Université de Montréal

Systematic prediction of feedback regulatory network motifs

par

Amruta Sahoo

Département de biochimie et médecine moléculaire
Faculté de médecine

Thèse présentée en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Bio-informatique

April 12, 2021

Université de Montréal

Faculté de médecine

Cette thèse intitulée

Systematic prediction of feedback regulatory network motifs

présentée par

Amruta Sahoo

a été évaluée par un jury composé des personnes suivantes :

Adrian Serohijos

(président-rapporteur)

Mohan Malleshaiah

(directeur de recherche)

Pascale Legault

(codirecteur)

Rafael Najmanovich

(membre du jury)

(examinateur externe)

(représentant du doyen de la FESP)

Résumé

Comprendre le câblage complexe de la régulation cellulaire reste un défi des plus redoutables. Les connaissances fondamentales sur le câblage et le fonctionnement du réseau d'homéostasie des protéines aideront à mieux comprendre comment l'homéostasie des protéines échoue dans les maladies et comment les modèles de régulation du réseau d'homéostasie des protéines peuvent être ciblés pour une intervention thérapeutique. L'étude vise à développer et à appliquer une nouvelle méthodologie de calcul pour l'identification systématique et la caractérisation des systèmes de rétroaction en homéostasie des protéines. La recherche proposée combine des idées et des approches issues de la science des protéines, de la biologie des systèmes de levure, de la biologie computationnelle et de la biologie des réseaux. La difficulté dans la tâche d'incorporer des données multi-plateformes multi-omiques est amplifiée par le vaste réseau de gènes, protéines et métabolites interconnectés qui se réunissent pour remplir une fonction spécifique. Pour ma thèse de maîtrise, j'ai développé un algorithme PBPF (Path-Based Pattern Finding), qui recherche et énumère les motifs de réseau de la topologie requise. Il s'agit d'un algorithme basé sur la théorie des graphes qui utilise la combinaison d'une méthode transversale de profondeur et d'une méthode de recherche par largeur ensuite pour identifier les topologies de sous-graphes de réseau requises. En outre, le fonctionnement de l'algorithme a été démontré dans les domaines de l'homéostasie des protéines chez *Saccharomyces cerevisiae*. Une approche systématique d'intégration des données de la biologie des systèmes a été orchestrée, qui montre l'identification systématique de motifs de rétroaction régulatrice connus dans l'homéostasie des protéines. Il revendique fortement la capacité d'identifier de nouveaux motifs de rétroaction réglementaire envahissants. L'application de l'algorithme peut être étendue à d'autres systèmes biologiques, par exemple, pour identifier des motifs de rétroaction spécifiques à l'état cellulaire dans le cas de cellules souches.

Mots clés: homéostasie, science des réseaux, motif de réseau, motif de rétroaction, biologie des systèmes

Abstract

Understanding the intricate wiring of cellular regulation remains a most formidable challenge. The fundamental insights into the wiring and functioning of the protein homeostasis network will help to better understand how protein homeostasis fails in diseases and how the regulatory patterns of protein homeostasis network can be targeted for therapeutic intervention. The study aims at developing and applying novel computational methodology for the systematic identification and characterization of feedback systems in protein homeostasis. The proposed research combines ideas and approaches from protein science, yeast systems biology, computational biology, as well as network biology. The difficulty in the task of incorporating multi-platform multi-omics data is amplified by the large network of inter-connected genes, proteins and metabolites that come together to perform a specific function. For my master's thesis, I developed a path-based pattern finding (PBPF) algorithm, which searches and enumerates network motifs of required topology. It is a graph theory based algorithm which utilizes the combination of depth-first transverse method and breadth-first search method to identify the required network sub-graph topologies. Further, the functioning of the algorithm has been demonstrated in the realms of protein homeostasis in *Saccharomyces cerevisiae.* A systematic approach of integration of systems biology data has been orchestrated, which shows the systematic identification of known regulatory feedback motifs in protein homeostasis. It claims the unique ability to identify novel pervasive regulatory feedback motifs. The application of the algorithm can be extended to other biological systems, for example, to identify cell-state specific feedback motifs in case of stem-cells.

Key words: homeostasis, network science, network motif, feedback motif, systems biology

Contents

List of tables

List of figures

List of acronyms and abbreviations

PQC	Protein quality control
PPI	Protein-protein interaction
GI	Genetic-interaction
GE	Gene-expression
SGD	*Saccharomyces* Genome Database
ND	Neurodegenerative disease
AD	Alzheimer's disease
PD	Parkinson's disease
PrD	Prion diseases
NGS	Next-generation sequencing
Hsp	Heat-shock protein

ER	Endoplasmic reticulum
UPR	Unfolded protein response
PBPF	Path-based pattern finding
PCs	Protein complexes
WGCNA	Weighted gene co-expression network analysis
MRKH	Mayer-Rokitansky-Küster-Hauser
CNS	Central nervous system
GO	Gene ontology
TF	Transcription factors
PFL	Positive-feedback loop
NFL	Negative-feedback loop
DNFL	Double-negative feedback loop

Remerciements

My journey through this project has been quite overwhelming. Through all the ups and downs, one thing that remained constant was ’learning’. There has been a plethora of instances that contributed to my growth, as a human as well as a student in science. Today, while writing this, I am taking a tour of the entire endeavor, and I am extremely thankful to a lot people, who helped me in different ways, and made this journey possible. I take this as an opportunity to thank each of them.

First of all, I would like to express my gratitude to my supervisor Mohan Malleshaiah, for accepting to support my project, for listening to my ideas and contributing with the invaluable discussions. I would like to thank all my lab members: Maxime Parisotto, Loick Joumier and Thulaj Meharwade, who helped me immensely through their questions to better communicate my work to people of different fields. I express my appreciations to my co-supervisor, Pascale Legault for helping me making the path towards finishing the project smooth through her constant effort to help me. I sincerely acknowledge the role of my previous supervisor, Sebastian Pechmann, for selecting me for this challenging project and allowing me to develop it and materialize it independently. I thank all my lab members from my previous lab: Musa Ozboyaci for making me acquainted with the life thousands of kilometres away from home; Nazli Kocatug for always being the go-to friend, and never ceasing to believe in me; A B M Shamim Ul Hasan for the mathematics discussions; Louis Gendron, Pedro Do Couto Bordignon, Savandara Besse and Yasmine Draceni for the interesting discussions during our daily group lunch.

There are people who helped me to a great extent to remain stable through this tupsy-turvy voyage and I wish to express my heartfelt gratitude to them. I thank Gunjan Dadhwal for becoming my family away from family; Pierre Dagenais for helping me believe in myself; Ali Mohamed and Geralyne Dionne for their kind words, just when I needed.

I will always remain indebted to my parents for inculcating scientific temperament of asking before accepting, which helped me immensely through the implementation of this project. I am grateful to my friends, who have always been in close contact with me through this. Lastly, but not the least, I thank my husband, Sidhartha Shankar Dash, for walking closely beside me, in all my decisions and tough times, without which, I would not have had the motivation and strength to complete this project.

Chapter 1

Introduction

Problem

Protein homeostasis is a complex regulatory network employed to maintain balance in the cell and is achieved through the integrated regulation of protein synthesis, folding, trafficking, aggregation, and degradation. Failure or dysregulation of protein homeostasis is associated with severe loss-of-function diseases (cystic fibrosis) and gain-of-toxic-function diseases (Alzheimer's, Parkinson's, and Huntington's disease). Thus, understanding the mechanisms of protein homeostasis will help us better understand and alleviate neurodegenerative diseases. In particular, while many protein quality control mechanisms have been studied in great detail, very little is known about how they integrate. The aim of my project is to systematically identify feedback systems in protein homeostasis in *Saccharomyces cerevisiae* (Yeast) by integrating different omics data to reveal fundamental insights into the wiring and functioning of the protein homeostasis network, which will eventually be helpful to better understand how protein homeostasis fails in diseases and how the protein homeostasis network can be targeted for therapeutic intervention.

Objectives

Our main objective is to propose a network-motif finding algorithm to identify putative functional feedback loops. Since searching network motifs is a computationally expensive task, the algorithm needs to focus on the type of network motif and optimise the search accordingly. The biological networks are usually huge in size with thousands of nodes and millions of interacting edges. Hence, we need to begin by restricting the search of motifs for only known functionally important proteins. To do so, the algorithm needs to provide flexibility of use.

Next, we seek to formalize and systematically establish the mechanistic basis of these feedback systems in yeast protein homeostasis. To identify feedback motifs in protein homeostasis, the first important criterion is that they must be connected. The aim of my project is to systematically identify feedback systems in protein homeostasis by integrating different omics data.

Experimental approach

I used bioinformatics and graph-theory to develop a search algorithm that identifies topologically connected motifs from interaction network. By systematically integrating biological information from available omics data, I aim to filter my search and add more confidence towards obtaining candidate feedback motifs which are potentially important in protein homeostasis. It is well-known that functional motifs have specific topology [**24**]. Therefore, the strategy is to make use of protein-protein interaction data to provide information on the physical connectedness between proteins. To reduce the computational expenses and obtain the most likely functional network motifs, list of proteins classified into known regulatory roles are to be used. The systematic integration of genetic-interaction data to add information on functional connectedness of the motifs is to be used to identify possible types of functional interaction (positive or negative). To further infer dynamics of the motifs, gene expression data is to be included and analysed in various ways to identify novel approach of integrating the expression data and better understand the feedback mechanisms. Thus, we hypothesize that systematic integration of omics data will help to identify key regulatory feedback motifs in protein quality-control systems.

Contribution

The study will provide a novel algorithm to integrate multi-omics data for predicting functionally relevant network motifs in protein homeostasis, which is anticipated to provide potential new and exciting connections between currently separate fields. Moreover, it will aid in better understanding how protein homeostasis fails in diseases and how the protein homeostasis network can be targeted for therapeutic intervention. Further, the analysis of the biological data over a general context and holistic study will provide an opportunity to better understand the significance of the information. Moreover, the simplicity and flexibility of the novel algorithm allows application on a broad range of data and can be extended to different usages in multiple fields.

Organisation of the thesis

The master's thesis is organized into the following six parts. The first chapter presents a brief introduction, the research objective and a concise mention of the experimental approach of the project and its perspectives in the field.In the next chapter, we make a critical synthesis of previous publications related to the subject of this thesis, from a biological point of view (2.1) and in bioinformatics (2.2). In the third chapter, we formulate the hypothesis and discuss the objectives. The fourth chapter contains the materials and methods in which we discuss the datasets used and the approach to be used for the study. The fifth chapter includes the results.Finally, the last chapter presents a general discussion which includes the conclusion and the future aspects of the project.

Chapter 2

Literature review

2.1. Biological aspects

Biological systems are characterised by extremely complex architecture involving enormous number of factors that carry out highly complicated, yet balanced inter-regulations to maintain a proper function. For a very long time, biological studies focused on the individual aspects of a biological system, leading to deep understanding of each of them and generation of ample genomics, transcriptomics and proteomics data. Because of their intrinsic strengths and weaknesses, no single approach can fully unravel the complexities of fundamental biological events. However, an appropriate combination of different information could lead to integrative analyses that would furnish new insights not accessible through a one-dimensional dataset. Further, we understand that the viability of an organism is contributed by the proper functioning of each of the involved factors, and the cross-communication between them. Also, we realise that the design principles of complex systems, in general, share universal design patterns which form the essential elements for building successful complex systems that can function, compete, survive, reproduce and evolve for long periods through multiple generations towards increased fitness and overall growth.

The aim of this thesis is to develop a framework for understanding regulation in complex systems, specifically, the goal is to identify regulatory motifs in biological systems. For the study, we have considered protein homeostasis to apply the pipeline on and perform the analysis in the realms of protein homeostasis, its functioning, and failure.

In this section, we introduce the basics of the protein quality control systems involved in protein homeostasis. More specifically, we present how the crosstalk between the protein quality control methods are achieved through integrated regulation of cellular pathways that govern protein synthesis, folding, trafficking, and degradation [**91**]. It forms a complicated network in the model organism which is a single-celled eukaryote [**20**], *Saccharomyces cerevisiae.*

2.1.1. Protein homeostasis

Homeostasis, which is a Greek word meaning "same" or "steady," was first coined by the physician Walter Cannon, in 1930 in his book, The wisdom of the body [**17**]. The term homeostasis refers to the process of maintaining a stable condition in living organisms that is necessary for their survival. Here, we focus specifically on the homeostasis of the proteome. The term protein homeostasis or proteostasis is attributed to a pervasive process that operates to maintain the stuctural and functional features of the proteome by preserving the correct concentration, conformation, and subcellular location of the proteins. It works through a coordinated system of protein synthesis, repair, and degradation where the state of dynamic equilibrium between protein synthesis and protein folding is balanced by degradation [**1**].

Proteins are made up of amino acids which are joined by peptide bonds to form a long chain of amino acids, which constitutes the primary structure of the protein. The systematic unique folding of these poly-peptide chains to form three-dimensional shapes, stabilizing their structure. This allows the exposure of only the specific sticky (hydro-phobic) positions, creating the final structure of the protein which is well-adapted for their functions.

Proteins continuously face perturbations increasing the risk of the organism to enter a number of disease states. As shown in figure 1, a failure in protein homeostasis leads to misfolding, acculmulation of damage, aggregation, and toxicity and disease [**26**]. Examples of neurodegenerative diseases (NDs) include formation of intracellular inclusions containing aggregated -synuclein in Parkinson's disease, huntingtin in Huntington's disease and extracellular -amyloid plaques in Alzheimer's disease [**123**]. NDs further include amyotrophic lateral sclerosis, spinocerebellar ataxias, frontotemporal dementia, corticobasal degeneration, progressive supranuclear palsy, chronic traumatic encephalopathy, multiple system atrophy, dementia with Lewy bodies, and prion diseases (PrD) [**110**]. NDs are related to oxidative stress [**26**], ER stress [**67**], apoptosis [**30**]. Moreover, proteins can also behave like an infectious pathogen and cause neurodegenerative (ND) diseases by degenerating the central nervous system [**92**]. Parkinson's disease (PD), a motor and cognitive neurodegenerative disorder for which an environmental exposure to the manganese (Mn) metal is a prominent risk factor has been shown to be altering the iron (Fe) and calcium (Ca) homeostasis, thereby interfering in the mitochondrial and ER health leading to accumulation of ER unfolded proteins and which may further compound toxicity [**5**]. To date, at least 37 proteins have been discovered which aggregates to form amyloids, in different organs of the body, like, the brain in AD, the pancreas in type II diabetes, the liver and the heart in systemic disorders [**77**]. Further, 7 out of the 37 have been found to form deposits in the central nervous system (CNS), giving rise to neurodegenerative conditions, whereas, 15 form aggregates in the tissues of heart, spleen, liver, and kidney, while the remaining 15 aggregate in specific tissues, causing conditions including, type II diabetes, atrial amyloidosis etc [**20**]. Amyloid-related diseases can be sporadic or hereditary in nature, which shows a late age of onset. It suggests that the process of protein aggregation and the onset of symptoms is closely connected to the loss of regulatory control.

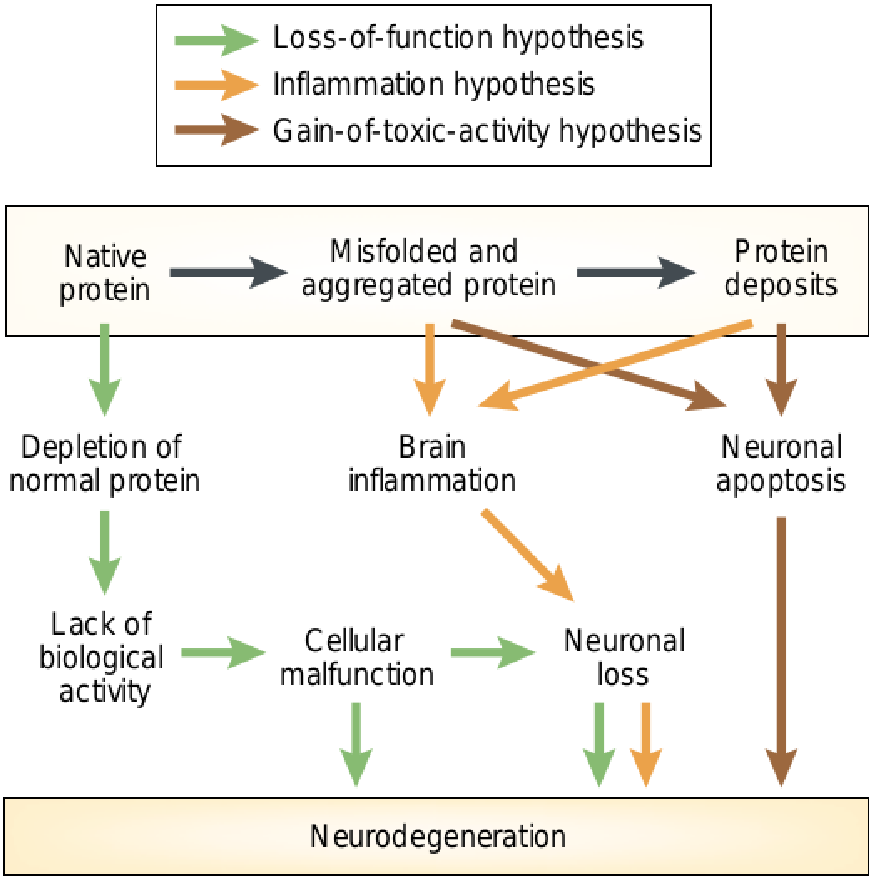

Loss-of-function hypothesis
Inflammation hypothesis
Gain-of-toxic-activity hypothesis
Native protein
Misfolded and aggregated protein
Protein deposits
Depletion of normal protein
Brain inflammation
Neuronal apoptosis
Lack of biological activity
Cellular malfunction
Neuronal loss
Neurodegeneration

Fig. 1. Models for the mechanism of neurodegeneration associated with protein misfolding and aggregation [**109**]

The figure 1 suggests three major models of possible pathways that leads to neurodegeneration. One of them is the **loss-of-function hypothesis**, a clinical condition that is related to specific mutation in which there is an alteration of gene product that lacks the molecular function of the wild-type gene. Series of such mutations could lead to non-functional misfolded proteins that causes cell impairment, further leading to neuronal loss or death. Accumulation of such misfolded (non-functional) proteins leading to inflammation in the regions of brain leading to neuronal apoptosis or neural loss is known as **inflammation hypothesis** [**109**]. Further, another pathway can be as simple as aggregation of misfolded proteins causing necrosis, which is called **gain-of-toxic-activity hypothesis** [**109**]. The predominant feature of these disorders is protein misfolding as manifested by the formation of intracellular and/or extracellular deposits of aggregated proteins.

2.1.1.1. *Saccharomyces cerevisiae*: an ideal model organism to study protein homeostasis.

Yeast is structurally simple and share similarity of essential cellular processes with humans. Further, the list of homologous proteins involved in neurodegenarative diseases of humans are known to be present in yeast. This remarkable homology despite of the huge evolutionary gap of millions of years is the reason why the study of yeast carries immense importance for the treatment of human diseases [**51**]. Also, the basic mechanisms and pathways underlying neurodegenerative diseases, such as mitochondrial dysfunction, transcriptional dysregulation, trafficking defects and proteasomal dysfunction, are remarkably well conserved between humans and yeast, and thus, study of yeast would provide molecular insights to the problem.

2.1.1.2. Protein quality control (PQC) systems.

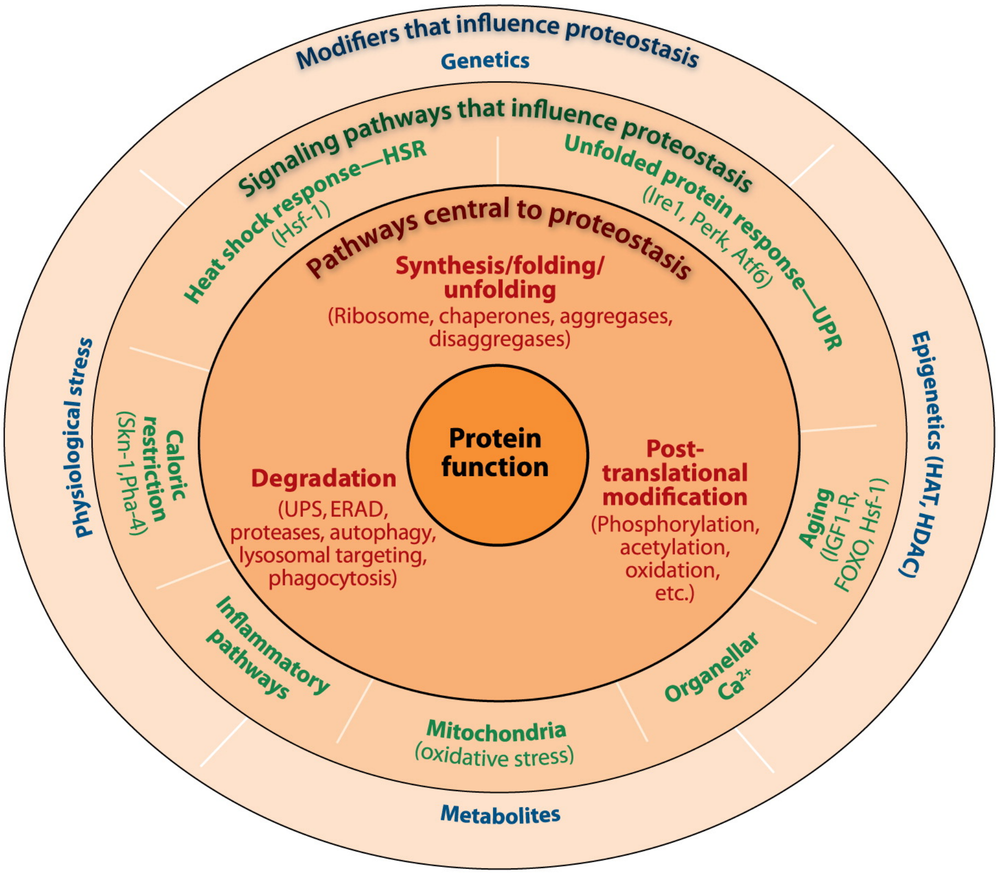
The cell of an organism contains several billion protein molecules, and to synthesize a proteome of this magnitude, millions of ribosomes work to translate the codons. Despite the complexity of mRNA synthesis, the chemical synthesis of proteins remains remarkably efficient: the vast majority of polypeptides that are produced from a single mRNA are perfect chemical copies of each other. By contrast, both the folding and maintenance of proteins in their functional, native, 3D conformations frequently fails. Quality control of the proteome is made more difficult by the high degree of heterogeneity present across populations of proteins, which prevents them from fitting into standardized categories of size, shape, or stability.

Organisms constantly face unexpected challenges with adverse environmental and cellular conditions, which have led to evolution of mechanisms to cope and adapt. One such a mechanism in protein homeostasis is the protein quality control system that comes to its rescue where molecular chaperones assist the (re)folding of proteins, facilitate their translocation and assembly into macromolecular complexes, and if necessary, target proteins for degradation [**38**], hence playing the central role in protein homeostasis. The macromolecular assistance helps in preventing protein aggregation and promotes its functionality, by controlling the rate of protein synthesis and influencing the folding-rate of the proteome. It also controls membrane trafficking patterns responsible for the compartmentalized localization, and initiates degradation when required to mitigate aggregation and enable protein turnover.

Multiple regulatory mechanisms are involved in the process of maintaining the quality of thousands of proteins in a cell. This complicated process is achieved through the cross-talk between the protein quality control methods. A perturbation or stress that disturbs the balance in protein homeostasis acts as a signal which is sensed, processed and a response is initiated to bring back the balance in the biological system to ensure the proper functioning of protein homeostasis.

In figure 2, different layers of proteostasis network is illustrated. It is demonstrated through three layers; in the first layer, there exist synthesis and maintenance (folding/unfolding) of the proteins which is directed by ribosome, chaperones, aggregases, and disaggregases, post-translational modifications that include phosphorylation, acetylation, oxidation etc., and degradation which consists of different pathways like, ubiquitin-proteasome system (UPS), endoplasmic reticulum (ER)-associated degradation (ERAD) systems, proteases, autophagic pathways, lysosomal/endosomal targeting pathways, and phagocytic pathways. Further, the second layer is made up of components that influence the first layer and represent the signalling system in proteostasis. The third layer makes the genetic and epigenetic pathways, physiologic stressors, and intracellular metabolites which affect the activities of the second and first layers.

The maintenance of PQC is carried out by complex regulatory homeostasis processes, which senses the folding state and signals for its modification required for proper functioning. In the Salvage pathway, to regulate the turnover of complex I (the largest mitochondrial respiratory chain complex) the NADH-oxidizing N-module of complex I is highly expressed. The mitochondrial matrix protease ClpXP plays an important role to recognize, disassemble, and rapidly degrade impaired core N-module proteins to safeguard against the accumulation of dysfunctional complex I [**112**]. Protein ubiquitination and SUMOylation have long been known to control regulated protein turnover in addition to labeling damaged proteins for degradation, thus regulating cell physiology [**40**]. Moreover, in case of interferon-induced oxidative stress, which leads to formation of defective ribosomal products and formation of aggresome-like induced structures are degraded by an up-regulation of immunoproteosome-induced ubiquitylation machinary to preserve cell viability [**100**]. In contrast, chaperones have been primarily associated with their roles in assisting the folding of proteins and



Powers ET, et al. 2009.
Annu. Rev. Biochem. 78:959–91

Fig. 2. Management of proteostasis

preventing their aggregation. By selectively stabilizing client proteins, and under dynamic cellular control, chaperones also fulfill important regulatory functions, which are however poorly characterized. An example is given by the chaperone controlled toxin-antitoxin system in the human pathogen M.tuberculosis where under stress the chaperone capacity is exhausted, which leads to destabilization and subsequent degradation of the anti-toxin as a protective mechanism [**98**]. Hsp70 co-chaperones can modulate spindle elongation [**72**], and steroid hormone receptor function [**54**]. Moreover, Hsp90 has been found to bind to many important regulatory and signaling proteins. In the pathogen C.albicans, a distinct role of Hsp90 to regulate temperature-dependent morphogenesis and cell cycle progression [**101**], chromatin architecture [**62**], and cell-wall remodeling [**61**] has been demonstrated. However, while broad consequences upon major perturbation of central and well-connected chaperones such as Hsp90 provide fundamental insights into potential mechanisms to target e.g. pathogens or cancers, a systematic understanding of how the different protein regulatory systems communicate to regulate important cellular pathways and networks remains completely lacking.

2.1.2. Feedback regulation in biological systems

Feedback occurs when outputs of a system either acted as or regulate inputs as part of a chain of cause- and-effect that forms a circuit or loop. A feedback system constitute of a detection mechanism which senses signal, a controller which measures the deviation and helps in deciding on which direction the system should respond and a mechanism which acts on the decision made. Feedback systems are omnipresent: both in natural and man-made systems. A thermostat is often cited as a simple example of feedback control; the device measures the temperature, compares that temperature to the desired set-point, and assesses the deviation to generate a control action e.g. to turn heat on when the temperature is too low and to turn it off when the temperature is too high. The mechanism of regulation of heat shock proteins (hsp) in a cell is similar in nature, where environmental stress acts as a signal which is detected by an increase in the rate of protein denaturation and compared to the amount of required functional protein as a control, to increase the expression of hsp-producing gene as an appropriate action.

Biological networks are composed of complicated interconnections among network modules and some subnetworks carrying out specific functions are often identified as network motifs [**48**]. Among such network motifs, feedback loops are thought to play an important role in decision making [**12**] to maintain the dynamic equilibrium of various biological processes such as cell fate specification, embryogenesis, circadian rhythms, cell cycle, and blood clotting, etc. [**52, 80**]. The genetic interaction data has been used to understand the biological details by using network biology. Triplet motifs are analyzed for the study of transcription-factors and target protein kinases [**43**]. Another study using triplet motifs is done for transcriptional networks where they tried to interpret the pattern based on genetic co-regulation [**57**]. Seminal work has established how biochemical networks are composed of distinct network motifs that define these feedback systems. Accordingly, the decomposition of complex regulatory networks into more tractable subnetworks [**106**] and network motifs [**117, 52, 27**] has revealed fundamental insights into the mechanistic basis of complex cellular processes such as adaptation [**70**].

2.1.2.1. Types of feedback systems in biological systems.

In a biological network, a feedback loop is a connected or interacting pattern between the network modules. A feedback loop in a cell governs the physiological responses of cells to external and internal stimuli. The cyclic topology of feedback systems gives rise to a non-linear type of interaction or argument which makes the response non-trivial. To maintain homeostasis, different types of feedback loops are may exist in a cell at different levels of cellular processes. The most common types of feedback loops are listed below 3.

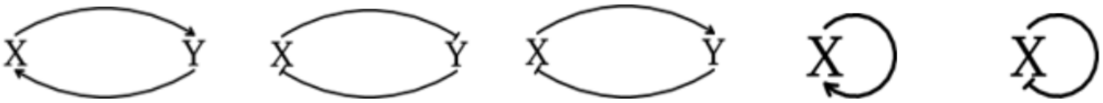


Fig. 3. Types of feedback motifs; from the left: PFL, DNFL, NFL and Self – regulation (positive and negative)

(1) **A positive feedback loop (PFL)**, in which the input-signal and output-response amplify each other in the recurring cycles. It is commonly observed in the case of cellular memory [5],[6] and differentiation [7]. A PFL has a tendency to amplify noise, also the time taken to reach the steady-state protein level is longer than in the case of an unregulated gene. The Rb-E2F-CycE system, that has role in cell-cycle, is an example of a three-component PFL [**16**].

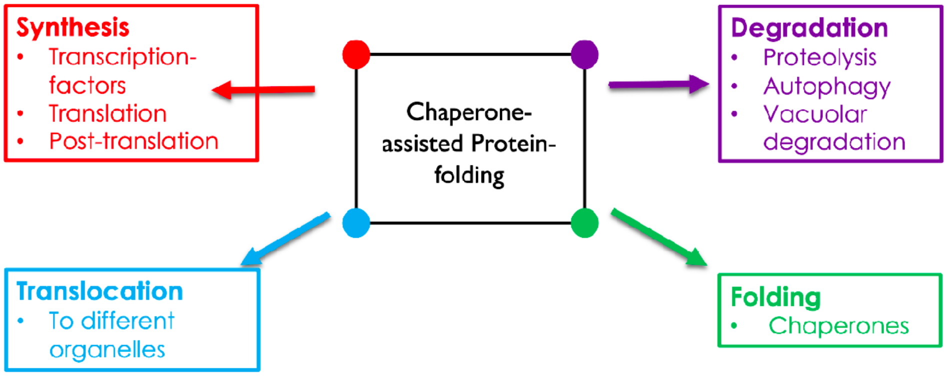
(2) **A double-negative feedback loop (DNFL)**, in which the input-signal and output-response damp each other in the recurring cycles. For example, the p53-dependent response to DNA damage in mammals involves the inactivation of p53 by binding to Mdm2, among other things [**81**]. One of the commonly observed examples for DNFL is, transcription-factor (TF) suppressing a miRNA whereas the TF itself is negatively regulated by the miRNA [**78**]. It tends to have a bi-stable region that can switch from one steady-state to the other by choosing the appropriate extrinsic noise source [**15**].

(3) **Negative feedback loops (NFL)** is a type of feedback in which amplification of one leads to dampening of another and vice versa. The mechanism that regulates protein synthesis involve binding of TF to DNA, mediated by enhancers and cofactors, is observed to be favouring the catalysis of the proteolytic inactivation of the TF in case of stress response mechanism that [**88**]. An NFL is known as a balancing loop, and it may be common to see oscillations in which a delayed negative feedback signal is used to maintain homeostatic balance in the system [**84**].

(4) **Self-regulation** can be either positive or negative, and can involve either a direct or indirect (multi-path or intermediate) regulation. One of the examples is the polymerization (positive self-regulation) and depolymerization (negative self-regulation) of actin filaments [**63**]. Positive autoregulation produce a switch-like response by increasing the sensitivity to signals. This leads to threshold-associated response in the system which helps to control dynamic stability by slowing down the response time and enhancing the variation [**31**]. Negative autoregulation on the other hand, speeds up the response time in gene circuits and promotes robustness to fluctuations in production rate [**97**].

2.1.2.2. Feedback regulation in PQC systems.

At the heart of the functioning of the PQC system, lies the intricate balance between failure in the maintenance of proteins and their recovery or removal from the system. In the PQC, there exists a constant functional dynamics (figure 4) of misfolding of proteins, which comes with a burden on protein homeostasis network and the processes that come together in its rescue by increasing the amount of chaperone in the system to enhance the possibility of re-folding of the protein, increase ER export-signal to mediate exit of the misfolded proteins from the ER [**53**], and increase proteases to direct degradation of the proteins in case it could not be rescued. Further, as a feedback, the synthesis of the protein decreases to stop misfolding and aggregation of proteins.


Synthesis

- Transcription-factors
- Translation
- Post-translation

Chaperone-assisted Protein-folding

Degradation

- Proteolysis
- Autophagy
- Vacuolar degradation

Translocation

- To different organelles

Folding

- Chaperones

Fig. 4. Folding-centered protein homeostasis

2.2. Bioinformatics framework

2.2.1. Network science

"I think the next century will be the century of complexity."
Stephen Hawking

To ensure our biological existence, the pre-requisite is the cooperative organization between the thousands of genes, proteins, and metabolites to form a cellular network. Our qualification to cognize comes from the systematic connection between billions of neurons, which decides the functioning of the brain. The society that we live in, requires cooperation and communication between billions of individuals. We interact with each other using communication devices, through wired internet connections or wireless links, that integrate billions of cell phones with computers and satellites. The transportation network: roadways, railways, airways, the electricity networks that involve generators and transmission lines, etc. are all important factors of our lives which work on the basis of network. The bottom line is, we are fundamentally surrounded by complex systems.

At the heart of complex systems, lies a network. A network is an emergent property i.e. a behavior that arise from a collective functioning of a system, but do not belong to any one part of that system. This makes the understanding of a network highly non-intuitive. The presence of such complex systems that play an important role in our daily lives makes it highly essential to have a set of methods to analyze, understand, and eventually be able to control such systems. Hence, the emergence of network science: *the science of the 21st century*, came into existence [**8**].

Fundamentals of Network science: A network is a representation of binary relationship between components of the system. In network science, the individual components are called **Nodes** and the interactions between the nodes are represented by **Links**. The analysis of the network properties can be categorized into the following three abstraction levels:

(1) **Element-level analysis:** It involves methods for identification of the most important nodes of the network. This includes, centrality measures such as degree centrality, eigenvector centrality, closeness centrality and betweenness centrality.

(2) **group-level analysis:** It comprises of methods to identify the closely connected groups of nodes in the network. It deals with calculation of densely connected groups

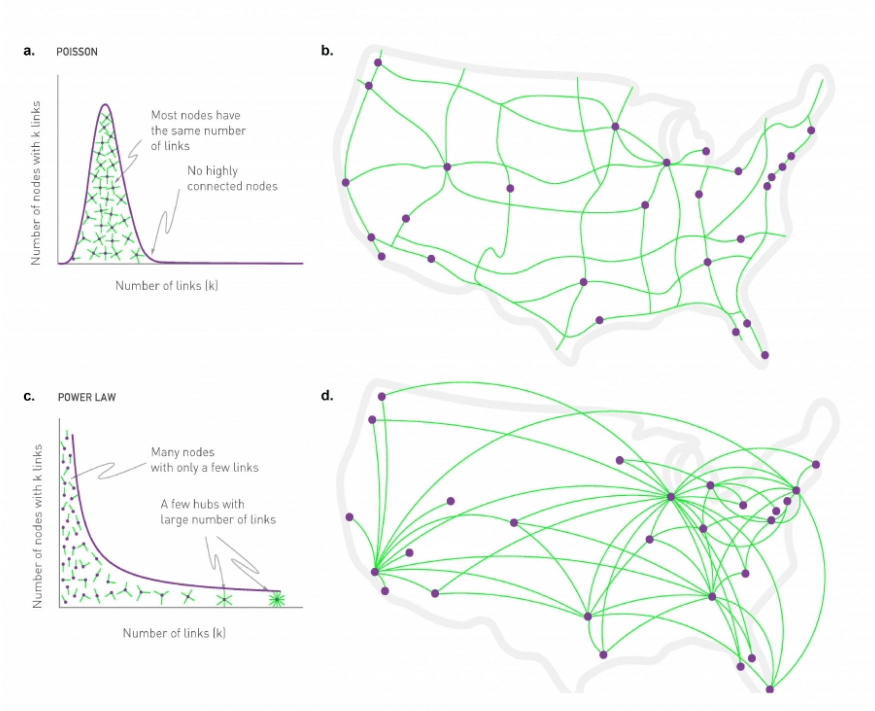

a.
POISSON
Number of nodes with k links
Most nodes have the same number of links
No highly connected nodes
Number of links (k)
b.
c.
POWER LAW
Number of nodes with k links
Many nodes with only a few links
A few hubs with large number of links
Number of links (k)
d.

Fig. 5. Random versus Scale-free network [**8**]

(clustering) and the computation of structural roles and positions.

(3) **network-level analysis:** The methods focuses on the topological properties of networks as a whole. Properties such as density, degree-distributions, transitivity, or reciprocity falls under this category.

As seen in figure 5(a), the degree distribution in case of a random graph follows a Poisson law, which suggests a more or less similar or small range in the variance of degree of the nodes in the graph. The same can be visualised through the national highway network in the United States, in the figure 5(b); the nodes are connected directly or indirectly, and it is hard to identify a hub among them because of the absence of highly connected nodes. On the contrary, the degree distribution of a scale-free network follows a power-law distribution (figure 5(c)), and some of its nodes have a very small degree whereas some are almost connected to all the nodes in the graph (figure 5(d)), as observed in the air traffic network in the United States.

2.2.2. Network biology

Biological entities are made up of intricately connected complex interactions. Network biology provides the framework to investigate the complex relationships and model them to characterize the enrichment patterns and understand the system-wide properties. The integrative and systems level approach of network science to uncover the informations administered by biology is highly significant. The complexity of biological phenomenons can be broadly represented as: i) Protein-protein interaction networks, where proteins are nodes and their interactions are edges, ii) Gene regulatory networks (DNA–protein interaction networks), iii) Gene co-expression networks (transcript–transcript association networks), iv) Metabolic networks, v) Signaling networks, vi) Neuronal networks, vii) Food web, viii) Between-species interaction networks, and ix) Within-species interaction networks.

Over the last decade, the advances in high-throughput genomic and proteomic profiling technologies, such as DNA microarrays, next-generation sequencing (NGS) and mass spectrometry based proteomics and metabolomics, has led to rapid developments in data acquisition of all kinds, namely, genomic, transcriptomic, proteomic and metabolomic data [**85, 68, 86**]. Although it gave a tremendous opportunity to better understand biological systems, there were huge challenges of bringing together the heterogeneous data and put in a framework that has the capacity to capture the high-dimensionality of the data [**104**]. The origin and evolution of network biology came handy as a tool with a holistic approach, with an ability to integrate multi-platform, multi-dimensional data, graphical(wiring diagram) representation to exhibit the interactions and functions that connects the dots to identify the fundamental mechanisms of biological systems [**120**].

Network biology has proven to be a powerful tool for the study of biological systems. It has been beneficial in resolving the inherent mechanisms of biological problems, starting from the basal activities of genes and proteins in cells [**8, 37**], to complex diseases like cancer [**129**], cardiovascular diseases [**22**] and neurodegenerative diseases [**76**]. Tools using network analysis, that uses integration of co-expression network and the human interactome network has been developed and shown to have efficient screening capacity to identify potential new disease- gene associations [**87**]. Single cell network biology has shown to identify regulatory programs specific to disease-associated cell types and cellular states using cell-type-specific gene networks [**18**]. A regulatory multi-omic network study has found the connection between lipid metabolism and glucose regulation in case of coronary artery disease [**22**]. The etiologies of a rare disease, Mayer-Rokitansky-Küster-Hauser (MRKH) syndrome has been identified using PPI network analysis of protein-coding genes found in the altered genomic regions [**90**]. To identify the underlying mechanisms behind biological functions, for example, the lactation process has been studied using weighted gene co-expression network analysis (WGCNA) [**29**]. The evolution of transcription network has been done using network analysis to identify the overlapping regulatory pathways [**108, 102**]. Fundamental studies like, the study of design principles in inter-layered biological networks in nine different species has been investigated using robustness analysis of the networks [**71**]. The study potentially provides a framework that can be extended to different molecular species to achieve a desired biological function.

2.2.2.1. Network biology in PQC systems.

Under certain stressful conditions protein folding can fail leading to misfolding or aggregation. There is a protein quality control system that comes to its rescue where molecular chaperones assist the (re)folding of proteins, facilitate their translocation, promote their assembly into macromolecular complexes, and if necessary, target proteins for degradation [9], hence playing the central role in protein homeostasis. Such macromolecular assistance to prevent aggregation and promote function controls rate of protein synthesis and influence the rate of folding of the proteome. In addition, membrane trafficking patterns responsible for compartmental localization can mitigate aggregation and mediate degradation to enables protein turnover.

The intricate play of inter-connections in PQC can be represented as network of proteins, where the connections can be defined as physical or functional in nature. An elaborate network of molecular chaperones and protein degradation factors continuously monitor and maintain the integrity of the proteome. Cellular protein quality control relies on three distinct yet interconnected strategies whereby misfolded proteins can either be refolded, degraded, or delivered to distinct quality control compartments that sequester potentially harmful misfolded species 4.

2.2.2.2. Modularity of PQC network.

One issue that has received a considerable amount of attention is the detection and characterization of community structure in networks, meaning the appearance of densely connected groups of vertices, with only sparser connections between groups (figure 6). The ability to detect such groups could be of significant practical importance [**82**].

On a mesoscopic level of protein-protein interaction networks, modules form the next functional and organisational building blocks above individual proteins. Therefore, modules in the network play an important role in getting insights into fundamental organizational principles of biological systems. Mathematically, there are several algorithms available to identify network modules, such as, Fast-greedy algorithm [**21**], Walktrap algorithm [**89**], Label propagation algorithm [**93**], Spinglass community algorithm [**94**], Multi-level community algorithm [**11**] and Correlation based hierarchical clustering [**115**]. According to the definition of modules (i.e. modules are functional building block of the network), we can also consider proteins involved in a particular function to constitute a module. These proteins inter-play between each other and with other functional modules to carry out

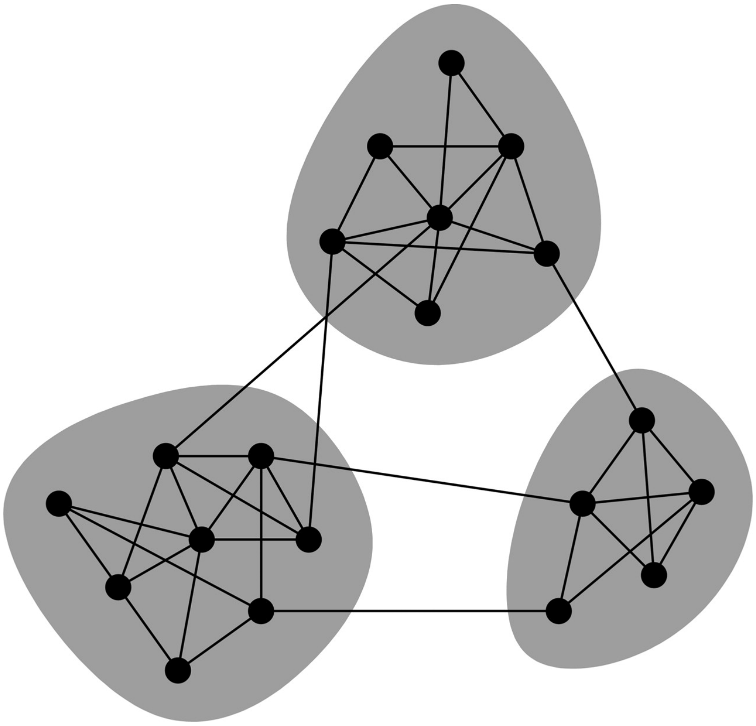


Fig. 6. Modularity in a network
The vertices in many networks fall naturally into groups or communities, sets of vertices (shaded) within which there are many edges, with only a smaller number of edges between vertices of different groups.

particular functions.

The interactions within and between proteins and metabolites in complex metabolic networks has been studied for different organisms, in the perspective of robustness of the network offered by its modular structure, to identify indispensable links present in the network [**35**]. These metabolic interactions are used as potential drug targets [**42**]. Yeast protein interactome has been studied in the context of modularity of the protein interaction network to understand intra-modular inetractions [**25**]. There has been successful application of a framework, MoNet, on yeast protein interaction weighted network to understand the organization of the biological system, where the modules obtained were found to be significantly enriched in proteins with related biological process Gene Ontology terms [**69**].

2.2.2.3. Functional module discovery in PQC network.

A functional module is nothing but a sub-network defined by a group of cellular components and their interactions that can be attributed a specific biological function [**39**]. There have been approaches that includes scoring functions to identify modules in gene-expression data [**130, 59**], metabolic pathways [**46**] and complexes [**44**]. The identification of functional network modules using network topology has been done in several ways; algorithms involved utilization of shortest-path length, betweenness, density of connections, and percolation clustering.

In PQC, the different processes of synthesis, which includes transcription, translation, folding, maintenance that consists of signal trafficking, re-folding and degradation cross-communicate in a huge network. In the PQC network, the cellular entities that govern the individual processes connect more densely and closely than their connection with each other. While studying the regulation of the network that is made up of thousands of nodes and edges, it is difficult to consider the enormous magnitude of factors and information. The modularity of the network can be played to reduce the dimensionality of the network without losing the information the network data offers.

2.2.3. Feedback motifs in PQC network

The term 'motif' is used in a variety of contexts conveying varied meanings in each field. The origin of the term dates back to 1848, which was used back then in French, specifically in an artistic or dramatic work where it meant "dramatic idea or theme". The most fundamental abstract definition of motif is, **a frequently occurring pattern with a salient significance**. The term carries its significance from arts, music and literature, to computer science and biochemistry.

In biochemistry, there are generally two types of motifs; (i) Sequence motifs and (ii) Structural motifs. Further, depending on the content of the sequence i.e. nucleotide or amino-acid, it can be DNA or RNA, or protein motif. The well-known regulatory sequences [**119**] such as transcription-factor binding sites or binding sites in general [**96**], promoter regions, and crossed-species conserved sequences are some of the examples of sequence motifs. It is clear from the examples that sequence motifs are generally functionally significant in nature [**127, 45**]. The discovery of sequence motifs is an important problem in bioinformatics. Among the currently available search algorithms, the Boyer-Moore algorithm, the Rabin-Karp, and the suffix trees are some of the most commonly used ones. The problem of sequence motif identification, in general, is based on sequence alignment, and the graphical representation of sequence motif is mostly done through sequence logo. The structural motif is a recurring patterns of three-dimensional spatial arrangement of amino acids or nucleotides [**56, 83, 74**]. A structural motif may [**126**] or may not be associated with a sequence motif. Stem-loop, D-loop, beta-hairpin, helix-turn-helix, and zinc finger are examples of structural motifs.

In the realm of biological networks, there exist network motifs that are nothing but repeatedly occurring statistically significant sub-graphs, with a frequency higher than the compatible randomized networks. The term, network motifs has been introduced by the group of Uri Alon [**75**] to better understand biological networks. The network motifs can be regarded as the structural and functional units of a network. They are known to have regulatory properties [**7, 49**]. As part of the network, the network motifs can be directed or un-directed, owing to the type of network it comes from. The detection of network motifs is computationally challenging. Several algorithms has been developed for the identification of specific network motifs, which is discussed in the next section.

2.2.3.1. Network motif discovery.

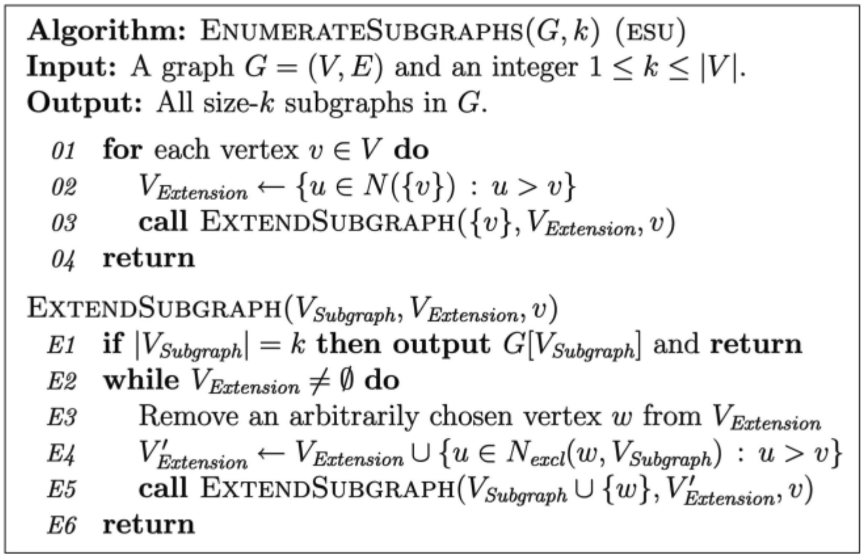
The task of searching and enumerating network motifs is time consuming due to their potentially large number, and the time consumption increases exponentially with an increase in the size of motif [**125**]. Also, it gets practically impossible for large and dense networks, for example, networks of biological systems. To tackle this problem, various methods has been developed, which utilizes different computational techniques like breadth-first search algorithm and isomorphism. The core goals of most of the motif finding algorithm can be summarized as follow:

(1) Search and enumeration of network subgraphs of certain size for a given input data.
(2) Classification of isomorphic network subgraphs.
(3) Calculation of frequency of network subgraphs in the network to identify network motif.

Below we discuss some of the well-known motif finding methods.

(1) **FANMOD:** Wernicke's ESU algorithm: Wernicke [**124**] presented the ESU algorithm, which enumerates all motifs of size k in a graph. The ESU algorithm starts with individual nodes in the graph and iteratively adds an additional node from the subgraph's neighborhood, until reaching subgraphs of size k (figure 7). It does consider open neighbourhood vertices in its graphlet search. Hence, the goal is to search and enumerate all the possible topologies for a given size of motif. (figure 8)

(2) **Mfinder:** The main aim of the algorithm used in MFINDER is to reduce the run time. In order to count all n-node they start with an edge, e1, and search for all



```
Algorithm: EnumerateSubgraphs($G, k$) (esu)
Input: A graph $G = (V, E)$ and an integer $1 \leq k \leq |V|$.
Output: All size-$k$ subgraphs in $G$.
    for each vertex $v \in V$ do
        $V_{Extension} \leftarrow \{u \in N(\{v\}) : u > v\}$
        call ExtendSubgraph($\{v\}, V_{Extension}, v$)
    return

ExtendSubgraph($V_{Subgraph}, V_{Extension}, v$)
    if $|V_{Subgraph}| = k$ then output $G[V_{Subgraph}]$ and return
    while $V_{Extension} \neq \emptyset$ do
        Remove an arbitrarily chosen vertex $w$ from $V_{Extension}$
        $V'_{Extension} \leftarrow V_{Extension} \cup \{u \in N_{excl}(w, V_{Subgraph}) : u > v\}$
        call ExtendSubgraph($V_{Subgraph} \cup \{w\}, V'_{Extension}, v$)
    return
```

Fig. 7. Pseudocode for ESU algorithm: Motif finding algorithm used in FANMOD [**124**]

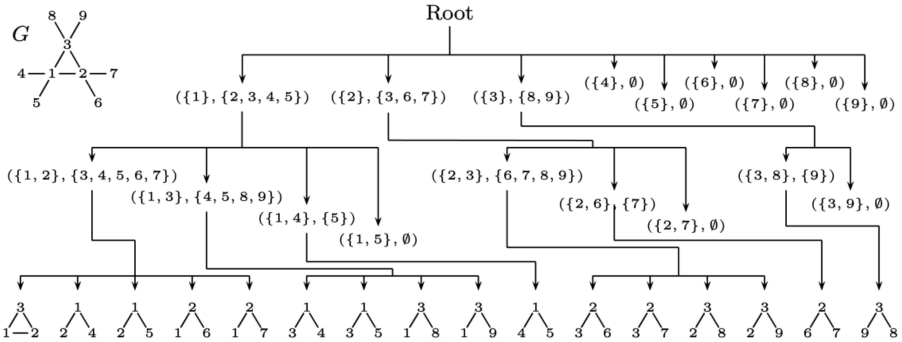


Fig. 4. Given the labeled graph in the left upperhand corner, the above ESU-tree corresponds to calling ENUMERATESUBGRAPHS$(G, 3)$. The tree has 16 leaves which correspond to the 16 size-3 subgraphs of G.

Fig. 8. ESU implementation for G(3) [**124**]

n-node subgraphs it participates in. All of the sets of nodes that were already visited are then stored in an array of hash tables. This saves time in searching because the traversing of a tree is stopped if a set or subsets of nodes have already been visited. When this process of edge e1 is finished, the hash tables are cleared and the next edge in the network, e2 is then searched. This process is repeated for all network edges and accumulates the counts for each subgraph type. At the end of this process each subgraph count is divided by the number of edges that the subgraph contains [**113**].

(3) **RAGE:** It is an efficient counting algorithm for 4-node size graphlets. Specifically, it present algorithms that count for each node, all non-induced tailed triangles, 4-node

cycles with chord (chordal cycles), and a path of length three graphlets [**73**].

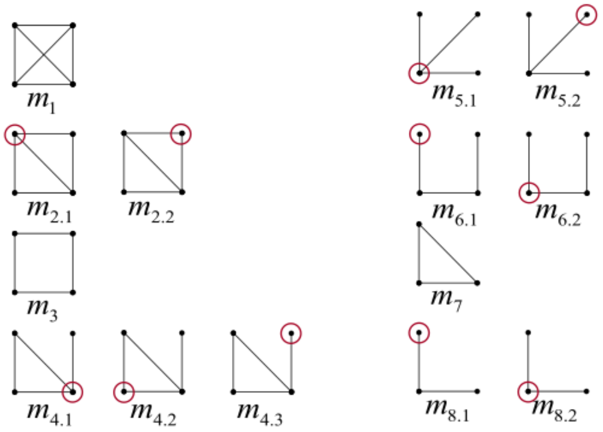


Fig. 9. All three and four node undirected position aware graphlets found using RAGE [**73**].

(4) **ORCA:** It uses a combinatorial method for counting graphlets and orbit signatures of network nodes. The algorithm builds a system of equations that connect counts of orbits from graphlets with up to five nodes, which allows to compute all orbit counts by enumerating just a single one. This reduces its practical time complexity in sparse graphs by an order of magnitude as compared with the existing pure enumeration-based algorithms (figure 10) [**41**].

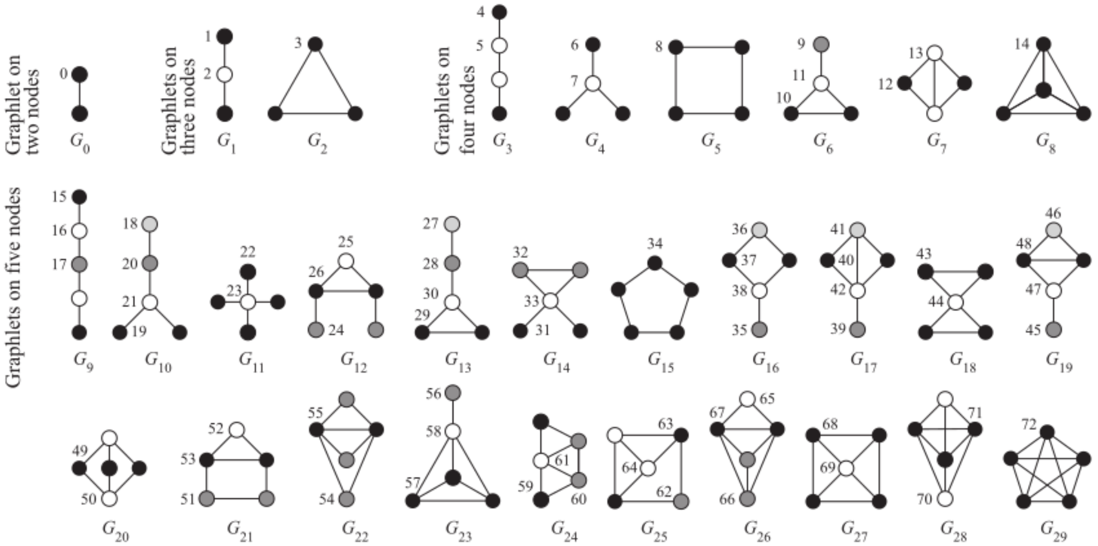


Fig. 10. Graphlets searched in ORCA with 2–5 nodes and automorphism orbits [**41**].

(5) **MAVisto:** It is meant to support both the search for motifs of any size under different frequency concepts (that is different ways of counting motif occurrences depending on the reuse of network elements) and powerful exploration of motif distribution and motif fingerprint. MAVisto is known as Motif Analysis and VISualization TOol [**99**]

(6) **Kavosh:** Kavosh makes use of "revolving door ordering" algorithm to obtain all the combinations of vertices present in the network. The aim here is to obtain statistically over-represented network sub-graphs [**47**]. The motifs, the method is concerned about, does not need to be connected or form a loop (figure 11).

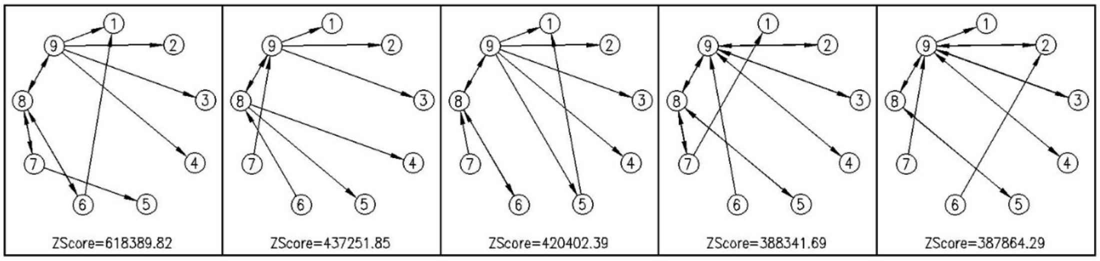


Fig. 11. 9-size motifs of *E.coli*, found by Kavosh [**47**].

(7) **NetMODE:** NetMODE can only perform motif detection for subgraphs of size less than 6. It is a method concerned towards specific improvement over Kavosh. The iteration procedure here does not involve revolving door algorithm and does not utilize Nauty for detecting isomorphs [**65**].

2.2.3.2. Regulatory motif discovery.

At the heart of a control or regulatory system, fundamentally lies a feedback regulation which controls the signal and compares to a desired reference signal and uses the discrepancy to compute corrective control action, and in the realms of engineering models, the modules that performed the actions are sensor, actuator and effector respectively [**58**]. As shown in the figure 12, the regulatory mechanisms are in a constant dynamics and biological control modules evolve and assemble into hierarchical modular systems to help the cell or organism survive in demanding environments.

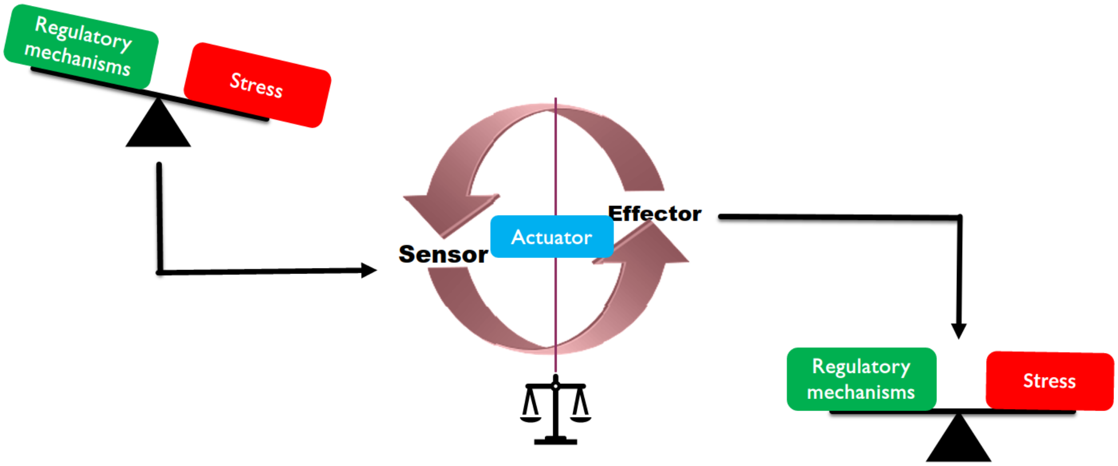


Fig. 12. Feedback is essential for house-keeping of self-regulation.

Chapter 3

Hypothesis and Objectives

3.1. Hypothesis

Since the protein homeostasis entirely relies on feedback regulation which in turn is driven by feedback motifs of closed-loop pattern, we hypothesize that a path-based algorithm that specifically searches for closed-loop pattern connections among the nodes (regulatory genes or proteins) will facilitate in unbiased identification of feedback motifs.

3.2. Objectives

3.2.1. Identification of network motifs with specific topology

As previously discussed, network motifs are essentially identified through enumeration and identification of the statistically significant ones. However, the challenge here is the computational expensiveness, and the aim is identification of functionally important network motifs (figure 13). Statistically proven network motifs are trivial solution for the problem of identification of network motifs. Also, the hub proteins have been claimed to be both physiologically more important and evolutionarily more stable, which suggests that they are less dispensable [**9**]. We aim at building a framework for identifying **key regulatory network motifs** present in a given network. We understand that rare presence could very well be essential for the proper functioning of a system. Moreover, a scarce yet essential entity has a higher possibility to break the system when it gets damaged, and hence constitute a key regulator. Therefore, it is important for us to not limit ourselves to the frequently available motifs.

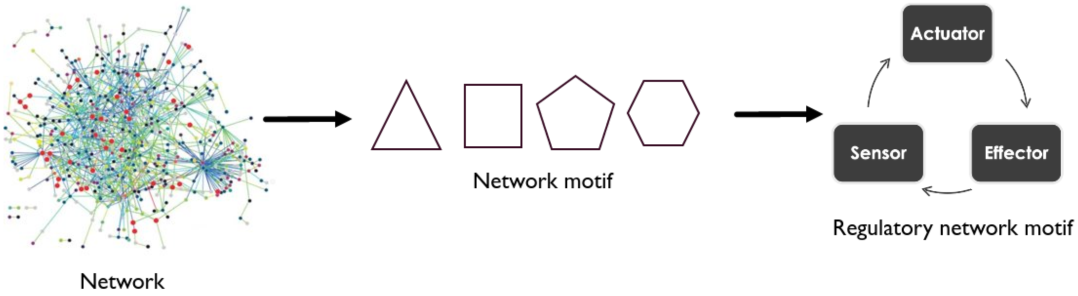


Fig. 13. Pipeline for regulatory motif discovery

It is important to figure out the characteristic properties that are to be used as essential filters for defining a network motif as a regulatory one. The filtering properties that are selected are given below:

(1) Physically connected cyclic motifs: The entities in the network motifs should be physically connected to each other, and the topology should be cyclic in nature to account for and select the motifs that form feedback loops. The size of network motifs are kept at 3 to 6, for now.

(2) Functionally connected: The connections between the entities of the network should be of a functional importance.

(3) Co-regulation: The entities in the network motifs are expected to be perturbed simultaneously under a similar condition.

In this section, we will focus on the first criteria i.e. the physically connected cyclic motifs. As discussed, there exist numerous motif finding algorithms. However, we lack a method to define the specific topology of the motifs, and above all, none of the known algorithms support the filtering steps at the core of this project. This leads us to our first objective, which is, building a methodological framework to identify cyclic-topology network motifs.

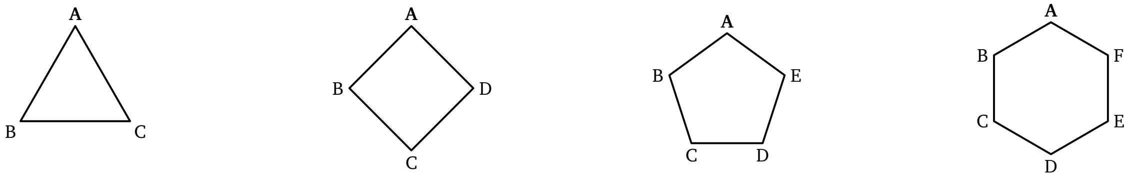


Fig. 14. Cyclic network motifs: (from left) Triplet, Quadruplet, Quintuplet and Sextuplet.

3.2.2. Identification of regulatory motifs

Fulfilling the previous objective will guarantee the topology of the candidate motifs. For the next filters, we aim at systematically integrating functional information to the candidate motifs. This is very relevant because understanding protein homeostasis is a problem which deals with different layers of biological informations and *Saccharomyces cerevisiae* is a well-studied model organism than we chose protein homeostasis in Yeast to apply the framework on and analyse the functioning of the data integration pipeline.

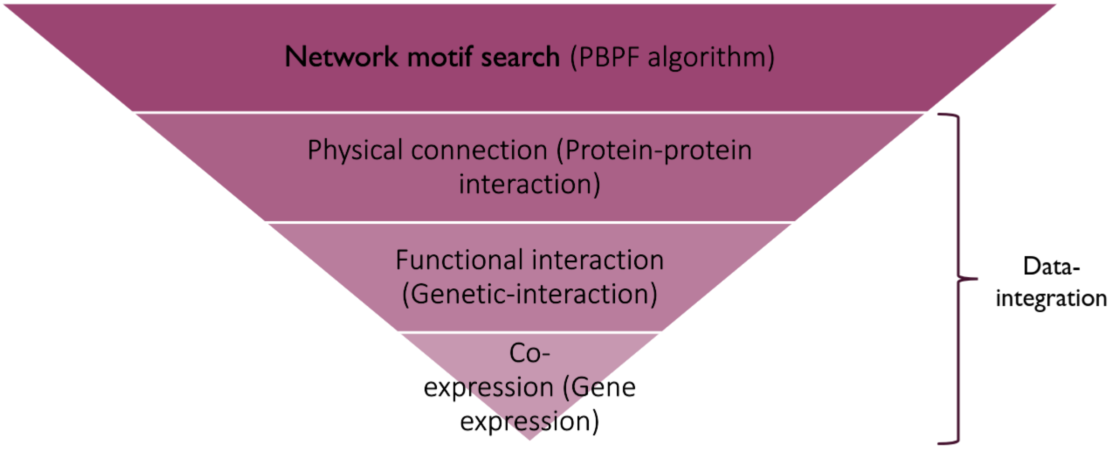

Network motif search (PBPF algorithm)
Physical connection (Protein-protein interaction)
Functional interaction (Genetic-interaction)
Co-expression (Gene expression)
Data-integration

Fig. 15. Data integration pipeline

For information on physical connection, protein-protein interaction data is used (figure 15) whereas for functional information, genetic interaction data is considered. Finally, to verify coregulation, gene-expression data is taken into account. In each step, a scoring function, available from the literature or explicitly derived, is used for the filter.

Chapter 4

Materials and methods: Network Data

4.1. Integration of information to study protein homeostasis in *Saccharomyces cerevisiae*

"The whole is greater than the sum of its parts"
Aristotle

The revolution in the research and applications in biology have made it easier and cheaper to generate ever-greater volumes and types of data which helped in the detailed study of the important parts of the biological systems. The continuous addition of the already existing data in biology has permitted the exploration of new areas in biology, namely, genomics, proteomics [**34**], metabolomics [**60**], lipidomics [**128**], and techniques charting epigenetic regulation [**64**] or chromatin structure [**14**]. However, the unrelated and the heterogeneous nature of the biological data obtained from different platforms through different experimental conditions and procedures makes it extremely challenging to map them for the purpose of integrating them. However, since the molecular complexity of protein homeostasis etiology exists at all different levels, the integrative analysis offers an effective way to increase the strength across multi-level omics data and can be more powerful than single-level analysis [**36, 66**]. An extensive integrated study of structural and functional aspects of transcription factors has been carried out [**3**] using bioinformatics approaches [**70, 121**]. It is evident despite the available data eligible for systems level analysis. It is due to the difficulty and complexity of the problem to put them in one single integrated framework, most of the large scale studies is often done in isolation, without considering the links that connect the underlying molecules. Although application of integration of data is known and has been used in many instances, an integrated framework for a holistic approach study of protein homeostasis remains unresolved. Hence, an important next step in systems biology is the identification of systematic approach for the integration of multi-omics data.

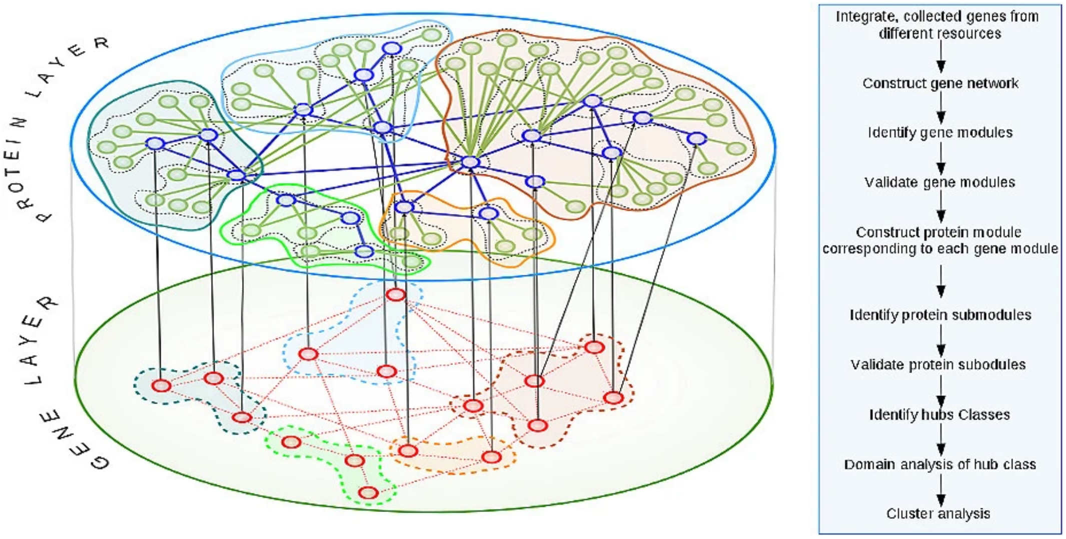

PROTEIN LAYER
GENE LAYER
Integrate, collected genes from different resources
Construct gene network
Identify gene modules
Validate gene modules
Construct protein module corresponding to each gene module
Identify protein submodules
Validate protein subodules
Identify hubs Classes
Domain analysis of hub class
Cluster analysis

Fig. 16. Schematic diagram representing two-layer network structure between genes and proteins [**111**].

Integration of omics data has been used to establish a framework that makes the identification of biologically correlated modules. It has been proven in the context of breast cancer [**111**], which can be extended to any biological function or disease. Figure 16 depicts the overall methodology of integration of the data combined with network analysis. Here, they have constructed two layers of the breast cancer network: the gene and the protein, where each gene module is extended to map out the network of expressed proteins and their interactions in order to identify submodules.

4.1.1. Functional module

The goal is to identify regulatory motifs, which could be achieved by considering a set of well-known regulatory proteins involved in the protein homeostasis network, and identifying all the connected motifs that these proteins make, and adding multiple filters to recognise the protein homeostasis based regulatory motifs. However, it is computationally very expensive to do the same for thousands of proteins. Also, most of the biological networks are scale-free in nature, where the degree follows a power-law distribution [**50**]. It suggests that the network consists of a few nodes which are connected to almost all rest of the nodes present in the network, and it is practically impossible to do the downstream analysis on such a huge number of candidates.

In the search of regulatory motif, we systematically choose functional modules based on their function and their representation in the various datasets used in the pipeline. This step is important, biologically and computationally. Through this decision, we increase the confidence of identifying candidate regulatory motifs, hence, it acts as a preliminary filter. Further, with specific targets, we skipped the unnecessary calculations that would have been otherwise required.

4.1.2. Protein-protein interaction data

PPIs are physical contacts between protein molecules, where the contact is mostly governed by electrostatic forces, hydrogen bonding and the hydrophobic effect. As shown in figure 15, the first step of integration is PPI data for *Saccharomyces cerevisiae* obtained from BioGrid. It consists of approximately 6000 proteins and 708411 inetractions, which maybe direct or indirect PPI. The protein - protein interactions are detected using several known experimental techniques such as, two-hybrid screening and different types of affinity capture. Interactions between proteins are also identified by enquiring about their biochemical activities, constitution of the complex they are a part of and identifying co-localized proteins.

To identify regulatory motifs in protein homeostasis, the first important criterion is that, the proteins involved must be physically connected. Further, the connection topology is to be predefined, and only the candidate motifs with particular patterns of connectivity will be identified. The PPI dataset is in the form of pairwise protein-protein interaction, which when taken together provides the information about connectivity in the interactome of the organism.

4.1.3. Genetic interaction data

Genetic interactions provide information on functional relationships among genes and the type of interaction depends on the type of genetic interaction. A large-scale genetic interaction study on *Saccharomyces cerevisiae* has been done [**33**]. The strength of genetic interaction is defined by deviation of a double-mutant organism's phenotype from the expected neutral phenotype. Genetic interactions occur when mutations in two or more genes combine to generate an unexpected phenotype. An extreme negative or synthetic lethal genetic interaction occurs when two mutations, neither lethal individually, combine to cause cell death (figure 17). Conversely, positive genetic interactions occur when two mutations produce a phenotype that is less severe than expected. Genetic interactions identify functional relationships between genes and can be harnessed for biological discovery and therapeutic target identification.

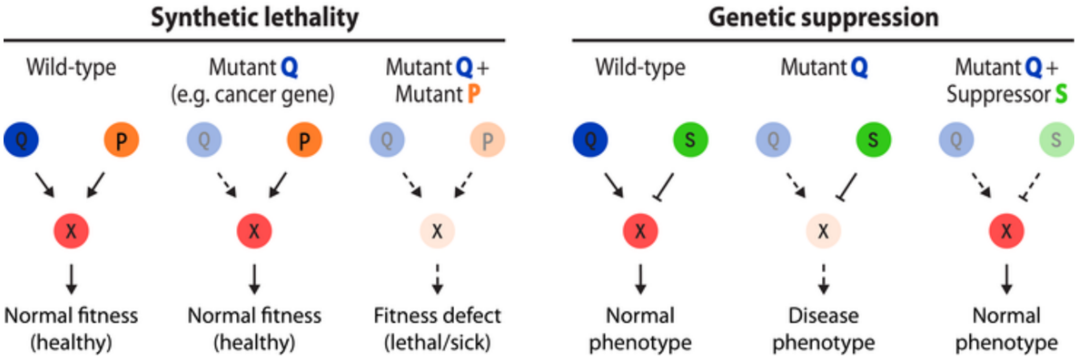


Fig. 17. Examples of genetic interactions [**118**]
abcdefgh

To identify regulatory motifs, it is important to integrate functional information to the previously obtained physically connected motifs. It ensures the regulation dependencies among the contained proteins in the candidate motifs. The genetic interaction information in *Saccharomyces cerevisiae* is taken from the article [**23**]. Further, in the realms of epistasis, the negative genetic interactions are known to be functionally related genes, whereas positive interactions might provide insights into general mechanisms of genetic suppression or resiliency.

4.1.4. Gene expression data

The data is obtained from the paper Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes [**33**]. The data captures the gene-expression in *Saccharomyces cerevisiae* under different environmental conditions. Gene expression is the process by which the genetic information in DNA is transcribed into mRNA. It is a tightly regulated process that governs the response of a cell against changes in environment. It controls the time when the protein is to be synthesized as well as the amount of proteins synthesized.

The data contains values from DNA microarrays. A microarray is a laboratory tool used to detect the expression of thousands of genes at the same time. DNA microarrays are microscope slides that are printed with thousands of tiny spots in defined positions, with each spot containing a known DNA sequence or gene. Often, these slides are referred to as gene chips or DNA chips. The DNA molecules attached to each slide act as probes to detect gene expression, which is also known as the transcriptome or the set of messenger RNA (mRNA) transcripts expressed by a group of genes. The data contained in the file represents the normalized, background-corrected log2 values of the Red/Green ratios measured on the DNA micro-arrays. The yeast gene expression data in its raw version has a dimension of 6,152 rows by 176 columns. The rows consists of the gene names whereas the columns display the description of the experimental condition, hence the data-point in the matrix is the expression values of a gene under a given environmental condition/stress.

The systematic integration of co-expression information will help to infer on the dynamics of the motifs. In the absence of condition-specific perturbation data, the gene-expression data screened under different environmental conditions is used to identify motifs that show co-expression.

Chapter 5

Results

5.1. Performance of Path-based pattern-finding (PBPF) algorithm

Multiple algorithms have been proposed to search and identify network motifs [**75**]. Since searching network motifs is a computationally expensive task, each algorithm focuses on the type of network motif and optimises the search accordingly. As we discussed above, the type of network motifs that are required to obtain the candidate feedback motifs constitutes of a closed-loop pattern (figure 14). Since none of the available algorithms effectively searches for this specific type of network motif, we propose a novel network motif finding algorithm which is based on the path that connects the nodes in a network, and we name it as Path-based pattern-finding (PBPF) algorithm.

5.1.1. PBPF algorithm

We developed an algorithm to search and enumerate connected patterns or graphlets in the yeast PPI network that meets the first requirement for a feedback motif. It is a path-based algorithm that utilizes the depth-first transverse method in which for a given input of required node ("source-node", here: protein) and the size of graphlet, the algorithm saves all the connected neighbors (direct and indirect) to the source node as well as all the paths connecting the source node and neighbors at different levels. The algorithm is divided into four different cases with different pattern sizes i.e. from 3 to 6. The details of the PBPF algorithm, for each graphlet type, is described in figure 18:

(1) **Triplet:** The 3-node graphlet is the simplest pattern discussed here. It takes into account all the immediate (level-1) neighbors. All pairwise unique interactions between nodes are constructed as a set of possible edges, which is then reviewed against the

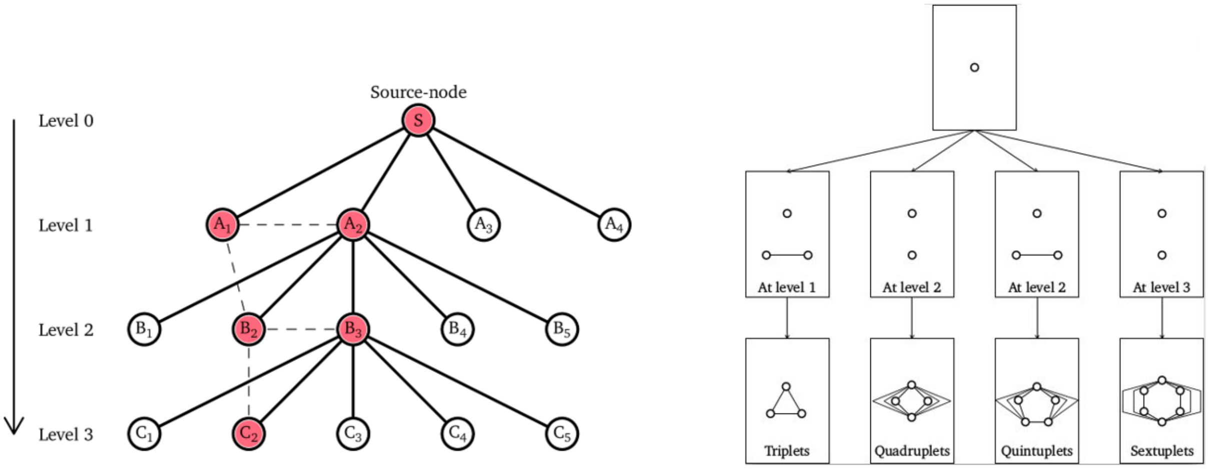


Fig. 18. Schematics for path-based pattern finding algorithm
To the left, it depicts the progression of the motif search algorithm for different sizes of motif. To the right, it represents, the step-by-step identification of network motif (with size=3 to 6) for a single source-node.

set of edges present in the PPI network, to keep the interactions that overlap with the PPI data. The source node along with the nodes forming such edges forms a triplet.

(2) **Quadruplet:** In the 4-node graphlet search, for a given input of node, all the nodes in the paths of length = 2 i.e until level-2 (end node) are curated. Further, for each node in level-2, the possible pairwise unique paths that connect the source and the end through the middle nodes(level-1) are taken. Hence, the source node, along with the middle nodes and the end node form a quadruplet.

(3) **Quintuplet:** The 5-node graphlet search is similar to that of the triplet search. The only difference is that the set of edges is obtained at the level-2 from the source. And the nodes of the edges connecting the source go through a path-length=2 i.e one middle node on each side for each node of the edge. The source node, the two middle nodes and their two neighboring nodes that are connected to each other, together form a quintuplet.

(4) **Sextuplet:** The 6-node graphlet search goes similar to that of quadruplet search, with a difference of the path-length between the source and end node i.e. three, hence including the indirectly connected level-3 nodes as the end nodes. It gives two layers of middle nodes which form a sextuplet along with the source and the end node. Fig 1.1 is a toy network which demonstrates an example for each of the case

mentioned above.

5.1.2. Comparison of PBPF-algorithm to FANMOD

FANMOD is an existing algorithm(discussed in 2.2.3.1) that shows the maximum ability to search for the closed-loop pattern network motif. However, FANMOD does not specifically focus on the required type of motif and hence there is a greater possibility that it does not render the best practically achievable solution to the problem. In the table given below, the properties which separates PBPF from FANMOD are listed. In comparison, the PBPF provides a greater stability by allowing a greater stability and flexibility of usage with respect to the problem.

Property	FANMOD	PBPF
Motif-size	Flexible	Flexible
Source-node	No flexibility	Flexible
Specific motif structure	No or low specificity	Highly specific
Output	Number of motifs	Motifs (nodes and edges), Number of motifs

Table 1. Comparison between FANMOD and PBPF algorithm.

The comparison is made on the essential properties that justifies the need of development of a novel algorithm i.e. the PBPF algorithm.

As shown in table 1, the only control possible in case of FANMOD is the size of the motif and the network in which the motif is to be searched. In contrast, the PBPF algorithm allows us to control the source-node, which reduces the required calculation as well as provides only the motifs that are of our interest. According to figure 8, for a given size of motif, all the possible types of motifs are produced (table 2). We are interested in only one out of the many types i.e. the closed-loop pattern (14). To obtain the specific closed-loop pattern motifs out of all the outputs, we would require another level of filtration which will add extra computing to the already done unnecessary calculations. Considering that, PBPF does only definite calculations to identify exclusively the closed-loop pattern motifs. FANMOD finds and enumerates motifs, checks for isomorphic motifs and classifies them, and compares the frequency of the motifs in the real network to the random network, which is generated using a random graph model. The important drawback here is, it only provides us with the frequency of the motifs. At this step, the obtained candidate motifs captures only the topological connectedness property, which is an important prerequisite but not enough to qualify as a feedback motif. For that, we need to further integrate biological properties to the candidate motifs. To do so, control over the nodes and edges that form the motifs is very important. On the other hand, PBPF gives motifs with constituent nodes and edges in closed-loops, satisfying all the required conditions of our purpose and justifies its application in identifying feedback motifs.

5.1.2.1. Application of PBPF and FANMOD algorithms on a test network.

The biological networks are usually huge in size with thousands of nodes and millions of interacting edges. Application of the graph-based algorithms on such networks could be very tricky. To avoid any nuisances, there exist a well established step that includes systematic validation of the algorithm, which is generally done using a small dummy graph because it is easier to visualise and faster to search and enumerate the motifs present in it.

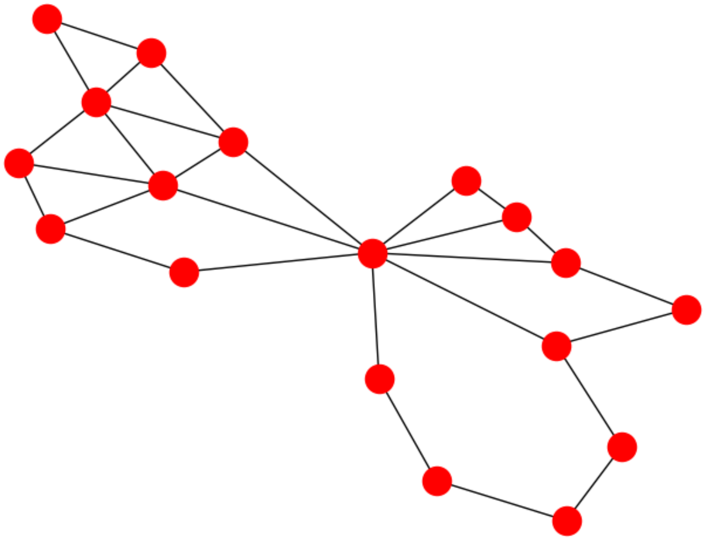


Fig. 19. Test network.
A dummy network to test the functioning of the algorithm on. It consists of the connections that are speculated to be potential cause of mis-counting of motifs.

Motif-size	FANMOD	PBPF
3	64	8
4	167	8
5	425	7
6	970	6

Table 2. FANMOD versus PBPF

Number of motifs obtained in the dummy network using FANMOD and PBPF algorithm.

As shown in figure 8, the motifs identified by FANMOD do not satisfy the topology of the search motifs. FANMOD is very lenient with respect to the type of topology, and it recognises chained structure, open-loop, along with closed-loop cycle for a given size of motif. It also accepts different permutations and combinations of connections as unique motifs. In contrast, PBPF is very stringent in its identification search for motif. The goal of the algorithm, which is, to identify potential feedback motifs, makes the filter conditions very stringent. Therefore, PBPF ensures only closed-loop cycles from size three to six to be the only possible motif topologies of motif to be selected. This is the reason behind the outcome presented in (table 2), which clearly shows a larger number of motifs while enumerated through FANMOD than PBPF. An increase in number of motifs with increase in size of the motif is also observed in case of FANMOD, which is definitely not the case in case of PBPF algorithm.

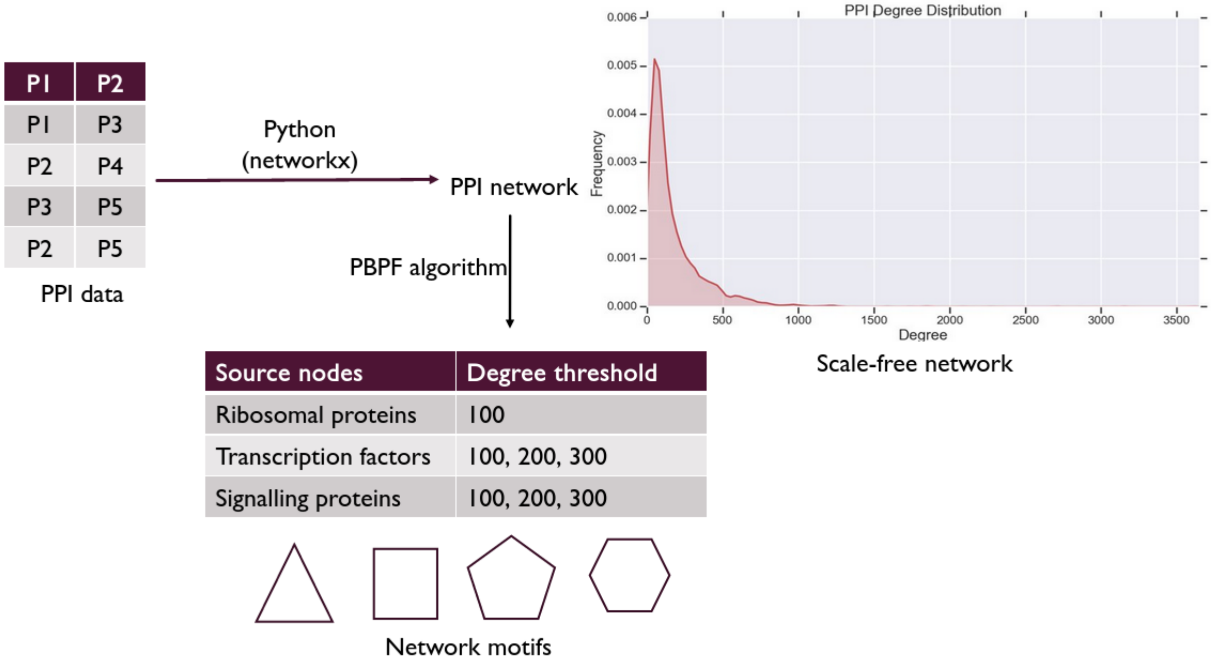
5.2. Application of PBPF-algorithm on protein homeostasis network of *Saccharomyces cerevisiae*

5.2.1. Identification of candidate motifs that are likely to be physically interacting

The proteins in the PPI data of *S.cerevisiae* form the nodes and the protein-protein pairwise interaction make the edges in the PPI network. The network is built using networkx in python. It is a huge network of 6000 nodes (proteins) and 708411 edges.

The degree distribution of the PPI network is calculated, and it was observed that there are few proteins with degree equivalent to total number of proteins in the network (figure 20). The **n** in the boxplot denotes the number of proteins that are forming motifs for each size. This suggests that these proteins form **hubs** in the network, the rest of the proteins in the network have a lower degree (figure 21). This confirms the PPI network to be scale-free in nature (figure 20), with the degree distribution following power-law function. As we have discussed earlier, it is computationally impractical and biologically less relevant for the analysis. Therefore, the hubs are required to be deleted from the network before starting any kind of analysis. To do so, a range of degree threshold is chosen, and all the nodes with degree beyond that threshold are deleted from the network. Moreover, this step provides the opportunity to keep a check on the number of motifs enumerated for each degree threshold. The aim is to keep more information but without losing the practical situation for the motif analysis i.e. the module that supplies less number of motifs for a higher degree threshold is the most favourable one.

As biologically significant and variant, ribosomal proteins, transcription factors and signalling proteins are taken as the functional modules, and are given as the input for the list of source nodes while applying on PBPF algorithm. It was observed that, for the list of source nodes as ribosomal proteins, the most suitable degree threshold is 100, whereas



P1	P2
P1	P3
P2	P4
P3	P5
P2	P5

Source nodes	Degree threshold
Ribosomal proteins	100
Transcription factors	100, 200, 300
Signalling proteins	100, 200, 300

Fig. 20. Pipeline for identification of network motifs formed by proteins in the PPI network. To the top-left, it is a representation of the format of the PPI data, which is represented as a network using networkx, a Python module. It treats the pair-wise PPIs as the connections or edges in the network. To the top-right, it is the degree-distribution plot of the PPI network. To the below, in the table, the degree-thresholds for different set of proteins are mentioned for which PBPF was applied on the network.

for TFs, although all the degree thresholds yield results for some of the proteins, but for some it gets terminated. The calculation pipeline runs smoothly for all the thresholds for signalling proteins. Also we observed that, for the proteins that yield results, the number of motifs for ribosomal proteins and TFs was extremely high, and hence are not the most convenient ones for the downstream analysis. Hence, we chose to proceed the study with signalling proteins.

The above observation can be explained as follows:

(1) Ribosomal proteins are highly expressed and central for almost all the cellular functions [**9**]. Therefore it is highly likely that they form hubs in the PPI network.

(2) TFs as well, being at the heart of every cellular function, makes it a popular regulator. This explains the huge number of motifs formed by TFs in PPI network.

(3) Signalling proteins enables interaction among other proteins, making them a core regulator of a cellular system. However, the functioning of signalling proteins depend on their free availability as well as their specificity of interaction through receptors. Probably, this is the reason behind obtaining feasible number of interactions with signalling proteins in a huge ball of densely connected network.

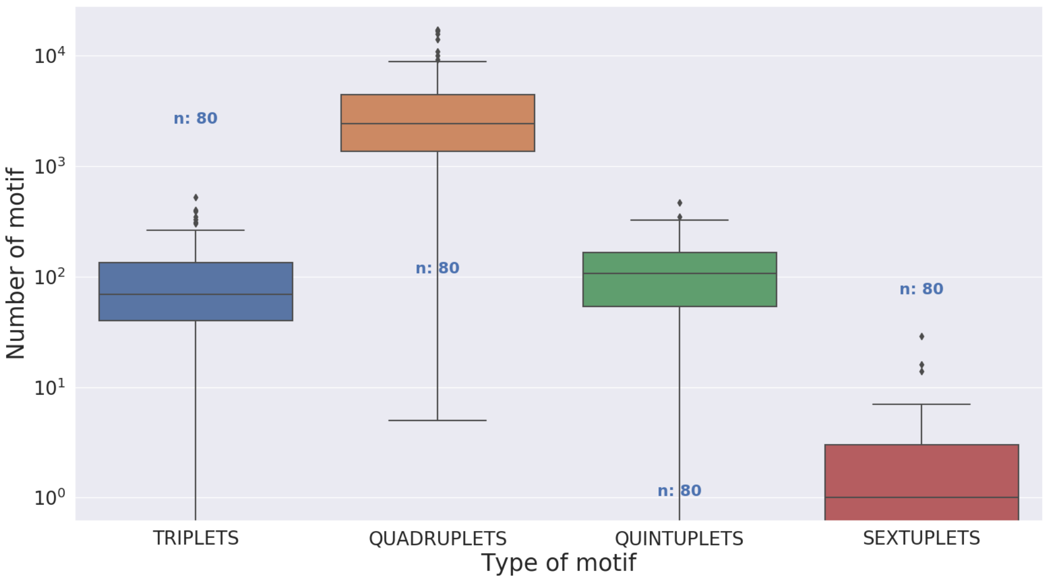


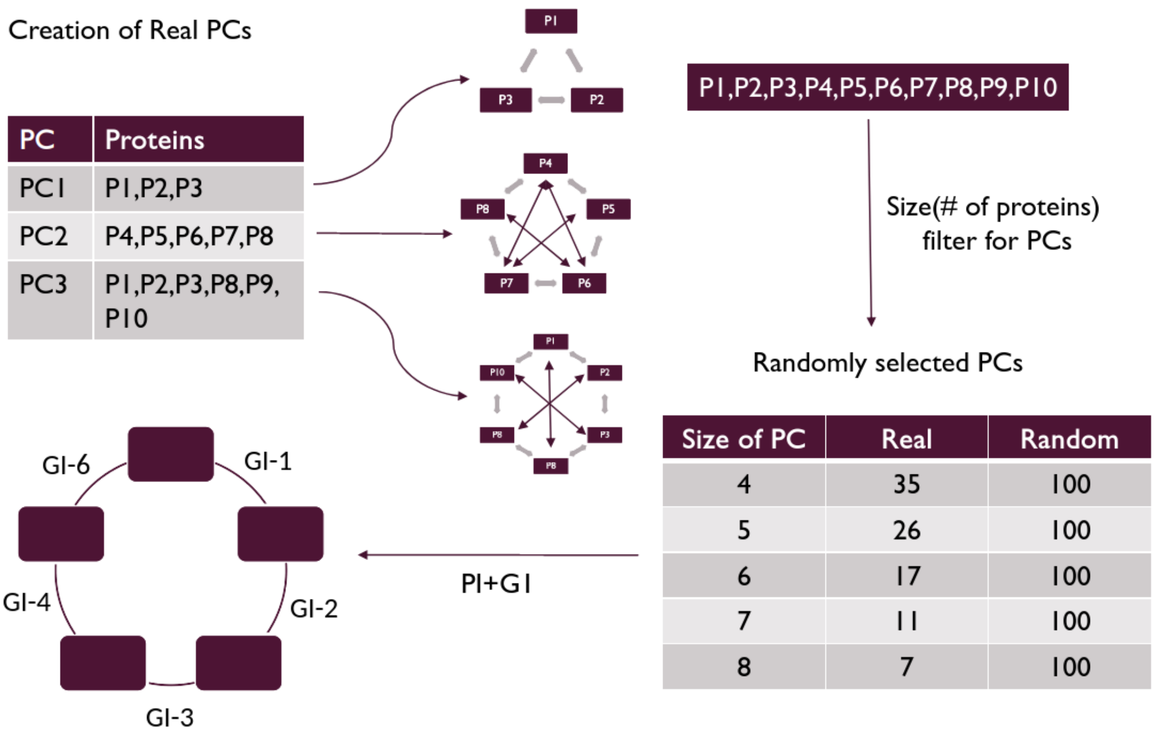
Fig. 21. Enumeration of network motifs formed by signalling proteins in the PPI network. Each data point is the representation of the number of a specific size of motif formed by a single signalling protein in the PPI network. The box-plot represents the distribution of numbers of a specific size motif for all the signalling proteins. Here, n=number of signalling proteins that form motifs.

In figure 21, the boxplot shows the distribution of the number of motifs formed by the signalling proteins, with a degree threshold of 300. And we see the number of motifs growing from triplets to quadruplets, which drops for quintuplets, and it is the least for sextuplets. The observation can be explained as the increase in possible combinations with an increase in size of motif, and also the conditions getting too stringent when the size of motif grows further.

5.2.2. Genetic interaction in Protein complexes (PCs)

PCs are biological entities, constructed through protein-protein or protein-ligand binding, hence consisting of group of proteins and ligands to perform specific cellular processes. PCs, being functional in nature, with proteins known to be physically connected within it, are considered as the control set for protein-protein interaction. The goal here is to check if PCs are able to show functional relevance when integrated with genetic interaction.

Out of the list of PCs, small connected sub-networks are created from the PPI network. As one of the approaches, the aggregate genetic-interaction score of the pairwise PPIs in the PCs is taken as a criterion to determine functionality of the PC. Based on the size of real



PC	Proteins
PC1	P1,P2,P3
PC2	P4,P5,P6,P7,P8
PC3	P1,P2,P3,P8,P9, P10

Size of PC	Real	Random
4	35	100
5	26	100
6	17	100
7	11	100
8	7	100

Fig. 22. Schematics explaining the creation of random PCs and integration of GIs to the PC based sub-networks.

To the top, the creation of PCs (real and random) is displayed. GI score is integrated to the constructed PCs. To the right-bottom, in the table, number of real and random PCs of different size of PC are mentioned.

protein-complexes, random protein complexes are created using randomly chosen proteins to create a background distribution (figure 22). The PCs are normalized and segregated according to their size to avoid over or under-representation of the aggregate scores.

A more negative score is observed for the real protein-complexes (figure 23), suggesting that a functional interaction tends to have a more negative genetic interaction score. A double mutant with a more extreme phenotype than expected bears a negative score which defines a synergistic interaction between the corresponding mutations or synthetic lethality, in the extreme case. This explains the negative score for the real PCs which are known to be functional in nature.

This validates the strategy of integrating PPI to genetic-interaction data, opening the door to extend the study and apply to the curated candidate motifs.

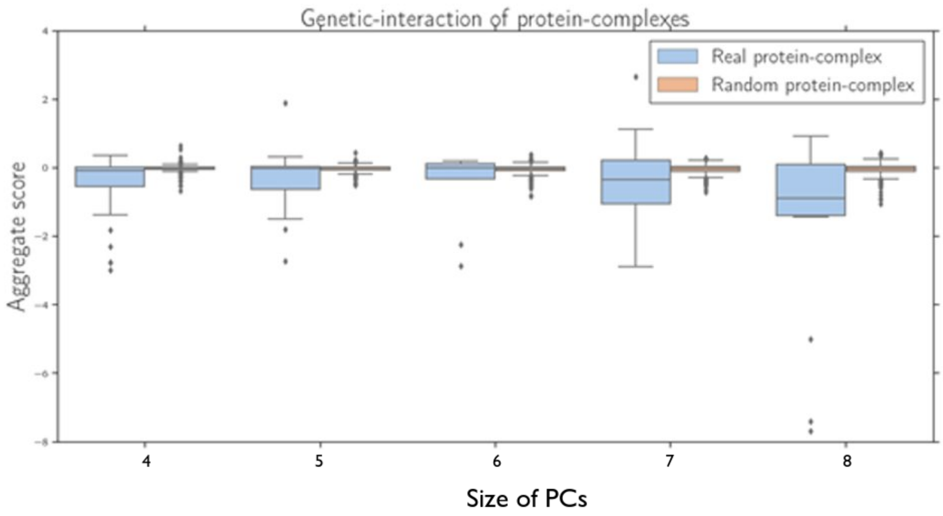


Fig. 23. Genetic interaction of PCs
An aggregate genetic-interaction score is plotted for the real versus random PCs of same size.

5.2.3. Identification of functionally interacting candidate motifs

The PPI-based physically connected candidate motifs is integrated with the genetic interaction score. From the literature [**23**] it is known that a genetic score above 0.08 or below -0.08 can be confidently defined as functional interaction.

At this step, we obtain candidate motifs which are likely physically connected and functionally interacting. After the filter of genetic interaction, we observed decrease in number of motifs (figure 24). Also, the number of signalling proteins forming any motif decreases dramatically for higher size motifs. We understand that PPI and GI data are orthogonal in nature, and this could explain the reduction in number of motifs. Below, the examples of cyclic motifs are drawn, with information of PPI and gene interaction. The GI information is integrated from the previously created GI-score matrix. The same information is available in the databases like *Saccharomyces* Genome Database (SGD); where the score might differ slightly because of the difference in approach to accommodate the multiple data points for each pairwise interaction. Thus, both of the factors are available in literature and in well-studied datasets, at pairwise interaction level. Here, the usage of the PBPF algorithm and data-integration techniques allows us to visualise and study at the level of cyclic motifs. A few cases of network motifs are identified, with one or many unavailable data point, suggesting that this approach has a potential to identify novel cyclic motifs from a given network.

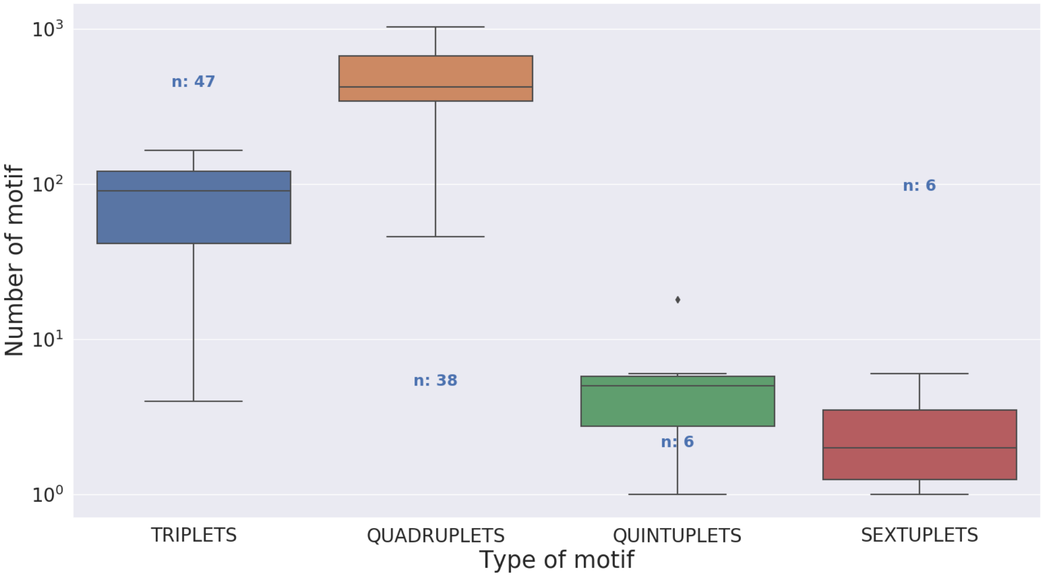


Fig. 24. Enumeration of network motifs likely functional motifs formed by signalling proteins in the PPI and Genetic interaction network.
Each data point is the representation of the number of a specific size of motif formed by a single signalling protein in the PPI network, integrated with GI score. The box-plot represents the distribution of numbers of a specific size motif for all the signalling proteins. Here, n=number of signalling proteins that form motifs.

5.2.4. Identification of candidate motifs that are likely active under the same condition

Heat shock response is a well studied topic in the realms of protein homeostasis. Hence, for testing co-expression, we chose temperature specific conditions, i.e. a steady-state growth condition at 25°C and a heat-shock temperature at 37°C. The former is represented as the normal condition and the later is considered as the stressed condition. Further, we keep a threshold of gene-expression = -/+ 2, to regard an expression as regulation; only the data points beyond the blue horizontal lines in figure 25 are considered as being relevant.

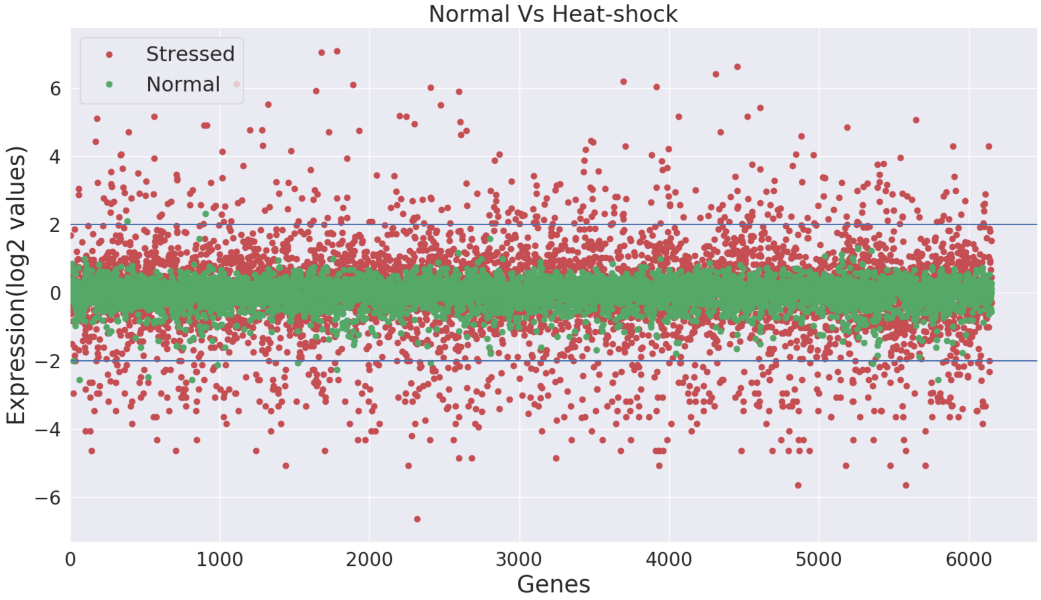


Fig. 25. Expression value of yeast genes compared over normal and stressed conditions. It is the distribution of GE log2 value for normal (25°C) and stressed (37°C) condition. The blue lines in the plot display the positive and negative GE threshold value.

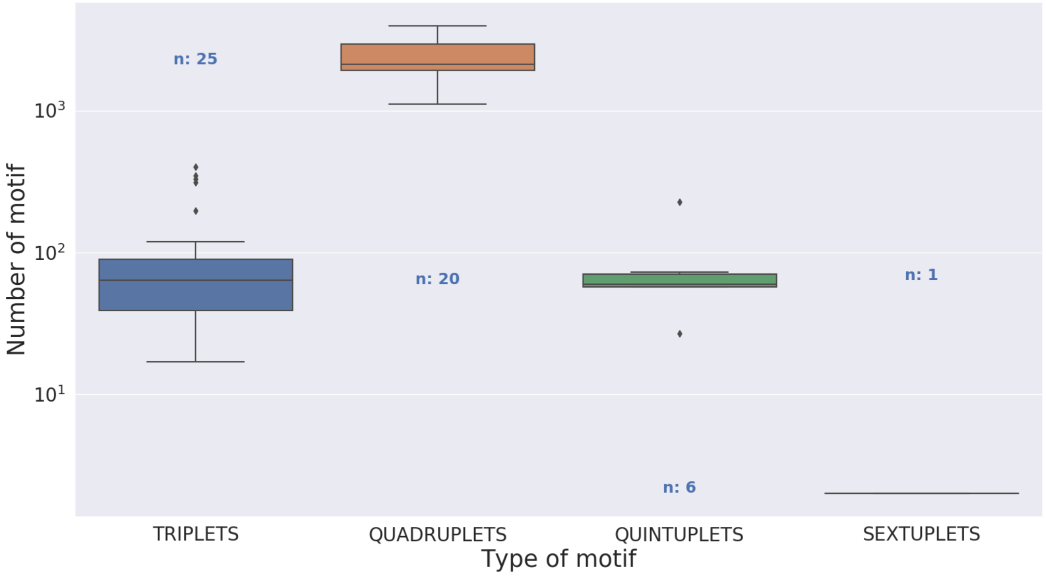


Fig. 26. Enumeration of network motifs likely co-expressed motifs formed by signalling proteins in the PPI and Genetic expression network in normal condition.

Each data point is the representation of the number of a specific size of motif formed by a single signalling protein in the PPI network, integrated with GE score at normal temperature. The box-plot represents the distribution of numbers of a specific size motif for all the signalling proteins. Here, n=number of signalling proteins that form motifs.

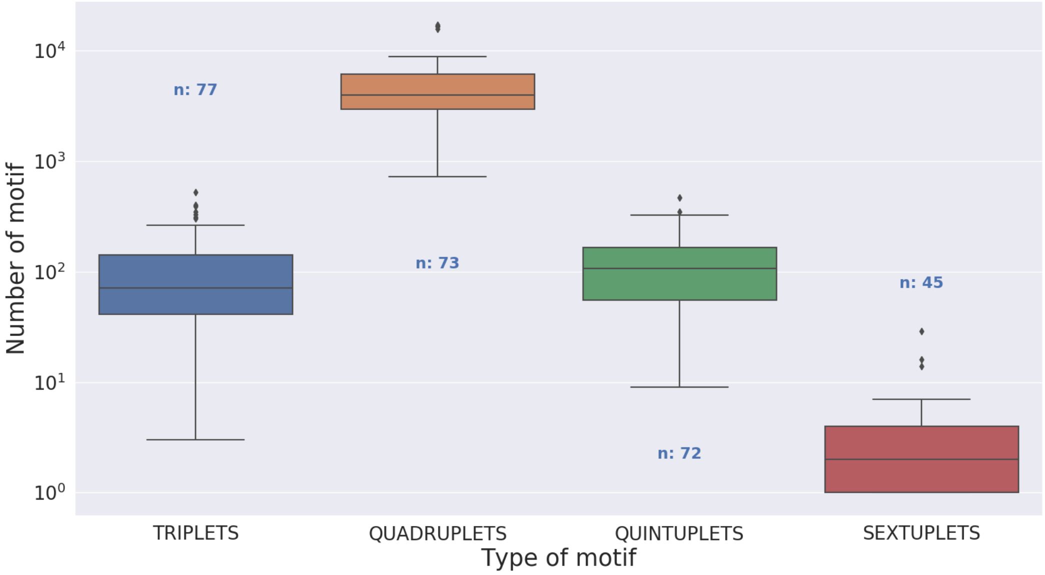
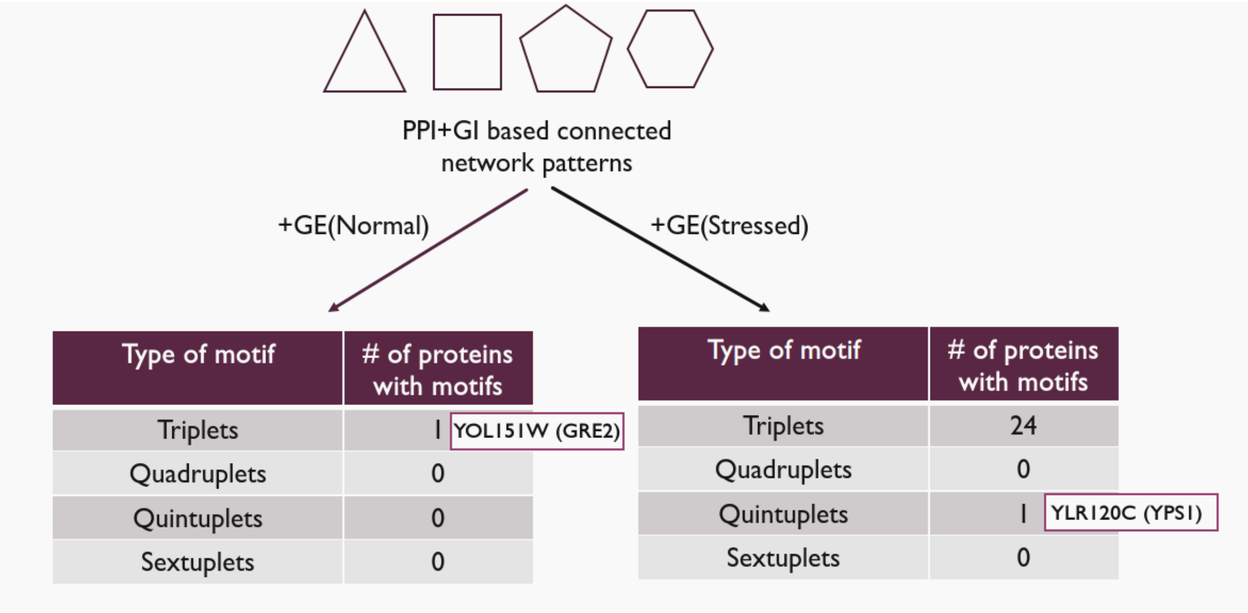


Fig. 27. Enumeration of network motifs likely co-expressed motifs formed by signalling proteins in the PPI and Genetic expression network in stressed condition.
Each data point is the representation of the number of a specific size of motif formed by a single signalling protein in the PPI network, integrated with GE score at stressed temperature. The box-plot represents the distribution of numbers of a specific size motif for all the signalling proteins. Here, n=number of signalling proteins that form motifs.

As shown in figure 26 and figure 27, the number of motifs is comparatively higher in stressed condition than in normal condition, suggesting that the stressed condition can be identified with an increase in the number of motifs and in number of proteins that are deferentially expressed to form motifs. Also, an observation worth noting is, the data points (number of proteins forming motifs) in the case of normal condition is far lower than in the stressed condition. The basal level expression in the normal condition could explain this observation. Since, the higher number is one of the core issues of motif enumeration problem, the decrease in the number with systematic integration of information could be a successful and meaningful way of dealing with the problem. In the examples of cyclic motifs, we added information of expression profile for each temperature condition to create two subsets of data (25°C and 37°C). We can see the change in expression for stressed condition compared to the normal suggesting that stressed conditions may provide additional information on regulatory motifs.

5.3. Identification of likely co-expressed motifs with physical connectivities and functional relations

Above, we saw pairwise integration of information, and by doing that we have systematically increased the confidence of candidate motifs which are potentially regulatory in nature. The ultimate step is where all the information is systematically integrated.



Type of motif	# of proteins with motifs
Triplets	1
Quadruplets	0
Quintuplets	0
Sextuplets	0

Type of motif	# of proteins with motifs
Triplets	24
Quadruplets	0
Quintuplets	1
Sextuplets	0

Fig. 28. Pipeline for enumeration of likely regulatory network motifs.
PPI+GI motifs are integrated with GE scores, normal and stressed, separately.

Here, the analysis pipeline makes sure that the candidate motifs that is obtained has gone through the three filters i.e. PPI, genetic interaction with +/- 0.08 threshold, and gene-expression with two-fold change. The network motifs at this step are putative regulatory motifs. The number of motifs is higher in case of stressed compared to the normal condition (figure 28). Above that, there is an extremely low number of proteins that crosses the three levels of filter, and among those, all are part of triplets, except for the unique occurrence of a quintuplet.

5.3.1. Identification of regulatory motifs

At this instance, we have successfully integrated the functional informations with the identified candidate motifs. In this section, we introduce to some of the examples of network motifs obtained through the framework of this project, some of which are found to be known, described in the literature, and some are novel putative regulatory motifs discovered through the stringent functionality filters of PBPF algorithm.

By carrying out further investigation of the candidate motifs, we discovered the presence of known regulatory motifs among the PBPF predicted candidate motifs (figure 29). Literature study of these motifs suggests their functional importance in protein homeostasis [**19, 105**]. Figure 29 is the graphical representation of the regulatory network of some of the signalling proteins (specifically, that of the BMH2), taken from SGD. BMH2 is a protein involved in the regulation of many processes including exocytosis, vesicle transport, Ras/MAPK signaling, and rapamycin-sensitive signaling [**10**]. It regulates many transcription factors, including the ones present in predicted motifs which play role in directly or indirectly initiating a cellular response against environmental stresses, specifically, response against heat [**13**]. The comparison of the known regulatory motifs with the regulatory network shows an overlap in the structure; the triplet motif BMH2-PKH1-PKH2 is one of such regulatory motif. This confirms that the discovered motif is regulatory in nature. It is a validation for the algorithm and the approach for successfully identifying the regulatory motifs.

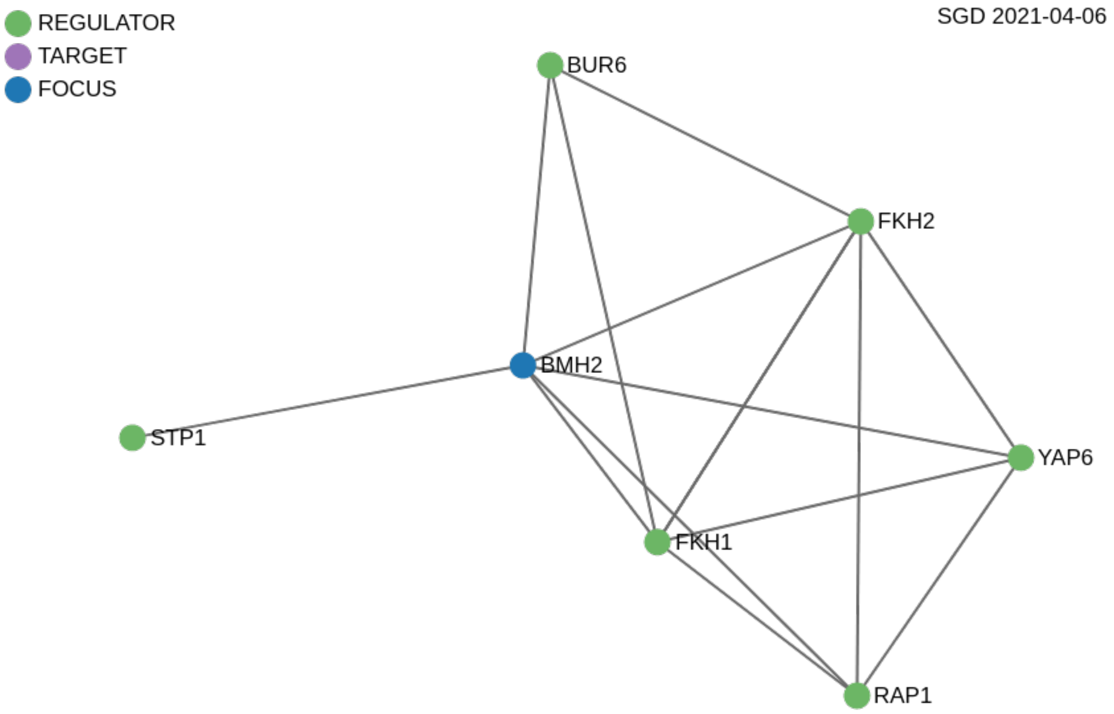


Fig. 29. Regulatory networks formed by BMH2 (YDR099W).

We tried to integrate GI data to the regulatory motifs (figure 29). Surprisingly, it was found that the GI data for the interactions present in the regulatory motifs (figure 29) does not exist. Firstly, it is a good news that the pipeline is able to capture the functionally important interactions. We learnt, every integration of information can yield important output, which may or may not be in alignment with the next integration.

The figure 30 demonstrates one example of each type of motif that are identified using PBPF-algorithm and the systematic integration of functional information. The PPI data carries the information of pairwise interaction. But the topological structure is obtained by treating the pairwise interaction data into graph or network, where the proteins form the nodes and the connection between them form edges. Therefore, it is now possible to carry out studies at the level of motifs.

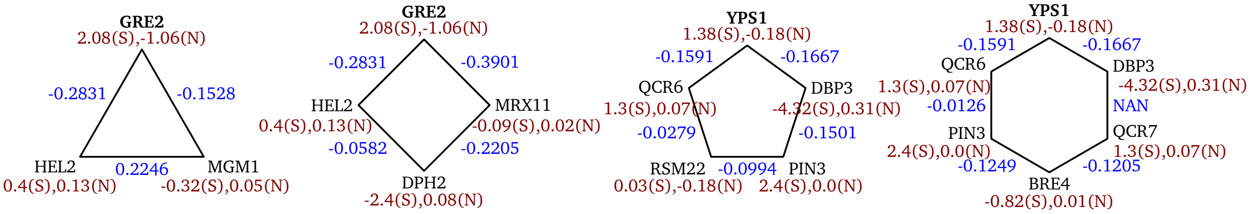


Fig. 30. Examples of likely regulatory network motifs.
Source nodes are signalling proteins. The figure represents network motifs formed by signalling proteins. GE-score is in red, GI-score is in blue.

In the figure 30, the nodes represent the names of proteins. Below them, in red, are mentioned their gene-expression values; S for the stressed condition and N for normal. At the edges, in blue, GI-score for the given interaction is given. The triplet and the quintuplet is derived from the final integration i.e. PPI+GI+GE. The triplet is taken from the specific integration of gene expression at steady-state growth condition whereas the quintuplet is obtained through that of the heat-shock condition. In both cases, surprisingly, the source-node selected are the only proteins that form motifs in the network. Therefore, we investigated the nature of these proteins as well as the proteins they form motif with. GRE2 is a stress response gene that plays an important role in restoring homeostasis in case of environmental perturbations, like, osmotic, ionic, oxidative, heat shock and heavy metals [**32**]. GRE2 is found to be involved in methylglyoxal reductase (NADPH-dependent) activity as well as 3-methylbutanol:NAD(P) oxidoreductase activity [**114**]. It is regulated by HOG MAPK signal transduction pathway [**95**]. Its relative gene-expression value for stressed condition over normal suggests its involvement in heat shock response. The immediate neighbour of GRE2 in the motifs are: HEL2, which is a ubiquitin ligase, which helps to control the protein quality by inducing degradation when required [**103**]; MGM1 is a mitochondrial GTPase, and plays role in maintaining the mitochondrial genome [**4**]; DPH1 is a protein involved in biosynthesis [**28**] and MRX11 is a mitochondrial ribosome: both regulate proteins involved in heat-shock response [**6**]. On the other hand, YPS1 is a MAPK activating protease. It regulates transcription factors which are active under heat shock condition [**107**]. QCR6 and QCR7 is a component of the mitochondrial inner membrane electron transport chain, and regulates proteins involved in cellular response against heat and during mitotic cell cycle [**2**]. RSM22 is also a component in mitochondrial subunit [**116**]. DBP3 is RNA-dependent ATPase [**122**]. PIN3 regulates cellular response against thermal stress [**79**]. BRE4 is a zinc finger protein with a role in endocytosis. Together, these suggest that the proteins involved in the motifs are functionally very different from each other and they often, belong to different pathway. Thus, suggesting that we conclude here is, there is an inter-play between functional modules in order to maintain homeostasis.

The function of individual gene or protein is known. In some cases, the interaction between them is also characterised. But as we discussed, at the heart of the functioning of biological systems lies communication and feedback systems which regulate each other in order to maintain or restore back a stable environment for the cell. In this project, we have moved a step closer in this direction by proposing to study biological systems at the level of network motifs.

Chapter 6

Discussion and Conclusion

PBPF: the proposed motif finding algorithm is written in Python 2.0 programming language, consisting of four parts, each for the triplets, quadruplets, quintuplets and sextuplets. PBPF can detect network motifs with a cyclic topology (figure 14). For a given network and an input node, the algorithm gives all the motifs in the form of list of sets of nodes that form the motifs; also, the nodes in the output are ordered in such a way that the adjacent nodes in the set of nodes of a motif represent the edges of the motifs. The whole project is set up on the aspect of protein homeostasis. It is achieved by the application of the algorithm on yeast PPI-network, and signalling proteins as the input nodes. The candidate motifs were integrated with functional information (genetic interaction and gene expression) to strengthen the confidence of the search pipeline. Since, there are thousands of proteins with millions of interactions, it is not feasible to do extensive analysis on the millions of motifs formed by them. The framework of this work, systematically adds regulatory information to the putative motifs, that reduces the number of motifs alongside. The motifs are identified and enumerated along the way, at each step of integration (that are: PPI, PPI+GI, PPI+GE, PPI+GI+GE). A few potential regulatory motifs are represented in the figures 30. It was also shown that, the framework of the PBPF algorithm, is able to retrieve known regulatory motifs.

We are aware, that we are not yet there to conclude a candidate motif to be a feedback motif. But I would like to propose a potential model at this stage, to help analyse the regulatory motifs as feedback motifs. As shown in figure 31, for a given three-node motif, depending on the expression value of the nodes (say, genes), could guide us to estimate the type of interaction each edge has. But since the directionality is not known, we will apply the logic bi-directionally. This might look simple, but, the fact that the motif that we are analysing are discovered after systematic filtration, and are a potential regulatory motif, adds confidence to the analysis. The model shown here, could be a simple feedback loop:

with edges 1, 2 and 6. It could also be a coherent feedforward loop: with edges 1, 3 and 2 when flow is from Y to Z.

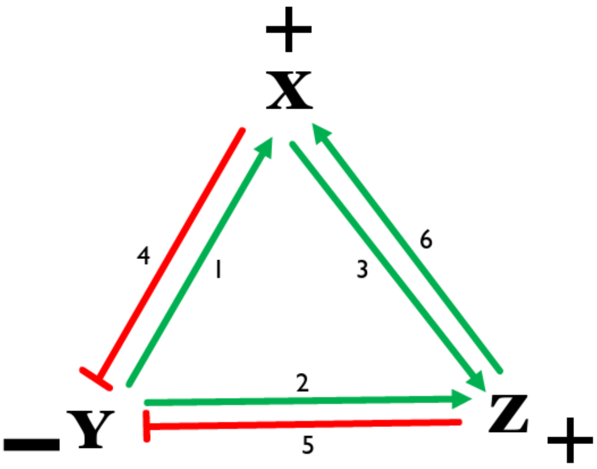


Fig. 31. Model for estimation of type of feedback
+ and - stand for positive and negative GI-score respectively. The connections between the nodes are represented bi-directionally to accommodate all the possible type of connections. Green represents activation regulation whereas red represents inhibitory regulation.

The PBPF holds potentials for improvement. We observed a scaling difficulty in the usage of the algorithm, because of which we could not include the whole network; we ignored the highly connected hubs from the network. The hubs create a huge amount of data creating memory problem, which can be fixed by adding a separate function that bins the nodes in a network according to their degree of connections, and deciding at a later stage which bins to be included for which calculation. Further, the intermediary steps, where the paths from the source node to the nth-level node is calculated, can be saved, which can be retrieved according to the requirement. This would help to avoid repetition of calculations. There is another possible limitation of the algorithm arising due to the very nature of the interaction-data. The interactions screened in the PPI data are present only in unique combinations. Therefore, the algorithm considers redundant combinations of interactions, to avoid losing information. This might lead to over-counting of number of motifs in some cases. In future work, this issue can be addressed by including a function that identifies isomorphic graphs and saves only the unique ones [**55**]. Since most of this work is dependent on the available data, and as we know, generally, these data are context dependent and in some cases the technology that is used might face challenges in capturing some of the reactions, is a drawback for our analysis.

The cyclic motifs could also represent feed-forward motifs and the PBPF algorithm has the potential to extract them from the network, as we summarize in the figure 31. But the above explained model for feedback motif cannot be conclusive about the feedback or the feed-forward behavior, since these systems are sensitive to the direction of interactions in the motif. Integration of functional information in combination with loss-of-function and gain-of-function perturbation experiments that gives simultaneous measurement of change in level of expression of all the nodes in a motif will provide information on direction of interaction.

The current study is carried out at a fundamental level and it immediately opens the possibility for multiple studies. The most on-the-spot recommendation would be to further strengthen the analysis to be able to identify known and novel feedback motifs from biological networks. Although we applied the PBPF-algorithm on yeast protein homeostasis, and limited the study to signalling proteins, it is possible to do the whole system-level analysis. But it can be computationally expensive and the method needs to improve to make it more robust. This model can be easily extrapolated in future work. Further, the topology of motifs could be extrapolated to accommodate complexes. Also, in signalling, directionality is quite well mapped, which can be integrated in the future work. Further, a more elaborate study on the impact of temperature change can be done by obtaining data for motifs formed for small and continuous shift in temperature. The identified motifs can be linked to the change in temperature to identify the early onset of motifs; late ones; the soft one; the rigorous ones; the ones which stay throughout and the ones that comes early and goes away in later stages. The dynamics of motifs and their numbers will be interesting to understand. Due to the unavailability of time resolved proteomics data during the time of integration, could not be added to the calculations. It could be another dataset to integrate in the future studies to strengthen the framework.

The fundamental structure of the framework allows its application in completely different biological systems, for example: single-cell data. The information or the analysis of the information on single-cell data can be quite robust and hence self-independent for studies like, identification of cell-state specific regulatory motifs. From the single-cell gene expression data, a weighted gene-regulatory network for each cell-state can be obtained. We can also, identify potential cell-state regulatory motifs by applying PBPF-algorithm to it. We have functional information such as, gene-expression and co-expression value between the genes. A scoring function or a threshold is to be identified, using which the candidate motifs can be filtered. Further, network topologies such as, change in centrality, neighbours or modularity, sepcific to cell-state can be used to identify functional modules for cell-state specific networks (figure 32).

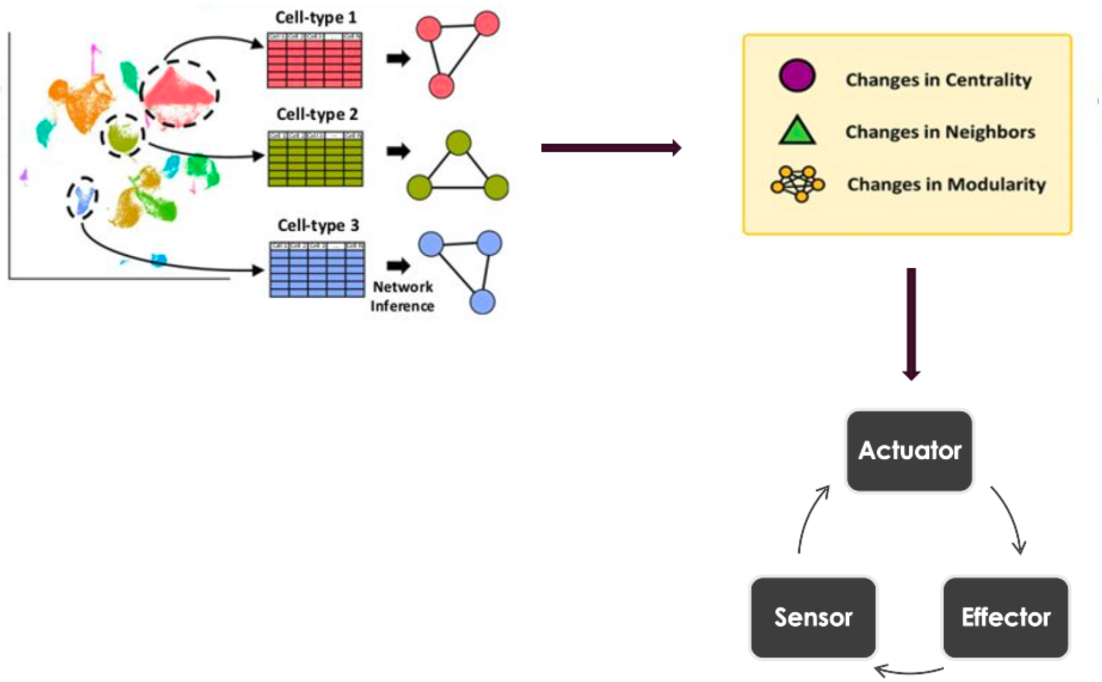


Fig. 32. Pipeline for identification of feedback motifs for cell-state specific stem-cells [**18**].

References

[1] Andrea Brigitta Alber and David Michael Suter. Dynamics of protein synthesis and degradation through the cell cycle. *Cell Cycle*, 18(8):784–794, 2019.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, K Roberts, and P Walter. Electron-transport chains and their proton pumps. *Molecular biology of the cell*, 4, 2002.

[3] Dominic J Allocco, Isaac S Kohane, and Atul J Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics*, 5(1):1–10, 2004.

[4] Boominathan Amutha, Donna M Gordon, Yajuan Gu, and Debkumar Pain. A novel role of mgm1p, a dynamin-related gtpase, in atp synthase assembly and cristae formation/maintenance. *Biochemical Journal*, 381(1):19–23, 2004.

[5] Suzanne Angeli, Tracy Barhydt, Ross Jacobs, David W Killilea, Gordon J Lithgow, and Julie K Andersen. Manganese disturbs metal and protein homeostasis in caenorhabditis elegans. *Metallomics*, 6(10):1816–1823, 2014.

[6] Syed H Askree, Tal Yehuda, Sarit Smolikov, Raya Gurevich, Joshua Hawk, Carrie Coker, Anat Krauskopf, Martin Kupiec, and Michael J McEachern. A genome-wide screen for saccharomyces cerevisiae deletion mutants that affect telomere length. *Proceedings of the National Academy of Sciences*, 101(23):8658–8663, 2004.

[7] Evren U Azeloglu, Simon V Hardy, Narat John Eungdamrong, Yibang Chen, Gomathi Jayaraman, Peter Y Chuang, Wei Fang, Huabao Xiong, Susana R Neves, Mohit R Jain, et al. Interconnected network motifs control podocyte morphology and kidney function. *Science signaling*, 7(311):ra12–ra12, 2014.

[8] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[9] Nizar N Batada, Laurence D Hurst, and Mike Tyers. Evolutionary and physiological importance of hub proteins. *PLoS Comput Biol*, 2(7):e88, 2006.

[10] Paula G Bertram, Chenbo Zeng, John Thorson, Andrey S Shaw, and XF Steven Zheng. The 14-3-3 proteins positively regulate rapamycin-sensitive signaling. *Current biology*, 8(23):1259–S1, 1998.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[12] Onn Brandman, James E Ferrell, Rong Li, and Tobias Meyer. Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science*, 310(5747):496–498, 2005.

[13] Astrid Bruckmann, Paul J Hensbergen, Crina IA Balog, André M Deelder, H Yde Steensma, and G Paul H van Heusden. Post-transcriptional control of the s accharomyces cerevisiae proteome by 14-3-3 proteins. *Journal of proteome research*, 6(5):1689–1699, 2007.

[14] TM Cafarelli, A Desbuleux, Yang Wang, Soon Gang Choi, David De Ridder, and Marc Vidal. Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Current Opinion in Structural Biology*, 44:201–210, 2017.

[15] Shuiming Cai, Peipei Zhou, and Zengrong Liu. Functional characteristics of a double negative feedback loop mediated by micrornas. *Cognitive neurodynamics*, 7(5):417–429, 2013.

[16] Laurence Calzone, Amélie Gelay, Andrei Zinovyev, François Radvanyi, and Emmanuel Barillot. A comprehensive modular map of molecular interactions in rb/e2f pathway. *Molecular systems biology*, 4(1):0174, 2008.

[17] Walter Bradford Cannon. The wisdom of the body. 1939.

[18] Junha Cha and Insuk Lee. Single-cell network biology for resolving cellular heterogeneity in human diseases. *Experimental & Molecular Medicine*, pages 1–11, 2020.

[19] Nagaraja Chappidi, Giuseppe De Gregorio, and Stefano Ferrari. Replication stress-induced exo1 phosphorylation is mediated by rad53/pph3 and exo1 nuclear localization is controlled by 14-3-3 proteins. *Cell division*, 14(1):1–9, 2019.

[20] Fabrizio Chiti and Christopher M Dobson. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annual review of biochemistry*, 86:27–68, 2017.

[21] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[22] Ariella T Cohain, William T Barrington, Daniel M Jordan, Noam D Beckmann, Carmen A Argmann, Sander M Houten, Alexander W Charney, Raili Ermel, Katyayani Sukhavasi, Oscar Franzen, et al. An integrative multiomic network model links lipid metabolism to glucose regulation in coronary artery disease. *Nature Communications*, 12(1):1–13, 2021.

[23] Michael Costanzo, Benjamin VanderSluis, Elizabeth N Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, Matej Usaj, Julia Hanchard, Susan D Lee, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306), 2016.

[24] Darren Davis, Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Aleksandar Stojmirovic, and Nataša Pržulj. Topology-function conservation in protein–protein interaction networks. *Bioinformatics*, 31(10):1632–1639, 2015.

[25] Géraldine Del Mondo, Damien Eveillard, and Irena Rusu. Homogeneous decomposition of protein interaction networks: refining the description of intra-modular interactions. *Bioinformatics*, 25(7):926–932, 2009.

[26] Christopher M Dobson. Protein misfolding, evolution and disease. *Trends in biochemical sciences*, 24(9):329–332, 1999.

[27] Chao-Yi Dong, Dongkwan Shin, Sunghoon Joo, YoonKey Nam, and Kwang-Hyun Cho. Identification of feedback loops in neural networks based on multi-step granger causality. *Bioinformatics*, 28(16):2146–2153, 2012.

[28] Min Dong, Xiaoyang Su, Boris Dzikovski, Emily E Dando, Xuling Zhu, Jintang Du, Jack H Freed, and Hening Lin. Dph3 is an electron donor for dph1-dph2 in the first step of eukaryotic diphthamide biosynthesis. *Journal of the American Chemical Society*, 136(5):1754–1757, 2014.

[29] Mohammad Farhadian, Seyed Abbas Rafat, Bahman Panahi, and Christopher Mayack. Weighted gene co-expression network analysis identifies modules and functionally enriched pathways in the lactation process. *Scientific Reports*, 11(1):1–15, 2021.

[30] Robert M Friedlander. Apoptosis and caspases in neurodegenerative diseases. *New England Journal of Medicine*, 348(14):1365–1375, 2003.

[31] Rong Gao and Ann M Stock. Overcoming the cost of positive autoregulation by accelerating the response with a coupled negative feedback. *Cell reports*, 24(11):3061–3071, 2018.

[32] Adriana Garay-Arroyo and Alejandra A Covarrubias. Three genes whose expression is induced by stress in saccharomyces cerevisiae. *Yeast*, 15(10A):879–892, 1999.

[33] Audrey P Gasch, Paul T Spellman, Camilla M Kao, Orna Carmel-Harel, Michael B Eisen, Gisela Storz, David Botstein, and Patrick O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12):4241–4257, 2000.

[34] Paul R Graves and Timothy AJ Haystead. Molecular biologist's guide to proteomics. *Microbiology and molecular biology reviews*, 66(1):39–63, 2002.

[35] Roger Guimerà, Marta Sales-Pardo, and Luis A Nunes Amaral. A network-based method for target selection in metabolic networks. *Bioinformatics*, 23(13):1616–1622, 2007.

[36] Steven Hahn and Elton T Young. Transcriptional regulation in saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, 189(3):705–736, 2011.

[37] Jing-Dong Jackie Han. Understanding biological functions through molecular networks. *Cell research*, 18(2):224–237, 2008.

[38] F Ulrich Hartl, Andreas Bracher, and Manajit Hayer-Hartl. Molecular chaperones in protein folding and proteostasis. *Nature*, 475(7356):324–332, 2011.

[39] Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray. From molecular to modular cell biology. *Nature*, 402(6761):C47–C52, 1999.

[40] Linda Hicke. Protein regulation by monoubiquitin. *Nature reviews Molecular cell biology*, 2(3):195–201, 2001.

[41] Tomaž Hočevar and Janez Demšar. A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4):559–565, 2014.

[42] Lei Huang, Fuhai Li, Jianting Sheng, Xiaofeng Xia, Jinwen Ma, Ming Zhan, and Stephen TC Wong. Drugcomboranker: drug combination discovery based on target network analysis. *Bioinformatics*, 30(12):i228–i236, 2014.

[43] Chi Nam Ignatius Pang, Apurv Goel, and Marc R Wilkins. Investigating the network basis of negative genetic interactions in saccharomyces cerevisiae with integrated biological networks and triplet motif analysis. *Journal of proteome research*, 17(3):1014–1030, 2018.

[44] Ronald Jansen, Dov Greenbaum, and Mark Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome research*, 12(1):37–46, 2002.

[45] Jia Jia, Eun Je Jeon, Mei Li, Dylan J Richards, Soojin Lee, Youngmee Jung, Ryan W Barrs, Robert Coyle, Xiaoyang Li, James C Chou, et al. Evolutionarily conserved sequence motif analysis guides development of chemically defined hydrogels for therapeutic vascularization. *Science advances*, 6(28):eaaz5894, 2020.

[46] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[47] Zahra Razaghi Moghadam Kashani, Hayedeh Ahrabian, Elahe Elahi, Abbas Nowzari-Dalini, Elnaz Saberi Ansari, Sahar Asadi, Shahin Mohammadi, Falk Schreiber, and Ali Masoudi-Nejad. Kavosh: a new algorithm for finding network motifs. *BMC bioinformatics*, 10(1):1–12, 2009.

[48] Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences*, 102(39):13773–13778, 2005.

[49] Arminja N Kettenbach, Kate A Schlosser, Scott P Lyons, Isha Nasa, Jiang Gui, Mark E Adamo, and Scott A Gerber. Global assessment of its network dynamics reveals that the kinase plk1 inhibits the phosphatase pp6 to promote aurora a activity. *Science signaling*, 11(530), 2018.

[50] Raya Khanin and Ernst Wit. How scale-free are biological networks. *Journal of computational biology*, 13(3):810–818, 2006.

[51] Vikram Khurana and Susan Lindquist. Modelling neurodegeneration in saccharomyces cerevisiae: why cook with baker's yeast? *Nature Reviews Neuroscience*, 11(6):436–449, 2010.

[52] Wooyoung Kim, Min Li, Jianxin Wang, and Yi Pan. Biological network motif detection and evaluation. *BMC systems biology*, 5(3):1–13, 2011.

[53] Margaret M Kincaid and Antony A Cooper. Misfolded proteins traffic from the endoplasmic reticulum (er) due to er export signals. *Molecular biology of the cell*, 18(2):455–463, 2007.

[54] Regina T Knapp, Michael JH Wong, Lorenz K Kollmannsberger, Nils C Gassen, Anja Kretzschmar, Jürgen Zschocke, Kathrin Hafner, Jason C Young, and Theo Rein. Hsp70 cochaperones hspbp1 and bag-1m differentially regulate steroid hormone receptor function. *PloS one*, 9(1), 2014.

[55] Johannes Kobler, Uwe Schöning, and Jacobo Torán. *The graph isomorphism problem: its structural complexity*. Springer Science & Business Media, 2012.

[56] Kaoru R Komatsu, Toshiki Taya, Sora Matsumoto, Emi Miyashita, Shunnichi Kashida, and Hirohide Saito. Rna structure-wide discovery of functional interactions with multiplexed rna motif library. *Nature communications*, 11(1):1–14, 2020.

[57] Wai Lim Ku, Geet Duggal, Yuan Li, Michelle Girvan, and Edward Ott. Interpreting patterns of gene expression: signatures of coregulation, the data processing inequality, and triplet motifs. *PloS one*, 7(2):e31969, 2012.

[58] Hiroyuki Kurata, Hana El-Samad, Rei Iwasaki, Hisao Ohtake, John C Doyle, Irina Grigorova, Carol A Gross, and Mustafa Khammash. Module-based analysis of robustness tradeoffs in the heat shock response system. *PLoS computational biology*, 2(7):e59, 2006.

[59] Manish P Kurhekar, Sudeshna Adak, Suchit Jhunjhunwala, and Karthik Raghupathy. Genome-wide pathway analysis and visualization using gene expression data. In *Biocomputing 2002*, pages 462–473. World Scientific, 2001.

[60] Zijuan Lai, Hiroshi Tsugawa, Gert Wohlgemuth, Sajjan Mehta, Matthew Mueller, Yuxuan Zheng, Atsushi Ogiwara, John Meissen, Megan Showalter, Kohei Takeuchi, et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature methods*, 15(1):53–56, 2018.

[61] Michelle D Leach, Susan Budge, Louise Walker, Carol Munro, Leah E Cowen, and Alistair JP Brown. Hsp90 orchestrates transcriptional regulation by hsf1 and cell wall remodelling by mapk signalling during thermal adaptation in a pathogenic yeast. *PLoS pathogens*, 8(12), 2012.

[62] Michelle D Leach, Rhys A Farrer, Kaeling Tan, Zhengqiang Miao, Louise A Walker, Christina A Cuomo, Robert T Wheeler, Alistair JP Brown, Koon Ho Wong, and Leah E Cowen. Hsf1 and hsp90 orchestrate temperature-dependent global transcriptional remodelling and chromatin architecture in candida albicans. *Nature communications*, 7(1):1–13, 2016.

[63] Sung Haeng Lee and Roberto Dominguez. Regulation of actin cytoskeleton dynamics in cells. *Molecules and cells*, 29(4):311–325, 2010.

[64] Travis A Lee and Julia Bailey-Serres. Lighting the shadows: methods that expose nuclear and cytoplasmic gene regulatory control. *Current opinion in biotechnology*, 49:29–34, 2018.

[65] Xin Li, Douglas S Stones, Haidong Wang, Hualiang Deng, Xiaoguang Liu, and Gang Wang. Netmode: Network motif detection without nauty. *PloS one*, 7(12):e50093, 2012.

[66] Xiao Liang, William Chad Young, Ling-Hong Hung, Adrian E Raftery, and Ka Yee Yeung. Integration of multiple data sources for gene network inference using genetic perturbation data. *Journal of Computational Biology*, 26(10):1113–1129, 2019.

[67] Dan Lindholm, Hanna Wootz, and Laura Korhonen. Er stress and neurodegenerative diseases. *Cell Death & Differentiation*, 13(3):385–392, 2006.

[68] Miaolong Lu and Xianquan Zhan. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *EPMA Journal*, 9(1):77–102, 2018.

[69] Feng Luo, Yunfeng Yang, Chin-Fu Chen, Roger Chang, Jizhong Zhou, and Richard H Scheuermann. Modular organization of protein interaction networks. *Bioinformatics*, 23(2):207–214, 2007.

[70] Xiaoke Ma and Lin Gao. Biological network analysis: insights into structure and functions. *Briefings in functional genomics*, 11(6):434–442, 2012.

[71] Tarun Mahajan and Roy D Dar. Internetwork connectivity of molecular networks across species of life. *Scientific reports*, 11(1):1–15, 2021.

[72] Taras Makhnevych and Walid A Houry. The control of spindle length by hsp70 and hsp110 molecular chaperones. *FEBS letters*, 587(8):1067–1072, 2013.

[73] Dror Marcus and Yuval Shavitt. Rage–a rapid graphlet enumerator for large networks. *Computer Networks*, 56(2):810–819, 2012.

[74] Martin Lee Miller, Lars Juhl Jensen, Francesca Diella, Claus Jørgensen, Michele Tinti, Lei Li, Marilyn Hsiung, Sirlester A Parker, Jennifer Bordeaux, Thomas Sicheritz-Ponten, et al. Linear motif atlas for phosphorylation-dependent signaling. *Science signaling*, 1(35):ra2–ra2, 2008.

[75] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[76] Z Moradimanesh, R Khosrowabadi, M Eshaghi Gordji, and GR Jafari. Altered structural balance of resting-state networks in autism. *Scientific reports*, 11(1):1–16, 2021.

[77] Abhisek Mukherjee and Claudio Soto. Prion-like protein aggregates and type 2 diabetes. *Cold Spring Harbor perspectives in medicine*, 7(5):a024315, 2017.

[78] Lila E Mullany, Jennifer S Herrick, Roger K Wolff, John R Stevens, Wade Samowitz, and Martha L Slattery. Microrna-transcription factor interactions and their combined effect on target gene expression in colon cancer cases. *Genes, Chromosomes and Cancer*, 57(4):192–202, 2018.

[79] Michael Mülleder, Enrica Calvani, Mohammad Tauqeer Alam, Richard Kangda Wang, Florian Eckerstorfer, Aleksej Zelezniak, and Markus Ralser. Functional metabolomics describes the yeast biosynthetic regulome. *Cell*, 167(2):553–565, 2016.

[80] Andreea Munteanu, James Cotterell, Ricard V Solé, and James Sharpe. Design principles of stripe-forming motifs: the role of positive feedback. *Scientific reports*, 4(1):1–10, 2014.

[81] S Nag, X Zhang, KS Srivenugopal, M-H Wang, W Wang, and R Zhang. Targeting mdm2-p53 interaction for cancer therapy: are we there yet? *Current medicinal chemistry*, 21(5):553–574, 2014.

[82] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[83] Sarah A Nordeen, Daniel L Turman, and Thomas U Schwartz. Yeast nup84-nup133 complex structure details flexibility and reveals conservation of the membrane anchoring alps motif. *Nature communications*, 11(1):1–12, 2020.

[84] Béla Novák and John J Tyson. Design principles of biochemical oscillators. *Nature reviews Molecular cell biology*, 9(12):981–991, 2008.

[85] Hiroyuki Ohashi, Mai Hasegawa, Kentaro Wakimoto, and Etsuko Miyamoto-Sato. Next-generation technologies for multiomics approaches including interactome sequencing. *BioMed research international*, 2015, 2015.

[86] Shane Thomas O'Donnell, R Paul Ross, and Catherine Stanton. The progress of multi-omics technologies: determining function in lactic acid bacteria using a systems level approach. *Frontiers in microbiology*, 10:3084, 2020.

[87] Paola Paci, Giulia Fiscon, Federica Conte, Rui-Sheng Wang, Lorenzo Farina, and Joseph Loscalzo. Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *npj Systems Biology and Applications*, 7(1):1–11, 2021.

[88] Marie Pireyre and Meike Burow. Regulation of myb and bhlh transcription factors: a glance at the protein level. *Molecular Plant*, 8(3):378–388, 2015.

[89] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005.

[90] Paola Pontecorvi, Laura Bernardini, Anna Capalbo, Simona Ceccarelli, Francesca Megiorni, Enrica Vescarelli, Irene Bottillo, Nicoletta Preziosi, Maria Fabbretti, Giorgia Perniola, et al. Protein–protein interaction network analysis applied to dna copy number profiling suggests new perspectives on the aetiology of mayer–rokitansky–küster–hauser syndrome. *Scientific reports*, 11(1):1–11, 2021.

[91] Evan T Powers, Richard I Morimoto, Andrew Dillin, Jeffery W Kelly, and William E Balch. Biological and chemical approaches to diseases of proteostasis deficiency. *Annual review of biochemistry*, 78:959–991, 2009.

[92] Stanley B Prusiner. Development of the prion concept. *COLD SPRING HARBOR MONOGRAPH SERIES*, 38:67–112, 1999.

[93] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007.

[94] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.

[95] Martijn Rep, Markus Proft, Fabienne Remize, Markus Tamás, Ramón Serrano, Johan M Thevelein, and Stefan Hohmann. The saccharomyces cerevisiae sko1p transcription factor mediates hog pathway-dependent osmotic regulation of a set of genes encoding enzymes implicated in protection from oxidative damage. *Molecular microbiology*, 40(5):1067–1083, 2001.

[96] Michael Rosenberg, Roy Blum, Barry Kesner, Eric Aeby, Jean-Michel Garant, Attila Szanto, and Jeannie T Lee. Motif-driven interactions between rna and prc2 are rheostats that regulate transcription elongation. *Nature Structural & Molecular Biology*, 28(1):103–117, 2021.

[97] Nitzan Rosenfeld, Michael B Elowitz, and Uri Alon. Negative autoregulation speeds the response times of transcription networks. *Journal of molecular biology*, 323(5):785–793, 2002.

[98] Ambre Sala, Virginie Calderon, Patricia Bordes, and Pierre Genevaux. Tac from mycobacterium tuberculosis: a paradigm for stress-responsive toxin–antitoxin systems controlled by secb-like chaperones. *Cell Stress and Chaperones*, 18(2):129–135, 2013.

[99] Henning Schwöbbermeyer and Röbbe Wünschiers. Mavisto: a tool for biological network motif analysis. In *Bacterial Molecular Networks*, pages 263–280. Springer, 2012.

[100] Ulrike Seifert, Lukasz P Bialy, Frédéric Ebstein, Dawadschargal Bech-Otschir, Antje Voigt, Friederike Schröter, Timour Prozorovski, Nicole Lange, Janos Steffen, Melanie Rieger, et al. Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress. *Cell*, 142(4):613–624, 2010.

[101] Heather Senn, Rebecca S Shapiro, and Leah E Cowen. Cdc28 provides a molecular link between hsp90, morphogenesis, and cell cycle progression in candida albicans. *Molecular biology of the cell*, 23(2):268–283, 2012.

[102] Mark L Siegal, Daniel EL Promislow, and Aviv Bergman. Functional and evolutionary inference in gene networks: does topology matter? *Genetica*, 129(1):83–103, 2007.

[103] Rakesh Kumar Singh, Melanie Gonzalez, Marie-Helene Miquel Kabbaj, and Akash Gunjan. Novel e3 ubiquitin ligases that regulate histone protein levels in the budding yeast saccharomyces cerevisiae. *PLoS One*, 7(5):e36295, 2012.

[104] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70:263–286, 2017.

[105] Aneta Smidova, Katerina Stankova, Olivia Petrvalska, Josef Lazar, Hana Sychrova, Tomas Obsil, Olga Zimmermannova, and Veronika Obsilova. The activity of saccharomyces cerevisiae na+, k+/h+ antiporter nha1 is negatively regulated by 14-3-3 protein binding at serine 481. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1866(12):118534, 2019.

[106] Kim Sneppen, Sandeep Krishna, and Szabolcs Semsey. Simplified models of biological networks. *Annual review of biophysics*, 39:43–59, 2010.

[107] Min Jeong Sohn, Doo-Byoung Oh, Eun Jung Kim, Seon Ah Cheon, Ohsuk Kwon, Jeong-Yoon Kim, Sang Yup Lee, and Hyun Ah Kang. Hpyps1 and hpyps7 encode functional aspartyl proteases localized at the cell surface in the thermotolerant methylotrophic yeast hansenula polymorpha. *Yeast*, 29(1):1–16, 2012.

[108] Trevor R Sorrells, Lauren N Booth, Brian B Tuch, and Alexander D Johnson. Intersecting transcription networks constrain gene regulatory evolution. *Nature*, 523(7560):361–365, 2015.

[109] Claudio Soto. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nature Reviews Neuroscience*, 4(1):49–60, 2003.

[110] Claudio Soto and Sandra Pritzkow. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nature neuroscience*, 21(10):1332–1340, 2018.

[111] Alok Srivastava, Suraj Kumar, and Ramakrishna Ramaswamy. Two-layer modular analysis of gene and protein networks in breast cancer. *BMC systems biology*, 8(1):1–16, 2014.

[112] Karolina Szczepanowska, Katharina Senft, Juliana Heidler, Marija Herholz, Alexandra Kukat, Michaela Nicole Höhne, Eduard Hofsetz, Christina Becker, Sophie Kaspar, Heiko Giese, et al. A salvage pathway maintains highly functional respiratory complex i. *Nature communications*, 11(1):1–18, 2020.

[113] NETWORKM TIF. mfinder tool guide.

[114] Johnny M Tkach, Askar Yimit, Anna Y Lee, Michael Riffle, Michael Costanzo, Daniel Jaschob, Jason A Hendry, Jiongwen Ou, Jason Moffat, Charles Boone, et al. Dissecting dna damage response pathways by analysing protein localization and abundance changes during dna replication stress. *Nature cell biology*, 14(9):966–976, 2012.

[115] Shailesh Tripathi, Salissou Moutari, Matthias Dehmer, and Frank Emmert-Streib. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC bioinformatics*, 17(1):1–18, 2016.

[116] Jiří Tỳč, Lucie Novotná, Priscilla Peña-Diaz, Dmitri A Maslov, and Julius Lukeš. Rsm22, mtysxc and pnkd-like proteins are required for mitochondrial translation in trypanosoma brucei. *Mitochondrion*, 34:67–74, 2017.

[117] John J Tyson and Béla Novák. Functional motifs in biochemical reaction networks. *Annual review of physical chemistry*, 61:219–240, 2010.

[118] Jolanda Van Leeuwen, Carles Pons, Joseph C Mellor, Takafumi N Yamaguchi, Helena Friesen, John Koschwanez, Mojca Mattiazzi Ušaj, Maria Pechlaner, Mehmet Takar, Matej Ušaj, et al. Exploring genetic suppression interactions on a global scale. *Science*, 354(6312), 2016.

[119] Eric L Van Nostrand, Peter Freese, Gabriel A Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M Blue, Jia-Yu Chen, Neal AL Cody, Daniel Dominguez, et al. A large-scale binding and functional map of human rna-binding proteins. *Nature*, 583(7818):711–719, 2020.

[120] Przemyslaw Waliszewski, Marcin Molski, and Jerzy Konarski. On the holistic approach in cellular and cancer biology: Nonlinearity, complexity, and quasi-determinism of the dynamic cellular network. *Journal of surgical oncology*, 68(2):70–78, 1998.

[121] Lin Wang, Wei Zheng, Hongyu Zhao, and Minghua Deng. Statistical analysis reveals co-expression patterns of many pairs of genes in yeast are jointly regulated by interacting loci. *PLoS Genet*, 9(3):e1003414, 2013.

[122] Paul L Weaver, Chao Sun, and Tien-Hsien Chang. Dbp3p, a putative rna helicase in saccharomyces cerevisiae, is required for efficient pre-rrna processing predominantly at site a3. *Molecular and cellular biology*, 17(3):1354–1365, 1997.

[123] C Wells, SE Brennan, M Keon, and NK Saksena. Prionoid proteins in the pathogenesis of neurodegenerative diseases. front mol neurosci 12: 271, 2019.

[124] Sebastian Wernicke and Florian Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[125] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. *Briefings in bioinformatics*, 13(2):202–215, 2012.

[126] Wang Xi and Michael A Beer. Loop extrusion model predicts ctcf interaction specificity. *bioRxiv*, 2020.

[127] Chao Xu, Jing Jin, Chuanbing Bian, Robert Lam, Ruijun Tian, Ryan Weist, Linya You, Jianyun Nie, Alexey Bochkarev, Wolfram Tempel, et al. Sequence-specific recognition of a pxlpxi/l motif by an ankyrin repeat tumbler lock. *Science signaling*, 5(226):ra39–ra39, 2012.

[128] Kui Yang and Xianlin Han. Lipidomics: techniques, applications, and outcomes related to biomedical sciences. *Trends in biochemical sciences*, 41(11):954–969, 2016.

[129] Le Yang, Runpu Chen, Steve Goodison, and Yijun Sun. An efficient and effective method to identify significantly perturbed subnetworks in cancer. *Nature Computational Science*, 1(1):79–88, 2021.

[130] Alexander Zien, Robert Küffner, Ralf Zimmer, and Thomas Lengauer. Analysis of gene expression data with pathway scores. In *Ismb*, volume 8, pages 407–417, 2000.