that are up-regulated in M2 macrophages as shown by negative fold change displayed in the bottom right (Frac Pos Fold Change = Fraction of genes with a positive fold change). The last three lines on the bottom right are GO BP terms that are enriched in the genes in this module. These annotations, and the HDACs and KATs in the module indicate that the module is related to epigenetic regulation.

two PINs, was found to be poor across GO slim network partitions. Further analysis must be performed to investigate whether the networks on which Ledgehop does appear to perform better than link clustering exhibit any structural similarities. Our results suggest that local clustering coefficients may affect the performance of both link clustering and Ledgehop, however the statistical significance of this signal must be further assessed.

The link clustering Ledgehop comparisons carried out in this appendix show Ledgehop outperforming link clustering only for high specificity GO slim partitions of BioGrid-AP. While this network is the most relevant to the pipeline developed in this project, the result does not suffice as evidence that Ledgehop is an improvement over link clustering for PINs in general. Testing more PIN data sets is necessary to draw such a conclusion. Based on the results presented in this appendix we can conclude that Ledgehop partitions networks in a different way to link clustering to generate functional modules that would not be found otherwise. Yet, our results also suggest that Ledgehop cannot replace link clustering for functional module detection as more enriched modules were found based on link clustering communities than based on Ledgehop communities.

The PIN analysis from which we draw most of our conclusions is based on NMI analysis with GO slim partitions. This simple evaluation of the performance of community detection methods is subject to certain issues such as disconnected label communities that may not include the majority of network nodes. Using the connected components of these label communities as in [213] may have improved the evaluation method. To compare Ledgehop and link clustering PIN partitions in more detail, it is also possible to compute the functionally significant communinities generated by these methods using the CommWalker framework and investigate summary statistics as in Section 6.4.4. This type of investigation can give us further insight into the benefits and limitations of Ledgehop community detection.

The additional information Ledgehop uses to attempt to improve link clustering's edge similarity scores, comes at the cost of its speed. Although Ledgehop scales well to currently available PIN sizes, it is considerably slower on some of the larger

networks tested in this appendix. Link clustering and Ledgehop scale with the number of connected edge pairs in the network. Thus, the difference in speed can be best seen on the example of the Texas84 network, which has the largest number of edge pairs of all networks tested here. While link clustering took several days to compute the edge similarity score matrix on this network, Ledgehop required approximately a month on a single core.

B

Data Sets

Contents

B.1 Pre-processing protein-protein interaction data 221
B.2 Microarray probe set mapping 222
B.3 Network statistics 224

This appendix contains additional information on network data sets. Specifically, we elaborate on how protein-protein interaction data from BioGrid and gene expression data from an Affymetrix microarray platform were pre-processed, and give network summary statistics of all social and biological networks used to test Ledgehop community detection in Chapter A.

B.1 Pre-processing protein-protein interaction data

As elaborated on in Section 3.1.1, protein interaction data obtained from several databases were divided into A-type and P-type interactions. As BioGrid [100] does not contain a division into these interaction types, the interaction data was split based on experimental evidence codes. To assign evidence codes to interaction data types we followed a study that divides the BioGrid Yeast interactome into A-type and P-type data [48].

To reflect additions to the evidence codes since the publication of [48], the list of P-type interactions was extended by FRET, and Proximity Label-MS. These experimental methods report binary interactions based on the proximity of two proteins, and thus do not require biochemical binding to occur. The experimental evidence codes "Protein-RNA", "Protein-peptide", and "Co-localization" could not be assigned to either category. Hence, these three evidence codes were not used and interactions which were only supported by these evidence codes were omitted in investigations performed with the BioGrid-AP data set. The low proportion of interactions reported based on these evidence codes make this omission unproblematic. Out of a total of 236,251 interactions quoted by separate publication in the BioGrid data set retrieved in August 2015 (cf. Section 3.1.1), there were 215 "Protein-RNA", 1,507 "Protein-peptide", and 1,547 "Co-localization" evidence codes.

The assignment of evidence codes to interaction data types is summarized in Table B.1.

Table B.1: BioGrid A-type & P-type data split.

A-type	P-type
Affinity Capture - Luminescense	Two-hybrid
Affinity Capture - Western	Reconstituted Complex
Affinity Capture - MS	PCA
Affinity Capture - RNA	Co-crystal structure
Biochemical Activity	FRET
Co-fractionation	Proximity Label-MS
Co-purification	
Far Western	

Division of experimental evidence codes for physical interactions in the BioGrid database into A-type and P-type data.

B.2 Microarray probe set mapping

In microarray experiments gene probes are used to measure the expression levels of a gene (cf. Section 2.6). These probes are in fact probe sets, each consisting of multiple probes which hybridize different parts of the targeted cDNA sequence.

An overall expression score for the probe set is obtained by fitting a statistical model to the individual probe measurements [235].

In the Affymetrix microarray platform used to measure the macrophage whole genome differential expression (cf. Section 3.3.2.2) not all probe sets are designed to the same quality criteria. The quality of a probe set can vary depending on its propensity for unspecific binding of cDNA sequences (cross-hybridization) or the number of probes in the probe-set. These different standards are reflected in their Affymetrix IDs. For example, IDs that end in "_x_at" denote probe sets where some probes are shared between multiple sequences and thus cross-hybridize. These probe set categories are described in more detail in Appendix B of [236].

Here, we built on previous work by collaborators at UCB Pharma to rank probe sets in the order:

1. "_at",
2. "_f_at",
3. "_s_at" and "_g_at",
4. "_x_at",
5. "_r_at",
6. "_i_at".

Using this ordering Affymetrix probe IDs were mapped to Entrez Gene IDs. For example, Entrez ID 5982 was mapped to probe sets 1053_at and 203696_s_at. As the quality of the "_at" probe set data is ranked higher than the quality of the "_s_at" probe set data, we used the differential expression data for the 1053_at probe set to represent the gene with Entrez ID 5982.

When several Affymetrix IDs of the same rank mapped to the same Entrez Gene ID, the most differentially expressed probe set was used.

B.3 Network statistics

In Chapter A we present Ledgehop, a community detection method that is built on link clustering. This community detection method is tested on several social and biological networks, including Facebook networks, other social networks, and PINs. Summary statistics for these networks are shown in Table B.2.

Table B.2: Network statistics.

Network	Nodes	Edges	Density	C	ALC	Comms
American75	6,386	217,662	0.0107	0.164	0.247	-
Amherst41	2,235	90,954	0.0364	0.233	0.315	-
Auburn71	18,448	973,918	0.0057	0.137	0.226	-
Baylor93	12,803	679,817	0.0083	0.155	0.213	-
BC17	11,509	486,967	0.0074	0.144	0.214	-
Berkeley13	22,937	852,444	0.0032	0.114	0.214	-
Bingham82	10,004	362,894	0.0073	0.160	0.227	-
Bowdoin47	2,252	84,387	0.0333	0.216	0.294	-
Brandeis99	3,898	137,567	0.0181	0.164	0.269	-
Brown11	8,600	384,526	0.0104	0.145	0.223	-
BU10	19,700	637,528	0.0033	0.121	0.195	-
Bucknell39	3,826	158,864	0.0217	0.202	0.281	-
Cal65	11,247	351,358	0.0056	0.162	0.233	-
Caltech36	769	16,656	0.0564	0.291	0.429	-
Carnegie49	6,637	249,967	0.0114	0.185	0.287	-
Colgate88	3,482	155,043	0.0256	0.207	0.271	-
Columbia2	11,770	444,333	0.0064	0.129	0.242	-
Cornell5	18,660	790,777	0.0045	0.136	0.225	-
Dartmouth6	7,694	304,076	0.0103	0.151	0.253	-
Duke14	9,895	506,442	0.0103	0.166	0.252	-
Emory27	7,460	330,014	0.0119	0.189	0.263	-
FSU53	27,737	1,034,802	0.0027	0.153	0.222	-
Georgetown15	9,414	425,638	0.0096	0.149	0.231	-
GWU54	12,193	469,528	0.0063	0.139	0.223	-
Hamilton46	2,314	96,394	0.036	0.219	0.302	-
Harvard1	15,126	824,617	0.0072	0.136	0.222	-
Haverford76	1,446	59,589	0.057	0.251	0.327	-
Howard90	4,047	204,850	0.025	0.173	0.233	-
Indiana69	29,747	1,305,765	0.003	0.135	0.208	-
JMU79	14,070	485,564	0.0049	0.131	0.205	-
Johns_Hopkins55	5,180	186,586	0.0139	0.193	0.279	-
Lehigh96	5,075	198,347	0.0154	0.190	0.270	-

Maine59	9,069	243,247	0.0059	0.145	0.246	-
Maryland58	20,871	744,862	0.0034	0.129	0.213	-
Mich67	3,748	81,903	0.0117	0.194	0.291	-
Michigan23	30,147	1,176,516	0.0026	0.133	0.216	-
Middlebury45	3,075	124,610	0.0264	0.211	0.288	-
Mississippi66	10,521	610,911	0.011	0.182	0.252	-
MIT8	6,440	251,252	0.0121	0.180	0.285	-
MSU24	32,375	1,118,774	0.0021	0.122	0.209	-
MU78	15,436	649,449	0.0055	0.152	0.218	-
Northeastern19	13,882	381,934	0.004	0.128	0.207	-
Northwestern25	10,567	488,337	0.0087	0.160	0.245	-
Notre_Dame57	12,155	541,339	0.0073	0.126	0.212	-
NYU9	21,679	715,715	0.003	0.108	0.201	-
Oberlin44	2,920	89,912	0.0211	0.174	0.269	-
Oklahoma97	17,425	892,528	0.0059	0.159	0.235	-
Penn94	41,554	1362,229	0.0016	0.098	0.217	-
Pepperdine86	3,445	152,007	0.0256	0.206	0.285	-
Princeton12	6,596	293,320	0.0135	0.164	0.244	-
Reed98	962	18,812	0.0407	0.221	0.330	-
Rice31	4,087	184,828	0.0221	0.203	0.300	-
Rochester38	4,563	161,404	0.0155	0.198	0.298	-
Rutgers89	24,580	784,602	0.0026	0.139	0.227	-
Santa74	3,578	151,747	0.0237	0.202	0.267	-
Simmons81	1,518	32,988	0.0287	0.212	0.325	-
Smith60	2,970	97,133	0.022	0.193	0.289	-
Stanford3	11,621	568,330	0.0084	0.157	0.253	-
Swarthmore42	1,659	61,050	0.0444	0.227	0.299	-
Syracuse56	13,653	543,982	0.0058	0.171	0.241	-
Temple83	13,686	360,795	0.0039	0.127	0.221	-
Tennessee95	16,979	770,659	0.0053	0.139	0.240	-
Texas80	31,560	1,219,650	0.0024	0.153	0.221	-
Texas84	36,371	1,590,655	0.0024	0.100	0.197	-
Trinity100	2,613	111,996	0.0328	0.228	0.295	-
Tufts18	6,682	249,728	0.0112	0.162	0.235	-
Tulane29	7,752	283,918	0.0095	0.190	0.261	-
UC33	16,808	522,147	0.0037	0.149	0.235	-
UC61	13,746	442,174	0.0047	0.177	0.271	-
UC64	6,833	155,332	0.0067	0.191	0.283	-
UCF52	14,940	428,989	0.0038	0.156	0.238	-
UChicago30	6,591	208,103	0.0096	0.155	0.262	-
UCLA26	20,467	747,613	0.0036	0.143	0.222	-
UConn91	17,212	604,870	0.0041	0.132	0.206	-
UCSB37	14,935	482,224	0.0043	0.157	0.227	-

UCSC68	8,991	224,584	0.0056	0.172	0.244	-
UCSD34	14,948	443,221	0.004	0.150	0.234	-
UF21	35,123	1,465,660	0.0024	0.121	0.225	-
UGA50	24,389	1,174,057	0.0039	0.144	0.214	-
UIllinois20	30,809	1,264,428	0.0027	0.141	0.219	-
UMass92	16,516	519,385	0.0038	0.123	0.214	-
UNC28	18,163	766,800	0.0046	0.116	0.206	-
UPenn7	14,916	686,501	0.0062	0.143	0.221	-
USC35	17,444	801,853	0.0053	0.144	0.221	-
USF51	13,377	321,214	0.0036	0.153	0.241	-
USFCA72	2,682	65,252	0.0181	0.191	0.276	-
UVA16	17,196	789,321	0.0053	0.135	0.215	-
Vanderbilt48	8,069	427,832	0.0131	0.182	0.255	-
Vassar85	3,068	119,161	0.0253	0.177	0.249	-
Vermont70	7,324	191,221	0.0071	0.152	0.236	-
Villanova62	7,772	314,989	0.0104	0.166	0.241	-
Virginia63	21,325	698,178	0.0031	0.105	0.225	-
Wake73	5,372	279,191	0.0194	0.204	0.280	-
WashU32	7,755	367,541	0.0122	0.172	0.267	-
Wellesley22	2,970	94,899	0.0215	0.173	0.269	-
Wesleyan43	3,593	138,035	0.0214	0.195	0.265	-
William77	6,472	266,378	0.0127	0.168	0.258	-
Williams40	2,790	112,986	0.029	0.207	0.296	-
Wisconsin87	23,842	835,952	0.0029	0.120	0.215	-
Yale4	8,578	405,450	0.011	0.151	0.242	-
Amazon	334,864	925,872	1.651×10^{-5}	0.205	0.430	75,149
DBLP	317,080	1,049,866	2.088×10^{-5}	0.306	0.732	13,477
Youtube	1,134,890	2,987,624	4.639×10^{-6}	0.006	0.172	8,385
HINT-P	10,927	49,301	0.0008	0.034	0.099	-
BioGrid-AP	15,405	165,343	0.0014	0.055	0.130	-

Table B.2: C refers to the clustering coefficient, ALC to the average local clustering coefficient (see Section 2.2.1), and Comms to the number of ground truth communities of size > 2 (only available for SNAP networks). The networks are split into three groups: The first 100 networks are parts of the Facebook100 data set, the following three are networks from SNAP, and the final two networks are PINs previously used in Chapters 4, 6, and 7. Facebook networks tend to have the highest densities and the SNAP networks have the most nodes and lowest densities, but the highest level of local clustering.

C

Disconnected communities in the Louvain algorithm

Communities are substructures in a network which interact more with each other than with the rest of the network. Thus, a fundamental characteristic of a community is that it is connected. Investigations of the communities generated by the Louvain implementation used (see Section 3.1.2) with both the configuration and Constant Potts null model showed a number of disconnected communities. To rule out bugs in the specific implementation of the algorithm, this observation was confirmed using the GenLouvain algorithm (community function in GenLouvain version 1.2, retrieved April 2014), which is a generalized Louvain implementation for Modularity Maximization [237].

Further investigation into individual disconnected communities showed that many could be reconnected by a single node. One such example is shown in Figure C.1.

Disconnected communities can trivially be shown to be non-optimal solutions to Modularity Maximization, as splitting such a community into its connected components would increase the Modularity of the partition (see Section 2.3.2). As two implementations of the Louvain algorithm independently reproduced this behaviour, it can be assumed that the issue of disconnected communities is inherent to the method rather than the individual implementations. Personal communication

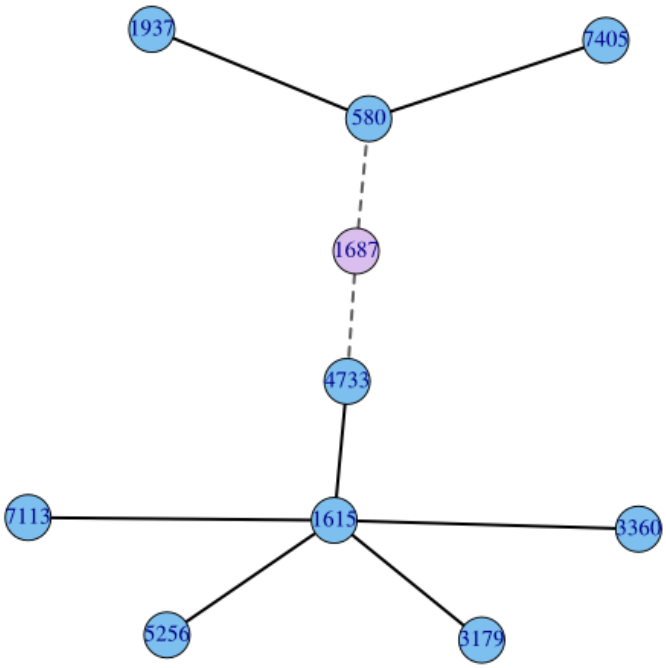

Figure C.1: Disconnected community generated by the Louvain algorithm. This community was generated by Louvain configuration model Modularity Maximization on HINT-P-14. The community was obtained at the resolution parameter, $\gamma = 0.251$. Nodes in blue represent the nodes in the community. Node 1687 (purple), which connects the two disconnected components, has degree 138. The only neighbour node of the community which is not shown, is a degree 199 hub node, interacting with node 4733.

with Dr. Vincent Traag, author of one of the implementations, confirmed this conclusion by testing the calculations performed by his implementation.

An explanation for this behaviour is likely to be found within the ordering of nodes in the first iteration of the Modularity optimization step of the algorithm (see Section 2.3.2.3). When a node is considered, it is assigned to a community with the neighbouring node that maximizes the overall Modularity. If several nodes are placed into a community with a central hub connecting all of these nodes, and subsequently the hub is removed as this step would maximize the Modularity, disconnected communities can be generated. As there is a heavy Modularity penalty for removing the hub node which scales with the number of separate connected components remaining, disconnected communities generally only contain two connected components.

In light of these connectivity issues, an updated version of the Louvain implementation that was received from Dr. Vincent Traag was used in this chapter. The

updated algorithm splits all disconnected communities formed after each Modularity optimization step to ensure no disconnected communities are obtained.

D

CommWalker Results

Contents

D.1 T-Value stability on BioGrid-AP 231
D.2 Coverage of Accepted modules 232
D.3 Module Statistics 233
D.4 Computational Module Validation 258

In this appendix we present additional results generated to analyse the efficacy of the CommWalker framework (cf. Chapter 6) on different data sets and to analyse the stability of T-values.

D.1 T-Value stability on BioGrid-AP

We investigated the stability of the T-values of nine randomly selected communities of size ≤ 35 from HINT-P in Section 6.3. The standard errors of the T-values of these communities were calculated over 100 CommWalker runs at varying node counts (Equation 6.2). This investigation showed an optimal trade-off between algorithm run time (proportional to node count) and T-value stability at a node count of 100,000. At this node count, HINT-P T-values should be assumed to have an error of ≈ 0.005. Here, we repeated this investigation with nine randomly selected communities from BioGrid-AP. Figure D.1 shows that the T-value stability behaves

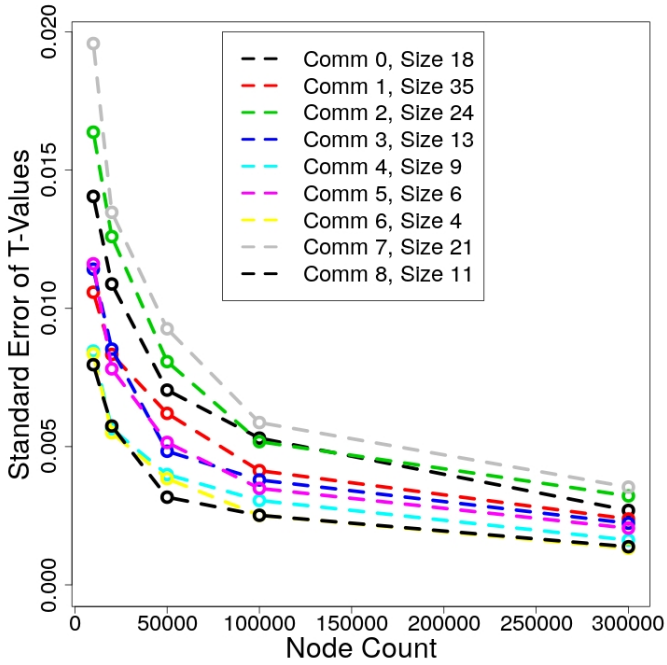

Figure D.1: BioGrid-AP T-value stability. The T-value stability analysis shown in Figure 6.3 was repeated for nine randomly selected communities of sizes ≤ 35 on BioGrid-AP. The number of random walks started at each node is calculated from the node count by Equation (6.2). The trade-off between algorith run time and node count is found to be optimal at a node count of 100,000 as in HINT-P. At this node count the standard errors of the nine communities lie in the range $[0.002, 0.007]$, which is consistent with the HINT-P results.

similarly on BioGrid-AP. At an optimal node count of 100,000, the investigated BioGrid-AP T-value standard errors lie in the range $\sigma(T_C) \in [0.002, 0.007]$ which is consistent with HINT-P results.

D.2 Coverage of Accepted modules

To investigate the efficacy of the CommWalker framework, we compared the communities that were evaluated as functionally significant by CommWalker and functional homogeneity without CommWalker. Here, functional significance was defined by the qualitatively similar thresholds discussed in Section 6.4.1. Specifically, the number of distinct proteins in these functionally significant communities was analysed across resolutions and data sets. The results to complement the analysis in Section 6.4.3 are shown in Figures D.2–D.5 for the Pandey measure, Figures D.6 and D.7 for simGIC, and Figures D.8 and D.9 for simUI.

These figures show that there tend to be more unique proteins in CommWalker accepted communities than in those accepted by functional homogeneity using the Pandey measure. In the case of simGIC this trend is reversed and the numbers are similar for simUI. This observation is expected given that the functional homogeneity threshold is more lenient for simGIC and simUI, as argued in Section 6.4.1.

D.3 Module Statistics

To ascertain whether the increase in coverage of functionally significant modules on the PINs was indeed due to an improved sensitivity in poorly-studied PIN regions, we analysed the communities accepted as functional modules by CommWalker versus functional homogeneity. As in Section D.2, acceptance as a functional module is determined by the qualitatively similar thresholds defined in Section 6.4.1. Using these thresholds, communities were divided into four sets: accepted by both methods, accepted only by CommWalker, accepted only by functional homogeneity, and rejected by both methods. Average network statistics were computed on these community sets to characterize the differences between the evaluation methods. The results presented in Figures D.10–D.13 complement those shown in Figures 6.6 and 6.7 in Section 6.4.4.

The network statistics used to analyse the community sets are the average community size, the average annotation of the communities' local environments (average random walk annotation fraction, cf. Section 6.2), and the average annotation of the communities (average community annotation fraction). As can be seen in the upper row of graphs in Figures D.10–D.13 some community sets make up only a very small fraction of the total communities. As such, the calculated statistics become very variable and thus unreliable to interpret. The small sample size for communities only accepted by CommWalker is specifically an issue for the simGIC data sets in Figures D.10 and D.11. To overcome this issue we focus our interpretation of the data on results at high resolutions where the network statistics are less variable.

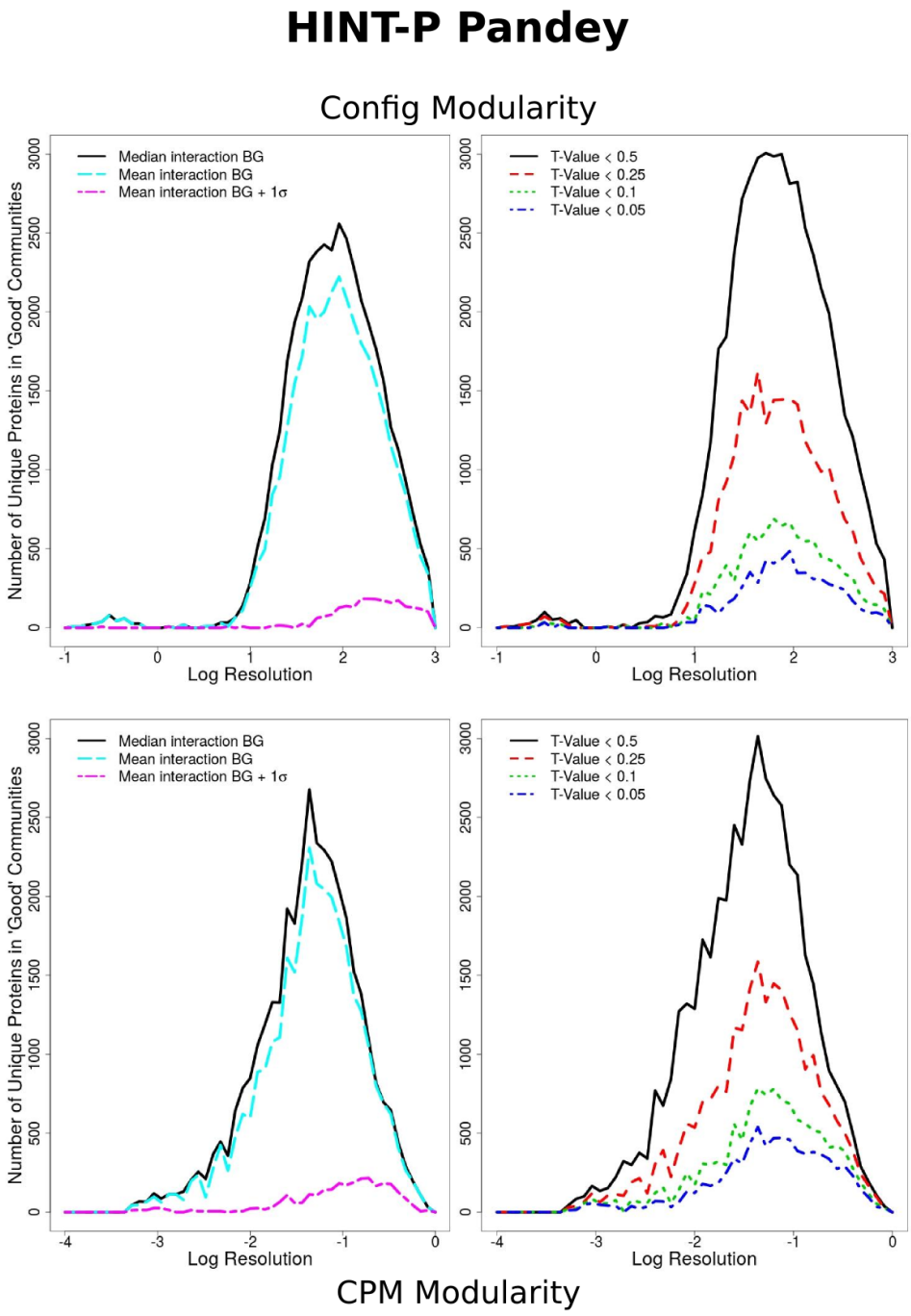

Figure D.2: HINT-P Modularity Maximization Pandey measure CommWalker coverage comparison. Comparison between the number of unique proteins in functionally significant communities of size 6 - 35 by functional homogeneity (left hand plots), and T-value thresholds (right hand plots) using the Pandey measure. The two cases can be compared for the qualitatively similar thresholds of a T-value of 0.5 and the median functional similarity of interacting proteins (Median Interaction BG – black, solid lines). Communities were generated by configuration model Modularity Maximization (Config Modularity) and Constant Potts model Modularity Maximization (CPM Modularity) community detection methods applied to HINT-P. The number of unique proteins in T-value-significant communities is consistently higher than in functional-homogeneity-significant communities for qualitatively similar thresholds. Using the mean functional similarity of interacting proteins (Mean interaction BG), the mean interaction BG with an added standard error of the interacting proteins' functional similarity scores (Mean interaction BG $+1\sigma$), and different T-values, we further highlight the number of unique proteins at different T-value and functional homogeneity thresholds.

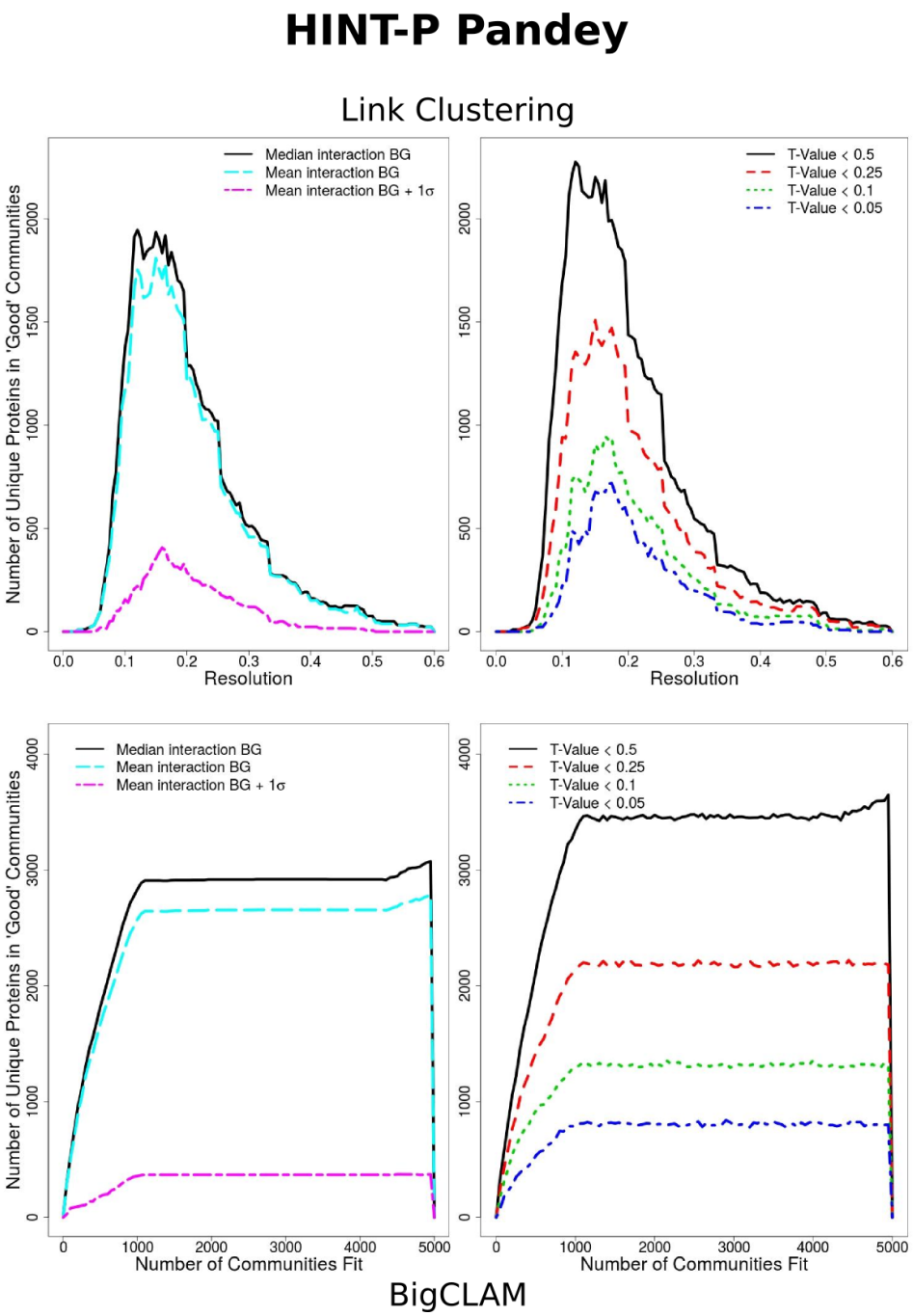

Figure D.3: HINT-P overlapping community detection Pandey measure CommWalker coverage comparison. Comparison between the number of unique proteins in functionally significant communities of size 6 - 35 by functional homogeneity (left hand plots), and T-value thresholds (right hand plots) using the Pandey measure. The two cases can be compared for the qualitatively similar thresholds of a T-value of 0.5 and the median functional similarity of interacting proteins (Median Interaction BG – black, solid lines). Communities were generated by the Link clustering and BigCLAM community detection methods applied to HINT-P. The number of unique proteins in T-value-significant communities is consistently higher than in functional-homogeneity-significant communities for qualitatively similar thresholds. Using the mean functional similarity of interacting proteins (Mean interaction BG), the mean interaction BG with an added standard error of the interacting proteins' functional similarity scores (Mean interaction BG $+1\sigma$), and different T-values, we further highlight the number of unique proteins at different T-value and functional homogeneity thresholds. BigCLAM results increase monotonically in the number of communities fit as neighbouring "resolutions" only differ in the additional 500 communities fit to the data.

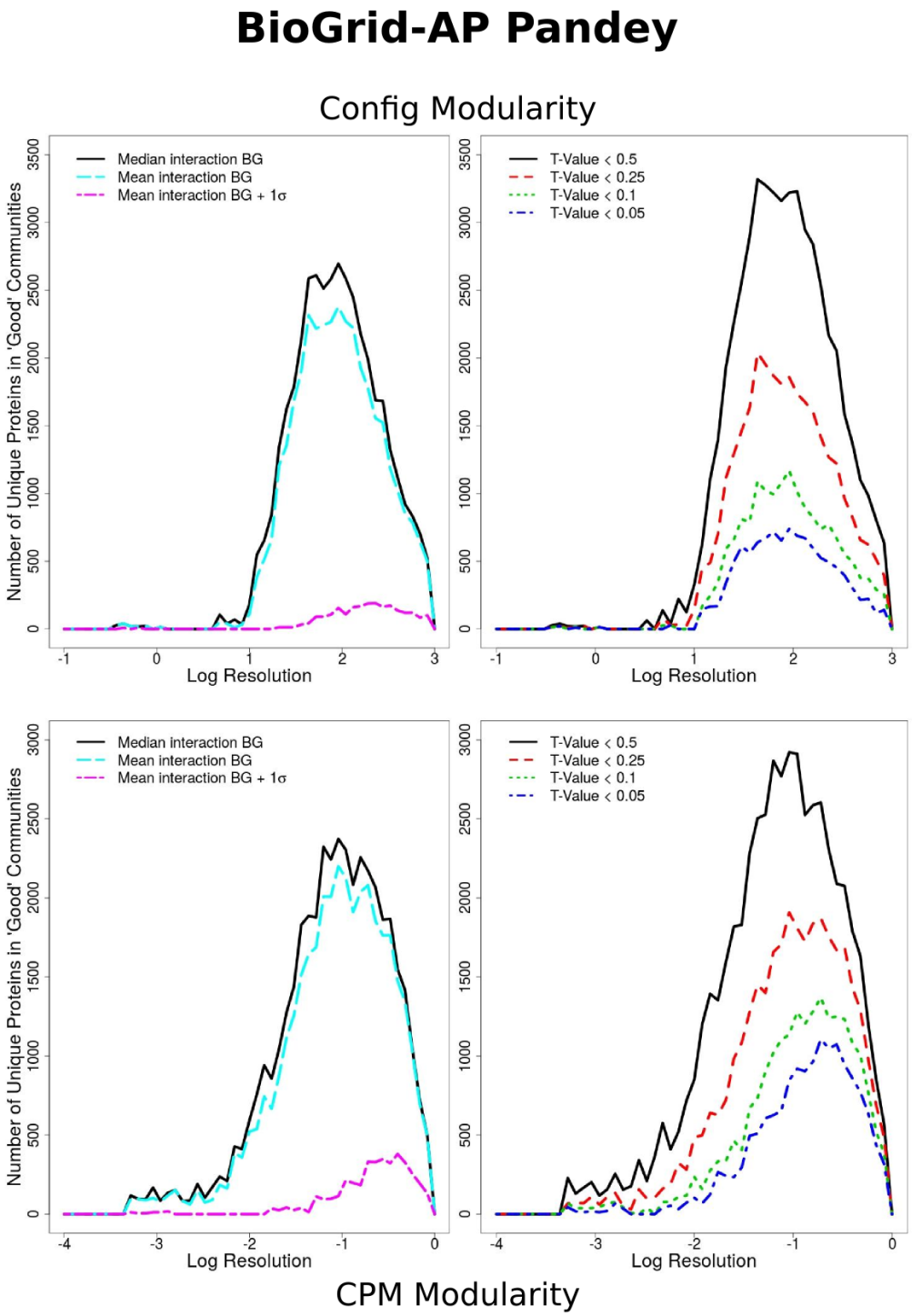

Figure D.4: BioGrid-AP Modularity Maximization Pandey measure CommWalker coverage comparison. Comparison of the number of unique proteins in functionally significant communities of size 6 - 35 by T-value and functional homogeneity. The data was generated as in Figure D.2 using the Pandey measure with BioGrid-AP. The number of unique proteins in T-value-significant communities is consistently higher than in functional-homogeneity-significant communities for the qualitatively similar thresholds indicated by the black lines. Using the mean functional similarity of interacting proteins (Mean interaction BG), the mean interaction BG with an added standard error of the interacting proteins' functional similarity scores (Mean interaction BG $+1\sigma$), and different T-values, we further highlight the number of unique proteins at different T-value and functional homogeneity thresholds.

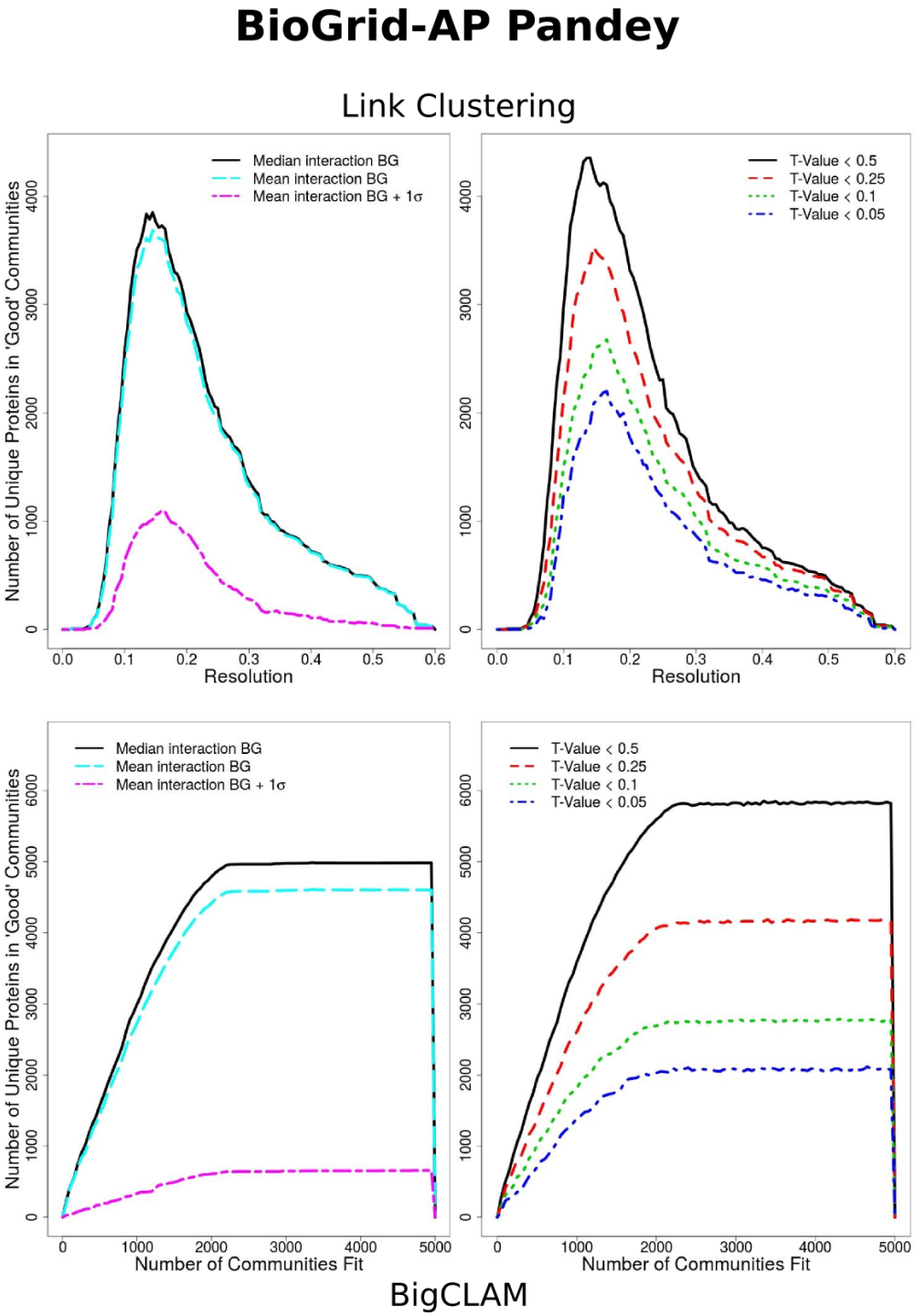

Figure D.5: BioGrid-AP overlapping community detection Pandey measure CommWalker coverage comparison. Comparison of the number of unique proteins in functionally significant communities of size 6 - 35 by T-value and functional homogeneity. The data was generated as in Figure D.3 using the Pandey measure with BioGrid-AP. The number of unique proteins in T-value-significant communities is consistently higher than in functional-homogeneity-significant communities for the qualitatively similar thresholds indicated by the black lines. Using the mean functional similarity of interacting proteins (Mean interaction BG), the mean interaction BG with an added standard error of the interacting proteins' functional similarity scores (Mean interaction BG $+1\sigma$), and different T-values, we further highlight the number of unique proteins at different T-value and functional homogeneity thresholds.

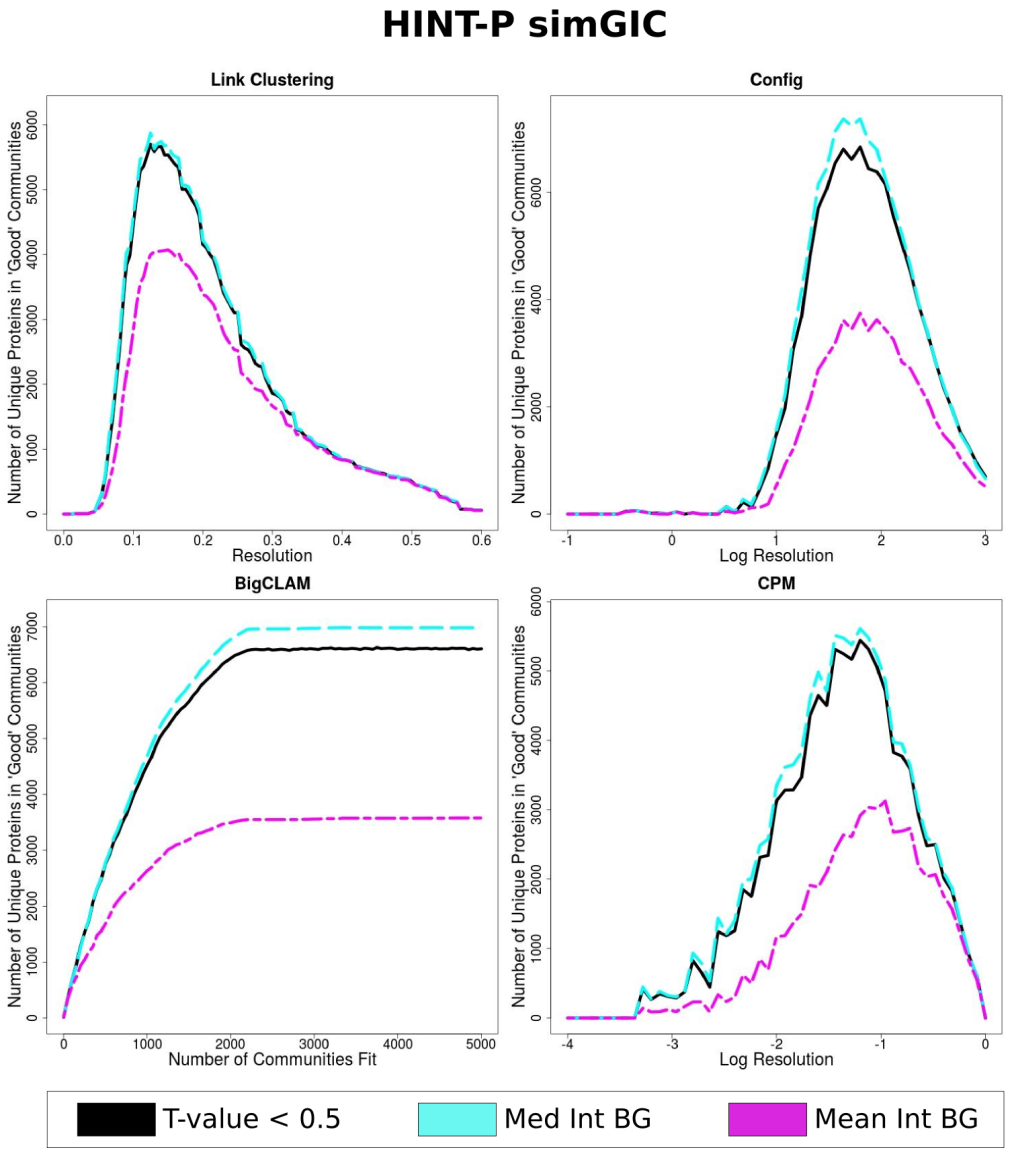

Figure D.6: HINT-P simGIC CommWalker coverage comparison. Comparison of the number of unique proteins in functionally significant communities of size 6 - 35 by T-value and functional homogeneity. The data was generated as in Figures D.2 and D.3 using the simGIC semantic similarity measure on HINT-P. Using simGIC, the number of unique proteins in T-value-significant communities tends to be slightly lower than the number in functional-homogeneity-significant communities for the qualitatively similar thresholds indicated by the black and cyan lines.

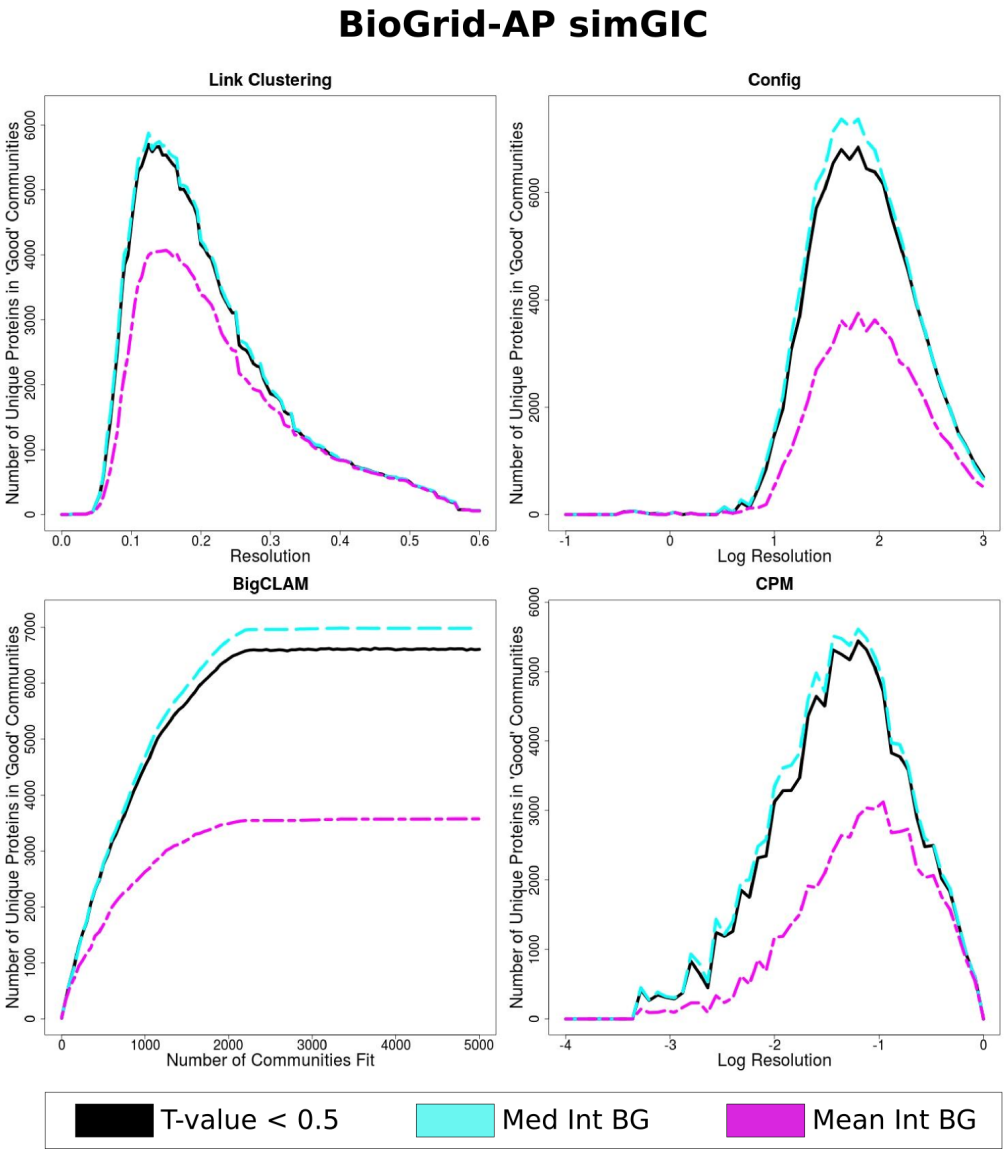

Figure D.7: BioGrid-AP simGIC CommWalker coverage comparison. Comparison of the number of unique proteins in functionally significant communities of size 6 - 35 by T-value and functional homogeneity. The data was generated as in Figures D.2 and D.3 using the simGIC semantic similarity measure on BioGrid-AP. Using simGIC, the number of unique proteins in T-value-significant communities tends to be slightly lower than the number in functional-homogeneity-significant communities for the qualitatively similar thresholds indicated by the black and cyan lines.

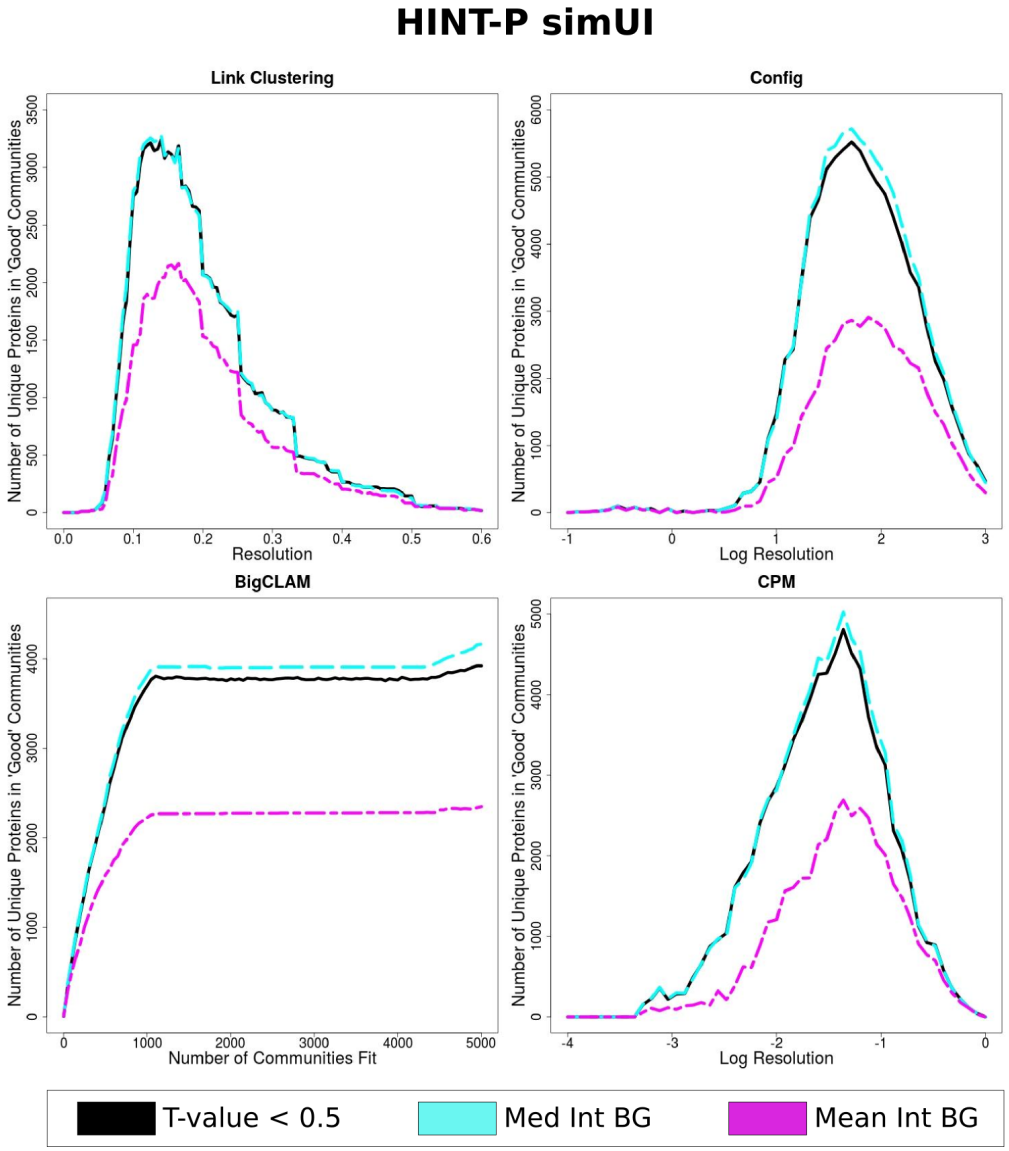

Figure D.8: HINT-P simUI CommWalker coverage comparison. Comparison of the number of unique proteins in functionally significant communities of size 6 - 35 by T-value and functional homogeneity. The data was generated as in Figures D.2 and D.3 using the simUI semantic similarity measure on HINT-P. Using simUI, the number of unique proteins in T-value-significant communities is similar to the number in functional-homogeneity-significant communities for the qualitatively similar thresholds indicated by the black and cyan lines.

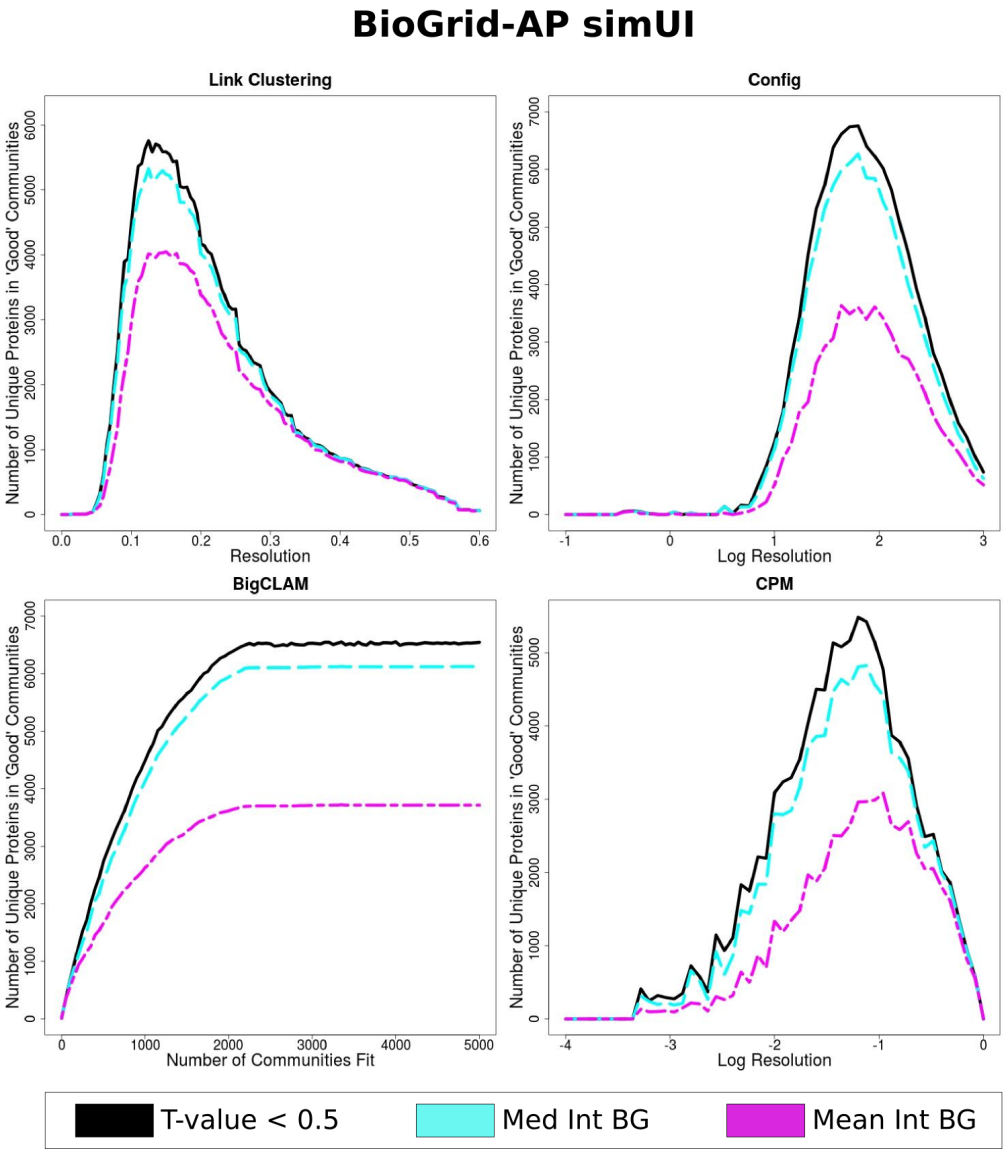

Figure D.9: BioGrid-AP simUI CommWalker coverage comparison. Comparison of the number of unique proteins in functionally significant communities of size 6 - 35 by T-value and functional homogeneity. The data was generated as in Figures D.2 and D.3 using the simUI semantic similarity measure on BioGrid-AP. Using simUI, the number of unique proteins in T-value-significant communities is similar to the number in functional-homogeneity-significant communities for the qualitatively similar thresholds indicated by the black and cyan lines.

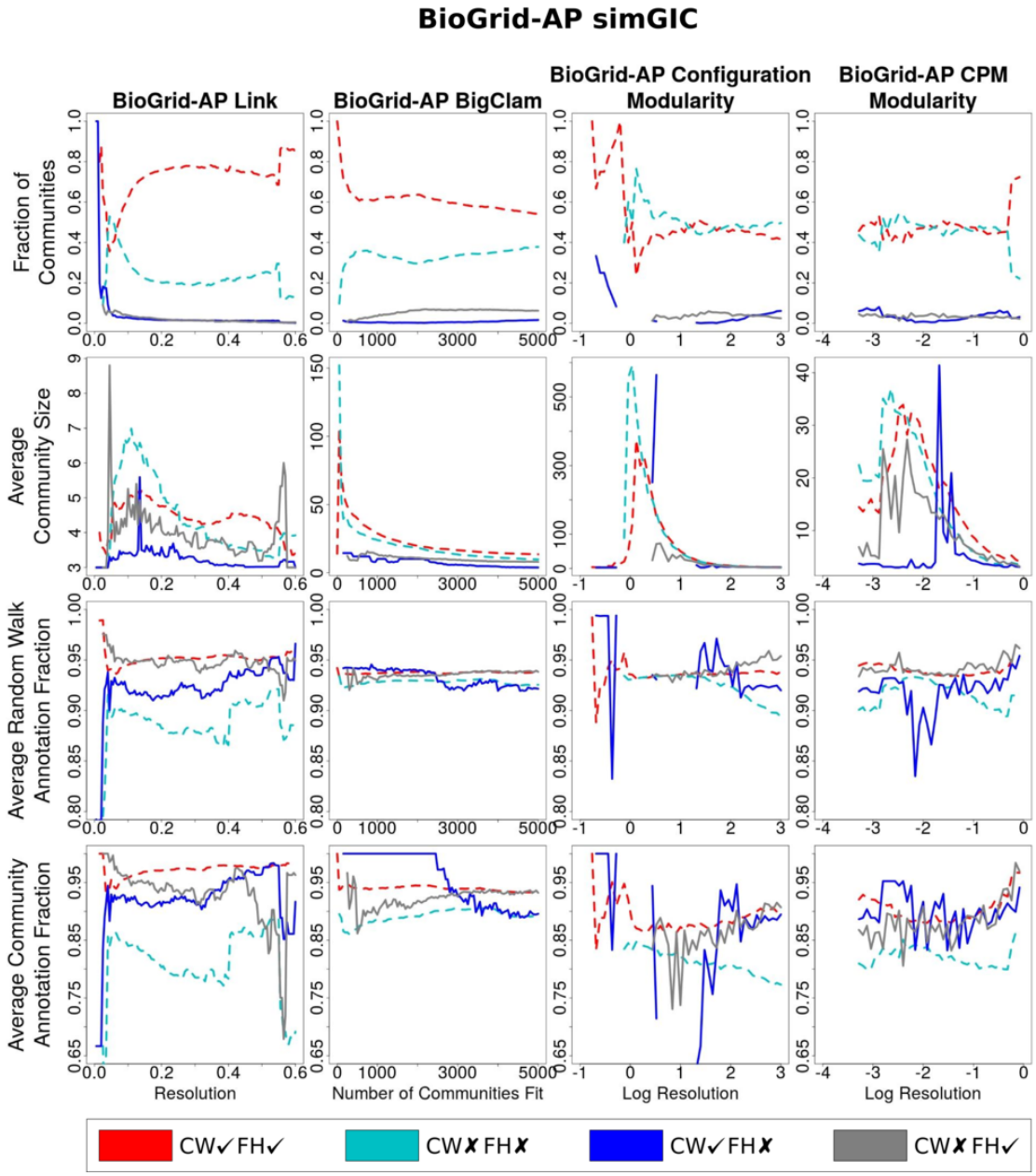

Figure D.10: BioGrid-AP simGIC community summary statistics. Community statistics for BioGrid-AP communities generated by four multi-resolution community detection methods. The data shown was generated as in Figure 6.6 using simGIC and suggests that functional homogeneity selects for smaller communities in well-annotated environments which also have a high level of annotation themselves. In contrast, T-value significant communities tend to have a broad distribution in the investigated statistics as seen by the lines representing CommWalker accepted communities (red dashed line and blue line).

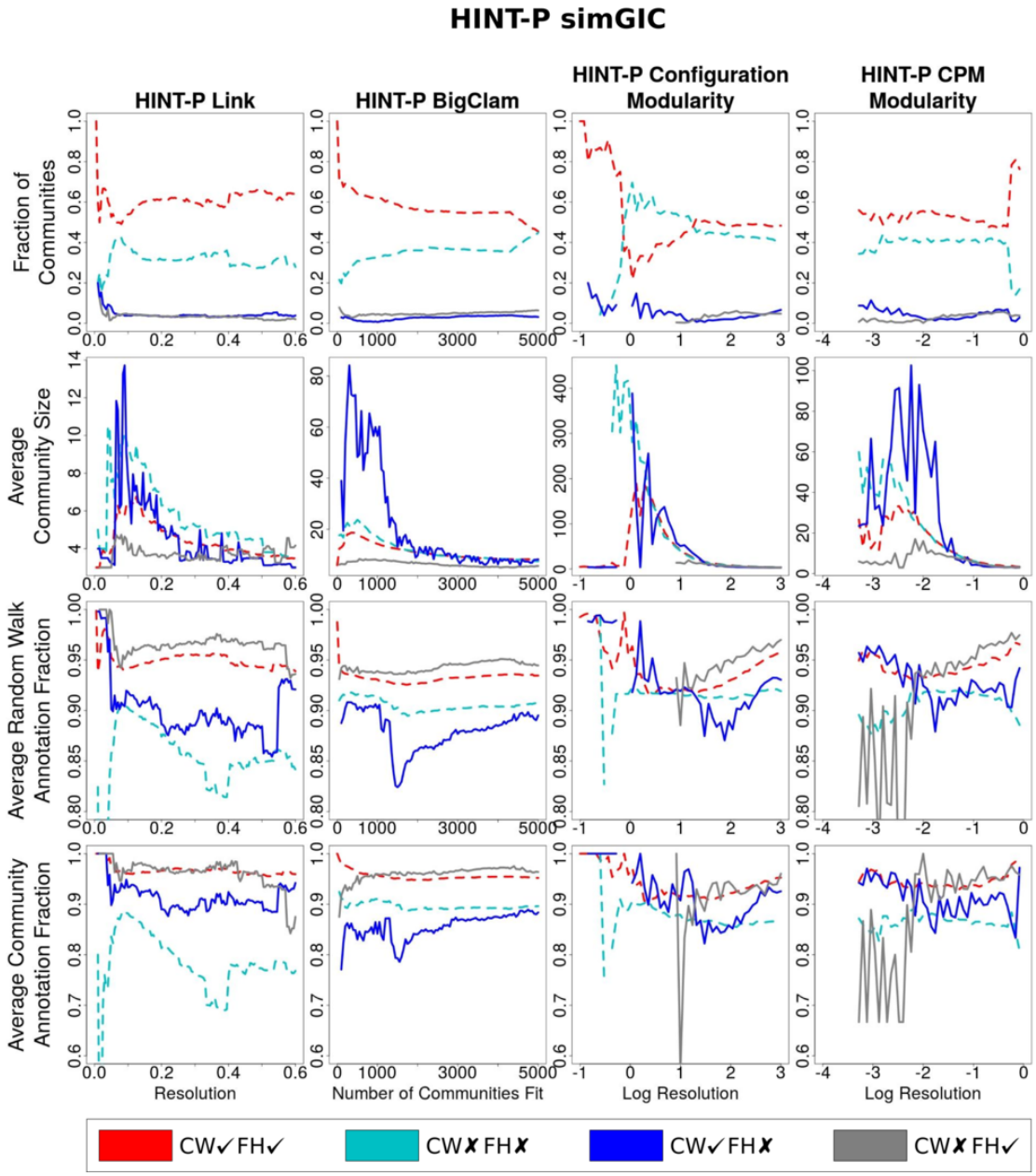

Figure D.11: HINT-P simGIC community summary statistics. Community statistics for HINT-P communities generated by four multi-resolution community detection methods. The data shown was generated as in Figure 6.6 using simGIC and suggests that functional homogeneity selects for smaller communities in well-annotated environments which also have a high level of annotation themselves. In contrast, T-value significant communities tend to have a broad distribution in the investigated statistics as seen by the lines representing CommWalker accepted communities (red dashed line and blue line).

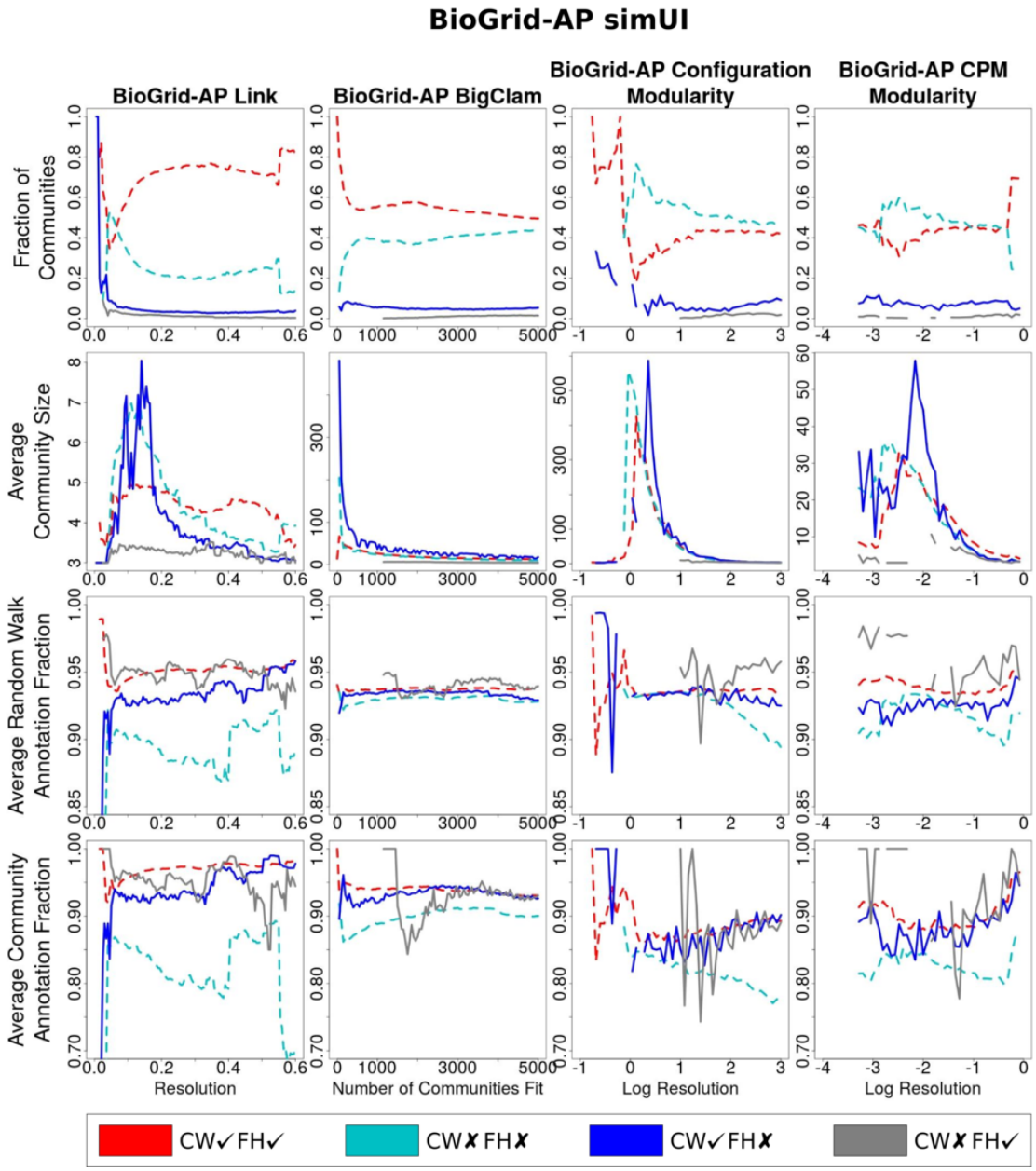

Figure D.12: BioGrid-AP simUI community summary statistics. Community statistics for BioGrid-AP communities generated by four multi-resolution community detection methods. The data shown was generated as in Figure 6.6 using simUI and suggests that functional homogeneity selects for smaller communities in well-annotated environments which also have a high level of annotation themselves. In contrast, T-value significant communities tend to have a broad distribution in the investigated statistics as seen by the lines representing CommWalker accepted communities (red dashed line and blue line).

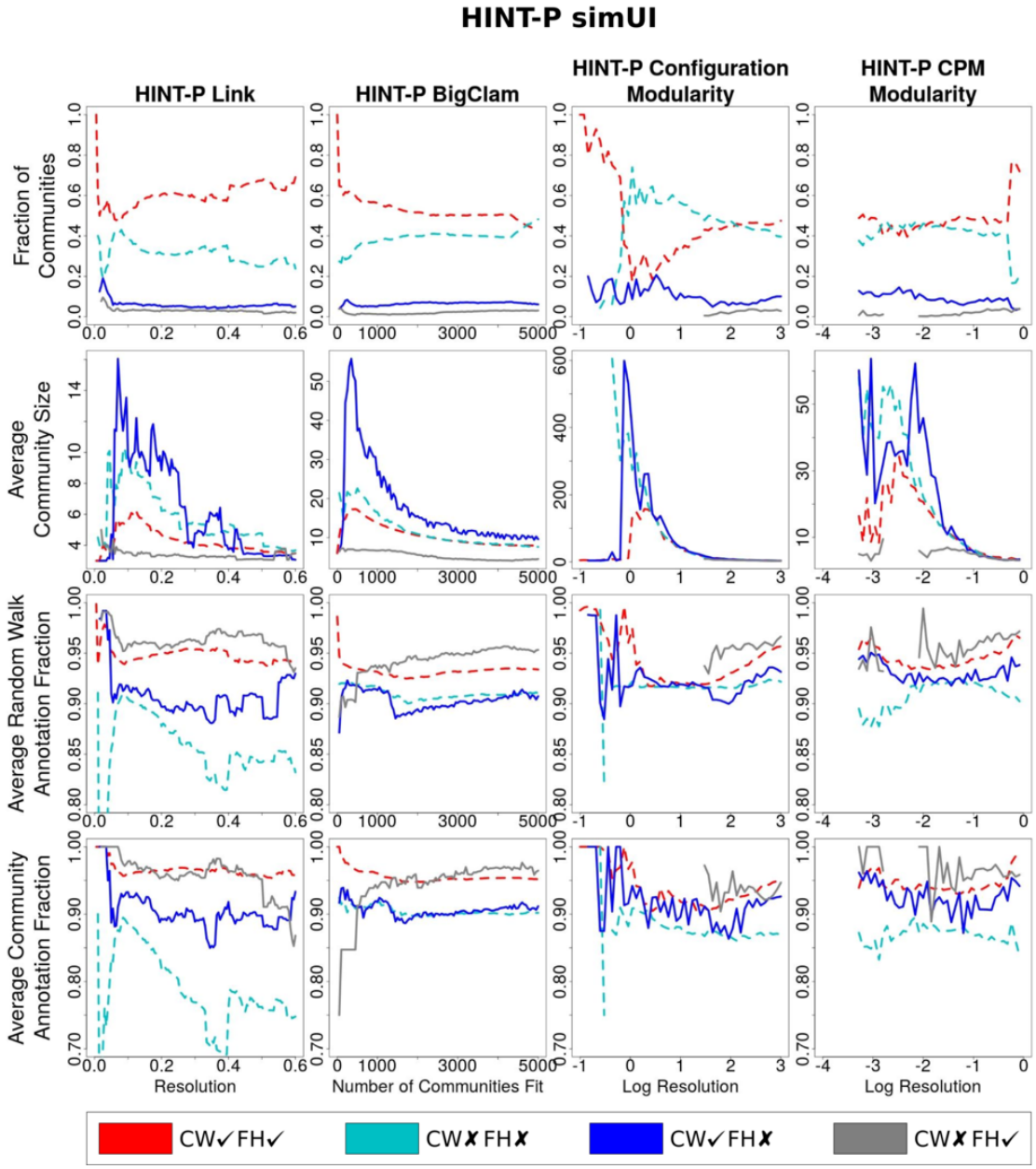

Figure D.13: HINT-P simUI community summary statistics. Community statistics for HINT-P communities generated by four multi-resolution community detection methods. The data shown was generated as in Figure 6.6 using simUI and suggests that functional homogeneity selects for smaller communities in well-annotated environments which also have a high level of annotation themselves. In contrast, T-value significant communities tend to have a broad distribution in the investigated statistics as seen by the lines representing CommWalker accepted communities (red dashed line and blue line).

Taking the small sample size issues into consideration, the data presented shows that functional homogeneity selects for smaller, well-annotated communities in well-annotated environments as compared to CommWalker. In contrast, communities accepted only by CommWalker show a broader distribution in these summary statistics. The broad distribution in the average annotation of the local network environment for CommWalker accepted communities is further visualized in Figures D.14–D.24.

Figures D.14–D.24 were generated in the same way as Figure 6.8 in Section 6.4.4 using different combinations of PIN, community detection method, and functional similarity measure. In these figures non-overlapping community data generated by configuration model and Constant Potts model Modularity Maximization on HINT-P and BioGrid-AP was used to show the distribution of proteins in accepted modules on the PINs for the three semantic similarity measures. In Figures D.14–D.24 the proteins are ordered by their functional similarity with their "vicinity", measured using random walks as described in Section 6.2. Proteins towards the left have higher similarity with their environment and will thus tend to be better studied (cf. investigation in Section 6.2). On this layout we show the distribution of proteins in communities that were accepted as modules by both methods (row b), only by CommWalker (row c), or only by functional homogeneity (row d). Figures D.14–D.24 show that across all data sets proteins in modules accepted by the standard functional homogeneity approach (rows b,d) tend to be distributed towards the well-studied left side of the figure. In contrast, modules accepted only by CommWalker (row c) reach further into the poorly-studied protein regions.

As the thresholds chosen for T-value and functional homogeneity are less similar for simGIC than for the Pandey measure and simUI, Figures D.17–D.20 show a distribution of proteins not seen in other data sets. In these figures proteins in CommWalker only accepted communities are rarely found in high functional similarity environments. This feature occurs as simGIC functional homogeneity accepts so many communities by the more lenient functional homogeneity threshold that the communities otherwise only accepted by CommWalker are now accepted by

both methods. Our conclusion that CommWalker accepted modules reach further into the poorly-studied protein regions are confirmed by this feature.

Due to ease of visualization only non-overlapping community data was used for this investigation, however judging from the module statistics displayed in Figures 6.6,6.7, and D.10–D.13 it can be concluded that CommWalker is equally successful in the overlapping case.

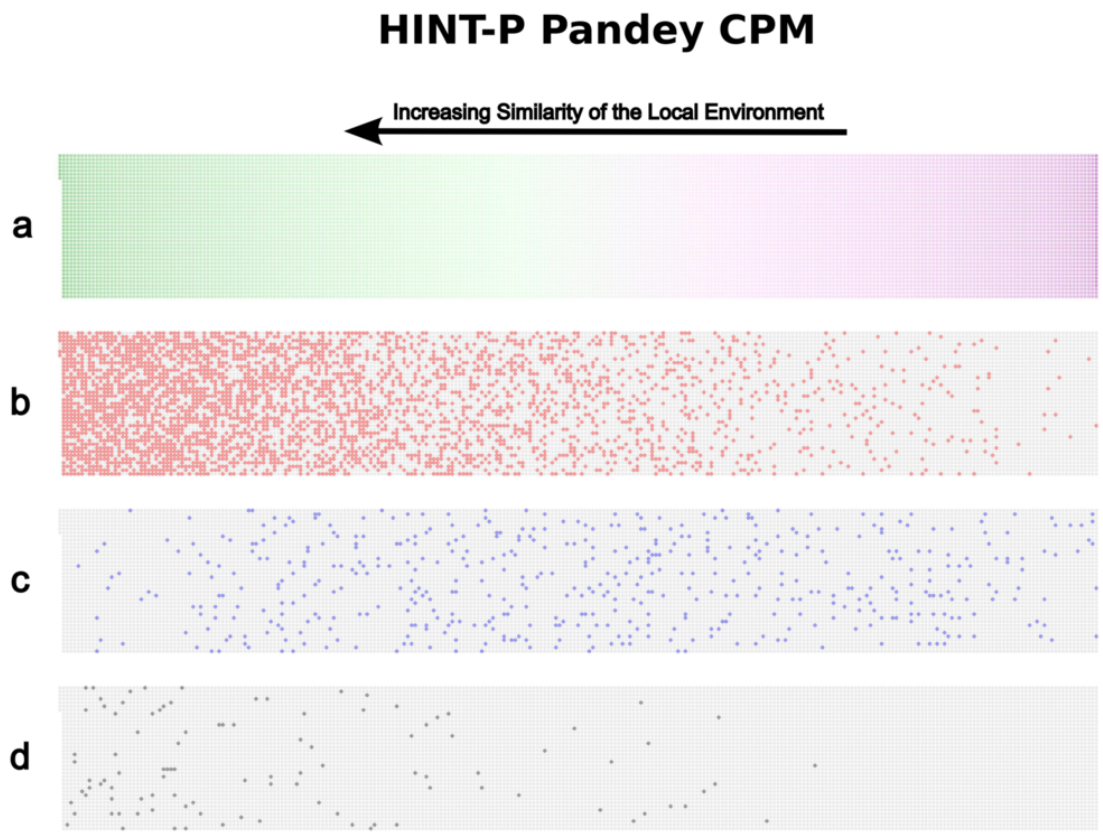

Figure D.14: Local environment comparison of functionally significant HINT-P CPM community proteins by Pandey. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. Nodes in HINT-P were ordered by their Pandey functional similarity with nodes in their vicinity as shown in (a). Communities were generated by Constant Potts model Modularity Maximization at the resolution where the maximum number of proteins are found in functionally significant communities by a T-value threshold of 0.5 (log resolution = −1.36, cf. Figure D.2). On this network layout the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

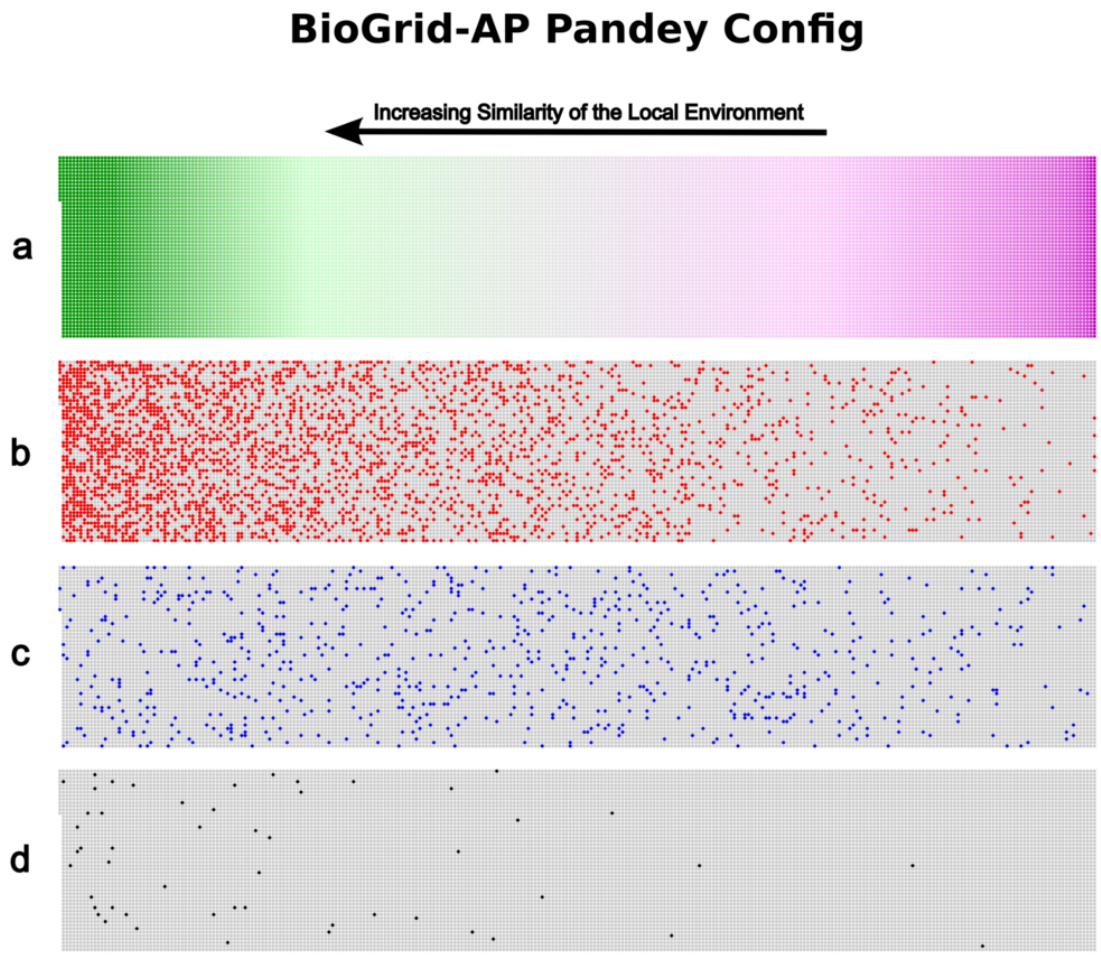

Figure D.15: Local environment comparison of functionally significant BioGrid-AP Config community proteins by Pandey. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using configuration model Modularity Maximization on BioGrid-AP with the Pandey measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

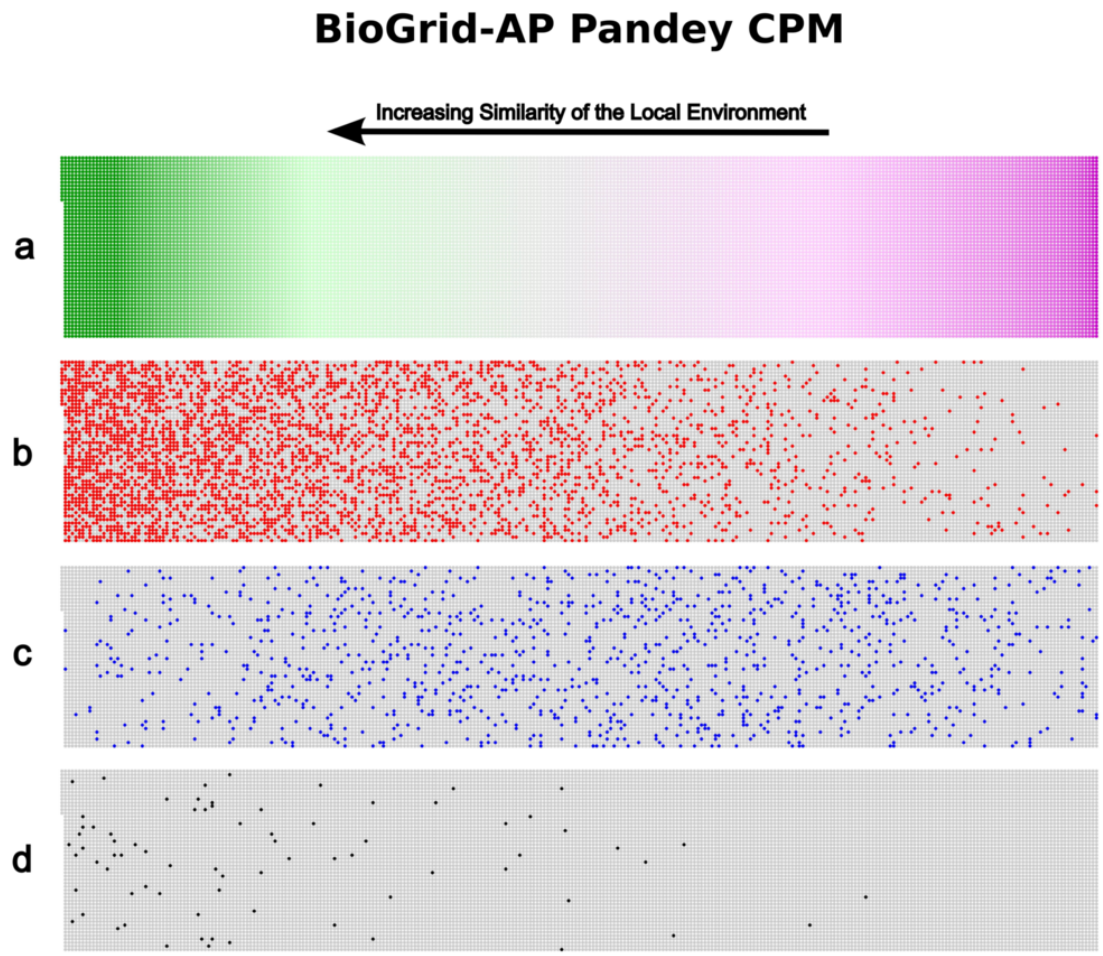

Figure D.16: Local environment comparison of functionally significant BioGrid-AP CPM community proteins by Pandey. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using Constant Potts model Modularity Maximization on BioGrid-AP with the Pandey measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

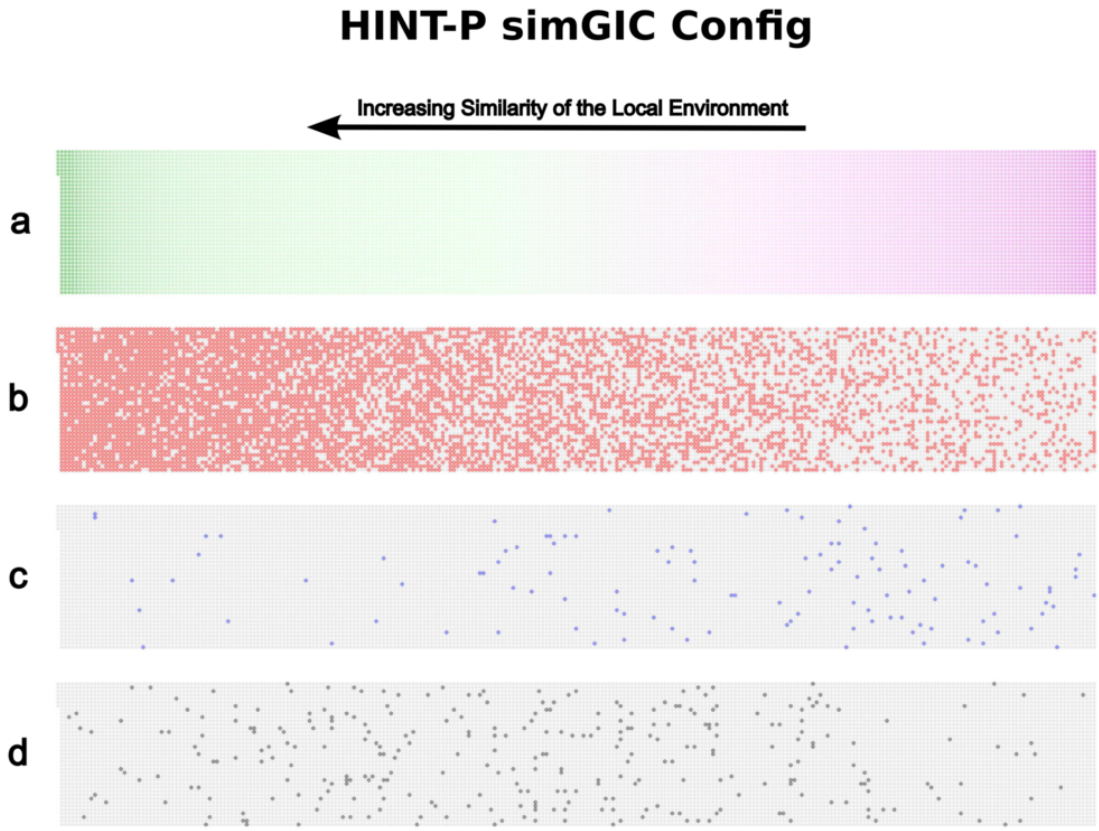

Figure D.17: Local environment comparison of functionally significant HINT-P Config community proteins by simGIC. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using configuration model Modularity Maximization on HINT-P with the simGIC functional similarity measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

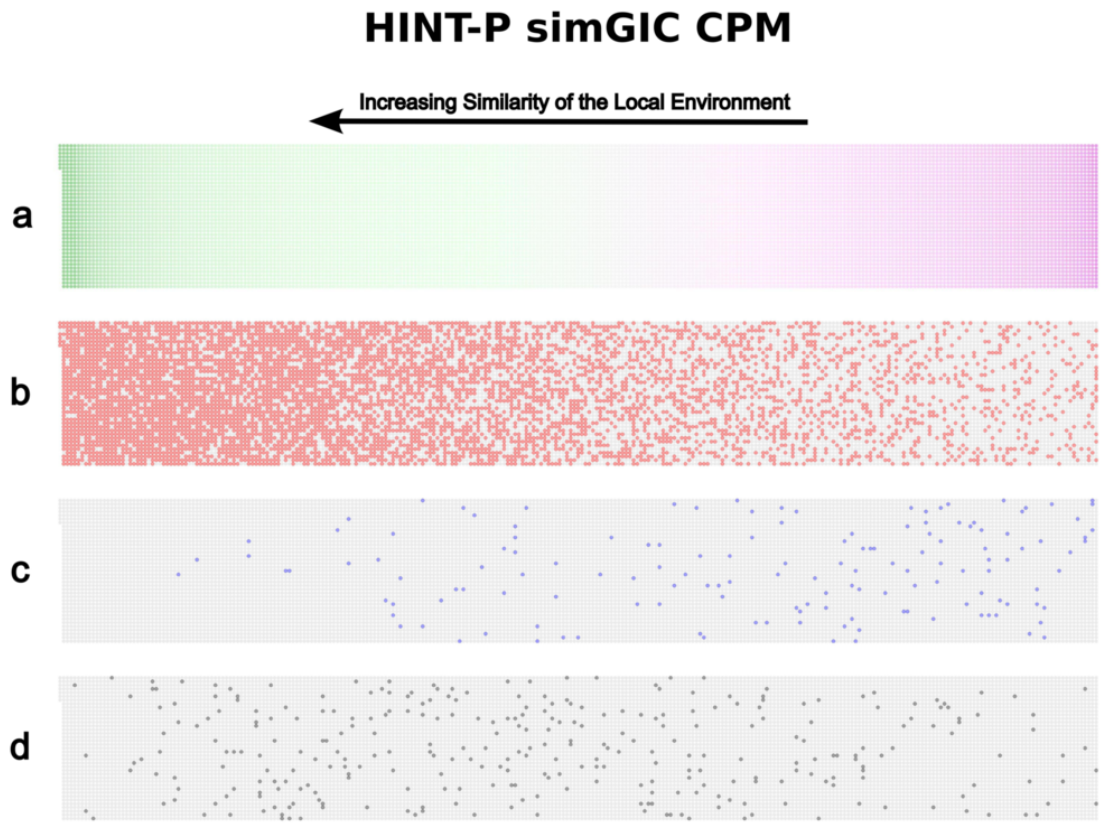

Figure D.18: Local environment comparison of functionally significant HINT-P CPM community proteins by simGIC. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using Constant Potts model Modularity Maximization on HINT-P with the simGIC functional similarity measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

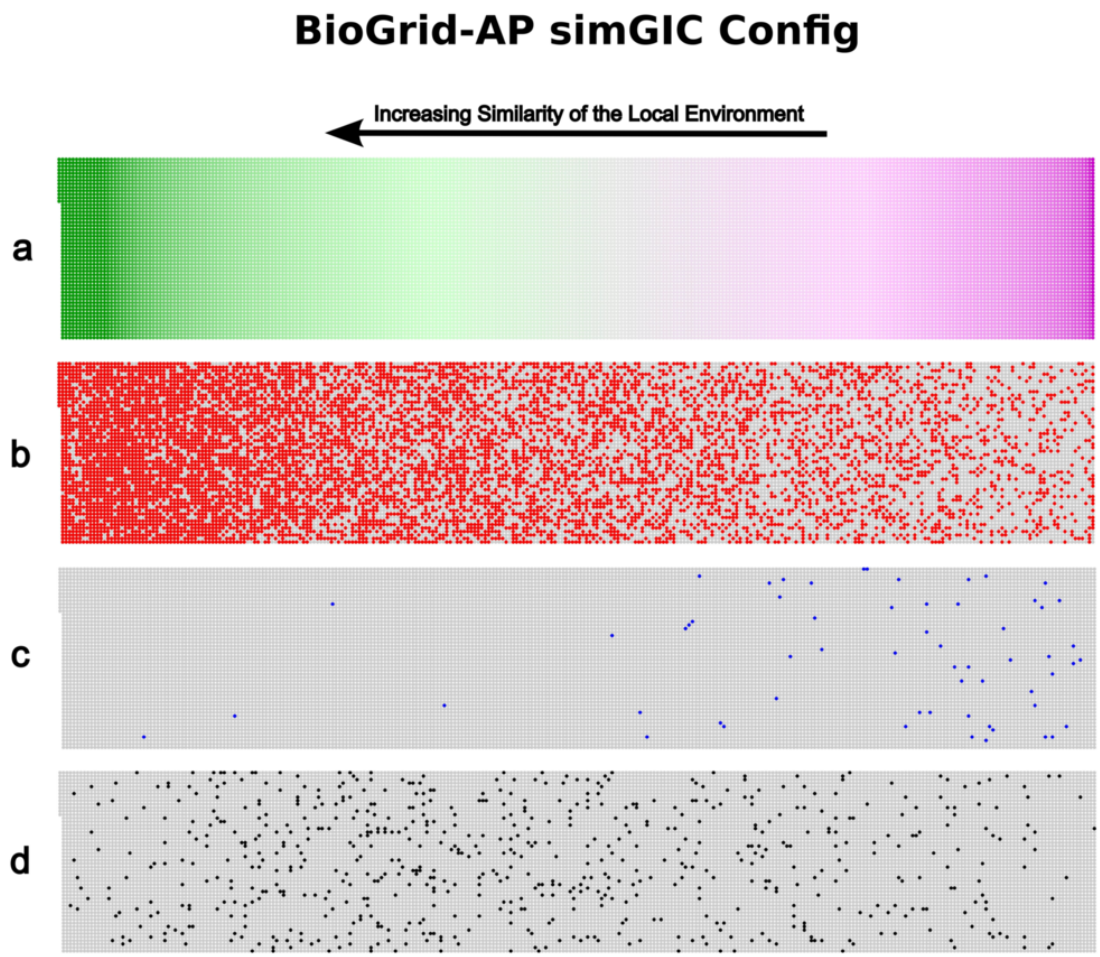

Figure D.19: Local environment comparison of functionally significant BioGrid-AP Config community proteins by simGIC. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using configuration model Modularity Maximization on BioGrid-AP with the simGIC functional similarity measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

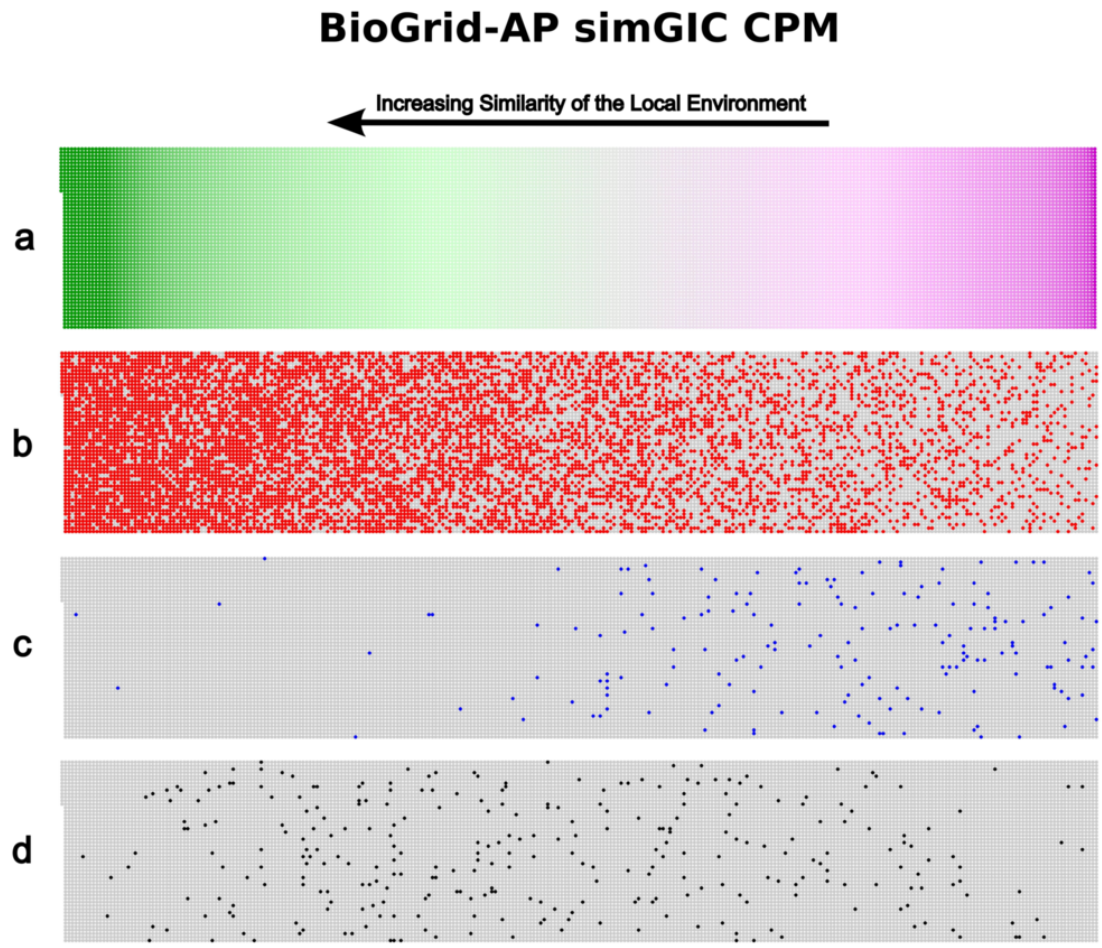

Figure D.20: Local environment comparison of functionally significant BioGrid-AP CPM community proteins by simGIC. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using Constant Potts model Modularity Maximization on BioGrid-AP with the simGIC functional similarity measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

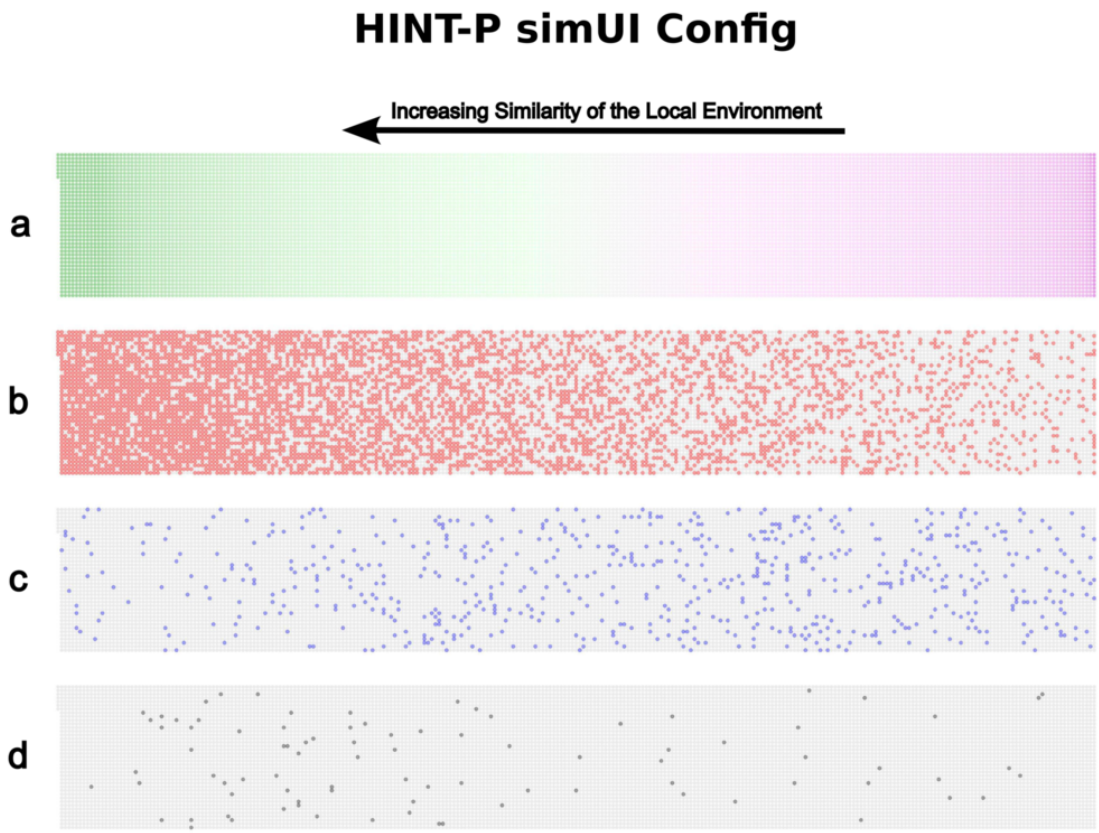

Figure D.21: Local environment comparison of functionally significant HINT-P Config community proteins by simUI. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using configuration model Modularity Maximization on HINT-P with the simUI functional similarity measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

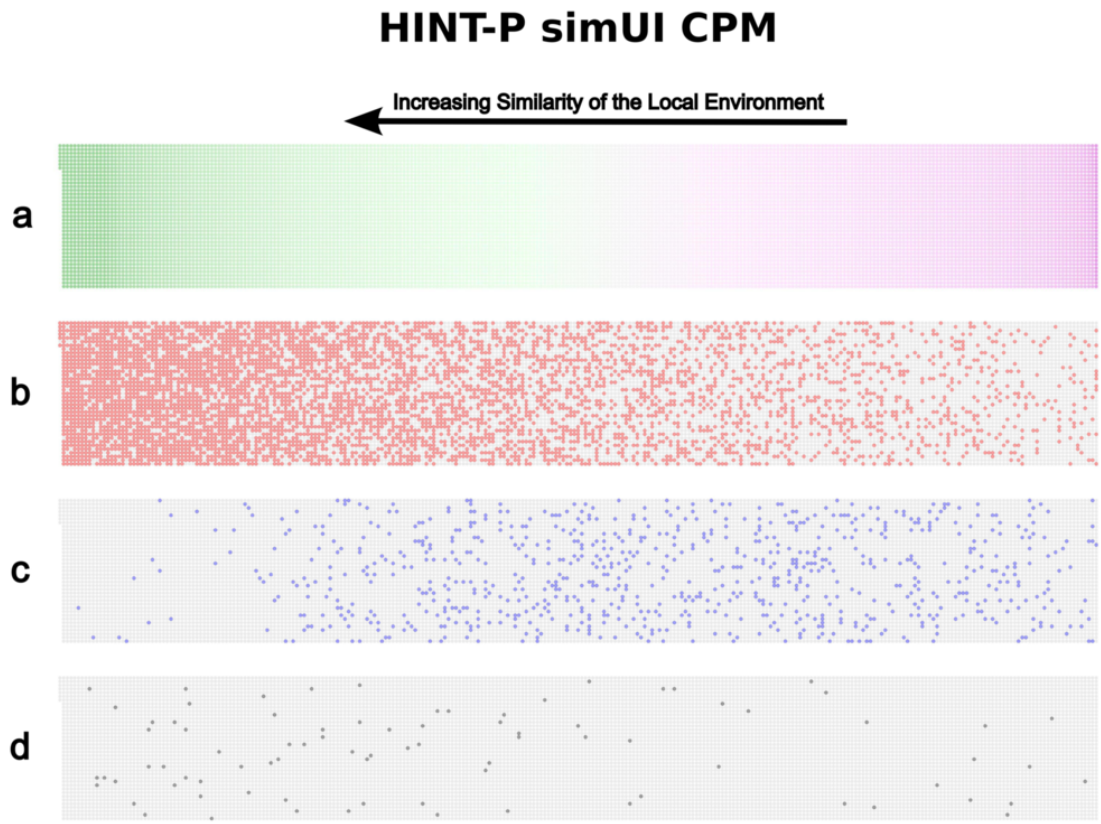

Figure D.22: Local environment comparison of functionally significant HINT-P CPM community proteins by simUI. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using Constant Potts model Modularity Maximization on HINT-P with the simUI functional similarity measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

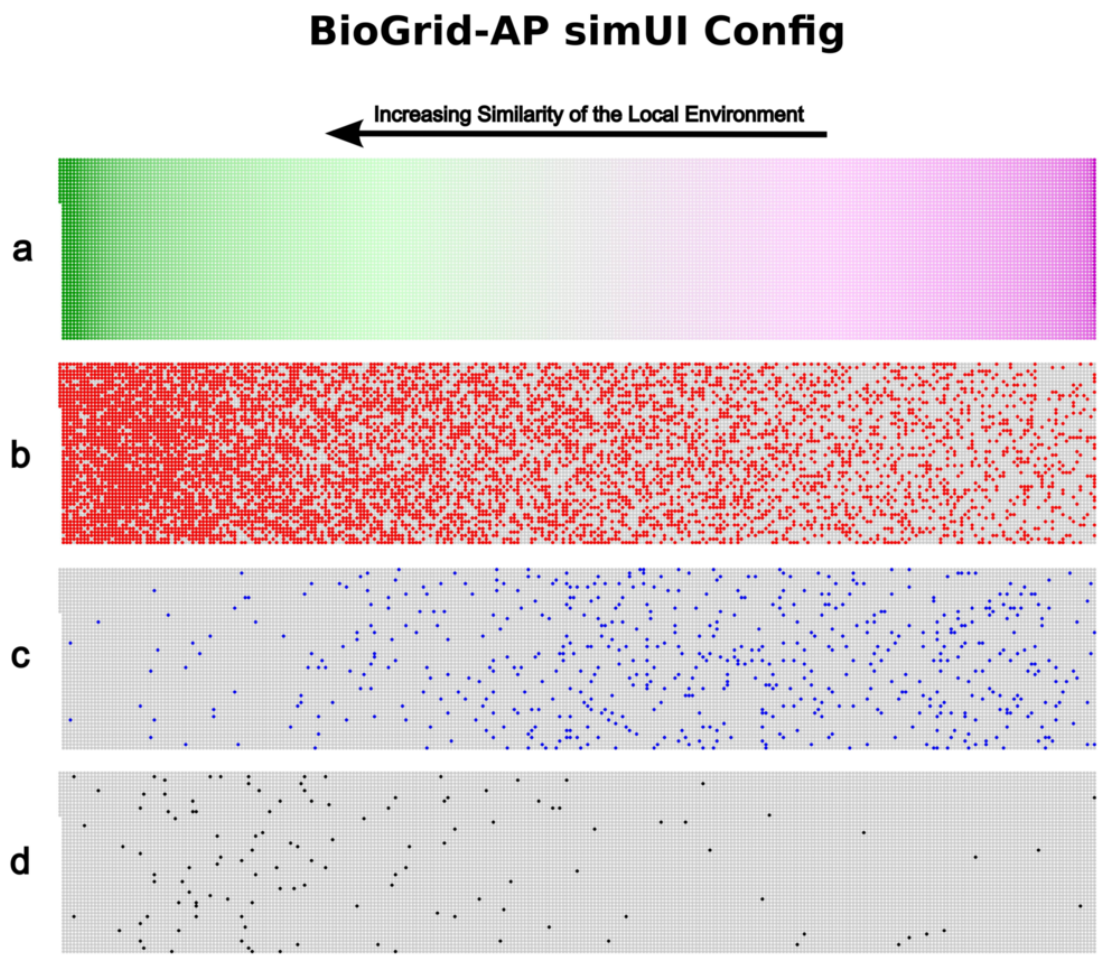

Figure D.23: Local environment comparison of functionally significant BioGrid-AP Config community proteins by simUI. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using configuration model Modularity Maximization on BioGrid-AP with the simUI functional similarity measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

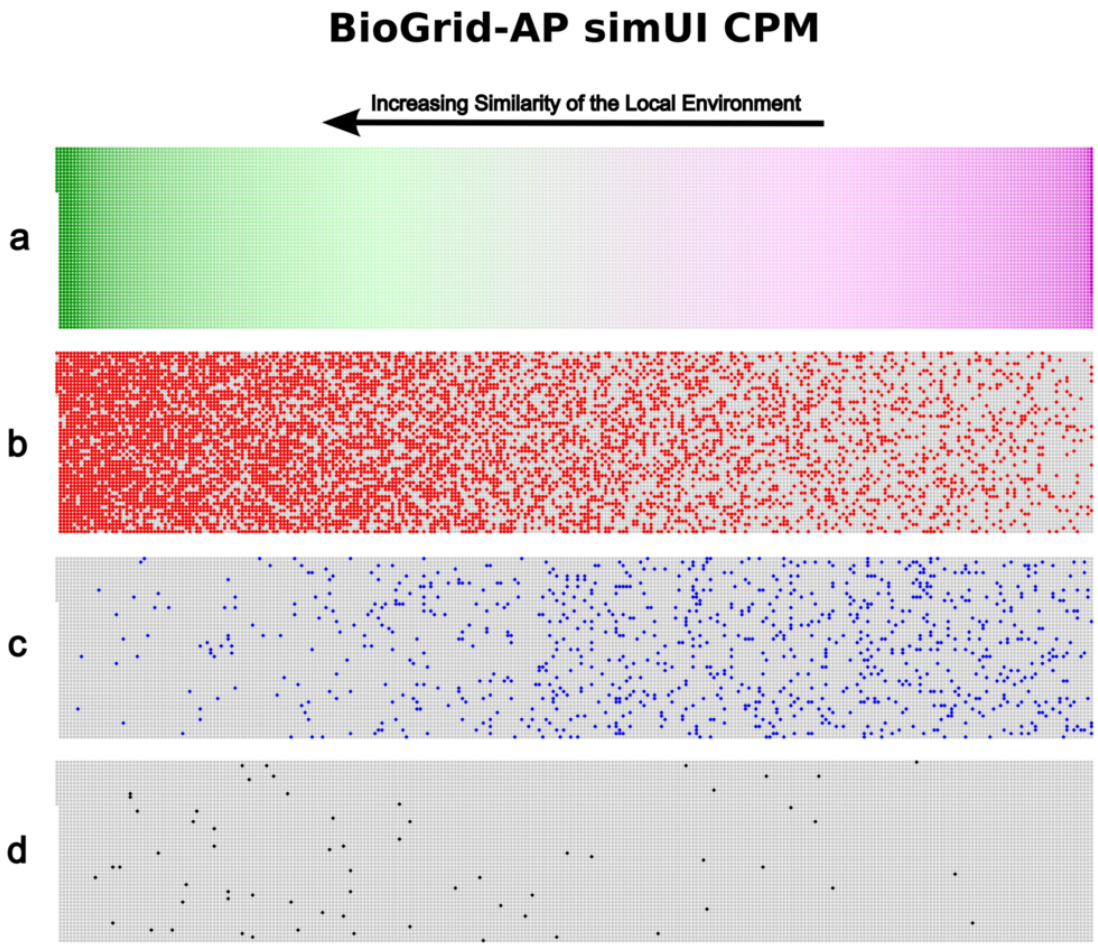

Figure D.24: Local environment comparison of functionally significant BioGrid-AP CPM community proteins by simUI. Comparison of the environments of proteins in communities identified as functionally significant by CommWalker and functional homogeneity. The data shown was generated as in Figure D.14 using Constant Potts model Modularity Maximization on BioGrid-AP with the simUI functional similarity measure. Proteins are ordered by their functional similairty with nodes in their vicinity (a). On this network layout, the proteins in communities identified as functionally significant by CommWalker and functional homogeneity in red (b), by only CommWalker in blue (c), and by only functional homogeneity in black (d) are shown. The further left the coloured nodes are, the higher the functional similarity of their environment.

D.4 Computational Module Validation

To computationally validate communities accepted by CommWalker, we assessed how co-expressed the genes are whose protein products are members of these communities (cf. Section 3.3.1). This validation was performed on the BioGrid-AP Link clustering data set for communities evaluated by the Pandey measure, which was found to best capture gene co-expression (c.f. Section 6.5.1). Figure 6.9 in Section 6.5.1 shows that the median co-expression level is substantially higher for communities only accepted by CommWalker, than for communities only accepted by functional homogeneity. To assess whether this result was due to different community sizes in the two community sets, we investigated the mean co-expression score of the community sets at different community sizes (Figure D.25).

Figure D.25 shows that the result holds true for most community sizes, apart from size 6, where the mean was calculated from only 90 observations of which several correspond to the same community at different resolutions. It should be noted that only mean expression values for community sizes of three and four for the community sets only accepted by functional homogeneity were based on over 1000 observations.

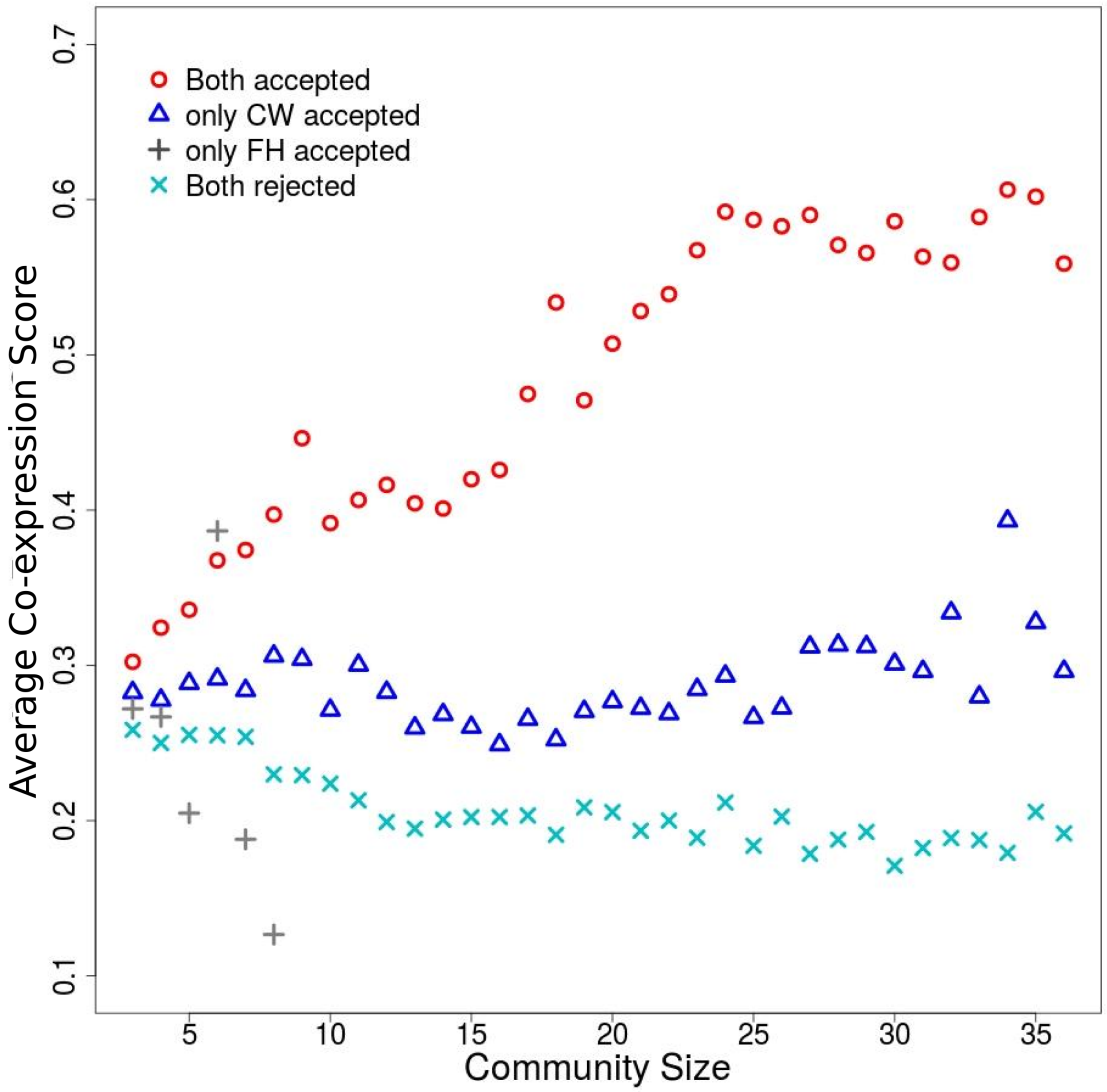

Figure D.25: Comparison of community evaluation methods by average module co-expression versus module size. Link clustering was used to partition BioGrid-AP into communities at multiple resolutions. Using the Pandey measure communities across all resolutions were divided into sets based on whether they were evaluated as functionally significant by both methods (red), only CommWalker (blue), only functional homogeneity (grey), or neither (turqoise). These sets are compared using their average community co-expression score for each community size. The data point marked at community size 36 contains all communities of size 36 - 1047 (the maximum community size for T-value calcuation in BioGrid-AP). Communities only accepted by CommWalker have higher co-expression scores than communities only accepted by functional homogeneity for most of the size range where both sets are populated. At higher community sizes we can still distinguish the co-expression scores of communities only accepted by CommWalker from those rejected by both methods.

Bibliography

[1] Wynn TA, Chawla A, Pollard JW, Macrophage biology in development, homeostasis and disease, *Nature* **496**(7446): pp. 445–455 [2013].

[2] Stuart G, Spruston N, Sakmann B, Häusser M, Action potential initiation and backpropagation in neurons of the mammalian CNS, *Trends in Neurosciences* **20**(3): pp. 125–131 [1997].

[3] Woodcock EA, Matkovich SJ, Cardiomyocytes structure, function and associated pathologies, *The International Journal of Biochemistry & Cell Biology* **37**(9): pp. 1746–1751 [2005].

[4] Nurse P, Reductionism: The ends of understanding, *Nature* **387**(6634): pp. 657–657 [1997].

[5] Hartwell LH, Hopfield JJ, Leibler S, Murray AW, From molecular to modular cell biology, *Nature* **402**(6761 Suppl): pp. C47–52 [1999].

[6] Wagner GP, Altenberg L, Perspective: Complex Adaptations and the Evolution of Evolvability, *Evolution* **50**(3): p. 967 [1996].

[7] Lauffenburger DA, Cell signaling pathways as control modules: complexity for simplicity?, *Proceedings of the National Academy of Sciences of the United States of America* **97**(10): pp. 5031–3 [2000].

[8] Wang PI, Marcotte EM, It's the machine that matters: Predicting gene function and phenotype from protein networks, *Journal of Proteomics* **73**(11): pp. 2277–2289 [2010].

[9] Lehner B, Genotype to phenotype: lessons from model organisms for human genetics, *Nature Reviews Genetics* **14**(3): pp. 168–178 [2013].

[10] Mezey JG, Cheverud JM, Wagner GP, Is the genotype-phenotype map modular? A statistical approach using mouse quantitative trait loci data, *Genetics* **156**(1): pp. 305–11 [2000].

[11] van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM, A text-mining analysis of the human phenome, *European Journal of Human Genetics* **14**(5): pp. 535–542 [2006].

[12] Titz B, Schlesner M, Uetz P, What do we learn from high-throughput protein interaction data?, *Expert Review of Proteomics* **1**(1): pp. 111–121 [2004].

[13] Barabási AL, Gulbahce N, Loscalzo J, Network medicine: a network-based approach to human disease, *Nature Reviews Genetics* **12**(1): pp. 56–68 [2011].

[14] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, *et al.*, The sequence of the human genome, *Science* **291**(5507): pp. 1304–51 [2001].

[15] Vidal M, Fields S, The yeast two-hybrid assay: still finding connections after 25 years, *Nature Methods* **11**(12): pp. 1203–1206 [2014].

[16] Lockhart DJ, Winzeler Ea, Genomics, gene expression and DNA arrays, *Nature* **405**(6788): pp. 827–36 [2000].

[17] Schulze A, Downward J, Navigating gene expression using microarrays–a technology review, *Nature Cell Biology* **3**(8): pp. 190–195 [2001].

[18] Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Séraphin B, The tandem affinity purification (TAP) method: a general procedure of protein complex purification, *Methods* **24**(3): pp. 218–229 [2001].

[19] Fields S, Song Ok, A novel genetic system to detect protein-protein interactions, *Nature* **340**(6230): pp. 245–246 [1989].

[20] Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U, Yeast two-hybrid, a powerful tool for systems biology, *International Journal of Molecular Sciences* **10**(6): pp. 2763–2788 [2009].

[21] Carpenter AE, Sabatini DM, Systematic genome-wide screens of gene function, *Nature Reviews Genetics* **5**(1): pp. 11–22 [2004].

[22] Li M, Wang J, Chen J, A Graph-Theoretic Method for Mining Overlapping Functional Modules in Protein Interaction Networks, *Bioinformatics Research and Applications*, edited by S Istrail, P Pevzner, M Waterman, pp. 208–219, Springer Berlin Heidelberg, Berlin, Heidelberg [2008].

[23] Pereira-Leal JB, Enright AJ, Ouzounis CA, Detection of functional modules from protein interaction networks, *Proteins* **54**(1): pp. 49–57 [2004].

[24] Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH, Modular organization of protein interaction networks, *Bioinformatics* **23**(2): pp. 207–14 [2007].

[25] Mete M, Tang F, Xu X, Yuruk N, A structural approach for finding functional modules from large biological networks, *BMC Bioinformatics* **9**(Suppl 9): p. S19 [2008].

[26] Spirin V, Mirny La, Protein complexes and functional modules in molecular networks, *Proceedings of the National Academy of Sciences of the United States of America* **100**(21): pp. 12123–12128 [2003].

[27] Tripathi S, Moutari S, Dehmer M, Emmert-Streib F, Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules, *BMC Bioinformatics* **17**(1): p. 129 [2016].

[28] Dunn R, Dudbridge F, Sanderson CM, The use of edge-betweenness clustering to investigate biological function in protein interaction networks, *BMC Bioinformatics* **6**(1): p. 39 [2005].

[29] Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A, Eisen M, Spellman P, *et al.*, Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors, *BMC Medical Genomics* **4**(1): p. 34 [2011].

[30] Chen H, Chen J, Muir LA, Ronquist S, Meixner W, Ljungman M, Ried T, Smale S, Rajapakse I, Functional organization of the human 4D Nucleome, *Proceedings of the National Academy of Sciences of the United States of America* **112**(26): pp. 8002—-8007 [2015].

[31] Amar D, Safer H, Shamir R, Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression, *PLoS Computational Biology* **9**(3): p. e1002955 [2013].

[32] Sharan R, Ulitsky I, Shamir R, Network-based prediction of protein function, *Molecular Systems Biology* **3**: p. 88 [2007].

[33] Zhang X, Wu D, Chen L, Li X, Yang J, Fan D, Dong T, *et al.*, RAID: a comprehensive resource for human RNA-associated (RNA-RNA/RNA-protein) interaction, *RNA* **20**(7): pp. 989–993 [2014].

[34] Sharma E, Sterne-Weiler T, O'Hanlon D, Blencowe BJ, Global Mapping of Human RNA-RNA Interactions, *Molecular Cell* **62**(4): pp. 618–626 [2016].

[35] Barabási AL, Oltvai ZN, Network biology: understanding the cell's functional organization, *Nature Reviews Genetics* **5**(2): pp. 101–113 [2004].

[36] Marcotte EM, Computational genetics: finding protein function by nonhomology methods, *Current Opinion in Structural Biology* **10**(3): pp. 359–365 [2000].

[37] Tornow S, Mewes H, Functional modules by relating protein interaction networks and gene expression, *Nucleic Acids Research* **31**(21): pp. 6283–6289 [2003].

[38] Mitra K, Carvunis AR, Ramesh SK, Ideker T, Integrative approaches for finding modular structure in biological networks, *Nature Reviews Genetics* **14**(10): pp. 719–732 [2013].

[39] Ames RM, MacPherson JI, Pinney JW, Lovell SC, Robertson DL, Modular Biological Function Is Most Effectively Captured by Combining Molecular Interaction Data Types, *PLoS ONE* **8**(5): p. e62670 [2013].

[40] Ideker T, Ozier O, Schwikowski B, Siegel AF, Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics* **18**(Suppl 1): pp. S233–S240 [2002].

[41] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T, Identifying functional modules in protein-protein interaction networks: an integrated exact approach, *Bioinformatics* **24**(13): pp. i223–31 [2008].

[42] Muraro D, Simmons A, An integrative analysis of gene expression and molecular interaction data to identify dys-regulated sub-networks in inflammatory bowel disease, *BMC Bioinformatics* **17**(1): p. 42 [2016].

[43] Chen J, Yuan B, Detecting functional modules in the yeast protein-protein interaction network, *Bioinformatics* **22**(18): pp. 2283–90 [2006].

[44] Gustafsson M, Nestor CE, Zhang H, Barabási AL, Baranzini S, Brunak S, Chung KF, *et al.*, Modules, networks and systems medicine for understanding disease and aiding diagnosis, *Genome Medicine* **6**(10): p. 82 [2014].

[45] Leiserson MDM, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, *et al.*, Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes, *Nature Genetics* **47**(2): pp. 106–114 [2014].

[46] Ji JZ, Jiao L, Yang CC, Lv JW, Zhang AD, MAE-FMD: multi-agent evolutionary method for functional module detection in protein-protein interaction networks, *BMC Bioinformatics* **15**(1): p. 325 [2014].

[47] Fellenberg M, Albermann K, Zollner A, Mewes HW, Hani J, Integrative analysis of protein interaction data, *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, vol. 8, pp. 152–61 [2000].

[48] Lewis A, Jones N, Porter MA, Deane CM, The function of communities in protein interaction networks at multiple scales, *BMC Systems Biology* **4**(1): p. 100 [2010].

[49] Chen J, Zhang S, Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data, *Bioinformatics* **32**(11): pp. 1724–1732 [2016].

[50] Cantini L, Medico E, Fortunato S, Caselle M, Detection of gene communities in multi-networks reveals cancer drivers, *Scientific Reports* **5**: p. 17386 [2015].

[51] Ideker T, Krogan NJ, Differential network biology, *Molecular Systems Biology* **8**(1): pp. 198–207 [2012].

[52] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America* **102**(43): pp. 15545–50 [2005].

[53] Gatenby RA, Gillies RJ, Why do cancers have high aerobic glycolysis?, *Nature Reviews Cancer* **4**(11): pp. 891–899 [2004].

[54] Semenza GL, The hypoxic tumor microenvironment: A driving force for breast cancer progression, *Biochimica et Biophysica Acta* **1863**(3): pp. 382–391 [2016].

[55] Indelicato M, Pucci B, Schito L, Reali V, Aventaggiato M, Mazzarino MC, Stivala F, Fini M, Russo MA, Tafani M, Role of hypoxia and autophagy in MDA-MB-231 invasiveness, *Journal of Cellular Physiology* **223**(2): pp. n/a–n/a [2010].

[56] Xu Y, Xia X, Pan H, Active autophagy in the tumor microenvironment: A novel mechanism for cancer metastasis (Review), *Oncology Letters* **5**(2): pp. 411–416 [2013].

[57] Stefater III JA, Ren S, Lang RA, Duffield JS, Metchnikoff's policemen: macrophages in development, homeostasis and regeneration, *Trends in Molecular Medicine* **17**(12): pp. 743–752 [2011].

[58] Hume DA, MacDonald KPA, Therapeutic applications of macrophage colony-stimulating factor-1 (CSF-1) and antagonists of CSF-1 receptor (CSF-1R) signaling, *Blood* **119**(8): pp. 1810–1820 [2012].

[59] Mantovani A, Sica A, Sozzani S, Allavena P, Vecchi A, Locati M, The chemokine system in diverse forms of macrophage activation and polarization, *Trends in Immunology* **25**(12): pp. 677–686 [2004].

[60] Martinez FO, Gordon S, The M1 and M2 paradigm of macrophage activation: time for reassessment, *F1000prime Reports* **6**: p. 13 [2014].

[61] Martinez FO, Gordon S, Locati M, Mantovani A, Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression, *Journal of Immunology* **177**(10): pp. 7303–7311 [2006].

[62] Hoesel B, Schmid JA, The complexity of NF-κB signaling in inflammation and cancer, *Molecular Cancer* **12**(1): p. 86 [2013].

[63] Chen S, Kammerl IE, Vosyka O, Baumann T, Yu Y, Wu Y, Irmler M, Overkleeft HS, Beckers J, Eickelberg O, Meiners S, Stoeger T, Immunoproteasome dysfunction augments alternative polarization of alveolar macrophages, *Cell Death and Differentiation* **23**(6): pp. 1026–1037 [2016].

[64] Xuan W, Qu Q, Zheng B, Xiong S, Fan GH, The chemotaxis of M1 and M2 macrophages is regulated by different chemokines, *Journal of Leukocyte Biology* **97**(1): pp. 61–69 [2015].

[65] Sahni N, Yi S, Zhong Q, Jailkhani N, Charloteaux B, Cusick ME, Vidal M, Edgotype: a fundamental link between genotype and phenotype, *Current Opinion in Genetics & Development* **23**(6): pp. 649–657 [2013].

[66] Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, Peng J, *et al.*, Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders, *Cell* **161**(3): pp. 647–660 [2015].

[67] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P, Comparative assessment of large-scale data sets of protein-protein interactions, *Nature* **417**(6887): pp. 399–403 [2002].

[68] Shoemaker BA, Panchenko AR, Deciphering protein-protein interactions. Part I. Experimental techniques and databases, *PLoS Computational Biology* **3**(3): p. e42 [2007].

[69] Das J, Yu H, HINT: High-quality protein interactomes and their applications in understanding human disease, *BMC Systems Biology* **6**(1): p. 92 [2012].

[70] Côté R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H, The Ontology Lookup Service: bigger and better, *Nucleic Acids Research* **38**(Web Server issue): pp. W155–60 [2010].

[71] Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature* **415**(6868): pp. 141–147 [2002].

[72] Collins MO, Choudhary JS, Mapping multiprotein complexes by affinity purification and mass spectrometry, *Current Opinion in Biotechnology* **19**(4): pp. 324–330 [2008].

[73] Aebersold R, Mann M, Mass spectrometry-based proteomics, *Nature* **422**(6928): pp. 198–207 [2003].

[74] Hakes L, Pinney JW, Robertson DL, Lovell SC, Protein-protein interaction networks and biology–what's the connection?, *Nature Biotechnology* **26**(1): pp. 69–72 [2008].

[75] de Las Rivas J, Fontanillo C, Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks, *PLoS Computational Biology* **6**(6): pp. 1–8 [2010].

[76] Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P, Estimating accuracy of RNA-Seq and microarrays with proteomics, *BMC Genomics* **10**(1): p. 161 [2009].

[77] Corominas R, Yang X, Lin GN, Kang S, Shen Y, Ghamsari L, Broly M, *et al.*, Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism, *Nature Communications* **5**: p. 3650 [2014].

[78] Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, Wrana JL, Blencowe BJ, Tissue-specific alternative splicing remodels protein-protein interaction networks, *Molecular Cell* **46**(6): pp. 884–92 [2012].

[79] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, *et al.*, Towards a proteome-scale map of the human protein-protein interaction network, *Nature* **437**(7062): pp. 1173–8 [2005].

[80] Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, *et al.*, A proteome-scale map of the human interactome network, *Cell* **159**(5): pp. 1212–26 [2014].

[81] Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, *et al.*, The BioPlex Network: A Systematic Exploration of the Human Interactome, *Cell* **162**(2): pp. 425–40 [2015].

[82] Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual JF, *et al.*, Literature-curated protein interaction datasets, *Nature Methods* **6**(1): pp. 39–46 [2009].

[83] Ali W, Deane C, Reinert G, Protein Interaction Networks and Their Statistical Analysis, *Handbook of Statistical Systems Biology*, edited by M Stumpf, D Balding, M Girolami, pp. 200–234, John Wiley & Sons, Ltd [2011].

[84] Srinivasa Rao V, Srinivas K, Sujini GN, Sunand Kumar GN, Protein-protein interaction detection: methods and analysis, *International Journal of Proteomics* **2014**: p. 147648 [2014].

[85] Washburn MP, There is no human interactome, *Genome Biology* **17**(1): p. 48 [2016].

[86] Ou-Yang L, Dai DQ, Li XL, Wu M, Zhang XF, Yang P, Detecting temporal protein complexes from dynamic protein-protein interaction networks, *BMC Bioinformatics* **15**(1): p. 335 [2014].

[87] Lewis A, *Communities and homology in protein-protein interactions*, Ph.D. thesis, University of Oxford, Statistics Department [2011].

[88] Hart GT, Ramani AK, Marcotte EM, How complete are current yeast and human protein-interaction networks?, *Genome Biology* **7**(11): p. 120 [2006].

[89] Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, *et al.*, An empirical framework for binary interactome mapping, *Nature Methods* **6**(1): pp. 83–90 [2009].

[90] Kelly WP, Stumpf MPH, Assessing Coverage of Protein Interaction Data Using Capture–Recapture Models, *Bulletin of Mathematical Biology* **74**(2): pp. 356–374 [2012].

[91] Salwinski L, Licata L, Winter A, Thorneycroft D, Khadake J, Ceol A, Aryamontri AC, *et al.*, Recurated protein interaction datasets, *Nature Methods* **6**(12): pp. 860–861 [2009].

[92] Royer L, Reimann M, Stewart AF, Schroeder M, Network compression as a quality measure for protein interaction networks, *PloS One* **7**(6): p. e35729 [2012].

[93] Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C, Estimating the size of the human interactome, *Proceedings of the National Academy of Sciences of the United States of America* **105**(19): pp. 6959–64 [2008].

[94] Mazandu GK, Mulder NJ, Information Content-Based Gene Ontology Functional Similarity Measures: Which One to Use for a Given Biological Data Type?, *PLoS ONE* **9**(12): p. e113859 [2014].

[95] Deane CM, Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations, *Molecular & Cellular Proteomics* **1**(5): pp. 349–356 [2002].

[96] Chen PY, Deane CM, Reinert G, Predicting and validating protein interactions using network structure, *PLoS Computational Biology* **4**(7): p. e1000118 [2008].

[97] Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, *et al.*, An experimentally derived confidence score for binary protein-protein interactions, *Nature Methods* **6**(1): pp. 91–97 [2009].

[98] Schaefer MH, Serrano L, Andrade-Navarro MA, Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types, *Frontiers in Genetics* **6** [2015].

[99] Mrowka R, Patzak A, Herzel H, Is there a bias in proteome research?, *Genome Research* **11**(12): pp. 1971–3 [2001].

[100] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M, BioGRID: a general repository for interaction datasets, *Nucleic Acids Research* **34**(Database issue): pp. D535–D539 [2006].

[101] Chatr-aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, *et al.*, The BioGRID interaction database: 2015 update, *Nucleic Acids Research* **43**(D1): pp. D470–D478 [2015].

[102] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, *et al.*, The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases, *Nucleic Acids Research* **42**(Database issue): pp. D358–63 [2014].

[103] Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, *et al.*, Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Research* **13**(10): pp. 2363–71 [2003].

[104] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, *et al.*, Human Protein Reference Database–2009 update, *Nucleic Acids Research* **37**(suppl. 1): pp. D767–72 [2009].

[105] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, *et al.*, STRING v10: protein-protein interaction networks, integrated over the tree of life, *Nucleic Acids Research* **43**(Database issue): pp. D447–52 [2015].

[106] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G, MINT: a Molecular INTeraction database, *FEBS Letters* **513**(1): pp. 135–140 [2002].

[107] Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G, MINT: the Molecular INTeraction database, *Nucleic Acids Research* **35**(Database issue): pp. D572–4 [2007].

[108] Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, *et al.*, Protein interaction data curation: the International Molecular Exchange (IMEx) consortium, *Nature Methods* **9**(4): pp. 345–50 [2012].

[109] Zhou X, Chen P, Wei Q, Shen X, Chen X, Human Interactome Resource and Gene Set Linkage Analysis for the Functional Interpretation of Biologically Meaningful Gene Sets, *Bioinformatics* **29**(16): pp. 2024–31 [2013].

[110] Martin T, Ball B, Newman MEJ, Structural inference for uncertain networks, *Physical Review E* **93**(1): p. 012306 [2016].

[111] Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA, Multilayer networks, *Journal of Complex Networks* **2**(3): pp. 203–271 [2014].

[112] Newman M, *Networks: An Introduction*, Oxford University Press, New York [2010].

[113] Yang J, Leskovec J, Structure and Overlaps of Ground-Truth Communities in Networks, *ACM Transactions on Intelligent Systems and Technology* **5**(2): pp. 1–35 [2014].

[114] Watts DJ, Strogatz SH, Collective dynamics of 'small-world' networks, *Nature* **393**(6684): pp. 440–442 [1998].

[115] Barabasi AL, Albert R, Emergence of scaling in random networks, *Science* **286**(5439): pp. 509–12 [1999].

[116] Barabási AL, Scale-free networks: a decade and beyond, *Science* **325**(5939): pp. 412–3 [2009].

[117] Stumpf MPH, Wiuf C, May RM, Subnets of scale-free networks are not scale-free: sampling properties of networks, *Proceedings of the National Academy of Sciences of the United States of America* **102**(12): pp. 4221–4 [2005].

[118] Han JDJ, Dupuy D, Bertin N, Cusick ME, Vidal M, Effect of sampling on topology predictions of protein-protein interaction networks, *Nature Biotechnology* **23**(7): pp. 839–844 [2005].

[119] Lima-Mendez G, van Helden J, The powerful law of the power law and other myths in network biology, *Molecular BioSystems* **5**(12): p. 1482 [2009].

[120] Rito T, Wang Z, Deane CM, Reinert G, How threshold behaviour affects the use of subgraphs for network comparison, *Bioinformatics* **26**(18): pp. i611—-i617 [2010].

[121] Rito T, Deane CM, Reinert G, The Importance of Age and High Degree, in Protein-Protein Interaction Networks, *Journal of Computational Biology* **19**(6): pp. 785–795 [2012].

[122] Rombach MP, Porter MA, Fowler JH, Mucha PJ, Core-Periphery Structure in Networks, *SIAM Journal on Applied Mathematics* **74**(1): pp. 167–190 [2014].

[123] Leskovec J, Lang KJ, Dasgupta A, Mahoney MW, Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters, *Internet Mathematics* **6**(1): pp. 29–123 [2009].

[124] Ali W, Rito T, Reinert G, Sun F, Deane CM, Alignment-free protein interaction network comparison, *Bioinformatics* **30**(17): pp. i430–i437 [2014].

[125] Yang J, Leskovec J, Overlapping community detection at scale, *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining - WSDM '13*, p. 587, ACM Press, New York, New York, USA [2013].

[126] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10): p. P10008 [2008].

[127] Ahn YY, Bagrow JP, Lehmann S, Link communities reveal multiscale complexity in networks, *Nature* **466**(7307): pp. 761–4 [2010].

[128] Porter MA, Onnela JP, Mucha PJ, Communities in networks, *Notices of the AMS* **56**(9): pp. 1082–1097 [2009].

[129] Fortunato S, Community detection in graphs, *Physics Reports* **486**(3-5): pp. 75–174 [2010].

[130] Newman MEJ, Communities, modules and large-scale structure in networks, *Nature Physics* **8**(1): pp. 25–31 [2011].

[131] Holland PW, Laskey KB, Leinhardt S, Stochastic blockmodels: First steps, *Social Networks* **5**(2): pp. 109–137 [1983].

[132] Girvan M, Newman MEJ, Community structure in social and biological networks, *Proceedings of the National Academy of Sciences of the United States of America* **99**(12): pp. 7821–7826 [2002].

[133] Newman M, Girvan M, Finding and evaluating community structure in networks, *Physical Review E* **69**(2): p. 026113 [2004].

[134] Clauset A, Newman M, Moore C, Finding community structure in very large networks, *Physical Review E* **70**(6): p. 066111 [2004].

[135] Reichardt J, Bornholdt S, Statistical mechanics of community detection, *Physical Review E* **74**(1): p. 16110 [2006].

[136] Brandes U, Delling D, Gaertler M, Gorke R, Hoefer M, Nikoloski Z, Wagner D, On modularity clustering, *Transactions on Knowledge and Data Engineering, IEEE* **20**(2): pp. 172–188 [2008].

[137] Newman MEJ, Community detection in networks: Modularity optimization and maximum likelihood are equivalent, *arXiv preprint arXiv:160602319* [2016].

[138] Bollobás B, A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs, *European Journal of Combinatorics* **1**(4): pp. 311–316 [1980].

[139] Fortunato S, Barthélemy M, Resolution limit in community detection, *Proceedings of the National Academy of Sciences of the United States of America* **104**(1): pp. 36–41 [2007].

[140] Traag VA, Van Dooren P, Nesterov Y, Narrow scope for resolution-limit-free community detection, *Physical Review E* **84**(1): p. 16114 [2011].

[141] Lancichinetti A, Fortunato S, Limits of modularity maximization in community detection, *Physical Review E* **84**(6): p. 066122 [2011].

[142] Good BH, de Montjoye YA, Clauset A, Performance of modularity maximization in practical contexts, *Physical Review E* **81**(4): p. 046106 [2010].

[143] Lancichinetti A, Fortunato S, Consensus clustering in complex networks, *Scientific Reports* **2**: pp. 47–97 [2012].

[144] Lancichinetti A, Fortunato S, Community detection algorithms: A comparative analysis, *Physical Review E* **80**(5): p. 056117 [2009].

[145] Yang J, Leskovec J, Community-Affiliation Graph Model for Overlapping Network Community Detection, *2012 IEEE 12th International Conference on Data Mining* pp. 1170–1175 [2012].

[146] Lee DD, Seung HS, Learning the parts of objects by non-negative matrix factorization, *Nature* **401**(6755): pp. 788–91 [1999].

[147] Zachary WW, An Information Flow Model for Conflict and Fission in Small Groups, *Journal of Anthropological Research* **33**(4): pp. 452–473 [1977].

[148] Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM, The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, *Behavioral Ecology and Sociobiology* **54**(4): pp. 396–405 [2003].

[149] Traud AL, Kelsic ED, Mucha PJ, Porter MA, Comparing Community Structure to Characteristics in Online Collegiate Social Networks, *SIAM Review* **53**(3): pp. 526–543 [2011].

[150] Traud AL, Mucha PJ, Porter MA, Social structure of Facebook networks, *Physica A: Statistical Mechanics and its Applications* **391**(16): pp. 4165–4180 [2012].

[151] Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B, Measurement and analysis of online social networks, *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC '07*, p. 29, ACM Press, New York, New York, USA [2007].

[152] Yang J, Leskovec J, Defining and evaluating network communities based on ground-truth, *Knowledge and Information Systems* **42**(1): pp. 181–213 [2015].

[153] Lancichinetti A, Fortunato S, Radicchi F, Benchmark graphs for testing community detection algorithms, *Physical Review E* **78**(4): p. 046110 [2008].

[154] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Others, Gene Ontology: tool for the unification of biology, *Nature Genetics* **25**(1): pp. 25–29 [2000].

[155] Reference Genome Group of the Gene Ontology Consortium and Others, The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species, *PLoS Computational Biology* **5**(7): p. e1000431 [2009].

[156] GO Consortium, Guide to GO Evidence Codes, *http://geneontologyorg/page/guide-go-evidence-codes* (accessed 24.08.2016) [2016].

[157] Guzzi PH, Mina M, Guerra C, Cannataro M, Semantic similarity analysis of protein data: assessment with biological features and issues, *Briefings in Bioinformatics* **13**(5): pp. 569–85 [2012].

[158] Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I, Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space, *PLoS Computational Biology* **9**(5): p. e1003063 [2013].

[159] Gene Ontology Consortium, The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Research* **32**(90001): pp. 258D–261 [2004].

[160] Alterovitz G, Xiang M, Mohan M, Ramoni MF, GO PaD: the Gene Ontology Partition Database, *Nucleic Acids Research* **35**(Database issue): pp. D322–7 [2007].

[161] Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M, Mewes HW, The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids Research* **32**(18): pp. 5539–5545 [2004].

[162] Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, *et al.*, The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data, *Nucleic Acids Research* **42**(Database issue): pp. D966–74 [2014].

[163] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M, KEGG as a reference resource for gene and protein annotation, *Nucleic Acids Research* **44**(D1): pp. D457–D462 [2016].

[164] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, *et al.*, The Reactome pathway knowledgebase, *Nucleic Acids Research* **42**(Database issue): pp. D472–7 [2014].

[165] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, *et al.*, The Reactome pathway Knowledgebase, *Nucleic Acids Research* **44**(D1): pp. D481–D487 [2016].

[166] Mewes HW, Frishman D, Mayer KFX, Münsterkötter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stümpflen V, MIPS: analysis and annotation of proteins from whole genomes in 2005, *Nucleic Acids Research* **34**(90001): pp. D169–D172 [2006].

[167] Mewes HW, MIPS: a database for genomes and protein sequences, *Nucleic Acids Research* **30**(1): pp. 31–34 [2002].

[168] Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, *et al.*, ClinVar: public archive of interpretations of clinically relevant variants, *Nucleic Acids Research* **44**(D1): pp. D862–D868 [2016].

[169] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Research* **33**(Database issue): pp. D514–7 [2005].

[170] Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, *et al.*, Principles for the post-GWAS functional characterization of cancer risk loci, *Nature Genetics* **43**(6): pp. 513–518 [2011].

[171] Schofield PN, Gkoutos GV, Gruenberger M, Sundberg JP, Hancock JM, Ackert-Bicknell C, Demissie S, *et al.*, Phenotype ontologies for mouse and man: bridging the semantic gap, *Disease Models & Mechanisms* **3**(5-6): pp. 281–9 [2010].

[172] Huang DW, Sherman BT, Lempicki RA, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Research* **37**(1): pp. 1–13 [2009].

[173] Khatri P, Sirota M, Butte AJ, Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges, *PLoS Computational Biology* **8**(2): p. e1002375 [2012].

[174] Langfelder P, Horvath S, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics* **9**(1): p. 559 [2008].

[175] Mclean C, He X, J IST, Armstrong D, Improved Functional Enrichment Analysis of Biological Networks using Scalable Modularity Based Clustering, *Journal of Proteomics & Bioinformatics* **9**(1) [2016].

[176] Brunet AC, Azais JM, Loubes JM, Amar J, Burcelin R, A new gene co-expression network analysis based on Core Structure Detection (CSD), *arXiv preprint arXiv:160701516* [2016].

[177] Dudoit S, Popper Shaffer J, Boldrick JC, Multiple Hypothesis Testing in Microarray Experiments, *Statistical Science* **18**(1): pp. 71–103 [2003].

[178] Shaffer JP, Multiple Hypothesis Testing, *Annual Review of Psychology* **46**(1): pp. 561–584 [1995].

[179] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM, Semantic similarity in biomedical ontologies, *PLoS Computational Biology* **5**(7): p. e1000443 [2009].

[180] Resnik P, Using information content to evaluate semantic similarity in a taxonomy, *arXiv preprint cmp-lg/9511007* [1995].

[181] Pandey J, Koyutürk M, Subramaniam S, Grama A, Functional coherence in domain interaction networks, *Bioinformatics* **24**(16): pp. i28–34 [2008].

[182] Pesquita C, Faria D, Bastos H, Ferreira AEN, Falcão AO, Couto FM, Metrics for GO based protein semantic similarity: a systematic evaluation, *BMC Bioinformatics* **9 Suppl 5**: p. S4 [2008].

[183] Gentleman R, Visualizing and Distances Using GO, *http://wwwbioconductororg/docs/vignetteshtml* (accessed 01.08.2016) [2005].

[184] Lin N, Wu B, Jansen R, Gerstein M, Zhao H, Information assessment on predicting protein-protein interactions, *BMC Bioinformatics* **5**(1): p. 154 [2004].

[185] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M, A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science* **302**(5644): pp. 449–53 [2003].

[186] Wang J, Zhou X, Zhu J, Zhou C, Guo Z, Revealing and avoiding bias in semantic similarity scores for protein pairs, *BMC Bioinformatics* **11**(1): p. 290 [2010].

[187] Pesquita C, Pessoa D, Faria D, Couto FM, CESSM: Collaborative Evaluation of Semantic Similarity Measures, *JB2009: Challenges in Bioinformatics* **157**: p. 190 [2009].

[188] Xu Y, Guo M, Shi W, Liu X, Wang C, A novel insight into Gene Ontology semantic similarity, *Genomics* **101**(6): pp. 368–75 [2013].

[189] Jansen R, Greenbaum D, Gerstein M, Relating whole-genome expression data with protein-protein interactions, *Genome Research* **12**(1): pp. 37–46 [2002].

[190] Li JJ, Biggin MD, Gene expression. Statistics requantitates the central dogma, *Science* **347**(6226): pp. 1066–7 [2015].

[191] Gygi SP, Rist B, Aebersold R, Measuring gene expression by quantitative proteome analysis, *Current Opinion in Biotechnology* **11**(4): pp. 396–401 [2000].

[192] Wang Z, Gerstein M, Snyder M, RNA-Seq: a revolutionary tool for transcriptomics, *Nature reviews Genetics* **10**(1): pp. 57–63 [2009].

[193] Vinciotti V, Wit EC, Jansen R, de Geus EJCN, Penninx BWJH, Boomsma DI, 't Hoen PAC, *et al.*, Consistency of biological networks inferred from microarray and sequencing data, *BMC Bioinformatics* **17**(1): p. 254 [2016].

[194] Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, *et al.*, Multiple-laboratory comparison of microarray platforms, *Nature Methods* **2**(5): pp. 345–50 [2005].

[195] Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, *et al.*, Standardizing global gene expression analysis between laboratories and across platforms, *Nature Methods* **2**(5): pp. 351–6 [2005].

[196] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, *et al.*, The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements, *Nature Biotechnology* **24**(9): pp. 1151–61 [2006].

[197] Oshlack A, Wakefield MJ, Transcript length bias in RNA-seq data confounds systems biology, *Biology Direct* **4**(1): p. 14 [2009].

[198] Zhou X, Kao MCJ, Wong WH, Transitive functional annotation by shortest-path analysis of gene expression data, *Proceedings of the National Academy of Sciences of the United States of America* **99**(20): pp. 12783–8 [2002].

[199] van Noort V, Snel B, Huynen MA, Predicting gene function by conserved co-expression, *Trends in Genetics* **19**(5): pp. 238–42 [2003].

[200] Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P, Coexpression analysis of human genes across many microarray data sets, *Genome Research* **14**(6): pp. 1085–94 [2004].

[201] Quackenbush J, Microarray data normalization and transformation, *Nature Genetics* **32 Suppl**(december): pp. 496–501 [2002].

[202] Scholtens D, Heydebreck AV, Analysis of Differential Gene Expression Studies, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, edited by R Gentleman, V Carey, W Huber, R Irizarry, S Dudoit, chap. 14, pp. 229–248, Springer [2005].

[203] Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub, Web Presence Working Group, AmiGO: online access to ontology and annotation data, *Bioinformatics* **25**(2): pp. 288–9 [2009].

[204] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, *et al.*, The Genotype-Tissue Expression (GTEx) project, *Nature Genetics* **45**(6): pp. 580–5 [2013].

[205] Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, *et al.*, Ensembl 2015, *Nucleic Acids Research* **43**(Database issue): pp. D662–9 [2014].

[206] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, *et al.*, NCBI GEO: archive for functional genomics data sets–update, *Nucleic Acids Research* **41**(D1): pp. D991–D995 [2013].

[207] Smyth GK, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, edited by MPH Stumpf, vol. 3, p. Article 3, De Gruyter [2004].

[208] Smyth GK, limma: Linear Models for Microarray Data, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, edited by R Gentleman, V Carey, S Dudoit, RA Irizarry, W Huber, pp. 397–420, Springer-Verlag, New York [2005].

[209] Davis D, Yaveroğlu ÖN, Malod-Dognin N, Stojmirovic A, Pržulj N, Topology-function conservation in protein-protein interaction networks, *Bioinformatics* **31**(10): pp. 1632–1639 [2015].

[210] Reboiro-Jato M, Arrais JP, Oliveira JL, Fdez-Riverola F, geneCommittee: a web-based tool for extensively testing the discriminatory power of biologically relevant gene sets in microarray data classification, *BMC Bioinformatics* **15**(1): p. 31 [2014].

[211] Dinkla K, El-Kebir M, Bucur CI, Siderius M, Smit MJ, Westenberg MA, Klau GW, eXamine: Exploring annotated modules in networks, *BMC Bioinformatics* **15**(1): p. 201 [2014].

[212] Leskovec J, Lang KJ, Mahoney M, Empirical comparison of algorithms for network community detection, *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, p. 631, ACM Press, New York, New York, USA [2010].

[213] Hric D, Darst RK, Fortunato S, Community detection in networks: Structural communities versus ground truth, *Physical Review E* **90**(6): p. 062805 [2014].

[214] Pearson K, *On further methods of determining correlation*, vol. 4, Dulau and Company [1907].

[215] Scrivens PJ, Noueihed B, Shahrzad N, Hul S, Brunet S, Sacher M, C4orf41 and TTC-15 are mammalian TRAPP components with a role at an early stage in ER-to-Golgi trafficking, *Molecular Biology of the Cell* **22**(12): pp. 2083–93 [2011].

[216] Tong H, Faloutsos C, Pan JY, Fast Random Walk with Restart and Its Applications, *Sixth International Conference on Data Mining (ICDM'06)*, pp. 613–622, IEEE [2006].

[217] Gray KA, Seal RL, Tweedie S, Wright MW, Bruford EA, A review of the new HGNC gene family resource, *Human Genomics* **10**: p. 6 [2016].

[218] Choudhry H, Schödel J, Oikonomopoulos S, Camps C, Grampp S, Harris AL, Ratcliffe PJ, Ragoussis J, Mole DR, Extensive regulation of the non-coding transcriptome by hypoxia: role of HIF in releasing paused RNApol2, *EMBO Reports* **15**(1): pp. 70–6 [2014].

[219] Kean MJ, Williams KC, Skalski M, Myers D, Burtnik A, Foster D, Coppolino MG, VAMP3, syntaxin-13 and SNAP23 are involved in secretion of matrix metalloproteinases, degradation of the extracellular matrix and cell invasion, *Journal of Cell Science* **122**(Pt 22): pp. 4089–98 [2009].

[220] Caswell PT, Spence HJ, Parsons M, White DP, Clark K, Cheng KW, Mills GB, *et al.*, Rab25 Associates with $\alpha 5\beta 1$ Integrin to Promote Invasive Migration in 3D Microenvironments, *Developmental Cell* **13**(4): pp. 496–510 [2007].

[221] Pike LRG, Singleton DC, Buffa F, Abramczyk O, Phadwal K, Li JL, Simon AK, Murray JT, Harris AL, Transcriptional up-regulation of ULK1 by ATF4 contributes to cancer cell survival, *The Biochemical Journal* **449**(2): pp. 389–400 [2013].

[222] Rawlings JS, Rosler KM, Harrison DA, The JAK/STAT signaling pathway, *Journal of Cell Science* **117**(Pt 8): pp. 1281–3 [2004].

[223] Rieser E, Cordier SM, Walczak H, Linear ubiquitination: a newly discovered regulator of cell signalling, *Trends in Biochemical Sciences* **38**(2): pp. 94–102 [2013].

[224] Shimizu Y, Taraborrelli L, Walczak H, Linear ubiquitination in immunity, *Immunological Reviews* **266**(1): pp. 190–207 [2015].

[225] Ehsani R, Drabløs F, TopoICSim: a new semantic similarity measure based on gene ontology, *BMC Bioinformatics* **17**(1): p. 296 [2016].

[226] Lancichinetti A, Fortunato S, Kertész J, Detecting the overlapping and hierarchical community structure in complex networks, *New Journal of Physics* **11**(3): p. 033015 [2009].

[227] Danon L, Díaz-Guilera A, Duch J, Arenas A, Comparing community structure identification, *Journal of Statistical Mechanics: Theory and Experiment* **2005**(09): pp. P09008–P09008 [2005].

[228] Gregory S, Finding overlapping communities in networks by label propagation, *New Journal of Physics* **12** [2010].

[229] Xie J, Szymanski BK, Towards Linear Time Overlapping Community Detection in Social Networks, *Advances in Knowledge Discovery and Data Mining*, pp. 25–36, Springer Berlin Heidelberg [2012].

[230] Rosvall M, Bergstrom CT, Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems, *PLoS ONE* **6**(4): p. e18209 [2011].

[231] Rosvall M, Bergstrom CT, Maps of random walks on complex networks reveal community structure, *Proceedings of the National Academy of Sciences of the United States of America* **105**(4): pp. 1118–23 [2008].

[232] Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S, Finding Statistically Significant Communities in Networks, *PLoS ONE* **6**(4): p. e18961 [2011].

[233] Onnela JP, Fenn DJ, Reid S, Porter MA, Mucha PJ, Fricker MD, Jones NS, Taxonomies of networks from community structure, *Physical Review E* **86**(3): p. 036104 [2012].

[234] Das Gupta K, Shakespear MR, Iyer A, Fairlie DP, Sweet MJ, Histone deacetylases in monocyte/macrophage development, activation and metabolism: refining HDAC targets for inflammatory and infectious diseases, *Clinical & Translational Immunology* **5**(1): p. e62 [2016].

[235] Millenaar FF, Okyere J, May ST, van Zanten M, Voesenek LACJ, Peeters AJM, How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results, *BMC Bioinformatics* **7**: p. 137 [2006].

[236] Affymetrix, GeneChip Expression Analysis: Data Analysis Fundamentals.

[237] Jutla IS, Jeub LGS, Mucha PJ, A generalized Louvain method for community detection implemented in MATLAB [2011].